Monitoring immune dynamics following infection and vaccination using B cell receptor sequencing

Velislava Petrova

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of

*Doctor of Philosophy*

Jesus College September 2017
Timshel!
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Velislava Petrova

September 2017
Acknowledgements

I would like to thank Carl Anderson for taking the responsibility of supervising my project during the most challenging stage of my PhD and for providing me with constant support and guidance every day thereafter. His support gave me the security and confidence that made the completion of this PhD possible. Thank you, Carl, for showing me that a good scientist can also be an incredibly selfless and giving human being. I am also very grateful for the mentorship and support of Colin Russell who nurtured my curiosity and passion for science from the start of my PhD and who taught me that every step of the journey has its worth no matter how challenging and demotivating it can be. Thank you, Colin, for teaching me to be brave and resilient and for inspiring me with your drive and dedication.

This PhD showed me the benefit of working in a supportive environment with smart and motivated people. Therefore, I would like to thank my collaborators Rik de Swart, Veronika von Messling and Bevan Sawatsky who were very generous with their time and resources and provided the unique samples and experimental models I had the chance to use in my work. The early stages of this work were conducted in close collaboration with Rachael Bashford-Rogers and Paul Kellam and I would like to thank them for sharing their experience and for providing advice during the methods development aspect of my PhD project.

As part of Team Anderson, I was lucky to work in a group of extremely kind and genuine people. Javi, Liz, Rebecca, Ben, thank you all for being a real joy to work with. I am particularly grateful to my colleague and dear friend Loukas Moutsianas who took the time to read my thesis and to provide useful comments, but more importantly who was always ready
to listen with care and patience, to share a good laugh and to inspire me with his bohemian perfectionsim.

Following this PhD journey took a lot of time, effort and dedication, but it gave me the invaluable gift of good friendships. Over the past few years I met some of the most intelligent, inspiring and giving people I know. I would like to thank Dobromir for giving me his continuous support and care, and for creating for me the feeling of home and family no matter how far away from them I am. I also thank my dearest friend Maria who became closer than a sibling for me and always made me feel understood and accepted with all of my flaws. Special thank you also to Rumen, Dimitris, Martin, Gianmarco and Liliana who were my very supportive crew and showed me the value of real friendships.

This journey started way before the PhD itself, so I would also like to thank those who believed in me way before I did myself. I thank Matt Cotten for teaching me to be pedantic about my work but to never forget the fun and the creative process behind doing science. I can also hardly express my gratitude to Emanuel Schenck who has been my personal and professional mentor, an understanding friend and a source of often sobering but always very genuine advice that could always bring me back the confidence and peace I needed.

Lastly, but most importantly, I would like to thank my parents and family for raising me with love, care and patience and for being there for me as my strongest support system in moments of doubt, despair and disappointment just as much as in moments of joy. I would like to thank them also for teaching me to be kind, honest and giving, to never take anything for granted, and to know that everything meaningful takes time. With their own example they have been my most important teachers of how to live my life with dignity, to work hard for my goals, and to always stay true to myself. I feel truly blessed to have them!

And finally, I am happy to say: “I may not have gone where I intended to go, but I think I have ended up where I needed to be.”
Abstract

Sequencing B and T cell receptor genes allows for detailed characterisation of the genetic diversity underlying adaptive immune responses in health and disease. In the context of infectious diseases this can act as a powerful tool for identification of pathogen-specific immune signatures and genetic determinants of immune memory, protection and response to re-exposure. As part of my PhD I developed and optimised a method for high-resolution profiling of B cell receptor (BCR) immune repertoires based on the barcoded sequencing of the human immunoglobulin genes. The use of molecular barcodes allowed for reduction of technical noise, which can lead to erroneous assignment of lymphocyte function. I applied this methodology to the study of natural infection with measles virus in unvaccinated children.

Childhood measles causes a profound immune suppression, which can last for weeks to months post infection, with large reductions in numbers of circulating B cells. Interestingly, long-term consequences of measles immune suppression result in increased incidence of secondary infections up to 3 years after resolution of measles. Vaccination against measles virus with the MMR vaccine has been a major factor in reducing direct and secondary childhood morbidity and mortality. The maintenance of sufficient global vaccine coverage, however, has been challenging due to the refusal of vaccination, mainly in religious communities, resulting in increasing number of outbreaks worldwide. In addition to the overall drop in measles virus herd immunity, measles-induced immune suppression can compromise immunity to other infectious pathogens, thus complicating global vaccination and surveillance efforts. The exact mechanisms underlying the prolonged immune-suppression associated with measles remain elusive and have not been investigated in humans.
I applied BCR sequencing to characterise the long-term immunological effects of natural measles virus infection in a cohort of unvaccinated children. Specifically, I addressed the re-structuring of immune memory and the possible loss of immunity to non-measles pathogens. My work provided evidence for previously hypothesised depletion of B cell memory pools, referred to as ‘immunological amnesia’. Loss of clonally expanded B memory populations lead to immune re-setting and convergence in repertoire diversity between measles-infected and control groups. In addition to the loss of individual-specific variation in immune memory, a subset of measles-infected individuals exhibited dramatic collapse in the diversity of their naïve B cell compartment, despite the recovery of normal B naïve cell counts. An effect of measles on serological immunity was also demonstrated in a ferret model of measles, where lymphotropic challenge lead to significant loss of vaccine-acquired immunity to influenza virus.

The work presented in this dissertation demonstrates the utility of BCR sequencing for understanding adaptive immune responses in the context of infectious diseases and highlights the potential of this approach to uncover novel mechanisms of immune (dys)function.
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>xvii</td>
</tr>
<tr>
<td>List of tables</td>
<td>xxi</td>
</tr>
<tr>
<td><strong>I Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1 Adaptive immune responses</td>
<td>3</td>
</tr>
<tr>
<td>1.1 Evolution and organisation of the adaptive immune response</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Development of B cell responses</td>
<td>4</td>
</tr>
<tr>
<td>2 Methods to study the evolution of B cell responses</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Molecular approaches</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Genetic approaches for characterisation of B cell diversity</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Characterisation of B cell responses using immunoglobulin sequencing</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Applications of BCR repertoire sequencing</td>
<td>32</td>
</tr>
<tr>
<td>3 Aims and outline of the thesis</td>
<td>37</td>
</tr>
<tr>
<td>4 List of Publications</td>
<td>39</td>
</tr>
<tr>
<td>**II Methods for correction of technical noise in immune repertoire sequencing data</td>
<td>41</td>
</tr>
</tbody>
</table>
# Table of contents

1 Introduction ......................................................... 42  
1.1 High-throughput sequencing of lymphocyte repertoires .......... 42  
1.2 Methods for immune repertoire sequencing ......................... 42  
1.3 The burden of technical noise in immune repertoire sequencing .... 46  
2 Methods .............................................................. 50  
2.1 Samples .................................................................. 50  
2.2 Reverse transcription and amplification with barcoded primers .... 50  
2.3 Next-generation sequencing, sequence quality control (QC) and barcode filtering ........................................ 53  
2.4 Network analysis ...................................................... 54  
2.5 Mismatch profiles of BCR reads in barcode bins ..................... 55  
2.6 Statistical methods .................................................... 55  
3 Results ................................................................. 56  
3.1 Sensitivity of BCR repertoire capture across library preparation methods ........................................... 56  
3.2 Reproducibility of capture of BCR diversity across library preparation methods ........................................ 59  
3.3 Effect of BCR amplification on the estimation of BCR diversity .... 61  
3.4 Method-specific PCR amplification biases result in distinct diversity estimates across amplification protocols ..................... 62  
3.5 Different PCR amplification strategies select for specific chemical properties of incorporated barcodes ..................... 67  
4 Discussion ............................................................ 70
### Table of contents

#### III BCR sequencing for characterisation of antigen-driven B cell evolution  
73

1 Introduction ................................................................. 74

1.1 Determinants of antigen specificity in the BCR molecule .... 74

1.2 Evolution of antibody effector functions ............................ 75

1.3 Antibody classes .......................................................... 76

1.4 Interaction between antibody variable region and isotype in the evolution of antigen-specific responses ................................. 78

1.5 Monitoring the evolution of isotype-specific antibody responses using BCR repertoire sequencing ............................. 79

1.6 Aims ............................................................................. 79

1.7 Colleagues ................................................................. 80

2 Methods ........................................................................ 80

2.1 Samples ........................................................................ 80

2.2 FACS sorting of B memory cell subsets ............................. 82

2.3 Library preparation for isotype-resolved sequencing ............ 83

2.4 Sequencing and barcode filtering .................................... 84

2.5 Network analysis .......................................................... 85

2.6 Transitional frequencies between isotypes ......................... 85

2.7 Conditional probability of a BCR being class-switched, given its variable gene family identity and mutational status .... 86

2.8 Conditional probability of a BCR being switched to a specific isotype, given its variable gene identity .......................... 86
Table of contents

2.9 Conditional probability of a BCR being switched to one or more isotypes, given its variable gene identity ........................................ 87
2.10 Calculation of isotype-specific VJ gene profiles .......................... 88
2.11 Hierarchical clustering of isotype-specific VJ gene profiles ......... 88
3 Results ..................................................................................... 89
3.1 Isotype-resolved BCR sequencing with 3’MPLX PCR amplification 89
3.2 Isotype-resolved BCR sequencing enables ‘genetic sorting’ of isotype-specific B cell populations .................................................. 90
3.3 Isotype composition of BCR repertoires at the extremes of variable gene diversity .............................................................. 92
3.4 Multi-switched clones and isotype co-evolution are characteristic of healthy repertoires but rare in CLL ........................................ 93
3.5 Switching to multiple isotypes is associated with higher SHM ....... 94
3.6 Multi-switched clones are generated after substantial clonal expansion 96
3.7 SHM and variable gene family of a BCR affect class-switch status . 98
3.8 Variable gene identity of BCRs affects class-switch trajectories to different isotype classes ...................................................... 99
3.9 Isotype-specific repertoires exhibit distinct signatures of variable gene usage ................................................................. 101
3.10 Isotype-specific variable gene repertoires are conserved across individuals ................................................................. 102
4 Discussion ............................................................................ 103

IV Characterisation of measles-associated immune suppression using B cell receptor sequencing 107
<table>
<thead>
<tr>
<th></th>
<th>1 Introduction</th>
<th>108</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1 Measles virus pathogenesis</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>1.2 Immunity to measles virus</td>
<td>110</td>
</tr>
<tr>
<td>1</td>
<td>1.3 Measles-associated immune-suppression</td>
<td>113</td>
</tr>
<tr>
<td>1</td>
<td>1.4 Aims</td>
<td>117</td>
</tr>
<tr>
<td>1</td>
<td>1.5 Colleagues</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>2 Methods</td>
<td>118</td>
</tr>
<tr>
<td>2</td>
<td>2.1 Samples</td>
<td>118</td>
</tr>
<tr>
<td>2</td>
<td>2.2 FACS sorting and library preparation</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>2.3 Human serology</td>
<td>121</td>
</tr>
<tr>
<td>2</td>
<td>2.4 Statistical analyses</td>
<td>122</td>
</tr>
<tr>
<td>2</td>
<td>2.5 Network analysis</td>
<td>123</td>
</tr>
<tr>
<td>2</td>
<td>2.6 Ferret model of immune suppression</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>3 Results</td>
<td>126</td>
</tr>
<tr>
<td>3</td>
<td>3.1 Effect of measles on lymphocyte population structure</td>
<td>126</td>
</tr>
<tr>
<td>3</td>
<td>3.2 Changes in B cell repertoire diversity following measles</td>
<td>127</td>
</tr>
<tr>
<td>3</td>
<td>3.3 Dynamics of the naïve B cell compartment following measles virus infection</td>
<td>128</td>
</tr>
<tr>
<td>3</td>
<td>3.4 Dynamics of the memory compartment B cell repertoire following measles virus infection</td>
<td>134</td>
</tr>
<tr>
<td>3</td>
<td>3.5 Changes in isotype composition of the B memory compartment post measles</td>
<td>134</td>
</tr>
<tr>
<td>3</td>
<td>3.6 Convergence of variable gene usage across individuals post measles</td>
<td>135</td>
</tr>
</tbody>
</table>
Table of contents

3.7 Loss of clonally expanded B cell populations as a source of convergence between memory repertoires post measles .................................................. 138

3.8 Clonotype reconstruction post measles .......................................................... 140

3.9 Maintenance of antigen-specific clones in the B cell memory compartment post measles .......................................................... 141

3.10 Immunological amnesia in a ferret model of measles ........................................ 144

4 Discussion ........................................................................................................ 148

V Discussion and future directions ....................................................................... 153

1 Cross-validation of experimental approaches and analysis tools for BCR sequencing ................................................................................................. 154

2 Development of novel methods for phylogenetic inference .................................... 156

3 Deriving predictive patterns of immune function using BCR sequencing .................. 158

3.1 Incomplete characterisation of germline diversity of human IGH loci 159

3.2 Limited models for inferring antigen-specificity from BCR data ........................ 159

4 BCR sequencing for discovery of novel immune phenotypes .................................. 160

4.1 Providing evidence for immune dysfunction ...................................................... 160

4.2 Uncovering novel mechanisms of immune dysfunction ...................................... 162

5 Final remarks ....................................................................................................... 164

References ........................................................................................................... 165

Appendix A Supplementary tables ........................................................................... 221
List of figures

I.1 Genetic composition of immunoglobulin loci. ........................................ 5
I.2 Genetic and amino acid structure of a human antibody. .......................... 8
I.3 Stages of B cell development before and after antigen exposure. ........... 12
I.4 Class-switch recombination for generation of different antibody classes. .. 16
I.5 Workflow of standard steps of bioinformatic analysis of BCR repertoire data. 22
I.6 Methods for correction of technical noise in BCR sequencing data. ........... 24

II.1 Strategies for library preparation for BCR sequencing. ............................ 48
II.2 Schematic of barcode filtering criteria .................................................. 54
II.3 Comparison of sensitivity of amplification methods for capture of diverse
     PBMC and highly clonal CLL repertoires. ........................................... 58
II.4 Hierarchical clustering of VJ gene frequency profiles across samples. ....... 59
II.5 Reproducibility of BCR repertoire capture across amplification methods. .. 60
II.6 Principle component analysis of network parameter of BCR repertoires. .... 62
II.7 Network properties of BCR repertoires across amplification methods. ....... 63
II.8 Barcode profiles across amplification methods. ........................................ 65
II.9 Nucleotide mismatch distribution within barcode groups. ........................ 66
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II.10</td>
<td>Nucleotide mismatch maps per barcode group</td>
<td>67</td>
</tr>
<tr>
<td>II.11</td>
<td>Sequence profiles of incorporated barcodes</td>
<td>69</td>
</tr>
<tr>
<td>III.1</td>
<td>Stages of evolution of antigen-specific B cell responses</td>
<td>75</td>
</tr>
<tr>
<td>III.2</td>
<td>Structure and functions of immunoglobulin isotypes</td>
<td>76</td>
</tr>
<tr>
<td>III.3</td>
<td>Library preparation steps of isotype-resolved BCR sequencing</td>
<td>90</td>
</tr>
<tr>
<td>III.4</td>
<td>Genetic sorting of isotype-specific B cell populations</td>
<td>91</td>
</tr>
<tr>
<td>III.5</td>
<td>Isotype composition of healthy and leukemic B cell repertoires</td>
<td>93</td>
</tr>
<tr>
<td>III.6</td>
<td>Frequency of transitions across isotypes in healthy and leukemic B cell repertoires</td>
<td>94</td>
</tr>
<tr>
<td>III.7</td>
<td>Relationship between SHM and class-switching in the evolution of B cell responses in health and disease</td>
<td>96</td>
</tr>
<tr>
<td>III.8</td>
<td>Deviation of naïve VJ gene profile in class-switched isotypes</td>
<td>97</td>
</tr>
<tr>
<td>III.9</td>
<td>Conditional probability of BCR class switch depending on V gene family and mutation status</td>
<td>99</td>
</tr>
<tr>
<td>III.10</td>
<td>Conditional probability of BCR class switch to different isotypes depending on V gene identity</td>
<td>100</td>
</tr>
<tr>
<td>III.11</td>
<td>Expression of V and J gene families across isotypes</td>
<td>102</td>
</tr>
<tr>
<td>III.12</td>
<td>Hierarchical clustering of isotype-specific VJ gene repertoires per individual</td>
<td>103</td>
</tr>
<tr>
<td>IV.1</td>
<td>Proposed model of 'immunological amnesia'</td>
<td>116</td>
</tr>
<tr>
<td>IV.2</td>
<td>Gating strategy for FACS sorting of B and T cell populations</td>
<td>121</td>
</tr>
<tr>
<td>IV.3</td>
<td>Lymphocyte frequencies across sampling timepoints</td>
<td>127</td>
</tr>
<tr>
<td>IV.4</td>
<td>Correlation of V gene frequencies of B naïve cells across individuals</td>
<td>129</td>
</tr>
</tbody>
</table>
## List of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV.5</td>
<td>Variable gene frequencies in the B naïve compartment across individuals.</td>
<td>130</td>
</tr>
<tr>
<td>IV.6</td>
<td>Correlation of VJ gene frequencies of B naïve cells across timepoints.</td>
<td>132</td>
</tr>
<tr>
<td>IV.7</td>
<td>BCR diversity of naïve repertoires of individuals B-35 and B-42.</td>
<td>133</td>
</tr>
<tr>
<td>IV.8</td>
<td>Changes in isotype composition of B memory cells post measles.</td>
<td>135</td>
</tr>
<tr>
<td>IV.9</td>
<td>Variable gene frequencies in the B memory compartment across individuals.</td>
<td>136</td>
</tr>
<tr>
<td>IV.10</td>
<td>Correlation of V gene frequencies of B memory cells across samples.</td>
<td>137</td>
</tr>
<tr>
<td>IV.11</td>
<td>Fold reduction in V gene frequency post measles.</td>
<td>139</td>
</tr>
<tr>
<td>IV.12</td>
<td>Restructuring of the B memory clonotype composition post measles.</td>
<td>141</td>
</tr>
<tr>
<td>IV.13</td>
<td>Clonal dynamics in the B memory compartment across timepoints.</td>
<td>143</td>
</tr>
<tr>
<td>IV.14</td>
<td>Clonal dynamics in the B memory compartment of individual B35.</td>
<td>144</td>
</tr>
<tr>
<td>IV.15</td>
<td>Effect of a lymphotropic challenge on vaccine-induced serological immunity to influenza virus.</td>
<td>146</td>
</tr>
<tr>
<td>IV.16</td>
<td>Effect of a lymphotropic challenge on the severity of secondary infection.</td>
<td>147</td>
</tr>
</tbody>
</table>
List of tables

II.1 Nucleotide sequences of primers used for BCR amplification . . . . . . . 52
II.2 List of samples and output of read QC steps . . . . . . . . . . . . . . . 57
II.3 Proportion of re-captured BCRs across replicates . . . . . . . . . . . . . 61
II.4 Network parameters of BCR repertoires before/after barcode correction. . 64
III.1 Sample filtering information for isotype-resolved sequencing . . . . . . . 80
III.2 Antibody panels for isotype-specific B cell sorting . . . . . . . . . . . . . 83
III.3 Primer sequences for isotype-resolved BCR sequencing . . . . . . . . . . . 84
IV.1 Sampling strategy for measles cases and control cohorts . . . . . . . . . . 119
A.1 Quality control of BCR sequencing reads . . . . . . . . . . . . . . . . . . 222
Chapter I

Introduction
Introduction

“It is not the most intellectual of the species that survives; nor is the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself.” Despite erroneously attributed to Darwin many times over, this quote from the American Professor of Management Leon C. Megginson, summarises Darwin’s evolutionary theory on the Origin of species and describes the fundamental principles behind the majority of physiological processes operating in the human body.

The human immune system is one of the best examples of a physiological system which acts as a major mediator between our bodies and the world around us, enabling us to recognise and respond to changes in the environment. Heritable mechanisms of defence are present at the level of unicellular organisms and have evolved to a complex innate immune response [1] and a highly organised adaptive immunity in jawless fish and vertebrates. The innate immune system confers protection in a non-specific manner. This is achieved by a constitutive production of generic receptors that recognise conserved molecular patterns on different classes of pathogens and ensure early control of infection [2, 3].

Adaptive immunity, by contrast, provides a targeted response to the encountered antigen by somatic diversification of specific receptor genes. This generates a vast repertoire of cells, able to recognise virtually any encountered antigen [4]. In addition to the specific recognition, adaptive immunity controls the magnitude of immune response and generates immune memory which protect us upon future re-exposure to the same antigen.

The synchronised action of the two branches of the immune system provides the speed and the specificity of host defense required to limit any potentially harmful interactions with the environment. The identification of new classes of cells with shared properties of both systems (innate lymphoid cells) [5] provides an additional level of complexity and demonstrates the high degree of functional crosstalk, essential for the successful response to an antigen.
1 Adaptive immune responses

1.1 Evolution and organisation of the adaptive immune response

The B and T lymphocytes are the major cell types of the adaptive immune system. They originate from a common haemopoietic stem cell progenitor in the bone marrow. T lymphocytes, responsible for cellular immune responses, undergo further maturation in the thymus, while B cells remain in the bone marrow and undergo genetic recombination of their receptor genes to produce a large pool of B cell clones with unique B cell receptor molecules (BCRs) [6]. After development in the primary lymphoid organs, B and T lymphocytes traffic to lymph nodes and the spleen, which serve to capture circulating antigens from lymph and blood, respectively. Immune responses originate in these secondary lymphoid organs and are mediated by signals from the innate immune system provided either directly by circulating antigens or indirectly by antigen-presenting cells (APCs).

T cells recognise peptides (or carbohydrates [7, 8]) processed by APCs and presented on their surface via Major Histocompatibility Complex (MHC) class II molecules. B cells recognise antigens directly or via APCs and become activated with (‘T-dependent’) or without (‘T-independent’) T cell help. B cell activation typically leads to maturation and somatic mutation of the B cell receptor to improve the affinity of antigen binding. Antigen-selected lymphocytes leave the spleen and lymph nodes and migrate to many sites in the body to exert effector functions. This trafficking is regulated by an array of adhesion molecules and chemokine receptors which determine the final homing location of lymphocytes (CCR6⁺ B cells traffic to the skin [9]; integrin α4β7⁺ B cells – to gastrointestinal tract) [10, 11]. For the purposes of this thesis, I will focus on the evolution of B cell responses and consider the T cell compartment where necessary for the explanation of B cell function.
Introduction

1.2 Development of B cell responses

1.2.1 Antigen-independent development in the bone marrow

The bone marrow is the primary site of B cell generation after birth [12, 13]. Commitment to the B cell lineage is dependent on the activity of transcription factors, such as PU.1, IKAROS, E2A, EBF (early B cell factor 1), PAX5 (paired box gene 5) and IRF8 (interferon regulatory factor 8) [14, 15]. Upon activation of this specific transcriptional profile, B cells undergo several stages of development, during which they acquire their antigen specificity, conferred by the genetic composition of their BCR.

1.2.1.1 Genetic basis of B cell diversity

BCR molecules exist in a membrane-bound form or in a secreted form as ‘antibodies’ or immunoglobulins (Igs). Human antibodies are formed by pairing of two identical heavy chains (HCs) with two identical light chains (LCs). Each chain has a variable domain at its N-terminal end responsible for antigen binding and a constant region at its C-terminal end involved in antibody effector functions (only the heavy chain). Each B cell has the genetic potential to simultaneously express two different heavy chains and four or more different light chains. However, as a result of allelic exclusion (the expression of only one functional allele), each cell usually expresses only one HC–LC pair [16]. This type of allelic exclusion differs from other forms of monoallelic gene expression as Ig genes are expressed from both alleles but under normal circumstances, only one of the two alleles is functional. Maintaining a unique specificity of each B cell is considered to be important for robust recognition of antigens and for the generation of a specific antibody response with reduced probability of self-reactivity.

The genetic composition of BCR molecules underlies the diversity of B cell responses and the ability to respond to virtually any encountered antigen. Genes that encode the BCR molecule are located at three primary loci in the human genome: IGH (heavy chain) - 14q32.33, IGK (kappa light chain) -2p11.2, and IGL 22q11.2 [17]. Each of these loci consists
Adaptive immune responses

of sets of variable (V), diversity (D; only for IGH), joining (J) and constant (C) exhibiting allelic and copy number variation across individuals [18–20]. Germline variation is the first source of genetic diversity of the BCR molecule and is also reflected in the inter-individual variation of the expressed naïve antibody repertoires [21] which shows a high degree of heritability [22] and can contribute to differences in antibody function [23](Figure I.1).

Figure I.1: Genetic composition of immunoglobulin loci.
Immunoglobulin loci are present on chromosomes 14 (heavy chain), chromosome 2 (kappa light chain) and chromosome 22 (lambda light chain). The heavy chain undergoes two steps of recombination: between D and J genes and then between V and DJ. Heavy chain locus contains 9 constant region genes which are carried through all recombination steps (only shown in the germline here). The kappa and lambda light chains do not contain D genes and recombine through a single step between V and J genes. Kappa light chain has a single constant region, while lambda has 4.

During B cell development in bone marrow individual V, D and J genes undergo somatic rearrangement at the DNA level. This process is referred to as ‘VDJ recombination’ and was first described by Tonegawa in 1974 [24]. This creates a layer of combinatorial diversity and
leads to production of B cell clones with a unique genetic composition of the BCR molecule. The joining of V, D and J segments is mediated by the enzymes recombination activating gene-1 and -2 (RAG1 and RAG2) [25, 26] and is guided by conserved noncoding DNA sequences adjacent to the recombination sites. These sequences consist of two conserved blocks of seven and nine nucleotides (5’CACAGTG3’ and 5’ACAAAAACC3’) separated by a non-conserved spacer sequence of 12 or 23 nucleotides. The conserved length of the spacer corresponds to one or two turns of the DNA double helix which brings the heptamer and nonamer sequences to the same side of the DNA helix where recombination can be initiated [27]. The heptamer-spacer-nonamer motif is referred to as a recombination signal sequence (RSS) [28]. Deviations from the optimal RSS sequence for certain genes may modulate the efficiency with which particular sites are used for recombination [29].

Recombination only occurs between gene segments located on the same chromosome and it follows the 12/23 rule which ensures that only a gene segment with a 12-base pair (bp) spacer can be joined to one flanked by a 23 bp spacer RSS. Thus, for the heavy chain, a D<sub>H</sub> gene segment can be joined to a J<sub>H</sub> and a V<sub>H</sub> to a D<sub>H</sub>, but direct V<sub>H</sub>-J<sub>H</sub> gene joining is not observed as both V<sub>H</sub> and J<sub>H</sub> gene segments are flanked by 23 bp spacers. Violation of this rule is demonstrated as direct joining of two D<sub>H</sub> gene. This has been proposed as a mechanism responsible for the unusually long hypervariable loops found in some heavy chains [30–32].

