THE UNIVERSITY *of* EDINBURGH

Sequencing B Cell Receptor Repertoires in Human Disease: Applications in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome and in Experimental Malaria Infection

Audrey Ryback

A thesis submitted in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy

University of Edinburgh

Institute for Infection and Immunity Research

2023

Declaration

I declare that this thesis contains my own work and any collaborative work has been explicitly stated in the text. This work has not been submitted, as a whole or in part, for any other degree to any other university or educational institution.

Audrey Andrea Ryback (19/08/2023)

Lay Summary

B cells recognise parts belonging to an infectious or harmful agent, so-called antigens, using specialised receptors on their cell surface called B cell receptors. Each B cell produces a unique B cell receptor, and the "B cell receptor repertoire" refers to all of the B cell receptors found in a sampled pool of B cells. When the immune system encounters an antigen, the B cells that have a receptor that can bind to the antigen will make copies of themselves and make their receptors even more specific to their targets by introducing mutations into the receptor. In this thesis, I studied these receptors in different diseases using DNA sequencing of B cells taken from the blood. Studying the BCR repertoires in this way allows us to see at a large scale how the immune system changes in disease.

First I wanted to know whether the BCRs from people with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) showed signs of active or chronic infection, or autoimmunity, compared to healthy controls and people with multiple sclerosis. I found that people with ME/CFS did not show any signs of ongoing infection. Using the same features of the BCR repertoire that have been previously patented as a diagnostic test for ME/CFS did not work as a diagnostic tool in our study. However, one gene segment which has been reported to be more abundant in people with ME/CFS in a published paper, was also increased in ME/CFS patients with mild/moderate disease in our cohort and could mean that these individuals were exposed to the same infectious agent.

In the second chapter, I looked at seven volunteers who were infected with malaria twice and sampled at one timepoint before, one timepoint during infection, and two timepoints after anti-malarial treatment. We found that after the first malaria exposure, volunteers had expansions of particular BCRs and BCRs with more mutations on the day they were diagnosed with malaria. After the second malaria exposure, we did not observe any expansion of B cells and instead the repertoires looked more diverse on the day the volunteers were diagnosed. Interestingly, BCRs that expanded in the first infection did

not return over the course of the second infection.

In the final chapter, I sought to develop a technique that would allow us to identify the specific target of each BCR at a large scale. To do this I attempted to link the two chains that make up the part of the BCR that recognises its antigen. I produced thousands of oil droplets containing single cells and the reaction mixes to pair the specific chains of each BCR. Although I made progress with optimising the protocol, I did not succeed in reliably linking the two chains. I also used 3D printing to create our own device to produce droplets.

Abstract

The human adaptive immune system has the capacity to respond to any potential pathogen, to fine-tune the specificity of this response upon encountering an antigen, and commit the effective B or T cells to immune memory. This specificity relies on selecting antigen-binders from a vastly diverse pool of B cell receptors (BCRs) produced by VDJ gene segment recombination and junctional diversification during B cell development, and affinity maturation upon encounter with a cognate antigen. Adaptive Immune Receptor Repertoire sequencing (AIRRseq) enables us to characterise features of B cell populations by sequencing BCRs. In this thesis AIRRseq was used to investigate properties of the human BCR repertoire in two different disease settings. We also attempted to improve on existing methods for BCR-antigen mapping, which would address a major limitation of current AIRRseq analyses.

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a common chronic illness with unknown aetiology and characterised uniquely by the exacerbation of symptoms following exertion. Chronic infection and autoimmunity have been proposed as two mechanisms that potentially underlie the pathology of ME/CFS. We compared the BCR repertoires of 25 patients with mild-moderate ME, 36 patients with severe ME, 21 healthy controls and 28 patients with Multiple Sclerosis to see if we could find signatures of infection or autoimmune responses. ME patients did not display increased clonality or differential somatic hypermutation compared to healthy controls and patients with Multiple Sclerosis. One of two V genes reported to be differentially used in ME patients in a previous study, was replicated in patients with mild/moderate disease. There were no obvious differences in affinity maturation in the ME cohort, but we observed skewing of the ratio of IgM to IgG BCRs in a majority of ME patients.

The second chapter explores a cohort of seven volunteers undergoing a first and second homologous challenge with *Plasmodium falciparum*. The BCR repertoires of volunteers

infected with malaria displayed clonal expansion and somatic hypermutation of repertoires in a primary challenge but, upon re-challenge, we did not observe any signatures of clonal expansion or recurrence of clones expanded in the first challenge. Twenty-eight days post challenge, volunteers showed a trend towards an enrichment of unmutated IgG B cell receptors in their repertoires and this signature was enhanced in the second infection. This was an unexpected finding that warrants further investigation.

Finally, we attempted optimisation of a protocol to pair native B cell receptor heavy and light chains as expression-ready scFv libraries for phage display at high throughput in a user-friendly microfluidics system. While significant progress was made with improving on existing protocols and developing the method, including making a low-cost alternative to a commercially available droplet generator to generate uniform and stable emulsions at high throughput, the full reactions to pair native heavy and light chains in single cell reactions were not achieved. The work described here provides a basis for future lab members to fully optimise the reactions and will allow the lab to interrogate the antigen specificity of sequenced BCR repertoires in future. Taken together, these three chapters explored the uses and limitations of state-of-the-art BCR repertoire sequencing, and generated and analysed two high-quality BCR repertoire datasets.

in a primary challenge but, upon re-challenge, we did not observe any signatures of clonal expansion or recurrence of clones expanded in the first challenge. Twenty-eight days post challenge, volunteers showed a trend towards an enrichment of unmutated IgG B cell receptors in their repertoires and this signature was enhanced in the second infection. This was an unexpected finding that warrants further investigation.

Finally, we attempted optimisation of a protocol to pair native B cell receptor heavy and light chains as expression-ready scFv libraries for phage display at high throughput in a user-friendly microfluidics system. While significant progress was made with improving on existing protocols and developing the method, including making a low-cost alternative to a commercially available droplet generator to generate uniform and stable emulsions at high throughput, the full reactions to pair native heavy and light chains in single cell reactions were not achieved. The work described here provides a basis for future lab members to fully optimise the reactions and will allow the lab to interrogate the antigen specificity of sequenced BCR repertoires in future. Taken together, these three chapters explore the uses and limitations of state-of-the-art BCR repertoire sequencing, and generated and analysed two high-quality BCR repertoire datasets.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I could not have completed this PhD without the help of many colleagues, mentors, family and friends. First of all I need to acknowledge the members of the Cowan Lab, past and present for their help, creative input and unwavering support. I have been continuously inspired by Catherine Sutherland, who has been my PhD buddy, as I have learned from her diligence, knowledge and tremendous capability with all things bioinformatics and value her patience answering my endless questions. Catherine shared many pieces of code with me, and I based many of my analyses in this thesis on her approaches. But most of all I appreciate our friendship which I am sure will continue as we move on to our next adventures.

Graeme Cowan has been a supportive, kind, and patient supervisor right from the start. Graeme gave me the space and freedom to explore my interests, learn the skills I wanted to build, and actually put in his own time and effort to help me realise my interests – the sign of a truly excellent mentor. I am tremendously grateful for the time I spent in his lab and have been inspired by his humble and thorough approach to research. In a short time, Prajitha Naddukkandy and I have become great friends and it has been a pleasure troubleshooting PCRs together. I would also like to thank Natasha Smith, who helped me learn the ropes when I first started, particularly with the CHMI project, and Ruth Shelton who patiently helped troubleshoot droplet-based overlap extension PCRs.

My thesis committee, Alex Rowe and Chris Ponting have been anchors throughout the process and I owe special thanks to Chris as well as Joshua Dibble for collaborating with us and sharing their samples for the ME/CFS chapter. Matt Roddis collected and shipped the T cell depleted samples for the ME/CFS chapter. I am also grateful to Phil Spence and his lab for sharing their precious samples and data and particularly to Phil for providing very helpful input with interpreting the results. I also have to thank the Wellcome Trust for their generous funding.

I would like to thank Prof Richard Milne who all those years ago suggested that I might consider doing a PhD and pointed out the HPGH PhD program to me. Without his encouragement, guidance and the hours of interview prep, I would certainly not have gotten here.

My friends in Edinburgh have been a constant source of joy and support. Sanjana Ravindran especially is an absolute anchor and I have so much admiration for her generous, strong, bold and viciously smart brain. And of course, Kynan and Casey, Anna and Georgia who Sanjana introduced me to and now count among my close friends here – I love going spontaneously to the theatre/comedy/cinema or just the pub and being surrounded by such kind and fun souls. Other friends and colleagues in Ashworth have made the PhD experience unforgettable- from the walks around Blackford Hill with Samer Halabi, coffee with the other members of the HPGH cohort, and everyone who keeps the happy hours and annual pantomime traditions alive.. There are too many names to write here, but I feel incredibly lucky to be a part of this vibrant community.

I am extremely grateful to my family who have worked hard to provide me with the opportunities that helped me reach this point and have always encouraged and supported my interests.

Finally I have to acknowledge the contributions of Charlie Hillier. Charlie is a source of endless support, both academic and personal and is someone who lives and breathes research with total ease. I would not have pursued a BSc, let alone a PhD if it weren't for all his help tutoring me with so much patience and enthusiasm through a significant degree change many years ago that was very daunting for me at the time. I aspire to be as positive and capable a scientist as he is.

Table of Acronyms and Abbreviations

Acronym	Full Word
5' RACE	5' Rapid Amplification of cDNA Ends
AID	Activation Induced Cytidine Deaminase
AMA1	Apical membrane antigen 1
AIRRseq	Adaptive Immune Receptor Repertoire Sequencing
atMBC	atypical memory B cells
BCR	B cell Receptor
cDNA	copy DNA
CDR	Complementarity Determining Region
CCC	Canadian Consensus Criteria
CHMI	Controlled Human Malaria Infections
CMV	Cytomegalovirus
CPET	Cardiopulmonary Exercise Testing
cMBCs	classical memory B cells
CTSD	cathepsin D
DALY	Disability Adjusted Life Years
DN2	Double Negative 2 B cells
DNA	Deoxyribonucleic Acid
EBV	Epstein-Barr Virus
Fc	Fragment Crystallisable
Fcγ RIIB	Fc receptor IIB for IgG
Fab	Fragment Antigen-Binding
GC	Germinal Centre
GPCRs	G-protein Coupled Receptors
HHV6	Human Herpesvirus 6

Continued on next page

Table 1 – continued from previous page

Acronym	Full Word
HLA	Human Leukocyte Antigen
IgD/M/G/A/E	Immunoglobulin D/M/G/A/E
IgH	Immunoglobulin Heavy Chain
IgK	Immunoglobulin Kappa Chain
IgL	Immunoglobulin Lambda Chain
IFN	Interferon
IMGT	International ImMunoGeneTics Information System
IOM	Institute of Medicine
LC	Long Covid
MANATEE	MANifold And Tray for Easy Emulsions
MMLV-RT	murine moloney leukaemia virus RTase
MSP1	Merozoite surface protein-1
MS	Multiple Sclerosis
NEB	New England Biolabs
NLRP6	NOD-like receptor family pyrin domain containing 6
N-Nucleotides	non-templated nucleotides
OE	Overlap-Extension PCR
OE RT-PCR	Overlap-Extension RT-PCR
PCR	Polymerase Chain Reaction
PAMP	pathogen-associated molecular pattern molecules
PEM	Post-exertional malaise
PfEMP1	Plasmodium falciparum erythrocyte membrane protein 1
PfSPZ	Plasmodium Falciparum Sporozoite
P-Nucleotides	palindromic nucleotides
RA	Rheumatoid Arthritis

Continued on next page

Table 1 – continued from previous page

Acronym	Full Word
RAG1/2	Recombination activating genes 1/2
ROC	Receiver operating characteristic
RT-PCR	Reverse transcription polymerase chain reaction
RSS	Recombination Signal Sequences
SBL	Systems Biology Laboratory
SLE	Systemic Lupus Erythematosus
Tbet	T-box expressed in T cells
TCR	T-Cell Receptor
TLR	Toll-Like Receptor
TdT	Terminal deoxynucleotidyl transferase
UMIs	Unique Molecular Identifiers
VH	Variable Region heavy chain
VK	Variable Region kappa chain
VL	Variable Region lambda chain

Chapter 1

General Introduction

1.1 Overview

B CELLS are one of two classes of lymphocytes that make up the adaptive immune system. Both B cells and T cells have homologous mechanisms for generating extremely diverse antigen and epitope-binding receptors by somatic recombination and junctional diversification. Adaptive immune receptor repertoire sequencing is a high-throughput methodology which allows us to observe and quantify dynamics of the adaptive immune system at a systems-wide scale. These new methodologies now allow us to capture thousands to millions of data-points about an individual's adaptive immune system. However, as this field has only emerged in recent years, experimental protocols and bioinformatic tools and analyses have not yet been systematically benchmarked and standardised, so care must be taken to consider the limitations of the technology and conclusions drawn from the data. Finally, deciphering antigen-antibody interactions still

represents a significant gap in our ability to translate and functionally interrogate adaptive immune signatures. This thesis explores the application of adaptive immune receptor repertoire sequencing to two disease settings, Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and controlled human malaria infection, and the development of a low-cost, high-throughput experimental method for deciphering antigen specificity of B cell receptors. Seeing projects through from experimental work to analysis has the added benefit of highlighting potential sources of technical bias in the data that we have attempted to address or acknowledge. The three chapters of this thesis apply and/or build on BCR repertoire sequencing methods to address the following questions:

1. Is there a B cell receptor repertoire signature in ME/CFS?
2. How does the B cell receptor repertoire respond to repeated *P. falciparum* infection in humans?
3. How can we leverage an existing microfluidics platform for high-throughput phage display of natively paired B cell receptors?

1.2 B cell Receptor Repertoires: An Overview of their Development and Function

1.2.1 A brief history

The study of antibodies and B cells has captured the fascination of scientists since the late 1800s and has arguably led to some of the most important advances in translational biomedical research. Vaccination and antibody-based therapies have opened unprecedented opportunities to prevent and treat disease, most recently illustrated with the expedited development of vaccines against COVID-19, which are estimated to have more than

halved the potential death toll linked to COVID-19 in the first year the vaccine was implemented (O. J. Watson et al. 2022). Immunoglobulin therapies have been used to treat patients with immunodeficiencies and cancer. Monoclonal antibodies have enabled the development of new laboratory technologies as well as precision medicine, and are currently the best selling class of drugs in the pharmaceutical industry (R.-M. Lu et al. 2020; Cyster and Allen 2019; Alfaleh et al. 2020).

These advances in translational research were made possible by the body of knowledge produced by many immunologists over the last century and a half. While early microscopist Walther Flemming observed foci of large lymphoid cells dividing in follicles in lymph nodes and coined the term "germinal centers" in 1885 (Nieuwenhuis and Opstelten 1984) it was not until the 1960s that B cells were identified as a distinct immune cell lineage to T cells. Max Cooper and colleagues discovered B cells in chickens originated in the Bursa of Fabricius, confirming their identity as a separate lineage to T cells (Cooper, R. D. Peterson, and R. A. Good 1965; Cooper 2015). Attempting to identify the equivalent origins of B cells in mammals, they later discovered that in the fetal liver B cells develop from hematopoietic stem cells in mice. Contemporaries, Gustav Nossal and Pierre Vassalli made similar discoveries in bone marrow. It was subsequently proven that in humans B cells originate from hematopoietic stem cells in the bone marrow and in the fetal liver (Gitlin and Nussenzweig 2015; Cooper 2015). However, the concept of antibody mediated immunity was developed long before the discovery of B cells. The notion of humoral immunity was first pioneered in the early 1900s by Paul Ehrlich, Shibasaburo Kitasato and Emil von Behring, all three of whom had worked with Robert Koch (Cooper 2015; Kaufmann 2017). In 1890 von Behring and Kitasato published their findings from the first passive immunisation studies using serum from animals that had been exposed to tetanus or diphtheria. They demonstrated that protection to these toxins could be transferred to infection-naive animals (Kaufmann 2017; Valent et al. 2016). This novel treatment

was soon trialled for treating diphtheria in humans. By 1895, serum therapy was used widely in German cities, leading to a reduction in mortality from diphtheria of more than 50 % (Kaufmann 2017). Paul Ehrlich coined the term "antibody" (Antikoerper) and also proposed an early theory of clonal selection, the "side-chain theory". This theory proposed that cells may produce receptors that match toxins, and when a toxin binds to the cell, the cell sheds these receptors and secretes some of the antitoxin-specific receptors into the serum (Ehrlich 1900; Valent et al. 2016). Later, Niels Jerne expanded on Ehrlich's side chain theory with an evolutionary perspective to suggest that antibody diversity precedes antigen exposure, that at random antibodies with the right specificity will encounter an antigen, and that this antigen-specific response is expanded and maintained through a process of selection (Jerne 1955). David Talmage and Frank Macfarlane Burnet built on this theory. In 1957 Burnet published a paper entitled "A modification of Jerne's theory of antibody production using the concept of clonal selection" (Burnet et al. 1957). At the time, ideas of somatic mutation were emerging from the study of cancer. Burnet's theory of clonal selection was also influenced by studies on selection and outgrowth of microbes and viruses from bacteriology and virology (Bernard 2016). Burnet proposed that somatic mutations could be introduced during selection to increase the affinity of antibodies. The tenets of clonal selection theory were based on the theory that antigen-specific immune responses arise from a pre-existing repertoire of diverse lymphocytes each expressing a unique receptor. Upon encountering an antigen, only the cells with a receptor specific to that antigen will become activated, multiply, and produce antibody. This theory was supported by early experimental evidence from Gustav Nossal and Joshua Lederberg in 1958, who immunised rats with two strains of *Salmonella* and subsequently isolated plasma cells from the rat's lymph nodes. They encapsulated the cells in single cell droplets in oil on microscope cover slips and stimulated them with antigens of the two strains. In a single cell droplet only one of the two strains was ever neutralised, demonstrating that a

single B cell only produces one specific antibody (Nossal and Lederberg 1958). Burnet's theory emphasised how T and B cell populations as a whole are influenced by the presence of antigens, which remains a key focus of many systems immunology studies in the present day, and a question many studies attempt to address using adaptive immune receptor repertoire sequencing.

An important missing piece of the puzzle of clonal selection theory was how this receptor diversity is generated in the first instance. The answer was first discovered by Susumu Tonegawa. At the time, the dominant theory was that each unique BCR was encoded by a separate gene. In 1976 Hozumi and Tonegawa published a paper in which they described the distances of hybridised RNA probes against variable and constant kappa chain genes in digested DNA from Balb/c mice early embryos and in a plasma cell tumor line which produces kappa chain (Hozumi and Tonegawa 1976). They realised that variable kappa and constant genes that were located at a substantial distance from each other in the embryo cells, were joined to form one contiguous segment in the DNA from the plasmacytes. Thus, the gene segments had been moved during lymphocyte differentiation. This was the first evidence of somatic gene rearrangement (Jung and Alt 2004). The same mechanism was later also found to generate T cell receptor diversity. This unique mechanism of generating receptor diversity is the only known instance of controlled somatic gene rearrangement in vertebrates. Despite these tremendous advances in our understanding of adaptive immunity over the last century and a half, we are still unravelling the complexities of clonal selection, central tolerance and immune memory in health and disease.

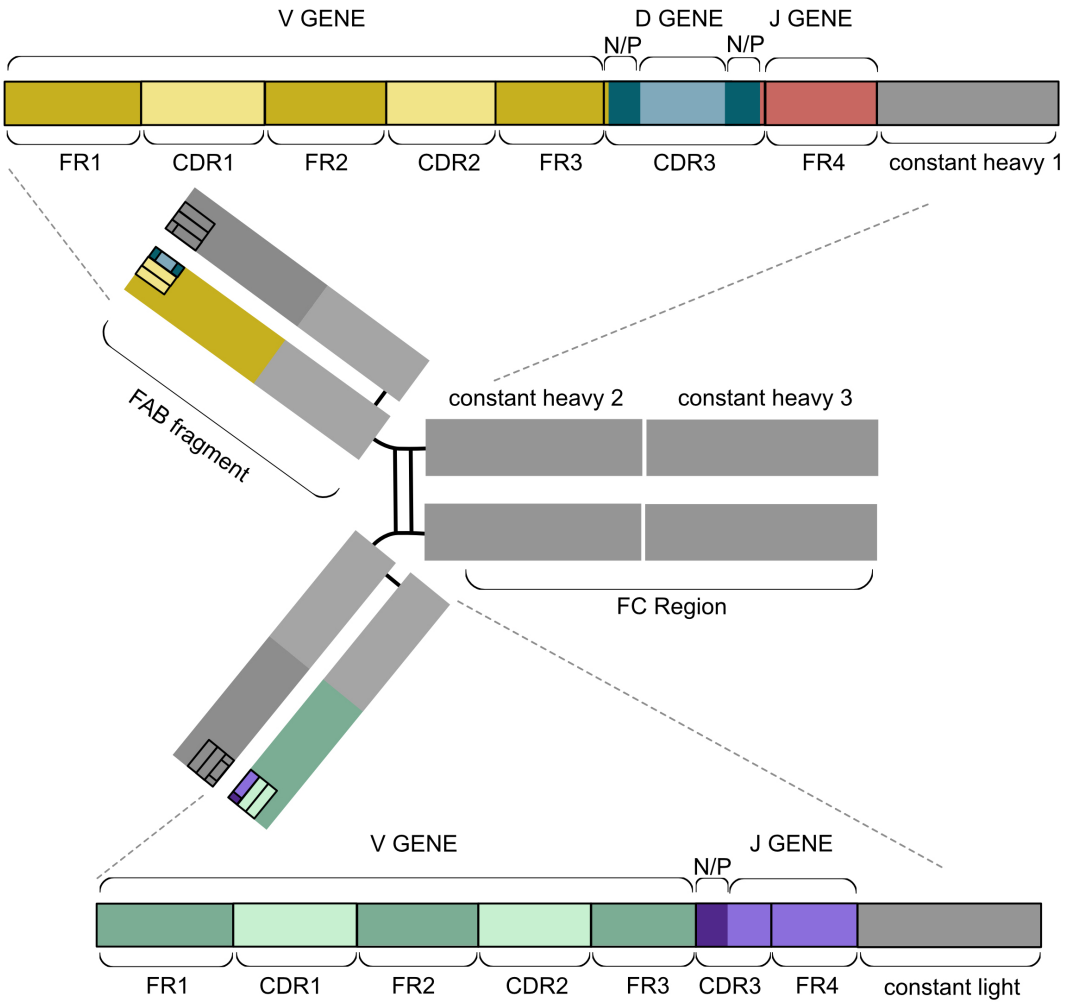


Figure 1: Antibody Structure. Abbreviations: FR = Framework Regions, CDR = Complementarity Determining Regions, N/P : Non-Templated/Palindromic Sequences. FAB: Fragment Antigen-Binding, FC = Fragment Crystallisable.

1.2.2 What is the BCR repertoire and how does it develop?

The Structure of the BCR

The antigen specificity of a B cell is determined by its B cell receptor. The B cell receptor is a globular protein made up of four polypeptides, two identical heavy chains and two identical light chains (Janeway et al. 2001; Chi, Yue Li, and Qiu 2020). Each light

chain is linked to a heavy chain by a disulphide bond, and the constant domains of two heavy chains are also connected by disulphide bonds. Antibodies can be separated into the antigen-binding fragment (Fab) containing the variable regions of the BCR that are antigen-specific, and the fragment crystallisable (Fc) domain of the antibody that can activate complement (C1q), or bind to Fc-receptors on phagocytes or cytotoxic cells (Vidarsson, Dekkers, and Rispens 2014). There are five main antibody isotypes (IgD, IgM, IgG, IgA and IgE) and BCRs can be expressed in a membrane bound form or be secreted with their conformations differing among isotypes. IgG and IgA can also be further divided into subclasses (IgG1-4, IgA1 and IgA2). IgG and IgE are secreted as monomers. The Joining chain, J-chain, promotes polymerisation of monomeric IgM or IgA to form multimeric secreted antibodies. IgM is secreted as a pentamer, and IgA can be secreted as a dimer (Janeway et al. 2001).

The different isotypes and subclasses can be associated with particular effector functions or antigen classes (Janeway et al. 2001; Vidarsson, Dekkers, and Rispens 2014). For example, IgG3 is a strong inducer of effector functions and tends to be short-lived and pro-inflammatory; IgG2 is the main subclass generated in response to polysaccharide antigens (Vidarsson, Dekkers, and Rispens 2014).

The variable region of the antibody is the primary site mediating antigen binding. On the heavy chain it consists of the V, D and J gene segments and the heavy constant region 1, while the light chain variable region consists of the V and J gene and the light constant gene segment (**Figure 1**)(Schroeder and Cavacini 2010). The Fab fragment can also be described in terms of which domains mediate antigen binding when the antibody is folded. The domains important for antigen binding are called "complementarity determining regions" (CDRs). Three CDRs on the heavy chain and three CDRs on the light chain create the unique antigen binding site (Schroeder and Cavacini 2010). The most diverse of these domains is the CDR3, since it spans the junction of the somatically recombined VDJ

or VJ gene segments. Additionally untemplated N/P (non-templated and palindromic) nucleotides are incorporated at the junction between the segments to generate additional diversity (Chi, Yue Li, and Qiu 2020; Janeway et al. 2001).

In the folded antibody, each Fab fragment consists of four beta sheets: Two formed of the constant domains of the light chain and the constant 1 domain of the heavy chain, and two consisting of the FR1-FR4 domains that, when folded, results in the CDR3s being located at the N-terminus of the "Y-shaped" antibody (Schroeder and Cavacini 2010). The combination of CDRs ultimately determine antigen specificity, although the heavy chain CDR3 is thought to be the most important determinant of antibody specificity.

Early B cell development

B cells develop from hematopoietic stem cells in the bone marrow in adults, or in the fetal liver (Y. Wang et al. 2020). B cells differentiate from common lymphoid progenitor (CLP) cells in these tissues. B cells developed in the liver during embryonic development mainly give rise to B1 cells which populate the spleen, intestine, the peritoneal cavity and pleural cavities. B1 cells produce natural antibody that is reported to have a more restricted repertoire and can be generated prior to antigenic exposure (Panda and Ding 2015). The majority of B cells developed in bone marrow are "B-2" cells and have an extremely diverse repertoire of BCRs (Y. Wang et al. 2020). The key determinant of the antigen specificity of a B cell is its B cell Receptor (BCR). The diversity of the BCR repertoire depends on the rearrangement of gene segments during B cell development. The process is tightly regulated and corresponds to specific developmental stages of B cells. During B cell differentiation in the bone marrow, the heavy chain locus is rearranged first, with D-J recombination occurring in CLPs and pre-pro B cells. Next, the rearranged DJ segment is recombined with a V gene segment to produce a functional heavy chain in pre-B cells at which point surrogate light chains and the IgH are expressed on the cell

surface as a pre-BCR (Y. Wang et al. 2020).

If the rearrangement is successful, expression of the pre-BCR acts as a signal to stop heavy-chain rearrangement and initiate several rounds of cell division followed by V-J rearrangement of light chain. If the BCR is out of frame or contains stop codons and does not result in a productive BCR, the B cell attempts to rearrange the locus on the other chromosome. If both rearrangements are unsuccessful, the B cell dies by apoptosis (Y. Wang et al. 2020; Janeway et al. 2001). Functional BCRs are initially expressed as IgM on the surface of immature naive B cells. There are several checkpoints to ensure quality of BCRs. If the BCR rearrangement is successful, there is only one heavy chain and it does not bind self-antigen; B cells undergo positive selection. This halts VDJ recombination and the naive B cell migrates to secondary lymphoid tissues (Chi, Yue Li, and Qiu 2020; Y. Wang et al. 2020). Unligated and successfully expressed BCRs are thought to transmit tonic signals that promote positive selection and subsequent B cell development and maturation (Kuo and Schlissel 2009; Nemazee 2017). If the heavy and light chain combination results in low expression, the B cell has two different heavy chains or if the BCR is activated by self antigen in the bone marrow, it can either be deleted or undergo receptor editing. Around 55%-75% of BCRs in early immature B cells and approximately 20% of mature naive B cells are reactive against self antigen (R. J. Bashford-Rogers, K. G. Smith, and David C. Thomas 2018). Receptor editing describes the process of continuing somatic gene rearrangement to rescue a BCR with low expression or self-reactivity. Receptor editing occurs in up to 20% of B cells and is thought to increase the diversity of the BCR repertoire (Nemazee 2017). Usually this process involves switching light chains, but on the heavy chain, an upstream V gene can also sometimes be exchanged with the original V to alter the BCR specificity (R. J. Bashford-Rogers, K. G. Smith, and David C. Thomas 2018).

Receptor editing is directional and will progress to downstream JK elements on the

kappa locus before attempting rearrangement on the lambda light chain locus. Lambda light chains are thought to be particularly effective in rescuing self-reactivity (A. M. Collins and C. T. Watson 2018). If the B cell remains self-reactive after editing it will apoptose (Nemazee 2017). Receptor editing can occur on both light chain loci which can lead to the expression of two light chains in the same B cell (Nemazee 2017).

B cell activation and maturation

Once naive IgM+ B cells have developed in the bone marrow, they migrate as transitional B cells to the spleen, lymph nodes, and other secondary lymphoid tissues where they complete their development and become dual-expressing IgM/IgD naive B cells.

These B cells patrol the secondary lymphoid tissues for antigen. In lymph nodes, specialised macrophages in the subcapsular region sample antigen from the lymph and present them to B cells. Follicular dendritic cells capture and display antibody and /or complement coated antigens. In addition to the BCR binding a cognate antigen, naive B cells generally require additional signalling to become fully activated (Young and Brink 2021). These secondary signals include T cell help via CD40 engagement, IL-4 and other cytokine signalling, Toll like receptors and other pattern recognition receptor signalling, or multivalent antigen binding. Upon B cell activation, additional diversity is introduced into the BCR by a process called somatic hypermutation (SHM). SHM is best understood in the context of germinal center reactions in T-dependent B cell activation.

In T-dependent responses a B cell encounters a cognate antigen, the BCR binds to the antigen and internalizes the BCR-antigen complex. The antigen is then processed and presented on the surface of the B cell on MHC class II. The B cell then migrates to the border of the T cell zone in the lymph node (Gatto and Brink 2010). If the B cell encounters a T helper cell that recognises the same antigen, signalling via CD40 and cytokines induces proliferation of the B cell. These B cells can then either differentiate

into short-lived plasma cells that produce low-affinity antibodies, or they can form a Germinal Centre with T-Follicular Helper cells. In the Germinal Centre, a B cell will undergo rounds of somatic hypermutation– the introduction of point mutations into the BCR – and selection for the BCRs with the highest-affinity for their antigen.

Germinal Centres (GCs) are structures which form between four and eight days after immune challenge within B cell follicles of secondary lymphoid tissues and are populated by an oligoclonal population of B cells, mainly consisting of the daughter cells of the B cells that seed the GC (Tas et al. 2016).

Germinal Centres are divided spatially into light and dark zones. The dark zone is mostly populated by lymphocytes and is where B cells proliferate and undergo somatic hypermutation. In the light zone lymphocytes, stromal cells and follicular dendritic cells mediate affinity selection. B cells that bind antigen with high affinity are given cytokine and direct survival signals by follicular dendritic and T follicular helper cells. B cells traffic between these two zones over rounds of affinity maturation (Young and Brink 2021; Gatto and Brink 2010). B cells also undergo class switch recombination during this process and a majority of B cells exiting germinal centres are class-switched from IgM to IgG, IgA or IgE. After several rounds of affinity selection, GC B cells differentiate into long-lived plasma cells or memory B cells.

Not all B cells undergo T-dependent activation. If signalling via antigen binding is very strong or if TLRs or other co-receptors are also engaged, B cells will typically give rise to short-lived unmutated plasma cells. Both unswitched and class switched memory B cells which also undergo SHM can be produced in extrafollicular B cell responses, however, the mechanisms of antigen-selection in extra follicular responses have not been studied extensively (Elsner and Shlomchik 2020).

1.2.3 Sources of Diversity in the BCR repertoire

Burnet's clonal selection theory acknowledged the need for a sophisticated diversification process to achieve a diverse repertoire against any potential pathogen:

"The theory requires at some stage in early embryonic development a genetic process for which there is no available precedent (Burnet 1957)."

Theoretically, the BCR repertoire is estimated to be able to generate at least 6000 possible permutations of V (variable), D (diversity) and J (joining) gene segments for the heavy chain and 320 potential combinations of V-J light gene segments (Nemazee 2017). This is much less diverse than the actual estimated diversity of the naive BCR repertoire: 10^{15} unique receptors(Briney et al. 2019). Additional diversity is introduced during maturation of the B cell response by somatic hypermutation. There are two sources of diversity during the process VDJ recombination: Combinatorial and junctional diversity.

Combinatorial diversity

Combinatorial diversity in the BCR repertoire is generated by somatic rearrangement of V,D, and J gene segments in the heavy chain and V-J gene segments in the light chain. There are approximately 38-46 functional IGHV genes belonging to 7 V gene families, 23 IGHD and 6 IGHJ genes located on chromosome 14 (Lefranc et al. 2005). In the kappa light chain locus there are 31-36 functional IGKV genes that form 5 V gene families and 5 IGKJ segments, located on chromosome 2. The lambda light chain locus, found on chromosome 22, contains 29 to 33 functional IGLV genes belonging to 10 V gene families and 4-5 IGLJ genes (Lefranc et al. 2005). Each individual will have two VDJ haplotypes, one inherited from each parent with different alleles and copy numbers of V,D and J genes.

An example of VDJ rearrangement is shown in **Figure 2** (simplified, for illustrative purposes). Somatic recombination depends on the Recombination-Activating Gene recom-

binase enzyme which consists of two proteins, RAG1 and RAG2 that produce double-strand breaks at specific sites.

These sites are called recombination signal sequences (RSS) and flank the V, D and J genes (**Figure 2A**). They have a highly conserved seven-base motif ($5'-CACAGTG-3'$) and a conserved nine-base motif ($5'-ACAAAAACC-3'$) separated by a spacer sequence of 12 or 23 nucleotides (Chi, Yue Li, and Qiu 2020). The length of the spacer sequence is important in ensuring the correct order of VDJ recombination. Both the 3' end of V fragment and the 5' end of J fragment have a 23 base spacer sequence and the D fragment contains a 12 base spacer. Efficient recombination is only thought to occur between RSS with different spacer lengths, the so-called "12/23 rule" which ensures the heavy chain J gene will recombine with a D gene but not a V **(Figure 2B-C)**. The DNA is nicked and cleaved by RAG at the junction between the coding segment and the RSS and a covalently linked hairpin is formed at each blunt end between the 3'OH group on the forward strand and the free 5' phosphate group on the reverse strand (**Figure 2D**). This results in two joints, the coding joint and the RSS joint containing the excised fragment of DNA (Chi, Yue Li, and Qiu 2020; Y. Wang et al. 2020).

Junctional Diversity

Additional diversity is introduced by nucleotide deletion and addition at the junction between the D and J gene, and V and D segments. After RAG cleaves the two gene segments, they are re-joined by nonhomologous DNA end joining. During the joining of the two coding ends, junctional diversity is generated by the addition of "N" (non-templated) and "P" (palindromic) nucleotides. The classical nonhomologous DNA end-joining (C-NHEJ) factors, including Artemis and DNA-PKcs, stabilise the two coding joints and open the hairpins at the ends of the joints. Nucleotides from the linearised hairpin can then be incorporated into the junction, introducing short Palindromic ("P")

sequences. Additionally, terminal deoxynucleotidyl transferase (TdT) can add up to twenty untemplated nucleotides ("N" nucleotides) to the junction **(Figure 2E)**. Nucleotides in the junction can also be trimmed by exonucleases and in some cases the D gene cannot be identified because so little of the D sequence remains (Chi, Yue Li, and Qiu 2020). Junctional diversification occurs at the ends of both cleaved segments (**Figure 2F**).

Finally, DNA ligation and repair enzymes then pair the two strands imperfectly, unpaired nucleotides are removed by exonucleases, and DNA polymerases fill the gaps on the complementary strand (**Figure 2G**). These mechanisms result in a tremendous theoretical diversity of the CDR3 (**Figure 2H**). However, junctional diversification can also frequently produce non-productive BCRs due to frame shifts introduced during N/P addition and nucleotide deletion (Chi, Yue Li, and Qiu 2020).

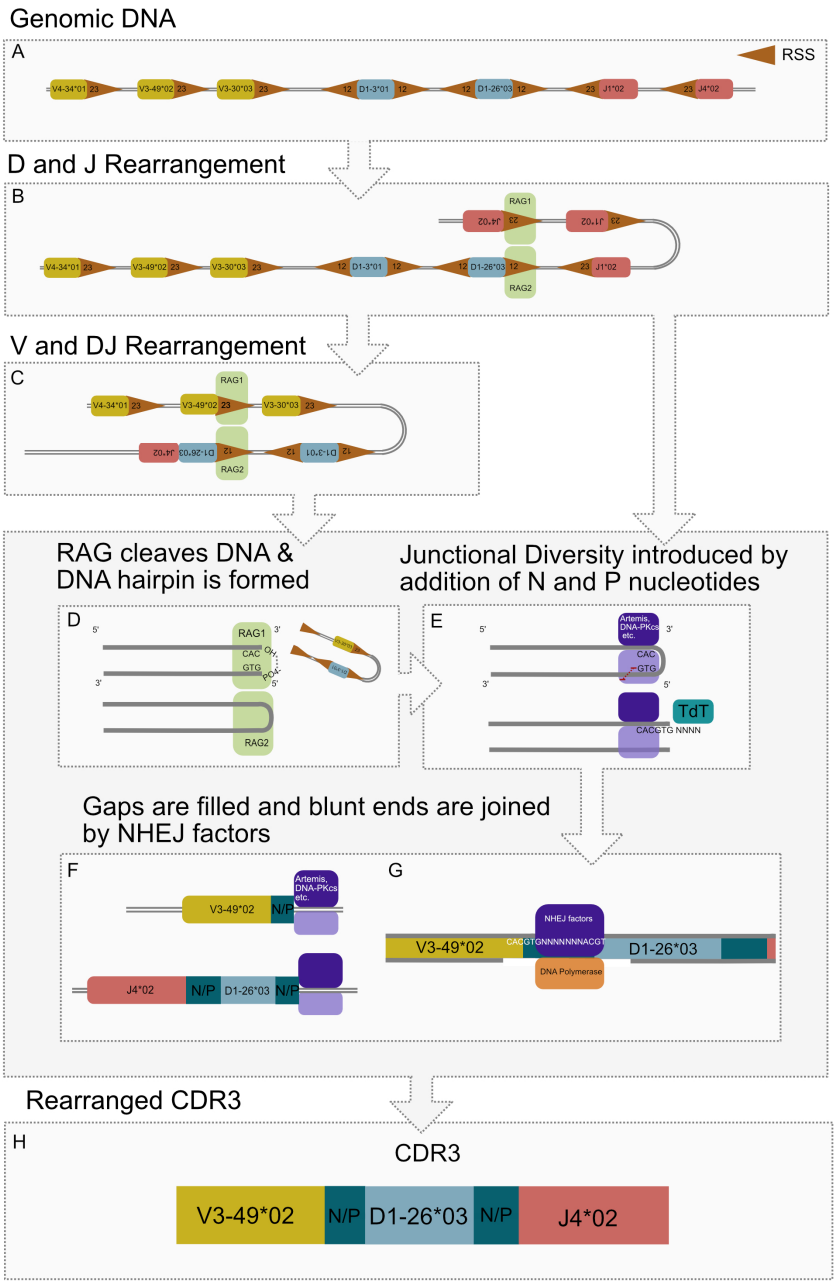


Figure 2: VDJ Recombination and Junctional Diversification. A) In unrearranged genomic DNA V,D, and J gene segments are flanked by RSS sequences. **B)** First the D and J genes are combined by RAG, following the 12/23 rule. **C)** Then a V gene is recombined with the DJ segment. **D)** RAG cleaves the two gene segments and the joint is closed by a DNA hairpin. **E)** Non-homologous end joining factors (NHEJ), including Artemis and DNA-PKcs stabilise the two coding joints and open the hairpins at the ends of the joints. Nucleotides from the complementary strand are incorporated into the junction (P), and TdT adds untemplated nucleotides (N), exonucleases can also trim bases. **F)** This process occurs on both ends of the junction. **G)** NHEJ and DNA polymerases fill the gaps and join the ends. **H)** The rearranged heavy chain CDR3 contains a V,D and J gene with N and P nucleotides in the junction.

Somatic Hypermutation

Somatic hypermutation (SHM)describes the process whereby point mutations are introduced into the variable regions of the BCR during affinity maturation, and the highest affinity antigen-binders are given survival and proliferation signals. Activation Induced Cytidine Deaminase (AID) is the key enzyme which mediates somatic hypermutation. Due to its mutagenic potential its expression is tightly controlled (Chi, Yue Li, and Qiu 2020). AID activity is induced by co-stimulation of B cells via CD40L and by engagement of TLRs.

AID is a member of the APOBEC family and converts Cytosine bases to Uracil by removing the amine group on Cytosine (Maul and Gearhart 2010). Although SHM only successfully mutates about 3% of the bases it targets, it still leads to 10^{-3} to 10^{-4} mutations per base (Gatto and Brink 2010). The inserted Uracils then trigger base excision repair pathways and Uracil DNA Glycosylase removes the base. Error-prone DNA polymerases then fill the gap but can also introduce mutations at neighboring bases (Maul and Gearhart 2010). Affinity maturation selects for nonsynonymous mutations in the CDRs as these are more likely to alter the strength of antigen-antibody binding (Yaari, Benichou, et al. 2015).

Somatic hypermutation does not occur uniformly across the sequence and there are known "hotspots" of SHM: Motifs containing WRC/GYW (W = A/T, R = A/G, Y = C/T) and when there are WGCW hotspots appearing opposite each other on both strands (Tang et al. 2020).

AID is also involved in class-switch recombination - the switching of constant regions. It does this by introducing mutations into the switch region preceding the different constant region genes (Maul and Gearhart 2010). C regions can be switched multiple times; however, only in the direction of the locus order (IgM, IgD, IgG3, IgG1, IgA1, IgG2, IgG4, IgE, and IgA2). For example IgM can switch to IgG1 and this B cell can class

switch to IgA1, but a BCR with an IgE constant region cannot switch to IgM because that gene will already have been excised (Chi, Yue Li, and Qiu 2020).

1.3 Adaptive Immune Receptor Repertoire Sequencing

Adaptive immune receptor repertoire sequencing (AIRRseq) uses high throughput sequencing to study the adaptive immune system at scale. At the core of AIRRseq is an attempt to determine how the adaptive immune repertoire is shaped by disease, whether infectious or non-communicable. It allows us to explore how B cell repertoires change in disease and, depending on the analysis conducted, it can sometimes indicate likely stages of B cell development at which these changes are occurring. The potential translational applications are many, but include the following: the discovery of monoclonal or broadly neutralising antibodies (B. Wang et al. 2018; Setliff, McDonnell, et al. 2018); the comparing and understanding of protective or ineffective vaccine responses (Avnir et al. 2016; Jacob D Galson, Clutterbuck, et al. 2015); and understanding how defects in central tolerance, clonal selection and germline susceptibility to auto-antigens shapes the BCR repertoire in autoimmunity (reviewed in R. J. Bashford-Rogers, K. G. Smith, and David C. Thomas 2018).

Despite the many potential applications, BCR repertoires are complex and the inability for us to comprehensively map antigen specificity back onto repertoire data still poses challenges to interpreting findings. The following section reviews the current best practices regarding wet-lab methods in repertoire sequencing, typical analyses, and their applications in infectious and non-communicable diseases.

1.3.1 Methods

Many different methodologies and analytical approaches have been used in AIRRseq. The field would greatly benefit from standardisation of experimental and analytical pipelines, which are gradually being established by the AIRR Community (Trück et al. 2021). BCR repertoires are often sequenced from mRNA encoding the BCRs, extracted from PBMCs or sorted B cells in humans. The use of mRNA as a template is mainly due to the fact that individual transcripts can be labelled using Unique Molecular Identifiers (UMIs) and subsequent PCR bias and errors can be corrected (Yaari and Kleinstein 2015). Amplification from genomic DNA has the advantage that Ig gene copy number is consistent between cells, but it is more challenging to label gDNA with UMIs and correct for technical biases. Furthermore, when using mRNA as a template the constant region can also be identified, providing useful information about the likely maturity of the response (**Figure 3A**). There are a number of technical challenges regarding PCR amplification and sequencing of BCRs due to the complexity of the rearranged BCR but improvements in the technology in recent years have partially addressed some of these issues.

5' RACE cDNA amplification

Due to combinatorial diversity, originally large cocktails of primers were used to amplify BCRs with all potential V genes. Different primers can be more or less efficient and it is difficult to match TMs across a wide set of primers. Using gene specific primers can introduce significant amplification bias. Therefore, using gene specific primers hampers estimation of clonality or V gene usage and only covers V genes already deposited in the reference databases.

The introduction of 5' RACE technology (Rapid Amplification of cDNA Ends) has overcome this substantial limitation since it permits unbiased amplification of V genes.

It makes use of a Moloney murine leukaemia virus RTase (MMLV-RT) which switches template when it reaches the end of the template molecule and adds a number of untemplated cytidines to the 5' end of the RNA template. A primer with complementary ribonucleic guanidine bases and a universal adapter can then bind to the poly-C tail and be incorporated into the nascent copy DNA (cDNA) molecule (**Figure 3B**). PCR amplification can then be performed using a primer complementary to the universal adapter and a primer that binds to the constant region. This significantly reduces the amplification bias seen when using cocktails of V gene specific primers, although 5' RACE is thought to be less efficient in capturing BCR transcripts and can therefore result in lower coverage of the repertoires (Chaudhary and Wesemann 2018).

Incorporation of Unique Molecular Identifiers

A second improvement has been the recent addition of unique molecular identifiers (UMIs) into the universal template oligonucleotide. The UMIs label each synthesised cDNA molecule with a unique DNA barcode enabling the researcher to correct for errors introduced during PCR amplification and sequencing errors. Consensus sequences can be identified by aligning reads containing the same UMI. The oligo contains several Uracil bases which are cleaved during Uracil DNA Glycosylase treatment immediately after cDNA synthesis. This prevents free oligos from introducing random UMIs into amplicons during subsequent rounds of PCR amplification. This is thought to improve drastically the accuracy of clonality estimates since samples can be normalised by down-sampling to the same number of UMIs, allowing for the comparison of equal numbers of mRNA transcription events between repertoires. BCR repertoires generated as part of this thesis used a 5' RACE with UMI library prep strategy. In Chapter 4, an approach using gene-specific primers was adopted to attempt to link heavy and light chains as expression-ready recombinant proteins.

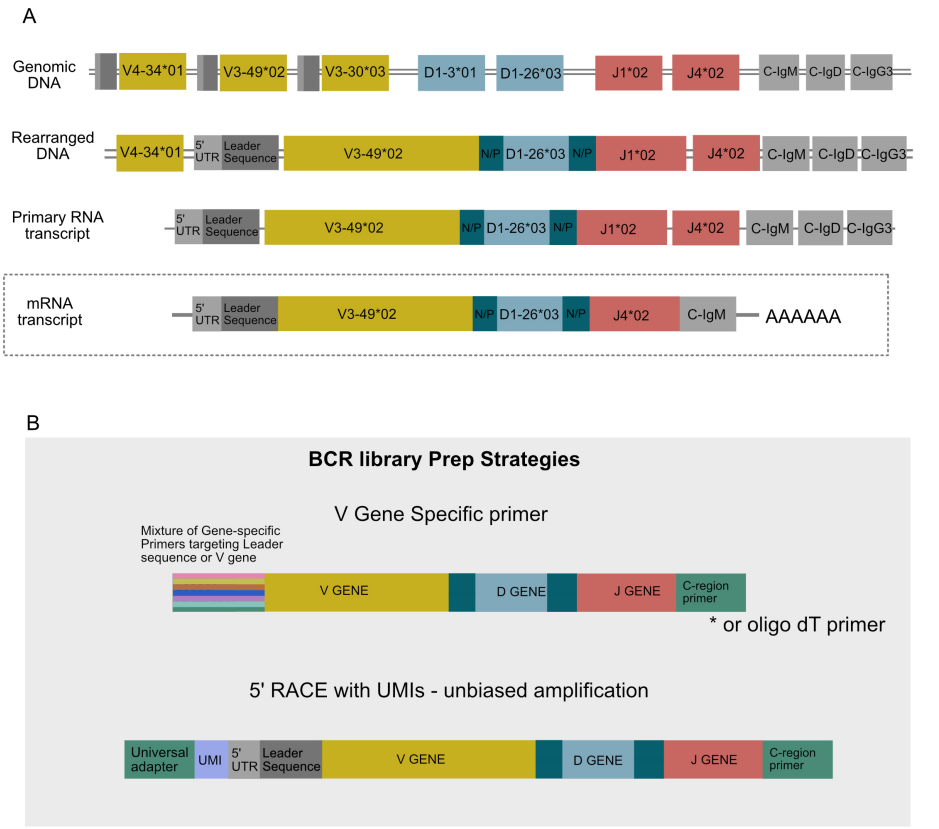


Figure 3: Repertoire sequencing from DNA or mRNA A) BCR libraries can be sequenced from rearranged DNA or from mRNA transcripts. When sequencing from DNA non-productive rearrangements may be sequenced and constant region cannot be identified. Sequencing from mRNA allows the expressed constant region to be identified and the rearranged BCR transcripts are more abundant. In this thesis BCRs were sequenced from mRNA **B)** V gene-specific primer cocktails can amplify all known V genes (this approach is used in Chapter 4) or 5' RACE with UMIs can be used to allow for unbiased amplification of all V genes (used in Chapters 2 and 3). See Chapter 2 for a detailed workflow of this BCR library prep strategy.

Considerations for sequencing

The length of the V,D, and J gene segments is approximately 300-400 nucleotides. When using 5' RACE approaches the leader sequence, 5' untranslated region and UMI will also be sequenced. Currently Illumina sequencing supports a maximum of 600 sequencing cycles. Since the UMI is incorporated at the 5' end of the molecule, stepping into the leader sequence is not an option. Therefore usually a strategy is used whereby sequencing is performed from the very end of the constant region, through the J, D, and V. While the

300 bases in one read will usually cover the CDR3 and most of the V segment, the full V segment may not be captured. Read1 and read2 often do not overlap, making it difficult to assemble a high-fidelity consensus read. We have used an asymmetric sequencing strategy whereby 400 bases are sequenced starting from the end of the constant or beginning of the J gene segment, and only 200 cycles are sequenced from the 5' end of the V, mainly to capture the UMI. This makes it more likely for the full V gene, or at least the majority, to be captured in read 1 which should make the V assignments more reliable. One disadvantage to this method is that most commercial sequencing facilities will not guarantee the sequencing run once custom parameters, such as adjusting the number of cycles per read, are used and sequencing quality drops towards the end of read 1. A different approach is to ligate on sequencing adapters instead of introducing them by PCR and to use 400 cycles in read 1. This has the advantage that both the 5' and the 3' end of the BCR may be sequenced for 400 cycles, providing additional coverage (Turchaninova et al. 2016). Another disadvantage to the 5' RACE sequencing approach is that stepping far into the constant region does not permit different isotype subclasses to be identified. IgG subclasses have different effector functions but can only be distinguished based on polymorphisms in particular portions of the constant region that would require substantial sections of the constant region to be sequenced. One study repurposed the P7 index reads on the Illumina Miseq platform to capture polymorphisms in the IgG subclasses, and used the read 1 primer to cover the remaining constant region so that read 1 would begin sequencing near the beginning of the J gene (Schanz et al. 2014). Finally, using two samples indices is preferable since it reduces the chances of index-hopping occurring between samples where adjacent clusters on the sequencing flow cell may erroneously be assigned the neighbouring spot's index read. Index hopping can introduce sequences that appear to be shared between individuals but have arisen from a technical artefact. Having two indices allows mis-assigned indices to be identified, and the chances of index hopping

occurring in both index reads is low. Advances in long-read sequencing technologies, such as the improving accuracy of the Oxford Nanopore Technology sequencing are likely to make these problems obsolete in future.

1.3.2 Analyses

Signatures in the BCR repertoire can be indicative of active infection, dysregulation of adaptive immune responses and autoimmunity. The following section outlines the AIRRseq workflow and some of the analyses used in this thesis and their relevance for characterising B cell responses in infection and chronic illness.

Bioinformatics workflow

BCR AIRRseq analyses have not been extensively benchmarked or standardised. Currently, most publications implement packages from the Immcantation suite for data pre-processing and use IgBlast for alignment. The bioinformatics pipeline for BCR repertoire analysis involves QC of the raw read data, followed by masking or trimming of primer or adapter sequences in the reads, UMI extraction, assembly of the paired-end reads (often reference-guided if the reads do not overlap), error correction and building a consensus sequence from reads with the same UMI, and alignment to isotype sequences for constant region identification. These assembled consensus sequences are then aligned to germline V,D and J reference databases from IMGT using IgBlast, which outputs an AIRR-format file. Among other parameters, information about the sequence such as V,D, and J call, the alignment and germline sequence, the CDR3 and constant region are contained within this file. These data are then used for downstream analyses described in the next section.

Diversity and Clonality

Since we are often interested in understanding how the B cell response is being shaped by a specific antigenic challenge, quantifying and comparing BCR diversity and clonality is important. In response to infection, we might expect to see clonal expansion of a particular BCR, likely with an IgM constant region early in infection, while boosted immune memory would more likely generate a clonal IgG or class-switched response (Akkaya, Kwak, and Pierce 2020). In autoimmunity, clonal expansions are often restricted to a particular tissue. For example, in Multiple Sclerosis oligoclonal B cell expansions have been observed in the Central Nervous System (CNS) and expressing these BCRs has revealed molecular mimicry between a protein (EBNA1) and a glial autoantigen (GlialCAM) suggesting antigen-driven autoimmune responses in the CNS (Lanz et al. 2022), but the peripheral repertoire remains diverse.

Many metrics for measuring species diversity have been developed in the field of ecology, with concepts borrowed from information theory, and these are often adopted for repertoire analysis. Commonly used metrics include Species Richness, Simpson Diversity, Shannon Entropy, Diversity Evenness 50 (DE50), the Gini Index of Inequality and Renyi Entropy (Greiff et al. 2015; Pelissier et al. 2023). Quantifying or even defining diversity is conceptually complicated and there is little consensus on which diversity metrics to use (Chao, C.-H. Chiu, and Jost 2014; Pelissier et al. 2023). Because different indices emphasise different aspects of diversity, such as the evenness of the repertoire (e.g. Gini Index), or the dominance of expanded clones (Simpson Diversity, Berger Parker Index), it is important to consider which metric is used and what the limitations of it are. For example, Simpson Diversity will mainly be influenced by large clonal expansions while Species Richness, the Gini Index and Shannon Entropy capture smaller differences in diversity but are strongly impacted by the number of unique BCRs captured in the sample (Greiff et al. 2015).

V Gene Analysis

V gene usage is the proportion of BCRs in the repertoire assigned to a particular V gene. It is often interpreted as preliminary evidence of clonal expansion or different individuals mounting a similar response to the same antigen. Because V genes will vary in their CDRs, they will have the propensity to form different binding sites for different antigens. Despite significant variation between individuals, certain V genes are consistently used more frequently than others in repertoires from healthy individuals (Boyd et al. 2010). Significant skewing of the V gene repertoire towards particular V genes is often observed in infection and vaccination: for example in COVID-19 strong convergent BCR signatures using the same V genes have been observed across individuals (Jacob D Galson, Schaetzle, et al. 2020).

Similarly, certain V genes are known to be self-reactive. For example, IGHV4-34 tends to react with red blood cell self antigens (Yucheng Li et al. 1996) and commensal bacteria (Schickel et al. 2017). This V gene is found more commonly in repertoires from individuals with autoimmune disease including SLE, Crohn's Disease and eosinophilic granulomatosis with polyangiitis (R. Bashford-Rogers et al. 2019; C. M. Tipton et al. 2015).

Germline V gene polymorphisms can also be associated with autoimmunity. Polymorphisms in VH2-5 have been associated with susceptibility to MS (Walter et al. 1991); a geographically-associated polymorphism in IGHV1-69 has been linked to susceptibility to rheumatoid arthritis (Vencovskỳ et al. 2002); and homozygous deletions of IGHV3 genes (specifically IGHV3-30*01 and IGHV3-30-3) have been found to be more common in patients with SLE with nephritis (M.-L. Cho et al. 2003).

Somatic Hypermutation

Somatic hypermutation is a unique feature of the BCR repertoire that can provide insight into how specific clonal populations evolve and also whether clonal selection is occurring

at the time of sampling. Somatic hypermutation can be quantified by comparing a given BCR sequence to the germline VDJ sequence it has been aligned to, and counting the number of mismatches. High levels of somatic hypermutation might suggest increased numbers of antigen-maturing B cells in the periphery. For example, high levels of SHM and substantial overlap in similar BCRs in patients with Alzheimer's disease is thought to reflect an antigen-driven response against tau proteins or plasma Aβ (Park et al. 2022). Abnormally low levels of SHM have been observed in rheumatoid arthritis patients (Cowan et al. 2019) and in SLE (C. M. Tipton et al. 2015) which have been hypothesised to be the source of auto-antibodies in these diseases, although their exact role in autoimmune pathogenesis remains to be confirmed. SHM can also be used to construct lineages of affinity maturing BCRs, and can be particularly useful when tracing the lineages of clones between tissues and blood (Stern et al. 2014).

N-Glycosylation in IgG Variable Regions

N-glycosylation sites in BCRs are positions where an oligosaccaride (glycan) is added as a post translation modification of the BCR in the endoplasmic reticulum. In healthy controls these sites are rare since few germline encoded BCRs have the correct motifs. Extensive *N*-glycosylation of variable region of BCRs is frequently observed in autoimmunity (Kissel et al. 2022). In rheumatoid arthritis anti-citrullinated protein antibodies are heavily *N*-glycosylated in the variable domain due to mutations introduced during SHM that create *N*-glycosylation sites. In patients on average over 80% of secreted ACPA-IgG in serum are glycosylated in the variable region not usually seen in healthy repertoires (Rochelle D. Vergroesen et al. 2019). This signature has also been observed in the autoimmune disease Myasthenia Gravis (MG) (Mandel-Brehm et al. 2021). The amino acid motif N-X-S/T (X can be any amino acid except for proline) results in *N*-glycosylation. The motif is rarely found in the germline repertoire but can occasionally be introduced into the repertoire by

IGHV1-8, IGHV4-34, IGHV5-10-1, IGLV3-12 and IGLV5-37 (Mandel-Brehm et al. 2021). Therefore it is thought that these motifs are mostly introduced by SHM in autoimmunity. Glycosylated BCRs have been shown to enhance B cell activation upon antigen binding (Kissel et al. 2022). *N*-glycosylation in the variable regions of BCRs is therefore thought to reflect a breakdown in selection during GC reactions and central tolerance.

1.3.3 Signatures of Infection and Autoimmunity in BCR Repertoires

BCR repertoires are often studied in the context of infection and vaccination, where adaptive immune responses are typically characterised by clonal expansion, class switching from IgM to IgG and accumulation of mutations in BCRs as B cells undergo affinity maturation and selection in germinal centres (Jacob D Galson, Clutterbuck, et al. 2015). However, depending on the antigenic challenge and host factors, different features of the BCR repertoire may be observed in a given disease context. Described below are a few features of BCR repertoires commonly observed in response to infectious disease, vaccination, and in autoimmunity.

Clonal Expansion in Infection

Clonal expansion is frequently observed in response to infection or vaccination. In influenza vaccination, a transient increase in plasmablasts, activated memory B cells, and resting memory B cells are observed around seven days post vaccination with expanded clones biased towards IgG1 (M. Wang et al. 2023). In COVID-19 vaccination, IgM BCRs increase transiently and increased clonality in IgG BCRs is observed with a narrow repertoire of clonally expanded BCRs that target mainly the receptor binding domain of the virus (Kotagiri et al. 2022).

Sequencing BCR repertoires in chronic Hepatitis C Virus infection has revealed that

chronic infection is associated with many heavily expanded B cell clones in the repertoire, particularly in IgM+ memory B cells (Tucci et al. 2018). BCR repertoires in HCV are also reported to be heavily skewed towards specific V genes such as IGHV1-69 and IGHV4-59.

V-gene Skew in Infection

Clonal expansion is often associated with skewing of the repertoire towards particular V genes. In some infections, convergence on specific V genes is reported, indicative of potential genetic predispositions to produce BCRs that target specific pathogens (Avnir et al. 2016). For example, survivors of Ebola virus infection have high IGHV4-39 gene usage in their repertoires (Stewart et al. 2022). Convergent V gene responses have been well characterised in influenza vaccination and infection. IGHV1-69 is a common V gene used in influenza-specific BCRs (Avnir et al. 2016; Katherine JL Jackson et al. 2014). This is thought to be due to certain alleles of IGHV1-69 that encode an unusually hydrophobic loop which enables the BCR to interact with hydrophobic pockets on antigens. This results in IGHV1-69 being a frequent feature of virus-specific repertoires (Crowe 2019).

Somatic Hypermutation Profiles

Mutations in BCRs are thought to accumulate with affinity maturation of BCRs. Levels of somatic hypermutation in memory B cell subsets increase throughout childhood and are thought to reflect the maturation of secondary memory B cell responses (Schatorjé et al. 2014). Conversely, lower global levels of SHM are sometimes a feature of autoimmune disease, such as in systemic lupus erythematosus and rheumatoid arthritis, and are hypothesised to be the result of defects in affinity maturation or extra-follicular B cell activation (Cowan et al. 2019; Ota et al. 2023).

Repertoire Features in Autoimmunity

Autoimmune diseases can present with heterogeneous repertoire signatures, depending on the organs affected and the site from which B cells are sampled. A comparative study of a monogenic, organ-specific autoimmune disease, autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED), with the polygenic systemic autoimmune disease SLE, demonstrated different peripheral BCR repertoire features (Clarke et al. 2023). In APECED, where antibodies target ectoderm-derived tissues, peripheral BCR repertoires had an oligoclonal and highly expanded BCR repertoire with a restricted but high-affinity repertoire of anti-cytokine antibodies (Clarke et al. 2023). In contrast SLE repertoires were diverse and less clonally expanded, and targeted a broader range of self antigens. The authors suggest that these repertoire differences may be explained by mainly T-independent B-cell activation occurring in SLE, while in APECED autoreactive T cells license T-dependent B-cell responses to self antigen resulting in an expanded and specific repertoire of BCRs targeting self-antigen.

In multiple sclerosis, BCR repertoires from peripheral blood do not display any stark signatures, however, clonal B cell expansions are found in the cerebral spinal fluid of patients (Lanz et al. 2022). A study of nearly 600 BCR repertoires, including 136 SLE patients, used BCRs reconstructed from RNA sequencing data to characterise global patterns of BCR repertoires in immune mediated disease (Ota et al. 2023). They identified shorter CDR3 lengths in naive B cells from patients with SLE and Sjogren's syndrome, suggesting skewing of the naive B cell compartment in these autoimmune diseases. Short CDR3 lengths in SLE have also been reported by others (S. Liu et al. 2017). Ota *et al.* also found increased IGHV4-34 gene usage in SLE, which has been reported in other studies. IGHV4-34 usage was skewed in plasmablasts and unswitched memory B cells in SLE in particular. IGHV4-34 is known to produce self-reactive BCRs that can bind to red blood cell self antigens (Yucheng Li et al. 1996). This V gene is frequently increased

in the repertoires of patients with SLE, Crohn's Disease and eosinophilic granulomatosis with polyangiitis (R. Bashford-Rogers et al. 2019; C. M. Tipton et al. 2015). In Ota *et al.*, patients treated with belimumab, which inhibits B cell activation factor (BAFF), displayed reduced IGHV4-34 gene usage in un-switched memory B cells. The authors interpreted this reduction in IGHV4-34 as an indication that treatment dampened extra-follicular B cell activation and partially restored peripheral tolerance.

Finally, *N*-glycosylated CDR3s are also more frequent in myasthenia gravis and rheumatoid arthritis and are thought to reflect defects in tolerance and affinity maturation as most *N*-glycosylation motifs are introduced by somatic hypermutation and are not frequently found in healthy controls (Rochelle D. Vergroesen et al. 2019).

Chapter 2

Characterising B cell Receptor Repertoires in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

2.1 Introduction

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) is a disease of unknown aetiology and pathophysiology that affects an estimated 250,000 people in the UK alone (NICE 2007). ME/CFS has a significant impact on quality of life. Individuals with ME/CFS experience on average greater disability than patients with type 2 diabetes, congestive heart failure, multiple sclerosis (MS) and most cancers (Falk Hvidberg et al. 2015). Despite the substantial impact ME/CFS has on quality of life, many patients experience stigma and disbelief surrounding their illness. People with ME/CFS experience

a range of symptoms that can vary among individuals, but generally include: Fatigue, autonomic dysfunction, cognitive impairment, sleep disturbances, pain, flu-like symptoms and sore throat. Post-exertional malaise (PEM) is the cardinal symptom of ME/CFS (Bateman et al. 2021). There are currently no effective treatments for ME/CFS and the illness is generally lifelong. Only approximately 5% of patients are reported to make a full recovery (Cairns and Hotopf 2005).

2.1.1 Epidemiology

Data from the USA estimates 836,000 to 2.5 million Americans suffer from ME/CFS, but only approximately 10% of cases have been diagnosed (Bateman et al. 2021). A study of 5,809 Norwegian ME/CFS patients reported two age peaks in the incidence of ME/CFS, one between the ages of 10-19 and a second peak between the ages of 30-39 (Bakken et al. 2014), with the second peak being more pronounced in female ME/CFS patients.

Women are more likely to develop ME/CFS than men. Bretherick *et al.* recently published an analysis of survey responses from >17,000 ME/CFS patients in the UK who are being recruited as participants in a Genome-Wide Association Study called "DecodeME". 83.5% of individuals in the study cohort are female, consistent with the literature (Bakken et al. 2014; Gallagher et al. 2004).

ME/CFS patients are often stratified according to the severity of their symptoms and functional impairment. Patients with "mild" ME/CFS are generally able to take care of themselves and can remain in employment or education but are likely to have reduced mobility and are often forced to cut out leisure and social activities to allow them to recuperate. People with moderate ME/CFS have usually stopped employment or education and require regular rest periods. Around 25% of ME patients are severely affected and are reliant on care for the activities of daily living (Pendergrast et al. 2016; Institute of Medicine 2015). Severe ME/CFS patients are mostly house- or bed-bound

and suffer from severe cognitive difficulties, severe light and sound sensitivity and may require a wheelchair for mobility. Very severe ME/CFS patients are completely bed-bound and depend on full-time care including assistance for personal hygiene and eating. Some patients are unable to eat or speak and require tube-feeding (Bateman et al. 2021; Dafoe 2021).

2.1.2 The Research Landscape in ME/CFS

Very few findings have been reproduced in ME/CFS research, leaving the molecular, cellular and genetic basis of the disease poorly understood. Severe under-funding in this field has undoubtedly exacerbated the challenges of studying a chronic, heterogeneous and, in many individuals, fluctuating illness.

A recent analysis of gender inequality in health research compared the ratio of research funding to disease burden across conditions that affect either women or men more (K. Smith 2023). While it identified a general trend of under-funding conditions predominantly affecting women, ME/CFS was highlighted as the most severely under-funded disease that predominantly affects women - it receives only 0.04 times the funding that would be commensurate with its burden. This funding gap is also mirrored in the number of publications in the field of ME/CFS. A literature search of articles deposited in the NCBI PubMed database since 1980, revealed that there are on average 0.11 times the number of publications related to ME/CFS compared to MS, and publications have not increased at the same rate as in MS (**Figure 1**). The ME/CFS Priority Setting Partnership (PSP) has set research priorities for the disease. They used participatory methods involving more than a thousand patients and their carers to produce eleven research priorities for ME/CFS which the government has endorsed (Tyson et al. 2022).

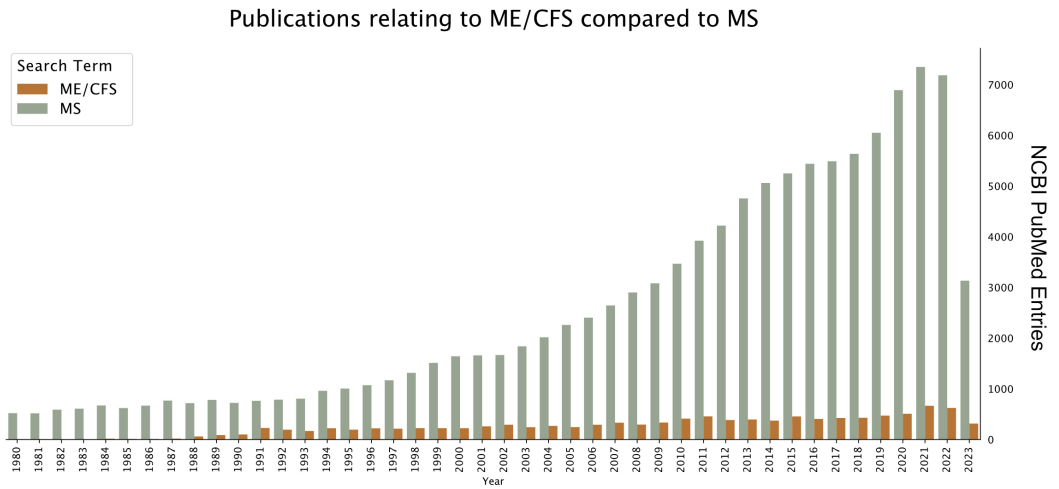


Figure 1: Publications related to ME/CFS compared to MS. Number of annual PubMed entries containing search terms "Multiple Sclerosis" and "Myalgic Encephalomyelitis" or "Chronic Fatigue Syndrome" downloaded from NCBI PubMed. Data are shown from 1980 onwards.

2.1.3 Diagnosis

Among the top three research priorities identified by the PSP is the need for a diagnostic marker. There are currently no validated biomarkers of ME/CFS and it is diagnosed based on the symptom picture. The use of different diagnostic criteria by different investigators, heterogeneous symptomatology and comorbidities all add to the challenge of identifying disease-specific biomarkers.

The terms "Myalgic Encephalomyelitis" and "Chronic Fatigue Syndrome" are used in combination and interchangeably in the literature and originate from different case definitions developed in the UK and USA (Lim and Son 2020). The first case definition for ME/CFS was proposed by Ramsay 1986 in the UK and described ME/CFS as a post-viral syndrome. Twenty-four different case definitions for ME/CFS have followed since, but only the most commonly implemented criteria used in current studies will be discussed here (**Figure 2**).

The criteria developed by Fukuda and colleagues (1994) require fatigue lasting for more than six months alongside at least four of the following symptoms: post exertional malaise, impaired memory or concentration, sore throat, tender axillary or cervical lymph nodes, muscle pain, joint pain (in multiple joints), new headaches or unrefreshing sleep (Fukuda et al. 1994). These criteria are still widely used today but have been criticised for being too non-specific, particularly since many of the symptoms included in the case definition overlap with symptoms of major depression (Institute of Medicine 2015). Additionally, the Fukuda criteria are a diagnosis of exclusion which means diagnoses are often slow to be made (Bateman et al. 2021).

Crucially, the Fukuda case definition does not require patients to experience Post Exertional Malaise (PEM), considered to be the hallmark symptom of ME/CFS. PEM describes the exacerbation of some or all of a patient's symptoms upon physical, physiological or cognitive exertion. PEM can vary among patients in symptoms, duration and the delay from the exerting event to its onset (Stussman et al. 2020). Symptoms of PEM can develop immediately or be delayed by hours or days post exertion and can last for hours, days or even weeks.

The physiological impact of PEM has been measured using repeat cardiopulmonary exercise testing (CPET) in ME/CFS. Healthy control subjects and individuals with other diseases will reproduce the same physiological measures (including oxygen consumption, heart rate, respiratory exchange rate etc.) by CPET on two successive days. ME/CFS patients, in contrast, consistently perform worse on the second day of CPET than their first performance in regard to physiological measures and duration (Keller, Pryor, and Giloteaux 2014; Nelson et al. 2019).

A revision of the diagnostic criteria by a panel of ME/CFS researchers, stakeholders and clinicians produced the Canadian Consensus Criteria (Carruthers et al. 2003). These criteria require PEM, persistent fatigue, pain, sleep disturbances and at least two symptoms related

to neurocognitive problems (cognitive dysfunction, motor-sensory disturbances or short-term memory issues), one symptom related to immune dysfunction (flu-like symptoms, infection susceptibility or food/chemical sensitivity), one symptom related to endocrine dysfunction (GI tract issues, genitourinary problems, or orthostatic intolerance) and one symptom related to autonomic disturbances (e.g. respiratory problems, cardiovascular issues, temperature intolerance, issues with thermo-regulation). These criteria were designed to emphasise the multi-system aspect of the illness and define a more specific cohort of ME/CFS by requiring a number of specific symptoms. Clinically, the CCC criteria are more difficult to implement than the Fukuda criteria. A third set of criteria, the Institute of Medicine (IOM) case definition was published in 2015 (Institute of Medicine 2015). The IOM criteria attempt to strike a balance between using criteria which are generalisable to a cohort with heterogeneous symptoms, whilst also being stringent enough to exclude patients who experience chronic fatigue due to other illnesses (CDC 2022).

PEM is the most common symptom experienced by ME/CFS patients, followed by unrefreshing sleep, confusion or brain fog, fatigue, muscle pain, and gastro-intestinal symptoms (Bretherick et al. 2023). Fatigue and reduced activity are frequently observed in both depression and anxiety, however, PEM and orthostatic intolerances are not commonly associated with either condition (Bateman et al. 2021; NIMH 2023). The most common comorbidities in ME/CFS patients include Irritable Bowel Syndrome (IBS), depression, fibromyalgia, hypothyroidism, anaemia and migraines (Chu et al. 2019; Bretherick et al. 2023). Approximately two-thirds of female and half of male ME/CFS patients experience symptoms of at least one active comorbidity in addition to their ME/CFS symptoms (Bretherick et al. 2023).

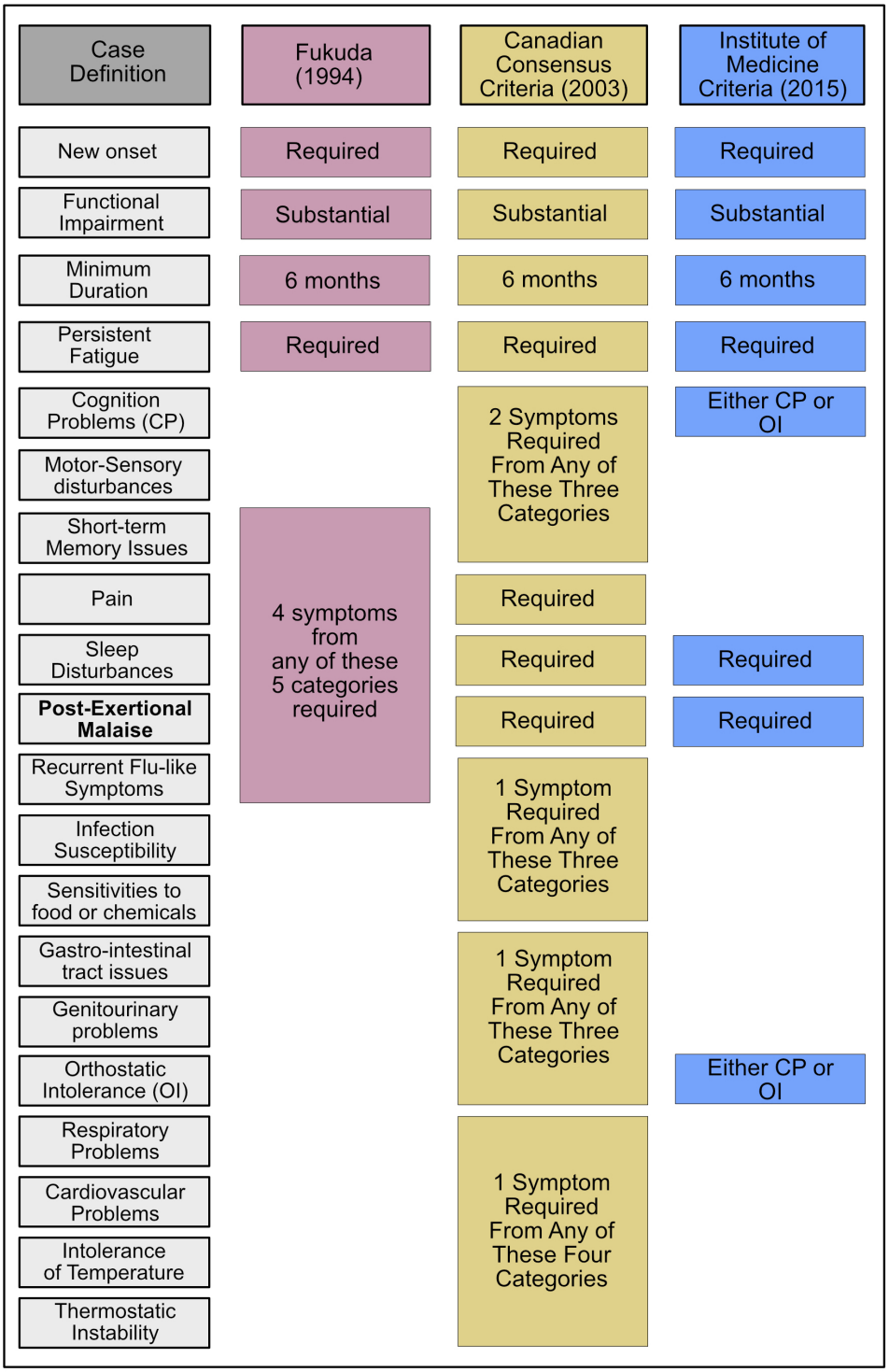


Figure 2: Common Case Definitions of ME/CFS. Summary table re-created from the Open Medicine Foundation Canada website (https://www.omfcanada.ngo/diagnosis-of-me-cfs/). Case definitions discussed in this thesis: Fukuda Criteria (1994), Canadian Consensus Criteria (2003) and Institute of Medicine Criteria (2015).

2.1.4 Potential Aetiologies of ME/CFS

Infection

While most studies into the pathophysiology of ME/CFS to date have not been reproducible, immune dysfunction has long been hypothesised to be important. Several studies have found that more than 60% of ME/CFS patients reported an infectious episode prior to the onset of their symptoms (Chu et al. 2019; Bretherick et al. 2023; Ghali et al. 2020).

A survey of the health records of the entire population of Norway from 2009-2012 investigated whether experiencing a swine flu infection was associated with a greater risk of developing ME/CFS. They compared the hazard ratios (HRs) of developing ME/CFS for people who were infected with H1N1 during the peak of the pandemic, to individuals who received a swine flu vaccine during that time. The HR of developing ME/CFS was 0.97 for individuals who received the H1N1 vaccine, while in people who experienced an H1N1 infection the hazard ratio of developing ME/CFS was doubled (HR: 2.04) (Magnus et al. 2015).

In the DecodeME cohort, 17.2% of respondents reported that glandular fever triggered their illness (Bretherick et al. 2023). In longitudinal follow-up studies of hundreds of individuals with EBV infectious mononucleosis (EBV-IM), an estimated 11% of EBV-IM cases had persistent ME/CFS symptoms six months (P. White et al. 1998), and 4% twenty-four months after infection (B. Z. Katz et al. 2009). Similar longitudinal studies have found Ross River virus and Q fever infections to be associated with ME/CFS (Hickie et al. 2006; Sandler et al. 2022). A recent pre-print which described a small cohort of adolescents and young adults with ME/CFS triggered by EBV infectious mononucleosis. Young adults who developed ME/CFS after EBV-IM reported more severe symptoms and lower chances of recovery than adolescents who developed ME/CFS from EBV (Pricoco et al. 2023). Furthermore, individuals who report glandular fever as the trigger for their

ME/CFS develop glandular fever on average 10 years after the median age of infectious mononucleosis which is 15-19 years in the UK (Bretherick et al. 2023). However, testing for active infectious agents, such as EBV, CMV, and HHV-6, in ME/CFS patients has not produced consistent evidence of active or reactivated pathogens (Bouquet et al. 2017; Shikova et al. 2020; Rasa et al. 2018).

"Leaky gut" has also been proposed as a potential driver of inflammation in ME/CFS and a recent study screened antibodies for reactivity against 244,000 antigens from pathogenic bacterial and viral agents as well as probiotics and commensal bacteria using a phage-displayed antigen library. Severe ME/CFS patients had distinct serum antibody epitope repertoires targeting *Lachnospiracaea* bacterial flagellins and other gut microbiome antigens (Vogl et al. 2022). This suggests that the microbiome could be driving inflammation in severe ME/CFS and may be associated with comorbid IBS and gut disturbances.

Long Covid

As of March 2023, 1.9 million individuals in the UK are living with self-reported Long Covid (Office for National Statistics 2023). Although Long Covid is an umbrella-term which describes a variety of long term conditions triggered by SARS-CoV2 infection, including cardiovascular, thrombotic and cerebrovascular disease; a substantial proportion of Long Covid patients experience symptoms compatible with an ME/CFS diagnosis, including PEM (Davis et al. 2023; Twomey et al. 2022; Kedor et al. 2022). Fatigue is the most common symptom reported (72%), followed by difficulty concentrating (51%), myalgia (49%) and shortness of breath (48%) (Office for National Statistics 2023). In the UK, self-reported Long Covid is most prevalent in people between the ages of 35-69 and females (Office for National Statistics 2023). Interestingly, increased antibody responses directed against other pathogens, particularly EBV and other herpesviruses, were reported

in Long Covid in a pre-print by Iwasaki and colleagues (Klein et al. 2022).

Genetic Predisposition

ME/CFS is thought to have a heritable component. One study on 941 ME/CFS patients found that their first, second and third degree relatives had relative risk ratios for developing ME/CFS, of 2.70, 2.34 and 1.93 respectively (Albright et al. 2011). A recent screen for genetic risk factors for ME/CFS using a factorial analysis to analyse data from the UK Biobank identified 14 candidate genes associated with ME/CFS which are, among other functions, involved in vulnerabilities to stress and infection, and autoimmune development (Das et al. 2022). A pre-print by the same group performed a follow-up study of genes associated with severe and long covid using the same approach. They found that nine of the genes identified in the ME/CFS study were replicated in the long-covid patients who experienced fatigue. The SNPs were linked to genes involved in circadian rhythm regulation and insulin regulation (Taylor et al. 2023).

Autoimmunity

ME/CFS has been proposed to be a potential autoimmune disease, although no specific immunological marker has been found consistently across studies. In most autoimmune diseases middle-aged women are more frequently affected. Autoimmunity tends to run in families, particularly autoimmune thyroid disease, SLE and rheumatoid arthritis (Criswell et al. 2005; Anaya et al. 2012). Autoimmune disease has been reported to be significantly more prevalent in first-degree relatives of ME/CFS patients (OR=5.30; 95%CI: 1.83-15.38; p=0.001), although this finding has yet to be validated in a large cohort (Moslehi, A. Kumar, and Dzutsev 2022).

Human Leukocyte Antigens (HLA) associations are frequently found in autoimmune diseases and a GWAS on 5000 individuals found both Class I and Class II HLA associations

(HLA-C*07:04 and HLA-DQB1*03:03) in 10% of ME/CFS patients and estimated that these were associated with a 1.5-2 fold increased risk of developing ME/CFS (Lande et al. 2020). An open-label clinical trial of cyclophosphamide treatment in ME/CFS has demonstrated modest efficacy, particularly in twelve individuals with HLA DQB1*03:03 and/or HLA-C*07:04 who had an 83% response rate, compared to the remaining cohort (24 patients) who had a 43% response rate (Ingrid G. Rekeland et al. 2020).

In ME/CFS elevated levels of natural antibodies against G-protein Coupled Receptors (GPCRs), in particular β adrenergic receptor 2 (β-2) and muscarinic receptors, have been reported in several studies(Loebel et al. 2016; Freitag et al. 2021; Hartwig et al. 2020; Bynke et al. 2020; Szklarski et al. 2021). Natural autoantibodies against GPCRs, however, are also found in healthy controls and are thought to have physiological roles in regulating homeostasis (Cabral-Marques et al. 2018). Levels of β adrenergeric receptor and muscarinic acetylcholine receptor autoantibody have been found to positively correlate with reported infectious onset and with symptom severity (Freitag et al. 2021). In a small study, β-2 adrenergic receptor activation by IgG was shown to be attenuated in antibodies derived from ME/CFS patients compared to IgG from healthy controls (Hartwig et al. 2020). However, autoantibodies have not consistently been found against the same adrenergic and cholinergic receptor targets (Bynke et al. 2020; Szklarski et al. 2021). Autoantibodies against GPCRs are frequently found in autoimmune diseases (reviewed in Wu et al. 2018). For example, in Myasthenia Gravis autoantibodies can target the second extracellular loop of β2-AR which amplifies T and B cell activity and is involved in the pathogenesis of MG (Lantsova, Gerasimov, and Sepp 2013).

Finally, the risk of developing MS increases by 32-fold after infection with EBV (Bjornevik et al. 2022). Given the frequent association of ME/CFS with EBV as a trigger, it is plausible that at least a subset of ME/CFS patients have an autoimmune disease.

2.1.5 B cells in ME/CFS

Several papers have investigated whether there are differences in B cells from peripheral blood of ME/CFS patients compared to healthy controls. Studies using Flow cytometry have not found any differences in overall CD19+ B cell abundance between ME/CFS and controls (Cliff et al. 2019). One study which characterised B cell subset by flow cytometry found increased numbers of naive and transitional B cells in ME/CFS (A. S. Bradley, Ford, and Bansal 2013). Anti-CD20 B-cell depletion therapy, though anecdotally effective in some ME/CFS patients (Fluge, Risa, et al. 2015), has not shown efficacy in ME/CFS in a double-blind placebo-controlled clinical trial (Fluge, I. Rekeland, et al. 2019).

One study that performed BCR repertoire sequencing in ME/CFS and healthy controls identified increased use of several V genes in ME/CFS patients. IGHV-3-30 and IGHV3-30-3 gene usage in particular was increased in patients who reported an infectious onset to their illness (Sato et al. 2021). The authors suggest this V gene signature may represent a common antigenic trigger among those ME/CFS patients. This result putatively fits a signature of dysregulated IGHV3-30/IGHV3-23 detected by plasma proteomic profiling reported by Milivojevic et al. 2020 (the two V genes are not distinguishable by Mass Spectrometry). Additionally IGHV3-49, IGHD1-26 and IGHJ-6 were found to be increased and replicated in a second cohort in the study by Sato *et al*. The BCR repertoire signature identified by this group has since been patented as a diagnostic of ME/CFS in Japan (WO2020040210A1). However, the evidence of a common antigenic trigger in ME/CFS is lacking. Furthermore, the study quantified repertoires based on in-frame reads rather than UMI counts, which may be more prone to technical artefacts. Additional features of the BCR repertoire, such as affinity maturation were not included in the paper and there were no disease controls. These results therefore require replication and validation.

2.2 Aims

The study described in this chapter applied high throughput sequencing of the peripheral B cell receptor repertoire in 25 mild/moderate and 36 severely affected ME patients and compared these to 21 healthy and 28 MS controls. We took advantage of samples from an existing study of T cell receptor (TCR) repertoires in ME/CFS to use the T-cell depleted samples for BCR repertoire analysis. The added value of this study to the question of immune involvement in ME pathophysiology is two-fold: firstly, to attempt to reproduce the findings of differential V gene usage from Sato *et al.* and secondly to examine whether BCR repertoires display evidence of infection (V gene skew, clonal expansion, somatic hypermutation) or autoimmunity (V gene skew, under-mutated BCR repertoires, increased *N*-glycosylation). Here I applied state-of-the-art BCR repertoire sequencing techniques using Unique Molecular Identifiers (UMIs) and 5' RACE amplification. The analyses I planned at the outset were:

- Does V gene usage differ between ME and healthy controls, in particular for IGHV3-30, IGHV3-30-3, IGHV3-49, IGHD1-26 and IGHJ-6?
- Do clonality or diversity differ between ME/CFS groups and controls?
- Does somatic hypermutation frequency differ ME/CFS between patients and controls?
- Does the frequency of *N*-glycosylation sites in BCRs differ between ME/CFS and controls?

Post-hoc analyses which were performed in addition to those included in the original analysis plan included examining IgM to IgG ratio.

2.3 Results

2.3.1 Repertoire Sequencing Strategy

The BCR library preparation and sequencing strategy was adapted from Turchaninova *et al.* (2016). This strategy employs two state-of-the-art techniques in AIRRSeq: 5' RACE (rapid-amplification of cDNA ends) and Unique Molecular Identifiers (UMIs) (**Figure 3A**). 5' RACE makes use of a Moloney Murine Leukaemia Virus reverse transcriptase that switches template and adds un-templated Cytidines to the end 5' end of the newly synthesised cDNA molecule. A "SMARTNNN" primer which contains ribonucleic Guanidine bases and the template switch motif at the 3' end (followed by the Unique Molecular Identifier and universal MISS_ext adapter) can then anneal to the poly-C template and the reverse transcriptase further extends the cDNA molecule to incorporate the UMI and MISS_ext adapter into the newly synthesised cDNA molecule. The Uridine bases of the SMARTNNN primers are later cleaved during Uracil DNA Glycosylase (UDG) treatment so that the UMIs cannot be incorporated in subsequent PCR reactions. The MISS_ext adapter allows for unbiased amplification of BCRs, as previously cocktails of V-gene specific primers had to be used to amplify BCRs, which resulted in primer biases and could skew mutation profiles in V genes or mask novel V alleles. UMIs drastically improve accuracy of AIRRseq analyses, because each UMI should represent a BCR originating from a single transcription event. This corrects for amplification biases and allows for identification and correction of sequencing and PCR errors, as reads containing the same UMI can be collapsed and used to build a consensus read. While an initial sequencing run performed early in 2020 on the Illumina Novaseq 6000 failed due to low nucleotide diversity and the read 2 primer TM being too low, I introduced improvements to the primer strategy to improve robustness: I elongated the read 2 primer sequence and I added UMIs of three different lengths (11, 12 and 13 N-nucleotides with 3, 4 and 4 constant Uridine bases) to introduce additional

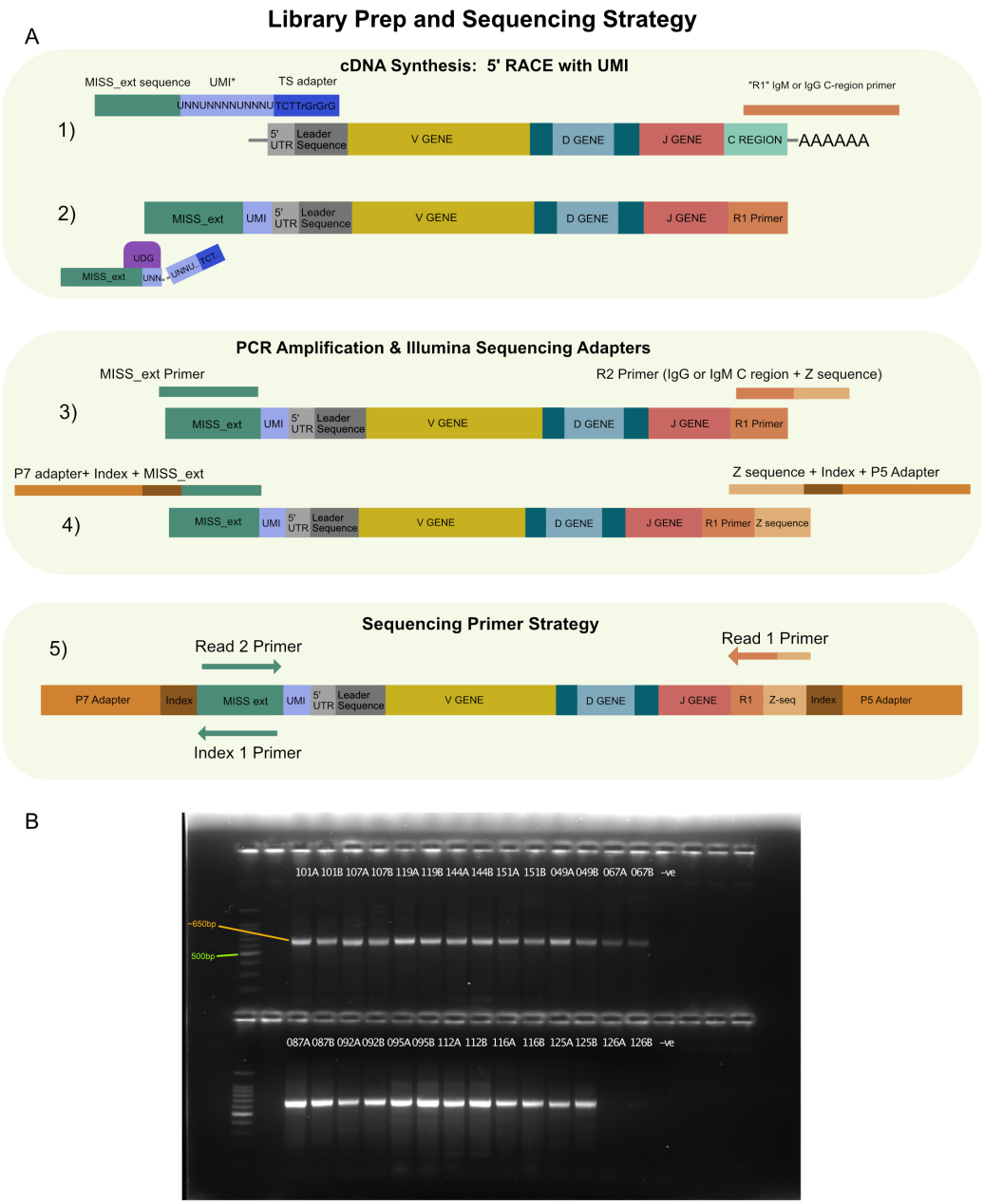


Figure 3: Library prep and sequencing strategy. A) cDNA synthesis was performed using **1)** IgM and IgG 3' constant region specific primers and 5' Rapid Amplification of cDNA Ends (RACE) with a "SMART" primer including a template switch motif with riboguanosine bases and a 12, 14 or 15bp UMI and a universal adapter sequence (MISS_ext) interspersed with Uracil bases so that primers **2)** can be cleaved by Uracil DNA Glycosylase after cDNA synthesis. This was followed by two round of PCR amplification: **3)** In the first round a primer for the universal MISS_ext sequence and "R2" primers for IgG and IgM were used to step further in to the constant region and introduce a universal adapter (Z sequence). **4)** In the second round of PCR two sample-specific barcoding primers were used to introduce Illumina sequencing adapters and sample indices. **5)** Custom Read 1, Read 2 and Index 1 primers were used with 400 cycles of sequencing in Read 1 and 200 cycles in Read 2. **B)** Example gel image of library preps, performed in replicate.

nucleotide diversity early in Read 2 and avoid low quality scores in the UMI. Finally, I originally planned to include IgA isotypes in the analysis, however in the first Novaseq run, repertoires were dominated by IgA (>90% reads) across all samples, which was likely due to amplification bias. It was deemed the additional reagent and sequencing costs required if the reactions were performed separately for each isotype were prohibitively expensive and it was decided to only proceed with IgM and IgG isotypes.

2.3.2 Study Design and Samples

Samples were obtained from a study examining T cell receptor repertoires in ME/CFS performed by Joshua Dibble under supervision of Prof. Chris Ponting at the University of Edinburgh. The aim of the TCR work was to replicate preliminary findings presented by Dr Mark Davis regarding clonal expansion in Lyme disease, MS and ME/CFS. We established a collaboration with the Ponting lab to allow us to examine the BCRs from the same ME and MS patients and healthy controls used in their TCR study. Because sample processing for the TCR study had already begun when we initiated the collaboration, not all samples from the original cohorts were obtained for BCR repertoire analysis. PBMCs were originally obtained from the CureME Biobank (LSHTM). ME/CFS patients were diagnosed as meeting one of, or both, the CCC and Fukuda criteria. All sampled individuals female and age-matched where possible (most individuals were between 40 and 60 years old at the time sampling) (Table 2.2). Fewer severe patients were available in the biobank, and therefore a few younger individuals (age group 18-29) were also included. The CureME Biobank have adopted the classifications of disease severity from the International Consensus Criteria for ME (ICC), which defines "mild" disease as approximately a 50% reduction in pre-illness activity levels, "moderately" affected patients as those who are mostly housebound, "severe" as mostly bed-bound and "very severe" as completely bed-bound and requiring full time care. Samples for TCR study were processed

with collaborators at the Systems Biology Laboratory (SBL) in Oxfordshire. This included positive selection for CD4+ and CD8+ T cells. T-cell depleted samples were stained for CD19+ cells, pelleted by centrifugation and snap frozen. We obtained cell pellets from SBL and performed RNA extraction directly from the T cell depleted pellets. Samples were randomised and blinded throughout processing. After data quality was checked and decisions finalised as to which samples should be excluded based on BCR library quality, I un-blinded samples with regards to which disease or control cohorts they belonged to **(Table 2.1)**.

Cohort	Samples Obtained	BCR libraries analysed
ME/CFSmm	31/40	25
ME/CFSsa	37/40	36
Healthy controls	24/40	21
MS controls	32/40	28

Table 2.1: Samples Obtained and Included

Age Group	18-29	30-39	40-49	50-60
Healthy control	0	1	8	12
ME/CFSmm	0	0	12	13
ME/CFSsa	7	9	5	15
MS control	0	0	8	20

Table 2.2: Age of Individuals Included in the Study

A threshold of 1500 UMIs was set since there was a range of input cell numbers (**Figure 4B**), read counts (**Figure 4C**) and UMI depths (**Figure 4D**) represented across the samples. The threshold was set prior to unblinding the samples. I chose the threshold because this would allow inclusion of samples with lower sampling depths while still permitting sufficient depth to characterise diversity and clonality for analyses where sub-sampling repertoires to equal UMI counts was necessary. This led to the exclusion of thirteen samples (**Table 2.1**, **Figures 4D,E**). Although samples spanned a range of UMI depths, the majority of samples included in the analysis had between 5000-15,000 UMIs (**Figure 4E**). Additionally sample SBL146 was excluded because it was evident

from the B cell counts 4×10^7 CD19+ cells) that this sample had a B cell malignancy. Repertoires from patients with B cell malignancies are heavily skewed regarding clonality and V gene usage and would be a significant outlier in all repertoire metrics. Upon examination of the repertoire for this individual, 96% of their BCRs used IGHV4-34 (**Table 2.3**) and more than 90% of the repertoire was made up of a single IgG CDR3: "CARVGAHYYYYYMDVW" (**Table 2.4**). The patient belonged to the mild/moderate ME group and could represent a lymphoma patient mis-diagnosed with ME/CFS or an ME/CFS patient who developed B cell lymphoma in addition to their ME/CFS.

V gene	% Repertoire
IGHV4−34	96.5
IGHV3−7	0.55
IGHV3−23	0.55
IGHV3−30	0.27
IGHV4−39	0.24

Table 2.3: Top Five V genes SBL146 (ME/CFSmm)

CDR3	% Repertoire
CARVGAHYYYYYMDVW	90.2
XARVGAHYYYYYMDVW	0.30
CARVGAXYYYYYMDVW	0.28
CARVGAHYYYYYMDVX	0.28
CGRGAGWWDYW	0.25

Table 2.4: Top Five CDR3s SBL146 (ME/CFSmm)

2.3.3 No Differences in Clonality or Diversity among groups

First I interrogated whether BCR repertoires in ME/CFS differed with regard to diversity or clonal structure. Diversity metrics can be sensitive to sampling depth, so I first confirmed that none of the disease groups differed significantly regarding the number of UMIs captured per sample (**Figure 5A**). UMI sampling depth was only weakly positively correlated with CD19+ cell counts (r-squared: 0.11) and the relationship was not statistically significant

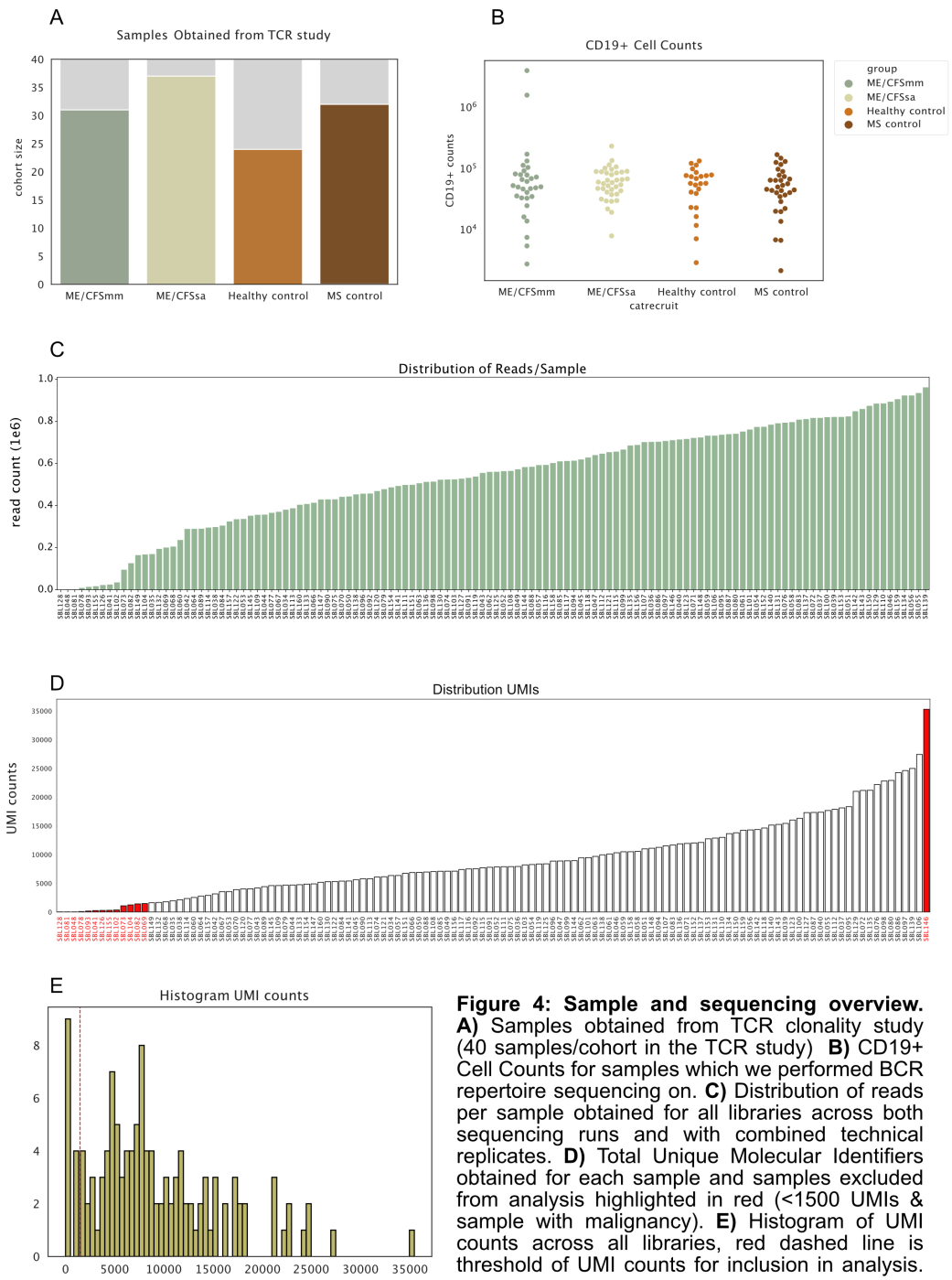


Figure 4: Sample and sequencing overview. A) Samples obtained from TCR clonality study (40 samples/cohort in the TCR study) **B)** CD19+ Cell Counts for samples which we performed BCR repertoire sequencing on. **C)** Distribution of reads per sample obtained for all libraries across both sequencing runs and with combined technical replicates. **D)** Total Unique Molecular Identifiers obtained for each sample and samples excluded from analysis highlighted in red (<1500 UMIs & sample with malignancy). **E)** Histogram of UMI counts across all libraries, red dashed line is threshold of UMI counts for inclusion in analysis.

(**Figure 5B**). Larger numbers of cells should theoretically yield more UMIs but perhaps this reflects limitations on the number of unique UMIs that can be captured at the sequencing depth used in this study. Clonal distributions, as assessed by the relative proportions of the top one, top 10, top 100 and top 1000 CDR3s of the total repertoire, did not appear to differ between groups, and varied between individuals within groups (**Figure 5C**).

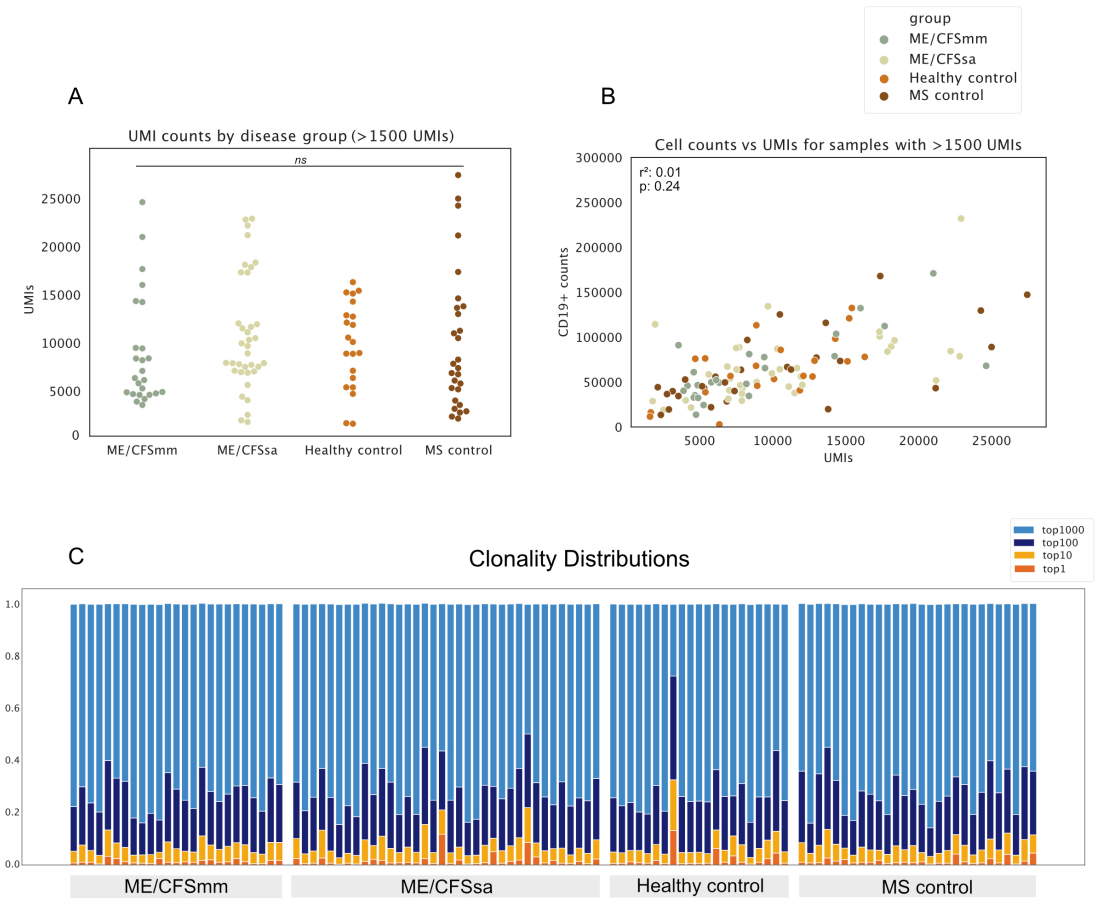


Figure 5: Library overview and clonality. **A)** UMI counts per library included in the analysis. Difference between the group means was tested using a one-way ANOVA (F statistic=0.49, p-value=0.69**)** using the python package scipy.stats. **B)** Correlation between CD19+ Cell Counts and UMI sampling depth (one outlier with high cell counts not shown on plot) - pearson correlation coefficient and p-value displayed on graph ($p > 0.05$, not significant). **C)** Distribution of clonotypes by sample - repertoires randomly subsampled to 1000 UMIs and proportion of repertoire made up of top one, ten, one hundred and one thousand clones calculated. *ns* = 'not significant' (alpha significance threshold $p < 0.05$).

To visualise BCR repertoires, they were plotted as Network Diagrams. Each dot represents a pre-clustered clonotype (based on CDR3 and V-J gene usage) and shared clonotypes are connected by edges. Only clonotypes represented by more than one UMI were included. Clonal distributions did not visually differ between groups and clonotypes shared between groups were extremely rare (**Figure 6A,B,C,D**). Repertoires were coloured based on V gene usage and revealed that V gene usage was diverse in expanded clonotypes across individuals. Colouring the network diagram by isotype demonstrated that most clonal expansions were restricted to one isotype, and that in mild/moderate ME/CFS there appeared to be more IgM. As different diversity metrics capture different aspects of "clonality", I used three different diversity metrics borrowed from economics and ecology to quantify clonality and diversity: the Gini Index of Inequality, Shannon Entropy and Simpson's Diversity Index. The Gini Index and, to a lesser extent, Shannon Entropy are known to be heavily influenced by the number of unique "species" in the sample (in this case, the number of unique CDR3s). To attempt to correct for this, all repertoires were subsampled to 1000 UMIs to compare repertoires of the same size. Additionally, I performed the diversity calculations over 1000 iterations for each individual, re-sampling the repertoire and estimating diversity parameters by taking the average over the 1000 iterations. To check whether this approach corrected for the effect of sampling depth on diversity measures, I first checked whether the Gini Index (calculated from subsampled repertoires) correlated with the UMI depth of the samples prior to sub-sampling (**Figure 7A**). There was a moderate and significant negative correlation (the more UMIs in the sample prior to downsampling, the lower the Gini Index, ie.: the more diverse the sample). This could be explained by samples with greater UMI depth capturing larger numbers of low-frequency BCRs, thereby increasing the diversity of the repertoires. I attempted to correct for this statistically: Testing differences between the mean Gini Index (**Figure 7C**), Shannon Entropy (**Figure 7E**), and Simpson Index (**Figure 7F**), using linear models,

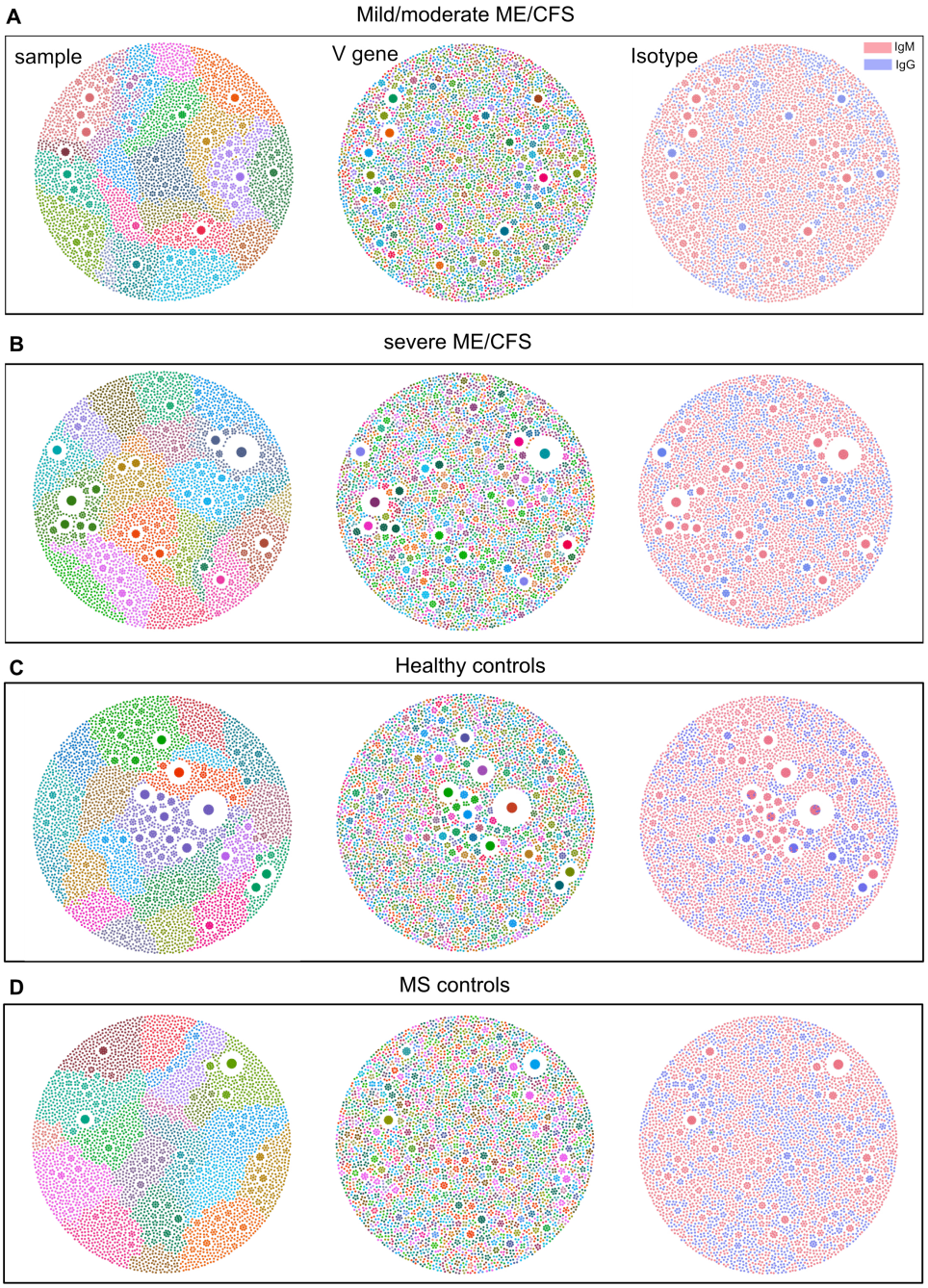


Figure 6: Clonotype network diagrams. BCRs were clustered into clonotypes based on V-J gene usage, CDR3 length and edit distance. Twenty-one individuals were randomly selected from each group and 800 UMIs sampled at random from each repertoire. Clonotypes with more than one UMI were selected and represented as network diagrams for each group. Each dot represents one BCR and BCRs belonging to the same clonotype are connected by an edge and are clustered in close proximity. Plots were coloured by sample (left), V gene (middle) and isotype (right). **A)** Mild/moderate ME, **B)** severe ME, **C)** Healthy controls **D)** MS controls.

with disease group as the explanatory variable and the diversity metric and original UMI sampling depth as fixed effects. None of the mean diversity estimates differed significantly between the groups, but UMI sampling depth was highly statistically significant.

Because BCRs undergo affinity maturation and clones may not represent identical CDR3s, I also performed the diversity analyses using pre-clustered "clonotypes". Clonotypes were generated by first grouping all BCRs by their V and J gene usage and CDR3 length, and then performing hierarchical clustering on these grouped BCRs to assign BCRs within an edit distance of 0.15 amino acids to the same clonotype. Performing diversity analyses with the clonotype-clustered BCRs did not change the correlation of the Gini Index with the original UMI count (**Figure 7B**). Neither the Gini Index (**Figure 7D**), Shannon Entropy (**Figure 7F**), or Simpson Diversity (**Figure 7H**) indices were statistically significantly different, although the repertoires appeared slightly more diverse in the healthy control samples than the ME groups or MS controls when using the Gini Index and Shannon Entropy.

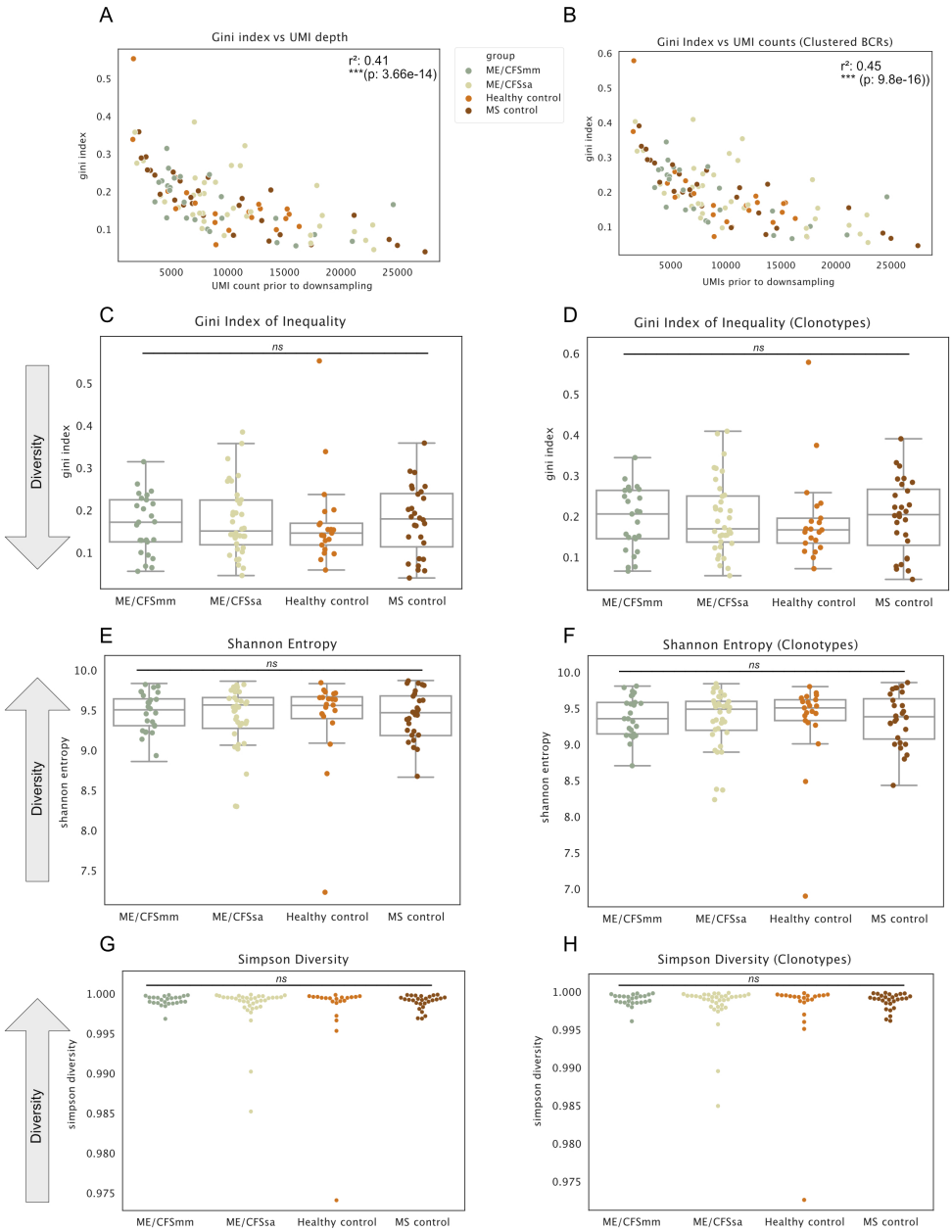


Figure 7: No Difference in Diversity Between Groups. Diversity measure were calculated by subsampling repertoires to 1000 UMIs over 1000 iterations and averaging diversity indices over the iterations. **A)** Correlation of UMI counts prior to downsampling and Gini Index . **B)** Correlation of UMI count and Gini Index for BCRs clustered into "clonotypes" prior to diversity analysis. **C)** Gini Index of Inequality for UMI-matched CDR3 distributions and **D)** clonotype-clustered samples. **E)** Shannon Entropy for CDR3 distributions and **F)** clonotype-clustered samples. **G)** Simpson diversity calculated for CDR3 distributions and **H)** clonotype-clustered distributions. Linear regression shown in A, B performed in python (OLS), in C,D and E were tested using a linear model: *lm(diversity_index~ factor(disease_group)+ UMI_count)* *** p<0.0001, ns = not significant (p>0.05)

2.3.4 V Gene Usage

In Sato et al. 2021, the authors identified four initial differences in V gene usage, three of which were replicated in a second cohort. First V gene usage for all V genes was examined (**Figure 8**). The distribution of V gene usage across repertoires is not random and certain V genes, such as IGHV3-23 (10% of the repertoire) or IGHV4-39 (7% of the repertoire) are used more commonly than others. Repertoires from all four groups followed similar patterns of V gene usage.

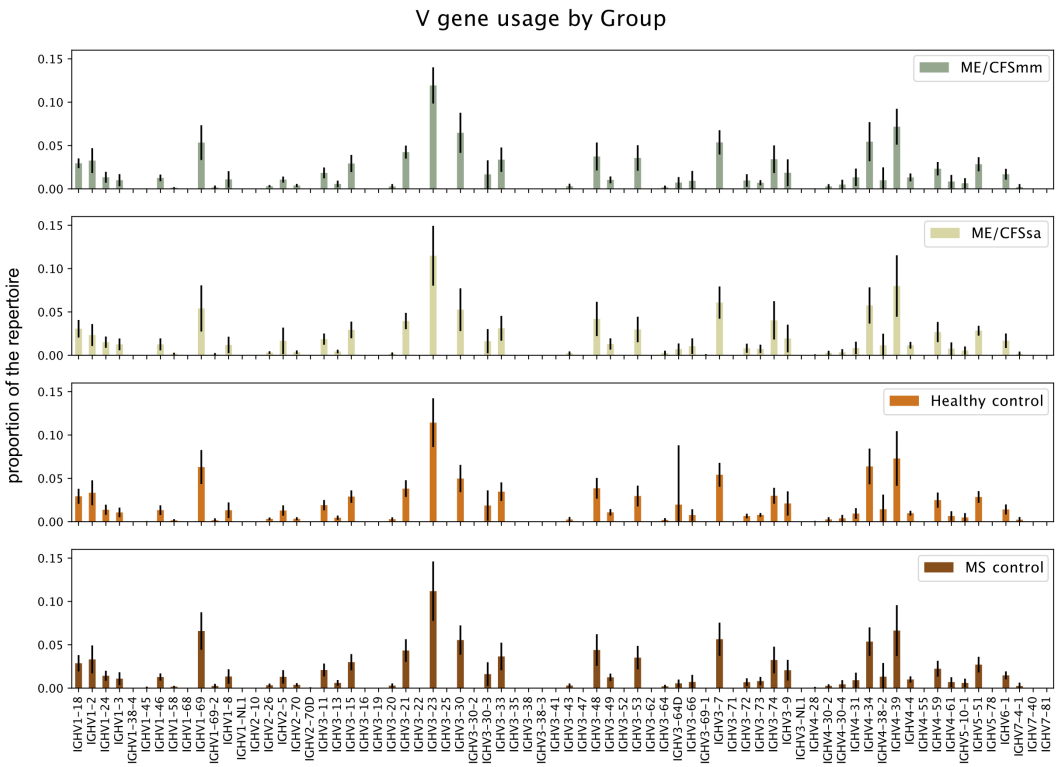


Figure 8: Mean V gene usage by Group. Mean proportion each V gene makes up of the repertoire for each group, averaged across individuals in each group. Error bars represent standard deviation.

While IgM+ BCRs will mostly originate from naive B cells, IgG+ BCRs are antigen-experienced and more likely to have undergone antigen selection and affinity maturation. For this reason V gene usage in IgG and IgM repertoires were examined separately (**Figure 9 and Figure 10**). The dendrogram on the cluster-maps did not separate ME/CFS patients from controls for either isotype. Thus I did not identify a specific set of V genes which differentiates ME/CFS patients from healthy or disease controls. Given the heterogeneity of BCR repertoires within healthy individuals, this is unsurprising.

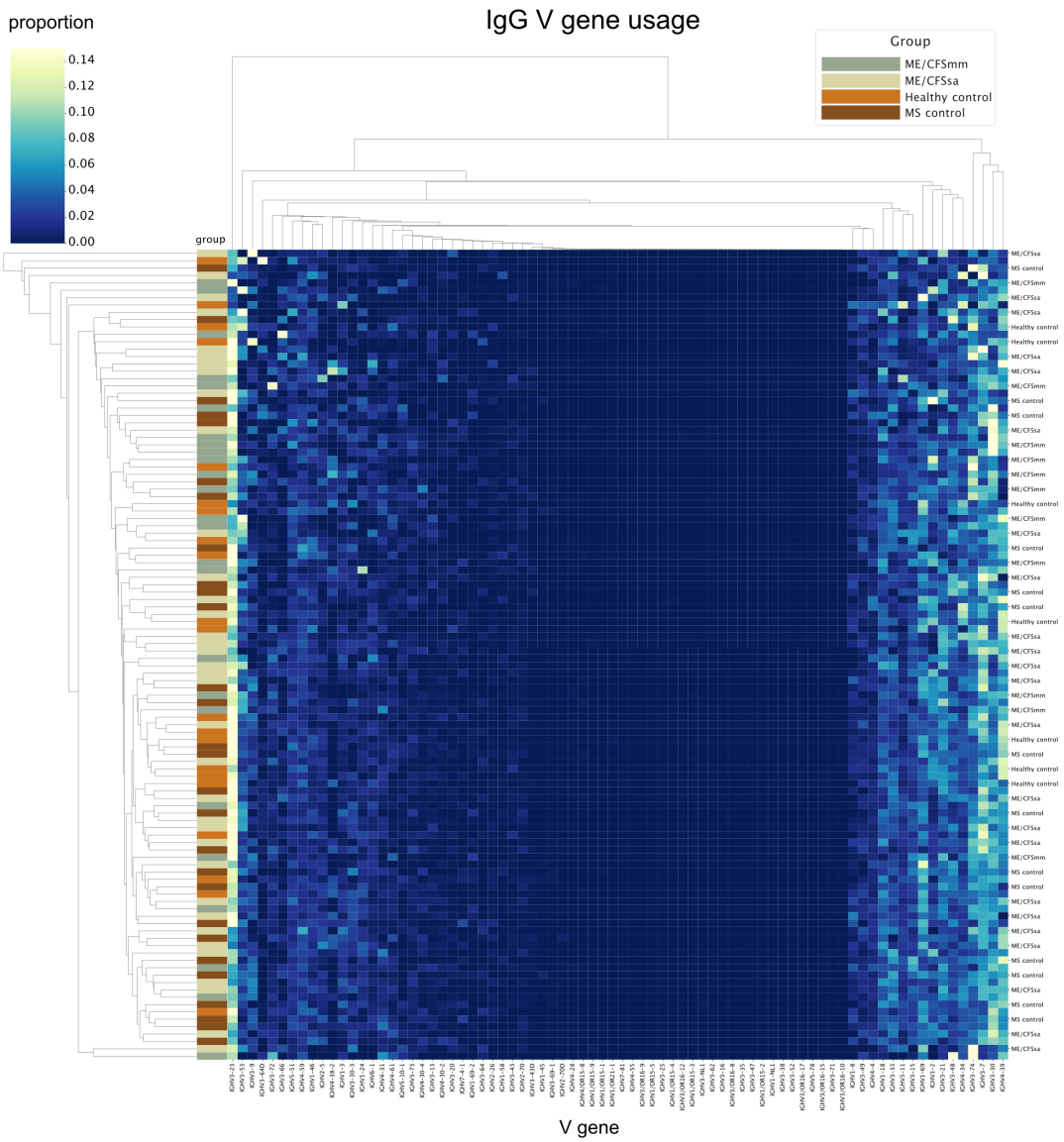


Figure 9: Samples clustered by IgG V gene usage. Heatmap of IgG repertoire V gene usage, clustered based on euclidian distance. Samples were subsetted to IgG BCRs based on their constant region calls. Scale bar represents proportion of the repertoire taken up by a given V gene. Each column is a V gene and each row is an individual's IgG repertoire. Colour bar on the left indicates which group the repertoire belongs to.

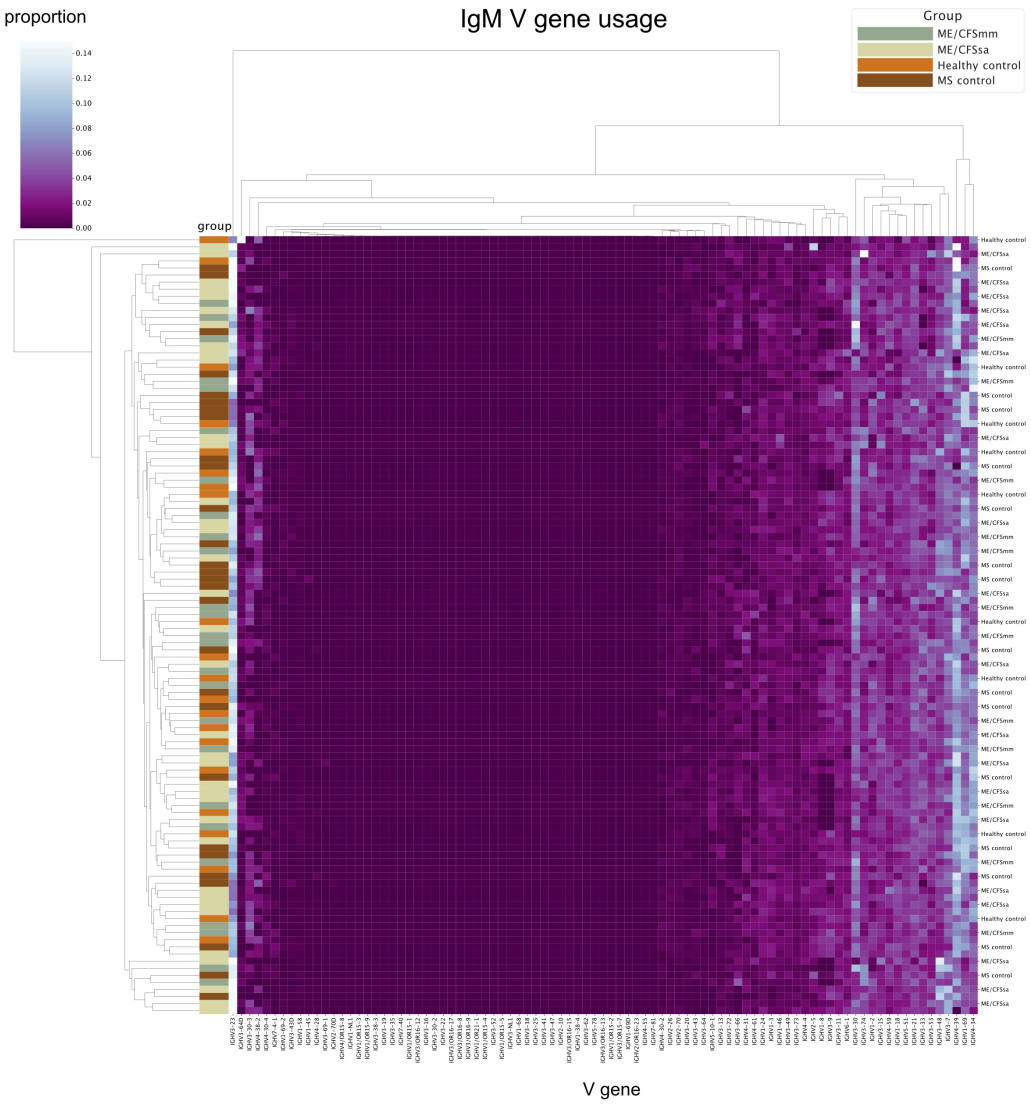


Figure 10: Samples clustered by IgM V gene usage: Heatmap of IgM repertoire V gene usage, clustered based on euclidian distance. Samples were subsetted to IgM BCRs based on their constant region calls. Scale bar represents proportion of the repertoire taken up by a given V gene. Each column is a V gene and each row is an individual's IgM repertoire. Colour bar on the left indicates which group the repertoire belongs to.

2.3.5 IGHV3-30 is Elevated in Mild-Moderate ME/CFS

Next I attempted to replicate findings reported by Sato *et al.* (2021) of increased usage of specific V, D and J in ME/CFS. The authors conclude that their results imply that B cell responses in ME/CFS are directed against the same antigens, skewing the BCR repertoire towards particular V,D and J genes. In their paper they initially identified a V gene signature of elevated IGHV1-3, IGHV3-30, IGHV3-30-3, IGHV3-49, IGHD1-26 and IGHJ6 in ME patients. In a second cohort, they replicated the IGHV3-30, IGHV3-30-3, IGHV3-49 and IGHD1-26, although they compared the patients to the same healthy controls, meaning that the replication was not fully independent. I tested these four V genes, the D- and the J-gene in our data. I used the same statistical tests as the authors and compared the mild/moderate ME and severe patients to the healthy controls separately (**Figure 11 A-F**) and combined (**Figure 11 G-L**). IGHV3-30 was statistically significantly elevated in the mild/moderate ME patients (**Figure 11 A**), but not the severe patients, when compared to the healthy controls. The effect size in our data was moderate (Cohen's D= 0.74), but close to the effect size calculated from the data shown in their paper (Cohen's D= 0.65).

2.3.6 Correcting Mis-assigned V calls

V genes can occasionally be mis-assigned to the wrong IMGT reference allele or even V gene and some individuals have novel alleles which have not yet been added to the reference databases. To address this, the repertoires were run through TIgGER, a novel allele discovery tool which is part of the Immcantation suite. In brief, TIgGER generates an inferred "genotype" for each individual, corrects V-assignments based on this genotype and identifies novel alleles if SNPs are identified which pass their quality thresholds. This revealed approximately 1-2% of the V genes were originally mis-assigned in a majority

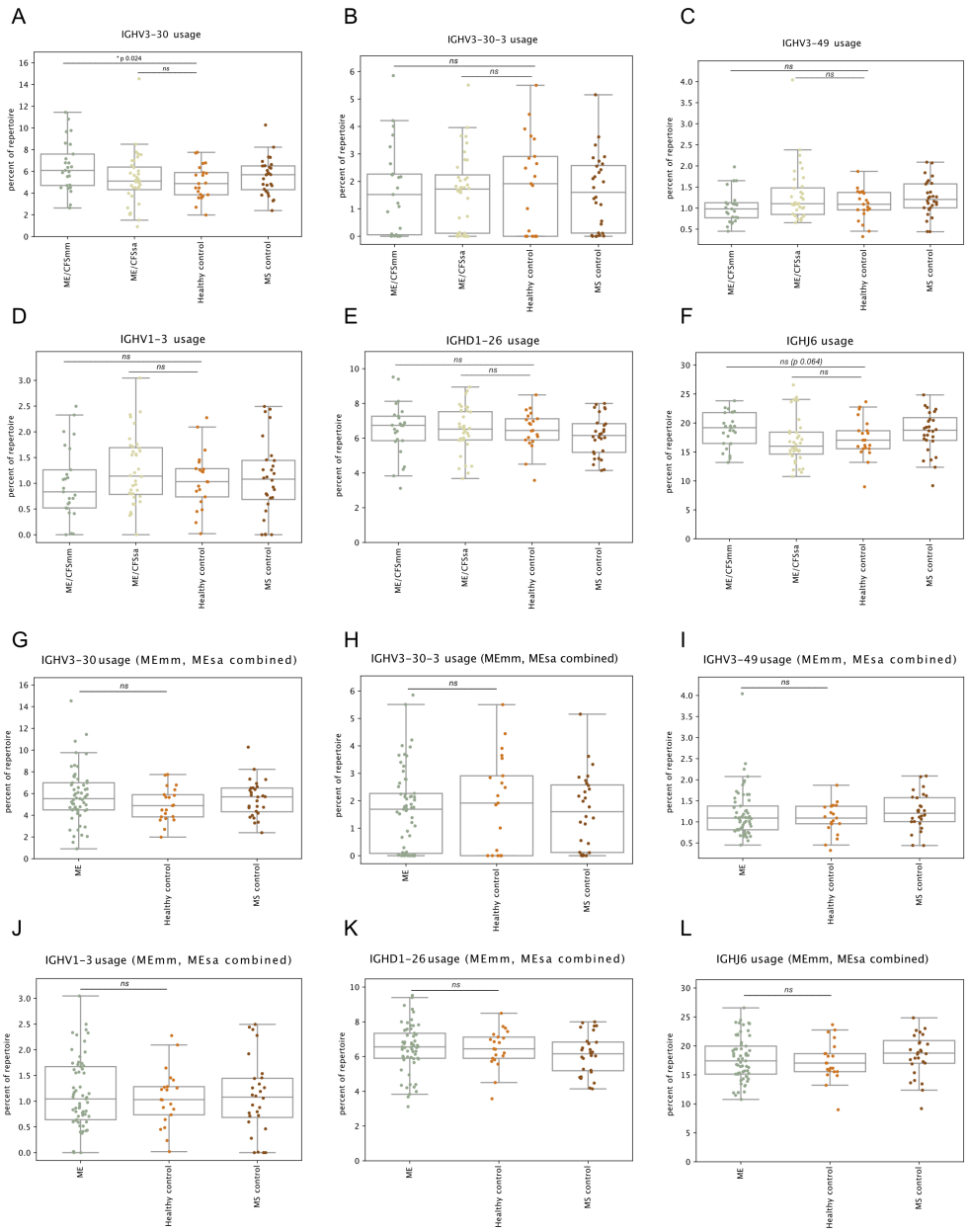


Figure 11: Testing V, D and J genes which were reported as being increased in ME/CFS in *Sato et al 2021*. Specific V, D and J genes of interest were tested for differential usage in ME/CFS compared to healthy controls. Differences to MS controls were not tested. Mann Whitney U tests were performed for comparisons as this was the statistical method applied in the *Sato et al.* paper. Repertoires for all four groups were compared for **A)** IGHV3-30, **B)** IGHV3-30-3, **C)** IGHV3-49, **D)** IGHV1-3, **E)** IGHD1-26, **F)** IGHJ-6. Mild/moderate and severe ME samples were also combined and differences between the ME cohort and healthy controls tested using a Mann Whitney U test. Again **G)** IGHV3-30, **H)** IGHV3-30-3, **I)** IGHV3-49, **J)** IGHV1-3, **K)** IGHD1-26 and **L)** IGHJ-6 genes were compared. * p <0.05, *ns* "not significant", p-values close to significance threshold (alpha 0.05) are shown.

of the repertoires and mis-assignments were equally uncommon across groups (**Figure 12A**).

Novel alleles were rare, but TIgGER did identify 12 novel SNPs across eight V-alleles (**Table 2.4, Figure 12B**). Several SNPs were found in different individuals, increasing the likelihood that these are true novel alleles. IGHV4-39*07 had the SNP 288C>A detected in four different individuals- two MS controls, one ME/CFSsa patient and one Healthy control. One SNP was found in IGHV4-38-2*02, 70A>G, one ME/CFSsa and one MS control and IGHV1-69*08 contained the SNP 191C>T identified in two MS controls. However, the corrected V calls and novel alleles did not affect the results of the V gene analysis described in **Figure 11** and IGHV3-30 remained the only V gene with a significant difference between ME/CFSmm and healthy controls (**Figure 12C**).