This evolution of the BCR molecules in bone marrow involves an additional level of diversity originating from the non-templated addition or removal of nucleotides between recombined gene segments. This joining diversity, originates from the imprecise ligation of the DNA double stranded breaks generated during VDJ recombination. The added nucleotides are known as either P or N nucleotides. P nucleotides generate palindromic sequences at the end of joined gene segments by formation of hairpins that are cleaved at a random position and then filled by complementary nucleotides. N nucleotides are also non-template encoded and are added by terminal deoxynucleotidyl transferase (TdT) enzyme to single-stranded ends of the coding DNA.
Genetic diversity is unevenly distributed along the sequence of the recombined variable region (Figure I.2). Three hypervariable regions are present for each heavy and light chain. The six hypervariable loops in total constitute the antigen-binding site of an antibody and form a specific amino acid complementary surface thus referred to as complementarity determining regions (CDRs), denoted CDR1, CDR2, and CDR3. The CDR3 falls between the V gene segment and the J gene segment, and in the heavy chain it is partially encoded by the D<sub>H</sub> gene segment. The diversity of the CDR3 region is significantly increased by the addition and deletion of nucleotides during the formation of the junctions between gene segments. As a result, the CDR3 is the most hypervariable region of the BCR molecule. The germline, the combinatorial and the junctional diversity of each chain, together with the pairing of the heavy and light chains, constitute the genetic basis of the extreme diversity of the naïve B cell clones.
Introduction

Figure I.2: Genetic and amino acid structure of a human antibody.
Schematic of IgG antibody structure. The heavy and the light chain in the top ‘Nucleotide’ panel reflect the germline V, D (only for heavy chain), J and C region. Non-templated N nucleotides are shown in red. These two chains reflect the 5’ to 3’ genetic composition of an antibody. In the bottom ‘Amino acid’ panel the complementarity-determining regions are indicated as coloured blocks along the heavy and the light chain. Framework regions on both chains are shown in white. The bottom panel reflects the N-terminal to C-terminal amino acid sequence. Dashed lines represent the disulfide bonds between chains and between gene segments. The figure was adapted from Georgiou et. al. [33]
1 Adaptive immune responses

1.2.1.2 Selection and maintenance of the naïve B cell compartment

Recombination and pairing of the heavy and light chain triggers tonic BCR signalling which promotes positive selection of B cell clones expressing IgM+ molecule on their surface. In the bone marrow microenvironment antigen exposure is largely restricted to self-antigens, and thus regulation at this stage is essential to prevent autoreactivity. B cells produced in the bone marrow can undergo apoptosis or several rounds of editing of their recombined BCR receptors as part of the central tolerance checkpoint. Central tolerance refers to the regulatory mechanisms that occur at the early stages of B cell development in the bone marrow, when B cells carry a surface antigen receptor of the IgM class but are not fully mature [34]. Tolerance mechanisms that occur in later developmental stages are referred to as peripheral tolerance and they regulate the survival and activation of B cells after they exit the bone marrow.

Most of the mechanisms of peripheral tolerance are reversible as B cell clones with autoreactive specificities can be activated by exogenous antigens with similar epitopes to self-antigens [35, 36]. As a result, autoreactive clones which are initially negatively selected in the periphery can undergo clonal expansion and lead to autoimmune phenotypes [37, 35]. Therefore, central tolerance has a key role in reducing the frequency of autoreactive cells in the naïve, pre-immune B cell repertoire. If a recombined BCR receptor is autoreactive, editing can lead to exchange of one functional light chain for another to generate a BCR receptor with minimal autoreactive properties [38, 39]. If this process of secondary rearrangement destroys the original light chain gene but does not replace it with another recombined one, B cells move back to the pre-B cell compartment [40]. Positively selected cells enter the transitional B cell stage where they lose the ability to edit the BCR and are highly susceptible to apoptosis [41].

Cells with successfully recombined BCR genes reach the immature naïve stage and exit the bone marrow to complete their development in the spleen, primarily developing into mature follicular B cells. Recirculating CD23+ follicular B cells include IgDhiIgMloCD21int and IgDhiIgMhiCD21int cells [42, 43]. These cells recirculate freely, homing to B cell areas or follicles in secondary lymphoid organs such as the spleen, lymph nodes, and Peyer’s patches.
Introduction

where they interact with antigen. This is signalled by the appearance of IgD in addition to IgM on the cell surface. This entire developmental sequence from early haemopoietic progenitor to and IgM\(^+\) IgD\(^+\) cell with recombined BCR genes occurs in the absence of any contact with exogenous antigen and is called antigen-independent B cell development.

1.2.2 Antigen-driven B cell activation

The second phase of B cell development occurs after encountering an antigen and undergoing activation and is called the antigen-dependent phase. As B cells have to respond to a wide range of potentially pathogenic challenges throughout the whole body, their interaction with antigens is confined to specialised sites to maximise the probability of successful recognition of an antigen. Such sites of B cell interaction with antigen are the lymph nodes, spleen and Payer’s patches, collectively known as secondary lymphoid organs (SLOs). These sites possess a highly organised microarchitecture consisting of an outer antigen-sampling zone, which is an APC-rich area that binds and delivers antigen to the B and T cell zones; the B cell activation zone (B cell follicle); and the T cell activation zone.

This highly conserved microanatomy is necessary for the orchestration of numerous cellular interactions and provides the optimal environment for the initiation of immune responses. Depending on the various contacts and cytokine stimuli received by the activated cell, it will become either a memory cell to be activated once again in the future or it will become a plasma cell (PC) producing large amounts of antibody. Within 24 hours lymphocytes that have been unsuccessful in their search for antigen recirculate throughout the body and between SLOs to dramatically increase the probability of encountering cognate antigen (antigen that can be recognised by direct contact with immune receptor) [44].

1.2.2.1 B cell activation via T-independent and T-dependent antigens

B cell responses are classified as T-dependent (T-D) or T-independent (T-I) based on the requirement for T cell help in antibody production [45, 46]. T-D antigens are usually proteins
that are processed and presented on MHC class II molecules for recognition by cognate helper T cells. T-I antigens can interact with Toll-like receptors [47] or cross-link multiple immunoglobulin receptors and thus deliver a partially activating signal to B cells. In the case of T-I antigens, B memory or plasma cells are generated only with additional signals provided by cytokines or other cell contacts (e.g. with dendritic cells), but not after T cell stimulation. [48].

The vast majority of antibody responses to proteins and glycoproteins involve T cell help for successful B cell activation. B cells require two principal types of signals to become activated. Signal 1 is delivered by cross-linking of the immunoglobulin receptor with a recognised antigen and, after engagement of CD45 and CD19 co-receptors, leads to activation of intracellular signalling pathways. This initial B cell priming renders the antigen-bound B cell capable of receiving signal 2 from T cells [49]. B cells can serve as APCs by expressing processed peptides on MHC class II molecules on their surface. The interaction between a B cell and a CD4⁺ T cell specific for such a peptide provides cognate help and activates the B cell. This initial interaction takes place at the margin between primary follicles and T cell areas in secondary lymphoid tissues.

The activated IgM⁺ B cells enter one of three pathways: to differentiate into early recirculating memory B cells, to become short-lived plasma cells secreting low-affinity antibody without somatic mutation, or to enter a follicle and to establish a germinal centre (GC)[50–52]. The two predominant pathways are extrafollicular plasma cell (PCs) fate and GC reaction (Figure I.3). Extrafollicular short-lived plasma cells secrete IgM antibodies or become IgM⁺ memory B cells[53]. GCs are defined microanatomical clusters within seconary lymphoid organs where B cells undergo proliferation, somatic hypermutation (SHM) and antigen-driven selection based on the sequence of their BCR [54, 55]. The mutation in the BCR can be coupled to changes in the isotype of the expressed BCR via class-switching [56].
Introduction

Figure I.3: Stages of B cell development before and after antigen exposure.

Stages of antigen-independent and antigen-dependent B cell development. Antigen-independent development in bone marrow begins from haemopoietic stem cell progenitor which undergoes D-J gene recombination in heavy chain locus resulting in a ProB cell. This is followed by IGHV-DJ gene recombination and generation of PreB cell. PreB cells recombine their kappa and lambda light chains (LCs) to produce Immature B cell with mature IgM B cell receptor (BCR) which is tested for autoreactivity as part of central tolerance checkpoint. Autoreactive B cells return to PreB stage and undergo receptor editing with exchange of LCs. Non-autoreactive B cells are released in peripheral blood as Mature IgD⁺IgM⁺ B cells and they home to B cell follicles in different secondary lymphoid organs. Follicular B cells are activated upon encounter of an antigen which leads to upregulated BCR expression. Activated B cells either undergo T cell-independent clonal expansion and generation of short-lived plasma cells with low degree of somatic hypermutation (SHM) or form germinal centres. As part of the germinal centre reaction activated B cells undergo clonal expansion and SHM with or without class-switch recombination (CSR) and are selected based on antigen affinity by interaction with follicular dendritic cells (FDCs) and T follicular helper cells (T<sub>FH</sub>). Cells that fail to bind with sufficient affinity undergo apoptosis. Positively selected B cells with high affinity form: 1) B memory cells which express the mutated form of their BCR with improved affinity; 2) Short-lived plasmablasts that secrete high-affinity antibodies identical to the BCR on the surface of B memory cells. Plasmablasts generate long-and short-lived plasma cells that home to bone marrow to establish long-term serological memory.
1 Adaptive immune responses

1.2.3 Germinal centre reaction and generation of B cell diversity

The average affinity of specific antibodies to a given antigen increases dramatically over the course of an immune response [57]. This phenomenon is known as *affinity maturation*, and is the result of a Darwinian process which includes somatic hypermutation (SHM) of immunoglobulin genes followed by selection and clonal expansion of B cells that have acquired mutations that improve their affinity to an antigen. Successive iterations of this cycle occur during germinal centre reaction and lead to generation of highly-specific antibodies able to efficiently recognise and neutralise a pathogen.

The germinal centre reaction was first described by Walther Flemming in 1884 as the site in secondary lymphoid organs where extensive lymphocyte replication takes place [58]. GCs were long believed to be the source of developing lymphocytes but their absence in germ free mice and their responsiveness to antigen dose revealed their role in evolution of antigen-specific responses [59]. Today the GC reaction is known to represent a key event in the adaptive immune response responsible for the production of affinity matured humoral immunity and long-term memory.

GCs contain up to a few thousand B cells and emerge in multiple copies within secondary lymphoid organs around 6-8 days after infection or immunisation [51, 60]. GCs serve as cellular factories in which low-affinity IgM antibodies are converted to high-affinity antibodies of other classes. GCs are polarized into two sectors of roughly the same size, a dark zone of rapidly proliferating B cells which undergo somatic hypermutation and a light zone where B cells undergo antigen-driven selection with the help of follicular dendritic cells (FDCs) and CD4+ follicular T helper cells (T<sub>FH</sub>).

Establishment of the GC was traditionally considered to result from oligoclonal expansion of several founder clones, which are gradually outcompeted by higher affinity ones [61, 62]. Recently developed Rosa<sub>26</sub>Confetti/Confetti mice carrying the Mx1-Cre transgene allowed for *in vivo* monitoring of the early events in GC formation using multiphoton imaging [60]. This model demonstrated that early GCs are usually highly diverse, containing up to hundreds
of distinct B cell clones and a stage of increasing clonality is gradually achieved following “clonal bursts” that lead to rapid expansion and further diversification of cells with improved affinity. This pattern of initial clonal diversity punctuated by sporadic clonal bursts ensures that the polyclonality of the GC response is maintained while high affinity clones can be heavily diversified and exported to effector or memory fates [63, 64].

During GC reaction B cells migrate rapidly from the dark zone to the light zone where they interact with FDCs and T_{FH}. Survival of B cells depends upon their affinity of binding to an antigen presented on FDCs and requires additional stimuli by cognate T_{FH} cells mediated via CD40-CD40L interaction and IL4, IL21 cytokine release. B cells compete for T_{FH} help based upon the amounts of antigen that they are able to bind through their BCRs [2,16]. Cells with high-affinity receptors capture more antigen [17] and are able to present greater amounts of peptide complexed with MHC class II. The high dose of presented peptide favours the productive interactions with a limiting number of T_{FH} cells. The outcome of this competition for access to T cell help is that the ‘helped’ B cells are selected to continue participating in the GC (cyclic re-entry), while the low affinity cells receive inadequate help and undergo cell death. Apoptosis of unselected B cells is important to maintain cell homeostasis and to prevent excessive lymphoproliferation. Recent work by Mayer et al. demonstrates that up to half of all GC B cells dying every 6 hours and programmed cell death is differentially regulated in the light zone and the dark zone. Light-zone B cells die by default if they are not positively selected, whereas dark-zone cells die when their antigen receptors are damaged by activation-induced cytidine deaminase [65]. Survivor B cells migrate back from the light zone to dark zone for repeated rounds of proliferation, mutation, and further selection of B cells with improved affinity to an antigen [66].

1.2.3.1 Somatic hypermutation

The modifications in BCR genes during affinity maturation are mediated by the enzyme activation induced cytidine deaminase (AID) [67]. This enzyme converts deoxycytidine (dC) to deoxyuridine (dU), which resembles deoxythymidine and results in dU:dG mismatches.
1 Adaptive immune responses

These lesions are then repaired by excision (by uracil DNA glycoslyase) or via the activity of non-templated error-prone polymerase (Pol\(\eta\)) which yields point mutations as well as insertions and deletions [68]. This process of mismatch repair or mutations in variable genes constitutes the SHM process underlying the increased affinity to the antigen. The mutation frequency during SHM has been estimated to be around \(10^{-2}\) to \(10^{-4}/\text{bp/cell division}\), or around a million times greater than the rate of random mutations in the rest of the genome [69].

All four bases can be subject to mutation but base targeting is biased in the context of Adenine (A):Thymine (T) base pairs with preferential mutation of A rather than T on coding strands [70]. Mutations are also non-random in terms of the preferred location and sequence context. This leads to mutational hot spots along a target region [71, 72]. A preferred motif for Guanine(G):Cytosine(C) pair mutation occurs within a WRCY consensus (where W = A/T, R = A/G, and Y = C/T) with AGCT being a preferred base combination. The sequences of germline V genes exhibit a biased codon usage such that regions implicated in antigen binding or affinity maturation are intrinsically more mutable than others (i.e. contain more hotspot motifs) [73, 74]. This is likely to reflect an evolutionary adaptation to optimise the targeting of amino acid substitutions during SHM and to achieve efficient antibody maturation.

1.2.3.2 Class-switch recombination

In addition to point mutations, AID-mediated deamination also leads to staggered double-stranded breaks in the intronic regions flanking constant region genes of Ig heavy chain (referred to as “switch regions”). These double-stranded breaks are repaired by non-homologous end joining recombination (NHEJ) which joins the recombined VDJ sequences to a C gene segment downstream from the first C\(\mu\) gene. This process is referred to as “class-switch recombination” (CSR) and results in a change in Ig class (from IgM and/or IgD to Ig\(A_{1/2}\), Ig\(G_{1/2/3/4}\) or IgE) determining the effector functions of secreted antibody (Figure I.4). The excision of the intervening introns before ligation of two DSBs makes the process of CSR
Introduction

unidirectional with consecutive class-switch events allowed only from proximal to distal C genes relative to the switch region [75–77]. The exact mechanisms underlying the class-switch fate of a B cell during GC reaction are incompletely understood but roles of specific cytokine stimulation [78–80] or epigenetic modifications [81, 82] have been described. The process of class-switch recombination results in production of antibodies with different constant region genes, referred to as immunoglobulin classes/isotypes. Antibody isotypes vary in molecular weight, abundance in serum and ability to trigger downstream signalling mediated by engagement of the Fc portion of their constant region genes.

Figure I.4: Class-switch recombination for generation of different antibody classes. Mechanism of class-switch recombination in the heavy chain constant region genes. The AID enzyme recognises switch regions located upstream of each constant region gene (except IGHD) and introduces double-stranded breaks. The breaks are repaired by non-homolougous end-joining recombination and the intervening sequence of the C genes between two targeted switch regions is removed. In the given example this leads to a class-switch from IgM to IgA2 isotype. Switch region is not present before the constant region encoding for IgD isotype and this immunoglobulin class is produced after alternative splicing (alternative RNA transcripts shown in red).
2 Methods to study the evolution of B cell responses

1.2.3.3 Outcomes of germinal centre reaction

B cells that have successfully undergone GC reaction have two possible fates: to become B memory cells or long-lived plasma cells (LLPCs). Multiple in vitro studies proposed that CD40-mediated signalling and/or cytokine signals could control this decision, but there is no consensus on the exact driving one pathway versus the other. High-affinity of bound antigen has been proposed as a possible driver of LLPC fate [83, 84], but these observations were later explained as an effect of increased overall proliferation-not a specific increase in antibody-secreting cells. [85] Mouse in vivo experiment revealed an additional level of complexity of post GC B cell fates where B cell subsets are generated in a sequential order: unswitched B memory cells, switched B memory cells and finally isotype-switched LLPCs homing to the bone marrow [86]. This suggests that single GC reaction can be a source of different subsets of antigen-specific effector B cells thus contributing to the generation of both B cell memory and serological memory to an antigen.

2 Methods to study the evolution of B cell responses

Maintaining the balance between strong immune activation and sufficient tolerance is fundamental for the successful function of the adaptive immune system. The genetics of the BCR molecule underlie the functional plasticity of B cell responses and enable B cells to respond to exogenous antigens while maintaining low self-reactivity. The mutational status of a BCR molecule together with its isotype identity reflect the stage of B cell evolution and can be used as indicators of B cell immunity generated in a response to infection, vaccination or any form of immune-mediated disease.

2.1 Molecular approaches

Understanding the genetic diversity of the BCR repertoires was originally based on molecular strategies that used probes for in situ hybridisation of specific immunoglobulin V genes
Introduction

This technique provided important information about the fundamental principles of generation and maintenance of B cell diversity. Persistence of B cell clones and their change in frequency with age was found to result from selection of BCRs with specific V genes [88, 89]. Generation of BCR diversity in bone marrow was shown to have non-random position-dependent IGHV gene expression, favouring usage of IGHV genes proximal to the D gene [90].

Less laborious techniques, developed later for large-scale analysis of lymphocyte repertoires, were junction length spectratyping, quantitative PCR, and micro-array analysis. A polymerase chain reaction (PCR)-based method of spectratyping of the heavy chain CDR3 region sequences was used to characterise the degree of clonality in BCR repertoires as reflected in the size distribution of the CDR3 region of immunoglobulin genes [91, 92]. Quantitative PCR (qPCR) strategies were developed in parallel to CDR3 spectratyping techniques and gave a more precise evaluation of repertoire diversity with detection of up to 2,000 V-C gene combinations [93]. With the addition of Taqman probes for real-time quantitative PCR, this approach was successfully used for monitoring of B cell malignancies and for detection of minimal residual disease [94, 95].

A scaled-up method for direct measurement of the BCR diversity was later employed in the form of a microarray with random nucleotides used to hybridise to RNA from different lymphocyte populations. The number of sites on the array undergoing successful hybridisation was used as a measure of immune receptor diversity. This method was successfully applied to characterise the dependence of T cell diversity on the circulating pool of B cell clones [96].

While now cheap and relatively easy to perform, these molecular techniques provide limited information about the sequence evolution of BCR genes and thus have minimal utility in understanding specific patterns of SHM and BCR diversification.
2 Methods to study the evolution of B cell responses

2.2 Genetic approaches for characterisation of B cell diversity

2.2.1 Sanger sequencing

Information about the exact nucleotide sequence of Ig V genes was possible with the use of Sanger sequencing. This method for DNA sequencing was first developed by Fred Sanger and enabled the sequencing of fragments of around 500 bp via chain termination reactions [97]. This method made it possible to determine the sequence identity of the recombined heavy and light chains in up to a few hundred B cells per experiment [98–100]. The ability to identify the full sequences of expressed Ig chains was subsequently used to clone, express and functionally characterise the entire Ig molecule from individual cells, thus generating information about antibody specificity. Single-cell cloning and antibody expression also revealed the genetic basis of central tolerance checkpoints and demonstrated for the first time the presence of autoreactive B cells that are subject to negative selection in the bone marrow [101]. The combination of cloning techniques with B cell immortalisation protocols lead to identification of pathogen-neutralising antibodies [102] and similar strategies have been used to isolate and characterise antibodies involved in autoimmunity [103, 104] and in responses to clinically important pathogens (Severe Acute Respiratory Syndrome (SARS) [105], influenza virus [106], human immunodeficiency virus (HIV) virus [107], dengue virus [108], and the malaria parasite *Plasmodium falciparum* [109]).

The use of Sanger sequencing provided important insights into the degree of sequence variation introduced by SHM and pioneered the field of antibody discovery. A major limitation of this technology, however, is the low throughput which enables the profiling of 100-1000 BCR sequences which constitute only a tiny proportion of the total B cell diversity.


Introduction

2.2.2 High-throughput sequencing

High-throughput sequencing (HTS) enables the profiling of millions of unique DNA molecules in a single reaction and thus is the only method available to date that can monitor a large proportion of the genetic diversity of the B cell populations. Several HTS systems have been used for immune repertoire sequencing. These include instruments from IonTorrent [110], Pacific Biosciences [111], Roche/454 Life Sciences [112], and Illumina [113, 114] and most recently 10X Chromium (released in March 2017). Each of these HTS platforms vary in cost, level of throughput and number and type of introduced errors. The Roche/454 platform was initially used due to the greater read length versus earlier releases of Illumina’s sequencing technology. However, this pyrosequencing approach along with the IonTorrent platform show high error rates in homopolymeric regions and a throughput of only up to $5 \times 10^6$ reads per run compared to $>1 \times 10^7$ for Illumina [115, 116]. Pacific Biosciences, with read lengths >1 kb, has been used for sequencing of paired HC/LCs where the amplification protocol uses a linker between the variable regions of the two chains [117]. The continuous advances in nanopore sequencing with read lengths >100kb enable the characterisation of the genetic diversity of the entire naïve immunoglobulin loci and when combined with the extremely portable MinIon machines holds promise for extensive use in clinical and diagnostic contexts[118]. The ability to study the entire heavy chain locus has also allowed for identification of novel mechanisms for generation of BCR diversity by chromosomal integrations into the variable regions [119].

Despite the wide range of available sequencing platforms, Illumina MiSeq technology has been by far the most commonly used due to its cost effectiveness (0.5$/Mbp), low error rate, and read lengths of up to 300 bp enabling the identification of both variable region and BCR isotype. New droplet-based technologies are increasingly employed to derive paired HC/LC data from thousands of B cells which has the potential to revolutionise antibody discovery (Briggs et al, May 2017, BioRxiv). The technical robustness of these technologies and their utility for reliable characterisation of BCR diversity remains to be fully tested.
2 Methods to study the evolution of B cell responses

2.3 Characterisation of B cell responses using immunoglobulin sequencing

High-throughput sequencing was first applied to the study of immunoglobulin diversity of Zebrafish in 2009 [120]. Since then the breadth of applications have expanded rapidly and BCR-repertoire sequencing has addressed fundamental questions in B cell biology and has shown promise for further use in the clinic. Despite the simple methodological concept, the accurate application of BCR-sequencing requires a careful consideration of experimental design, methods for control of technical noise, robust bioinformatics tools for mining of the large output of data and ultimately a testing framework that would associate variation in BCR repertoire with clinically relevant immune phenotypes.

2.3.1 Bioinformatics analysis of BCR repertoire data

The data derived from sequencing of the BCR repertoire can be used to describe a number of aspects of B cell genetic diversity: overall transcriptional profile defined by frequency of expression of V, D and J genes; length distribution and amino acid usage of BCR CDR3 regions as a metric of breadth of possible antigen reactivities; abundance of unique BCR molecules as an estimate of repertoire diversity; mutational status of clonally expanded B cell populations as a metric of the evolution of an antigen-specific response; frequency of known antigen-specific sequences over time and many others depending on the sample source and experimental design. The derivation of accurate inferences, however, requires multiple steps of data manipulation and quality control to remove the biases of technical noise and to detect accurately biological variation. Figure I.5.
Figure I.5: Workflow of standard steps of bioinformatic analysis of BCR repertoire data.
BCR reads are first filtered for sequencing quality based on their Phred score. The high-quality reads are then corrected for PCR errors. Corrected reads are then filtered for productive rearrangements and mapped to reference V, D, J genes for gene assignment. Reads with successfully assigned immunoglobulin genes are used for different downstream analyses (several examples given, but other analyses are possible depending on the study design). This workflow represents standard analysis steps using Illumina sequencing platform and might vary if other sequencing strategies are used.

2.3.1.1 Read quality control

The typical starting point for processing of BCR repertoire data is the generation of Fastq (or Fasta) files for each sample. These contain information about base quality derived from the sequencing platform (typically Illumina) [121]. Individual reads can vary in quality which is quantified by their Phred-score [122] which estimates the probability of error along
2 Methods to study the evolution of B cell responses

a sequence of given length \( (p = 10^{-Q/10}, p – \text{probability of an erroneous base call}, Q – \text{the Phred score}) \). Although read quality usually drops towards the end of the reads, an overall Phred score of >30 is required to ensure that no erroneous base calls are made in 1000 bp of sequence. The stringent quality cut-offs reduce the number of sequences but they are essential to minimise the number of single nucleotide sequencing errors which present a serious burden for BCR repertoire analysis. Reads with sufficient quality are selected for successful primer match based on alignment to the used primer sets. If primers contain isotype-specific information, this can be used for read annotation. The part of the sequence that matches the primer should then be removed because primers can bind with a degree of mismatch to the BCR template and lead to erroneous estimates of SHM. Multiple software tools for development of these early processing steps have been developed with some examples being: pRESTO [123], PANDAseq [124], PEAR [125] and VDJTools [126].