Table 2.5: Novel Alleles

V-allele	SNP	Sample	Group
IGHV4-39*01	66C>G	SBL045	ME/CFSsa
IGHV4-39*07	288C>A	SBL051	MS control
IGHV4-39*07	288C>A	SBL071	ME/CFSsa
IGHV4-39*07	288C>A	SBL075	ME/CFSsa
IGHV4-38-2*02	70A>G	SBL080	ME/CFSsa
IGHV4-39*07	288C>A	SBL084	Healthy control
IGHV3-30*02	201T>C	SBL086	MS control
IGHV4-38-2*02	70A>G	SBL086	MS control
IGHV5-51*01	45C>G	SBL086	MS control
IGHV1-69*08	191C>T	SBL106	MS control
IGHV3-21*01	34C>T,40A>C,90C>T,112A>T ,114C>G,119A>G,210A>C	SBL134	MS control
IGHV1-69*08	191C>T	SBL139	MS control

2.3.7 V, D and J Gene Signature Reported in Sato *et al.* Does Not Predict ME/CFS in Our Data

In Sato et al. 2021 the authors performed a Receiver Operating Characteristic (ROC) analysis to assess the predictive power of these BCR attributes and found AUCs of 0.85-0.9,

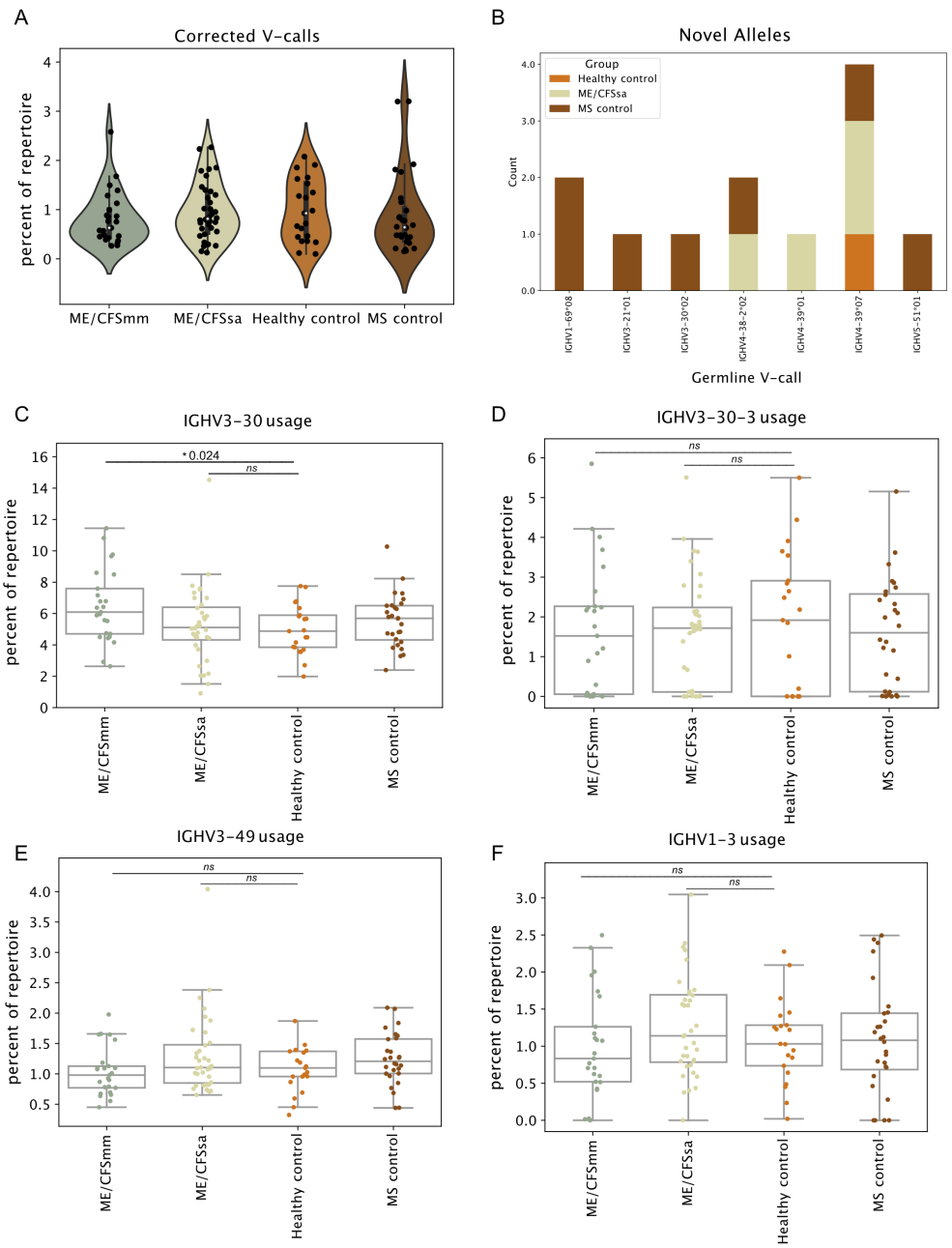


Figure 12: V allele assignment correction. V-calls were corrected based on individual's inferred genotype using the TIgGER workflow and novel alleles identified and tested using the same framework. **A)** Percent of V-calls which were corrected after genotyping. **B)** Alleles in which novel SNPs were identified and passed quality thresholds, coloured by group. **C)** Percent of the repertoire occupied by IGHV3-30, **D)** IGHV3-30-3, **E)** IGHV3-49 and **F)** IGHV1-3. Differences in mean V gene usage were tested using a Mann-Whitney U test, mild/moderate ME and severe ME were each compared to healthy controls. * = $p < 0.05$, ns = not significant.

demonstrating a high sensitivity and specificity to distinguish ME/CFS patients from healthy controls. The method has since been patented as a diagnostic of ME/CFS, filed by Repertoire Genesis (patent number: WO2020040210A1), the company which performed repertoire sequencing in the original paper. Although the methods in the paper do not specify how the ROC analysis was conducted, the patent states "univariate or multivariate logistic regression on the variable using a patient / normal person classification as an objective variable" was used on the same six repertoire features (in addition to B cell or Treg cell counts) from figures provided in the patent. The raw data was not available to replicate the analysis from the Sato *et al.* paper (2021) and apply the model to our data set. To see if I could use the same set of variables to identify ME/CFS patients in our data, I split our own data into a training (60%) and test (40%) data set and used the training data to model a multivariate logistic regression on the six parameters of interest. First I used both the mild/moderate ME patients and severe patients to train and test the classifier. PCA analysis on these six features did not reveal any separation of the ME cohorts compared to the healthy controls (**Figure 13A**). IGHV3-30 gene usage was statistically significant in the model and IGHV3-49 usage was close to significance (p 0.08) (**Figure 13B**). The training data had a modest predictive power (AUC 0.78), however the model performed very poorly on classifying the test data (AUC 0.6) (**Figure 13C**). Next, I hypothesised that ME patients included in the Sato *et al.* paper were more likely to be mild/moderate than severe since the samples were collected at a clinic, and therefore were unlikely to have samples from severe patients who are house- or bed-bound. To make the data more comparable to the samples used in their paper, I performed the analysis on the same training and test data set, this time using only the mild/moderate ME patients and healthy controls. Again, I did not observe any separation of the samples by PCA (**Figure 13D**). In the logistic regression IGHV3-30 was statistically significant and IGHD1-26 was close to the significance threshold (p-value: 0.07). The Area Under

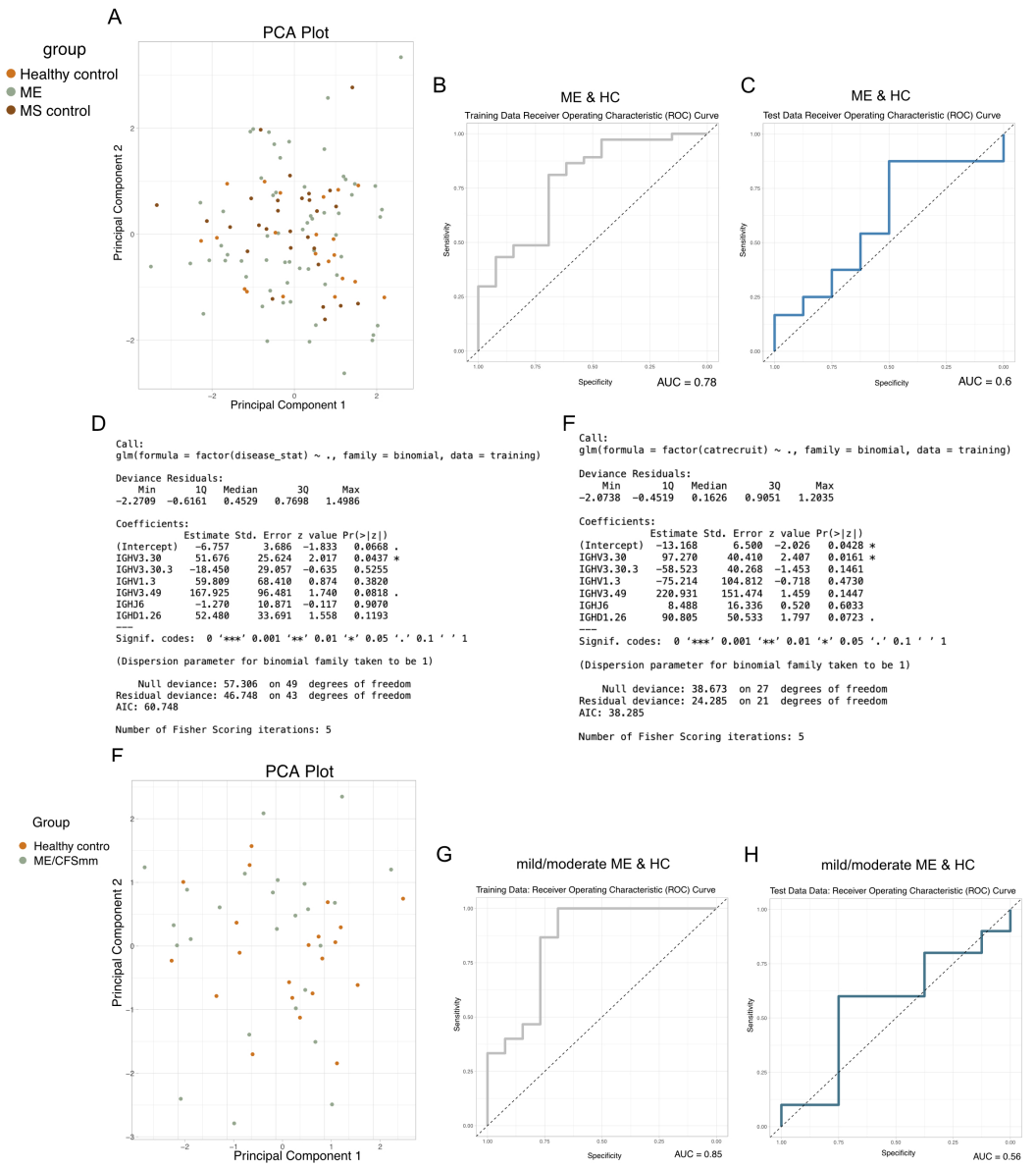


Figure 13: Testing predictive power of *Sato et al. 2021* BCR signature. A) Principal component analysis (PCA) was performed on the six repertoire attributes (IGHV1-3, IGHV3-30, IGHV3-30-3, IGHV3-49, IGHD1-26 and IGHJ6 gene usage) on mild/moderate ME/CFS, severe ME/CFS patients and health controls. Data was split into training (60%) and test datasets (40%), with each group represented in equal proportions in test and training data. **B)** A multi-variate logistic regression model was run on the training dataset and ROC analysis performed (AUC 0.85). **C)** ROC-curve of test data classified using the trained model (AUC 0.56). **D)** Model summary showing coefficients, standard errors and p values for multivariate logistic regression model that was used to train the classifier. IGHV1-3, IGHV3-30, IGHV3-30-3, IGHV3-49, IGHD1-26 and IGHJ6 gene usage were used as independent variables and disease group as the dependent variable. **E)** Model was trained and tested on the same training and test dataset, this time subsetted to include only the Mild/moderate ME patients and healthy controls. **F)** PCA of data from Mild/moderate ME patients and healthy controls used to train the model. **G)** ROC curve of training data (AUC 0.85) and **H)** of the test data (AUC 0.56).

the ROC Curve (AUC) obtained from ROC analysis of the training data was higher than for the previous model, 0.85 (**Figure 13G**), but the classifier performed worse on the test data (AUC 0.56) (**Figure 13H**). Therefore I concluded that the V, D and J gene signature identified in Sato *et al. 2021* does not predict ME/CFS status in our data set.

2.3.8 Ratio of IgM Increased

Upon encounter with a cognate antigen, B cells usually undergo a germinal centre (GC) reaction during which point mutations are introduced in the BCRs and high-affinity antigen-binders selected. These point mutations can be quantified by counting the number of mutations in the sequence compared to the germline reference. In some autoimmune conditions a decrease in mutation frequency has been reported, such as in rheumatoid arthritis or in SLE. This has not been investigated in ME/CFS or in MS before. Comparing the overall mutation frequency between the four groups, I found no differences in mean mutation count (**Figure 14A**). When conducting initial quality checks of the data for SHM analysis, I examined the ratio of IgM to IgG to determine whether mutation frequency could be performed for both IgM and IgG together, or whether the isotypes needed to be analysed separately. The proportion of IgM BCRs was higher in both of the ME groups, but the difference was more pronounced in the mild/moderate ME cohort (**Figure 14B**). The result was statistically significant when performing a Kruskal-Wallis test. Post-hoc testing using Dunn's test found the differences between ME/CFSmm and Healthy controls, ME/CFSmm and MS controls and ME/CFSsa and MS controls to be statistically significant, however, when multiple testing correction was applied (Benjamini-Hochberg false discovery rate), only the difference between ME/CFSmm and MS controls remained significant. Combining the ME/CFSmm and ME/CFSsa cohorts produced the same results, with initial differences between the "ME" cohort and healthy controls, as well as the MS controls being statistically significant and only the differences

between ME and MS controls remaining significant after correction for multiple testing (**Figure 14C**). This analysis was not part of our original plan and should be treated as a preliminary result which would need to be replicated.

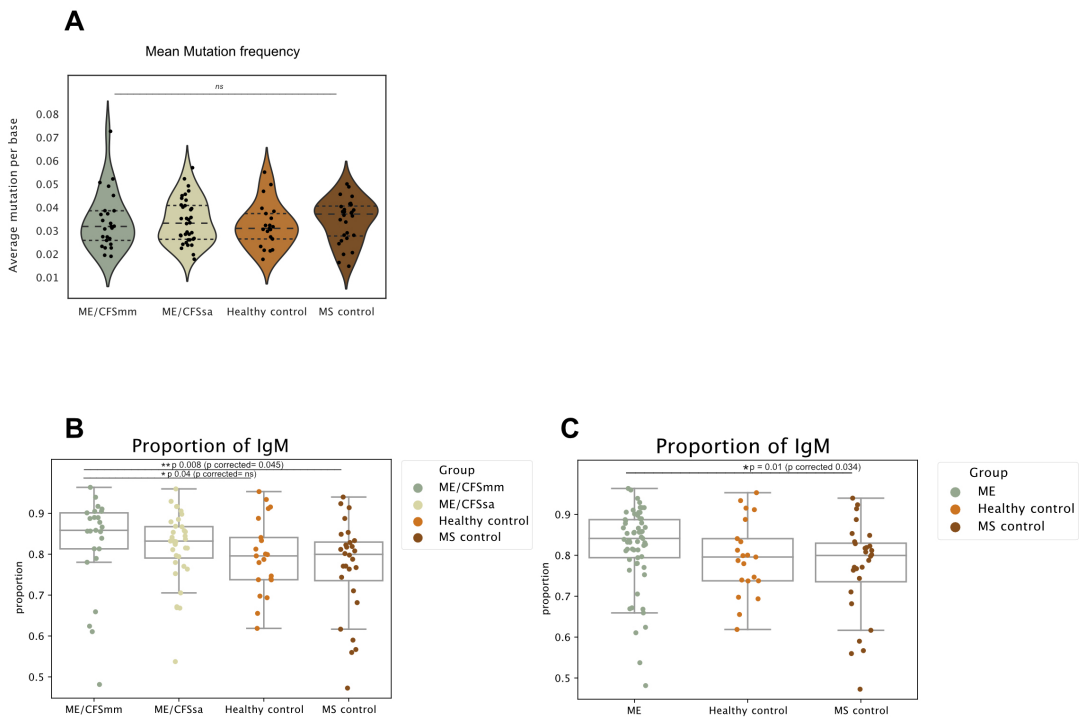


Figure 14: Increased IgM in ME/CFSmm. A) Mean mutation frequency in BCRs **B)** Proportion of BCRs that have IgM constant region for each group **C)** Proportion of BCRs that have IgM constant region - MEmm and MEsa combined compared to healthy and MS controls. Differences in the median proportions of IgM were compared between groupsusing a Kruskal Wallis test (p value: *0.034). Post-hoc testing was performed using Dunn's test. Upon application of Benjamini-Hochberg correction for multiple testing, only differences between ME/CFSmm and MS controls, or grouped ME and MS controls remained statistically significant. * = p <0.05, ns = not significant.

2.3.9 Somatic Hypermutation Does Not Differ Between Patients and Controls

Next, I examined the IgM and IgG repertoires separately to account for the fact that a majority of IgM transcripts are expected to come from naive B cells in the periphery

and will not have any mutations, while mutations in the IgG repertoires follow a normal distribution (**Figure 15**). First I quantified mutation frequency in IgM, calculated by counting the number of mutations in a given sequence compared to the germline reference and dividing it by the length of that sequence. As expected, the distribution of mutation counts in IgM was very right-skewed as a majority of IgM is unmutated in the peripheral repertoire (**Figure 15A**). However, IgM can undergo somatic hypermutation and is represented in the memory B cell compartment. The mean mutation frequency in IgM was similar across the four groups, although the healthy controls appeared to have a trend towards slightly lower mutation frequencies than all three of the disease groups. These differences were not statistically significant. Similarly in IgG there were no significant differences in mean mutation frequency, although the ME/CFSmm patients had on average a slightly higher mutation frequency than the healthy controls. In the kernel density plots, the repertoires from multiple sclerosis patients appeared to be right-skewed suggesting a potential enrichment in unmutated IgG (**Figure 15B**). Therefore the percentage of the IgG repertoire which had one or more mutations relative to germline was quantified. There was no significant difference between the groups. Upon inspection of individual distributions, nine individuals had very right-skewed IgG repertoires in the MS cohort, some of which were extreme; however, repertoires with similar skew were also observed in six of the health controls (data not shown).

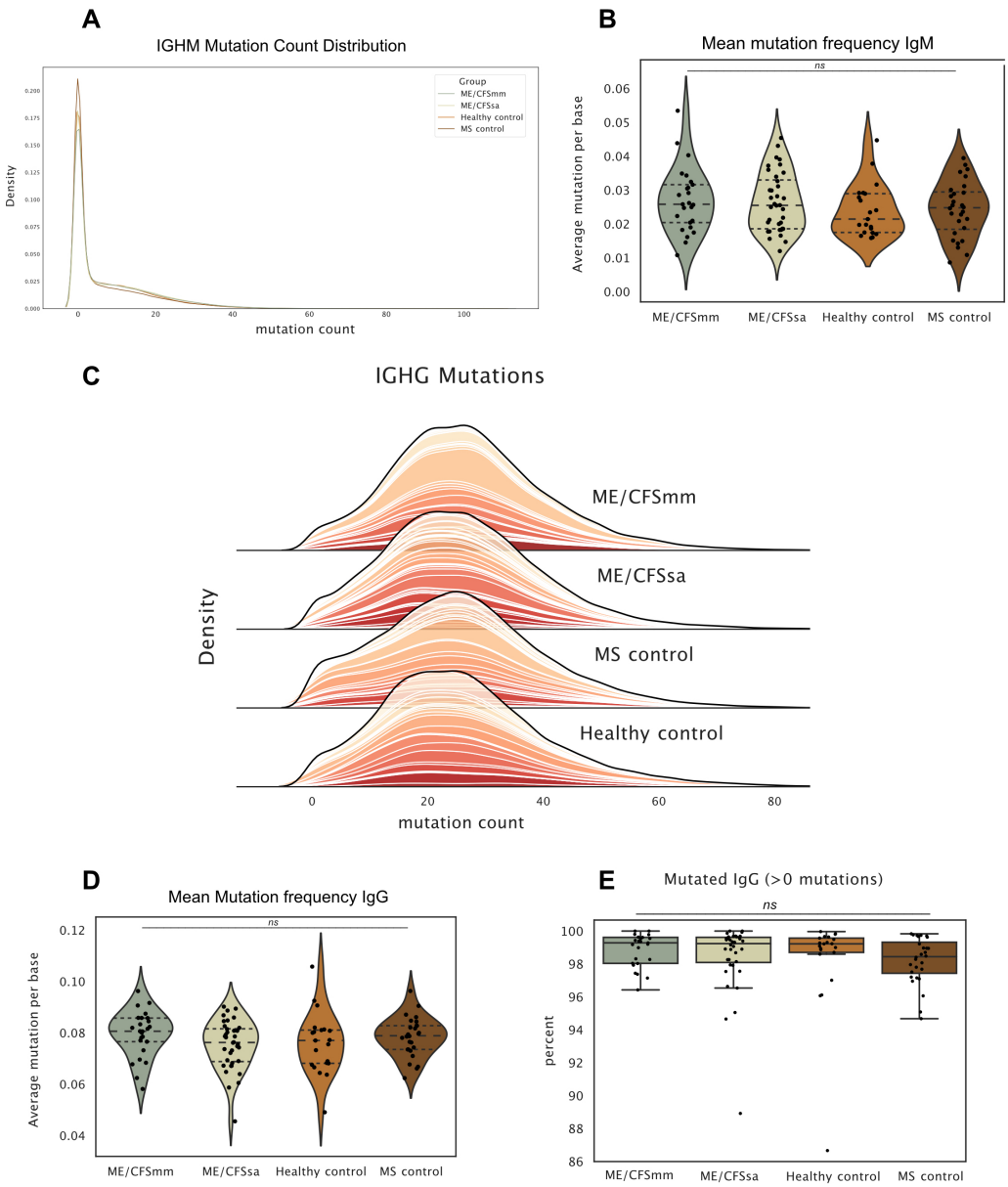


Figure 15: Mutation Frequency In IgM and IgG. A) IgM Density plots of mutation frequency by group. **B)** Mean mutation frequency IgM (median and upper and lower quartiles shown). **C)** Ridge plots of mutation count normalised to mode for each group. **D)** Mean IgG mutation frequency for each group. **E)** Percentage of IgG repertoire which is mutated. Differences in median mutation frequencies were tested for statistical significance using Kruskal-Wallis test. *ns:* not significant.

2.3.10 Frequency of N-Glycosylation Sites Does Not Differ Between Groups

Lastly, I quantified the number of *N*-glycosylation sites per sequence by translating the aligned sequence and identifying the *N*-glycosylation motif (motif N-X-S/T where X can be any amino acid except proline). There were no statistically significant differences in *N*-glycosylation site frequency in any of the groups (**Figure 16A**), although the Healthy controls had, on average, more N-Glycosylated BCRs and the MS controls had the fewest but the variance was high in all three disease groups. *N*-glycosylation sites in IgM followed the same trend (**Figure 16B**). In IgG repertoires the frequency of *N*-glycosylation sites was very similar across all repertoires (**Figure 16C**).

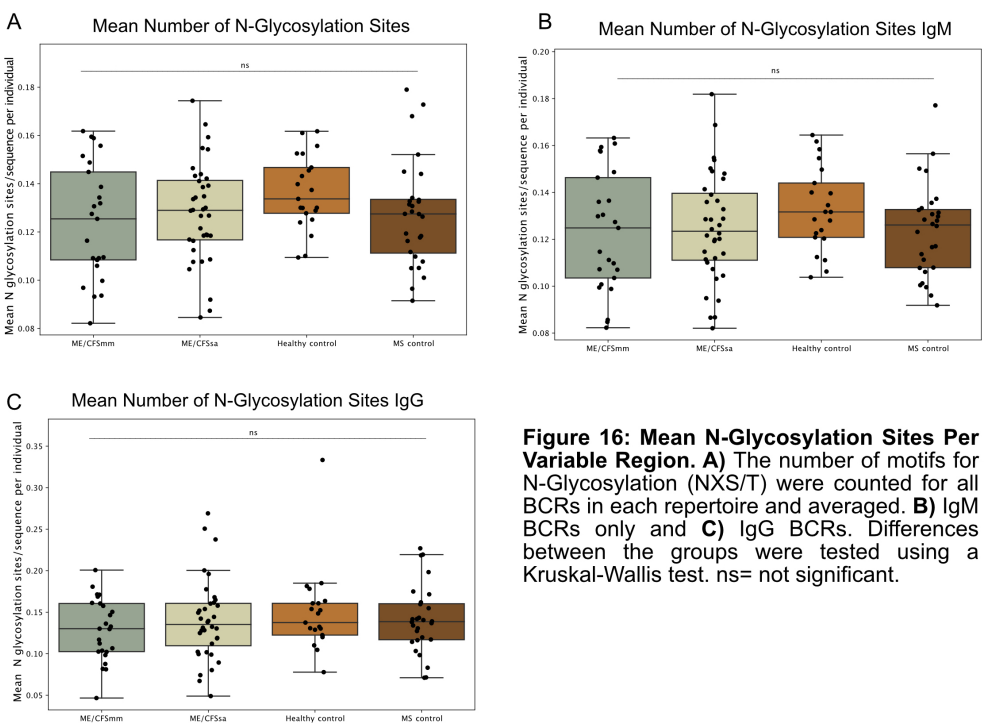


Figure 16: Mean N-Glycosylation Sites Per Variable Region. A) The number of motifs for N-Glycosylation (NXS/T) were counted for all BCRs in each repertoire and averaged. **B)** IgM BCRs only and **C)** IgG BCRs. Differences between the groups were tested using a Kruskal-Wallis test. ns= not significant.

2.4 Discussion

2.4.1 Summary

In this chapter I performed comprehensive repertoire sequencing using state-of-the-art methods to characterise BCR repertoires in two ME/CFS groups and two control groups. The analyses performed were aimed at identifying potential signatures of infection, which we would expect to be associated with increased clonality, differential V gene usage and increases in somatic hypermutation, or autoimmunity, which can also manifest in V gene signatures, decreased somatic hypermutation in IgG and increases in *N*-glycosylation sites in BCRs. I did not observe any striking differences between ME/CFS repertoires and healthy controls indicative of either of these diseased states, and indeed between MS and healthy controls. Repertoire diversity, like in the Sato *et al.* paper, did not differ between the groups. I did replicate one of the reported differences in IGHV gene usage, IGHV3-30. Using the six features of the BCR repertoire from their paper to predict ME/CFS status using logistic regression was unsuccessful in our dataset. Somatic hypermutation profiles were comparable between the groups and heterogeneous among individuals, for both IgM and IgG. There was a higher proportion of IgM in the ME/CFS repertoires, particularly in the mild/moderate ME/CFS patients, but this result was not included in the original hypotheses to be tested, did not survive multiple testing correction, and would need to be validated by other methods such as flow-cytometry. Finally, the number of *N*-glycosylation sites per CDR3 did not differ between the groups. Taken together, I did not observe any striking differences in the peripheral repertoire of mild/moderate ME patients or severe ME patients compared to healthy and MS controls. This is the second BCR repertoire study to be conducted on ME patients, and the first to include a disease control cohort and to assess somatic hypermutation and *N*-glycosylation site frequency.

2.4.2 Partial Replication of Increased IGHV3-30 Gene Usage

One of the specific hypotheses I tested was whether gene usage of the V,D and J genes reported by Sato *et al.* differed between ME patients and healthy controls in our cohort. The result was only statistically significant when the mild/moderate ME patients were compared to healthy controls separately from the severe ME patients. However, the cohorts included in the Sato *et al.* paper are likely to have consisted of mostly mild/moderate ME patients since samples were collected at a hospital which would be a significant barrier to participating in the study for severe patients. The authors did not specify what disease severity the patients had.

Furthermore, the authors found that patients who reported an infection prior to the onset of their ME symptoms had higher levels of IGHV3-30 and IGHV3-30-3. They also reported an inverse relationship between disease duration and levels of IGHV3-30 and IGHV3-30-3 gene usage, and a recent analysis by Bretherick et al (2023) suggested that ME severity is positively correlated with disease duration. This could provide one explanation for why our severe cohort had more variance in IGHV3-30 usage and did not recapitulate the findings in the mild/moderate cohort. The severe cohort was also skewed towards younger ME/CFS patients which could have influenced V gene usage in those repertoires.

The partial replication of increased IGHV3-30 gene usage is interesting since this is also one of the V genes which was detected as significantly different in ME patients in a plasma proteomics study by Milivojevic et al. 2020. The study compared the entire plasma proteome between ME/CFS patients and controls. The authors described a signature of plasma IGHV3-30 or IGHV3-23 antibody (they could not distinguish between the two by Mass Spectrometry) which was either lower or higher in ME patients, whereas the BCR repertoire signature displayed increased use of IGHV3-30. However, given the inability to distinguish between IGHV3-30 and IGHV3-23 by their method, it is possible that one V

gene was elevated and one was reduced compared to controls. Furthermore, repertoire signatures from the peripheral B cell repertoire might not translate to antibody signatures, as the plasma cell compartment will consist of a very restricted subset of BCRs which have terminally differentiated into plasma cells while the peripheral repertoire contains a mixture of naive and antigen-experienced B cells. Sato et al. 2021 conclude that the V gene signature they identified is consistent with a potential common antigen exposure among ME/CFS patients. While the increased use of IGHV3-30 could be suggestive of a common antigen exposure, in the absence of any data about the antigen-specificity of these BCRs, there is no evidence to suggest a common antigen exposure in the ME/CFS repertoires we have sampled. Additional data about whether patients experienced an infection immediately prior to onset of their symptoms might reveal subgroups of patients with clonal or with more homogenous V gene signatures. Despite the moderate effect size (0.71) it is an interesting validation since this was one of only four IGHV genes which I specifically tested for differential V gene usage.

2.4.3 The BCR Repertoire in ME/CFS as a Diagnostic

The six BCR repertoire features (IGHV1-3, IGHV3-30, IGHV3-30-3, IGHV3-49, IGHD1-26 and IGHJ-6) described in Sato *et al.* (2021) did not predict disease status in our data. Having attempted to copy their analysis as closely as possible with the information available to us, the model trained using these features could not classify our test data accurately. A direct comparison of how their trained model works on our data would be necessary to draw any firm conclusions about the sensitivity and specificity of their diagnostic, particularly as the patent reported using either B cell or T-reg cell counts as an additional feature in the multi-variate model. Nonetheless, our analysis should be comparable to the model reported in the paper, and I was unable to replicate those findings. However, V gene usage and alleles can be heterogeneous among different

geographic and ethnic groups, as has been demonstrated with differential B cell responses elicited by the Influenza vaccine across different populations. Furthermore, the current reference databases are biased towards caucasian haplotypes. Interestingly, IGHV3-30 and IGHV3-30-3 are very closely related and in some instances cannot be distinguished from one another. For example, the IGHV3-30*04 allele has the identical nucleotide sequence as IGHV3-30-3*03. This is one reason why I decided to repeat the V gene usage analysis using the corrected V allele assignments using TIgGer, in case IGHV3-30 and IGHV3-30-3 calls had been confounded. The Sato *et al.* paper uses a proprietary method for BCR repertoire sequencing and analysis, making it difficult to directly compare potential differences in our analysis approaches which may have resulted in different V gene usage profiles, but it is possible that novel alleles in the BCR repertoires of the ME patients sampled in their study may have affected their IGHV3-30 and IGHV3-30-3 gene usage profiles, or that the populations they sampled have different V gene usage biases to our cohorts. These factors could explain why I did not fully replicate their findings. One advantage of our approach over the method used by Sato *et al.* (ligation of adapters for PCR amplification), is the use of Unique Molecular Identifiers to label each unique cDNA molecule contributing to the BCR libraries and therefore correct for PCR bias, PCR errors and sequencing errors in our data.

2.4.4 Increased Ratio of IgM in ME Repertoires

One unexpected finding was an increased ratio of IgM in the ME repertoires. This analysis was conducted post-hoc. While this result did not survive multiple testing correction, it could warrant further investigation in the future. We did not sort our B cell populations prior to repertoire sequencing, so it is possible that this reflects a higher number of naive B cells in the ME/CFS samples, or a technical artefact resulting in preferential amplification of IgM BCR transcripts. Since the repertoires were processed blindly, however, it is unlikely

that any technical artefact would account fully for this result, unless it was due to sample collection or PBMC processing. One paper found a greater proportion of naive B cells and a greater proportion of transitional B cells in ME/CFS patients compared to controls. Naive and transitional B cells both express IgM which could explain our finding of an increase in IgM (A. S. Bradley, Ford, and Bansal 2013). However, validation of this increased ratio of IgM to IgG by Flow Cytometry, alongside other markers to confirm the identity of these B cell subsets would be required to confirm the validity of these findings. It is not clear what role an increase in naive and transitional B cells would play in ME/CFS.

2.4.5 MS Repertoires Do Not Have a Peripheral BCR Signature

The inclusion of disease controls is important in ME/CFS, in part to account for the effects of de-conditioning on any phenotype observed in ME, and to validate the specificity of any signatures observed in ME. MS patients were included as disease controls in this study because they are available via the CureME Biobank and have been collected and stored in the same manner as the other samples. Furthermore, MS is often used as a comparator group to ME/CFS since they experience similar symptoms. In this analysis there were no specific BCR repertoire signatures in MS patients, but this is likely because clonally expanded B cells are mostly found in the CNS in MS (Lanz et al. 2022; Büdingen et al. 2012; Colombo et al. 2000; Owens et al. 2001). Clonal expansion, somatic hypermutation and V gene skews have been observed in B cells sampled from the cerebral spinal fluid of MS patients (Lanz et al. 2022; Johansen et al. 2015). This illustrates the subtlety of many BCR signatures: Even well-characterised autoimmune conditions which are known to involve B cells, do not necessarily have a stark peripheral B cell repertoire signature since the site of pathogenesis is often in specific tissues. One paper investigated BCR

repertoires of regulatory B cell subsets in peripheral blood in MS patients since Bregs are widely reported to be dis-regulated in MS. They found transitional Bregs from patients with highly active MS had fewer mutations in BCRs than healthy donors and concluded that this provided evidence that these Bregs were at an early maturation stage (Lomakin et al. 2022). In our data there appeared to be a trend towards lower levels of somatic hypermutation in IgG in MS patients, and a slightly higher percentage of IgG BCRs overall, although neither of these differences were statistically significant compared to healthy controls. Several patients displayed a severe right-skew of mutation frequency in IgG, similar to what has been reported in rheumatoid arthritis and in SLE but this was not statistically significant and similar skews were observed in individuals from the other cohorts.

2.4.6 Other Observations

One mild/moderate ME patient with an apparent B cell malignancy was excluded from this study. It is interesting that this was observed since ME/CFS patients are reported to have a higher risk of developing non-Hodgkin's lymphoma (Chang, Warren, and Engels 2012). Autoimmunity and infectious mononucleosis are both considered risk factors for developing non-Hodgkin's lymphoma. Analysis of health insurer's records for 100,000 cancer and non-cancer controls revealed that ME/CFS was associated with an increased risk of developing non-Hodgkin's lymphoma (odds ratio 1.29), in particular with diffuse large B cell lymphoma and marginal zone lymphoma (Chang, Warren, and Engels 2012).

2.4.7 Future Prospects for the Study of BCR Repertoires in ME/CFS

There is a significant gap in our current understanding of the aetiology and pathophysiology of ME/CFS. This is largely due to dire under-funding of this research area, historically

and in the present. Many patients with long-covid now qualify for an ME/CFS diagnosis based on their symptoms and disease duration. Research on this group with a known infectious trigger where we also know the pathogen, will hopefully provide valuable insight into the immune and homeostatic dysfunctions underlying this disabling disease. While there is not a striking signature apparent from the peripheral B cell repertoire with the current methods we have applied, if we knew the antigen-specificity of BCRs this could provide us with substantial insight into an individual's infection history. This is likely to be relevant in ME/CFS, where infection and particularly EBV infection, appear to trigger disease in significant subset of patients. Being able to stratify patients based on the infectious triggers of their disease would likely be helpful in reducing the heterogeneity among patients, since Bretherick *et al.* reported different symptom and disease patterns associated with different onset types in questionnaire data from 17,000 ME/CFS patients. Very few findings have been replicated in ME/CFS and therefore the observation of increased usage of IGHV3-30 in two BCR repertoire studies and one plasma proteomics paper is noteworthy. It may reflect a common antigen exposure, or a propensity to form auto-reactive B cells. In future it could be valuable to identifiy whether IGHV3-30 gene usage is associated with a particular cell population in ME/CFS. This could be achieved by sorting B cell populations prior to repertoire sequencing. Analysing BCR repertoires from sorted B cell populations would substantially improve the power to detect and interpret potential BCR signatures, as many BCR signatures reported in infection and autoimmunity are restricted to particular B cell subsets.

2.5 Methods

2.5.1 Samples

Samples were obtained from the CureME Biobank, via a collaboration with Prof. Chris Ponting. Ethical approval was given by the University College London Biobank Ethical Review Committee (RFL B-ERC) (ref. EC.2018.006) and the study was sponsored by the University of Edinburgh. Samples were originally sourced from the CureME biobank as frozen PBMC stocks isolated from blood. ME/CFS patients were included if they had a previous diagnosis of ME and met either the Canadian Consensus Criteria or the Fukuda Criteria. Severe patients were sampled in their homes. Samples were initially processed for a TCR repertoire sequencing project by Systems Biology Laboratory (SBL) in Oxfordshire. CD8, CD4 and $\gamma\delta$T cells were obtained by MACS sorting and the remaining cells were stained for CD19+ B cells and snap frozen as cell pellets.

2.5.2 BCR library prep

Blinded T cell depleted samples were obtained from SBL. The library preparation strategy was performed as described in section 2.3.1, adapted from Turchaninova et al (2016). All RNA and cDNA reaction set-up were performed in a hood in an amplicon-free clean room. RNA extraction was performed using the Zymo Quick-RNA Miniprep Plus Kit (Zymo Research#R1058) as per the manufacturer's instructions. RNA lysis buffer was added directly to cell pellets upon thawing and the DNA digestion step was included. cDNA synthesis was performed using the Takara SMARTScribe cDNA synthesis kit (#639538) using constant-region specific primers ("R1 primers", see Table 2.6) to prime the cDNA synthesis reaction. R1 primers were mixed 1:1, resuspended as a 10µM working stock, and 2µl of the primer working stock added to 8µl of the freshly isolated RNA in a sterile

thin-walled 0.2ml lidded reaction tube (Corning #CLS3745) on a cooling block at 4°C. The tubes were then immediately placed in a thermocycler for 2min at 70°C and followed by a 42°C incubation step to anneal the synthesis primers for 1-3 minutes. In the meantime cDNA Synthesis Mix reagents were assembled on a cooling block at 4°C(see Table 2.7). SMARTNNN_ext primers were pooled 1:1:1 to incorporate UMIs of the three different lengths. The 12µl of cDNA synthesis mastermix were then added directly to sample tubes still in the thermal cycler. The components were mixed by pipetting. Samples were then incubated for 60mins at 42°C, before incubating at 70°C for 10 minutes to terminate the cDNA synthesis reactions. Following this, 1µl of Uracil DNA glycosylase (5U/µl) (NEB #M0280L) was added directly into the reaction tube for a 15 minute incubation at 37°C. 2µl of product from the cDNA reaction was added to PCR1 mastermix (see Table 2.8) assembled in the clean room hood in a sterile thin-walled 0.2mL reaction tube on a cooling block. Samples were then brought on ice to a different lab for PCR thermocycling. PCR cycling was performed as follows: Initial denaturation at 98°C for 2 minutes, followed by 18 cycles of 98°C for 10s, 72°C for 15s and 72°C for 25s. This was followed by a final extension of 72°C for 4 minutes. Reagents for the PCR 2 mastermix were again assembled in the clean room on a cooling block (see Table 2.9) and brought to the main lab on ice, where 2µl of PCR1 product was added in a PCR-clean working area. PCR cycling was performed as follows: Initial denaturation at 98°C for 2 minutes, followed by 18 cycles of 98°C for 10s, 72°C for 15s and 72°C for 25s. This was followed by a final extension of 72°C for 4 minutes. (NB.: samples SBL093, SBL102, SBL069 and SBL073 had faint bands and reactions were repeated with 20 cycles in PCR2 for these).

2.5.3 Sequencing

BCR libraries were run on a 2% agarose gel, visualised with 0.5X SYBR-Safe and a band between 400-800 bp excised. A DNA gel extraction was performed using the NEB Monarch

Table 2.6: PCR and Sequencing Primers (5'-3')

cDNA synthesis	
hIGG_r1	GAAGTAGTCCTTGACCAGGCA
hIGM_r1	GTGATGGAGTCGGGAAGGAAG
SMARTNNNext_12ntUMI	AAGCAGUGGTAUCAACGCAGAGTGCUNNNNUNNNNUNNNNUCTTrGrGrG
SMARTNNNext_11ntUMI	AAGCAGUGGTAUCAACGCAGAGTGCNUNNNNUNNNNUNNUCTTrGrGrG
SMARTNNNext_10ntUMI	AAGCAGUGGTAUCAACGCAGAGUGCNNUNNNNUNNNNUCTTrGrGrG
PCR 1	
M1SS_ext	GGCGAAGCAGTGGTATCAACGCAGAGTGC
hIGGE_r2	ATTGGGCAGCCCTGATTARGGGGAAGACSGATG
hIGM_r2	ATTGGGCAGCCCTGATTAGGGGGAAAAGGGTTG
PCR2	
P7+M1SS	CAAGCAGAAGACGGCATACGAGATNNNNNNGGCGAAGCAGTGGTATCAACGCAGAGT
P5+Z	AATGATACGGCGACCACCGAGATCTACACNNNNNNATTGGGCAGCCCTGATT
Sequencing primers	
hIGG_read1ext	ATTGGGCAGCCCTGATTARGGGGAAGACSGATG
hIGM_read1ext	ATTGGGCAGCCCTGATTAGGGGGAAAAGGGTTG
hBCR_read_2ext	GGCGAAGCAGTGGTATCAACGCAGAGTGC
hBCR_Index_1ext	GCACTCTGCGTTGATACCACTGCTTCGCC

Table 2.7: cDNA Synthesis Mix

Component	Supplier and catalogue nr	Volume per reaction
First strand buffer	Takara (#639538)	4ul
DTT	Takara (#639538)	1ul
SMARTNNNext (12uM)	IDT	2ul
dNTP (10mM)	ThermoFisher Scientific (#R0192)	2ul
SMARTscribe RTase	Takara (#639538)	2ul
RNAse inhibitor	Takara (#2313A)	1ul

Table 2.8: PCR1 Mix

Component	Supplier and catalogue nr	Volume per reaction
Nuclease-free water	IDT	4ul
Primer MISS_ext (10uM)	IDT	2ul
R2 primer mix (IgM+IgG combined) (10uM)	IDT	2ul
Phusion flash High Fidelity PCR Mastermix	Thermo Fisher #F548S	10ul

Table 2.9: PCR2 Mix

Component	Supplier and catalogue nr	Volume per reaction
Nuclease-free water	IDT	4ul
Primer P7-SMARTamp w. sample index(10uM)	IDT	2ul
Primer BCR_P5-Rev w. sample index (10uM)	IDT	2ul
Phusion flash High Fidelity PCR Mastermix	Thermo Fisher #F548S	10ul

DNA Gel Extraction kit (catalogue nr T1020S) and the DNA concentration quantified using a Nanodrop. Libraries were submitted to Genewiz (Azenta) for sequencing on the Illumina Miseq v3, with paired-end asymmetric sequencing. 400 cycles were sequenced in Read 1 and 200 cycles in Read 2 as well as two index reads. Raw fastq files were obtained from Genewiz.

2.5.4 Data pre-processing

FastQC (Andrews et al. 2010) was used to check read 1 and read 2 quality for a handful of libraries. Fastq files were pre-processed in pRESTO (Vander Heiden et al. 2014), to remove low-quality reads, pair reads, extract UMIs and build consensus sequences from reads with the same UMIs. These processed consensus sequences were then aligned to IMGT reference databases of IGHV, IGHD and IGHJ genes using IgBlast using the IgBlast-wrapper in change-O (N. T. Gupta et al. 2015). The output .tsv files were then parsed in python and sequences with unproductive BCRs were removed. Libraries with fewer than 1500 unique UMIs associated with a productive BCR were then excluded from further analysis. These decisions were made prior to unblinding the data.

2.5.5 V Allele Reassignment

V alleles were corrected and reassigned in TIgGER (Gadala-Maria et al. 2015). The threshold for the minimum number of sequences required to call a novel allele was set to 50 sequences.

2.5.6 Data Analysis

All analyses were performed in Python (version 2.7.5) using custom scripts in base Python and the pandas package for data manipulation (version 1.3.5) (The pandas development team 2020). The seaborn (version 0.12.0)(Waskom 2021) and matplotlib (version 3.5.3)

(J. D. Hunter 2007) packages were used for plotting unless specified. Analysis scripts will be made available upon publication of the findings.

Clonotype clustering

Clonotype clustering was performed with custom python scripts and validated on a test sample. Clonotypes were clustered using all BCRs from all individuals to allow for the identification of shared clonotypes between different individuals. First, clones were grouped based on their V-J gene call and CDR3 length. Next clones within each of these groups were clustered within a hamming edit distance of 0.15 amino acid changes using hierarchical clustering in scikit.learn (Pedregosa et al. 2011). Clones within each of these clusters were then assigned a "clonotype id" consisting of their V-J gene, CDR3 length and a number assigned to each unique cluster identified using hierarchical clustering.

Diversity analysis

Repertoires were subsampled to 1000 UMIs and Gini Index of Inequality, Shannon Entropy and Simpson's Diversity calculated using custom python scripts. This process was repeated over 1000 iterations and diversity obtained by averaging across all iterations for each repertoire. Diversity was calculated either on unique CDR3s or on pre-clustered clonotypes. Diversity indices were caluclated as follows:

Simpson's diversity:

$$D = 1 - \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)}$$

Where S is the number of species, n_i is the frequency of each species, and N is the sum of the abundances of all species in the distribution.

Shannon Diversity:

$$H = \sum_{i=1}^{S} -(P_i \times lnP_i)$$

where S is the number of species and P is the proportion each species makes up of the population.

Gini Inequality:

$$G = \frac{\sum_{i=1}^{S}(2i - S - 1) \cdot P_i}{n \cdot \sum_{i=1}^{S} P_i}$$

where S is the number of species and P_i is the proportion each species makes up of the population.

2.5.7 Statistical testing

Standard linear regression, Kruskal-Wallis and Mann-Whitney U-tests were performed in Python using the scipy.stats package. Mann-Whitney U-tests were performed for V gene usage testing to make the results comparable to those described in Sato *et al.* (2021). Otherwise, Kruskal-Wallis tests were used to compare other parameters unless specified. This was due to the fact that the variances differed between groups, so a non-parametric test was chosen. Post-hoc tests were performed using Dunn's test in the scikit_posthocs package (Terpilowski 2019) and p-values with and without multiple testing correction reported. Multi-variate linear models and generalised linear models were run using the lm and glm in base R (R version 4.2.0 (2022-04-22)) (R Core Team 2022).

2.5.8 ROC analysis and PCA

Data was first split into a training and test set, with each group being represented in the same proportions as in the original data set. 60% of the data was used for training the model and 40% used to test the model. The same test and training dataset was used for the analysis with both severe and mild/moderate ME patients and with mild/moderate ME patients only. ROC analysis was conducted in R, using a Generalised Linear Model to perform multivariate logistic regression:

```
model <- glm(factor(disease_stat) ~., data = training, family = binomial)
```

Predictions were made using the "predict" function using and ROC analysis was performed using the "roc" function from the "pROC" library in R (Robin et al. 2011). PCA was performed using the "prcomp" function in base R. Results were plotted using ggplot2 (Wickham 2016).

2.5.9 Network Diagrams

Repertoires were randomly subset to the same number of samples for each group and subsampled to 800 UMIs per sample. Next, all singleton reads were excluded. Clonotypes with more than one UMI associated with it were plotted for each group in Gephi (Bastian, Heymann, and Jacomy 2009), with random x,y coordinates assigned to each individual to space the samples out evenly across the graph. Next the Fruchterman Reingold algorithm was used to cluster the datapoints. Graphs were coloured according to IgM and IgG isotype, V gene usage or sample id.

2.5.10 Supplementary Material

Table 2.10: P7 Adapter and Index Primers

Sequence Name	Sequence 5'-3'
P7-SMARTamp1_new	CAAGCAGAAGACGGCATACGAGATAGCTCTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp2_new	CAAGCAGAAGACGGCATACGAGATGATCCTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp3_new	CAAGCAGAAGACGGCATACGAGATCTAGCTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp4_new	CAAGCAGAAGACGGCATACGAGATTCGACTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp5_new	CAAGCAGAAGACGGCATACGAGATCAGTGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp6_new	CAAGCAGAAGACGGCATACGAGATTGACGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp7_new	CAAGCAGAAGACGGCATACGAGATACTGGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp8_new	CAAGCAGAAGACGGCATACGAGATGTCAGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp9_new	CAAGCAGAAGACGGCATACGAGATTACGATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp10_new	CAAGCAGAAGACGGCATACGAGATGACTTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp11_new	CAAGCAGAAGACGGCATACGAGATAGTCTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp12_new	CAAGCAGAAGACGGCATACGAGATTCAGTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp13_new	CAAGCAGAAGACGGCATACGAGATCTGATCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp14_new	CAAGCAGAAGACGGCATACGAGATCCTTCCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp15_new	CAAGCAGAAGACGGCATACGAGATAAGGCCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp16_new	CAAGCAGAAGACGGCATACGAGATGGAACCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp17_new	CAAGCAGAAGACGGCATACGAGATATATGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp18_new	CAAGCAGAAGACGGCATACGAGATTGGTACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp19_new	CAAGCAGAAGACGGCATACGAGATCAACACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp20_new	CAAGCAGAAGACGGCATACGAGATGTTGACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp21_new	CAAGCAGAAGACGGCATACGAGATACCAACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp22_new	CAAGCAGAAGACGGCATACGAGATACGTTGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp23_new	CAAGCAGAAGACGGCATACGAGATGTACTGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp24_new	CAAGCAGAAGACGGCATACGAGATCTAGTGGGCGAAGCAGTGGTATCAACGCAGAGT

Continued on next page

Table 2.10 – continued from previous page

Sequence Name	Sequence 5'-3'
P7-SMARTamp25_new	CAAGCAGAAGACGGCATACGAGATTGCATGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp26_new	CAAGCAGAAGACGGCATACGAGATGCCGCGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp27_new	CAAGCAGAAGACGGCATACGAGATATTACGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp28_new	CAAGCAGAAGACGGCATACGAGATGGTTGGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp29_new	CAAGCAGAAGACGGCATACGAGATAACCGGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp30_new	CAAGCAGAAGACGGCATACGAGATCCAAGGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp31_new	CAAGCAGAAGACGGCATACGAGATCTCTAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp32_new	CAAGCAGAAGACGGCATACGAGATTCTCAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp33_new	CAAGCAGAAGACGGCATACGAGATAGAGAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp34_new	CAAGCAGAAGACGGCATACGAGATGAGAAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp35_new	CAAGCAGAAGACGGCATACGAGATCGATTAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp36_new	CAAGCAGAAGACGGCATACGAGATGCTATAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp37_new	CAAGCAGAAGACGGCATACGAGATGTGTCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp38_new	CAAGCAGAAGACGGCATACGAGATACACCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp39_new	CAAGCAGAAGACGGCATACGAGATTGTGCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp40_new	CAAGCAGAAGACGGCATACGAGATCACACAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp41_new	CAAGCAGAAGACGGCATACGAGATTCCTGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp42_new	CAAGCAGAAGACGGCATACGAGATCTTCGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp43_new	CAAGCAGAAGACGGCATACGAGATGAAGGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp44_new	CAAGCAGAAGACGGCATACGAGATAGGAGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp45_new	CAAGCAGAAGACGGCATACGAGATAATTAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp46_new	CAAGCAGAAGACGGCATACGAGATGGCCAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp47_new	CAAGCAGAAGACGGCATACGAGATCCGGAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp48_new	CAAGCAGAAGACGGCATACGAGATTAGCTAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp49_new	CAAGCAGAAGACGGCATACGAGATATCGTAGGCGAAGCAGTGGTATCAACGCAGAGT

Continued on next page

Table 2.10 – continued from previous page

Sequence Name	Sequence 5'-3'
P7-SMARTamp50_new	CAAGCAGAAGACGGCATACGAGATTAATCGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp51_new	CAAGCAGAAGACGGCATACGAGATCGGCCGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp52_new	CAAGCAGAAGACGGCATACGAGATGCGCGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp53_new	CAAGCAGAAGACGGCATACGAGATCGCGGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp54_new	CAAGCAGAAGACGGCATACGAGATTATAGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp55_new	CAAGCAGAAGACGGCATACGAGATGCATATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp56_new	CAAGCAGAAGACGGCATACGAGATATGCATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp57_new	CAAGCAGAAGACGGCATACGAGATCGTAATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp58_new	CAAGCAGAAGACGGCATACGAGATGCCTCTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp59_new	CAAGCAGAAGACGGCATACGAGATGCGAGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp60_new	CAAGCAGAAGACGGCATACGAGATGATGCCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp61_new	CAAGCAGAAGACGGCATACGAGATCGAGTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp62_new	CAAGCAGAAGACGGCATACGAGATACGTACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp63_new	CAAGCAGAAGACGGCATACGAGATCGCTCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp64_new	CAAGCAGAAGACGGCATACGAGATATCTGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp65_new	CAAGCAGAAGACGGCATACGAGATATCTATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp66_new	CAAGCAGAAGACGGCATACGAGATAGCGATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp67_new	CAAGCAGAAGACGGCATACGAGATATCAGGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp68_new	CAAGCAGAAGACGGCATACGAGATAGTGACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp69_new	CAAGCAGAAGACGGCATACGAGATCGAAGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp70_new	CAAGCAGAAGACGGCATACGAGATCTGACAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp71_new	CAAGCAGAAGACGGCATACGAGATTCAACGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp72_new	CAAGCAGAAGACGGCATACGAGATAGTACCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp73_new	CAAGCAGAAGACGGCATACGAGATGACACTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp74_new	CAAGCAGAAGACGGCATACGAGATGTAATCGGCGAAGCAGTGGTATCAACGCAGAGT

Continued on next page

Table 2.10 – continued from previous page

Sequence Name	Sequence 5'-3'
P7-SMARTamp75_new	CAAGCAGAAGACGGCATACGAGATGTCGAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp76_new	CAAGCAGAAGACGGCATACGAGATACTGAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp77_new	CAAGCAGAAGACGGCATACGAGATTGAGACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp78_new	CAAGCAGAAGACGGCATACGAGATCAGGATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp79_new	CAAGCAGAAGACGGCATACGAGATTGAAGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp80_new	CAAGCAGAAGACGGCATACGAGATCATAGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp81_new	CAAGCAGAAGACGGCATACGAGATGTGGATGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp82_new	CAAGCAGAAGACGGCATACGAGATTTAGCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp83_new	CAAGCAGAAGACGGCATACGAGATCCAGTAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp84_new	CAAGCAGAAGACGGCATACGAGATTTGGTGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp85_new	CAAGCAGAAGACGGCATACGAGATAACGTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp86_new	CAAGCAGAAGACGGCATACGAGATGGTGTTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp87_new	CAAGCAGAAGACGGCATACGAGATCGTCAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp88_new	CAAGCAGAAGACGGCATACGAGATTACCAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp89_new	CAAGCAGAAGACGGCATACGAGATTAGCACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp90_new	CAAGCAGAAGACGGCATACGAGATGAGTGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp91_new	CAAGCAGAAGACGGCATACGAGATTACATTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp92_new	CAAGCAGAAGACGGCATACGAGATAATCGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp93_new	CAAGCAGAAGACGGCATACGAGATTACCGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp94_new	CAAGCAGAAGACGGCATACGAGATAATCTAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp95_new	CAAGCAGAAGACGGCATACGAGATAGACCTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp96_new	CAAGCAGAAGACGGCATACGAGATCGAACAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp97_new	CAAGCAGAAGACGGCATACGAGATAGACTGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp98_new	CAAGCAGAAGACGGCATACGAGATTCTCTCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp99_new	CAAGCAGAAGACGGCATACGAGATCTCCTTGGCGAAGCAGTGGTATCAACGCAGAGT

Continued on next page

Table 2.10 – continued from previous page

Sequence Name	Sequence 5'-3'
P7-SMARTamp100_new	CAAGCAGAAGACGGCATACGAGATTCGTAAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp101_new	CAAGCAGAAGACGGCATACGAGATCAACTGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp102_new	CAAGCAGAAGACGGCATACGAGATCACTGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp103_new	CAAGCAGAAGACGGCATACGAGATATGCGAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp104_new	CAAGCAGAAGACGGCATACGAGATTTATACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp105_new	CAAGCAGAAGACGGCATACGAGATGTGTGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp106_new	CAAGCAGAAGACGGCATACGAGATGGATCAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp107_new	CAAGCAGAAGACGGCATACGAGATAAGTCGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp108_new	CAAGCAGAAGACGGCATACGAGATCTCATAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp109_new	CAAGCAGAAGACGGCATACGAGATTGTCACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp110_new	CAAGCAGAAGACGGCATACGAGATCTGAAGGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp111_new	CAAGCAGAAGACGGCATACGAGATAGAGTAGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp112_new	CAAGCAGAAGACGGCATACGAGATAACAGTGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp113_new	CAAGCAGAAGACGGCATACGAGATATGAGCGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp114_new	CAAGCAGAAGACGGCATACGAGATGATTACGGCGAAGCAGTGGTATCAACGCAGAGT
P7-SMARTamp115_new	CAAGCAGAAGACGGCATACGAGATTTACGAGGCGAAGCAGTGGTATCAACGCAGAGT

Table 2.11: P5 Adapter and Index Primers

Sequence Name	Sequence 5'-3'
BCR_P5-Rev1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GAG CTA TTG GGC AGC CCT GAT T
BCR_P5-Rev2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GGA TCA TTG GGC AGC CCT GAT T
BCR_P5-Rev3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GCT AGA TTG GGC AGC CCT GAT T

Continued on next page

Table 2.11 – continued from previous page

Primer Name	Sequence 5'-3'
BCR_P5-Rev4	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GTC GAA TTG GGC AGC CCT GAT T
BCR_P5-Rev5	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CAC TGA TTG GGC AGC CCT GAT T
BCR_P5-Rev6	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CGT CAA TTG GGC AGC CCT GAT T
BCR_P5-Rev7	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CCA GTA TTG GGC AGC CCT GAT T
BCR_P5-Rev8	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CTG ACA TTG GGC AGC CCT GAT T
BCR_P5-Rev9	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TCG TAA TTG GGC AGC CCT GAT T
BCR_P5-Rev10	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG AAG TCA TTG GGC AGC CCT GAT T
BCR_P5-Rev11	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG AGA CTA TTG GGC AGC CCT GAT T
BCR_P5-Rev12	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG ACT GAA TTG GGC AGC CCT GAT T
BCR_P5-Rev13	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG ATC AGA TTG GGC AGC CCT GAT T
BCR_P5-Rev14	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG GAA GGA TTG GGC AGC CCT GAT T
BCR_P5-Rev15	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG GCC TTA TTG GGC AGC CCT GAT T
BCR_P5-Rev16	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG GTT CCA TTG GGC AGC CCT GAT T
BCR_P5-Rev17	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CAT ATA TTG GGC AGC CCT GAT T
BCR_P5-Rev18	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TAC CAA TTG GGC AGC CCT GAT T
BCR_P5-Rev19	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TGT TGA TTG GGC AGC CCT GAT T
BCR_P5-Rev20	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TCA ACA TTG GGC AGC CCT GAT T
BCR_P5-Rev21	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TTG GTA TTG GGC AGC CCT GAT T
BCR_P5-Rev22	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC AAC GTA TTG GGC AGC CCT GAT T
BCR_P5-Rev23	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC AGT ACA TTG GGC AGC CCT GAT T
BCR_P5-Rev24	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC ACT AGA TTG GGC AGC CCT GAT T
BCR_P5-Rev25	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC ATG CAA TTG GGC AGC CCT GAT T
BCR_P5-Rev26	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GCG GCA TTG GGC AGC CCT GAT T

Continued on next page

Table 2.11 – continued from previous page

Primer Name	Sequence 5'-3'
BCR_P5-Rev27	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GTA ATA TTG GGC AGC CCT GAT T
BCR_P5-Rev28	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC CAA CCA TTG GGC AGC CCT GAT T
BCR_P5-Rev29	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC CGG TTA TTG GGC AGC CCT GAT T
BCR_P5-Rev30	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC CTT GGA TTG GGC AGC CCT GAT T
BCR_P5-Rev31	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TAG AGA TTG GGC AGC CCT GAT T
BCR_P5-Rev32	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TGA GAA TTG GGC AGC CCT GAT T
BCR_P5-Rev33	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TCT CTA TTG GGC AGC CCT GAT T
BCR_P5-Rev34	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TTC TCA TTG GGC AGC CCT GAT T
BCR_P5-Rev35	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT AAT CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev36	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATA GCA TTG GGC AGC CCT GAT T
BCR_P5-Rev37	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT GAC ACA TTG GGC AGC CCT GAT T
BCR_P5-Rev38	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT GGT GTA TTG GGC AGC CCT GAT T
BCR_P5-Rev39	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT GCA CAA TTG GGC AGC CCT GAT T
BCR_P5-Rev40	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT GTG TGA TTG GGC AGC CCT GAT T
BCR_P5-Rev41	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CAG GAA TTG GGC AGC CCT GAT T
BCR_P5-Rev42	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CGA AGA TTG GGC AGC CCT GAT T
BCR_P5-Rev43	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CCT TCA TTG GGC AGC CCT GAT T
BCR_P5-Rev44	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTC CTA TTG GGC AGC CCT GAT T
BCR_P5-Rev45	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT TAA TTA TTG GGC AGC CCT GAT T
BCR_P5-Rev46	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT TGG CCA TTG GGC AGC CCT GAT T
BCR_P5-Rev47	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT TCC GGA TTG GGC AGC CCT GAT T
BCR_P5-Rev48	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT AGC TAA TTG GGC AGC CCT GAT T
BCR_P5-Rev49	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ACG ATA TTG GGC AGC CCT GAT T

Continued on next page

Table 2.11 – continued from previous page

Primer Name	Sequence 5'-3'
BCR_P5-Rev50	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GAT TAA TTG GGC AGC CCT GAT T
BCR_P5-Rev51	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC GGC CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev52	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CGC GCA TTG GGC AGC CCT GAT T
BCR_P5-Rev53	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CCG CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev54	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CTA TAA TTG GGC AGC CCT GAT T
BCR_P5-Rev55	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TAT GCA TTG GGC AGC CCT GAT T
BCR_P5-Rev56	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TGC ATA TTG GGC AGC CCT GAT T
BCR_P5-Rev57	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TTA CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev58	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GAG GCA TTG GGC AGC CCT GAT T
BCR_P5-Rev59	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CTC GCA TTG GGC AGC CCT GAT T
BCR_P5-Rev60	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG GCA TCA TTG GGC AGC CCT GAT T
BCR_P5-Rev61	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG ACT CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev62	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TAC GTA TTG GGC AGC CCT GAT T
BCR_P5-Rev63	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT GAG CGA TTG GGC AGC CCT GAT T
BCR_P5-Rev64	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CAG ATA TTG GGC AGC CCT GAT T
BCR_P5-Rev65	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TAG ATA TTG GGC AGC CCT GAT T

Chapter 3

Longitudinal Dynamics of B cell Receptor Repertoires in Controlled Human Malaria Infection

3.1 Introduction

FORTY percent of the world population is at risk of malaria infection and there were an estimated 247 million cases of malaria in 2021 (Q. Liu et al. 2021; WHO 2022). The greatest burden in mortality is caused by the *Plasmodium falciparum* parasite, which resulted in 593 000 deaths in the African Region in 2021, predominantly in children under the age of five (WHO 2022). Furthermore, the morbidity and Disability Adjusted Life Years (DALYs) associated with frequent reinfection have a significant economic impact on working households (Andrade et al. 2022). Severe disease usually occurs in the first few infections, yet immunity to malaria only develops over repeated exposure. Sterilising immunity, as seen in response to many bacterial and viral pathogens, is not achieved

(S. J. Gonzales et al. 2020; Crompton, Moebius, et al. 2014; Langhorne et al. 2008). B cells are known to play a central role in controlling parasitaemia and severe disease, but the mechanisms of acquired B cell immunity to *P. falciparum* are only partially understood (Pérez-Mazliah, Francis M. Ndungu, et al. 2019; Ly and Hansen 2019). This chapter leverages adaptive immune receptor repertoire sequencing to characterise B cell dynamics in a controlled human malaria infection trial across two homologous *P. falciparum* challenges. While BCR repertoire sequencing has been performed on individuals from malaria endemic regions (Ashley E Braddom et al. 2021; Holla et al. 2021; Portugal et al. 2015; Muellenbeck et al. 2013; Zinöcker et al. 2015; Ben S. Wendel et al. 2017) and in vaccinees against a range of malaria antigens (Coelho, Nadakal, et al. 2020; Coelho, Jacob D. Galson, et al. 2022; McNamara et al. 2020) to better understand the responses to this antigenically complex pathogen, BCR repertoires have not been studied in a first and second experimental blood stage malaria challenge in humans before.

3.1.1 Immunity to Malaria

Although *P. falciparum* undergoes multiple life stages in the human host, the skin and liver stages of infection are clinically silent. The symptoms of malaria are driven by asexual parasites as they undergo rounds of replication or sexually mature in the blood. In brief, merozoites infect erythrocytes, and can either produce more merozoites or sexually mature gametocytes. Replicated parasites egress from red blood cells and a new cycle of replication begins, with intermittent fevers corresponding to cyclical release of parasites from infected erythrocytes. The key drivers of pathogenesis in malaria are thought to be sequestration of infected red blood cells and a systemic inflammatory response. Infected red blood cells adhere to the vascular endothelium, a mechanism to avoid clearance of parasites by the spleen, causing obstruction, ischaemia and inflammation in severe disease and can lead to cerebral malaria which is often lethal (Crompton, Moebius, et al. 2014;

Pérez-Mazliah, Francis M. Ndungu, et al. 2019). During blood-stage infection parasite GPI-anchors, DNA, and hemozoin are all thought to activate pattern recognition receptors including toll like receptors. This triggers a strong and conserved pro-inflammatory interferon-γ response, involving Natural Killer cells, monocytes and $\gamma\delta$T cells (Pohl and Cockburn 2022; Scholzen and Sauerwein 2016). Malaria infection was used historically to treat syphilis and examination of historical studies provided early evidence that severe disease usually only occurs within the first exposures (Jeffery and W. E. Collins 1999). Studies of both endemic populations and serial human experimental infections have shown that blood-stage infections develop into clinical malaria over many repeat exposures, manifesting as a mild fever, or no symptoms at all, while parasites replicate in the blood. This is thought to be a form of tolerance to Plasmodium-induced inflammation. Passive immunisation studies from the 1960s, whereby children with severe malaria were given antibody from immune adults, demonstrated rapid reductions in parasitemia and fever (Cohen, McGregor, and Carrington 1961), highlighting the importance of humoral responses in protection against disease. However, while antibody responses in malaria are protective, they are reportedly short lived (Crompton, Moebius, et al. 2014). The reasons for rapid waning and inconsistent boosting of protective antibodies are still poorly understood and have hampered the design of effective vaccines that induce protective and long-lasting immunity.

3.1.2 The Role of B cells in Immunity to Malaria

The B cell responses to *P. falciparum* infection in individuals living in malaria endemic areas are distinguished by two main characteristics: inefficiency in producing long-lived plasma cells and classical memory B cells (cMBCs), and the large expansion of atypical memory B cells (atMBCs).

Generation of Antibody Responses in Malaria

P. falciparum has more than 5,000 proteins, many of which are used for host-parasite interactions and immune escape. For example, the parasite can switch variable surface antigens, resulting in many potential antigenic targets for antibodies (M. J. Gardner et al. 2002; Rénia and Goh 2016). Antibody levels against extracellular, plasma membrane proteins, highly abundant parasite proteins, and those that lack human orthologs typically elicit stronger responses (Yaohui Liu et al. 2018). Protective antibody responses mainly target antigens that are expressed on the surface of merozoites such as Merozoite Surface Protein-1 (MSP1) and Apical Membrane Antigen 1 (AMA1), and variable surface antigens, such as *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1), that are parasite receptors expressed on the surface of infected erythrocytes mediating cytoadherence and rosetting of infected red blood cells (S. J. Gonzales et al. 2020). One leading theory is that humoral responses protect against severe disease by generating antibodies against virulent variants of parasite variable genes, including those which mediate rosetting by PfEMP1 (J.-A. Chan et al. 2019; Cavanagh et al. 2004; Ofori et al. 2002). Additionally, antibodies reduce malaria pathogenesis by preventing merozoites from invading red blood cells by neutralization, opsonisation and activating complement to recruit phagocytes for Fc receptor and complement mediated phagocytosis (Ly and Hansen 2019). Observations from malaria-endemic countries suggest that parasite-specific antibody titres wane rapidly to low or undetectable levels within months or even weeks after infection, despite high initial titres but increase with age and transmission intensity (Akpogheneta et al. 2008; Kinyanjui et al. 2007; Crompton, Kayala, et al. 2010; Yman et al. 2019). Antibody secreting cells were detected in a paediatric cohort in Uganda immediately following acute malaria and were increased after a second clinical infection, but the half-life of these antibodies was short, between 2-10 days (Michael T. White et al. 2014). One theory is that short-lived plasma cell responses consist mainly of IgM-secreting

plasma cells. These short-lived plasma cells are hypothesised to arise from extra-follicular marginal zone B cells outside of germinal centres, or from un-switched memory B cells within germinal centres (Ly and Hansen 2019).

Atypical Memory B cells in Malaria

Atypical memory B cells are found to be consistently elevated in individuals from malaria-endemic countries compared to malaria naïve individuals (Portugal et al. 2015; Ashley E. Braddom et al. 2020; Pérez-Mazliah, P. J. Gardner, et al. 2018; Greta E. Weiss, Crompton, et al. 2009). While central memory B cells (cMBCs) are generated upon malaria infection and appear to be long-lived, their reported prevalence is low among adults (30-50%)(Portugal et al. 2015; Michael T. White et al. 2014; Francis Maina Ndungu et al. 2012; Greta E. Weiss, Traore, et al. 2010). For example, age-matched children in rural Kenya who were exposed to *P. falciparum* displayed expansion of atypical MBC compared to children in unexposed communities (Illingworth et al. 2013). The broadest classification of atMBCs includes B cells expressing CD19+CD21−CD27− (often with high expression of CD19 and CD20). atMBCs also frequently additionally express Tbet and FcRL5(Obeng-Adjei et al. 2017; Portugal et al. 2015; Sutton et al. 2021). Similar B cell subsets which are CD19+CD21−CD27− have been described in other chronic infections such as HIV (Moir et al. 2008) and are associated with autoimmune disease (S. A. Jenks et al. 2018), however, these cells have also been suggested to play a part in normal vaccine responses (Sutton et al. 2021). Single cell RNA seq has demonstrated that atMBCs in malaria transcriptionally resemble autoimmune-associated Double Negative 2 (DN2) B cells (Holla et al. 2021). DN2 cells defined as CD19+IgD−CD27−CD21−CD11c+T-bet+CXCR5−, have been observed in SLE, Sjogrens syndrome and rheumatoid arthritis and can differentiate into auto-antibody secreting cells (S. A. Jenks et al. 2018). There are several developmental pathways which appear to give rise to atypical memory B cells

(**Figure 1**). These include, IFN-γ signalling, TLR engagement (TLR7/9 in particular in malaria), BCR cross-linking, and Th1-polarised germinal centre responses. IFN-γ is an important cytokine in the induction of atMBCs.