2.3.1.2 Correction of PCR errors

The stringent read quality cut-offs reduce the number of sequencing errors but do not account for errors during library preparation. The burden of technical noise is particularly detrimental in the context BCR sequence data, as SHM and single nucleotide errors introduced by PCR are practically indistinguishable. Multiple approaches for reducing this bias have been explored and include both in silico data correction or choice of a specific library preparation strategy. Some of the most commonly used approaches are explained below. (Figure I.6).
Introduction

Figure I.6: Methods for correction of technical noise in BCR sequencing data.
Correction of PCR and sequencing errors in BCR repertoires can be performed using different approaches. Correction using technical repicates compares the frequency-based ranking of BCR sequences present in both replicates to select a frequency threshold for acceptance of a single sequence. The number shown next to each read represents the frequency of the read as percentage of the total repertoire. Clustering based correction combines BCR reads into groups based on identity of the CDR3 region (CDR3_x, CDR3_y, CDR3_z, represent different CDR3 sequences). Additional criteria for sharing the same V and/or J gene with a certain identity can also be used to define a cluster. The consensus sequence from each cluster is taken as a representative and the the number of reads for the remaining sequences are added to calculate cluster abundance. The number next to each read represents the number of reads supporting each sequence. Sequences of the same colour share the same CDR3 sequence. Correction with unique molecular identifiers (UMIs) is based on incorporation of barcoded primers during reverse transcription that tag each RNA transcript with a unique barcode. Reads are then grouped based on their barcode sequence into barcode bins. A consensus read is created for each bin and taken as a representative. Bins can be removed if the level of identity between reads is below a set threshold. Black ellipses represent PCR errors.
The use of technical replicates from the same biological sample and cross-validation of the obtained sequencing results has been used as a method to establish QC thresholds for erroneous reads [127]. In this approach reads are ranked by abundance and compared across replicate data sets, retaining only those BCR sequences present at a specific abundance threshold in all replicates. When performed on the entire variable region (VDJ), this approach removes around 50% of the total reads [128] so sequencing depth needs to be doubled to capture sufficient degree of BCR diversity. This approach also cannot detect rare sequences, that are not shared across replicates and is prone to the effect of hotspots of PCR mutations which would result in highly abundant clones shared across replicates [129, 130]. The requirement for enough starting material for multiple technical replicates is also a limitation, particularly when clinical samples are analysed.

An alternative strategy for removal or erroneous reads is the *in silico* correction of the data by clustering of highly similar sequences [131]. In this case technical replicates are not required and all derived sequencing reads can be used. Clustering is often performed by grouping together sequences based on their CDR3 or entire VDJ sequence using a distance metric such as the Hamming or Levenshtein distance [132, 133]. Each cluster is then represented by a consensus sequence of all reads with a given cluster definition. The application of this method to BCR sequencing data is complicated by the effect of SHM which can lead to BCR sequences with identical CDR3s but with mutations introduced at different parts of the sequence. Thus, correction based on CDR3 sequence would remove multiple unique variants. An attempt to account for this has been made by MiXCR method for multilayer clustering, but this approach has not been formally compared with other clustering algorithms [134].

A commonly used error-correction strategy is based on unique molecular identifiers (UMIs) introduced during library preparation [135, 136]. This approach has the added benefit of correcting not only single nucleotide errors, but also biases in the amplification of individual RNA molecules that can lead to erroneous estimation of clonality. The basic principle of this approach is the introduction of degenerate barcodes of random nucleotides
Introduction
to gene-specific primers used during reverse transcription [137] or during template-switching [138] resulting in tagging of each cDNA molecule with a unique barcode. The size of the UMIs is selected to ensure that the barcode diversity exceeds substantially the expected diversity of the RNA/DNA pool in the sample to avoid tagging of more than one nucleic acid with the same barcode. On this basis, sequencing reads are grouped based on their UMI identity and all reads with identical UMIs should result from the amplification of a single RNA transcript and should be identical. Any reads with identical UMIs but with mutations away from the consensus sequence, can be corrected or removed. Since each molecule is uniquely tagged, this method provides additional qualitative information about transcript abundances which aids the correction of any overamplification events.

A possible disadvantage of this method is the increased amplicon size as a result of the added barcode sequences which makes the end PCR amplicon difficult to sequence with 2×300bp Illumina reads. This is of particular concern in the context of template-switch amplification strategies where the amplicon contains the entire 5’UTR sequence and thus reaches around 500bp even without addition of UMIs. Additional consideration important for the use of this method is the possible errors in the barcode sequence which can result from extended amplification cycles required if the input material is too low. This can be addressed by performing a barcode filtering step that identifies “barcode offspring” resulting from a PCR error in the UMI sequence [139]. Alternatively, an additional UMI can be added during PCR amplification resulting of double barcoding of each starting nucleic acid (‘molecular amplification fingerprinting’) [140] which enables the normalisation of ratios of forward and reverse UIDs and barcode correction. Again, the addition of an extra barcode increases the size of the amplicon and reduces the efficiency of read joining if Illumina Miseq protocols are used for sequencing. Available tools that perform error correction of BCR data based on UMIs are MiGEC [138] and pRESTO [123] and VDJtools [126].
2 Methods to study the evolution of B cell responses

2.3.1.3 Assignment of V(D)J genes

Major utility of BCR sequencing is the ability to identify SHM events and thus infer the clonal evolution of B cell populations. The detection of SHM events requires initial identification of the closest germline V(D)J sequence for each BCR read. This is performed by alignment of each read to a database of known VDJ sequences and selection of the closest one based on alignment score. Multiple methods for accurate assignment of V, D, J genes have been developed to date and they vary in the statistical framework used to calculate probability of each gene segment, in the required computing time and in the certainty of D gene assignment [141–144]. The D gene presents the greatest challenge for assignment as it falls in the CDR3 region and is subjected to non-templated addition of nucleotides after joining to J and then to V gene. Therefore, the diversity associated with the D gene in the immunoglobulin heavy chain is commonly ignored in BCR sequencing analysis.

Due to the presence of mutations in the sequenced BCR repertoires and the high similarity between V gene subfamilies, very often more than one V gene assignment can be equally likely complicating the assignment of a single gene. A possible strategy to overcome this is to ignore the positions in the BCR sequence that differ between the two possible V gene segments. In practice, however, usually only one of the V genes is selected arbitrarily or reads with ambiguous assignments are ignored. The most extensive database for identification of VDJ germline alleles is IMGT [145] which contains annotated Ig sequences from human and other mammalian species, as well as fish. Importantly, the accurate estimation of SHM in BCR repertoire data is dependent on the breadth of available alleles in the germline database. If a given allele has not been previously reported, it will be assigned as a somatically mutated sequence.

Novel V gene alleles are continuously identified for humans and other species highlighting the likely underrepresentation of the shuman IGH alleles in the available database [146–149]. Therefore, the development of methods for identification of novel alleles and for determination of Ig gene haplotypes is an active area of research [150, 151]. Direct derivation of alleles from whole genome sequencing data has not been possible so far due to the highly
repetitive nature of the Ig loci which complicates their alignment. Long-read sequencing platforms could be used in the future to provide better capture of the germline variation in this region.

2.3.1.4 Identification of B cell clones

BCR sequencing can reconstruct the evolutionary history of B cells by inferring clonal expansion events. This requires initial grouping of BCR sequences into expected clonal lineages (‘clonotypes’). Due to acquired SHMs, members of the same clonal lineage do not share identical BCR sequences and thus the inference of clonal evolution of B cells is not trivial [152]. Moreover, sampling of the peripheral blood captures a population of cells which are derived from multiple germinal centre reactions in response to different or identical antigens with variable level of convergence. This further complicates the deconvolution of clonal history. The most basic approach used to date is based on conserved definition of a clone encompassing BCR sequences with identical V, J gene assignment and nucleotide length across the CDR3 region. The clonal relatedness between these sequences is determined on the basis of a defined threshold of sequence similarity of their CDR3 regions (usually >90%) [153].

Non-conservative distance measures that account for the intrinsic biases of SHM have also been developed. Sequences that differ at a SHM hotspot positions are considered more similar than those that are separated by differences in positions unlikely to be targeted by SHM [154]. Sequences in defined clonal lineages can be further compared using hierarchical clustering or used as an input for generation of lineage trees which represent evolutionary relationships between BCR sequences based on maximum likelihood or maximum parsimony methods [155]. Both of these methods make assumptions about the nature of sequence diversity that do not apply to B cell evolution (e.g. long evolutionary timescales, independent evolution of each nucleotide). A recently developed time-resolved simulation framework demonstrates clear limitations of current phylogenetic methods in accurately predicting BCR
Methods to study the evolution of B cell responses

evolution [156]. Therefore, development of improved phylogenetic frameworks for reliable reconstruction of the B cell lineages is still an active area of research [157, 158].

To derive a more meaningful model of ancestral relationships, lineage trees are often combined with information about isotype [153] to infer affinity maturation, class-switching and cell trafficking across immune compartments [153, 159, 160]. An important consideration in estimating diversity of BCR repertoires is the assumption that the observed clonal expansions reflect the number of B cells derived from a single common ancestor. Yet, this relationship is difficult to infer from bulk cell data, because B cell subsets can express different levels of BCR molecules depending on their activation status [161]. As a consequence, high frequency of the same molecule can result from multiple clonally expanded cells or from multiple sequenced transcripts from a single antigen-secreting cell. Thus, sorting of B cell population based on their activation status prior to sequencing could provide a better assessment of the expected ratio of RNA transcripts per cell. The optimal approach to overcome this bias is the use of single cells which would allow for estimation of the burden of errors during transcription to the inferred BCR diversity [162].

2.3.1.5 Calculation of SHM and BCR selection

The degree of SHM of B cell clones is an important metric of their evolution and exposure to an antigen. Calculation of SHM from BCR repertoire data is usually performed based on the number of single nucleotide mutations from the closest germline V gene [145]. SHM is highly dependent on sequence context and targets different regions of the BCR sequence with different probability, making the estimation of overall SHM rate along the entire V gene an incomplete metric of BCR evolution [163, 164]. Different mutability models have been proposed to give distinct weight of mutations occurring in sequence motifs targeted for mutation by AID (e.g. WRCY) compared to known coldspot motifs (e.g. SYC) and attempt to account for mutations that can result from antigen-driven selection [165, 164]. Considering the expected distribution of mutations across the BCR sequence (high mutation rate in the CDR regions and low in framework regions), the observed SHM pattern can be
used to calculate selection by virtue of non-synonymous substitutions [166]. Higher than expected rate of non-synonymous mutations in AID hotspots is considered a hallmark of antigen-driven evolution. Yet, the relationship between BCR mutation level and affinity to an antigen is not always clear. It likely depends on the biochemical properties of the recognised epitope, so direct detection of potentially high-affinity antibodies based on high SHM still requires empirical validation.

2.3.1.6 Estimation of diversity

The quantification of degree of clonal expansion as well as repertoire diversity are important aspects of BCR sequencing analysis as they can reflect the successful generation of antigen-specific responses. The majority of diversity metrics used in immune repertoire sequencing are derived from ecology with the most commonly used ones being species richness, the Shannon entropy, and the Gini-Simpson indices. Each metric captures different aspects of the diversity depending on the size and abundance distribution of the B cell population. This may lead to contradicting results depending on the applied metric [167]. To overcome this, a set of metrics (rather than as single one) as a total diversity profile is increasingly used [154, 168]. Diversity measures are affected by the number of sampled B cells. Therefore, comparing the diversity between two samples often requires subsampling of the larger repertoire [137], or rarefaction analysis [169]. Diversity calculations of BCR repertoires are useful for comparative analysis between individuals, across timepoints or to estimate relative changes in diversity after perturbation.

2.3.1.7 Identifying HC:LC pairs

Identification of HC:LC pairing is important for understanding antigen-specific B cell responses and for antibody discovery. The direct derivation of this pairing information from bulk B cell populations has been challenging and the majority of sequencing studies to date focus on estimation of B cell diversity by virtue of the sequence evolution in the heavy
2 Methods to study the evolution of B cell responses

Chain. Two possible approaches have been demonstrated so far: 1) selection of pairs based on similar rank order of their frequencies; 2) phylogenetic inference of pairs based on tree topology. These methods have lead to accurate identification of antigen-specific HC:LC pairs and for discovery of novel anti-HIV antibodies [170]; [171].

Single-cell sequencing has been increasingly used as a preferred method for generation of paired Ig sequences as it provides higher certainty of correct chain combinations. Typically, single B cells are sorted into single wells and amplification of paired Ig chains is performed using primers with well and/or column specific barcodes that enable the multiplexing of cells during sequencing. The maximum number of single cells processed with this approach are $4.6 \times 10^4$ with around 60% efficiency of derived antibody pairs [172, 173]. With this approach one can combine the cell surface marker information for each cell derived from the sorting with the sequence of paired chains which could provide important insights in the activation stage of B cells.

Droplet emulsion based technologies enable up to $10^7$ cells to be processed with reduced reaction volumes and without the need for single-cell sorting. In this case the reverse transcription (RT) and the PCR steps of library preparation are entirely performed in a droplet containing a single cell that is lysed by heat-shock to allow priming [174]. To overcome the possible inhibition of the RT enzyme during the heat-shock step, a separate lysis reaction can be performed either as a pre-processing a microtiter chip (60,000 wells/plate) followed by droplet encapsulation of each bead [111] or by two-step droplet encapsulation in which transcripts from lysed cells are recovered by magnetic separation [175].

A growing area of research is the combined derivation of the whole transcriptome of single cells together with paired HC:LC sequences. This provides opportunities to derive antigen specificity and transcriptional signatures of individual cells. Immune receptor data can be derived from total transcriptome of single T cells using TraCeR software [176], but the same approach is still to be demonstrated for B cells.
2.4 Applications of BCR repertoire sequencing

2.4.1 Application in understanding the basic biology of the immune system

Since the first application of BCR sequencing for characterisation of the immunoglobulin repertoire of zebrafish, this method has been successfully used to study of immune receptor genes of multiple species [177–179]. Sequencing of Ig genes has characterised a high degree of allelic diversity of the human naïve repertoire [21, 151] and substantial skewing in the frequencies of individual VJ gene combinations in naïve cells that are reflected in the antigen-experienced repertoire [180, 181, 32]. The frequency of V,D and J genes in the naïve repertoire has been shown to be identical in homozygotic twins, suggesting genetic heritability [22]. This role of genotype, however, is not observed in the context of antigen-experienced B cells [153]. This proposed a model of genetic predetermination that affects the pre-immune repertoire, but is outcompeted by antigen-driven selection after B cell activation [182]. Analysis of the BCR diversity of different tissues has led to important insights in mucosal immunity and to the identification of the intestinal lamina propria as an additional site of B cell development (demonstrated in mice [183]). Comparison of BCR diversity of different B cell subsets aided the important distinction of evolutionary pathways leading to the development of B cell memory [184].

2.4.2 Immune repertoire sequencing in infectious diseases

Immune receptor sequencing has been extensively applied in the field of infectious diseases and has shown increasing potential for future use in clinical contexts for identification of immune signatures as diagnostic biomarkers, for design of therapeutic antibodies, and for optimisation of vaccine design.
2 Methods to study the evolution of B cell responses

2.4.2.1 Identification of pathogen – specific antibodies

Knowing the exact antigenic stimulation in the context of infection improves the power of immune repertoire sequencing for monitoring the evolution of antigen-specific B cell responses. B cell diversity has been shown to decrease early after exposure to HIV, dengue virus and herpesviruses [185–187]. This is consistent with the expected focusing of B cell responses and clonal expansions to the specific pathogen. The majority of studies using BCR sequencing to follow B cell responses to natural infection focus on identification of a small number of antigen-specific antibodies and characterisation of their neutralisation properties.

Antigen-specific antibodies can be isolated from peripheral blood via several mechanisms: 1) via hybridoma cell lines [188, 106]; 2) via immortalisation of B cells derived from infected subjects (without previous antigen-specific cell sorting [108, 102, 189] 3) via selection of antigen-specific B cells [190]); via single-cell sorting of plasmablasts and PCR amplification of paired HC/LCs [191].

Although lymphocytes isolated from peripheral blood are the most commonly used sample type in human studies, antigen-specific antibody sequences can also be derived from serum. In this approach, a fluorescently labelled probe of the antigen of interest is used for capture of antigen-specific antibodies in serum of infected individual [192, 193]. The sequence of antigen-specific antibodies is then interrogated using shotgun liquid chromatography-tandem mass spectrometry (LC-MS/MS) to derive the CDR3 sequences of antigen specific antibodies. This information is then used in combination with total BCR repertoire data from the same individual which enables the mapping of the serum-derived CDR3 regions to specific antibody sequences and thus following the clonal evolution in response to infection. This approach provides an indirect way to investigate the function of the plasma cell compartment which is otherwise difficult to study in humans because bone marrow samples are rarely available. When combined with BCR sequencing of B cells in peripheral blood, this approach enables the complete profiling of both cellular and serological immunity generated in response to an antigen and provides insights into the interplay between cellular compartments for maintenance of long-term memory after pathogen exposure [194].
Introduction

When applied to convalescent sera from people recovering from disease, this method has the potential to identify antibodies involved in resolution of disease that are good candidates for future development in a therapeutic context.

Direct prediction of antigen specificity from BCR sequence data alone has not yet been accomplished. However, a possible approach is the use of already characterised antigen-specific antibodies in order to identify clones with similar amino acid composition in the total sequenced BCR repertoire. This strategy has been used in the context of influenza infection where sequences with high identity to previously characterised neutralising antibodies have been identified and later demonstrated to have neutralising activity [195]. In this case the power to detect such sequences is dependent on the availability of convergent clones shared between individuals.

2.4.2.2 Understanding immune response to vaccination

In addition to the knowledge of the specific antigen, vaccination enables selection of well-defined time points for sampling to increase the probability of detection of the vaccine-specific response. Thus, studies with pre-defined vaccination schedules are an optimal experimental design for application of immune repertoire sequencing [116]. Similar to studies on natural infection, cloning of paired HC/LC sequences has also been used to identify pathogen-specific immunity after vaccination. This has been demonstrated for trivalent inactivated influenza vaccine (TIV) [196], tetanus toxoid (TT) vaccine [197], Streptococcus pneumoniae vaccine [198].

In the context of influenza, BCR repertoire sequencing has been used to identify of recall responses and for demonstration of the phenomenon of original antigenic sin [196, 136]. In a study of B cell responses to Hepatitis B vaccination the peak antigen-specific response was observed seven days post infection but vaccine-induced immune responses showed highest degree of convergence across individuals at 14 and 21 days post vaccination. This suggests that later sampling timepoints might be required for detection of common immune features across individuals [199]. Characterisation of responses to varicella zoster vaccination in
monoyzygotic twins showed that genotype contributes to general immune features (gene segment usage, junctional features, and mutation rates) but has little influence on the antigen-specific acute responses which are largely private, even in identical twins [200].

### 2.4.3 Understanding the mechanisms underlying immune dysfunction

Immunodeficiencies or autoimmunity are forms of immune dysfunction that are increasing in prevalence but the immunological mechanisms underlying their development remain largely unknown. BCR analysis of PBMCs from patients with common variable immunodeficiency (CVID) identified decreased diversity of the naïve B cell pool and low level of somatic hypermutation as possible mechanism of the observed disease phenotype despite normal B cell counts [201]. Characterisation of BCR repertoires of cerebrospinal fluid of patients with multiple sclerosis showed strong evidence of B cell activation, suggesting antigenic stimulation and a likely antibody-mediated response involved in disease-associated pathology [202]. Similarly, in idiopathic IgG4-associated cholangitis, BCR analysis revealed the high levels of IgG4 in serum are caused by expansions of IgG in peripheral blood which were resolved after corticosteroid treatment. In this context, the emergence of clonal expansions can be used as an indication of a relapsing disease or resistance to treatment [203].

### 2.4.4 Early diagnosis and monitoring of B cell malignancies

B cell malignancies (leukemias, lymphomas and multiple myeloma) arise at different stages of B cell development and thus the BCRs associated with a malignant expansion can be used as a biomarker to track the dynamics of the cancer population. This is best applied in the context of B cell chronic lymphocytic leukemias (B-CLLs) which have highly clonal and V gene restricted B cell populations [204]. The clonally expanded B cells can be tracked in peripheral blood following treatment to detect minimal residual disease [205] without the need to design of patient-specific PCR assay for detection of clone size post treatment. The utility of such approaches has also been demonstrated in pediatric patients with B
Introduction

cell acute lymphoblastic leukemia [206] (B-ALL) and in non-Hodgkins lymphoma [207]. Importantly, CLL and other hematologic malignancies, exhibit shared heavy chain CDR3s across individuals. The detection of such malignancy-associated clonotypes, present at a low frequency in people without clinical symptoms can act as an useful tool for early diagnosis.

2.4.5 Derivation of immune signatures as potential diagnostic markers

BCR sequencing has the potential to identify immune signatures that can act as biomarkers for diagnosis and monitoring of infectious disease. This application requires shared patterns of immune response between individuals. Such stereotyped features can be VJ gene expression frequencies, common CDR3 sequences, or CDR3 spectratypes. Common patterns of response to infection have been shown in dengue virus infection where the acute onset of disease is characterised with stereotyped CDR3 sequences that are not found in healthy individuals or in those who have recovered successfully after infection [186]. The basis of this repertoire convergence and its potential use for the development of better vaccine candidates remains to be explored.

The greatest hurdle for identification of shared immune signatures across individuals is the high heterogeneity in specific BCR sequences responding to identical antigens even in the context of vaccination. The minimal number of public clones makes biomarker design very challenging. However, it is also likely that BCR sequencing has not yet been applied to a large enough cohort of individuals to detect shared clones. With the improvements in the performance and standartisation of the technology, its use for routine immune screening in clinical setting might provide an opportunity to generate more sequence patterns to be used as biomarkers.
3 Aims and outline of the thesis

The focus of this dissertation is to use immune repertoire sequencing for characterisation of the B cell responses in the context of infectious disease. Specifically, my aim is to follow the changes in diversity of B cell populations in response to a lymphotropic infection with measles virus and to describe how these changes can contribute to the observed phenotype of prolonged immune suppression after recovery from the measles.

In Chapter II I compared strategies for the correction of technical noise based on primer barcoding. I evaluated the introduced method-specific biases and their impact on the ability to make accurate biological inferences using BCR sequencing. As part of this work, I developed and optimised an amplification strategy for reliable characterisation of the genetic complexity of BCR repertoires from both clonal and highly diverse B cell populations. I then demonstrated the utility of this strategy for capturing the diversity of BCR repertoires in a highly sensitive and reproducible manner with minimal amplification biases.

In Chapter III I built on the optimised strategy for barcoded library preparation by introducing isotype-specific primers which enabled detection of all immunoglobulin isotypes in a single reaction. Using this strategy, I characterised the evolution of antibody responses in healthy repertoires and their aberration in the context of B cell malignancy. With the use of isotype-resolved BCR sequencing I followed simultaneously the processes of somatic hypermutation and class-switch recombination during antigen-driven B cell evolution. When applied to diverse B cell repertoires, this strategy uncovered an important relationship between the genetic background of a BCR and the probability of a B cell to class-switch to different isotypes. This work demonstrated the presence of a genetic substructure underlying the diversity of antigen-experienced B cells and highlighted the importance of monitoring both the variable and the constant region genes for characterisation of B cell evolution.

Having identified a BCR sequencing strategy that reliably captures the variable gene diversity and the isotype composition of B cell repertoires, in Chapter IV I used this approach to investigate the dynamics of B cell populations after infection with measles virus. I
investigated the immunological mechanisms underlying the phenomenon of prolonged immune suppression and increased susceptibility to secondary infections after measles. Using isotype-resolved BCR sequencing I followed the changes in diversity of the B naïve and the B memory populations after measles-associated lymphopenia in unvaccinated individuals. I compared the observed patterns of immune restructuring in measles to the changes in B cell diversity in vaccinated and uninfected controls. This study design lead to the identification of measles-specific effects on both the naïve and memory compartments and uncovered two possible immunological mechanisms, likely underlying the observed immune dysfunction following measles. In this chapter I demonstrated the utility of BCR sequencing to uncover aspects of B cell diversity in response to infection that have not been possible to address previously with traditional molecular assays.

In Chapter V I provide a summary of my findings and discuss their importance in the context of the most recent developments and outstanding challenges in the field of immune repertoire sequencing.
4 List of Publications

1. “Isotype-resolved BCR sequencing reveals variable gene effects on B cell class-switch fate”

2. “Optimisation of ex vivo Memory B cell Expansion/Differentiation for Interrogation of Rare Peripheral Memory B Cell Subset Responses”
   Luke Muir, Paul F. McKay, Velislava N. Petrova, Oleksiy V. Klymenko, Sven Kraitochvil, Christopher L. Pinder, Paul Kellam and Robin J. Shattock (under review)

3. “Re-setting of immune memory following measles”

Chapter II

Methods for correction of technical noise in immune repertoire sequencing data
1 Introduction

1.1 High-throughput sequencing of lymphocyte repertoires

The extreme diversity of the human B cell receptor (BCR) and T cell receptor (TCR) repertoires requires a high-throughput approach for its characterisation and for better understanding of the role of the complex genetic architecture in lymphocyte function and antigen specificity. High-throughput receptor gene sequencing (sometimes referred to as “Rep-Seq” [208]) has been increasingly applied to the study of adaptive immune repertoires. Various methods of immune receptor profiling have been successfully used for the characterisation of inter-individual variation in B cell diversity of healthy repertoires [21, 209], for understanding of adaptive immune responses in natural infection [210, 211] and in vaccination [212, 213], for monitoring of cancer progression and response to therapy [214–216], and for improving the discovery of therapeutic antibodies [217, 218]. The continuous increase in the number of available methodologies has opened an array of possible clinical applications: for prediction of transplantation outcomes [219], for monitoring immune reconstitution after antiretroviral therapy [220], identification of disease etiology in multiple sclerosis [221]; for early identification of relapse in B cell malignancies [222]

1.2 Methods for immune repertoire sequencing

Sequence profiling of immune receptor repertoires can be applied to any human or animal tissue containing B and/or T lymphocytes. Reliable characterisation of the immune receptor diversity, however, requires a careful consideration of the cell composition of the investigated immune compartment, the experimental design used for sample collection and the technical biases introduced during library preparation and immune receptor sequencing. Very similar practical considerations apply for both B and T cell receptor sequencing but for the purposes of this chapter, I will focus on the characterisation of BCR repertoires.
1.2.1 Technical considerations for performing BCR sequencing

Despite the breadth of generated data from BCR sequencing, its biological relevance is dependent on the initial experimental set up and the in-depth understanding of the studied B cell compartment. Sorting of B cell populations based on surface markers provides prior knowledge of expected status of antigen experience, secretory or memory phenotype which can aid the interpretation of BCR repertoire data. Due to the continuous exposure to antigens from the environment, the use of longitudinal data and control cohorts is essential for detection of immunological patterns associated with specific perturbation (e.g. infection, vaccination, treatment). The accuracy of biological inferences is also dependent on the optimal biological and technical sampling of the B cell compartment.

Peripheral blood is the most commonly sampled immune compartment in humans largely due to the limited access to secondary lymphoid organs. The contribution of peripheral blood B cells to the total B cell in an individual \((\approx 1.5 \times 10^{11})\) is estimated to be only 2% with the majority of the B cells present in the bone marrow \((\approx 17\%)\) or in secondary lymphoid organs (28% in lymph nodes, 23% in the spleen and on mucosal surfaces) [223]. Thus, in the absence of a specific immune perturbation which would require B cell recruitment to peripheral blood, the circulating B cell pool likely provides a limited representation of the total B cell diversity of an individual. By contrast, in infection or immune-mediated diseases, where immune responses in the blood can play a major role, B cell diversity can be represented much more reliably by sampling the peripheral blood. To this end, the use of BCR repertoire sequencing for characterisation of B cell immune responses in health and disease requires careful experimental design and sampling of the immune compartments most relevant to the investigated immune phenotype.

Technological limitations of sampling the B cell compartment affect the detection of clones shared across individuals (shown for T cells) [224, 225]. Due to the difference in the size of the lymphocyte population involved in a specific response, there are no set standards for sufficient biological sampling. However, depending on the studied B cell population, the sequencing depth has to be defined in a way to provide more sequence reads than expected B
Methods for correction of technical noise in immune repertoire sequencing data

cell clones, accounting for the fact that low frequency clones will always require greater depth for detection. Knowing the clonal composition of a B cell population is often challenging, especially when a specific antigen and the degree of clonal expansion is to be determined by BCR sequencing itself. Therefore, the only general guideline that can be used is the prior knowledge of expected clonal evolution in antigen experienced versus naïve lymphocyte populations. B memory cells and plasmablasts have been generated after clonal expansion to an antigen and thus have clone-to-cell ratio lower than one, requiring less sequencing depth to cover clonal diversity [226, 227]. B cells in the naïve compartment are not expected to have undergone expansion, so each clone will be represented by a single cell [228] and would require higher sequencing depth for sufficient diversity capture.

Importantly, the biological reasons for undersampling can only be assessed if sufficient technical sampling is achieved [127]. Therefore, any developed experimental strategy for library preparation and sequencing of BCR repertoires would require prior testing over several biological [136] and technical replicates [229] to ensure minimal technical biases.

1.2.2 Sample preparation and selection of input nucleic acid

Characterisation of BCR diversity is based on amplification and sequencing of immunoglobulin genes from genomic DNA (gDNA) or RNA of bulk or single B cell samples. B cells with a specific phenotype and activation status can be selected prior to sequencing based on surface marker expression which enables stratification of naïve, memory, antigen-experienced or tissue resident cells [230]. In absence of such prior selection, B cell activation status can be inferred from the degree of somatic mutation of the sequenced BCR genes. The choice of gDNA or RNA as an input nucleic acid is dependent on the specific study and biological question being addressed. gDNA captures both productive and non-productive recombinations of immunoglobulin genes and thus it is not directly representative of the expressed antibody repertoire. However, the non-productive out-of-frame genes which were silenced by allelic exclusion [231] can be used to infer neutral mutational processes (as opposed to antigen-driven selection) during B cell affinity maturation [232]. Sequencing
the immunoglobulin mRNA enables the detection of the expressed BCR genes and thus for profiling of both receptor and antibody repertoires with estimation of relative abundance of individual Variable (V), Diversity (D) and Joining (J) genes. Naïve and antigen-experienced B cells differ substantially in the rate of expression of immunoglobulin genes [233]. This complicates the estimation of clonal frequencies because it is impossible to distinguish between multiple clonally expanded cells expressing the same BCR, and a single cell with highly abundant BCR transcripts. Therefore, additional cell phenotyping, commonly based on surface marker expression, is often used to stratify cell populations before immune repertoire sequencing.

1.2.3 Library preparation strategies for sequencing of immune repertoires

The majority of commonly used approaches for capturing of immune receptor repertoires are based on targeted amplification of immunoglobulin heavy and light chains with variations in priming strategies and library preparation protocols both for bulk and for single-cell sequencing. If RNA is used as the input nucleic acid, library preparation starts with a reverse transcription (RT) step priming either with non-specific poly-dT primers or with a J gene specific primer (for human IGHJ, a single J primer can capture all 6 functional J genes). The resulting cDNA is then subject to PCR amplification. The most commonly employed strategies for 5’ priming of BCR cDNA and consecutive PCR amplification are summarised below:

1. Multiplex priming of variable genes

The human BCR repertoire consists of ≈55 functional V genes [234] which presents a challenge for capture of the complete variable gene diversity in a single amplification reaction and requires the use of multiplex primer sets. To reduce the number of used V-gene specific primers, most commonly employed strategies for BCR priming target the conserved framework regions (FR1, 2, 3) which enables capture of all V gene families with 6 FR-specific primers [136]. Due to the organisation of the FRs along the V gene, the choice of FR1, 2 or 3 as a priming location affects the size of the
1. Methods for correction of technical noise in immune repertoire sequencing data

resulting PCR amplicon and thus it is usually selected based on the read length of the sequencing platform used downstream.

2. Priming upstream of variable genes

An alternative approach to the multiplex V gene priming strategies is the use of primers before the start of the V gene ORF – by priming of the leader sequence or by a 5’ Rapid Amplification of cDNA Ends (5’RACE) template-switch reaction which primes the end of the cDNA by introduction of a switch oligonucleotide. Both of these approaches enable sequencing of the complete variable gene and reduce any possible primer annealing biases associated with the use of multiplex primers. The capture of the full-length of the IGHV/IGLV genes is important for downstream antibody cloning and therefore, these amplification approaches are preferable for paired heavy and light chain sequencing from single cells. An important consideration for 5’RACE amplification is that the annealing of a switch oligonucleotide at the 5’end of the cDNA molecule leads to amplification of the 5’UTR which increases the size of the PCR amplicon and thus requires reads longer than 250bps for capture of the entire variable region.

1.3 The burden of technical noise in immune repertoire sequencing

Despite the variation across strategies for repertoire sequencing, the majority of available methods are reliant on initial PCR amplification of receptor sequences. This enables identification of receptor genes (or their transcripts) from rare cell populations or single cells [235]. Estimated error rates of most commonly used PCR enzymes are 1 in $10^3$ to 1 in $10^5$ for reverse transcriptases [236] and around 1 in $4.4 	imes 10^7$ nucleotides for polymerising enzymes (as reported by Finnzymes/Thermo Scientific). These rates are also influenced by reaction conditions, GC content and secondary structure of the RNA template. Errors occurring in the reverse transcription and early in the amplification step are expanded exponentially. This means that around 1% of PCR products of 500bp size after 36 amplification cycles

1https://www.thermofisher.com
will have at least 1 PCR error. This is further magnified by the library preparation and the sequencing platform-specific error rate. While this might pose a minor error burden for some sequencing applications, it significantly complicates the BCR repertoire analysis as it makes the PCR errors and the immunoglobulin diversity indistinguishable. The diversity generated during somatic hypermutation results in a multitude of almost identical sequences with few polymorphisms and present at different frequencies. As a result, minor unique clones could be erroneously defined as variants of a major clone, making the accurate analysis of hypermutated immunoglobulin genes extremely challenging [237–239].

Current efforts to tackle this technical limitation are focused on the development of methodologies for PCR error correction using unique molecular identifiers (UMIs) added to the immunoglobulin-specific primers during reverse transcription and/or the PCR amplification step. This enables “tracking” of individual BCR molecules and identification of errors that have occurred during the amplification process. The error-correction is achieved by grouping the BCR reads according to their barcode based on the assumption that sequences with the same barcode should originate from the same starting template [210, 240].

1.3.1 Correction of technical biases using unique molecular identifiers

Unique molecular identifiers (barcodes) have been previously used in several research contexts: in pathogen sequencing for accurate identification of minority sequence variants and for monitoring drug-resistance mutations [212, 241]; in single-cell transcriptomics and epigenomics, for quantifying gene expression [213, 215, 214, 242] and in immune repertoire sequencing for estimating B cell receptor (BCR) diversity and PCR error correction [210, 136]. DNA barcoding relies on the incorporation of a unique molecular identifier (barcode) to a single target nucleic acid, which is subsequently amplified during PCR cycles whilst retaining the barcode. Stochastic single-nucleotide errors occurring during amplification or sequencing can then be detected as differences from the consensus sequence of each barcode-group and corrected. Barcodes can be introduced in different ways during the amplification process: via template-switching in 5’ RACE amplification [213, 210], via
Methods for correction of technical noise in immune repertoire sequencing data

barcoded gene-specific primers and nested PCR [136], or during randomly primed cDNA synthesis [214]. As the range of available methodologies increases, there is a need for a better understanding of the limitations and challenges associated with barcode counting, incorporation efficiency, and error correction [212, 216, 231].

In this Chapter, I use RNA from peripheral blood mononuclear cells (PBMCs) of healthy donors, and from lymphoblastoid cell lines (LCLs) to compare the amplification efficiency, sensitivity and reproducibility of three methodologies of library preparation of BCR repertoires: 1) 3’multiplex (3’MPLX) reverse primer barcoding; 2) 5’RACE method based on template switching and 3) 5’multiplex (5’MPLX) method for barcoding during PCR (Figure II.1).

Figure II.1: Strategies for library preparation for BCR sequencing.
Comparison of methods for library preparation with three different barcode incorporation strategies. Each of the three methods uses an identical 15 nt barcode (shown in orange). For 3’Multiplex (3’MPLX) method the barcode is introduced during reverse transcription (RT) as part of the reverse J gene primer. The 5’ Multiplex PCR (5’MPLX) method uses a two-step PCR and the barcode is introduced on each of the 6 V gene primers during the first PCR step. In the 5’RACE approach the barcode is incorporated via a template-switch during RT; The 3’MPLX and 5’MPLX methods generate PCR products of approx. 400 – 450 bp. 5’RACE – based amplification produces longer amplicons (≈500 – 550bp) due to amplification of 5’UTR; All strategies use the Illumina MiSeq 300 Paired End platform for sequencing of the barcoded amplicons.
1.3.2 Aims:

1. To test a range of strategies for library preparation and incorporation of primer barcodes during B cell receptor (BCR) repertoire sequencing.

2. To select a library preparation protocol that enables reliable correction of technical noise during BCR sequencing and provides the most sensitive and reproducible capture of BCR repertoire diversity.

1.3.3 Colleagues:

The analysis on method-specific biases in chemical properties of incorporated barcodes was performed in collaboration with Dr Rachael Bashford-Rogers. The rest of the presented data is a result of my own work unless stated otherwise.
Methods for correction of technical noise in immune repertoire sequencing data

2 Methods

2.1 Samples

Peripheral blood mononuclear cells (PBMCs) and Lymphoblastoid cell lines (LCL) were obtained under appropriate ethics approval from Cambridge Blood and Stem Cell Biobank (CBSB) and from Coriell Institute respectively. Single-cell flow sorting was performed using CD20-FITC, CD19-PE, CD5-APC, IgG-V450 (BD), Aqua (for live-dead cell detection, Invitrogen) into 96 well plates from frozen healthy PBMCs.

2.2 Reverse transcription and amplification with barcoded primers

RNA extraction was performed using RNeasy Mini Kit (Qiagen) according to manufacturer’s protocol. Reverse transcription (RT) was run as a 20μL reaction with SuperScript® III (Thermo Fisher) for 3’MPLX and 5’MPLX reactions and SuperScript® II (InvitrogenTM) for 5’RACE reactions. cDNA was cleaned-up with Agencourt AMPure XP beads in each of the three methods. Reagents for each RT step were divided in two mixes. For 5’RACE protocol Mix1 (RNA template, 12μM reverse primer, nuclease-free water) was incubated for 1 min at 70°C and then immediately transferred on ice for 1 min. Mix2 (3.5 μL, 5× FS buffer, 1 μL DTT (20mM), 1 μL dNTP (10mM), 1 μL RNaseOut® , 1 μL SuperscriptII) was added and incubated at 42°C for 30 min. After 30 min reactions were removed from the thermal cycler and 1 μL of 5’ oligo (12nM) was added. Reverse transcription was resumed at 42°C for 90 min, followed by inactivation at 70°C for 10 min. Cleaned cDNA was amplified using KAPA Real-Time Library Amplification Kit (KAPA Biosystems) with 10μM 5’Universal primer and 10μM J-gene specific primer and the following thermal cycling conditions: 1 step (95°C–5 min); 5 cycles (98°C–5 sec; 72°C–2 min); 5 cycles (65°C–10 sec, 72°C–2 min); 25 cycles (98°C–20sec, 60°C–1 min, 72°C–2 min); 1 step (72°C–10 min). For the 3’MPLX protocol Mix1 (RNA template, 12 μM reverse primer, nuclease-free water) was incubated for 1 min at 70°C and then immediately transferred on ice for 1 min. Mix2 (4 μL 5× FS buffer,
1 µL DTT (0.1M), 1 µL dNTP (10mM), 1 µL SuperscriptIII) was added and incubated at 50°C for 60 min followed by inactivation at 70°C for 15 min. Cleaned cDNA was amplified with V-gene multiplex primer mix (10µM each forward primer) and 3’ universal reverse primer (10µM) using KAPA protocol and the thermal cycling conditions described above for 3’MPLX method. For the 5’MPLX RT step was carried out as for 3’MPLX method but with non-barcoded J-gene reverse primer. cDNA was cleaned up and amplified for 5 cycles with KAPA enzyme and barcoded forward V gene primer mix (2.5µM each primer) and non-barcoded J-gene reverse primer with the following thermal cycling conditions: 1 step (95°C - 5 min); 5 cycles (98°C - 20sec, 60°C - 1 min, 72°C - 2 min); 1 step (72°C - 10 min). The PCR product was cleaned-up and amplified further with 5’universal forward primer (10µM) and non-barcoded J-reverse primer (10µM). Thermal cycling conditions were the same as for the 3’MPLX method.

Nucleotide sequences of all used primers are provided in Table II.1:
Table II.1: Nucleotide sequences of primers used for BCR amplification

<table>
<thead>
<tr>
<th>Primer name</th>
<th>IGH region</th>
<th>Primer sequence</th>
<th>Barcoding pattern</th>
<th>Direction</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>JH</td>
<td>J region</td>
<td>CTTACCTGAGGA GACCTG ACC</td>
<td>non-barcoded</td>
<td>reverse</td>
<td>5'MPLX, 5'RACE</td>
</tr>
<tr>
<td>JH_BC</td>
<td>J region</td>
<td>TGUCCAG CACGCTU CAGG CUN NNNUNN NNUNNNN CTTCAGT GAGG ACGGTG ACC</td>
<td>3' barcoded</td>
<td>reverse</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH1-FR1</td>
<td>V region</td>
<td>GCCTCCT GTAGA GGTCTCTGCAAG</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH2-FR1</td>
<td>V region</td>
<td>GTCTGG TCTAC CGCTGGTG AAACC</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH3-FR1</td>
<td>V region</td>
<td>CTGGGGGGTCTGGAGACTCTCTCTGT</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH4-FR1</td>
<td>V region</td>
<td>CTCTGGAGACCCTGTCCCTCACCTG</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH5-FR1</td>
<td>V region</td>
<td>CGGGGAGTCTCTGAACTCTCTG</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH6-FR1</td>
<td>V region</td>
<td>TGCCAGA CACCTCTG CTCACTCTGT</td>
<td>non_barcoded</td>
<td>forward</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>VH1-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>VH2-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>VH3-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>VH4-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>VH5-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>VH6-FR1_BC</td>
<td>V region</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>5' barcoded</td>
<td>forward</td>
<td>5'MPLX</td>
</tr>
<tr>
<td>5' universal</td>
<td>N/A</td>
<td>AGCAGTGGTATCA AGCGA</td>
<td>universal</td>
<td>forward</td>
<td>5'MPLX, 5'RACE</td>
</tr>
<tr>
<td>3' universal</td>
<td>N/A</td>
<td>TGCCAG CACGCTU CAGCC</td>
<td>universal</td>
<td>reverse</td>
<td>3'MPLX</td>
</tr>
<tr>
<td>switch_oligo</td>
<td>N/A</td>
<td>AGCAGUGTAUCA AGCCAGA GUUNNN UNNNUNN KNUUGCC TCTGAG AAGTCTCTGCA AG</td>
<td>universal</td>
<td>forward</td>
<td>5'RACE</td>
</tr>
</tbody>
</table>
2 Methods

2.3 Next-generation sequencing, sequence quality control (QC) and barcode filtering

MiSeq libraries were prepared using Sanger Bespoke Sequencing Team protocols and sequenced using 300bp paired-ended MiSeq (Illumina) reads with 20% PhiX spike-in sequences for improved library complexity. Amplicons were not sheared prior to ligation of Illumina adaptors to retain the barcode sequence (unique molecular identifier (UMIs)). Raw MiSeq reads were filtered for base quality (median Phred score >34) using the QUASR program (http://sourceforge.net/projects/quasr/) [222]. MiSeq forward and reverse reads were merged together if they contained identical overlapping region of >50bp, or otherwise discarded. Universal barcoded regions were identified in reads and orientated to read from V-primer J-region primer. The barcoded region within each primer was identified and checked for conserved bases (i.e. the T’s in NNNNTNNNNTNNNNTN). Forward and reverse primers were trimmed from each sequence, and sequences were grouped into barcode bins based on the identity of their barcode sequence. A consensus sequence was generated for each barcode group. A barcode bin was retained if there was >80% sequence certainty between all sequences obtained with the same barcode. For barcode bins in which there was >80% and <100% identity between sequences, the consensus sequence was taken as a representative. Barcode bins with with less than five reads per bin were only accepted if no errors were present in any of the sequences, otherwise rejected Figure II.2. Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IgHV and J genes from the IMGT database were retained using BLAST [220]. Due to the difference in length of the different gene families, different BLAST E-value thresholds were used for the IGHV, IGHD, and IGHJ genes (10⁻⁷⁰, 10⁻³, and 10⁻²⁰, respectively).
Methods for correction of technical noise in immune repertoire sequencing data

Figure II.2: Schematic of barcode filtering criteria
Each baced primer contains a gene-specific sequence, a 15 nt UMI and a universal sequence. Joined reads are grouped into bins based on the sequence of their barcode (UMI). Sequences with 5 or more reads in a bin are filtered based on the number of sequences with mismatches from the consensus sequence for each bin. Bins with >80% identical sequences are accepted and reduced to the consensus. Bins with <80% identical sequences are additionally filtered on the basis of 80% certainty of each base. If there was a single base with <80% certainty, the bin was rejected. Bins with <5 reads were only accepted if no errors were identified.

2.4 Network analysis
After barcode filtering each unique BCR sequence was represented as a vertex and the number of identical BCR sequences defines the vertex size. Edges are created between vertices that differ by one nucleotide. Clusters are groups of interconnected vertices. The network analyses were performed using igraph implemented in R (http://igraph.sourceforge.net/index.html). The distribution of mismatches within a single network cluster were determined by aligning the sequence representing the largest vertex with the sequences to which it is connected, and the positions of mismatches were determined along the sequences. Two-sided t-tests were performed in R. Gini Index was used as a diversity measure and was calculated by ordering
the cluster or vertex sizes from the largest to smallest and creating a cumulative frequency
distribution, where \( R = \{ r_1, r_2, \ldots, r_n \} \times r_i \) is the cumulative size of all of the largest clusters
until the \( i \)th largest cluster and normalized such that \( r_n = 1 \). The Gini index is:

\[
Gini \text{ index}(g) = \sum_{i=1}^{N} \frac{r_i - \left( \frac{i}{N} \right)}{N}
\]

where \( N \) is the number of clusters[243]. Gini index values were in the range between 0 and 1
and represent repertoires with increasing clonality.

### 2.5 Mismatch profiles of BCR reads in barcode bins

Mismatch profiles were calculated for all barcode bins (group of BCR sequences with the
same barcode) per sample. Total frequency of mismatches at each position of the joined BCR
reads were calculated as follows:

\[
\text{mismatch at position } x \times 100% = \frac{\text{mismatch at position } x}{\text{total number of mismatches per barcode bin}} \times 100%
\]

### 2.6 Statistical methods

The relationships among estimated VJ gene frequencies across all four replicates of PBMC
sample H1 were estimated using Pearson’s correlation coefficient. Gene frequency for each
VJ gene combination was first calculated as the percentage of the total captured combinations
per sample. Principle component analysis (PCA) of network parameters (number of unique
BCR sequences, largest cluster size, second largest cluster size, vertex Gini index, cluster
Gini index) was performed after log\(_2\)-transformation of each parameter value and using
spectral decomposition approach as part of the prcomp function in R. PCA results were
plotted using ggbiplot2 with observations and variables scaled to 1.
3 Results

3.1 Sensitivity of BCR repertoire capture across library preparation methods

The theoretical maximum of RNA molecules primed with unique barcodes with 12 random nucleotides is $1.7 \times 10^7$, assuming no barcode sequence biases and an exclusive priming of each unique barcode to a single unique RNA molecule. However, the size of captured BCR repertoire can be affected by the efficiency of the amplification process and the degree of bias this introduces. I compared the repertoire diversity captured by each amplification strategy by amplifying in parallel three PBMC samples from healthy individuals and four LCL samples with each of the amplification methods, starting from the same concentration of RNA.

On average, the number of unique BCR molecules captured by the 3’MPLX method captured between 9 and 90 times larger immune repertoires than the 5’RACE and the 5’MPLX methods across all samples, particularly in healthy PBMCs (Table II.2). For the 5’RACE reactions, up to 92% of the sequences were lost at the read joining step due to insufficient overlap. This effect likely results from the increased size of the PCR amplicon (500 – 550bp) covering the 5’UTR and leader sequence of the immunoglobulin gene. Additionally, the 5’RACE method had significantly increased rates (p-value < 0.002, Mann-Whitney) of read failure of Open Reading Frame (ORF) identification with an average of 56% of the primer-matched reads removed at this step, compared to 13% and 6% for the 5’MPLX and 3’MPLX methods, respectively.
Table II.2: List of samples and output of read QC steps

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (Fwd)</th>
<th>Raw reads (Rev)</th>
<th>Joined reads</th>
<th>Primer matched reads</th>
<th>Reads with ORF</th>
<th>Unique BCRs</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1_a_3MPLX</td>
<td>543808</td>
<td>523060</td>
<td>403918</td>
<td>30199</td>
<td>30372</td>
<td>27863</td>
<td>EGAN00001588508</td>
</tr>
<tr>
<td>H1_a_5MPLX</td>
<td>176177</td>
<td>173523</td>
<td>112792</td>
<td>1631</td>
<td>1484</td>
<td>1384</td>
<td>EGAN00001588509</td>
</tr>
<tr>
<td>H1_a_5RACE</td>
<td>611997</td>
<td>617202</td>
<td>111029</td>
<td>3186</td>
<td>1300</td>
<td>1176</td>
<td>EGAN00001588510</td>
</tr>
<tr>
<td>H1_b_3MPLX</td>
<td>200119</td>
<td>104411</td>
<td>82730</td>
<td>31806</td>
<td>30719</td>
<td>18858</td>
<td>EGAN00001588514</td>
</tr>
<tr>
<td>H1_b_5MPLX</td>
<td>181880</td>
<td>105941</td>
<td>79652</td>
<td>573</td>
<td>553</td>
<td>312</td>
<td>EGAN00001588515</td>
</tr>
<tr>
<td>H1_b_5RACE</td>
<td>169730</td>
<td>127564</td>
<td>31202</td>
<td>6019</td>
<td>3485</td>
<td>397</td>
<td>EGAN00001588516</td>
</tr>
<tr>
<td>H1_c_3MPLX</td>
<td>379587</td>
<td>228775</td>
<td>144743</td>
<td>63556</td>
<td>60689</td>
<td>49645</td>
<td>EGAN00001588523</td>
</tr>
<tr>
<td>H1_c_5MPLX</td>
<td>353715</td>
<td>231557</td>
<td>167611</td>
<td>693</td>
<td>663</td>
<td>303</td>
<td>EGAN00001588524</td>
</tr>
<tr>
<td>H1_c_5RACE</td>
<td>317716</td>
<td>256045</td>
<td>31202</td>
<td>6019</td>
<td>3485</td>
<td>397</td>
<td>EGAN00001588525</td>
</tr>
<tr>
<td>H1_d_3MPLX</td>
<td>382270</td>
<td>104411</td>
<td>82730</td>
<td>31806</td>
<td>30719</td>
<td>18858</td>
<td>EGAN00001588526</td>
</tr>
<tr>
<td>H1_d_5MPLX</td>
<td>342769</td>
<td>195067</td>
<td>115877</td>
<td>853</td>
<td>813</td>
<td>445</td>
<td>EGAN00001588527</td>
</tr>
<tr>
<td>H1_d_5RACE</td>
<td>342844</td>
<td>289558</td>
<td>60133</td>
<td>10357</td>
<td>4067</td>
<td>3968</td>
<td>EGAN00001588528</td>
</tr>
<tr>
<td>H2_3MPLX</td>
<td>186565</td>
<td>102495</td>
<td>77076</td>
<td>30991</td>
<td>29644</td>
<td>25742</td>
<td>EGAN00001588517</td>
</tr>
<tr>
<td>H2_5MPLX</td>
<td>203347</td>
<td>118624</td>
<td>69636</td>
<td>237</td>
<td>223</td>
<td>123</td>
<td>EGAN00001588518</td>
</tr>
<tr>
<td>H2_5RACE</td>
<td>191504</td>
<td>154431</td>
<td>41358</td>
<td>3891</td>
<td>1480</td>
<td>1421</td>
<td>EGAN00001588519</td>
</tr>
<tr>
<td>H3_3MPLX</td>
<td>182958</td>
<td>103150</td>
<td>77521</td>
<td>31429</td>
<td>30017</td>
<td>27484</td>
<td>EGAN00001588520</td>
</tr>
<tr>
<td>H3_5MPLX</td>
<td>178516</td>
<td>103805</td>
<td>78254</td>
<td>408</td>
<td>362</td>
<td>177</td>
<td>EGAN00001588521</td>
</tr>
<tr>
<td>H3_5RACE</td>
<td>171018</td>
<td>137021</td>
<td>39307</td>
<td>4913</td>
<td>2322</td>
<td>2238</td>
<td>EGAN00001588522</td>
</tr>
<tr>
<td>LCL1_3MPLX</td>
<td>191009</td>
<td>84295</td>
<td>66836</td>
<td>34801</td>
<td>33161</td>
<td>6648</td>
<td>ERS939080</td>
</tr>
<tr>
<td>LCL1_5MPLX</td>
<td>177650</td>
<td>137102</td>
<td>83503</td>
<td>297</td>
<td>259</td>
<td>92</td>
<td>ERS939081</td>
</tr>
<tr>
<td>LCL1_5RACE</td>
<td>166344</td>
<td>128303</td>
<td>43528</td>
<td>7604</td>
<td>4040</td>
<td>3086</td>
<td>ERS939082</td>
</tr>
<tr>
<td>LCL2_3MPLX</td>
<td>633249</td>
<td>291988</td>
<td>210576</td>
<td>36538</td>
<td>28834</td>
<td>3804</td>
<td>ERS939077</td>
</tr>
<tr>
<td>LCL2_5MPLX</td>
<td>175653</td>
<td>96821</td>
<td>54327</td>
<td>251</td>
<td>54</td>
<td>25</td>
<td>ERS939078</td>
</tr>
<tr>
<td>LCL2_5RACE</td>
<td>264204</td>
<td>226759</td>
<td>49757</td>
<td>10067</td>
<td>3652</td>
<td>2651</td>
<td>ERS939079</td>
</tr>
<tr>
<td>LCL3_3MPLX</td>
<td>194448</td>
<td>108681</td>
<td>89420</td>
<td>62429</td>
<td>61469</td>
<td>5477</td>
<td>ERS939074</td>
</tr>
<tr>
<td>LCL3_5MPLX</td>
<td>167039</td>
<td>96877</td>
<td>60067</td>
<td>361</td>
<td>344</td>
<td>101</td>
<td>ERS939075</td>
</tr>
<tr>
<td>LCL3_5RACE</td>
<td>173134</td>
<td>144287</td>
<td>71538</td>
<td>12085</td>
<td>6123</td>
<td>2517</td>
<td>ERS939076</td>
</tr>
<tr>
<td>LCL4_3MPLX</td>
<td>1008500</td>
<td>722194</td>
<td>514525</td>
<td>23945</td>
<td>23010</td>
<td>2093</td>
<td>ERS939071</td>
</tr>
<tr>
<td>LCL4_5MPLX</td>
<td>182146</td>
<td>71326</td>
<td>51533</td>
<td>1792</td>
<td>1704</td>
<td>407</td>
<td>ERS939072</td>
</tr>
<tr>
<td>LCL4_5RACE</td>
<td>182776</td>
<td>155343</td>
<td>34395</td>
<td>12276</td>
<td>3867</td>
<td>2854</td>
<td>ERS939073</td>
</tr>
</tbody>
</table>

1. Sample name with method-specific suffix
2. Number of row forward MiSeq reads
3. Number of row reverse MiSeq reads
4. Number of joined forward and reverse reads (50nt overlap)
5. Number of reads matched to primers with 100% identity
6. Number of reads with ORF identified on IMGT
7. Number of unique BCR molecules
8. ENA or EGA accession number

57
Methods for correction of technical noise in immune repertoire sequencing data

The increased sampling depth of the 3’MPLX approach improved the functional characterisation of the immune repertoire. The 3’MPLX method captured a wide dynamic range of VJ gene frequencies. Between 16% and 95% of the VJ gene combinations captured by 3’MPLX in healthy samples remained undetected by 5’RACE or 5’MPLX (average: 5’RACE=43%, 5’MPLX=73%). The decreased sensitivity affected primarily genes present at lower frequency, suggesting limited ability of 5’RACE and 5’MPLX methods to detect rare BCR transcripts (Figure II.3).