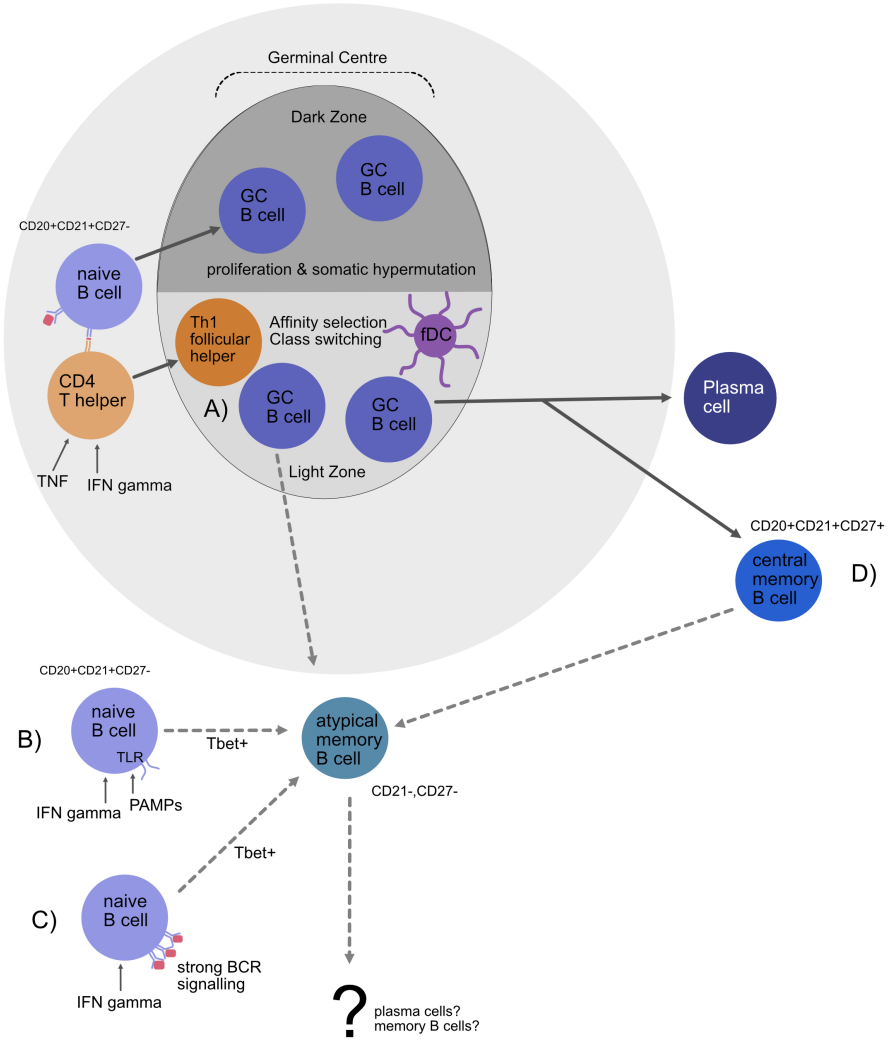


Figure 1: Graphical Summary, Atypical Memory B cells. Cartoon adapted from Braddom *et al* 2020 *"Potential functions of atypical memory B cells in Plasmodium-exposed individuals"*. Atypical memory B cells can arise **A)** as a result of impaired T cell help from T follicular helper cells during Germinal Centre reactions. In response to IFN-gamma and TNF signalling, T follicular helper cell precursors polarise towards a Th1 phenotype and result in impaired T cell help in GC reactions. **B)** PAMPs, such as parasite DNA released during merozoite rupture, and IFN-gamma signalling can signal vial TLRs and other pattern recognition receptors on naive B cells and direct them towards an atMBC phenotype **C)** IFN-gamma signalling along with BCR cross-linking on naive B cells can induce polarisation towards Tbet+ atMBCs. **D)** cMBCs can also give rise to atMBCs particularly IgG+ atMBCs.

In Holla et al. 2021, they performed scRNAseq of total B cells from three malaria-exposed Malian adults and three unexposed, healthy U.S. adults. Malaria-exposed individuals had a cluster of expanded atMBCs which were transcriptionally distinct and were not present in unexposed individuals. Trajectory analysis revealed that these cells had differentiated from naive B cells and were modulated by IFN-γ signaling. Furthermore, when the B cell clusters were compared to a publicly available dataset of B cells treated in vitro with IFN-γ, the atMBC cluster fitted the transcriptional signature best. In agreement with this, Obeng and colleagues found that atMBCs could be induced *in vitro* upon stimulating tonsillar or peripheral naive B cells with the supernatant from PBMCs co-cultured with malaria-infected red blood cells and stimulating these B cells with anti-IgM agonists. When they neutralised IFN-γ in the supernatant, or blocked IFN-γ receptors on the cells, the cells no longer differentiated into Tbet-high atMBCs (Obeng-Adjei et al. 2017). IFN-γ has also been shown to polarise T follicular helper cells towards a Th1 phenotype which provides less effective T cell help during germinal centre reactions. Ryg-Cornejo and colleagues demonstrated in a mouse model of malaria, that infection induces high frequency of Th1 polarised T follicular helper cell precursors (which were also T-bet high). Blocking TNF and IFN-γ, or deleting T-bet, rescued normal Tfh differentiation and resulted in germinal centre responses to malaria (Ryg-Cornejo et al. 2016). Furthermore, atMBCs which express T-bet have been reported to have reduced BCR signalling relative to naïve and cMBCs in malaria. Differentiation of atMBCs can occur from cMBCs and atMBCs have been shown to be anergic to activation by BCR signalling (Portugal et al. 2015), suggesting this may represent a less antigen-responsive population. Conversely, cloning and expressing BCRs from atMBCs demonstrated that BCRs from atMBCs can target *P. falciparum* antigens, and these same BCRs were also identified as secreted antibodies by Mass Spectrometry (Muellenbeck et al. 2013). While atMBCs are undoubtedly associated with malaria infection, it is unclear whether atMBCs

represent a dysfunctional B cell subset or are part of a normal adaptive response to an inflammatory challenge (Sutton et al. 2021).

3.1.3 Controlled Human Malaria Infection as a Model to Study Malaria Immunity

Murine models of malaria have been widely used to understand tissue and organ-wide immune responses across different stages of infection, with different parasite strains and co-infections with other pathogens (Wykes and M. F. Good 2009; Scholzen and Sauerwein 2016). However, findings from mice are not necessarily translatable to humans. This was the case with a TCR repertoire study from the Cowan lab which identified a conserved V gene signature in response to *P. chabaudii* infection (Natasha L. Smith et al. 2020), but no such public TCR signatures were observed in humans. While many studies to characterise malaria immune responses in humans are conducted on individuals living in endemic areas, it is difficult to obtain samples from subjects prior to malaria exposure, particularly because children are often exposed from a very early age and are transiently protected by maternal antibodies in the first few months of life. Using Controlled Human Malaria Infections as a study system allows control of infection and sampling times, including sampling pre-challenge timepoints, and longitudinal follow-up at timepoints of interest post infection. Detailed knowledge of volunteers' previous exposure history and comorbidities can help unpick heterogeneous responses to infection (Scholzen and Sauerwein 2016). Finally, infecting volunteers with a known strain at a standardised inoculum reduces some of the variation which would be encountered in field studies. The Spence Lab have performed extensive phenotyping of CHMI volunteers enrolled in vaccine studies at the Jenner Institute and this has produced a rich dataset on the immune responses to a primary malaria infection and subsequent re-challenges.

3.2 Aims

This chapter contributes to the Spence Lab's wider effort to understand how malaria develops in the first few infections of life. Here we characterise BCR repertoires longitudinally in CHMI. Specifically, we set out to address the following questions:

- How does the BCR repertoire differ between a first and second malaria challenge?
- Do we observe evidence of BCR memory boosting upon re-challenge?
- What signatures of affinity maturation do we observe in response to malaria challenge?

3.3 Results

3.3.1 CHMI Study Design and Sample Characteristics

This experiment made use of PBMCs from volunteers undergoing controlled human malaria infections (CHMI) at the Jenner Institute (Oxford) which were obtained from a collaboration with the Spence Lab at the University of Edinburgh. Volunteers were infected with parasitised erythrocytes from a single volunteer previously infected with 3D7 *P. falciparum*. The parasites were from a recently mosquito-transmitted line (< 3 blood cycles from liver egress), since vector transmission attenuates parasite growth and pathology in the mammalian host (Spence et al. 2013). PBMCs were sampled a day prior to infection ("challenge -1") and at three key timepoints post infection: day of diagnosis, 28 days post challenge and 90 days post challenge, for a homologous primary and secondary malaria infection **(Figure 2)**. Treatment with anti-malarial drugs was initiated once participants had more than 10,000 parasites/ml of blood, or if they had more than 5,000 parasites/ml of blood and symptoms consistent with malaria infection. If

they had less than 5,000 parasites/ml of blood and symptoms of malaria, treatment was not yet initiated. The day treatment was initiated is recorded as "day of diagnosis" in this study. The same PBMC samples used to perform the BCR repertoire analysis described in this chapter, were also used to interrogate TCR repertoire dynamics in these volunteers (see N. L. Smith 2022).

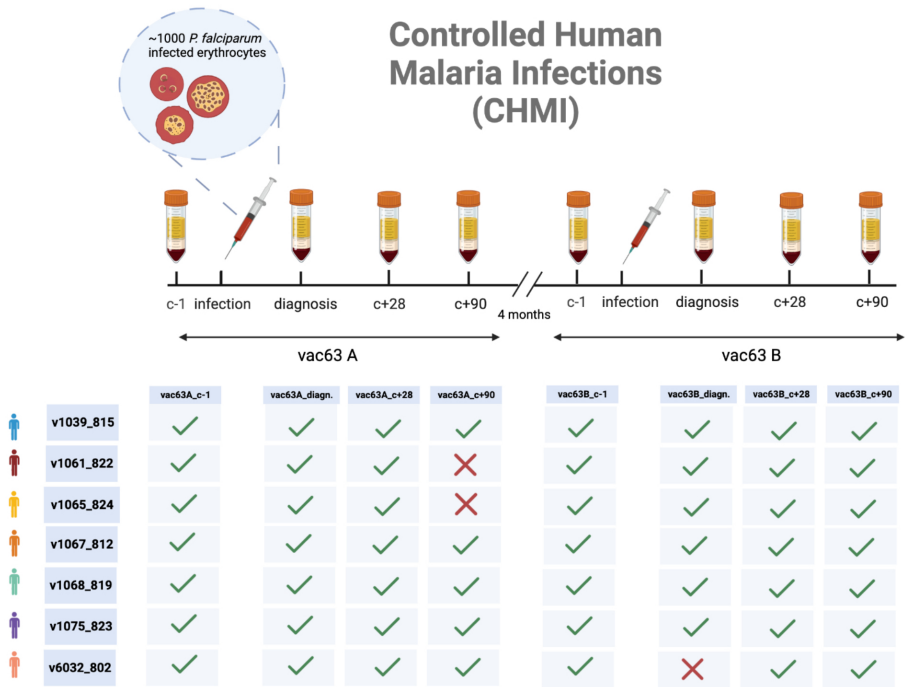


Figure 2: Overview of Study Design. Seven malaria-naive individuals were given two homologous challenges with *P. falciparum* (3D7 strain) as part of the placebo arm of the RH5.1/AS01 vaccine trial and a comprehensive phenotyping study conducted by the Spence Lab. Each individual was sampled at four timepoints in each infection: a day before challenge (c-1), the day they were diagnosed (diagnosis), day 28 post challenge (c+28) and day 90 post challenge (c+90). Samples were taken at these timepoints for a first (vac63A) and second (vac63B) challenge with a four month gap between trials. PBMCs were isolated from each blood sample and used for RNA isolation and subsequent BCR repertoire sequencing. Red crosses indicate samples that dropped out.

Table 3.1: CHMI Volunteer Characteristics*

Volunteer	Age(years)	BMI	Smoking	CMV	EBV	Ethnicity	Treatment	Gender	PMR
v1039_815	32	26.2	N	Positive	Positive	White Portuguese	Riamet	F	8.6
v1061_822	24	25.2	Y	Negative	Positive	White British	Riamet	M	10.87
v1065_824	20	20.8	N	Positive	Not tested	White British	Riamet	M	11.29
v1067_812	21	26.7	N	Negative	Not tested	White British	Malarone	M	9.01
v1068_819	33	25.4	N	Positive	Positive	South Korean	Riamet	M	13.15
v1075_823	22	33.4	N	Negative	Positive	White British	Riamet	M	6.1
v6032_802	22	19.5	N	Negative	Positive	White British	Riamet	M	9.61

* Abbreviations: BMI = Body mass index, CMV = Cytomegalovirus status, EBV = Epstein-Barr Virus status, PMR = parasite multiplication rate

3.3.2 Lymphopenia at Day of Diagnosis

The Spence Lab have previously reported lymphopenia at day of diagnosis in these volunteers (Sandoval et al. 2021) (**Figure 3A**). Lymphocyte data was obtained from full blood counts which were collected for purposes of monitoring the infections. Relative to the pre-challenge baseline (c-1) lymphocyte counts dropped to lower levels at day of diagnosis in the second infection than in the first (**Figure 2B**).

3.3.3 Sequencing Data Generation and Quality Control

BCR libraries were generated as described in detail in Chapter 2. Instead of using T cell depleted cell pellets as in the previous chapter, here RNA extraction was performed from PBMCs. Library preps were performed in replicate for each sample. cDNA synthesis was performed using a 5' RACE strategy, UMIs of three different lengths and IgM and IgG specific constant region primers (**Figure 4A**). cDNA libraries underwent two rounds of PCR amplification. Each sample had a P5 and P7 index. We performed two rounds of sequencing on the Illumina Miseq V3 platform. After the first run, the level of UMI coverage was assessed by generating rarefaction curves of unique UMIs by read depth

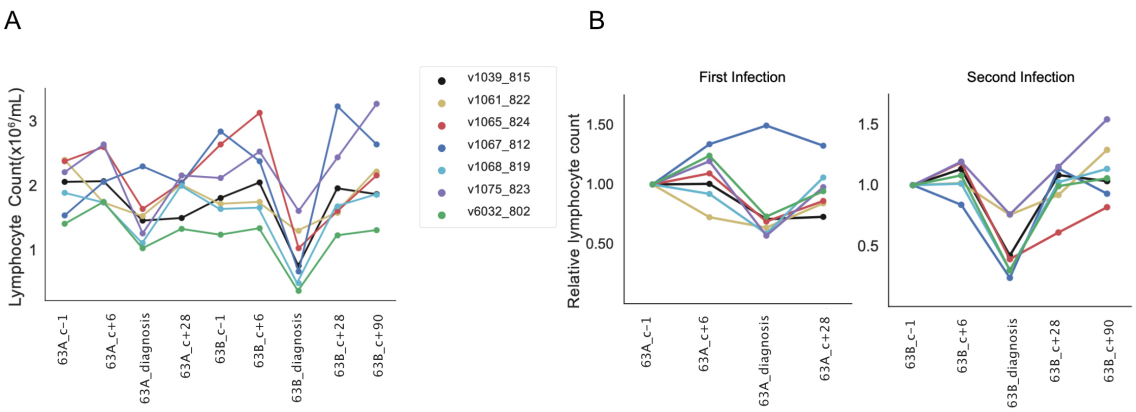


Figure 3: Lymphopenia at day of diagnosis. A) Lymphocyte counts from full blood counts across key timepoints in the first and second infection. **B)** Lymphocyte counts relative to c-1 (both plots are from Natasha Smith's thesis, see Smith 2022, "*Decoding malaria T-cell responses using adaptive immune receptor repertoire sequencing*").

for a random selection of libraries with low, medium and high numbers of reads (**Figure 4B,C,D**). This revealed that UMI coverage had not reached saturation, so a second round of sequencing was performed which yielded additional depth for each sample (**Figure 4E**). UMIs were the anticipated lengths, with 15 million unique UMIs of the correct lengths identified and one million reads where no template switch adapter was matched (**Figure 4F**). FastQC analysis also demonstrated good quality (Phred score $>$ 20) throughout most of read 1 and read 2, with read quality dropping towards the end of read 1 as expected (**Supplementary 1**).

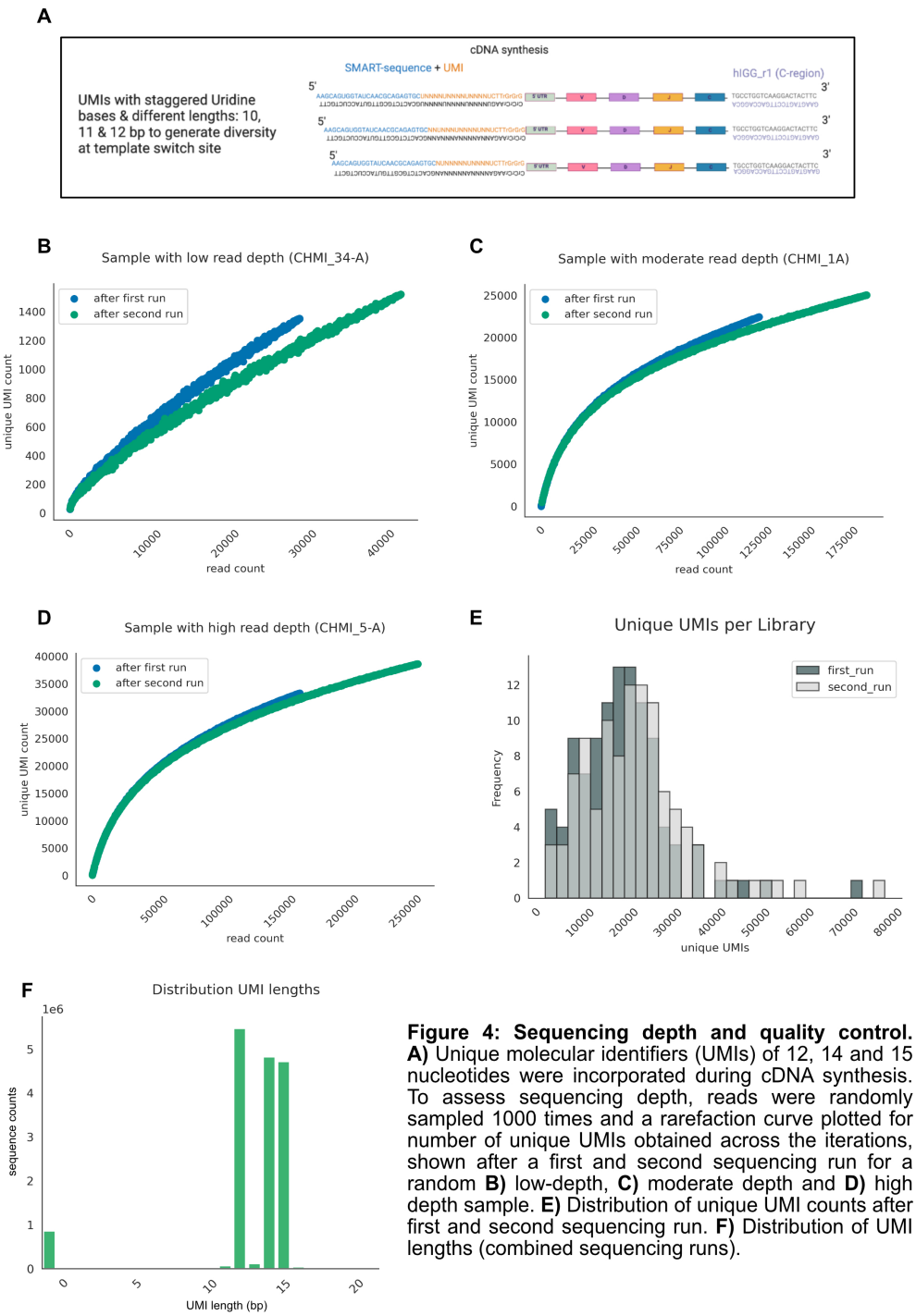


Figure 4: Sequencing depth and quality control. A) Unique molecular identifiers (UMIs) of 12, 14 and 15 nucleotides were incorporated during cDNA synthesis. To assess sequencing depth, reads were randomly sampled 1000 times and a rarefaction curve plotted for number of unique UMIs obtained across the iterations, shown after a first and second sequencing run for a random **B)** low-depth, **C)** moderate depth and **D)** high depth sample. **E)** Distribution of unique UMI counts after first and second sequencing run. **F)** Distribution of UMI lengths (combined sequencing runs).

3.3.4 Repertoire Overview

The sequenced libraries we obtained spanned a range of UMI sampling depths (smallest 1256 to largest 52430 UMIs)**(Figure 5A)**. To confirm that technical replicates were consistent between samples and to check for barcoding errors, we wanted to assess repertoire similarity between technical replicates and across timepoints in a given volunteer. The composition of unique V genes found in a repertoire, should be consistent within an individual. We used the Manhattan Distance to compare repertoire similarity based on V gene composition between all of the libraries. Pairwise distances were calculated and clustered using UPGMA hierarchical clustering (**Figure 5B**). The resulting dendrogram illustrates, firstly, that technical replicate clustered together in pairs, and secondly that libraries clustered by individual, with the exception of the two libraries with lowest UMI depths - CHMI 34 (corresponds to v1065_824, time-point 63A_c+28) and CHMI 98 (corresponds to v6032_802, timepoint 63B_c+90) which were both outliers. However, the technical replicates were still consistent. Because a threshold for UMI sampling depth was not determined a-priori, and the lowest-depth repertoire still contained >1800 unique UMIs when combining technical replicates, all samples were included in the analysis, with the exception of diversity analyses which are sensitive to sampling depth. Data pre-processing steps (pRESTO and changeO) were performed using the Immcantation framework and were performed separately for each library, and technical replicate libraries were combined prior to analysis.

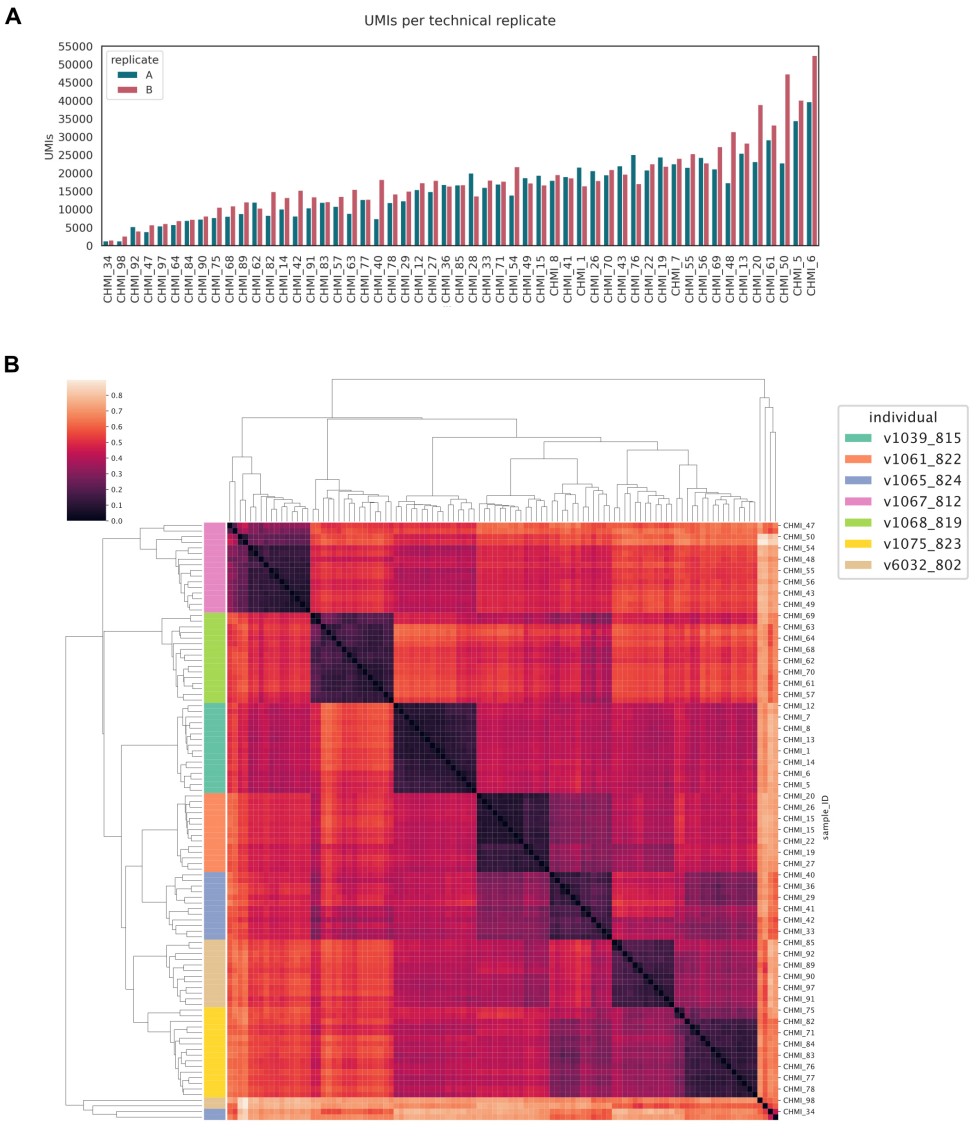


Figure 5: Overview of BCR Libraries. A) Counts of unique UMIs obtained per library and technical replicate. **B)** Pairwise Manhattan Distance between all BCR libraries was calculated based on V gene proportion using singleton CDR3s and plotted as a cluster map (colour bar indicates individual).

3.3.5 V Gene Usage is Individual-Specific

Changes in V gene usage upon infection can be indicative of an antigen-specific immune response, and particular V genes can be associated with protective of pathogenic responses. We therefore investigated whether *P. falciparum* infection was associated with any changes in V gene usage in our samples. We first wanted to examine whether malaria infection resulted in convergent V gene usage between individuals at the post-infection timepoints. To test this, the proportion of the repertoire using a given V gene was calculated for each repertoire. Principal Component Analysis (PCA) appeared to separate the repertoires mainly by individual, although some of the clusters overlapped (**Figure 6A**). Clustering V gene repertoires using UMAP, which can capture non-linear relationships, clustered the repertoires clearly by individual (**Figure 6B**), with the exception of volunteer "v1065_824", whose repertoire split into two halves. Therefore, any malaria-specific V gene signature did not skew the repertoires enough to overcome the individual-specific V gene signature. This is unsurprising, given both the heterogeneity of immune repertoires, different MHC haplotypes represented in the sample set, the volunteers' potentially diverse pathogen exposure histories and the fact that most activated B cells would likely home to secondary lymphoid tissues rather than remain in the blood. To compare whether V gene usage differed between the same timepoints in the first and second infection, V gene proportions were log transformed and mean log 10 proportions compared between first and second infections (**Figure 7**). While V gene usage between first and second infection was tightly correlated at all timepoints (R-squared values >0.9), several moderately-abundant V genes were differentially used in second infection at day of diagnosis (**Figure 7B**) and at c+28 (**Figure 7C**). The challenge +90 timepoint (**Figure 7D**) looked the most distinct between first and second infection and had the weakest correlation between first and second challenge of the four timepoints (r-squared 0.9), suggesting that V gene usage did not return to the same baseline after the second infection.

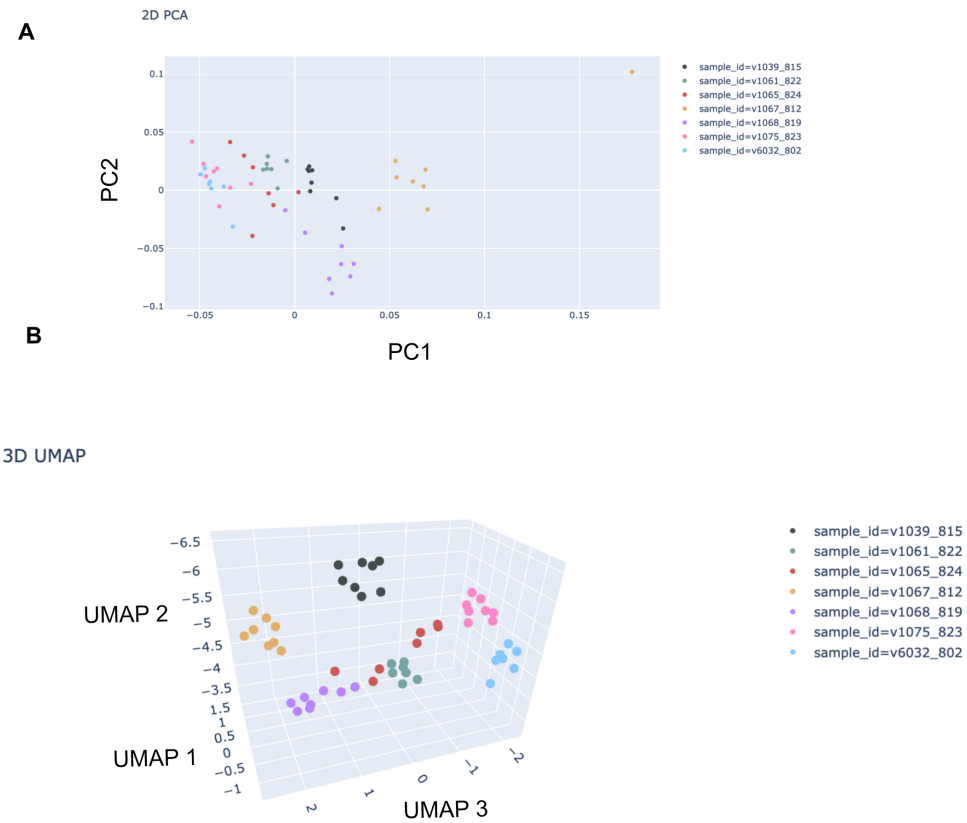


Figure 6: V Gene Usage Clusters by Individual. A) A principal component analysis (PCA) was performed using V gene proportions for each sample as an input, coloured by volunteer **B)** The same data was clustered using a 3D UMAP.

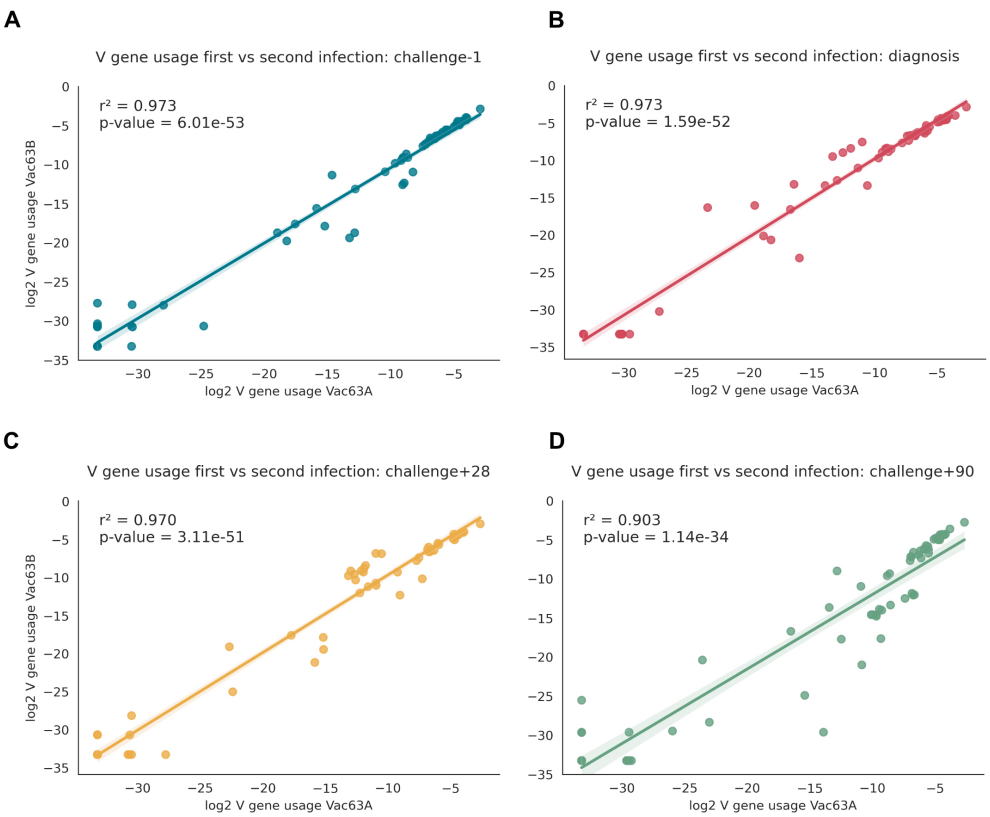


Figure 7: V Gene Usage by Infection. Mean V gene usage for each V gene was calculated for the first and second infection (subsetted to the individuals available at each timepoint to account for samples that dropped out) and Pearson correlations of log2 mean V gene usage were calculated for each comparison. **A)** Challenge -1 **B**) day of diagnosis, **C)** challenge + 28, and **D)** challenge + 90 in the first versus second infection.

3.3.6 V Gene Usage in IgM and IgG

With the maturation of an adaptive response, we would expect to initially observe changes in the IgM repertoires after the first infection, followed by changes to the IgG repertoires in the second infection. We therefore decided to examine V gene usage in each volunteer's repertoire for IgM and IgG separately. V gene usage for each isotype was plotted on a heatmap and the rows ordered by timepoint. We would expect the IgM repertoires to represent mostly naive B cells, containing some IgM memory B cells, and all of the IgG repertoires to be antigen-experienced. There were no obvious differences in V gene usage in the IgG repertoires (**Figure 8**). In IgM we detected a slight increase in IGHV3-7 gene usage at the timepoint immediately following the first malaria exposure- the day of diagnosis, in the first infection only.

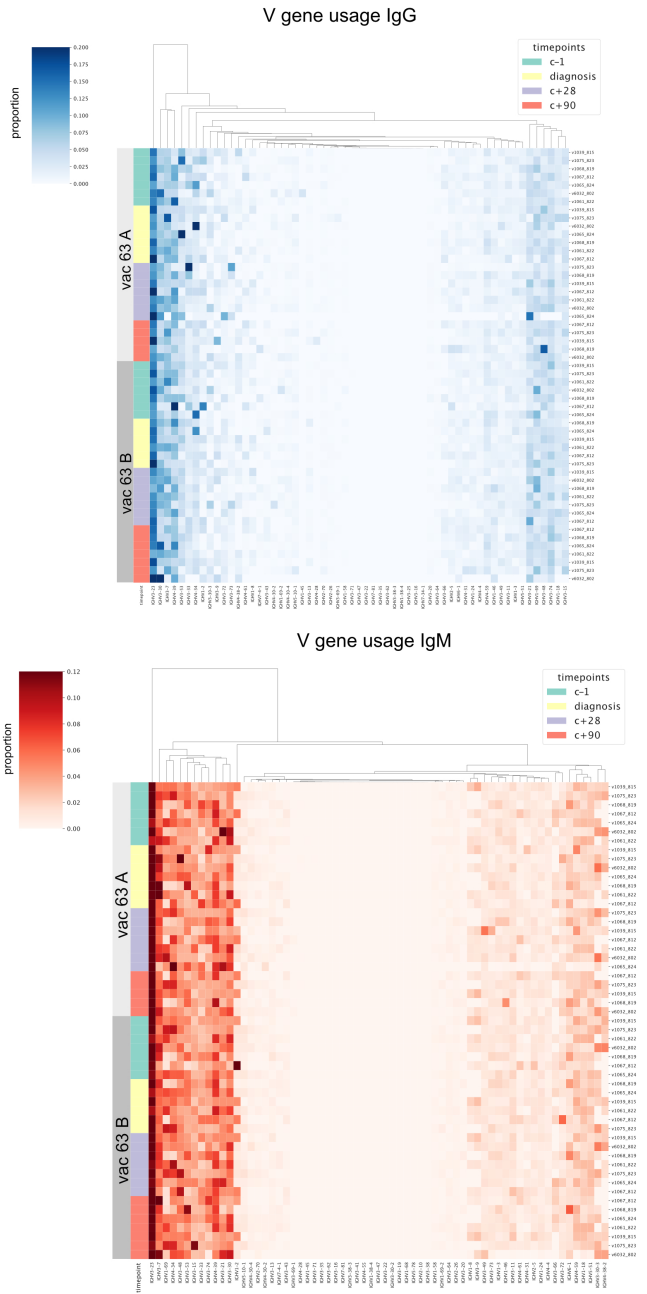


Figure 8: V Gene Usage by Isotype. Repertoires were subsetted to **A)** IgG or **B)** IgM BCRs and heatmap clustered by columns (V genes) was generated for V gene usage (each column is one V gene, each row is one individual's repertoire at a specific timepoint). Vertical colour bar indicates timepoint for each infection, infection indicated by light and dark grey colour bars.

3.3.7 IgM IGHV3-7 Usage Increases at Day of Diagnosis in the First Challenge

To test whether IGHV3-7 was increased at day of diagnosis in the first challenge, we first examined usage of the IGHV3 family more generally and did not observe any other V genes which increased post-challenge (**Figure 9A**). Change in IGHV3-7 gene usage in IgM repertoires relative to baseline (challenge-1) was plotted by individual and tested across time-points using a mixed effect model with the effect of individual treated as random (**Figure 9B**). Six out of seven volunteers displayed an increase of IGHV3-7 at day of diagnosis in the first infection and this result was statistically significant (p-value: 0.015), but this significance would be lost if multiple testing correction were applied. After day of diagnosis, IGHV3-7 usage returned to baseline or lower levels across all subsequent timepoints except in two individuals where this initial expansion was followed by an increase in IGHV3-7 usage at the final timepoint (63B_challenge + 90). While this conserved V gene signature could potentially indicate a very early malaria-specific B cell response, the specificity of BCRs using IGHV3-7 at this timepoint cannot be inferred from this data.

3.3.8 Distinct Diversity Profiles in First and Second Infection

Next we wanted to investigate whether repertoire diversity or clonality were impacted by malaria infection. An overview of clonotype distributions for each repertoire was plotted displaying the proportion the top 1, 10, 100 and 1000 CDR3s made up of each repertoire (subsampled to 1000 UMIs)**(Figure 10A)**. This demonstrated that there was substantial heterogeneity of clonal distributions among participants, and even pre-infection repertoires contained expanded clonotypes in some volunteers. Repertoires from day of diagnosis in the first infection and c+90 in the second infection contained multiple dominant clonal

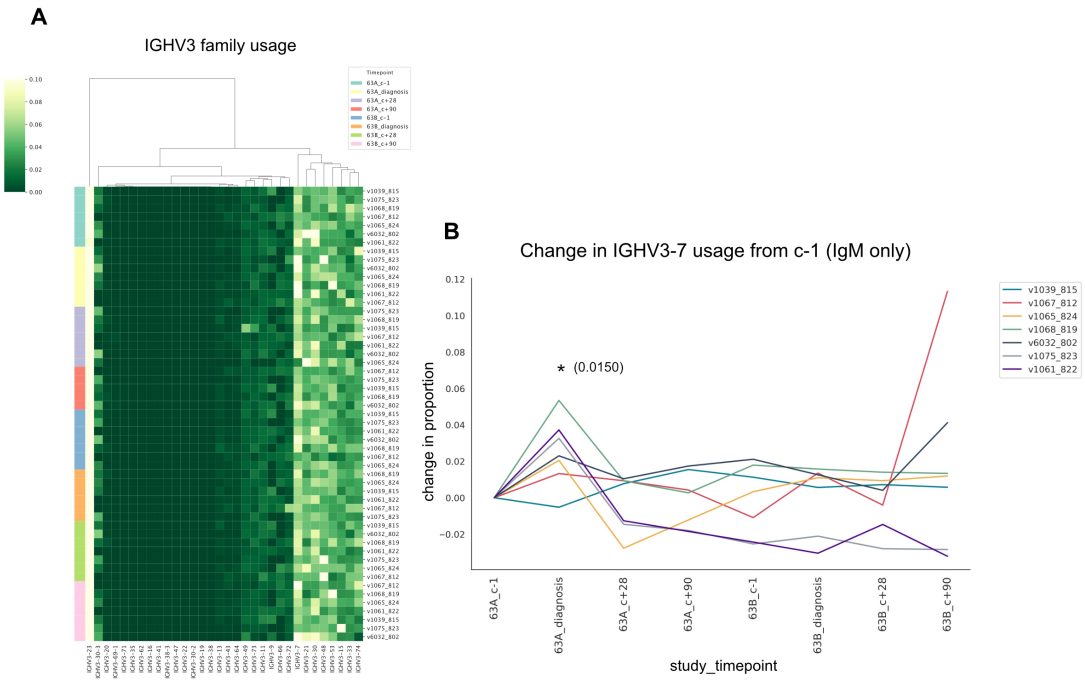


Figure 9: IGHV3-7 Usage Increased in IgM. A) Heatmap of IgM V genes were subsetted to the IGHV3 family. **B)** Changes in IGHV3-7 proportions relative to the 63A_c-1 timepoint were plotted. glmmTMB was used to analyse change in IGHV3-7 usage, with study_timepoint as a fixed effect and the effect of individual treated as random. P-value was not corrected for multiple testing.

expansions in several volunteers. A range of diversity and clonality indices have been applied to AIRR-seq data to capture changes in clonotype distributions, some of the most common ones being Simpson's Diversity, Shannon Index and the Gini Index of Inequality. However, these single metrics are biased either towards the total number of unique BCRs or the distribution of each BCR (Chiffelle et al. 2020). Renyi Entropy has been proposed as a useful metric of diversity to capture a more complete profile of the diversity composition of the repertoire (Chao and Jost 2015). Renyi Entropy at different alpha values reflect a sample's Species Richness (alpha ==0), Shannon Index (alpha ==1), Simpson's Diversity (alpha ==2) and the dominance index- the Berger Parker Index (alpha approaching infinity). The higher the Renyi Entropy, the more diverse the repertoire and as the value of alpha increases, Renyi Entropy is more heavily weighted towards the most abundant clonotypes in the repertoire. We decided to perform the diversity analyses on BCRs clustered into "clonotypes" to account for the fact that somatic hypermutation might result in related BCRs which are not identical. We clustered clonotypes based on V-J gene usage, CDR3 length and a threshold of a 0.15 amino acid edit difference between CDR3s. We generated Renyi Entropy curves for the clonotype distributions in each repertoire by downsampling all of the repertoires to 1500 UMIs over 1000 iterations and averaged the Renyi Entropy values across the iterations. To visualise the curves, we plotted the average Renyi Entropy curves for each timepoint (averaged across individuals) with 95% confidence intervals determined by bootstrapping. We observed that in the first infection the diagnosis and c+28 timepoints were, on average, more clonal than the c-1 and c+90 repertoires, but the 95% confidence intervals overlapped substantially (**Figure 10B**). In the second infection, however, the trend was reversed: The c+28 and day of diagnosis curves had, on average, higher Renyi Entropy than the c-1 and c+90 curves, suggesting repertoires were more diverse at those timepoints and the confidence interval for the average Renyi Entropy curve at day of diagnosis did not overlap with that of any other

timepoint, suggesting it was more diverse than all other timepoints (**Figure 10C**).

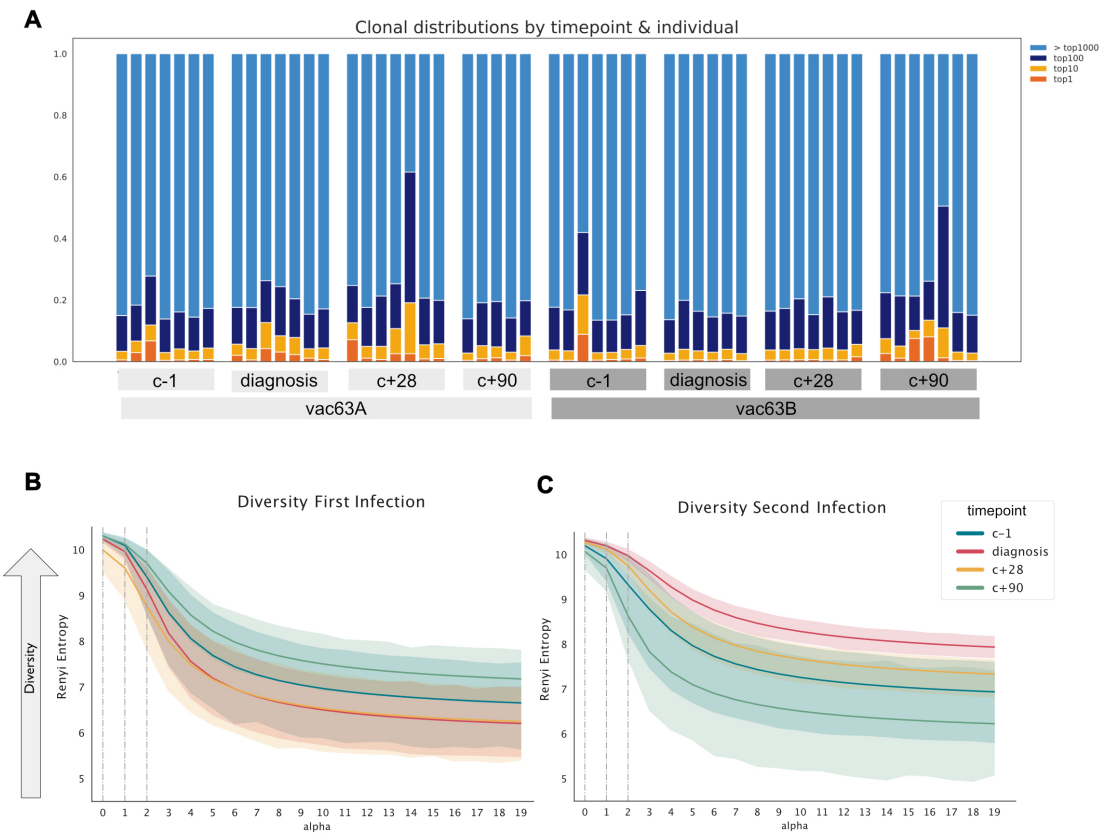


Figure 10: Clonality overview. A) Proportion of the repertoire that the top 1, 10, 100 and >1000 CDR3s make up of each volunteer's repertoire at each timepoint for the first and second infection (each vertical bar represents one repertoire). **B)** Renyi Entropy curves were generated by subsampling each repertoire to 1500 UMIs at random, calculating Renyi Entropy for alpha 0-20 and repeated over 1000 iterations. Values obtained across the alpha range were averaged across iterations for each sample. Renyi Entropy curves for all individuals were plotted and averaged for each timepoint in the first infection and **C)** in the second infection. Shaded areas represent 95% confidence intervals for mean Renyi Entropy. Dotted lines represent key alpha values which relate to Species Richness (alpha =0), Shannon Index (alpha=1), Simpson Diversity (alpha =2).

Comparing the timepoints specifically between the first and second infection, we found that at c-1 and c+90 the average Renyi Entropy was similar between first and second infection, with wide and overlapping confidence intervals, suggesting individual repertoires varied substantially in their diversity (**Figure 11A,D**). However, at day of diagnosis repertoires were strikingly diverse upon re-challenge compared to the first infection and

the 95% confidence intervals did not overlap at alpha values larger than 1, which reflects the Shannon Index (**Figure 11B**). This trend of increased diversity in repertoires sampled during the second challenge was also observed at challenge +28 compared to the first infection, but the confidence intervals overlapped slightly at the higher alpha values (**Figure 11C**).

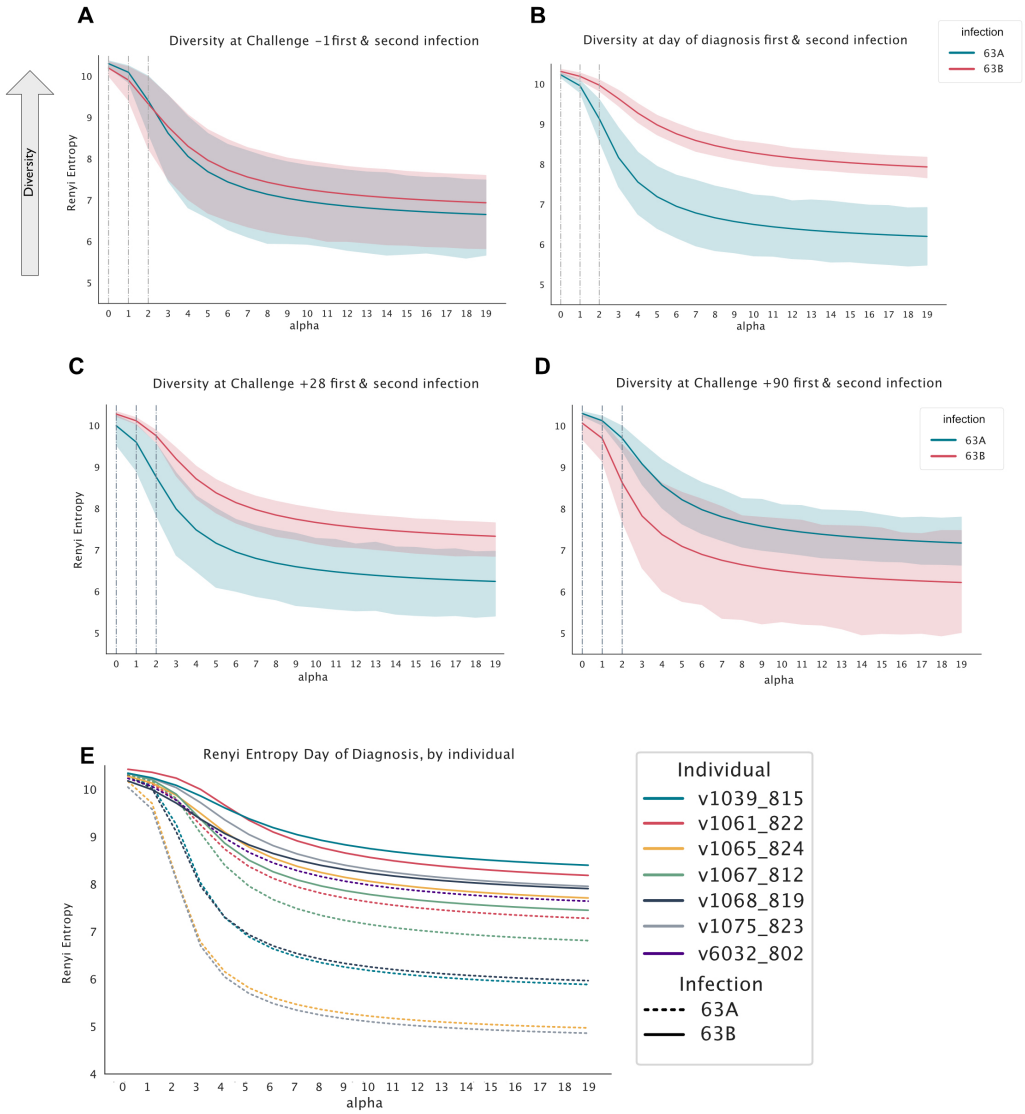


Figure 11 : Diversity profiles differ in the first and second infection. A) Repertoires were subsampled to 1500 UMIs over 1000 iterations and Renyi Entropy calculated for alpha 0-20. Values were averaged across iterations for each sample. Renyi Entropy curves were plotted comparing each timepoint for the first and second infections: A**)** challenge -1, **B)** diagnosis, **C)** challenge +28 and **D)** challenge + 90. Shaded areas represent 95% confidence intervals for mean Renyi Entropy. Dotted lines represent key alpha values which relate to Species Richness (alpha =0), Shannon Index (alpha=1), Simpson Diversity (alpha =2). **E)** Individual Renyi Entropy Curves coloured by individual, at day of diagnosis.

3.3.9 Clonal Expansion of IgM Repertoires in First Infection

In a re-challenge model of infection the prediction would be that the first infection is dominated by an IgM response, and upon re-challenge, more expansion of IgG BCRs is observed. To examine clonal dynamics in more detail, we calculated diversity metrics for clonotypes from IgM and IgG repertoires separately (**Figure 12**). Each sample was split into IgM and IgG repertoire based on constant region calls, and down-sampled to 700 UMIs. Two repertoires had very low numbers of IgG clonotypes, and those samples were excluded from this particular analysis. Simpson Diversity, Shannon Index and Gini Index, were calculated, again iteratively re-sampling the repertoire 1000 times and averaging across iterations. Treemaps for IgM (**Supplementary Figure 2**) and IgG (**Supplementary Figure 3**) repertoires were generated for all timepoints and volunteers to examine clonal composition of IgM and IgG repertoires. 700 UMIs were randomly sampled from each repertoire, tiles were scaled according to the proportion of the repertoire made up by that CDR3 and coloured by V gene. IgG repertoires had larger clonal expansions than IgM and were diverse regarding V gene usage, but did not display any particular trends. IgM repertoires appeared to show very modest oligoclonal expansion after the first challenge, but remained very diverse upon re-challenge. In the first infection, IgM repertoires underwent modest clonal expansion at day of diagnosis and day 28 post challenge, with most repertoires returning to a polyclonal distribution similar to pre-challenge by day 90 post first challenge. In the second infection, the kinetics appear quite different in IgM: There are no apparent changes in clonality at day of diagnosis or challenge + 28 **(Figure 12A,B,C)**. Treemaps of CDR3 distributions for all volunteers at selective timepoints support the finding that modest oligoclonal expansions occur at day of diagnosis and c+28 in the first challenge, and c+90 in the second challenge **(Figure 12D)**, however, there is also substantial heterogeneity among volunteers. Of note, the challenge +90 timepoint in the second infection seemed to show unexpected clonality relative to the earlier timepoints.

Only the Gini Index was statistically significant, but the trend of modest clonal expansion in IgM after the first challenge was consistent between different diversity indices. The Gini Index of Inequality is sensitive to small differences in clonality, as it quantifies the degree to which the distribution deviates from a perfectly even distribution, so would be the most suitable index to detect oligoclonal expansions.

We hypothesised that modest clonal expansion in the first challenge but not the second may be due to a shift from an IgM to an IgG B cell response, and subsequently investigated the same metrics in the IgG BCRs **(Figure 12 E,F,G)**. In IgG repertoires only the Simpson's Diversity Index was statistically significantly different at c+28 in the first infection, however, this result looks like it may have been impacted by one outlier sample at that timepoint **(Figure 12H)** and the trend was not observed using any of the other diversity metrics. Taken together with the previous results, we observe an IgM-mediated response to primary *P. falciparum* infection, no evidence of clonal expansion or V gene skew in the IgG repertoires.

3.3.10 Clonotypes Expanding in First Infection Do Not Recur Upon Re-Challenge

Next we wanted to know whether clonotypes which had expanded in the first infection were being boosted in a second infection, even if overall the repertoires looked more diverse in the second infection. To gain an overview of whether expanded clones were persisting or being boosted upon re-challenge, we selected the top 25 most abundant clonotypes for each timepoint in each volunteer, and examined whether they were found at any subsequent timepoints **(Figure 13A-G)**. We found that expanded clones were remarkably unstable within individual participants, with very few clonotypes recurring at more than one timepoint. v1068_819 had completely unique repertoires of top 25 clonotypes at each sampled timepoint **(Figure 13F)**. Next we hypothesised that clonotypes that showed

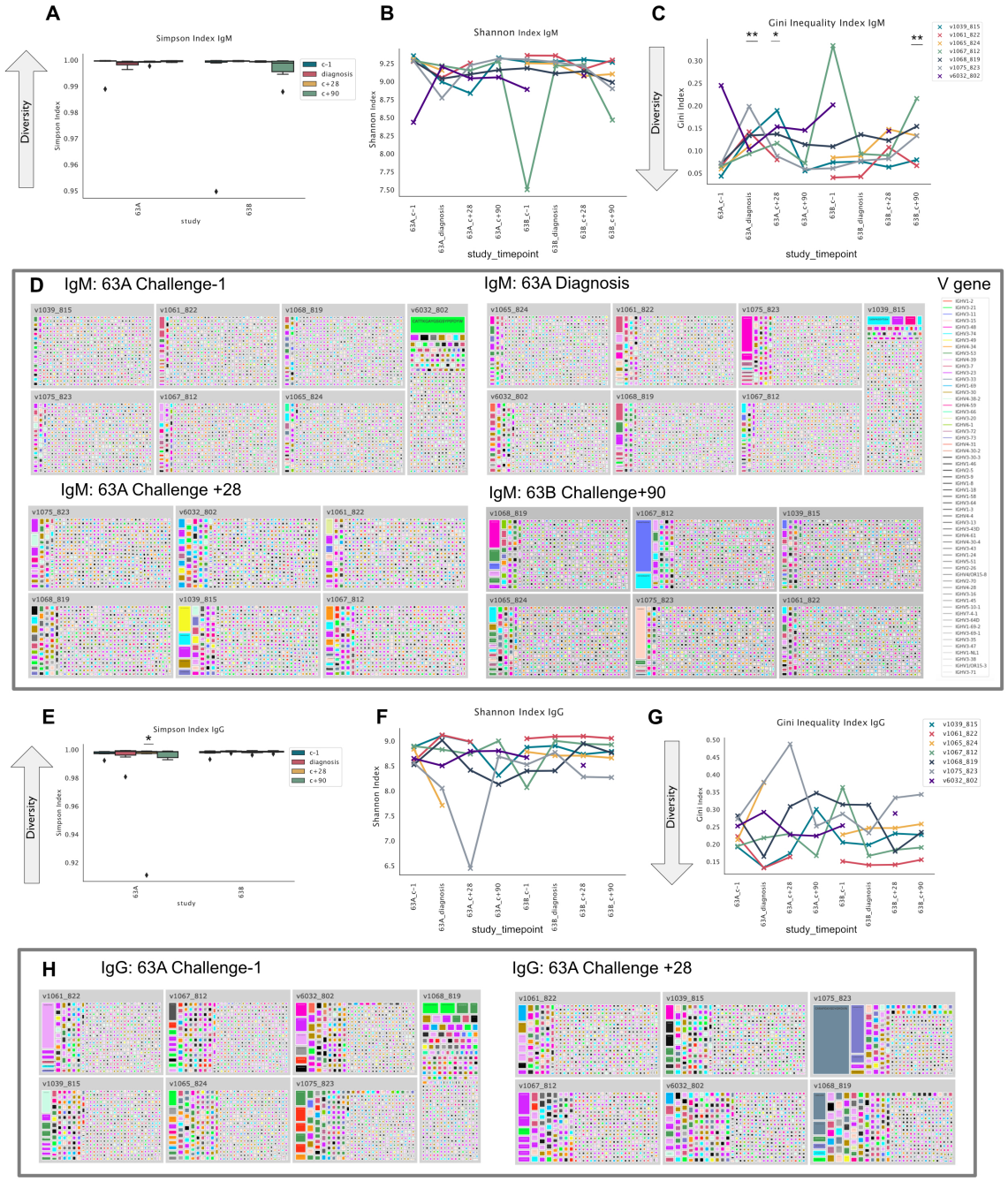


Figure 12: Diversity in IgM and IgG. Repertoires were randomly subsampled to 700 UMIs. For diversity indices, repertoires were re-sampled 1000 times and the diversity indices averaged across iteration. Two samples were excluded because of insufficient IgG for diversity analysis (v1065_824 at 63A_c+28 and v6032_802 at 63B_c+90). **A)** Simpson's Diversity Index, **B)** Shannon Index and **C)** Gini Index of Inequality coloured by individual for IgM repertoires. **D)** Treemap plots shown for timepoints of interest. **E)** Simpson's Diversity Index, **F)** Shannon Index and **G)** Gini Index of Inequality coloured by individual for IgG repertoires. Differences in diversity relative to c-1 were tested using a linear mixed model with study_timepoint as a fixed effect and the effect of individual treated as random. None of the diversity metrics met the assumptions of normality. Simpson's Diversity and Shannon Index were modelled using gamma distributions. Gini Inequality was log-transformed and modelled with a gaussian distribution. P values were not adjusted for multiple comparisons (only p * < 0.05, ** < 0.01 shown, differences all other timepoints were not statistically significant).

a large degree of expansion or contraction were likely to be changing in response to infection. We identified clonotypes that displayed the highest degree of variance in their proportions across the timepoints (>2 standard deviations above each volunteer's mean variance). These analyses were performed without considering isotype, since BCRs may undergo class switch recombination between timepoints. BCRs were assigned to clonotypes independently of constant region. Plotting the proportions of high-variance clonotypes across timepoints revealed that expansions in the first infection were not recurring upon re-challenge **(Figure 13H)**. Again, few expansions in the second infection were observed apart from at the c+90 timepoint.

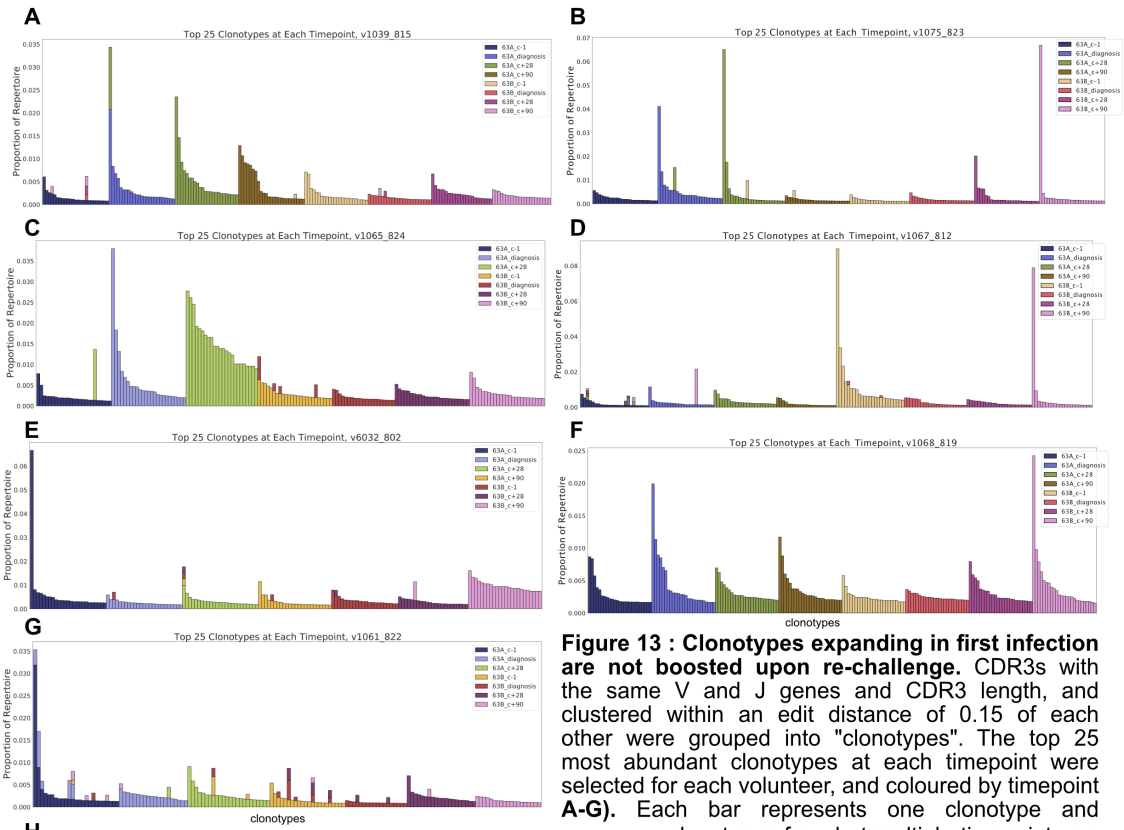
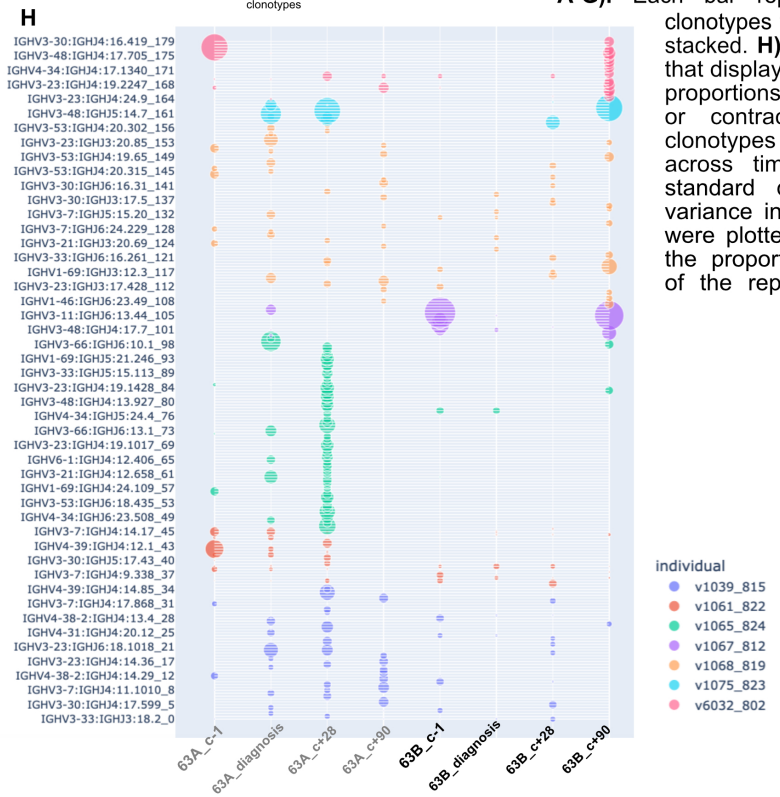


Figure 13 : Clonotypes expanding in first infection are not boosted upon re-challenge. CDR3s with the same V and J genes and CDR3 length, and clustered within an edit distance of 0.15 of each other were grouped into "clonotypes". The top 25 most abundant clonotypes at each timepoint were selected for each volunteer, and coloured by timepoint **A-G).** Each bar represents one clonotype and clonotypes found at multiple timepoints are stacked. **H)** Next we identified clonotypes that displayed the highest variation in their proportions across timepoints (expanding or contracting). For each volunteer, clonotypes for which the proportions across timepoints were more than 2 standard deviations above the mean variance in proportion for that volunteer, were plotted. Dots are scaled relative to the proportion the clonotype makes up of the repertoire at a given timepoint.



3.3.11 Clonal Trajectories

Since the lack of clonal boosting or recall observed in the previous analysis was surprising, we used an alternative approach to examine which clonal trajectories B cells followed. We adapted an analysis from Minervina *et al.* (2021) to identify distinct TCR trajectories. We were able to replicate their analyses to identify three TCR trajectories using their data (Supplementary Figure 4). Briefly, we sampled the top 1000 most abundant clonotypes across any of the timepoints and normalised each clonotype's abundance at other timepoints relative to the timepoint where it made up the largest proportion of the repertoire. We then performed a principal component analysis and used the variance ratios (**Figure 14A**) and the skreeplot (**Figure 14B**) to determine the number of clusters (8). We then used kmeans clustering to identify clonotypes which followed similar trajectories. In our data, the PCA displayed clear artefacts of under-sampling with most clonotypes only being highly abundant at one timepoint (**Figure 14C**). Plotting the proportion clonotypes from each cluster made up of the repertoires at a given timepoint confirmed that each cluster was defined mainly by clonotypes expanded at a single timepoint (**Figure 14D**).

We expected to see boosting of existing immune memory at the c+28 timepoint in the second infection, so we were particularly interested the trajectories of clonotypes expanded at c+28 in the second infection. Cluster 2 identified from the trajectory analysis contained clonotypes expanded at c+28 in the second infection (**Figure 14 E,F**). There was no strong evidence suggesting that clonotypes expanded in the first challenge were being boosted upon re-challenge. Thus, clonal trajectories could not be reliably identified in our data which could be due to sampling depth being too low or the peripheral BCR repertoire being too diverse making re-sampled clonotypes rare.

3.3.12 Clonotypes Do Not Expand Between Timepoints

Because the lack of apparent clonal boosting between the first and second infection was surprising, we used a third approach to confirm that this signal was not detected. For clonotypes which were found in more than one timepoint (re-sampled clonotypes), we calculated the fold change in proportion of each clonotype between the pairs of timepoints it was found in, and averaged these fold changes across participants (**Figure 15A**). Clonotypes sampled at both day of diagnosis and challenge +28 in infection one were expanding. There were also expanding clones between nearly all timepoints and challenge + 90 in the second infection, except for 63B_c-1 and 63B_diagnosis. Next we examined these dynamics in each volunteer separately, as well as the proportion of the repertoire made up by the expanded clonotypes (Figure 15 B-H). While there was substantial heterogeneity among volunteers, four of the seven volunteers had clonotypes which expanded from day of diagnosis to c+28 in the first infection and six of seven volunteers had clones from several timepoints expanding to c+90 in the second infection.

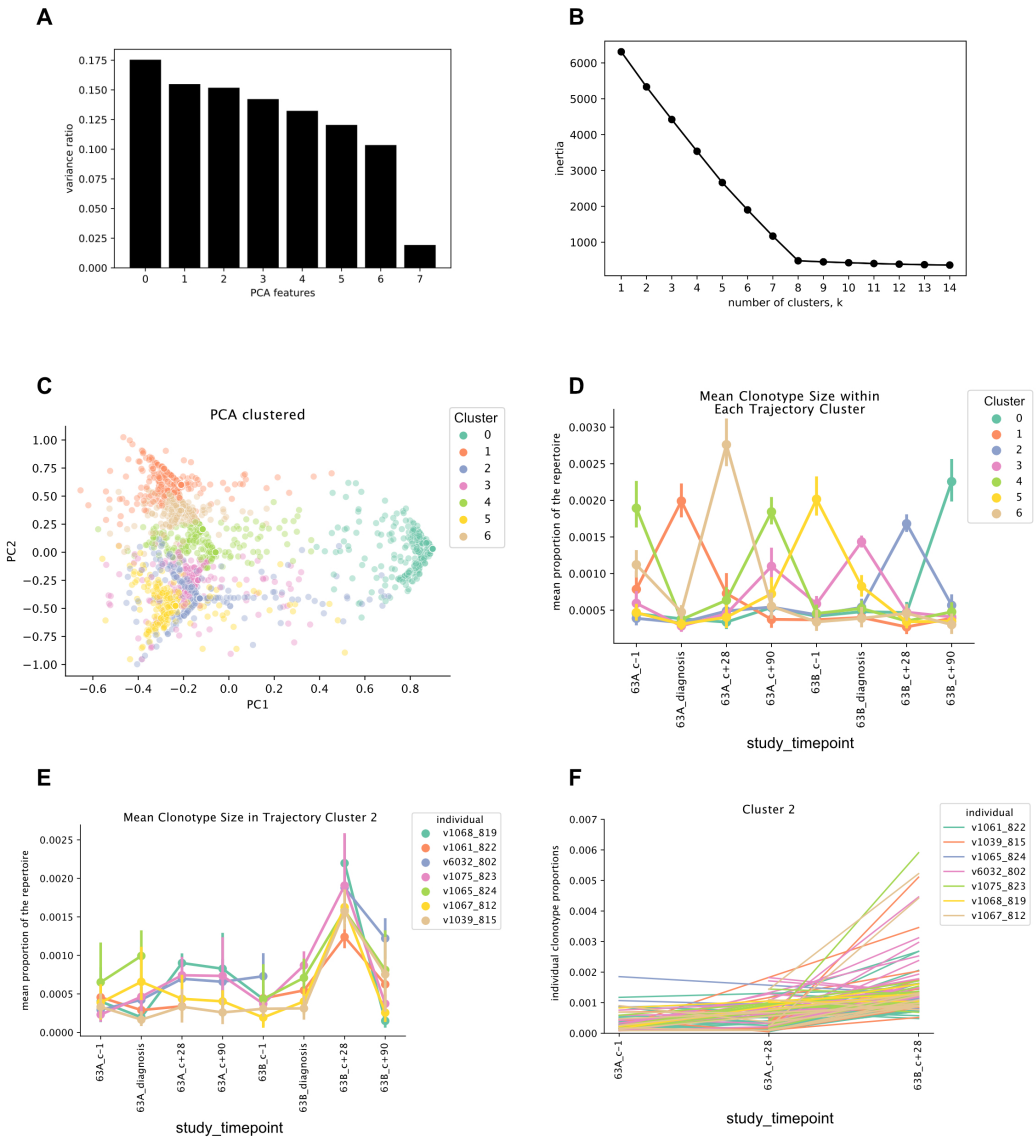


Figure 14: Clonotype Trajectories. Clonal trajectory analysis was adapted from Minervina *et al* 2021: First, for each individual, clonotypes were selected that were present among the top 1000 most abundant in any of the timepoints. Next, the proportions of these top 1000 clonotypes were normalised by dividing the abundance of each clonotype at a given timepoint by the largest proportion for that clonotype at any of the timepoints. PCA was performed on the resulting normalized clonal trajectory matrix and the number of clusters was determined using a skreeplot. K-means clustering was performed on the trajectories to identify clonotypes with similar trajectories. **A)** Variance ratio of PCA. **B)** Skreeplot to determine number of clusters. **C)** PCA coloured by trajectory clusters identified from k-means clustering. **D)** Mean proportion of the repertoire occupied by clonotypes for each trajectory cluster. **E)** Cluster 2 highlighted to demonstrate that clonotypes abundant at 63B c+28 do not expand in the first infection before being boosted upon re-challenge. **F)** Proportions of all individual clonotypes assigned to Trajectory Cluster 2, for c-1, c+28 in the first, and c+28 in the second infection, coloured by individual.