Figure II.3: Comparison of sensitivity of amplification methods for capture of diverse PBMC and highly clonal CLL repertoires.
Frequency of VJ gene combinations (shown as read counts after barcode correction) captured by sequencing of each of the PBMC and CLL samples with the three amplification strategies. Data for sample H1 represents the pooled repertoires after resequencing of the same individuals from different PBMC samples and RNA preparations (H1_a, H1_b, H1_c, H1_d; See Table II.2).
Interestingly, hierarchical clustering of the VJ gene frequency distributions across samples clustered the highly clonal LCL repertoires by sample of origin, and the diverse PBMC repertoires by barcoding method. This suggests a strong influence of amplification strategy on the capture of highly diverse repertoires with multiple low-frequency BCR clones, but only a moderate effect on highly clonal repertoires where a single major clone dominates the B cell population (Figure II.4).

![Hierarchical clustering of VJ gene frequency profiles across samples.](image)

**Figure II.4: Hierarchical clustering of VJ gene frequency profiles across samples.**
VJ gene frequency profiles were generated for each sample by calculating the frequency of individual VJ gene combinations as percentage of the total BCR repertoire. Distance matrix was generated after comparison of VJ gene profiles across samples using Euclidean distance metric. Sample names are coloured by sample origin for LCL samples and by amplification method for PBMC samples to highlight the specific patterns in clustering.

### 3.2 Reproducibility of capture of BCR diversity across library preparation methods

Re-sequencing the same healthy individual from different PBMC samples, RNA or cDNA preparations showed virtually identical VJ gene frequency (as a percentage of total repertoire) for 3’MPLX method \( R^2 = 0.93 \) compared to moderate \( R^2 = 0.80 \) and low \( R^2 = 0.40 \) reproducibility of the 5’RACE and the 5’MPLX methods respectively (Figure II.5).
Methods for correction of technical noise in immune repertoire sequencing data

Two biological replicates (PBMC1/PBMC2) from the same healthy individual were processed in parallel into 3 RNA samples (RNA1, RNA2, RNA3). Equal quantities (100ng) of each RNA sample was used for downstream RT-PCR with each of the 3 amplification methods. The exemplified workflow was performed for each method. Sample RNA2 was divided into two RT-PCR reactions for each method. The amplicons generated after amplification of the 3 RNA replicates were sequenced on separate lanes (Run1, Run2, Run3). Pearson correlation of VJ gene frequencies (as % of total repertoire) across the sequenced replicates ‘H1_a’, ‘H1_b’, ‘H1_c’, ‘H1_d’ for each of the three library preparation methods. The plotted biological replicates represent V J gene frequencies of reaction PCR1 (x-axis) vs mean VJ frequencies for reactions PCR2, 3, 4 (y-axis). Technical RNA and cDNA replicates are plotted as ‘RNA1’ vs ‘RNA2’, ‘RNA1’ vs ‘RNA3’, ‘RNA2’ vs RNA3’ and ‘cDNA1’ vs ‘cDNA 2’, ‘cDNA 1’ vs ‘cDNA3’, ‘cDNA2’ vs ‘cDNA3’.

Gene combinations present at 0.005% of the repertoire were successfully captured across the 4 sample replicates by the 3’MPLX method compared to lowest reproducible frequency detection of 0.03% and 0.5% for the 5’RACE and the 5’MPLX methods. Only 31% (21/68) of the genes present at average replicate frequency below 0.01% in the 3’MPLX were captured in at least one of the 5’RACE replicates and 0.3% (2/68) in the 5’MPLX. In terms of individual RNA transcripts, >27.4% of BCRs from 3’MPLX were recaptured between at least
3 Results

one technical replicate and >13.8% of the 3’MPLX BCRs were observed in all 4 replicates, compared to only 0.556% and 0% of 5’RACE and 5’MPLX respectively (Table II.3).

Table II.3: Proportion of re-captured BCRs across replicates

<table>
<thead>
<tr>
<th>Replicates</th>
<th>3’MPLX</th>
<th>5’RACE</th>
<th>5’MPLX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.725</td>
<td>0.956</td>
<td>0.935</td>
</tr>
<tr>
<td>2</td>
<td>0.041</td>
<td>0.028</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.095</td>
<td>0.009</td>
<td>0.013</td>
</tr>
<tr>
<td>4</td>
<td>0.138</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td>Total BCRs</td>
<td>192354</td>
<td>13837</td>
<td>3513</td>
</tr>
</tbody>
</table>

3.3 Effect of BCR amplification on the estimation of BCR diversity

Next, I tested if the observed differences across methods with respect to the sensitivity for repertoire capture affect the estimation the total repertoire diversity. I used the technical and biological replicates of PBMC and CLL samples (described in Figure II.5) to characterise the diversity of the captured unique BCRs. Network analysis was used to derive cluster and vertex Gini indices as measures of repertoire diversity (see ‘Methods’ section). The derived network parameters (unique RNA molecules, Largest cluster size, Second Largest cluster size, vertex and cluster Gini indices) were used in a principle component analysis to test if the derived repertoire diversity was biased by the amplification strategy used. Indeed, in healthy PBMC repertoires, diversity metrics clustered by method of library preparation instead of biological sample of origin. The total number of unique BCR RNA molecules was the main factor driving the observed clustering and was positively correlated with the first principle component which explains ≈80% of the observed variance. Consistent with the effect of amplification strategy observed on VJ gene frequency distribution, in the case the LCL samples BCR read number contributed less to the observed variance, likely due to the highly clonal repertoire that does not require higher molecule sampling depth to be captured successfully (Figure II.6).
Methods for correction of technical noise in immune repertoire sequencing data

Figure II.6: Principle component analysis of network parameter of BCR repertoires.
Principle Component Analysis of network parameters: a) Vertex Gini Index; b) Cluster Gini index; c) Largest cluster size; d) Second largest cluster size) derived from the captured repertoires of samples H1-3 in ‘PBMC’ panel and samples LCL1-4 in ‘LCL’ panel. The ‘RNA’ label designates the number of unique BCR RNA molecule. Arrows represent the eigenvectors of the different network parameters contributing to the observed variance. The numbers shown in brackets for each axis label represent the percentage of variance explained by the first (x-axis) and the second (y-axis) principle components.

3.4 Method-specific PCR amplification biases result in distinct diversity estimates across amplification protocols

The quantitative and qualitative differences in the BCR repertoires captured by the three library preparation methods suggested method-specific effects of PCR amplification which affect the size and the diversity of the captured repertoires. Therefore, I next investigated the potential sources of such amplification biases by characterising the clonal architecture of a healthy PBMC repertoire across methods and the degree of technical noise removed after barcode filtering (Figure II.7).
Figure II.7: Network properties of BCR repertoires across amplification methods.

Network analysis was performed as explained in ‘Methods’ section. Each vertex represents a unique BCR sequence and the size of the vertices corresponds to the count of BCR reads corresponding to this sequence. Edges were drawn between vertices that differ in 1 nt. The three top panels represent H1_a sample sequenced with 3’MPLX, 5’RACE, 5’MPLX methods respectively (left-right) and networks were generated without barcode filtering to remove PCR errors. The lower panels represent the same BCR repertoires but networks were generated after read filtering based on primer barcodes (see ‘Methods’).

Differences in size and diversity of the sequenced repertoires were present before barcode filtering, confirming the presence of method-specific effects. Starting from the same RNA concentration, the 5’MPLX and 5’RACE methods captured 94,320 and 42,619 total BCRs respectively compared to 362,891 sequences from the 3’MPLX method. In addition to the reduced size of the captured repertoire, the 5’MPLX and 5’RACE methods also showed high rates of clonal expansion and somatic hypermutation (as measured by the high value of the cluster Gini index). After correction of technical noise, the estimated high clonality of BCR repertoires in 5’MPLX and 5’RACE was removed resulting in reduced maximum cluster sizes and low network complexity (Table II.4). By contrast, despite the 15-fold reduction in unique RNA molecules for the 3’MPLX method, the removal of technical noise did not
Methods for correction of technical noise in immune repertoire sequencing data

change the overall diversity of the captured repertoire, suggesting minimal amplification biases in the 3’MPLX amplification strategy.

Table II.4: Network parameters of BCR repertoires before/after barcode correction.

<table>
<thead>
<tr>
<th>Error correction</th>
<th>Method</th>
<th>Read count(^1)</th>
<th>Vertex size(^2)</th>
<th>Vertex Gini Index(^3)</th>
<th>Cluster Gini Index(^4)</th>
<th>Largest Cluster (%): (^5)</th>
<th>2nd Largest Cluster (%): (^6)</th>
<th>% Vertices in largest cluster (^7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before 3’MPLX</td>
<td>362891</td>
<td>256909</td>
<td>0.2799</td>
<td>0.2777</td>
<td>0.3474</td>
<td>0.1992</td>
<td>0.1047</td>
<td></td>
</tr>
<tr>
<td>Before 5’MPLX</td>
<td>94320</td>
<td>60688</td>
<td>0.3510</td>
<td>0.3461</td>
<td>2.6865</td>
<td>2.1564</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>Before 5’RACE</td>
<td>42619</td>
<td>32384</td>
<td>0.2372</td>
<td>0.2822</td>
<td>1.8442</td>
<td>1.4007</td>
<td>0.0030</td>
<td></td>
</tr>
<tr>
<td>After 3’MPLX</td>
<td>63238</td>
<td>30372</td>
<td>0.2291</td>
<td>0.0605</td>
<td>0.3921</td>
<td>0.2862</td>
<td>0.0020</td>
<td></td>
</tr>
<tr>
<td>After 5’MPLX</td>
<td>1588</td>
<td>1384</td>
<td>0.1257</td>
<td>0.1288</td>
<td>1.7632</td>
<td>1.1964</td>
<td>0.0722</td>
<td></td>
</tr>
<tr>
<td>After 5’RACE</td>
<td>1300</td>
<td>1176</td>
<td>0.0920</td>
<td>0.0541</td>
<td>0.7692</td>
<td>0.6923</td>
<td>0.0850</td>
<td></td>
</tr>
</tbody>
</table>

1 Total number of ORF and barcode filtered reads
2 Total number of unique BCR molecules (total number of vertices)
3 Measure of repertoire diversity
4 Measure of degree of somatic hypermutation within a clone
5 Cluster consisting of the largest number of sequences within the network
6 Cluster consisting of the second largest number of sequences within the network
7 Percentage of unique BCR molecules within the largest cluster

The difference in amplification biases across methods was also evident from the barcode profiles generated during the filtering of each BCR read (Figure II.8). The number of joined reads associated with a unique barcode (‘barcode multiplicity’) was used to calculate the level of skewing in transcript amplification in healthy PBMC repertoires. The 5’MPLX method showed the greatest amplification bias with an average of 12 reads associated with the same barcode. By comparison, 5’RACE averaged 2.5 reads and the 3’MPLX averaged 1.2 reads. Importantly, in the 5’MPLX and 5’RACE samples barcodes were associated with 1834 and 391 maximum reads per barcode compared to only 41 in PBMC and 154 in LCL samples for the 3’MPLX method which showed least amplification skewing. The observed barcode profile in the 3’MPLX method was most successful in the capture of both diverse and clonal repertoires with close to 1:1 ration of unique barcodes to RNA molecules in the PBMC samples, and highest ratio of barcodes to RNA in LCL improving the capacity to capture the repertoires with extensive clonal expansion.
Figure II.8: Barcode profiles across amplification methods.
Profiles of primer-associated barcodes for each of the three amplification strategies: 3’MPLX – red, 5’MPLX – green, 5’RACE – blue. “Multiplicity” was calculated as number of BCR reads associated with a single barcode. Barcodes/BCR ratio represents the number of unique barcodes associated with a unique BCR RNA. ‘***’ – p-value < 0.001; ‘*’ – p-value < 0.01 (two-sided Wilcoxon signed-rank test).

Interestingly, in addition to the increased amplification bias, the 5’MPLX and 5’RACE methods also showed higher rate of nucleotide mismatch from consensus for reads within the same barcode group compared to the 3’MPLX method (Figure II.9). Since all three methods use identical DNA polymerising enzyme, the difference in mismatch frequency is unlikely to reflect a PCR error rate.
Methods for correction of technical noise in immune repertoire sequencing data

Figure II.9: Nucleotide mismatch distribution within barcode groups.
Percentage of reads with single nucleotide mismatches from consensus sequence for each barcode group. Legend represents the number of mismatches from 0 to 5. The category with 5 mismatches includes 5 or more mismatches per bin.

To investigate the source of mismatch bias across methods, I generated maps of nucleotide mismatch frequency along the BCR sequence for each amplification strategy (Figure II.10). The mismatch profile of the 3’MPLX method showed uniform distribution of mismatches from consensus along the BCR sequence, consistent with the expected stochastic nature of PCR error. By contrast, the 5’MPLX method showed an enrichment of mismatches within the CDR3 region relative to the rest of the BCR sequence. Such mismatch profile is likely associated with the overlapping nature of the variable gene priming resulting in redundant use of barcodes for transcripts generated from different recombination and somatic hypermutation events. The SNP profile of the 5’RACE reaction was distinct from the other two methods and showed an enrichment around position 200 of the BCR within the V gene segment of the transcript. The increased mismatch frequency in this region of the V gene could represent a PCR recombination event or an increased error rate in this part of the variable gene associated with the joining of forward and reverse read during QC.
Figure II.10: Nucleotide mismatch maps per barcode group.
Frequencies of single nucleotide mismatches at each position of BCR reads from barcode groups for healthy PBMC repertoires. The x-axis represents the nucleotide coordinates of joined BCR read in V-J gene orientation. The heat colours represent frequency of mismatches at any given position as a percentage of the total number of mismatches detected for each sample (blue–low frequency; red–high frequency).

3.5 Different PCR amplification strategies select for specific chemical properties of incorporated barcodes

Next, I tested if the observed amplification biases across methods result in a non-random incorporation of barcodes during library preparation. I used a simulated barcode dataset with 1,000,000 randomly generated primers per sample. Each of this primers had a randomly selected bases added into the variable positions of the UMI. Using this fully random dataset I compared the sequence of observed vs. randomly generated barcodes to identify any method-specific biases in barcode incorporation.

The 3’MPLX and 5’MPLX barcodes showed significantly lower % GC content in the observed primer set compared to randomly simulated barcodes, while the 5’RACE- barcoded primers had an enrichment in GC content compared to random simulation (Figure II.11). This suggests that higher GC content may be preferable for template switching but not for multiplex amplification. Furthermore, the 3’MPLX and 5’MPLX barcodes showed prefer-
Methods for correction of technical noise in immune repertoire sequencing data

...ually higher incorporation of regions of three identical adjacent bases (homopolymers). However, there was no significant trend with the 5’RACE-barcode primers.

Interestingly, each amplification method had a different nucleotide triplet profile, suggesting distinct barcode features are preferred between the different amplification methods. These method-specific barcode profiles were consistent across samples amplified with the 3’MPLX and 5’RACE methods but varied considerably in the 5’MPLX, suggesting minimal contribution of barcode sequence for primer efficiency in 5’MPLX. Together these show that both 3’MPLX and 5’RACE amplification strategy can select for particular barcode sequence motifs and features and those observed in 3’MPLX method are optimal for high molecule sampling depth and minor amplification biases.
Figure II.11: Sequence profiles of incorporated barcodes. Biases in sequence profiles of incorporated barcodes compared to a simulated barcode set of 1,000,000 barcodes. “GC content” panel represents the percentage of G an C bases in observed vs. simulated barcode sets for each method. “Homopolymers” panel represents the mean number of adjacent identical nucleotides for the barcodes in each amplification method. The heatmap shows the frequency of different nucleotide triplet combinations within the observed barcode sequences across methods. The colour blocks represent the corresponding method: red (3'MPLX), blue (5’RACE), green (5'MPLX). Distance matrix was build using Euclidean distance and hclust package in R.
4 Discussion

In this chapter, I performed a comparison of three library preparation strategies for sequencing of BCR repertoires. Each of the compared methods was based on a different strategy for targeted PCR amplification of the immunoglobulin heavy chain. The incorporation of barcoded primers during library preparation enabled the correction of technical noise and the assessment of the introduced method-specific biases.

The choice of an amplification method considerably affected the capture of diverse PBMC repertoires. The reduced sampling depth of the 5’RACE and 5’MPLX methods resulted in limited ability to detect low frequency VJ gene combinations. Such limitation would be less important for the study of highly clonal repertoires (e.g. in leukaemia [222]) but presents a serious challenge for the characterisation of rare B cell populations and for monitoring dynamic changes in polyclonal responses.

In addition to improved sensitivity, the 3’MPLX method also provided considerable reduction of amplification biases compared to the 5’MPLX and the 5’RACE methods. The level of introduced bias during amplification is important for the study of B cell clonal evolution during antigen-driven responses. The skewing in transcript amplification leads to artificially increased clonality as observed for the 5’RACE and the 5’MPLX methods. This highlights the importance of barcode correction of technical noise (especially in the context of 5’MPLX and 5’RACE amplification) to distinguish between amplification biases and true clonal expansion events.

The observed differences in amplification biases across methods were reflected in the patterns of technical noise removed after barcode filtering. The 5’MPLX method showed the most prominent skewing in transcript capture and an increased rate of mismatches between sequences with the same barcodes. This lead to removal of large proportion of sequenced reads after all filtering steps (less than 1% of reads were retained). The pattern of mismatch distribution along the sequence of the BCR reads showed enrichment of mismatches within the CDR3 region. This likely results from priming of more than one RNA molecule with
the same barcoded primer. Such barcode mis-assignment event is possible due to the staggered overlapping nature of the multiplex V region primers. As a result, already barcoded transcripts can be further primed during the first 5 cycles of PCR amplification resulting in primer redundancy and limited diversity of captured BCR molecules. This pattern of technical noise suggests that the addition of UMIs to multiplex primer sets is not a reliable approach for correction of technical noise because it limits the priming of each unique RNA with a unique barcode.

The barcode profile observed for the 5’RACE method showed a peak in mismatch accumulation around 200bp within the BCR sequence (within the V gene). This mutational footprint could result from a hotspot for a recombination event specific to the template-switching reaction. An alternative explanation of such a mismatch profile could be the increased length of the 5’RACE amplicons. This lead to overlap between the forward and reverse reads only at the ends of the sequence reads where the Phred score of base quality gradually decreases and sequencing errors are more likely. This suggests that read lengths higher than 300bp might be required for accurate capture of BCR diversity using 5’RACE library preparation approach.

Given the aim of this research chapter to identify an optimal method for amplification and reliable sequencing of BCR repertoires, the exact origins of amplification biases associated with the chemistry of 5’MPLX and 5’RACE reactions were not investigated in further detail. The results presented here constitute an interesting basis for possible further analysis of such method-specific biases and highlight the importance of a careful choice of an amplification strategy for reliable capture of immune repertoire diversity.

The 3’MPLX method for reverse primer barcoding overcomes the majority of biases and limitations of alternative barcoding approaches and is the optimal method for immune repertoire sequencing in terms of sensitivity, reproducibility and low bias in repertoire capture. Therefore, this priming strategy was used for the development of an isotype-specific primer set for BCR sequencing described in Chapter III.
Chapter III

BCR sequencing for characterisation of antigen-driven B cell evolution
BCR sequencing for characterisation of antigen-driven B cell evolution

1 Introduction

The adaptive immune system has mechanisms for generation of diverse populations of B and T cells with unique receptor molecules. This enables the specific recognition of virtually any encountered antigen (estimated unique B cell receptors (BCRs) in an individual \(\approx 10^{11−13}\)) and limits strong autoreactive specificities [181]. Any dysregulation of the balance between detection of foreign antigens and self-recognition can lead to immunodeficiency, autoreactivity, and malignancy.

1.1 Determinants of antigen specificity in the BCR molecule

In the case of B cells, the recognition of a specific antigen is encoded in the genetics of the BCR and involves several layers of genetic diversity generated before and after antigen exposure. Recombination of BCR genes in the bone marrow and pairing of the heavy and light chains generates the initial pool of immune diversity characteristic of the naïve B cells. These B lymphocytes are generated without antigen exposure and have not undergone clonal expansion. Therefore, each cell carries a unique unmutated BCR with the potential to mature into distinct antibody producing clonal lineages [244].

The specific binding of a BCR to an antigen, along with other auxiliary signals (e.g. from T cells, cytokines and TLR ligands) leads to B cell activation, which is characterised by clonal expansion and somatic hypermutation (SHM) of immunoglobulin variable genes [245]. This subsequently leads to selection of B cells based on the affinity of their BCR molecule for a specific antigen [246]. B cells with high affinity are stimulated to proliferate into expanded populations of cells with identical BCRs which ultimately differentiate into antibody-producing plasma cells and/or memory B cells specific to the same antigen. Together, the naïve B cell diversity and the antigen-driven SHM underlie the breadth of antigen specificities encoded within the variable domains of BCR molecules (Figure III.1).
1 Introduction

Figure III.1: Stages of evolution of antigen-specific B cell responses
Naïve B cells co-express unmutated IgD and IgM immunoglobulin receptors on their surface as they exit the bone marrow. Upon exposure to an antigen, B cells with high-affinity BCRs are clonally expanded and undergo further affinity maturation which involves somatic hypermutation (SHM) of their immunoglobulin variable genes and in most cases class-switch recombination (CSR) to different isotypes. Affinity maturation results in production of antibody-secreting plasmablasts and B memory cells with mutated BCRs.

1.2 Evolution of antibody effector functions

In addition to the recognition of diverse antigens, effective B cell responses require optimal antibody effector functions reflecting the nature of the antigen and the strategy required for neutralisation. This functional plasticity of antibody molecules is encoded in the constant region segments of the immunoglobulin genes which can diversify by class-switching. Class-switching occurs after antigen exposure and results in the production of antibodies of different classes (isotypes) (IgD, IgM IgG1–4, IgA1–2, IgE) which vary in their functional phenotype, abundance and location in the body [247, 6]. The expression of a specific isotype determines
BCR sequencing for characterisation of antigen-driven B cell evolution

the function of the antibody, both in terms of avidity, specific Fc receptor binding and recruitment of different immune effector cells.

1.3 Antibody classes

Different isotype expression reflects the maturation stage of a B cell. Naïve B cells express IgM+ and IgD+ classes with unmutated variable genes, which are produced from the same initial transcript following alternative splicing [248]. Expression of other BCR isotypes is a marker of antigen-activated B cells that have undergone class-switch recombination (CSR), usually after T cell help in a germinal centre reaction. Antibodies of different classes differ in their location around the body. They also appear at different stages of an adaptive immune response and are often specialised to predominantly response to specific types of antigens via engagement of distinct soluble factors or immune effector cells (Figure III.2).

Figure III.2: Structure and functions of immunoglobulin isotypes.
Schematic of protein structure of antibodies of different classes (IgM, IgD, IgG, IgE, IgA). Glycosylation patterns are shown as black circles along the heavy chain of each isotype. Disulphide bonds (S-S) link the monomers of the IgM pentamer. J chains are present in IgM and IgA isotypes to assist antibody secretion. The secretory component of IgA antibodies enables secretion across mucosal surfaces. A summary of main functional differences between isotypes is provided as a table.
IgM is the first immunoglobulin expressed during B cell development as a monomer on the surface of B naïve cells. Upon antigenic stimulation, IgM⁺ B cells secrete multimeric IgM, where five Ig monomers are linked via disulphide bonds (Figure III.2). The pentamer also contains a polypeptide J-chain which links two of the monomers and facilitates secretion at mucosal surfaces. The pentameric structure of IgM antibodies makes them efficient at binding antigens with repetitive epitopes (e.g. bacterial capsule, viral capsid) and activation of complement cascade. As IgM antibodies are expressed early in a B cell response, they are rarely highly mutated and have broad antigen reactivity thus providing an early response to a wide range of antigens without the need for T cell help [249]. IgD isotypes are expressed on naïve B cells as they leave bone marrow and populate secondary lymphoid organs. The levels of surface expression of IgD isotype has been associated with differences in B cell activation status [248] but their role in serum is poorly understood.

The IgG, IgE and IgA antibody isotypes are generated following class-switching during germinal centre reaction and provide different effector functions in response to specific antigens. IgG is the most abundant antibody class in the serum and it is divided into 4 subclasses based on differences in the structure of the constant region genes and the ability to trigger different effector functions. Despite the high sequence similarity (90% identical on the amino acid level), each subclass has a different half-life, a unique profile of antigen binding and distinct capacity for complement activation. IgG1 antibodies are the most abundant IgG class and dominate the responses to protein antigens. Impaired production of IgG1 is observed in some cases of immunodeficiency and is associated with recurrent infections [250]. The IgG responses to bacterial capsular polysaccharide antigens are mediated primarily via IgG2 subclass, and deficiencies in this subclass result in susceptibility to certain bacterial species [251–253]. IgG2 represents the major antibody subclass reacting to glycan antigens but IgG1 and IgG3 subclasses have also been observed in such responses, particularly in the case of protein-glycan conjugates [254]. IgG3 is an efficient activator of pro-inflammatory responses by triggering the classical complement pathway [255]. It has the shortest half-life compared to the other IgG subclasses [256] and is frequently present together with IgG1 in response to protein antigens in particular after viral infections [257]. IgG4 is the least
BCR sequencing for characterisation of antigen-driven B cell evolution

abundant IgG subclass in the serum and is often generated following repeated exposure to the same antigen or during persistent infections [258, 259].

IgA antibodies are secreted in the respiratory or the intestinal tract and act as the main mediators of mucosal immunity [260]. They are monomeric in the serum, but appear as a dimer termed secretory IgA (sIgA) at mucosal surfaces. The secretory IgA is associated with a J-chain and another polypeptide chain called the secretory component [261]. IgA antibodies are divided into two subclasses that differ in the size of their hinge region [262]. IgA1 has a longer hinge region which increases its sensitivity to bacterial proteases [263]. Therefore, this subclass dominates the serum IgA, while IgA2 is predominantly found in mucosal secretions. Complement fixation by IgA is not a major effector mechanism at the mucosal surface but IgA receptor is expressed on neutrophils which may be activated to mediate antibody-dependent cellular cytotoxicity [264]. sIgA has also been shown to potentiate the immune response in intestinal tissue by uptake of antigen together with the bound antibody by dendritic cells [265].

IgE antibodies are present at lowest concentrations in peripheral blood but constitute the main antibody class in allergic responses through the engagement of mast cells, eosinophils and Langerhans cells [6]. Responses to specific helminths are also characterised with elevated levels of IgE antibodies [247].

1.4 Interaction between antibody variable region and isotype in the evolution of antigen-specific responses

In addition to the activation of distinct effector mechanisms, increasing number of studies attribute a direct role of antibody isotype in the recognition of an antigen by affecting antibody secondary structure [266, 267]. Interestingly, this effect appears to be restricted to specific phylogenetically related variable genes [266]. These observations prompt a re-visit of the initially accepted bipartite nature of antibodies and suggests that a specific combination between BCR variable genes and antibody isotype might be required for the optimal antigen
recognition and efficient neutralisation. Thus, gaining a better understanding of the evolution of B cell responses to an antigen requires simultaneous characterisation of both variable region diversity and isotype composition of BCR repertoires.

1.5 Monitoring the evolution of isotype-specific antibody responses using BCR repertoire sequencing

BCR sequencing enables the monitoring of B cell responses to an antigen by characterisation of the diversity in BCR genes. This has been demonstrated successfully in the context of infection [186], vaccination [200, 268] and immune-mediated disease [137]. The majority of the studies, however, focus only on the degree of SHM in variable genes as a metric of B cell clonal evolution. This approach has limited capacity to detect any possible effects of the isotype composition of BCR repertoires.

Therefore, in this chapter I used isotype-resolved barcoded BCR sequencing to characterise the relationship between SHM and isotype switching in healthy individuals’ repertoires and in cases of B cell malignancy. I identify significant differences in isotype frequencies and class-switch patterns between healthy and leukemic repertoires and a significant effect of BCR variable gene identity and mutation status on the class-switch fate of B cells. The work presented in this chapter demonstrates that BCR sequencing can be used to study the evolution of antigen-experienced B cell populations and the generation of antibodies of different isotypes.

1.6 Aims

1. To test the utility of BCR sequencing for genetic stratification of isotype-specific B cell populations.

2. To investigate the relationship between somatic hypermutation and class-switch recombination in the evolution of antibodies of different classes.
BCR sequencing for characterisation of antigen-driven B cell evolution

1.7 Colleagues

The work presented in this chapter was done in collaboration with Dr Rachael Bashford-Rogers who provided the sequencing data on the chronic lymphocytic leukemia samples and a subset of the healthy individuals. This work contributed to a joint publication which is currently under review. The technology for isotype-resolved BCR sequencing was used for a patent application which is currently under review by the European Patent Office. The cell sorting of isotype-specific B cells was performed by Luke Muir. Everything else is a result of my own work unless stated otherwise.

2 Methods

2.1 Samples

Peripheral blood mononuclear cells (PBMCs) were isolated from 10 mL of whole blood from nineteen healthy volunteers and six CLL patients using Ficoll gradients (GE Healthcare) for bulk-sequencing experiments. Research was approved by the relevant institutional review boards and ethics committees (07/MRE05/44). B memory cells used for FACS sorting were derived from two unlinked anonymous patient donors under the following ethics approval: 12/WA/0196. Sequencing was approved by Sanger Institute Ethics committee under HMDMC form 17/085. Read processing information for each sample is provided in Table III.1.

Table III.1: Sample filtering information for isotype-resolved sequencing

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Days after first sample</th>
<th>Timepoint</th>
<th>Number of barcoded reads</th>
<th>Number of unique barcodes</th>
<th>Number of reads after QC</th>
<th>Number of unique sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy 1</td>
<td>0</td>
<td>0</td>
<td>28828</td>
<td>28696</td>
<td>27261</td>
<td>24111</td>
</tr>
<tr>
<td>Healthy 2</td>
<td>0</td>
<td>0</td>
<td>163021</td>
<td>161947</td>
<td>153137</td>
<td>132937</td>
</tr>
<tr>
<td>Healthy 3</td>
<td>0</td>
<td>0</td>
<td>12444</td>
<td>12434</td>
<td>11817</td>
<td>11296</td>
</tr>
<tr>
<td>Healthy 4</td>
<td>0</td>
<td>0</td>
<td>73520</td>
<td>73341</td>
<td>69812</td>
<td>63686</td>
</tr>
<tr>
<td>Healthy 5</td>
<td>0</td>
<td>0</td>
<td>63183</td>
<td>62937</td>
<td>58794</td>
<td>53890</td>
</tr>
</tbody>
</table>
### Table III.1: Sample filtering information for isotype-resolved sequencing (continued)

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Days after first sample</th>
<th>Timepoint</th>
<th>Number of barcoded reads</th>
<th>Number of unique barcodes</th>
<th>Number of reads after QC</th>
<th>Number of unique sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy 6</td>
<td>0</td>
<td>0</td>
<td>15053</td>
<td>15039</td>
<td>14356</td>
<td>13022</td>
</tr>
<tr>
<td>Healthy 7</td>
<td>0</td>
<td>0</td>
<td>12454</td>
<td>12449</td>
<td>11858</td>
<td>11232</td>
</tr>
<tr>
<td>Healthy 8</td>
<td>0</td>
<td>0</td>
<td>13843</td>
<td>13833</td>
<td>13021</td>
<td>12604</td>
</tr>
<tr>
<td>Healthy 9</td>
<td>0</td>
<td>0</td>
<td>20242</td>
<td>20228</td>
<td>19219</td>
<td>18263</td>
</tr>
<tr>
<td>Healthy 10</td>
<td>0</td>
<td>0</td>
<td>22513</td>
<td>22496</td>
<td>21357</td>
<td>20186</td>
</tr>
<tr>
<td>Healthy 11</td>
<td>0</td>
<td>0</td>
<td>24041</td>
<td>24023</td>
<td>22901</td>
<td>20914</td>
</tr>
<tr>
<td>Healthy 12</td>
<td>0</td>
<td>0</td>
<td>28510</td>
<td>28485</td>
<td>27062</td>
<td>26206</td>
</tr>
<tr>
<td>Healthy 13</td>
<td>0</td>
<td>0</td>
<td>163021</td>
<td>161947</td>
<td>153137</td>
<td>132937</td>
</tr>
<tr>
<td>Healthy 14</td>
<td>0</td>
<td>0</td>
<td>8626</td>
<td>8622</td>
<td>8123</td>
<td>7908</td>
</tr>
<tr>
<td>Healthy 15</td>
<td>0</td>
<td>0</td>
<td>21152</td>
<td>21106</td>
<td>20036</td>
<td>19333</td>
</tr>
<tr>
<td>Healthy 16</td>
<td>0</td>
<td>0</td>
<td>54234</td>
<td>54051</td>
<td>51886</td>
<td>44208</td>
</tr>
<tr>
<td>Healthy 17</td>
<td>0</td>
<td>0</td>
<td>42484</td>
<td>42408</td>
<td>40610</td>
<td>34804</td>
</tr>
<tr>
<td>Healthy 18</td>
<td>0</td>
<td>0</td>
<td>62985</td>
<td>62853</td>
<td>61246</td>
<td>52353</td>
</tr>
<tr>
<td>Healthy 19</td>
<td>0</td>
<td>0</td>
<td>68705</td>
<td>67901</td>
<td>64264</td>
<td>51679</td>
</tr>
<tr>
<td>CLL 1</td>
<td>0</td>
<td>0</td>
<td>72700</td>
<td>72256</td>
<td>69479</td>
<td>7941</td>
</tr>
<tr>
<td>CLL 1</td>
<td>0</td>
<td>0</td>
<td>65359</td>
<td>64735</td>
<td>61147</td>
<td>13795</td>
</tr>
<tr>
<td>CLL 1</td>
<td>182</td>
<td>1</td>
<td>45488</td>
<td>45281</td>
<td>43366</td>
<td>13221</td>
</tr>
<tr>
<td>CLL 1</td>
<td>182</td>
<td>1</td>
<td>51364</td>
<td>51134</td>
<td>48335</td>
<td>24810</td>
</tr>
<tr>
<td>CLL 2</td>
<td>0</td>
<td>0</td>
<td>56272</td>
<td>55944</td>
<td>52741</td>
<td>6529</td>
</tr>
<tr>
<td>CLL 2</td>
<td>0</td>
<td>0</td>
<td>104131</td>
<td>103030</td>
<td>97190</td>
<td>13552</td>
</tr>
<tr>
<td>CLL 2</td>
<td>28</td>
<td>1</td>
<td>57472</td>
<td>57066</td>
<td>53442</td>
<td>6960</td>
</tr>
<tr>
<td>CLL 2</td>
<td>28</td>
<td>1</td>
<td>107890</td>
<td>106548</td>
<td>100122</td>
<td>14474</td>
</tr>
<tr>
<td>CLL 2</td>
<td>119</td>
<td>2</td>
<td>79318</td>
<td>78714</td>
<td>74472</td>
<td>16708</td>
</tr>
<tr>
<td>CLL 2</td>
<td>119</td>
<td>2</td>
<td>152276</td>
<td>150028</td>
<td>141698</td>
<td>32051</td>
</tr>
<tr>
<td>CLL 3</td>
<td>0</td>
<td>0</td>
<td>62403</td>
<td>62035</td>
<td>60098</td>
<td>6919</td>
</tr>
<tr>
<td>CLL 3</td>
<td>0</td>
<td>0</td>
<td>130905</td>
<td>129252</td>
<td>124889</td>
<td>17211</td>
</tr>
<tr>
<td>CLL 3</td>
<td>255</td>
<td>1</td>
<td>50283</td>
<td>49977</td>
<td>48146</td>
<td>5989</td>
</tr>
<tr>
<td>CLL 3</td>
<td>255</td>
<td>1</td>
<td>101888</td>
<td>100789</td>
<td>97136</td>
<td>13871</td>
</tr>
<tr>
<td>CLL 4</td>
<td>0</td>
<td>0</td>
<td>65182</td>
<td>63548</td>
<td>59123</td>
<td>6685</td>
</tr>
<tr>
<td>CLL 4</td>
<td>0</td>
<td>0</td>
<td>160412</td>
<td>150593</td>
<td>138060</td>
<td>15627</td>
</tr>
</tbody>
</table>
Table III.1: Sample filtering information for isotype-resolved sequencing (continued)

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Days after first sample</th>
<th>Timepoint</th>
<th>Number of barcoded reads</th>
<th>Number of unique barcodes</th>
<th>Number of reads after QC</th>
<th>Number of unique sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLL 4</td>
<td>28</td>
<td>1</td>
<td>62743</td>
<td>62163</td>
<td>58289</td>
<td>7052</td>
</tr>
<tr>
<td>CLL 4</td>
<td>28</td>
<td>1</td>
<td>155356</td>
<td>151363</td>
<td>141261</td>
<td>16952</td>
</tr>
<tr>
<td>CLL 5</td>
<td>0</td>
<td>0</td>
<td>92180</td>
<td>91411</td>
<td>89226</td>
<td>11497</td>
</tr>
<tr>
<td>CLL 5</td>
<td>0</td>
<td>0</td>
<td>175087</td>
<td>172400</td>
<td>166464</td>
<td>33456</td>
</tr>
<tr>
<td>CLL 5</td>
<td>140</td>
<td>1</td>
<td>81703</td>
<td>80723</td>
<td>78872</td>
<td>10398</td>
</tr>
<tr>
<td>CLL 5</td>
<td>140</td>
<td>1</td>
<td>159559</td>
<td>155969</td>
<td>150195</td>
<td>28847</td>
</tr>
<tr>
<td>CLL 6</td>
<td>0</td>
<td>0</td>
<td>60381</td>
<td>60052</td>
<td>59344</td>
<td>7931</td>
</tr>
<tr>
<td>CLL 6</td>
<td>0</td>
<td>0</td>
<td>149933</td>
<td>148020</td>
<td>145238</td>
<td>21585</td>
</tr>
<tr>
<td>CLL 6</td>
<td>151</td>
<td>1</td>
<td>47853</td>
<td>47418</td>
<td>46662</td>
<td>8704</td>
</tr>
<tr>
<td>CLL 6</td>
<td>151</td>
<td>1</td>
<td>117827</td>
<td>115511</td>
<td>112477</td>
<td>21548</td>
</tr>
<tr>
<td>D1_IgD⁺IgM⁺</td>
<td>–</td>
<td>–</td>
<td>56352</td>
<td>23544</td>
<td>20797</td>
<td>10722</td>
</tr>
<tr>
<td>D1_IgD⁺</td>
<td>–</td>
<td>–</td>
<td>139069</td>
<td>20610</td>
<td>16531</td>
<td>7552</td>
</tr>
<tr>
<td>D1_IgM⁺</td>
<td>–</td>
<td>–</td>
<td>91127</td>
<td>29352</td>
<td>25492</td>
<td>8883</td>
</tr>
<tr>
<td>D1_IgA⁺IgE⁺</td>
<td>–</td>
<td>–</td>
<td>98236</td>
<td>21937</td>
<td>17289</td>
<td>7416</td>
</tr>
<tr>
<td>D1_IgG⁺</td>
<td>–</td>
<td>–</td>
<td>121162</td>
<td>11731</td>
<td>9138</td>
<td>3926</td>
</tr>
<tr>
<td>D2_IgD⁺IgM⁺</td>
<td>–</td>
<td>–</td>
<td>67727</td>
<td>1643</td>
<td>1035</td>
<td>650</td>
</tr>
<tr>
<td>D2_IgD⁺</td>
<td>–</td>
<td>–</td>
<td>48970</td>
<td>11429</td>
<td>8600</td>
<td>4679</td>
</tr>
<tr>
<td>D2_IgM⁺</td>
<td>–</td>
<td>–</td>
<td>68656</td>
<td>18233</td>
<td>15192</td>
<td>5678</td>
</tr>
<tr>
<td>D2_IgA⁺IgE⁺</td>
<td>–</td>
<td>–</td>
<td>92692</td>
<td>17512</td>
<td>12749</td>
<td>5416</td>
</tr>
<tr>
<td>D2_IgG⁺</td>
<td>–</td>
<td>–</td>
<td>155919</td>
<td>53417</td>
<td>42359</td>
<td>16468</td>
</tr>
</tbody>
</table>

2.2 FACS sorting of B memory cell subsets

Isolated PBMC were incubated with 100 µL Aqua Live/Dead viability dye (Thermo Scientific, UK) per 1 × 10^6 cells for 20 minutes at room temperature in the dark. Viability dye was prepared by diluting 1 : 400 in PBS. Cells were then washed with PBS, and incubated with the pre-titrated B cell phenotyping panel shown in Table III.1 for 30 minutes at room temperature in the dark. Following incubation, cells were washed with PBS and re-suspended in warm RPMI 1640 medium (Life Technologies, UK) supplemented with 100 U/mL penicillin (Life
Technologies, UK), 100 µg/mL streptomycin (Life Technologies), 2% FCS and 2 mM L-glutamine (Life Technologies, UK) [R2 medium]. CD27+ Memory B cells were sorted into four distinct populations based on IgM, IgD and IgG expression (IgD+ IgM+, IgD− IgM+, IgD− IgM− IgG−, IgD− IgM− IgG+). Cells were sorted directly into 1.5 mL microcentrifuge tubes containing RPMI 1640 medium supplemented with 100 U/mL penicillin, 100 µg/mL streptomycin, 2 mM L-glutamine, and 10% FCS. Sorting was carried out using a BD FACSria III and data analysed using FlowJo software 9.7.5 (TreeStar, Ashland, OR, USA). Antibody panels used are shown in Table III.2 below:

Table III.2: Antibody panels for isotype-specific B cell sorting

<table>
<thead>
<tr>
<th>Marker</th>
<th>Fluorophore</th>
<th>Laser</th>
<th>Channel</th>
<th>Volume</th>
<th>Clone</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD3</td>
<td>V500</td>
<td>Violet</td>
<td>405-525/50</td>
<td>1.25 µL</td>
<td>UCHT1</td>
<td>561416, BD Biosci</td>
</tr>
<tr>
<td>CD4</td>
<td>V500</td>
<td>Violet</td>
<td>405-525/50</td>
<td>1.25 µL</td>
<td>RPA-T4</td>
<td>560768, BD Biosci</td>
</tr>
<tr>
<td>CD14</td>
<td>V500</td>
<td>Violet</td>
<td>405-525/50</td>
<td>2.50 µL</td>
<td>M5E2</td>
<td>561391, BD Biosci</td>
</tr>
<tr>
<td>CD19</td>
<td>BV605</td>
<td>Violet</td>
<td>405-605/12</td>
<td>1.25 µL</td>
<td>SJ25C1</td>
<td>363024, Biolegend</td>
</tr>
<tr>
<td>CD27</td>
<td>PE Cy7</td>
<td>Yellow Green</td>
<td>561-780/60</td>
<td>2.50 µL</td>
<td>M-T271</td>
<td>356412, Biolegend</td>
</tr>
<tr>
<td>CD38</td>
<td>APC</td>
<td>Red</td>
<td>640-670/14</td>
<td>0.60 µL</td>
<td>HB-7</td>
<td>356606, Biolegend</td>
</tr>
<tr>
<td>CXCR4</td>
<td>PE</td>
<td>Yellow Green</td>
<td>561-582/15</td>
<td>0.60 µL</td>
<td>12G5</td>
<td>306506, Biolegend</td>
</tr>
<tr>
<td>IgM</td>
<td>FITC</td>
<td>Blue</td>
<td>488-530/30</td>
<td>1.25 µL</td>
<td>MHM-88</td>
<td>314506, Biolegend</td>
</tr>
<tr>
<td>IgD</td>
<td>PE-CF594</td>
<td>Yellow Green</td>
<td>561-610/20</td>
<td>1.25 µL</td>
<td>IA6-2</td>
<td>562540, BD Biosci</td>
</tr>
<tr>
<td>IgG</td>
<td>APC-H7</td>
<td>Red</td>
<td>640-780/60</td>
<td>1.25 µL</td>
<td>G18-145</td>
<td>561297, BD Biosci</td>
</tr>
<tr>
<td>R2 medium</td>
<td></td>
<td></td>
<td></td>
<td>86.3 µL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3 Library preparation for isotype-resolved sequencing

RNA extraction was performed using RNeasy Mini Kit (Qiagen) according to manufacturer’s protocol. Reverse transcription (RT) was run as a 20µL reaction with SuperScript® III (Thermo Fisher). cDNA was cleaned-up with Agencourt AMPure XP beads. Reagents for each RT step were divided in two mixes. Mix1 (RNA template, 12µM reverse primer, nuclease-free water) was incubated for 1 min at 70°C and then immediately transferred on ice for 1 min. Mix2 (4µL 5×FS buffer, 1µL DTT (0.1M), 1µL dNTP (10mM), 1µL
BCR sequencing for characterisation of antigen-driven B cell evolution

SuperScript® III was added and incubated at 50°C for 60 min followed by inactivation at 70°C for 15 min. Cleaned cDNA was amplified using KAPA Real-Time Library Amplification Kit (KAPA Biosystems) with V gene primer mix (10μM each primer) and 10μM 3’Universal primer and the following thermal cycling conditions: 1 step (95°C - 5 min); 5 cycles (98°C – 5 sec; 72°C – 2 min); 5 cycles (65°C – 10 sec, 72°C – 2 min); 25 cycles (98°C – 20sec, 60°C – 1 min, 72°C – 2 min); 1 step (72°C – 10 min). Primer sequences are provided in Table III.3 below:

### Table III.3: Primer sequences for isotype-resolved BCR sequencing

<table>
<thead>
<tr>
<th>Primer name</th>
<th>IGH region</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>VH1–FR1</td>
<td>V region</td>
<td>GGCCTCAGTGAAGGTCTCCTGCAAG</td>
</tr>
<tr>
<td>VH2–FR1</td>
<td>V region</td>
<td>GTCTGGTCCCTACGCTGGTGAAACCC</td>
</tr>
<tr>
<td>VH3–FR1</td>
<td>V region</td>
<td>CTGGGGGTCTCCCTGAGACCTCTCTTG</td>
</tr>
<tr>
<td>VH4–FR1</td>
<td>V region</td>
<td>CCTCGGAGACCCCTGTCCTCCTACCTTG</td>
</tr>
<tr>
<td>VH5–FR1</td>
<td>V region</td>
<td>CGGGGAGTCTCTGACATCTCTCTGT</td>
</tr>
<tr>
<td>VH6–FR1</td>
<td>V region</td>
<td>TC5CA6A6CCCTCTCACTCACTCTGTG</td>
</tr>
<tr>
<td>3’universal</td>
<td>n/a</td>
<td>TGUCCAGCACGCTUCAGGC</td>
</tr>
<tr>
<td>IGH A</td>
<td>C region</td>
<td>TGTC6CA6GACGCTCA6GCTN6NNTN6NNTN6NNTN6NNGAYGACCA6CGGCT6CCC6ATCT</td>
</tr>
<tr>
<td>IGH M</td>
<td>C region</td>
<td>TGTC6CA6GACGCTCA6GCTN6NNTN6NNTN6NNTN6NNTN6NAGACGTCACGGAGGTGGCATT</td>
</tr>
<tr>
<td>IGH D</td>
<td>C region</td>
<td>TGTC6CA6GACGCTCA6GCTN6NNTN6NNTN6NNTN6NNTN6NAGTAGCTC6CTGACCAGG</td>
</tr>
<tr>
<td>IGH E</td>
<td>C region</td>
<td>TGTC6CA6GACGCTCA6GCTN6NNTN6NNTN6NNTN6NAGTACGTC6CTGACCAGG</td>
</tr>
<tr>
<td>IGH G</td>
<td>C region</td>
<td>TGTC6CA6GACGCTCA6GCTN6NNTN6NNTN6NNTN6NAGTACGTC6CTGACCAGG</td>
</tr>
</tbody>
</table>

2.4 Sequencing and barcode filtering

MiSeq libraries were prepared using Illumina protocols and sequenced using 300bp paired-ended MiSeq (Illumina). Raw MiSeq reads were filtered for base quality (median Phred score >34) using the QUASR program (http://sourceforge.net/projects/quasr/) [222] MiSeq forward and reverse reads were merged together if they contained identical overlapping region of >50bp, or otherwise discarded. Universal barcoded regions were identified in reads and orientated to read from V-primer constant region primer. The barcoded region within each primer
was identified and checked for conserved bases (i.e. the T’s in NNNNTNNNNTNNNNT). The reads were checked for homology to the first 50bp of the reference constant region genes from the IMGT database by k-mer matching (where k=10bp). The closest matching constant region allele were identified, and information retained throughout the analysis. Primers and constant regions were trimmed from each sequence, and sequences were retained only if there was >80% sequence certainty between all sequences obtained with the same barcode, otherwise discarded. Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IgHV and J genes from the IMGT database were retained using BLAST.

2.5 Network analysis

The definition of BCR clusters used in Figure III.7 was derived as previously described by Bashford-Rogers et al. [269]. Briefly, each vertex represents a unique sequence in which relative vertex size is proportional to the number of identical sequence reads after barcode correction. Edges are generated between vertices that differ by single-nucleotide, non-indel differences and clusters are collections of related, connected vertices.

2.6 Transitional frequencies between isotypes

Frequency of unique BCR sequences was derived as percentage of total BCR repertoire per sample. Sequences with 0 mutation in variable genes and associated with IgD and IgM isotypes were considered to be derived from naïve B cell clones. Transitions between naïve and any switched isotype were calculated as number of BCR sequences associated with IgD/IgM plus any other switched isotype. Transitions between switched isotypes were calculated as number of BCR sequences associated with any pair of switched isotypes. Relative transitional frequency was calculated as number of BCRs representing a given transitional pair divided by the total number of transitions observed.
2.7 Conditional probability of a BCR being class-switched, given its variable gene family identity and mutational status

Conditional probabilities were calculated for V genes represented by more than 100 reads (in total across samples).

\[ f_x(\text{total repertoire}) = \text{total number of reads filtered as described above} \]

\[ P(\text{Mutated}|\text{VgeneX}) = \frac{f_x(\text{VgeneX with 1 or more nucleotide mutations in V gene})}{f_x(\text{total repertoire})} \]

\[ P(\text{Unmutated}|\text{VgeneX}) = \frac{f_x(\text{VgeneX with no mutations in V gene})}{f_x(\text{total repertoire})} \]

\[ P(\text{Switched}|\text{VgeneX,unmutated}) = \frac{f_x(\text{VgeneX AND switched})}{P(\text{Unmutated}|\text{VgeneX})} \]

\[ P(\text{Switched}|\text{VgeneX,mutated}) = \frac{f_x(\text{VgeneX AND switched})}{P(\text{Mutated}|\text{VgeneX})} \]

2.8 Conditional probability of a BCR being switched to a specific isotype, given its variable gene identity

Conditional probabilities were calculated for V genes represented by more than 100 reads (in total across samples) and present in more than 5 samples from healthy PBMCs. BCR reads were divided by associated isotype. If part of multi-switched clones BCR reads were considered separately for each isotype.
2 Methods

\[ f_x(\text{total repertoire}) = \text{total number of reads filtered as described above} \]

\[ P(\text{Switched}|V_{\text{geneX}}) = \frac{f_x(V_{\text{geneX}} \text{ in any switched isotype})}{P(\text{total repertoire})} \]

\[ P(\text{IsotypeY}|V_{\text{geneX}}) = \frac{f_x(V_{\text{geneX}} \text{ AND switched to Isotype Y})}{P(\text{Switched}|V_{\text{geneX}})} \]

2.9 Conditional probability of a BCR being switched to one or more isotypes, given its variable gene identity

Conditional probabilities were calculated for V genes represented by more than 100 reads (in total across samples) and present in more than 5 samples from healthy PBMCs. BCR reads were divided depending on the number of associated isotypes: 1, 2, 3, >3. Conditional probability of switching to a given number of isotypes (n) was calculated as follows:
BCR sequencing for characterisation of antigen-driven B cell evolution

\[ n = 1, 2, 3, \text{ or } > 3 \]

\[ f_x(\text{total repertoire}) = \text{total number of reads filtered as described above} \]

\[ P(\text{Switched to } n | V\text{geneX}) = \frac{f_x(V\text{geneX in n isotypes})}{P(\text{total repertoire})} \]

\[ P(\text{Isotype } n | V\text{geneX}) = \frac{f_x(V\text{geneX AND switched to Isotype } n)}{P(\text{Switched} | V\text{geneX})} \]

2.10 Calculation of isotype-specific VJ gene profiles

Isotype-specific VJ gene frequency profiles were calculated for each healthy individual based on the frequency (%) of all VJ gene combinations (no filtering for number of reads) in a total number of reads associated with a given isotype.