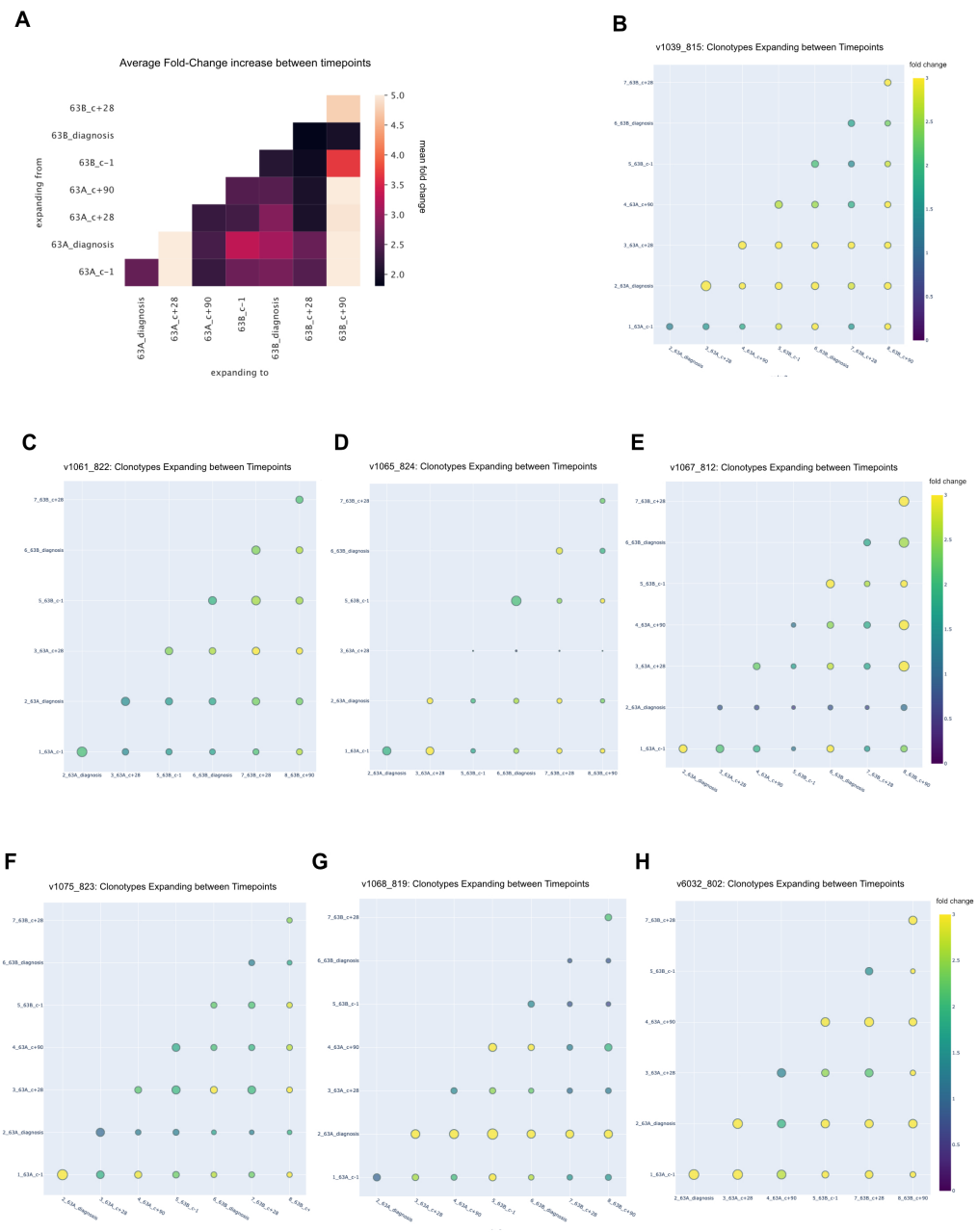


Figure 15: Clonotypes Expanding between Timepoints. A) Heatmap of average log fold change of clonotypes shared between two timepoints (averaged across individuals) **B-H)** Clonotypes shared between two timepoints plotted for each individual, size of dot corresponds to proportion of the repertoire the clones make up at the timepoint on the x-axis, colour corresponds to the average fold-change of clonotypes shared between the two timepoints. One plot shown for each individual.

3.3.13 More Mutated BCRs in the Periphery at Day of Diagnosis in the First Infection

Next we wanted to investigate whether there was any evidence of affinity maturation in BCRs post challenge. Levels of somatic hypermutation in the BCR can indicate that B cells are undergoing affinity maturation and would be expected to increase with more "mature" memory responses. We quantified the average number of mismatches between the germline and BCR sequences (**Figure 16A**). We observed an increase in the average number of mutations per BCR at day of diagnosis in the first infection and levels of SHM similar to the pre-challenge time-point at day of diagnosis in the second infection.

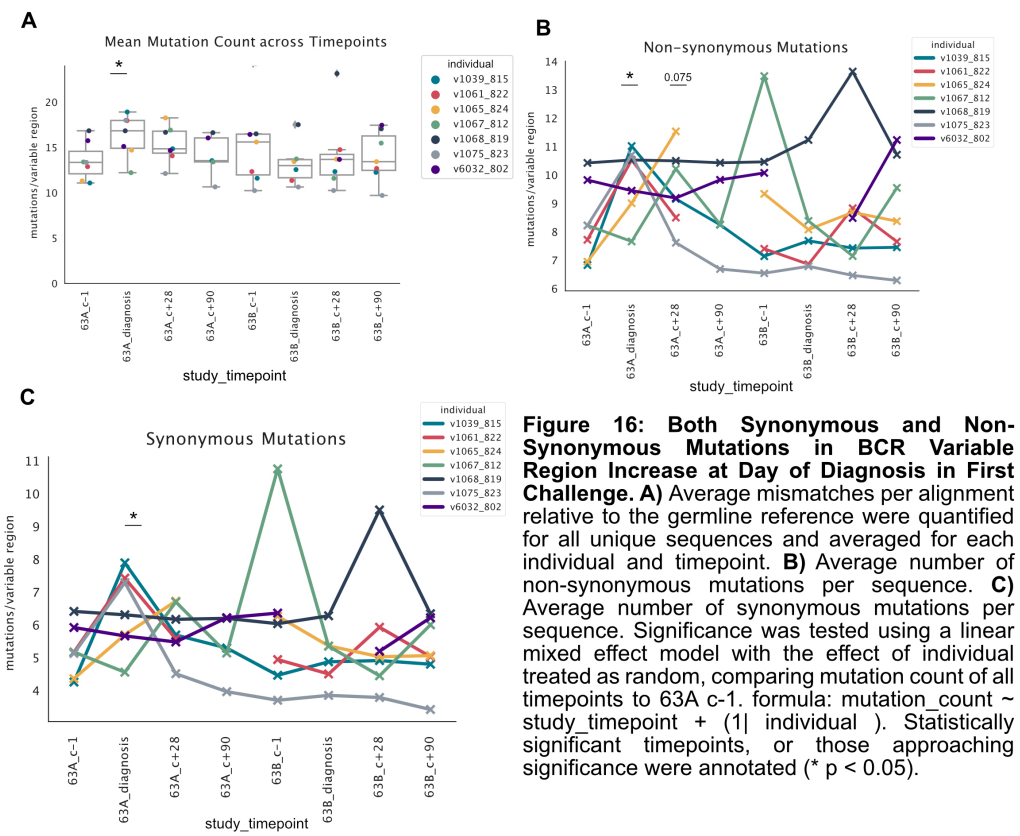


Figure 16: Both Synonymous and Non-Synonymous Mutations in BCR Variable Region Increase at Day of Diagnosis in First Challenge. A) Average mismatches per alignment relative to the germline reference were quantified for all unique sequences and averaged for each individual and timepoint. **B)** Average number of non-synonymous mutations per sequence. **C)** Average number of synonymous mutations per sequence. Significance was tested using a linear mixed effect model with the effect of individual treated as random, comparing mutation count of all timepoints to 63A c-1. formula: mutation_count ~ study_timepoint + (1| individual). Statistically significant timepoints, or those approaching significance were annotated (* $p < 0.05$).

Affinity maturation tends to select for BCRs that contain mutations which result in a change in amino acid, so-called nonsynonymous (NS) mutations. Examining the average number of both non-synonymous (**Figure 16B**) and synonymous mutations (**Figure 16C**), we observed an increase in both at day of diagnosis. Upon re-challenge the prediction would be that an antigen-maturing response would accumulate NS mutations as BCRs undergo rounds of affinity maturation and antigen selection, however we did not observe any consistent increase in mutation burden in BCRs at the later timepoints.

3.3.14 Increased Affinity Maturation in IgM Compartment

Next we looked at SHM in IgM and IgG separately. While most IgM is expected to be unmutated and derived from naive B cells, some mutated IgM from IgM memory B cells can be sampled in the periphery. The distribution for mutation frequency in IgM is therefore heavily weighted to zero (**Figure 17A**). We subsetted the repertoires to IgM with more than one mutation to analyse non-naive IgM. Comparing the distributions of the number of mismatches per BCR in mutated IgM between the two infections at the four timepoints, repertoires increased their average numbers of mutations per base, ie. mutation frequency, at day of diagnosis and c+28 in the first infection, and a sharp increase in SHM in two samples at c+90 in the second infection (**Figure 17B**). The increase at day of diagnosis in the first infection was present in all but one individual and was also observed at c+28 in most volunteers and this difference relative to c-1 was statistically significant for both timepoints (**Figure 17C**).

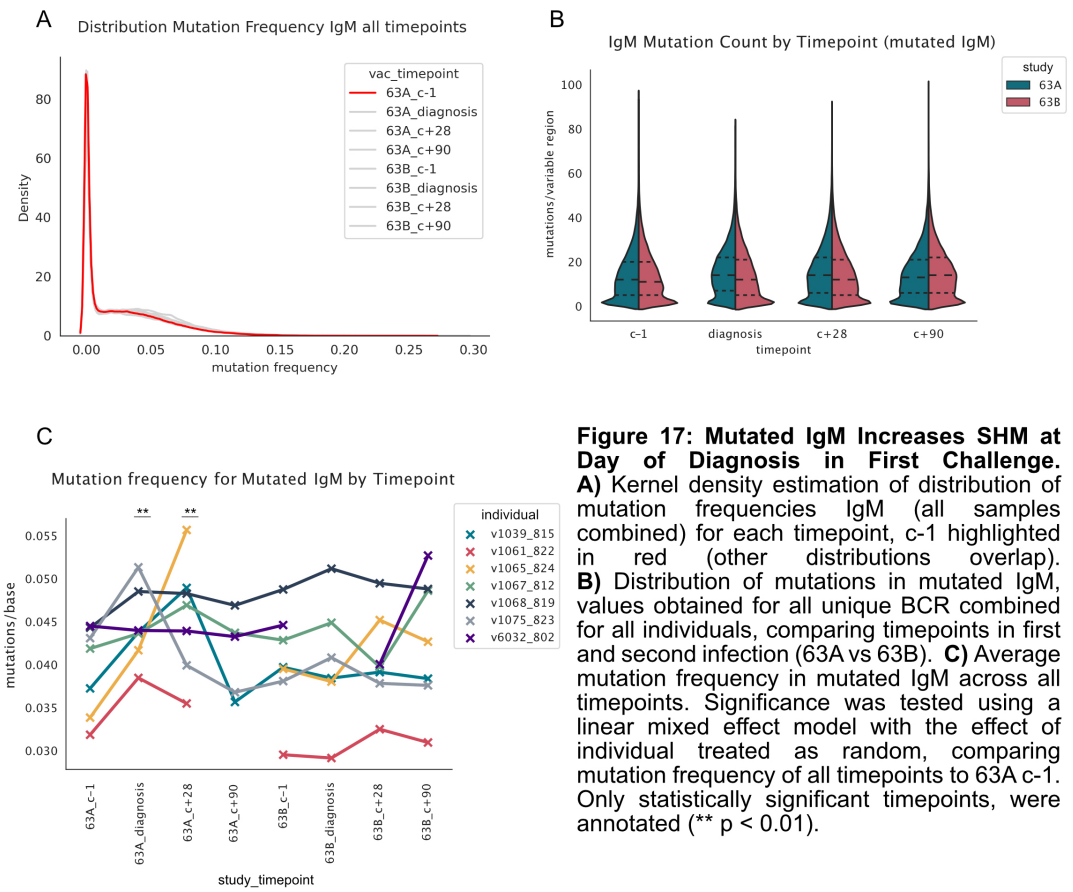


Figure 17: Mutated IgM Increases SHM at Day of Diagnosis in First Challenge. **A)** Kernel density estimation of distribution of mutation frequencies IgM (all samples combined) for each timepoint, c-1 highlighted in red (other distributions overlap). **B)** Distribution of mutations in mutated IgM, values obtained for all unique BCR combined for all individuals, comparing timepoints in first and second infection (63A vs 63B). **C)** Average mutation frequency in mutated IgM across all timepoints. Significance was tested using a linear mixed effect model with the effect of individual treated as random, comparing mutation frequency of all timepoints to 63A c-1. Only statistically significant timepoints, were annotated (** $p < 0.01$).

3.3.15 Potential Signature of Decreased SHM in IgG

B cells that have undergone a germinal centre reaction can switch their isotype from IgM and IgD to IgG, and undergo affinity maturation. Mutation frequency in IgG generally follows a gaussian distribution, which was observed in our data. Comparing IgG mutation counts between timepoints in the first and second infection revealed that at c+28 a small peak of BCRs with very few mutations relative to germline were apparent (**Figure 18A**). While the mean mutation frequency did not differ across timepoints (**Figure 18B**), an enrichment of unmutated BCRs was observed at c+28 in the first infection (**Figure 18C**), and at diagnosis and c+28 in the second infection (**Figure 18D**). Overall there was a greater frequency of unmutated IgG in the second challenge compared to the first (**Figure 18E**). Examining the percent of the IgG repertoire with more than one mutation relative to germline confirmed a drop in mutated IgG at the c+28 timepoints and these differences were statistically significant in the second infection (**Figure 18F**). Thus, a larger proportion of the repertoire contains BCRs identical to germline at c+28 post infection, which was an unexpected observation.

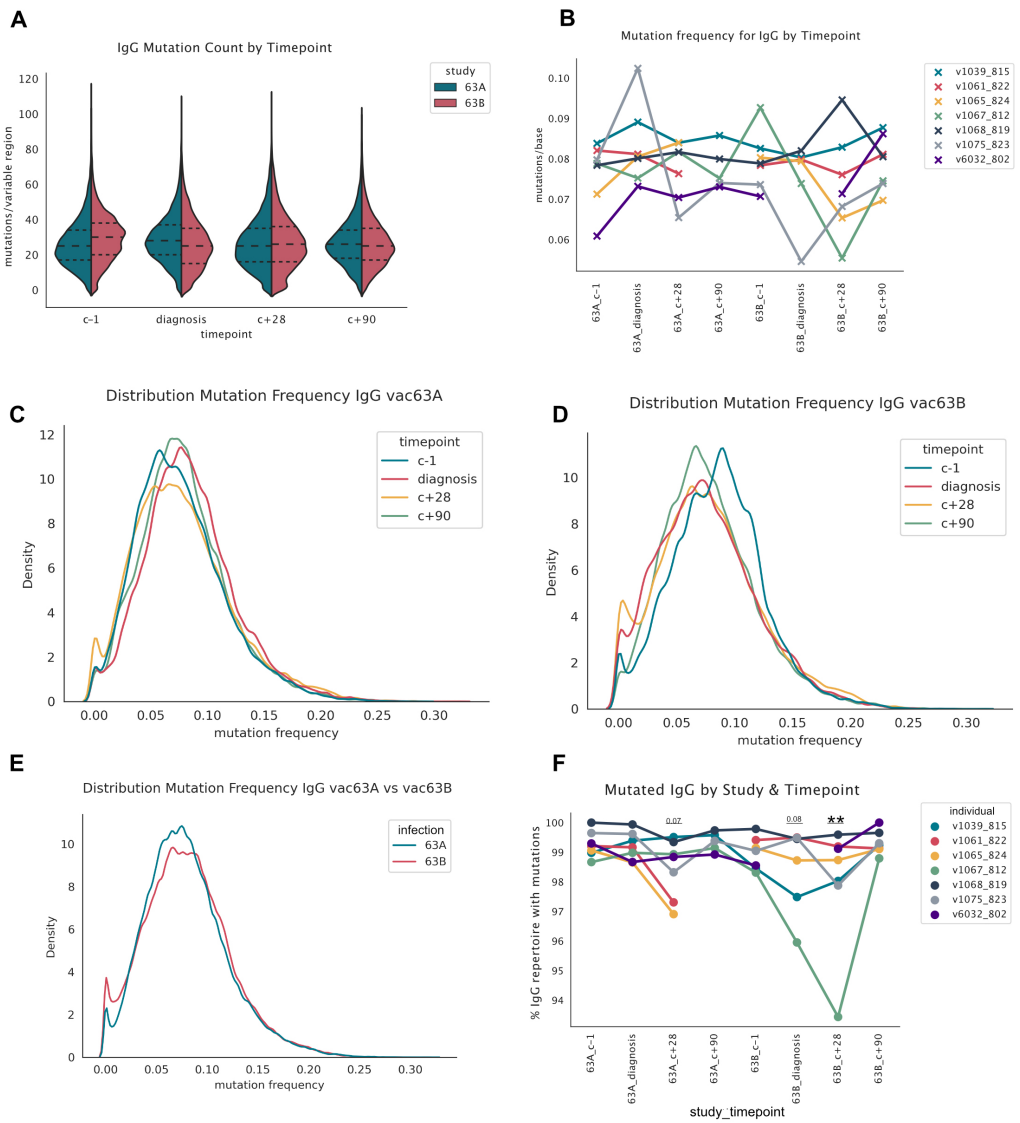


Figure 18: Increase in Unmutated IgG Upon Re-Challenge. A) Total mismatches for each unique CDR3 relative to germline were quantified for IgG and compared between timepoints in the first and second infection. **B)** Mean mutation frequency for IgG across timepoints (tested with mixed effect model, no timepoints were significantly different to 63A c-1). **C)** Density plot of mutation frequency for all timepoints in first infection (vac63A), **D)** in the second infection (vac63B) and **E)** between first and second infection.**F)** Percent of the IgG repertoire with > 0 mutations relative to germline. Timepoints with statistically significant difference, or approaching significance are annotated (** p < 0.010).

3.3.16 Integrating Clonal Expansion, Shared Clonotypes Across Timepoints, Isotype Usage and Affinity Maturation Using Network Diagrams

To understand relationships between clonal expansion, isotype usage, and somatic hypermutation, we generated network diagrams of clonotypes across all of the timepoints for each volunteer and coloured them by isotype or somatic hypermutation frequency. We randomly subsampled each repertoire at each timepoint to 1900 unique UMIs and removed any clonotypes which only had one associated UMI ("singletons"). One network was generated for each volunteer across all of the timepoints. Each point represents one BCR identified by a unique UMI. BCRs which belong to the same clonotype are linked by edges and clustered using the Fruchterman Reingold algorithm. These networks were then coloured by timepoint (left), constant region (middle) and shm (right) (**Figure 19A and 19B**). While the trends seen in other analyses, such as repertoires being very diverse at the day of diagnosis and c+28 timepoints in the second infection were observed, this visualisation did reveal some interesting additional trends: most expanded clonotypes had modest levels of somatic hypermutation, and tended to be restricted to one isotype, usually IgM, although occasionally mixed clusters were observed. Strikingly, very few clonal groups contained BCRs from multiple timepoints and those that did were usually shared between adjacent timepoints. Most IgG expansions were found at the later timepoints in the first infection (c+28 and c+90) and levels of SHM varied substantially among clonotype clusters. Volunteer v1061_822 had a repertoire which was heavily skewed towards IgG with 60% of the BCRs having an IgG constant region call and volunteer v6032_802 and v1065_824 both had oligoclonal IgM expansions at one timepoint (those timepoints were also the ones that were excluded from the isotype-specific diversity analysis in Figure 12 because they had low levels of IgG).

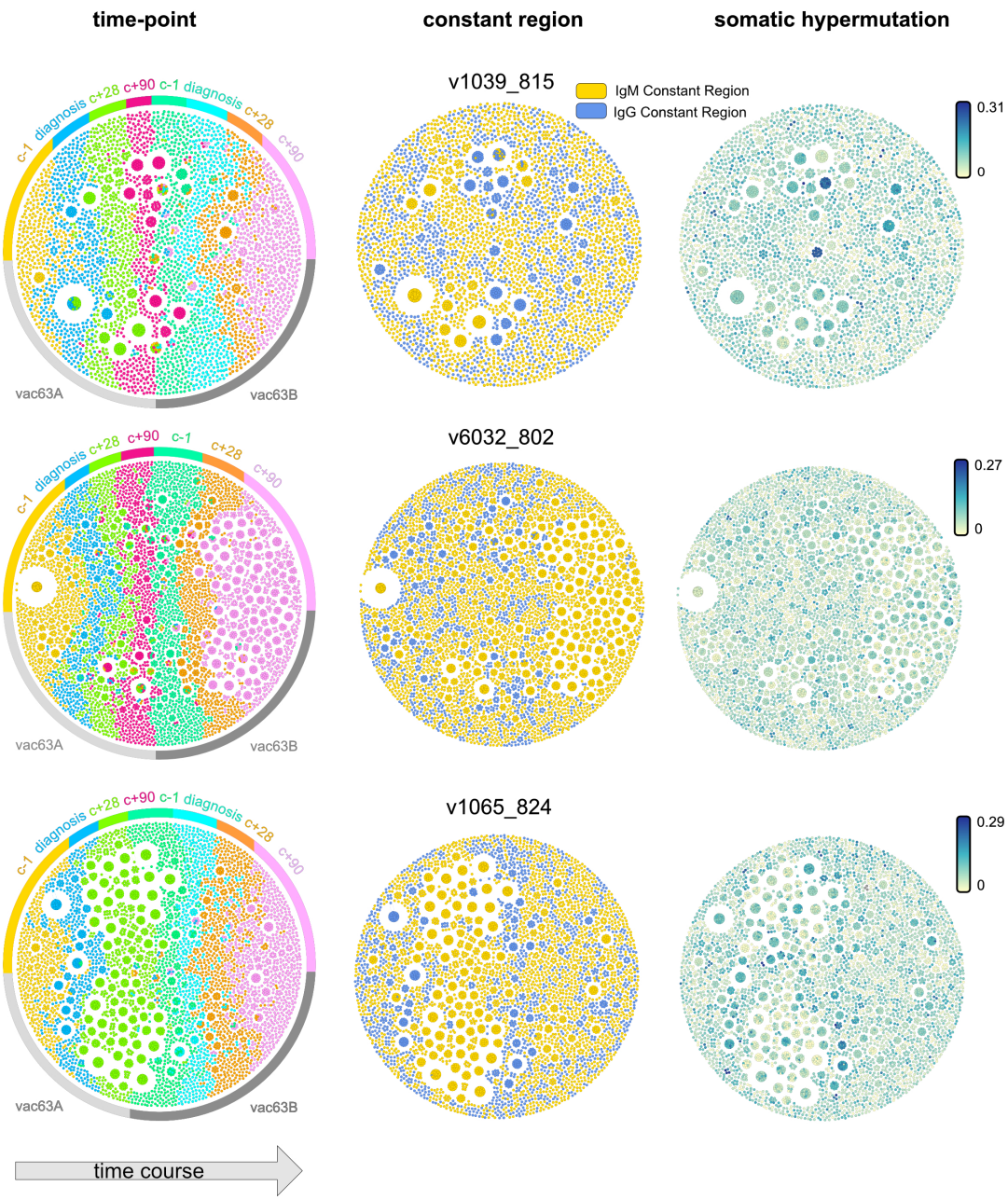


Figure 19: Network diagrams of BCR repertoires for each individual. 1900 UMIs were randomly sampled from each repertoire and clonotypes with less than two different UMIs ("singletons") were removed. Clonotypes from all timepoints were combined for each individual and were clustered using the Fruchterman-Reingold algorithm and coloured by sampling time-point (left), constant region (middle) and somatic hypermutation frequency (right). Each dot represents one BCR and BCRs belonging to the same clonotype are connected by an edge and clustered in close proximity. BCRs were coloured by timepoint (left), constant region (middle), or somatic hypermutation frequency (right). Remaining individuals are shown in the next Figure.

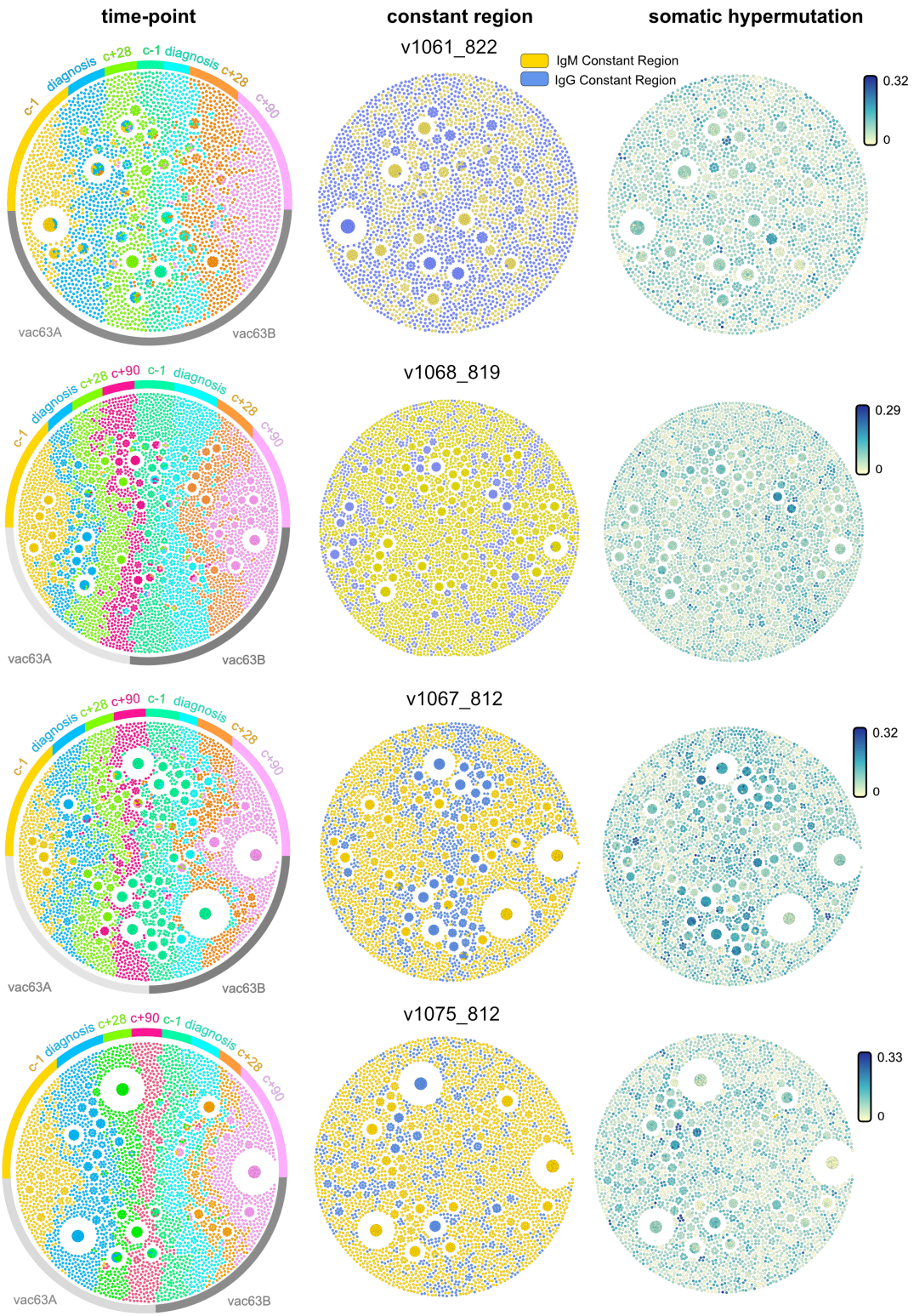


Figure 19 (continued): Network diagrams of BCR repertoires for each individual.

3.3.17 Lymphopenia Affects B and T cells in Equal Proportion in *P. vivax* Challenge

One limitation of this study is that we do not have phenotypic information about the kinetics of the B cell populations and subsets across the timepoints. While we know that the volunteers have severe lymphopenia at day of diagnosis in both infections, we do not know whether this is due to T cells homing to the spleen or whether both T and B cells are lost from the periphery. Sandoval et al. 2021 characterised T cell responses in these same volunteers and reported that 30-70% of circulating T cells are lost from the periphery at the peak of infection and home to the spleen. No B cell markers were recorded for these volunteers. However, a marker for CD20 was included in a mass cytometry (cytometry by time of flight-CyTOF) panel used to characterise immune responses to a CHMI *P. vivax* trial led by the Spence Lab. Since the lymphocyte populations are reported to have similar dynamics in response to both malaria parasites, we used the CyTOF data to estimate whether B cell kinetics were likely similar to the T cells. We first gated singlets, then CD45+ cells to identify Leukocytes and then gated for CD3+ T cells and CD20+ B cells. We included the pre-infection timepoint, day of diagnosis and a post-treatment timepoint (T+6). Lymphocytes were severely reduced in the periphery at day of diagnosis in response to *P. vivax* infection as seen by full blood counts (**Figure 20 A,B**) and by Mass Cytometry (**Figure 20 C**), mirroring the observations in *P. falciparum*. The ratio of B to T cells was the same across all timepoints, suggesting that B cells are lost or displaced in equal proportion to the T cells at day of diagnosis (**Figure 20 D**). Finally, gating for CD27+ B cells revealed a trend towards a smaller percentage of CD27+ B cells in the periphery post CHMI challenge. Ideally B cell subsets will be characterised in future CHMI studies of *P. falciparum* to confirm whether this is the case.

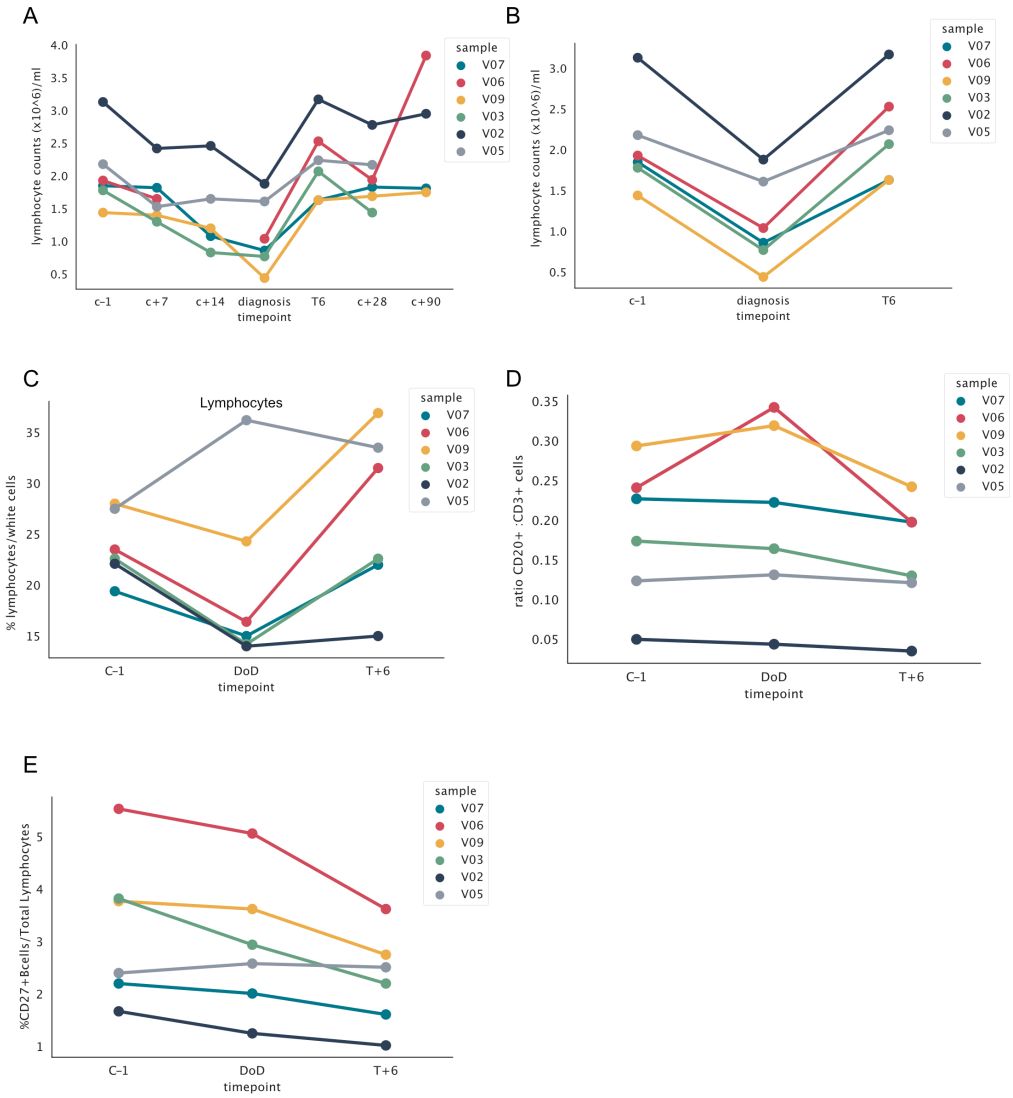


Figure 20: Both B and T cells disappear from the periphery at day of diagnosis in malaria infection. A) Since lymphocyte dynamics have been shown to be similar between the *P. falciparum* and *P. vivax* challenge models, data from *P. vivax* challenge was used to examine B cell dynamics in a first malaria infection. Lymphocyte counts from vac069 *P. vivax* clinical trial, obtained from full blood counts with a 5-part differential white cell count, across time-course of infection. **B)** Lymphocyte counts subsetted to timepoints analysed in CyTOF data. **C)** Lymphocytes as a % of leukocytes (CD45+) and **D)** ratio of B:T cells (C20+ : CD3+ cells) gated from CD45+ population. **E)** CD27+ B cells as a percentage of total Lymphocytes. Abbreviations= c-1: challenge -1, c+7: challenge+7, c+14: challenge+14, day of diagnosis: DoD, T+6: Treatment+6, c+28: challenge+28, c+90: challenge +90).

3.4 Discussion

This study was intended as an exploratory examination of BCR repertoires in two controlled human malaria infection trials but the results were nonetheless somewhat unexpected. The key observations made in this chapter were:

1. Clonal expansion at day of diagnosis in the first infection, but increased diversity compared to even the pre-challenge timepoints at day of diagnosis in the second infection.

2. In the first infection, IgM repertoires clonally expanded, increased mutation frequency and skewed their V gene repertoire early post-challenge.

3. In the second infection, IgM repertoires did not show evidence of clonal expansion until the c+90 timepoint.

4. IgG repertoires did not display signs of clonal expansion or affinity maturation. However, an enrichment of unmutated BCRs was observed at c+28 in the first, and to a greater extent in the second infection.

5. We could not detect evidence of clonotypes from the first infection being boosted upon re-challenge.

IgM Signatures Were Detected at Early Timepoints in the First Infection

In the first challenge, IgM repertoires underwent clonal expansion at day of diagnosis, increased somatic hypermutation in mutated IgM BCRs, as well as increased IGHV3-7 gene usage. Day of diagnosis, which was between day 8-12 post challenge across participants, is early post-challenge to be detecting B cell responses, but other studies using CHMI have found serum IgM against MSP2 from day 14 post primary infection (Boyle et al.

2019) which is only 2-6 days after we observed the IgM BCR signatures in our cohort and could be consistent with a rapid and short lived initial B cell response. In the second infection, four of the volunteers also had increased IgM clonality at the c+90 timepoint. This timepoint is so long after challenge that it is unlikely to be a malaria-related signature and may reflect other seasonal or infectious exposures.

Increased Diversity at Day of Diagnosis Upon Re-challenge

Blood counts from clinical monitoring of the volunteers showed that lymphopenia was more severe in the second infection compared to the first, so there were likely fewer B cells sampled. We examined B cell markers from *P. vivax* infection, which have been shown by Bach *et al.* to mirror the acute phase response of *P. falciparum* infection closely (F. A. Bach et al. 2021). The relative proportions of B and T cells were unchanged at day of diagnosis when volunteers were lymphopenic. However, the authors also found that T cell activation is more severe in *P. falciparum* infection than in *P. vivax* (F. A. Bach et al. 2021). If we assume that lymphopenia at day of diagnosis in the second infection affects B cells as severely as T cells, the increased diversity at day of diagnosis upon re-challenge may simply reflect a more naive repertoire remaining in the blood during the acute-phase malaria response. Clonally expanded B cells from the first challenge could be absent due to the cells being recruited into the secondary lymphoid tissues. However, we might expect B cells which expanded in the first infection to be captured at any of the other timepoints in the second infection. Instead we do not see many clonotypes reappearing over the course of the re-challenge experiment. An explanation for this lack of boosting could be that short-lived B cell responses rapidly expand early in the first infection but that these cells do not survive until the re-challenge.

Increase in Unmutated IgG BCRs at c+28 in the Second Challenge

Finally, the moderate increase in unmutated, class switched IgG BCRs which we observed in a subset of volunteers in the first challenge at c+28, and to a greater extent upon re-challenge at the same timepoint was unexpected. This signature could be explained by polyclonal B cell activation, or by class switched B cells exiting the Germinal Centres early. In future it would be interesting to characterise these unmutated IgG B cells phenotypically and follow their trajectories to see whether they are maintained in B cell memory or are deleted. A colleague in the Cowan Lab has identified a B cell marker associated with under-mutated IgG BCRs (Sutherland, unpublished), which, if validated as a flow cytometry marker, could provide a way of identifying these unmutated class-switched B cells and tracking their dynamics. Under-mutated IgG BCRs have also been observed in autoimmune context, such as rheumatoid arthritis (Cowan et al. 2019). Interestingly these diseases are also associated with DN2 B cells (IgD−CD21−CD27−T-bet+CD11c+CXCR5−) that are transcriptionally similar to atMBCs in malaria (Holla et al. 2021). According to CyTOF data from the *P. vivax* trial that the Spence Lab are studying (F. A. Bach et al. 2021), the proportion of CD20+, CD27+ B cells decrease over the course of infection which could be consistent with malaria infection being skewed towards an atypical memory response rather than a central memory response. IgD was not included as a CyTOF marker so the dynamics of IgD-CD27- could not specifically be investigated, but would be of interest to examine in future studies.

Interferon-gamma Response at Day of Diagnosis May Drive atMBC Phenotype in These Volunteers

Other data recorded from these volunteers demonstrated a potent myeloid and type-1 interferon response (Sandoval et al. 2021). RNA-seq analysis performed by the Spence Lab on these volunteers demonstrated a strong IFN-gamma response at day of diagnosis in

the first, and to a greater extent in the second and third challenges (Sandoval et al. 2021). They reported Th-1 polarised T cell phenotypes. These signatures could be consistent with mechanisms for inducing atypical memory B cell development. Furthermore, a previous study of the TCR repertoires by a former PhD student on this same cohort found non-specific recruitment of T cells from the periphery at day of diagnosis (N. L. Smith 2022). Previously established and persistent TCR clones, including CMV and EBV-specific clones disappeared out of peripheral circulation at day of diagnosis before returning to their baseline abundance following treatment. Taken together, this evidence could support the hypothesis that in these volunteers type-1 interferon responses and TLR engagement result in polyclonal B cell activation which could lead to atMBC induction, perhaps as a mechanism of dampening an early pro-inflammatory response. Perhaps the reason we did not observe clonal boosting upon re-challenge was because the majority of cells captured were polyclonally activated B cells making the antigen-specific signal difficult to detect. Data from other controlled human infection trials have found that acute infection up-regulates B cell activation factors and B cell survival signals which can result in T-independent, or non-specific, polyclonal memory B cell activation (Ly and Hansen 2019).

Volunteers Produce and Boost Malaria-Specific Antibodies

Interestingly, despite the lack of clonal signature detected in the second infection, antibodies against AMA1 and MSP1-19 were produced by the participants in our study, and boosted upon re-challenge (**Figure 21A,B and 21C,D**- data from Salkeld et al. 2022). This could be explained by a small subset of antigen-specific B cells committing to plasma and memory B cell lineages that we cannot identify from the sequencing data, or that the timepoints at which plasma cells that produce these antibodies circulated, were missed.

CHMI and field studies have found that malaria antibody responses are generated

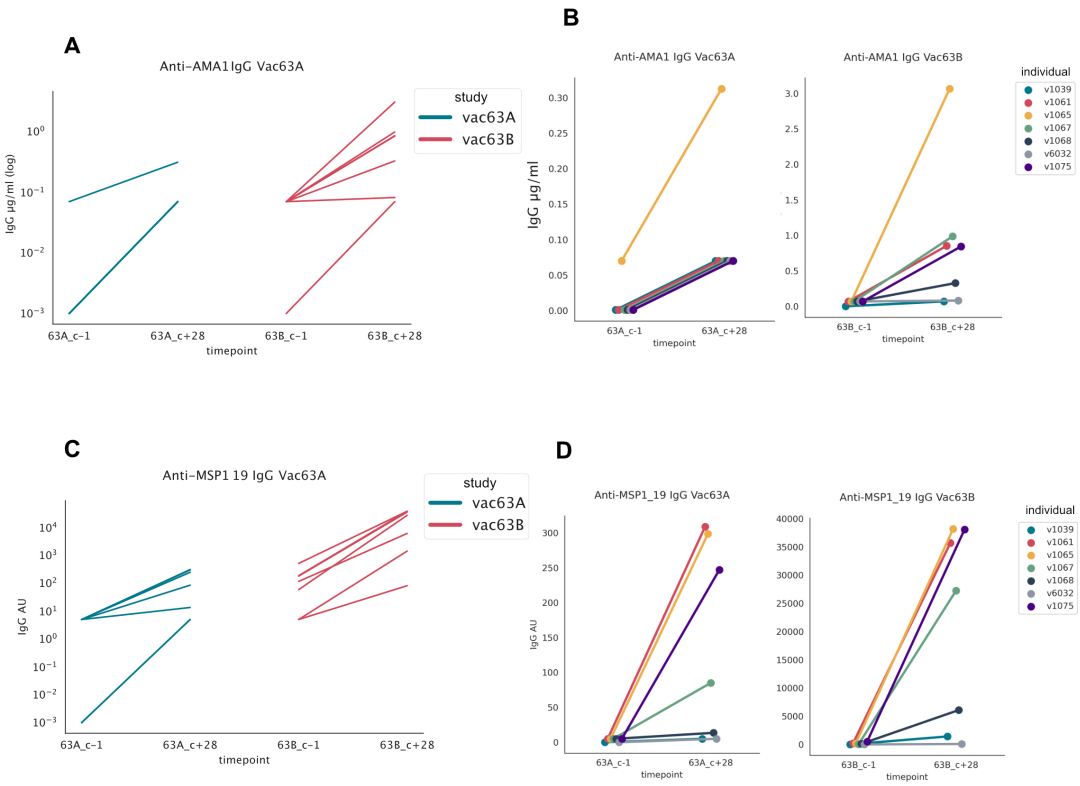


Figure 21: Antibodies against parasite antigens are boosted after a second challenge. Data re-used with permission from Spence Lab, published in Salked *et al.* (2022). **A)** IgG antibody titres against AMA1 (µg/ml - shown on a log scale) after a first (vac63A) and second infection (vac63B), measured by ELISA. **B)** Anti-AMA1 IgG titres shown for each individual (µg/ml linear scale). **C)** IgG antibodies against MSP1-19 (Arbitrary Units) after a first and second challenge measured by ELISA. **D)** Anti-MSP1-19 IgG shown for each individual and for the first (vac63A) and second (vac63B) challenge.

after a first infection and are boosted upon re-exposure (Scholzen and Sauerwein 2016). The polyclonal B cell responses to malaria might therefore not reflect defects in memory generation or maintenance, but may instead be influenced by the polymorphic and diverse malaria antigens.

3.4.1 Studies Report Diverse BCR Repertoires Features

Results from different BCR repertoire studies of human malaria do not report conserved BCR repertoire signatures, but most studies report that atMBC and cMBC BCRs are similar. One group compared the repertoires of atMBCs and cMBCs in three malaria-experienced children and found that the repertoires do not differ substantially in terms of V gene usage, SHM and CDR3 length and amino acid properties (Zinöcker et al. 2015). Another study compared BCR repertoires of seven Plasmodium-exposed adults from an endemic country with 13 malaria-naive American adults. They found that IgM+ atMBCs closely resemble naive B cells and IgG+ atMBCs resembled IgG+ cMBCs regardless of malaria exposure (Ashley E Braddom et al. 2021). These findings agree with data from Portugal *et al.* which found that VH gene usage and somatic hypermutation rates of atMBCs and cMBCs were indistinguishable in four Malian adults and that classical and atypical MBCs were similarly clonally expanded and 10% of clones were found in both populations (Portugal et al. 2015). A study comparing infants and children pre-malaria exposure and during acute infection found unexpectedly high levels of SHM in both age groups and levels of mutation in IgM were particularly increased during acute infection. Furthermore, diversity and sizes of B cell lineages increased during acute infection (Ben S. Wendel et al. 2017). Braddom *et al.* (2021) also observed that in malaria-experienced adults rates of SHM were increased in atMBCs, and found increased use of IGHV3-73 both in IgG+ cMBCs and IgM+ and IgG+ atMBCs.

V gene biases have been reported in malaria, albeit in very divergent study systems:

A study following 41 vaccinees receiving the Pfs25-EPA/Alhydrogel vaccine, compared to hepatitis B and meningococcal vaccinees, found that V gene usage in total B cells was unchanged after 4 vaccine doses. When examining Pfs25-IG-specific antibodies specifically, they observed preferential usage of IGHV4 (Coelho, Jacob D. Galson, et al. 2022). In studies sampling endemically exposed individuals, IGHV4-34 has been reported to be enriched in IgD+IgM+ atMBCs (Holla et al. 2021). The authors also observed this same V gene bias in a previous study using an antibody to stain autoreactive IGHV4-34 B cells and identified autoreactive serum IgG that used the IGHV4-34 gene segment. They saw an increase in this self-reactive IgG following acute febrile malaria but it did not increase with age or correlate with protection from acute malaria (Hart et al. 2016).

We do not know which BCRs came from naive, memory or atypical B cells in our dataset, making it difficult to directly compare our repertoire signatures to those published in the literature. However it is clear that *P. falciparum* elicits an antigenically complex immune response and no single "public" response has been reported in the literature. Increased somatic hypermutation appears to be a characteristic of chronic malaria infection. All of these studies share the limitation that sequencing BCR repertoires from peripheral blood will likely miss most of the cells actively involved in responding to infection - the spleen or bone marrow would be preferable sampling sites, but these are difficult tissues to access in humans.

3.4.2 atMBCs: Protective or Pathogenic in Malaria?

Although the expansion of atypical MBCs, particularly IgM+ atMBCs, appears to be triggered by malaria infection (Illingworth et al. 2013; Muellenbeck et al. 2013; Greta E Weiss et al. 2011), likely due to early IFN-γ signalling, their role in hampering or regulating adaptive immune responses in malaria is debated. While atMBCs have been shown to be able to produce parasite-specific B cells and affinity-mature their BCRs, studies have also

found that atMBCs differentiating from cMBCs in malaria become anergic to activation by BCR signalling and express inhibitory receptors such as FcγRIIb, which inhibits Fc receptor signalling (Portugal et al. 2015; Obeng-Adjei et al. 2017). Some groups have identified parasite-specific B cells in the atMBC compartment (Muellenbeck et al. 2013). Finally, another theory proposes that dysregulation of B cell memory is the result of parasite products activating B cells via T-independent responses. Meanwhile, atMBCs have also been observed in both influenza and in malaria vaccine responses. In one study, Sutton *et al.* describe an alternative lineage of B cells which includes atypical B cells (CD21− CD27−), and found this lineage to increase after immunisation and contain antigen-specific B cells. As a result, these cells were proposed as a normal B cell lineage that participates in infection and vaccine responses (Sutton et al. 2021).

While atMBCs have been suggested to be an "exhausted" B cell subset (Greta E. Weiss, Traore, et al. 2010), another possible explanation for the increase in atMBCs in malaria proposed by Holla *et al.* and others, is that they contribute to the control of inflammation during acute malaria: The authors found that atMBCs expressed more IL-10 compared to naïve B cells and classical MBCs and expressed high levels of IRF8 which plays a role in maintaining peripheral tolerance and anergy. Consistent with other studies, they described high antigen affinity thresholds required for activation of atMBCs, in particular IgD+IgM-lo atBCs, which they found to be expanded in children during acute febrile malaria (Holla et al. 2021). atMBCs express Tbet in response to pathogen-associated molecular pattern molecules (PAMPs), IFN-γ and strong BCR engagement, which results in the cells being less responsive to BCR engagement (Obeng-Adjei et al. 2017). Indeed, stimulating malaria atMBCs in vitro with CpG and cytokines did not induce differentiation into plasma cells. In contrast cMBCs stimulated with these same agonists readily differentiated into plasma cells. Similarly, DN2 cells differentiated *in vitro* from naive B cells from autoimmune patients do not readily differentiate into plasma cells when

BCRs are cross-linked, but will do so when BCRs are transiently engaged (Zumaquero et al. 2019).

The atMBC population peaks early in malaria infection, around 10 days after diagnosis but contracts after the infection is cleared (Sundling et al. 2019). The frequent exposure to malaria in endemic regions may result in an increase in atMBCs that are less responsive to antigen, PAMPs and cytokines, to reduce polyclonal B cell activation and immune mediated pathology. Since many PAMPs and self antigens will be released into the blood-stream upon merozoite rupture, dampening polyclonal B cell activation may in fact reduce immunopathology, perhaps at the cost of generating mature germinal centre B cell responses and efficient immune memory. While the role of atMBCs and even their defining surface markers still appear to be debated in the literature, perhaps they simply represent a more inactive cell state. In highly pro-inflammatory environments with lots of antigen stimulation, perhaps these cells partially dampen an otherwise overwhelming and potentially self-reactive immune response, at the cost of less efficient clonal selection and affinity maturation.

Investigating whether atMBCs are induced in this study system, and how they relate to the clonal dynamics and somatic hypermutation signatures we observed would be of interest. In future longitudinal studies of CHMI, FACS-sorting B cells into naive (CD19+ CD21+ CD27−), classical memory B cells (CD19+ CD21+ CD27+), and atypical memory B cells (CD19+ CD21− CD27−, IgD−), prior to repertoire sequencing could help us to identify whether the increase diversity at day of diagnosis upon re-challenge is driven by changes in the proportion of naive or memory B cell populations. It could also help us understand whether the increased frequency of mutations in mutated IgM and the increase in un-mutated BCRs in IgG is associated with particular B cell subsets.

3.4.3 Antigen-Complexity and Diversity May Alter Clonal Selection in Malaria

The malaria parasite elicits complex immune responses, both because individual antigens are structurally complex, but also because of the abundance of parasite antigens encountered during infection and antigenic variation (Rénia and Goh 2016).

Murugan *et al.* examined affinity maturation dynamics of BCRs that bind to the immunodominant epitope of Pf circumsporite protein (NANP repeat) in Pf-naïve volunteers infected with attenuated Pf sporozoites (PfSPZ Challenge) under chloroquine prophylaxis. This challenge system produces a liver-stage infection, but does not progress to blood-stage infection. They used single cell BCR sequencing and produced BCRs as recombinant monoclonal antibodies. Interestingly, they observed that, over repeated challenges, volunteers were more likely to select germline and memory B cell precursors against the PfCSP antigen than affinity mature existing responses. Over the three PfSPZ challenges, the affinity of the antibodies for PfCSP, and the capacity for the recombinant antibodies to inhibit *P. falciparum*, increases, but the number of mutations in the BCRs does not affect binding affinity and indeed many unmutated BCRs accumulate by the third re-challenge. Using mathematical modeling, they show that with increased antigen complexity, affinity maturation become less efficient. They hypothesise that without continuous exposure to antigen, anti-PfCSP responses are more likely to select precursor cells which already have high affinity for the antigen, rather than mature existing responses (Murugan, Buchauer, Triller, Kreschel, Costa, Pidelaserra Marti, et al. 2018). Although their experimental system is very different to the one used in this study and they only examined maturation of responses to a single antigen, their proposed model may explain both the lack of clonal recall in the second infection, despite antibody titres being boosted, and the increase in unmutated IgG observed at c+28 in the second infection.

In future CHMI studies it would be interesting to investigate whether antigen-specific clones persist and mature between a primary, secondary and tertiary challenge, or whether novel clonotypes specific for malaria antigens are selected from the repertoire instead. This could be achieved by sorting malaria antigen-specific B cells and sequencing the antigen-specific repertoires longitudinally. It could demonstrate whether antigen-specific B cells are maintained or replaced with new clonotypes upon re-challenge.

LIBRA-seq, whereby up to nine antigens can be labelled with a DNA barcode and antigen-specific B cells sorted and sequenced using single cell RNA sequencing, could also provide valuable insight into antigen-specific clonal dynamics in this experimental system. Sorting B cells labelled with malaria antigens and sampling cells across multiple timepoints would reveal whether antigen-specific B cells are matured upon re-challenge or are selected from new clonotypes. Integrating this with transcriptomic data would additionally permit identification of whether these cells are atypical or classical B cells. RNA pseudotime analysis alongside B cell lineage analysis could be used to infer the developmental trajectories these B cells have arisen from. Including a few common antigens like CMV or EBV in the analysis could also be useful to observe whether B cells undergo polyclonal activation upon malaria challenge. Finally, in future CHMI studies participants could drink heavy water that incorporates into newly synthesised DNA and allows the estimation of cell division rates, up until the first treatment timepoint (small quantities of heavy water are not harmful to human health). Deuterium labelled cells from the first challenge could then be tracked across subsequent infections to identify whether B cells from the first challenge persist or multiply upon re-challenge. Combining this with flow-cytometric B cell phenotyping could reveal which B cell subsets actively multiply upon re-challenge.

3.4.4 Conclusion

Immune responses to malaria are complex and the dynamics of affinity maturation and clonal selection are not yet fully understood. Field studies and CHMI models paint a complex picture of malaria adaptive immunity, but it is clear that clonal selection and affinity maturation follow different kinetics and constraints in malaria, perhaps due to the strong pro-inflammatory signals early in infection and the antigenic complexity of the parasite. If comprehensive profiling of antigen-specificity of malaria repertoires were possible, it would likely clarify how the BCR repertoire is shaped by infection and how B cell memory to malaria is acquired.

3.5 Methods

3.5.1 Study Cohort and Sample Collection

All volunteers included in this study were healthy and malaria-inexperienced adults between the ages of 18 and 50 years. They were enrolled in two clinical trials VAC063A (November 2017) and VAC063B (March 2018) that both followed the VAC63 protocol. It was an open label, non-randomised phase I/IIa clinical trial for the RH5.1/AS01 B malaria vaccine (recombinant blood-stage malaria protein RH5.1 in AS01B adjuvant (GSK)) at Oxford University. VAC063 had ethical approval from the UK NHS Research Ethics Service (Oxfordshire Research Ethics Committee A, reference 16/SC/0345) and was registered on ClinicalTrials.gov (reference NCT02927145). For details of the clinical trial and volunteer characteristics see Minassian et al. 2021. From one day after infection onwards, blood samples were taken from volunteers every 12h and PMR (parasite density) was quantified using RT-qPCR (target gene = 18S ribosomal RNA) and, in the vac63A thick blood films were also evaluated to identify parasites (diagnosed if one viable parasite was detected

in 200 fields). Once volunteers developed parasitaemia above 10,000 parasites/ml, or developed symptoms with parasitaemia above 5000 parasites/ml, they were diagnosed and treated with either Riamet or Malarone.

Table 3.2: Day of Diagnosis

Volunteer	Day of diagnosis 63A	Day of diagnosis 63B
v1039_815	9	9
v1061_822	10	9.5
v1065_824	10.5	10.5
v1067_812	9	9.5
v1068_819	8.5	9
v1075_823	12.5	13.5
v6032_802	9	9

3.5.2 Library Preparation for 5' RACE Sequencing with UMIs

The BCR repertoire sequencing was performed as described in Chapter 2. In brief, a cocktail of constant region specific primers (for IgM and IgG BCR isotypes) bind to the 3' end of the mRNA molecule and MMLV RT adds several non-templated deoxycytidines to the 3' end of a newly synthesised cDNA strand when it reaches the 5' end of the RNA template. An oligonucleotide with several riboguanosines at the 3' end can base pair with the stretch of Cs (Matz et al., 1999) to introduce a 12 base unique molecular identifier (UMI) that labels the individual cDNA molecule as well as adding an adapter sequence to amplify the molecule in subsequent PCRs. Newly synthesised cDNA was treated with 1uL of Uracil DNA Glycosylase (5 U/uL, New England Biolabs) which cleaves the uridine bases in unused UMI oligos, so they are not incorporated during subsequent amplification steps. cDNA was amplified in two rounds of PCR to step further into the constant region (IgM_R2 and IgG_R2 primers), introduce Illumina sequencing adapters (P7/P5) and a unique pair of sample barcodes (one index at either end of the molecule) to allow for multiplexing on the sequencing run.

3.5.3 Sequencing

BCR libraries were sequenced at GENEWIZ (Azenta) on an Illumina Miseq V3 flow cell using asymmetric sequencing with 400 cycles in read 1 and 200 cycles in read and both P5 and P7 index reads (dual-indexed), using custom read 1, read 2 and index 1 primers and index 2 was read directly off of the P5 adapter. Sequencing depth was assessed by rarefaction analysis of UMIs to determine UMI coverage - reads were subsampled to a random number of reads over a thousand iterations and the number of unique UMIs captured each time was counted and used to generate a species accumulation curve. Since curves had not plateaued, we decided to re-sequence the libraries on a second Miseq V3 run to obtain additional UMI sampling depth and coverage.

3.5.4 Data Pre-Processing

Sequencing data from both Illumina runs was combined for each unique library prior to running the data pre-processing and alignment steps, such that the two sequencing runs were treated as one. This was so that UMIs re-sequenced across the two runs for the same library would be collapsed down into the same consensus read. Fastq pre-processing and read alignment were performed with the pRESTO and changeO packages from Immcantation. In brief, low-quality reads (<Q20) were filtered and removed. Next the UMIs were extracted from read2 by matching to the template switch motif "TCTTGGG", extracting the preceding sequence and annotating the sequence with the UMI barcode. A consensus sequence built for each of the reads and read pairs were assembled. Constant regions were identified by matching "internal" sequences for IGHG and IGHM constant regions to the reads.

3.5.5 Repertoires Analysis

For all analyses, non-productive BCRs were filtered out by filtering for "productive == T" in the output .tsv files from IgBlast. BCRs with UMIs which were not 12,14 or 15 bp long were also excluded from the analyses. Technical replicates for each repertoire were combined and treated as one sample.

Diversity and Clonality

Each repertoire was down-sampled to 1500 unique UMIs over 1000 iterations and Renyi Entropy averaged across the iterations. Renyi entropy was calculated for alpha values in the range of 0-20, using the from dit.other "renyi_entropy" function in the "dit" python package (R. G. James, Ellison, and Crutchfield 2018).

For isotype-specific diversity analysis, repertoires were first subsetted to IgM or IgG based on the cregion call. Two samples with low IgG counts (=<101 UMIs) were excluded from both isotype-specific diversity analyses. IgM and IgG repertoires were subsampled to 700 UMIs over 1000 iterations and custom python scripts were used to calculate Simpson's Diversity, Shannon Entropy and Gini Index of Inequality in each iteration. The values obtained were averaged across all of the iterations. Statistical analysis was performed in R using the glmmTMB package. The diversity indices were calculated as follows:

Simpson's Diversity:

$$D = 1 - \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)}$$

Where *S* is the number of species, n_i is the frequency of each species, and *N* is the sum of the abundances of all species in the distribution.

Shannon Index:

$$H = \sum_{i=1}^{S} -(P_i \times lnP_i)$$

where S is the number of species and P is the proportion each species makes up of the population.

Gini Index of Inequality:

$$G = \frac{\sum_{i=1}^{S}(2i - S - 1) \cdot P_i}{n \cdot \sum_{i=1}^{S} P_i}$$

where S is the number of species and P_i is the proportion each species makes up of the population.

Clonotype Clustering

Clonotype clustering was performed by first grouping BCRs by V-J gene call and CDR3 length, and using the hierarchical clustering functionality in scikit.learn to assign CDR3s within a hamming distance of 0.15 amino acids to the same clonotype cluster. This analysis was performed to account for somatic hypermutation.

V gene usage

V gene usage was calculated by adding counts of functional BCRs which belonged to a particular V gene and dividing them by the total number of BCRs for that sample. Manhattan distance was calculated using the scikit-learn python package.

Somatic Hypermutation

Somatic hypermutation was calculated by counting the number of mismatches between germline and aligned BCR sequences, excluding any non-templated nucleotides in the junction, "N" bases and IMGT gaps in the sequence alignment.

Trajectory Analysis

Trajectory analysis was adapted from Minervina et al 2020. In brief, proportions of the top 1000 clonotypes across all timepoints were selected and normalised by dividing all other timepoints by the proportion of the largest clonotype. PCA and k-means clustering was then performed using the SciPy library.

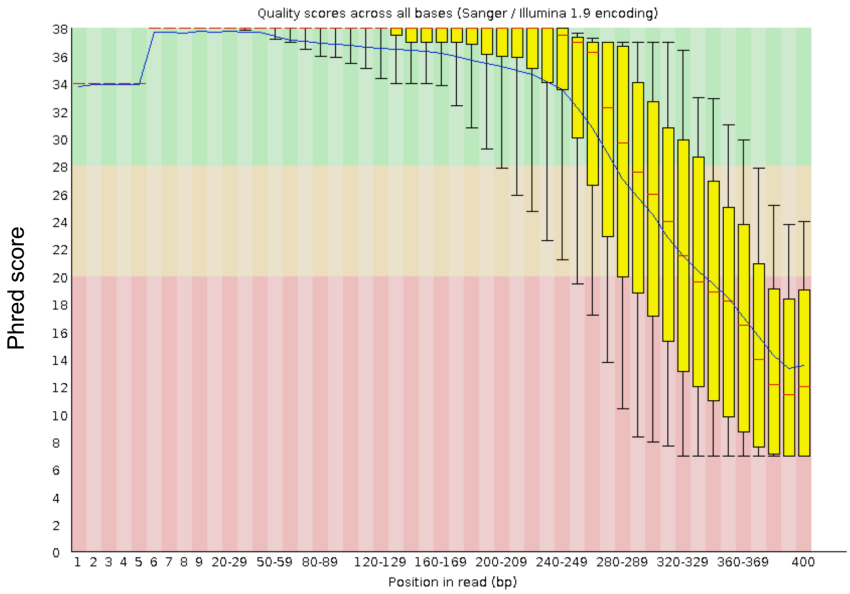
Mass Cytometry

Details of CyTOF sample collection, experimental protocol and data analysis are reported in F. A. Bach et al. 2021. We obtained .fcs files from this experiment and data were analysed using the FlowJo software using the gating strategy from the Spence Lab. We gated for 191Ir (intercalates with DNA) to label singlets, White cells (CD45+), lymphocytes (CD3+ and CD20+) and CD3+ cells to identify T cells and CD20+ cells to identify B cells. An additional gate was drawn on the CD20 population to identify the CD27+ B cells. Relative proportions of CD20+ and CD3+ cells were calculated and plotted.

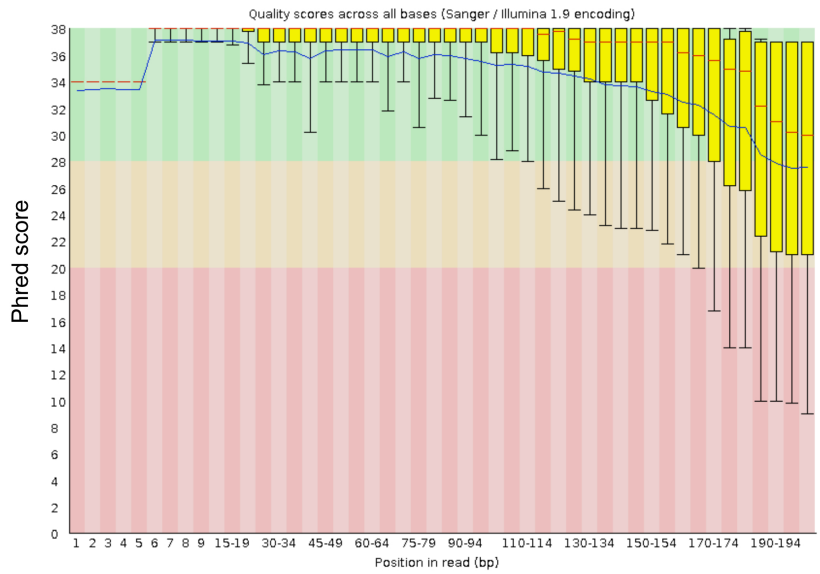
Statistical analysis

Mixed linear models were run using the R package glmmTMB, with the effect of individual treated as random and the effect of timepoint as fixed. Linear regression was performed using the python OLS functions in the scipy.stats package

Read 1 (400bp)



Read 2 (200bp)



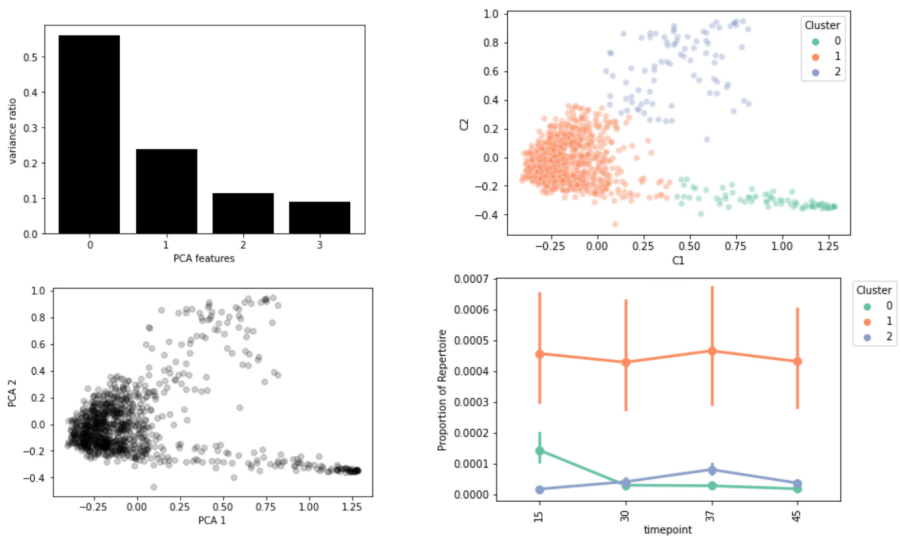
Supplementary Figure 1: Phred score plots for a representative sample (CHMI12-A)

Supplementary Figure 2A: Treemaps of top CDR3s IgM. Each tile represents one individual sampled at a specific timepoint. Repertoires were downsampled to 700 random UMIs. Size of the coloured blocks represent the proportion of the repertoire made up by that specific CDR3, coloured by its V gene (same V gene colour scheme used for each timepoint). Plots generated in plotly.

Supplementary Figure 2B: Treemaps of top CDR3s IgM (continued).

Supplementary Figure 3A: Treemaps of top CDR3s IgG. Each tile represents one individual sampled at a specific timepoint. Repertoires were downsampled to 700 random UMIs. Size of the coloured blocks represent the proportion of the repertoire made up by that specific CDR3, coloured by its V gene (same V gene colour scheme used for each timepoint).

Supplementary Figure 3B: Treemaps of top CDR3s IgG (continued)



Supplementary 4: Trajectory analysis proof of principle: TCR trajectories from Minervina *et al.* 2021 were replicated using custom functions and same functions were used to analyse BCR data shown in Figure 14.

Chapter 4

Optimising an Accessible High Throughput Protocol for Phage Display of Cognate BCRs

4.1 Introduction

One limitation of AIRRseq data is that when we observe signatures of response in the B cell receptor repertoire, such as clonal expansion, V gene usage skewing or even shared public responses, we cannot infer antigenic targets of these responses with confidence. While databases of curated TCR epitopes, such as VDJdb (Shugay et al. 2018) and IEDB (Vita et al. 2009) are now routinely incorporated into TCR repertoire analysis to infer TCR epitope specificity, equivalent databases of BCR specificities are very limited. Conversely, when it is possible to assign antigen specificities to BCRs in AIRRseq analyses, it enriches our understanding of adaptive immune dynamics in response to infection and can be used to identify monoclonal antibodies rapidly for therapeutic

purposes.