2.11 Hierarchical clustering of isotype-specific VJ gene profiles

Isotype-specific VJ gene frequencies were derived for each individual and distance matrix of VJ gene frequency profiles was calculated using Euclidean distance. Significance of co-clustering between isotype classes was calculated using Wilcoxon signed rank test between the inter-isotype class distances compared to the intra-isotype class distances).
3 Results

3.1 Isotype-resolved BCR sequencing with 3’MPLX PCR amplification

Based on the comparison of library preparation methods described in Chapter II, the 3’MPLX method of barcode incorporation demonstrated the highest sensitivity and reproducibility for capture of both diverse and highly clonal BCR repertoires with minimal amplification biases. Therefore, I next used this library preparation strategy as a basis of an extended method called IsoTyper which enables the amplification of BCRs of all immunoglobulin classes and subclasses in a single reaction. In this approach 12 nt barcodes were added to a mixture of 5 reverse primers corresponding to each of the 5 Ig classes. IgG and IgA - specific primers were designed to capture all respective subclasses. The constant region primer set was designed to capture both the Ig VDJ domain and sufficient constant region sequence to allow unambiguous isotype assignment.
BCR sequencing for characterisation of antigen-driven B cell evolution

Figure III.3: Library preparation steps of isotype-resolved BCR sequencing
Amplification strategy for simultaneous amplification of Ig variable and constant region using 3’ multiplex primer set to capture all immunoglobulin (Ig) classes and subclasses in a single reaction; Each reverse primer has a 15nt unique molecular identifier (UMI) for correction of sequencing and amplification biases and quantification of BCR abundance. PCR amplification of barcoded cDNA uses a multiplex primer set of 6 FR1-specific primers and a 5’ universal primer for amplification across the UMI. The size of the resulting PCR amplicon is ≈450 bp and is sequenced by 300 bp paired-end MiSeq libraries.

3.2 Isotype-resolved BCR sequencing enables ‘genetic sorting’ of isotype-specific B cell populations

As a validation of the utility of the IsoTyper library preparation strategy, I next performed isotype-resolved BCR sequencing of FACS sorted B cell populations expressing different Ig classes. CD19⁺CD27⁺ B cells from two individuals were sorted into six different populations based on the expression of IgD, IgM, IgG surface markers. The isotype-composition reflected in the BCR transcriptome of each population was calculated as percentage of BCR sequences associated with a specific isotype. The complete isotype-specific primer set for each cell population reliably captured the expected isotype identity of each population and the majority of the detected BCR transcripts corresponded to the surface expression profile of each population.
3 Results

The only cell population where the obtained isotype structure of BCR repertoire did not reflect the surface isotype expression in both individuals were IgD⁺/IgM⁻ cells which showed significant and detectable level of IgM RNA, which was even present at higher frequency than the IgD class in donor 1 (55% IgM⁺ BCRs vs 25% IgD⁺) (Figure III.4). This discordance between transcriptome and surface expression is most likely explained by the expression of IgM and IgD isotypes on a single transcript [270]. The predominance of IgM⁺ BCR transcripts in IgD⁺IgM⁺ B cell population is consistent with a previous by Yuan et al. of 7:1 ratio of μ to δ mRNA levels in resting B cells the despite the higher surface levels of IgD than IgM[271].

Based on these results, the isotype-resolved BCR sequencing method has a limited ability to differentiate between IgD⁺IgM⁻ and IgD⁺IgM⁺ but can reliably reflect the isotype identity of class-switched B cell populations. This allows to perform ‘sorting’ class-switched B cells from total PBMC based on the genetics of the BCR receptor, without the requirement of FACS sorting.

Figure III.4: Genetic sorting of isotype-specific B cell populations
Frequency of captured BCRs from each isotype as a percentage of the total repertoire for each of the 6 FACS sorted populations of CD19⁺CD27⁺ B cells for Donors 1 and 2 (‘D1’ and ‘D2’ samples, Table T1); x-axis for both panels represent the respective surface markers used for isotype-specific sorting in addition to CD19⁺CD27⁺
3.3 Isotype composition of BCR repertoires at the extremes of variable gene diversity

Having identified that isotype-resolved BCR sequencing can be used to reliably stratify class-switched B cell populations, I next tested how the isotype structure of BCR repertoires changes depending on the diversity of variable genes. Using the IsoTyper approach, I determined the BCR repertoires of 19 healthy individuals and 6 CLL patients which represent examples of high variable gene diversity and high clonality respectively (Table III.1). Using the unique molecular barcodes, I counted the number of BCR clones of a particular isotype across all samples. Here, I limited the definition of B cells clones as 100% identical at the sequence level. BCR repertoires of healthy individuals were dominated by IgM, IgA1/2 and IgG1/2 subtypes (Figure III.5) consistent with a previous report where IgM⁺ naïve cells together with IgA⁺, IgG⁺ and IgM⁺ B memory cells constitute more than 90% of the immune cells in peripheral blood $citeRN95$. By contrast, the clonal CLL repertoires were primarily un-switched and showed significant over-representation of IgM⁺ and IgD⁺ isotypes coupled to significantly lower representation of IgA1/2 and IgHG1/2 isotypes (p-value<0.005).
3 Results

Figure III.5: Isotype composition of healthy and leukemic B cell repertoires
The frequency of each Ig isotype class was calculated as percentage of BCR reads per sample associated with a specific isotype. IGHG4 sub-class was not detected in leukemic repertoires.

3.4 Multi-switched clones and isotype co-evolution are characteristic of healthy repertoires but rare in CLL

Next, I investigated the evolution of multi-switched BCR clones in healthy and in leukemic repertoires by calculating the frequencies of isotype pairs in clones associated with more than one isotype. In addition to the distinct isotype architecture, healthy and CLL repertoires also differed in the degree of observed isotype transitions from unswitched state and between switched isotypes. Dually expressing IgD/IgM clones were most commonly switched to IgG1/2 isotypes (≈30% of two-isotype clones) or to IgA1/2 (≈25% of two-isotype clones). Transitions between IgG1/2 and IgA1/2+ and were also common, contributing to ≈20% of two-isotype clones. Despite the predominance of un-switched IgM BCRs in
the highly clonal CLL repertoires class-switching events were also observed. Interestingly, IgD/IgM CLL cells were most commonly switched to IgA1/2 isotypes and these IgD/IgM to IgA1/2 transitions were significantly enriched in CLL repertoires compared to healthy repertoires. Transitions from IgD/IgM to any other isotype, as well as secondary class-switch events (transitions between switched isotypes) were significantly less common in leukemic repertoires (Figure III.6).

Figure III.6: Frequency of transitions across isotypes in healthy and leukemic B cell repertoires
Frequency of BCR clones representing transitions between different isotypes in healthy and in leukemic repertoires. The thickness of black lines represents the frequency of transitions between isotype pairs. The direction of the arrows represents the possible direction of transitions depending on the location of the respective C gene on IGH locus. Coloured lines and arrows represent transitions significantly reduced (blue) or increased (red) in CLL compared to healthy repertoires. Boxplots represent the direct comparison between healthy and leukemic repertoires across samples; ** denotes p-value < 0.0005; Wilcoxon-ranked test;

3.5 Switching to multiple isotypes is associated with higher SHM

The inference of B cell clonal evolution is commonly based on a conservative definition of a clone assuming a certain degree of sequence similarity between BCR sequences originating from a clonal expansion event. Here, we used an extended definition of a clone by using the combination of sequence diversity in variable genes [269] together with isotype profile of
3 Results

each clone to genetically stratify naïve and antigen experienced B cell clones and to follow their evolution across different isotype lineages. By definition, BCR sequences with zero mutations from germline have not undergone affinity maturation and should be associated with naïve or un-mutated but antigen-experienced (T-independent) B cell clones. Such un-mutated sequences were significantly enriched (p-value < 0.001) in un-switched (IgD+ IgM+) BCR clusters compared to non IgD+ IgM+ switched IgA1/2 and IgG1-4 clusters.

Under the assumption that the IgM+IgD+ un-mutated pool of BCRs represents primarily the naïve B cell population, I investigated the relationship between mutation of variable genes and class-switching on the process of affinity maturation and immune-repertoire evolution from naïve to antigen-experienced states. Here, B cell/BCR clusters (defined as a group of BCR sequences with 0 or cumulative 1 nucleotide differences[269]) of >2 isotypes where one isotype was IgM+IgD+ were significantly more mutated than IgM+IgD+ only BCRs that had not class-switched (mean number of mutations from germline 12.965 versus 5.205 respectively). Loss of IgD+ and IgM+ expression in clusters of >2 isotypes (i.e. ‘IgD−IgM− clones’) had the highest level of BCR mutation (mean number of mutations from germline = 18 nucleotides), suggesting a trajectory of mutations and class-switching away from the naïve B cell (Figure III.7).
BCR sequencing for characterisation of antigen-driven B cell evolution

Figure III.7: Relationship between SHM and class-switching in the evolution of B cell responses in health and disease

Boxplots of the mean number of mutations in clusters exhibiting 2 isotype classes or greater than 2 isotype classes that are either IgM+IgD+ or IgM−IgD−. Mutations were calculated as number of mutations in V genes away from the closest germline V gene as determined by IMGT-HighVQuest. Number of clusters representing each of the isotype groups is added as a value of ‘n’. *** denotes p-value <0.0005. (Wilcoxon ranked test)

3.6 Multi-switched clones are generated after substantial clonal expansion

Interestingly, individual antibody classes showed significant differences in the VJ gene usage from naïve Ig gene usage frequencies suggesting isotype-specific trajectories of clonal expansion and variable gene usage. BCRs that had undergone single class-switch to IgA1/2 and IgG1/2 isotypes showed highest correlation of their VJ gene frequency profile with
un-mutated IgD⁺/IgM⁺, suggesting substantial clonal diversification has not taken place to affect the observed VJ frequency distribution. By contrast, class-switches to IgG3 and IgE isotypes were characteristic of clones with low correlation (R₀ = 0.5 and 0.3 respectively) to naïve VJ gene usage profile. BCRs switched to two or more isotypes show higher deviation from naïve VJ gene profiles (p-values < 0.0005). Together the observed level of mutation and the deviation from naïve variable gene expression suggesting that B cell clones switch to multiple isotypes only after substantial SHM antigen driven clonal expansion which leads to overrepresentation of specific VJ gene combinations and shifts variable gene diversity of B cell populations away from naïve state (Figure III.8).

Figure III.8: Deviation of naïve VJ gene profile in class-switched isotypes
Boxplot of Pearson correlation coefficients (R²-values) between the naïve BCR (IgM⁺IgD⁺ unmutated) repertoire heavy chain VJ gene usages and that of BCR clusters associated with 1, 2 or more different isotypes. VJ gene usage was calculated as percentage of each VJ gene combination from total BCR repertoire for each cluster from a given isotype group. ** denotes p-values < 0.005, *** denotes p-values < 0.0005 and **** denotes p-values < 0.00005.
3.7 SHM and variable gene family of a BCR affect class-switch status

Next I tested if the class-switch fate of a BCR clone is affected by the identity and the mutation load of its variable gene in healthy individuals. First, I used the variable gene identity and the conditional class-switch probability as predictor and response variables in a linear regression model. I found that variable gene had a significant effect on the probability of a BCR to undergo class-switching (p-value: 3.759e\(^{-06}\)).

Then I tested the interaction between mutational status and class-switch fate of BCRs from different V gene families. IGHV5 and IGHV6 gene families had the highest probability of being associated with a class-switched isotype when in unmutated form, while IGHV1 was rarely class-switched without SHM (Figure III.9). As expected, I found that across all V gene families, BCRs have significantly increased probability of class-switch in a mutated form. Using a two-way ANOVA model, I tested if the interaction between the mutational status of a BCR and its V gene identity was associated with the class-switch fate of a B cell. Indeed, the degree of mutation and V gene identity of a BCR individually acted as good predictors of class-switch fate of a BCR (p-value: <2e\(^{-16}\)). Surprisingly, the class-switch probability of a BCR clone was also affected by the interaction between these two predictors (p-value: 0.0238), suggesting that the V gene context in which mutation occurs may also be important for BCR clone evolution and class-switching.
3.8 Variable gene identity of BCRs affects class-switch trajectories to different isotype classes

To investigate this relationship further, I next tested if the V gene effect on B cell switch fate is isotype-specific. I calculated the conditional probability of BCR being associated with 1 or multiple isotypes and of being of a specific isotype class (and subclass) given its V gene state. Indeed, the V gene identity was significantly associated with the number of isotypes a BCR clone is switched to (One-way ANOVA, p-values: 1 isotype – 0.003; 2 isotypes – $1.02e^{-06}$; 3 isotypes – $7.19e^{-08}$; >3 isotypes – 0.0038) (Figure III.10). The distribution of class-switch probabilities across all variable genes showed a distinct isotype-specific pattern with variable gene identity, significantly associated with class-switching to IgA1/2, IgE and to IgG1/G2.
BCR sequencing for characterisation of antigen-driven B cell evolution

(One-way ANOVA, IgA – $7.28e^{-09}$; IgE – 0.00394; IgG3 – 0.073; IgG1/2 – $2.45e^{-07}$; IgG4 – 0.69).

Figure III.10: Conditional probability of BCR class switch to different isotypes depending on V gene identity
Heatmap of probabilities of BCRs to be associated with one or more isotypes (left) or to be of a specific isotype class (right) depending on their V gene identity. Individual values are scaled by row; key represents row-Z score; distance matrix between V genes was generated using Euclidian distance; only V genes with >100 reads and present in >5 healthy samples were considered. ** denotes p-values < 0.005, *** denotes p-values < 0.00005 (One-way ANOVA).
3 Results

3.9 Isotype-specific repertoires exhibit distinct signatures of variable gene usage

To test if the observed effect of V gene identity on the probability of BCR class-switching to individual isotypes leads to differential V gene frequency across isotypes, I compared the variable gene usage (frequencies of all V and J genes) across the sampled Ig classes for each of the 19 healthy individuals (Figure III.11). IGHV3 and IGHJ4 were the most abundant V gene families across all isotypes but individual Ig isotypes showed significant differences in the variable and joining gene usage (except for IGHV1 family), for example IGHV3 is more common in IgA1/2 than IgE but IGHV6 is the most common IgE variable gene. Each isotype class exhibited a distinct and non-random V gene expression signature when compared to naïve IgD/IgM suggestive that certain V-C combinations are maintained or favoured in healthy mature memory B cell repertoires (2 × 2 Chi-squared test of non-randomness: p.values: IgA – 0.01; IgE – < 2.2e−16; IgG3 – < 2.2e−16; IgG1/2 – 0.0156; IgG4 – < 2.2e−16).
BCR sequencing for characterisation of antigen-driven B cell evolution

Figure III.11: Expression of V and J gene families across isotypes
Differences in V (upper panel) and J (lower panel) gene family usages between different isotype classes in healthy individuals; frequencies are calculated as percentage of BCR reads from each V or J gene family associated with the respective isotype; * denotes p-values <0.05, ** denotes p-values <0.005, *** denotes p-values <0.0005 and **** denotes p-values <0.00005. (Wilcoxon signed rank test).

3.10 Isotype-specific variable gene repertoires are conserved across individuals

Based on the conditional class-switch probability and frequency of individual V genes per isotype, each Ig class appeared to have a distinct V gene expression profile. Interestingly, the observed isotype-specific signatures were consistent across healthy individuals. Hierarchical
clustering of VJ gene repertoires (per isotype) for each of the tested healthy individuals showed that generally V-J gene usage frequencies by isotype class clustered independently of the individual but within isotype class (co-clustering p-value $< 10^{-10}$) indicative of an inherent preference for certain variable gene and isotype combinations which are independent of variations in antigen exposure between individuals (Figure III.12).

Figure III.12: Hierarchical clustering of isotype-specific VJ gene repertoires per individual
Hierarchical clustering of heavy chain VJ gene usage frequencies between different class isotypes for healthy individuals (represented as numbers in square brackets); p-value of co-clustering between isotype classes was $< 10^{-10}$ (as calculated from Wilcoxon test between the inter-isotype class distances compared to the intra-isotype class distances).

4 Discussion

The mechanism which enables a humoral immune response to link antibody binding diversity and isotype effector functions in order to resolve an infection, maintain a protective antibody
response and generate B cell memory is central to understanding infectious disease immunity and vaccine efficacy. The combined capture of deep BCR sequence diversity together with the isotype architecture of B cell populations may provide a comprehensive picture of this. Using isotype-resolved BCR sequencing in this chapter I performed genetic sorting of isotype-specific B cell populations and characterised the interplay between BCR mutation and class-switching during the evolution of a B cell response to an antigen. The highly clonal leukemic repertoires exhibit an isotype profile distinct from health and are largely restricted to primary switch event with minimal transitions between switched Ig classes. By contrast, the mature healthy B cell clones are multi-isotypic with high frequency of class switching from IgD/M to IgG1/2 and IgA1/2 isotypes in agreement with Horns et al. [153]. Class-switching to multiple isotypes in healthy repertoires is associated with increased rate of SHM suggesting that affinity maturation precedes the generation of multi-switched clones. Although the usage distribution of IGHV genes is correlated between IgD/M naïve B cells clones and B cell clones with only one isotype switch event, the IGHV and IGHJ gene usage frequency becomes significantly skewed with greater B cell maturity. This is perhaps unsurprising, as different antigenic exposures should shape different mature repertoires. However, VJ gene usage frequencies by isotype class cluster independently of the individual, indicating there is a preference for some V and J genes for different isotypes (Figure III.12).

The isotype-specific signatures of variable gene usage and their high conservation across individuals suggest an important relationship between the variable gene identity of a B cell clone and its probability to populate a specific class-switched state in peripheral blood. From these data, it is not possible to determine if the variable gene identity affects the class-switch decisions during affinity maturation in the germinal centre or if the variable gene signatures are shaped during population of the memory B cell pool. However, the characterisation of class-switching leading to generation of different antibody classes in vitro [153] shows high concordance of class-switch trajectories of sister B cells (same VDJ), suggesting that class-switch decisions can be affected by cell intrinsic mechanisms, supporting the hypothesis that the identity of a variable gene can affect class-switch decisions during B cell maturation. Interestingly, the observed isotype-specific signatures of variable gene usage in our study...
were affected by the degree of mutation of variable genes consistent with the reported discordance of class-switch fate upon increase of SHM in paired BCRs originating from the same progenitor [153]. This supports a model in which the early class-switch decisions of a B cell upon antigen stimulation have a degree of variable gene dependency which can be later outcompeted by antigen-driven SHM and cytokine stimulation leading to some poly-switched clones. Such a role of variable gene status is interesting in the context of recent work by Greiff et al. [182] which describes a high degree of pre-determination of BCR repertoires and suggests that a specific genetic substructure underlies the diversity of BCR repertoires and can be used as a predictor of B cell evolution. In the light of our data, this substructure may in part be formed by IGHV gene preference for certain isotypes.

In the context of infectious diseases, relationship between antibody isotype and variable gene usage can lead to the establishment of isotype-specific response to an antigen and determine the success of pathogen neutralisation and generation of long-term immunity. Although vaccines should produce a polyclonal antibody response, subunit vaccines maybe more restricted in eliciting distinct antibody isotype profiles. The recent analysis of correlates of limited protection in the HIV vaccine trial RV144 by Chung et al. [272] showed a protective immune response is only present after generation of IgG3, but not IgG4 Abs and is independent of T cell cytotoxicity or antibody neutralisation properties. Vaccine-directed expansion of IgG3 over IgG4 antibodies to antigenic epitopes may require IGHV compatibility which may be difficult to achieve at sufficient frequency for a sustained humoral response if certain IGHV genes form the majority of the epitope specific response.

Vaccination to generate broadly neutralising antibody (bNAb) responses is an alternative strategy with potential utility for protection against antigenically variable pathogens such as HIV and influenza viruses. However, the abundance of such naturally occurring bNAbs in vivo is often low suggesting priming and expanding such a B cell response by vaccination may be difficult. The low abundance of some HIV bnAbs maybe in part explained by the frequency of the required IGHV switched to an appropriate isotype. Many HIV bNAbs isolated from infected individuals are of IgG1 isotype and use members of VH1 gene family.
BCR sequencing for characterisation of antigen-driven B cell evolution

The analyses of healthy repertoires shows VH1 genes have the lowest frequency of expression in an IgG1 context. The identification of glycan dependent CD4 binding bNAb using more abundant and universally switch IGVH3-21 and IGVL3-1 [273] indicates some epitopes may be more amenable than others for working with the flow of B cell evolution. Indeed, of 21 known HIV bNAb only 6 (28%) utilise the abundant IGVH3 and 10 (48%) utilise IGHJ6 a J gene more associated with IgE class switching (Figure III.9 lower panel). Simultaneous monitoring of the BCR and isotype composition in a response to an antigen may therefore provide a ore accurate assessment of B cell evolution as a surrogate for vaccination response.
Chapter IV

Characterisation of measles-associated immune suppression using B cell receptor sequencing
Characterisation of measles-associated immune suppression using B cell receptor sequencing

1 Introduction

Measles virus (MeV) is an enveloped virus with a single-stranded, non-segmented negative sense RNA genome. It is a member of Morbillivirus genus, Paramyxoviridae family and exhibits a narrow host tropism, causing disease exclusively in old- and new-world non-human primates (NHPs) and humans. MeV is transmitted via the respiratory route and has a basic reproductive number (R₀) of ≈12–16, making it one of the most contagious human viruses [274, 275]. Clinical symptoms of measles appear 9–19 days after infection and start with a prodromal phase of high fever and malaise followed by cough, coryza (inflammation of the mucosal surfaces in the nasopharynx) and conjunctivitis. Koplik’s spots appear on the buccal mucosa in a subset of individuals and precede the onset of the characteristic maculopapular rash which affects the ears, the face, the trunk and extremities [276]. MeV infection is usually self-limiting, and recovery is followed by lifelong immunity to measles[277]. Waning of immunity likely occurs with age but the effect on susceptibility has not been quantified. If present, immune waning is likely to affect vaccine-acquired serological immunity more dramatically, as a drop in antibody titers occurs faster after vaccination compared to natural infection[278].

In rare cases, severe measles-associated central nervous system (CNS) complications may develop: Acute disseminated encephalomyelitis (ADEM), measles inclusion body encephalitis (MIBE) or subacute sclerosing panencephalitis (SSPE). As a lymphotropic virus, MeV causes transient lymphopenia, associated with cytotoxic killing of infected cells [279] and bystander apoptosis [280]. Despite the resolution of lymphopenia shortly after disappearance of rash, measles - associated immune suppression could last months to years post infection, increasing the risk of secondary infections [270]. Vaccination against measles virus via the MMR vaccine has been a major factor in reducing childhood morbidity and mortality with ≈75% reduction in measles cases between 2000 and 2013 (estimated 15.6 million lives saved) [281]. The rise of anti-vaccination campaigns and the presence of non-vaccinating religious communities challenge the maintenance of herd immunity to measles. Despite the availability of a vaccine, outbreaks of measles, primarily in unvaccinated
1 Introduction

communities, have led to 134,200 measles deaths globally in 2015, making measles one of
the leading causes of infectious-disease associated childhood mortality [282–284].

1.1 Measles virus pathogenesis

MeV is a human-restricted virus without a known animal reservoir that is transmitted
efficiently by aerosol or respiratory droplets. Natural infection of non-human primates is not
observed likely due to the insufficient size of primate populations to sustain transmission.
Therefore, studies of experimentally infected macaques with wild-type strains of MeV or
ferrets infected with canine distemper virus (another member of *Morbillivirus* genus) are a
major source of the current knowledge of measles pathogenesis [285].

Measles pathogenesis proceeds in three main stages: 1) infection of immune cells in
the upper respiratory tract; 2) viremia and virus replication in lymphoid organs; 3) entry
to epithelial cells and transmission. The early immune cell tropism is associated with
the expression of signalling lymphocyte activation molecule family member 1 (SLAMF1,
also known as CD150) which has been identified as the primary receptor for MeV [286].
CD150 is a membrane protein present on subsets of thymocytes, dendritic cells (DCs),
haematopoietic stem cells (HSCs), macrophages, T and B cells. Macaque models of measles
virus infection identify CD11c+ myeloid cells, alveolar macrophages and dendritic cells
(DCs) as the primary targets of the virus and thus important for the initial establishment of
infection [274, 287, 288]. Entry of host cells occurs either through infection of CD150+
cells in alveolar space or via attachment to dendrites of DC-specific intercellular adhesion
molecule-3 grabbing non-integrin (DC-SIGN+) submucosal DCs.

Infected cells spread the virus systemically to the gastrointestinal tract, kidneys, liver and
skin. Virus dissemination is mediated primarily via cell-cell dendrites and minimal release of
cell-free virus [289, 290]. Primary (bone marrow and thymus), secondary (spleen, tonsils,
lymph nodes) and tertiary (e.g. bronchus-associated lymphoid tissue (BALT)) lymphoid

[^1]: http://www.who.int/mediacentre/factsheets/fs286/en/
Characterisation of measles-associated immune suppression using B cell receptor sequencing

tissues which are rich in CD150$^+$ lymphocytes act as major sites of MeV replication in vivo [291–293, 285]. Experimentally infected NHPs show high MeV replication in B cell follicles [291] and formation of multinucleated giant cells [294] or syncytia of fused B cells [287] in upper respiratory tract (URT). In addition to B cells, MeV replication is observed in both CD4$^+$ and CD8$^+$ CD150$^+$ memory T cells [279]. Replication in lymphatic tissue is efficient, and infected cells enter the circulation which leads to detection of MeV in peripheral blood mononuclear cells (PBMCs) around 7–9 days after infection [295, 296].