4.1.1 Low Throughput Methods to Link BCR Sequence Data to Antigen Specificity

The most robust method for characterising BCR-antigen interactions is by solving 3D structures of BCR-antigen complexes and identifying which amino acids are involved in the interaction, but this approach is costly and time consuming. Other, more commonly applied methods include sorting single cells with antigen baits prior to repertoire sequencing. Wardemann and colleagues have used the approach of labelling B cells with antigen baits and sorting antigen-specific cells into 96 or 384 well plates. They then perform barcoded single-cell RT and PCR of full-length heavy and light chains using a primer matrix of row and column specific barcoding primers (Tiller et al. 2008; Murugan, Imkeller, et al. 2015; Murugan, Buchauer, Triller, Kreschel, Costa, Martí, et al. 2018). Using this approach they demonstrated that the repertoires of atypical memory B cells (atMBCs) from malaria-experienced individuals produce broadly neutralizing *P. falciparum*-specific BCRs (Muellenbeck et al. 2013). By comparing BCR repertoires pre and post- HIV infection, Setliff *et al.* were able to identify public broadly neutralising antibodies from HIV infected individuals, which can also be found in infection-naïve individuals (Setliff, McDonnell, et al. 2018). Similar approaches have also been used in autoimmunity: Mueller and colleagues used tetramers with citrullinated peptide to capture and sort single autoreactive B cells into 96 well plates for heavy and light chain pairing and expression as mAbs. BCR sequencing these cognate pairs revealed a bias in V gene usage, mAbs generated from clonally expanded BCRs were cross-reactive towards various citrullinated peptides, and some of these autoreactive mAbs were shown to promote arthritis in mice (Titcombe et al. 2018). Zost et al. 2020, performed functional screening of antibodies prior to cloning using the Berkeley Lights' Beacon optofluidic system. Light is used to

transfer thousands of plasma cells into individual nanoliter-volume chambers (NanoPens). Antibodies which could block human ACE2 receptor binding to RBD were identified by incubating protein-conjugated beads in chambers adjacent to the NanoPens where plasma cells secreting antigen against SARS-CoV-2 Spike or RBD protein sequestered a fluorescent secondary antibody. They subsequently exported >200 antigen-specific cells of interest to 96-well plates for RT-PCR and BCR sequencing and successfully cloned 78 antigen-reactive mAbs.

4.1.2 High Throughput Methods to Link BCR Sequence Data to Antigen Specificity

Several high-throughput options exist to sequence antigen specific BCRs and, broadly, perform single cell reactions in emulsions or in micro-or nanowell plates. Cao *et al.* identified 8,558 SARS-Cov 2 binding BCRs of which 14 were found to be potent neutralising antibodies by sorting antigen-specific B cells from 60 convalescent patients prior to 10X Genomics single cell sequencing (Cao et al. 2020). They performed scRNA/VDJ sequencing and characterised the binding properties of these BCRs further by cloning and expressing cognate heavy and light chains as IgG and identified a mAb which showed therapeutic and prophylactic efficacy in mice. They also found that neutralising antibodies could be identified based on similarities of their predicted heavy chain CDR3 structures to those of SARS-CoV2-neutralizing antibodies. The 10X Genomics platform is also used in the LIBRA-seq workflow that multiplexes up to nine DNA barcoded antigens. VDJ-enriched single cell libraries with the antigen barcode are sequenced and BCR sequences are assigned to the known antigen barcodes (Setliff, Shiakolas, et al. 2019). Antigens are also fluorescently labelled to allow for antigen-specific cell sorting prior to encapsulation. An advantage of this approach is that it also allows profiling of the cell transcriptome and is now commercially available via BioLegend. Limitations of this approach are the high cost,

the fact that only few antigens can be multiplexed, and that BCRs need to be synthesised to characterise the binding properties of these BCRs such as affinity or cross-reactivity. DeKosky and colleagues have made use of custom microfluidics setups with flow-focusing devices to perform two rounds of emulsion-based reactions for cognate BCR pairing: One to encapsulate single B cells with mRNA capture beads and perform cell lysis, and a second to perform cDNA synthesis and OE PCR on the beads (DeKosky, Kojima, et al. 2015; DeKosky, Lungu, et al. 2016). A recent improvement on this approach has involved the use of a cell lysate resistant xenopolymerase (RTX) that can perform in-droplet RT and PCR to pair cognate heavy and light chains (Tanno et al. 2020). However, while they can sequence the cognate paired libraries, their primer and linking strategy does not allow for cloning and expression of the BCRs directly from linked product. In B. Wang et al. 2018 they applied this microfluidics approach to encapsulate millions of B cells into single cell emulsions and clone natively paired VH:VL libraries as Fab fragments in a yeast display system. They identified bnAbs against HIV-1, Ebola virus glycoprotein and influenza hemagglutinin. A limitation of this method is that, to identify the sequence of the Fab fragment, yeast colonies are screened individually and selected for sequencing to confirm the identity of the BCR. The high-throughput nanowell platform, SeqWell, has been used to identify epitope-specific TCRs at high throughput by performing single cell cognate pairing (Tu et al. 2019). This approach has not yet been applied to BCRs but would likely be a suitable platform.

4.1.3 Phage Display as a Powerful Platform for Identifying Antigen-Specific BCRs

Alongside hybridoma technologies, conventional phage display, pioneered by Nobel Laureates George Smith and Greg Winter, has been extensively used since the 1990s to identify monoclonal antibodies for research and clinical use. Therapeutic antibodies derived from

phage display are used to treat a wide range of conditions including autoimmune conditions (Adalimumab -anti TNFAalpha, Belimumab – anti BLyS), and cancers (Avelumab- anti PDL-1) (for a comprehensive list see Alfaleh et al. 2020). Random combinations of the FR1-FR4 regions of the heavy and light chains are linked by a flexible polypeptide linker, most commonly a (G4S)3 linker, and expressed as Single Chain Fragment Variable (scFv) in place of some of the five copies of the minor coat protein (pIII) on the surface of the M13 filamentous bacteriophage. These "combinatorial" libraries make use of the tremendous diversity generated by VDJ recombination to screen and select for antigen-specific scFvs by multiple rounds of panning of phage libraries against a target antigen (Marks et al. 1991). An advantage of this approach is that phages both express the scFv and contain the heavy and light chain sequence in the phagemid, which allows for functional characterisation and easy genetic sequencing. These scFvs can also be reformatted into Fab fragments or full antibodies in other expression systems. However, the diversity of scFvs generated by these combinatorial libraries is vast (up to 10^{11}, André et al. 2022). For therapeutic purposes, cognate libraries have shown to yield monoclonals with greater sensitivity and specificity than combinatorial libraries (Adler et al. 2017) and are thought to avoid selection of potentially self-reactive or unstable antibodies. Emulsion and nano-well based platforms to perform single cell reactions at high throughput provide an opportunity to obtain scFvs from single cells to preserve their natural heavy and light chain pairing and express them by phage display.

4.1.4 Scalable Single Cell Reactions Provide New Avenues for BCR-Antigen Mapping

The Cowan Lab wanted to perform high-throughput expression of cognate BCRs to identify the antigen-specificity of BCRs at scale. I initially set out to combine conventional phage display with emulsion-based RT-PCR methods to produce natively linked heavy and light

chains at high throughput. In Rajan *et al.* (2018), the authors report generating millions of natively paired scFv fragments by performing one-step RT-PCR in single cell emulsions generated using the Dolomite Microfluidics system (Rajan et al. 2018). We wanted to adapt this protocol to our repertoire sequencing projects in the lab. A similar approach has been used to identify monoclonal antibodies against influenza A and pneumococcus in yeast display systems (Adler et al. 2017). One limitation of these approaches is that they use costly glass microfluidics chips (£800) which allow for cells from only one sample to be processed at a time and are difficult to re-use as they require washing and treatment with hydrophobic agents between runs. Furthermore, the setup is time-consuming and requires some technical expertise to achieve the correct flow rates for oil and aqueous solution. Having searched the literature for microfluidics platforms, BioRad microfluidics chips, used in the QX200 ddPCR digital quantitative PCR platform, were an attractive option as they are less costly (£8/chip) and disposable with relatively high throughput (>180,000 droplets/chip). Our initial goal was to replicate the protocol used in Rajan *et al.* (2018) in a more user-friendly microfluidics system, however, I was not able to replicate their results reliably. This chapter demonstrates the optimisations and issues identified with the strategy used in Rajan et al. and lays the foundation for future improvements on the methodology currently being implemented by lab members.

4.2 Aims

- Replicate RT-PCR reactions for Overlap Extension as per Rajan *et al.* using the BioRad microfluidics system
- Demonstrate native pairing is maintained by co-encapsulating a clonal B cell line with healthy donor B cells and quantifying mis-pairing frequency

4.3 Results

4.3.1 Optimisations with BioRad Droplet Generator

The initial aim was to reproduce the workflow published by Rajan *et al.* (2018) (**Figure 1A**) but simplifying the workflow by repurposing the Biorad ddPCR platform (the QX200 microfluidics chips and droplet generator) to produce single cell emulsions instead of the Dolomite microfluidics system (**Figure 1B**). The proposed workflow consists of MACS sorting B cells and encapsulating a single cell suspension with RT-PCR reagents in nanolitre size droplets. Within the droplets, cells are lysed by heating to 50 degrees followed by cDNA synthesis and overlap-extension PCR. Emulsions are then broken, PCR1 product run on an agarose gel and products 650bp-1000bp size selected and extracted. A nested PCR using only 5' heavy chain primers and 3' light chain primers is then performed in solution to amplify the product. These products would subsequently be cloned into phagemids and expressed in *E. coli*. The resulting phage display libraries could then be selected on specific antigen arrays or used for BCR-antigen pull-down. We considered that using disposable and affordable microfluidics chips would make it easier to multiplex samples and require less technical expertise than the Dolomite microfluidics system that had been trialled in the lab by a previous student. After consultation with BioRad technical representatives, they agreed to lend us a droplet generator to test the protocol for several months. The platform is designed for quantitative emulsion based PCR but not validated for used with single cells. Upon obtaining the BioRad droplet generator, our initial objectives were:

1. Test whether stable and uniform emulsions containing single cells could be produced using the BioRad microfluidics system

2. If successful, produce scFvs from natively linked heavy and light chain in single cell emulsions

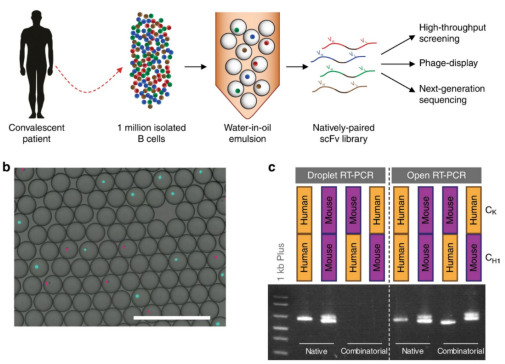
The primer strategy was adopted from Rajan *et al.* (**Figure 1C**) and makes use of gene specific primer pools. In brief, cDNA synthesis is performed using gene-specific primer cocktails (VH_in_3, VL_out_3 and VK_out_3) which bind to the J region of the heavy chain and the J gene-constant region splice junction of the lambda or kappa chains or within the first 50bp of the constant region. Following the RT step, PCR1 is directly performed in the droplets without the addition of any further reagents. A set of "inner" primers that have a complementary overlap introduce half of the (G4S)3 linker sequence to the 3' of the heavy chain at the end of the J gene and the other half to the 5' of the kappa/lambda chain at the first open reading frame of the V gene. These linker primers are added at 10 fold lower concentration than the "outer" primers which target the splice junction of the leader sequence at the 5' of the heavy chain and 3' end of the J-constant splice junction of the light chain. In PCR 2 primers are used to step into the start of the V gene of the heavy chain starting with the open reading frame and end of the J of the light chains and should yield in-frame scFvs.

4.3.2 Generating Stable & Uniform Emulsions in the BioRad Microfluidics System

In Rajan *et al.* the authors emphasised that of 10 RT-PCR kits they trialled only one could yield product in emulsion PCR. We therefore sought to adapt the published protocol to the Biorad microfluidics using the same reagents wherever possible. To test whether we could produce stable emulsions with the reagents used in the paper, in the BioRad ddPCR microfluidics system **(Figure 1 A,B)**; I compared emulsions generated with the BioRad RT-ddPCR reagents to those generated with the reagents used in Rajan et al., using either PBS or an "encapsulation buffer" (hypo-osmolar and high density buffer). Emulsions generated with the reagents from Rajan et al. were uniform and yielded slightly lower volumes to those produced with the BioRad reagents **(Figure 2 A, B)**, however,

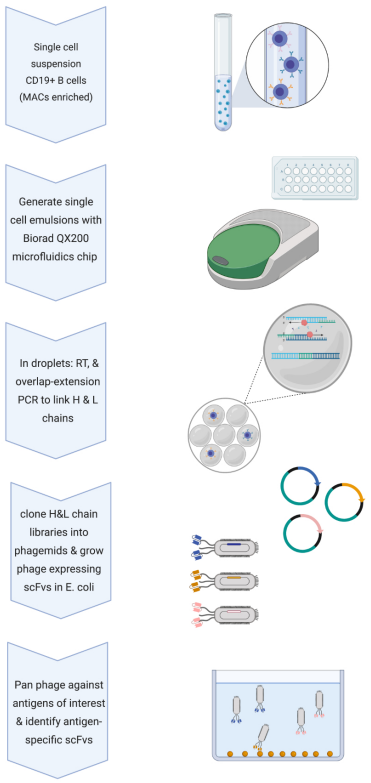
Figure 1 : Proposed workflow: Phage display of cognate BCRs (A) Workflow as described in Rajan *et al.* 2018, **(B)** adapted here for use with BioRad QX200 ddPCR microfluidics system. **(C)** PCR strategy adopted from Rajan *et al.* consists of one-step RT-PCR and overlap extension PCR in droplets occupied by a single cell. Complementary primers that bind to gene-specific regions and introduce the (G4S)3 linker are added at 10x lower concentrations than the "outer" primers so amplification of the full length, overlap extended (OE), product is favoured once linker primers have been exhausted. After size-selection and PCR cleanup, a second round of amplification using only 5' heavy chain and 3' light chain primers is performed in solution.

A

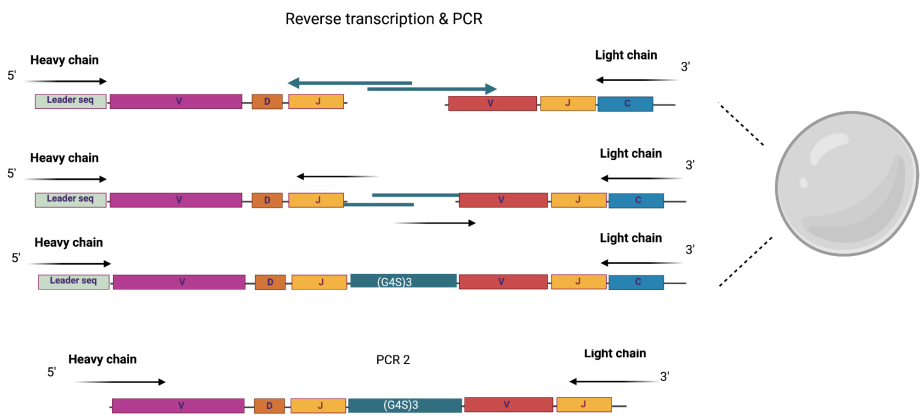


Rajan *et al* 2018

B



C



droplets showed signs of coalescence upon thermocycling **(Figure 2 D, E)**, compared to the BioRad reagents which remained intact **(Figure 2 C)**. The type of oil, surfactant, concentration of BSA, channel size and geometry and vacuum conditions all affect the uniformity and stability of emulsions. Having only had access to the droplet generator for a limited period, I decided to attempt optimisation of the single cell RT-PCR reactions using reagents from the Biorad ddPCR "RT-PCR Advanced Kit for Probes".

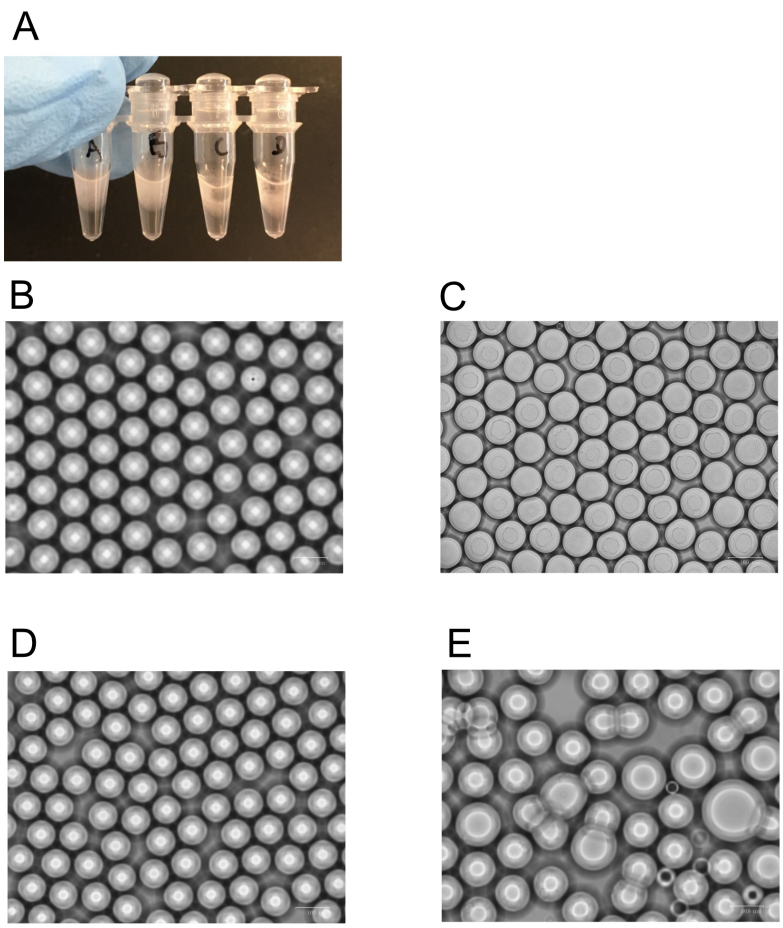


Figure 2 : Testing Emulsion Stability. Emulsions generated using DG8 cartridges and QX200 droplet generator with standard reagents or custom reagents were assessed visually. **A)** Emulsion (turbid fraction floating on top of oil) shown from left to right: "A,B"- BioRad RT-PCR reagents, "C,D" - Roche Titan RT-PCR reagents. Emulsions generated with BioRad regents imaged prior **(B)** and post thermocycling **(C).** Emulsions generated with Roche Titan RT-PCR reagents imaged prior to **(D)** and post **(E)** thermocycling.

4.3.3 Single Cells Can Be Encapsulated in Emulsions

To check that cells would not immediately lyse when they were mixed with the Biorad RT-PCR reagents prior to encapsulation, I incubated MACS isolated B cells for 15 minutes in PBS or RT-PCR mastermix and imaged them (**Figure 3A)**. We found that cells maintained their morphology and did not lyse in the RT-PCR buffer, although some cells showed signs of granularity. This did not raise significant concerns as cells had also been recently thawed and MACS sorted which may have caused stress and cells would only need to maintain their integrity in RT-PCR buffer in solution for 3-5 minutes to allow for encapsulation. To demonstrate that these cells can be encapsulated in droplets, primary B cells were stained with a live cell stain, resuspended in RT-PCR buffer and encapsulated in droplets in RT-PCR (**Figure 3B**). To confirm that cells could travel through the channels and be encapsulated in droplets using this chip I loaded stained live B cells and imaged them in droplets (**Figure 3 C**).

Loading appropriate cell numbers is important to avoid droplets being occupied by multiple cells, which would lead to mixing of heavy and light chains between different cells. As a rule of thumb, 1/10 droplets should be occupied by a cell. Each well produces approximately 23,000 droplets so the doublet rate at different cell loading concentrations can be estimated using a Poisson distribution. Modelling droplet occupancy in 23,000 droplets with 2000-20,000 cells **(Figure 4A)** demonstrated loading up to 3000 per well should result in less than 1% of droplets containing more than one cell **(Figure 4B)**. We loaded B cells at concentrations between 2000-20,000 cells and observed droplets with multiple cells when more than 3000 cells were loaded per well **(Figure 4C)**.

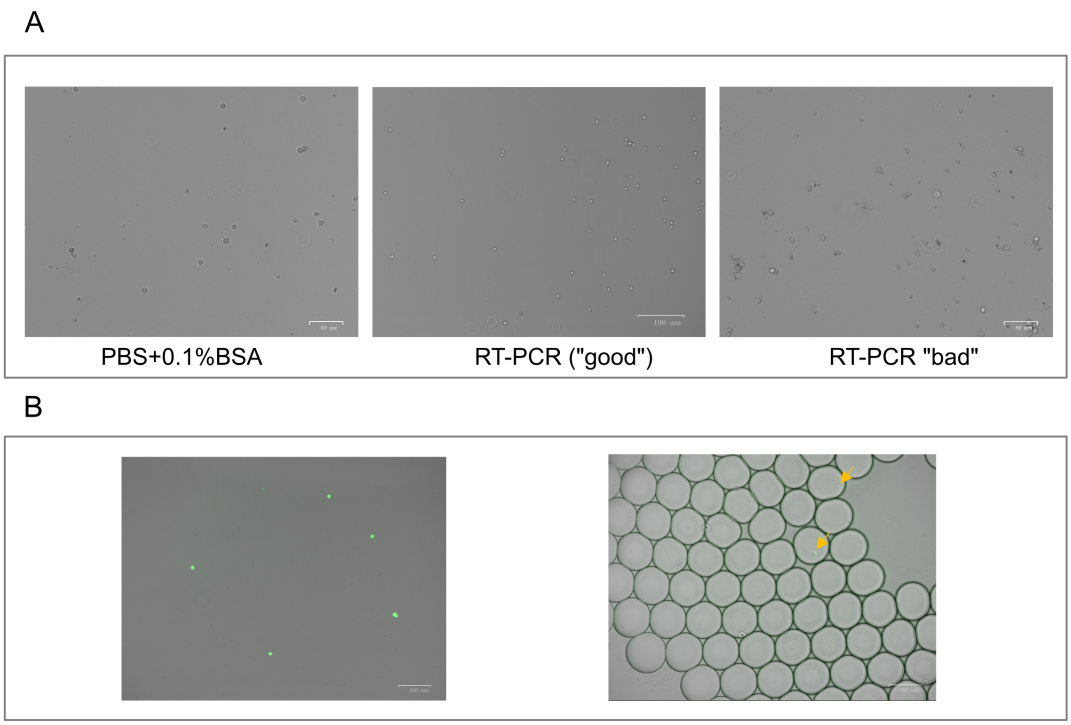
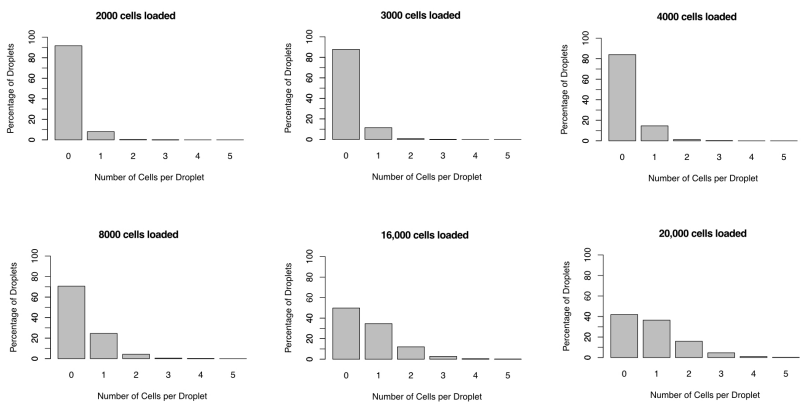
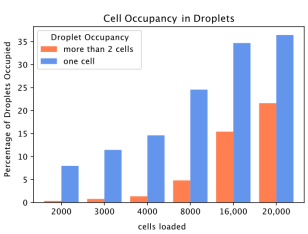


Figure 3: Single cell emulsions can be generated with BioRad QX200 system. B cells isolated from a healthy donor were incubated for 15 minutes in PBS +0.1% BSA or RT-PCR mastermix and imaged **(A)** - "good" and "bad" representative image shown. **(B)** Primary B cells labelled with CM Green (1:1000) in RT-PCR buffer in solution and encapsulated in droplets (highlighted with yellow arrows). (Scale bar 100μm.

A



B



C

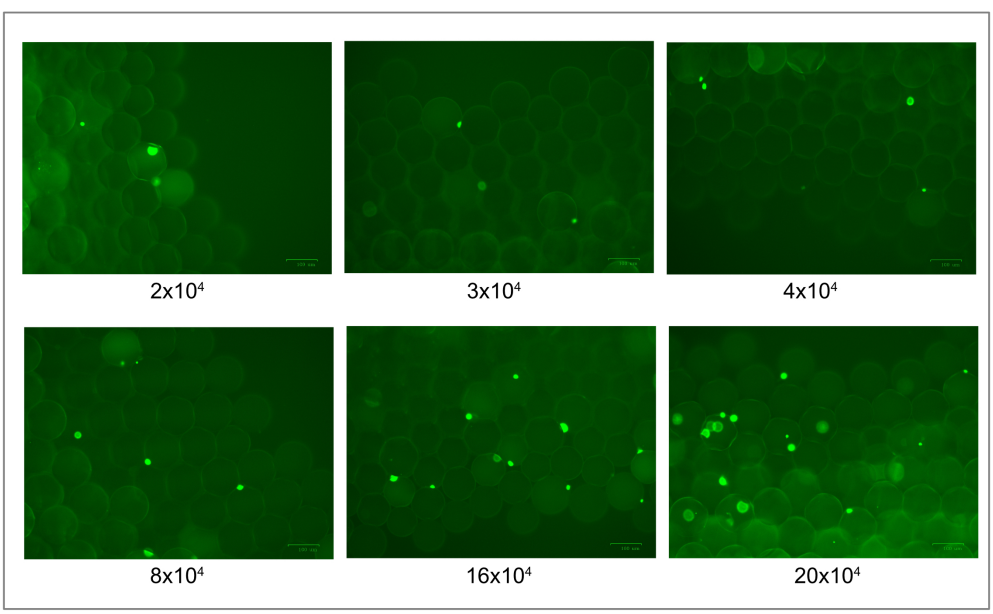


Figure 4: Determining optimal cell loading concentrations. (A) Modelling number of cells occupying each droplet for 2000-20,000 cells. A poisson distribution was used to model occupancy events (assuming 23,000 droplets in a well). (**B)** Droplet occupancy for singletons or more than one cell, across different numbers of cells loaded (assuming 23,000 droplets). (**C)** B cells were stained with CM Green, encapsulated in PBS at increasing cell loading concentrations per well. Images captured on a ZOE™ Fluorescent Cell Imager (BioRad). Scale bar 100μm.

4.3.4 Heavy and Light Chains, but No Linked Products, Obtained by Emulsion RT-PCR from Bulk RNA

Initial attempts at overlap-extension RT-PCR from single cell emulsions were unsuccessful and yielded no visible bands (data not shown). To simplify the workflow, I used bulk RNA from healthy donor PBMCs as template. Bands of the expected size for heavy (450-500bp) and light chains (300-400bp) were successfully obtained in emulsion RT-PCR from bulk RNA **(Figure 5A)**, however, overlap-extended product was not amplified. Products from successful heavy and kappa/lambda light chain amplifications were mixed 1:1 and overlap-extension attempted across a range of annealing temperatures in PCR2 **(Figure 5A)**. Since individual heavy and light chains were readily amplified, we hypothesised that some of the primers may be interacting and inhibiting the PCRs when both the heavy and light chain primer sets are present in PCR1. To check this, heavy and light chains were amplified in Reaction 1 and individual forward, reverse and forward+reverse primer mixes for the other chain spiked into the mastermix **(Figure 5B)**. None of the spike-ins inhibited individual amplification of heavy and light chains; however, overlap extended product was still not amplified.

Next, we hypothesised that the issue may lie with PCR2. To test whether the issue lay with the PCR2 primers that step into the V gene and start of the J gene, I amplified products from emulsion RT-PCR using the same 5' heavy chain primers and 3' light chain primers from the RT-PCR reaction in the second PCR, in solution. We also tried two different polymerases for amplification in PCR2, Q5 (NEB) polymerase and GoTaq (Promega). Performing the second round of amplification with reaction 1 primers revealed that a product of approximately the correct size for overlap-extended amplicon could be obtained using GoTaq. To confirm whether this amplicon contained overlap-extended product, the 600-800 bp band was excised and sequenced on an Oxford Nanopore MinIon

with a Flongle flow cell.

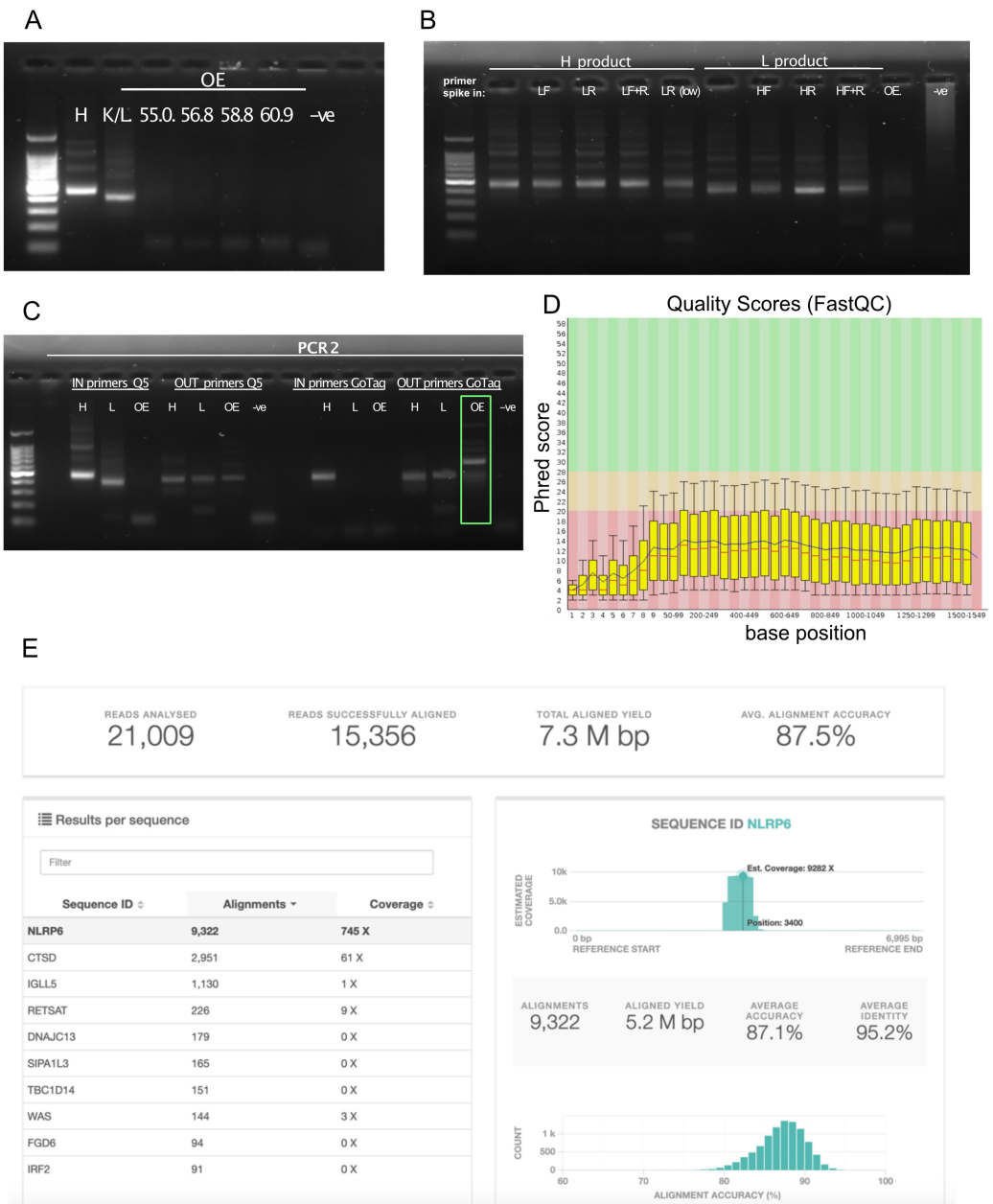


Figure 5: Troubleshooting reveals off target amplification in PCR1. Initial attempts to amplify individual heavy (H) and light (L) chains in emulsion PCR using bulk RNA as a template were successful. **A)** Overlap-extended (OE) from heavy and light chains was attempted across a range of annealing temperatures using (55-60.9ºC). **B)** Forward and reverse primer cocktails for light chains were spiked into heavy chain reactions and vice verse to check for primer inhibition in PCR1. Reverse L primers were added either at normal concentrations (416nM: LR) or lower concentrations (42nM: LR (low). OE amplification was also attempted as usual. **(C)** Second round of amplification was either performed with usual nested PCR2 primer cocktails (IN primers) or with the H forward and L reverse PCR1 primers ("OUT" primer) using either Q5 or GoTaq mastermix. "OE" product highlighted in green box was sequenced using an ONT Flongle flow cell. **(D)** Representative fastQC plot from sequencing run. **(E)** EPI2ME alignment results of products >500 bp.

4.3.5 Troubleshooting OE-PCR Using ONT Flongle Sequencing

BCR repertoires amplified by PCR produce highly diverse libraries of heavy and light chain amplicons making it difficult to confirm the products by Sanger sequencing as it requires a single template. This can be obtained by cloning products into plasmids, transforming bacteria and picking a single colony from which to amplify plasmid. Instead, I used Oxford Nanopore Flongle Sequencing as a convenient platform for troubleshooting PCR reactions because this enabled us to sequence the diverse pools of amplicon at relatively low cost (£72/flow cell) with a quick turnaround time since library preps and sequencing could be performed ourselves, enabling me to obtain sequence data from amplicon within 24h. A drawback of this approach is that Flongle sequencing (at the time) had a high error rate (up to 20% bases miscalled) (**Figure 4D**) and was particularly prone to skipping repeat bases, resulting in many frameshift errors being introduced by sequencing error. While the standard pipelines for ONT sequence analysis correct for the high error rate by building higher quality consensus reads by correction, trimming and assembly using tools like canu or aligning reads to a reference using minimap2, the recombinant sequences produced by VDJ recombination and the untemplated nucleotides in the CDR3 as well as somatic hypermutation pose a challenge to adapting these pipelines to adaptive immune receptor repertoire data. As a result, the analyses I performed with these data were rudimentary: their purpose was to provide a general overview of the contents of our amplicon libraries and troubleshoot unexpected results. Illumina sequencing would have been used to characterise scFv libraries for proof of principle experiments. We first aligned the sequences to immunoglobulin reference databases using IMGT-High V quest, using the scFv alignment parameter. Of 88646 reads, only 96 (0.1%) reads contained linked heavy and light chain products and of these 85% of the linkers were longer than expected

with a mean length of 91 bp and mode of 58 bp, suggesting they were the wrong product (see **Supplementary 1**). To identify whether any particular off-target genes were being amplified by our primer sets, sequences longer than 550 bp were aligned to the human exome reference database using EPI2ME and showed that 44% of reads mapped to the NOD-like receptor family pyrin domain containing 6 (NLRP6) gene on chromosome 11, followed by 14% reads mapping to the cathepsin D (CTSD) gene on chromosome 11 **(Figure 5E)**.

4.3.6 Lambda Light Chain Primers Amplify Off-Target Products in PCR1

Examining a few sequences manually and checking for matches with our primer sets revealed that lambda light chain primers VL_out_3_01 and VL_out3_03 appeared to be amplifying off target sequences **(Figure 6A)**. Using fuzzy string matching in python, I checked all of our primers against reads >550 bp, allowing for up to 3 mismatches to account for the high error rate of Flongle sequencing. If multiple primers matched the sequence, the match within the smallest edit distance was kept. This revealed that VL_out_3_01, VL_out_3_03 and VL_out_3_04 matched by far the largest number of sequences **(Figure 6B)**. Examining sequences which had both forward and reverse primer matches revealed that a large proportion of sequences matched to the VL_out_3 primers as both forward and reverse primers, with the combination of VL_out_3_01 and VL_out_3_03 being the most common **(Figure 6C)**. Selecting sequences with these primer matches and running them through BLAST **(Figure 6D)** confirmed that 78% of these reads aligned to the NLRP6 gene (gene ID NG_050573.1), followed by 20% of reads which aligned to CTSD (gene id NM_001909.5). This suggests that the lambda light chain primers were amplifying two off-target genes of a similar size to overlap-extended product. We tested the hypothesis that VL "out" primers were priming both sense and

antisense strands of an off-target product by amplifying PCR1 product from light chain or "overlap-extended" product using only the VL_out_3' primer cocktail, with no forward primers added. The VL_out_3' primers were sufficient to amplify a high molecular weight band of approximately 700bp from light chain and "OE" template **(Figure 6E)**.

Upon inspection of the primers and their target sites in the reference databases, they bound to the constant region of lambda light chain, in some instances with three of the four primers priming the same lambda constant region from different positions (**Supplementary Figure 2**). These observations led us to discard and redesign the VL_out_3 primer sets and to proceed with our optimisations using only the kappa chain primers for the limited time during which I had access to the droplet generator. At a later time, we redesigned lambda light chain primers to span the J-C exon-exon boundary to favour amplification from mRNA (**Supplementary Figure 3**).

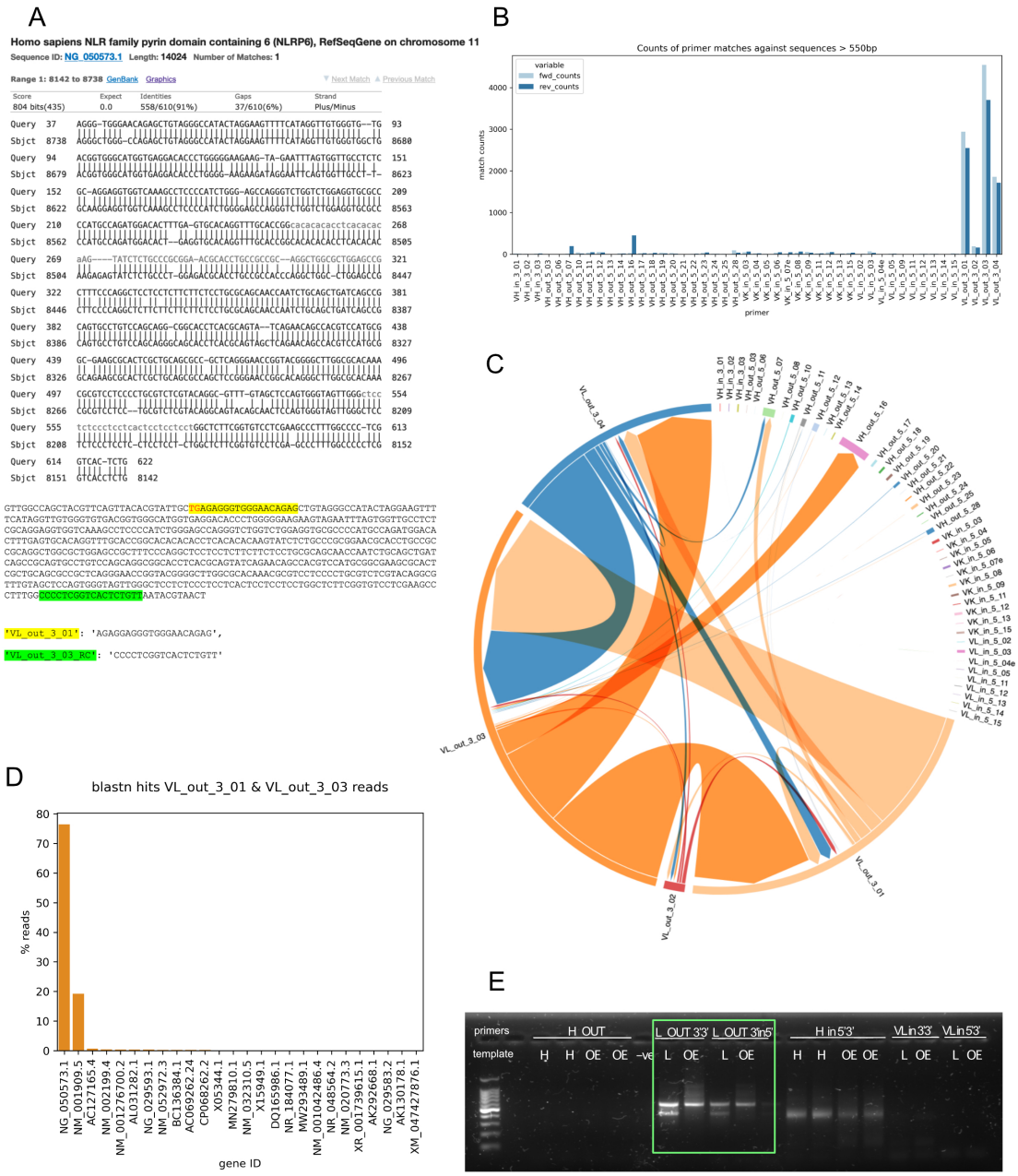


Figure 6: Flongle sequencing reveals off-target amplification from VL out primer set. (A) Example of sequence aligned to NLRP6 gene and corresponding forwards (yellow) and reverse (green) primer matches highlighted in sequence. **(B)** Barplot displaying counts of closest primer matches against sequences > 500 bp **(C)** Chord diagram displaying frequency of closest forward and reverse primer combinations found in sequences. **(D)** Top BLASTN hits for sequences with matches for VL_out_3_01 + VL_out_3_03 primer combination. Gene IDs: NG_050573.1 = "NLRP6", NM_001909.5 = "CTSD". **(E)** DNA gel confirming aberrant amplification of light chain OUT 3'primers using lambda light chain template (L) or "overlap-extended" template (OE). Products highlighted in green box.

4.3.7 Overlap Extended Bands Obtained from Heavy and Kappa Chains

Using only heavy and kappa light chain primer sets for PCR2, overlap-extended bands were successfully amplified by pairing previously amplified heavy and kappa chains (**Figure 7A**). We were also able to obtain OE products directly from emulsions containing single cells **(Figure 7B)**. For the latter, 16000 cells were loaded in total across 8 wells and RT-PCR products obtained from single cell emulsions were pooled, size-selected on a DNA gel (600-1000bp region excised). This size-selected template was used as input for PCR2 to enrich for product that had been linked in single cell droplets and avoid pairing of un-linked heavy and light chains in PCR2. We sequenced size-selected products from the combinatorial and single cell emulsions on Flongle flow cells and aligned them to reference databases using IMGT High V-quest. 17.2% of reads (3730 reads) in the combinatorial and 8.4% reads (5601 reads) in the single cell library were called as scFvs. In both the combinatorial **(Figure 7C)** and single cell libraries **(Figure 7D)**, a majority of scFvs were unproductive, or contained one chain with no identified rearrangement. Due to the propensity of ONT data to contain sequencing indel errors, the high sequencing error rate and lack of any error correction or quality filtering in our analysis, it was unsurprising that most of the sequences were "unproductive". The profiles of CDR3 functionality were comparable between the combinatorial and single cell libraries. Aligning reads from the single cell library to the human exome using EPI2ME demonstrated the off-target amplification of NLRP6 and CTSD was avoided, with only 95/60466 reads aligning to the human exome database (data not shown). Despite the single cell and combinatorial libraries being obtained from the same healthy donor, the V gene frequencies for heavy-kappa chain combinations showed only a moderate positive correlation ($r = 0.7$) between the two libraries **(Figure 7E)**, with the single cell library displaying skewing towards

certain VH-VK combinations. While this could potentially be explained by enrichment of specific native heavy and light chain combinations in the single cell library and random pairing in the combinatorial library, these results are only preliminary and cannot be taken as definitive evidence due to the high error rate of the sequencing data and fact that input cell numbers were not matched.

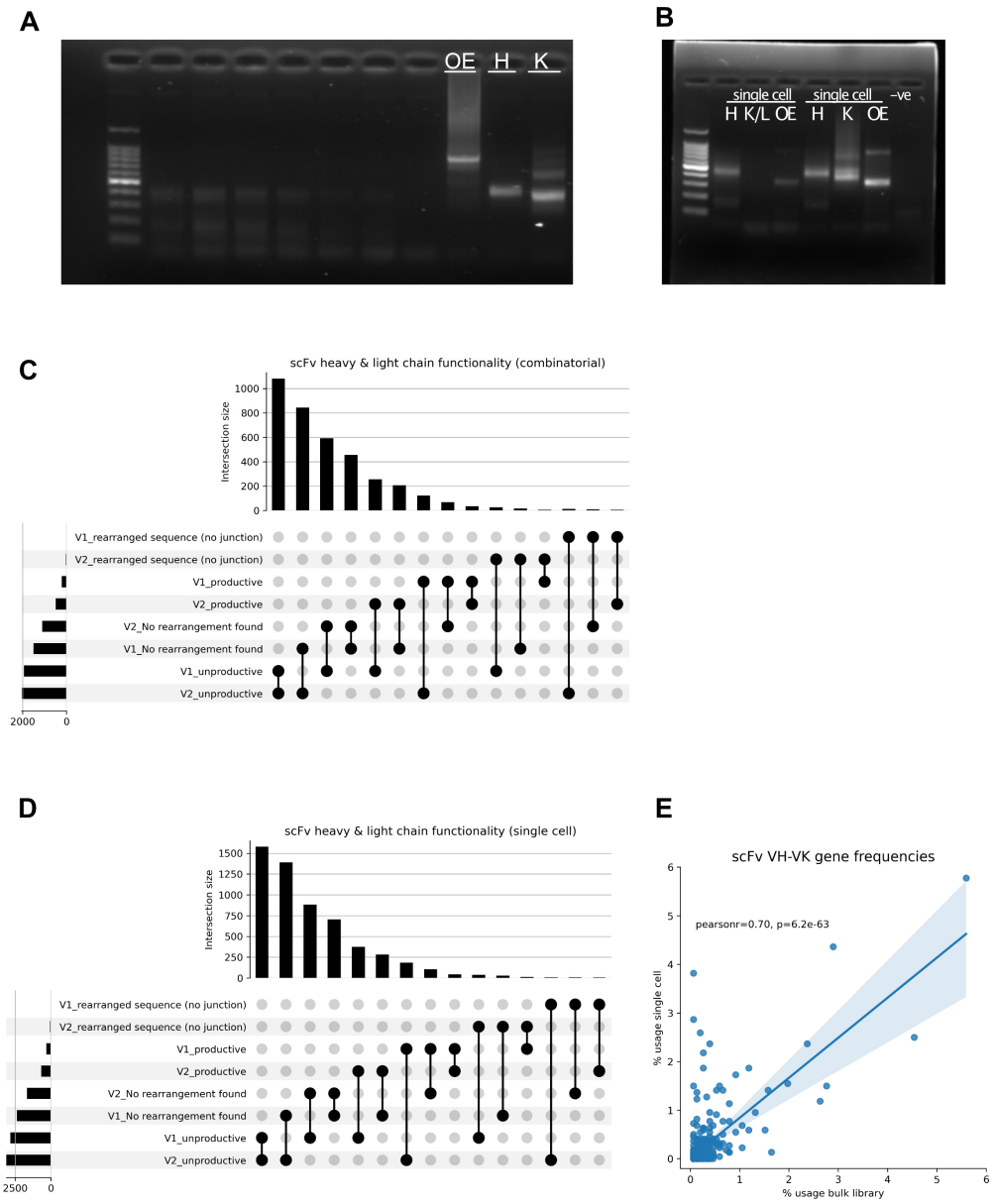


Figure 7: Overlap-extended product amplified from combinatorial and single cell emulsions. **(A)** Heavy and kappa chains paired by OE-PCR.**(B)** OE product obtained from single cell emulsions. (**C**)Upset Plots displaying variable segment functionality for combinatorial and **(D)** single cell libraries. **(E)** Correlation between VH-VK gene frequencies from scOE and combinatorial library from the same donor.

4.3.8 Characterising Overlap-Extended Products Obtained from Single Cell Emulsion RT-PCR

Next I sought to check whether the overlap-extended products from the single cell emulsion library contained the correct product. The sequences were approximately 800 bp long **(Figure 8A)** and heavy and light chain alignment lengths were distributed around 300 and 350bp as expected **(Figure 8B)**. We identified the (G4S)3 linker tag in sequences using fuzzy string matching (allowing for 0.2 error rate) in python. The (G4S)3 tag was located around 400bp on the sense strand and 300bp on the antisense strand **(Figure 8C)**. The scFvs made use of diverse heavy and kappa chain genes **(Figure 8D)**. Overall, this data was consistent with what we expected a library of overlap-extended heavy and light chains to look like.

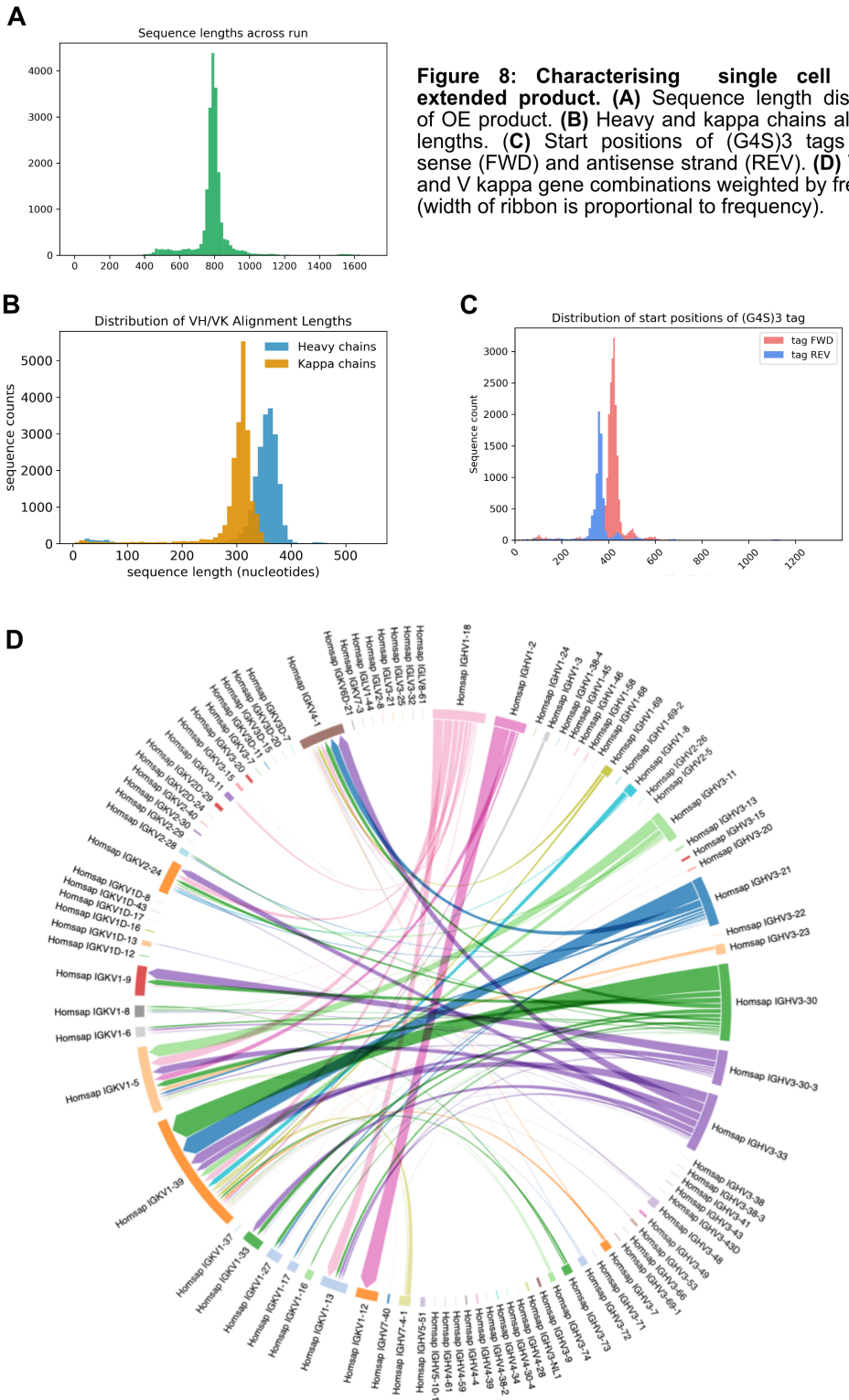


Figure 8: Characterising single cell overlap extended product. (A) Sequence length distribution of OE product. **(B)** Heavy and kappa chains alignment lengths. (**C**) Start positions of (G4S)3 tags on the sense (FWD) and antisense strand (REV). **(D)** V heavy and V kappa gene combinations weighted by frequency (width of ribbon is proportional to frequency).

4.3.9 Investigating "Unrearranged" BCR Sequences

In both the combinatorial and single cell libraries a majority of scFvs contained heavy or light chains with "no rearrangement", which raised concerns that V genes may have been amplified from un-rearranged genomic loci. As our reactions are performed in the presence of genomic DNA, it is possible that our primers could amplify from gDNA rather than cDNA. To check whether any particular V genes were being amplified from gDNA by our primers, I examined V gene usage in sequences according to whether they were rearranged or not **(Figure 9A,B)**. No particular V genes were over-represented in unrearranged sequences and the most abundant V genes (IGHV3-21, IGHV1-18, IGHV3-30, IGHV3-33) also had the most rearranged and unrearranged sequences. V gene counts showed a positive correlation between BCRs called as rearranged and unrearranged (**Figure 9C**). We hypothesised that if unrearranged heavy chains had been linked to light chains by PCR, the position of the (G4S)3 linker sequence would be further upstream in the sequence. Subsetting our reads to sequences identified as having "no rearrangement", I found that 64% of them still contained the (G4S)3 tag and the start positions were located at the anticipated sites **(Figure 9D)**. Upon aligning a handful of these "unrearranged" sequences using IgBlast, VDJ calls were identified **(Figure 9E)**, however the majority of alignment was out of frame suggesting the issue may be due to differences in how IMGT and IgBlast handle indels rather than amplification of unrearranged genomic template.

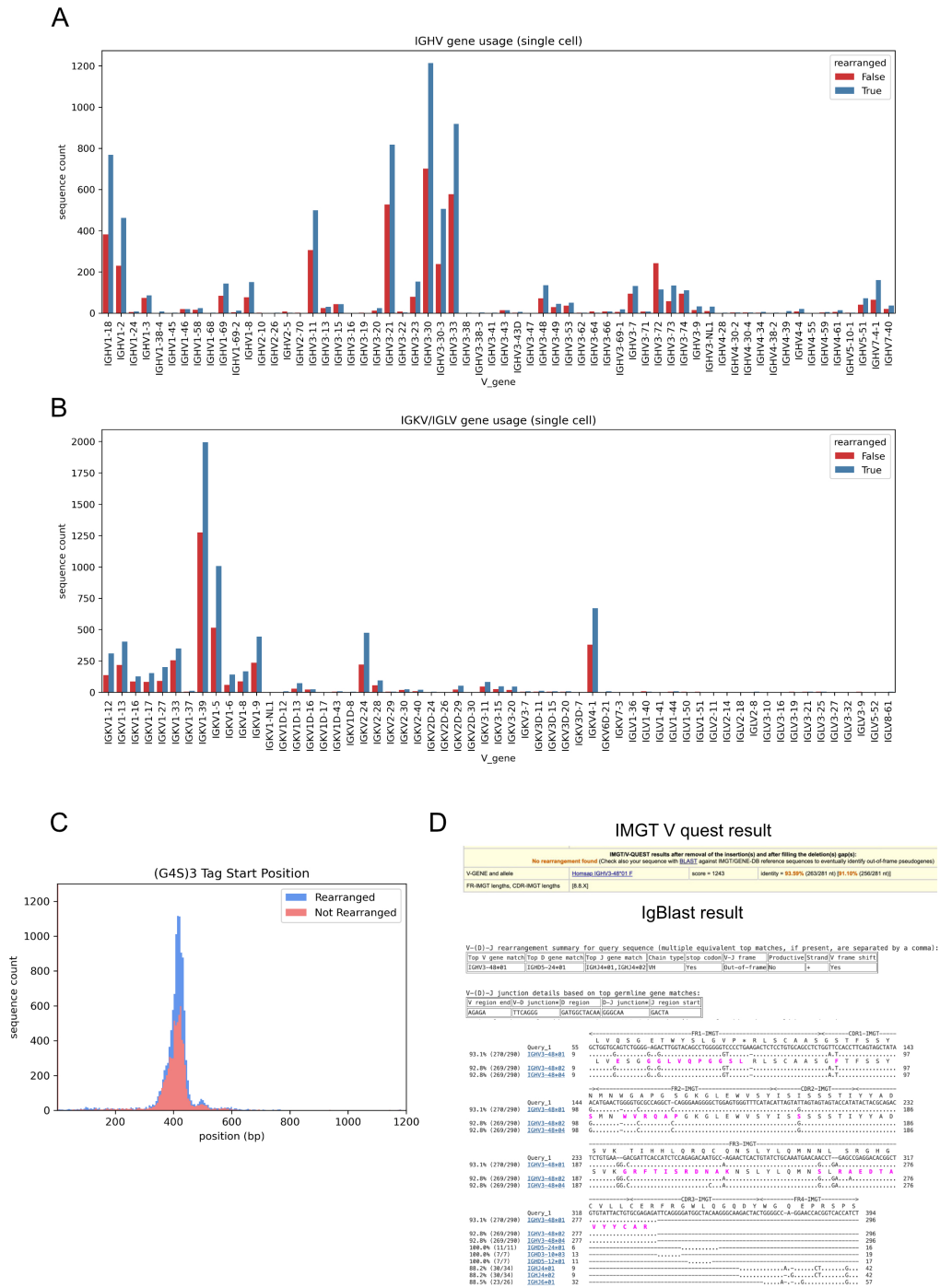


Figure 9: Rearranged and Non-Rearranged segments linked as scFvs **(A)** V gene abundance for single cell OE libraries by rearrangement status for heavy chain **(B)** and light chain. **(C)** Distribution of (G4S)3 start positions according to whether sequence was identified as "rearranged" or "not rearranged" by IMGT. **(D)** Example alignment results from sequence identified as having "no rearrangement" by IMGT and same sequence with CDR3 identified by IgBlast.

4.3.10 Raji Cell Line BCR Identification

Our planned proof of principle experiment to demonstrate that our protocol maintained cognate pairing, was to mix a diverse B cell suspension from a healthy donor with a clonal B cell line and quantify the rate of heavy and light chain mispairings. In anticipation of this experiment, we obtained a cancer transformed B cell line (Raji), however, we could not find information on what BCR the cell line produces (if any). Therefore I performed cDNA synthesis and PCR amplified the heavy and light chains from RNA isolated from cultured cells and sequenced the libraries. In the sequencing data, 84% of heavy chain BCRs were called as IGHV3-21 and 84% to IGHJ4 **(Figure 10A,B)**, while 93% of light chains aligned to IGKV3-20 and 49% to IGKJ2 **(Figure 10C,D)**. Sequence logos based on amino acid frequency were generated using kpLogo for the heavy and the light chain. Junctions with stop codons were filtered out and sequences were subsetted to those with the average amino acid length (21 aa heavy chain, 12 aa light chain), to account for the frequent frame-shift mutations in the data. The consensus amino acid CDR3 for the Raji heavy chain is "CARQRNDFSDNNSYYSNFDFW" and for the light chain is "CQQYASSTLFTF" **(Figure 10E,F)**.

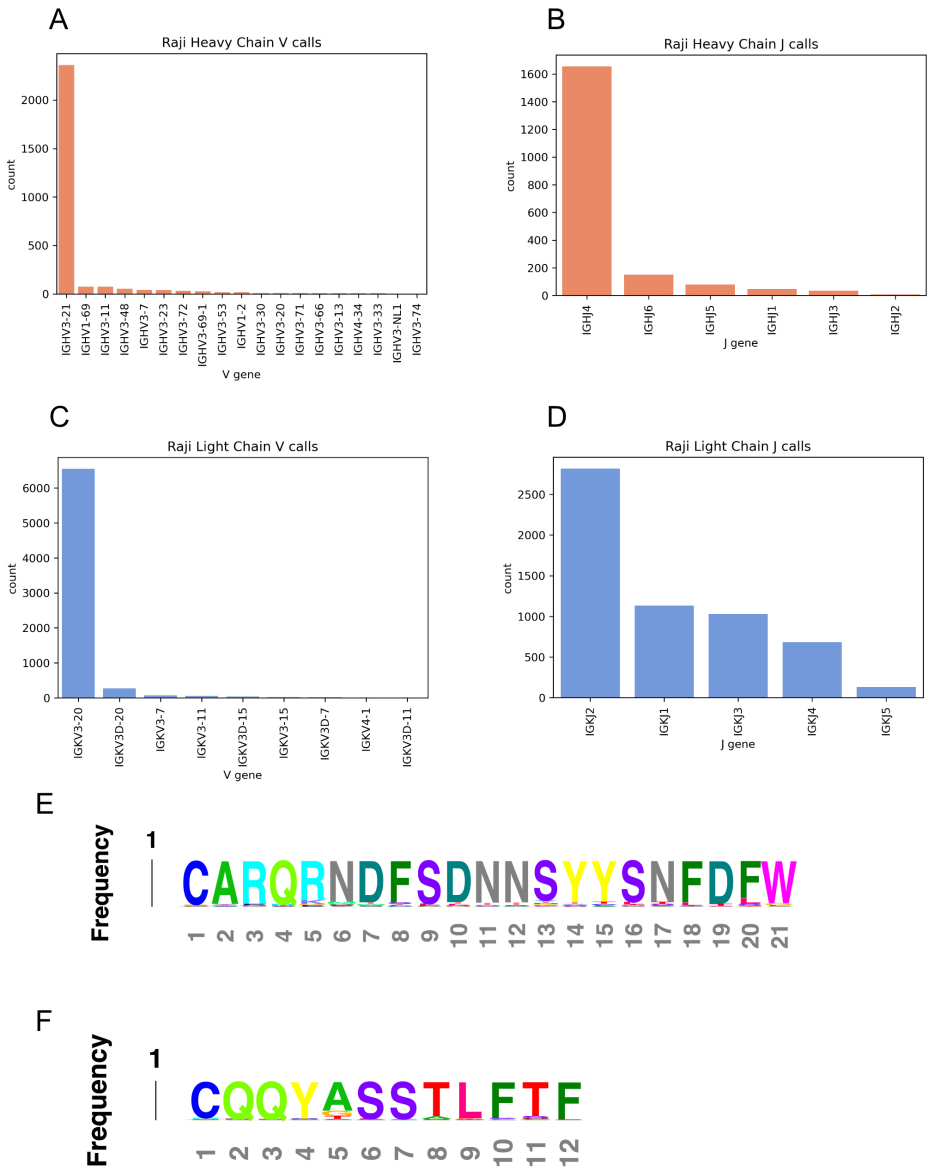


Figure 10: Confirming BCR identity of Raji Cell Line. Raji cells were sequenced on a Flongle flow cell and sequences aligned to reference databases using IMGT High V quest. **(A)** V gene and **(B)** J gene calls for heavy chains were tallied and genes represented by more than 2 reads were plotted on a barplot. Light chain V calls **(C)** and J calls **(D)** were also analysed. Amino acid sequence logos for heavy chain CDR3s **(E)** and light chain CDR3s **(F)** were generated using kpLogo (Xuebing Wu 2017).

4.3.11 Building a Custom Droplet Generation System

Shortly after the successful amplification of overlap-extended product with heavy and kappa chains from a single cell emulsion, the droplet generator had to be returned to BioRad before we were able to conduct this proof of principle experiment. Due to coronavirus working restrictions in place at the time, I was unable to access labs in other institutes in Edinburgh with ddPCR droplet generators. Instead, we decided to produce our own emulsions using the commercially available BioRad microfluidics chips, oil and RT-ddPCR reagents. An added benefit of this approach is that other labs attempting to reproduce this protocol would not be required to purchase a droplet generator (£20,000) for this purpose. Although the mechanism of droplet generation for BioRad ddPCR system is proprietary, we studied the microfluidics chip and other similar systems and hypothesised that applying a vacuum to the outlet wells should pull the oil and aqueous solution through the microfluidics channels across a T-junction to produce emulsions. A simple attempt at applying a vacuum to one of the outlet wells using a syringe demonstrated that emulsions could be made by this method **(Figure 11A)**. However, the emulsions were uneven, and it was difficult to control the vacuum applied with this method. Using Computer Assisted Design software (AutoDesk Fusion 360) I designed a prototype manifold and tray that would fit over the microfluidics chip and hold the rubber gasket in place to allow a sealed vacuum to be applied to the outlet wells **(Figure 11B, C)**. Prototypes were 3D printed and several adjustments were made to the original design to improve how it sealed off the outlet wells. The final version, the MANifold And Tray for Easy Emulsions (MANATEE) was 3D printed from Shapeways using selective laser sintering (SLS) with high grade thermostable and durable nylon to achieve a smooth finish **(Figure 11D)**. Initially a low-level vacuum was generated using a syringe and emulsions successfully generated **(Figure 11E)** however, I found that the droplets produced were still variable in terms of droplet size **(Figure 11F)**. Because cells need to be encapsulated at a predictable

rate (1 in 10 droplets occupied by cells), the number of droplets generated needs to be uniform to ensure the appropriate concentration of cells is loaded.

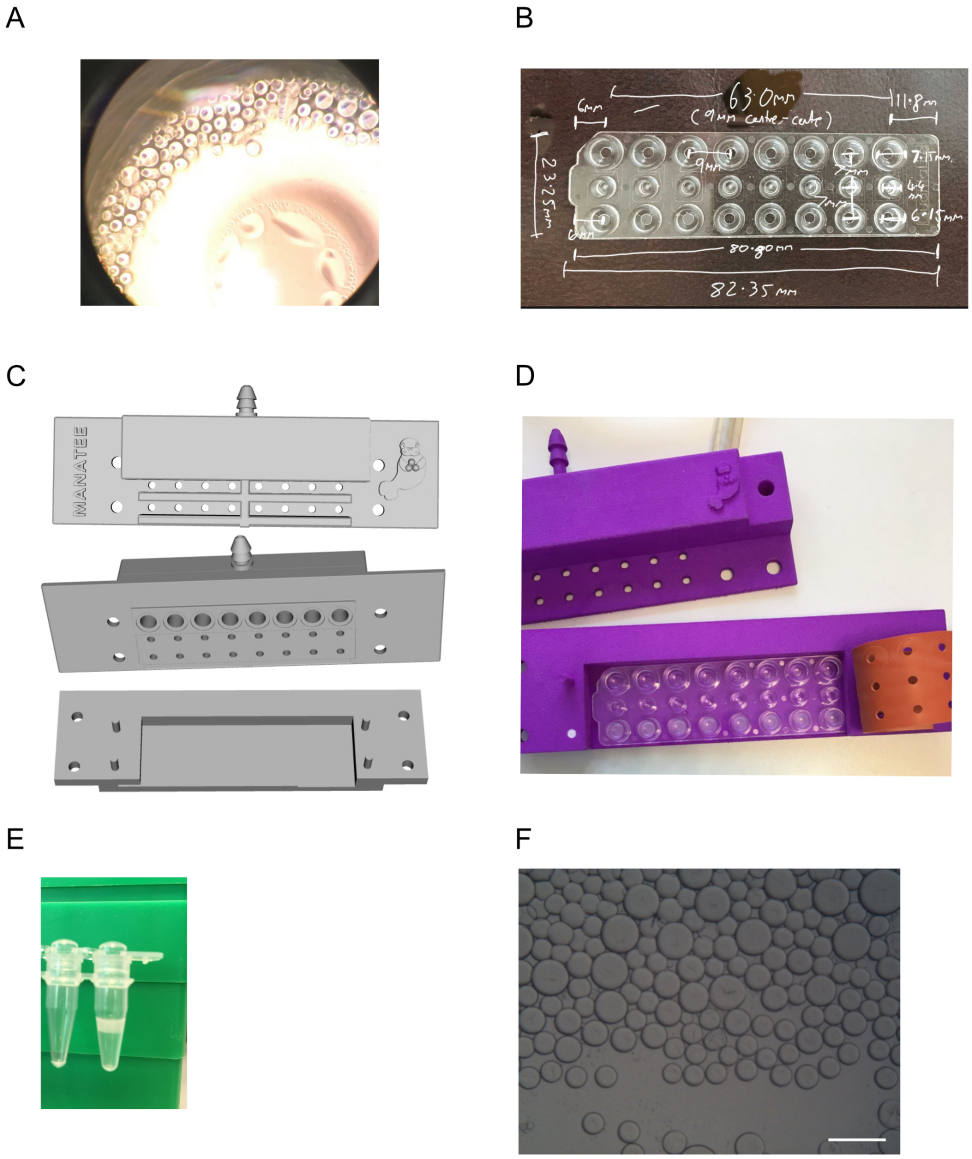


Figure 11: Designing "MANATEE" droplet generation setup.
(A) Image down the lens of a microscope of emulsion generated by applying a vacuum to a single outlet well of the BioRad Qx200 microfluidics chip. **(B)** Dimensions of BioRad QX200 microfluidics chip (from Graeme Cowan). **(C)** "MANifold Adaptor and Tray for Easy Emulsions" (MANATEE) was designed to fit the dimensions of the Biorad QX200 microfluidics chips: A tray and manifold were designed using CAD software to produce a sealed vacuum chamber over the outlet well of the chip and hold the rubber gasket in place over the wells**. (D)** 3D printed MANATEE with Biorad QX200 microfluidics chip and rubber gasket (generated emulsions in the bottow wells). **(E)** Emulsion containing BioRad RT-PCR reagents was generated by applying a vacuum to the MANATEE with a syringe and **(F)** droplets imaged (scale bar 200 microns).

To achieve better control over and monitor the vacuum conditions, a simple vacuum monitoring and release system was assembled and the hardware controlled using an Arduino (an open-source electronics platform)(**Figure 12A**). The circuit consists of a pressure gauge which measures the difference in vacuum between the MANATEE and atmospheric pressure, and a valve which is controlled by the Arduino and connected to the vacuum source and the MANATEE. When the valve is powered on, it blocks the vacuum source, and when it is powered off it releases vacuum into the system, allowing the vacuum levels to be controlled and leakage to be compensated. The pressure gauge is a variable resistor (ie output voltage changes according to the pressure differential detected) and was wired and calibrated according to the manufacturer's datasheets (**Figure 12B**). To confirm the gauge was working, the output voltage from the pressure gauge was measured with a multimeter and compared across different levels of vacuum applied (**Figure 12C**). To check that the measurements and calibrations were accurate, the theoretical changes in vacuum with different volumes of air removed were calculated using Boyle's Law and compared to experimentally measured pressure when equal volumes of air were removed with a syringe (**Figure 12D**). The thresholds at which the solenoid valve is switched on (block vacuum) or off (release vacuum) can therefore be set by the user and allowed us to trial a range of different pressures. This served as a low-cost solution ($<$£15) to test whether optimising vacuum conditions would yield more consistent emulsions. In the final setup, the MANATEE and vacuum control circuit were attached to a plate-washer (but any vacuum source would work). This allowed for the vacuum applied to the microfluidics chip to be maintained consistently, compensate for vacuum leakage and allowed us to test many different vacuum conditions when trialling PCR reagents with different properties and viscosities.

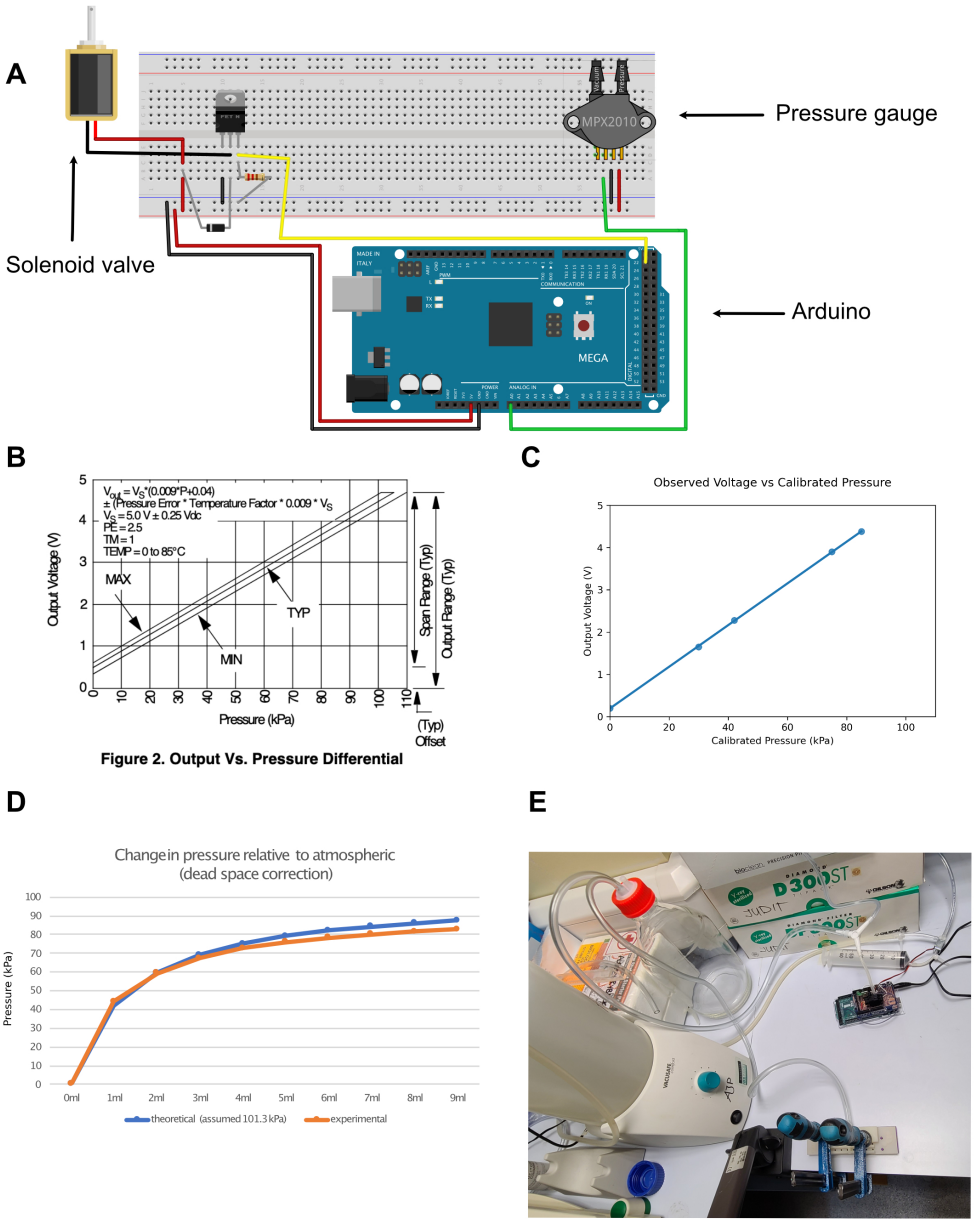


Figure 2. Output Vs. Pressure Differential

Figure 12 Controlling low-level vacuum in MANATEE. (A) Arduino-controlled pressure monitoring and vacuum release circuit were designed to control low level vacuum in MANATEE system. **(B)** Pressure gauge sensor output signal (voltage) relative to pressure input from data sheet (https://www.farnell.com/datasheets/37723.pdf). **(C)** Observed pressure readings according to multimeter voltage readings to confirm correct calibration. **(D)** Calibration was validated by comparing theoretical vacuum applied by removing increasing volumes of air with a syringe (calculated using Boyle's Law) to experimentally observed pressure at these volumes (assuming 101.3kPa atmospheric pressure with a 0.404ml dead space correction). **(E)** Final setup with vacuum source (plate-washer), Arduino-controlled pressure release system and MANATEE (clamped to the bench).

4.3.12 Validating Emulsions Generated Using MANATEE

Emulsions generated using the MANATEE system were evaluated for droplet uniformity by microscopy and the volume of the emulsion generated, measured by volume taken up in pipette. Droplet uniformity and emulsion volume still varied between wells across individual microfluidics chips and across different wells. A large Duran bottle was incorporated to buffer the vacuum and the addition of sealing grease at the junctions and tube connections greatly improved the consistency of the emulsions. Pressures from 5kPa-50kPa were tested and applying an 8 kPa vacuum yielded the most uniform emulsions using the BioRad RT-PCR reagents. Comparing droplet volumes pre and post optimisation of the vacuum system setup **(Figure 13A, B)**, and droplet uniformity **(Figure 13C, D)** demonstrated the optimised vacuum conditions yielded consistent and uniform emulsions **(Figure 13E)**. Single B cells were again successfully encapsulated into emulsions using this system **(Figure 13F)**.

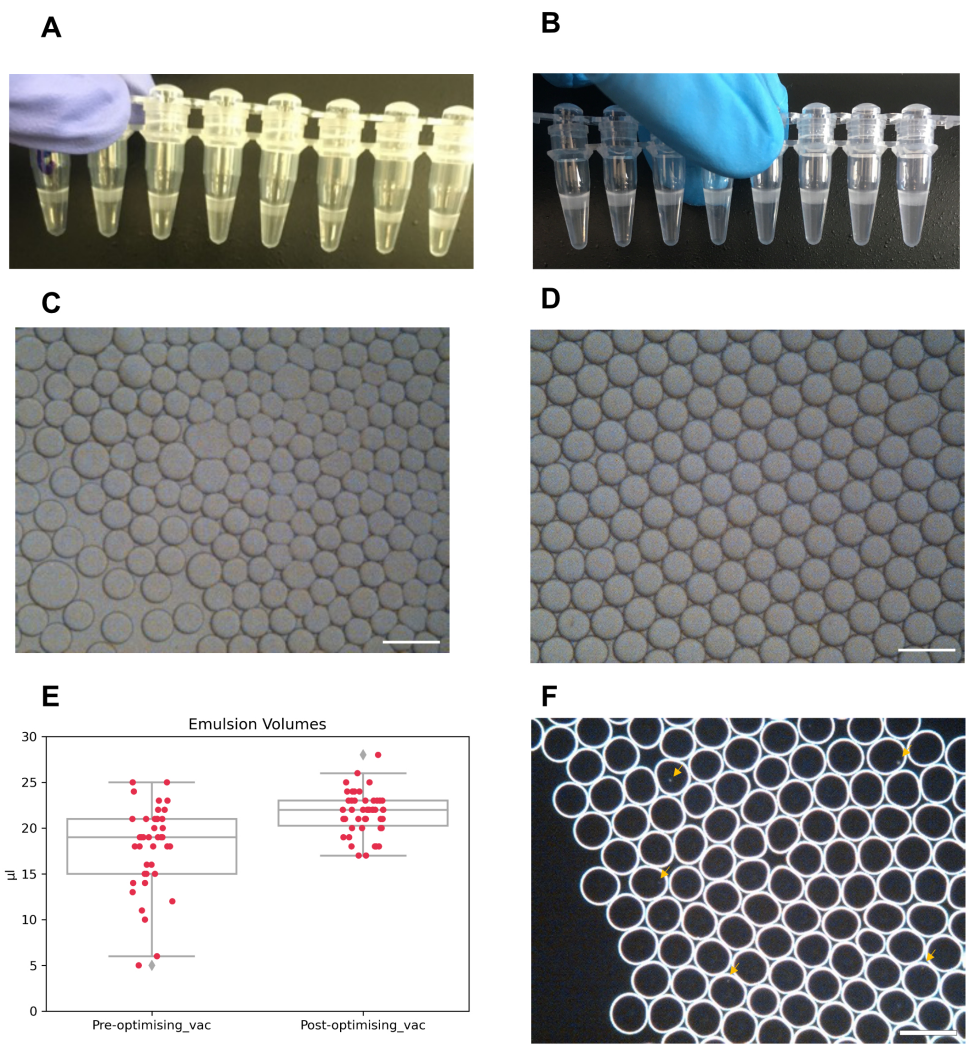


Figure 13: Optimising MANATEE and vacuum conditions. Representative emulsions pre- **(A,C)** and post **(B,D)** optimisations. Droplet uniformity was assessed visually and emulsion volume measured using a pipette.**(E)** Emulsion volumes recorded for 40 emulsions pre-optimisation and 46 emulsions post-optimisation of vacuum conditions. **(F,G)** Primary B cells encapsulated in droplets (in PBS) at 2500 cells/ well and imaged with phase contrast - individual cells highlighted with yellow arrows. Scale bar: 200 µm.

4.3.13 Troubleshooting BioRad RT-PCR Reactions

Having validated the new setup for generating emulsions, I tried to replicate our single cell RT-PCR reactions, without success. Given the issues I encountered with VL_OUT_3 primers, I revisited the remaining primers adopted from Rajan *et al*. The TMs of the primer sets were mismatched (54°C – 62°C) and some primers were redundant. For example, VH_out_5_10: GGTGGCAGCTCCCAGATGG, is identical to VH_out_5_12: GTGGCAGCTCCCAGATGG except that it is one base longer at the 5' end. We extended the VH OUT forward and VK OUT reverse primers by several bases to increase and match the TMs of these primer sets. Primers were validated in silico by matching against all recorded V genes in the IMGT reference databases. Primer sets that introduce the linker sequence (VH_in_3 and VK_in_5) were also redesigned to incorporate a (G4S)4 tag with a longer overlapping sequence (25 bp, TM 72°C). When testing these primers, I found that for the heavy chains the extended VH_OUT (VH_OUT_n) primers produced brighter bands but the new linker primers, VH_in_3, prevented amplification, likely because secondary structures are formed due to their length **(Figure 14A)**. We therefore proceeded with using the new VH_OUT primer sets and original linker primers. We were able to generate heavy chain and kappa light chain products (**Figure 14B, C**), and pair these products by overlap-extension in emulsions (**Figure 14D**). We could also obtain overlap-extended product from bulk RNA in droplets and attempted to improve the yield of overlap-extended product by performing a temperature gradient of the reverse transcription step **(Figure 14E)**. We tried reducing the concentration of linker primers (VH_in_3 and VK_in_5 primer pools) so these would be exhausted more rapidly and favour PCR priming from the complementary (G4S)3 linkers **(Figure 14F)**. We tried adding a single stranded DNA binding protein to reduce non-specific amplification during reaction-set up **(Figure 14F)**. We also attempted to increase yield in PCR2 by optimising the Magnesium Chloride concentration **(Figure 14G)** and found 1uM to produce the

brightest bands. However, the reactions from single cell emulsions consistently failed and even the RT-PCR from bulk RNA proved variable despite using single-use aliquots for all reagents and primers. We hypothesised that product may be lost during the gel extraction step between PCR1 and PCR2 and found that, when amplifying from bulk RNA, the overlap-extended template was lost during the gel extraction step **(Figure 14H)**. Even when omitting the size selection step after PCR1, I did not obtain overlap extended product from single cell emulsions (blank gel- not shown). Our conclusion was that it is possible for overlap-extension to occur from RNA and even single cells, however the reactions may be extremely inefficient and variable.

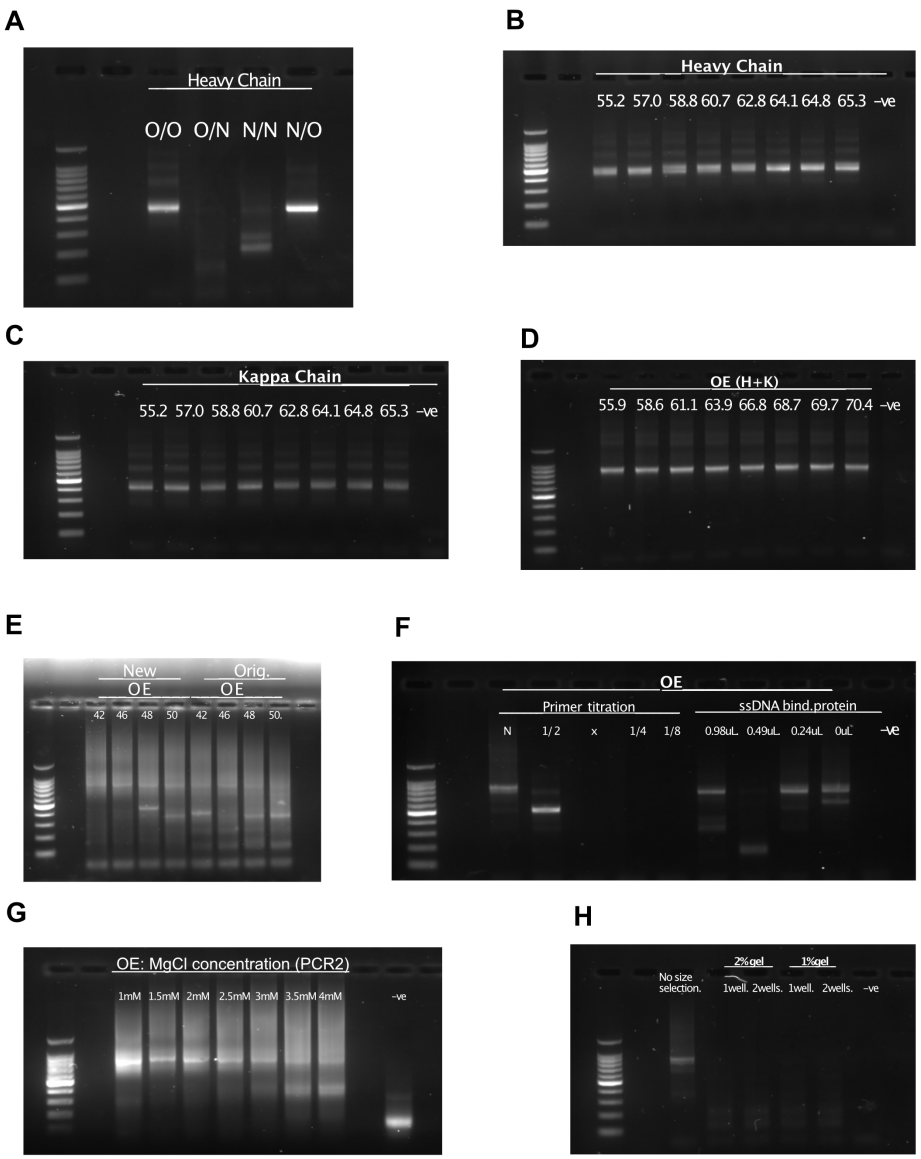


Figure 14: Troubleshooting Biorad One-Step RT PCR.
(A) Original and newly designed forward (fwd) and reverse (rev) heavy chain primers tested: O/O - original fwd + original rev, O/N: original fwd, new rev, N/N: new fwd, new rev, N/O: new fwd, original reverse. Note - new forward primers designed to have higher TM and new reverse primers contained a (G4S)4 linker instead of (G4S)3 linker. (**B)** RT-PCR using bulk RNA yielded heavy chains, **(C)** kappa light chains, and products from these two reactions could be used to generate overlap-extended(OE) products**(D). (E)** Temperature gradient for Reverse Transcription step, using "New" (VH_OUT_N, VK_OUT_N and original linker primers) or original VH and VK primer sets as per Rajan et al ("Orig.") and bulk RNA. **(F)** Titrating linker primer concentrations N = "normal", 1/2, 1/4 and 1/8 of concentrations used in Rajan et al. ssDNA binding protein (HotStart-IT) titrated. **(G)** Magnesium chloride concentrations were titrated in PCR2 using successfully amplified OE product from PCR1 as input (from bulk RNA). **(H)** Overlap extended product obtained with or without size selection after PCR1 for 20uL (1 well) or 40uL (2 wells) product on a 2% or 1% agarose gel.

4.3.14 Attempting OE RT-PCR with Alternative Reagents

Certain polymerases are more amenable to difficult templates but as the BioRad RT-PCR reagents come as a ready-made mastermix, there is little scope for optimisation. We hypothesised that perhaps the polymerases in the Biorad RT-PCR mastermix did not have proofreading activity and the addition of 3' A overhangs could interfere with the pairing of the complementary overhangs. We therefore attempted to spike-in a range of additional polymerases into PCR1: Q5, PWO, PFU and Expand™ High Fidelity enzyme mix (Taq and a proprietary proofreading polymerase) to see if they improved the yield of overlap-extended product from RNA. We found that none of the polymerase spike-ins yielded more product (**Figure 15A**). With the help of an MSc rotation student (Ruth Shelton), we revisited the reagents used in Rajan et al. 2018 and found that the addition of BSA improved the thermostability of the emulsions while Pluronic F-68 Non-ionic Surfactant (usually used in cell culture) also improved droplet uniformity (**Figure 15B,C**). When attempting to pair spiked-in heavy and light chain amplicon in emulsions, the Roche Titan reagents failed to produce an overlap-extended product, while heavy and light chains were amplified readily (**Figure 15E**). We also attempted using a custom RT-PCR mastermix described in Ma *et al.* (2021) for single-cell multiplex gene expression analysis. Thermostable emulsions were again readily optimised by titrating BSA and surfactant and heavy and light chain amplicons were successfully obtained using these methods. We could pair heavy and light chain amplicons using the custom mastermix and HIFI polymerase, but overlap-extension from bulk RNA failed (**Figure 15D**). We also observed that high concentrations of enzymes were required for successful RT-PCR in emulsions (data not shown). Due to time constraints the work could not be carried forward further, but a current lab member is performing further optimisations based on the lessons learnt in this chapter. While we did not reach our overall objective of obtaining OE product from single cell emulsions by RT-PCR, these experiments did demonstrate that a range of

commercial and custom reagents can be adapted for droplet PCR by titrating BSA and surfactant which produce thermostable emulsions.

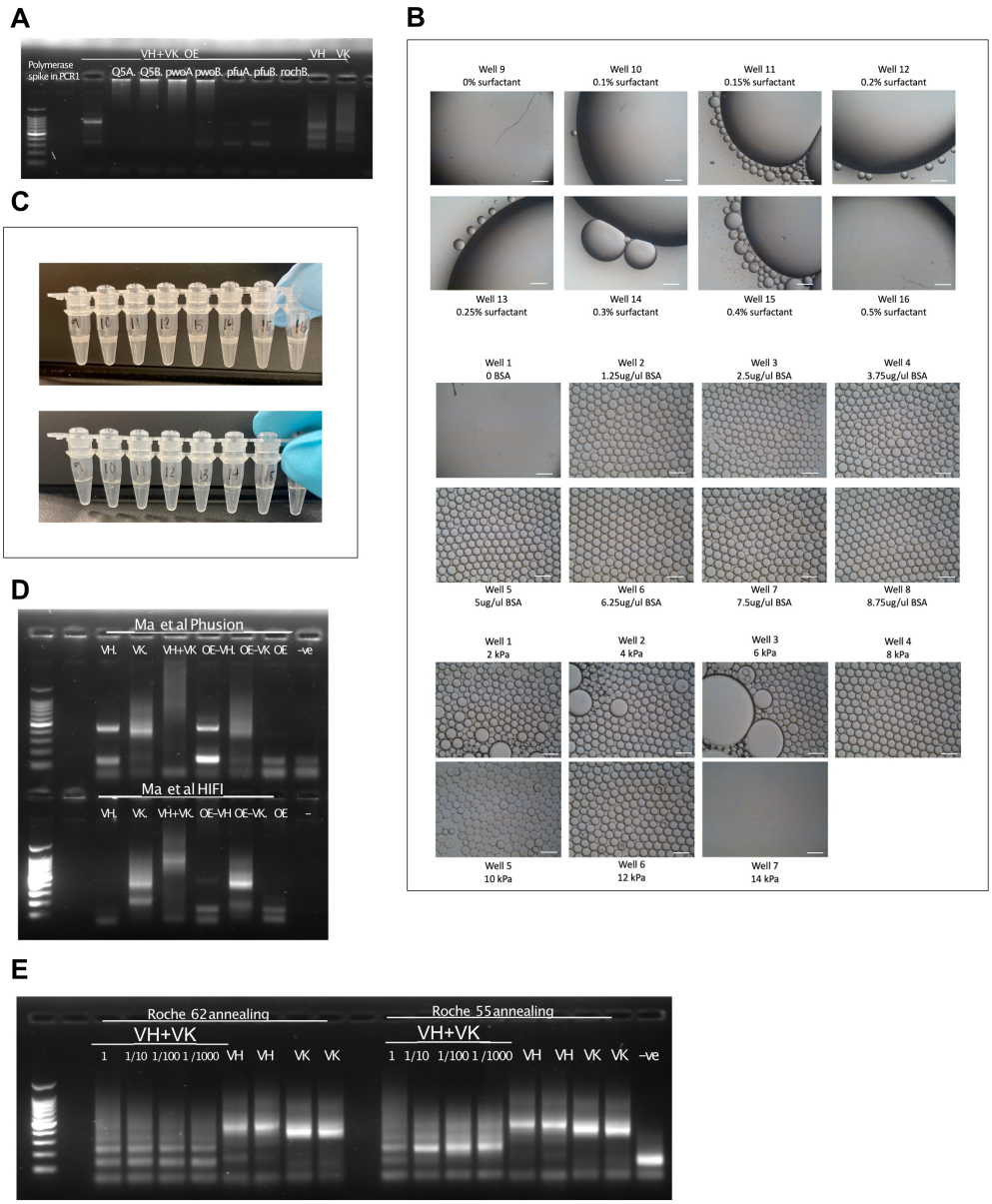


Figure 15: Attempting overlap-extension PCR with alternative reagents. (A) OE RT-PCR from bulk RNA using Biorad ddPCR reagents with other polymerase spike-ins in PCR1 (A = 0.5uL, B: 0.25uL). **(B)** Emulsions generated with Roche Titan reagents were imaged post thermo-cycling to assess droplet stability. Emulsion stability was optimised by adding 0-16% surfactant to mastermix (top panel), 0-8.75ug/uL BSA added to master mix (middle panel) and emulsions with optimised BSA and surfactant concentrations were generated with 2-14kPa vacuum (bottom panel) - scale bar 200 micros. **(C)** Roche Titan reagents, with surfactant added, pre- and post thermocycling. **(D)** Custom RT-PCR mastermix adapted from Ma et al (2021) using Superscript 3 RTase and Phusion or HIFI polymerases. Attempting to pair spike-ins (VH+VK) or attempting OE reactions from bulk RNA ("OE"). **(E)** Attempting overlap-extension from VH and VK spike-ins using Roche Titan RT-PCR system, across four different dilutions of template and using annealing temperatures as per Rajan et al (62ºC) or annealing temperature used with successful BioRad reactions (55ºC).

4.4 Discussion

4.4.1 Summary

Our attempts to replicate results from Rajan et al. 2018 were unsuccessful. While we cannot exclude the possibility that our microfluidics setup made a difference to the reaction efficiency (for example, different sized droplets, surfactant that was incompatible with the reagents they used), we did identify several concerns about the paper. Even when using the Roche Titan RT-PCR mastermix –which they emphasised was the only mastermix to successfully amplify a housekeeping gene in emulsion RT-PCR– and using their primers, we were unable to obtain OE product. In their paper they used pairing of constant region genes between mouse and human cell lines to demonstrate proof of principle that their method preserves cognate pairing, but nowhere do they show a gel demonstrating successful amplification of heavy chain, light chain and overlap-extended product. In the supplementary material a gel showing scFvs is provided, however there is no information about the size of the products and no gel demonstrating that heavy and light chains were successfully amplified individually. It is also worth noting that their PCR strategy included three rounds of amplification, totalling 100-120 cycles of PCR, which suggests very low efficiency of the PCRs. We also encountered issues with the primers provided in their paper including off-target amplification, and identified primers with redundant sequences. Finally, the method has not been applied by the lab in any follow-up work or adopted by other groups, despite there being demand for such a protocol. While we successfully obtained apparent overlap-extended product on one occasion, the reactions proved very inconsistent despite our best efforts to replicate results and control for experimental factors, such as using single-use aliquots for all reagents.

4.4.2 Why Did the PCRs Not Work?

Performing one-step RT-PCR in single cell emulsions is an ambitious goal with many potential issues. Cell lysate inhibition is a known challenge when working with small volumes and single cell reactions – many cytosolic proteins can inhibit or interact with the enzymes. Cell-lysate resistant polymerases (Tanno et al. 2020) and using the bacteriophage T7 gene 2.5 protein, which is also a single stranded DNA binding protein, as a PCR additive (Ma et al. 2021) have been used to overcome these issues. We have attempted adding a commercial ssDNAbp, used by Ma *et al.*, to our reactions but found it made no difference. Genomic DNA amplification was an issue when using lambda light chain primers, but may also have posed a problem with other primer sets, particularly if reverse transcription was inefficient and many PCR cycles were performed. Secondly, cDNA synthesis was performed with the gene specific primers in the same reaction as the PCR. We encountered particular issues with amplifying heavy chains. This may be due to inefficient cDNA synthesis using the VH-in_3' primers containing long linker sequences which were likely to form secondary structures and anneal to off-target sequences. Due to the complex primer pools needed to amplify all known heavy and light chain genes as expression-ready libraries (96 primers were used in total for these reactions), variable TMs, and long primers required to introduce (G4S)3 linkers, it is very likely that many of these primers interacted and hampered efficient amplification. Particularly the primers containing the linker tags were >50 bp long and were predicted to form many secondary structures. Finally the need for size-selection post RT-PCR may have posed an issue, as we were routinely able to amplify overlap extended product from bulk RNA from unpurified product. However, when DNA gel size-selection was performed our reactions were often unsuccessful. We hypothesised that this was due to either products not linking in the first reaction, or the PCRs being so inefficient that we lost our template during the gel extraction step.

4.4.3 Next Steps and Improvements

Current members of the Cowan Lab are adapting the protocols described here to perform bead-based mRNA capture with the BD Rhapsody system and bead-bound cDNA synthesis followed by encapsulation of single beads into emulsions using the MANATEE system. This amended workflow simplifies the reactions performed by emulsions PCR and eliminates the issue of genomic amplification. In Adler et al (2017) a similar approach is adopted for linking mouse heavy and light chains, by performing cell lysis and mRNA capture on oligo-dT beads in emulsions, followed by extraction of the beads from emulsions and encapsulation of beads into a second emulsion to perform RT-PCR. Initial results have demonstrated that emulsions can be optimised for encapsulation of the Rhapsody cDNA beads and currently PCRs to perform heavy and light chain overlap-extensions are being optimised. Separating the cDNA synthesis step from the overlap extension PCR also allows for the primer strategy to be amended as the linker tag could be introduced during cDNA synthesis or PCR. Having established that many standard PCR reagents can be adapted for use in thermostable emulsions by titrating BSA and surfactant in this work, we have many options for screening the most efficient and robust PCR reagents. An advantage of the MANATEE system over the Biorad ddPCR QX200 system is that the reactions and vacuum conditions can be customised and optimised, making it more amenable to different reagents which would not be possible otherwise. Another improvement currently being adopted in the lab is to use the Raji cells to optimise the reactions, which limits the primers used to 4 individual primers instead of complex primer pools and including a housekeeping gene (GAPDH) as a positive control for efficient cDNA synthesis and PCR. Nonetheless these reactions are proving challenging to optimise as emulsion PCR requires different conditions to PCR in solution and generating and working with emulsions remains time-consuming.

4.4.4 Prospects and Limitations for Phage Display of Native scFvs

If the lab does successfully develop the protocol, there are countless potential applications of the technology. This pipeline would enable rapid identification of monoclonal antibodies against pathogens from convalescent patients whilst also generating an archive of BCR-antigen specificity. Phage libraries could be revisited to screen for cross-reactive scFvs or to test affinity of the interactions by increasing salt concentration to disrupt binding (Hawkins et al 1992). BCRs with different sequences but convergent specificities could be compared to better understand determinants of antigen specificity. In Hemadou et al. 2018 scFvs from a human combinatorial library were used for *in vivo* biopanning against lesions from a rabbit model of atherosclerosis; high affinity binders were identified by sequencing phage bound to endothelial surfaces and aortic tissue after sacrifice. Similar approaches could be taken to identify phage which binds, for example, to autoantigens in tissue biopsies or post-mortem tissue samples from patients with autoimmune conditions. Phage display libraries could also be incubated with tissue lysates and autoantigen bound to scFvs could be pulled down.

Combining native antibody phage display libraries with high-throughput peptide library or antigen screening could potentially allow broad characterisations of repertoire specificities. For example, a similar approach to that taken in Yang *et al.* (2019) using two sets of yeast expression systems to display both antibodies and membrane proteins, could be taken by combining phage display of scFvs with a yeast expression library that expresses all human proteins. After co-incubation and several rounds of washes, a phage-specific antibody conjugated to a fluorophore could be used to sort yeast cells with bound phage into a 384 well plate and barcoded with a primer matrix for sequencing. This would allow screening of 384 antigens and their cognate BCRs in parallel. An advantage of this

approach would be that membrane-bound proteins, which are otherwise difficult to study in solution, could be characterised alongside any other proteins of interest.

While phage display is a powerful platform for scFv expression and screening, there are some limitations which may be encountered. Due to different codon usage between pro- and eukaryotes, some eukaryotic proteins will be expressed at low levels or with improper folding in bacteria and may interfere with the assembly of the phage or stability of the scFv (Kang and Seong 2020). scFvs have a propensity to aggregate under thermal stress and bacteria lack machinery for eukaryotic post translational modifications such as disulfide bond formation and glycosylation. VH and VL domains have two conserved disulfide bonds that connect the two β-sheets. In scFvs produced in E coli, these disulfide bonds are often reduced which in turn affects the Δ G of scFv folding and can affect its antigen binding affinity, susceptibility to proteolysis, and propensity to from aggregates (M. J. Seo et al. 2009). Other post translational modifications which are known to affect CDR3 affinity include *N*-glycosylation sites in CDR3s. These have been reported to be more common in repertoires of patients with rheumatoid arthritis (Rochelle D. Vergroesen et al. 2019; Houde et al. 2010) and have been found in up to 3% of plasma cells in healthy donors (Bondt et al. 2021).

The difficulties of expressing eukaryotic proteins in prokaryotes can be overcome by switching to eukaryotic expression systems such as mammalian cell lines (Vendel et al. 2012). Furthermore, we cannot assume that scFvs will necessarily mirror the antigen-binding affinities occurring in vivo because BCRs are expressed as dimers which can stabilise antigen-BCR binding by antigen cross-linking. Membrane bound BCRs may also have increased affinity due to receptor clustering: Lingwood et al (2012) reverted three broadly neutralising antibodies that recognise influenza haemagglutinin (HA) protein to their germline precursors. When expressing these pre-affinity matured receptors as IgG and decameric soluble IgM they lost their binding capacity to HA. However, when these

same receptors were expressed as cell surface IgM all three BCRs bound HA and triggered BCR signalling, suggesting that the affinity of the membrane bound BCRs was increased by BCR receptor clustering. B cell subsets and maturation stage should therefore also be taken into account when characterising affinity.

4.4.5 *In Silico* Approaches to Map BCR Sequence Data to Antigen Specificity

Despite the development of many sequence and structure-based tools to predict epitope-paratope interactions, predictive models are currently computationally intensive and most still have low accuracy (AUC values between $0.6-0.75$) (Fleri et al 2017). This is partly due to the limited availability of experimental data to train these models. Akbar et al. 2021 used solved structures for 825 antibody-antigen complexes and concluded that paratope-epitope interactions are predictable. They estimated that based on available data BCR-antigen interactions use a constrained repertoire of interaction motifs (approximately 10^4). It would be valuable to validate these predicted interaction motifs by comparing scFv-antigen interactions identified experimentally with their predicted targets. Building an archive of antigen-BCR specificities will further improve predictive models of CDR3 specificity as well as enrich AIRRseq data analyses by allowing assignment of BCR specificities based on previously recorded antigen-BCR interaction. If repertoires could be annotated comprehensively according to their antigen specificity, this could provide insight into an individual's infection history and improve our understanding of the dynamics of the adaptive immune repertoire response to different stimuli including responses to chronic and reactivated infections.

4.5 Materials and Methods

4.5.1 Biorad RT-PCR Reactions ("Reaction 1")

All reaction were assembled in an amplicon-free clean room on cool blocks and kept on ice prior to encapsulation. Primer cocktails for RT-PCR 1 were made up for each set (VH-out_5, VKVL_out_3, VH_in_3, VKVL_in_5) by mixing individual primers 1:1 and stored at 100uM in single use aliquots. Supermix, and eventually, DTT were also stored as single use aliquots. Concentration of primer cocktails were used as per Rajan et al, 2018 for a 2X mastermix (combined 1:1 with RNA and water or cell suspension): First, fresh working stocks of primers were made up (from a single-use 100uM aliquot each time)

Table 4.1: Primer Working Stocks

Primer cocktail	Working stock	Nuclease free water	100uM primer
VH_out_5	10uM	45uL	5uL
VK/VL_out_3	10uM	45uL	5uL
VH_in_3	1uM	99uL	1uL
VK/VL_in_5	1uM	99uL	1uL

For each ddPCR microfluidics chip, mastermix was made up for 9 wells (8+1 extra). Primer mixes were made up as follows:

Table 4.2: Primer Cocktails

Primer	Final conc./Rx	Working stock	Volume 1 Rx	Total volume (x9)
VH_out_5	139 nM	10uM	0.139 uL	1.251 uL
VKVL_out_3	416 nM	10uM	0.416 uL	3.744 uL
VH_in_3	39 nM	1uM	0.39 uL	3.51 uL
VKVL_in_5	13 nM	1uM	0.13 uL	1.17uL
H20			0.925 uL	8.325 uL

Mastermix was assembled in an amplicon-free clean room, and template (single cell suspension or RNA) added immediately prior to loading microfluidics chip and briefly mixed by flicking the tube or vortexing. Reagents were supplied from One-Step RT-ddPCR Advanced Kit for Probes (BioRad #1864021):

Table 4.3: RT-PCR Mix

Reagent	Volume for 1 Rx (VF: 20uL)	Volume (x9)
Supermix	5uL	45uL
DTT (100uM)	1uL	9uL
RTase	2uL	18uL
Primer Mix	2uL	18uL
RNA or single cell suspension	5uL	45uL
H2O or PBS (+0.1% BSA)	5uL	45uL

20uL of Mastermix was loaded using a P20 pipette into each of the eight sample wells of a Biorad QX200 DG8 microfluidics chip (BioRad #1864008), taking care to avoid introducing air bubbles at the bottom of the well and subsequently 70uL of BioRad EvaGreen Oil (#1864005) were loaded into the oil (bottom) well. The chip was sealed with a gasket (BioRad #1863009) and the BioRad Qx200 droplet generator or MANATEE system were used to generate emulsions (vacuum applied until all oil was pulled through). Emulsions were then gently transferred with a P200 pipette tip, pipetting from the top of the well downwards to avoid compressing the droplets, to a PCR strip for thermocycling. Standard thermocycling conditions were: 60 mins at 50°C, 10 mins at 95°C, followed by 40 cycles of 30s at 95°C, 1 minute at 55°C and 1 minute at 72°C, and a final extension for 7 minutes at 72°C and 10 minutes at 98°C to terminate the reactions. For reactions performed with single cell emulsions I extended the cDNA synthesis step to 90 minutes to allow for cell lysis at 50°C

4.5.2 Breaking Emulsions

To break the emulsions, the entire volume of droplets and oil were transferred into a 1.5ml Eppendorf, combining replicates if needed. As much of the oil phase was removed as possible by carefully pipetting the oil from below the emulsion. 20uL of TE buffer was then added to each Eppendorf, followed by 70uL of molecular grade chloroform per emulsion. If multiple emulsions were combined the volumes of TE buffer and chloroform were multiplied by the number of emulsions. Tubes were then vortexed at maximum speed for 1 minute and centrifuged for 10 minutes at 15,500xG. After centrifugation the mixture would separate into three distinct phases. The top phase, containing the aqueous solution, was removed carefully by pipetting, and transferred to a fresh tube. This RT-PCR product was then used for subsequent reactions, purifications, or stored at -20°C.

4.5.3 PCR Clean-Up

For optimisations which did not require size-selection, column-based PCR clean-up (Monarch® PCR DNA Cleanup NEB #T1030) was performed to remove excess primers and contaminants as per the manufacturer's instructions and eluted in nuclease free water. For reactions which required size-selection, 20uL of PCR1 product per sample was run on a 2% agarose gel with 0.5X SybrSafe (ThermoFisher #S33102) and the band between 600-1000bp excised and extracted using the Monarch® DNA Gel Extraction Kit (NEB #T1020) as per the manufacturer's instructions and eluted in 10uL of nuclease free water.

4.5.4 PCR 2 "Reaction 2"

PCR2 was used to step in to the 5' end of the VH chain and the 3' end of the VK/VL chain to amplify the overlap-extended product. For amplifying VH chains, VH_in_5 primer pools and VH_in_3' (used in PCR1) were used. For amplifying VK/VL chains

only, I used VK/VL_in_5' primer pools (used in PCR1) and the VK/VL_in_3' primer pools. Standard optimisations were performed using GoTaq Flexi (Promega, #M8295) as per the manufacturer's protocol. PCR reactions were set up in an amplicon-free clean room, and PCR1 product added immediately prior to thermocycling in the PCR lab. Cycling conditions: 94° 2 mins 40 cycles of: 94° 30 sec 55° 30 sec 72° 1min + 5 min extension: 72

Table 4.4: Reagents and Volumes for PCR2.

Reagent	Volume for 1 reaction (VF 25uL)	Final concentration
Fwd primer (1uM)	2.5uL	0.1uM
Rev primer (1uM)	2.5uL	0.1uM
5X GoTaq Flexi Buffer	5uL	1X
dNTPs (10mM)	0.5uL	0.2mM
MgCl2 (25mM)	2uL	2.5mM
Polymerase	0.125uL	0.025U/uL
H2O	9.875uL	-
PCR1 product*	2.5uL	-

* add in upstairs PCR lab

4.5.5 Agarose Gels

Amplicons were visualised by mixing 5uL of PCR product with 2uL of 5X loading dye on a 2% Agarose gel with 0.5X SybrSafe run at 120V for 30 mins and product sizes compared to a 100bp ladder (Promega # G2101). Gels were imaged using D-Digit LiCor.

4.5.6 Thawing PBMCs

Cryopreserved PBMCs were thawed according to the 10X Genomics thawing protocol: "Demonstrated Protocol - Fresh Frozen Human PBMCs for Single Cell RNA Sequencing – Rev D" (CG00039)

4.5.7 MACS Sorting B cells (Positive Selection)

Thawed PBMCs (described above) were counted, strained through a 40uM strainer, and spun at 300g for 10 minutes and all media removed (invert and pipette off any liquid that remains on lip of tube). For 10^7 cells, cells were resuspended in 80uL of MACS buffer (PBS + 0.5% BSA + 2mM EDTA) and 20uL of CD19 MicroBeads human (Miltenyi, 130-050-301). Cells were incubated in the fridge (4°C) for 15 minutes. 2ml MACS buffer was added and Falcon centrifuged at 300g for 10 minutes. After the spin, supernatant was removed and resuspended in 500uL MACS buffer (for up to 10^8 cells). MS columns were attached to Miltenyi Magnet and 500uL of MACS buffer was added to rinse the column, the wash discarded, and a fresh tube used to collect effluent. Cells were pipetted onto the column and allowed to pass over the column, followed by 3 washes with 500uL MACS buffer. Once all washes passed over the column, the column was removed from the magnet and 1ml MACS buffer added to the column and buffer forced through with the plunger. CD19 positive cells were found in the flushed-out fraction.

4.5.8 Live Cell Staining

MACS sorted B cells were resuspended gently in pre-warmed CellTRacker Working Solution (HBSS + 1:1000 CM Green). Cells were incubated for 30 minutes at 37°C), 5% CO2 in a cell culture incubator. Cells were pelleted at 300g for 5 minutes and supernatant removed, followed by a wash with 10ml HBSS and a further spin at 300g for 20 minutes. Cells were resuspended at a concentration of 200-2000 cells/uL in either RT-PCR mastermix or PBS + 0.1% BSA, 10uL of cell suspension loaded alongside mastermix or PBS and imaged using the BioRad ZOE Fluorescent Cell Imaging platform.

4.5.9 Imaging Droplets

Emulsions were imaged on glass microscopy slides pre-treated with a hydrophobic agent (AquaPel) to preserve droplets. Fluorescence imaging was performed using BioRad ZOE Fluorescent Cell Imager. Standard light microscopy was performed using a Zeiss IM35 microscope at 3.2X magnification.

4.5.10 Oxford Nanopore Sequencing

Oxford Nanopore Sequencing was performed using the Flongle v3 flow cells and chemistry. Amplicon was first purified using Ampure (XP) beads as per manufacturer's instructions. The LSK109 ligation sequencing kit was used for library preps (using short fragment buffer) adopting the library preparation strategy as described in the supplementary data in Eaton et al (2020). Minion sequencing and basecalling was performed using MinKNOW and "fast" base calling with Guppy.

4.5.11 Sequencing Data Analysis

Fastq files which passed the quality threshold in MinKNOW were first concatenated into a single file with all sequences and reformatted into FASTA files in linux command line. The fasta file was uploaded to IMGT-High V quest for alignment to germline reference sequences. Default parameters were used (looking for indels) and the scFv alignment option was selected. Custom python scripts were used to analyse the outputs of IMGT-High V quest. In particular "1_Summary.txt" files were used to generate summary statistics of functionality of assignments and "12_scfv.txt" files were analysed to analyse linker lengths and paired heavy and light chain V calls. For sequence alignment using IgBlast, the (G4S)3 tag was located using fuzzy string matching with the Python "fuzzysearch" package using the "find_near_matches" function and sequences split into two files containing the first or

second half of the scFv. The IgBlast wrapper provided by changeO in the Immcantation suite was used to align each half of the scFv to IMGT reference databases.

4.5.12 CAD Design & 3D Printing

The tray was originally designed in FreeCAD and the Manifold was designed using Autodesk Fusion 360. Objects were uploaded to Shapeways and QC conducted using their 3D Tools suite and adjustments to wall thickness and other parameters made accordingly. The final version was 3D printed from Shapeways using selective laser sintering (SLS) in nylon (plastic polyamide 11: PA 11), selected due to being chemically and mechanically heat-resistant and durable. Designs have been made freely available for download and re-use on Thingiverse (available at: https://www.thingiverse.com/thing:6174004)

4.5.13 Vacuum Release Circuit & Components

Code for calibrating pressure gauge was adapted from: https://circuits4you.com/2016/05/13/arduino-pressure-measurement/. The remaining code was written as a custom script. I have made a full description of the project and how it was assembled available on GitHub at: https://github.com/aaryback/MANATEE

4.5.14 RT-PCR with Roche Titan Reagents

RT-PCR reactions were performed as per the manufacturer's instructions, with the addition of 5ug/uL BSA (in place of nuclease-free water) and using annealing temperatures of either 62C or 55C.

4.5.15 RT-PCR with Ma *et al.* Reagents

Reagents were adapted from Ma *et al.* (2021) using "Crude Cell Multiplex" reagents (https://www.nature.com/articles/s41598-021-86087-4/tables/2), with GoTaq polymerase instead of their polymerase (they used a custom-made one) and reduced the concentration of HotStart-IT Binding Protein. For mastermix components see Table 4.6. cDNA synthesis was performed for 45 mins at 55C, and PCR1 cycling conditions: 94° 2 mins, 45 cycles of: 94° 30 sec, 55° 30 sec, 72° 1min, + 5 min extension: 72.

The second round of PCR amplification was performed with 2 uL of PCR1 product using GoTaq reagents as described in section: PCR 2 "Reaction 2". PCR cycling conditions: 94° 2 mins, 50 cycles of: 94° 30 sec, 55° 30 sec, 72° 1min, + 5 min extension: 72.

Table 4.5: Components for MANATEE Control Circuit

Part	Catalogue Numbers / specifications	Function
Arduino	Arduino Mega 2560	Control circuit
Pressure Gage	MPX5100DP	Measures vacuum in MANATEE relative to atmospheric pressure
Solenoid Valve	DFRobot 6V 2-Position 3-Way Air Valve for Arduino (DFR0866)	Release of vacuum from vacuum source into MANATEE (controlled by Arduino, threshold set in code)
N channel power MOSFET (30V 60A)	IRLB8721PbF	Controls whether solenoid valve is "on" or "off"
Flyback diode	Vishay 1N4007E354	Shorts the circuit in case solenoid sends surge of current through the circuit in the "wrong" direction
Green LED		Visual signal to see that circuit is powered on and working (blinks whenever solenoid valve is "open")
Resistors	220 Ω Resistor	
Prototyping PCB board	Arduino UNO R3 ProtoShield (483150)	Allows soldering of components to circuit board to keep wires and circuits in place and functioning.

Table 4.6: RT-PCR from Ma et al. 2021

Reagents	Unit	Concentration
TrisHCl, pH 8.0	mM	30
KCl	mM	70
MgCl2	mM	2.5
Recombinant RNase Inhibitor (Takara)	U/µL	1
Go Taq polymerase (Promega)*	µL	0.6
HotStart-IT Binding Protein (Thermo Scientific)**	ng/µL	475
dNTPs	mM	0.5
Forward and reverse primers (IDT)	µM	0.6
Nonidet P-40	% vol	0.14
Pluronic F-68 Non-ionic Surfactant (Gibco)	% vol	0.2
SuperScript III reverse transcriptase (Invitrogen)	U/µL	1
Bovine serum albumin (Roche)	ng/µL	1380
DTT (Invitrogen)	mM	1
*used this instead of their home-made Polymerase		
**used half the recommended concentration due to cost		

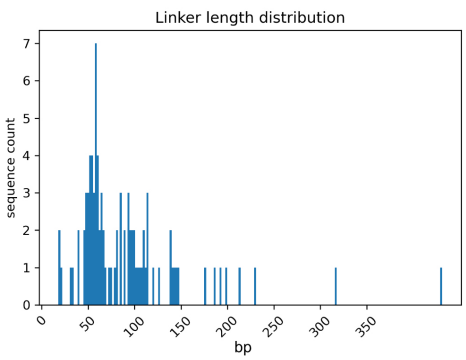
4.6 Supplementary Information Chapter 4

Arduino Code

```
// Prints average pressure to the serial monitor
// Typically the input is biased at Vcc/2 for a reading of ~512 with no signal
// Reads input A0
// sensor offset
const float SensorOffset = 38.0;
// digital pin 26 has a mosfet controlling a solenoid attached to it.
const int solenoid = 26;
const float pressure= 8.0; //change this to change vacuum
// Global variables
int Analog;
float Sum;
float Average;
void setup() {
  // put your setup code here, to run once:
   Serial.begin(9600);
   pinMode(solenoid, OUTPUT);
   delay(100);       //"Stabilization time".
   Serial.print("Hiya, I'm a pressure sensor \n");
}
void loop() {
  // put your main code here, to run repeatedly:
Sum = 0;        //Initialize/reset
//Take 1000 readings, find min, max, and average.  This loop takes about 100ms.
```

```
for (int i = 0; i < 1000; i++)
{
  Analog = analogRead(A0);
  Sum = Sum + Analog;   //Sum for averaging
  }
Average = (Sum/1000);
//Calibration
float Average_calib= (Average-SensorOffset)/10.0;
if (Average_calib >= pressure) {
  delay(100);
  digitalWrite(solenoid,LOW); // if the gauged pressure is higher
//than the cutoff, send a LOW V signal to the mosfet (valve closed)
} else {
  delay(20);
  digitalWrite(solenoid,HIGH);
}
 // if the gauged pressure is lower than the pressure cutoff,
//send a HIGH V signal to the mosfet (ie Valve open)
Serial.print (" Average = ");
Serial.println (Average_calib);}
```

Supplementary Figure 1: Length of scFv linkers from "overlap-extended" product amplified with original lambda light chain primers



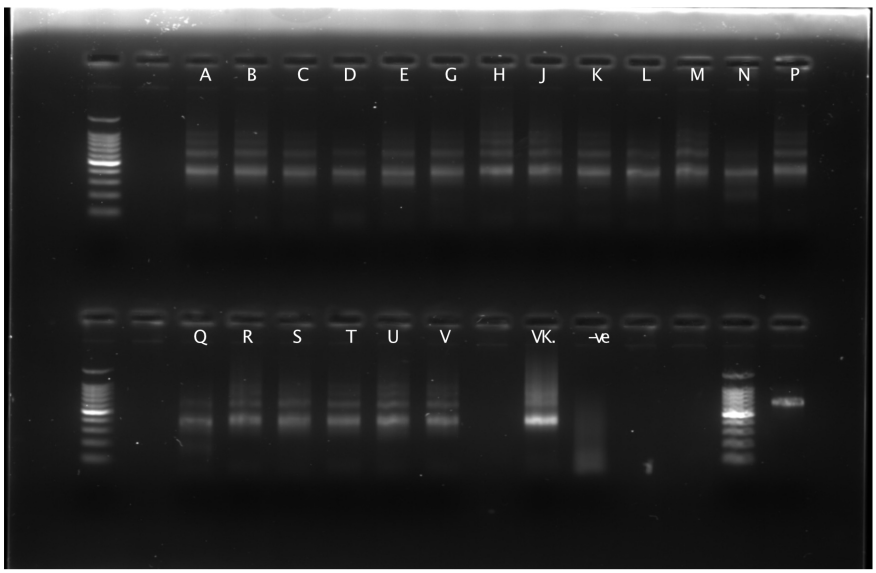
Supplementary Figure 2: Original lambda 3' "out" light chain primer binding sites

Primers (reverse complemented):

VL_out_3_01 CTCTGTTCCCACCCTCCTCT
VL_out_3_02 CGCCTGAGCAGTGGAAGT
VL_out_3_03 CCCCTCGGTCACTCTGTT
VL_out_3_04 CTCTGTTCCCGCCCTCCTCT

>J00254|IGLC3*01|Homo sapiens|F|C-REGION|1..312|312 nt|1| | | |
|312+0=312|partial in 5'| |
cccaaggctgccccctcggtcactctgttcccaccctcctctgaggagcttcaagccaac
aaggccacactggtgtgtctcataagtgacttctacccgggagccgtgacagttgcctgg
aaggcagatagcagccccgtcaaggcgggggtggagaccaccacaccctccaaacaaagc
aacaacaagtacgcggccagcagctacctgagcctgacgcctgagcagtggaagtcccac
aaaagctacagctgccaggtcacgcatgaagggagcaccgtggagaagacagttgccccct
acggaatgttca
>K01326|IGLC3*02|Homo sapiens|F|C-REGION|g,203..519|318 nt|1|+1| | |
|318+0=318| | |
ggtcagcccaaggctgccccctcggtcactctgttcccaccctcctctgaggagcttcaa
gccaacaaggccacactggtgtgtctcataagtgacttctacccggggccagtgacagtt
gcctggaaggcagatagcagccccgtcaaggcgggggtggagaccaccacaccctccaaa
caaagcaacaacaagtacgcggccagcagctacctgagcctgacgcctgagcagtggaag
tcccacaaaagctacagctgccaggtcacgcatgaagggagcaccgtggagaagacagtg
gcccctacggaatgttca
>X06876|IGLC3*03|Homo sapiens|F|C-REGION|g,98..414|318 nt|1|+1| | |
|318+0=318| | |
ggtcagcccaaggctgccccctcggtcactctgttcccaccctcctctgaggagcttcaa
gccaacaaggccacactggtgtgtctcataagtgacttctacccgggagccgtgacagtg
gcctggaaggcagatagcagccccgtcaaggcgggagtggagaccaccacaccctccaaa
caaagcaacaacaagtacgcggccagcagctacctgagcctgacgcctgagcagtggaag
tcccacaaaagctacagctgccaggtcacgcatgaagggagcaccgtggagaagacagtg
gcccctacagaatgttca
>D87017|IGLC3*04|Homo sapiens|F|C-REGION|g,3455..3771|318 nt|1|+1| | |
|318+0=318| | |
ggtcagcccaaggctgccccctcggtcactctgttcccgccctcctctgaggagcttcaa
gccaacaaggccacactggtgtgtctcataagtgacttctacccgggagccgtgacagtg
gcctggaaggcagatagcagccccgtcaaggcgggagtggagaccaccacaccctccaaa
caaagcaacaacaagtacgcggccagcagctacctgagcctgacgcctgagcagtggaag
tcccacagaagctacagctgccaggtcacgcatgaagggagcaccgtggagaagacagtg
gcccctacagaatgttca

Supplementary Figure 3: Testing redesigned lambda "out" chain primers. ILJC_OUT_3 primers A-V were tested individually in PCR1 using cDNA synthesised with oligodT as template, followed by a nested PCR with VL_in_3' primers from Rajan et al. Bright band on ladder is 500bp, kappa chain also amplified as a positive control



4.6.1 Primer Tables

Primers used in this Chapter are shown below. Gene-specific parts of the primer are written in upper case while G4S3 linkers and sequencing adapters are in lower case. All primers sequences are shown 5' 3'.

Table 4.7: Original PCR1 Primer Sets from Rajan *et al.*: VH_OUT_5,VH_IN_3,VK_OUT_3, VK_IN_5, VL_OUT_3 and VL_in_5

Primer Name	Primer Sequence
VH_out_5_01	AAAAGGTGTCCAATGTGAGGTGC
VH_out_5_03	AAGGTGTCCAGTGTGAGGT
VH_out_5_04	ACAGGTGYCCACTCCCARR
VH_out_5_05	ACAGATGCCTACTCCCAGATGC
VH_out_5_06	AAAGCTGTCCAGTGTCAGGT
VH_out_5_07	CAGCAGCTACAGGTGTCCA
VH_out_5_08	AAGAGGTGTCCAGTGTCAGGT
VH_out_5_10	GGTGGCAGCTCCCAGATGG
VH_out_5_11	CTGACCACCCCTTCCTGG
VH_out_5_12	GTGGCAGCTCCCAGATGG
VH_out_5_13	ATGGGGTGTCCTGTCACAG
VH_out_5_14	AAGTGCCCACTCCCAGGT
VH_out_5_15h	GCTATTTTAAAAGGTGTCCAGWGTG
VH_out_5_16	AGCAGCTACAGGCACCCA
VH_out_5_17	CCATGGGTGTCCTGTCACA

Continued on next page

Table 4.7 – continued from previous page

Primer Name	Sequence
VH_out_5_18	ACTGACTGTCCCGTCCTGG
VH_out_5_19	ACAGGTGCCCACTCCCAG
VH_out_5_20	TGGTTCTTCCTCCTGCTGG
VH_out_5_21	CCCCTCCACAGTGAGAGTC
VH_out_5_22	AGCTCCAGGTGCTCACTCC
VH_out_5_23	CAGCCACAGGAGCCCACT
VH_out_5_24	AGGTGTCCAGTGTCAGGTG
VH_out_5_25	CTGCTGACCATCCCTTCATG
VH_out_5_26	CCTTGTTGCTATTTTAAAAGGTGTCC
VH_out_5_27	AAGGAGTCTGTKCCGAGGTG
VH_out_5_28	CTGGCTGTAGCACCAGGT
VH_in_3_01	gagccacctccgccgctaccgccgcctccagaGGAGACGGTGACCGTGG
VH_in_3_02	gagccacctccgccgctaccgccgcctccagaGGAGACAGTGACCAGGGTG
VH_in_3_03	gagccacctccgccgctaccgccgcctccagaGGAGACGGTGACCAGGGT
VH_in_3_04h	gagccacctccgccgctaccgccgcctccagaAGAGACGRTGACCATTGTCC
VK_out_3_01	CCACAGTTCGTTTRATHTCCAS
VK_in_5_01e	agcggcggaggtggctcaggcggtggcggaagtGAAATWGTGWTGACRCAGTCTCCA
VK_in_5_02h	agcggcggaggtggctcaggcggtggcggaagtRACATCCAGATGACCCAGTYTC
VK_in_5_03	agcggcggaggtggctcaggcggtggcggaagtGCCATCCGGATGACCCAG
VK_in_5_04	agcggcggaggtggctcaggcggtggcggaagtGAAATAGTGATGATGCAGTCTCCAG
VK_in_5_05	agcggcggaggtggctcaggcggtggcggaagtGAAACGACACTCACGCAGTC
VK_in_5_06	agcggcggaggtggctcaggcggtggcggaagtGACATCGTGATGACCCAGTCT
VK_in_5_07e	agcggcggaggtggctcaggcggtggcggaagtGAGATTGTGATGACCCAGACTCCA

Continued on next page

Table 4.7 – continued from previous page

Primer Name	Sequence
VK_in_5_08	agcggcggaggtggctcaggcggtggcggaagtGACATCCAGTTGACCCAGTCT
VK_in_5_09	agcggcggaggtggctcaggcggtggcggaagtGTCATCTGGATGACCCAGTCTC
VK_in_5_11	agcggcggaggtggctcaggcggtggcggaagtGATATTGTGATGACTCAGTCTCCAC
VK_in_5_12	agcggcggaggtggctcaggcggtggcggaagtGATGTTGTGATGACTCAGTCTCCA
VK_in_5_13	agcggcggaggtggctcaggcggtggcggaagtGAAATTGTAATGACACAGTCTCCAGC
VK_in_5_14	agcggcggaggtggctcaggcggtggcggaagtGCCATCCAGWTGACCCAGT
VK_in_5_15	agcggcggaggtggctcaggcggtggcggaagtGATATTGTGATGACCCAGACTCCA
VL_out_3_01	AGAGGAGGGTGGGAACAGAG
VL_out_3_02	ACTTCCACTGCTCAGGCG
VL_out_3_03	AACAGAGTGACCGAGGGG
VL_out_3_04	AGAGGAGGGCGGGAACAGAG
VL_in_5_01	agcggcggaggtggctcaggcggtggcggaagtCAGRCTGTGGTGACYCAGG
VL_in_5_02	agcggcggaggtggctcaggcggtggcggaagtTCCTATGAGCTGACTCAGCCA
VL_in_5_03	agcggcggaggtggctcaggcggtggcggaagtCAGTCTGTGCTGACGCAG
VL_in_5_04e	agcggcggaggtggctcaggcggtggcggaagtCAGGCAGGGCTGACTCAGCCA
VL_in_5_05	agcggcggaggtggctcaggcggtggcggaagtAATTTTATGCTGACTCAGCCCC
VL_in_5_06e	agcggcggaggtggctcaggcggtggcggaagtTCCTATGAGCTGAYRCAGCCAYC
VL_in_5_07	agcggcggaggtggctcaggcggtggcggaagtCWGSCTGTGCTGACTCAGC
VL_in_5_08	agcggcggaggtggctcaggcggtggcggaagtCAGCYTGTGCTGACTCAATCR
VL_in_5_09	agcggcggaggtggctcaggcggtggcggaagtCAGTCTGCCCTGACTCAGC
VL_in_5_10	agcggcggaggtggctcaggcggtggcggaagtCAGTCTGTSKTGACGCAGC
VL_in_5_11	agcggcggaggtggctcaggcggtggcggaagtCAGACTGTGGTGACTCAGGAG
VL_in_5_12	agcggcggaggtggctcaggcggtggcggaagtCAGTCTGTGCTGACTCAGCC

Continued on next page

Table 4.7 – continued from previous page

Primer Name	Sequence
VL_in_5_13	agcggcggaggtggctcaggcggtggcggaagtTCTTCTGAGCTGACTCAGGACC
VL_in_5_14	agcggcggaggtggctcaggcggtggcggaagtTCCTATGAGCTGACACAGCTAC
VL_in_5_15	agcggcggaggtggctcaggcggtggcggaagtTCCTATGTGCTGACTCAGCC

Table 4.8: Original PCR2 Rajan et al. Primer Sequences, VH_IN_5, VK_IN_3 and VL_in_3

Primer	Sequence
VH_in_5_01f	gaagacggcatacgagatggcccagccggccatggccCAGGTGCAGCTACAACAGTG
VH_in_5_02f	gaagacggcatacgagatggcccagccggccatggccCAGGTRCAGCTRCAGSAGT
VH_in_5_03f	gaagacggcatacgagatggcccagccggccatggccCAGGTCACCTTGAAGGAGTCT
VH_in_5_04f	gaagacggcatacgagatggcccagccggccatggccCAGATCACCTTGAAGGAGTCTGG
VH_in_5_05f	gaagacggcatacgagatggcccagccggccatggccCAGGTGCAGTCTGGTGGAGT
VH_in_5_06f	gaagacggcatacgagatggcccagccggccatggccGARGTGCADCTGGTGGAGWC
VH_in_5_07f	gaagacggcatacgagatggcccagccggccatggccCAGGTYCAGCTKGTGCAGT
VH_in_5_08f	gaagacggcatacgagatggcccagccggccatggccCAGGTACAGCTGGTGGAGTC
VH_in_5_09f	gaagacggcatacgagatggcccagccggccatggccCGGCTGCAGCTGCAGG
VH_in_5_10f	gaagacggcatacgagatggcccagccggccatggccCAGSTGCAGCTGCAGGA
VH_in_5_11f	gaagacggcatacgagatggcccagccggccatggccGAGGTGCAGCTGGTGCA
VH_in_5_12f	gaagacggcatacgagatggcccagccggccatggccSAGGTGCAGCTGTTGGAGT
VH_in_5_13f	gaagacggcatacgagatggcccagccggccatggccSAGGTCCAGCTGGTACAGTC
VH_in_5_14f	gaagacggcatacgagatggcccagccggccatggccCAGGTCACCTTGAGGGAGTC
VH_in_5_15f	gaagacggcatacgagatggcccagccggccatggccCARATGCAGCTGGTGCAGT
VH_in_5_16f	gaagacggcatacgagatggcccagccggccatggccCAGGTSCAGCTGGTGSA
VH_in_5_17f	gaagacggcatacgagatggcccagccggccatggccGAAGTGCAGCTGGTGCAGT
VH_in_5_18f	gaagacggcatacgagatggcccagccggccatggccCGGGTCACCTTGAGGGAG
VH_in_5_19f	gaagacggcatacgagatggcccagccggccatggccCAGGTGCGGCTGCAGG
VK_in_3_01g	tacgccaagctttggagccgcggccgcTTTGATCTCCACCTTGGTCCCTCCGCCGAAMGT
VK_in_3_02g	tacgccaagctttggagccgcggccgcTTTGATTTCCACCTTGGTCCCTTGGCCGAACGT
VK_in_3_03g	tacgccaagctttggagccgcggccgcTTTAATCTCCAGTCGTGTCCCTTGGCCGAAGGT
VK_in_3_04g	tacgccaagctttggagccgcggccgcTTTGATATCCACTTTGGTCCCAGGGCCGAAAGT
VK_in_3_05g	tacgccaagctttggagccgcggccgcTTTGATCTCCAGCTTGGTCCCCTGGCCAAAAST
VL_in_3_01g	tacgccaagctttggagccgcggccgcTAGGACGGTCAGCTTGGTCCCTCCGCCGAAYAC
VL_in_3_02g	tacgccaagctttggagccgcggccgcGAGGRCGGTCAGCTGGGTGCCTCCTCCGAACAC
VL_in_3_03g	tacgccaagctttggagccgcggccgcTAGGACGGTGACCTTGGTCCCAGTTCCGAAGAC
VL_in_3_05g	tacgccaagctttggagccgcggccgcGAGGACGGTCACCTTGGTGCCACTGCCGAACAC

Table 4.9: Redesigned Lambda VL OUT 3' Primers

Primer	Sequence
IGLCJ_out3_A	CTTGGGCTGACcKaggRcg
IGLCJ_out3_B	GGTTGGCCTTGGGcKaggRc
IGLCJ_out3_C	GGTGTCTCAGTGACCAGcKagg
IGLCJ_out3_D	CCCTGAAGAATGTTCTTAGcKag
IGLCJ_out3_E	GGCGTCAGGCTCAGGcKaggRc
IGLCJ_out3_G	TGAAGGTGTCTCAGTGACCcKaggRc
IGLCJ_out3_H	GGCCTTGGGCTGcKaggRc
IGLCJ_out3_J	GGGTTGGCCTTGGGctaaaatg
IGLCJ_out3_K	GTGTCTCAGTGACCAGctaaaatg
IGLCJ_out3_L	GGCCTTGAGCGGACctaaaatg
IGLCJ_out3_M	CTGAAGAATGTTCTTAGctaaaatg
IGLCJ_out3_N	GCGTCAGGCTCAGGctaaaatg
IGLCJ_out3_P	AAGGTGTCTCAGTGACctaaaatg
IGLCJ_out3_Q	GTGGCCTTGGGCTGctaaaatg
IGLCJ_out3_R	CTTGGGCTGACCcKaggRcg
IGLCJ_out3_S	CTTGGGCTGACctaaaatg
IGLCJ_out3_T	CCTTGGGCTGACCctaaaatg
IGLCJ_out3_U	GGGCAGCCTTGGGcKaggRcg
IGLCJ_out3_V	GCCTTGAGCGGACcKaggRcg

Primers designed to span junction between the J and C. J-region sequence is shown in lower case, C region sequence in upper case

Table 4.10: Redesigned VH_OUT_5' Primers with Higher and More Closely Matched TMs

Primer	Sequence	TM
>fwd_VH_out_5_01_n	AAAAGGTGTCCAATGTGAGGTGCAGCTG	63 ºC
>fwdVH_out_5_03_D_n	AAGGTGTCCAGTGTGAGGTGCAGS	∼63.2 ºC
>fwdVH_out_5_04_n	GCSACAGGTGYCCACTCCCARR	∼64 ºC
>fwdVH_out_5_05_n	GCCACAGATGCCTACTCCCAGATGC	63.3 ºC
>fwdVH_out_5_06_n	AAAGCTGTCCAGTGTCAGGTGCAGTCT	63.6 ºC
>fwdVH_out_5_07_n	GGCAGCAGCTACAGGTGTCCAST	63.4 ºC
>fwdVH_out_5_08_n	AAGAGGTGTCCAGTGTCAGGTRCAGC	∼63.1 ºC
>fwdVH_out_5_10_n	GGTGGCAGCTCCCAGATGGG	63.2 ºC
>fwdVH_out_5_11_n	GCTGACCACCCCTTCCTGGGT	64.1 ºC
>fwdVH_out_5_13_n	CTCCCATGGGGTGTCCTGTCACAG	63.2 ºC
>fwdVH_out_5_14_n	AAGTGCCCACTCCCAGGTGCA	63.9 ºC
>fwdVH_out_5_15h_n	GCTATTTTAAAAGGTGTCCAGWGTGARGTGCAGS	∼62.6 ºC
>fwdVH_out_5_16_n	AGCAGCTACAGGCACCCACGC	64.8 ºC
>fwdVH_out_5_17_n	GCCTCCATGGGTGTCCTGTCACA	63.3 ºC
>fwdVH_out_5_18_n	ACTGACTGTCCCGTCCTGGGTCT	63.7 ºC
>fwdVH_out_5_19_n	ACAGGTGYCCACTCCCAGGT	62.1 ºC
>fwdVH_out_5_20_n	TGGTTCTTCCTCCTGCTGGTGGC	63.5 ºC
>fwdVH_out_5_21_n	CCCCTCCACAGTGAGAGTCTGTGC	63 ºC
>fwdVH_out_5_22_n	GCTGTAGCTCCAGGTGCTCACTCC	63.2 ºC
>fwdVH_out_5_23_n	CAGCCACAGGAGCCCACTCC	63.0 ºC
>fwdVH_out_5_24_n	AGGTGTCCAGTGTCAGGTGCAGC	63.6 ºC
>fwdVH_out_5_25_n	GCTGCTGACCATCCCTTCATGGGT	63.4 ºC
>fwdVH_out_5_26_n	AGCTGGGTTTTCCTTGTTGCTATTTTAAAAGGTGTCC	63 ºC
>fwdVH_out_5_27_n	AAGGAGTCTGTKCCGAGGTGCAGC	∼64.1 ºC
>fwdVH_out_5_28_n	CTGGCTGTAGCACCAGGTGCC	63 ºC
>fwdVH_out_5_03_C_n	AGAAGGTGTCCAGTGTGAGGTGCA	62.2 ºC

Chapter 5

General Discussion

5.1 Summary

B CELL receptor repertoire sequencing is a promising technique to discern how the adaptive immune system responds to acute and chronic infection and immune dysfunction. In this thesis BCR repertoire sequencing was applied to look for evidence of infection or immune dysfunction in a disease of unknown etiology, ME/CFS, and to assess longitudinal dynamics of malaria B cell responses in a human re-challenge model. Finally, progress towards developing a protocol for high-throughput BCR-antigen mapping was made, though the final protocol was not successfully optimised.

In this work I partially replicated differential IGHV3-30 gene usage in ME/CFS and observed increased use of IgM BCRs in mild/moderate ME patients. While both findings had modest effect sizes, it is interesting that the previously published increase in IGHV3-30 was reproducible, particularly since IGHV3-30 in antibody has also been detected as

being disregulated in ME/CFS patients using plasma proteomics (Milivojevic et al. 2020). This could reflect a shared foreign or self antigen exposure, such as EBV infectious mononucleosis, although I did not observe other signals associated with recent or ongoing infection or an active B cell response such as clonal expansion or somatic hypermutation. IGHV3-30 gene usage was variable across the severe ME/CFS patients, but it would be valuable to repeat this analysis with the severe patients stratified by age since severe ME/CFS patients tended to belong to the younger age groups. Mild-moderate ME/CFS patients also had a greater proportion of IgM BCRs that could reflect a more naive B cell population, consistent with findings from a previous study (A. S. Bradley, Ford, and Bansal 2013). It would be valuable to investigate BCR repertoires in Long Covid patients who meet ME/CFS diagnostic criteria, given that the infectious trigger is known in these patients. Sorting antigen-specific B cells against SARS-CoV2 as well as characterising the BCR repertoire of their naive, memory and atypical B cell subsets could reveal whether this disease phenotype is associated with B cell dysfunction. Furthermore, characterising EBV-specific B and T cells in terms of repertoire and, phenotype by flow cytometry, could be insightful given that EBV-infection has been proposed as a potential trigger for ME/CFS, EBV-reactivation can occur upon SARS-CoV2 infection (Gold et al. 2021; Bernal and Whitehurst 2023) and serological evidence of recent EBV reactivation has been associated with fatigue in Long Covid patients (Peluso et al. 2023).

With so little known about the pathophysiology of ME/CFS it is difficult to speculate on the biological underpinnings of the observations made in our study. A large-scale Genome Wide Association Study on ME/CFS due to be published in 2024 will hopefully begin to shed light on what future research priorities should be pursued in this disease. Nonetheless, the finding that BCR repertoires are not suitable as a general diagnostic of ME/CFS is valuable.

In Chapter 3, I observed different kinetics of clonal expansion in a first malaria

challenge compared to a second challenge. I also observed a potential enrichment of un-mutated IgG+ B cells at the c+28 timepoint. Crucially, I did not observe boosting of clones from the first infection upon re-challenge which was an unexpected result. However, these observations may fit with other findings of impaired affinity maturation and selection of BCRs against malaria antigens (Murugan, Buchauer, Triller, Kreschel, Costa, Pidelaserra Marti, et al. 2018). Additionally, early GCs can be highly diverse containing hundreds of different clonotypes (Tas et al. 2016), and naive B cells have been shown to invade established GCs and progressively replace founder clonotypes (Hägglöf et al. 2023; Carvalho et al. 2023), painting a dynamic picture of affinity maturation. How clonal selection and affinity maturation are coordinated in the context of many antigens with significant structural complexity like in malaria is an interesting area of future study.

Our findings of an early IgM mediated response that is not boosted upon re-challenge could also fit with atypical memory B cell responses, which may represent a less active B cell subset, being elicited by malaria challenge. In the absence of B cell phenotype data, this interpretation is speculative. We also cannot exclude the possibility that the timepoints we sampled missed the clones from the first challenge that were expanding after the second challenge. Future studies in the same experimental system would benefit from performing longitudinal B cell phenotyping to identify whether atypical memory B cells are induced and additionally sample the diagnosis+6 timepoint where a transcriptional signature of B cell genes that were dramatically down-regulated was observed by RNA sequencing in a different study of this same cohort (Spence Lab, communications). Finally, our results were not subjected to multiple testing correction and should be treated as a preliminary hypothesis-generating study.

While our analyses did not reveal many stark BCR repertoire changes in either datasets, the BCR data we generated is of good quality and will be deposited in AIRRseq databases. Many tools are still being developed, benchmarked, and validated on publicly available

AIRRseq datasets. It is possible that these data may be useful to other researchers in future. If comprehensive databases of BCR-antigen specificities are eventually developed, the repertoires analysed in this thesis could be annotated with their antigen specificities which would significantly improve the inferences we can draw from the data.

In the fourth chapter, I made progress towards addressing the aforementioned BCR-antigen specificity data gap by optimising reactions for pairing native BCR heavy and light chains for phage display. While I could not replicate the methods used in a published protocol, I successfully amplified heavy and light chains in single cell reactions and obtained paired heavy and light chains from bulk RNA in emulsion RT-PCR. We addressed several issues with the original primer design, including replacing the lambda light chain primers and increasing and matching the TM of the first heavy chain primer set. Additionally, access to equipment to generate single cell emulsions was a barrier for us to progress with the project. I developed a low-cost system for generating emulsions compatible with a range of different PCR reagents in a 3D printable device and a simple vacuum release circuit. This has since allowed members of the lab to try to optimise the reactions using different PCR reagents and makes the platform for generating emulsions much more flexible while still producing consistent emulsions. I have made the technical details of the droplet generator, including the 3D printing files available publicly on GitHub and Thingiverse, which should permit anyone to recreate the droplet generation setup for their own use (see Methods, Chapter 4).

5.2 Limitations and Potential Future Improvements

5.2.1 Technical Limitations

Several experimental improvements could be made to the methods used in the BCR repertoire studies described in this thesis. When sampling PBMCs, we only expect to

capture cells travelling between lymphoid tissues in the blood and therefore it is easy to miss the cells involved in active responses. Longitudinal, rather than cross-sectional sampling of B cells is beneficial since even in healthy controls spontaneous B cell expansions are frequently captured - as was seen in some healthy controls from our ME/CFS and CHMI repertoire studies. These could be the result of a recent subclinical infection or antigen encounter. BCR repertoire studies already have low power to detect the signal of interest because we are more likely to sample naive B cells in blood than activated or memory cells (Waltari et al. 2019). Currently we depend on the ability to detect very stark clonal expansions in order to draw inferences from our data. Sampling the lymph nodes, spleen or other secondary lymphoid tissues in humans is not possible for a majority of studies.

Another limitation of our current protocol is that mRNA content can vary substantially between different cell subsets, particularly between plasma cells and naive B cells. Plasma cells are thought to produce 100 times the number of immunoglobulin transcripts than naive B cells (Eugster et al. 2022). Our current protocols do not account for the fact that different B cells will contribute different quantities of mRNA to the reactions, and therefore, even though we can normalise UMI counts between samples, we cannot accurately estimate diversity: a large clonal expansion seen in the repertoire may be the result of having captured a plasma cell, or a genuine expansion of a particular clonotype. A potential solution is to sequence BCRs from DNA rather than RNA, however, DNA-based methods use V and J gene primers that can introduce amplification bias and also do not provide information about which constant region is expressed (Eugster et al. 2022; Yaari and Kleinstein 2015).

Difficulties in Measuring Diversity

We observed a correlation of increased diversity with larger sampling depth, even when repertoires were downsampled to the same number of UMIs, particularly with the Gini Index. Simulating repertoires that are diverse (**Figure 1A**), clonal (**Figure 1B**) and very clonal (**Figure 1C**), reveals that the Gini Index calculated from clonal repertoires over-estimates the repertoire's clonality when these repertoires are sub-sampled (**Figure 1D**). Both the distributions with clonal expansions had a statistically significant negative correlation of Gini Index with increased sampling depth, although the effect was much stronger for the very uneven distribution (simulated in **Figure 1C**). While these simulations likely do not reflect the true complexity of clonality in human repertoires, it does illustrate that the Gini Index can be affected differently by downsampling depending on the shape of the underlying clonal distributions. Shannon Entropy and Simpson Diversity also do not scale linearly with UMI sampling depth, making downsampled repertoires difficult to compare since the diversity index obtained from a downsampled repertoire can be biased depending on the underlying clonotype distributions (Chao and Jost 2015; Hoehn et al. 2015). In our data, the relationship between increased diversity and UMI depth could be explained by several technical sampling artefacts: Firstly, there was a positive correlation of B cell number with UMI count. It is possible that in samples with fewer B cells, multiple mRNA transcripts from a single B cell were reverse transcribed during cDNA synthesis, which would make the repertoire appear more clonal because multiple UMIs would be associated with a given BCR. Conversely, samples with large numbers of B cells would have more diverse mRNA pools and the likelihood of sampling multiple transcripts from one cell would be lower. Thus repertoires with low numbers of cells may have higher UMI coverage per cell than repertoires with many cells where mRNA from a given cell is unlikely to be sampled twice.

Secondly, samples with lower concentrations of mRNA were perhaps more likely to

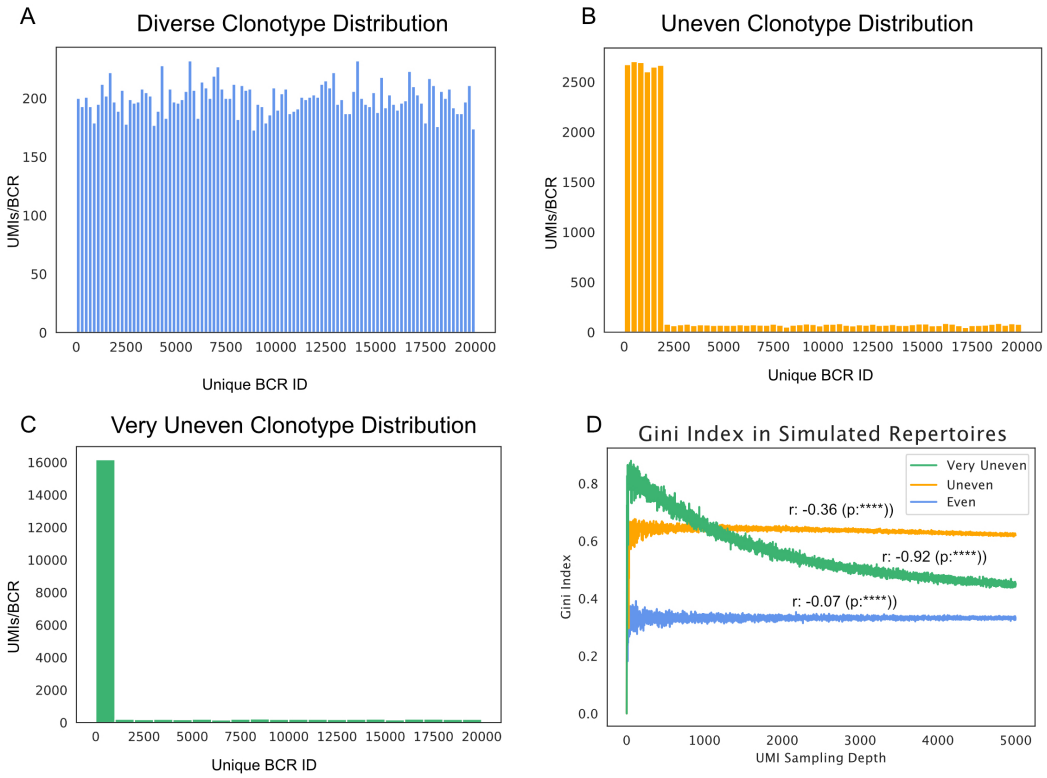


Figure 1: Relationship Between Gini Index and UMI Sampling Depth in Simulated Data. Repertoires with 20,000 potential unique BCRs were simulated for **A)** a very diverse distribution, **B)** an uneven distribution and **C)** a very uneven clonotype distribution. **D)** These simulated repertoires were then repeatedly sampled at different sampling depths, and the Gini Index calculated for each sub-sampled repertoire. Pearson correlation coefficients and associated P-values are shown for each curve. **** $p \leq 0.0001$

have multiple transcription events per mRNA molecule. This again would artifically inflate the number of UMIs associated with a unique BCR.

Finally, in small samples the effect of capturing activated B cells or plasma cells with high mRNA content may have a bigger impact on clonality than in samples with large numbers of cells, due to sampling bias and coverage of the repertoire.

It is difficult to know at which stage these technical or sampling biases were introduced, but possible improvements could be normalising input mRNA concentrations for each repertoire by Nanodrop quantification or using a Qubit. Performing serial dilutions on the same mRNA sample and assessing UMI sampling depth and clonality across these

diluted reactions could also shed light on whether clonality is artificially inflated when lower concentrations of RNA are used, due to multiple transcription events arising from the same mRNA molecule. A second bench-marking experiment could examine whether sampling multiple mRNA transcripts from the same cells is more likely to occur with fewer input cells. This could be achieved by sorting naive B cells from the same donor and performing repertoire sequencing on serially diluted pools of B cells. Using naive B cells would make it less likely for individual cells to have substantially higher mRNA content. Samples with lower B cell numbers would be expected to have more UMIs associated with each BCR (ie. higher coverage). Finally, spiking in known numbers of a clonal B cell line (for example Raji B cells) could be useful to include in our future repertoire sequencing experiments to allow normalisation of coverage between samples, based on UMIs captured for the known spike-in. While cell and RNA spike-ins have been used to benchmark TCR and BCR repertoire sequencing methods, these methods are not routinely incorporated into BCR repertoire studies (Barennes et al. 2021; K. Peng, Nowicki, et al. 2022; Khan et al. 2016).

Diversity metrics other than those that capture large clonal expansions- such as Simpson's Diversity or the Berger-Parker Index - can be challenging to interpret biologically in the context of BCR repertoires. While large clonal expansions likely represent a recent or ongoing immune challenge or a malignancy, it is less clear what smaller expansions detected only by Shannon Entropy or the Gini Index might represent biologically (Hoehn et al. 2015). It would be interesting to combine longitudinal sampling of BCR repertoires from the blood in a mouse model of immunisation with haptens, vaccines or complex antigens such as parasites, and assess how the detection of clonal signatures in these different immune challenges scales with cell and UMI sampling depth: How many daughter cells are required to observe a clonal signature using different diversity metrics and what minimum thresholds of cell and UMI sampling depth are optimal to detect clonal expansion.

Defining Clones

The AIRRseq community would greatly benefit from a rigorous definition of clonality in BCR repertoires. There is currently no consensus on which diversity metrics are used or which are the most biologically meaningful considering the biology of B cell clonal expansion. Furthermore, there is no consensus definition of a "clone" in BCR repertoire studies. In some papers diversity is quantified using unique CDR3s, in others, BCRs are clustered based on parameters like V-J gene usage, CDR3 length and edit distance to assign BCRs to clonotypes. It is probably more meaningful to assign highly similar BCRs to clonotypes, since we know that SHM is likely to introduce nonsynonymous mutations into the CDR3. We also cannot exclude the possibility that CDR3s with different amino acids target the same antigen. Therefore, we might underestimate the extent to which a repertoire has been skewed towards a pathogen-specific response when quantifying BCR or clonotype diversity based on sequence identity alone. An alternative approach is to identify BCRs with structural similarities and cluster BCRs with similar "paratopes" instead of amino acid or nucleotide sequences (Leem et al. 2016). The BCR structures of SARS-CoV2 neutralising antibodies have been used to identify additional SARS-CoV2 neutralising antibodies by selecting BCRs with similar CDR3 structures (Cao et al. 2020). Annotating BCRs with information about their CDR structure has shown that antigen-naive B cell paratopes are highly conserved across individuals, while differentiated B cell types use more personalised and divergent CDR structures (Kovaltsuk et al. 2020). Given the diversity of BCR repertoires, more experimental evidence is needed to support whether predicted structural similarities consistently result in similar antigen-specificities and whether BCRs can be grouped reliably based on their paratopes.

Improving V Gene Reference Databases

Studying V gene usage is inherently limited by the completeness of reference databases. IMGT is the most frequently used repository for V genes, but novel alleles that have not yet been added to the databases are frequently identified. Several putative SNPs in V genes were identified in Chapter 2. V gene misassignment can impact both on V gene usage analysis and the accuracy of quantifying somatic hypermutation, particularly in populations that are under-represented in the reference databases. In future, sequencing repertoires and genomic DNA from individuals from diverse populations will improve the accuracy of the V, D and J gene segment reference databases (K. Peng, Safonova, et al. 2021; Mikocziova, Greiff, and Sollid 2021).

Finally, comprehensive bench-marking and validated workflows need to be generated for AIRRseq data. There are a plethora of tools, protocols and analyses that would benefit from standardisation.

5.2.2 Potential Technical Improvements for Future Studies

Knowing the cell subsets that the BCRs originated from would make it simpler to interpret repertoire signatures. This could easily be achieved by, for example, sorting B cells into naive, memory, atypical and plasma cell subsets prior to performing repertoire sequencing. This would likely improve the power to detect biologically meaningful signals in the repertoire, particularly by excluding naive B cells from clonality and somatic hypermutation analyses. Interpreting differences in V gene usage, like the IGHV3-30 skew observed in ME/CFS, can also be challenging in the absence of strong clonal signatures. Sorting cells into sub-populations would allow us to observe whether V gene skews are found in the naive BCR repertoire and therefore represent a potential genetic or environmental predisposition resulting in a skewed BCR repertoire, or whether signatures are found in the memory population and are more likely the result of a shared pathogen or antigen

exposure. In our data, the difference in clonal expansion at day of diagnosis between the first and second infection in the CHMI repertoires could simply be explained by a more naive B cell population remaining in the blood at day of diagnosis in the second infection. Separating naive from antigen-experienced B cells as a minimum is likely useful in improving the inferences drawn from BCR repertoire signatures.

The work described in Chapter 4 could have enriched the repertoire sequencing work we performed in other projects. For example, if we had the native BCR phage display system working we might have been able to use blood from an additional timepoint in the CHMI trial to identify malaria-specific clonotypes and examine their trajectories in the rest of the data. This could have helped us to identify whether novel antigen-specific clonotypes are selected upon re-challenge, or whether antigen-specific clones from the previous infection are maintained or boosted in low number upon re-challenge. We could also have identified the antigenic targets of the unmutated IgG BCRs observed at the c+28 timepoints. This would have significantly improved the utility of our observations.

Another method that could be applied is to use single cell transcriptomics alongside BCR sequencing to help us understand how the repertoire is changing alongside the B cell transcriptional phenotype in response to infection or in autoimmunity (Z. Zhang et al. 2022). Clonal lineage analysis can also validate developmental trajectories inferred from pseudo-time analysis. Another advantage of single cell sequencing is that it allows the identification of the cognate heavy and light chain in a given B cell so if particular clonotypes are of interest, these antibodies can be synthesised, cloned into expression vectors and tested for their specificity using peptide or antigen arrays (Cao et al. 2020).

5.3 The Promise of AIRRseq

Currently AIRRseq is best applied in experiments where other data such as flow cytometry or RNA sequencing can help contextualise the findings. If we were able to map antigen

specificity onto BCR repertoires at scale, it would enable us to understand which signatures were specific to the antigens of interest, as well as understanding how antigen exposure alters BCR repertoires depending on what infections they experience or how adaptive immunity changes with age. For example, some vaccines and infections confer lifelong immunity, such as measles, mumps and varicella zoster, while others such as tetanus require intermittent boosting (Amanna, Carlson, and Slifka 2007). In malaria frequent infection is required to achieve non-sterilising immunity. Examining the specificities of repertoires generated in response to vaccines which confer sterilising immunity and those which produce less efficient antibody responses would help reveal whether defects lie with non-specific B cell activation or affinity maturation. Knowing the specificity of BCR repertoires would unlock the full translational potential of AIRRseq data and facilitate the development of vaccines, monoclonal antibodies and precision medicine. It would also allow us to revisit archived repertoire data and understand more about the complexity of adaptive immune responses across many disease settings.

While the principles of Burnet's clonal selection theory still hold almost 70 years after it was first proposed, and our understanding of immunology has since been revolutionised by new findings and technologies, it is clear that clonal selection and adaptive immune memory development are complex processes and there is still much to be discovered. AIRRseq is a technology that allows us to follow the dynamics of adpative immune responses at high throughput and its full potential will be unlocked once repertoires and their antigen-specificities can be studied side by side.

References

Adler, Adam S. et al. (Nov. 2017). "Rare, high-affinity anti-pathogen antibodies from human repertoires, discovered using microfluidics and molecular genomics". In: *mAbs* 9 (8), pp. 1282–1296. ISSN: 19420870. DOI: 10.1080/19420862.2017.1371383/SUPPL_FILE/KMAB_A_1371383_SM8285.ZIP.

Akbar, Rahmad et al. (Mar. 2021). "A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding". In: *CellReports* 34 (11), p. 108856. DOI: 10.1016/j.celrep.2021.108856.

Akkaya, Munir, Kihyuck Kwak, and Susan K Pierce (2020). "B cell memory: building two walls of protection against pathogens". In: *Nature Reviews Immunology* 20.4, pp. 229–238.

Akpogheneta, Onome J. et al. (Apr. 2008). "Duration of naturally acquired antibody responses to blood-stage Plasmodium falciparum is age dependent and antigen specific". In: *Infection and Immunity* 76 (4), pp. 1748–1755. ISSN: 00199567. DOI: 10.1128/IAI.01333-07.

Albright, Frederick et al. (2011). "Evidence for a heritable predisposition to Chronic Fatigue Syndrome". In: *BMC neurology* 11, pp. 1–6.

Alfaleh, Mohamed A. et al. (2020). *Phage Display Derived Monoclonal Antibodies: From Bench to Bedside*. DOI: 10.3389/fimmu.2020.01986.

Amanna, Ian J, Nichole E Carlson, and Mark K Slifka (2007). "Duration of humoral immunity to common viral and vaccine antigens". In: *New England Journal of Medicine* 357.19, pp. 1903–1915.

Anaya, Juan-Manuel et al. (2012). "The multiple autoimmune syndromes. A clue for the autoimmune tautology". In: *Clinical reviews in allergy & immunology* 43, pp. 256–264.

Andrade, Mônica V. et al. (Dec. 2022). "The economic burden of malaria: a systematic review". In: *Malaria Journal* 21 (1), p. 283. ISSN: 14752875. DOI: 10.1186/S12936-022-04303-6. URL: /pmc/articles/PMC9533489/%20/pmc/articles/PMC9533489/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9533489/.

André, Ana S. et al. (Sept. 2022). "In vivo Phage Display: A promising selection strategy for the improvement of antibody targeting and drug delivery properties". In: *Frontiers in Microbiology* 13, p. 3704. ISSN: 1664302X. DOI: 10.3389/FMICB.2022.962124/BIBTEX.

Andrews, Simon et al. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Avnir, Yuval et al. (2016). "IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity". In: *Scientific reports* 6.1, p. 20842.

Bach, Florian A et al. (2021). "Mapping T cell activation and differentiation at single cell resolution in naive hosts infected with Plasmodium vivax". In: *medRxiv*, pp. 2021–03.

Bakken, Inger Johanne et al. (2014). "Two age peaks in the incidence of chronic fatigue syndrome/myalgic encephalomyelitis: a population-based registry study from Norway 2008-2012". In: *BMC medicine* 12.1, pp. 1–7.

Barennes, Pierre et al. (2021). "Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases". In: *Nature biotechnology* 39.2, pp. 236–245.

Bashford-Rogers, Rachael JM, Kenneth GC Smith, and David C. Thomas (2018). "Antibody repertoire analysis in polygenic autoimmune diseases". In: *Immunology* 155.1, pp. 3–17.

Bashford-Rogers, RJM et al. (2019). "Analysis of the B cell receptor repertoire in six immune-mediated diseases". In: *Nature* 574.7776, pp. 122–126.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bateman, Lucinda et al. (2021). "Myalgic encephalomyelitis/chronic fatigue syndrome: essentials of diagnosis and management". In: *Mayo clinic proceedings*. Vol. 96. 11. Elsevier, pp. 2861–2878.

Bernal, Keishanne Danielle E and Christopher B Whitehurst (2023). "Incidence of Epstein-Barr virus reactivation is elevated in COVID-19 patients". In: *Virus Research* 334, p. 199157.

Bernard, Nicholas J. (Dec. 2016). "When humoral became cellular". In: *Nature Immunology* 17 (S1), S9–S9. ISSN: 1529-2908. DOI: 10.1038/ni.3604.

Bjornevik, Kjetil et al. (2022). "Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis". In: *Science* 375.6578, pp. 296–301.

Bondt, Albert et al. (2021). "Human plasma IgG1 repertoires are simple, unique, and dynamic". In: *Cell Systems* 12 (12). ISSN: 24054720. DOI: 10.1016/j.cels.2021.08.008.

Bouquet, Jerome et al. (2017). "RNA-seq analysis of gene expression, viral pathogen, and B-cell/T-cell receptor signatures in complex chronic disease". In: *Clinical Infectious Diseases* 64.4, pp. 476–481.

Boyd, Scott D et al. (2010). "Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements". In: *The Journal of Immunology* 184.12, pp. 6986–6992.

Boyle, MJ et al. (2019). *IgM in human immunity to Plasmodium falciparum malaria. Sci Adv 5: eaax4489*.

Braddom, Ashley E et al. (2021). "B cell receptor repertoire analysis in malaria-naive and malaria-experienced individuals reveals unique characteristics of atypical memory B cells". In: *Msphere* 6.5, e00726–21.

Braddom, Ashley E. et al. (Nov. 2020). "Potential functions of atypical memory B cells in Plasmodium-exposed individuals". In: *International Journal for Parasitology* 50 (13), pp. 1033–1042. ISSN: 00207519. DOI: 10.1016/j.ijpara.2020.08.003.

Bradley, A S, B Ford, and A S Bansal (Apr. 2013). "Altered functional B cell subset populations in patients with chronic fatigue syndrome compared to healthy controls." In: *Clinical and experimental immunology* 172 (1), pp. 73–80. ISSN: 1365-2249. DOI: 10.1111/cei.12043.

Bretherick, Andrew D et al. (2023). "Typing myalgic encephalomyelitis by infection at onset: A DecodeME study". In: *NIHR Open Research* 3, p. 20.

Briney, Bryan et al. (2019). "Commonality despite exceptional diversity in the baseline human antibody repertoire". In: *Nature* 566.7744, pp. 393–397.

Büdingen, H-Christian von et al. (2012). "B cell exchange across the blood-brain barrier in multiple sclerosis". In: *The Journal of clinical investigation* 122.12, pp. 4533–4543.

Burnet, Frank Macfarlane et al. (1957). "A modification of Jerne's theory of antibody production using the concept of clonal selection." In: *Australian Journal of Science* 20.3, pp. 67–9.

Bynke, Annie et al. (2020). "Autoantibodies to beta-adrenergic and muscarinic cholinergic receptors in Myalgic Encephalomyelitis (ME) patients–A validation study in plasma and cerebrospinal fluid from two Swedish cohorts". In: *Brain, Behavior, and Immunity-Health* 7, p. 100107.

Cabral-Marques, Otavio et al. (2018). "GPCR-specific autoantibody signatures are associated with physiological and pathological immune homeostasis". In: *Nature communications* 9.1, p. 5224.

Cairns, R and M Hotopf (2005). "A systematic review describing the prognosis of chronic fatigue syndrome". In: *Occupational medicine* 55.1, pp. 20–31.

Cao, Yunlong et al. (2020). "Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells". In: *Cell* 182 (1). ISSN: 10974172. DOI: 10.1016/j.cell.2020.05.025.

Carruthers, Bruce M et al. (2003). "Myalgic Encephalomyelitis/ Chronic Fatigue Syndrome: Clinical Working Case Definition, Diagnostic and Treatment Protocols". In: DOI: 10.1300/J092v11n01_02. URL: http://www.HaworthPress.com.

Carvalho, Renan VH de et al. (2023). "Clonal replacement sustains long-lived germinal centers primed by respiratory viruses". In: *Cell* 186.1, pp. 131–146.

Cavanagh, David R. et al. (Nov. 2004). "Antibodies to the N-Terminal Block 2 of *Plasmodium falciparum* Merozoite Surface Protein 1 Are Associated with Protection against Clinical Malaria". In: *Infection and Immunity* 72 (11), pp. 6492–6502. ISSN: 0019-9567. DOI: 10.1128/IAI.72.11.6492-6502.2004.

CDC (2022). *IOM 2015 Diagnostic Criteria*. URL: https://www.cdc.gov/me-cfs/healthcare-providers/diagnosis/iom-2015-diagnostic-criteria.html.

Chan, Jo-Anne et al. (Feb. 2019). "Antibody Targets on the Surface of Plasmodium falciparum– Infected Erythrocytes That Are Associated With Immunity to Severe Malaria in Young Children". In: *The Journal of Infectious Diseases* 219 (5), pp. 819–828. ISSN: 0022-1899. DOI: 10.1093/infdis/jiy580.

Chang, Cindy M., Joan L. Warren, and Eric A. Engels (Dec. 2012). "Chronic fatigue syndrome and subsequent risk of cancer among elderly US adults". In: *Cancer* 118 (23), pp. 5929–5936. ISSN: 0008543X. DOI: 10.1002/cncr.27612. URL: http://doi.wiley.com/10.1002/cncr.27612.

Chao, Anne, Chun-Huo Chiu, and Lou Jost (2014). "Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through

Hill numbers". In: *Annual review of ecology, evolution, and systematics* 45, pp. 297–324.

Chao, Anne and Lou Jost (2015). "Estimating diversity and entropy profiles via discovery rates of new species". In: *Methods in Ecology and Evolution* 6.8, pp. 873–882.

Chaudhary, Neha and Duane R Wesemann (2018). "Analyzing immunoglobulin repertoires". In: *Frontiers in immunology* 9, p. 462.

Chi, Xiying, Yue Li, and Xiaoyan Qiu (July 2020). "V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation". In: *Immunology* 160 (3), pp. 233–247. ISSN: 0019-2805. DOI: 10.1111/imm.13176.

Chiffelle, Johanna et al. (2020). "T-cell repertoire analysis and metrics of diversity and clonality". In: *Current opinion in biotechnology* 65, pp. 284–295.

Cho, Mi-La et al. (2003). "Association of homozygous deletion of the Humhv3005 and the VH3-30.3 genes with renal involvement in systemic lupus erythematosus". In: *Lupus* 12.5, pp. 400–405.

Chu, Lily et al. (Feb. 2019). "Onset Patterns and Course of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome". In: *Frontiers in Pediatrics* 7, p. 12. ISSN: 2296-2360. DOI: 10.3389/fped.2019.00012.

Clarke, Thomas et al. (2023). "Autoantibody repertoire characterization provides insight into the pathogenesis of monogenic and polygenic autoimmune diseases". In: *Frontiers in Immunology* 14, p. 1106537.

Cliff, Jacqueline M et al. (2019). "Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)". In: *Frontiers in immunology*, p. 796.

Coelho, Camila H., Jacob D. Galson, et al. (Oct. 2022). "B cell clonal expansion and mutation in the immunoglobulin heavy chain variable domain in response to Pfs230 and Pfs25 malaria vaccines". In: *International Journal for Parasitology* 52 (11), pp. 707–710. ISSN: 00207519. DOI: 10.1016/j.ijpara.2021.11.008.

Coelho, Camila H., Steven T. Nadakal, et al. (Nov. 2020). "Antimalarial antibody repertoire defined by plasma IG proteomics and single B cell IG sequencing". In: *JCI Insight* 5 (22). ISSN: 23793708. DOI: 10.1172/JCI.INSIGHT.143471.

Cohen, S., I. A. McGregor, and S. Carrington (1961). "Gamma-globulin and acquired immunity to human malaria". In: *Nature* 192 (4804), pp. 733–737. ISSN: 00280836. DOI: 10.1038/192733A0.

Collins, Andrew M and Corey T Watson (2018). "Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire". In: *Frontiers in immunology* 9, p. 2249.

Colombo, Monica et al. (2000). "Accumulation of clonally related B lymphocytes in the cerebrospinal fluid of multiple sclerosis patients". In: *The Journal of Immunology* 164.5, pp. 2782–2789.

Cooper, Max D (2015). "The early history of B cells". In: *Nature Reviews Immunology* 15.3, pp. 191–197.

Cooper, Max D, Raymond DA Peterson, and Robert A Good (1965). "Delineation of the thymic and bursal lymphoid systems in the chicken". In: *Nature* 205.4967, pp. 143–146.

Cowan, Graeme JM et al. (2019). "Rheumatoid arthritis patients express a skewed repertoire of polyclonal, hypomutated B-cell receptors". In: *bioRxiv*, p. 771949.

Criswell, Lindsey A et al. (2005). "Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes". In: *The American Journal of Human Genetics* 76.4, pp. 561–571.

Crompton, Peter D., Matthew A. Kayala, et al. (Apr. 2010). "A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray". In: *Proceedings of the National Academy of Sciences* 107 (15), pp. 6958–6963. ISSN: 0027-8424. DOI: 10.1073/pnas.1001323107.

Crompton, Peter D., Jacqueline Moebius, et al. (2014). "Malaria immunity in man and mosquito: insights into unsolved mysteries of a deadly infectious disease". In: *Annual review of immunology* 32, p. 157. ISSN: 15453278. DOI: 10.1146/ANNUREV-IMMUNOL-032713-120220.

Crowe, James E (2019). "Influenza virus–specific human antibody repertoire studies". In: *The Journal of Immunology* 202.2, pp. 368–373.

Cyster, Jason G. and Christopher D.C. Allen (Apr. 2019). "B Cell Responses: Cell Interaction Dynamics and Decisions". In: *Cell* 177 (3), pp. 524–540. ISSN: 00928674. DOI: 10.1016/j.cell.2019.03.016.

Dafoe, Whitney (2021). "Extremely severe ME/CFS—a personal account". In: *Healthcare*. Vol. 9. 5. MDPI, p. 504.

Das, Sayoni et al. (2022). "Genetic risk factors for ME/CFS identified using combinatorial analysis". In: *Journal of Translational Medicine* 20.1, pp. 1–20.

Davis, Hannah E et al. (2023). "Long COVID: major findings, mechanisms and recommendations". In: *Nature Reviews Microbiology* 21.3, pp. 133–146.

DeKosky, Brandon J., Takaaki Kojima, et al. (2015). "In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire". In: *Nature Medicine* 21 (1). ISSN: 1546170X. DOI: 10.1038/nm.3743.

DeKosky, Brandon J., Oana I. Lungu, et al. (2016). "Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires". In: *Proceedings of the National Academy of Sciences of the United States of America* 113 (19). ISSN: 10916490. DOI: 10.1073/pnas.1525510113.

Ehrlich, Paul (1900). "Croonian lecture.—On immunity with special reference to cell life". In: *Proceedings of the royal Society of London* 66.424-433, pp. 424–448.

Elsner, Rebecca A and Mark J Shlomchik (2020). "Germinal center and extrafollicular B cell responses in vaccination, immunity, and autoimmunity". In: *Immunity* 53.6, pp. 1136–1150.

Eugster, Anne et al. (2022). "AIRR Community Guide to Planning and Performing AIRR-Seq Experiments". In: *Methods and Protocols*, p. 261.

Falk Hvidberg, Michael et al. (2015). "The health-related quality of life for patients with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)". In: *PloS one* 10.7, e0132421.

Fluge, Øystein, Ingrid Rekeland, et al. (2019). "B-lymphocyte depletion in patients with myalgic encephalomyelitis/chronic fatigue syndrome: a randomized, double-blind, placebo-controlled trial". In: *Annals of internal medicine* 170.9, pp. 585–593.

Fluge, Øystein, Kristin Risa, et al. (2015). "B-lymphocyte depletion in myalgic encephalopathy/chronic fatigue syndrome. An open-label phase II study with rituximab maintenance treatment". In: *PloS one* 10.7, e0129898.

Freitag, Helma et al. (Aug. 2021). "Autoantibodies to Vasoregulative G-Protein-Coupled Receptors Correlate with Symptom Severity, Autonomic Dysfunction and Disability in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome". In: *Journal of Clinical Medicine 2021, Vol. 10, Page 3675* 10 (16), p. 3675. DOI: 10.3390/JCM10163675.

Fukuda, Keiji et al. (Dec. 1994). "The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study". In: *Annals of Internal Medicine* 121 (12), p. 953. ISSN: 0003-4819. DOI: 10.7326/0003-4819-121-12-199412150-00009.

Gadala-Maria, Daniel et al. (2015). "Automated analysis of high-throughput B cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles." In: *Proceedings of the National Academy of Sciency of the United States of America* 122, E862–70.

Gallagher, Arlene M et al. (2004). "Incidence of fatigue symptoms and diagnoses presenting in UK primary care from 1990 to 2001". In: *Journal of the Royal Society of Medicine* 97.12, pp. 571–575.

Galson, Jacob D, Elizabeth A Clutterbuck, et al. (2015). "BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines". In: *Immunology and cell biology* 93.10, pp. 885–895.

Galson, Jacob D, Sebastian Schaetzle, et al. (2020). "Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures". In: *Frontiers in immunology* 11, p. 605170.

Gardner, Malcolm J. et al. (Oct. 2002). "Genome sequence of the human malaria parasite Plasmodium falciparum". In: *Nature* 419 (6906), pp. 498–511. ISSN: 0028-0836. DOI: 10.1038/nature01097.

Gatto, Dominique and Robert Brink (2010). "The germinal center reaction". In: *Journal of Allergy and Clinical Immunology* 126.5, pp. 898–907.

Ghali, Alaa et al. (2020). "Epidemiological and clinical factors associated with post-exertional malaise severity in patients with myalgic encephalomyelitis/chronic fatigue syndrome". In: *Journal of Translational Medicine* 18.1, pp. 1–8.

Gitlin, Alexander D. and Michel C. Nussenzweig (Jan. 2015). "Immunology: Fifty years of B lymphocytes". In: *Nature* 517 (7533), pp. 139–141. ISSN: 0028-0836. DOI: 10.1038/517139a.

Gold, Jeffrey E et al. (2021). "Investigation of long COVID prevalence and its relationship to Epstein-Barr virus reactivation". In: *Pathogens* 10.6, p. 763.

Gonzales, S. Jake et al. (Oct. 2020). "Naturally Acquired Humoral Immunity Against Plasmodium falciparum Malaria". In: *Frontiers in Immunology* 11, p. 594653. ISSN: 16643224. DOI: 10.3389/FIMMU.2020.594653/BIBTEX.

Greiff, Victor et al. (2015). "A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status". In: *Genome medicine* 7.1, pp. 1–15.

Gupta, Namita T et al. (2015). "Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data". In: *Bioinformatics* 31.20, pp. 3356–3358.

Hägglöf, Thomas et al. (2023). "Continuous germinal center invasion contributes to the diversity of the immune response". In: *Cell* 186.1, pp. 147–161.

Hart, Geoffrey T et al. (2016). "The regulation of inherently autoreactive VH4-34–expressing B cells in individuals living in a malaria-endemic area of West Africa". In: *The Journal of Immunology* 197.10, pp. 3841–3849.

Hartwig, Jelka et al. (Mar. 2020). "IgG stimulated β2 adrenergic receptor activation is attenuated in patients with ME/CFS". In: *Brain, Behavior, and Immunity - Health* 3, p. 100047. ISSN: 2666-3546. DOI: 10.1016/J.BBIH.2020.100047.

Hemadou, Audrey et al. (Dec. 2018). "An innovative flow cytometry method to screen human scFv-phages selected by in vivo phage-display in an animal model of atherosclerosis". In: *Scientific Reports* 8 (1). ISSN: 20452322. DOI: 10.1038/s41598-018-33382-2.

Hickie, Ian et al. (2006). "Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study". In: *Bmj* 333.7568, p. 575.

Hoehn, Kenneth B et al. (2015). "Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1676, p. 20140241.

Holla, Prasida et al. (May 2021). "Shared transcriptional profiles of atypical B cells suggest common drivers of expansion and function in malaria, HIV, and autoimmunity". In: *Science Advances* 7 (22). ISSN: 2375-2548. DOI: 10.1126/sciadv.abg8384.

Houde, Damian et al. (Aug. 2010). "Post-translational modifications differentially affect IgG1 conformation and receptor binding." In: *Molecular and cellular proteomics : MCP* 9 (8), pp. 1716–28. ISSN: 1535-9484. DOI: 10.1074/mcp.M900540-MCP200.

Hozumi, N and S Tonegawa (Oct. 1976). "Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions." In: *Proceedings of the National Academy of Sciences* 73 (10), pp. 3628–3632. ISSN: 0027-8424. DOI: 10.1073/pnas.73.10.3628.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Illingworth, Joseph et al. (Feb. 2013). "Chronic Exposure to *Plasmodium falciparum* Is Associated with Phenotypic Evidence of B and T Cell Exhaustion". In: *The Journal of Immunology* 190 (3), pp. 1038–1047. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1202438.

Institute of Medicine (2015). *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*. THE NATIONAL ACADEMIES PRESS.

Jackson, Katherine JL et al. (2014). "Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements". In: *Cell host & microbe* 16.1, pp. 105–114.

James, Ryan G, Christopher J Ellison, and James P Crutchfield (2018). ""dit": a Python package for discrete information theory". In: *Journal of Open Source Software* 3.25, p. 738.

Janeway, CA Jr et al. (2001). *Immunobiology: The Immune System in Health and Disease*. 5th. Garland Sciene.

Jeffery, Geoffrey M. and William E. Collins (July 1999). "A Retrospective Examination of Sporozoite- and Trophozoite-Induced Infections with Plasmodium Falciparum in Patients Previously Infected with Heterologous Species of Plasmodium: Effect on

Development of Parasitologic and Clinical Immunity". In: *The American Journal of Tropical Medicine and Hygiene* 61 (1_Supplement), pp. 36–43. ISSN: 0002-9637. DOI: 10.4269/tropmed.1999.61-036.

Jenks, Scott A et al. (2018). "Distinct effector B cells induced by unregulated toll-like receptor 7 contribute to pathogenic responses in systemic lupus erythematosus". In: *Immunity* 49.4, pp. 725–739.

Jerne, Niels K (1955). "The natural-selection theory of antibody formation". In: *Proceedings of the National Academy of Sciences* 41.11, pp. 849–857.

Johansen, Jorunn N et al. (2015). "Intrathecal BCR transcriptome in multiple sclerosis versus other neuroinflammation: equally diverse and compartmentalized, but more mutated, biased and overlapping with the proteome". In: *Clinical Immunology* 160.2, pp. 211–225.

Jung, David and Frederick W Alt (2004). "Unraveling V (D) J recombination: insights into gene regulation". In: *Cell* 116.2, pp. 299–311.

Kang, Tae Hyun and Baik Lin Seong (Sept. 2020). "Solubility, Stability, and Avidity of Recombinant Antibody Fragments Expressed in Microorganisms". In: *Frontiers in Microbiology* 11, p. 1927. ISSN: 1664302X. DOI: 10.3389/FMICB.2020.01927/BIBTEX.

Katz, Ben Z et al. (2009). "Chronic fatigue syndrome after infectious mononucleosis in adolescents". In: *Pediatrics* 124.1, pp. 189–193.

Kaufmann, Stefan H. E. (June 2017). "Emil von Behring: translational medicine at the dawn of immunology". In: *Nature Reviews Immunology* 17 (6), pp. 341–343. ISSN: 1474-1733. DOI: 10.1038/nri.2017.37.

Kedor, Claudia et al. (2022). "A prospective observational study of post-COVID-19 chronic fatigue syndrome following the first pandemic wave in Germany and biomarkers associated with symptom severity". In: *Nature communications* 13.1, p. 5104.

Keller, Betsy A, John Luke Pryor, and Ludovic Giloteaux (2014). "Inability of myalgic encephalomyelitis/chronic fatigue syndrome patients to reproduce VO2peak indicates functional impairment". In: *Journal of translational medicine* 12.1, pp. 1–10.

Khan, Tarik A et al. (2016). "Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting". In: *Science advances* 2.3, e1501371.

Kinyanjui, Samson M et al. (Dec. 2007). "IgG antibody responses to Plasmodium falciparum merozoite antigens in Kenyan children have a short half-life". In: *Malaria Journal* 6 (1), p. 82. ISSN: 1475-2875. DOI: 10.1186/1475-2875-6-82.

Kissel, Theresa et al. (2022). "Surface Ig variable domain glycosylation affects autoantigen binding and acts as threshold for human autoreactive B cell activation". In: *Science advances* 8.6, eabm1759.

Klein, Jon et al. (2022). "Distinguishing features of Long COVID identified through immune profiling". In: *MedRxiv*, pp. 2022–08.

Kotagiri, Prasanti et al. (2022). "B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination". In: *Cell reports* 38.7.

Kovaltsuk, Aleksandr et al. (2020). "Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice". In: *PLoS computational biology* 16.2, e1007636.

Kuo, Tracy C and Mark S Schlissel (2009). "Mechanisms controlling expression of the RAG locus during lymphocyte development". In: *Current opinion in immunology* 21.2, pp. 173–178.

Lande, Asgeir et al. (Dec. 2020). "Human Leukocyte Antigen alleles associated with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)". In: *Scientific Reports* 10 (1), pp. 1–8. ISSN: 20452322. DOI: 10.1038/s41598-020-62157-x.

Langhorne, Jean et al. (July 2008). "Immunity to malaria: more questions than answers". In: *Nature Immunology 2008 9:7* 9 (7), pp. 725–732. ISSN: 1529-2916. DOI: 10.1038/ni.f.205. URL: https://www.nature.com/articles/ni.f.205.

Lantsova, VB, AS Gerasimov, and EK Sepp (2013). "The role of ADRB2 in myasthenia: genetic and immunological factors". In: *Bulletin of experimental biology and medicine* 154, pp. 351–353.

Lanz, Tobias V et al. (2022). "Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM". In: *Nature* 603 (7900), pp. 321–327. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04432-7.

Leem, Jinwoo et al. (2016). "ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation". In: *MAbs*. Vol. 8. 7. Taylor & Francis, pp. 1259–1268.

Lefranc, Marie-Paule et al. (2005). "IMGT, the international ImMunoGeneTics information system®". In: *Nucleic acids research* 33.suppl_1, pp. D593–D597.

Li, Yucheng et al. (1996). "The I binding specificity of human VH4-34 (VH4-21) encoded antibodies is determined by both VHFramework region 1 and complementarity determining region 3". In: *Journal of molecular biology* 256.3, pp. 577–589.

Lim, Eun-Jin and Chang-Gue Son (2020). "Review of case definitions for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)". In: *Journal of translational medicine* 18.1, pp. 1–10.

Liu, Qiao et al. (July 2021). "Trends of the global, regional and national incidence of malaria in 204 countries from 1990 to 2019 and implications for malaria prevention". In: *Journal of Travel Medicine* 28 (5). ISSN: 17088305. DOI: 10.1093/JTM/TAAB046. URL: https://dx.doi.org/10.1093/jtm/taab046.

Liu, S et al. (2017). "Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus". In: *Genes & Immunity* 18.1, pp. 22–27.

Liu, Yaohui et al. (Nov. 2018). "High-throughput reformatting of phage-displayed antibody fragments to IgGs by one-step emulsion PCR". In: *Protein Engineering, Design and Selection* 31 (11). Ed. by Anna Wu, pp. 427–436. ISSN: 1741-0126. DOI: 10.1093/protein/gzz004. URL: https://academic.oup.com/peds/article/31/11/427/5433372.

Loebel, Madlen et al. (2016). "Antibodies to β adrenergic and muscarinic cholinergic receptors in patients with Chronic Fatigue Syndrome". In: *Brain, behavior, and immunity* 52, pp. 32–39.

Lomakin, Yakov A et al. (2022). "Deconvolution of B cell receptor repertoire in multiple sclerosis patients revealed a delay in tBreg maturation". In: *Frontiers in immunology* 13, p. 803229.

Lu, Ruei-Min et al. (2020). "Development of therapeutic antibodies for the treatment of diseases". In: *Journal of biomedical science* 27.1, pp. 1–30.

Ly, Ann and Diana S. Hansen (Apr. 2019). "Development of B cell memory in malaria". In: *Frontiers in Immunology* 10 (APR), p. 435267. ISSN: 16643224. DOI: 10.3389/FIMMU.2019.00559/BIBTEX.

Ma, Jennifer et al. (Mar. 2021). "Microdroplet-based one-step RT-PCR for ultrahigh throughput single-cell multiplex gene expression analysis and rare cell detection". In: *Scientific Reports 2021 11:1* 11 (1), pp. 1–18. ISSN: 2045-2322. DOI: 10.1038/s41598-021-86087-4.

Magnus, Per et al. (2015). "Chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) is associated with pandemic influenza infection, but not with an adjuvanted pandemic influenza vaccine". In: *Vaccine* 33.46, pp. 6173–6177.

Mandel-Brehm, Caleigh et al. (2021). "Elevated N-linked glycosylation of IgG V regions in myasthenia gravis disease subtypes". In: *The Journal of Immunology* 207.8, pp. 2005–2014.

Marks, James D. et al. (Dec. 1991). "By-passing immunization". In: *Journal of Molecular Biology* 222 (3), pp. 581–597. ISSN: 00222836. DOI: 10.1016/0022-2836(91)90498-U.

Maul, Robert W and Patricia J Gearhart (2010). "Controlling somatic hypermutation in immunoglobulin variable and switch regions". In: *Immunologic research* 47, pp. 113–122.

McNamara, Hayley A. et al. (Oct. 2020). "Antibody Feedback Limits the Expansion of B Cell Responses to Malaria Vaccination but Drives Diversification of the Humoral Response". In: *Cell Host and Microbe* 28 (4), 572–585.e7. ISSN: 19313128. DOI: 10.1016/j.chom.2020.07.001.

Mikocziova, Ivana, Victor Greiff, and Ludvig M Sollid (2021). "Immunoglobulin germline gene variation and its impact on human disease". In: *Genes & Immunity* 22.4, pp. 205–217.

Milivojevic, Milica et al. (July 2020). "Plasma proteomic profiling suggests an association between antigen driven clonal B cell expansion and ME/CFS". In: *PLOS ONE* 15 (7), e0236148. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0236148.

Minassian, Angela M et al. (2021). "Reduced blood-stage malaria growth and immune correlates in humans following RH5 vaccination". In: *Med* 2.6, pp. 701–719.

Moir, Susan et al. (2008). "Evidence for HIV-associated B cell exhaustion in a dysfunctional memory B cell compartment in HIV-infected viremic individuals". In: *The Journal of experimental medicine* 205.8, pp. 1797–1805.

Moslehi, Roxana, Anil Kumar, and Amiran Dzutsev (2022). "Increased risks of cancer and autoimmune disease among the first-degree relatives of patients with myalgic encephalomyelitis (ME)/chronic fatigue syndrome (CFS)". In: *Cancer Research* 82.12_Supplement, pp. 34–34.

Muellenbeck, Matthias F. et al. (2013). "Atypical and classical memory B cells produce plasmodium falciparum neutralizing antibodies". In: *Journal of Experimental Medicine* 210 (2). ISSN: 00221007. DOI: 10.1084/jem.20121970.

Murugan, Rajagopal, Lisa Buchauer, Gianna Triller, Cornelia Kreschel, Giulia Costa, Gemma Pidelaserra Martí, et al. (2018). "Clonal selection drives protective memory B cell responses in controlled human malaria infection". In: *Science Immunology* 3 (20). ISSN: 24709468. DOI: 10.1126/sciimmunol.aap8029.

Murugan, Rajagopal, Lisa Buchauer, Gianna Triller, Cornelia Kreschel, Giulia Costa, Gemma Pidelaserra Marti, et al. (2018). "Clonal selection drives protective memory B cell responses in controlled human malaria infection". In: *Science immunology* 3.20, eaap8029.

Murugan, Rajagopal, Katharina Imkeller, et al. (2015). "Direct high-throughput amplification and sequencing of immunoglobulin genes from single human B cells". In: *European Journal of Immunology* 45 (9). ISSN: 15214141. DOI: 10.1002/eji.201545526.

Ndungu, Francis Maina et al. (May 2012). "Memory B cells are a more reliable archive for historical antimalarial responses than plasma antibodies in no-longer exposed children". In: *Proceedings of the National Academy of Sciences* 109 (21), pp. 8247–8252. ISSN: 0027-8424. DOI: 10.1073/pnas.1200472109.

Nelson, Maximillian J et al. (2019). "Diagnostic sensitivity of 2-day cardiopulmonary exercise testing in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome". In: *Journal of translational medicine* 17.1, pp. 1–8.

Nemazee, David (2017). "Mechanisms of central tolerance for B cells". In: *Nature Reviews Immunology* 17.5, pp. 281–294.

NICE (2007). "Introduction | Chronic fatigue syndrome/myalgic encephalomyelitis (or encephalopathy): diagnosis and management | Guidance | NICE". In.

Nieuwenhuis, P. and D. Opstelten (1984). "Functional anatomy of germinal centers". In: *American Journal of Anatomy* 170.3, pp. 421–435. DOI: https://doi.org/10.1002/aja.1001700315. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aja.1001700315.

NIMH (2023). *Depression*. URL: https://www.nimh.nih.gov/health/topics/depression.

Nossal, Gustav JV and Joshua Lederberg (1958). "Antibody production by single cells". In: *Nature* 181.4620, pp. 1419–1420.

Obeng-Adjei, Nyamekye et al. (2017). "Malaria-induced interferon-γ drives the expansion of Tbethi atypical memory B cells". In: *PLoS pathogens* 13.9, e1006576.

Office for National Statistics (July 2023). *Self-reported long COVID symptoms, UK: 10 July 2023*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins%5C/selfreportedlongcovidsymptom 10july2023.

Ofori, Michael F. et al. (June 2002). "Malaria-Induced Acquisition of Antibodies to Plasmodium falciparum Variant Surface Antigens". In: *Infection and Immunity* 70 (6), pp. 2982–2988. ISSN: 0019-9567. DOI: 10.1128/IAI.70.6.2982-2988.2002.

Ota, Mineto et al. (2023). "Multimodal repertoire analysis unveils B cell biology in immune-mediated diseases". In: *Annals of the Rheumatic Diseases* 82.11, pp. 1455–1463.

Owens, Gregory P et al. (2001). "The immunoglobulin G heavy chain repertoire in multiple sclerosis plaques is distinct from the heavy chain repertoire in peripheral blood lymphocytes". In: *Clinical Immunology* 98.2, pp. 258–263.

Panda, Saswati and Jeak L Ding (2015). "Natural antibodies bridge innate and adaptive immunity". In: *The journal of immunology* 194.1, pp. 13–20.

Park, Jong-Chan et al. (2022). "Association of B cell profile and receptor repertoire with the progression of Alzheimer's disease". In: *Cell Reports* 40.12.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pelissier, Aurelien et al. (2023). "Exploring the impact of clonal definition on B-cell diversity: implications for the analysis of immune repertoires". In: *Frontiers in Immunology* 14, p. 1123968.

Peluso, Michael J et al. (2023). "Chronic viral coinfections differentially affect the likelihood of developing long COVID". In: *The Journal of Clinical Investigation* 133.3.

Pendergrast, Tricia et al. (2016). "Housebound versus nonhousebound patients with myalgic encephalomyelitis and chronic fatigue syndrome". In: *Chronic illness* 12.4, pp. 292–307.

Peng, Kerui, Theodore Scott Nowicki, et al. (2022). "Rigorous benchmarking of T cell receptor repertoire profiling methods for cancer RNA sequencing". In: *medRxiv*, pp. 2022–03.

Peng, Kerui, Yana Safonova, et al. (2021). "Diversity in immunogenomics: the value and the challenge". In: *Nature methods* 18.6, pp. 588–591.

Pérez-Mazliah, Damián, Peter J Gardner, et al. (Nov. 2018). "Plasmodium-specific atypical memory B cells are short-lived activated B cells". In: *eLife* 7. ISSN: 2050-084X. DOI: 10.7554/eLife.39800.

Pérez-Mazliah, Damián, Francis M. Ndungu, et al. (Jan. 2019). "B-cell memory in malaria: Myths and realities". In: *Immunological Reviews* 293 (1), pp. 57–69. ISSN: 1600-065X. DOI: 10.1111/IMR.12822.

Pohl, Kai and Ian A. Cockburn (Aug. 2022). "Innate immunity to malaria: The good, the bad and the unknown". In: *Frontiers in Immunology* 13. ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.914598.

Portugal, Silvia et al. (May 2015). "Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function". In: *eLife* 4 (MAY). ISSN: 2050084X. DOI: 10.7554/ELIFE.07218.

Pricoco, Rafael et al. (2023). "One-Year Follow-up of Young People with ME/CFS Following Infectious Mononucleosis by Epstein-Barr Virus". In: *medRxiv*. DOI: 10.1101/2023.07.24.23293082. eprint: https://www.medrxiv.org/content/early/2023/07/27/2023.07.24.23293082.full.pdf.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rajan, Saravanan et al. (Dec. 2018). "Recombinant human B cell repertoires enable screening for rare, specific, and natively paired antibodies". In: *Communications Biology* 1 (1), p. 5. ISSN: 2399-3642. DOI: 10.1038/s42003-017-0006-2. URL: http://www.nature.com/articles/s42003-017-0006-2.

Ramsay, AM (1986). "Postviral fatigue syndrome: the saga ofthe Royal Free Disease". In: *London: Gower Medical*.

Rasa, Santa et al. (2018). "Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)". In: *Journal of Translational Medicine* 16.1, pp. 1–25.

Rekeland, Ingrid G. et al. (Apr. 2020). "Intravenous Cyclophosphamide in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. An Open-Label Phase II Study". In: *Frontiers in Medicine* 7, p. 162. ISSN: 2296-858X. DOI: 10.3389/fmed.2020.00162. URL: https://www.frontiersin.org/article/10.3389/fmed.2020.00162/full.

Rénia, Laurent and Yun Shan Goh (Nov. 2016). "Malaria Parasites: The Great Escape". In: *Frontiers in Immunology* 7. ISSN: 1664-3224. DOI: 10.3389/fimmu.2016.00463.

Robin, Xavier et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.

Ryg-Cornejo, Victoria et al. (2016). "Severe malaria infections impair germinal center responses by inhibiting T follicular helper cell differentiation". In: *Cell reports* 14.1, pp. 68–81.

Salkeld, Jo et al. (2022). "Repeat controlled human malaria infection of healthy UK adults with blood-stage Plasmodium falciparum: safety and parasite growth dynamics". In: *Frontiers in Immunology*, p. 4669.

Sandler, Carolina X et al. (2022). "Predictors of Chronic Fatigue Syndrome and Mood Disturbance After Acute Infection". In: *Frontiers in Neurology* 13, p. 935442.

Sandoval, Diana Muñoz et al. (2021). "Adaptive T cells regulate disease tolerance in human malaria". In: *MedRxiv*, pp. 2021–08.

Sato, Wakiro et al. (July 2021). "Skewing of the B cell receptor repertoire in myalgic encephalomyelitis/chronic fatigue syndrome". In: *Brain, Behavior, and Immunity* 95, pp. 245–255. ISSN: 0889-1591. DOI: 10.1016/J.BBI.2021.03.023.

Schanz, Merle et al. (2014). "High-throughput sequencing of human immunoglobulin variable regions with subtype identification". In: *PloS one* 9.11, e111726.

Schatorjé, EJH et al. (2014). "Levels of somatic hypermutations in B cell receptors increase during childhood". In: *Clinical & Experimental Immunology* 178.2, pp. 394–398.

Schickel, Jean-Nicolas et al. (2017). "Self-reactive VH4-34–expressing IgG B cells recognize commensal bacteria". In: *Journal of Experimental Medicine* 214.7, pp. 1991–2003.

Scholzen, Anja and Robert W. Sauerwein (Feb. 2016). "Immune activation and induction of memory: lessons learned from controlled human malaria infection with *Plasmodium falciparum*". In: *Parasitology* 143 (2), pp. 224–235. ISSN: 0031-1820. DOI: 10.1017/S0031182015000761.

Schroeder, Jr Harry W and Lisa Cavacini (2010). "Structure and function of immunoglobulins". In: *Journal of allergy and clinical immunology* 125.2, S41–S52.

Seo, Min Jeong et al. (Feb. 2009). "Engineering antibody fragments to fold in the absence of disulfide bonds". In: *Protein Science : A Publication of the Protein Society* 18 (2), p. 259. ISSN: 09618368. DOI: 10.1002/PRO.31.

Setliff, Ian, Wyatt J. McDonnell, et al. (2018). "Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection". In: *Cell Host and Microbe* 23 (6). ISSN: 19346069. DOI: 10.1016/j.chom.2018.05.001.

Setliff, Ian, Andrea R. Shiakolas, et al. (Dec. 2019). "High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity". In: *Cell* 179 (7), 1636–1646.e15. ISSN: 10974172. DOI: 10.1016/j.cell.2019.11.003.

Shikova, Evelina et al. (2020). "Cytomegalovirus, Epstein-Barr virus, and human herpesvirus-6 infections in patients with myalgic encephalomyelitis/chronic fatigue syndrome". In: *Journal of medical virology* 92.12, pp. 3682–3688.

Shugay, Mikhail et al. (Jan. 2018). "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity". In: *Nucleic acids research* 46 (D1), pp. D419–D427. ISSN: 1362-4962. DOI: 10.1093/NAR/GKX760. URL: https://pubmed.ncbi.nlm.nih.gov/28977646/.

Smith, Kerri (2023). "Women's health research lacks funding—in a series of charts". In: *Nature* 617.7959, pp. 28–29.

Smith, N. L. (2022). "Decoding malaria T-cell responses using adaptive immune receptor repertoire sequencing". PhD thesis. University of Edinburgh.

Smith, Natasha L. et al. (Nov. 2020). "A Conserved TCRβ Signature Dominates a Highly Polyclonal T-Cell Expansion During the Acute Phase of a Murine Malaria Infection". In: *Frontiers in Immunology* 11. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.587756.

Spence, Philip J et al. (2013). "Vector transmission regulates immune control of Plasmodium virulence". In: *Nature* 498.7453, pp. 228–231.

Stern, Joel N.H. et al. (Aug. 2014). "B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes". In: *Science Translational Medicine* 6 (248), 248ra107. ISSN: 19466242. DOI: 10.1126/scitranslmed.3008879.

Stewart, Alexander et al. (2022). "Pandemic, epidemic, endemic: B cell repertoire analysis reveals unique anti-viral responses to SARS-CoV-2, Ebola and Respiratory Syncytial Virus". In: *Frontiers in Immunology* 13, p. 807104.

Stussman, B et al. (2020). "Characterization of Post-exertional Malaise in Patients With Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Front Neurol. 2020; 11: 1025". In.

Sundling, Christopher et al. (2019). "B cell profiling in malaria reveals expansion and remodeling of CD11c+ B cell subsets". In: *Jci Insight* 4.9.

Sutton, Henry J. et al. (Feb. 2021). "Atypical B cells are part of an alternative lineage of B cells that participates in responses to vaccination and infection in humans". In: *Cell Reports* 34 (6), p. 108684. ISSN: 22111247. DOI: 10.1016/j.celrep.2020.108684.

Szklarski, Marvin et al. (2021). "Delineating the association between soluble CD26 and autoantibodies against G-protein coupled receptors, immunological and cardiovascular parameters identifies distinct patterns in post-infectious vs. non-infection-triggered Myalgic Encephalomyelitis/Chronic Fatigue Syndrome". In: *Frontiers in Immunology*, p. 1077.

Tang, Catherine et al. (2020). "AID overlapping and Polη hotspots are key features of evolutionary variation within the human antibody heavy chain (IGHV) genes". In: *Frontiers in immunology* 11, p. 788.

Tanno, Hidetaka et al. (Apr. 2020). "A facile technology for the high-throughput sequencing of the paired VH:VL and TCRβ:TCRα repertoires". In: *Science Advances* 6 (17), eaay9093. ISSN: 23752548. DOI: 10.1126/sciadv.aay9093.

Tas, Jeroen MJ et al. (2016). "Visualizing antibody affinity maturation in germinal centers". In: *Science* 351.6277, pp. 1048–1054.

Taylor, Krystyna et al. (2023). "Genetic Risk Factors for Severe and Fatigue Dominant Long COVID and Commonalities with ME/CFS Identified by Combinatorial Analysis". In: *medRxiv*, pp. 2023–07.

Terpilowski, Maksim A (2019). "scikit-posthocs: Pairwise multiple comparison tests in Python". In: *Journal of Open Source Software* 4.36, p. 1169.

The pandas development team (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Tiller, Thomas et al. (2008). "Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning". In: *Journal of Immunological Methods* 329 (1-2). ISSN: 00221759. DOI: 10.1016/j.jim.2007.09.017.

Tipton, Christopher M et al. (2015). "Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus". In: *Nature immunology* 16.7, pp. 755–765.

Titcombe, Philip J. et al. (Dec. 2018). "Pathogenic Citrulline-Multispecific B Cell Receptor Clades in Rheumatoid Arthritis". In: *Arthritis & rheumatology (Hoboken, N.J.)* 70 (12), pp. 1933–1945. ISSN: 2326-5205. DOI: 10.1002/ART.40590.

Trück, Johannes et al. (2021). "Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling". In: *Elife* 10, e66274.

Tu, Ang A et al. (Nov. 2019). "Recovery of Paired T Cell Receptors from Single-cell Seq-Well Libraries". In: DOI: 10.21203/RS.2.13685/V1. URL: https://www.researchsquare.com%20https://protocolexchange.researchsquare.com/article/pex-702/v1.

Tucci, Felicia A et al. (2018). "Biased IGH VDJ gene repertoire and clonal expansions in B cells of chronically hepatitis C virus–infected individuals". In: *Blood, The Journal of the American Society of Hematology* 131.5, pp. 546–557.

Turchaninova, M. A. et al. (2016). "High-quality full-length immunoglobulin profiling with unique molecular barcoding". In: *Nature Protocols*. ISSN: 17502799. DOI: 10.1038/nprot.2016.093.

Twomey, Rosie et al. (2022). "Chronic fatigue and postexertional malaise in people living with long COVID: an observational study". In: *Physical therapy* 102.4, pzac005.

Tyson, Sarah et al. (2022). "Research priorities for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): the results of a James Lind alliance priority setting exercise". In: *Fatigue: Biomedicine, Health & Behavior* 10.4, pp. 200–211.

Valent, Peter et al. (2016). "Paul Ehrlich (1854-1915) and His Contributions to the Foundation and Birth of Translational Medicine". In: *Journal of Innate Immunity* 8 (2), pp. 111–120. ISSN: 1662-811X. DOI: 10.1159/000443526.

Vander Heiden, Jason A et al. (2014). "pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires". In: *Bioinformatics* 30.13, pp. 1930–1932.

Vencovskỳ, J et al. (2002). "Polymorphism in the immunoglobulin VH gene V1-69 affects susceptibility to rheumatoid arthritis in subjects lacking the HLA-DRB1 shared epitope". In: *Rheumatology* 41.4, pp. 401–410.

Vendel, Michelle C. et al. (Oct. 2012). "Secretion from bacterial versus mammalian cells yields a recombinant scFv with variable folding properties". In: *Archives of Biochemistry and Biophysics* 526 (2), pp. 188–193. ISSN: 0003-9861. DOI: 10.1016/J.ABB.2011.12.018.

Vergroesen, Rochelle D. et al. (2019). "N-Glycosylation Site Analysis of Citrullinated Antigen-Specific B-Cell Receptors Indicates Alternative Selection Pathways During

Autoreactive B-Cell Development". In: *Frontiers in Immunology* 10. ISSN: 16643224. DOI: 10.3389/fimmu.2019.02092.

Vidarsson, Gestur, Gillian Dekkers, and Theo Rispens (2014). "IgG subclasses and allotypes: from structure to effector functions". In: *Frontiers in immunology* 5, p. 520.

Vita, Randi et al. (2009). "The Immune Epitope Database 2.0". In: *Nucleic Acids Research* 38 (SUPPL.1). ISSN: 03051048. DOI: 10.1093/nar/gkp1004.

Vogl, Thomas et al. (2022). "Systemic antibody responses against human microbiota flagellins are overrepresented in chronic fatigue syndrome patients". In: *Science advances* 8.38, eabq2422.

Waltari, Eric et al. (2019). "Functional enrichment and analysis of antigen-specific memory B cell antibody repertoires in PBMCs". In: *Frontiers in immunology* 10, p. 1452.

Walter, Michael A et al. (1991). "Susceptibility to multiple sclerosis is associated with the proximal immunoglobulin heavy chain variable region." In: *The Journal of clinical investigation* 87.4, pp. 1266–1273.

Wang, Bo et al. (2018). "Functional interrogation and mining of natively paired human v H:V L antibody repertoires". In: *Nature Biotechnology* 36 (2). ISSN: 15461696. DOI: 10.1038/nbt.4052.

Wang, Meng et al. (2023). "High-throughput single-cell profiling of B cell responses following inactivated influenza vaccination in young and older adults". In: *Aging (Albany NY)* 15.18, p. 9250.

Wang, Ying et al. (2020). "B cell development and maturation". In: *B Cells in Immunity and Tolerance*, pp. 1–22.

Waskom, Michael L. (2021). "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

Watson, Oliver J et al. (2022). "Global impact of the first year of COVID-19 vaccination: a mathematical modelling study". In: *The Lancet Infectious Diseases* 22.9, pp. 1293–1302.

Weiss, Greta E et al. (2011). "A positive correlation between atypical memory B cells and Plasmodium falciparum transmission intensity in cross-sectional studies in Peru and Mali". In: *PloS one* 6.1, e15983.

Weiss, Greta E., Peter D. Crompton, et al. (Aug. 2009). "Atypical Memory B Cells Are Greatly Expanded in Individuals Living in a Malaria-Endemic Area". In: *The Journal of Immunology* 183 (3), pp. 2176–2182. ISSN: 0022-1767. DOI: 10.4049/jimmunol.0901297.

Weiss, Greta E., Boubacar Traore, et al. (May 2010). "The Plasmodium falciparum-Specific Human Memory B Cell Compartment Expands Gradually with Repeated Malaria Infections". In: *PLoS Pathogens* 6 (5), e1000912. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1000912.

Wendel, Ben S. et al. (Sept. 2017). "Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children". In: *Nature Communications* 8 (1), p. 531. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00645-x.

White, Michael T. et al. (Oct. 2014). "Dynamics of the Antibody Response to Plasmodium falciparum Infection in African Children". In: *The Journal of Infectious Diseases* 210 (7), pp. 1115–1122. ISSN: 0022-1899. DOI: 10.1093/infdis/jiu219.

White, PD et al. (1998). "Incidence, risk and prognosis of acute and chronic fatigue syndromes and psychiatric disorders after glandular fever". In: *The British Journal of Psychiatry* 173.6, pp. 475–481.

WHO (2022). *World malaria report 2022*. URL: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wu, Li et al. (2018). "Bidirectional role of β2-adrenergic receptor in autoimmune diseases". In: *Frontiers in pharmacology* 9, p. 1313.

Wykes, Michelle N. and Michael F. Good (Aug. 2009). "What have we learnt from mouse models for the study of malaria?" In: *European Journal of Immunology* 39 (8), pp. 2004–2007. ISSN: 00142980. DOI: 10.1002/eji.200939552.

Yaari, Gur, Jennifer IC Benichou, et al. (2015). "The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1676, p. 20140242.

Yaari, Gur and Steven H Kleinstein (2015). "Practical guidelines for B-cell receptor repertoire sequencing analysis". In: *Genome medicine* 7, pp. 1–14.

Yman, Victor et al. (Dec. 2019). "Antibody responses to merozoite antigens after natural Plasmodium falciparum infection: kinetics and longevity in absence of re-exposure". In: *BMC Medicine* 17 (1), p. 22. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1255-3.

Young, Clara and Robert Brink (2021). "The unique biology of germinal center B cells". In: *Immunity* 54.8, pp. 1652–1664.

Zhang, Ze et al. (2022). "Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse". In: *Nature Machine Intelligence* 4.6, pp. 596–604.

Zinöcker, Severin et al. (Feb. 2015). "The V Gene Repertoires of Classical and Atypical Memory B Cells in Malaria-Susceptible West African Children". In: *The Journal of Immunology* 194 (3), pp. 929–939. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1402168.

Zost, Seth J. et al. (July 2020). "Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein". In: *Nature Medicine*

2020 26:9 26 (9), pp. 1422–1427. ISSN: 1546-170X. DOI: 10.1038/s41591-020-0998-x.

Zumaquero, Esther et al. (2019). "IFNγ induces epigenetic programming of human T-bethi B cells and promotes TLR7/8 and IL-21 induced differentiation". In: *Elife* 8, e41641.