Extensive virus production in lymphoid tissues is followed by infection of lymphocytes and DCs in the skin and the epithelial submucosa. Infected lymphocytes and DCs transmit the virus to the neighbouring epithelial cells or keratinocytes via nectin-4 receptor expressed on the basolateral surface of epithelial cells [297, 298]. This stage of MeV pathogenesis is responsible for the appearance of Koplik spots and maculopapular rash [299]. Skin lesions are characterised with intracellular oedema, cell necrosis and mononuclear cell infiltration of the epidermal keratinocytes [300]. Host immune response is important for the development of rash and this disease manifestation is not observed in immunocompromised patients, despite the severe or often lethal course of measles [301].

1.2 Immunity to measles virus

1.2.1 Innate immunity

A characteristic feature of innate responses to RNA viruses is the production of Types I and III interferons (IFNs) by infected cells [302]. Induction of IFN constitutes the primary host immune defence mechanism that limits virus replication. In the context of MeV infection, the induction of IFNs is largely suppressed by the combined activity of MeV proteins P, C and V [303–305] which enables the virus to replicate and spread during the prodromal phase of the infection without any clinical symptoms. Despite the minimal Type I/III IFN induction and signalling, MeV infection activates the innate immune system via the stress response and the inflammasome pathways. IFN-independent induction of stress response occurs in infected
1 Introduction

antigen presenting cells upon virus interaction with DG-SIGN [306]. In vitro studies of myeloid cells [307] and transcriptomic analysis of PBMCs [308] demonstrate the assembly of the inflammasome complex and increased expression of the inflammasome products IL1β and IL18 following MeV infection. This mode of engagement of the innate immune response facilitates strong immune activation without establishment of IFN-mediated antiviral state that would limit virus replication. Therefore, the major role in the control of infection is played by the adaptive immune system.

1.2.2 Adaptive immunity

The prodromal phase of measles, which is characterised by fever, conjunctivitis, and maculopapular rash appears 10–14 days after infection and is coincident with the appearance of the adaptive immune response. Both MeV-specific antibodies and activated T cells are detectable in circulation at the time of the rash [309–312].

1.2.2.1 Humoral immunity

MeV-specific neutralising antibodies are elicited against each of the viral envelope glycoproteins but protection is conferred primarily via haemagglutinin (H) protein-specific antibodies [313, 314], with higher titers and higher avidity needed to prevent infection than to prevent disease (rash)[315, 316]. While important as a correlate of protection [317], the role of neutralising antibodies during infection is unclear as children with hypogammaglobinaemia (immunodeficiency resulting in very low level of IgG immunoglobulins) do not exhibit an atypical clinical course and excessive morbidity [318]. However, antibody-dependent cellular cytotoxicity in measles patients, is correlated with the end of viremia [319], and binding of antibody to infected cells in vitro downregulates intracellular viral replication [320, 321]. Therefore, despite the lack of established correlation of neutralising antibody with clearance of the virus, humoral immunity is likely to play a role in the control of virus replication.
Characterisation of measles-associated immune suppression using B cell receptor sequencing

1.2.2.2 Cellular immunity

The importance of cellular immunity in the clearance of MeV infection was implicated by the clinical observation of severe measles in children with cellular immunodeficiencies [318] and the prolonged viraemia in CD8\(^+\) T cell depleted macaques, independent of B cell function [322]. Biopsy of the maculopapular skin lesion shows infiltration of CD4\(^+\) and CD8\(^+\) T cells in areas of MeV-infected epithelial cells [322]. Plasma from MeV-infected individuals has elevated levels of soluble CD8, and β2 microglobulin as well as activated CD4\(^+\) T [323, 324]. Within a few days after the appearance of the rash, the viremia is cleared and MeV can no longer be recovered from PBMCs by co-cultivation. Numbers of activated CD8\(^+\) T cells in circulation and plasma levels of IFN-γ decrease rapidly after clearance of infectious virus. This is followed by a slower decline in CD4 which is detectable several weeks after resolution of rash [325–327]. A MeV-specific memory CD8\(^+\) T-lymphocyte pool is established after infection [328, 329], with specificity for several MV proteins [323, 330, 331].

In the absence of a robust cellular immune response rash is not observed and progressive infection of the lungs or the nervous system may result in a fatal giant cell pneumonia or inclusion body encephalitis [223, 332]. In rare instances, immunologically normal children infected at a young age can also develop persistent infection of the nervous system leading to SSPE [333]. Although infectious virus is no longer detectable after resolution of the rash, viral RNA can be detected in PBMCs, as well as in respiratory secretions and urine for several weeks after apparent recovery [334, 335]. This persistence of viral RNA in infected cells has been proposed as a mechanism for maintenance of a plasmablasts pool of B cells producing MeV-specific antibodies and thus establishing a life-long serological immunity to measles [336]. The mechanism by which MeV is eventually cleared from its many sites of replication and the role of immune-mediated clearance to immunosuppression are poorly understood.
1.3 Measles-associated immune-suppression

In addition to the strong immune activation and generation of lifelong immunity, MeV pathogenesis is associated with a profound immune suppression. This duality in the immune phenotype to measles is referred to as “measles paradox” [279]. Measles was the first disease demonstrated to increase susceptibility to other infections, and it is now recognised that the majority of measles deaths are due to secondary infections that occur as a consequence of immune suppression [337, 338]. A number of different potential indirect and direct mechanisms for MeV-induced immunosuppression have been proposed, including lymphopenia, suppression of lymphocyte proliferation, type 2 skewing of cytokine responses and a loss of immunological memory. Somewhat conflicting results from early in vitro studies and more recent macaque experiments maintain the controversy regarding the exact mechanisms underlying this immunosuppression and imply a likely multi-factorial aetiology.

1.3.1 Lymphopenia

A profound lymphopenia is observed during the acute phase of MeV infection and affects CD4+ T lymphocytes, CD8+ T lymphocytes, B lymphocytes, and monocytes [339–342]. MeV RNA is detectable in PBMCs from infected individuals for one week or more following the appearance of the rash [343]. Mononuclear cells isolated from infected individuals undergo spontaneous proliferation in vivo and are proposed as a source of target cells for productive MeV infection [342, 344]. Detection of viral RNA via in-situ hybridisation of PBMCs from infected macaques identified only 5-10% virus positive cells [345]. This questioned the direct viral killing as a sole source of lymphopenia and proposed that reduced lymphocyte counts may reflect cellular trafficking from the periphery into secondary lymphoid tissue [346, 325] or bystander apoptosis [280, 347].
1.3.2 Suppression of lymphocyte proliferation

Suppression of lymphocyte proliferation has been proposed as an alternative mechanism, underlying the prolonged immune suppression after resolution of lymphopenia. The generalised immune dysfunction is evidenced by the disappearance of tuberculin skin reaction [348] and some autoimmune phenotypes [349, 350]. Multiple studies suggest a direct effect of MeV on lymphocyte function. In the context of T cells, MeV suppresses T lymphocyte proliferation to mitogens [351] cytotoxic function [352], presentation of unprocessed protein antigens [353], IL-2 responses [354], and associated downstream signalling [355, 356]. In vitro infection of unstimulated B cells induce cell cycle arrest in G1 phase [357] and impairment of immunoglobulin production [352, 358]. Cell-cell interactions mediated through MeV glycoproteins expressed on host cell membranes has also been proposed as a mechanism to initiate a signalling that inhibits lymphocyte proliferation [359–361].

1.3.3 Altered cytokine profile

Global, long-term changes in cytokine secretion by immune cells during measles have been proposed as a third mechanism of measles immune suppression. Despite the strong lymphocyte activation during the acute phase of MeV infection [324, 344], a predominant Th2 CD4+ T-lymphocyte cytokine profile is observed for MeV-infected individuals during the convalescent stage of disease. This state is characterised by increased IL-4 and IL-6 levels, as well as low levels of IL-2 and IFN-γ in plasma and mitogen-stimulated PBMC supernatant fluid [362]. Th2 cytokine predominance after resolution of the rash creates an environment favouring B cell maturation and establishment of humoral memory. However, this could compromise immune activation and induction of type 1 responses required for control of new pathogens. This hypothesis of a cytokine origin of immune suppression was not confirmed in infection of cotton rats with an IL-4-producing recombinant MeV and lymphocyte proliferation was not compromised [363]. In addition, dendritic cells primed with different MeV strains promote both Th1 and mixed IFN-γ and IL-4 production of naive T lymphocytes [364] suggesting that the cytokine effect might not be universal. Therefore,
while a skewed cytokine profile is an immunologically plausible explanation for measles-induced immune suppression, the role of Th2 responses in the suppression of lymphocyte activation following measles remains controversial.

1.3.4 ‘Immunological amnesia’

The macaque model of measles using fluorescently-labelled MeV enables the in vivo monitoring of all stages of MeV pathogenesis and interaction with host immune system [279]. In this model lymphopenia coincides with the peak of viraemia and is associated with cytotoxic T cell infiltration in the lymph nodes. Killing of infected lymphocytes occurs without detectable caspase activity and apoptosis [279] thus challenging the hypothesis of a non-cytotoxic origin of lymphopenia. The observations of suppressed lymphocyte function in in vitro studies are also not confirmed in vivo where measles results in a dramatic expansion of MV-specific lymphocytes followed by resolution of viremia and lymphopenia [279, 344, 365]. The macaque model further demonstrates higher rates of infected lymphocytes than originally estimated from in vitro models, especially in secondary and tertiary lymphoid organs [287].

Together, the lack of substantial evidence of suppressed lymphoproliferative responses in vivo and the profound MeV-associated lymphopenia suggest that an altered composition of peripheral lymphocyte populations rather than impaired functionality is the likely cause of measles immune suppression [366]. A possible basis of immune restructuring is the infection and loss of a large proportion of B and T memory cells in the macaque model [279] which cannot be restored without re-exposure to the same antigen. This loss in previous immune memory, is masked by the generation of measles-specific memory which restores normal cell counts. Therefore, after the resolution of lymphopenia, MeV infected individuals have normal lymphocyte counts but a depleted memory pool to previously encountered pathogens thus are more likely to be subjected to secondary infections (Figure IV.1).
Characterisation of measles-associated immune suppression using B cell receptor sequencing

Figure IV.1: Proposed model of 'immunological amnesia'.
Before infection individuals have B naïve pools (shown in blue) produced by B cell development in the bone marrow and B memory pools (shown in red) resulting from the individual-specific exposure to different antigens. Following measles-associated lymphopenia, B memory pools are depleted and the recovery of the normal B cell counts results from expansion of measles-specific clones (shown in light green). The recovery of the full breadth of antigen reactivities requires re-exposure to the same antigens and during this period an individual is susceptible to pathogens for which immunity was present before measles. The estimated time required for recovery is up to 3 years based on the predicions of Mina et al [270]. The effect on the naïve compartment is not considered in this model.

This hypothesis of ‘immunological amnesia’ was proposed by de Vries et al. [279] following characterisation of MeV pathogenesis in the macaque. The effect of measles-associated immune suppression on incidence of secondary infections was further quantified by Mina et al. who demonstrated a significant increase in non-measles infectious disease burden up to three years post measles and speculated that this is likely the time required for immune memory responses to recover the full spectrum of antigen specificities [270].

Formal testing of these hypotheses, however, requires characterisation of the lymphocyte composition of individuals before and after infection with MeV and estimation of the degree of genetic re-structuring which persists beyond the period of lymphopenia. Traditional immune phenotyping using cell surface markers to stratify lymphocyte populations cannot distinguish between previously available and recently re-populated B and T cell memory pools. Immune receptor sequencing, however, enables the use of the BCR/TCR as a barcode
unique for each cell. As a result, B cells can be tracked before and after infection which can provide a detailed view of the lymphocyte dynamics associated with measles.

Therefore, in this chapter, I use the methods for profiling of BCR repertoires described in Chapter II and Chapter III to characterise the changes in diversity and lymphocyte composition in B memory and B naïve compartments of individuals sampled before and after measles. In order to monitor the patterns of immune suppression during a natural infection, I performed analysis of paired samples from unvaccinated individuals recruited in a prospective study during a measles outbreak in 2013. Using isotype-resolved BCR sequencing I tested if MeV infection can cause substantial restructuring in the B cell compartment detectable beyond the period of lymphopenia and how this change in immune composition can explain the prolonged measles-associated immune suppression [270]. To test directly the immune amnesia hypothesis I also use a ferret model of *Morbillivirus* infection [367] to confirm that lymphotropic challenge could deplete vaccine acquired immunity and lead to higher severity of secondary infections.

### 1.4 Aims

1. To characterise the effect of measles on the diversity of B naïve and B memory pools after resolution of lymphopenia.

2. To test if measles-associated lymphopenia can deplete previously acquired immunity and increase the severity of secondary exposure to previously encountered pathogens.

### 1.5 Colleagues

The work presented in this chapter was done in collaboration with Dr. Rik de Swart who provided all human samples from the measles cases and the vaccinated and uninfected control groups. The ferret experiment was funded and conducted by Dr. Veronika von Messling and
Dr. Bevan Sawatsky. The data on influenza-specific antibody titres was generated by Lisa Walz. The rest of the presented analyses are result of my own work unless stated otherwise.

2 Methods

2.1 Samples

Three study groups were ascertained for BCR repertoire analysis: measles-infected children, uninfected children and vaccinated adults. Children samples were part of a non-vaccinating orthodox protestant cohort recruited by Dr. Rik De Swart at Erasmus Medical Centre, Rotterdam. The collection of the samples was approved by the Medical Ethical Committee of the Erasmus Medical Centre under REC approval number MEC-2013-302 (NL45323.078.13) on 13th July 2013. Sequencing of the samples was approved by Sanger Institute Ethics committee under REC approval number 14/081 on 30th Oct 2014.

Children between seven and fourteen years of age, without previous history of measles infection or MMR vaccination and without any chronic disease or immune suppression were included in the cohort. Twenty children sampled before and after measles infection and three seronegative children sampled across two timepoints were used as measles disease and control groups respectively. Sampling timepoints and measles seroconversion titers are provided in Table IV.1. PBMC samples from seven adults sampled before and forty days after vaccination with trivalent inactivated influenza vaccine (TIV) were collected by Bri Laksono at Erasmus Medical Centre and used as vaccine controls. PBMCs were isolated from 1 or 3 µL of blood (children vs. adult cohort respectively) using Ficoll Density Centrifugation according to manufacturer’s protocol. Frozen PBMCs were prepared by the my collaborators at Erasmus Medical Centre.
## Table IV.1: Sampling strategy for measles cases and control cohorts

<table>
<thead>
<tr>
<th>Sample</th>
<th>Age</th>
<th>Timepoint 1 (T1)</th>
<th>Timepoint 2 (T2)</th>
<th>IgM (post)</th>
<th>IgG (pre)</th>
<th>IgG (post)</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-05</td>
<td>7</td>
<td>17</td>
<td>40</td>
<td>6.4</td>
<td>0.05</td>
<td>1.01</td>
<td>Measles</td>
</tr>
<tr>
<td>B-07</td>
<td>11</td>
<td>17</td>
<td>40</td>
<td>6.1</td>
<td>0.06</td>
<td>1.3</td>
<td>Measles</td>
</tr>
<tr>
<td>B-08</td>
<td>9</td>
<td>15</td>
<td>42</td>
<td>4.5</td>
<td>0.04</td>
<td>1.32</td>
<td>Measles</td>
</tr>
<tr>
<td>B-09</td>
<td>7</td>
<td>14</td>
<td>43</td>
<td>6.2</td>
<td>0.04</td>
<td>1.34</td>
<td>Measles</td>
</tr>
<tr>
<td>B-12</td>
<td>10</td>
<td>16</td>
<td>41</td>
<td>2.5</td>
<td>0.02</td>
<td>1.03</td>
<td>Measles</td>
</tr>
<tr>
<td>B-13</td>
<td>7</td>
<td>14</td>
<td>43</td>
<td>1.3</td>
<td>0.03</td>
<td>1.24</td>
<td>Measles</td>
</tr>
<tr>
<td>B-16</td>
<td>7</td>
<td>15</td>
<td>41</td>
<td>2.1</td>
<td>0.14</td>
<td>0.89</td>
<td>Measles</td>
</tr>
<tr>
<td>B-25</td>
<td>25</td>
<td>15</td>
<td>41</td>
<td>5.9</td>
<td>0.08</td>
<td>1.19</td>
<td>Measles</td>
</tr>
<tr>
<td>B-27</td>
<td>6</td>
<td>22</td>
<td>34</td>
<td>6.2</td>
<td>0.07</td>
<td>1.1</td>
<td>Measles</td>
</tr>
<tr>
<td>B-35</td>
<td>10</td>
<td>20</td>
<td>38</td>
<td>3.7</td>
<td>0.03</td>
<td>1.22</td>
<td>Measles</td>
</tr>
<tr>
<td>B-42</td>
<td>7</td>
<td>19</td>
<td>32</td>
<td>8.2</td>
<td>0.02</td>
<td>1.27</td>
<td>Measles</td>
</tr>
<tr>
<td>B-54</td>
<td>8</td>
<td>21</td>
<td>37</td>
<td>5.3</td>
<td>0.02</td>
<td>1.27</td>
<td>Measles</td>
</tr>
<tr>
<td>B-71</td>
<td>8</td>
<td>22</td>
<td>35</td>
<td>7.1</td>
<td>0.01</td>
<td>1.02</td>
<td>Measles</td>
</tr>
<tr>
<td>B-75</td>
<td>5</td>
<td>14</td>
<td>43</td>
<td>5.9</td>
<td>0.02</td>
<td>1.06</td>
<td>Measles</td>
</tr>
<tr>
<td>B-78</td>
<td>11</td>
<td>70</td>
<td>34</td>
<td>7.6</td>
<td>0.02</td>
<td>1.17</td>
<td>Measles</td>
</tr>
<tr>
<td>B-83</td>
<td>11</td>
<td>67</td>
<td>37</td>
<td>8.5</td>
<td>0.01</td>
<td>1.15</td>
<td>Measles</td>
</tr>
<tr>
<td>B-88</td>
<td>7</td>
<td>59</td>
<td>45</td>
<td>5.7</td>
<td>0.02</td>
<td>1.16</td>
<td>Measles</td>
</tr>
<tr>
<td>B-90</td>
<td>8</td>
<td>21</td>
<td>41</td>
<td>1.5</td>
<td>0.02</td>
<td>1.18</td>
<td>Measles</td>
</tr>
<tr>
<td>B-94</td>
<td>8</td>
<td>61</td>
<td>43</td>
<td>7.2</td>
<td>0.02</td>
<td>1.32</td>
<td>Measles</td>
</tr>
<tr>
<td>B-95</td>
<td>12</td>
<td>80</td>
<td>24</td>
<td>6.8</td>
<td>0.04</td>
<td>1.07</td>
<td>Measles</td>
</tr>
<tr>
<td>B-55</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>0.2</td>
<td>0.03</td>
<td>0.03</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B-56</td>
<td>6</td>
<td>0</td>
<td>56</td>
<td>0.5</td>
<td>0.01</td>
<td>0.02</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B-99</td>
<td>12</td>
<td>0</td>
<td>98</td>
<td>0.3</td>
<td>0.02</td>
<td>0.03</td>
<td>Uninfected</td>
</tr>
<tr>
<td>FLU1</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU2</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU3</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU4</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU5</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU6</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU7</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
<tr>
<td>FLU8</td>
<td>–</td>
<td>0</td>
<td>40</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Vaccinated</td>
</tr>
</tbody>
</table>

1. Days before rash (parent reported) or timepoint 1 for uninfected controls
2. Days after rash (parent reported) or days post timepoint 1 for uninfected controls.
3. Measles-specific IgM at T2 (Microimmune capture ELISA; Positive threshold: OD(sample)/OD(positive control) > 1.1)
4. Measles-specific IgG at T1 (Indirect ELISA; 1:300 serum dilution; OD450)
5. Measles-specific IgG at T2 (Indirect ELISA; 1:300 serum dilution; OD450)
2.2 FACS sorting and library preparation

Frozen PBMCs were thawed in 7µL R10/DNAse solution (RPMI media, 10% FBS, 10µL/500mL R10 DNAse inhibitor), pelleted at 2000rpm for 5’ at 4°C and washed in 4µL PBS twice to remove residual DMSO. After the second washing step, supernatant was removed and cells were incubated with antibody staining mix (see below) for 20 min. DAPI stain was added to each sample immediately before sorting. Antibody staining mix contained the following antibodies: CD19-FITC (BD, Cat: 555412), PeCy7-CD3 (BD, Cat: 557749), PE-CD45Ra (BD, Cat: 555489), AF700-CD45Ro (BD, Cat: 561136), APC-CD27 (BD, Cat: 558664), BV711-CD38 (BD, Cat: 563965). CYTO-COMPTM beads (Backman Coulter, Cat: 6607023). Samples were sorted using BD Influx™ machine into 5 cell populations: CD19+CD27− (B naïve cells); CD19+CD27+ (B memory cells); CD19+CD27+ CD38+ (plasmablasts); CD3+CD45Ra+ (T-naïve cells); CD3+CD45Ro+ (T memory cells). RNA from sorted B cell populations was used for amplification of IGH gene using the isotype-resolved BCR sequencing protocol described in Chapter III. RNA from T cell populations was extracted and returned back to Dr Rik de Swart.
2 Methods

Figure IV.2: Gating strategy for FACS sorting of B and T cell populations.
Example of gating strategy used for sorting of peripheral blood mononuclear cells (PBMCs). DAPI was used as a marker of cell viability. Viable cells were gated into CD3⁺ T cells and CD19⁺ B cells. CD3⁺ cells were further gated into CD45Ra⁺ T naïve cells and CD45Ro⁺ T memory cells. CD19⁺ B cells were sorted into CD27⁻ B naïve and CD27⁺ B memory cells. CD19⁺CD27⁺ B cells were used to gate the CD38⁺ plasmablasts. Gating strategy does not differentiate between CD19⁺CD27⁻ naïve and memory cells.

2.3 Human serology

MeV IgG titers were determined by an in-house method using indirect IgG ELISA. Each plasma sample was measured in a single dilution (1:300). For coating, MV HA protein (Bio Products Laboratory) was reconstituted in 0.5 ml distilled water, and diluted 1:100 in PBS, of which 100µL was added to each well. The plates were incubated at 4°C overnight. Plates were washed with PBS supplemented with 0.05% tween-20. Plasma- or serum samples were diluted in ELISA buffer to reach the desired dilution (1:300). 100µL of each diluted sample
Characterisation of measles-associated immune suppression using B cell receptor sequencing

was transferred to a well (in duplicate). The plates were incubated at 37 °C for 1 hour. Plates were washed with PBS supplemented with 0.05% tween-20. Goat-anti-human IgG-HRPO conjugate is diluted 1:5000 in ELISA buffer, and 100µL is added to each well. The plates were incubated at 37 °C for 1 hour. Plates were washed with PBS supplemented with 0.05% tween-20. Tetramethyl-benzidine (TMB) substrate was diluted 1:10 in TMB diluent (Meddens), and 100µL is added to each well. The plates were incubated at room temperature for 10 minutes. Reaction was stopped by addition of sulphuric acid (100µL/well). Plates were measured in an ELISA reader at wavelength 450nm (reference wavelength 620nm). The assay has no formal threshold for positivity. However, the fold increase in titer post-measles por each patient was considered as formal evidence for MeV infection.

The MeV IgM titers were derived using the commercial MicroImmune capture IgM assay EIA (Cat.No:MeVM010) according to manufacturer’s protocol. The results were expressed as a ratio of OD450(tested sample)/OD450(positive control). Values below 0.9 were considered negative. Values between 0.9 and 1.1 are equivocal.

2.4 Statistical analyses

Comparison of V gene, VJ gene, clonotype frequencies across samples was performed by calculating the V, VJ gene or clonotype frequency as percentage of the total repertoire for each individual from a measles-infected or control groups. The derived frequencies were then compared using two-sided non-parametric Wilcoxon ranked-signed test as it does not assume that data in each category is normally distributed. The test is based on three assumptions: 1) the dependent variable is measured in a continuous level (which is the case for V(J) gene and clonotype frequencies); 2) the categorical variable consists of matched pairs (which is the case in this dataset as the same individuals were sampled across 2 timepoints); 3) the distribution of differences between the two related groups is symmetrical (this is ensured by the representation of V(J) gene frequencies as percentage). Correlation in frequencies across timepoints was performed using Pearson’s correlation coefficient.
2.5 Network analysis

Network analysis was performed as described in Chapter III. In brief, error-corrected BCR sequences were represented as vertices and edges were drawn between any two vertices with 1nt difference within the full length of the BCR sequence (without considering indels).

2.6 Ferret model of immune suppression

2.6.1 Influenza vaccinations and infection with canine distemper virus

Eight male ferrets each were vaccinated with the tetravalent seasonal inactivated influenza vaccine (Influsplit Tetra 2016/2017; GlaxoSmithKline) and boosted four weeks later. Another group of eight male ferrets each were vaccinated with the tetravalent seasonal live-attenuated influenza vaccine (Fluenz Tetra; AstraZeneca) by intranasal administration at the same time as the boost in the other inactivated vaccine group. One group of four male ferrets was not vaccinated. Four weeks after boost with the inactivated vaccine or immunization with the live-attenuated vaccine, four ferrets each from the inactivated and live-attenuated vaccine groups were infected intranasally with $2 \times 10^5$ TCID$_{50}$ of recombinant canine distemper virus (CDV) that does not recognise the morbillivirus epithelial receptor nectin4 (nectin4-blind). Ten weeks after the CDV nectin4-blind infection, two of the animals from each treatment condition were infected intranasally with $5 \times 10^5$ TCID$_{50}$ of the virulent 2009 pandemic H1N1 influenza virus strain MX10. The remaining two animals from each treatment condition were challenged with Influenza A two weeks later.

2.6.2 Blood sampling

Whole blood was collected from ferrets under general anesthesia in two lithium-heparin tubes per animal. Total leukocyte counts were obtained by counting cells after mixing whole blood (10 µL) with 3% acetic acid (990 µL). Heparinised blood was centrifuged for 3,000 rpm for 15 min to collected plasma, and the remaining cells were mixed with 3
volumes of PBS and overlaid onto Histopaque-1077 cushions and centrifuged for 40 min at 400×g. Purified PBMCs were collected and washed three times with PBS. PBMCs were then resuspended in RMPI 1640 containing 10% FBS and 1% L-glutamine and counted. The purified PBMCs were then pelleted and resuspended in RNAlater before being frozen at -80°C. Major blood sampling times were as follows: day of vaccination (2016.10.07); day of boost/vaccination (2016.10.31); day of nectin4-blind CDV infection (2016.11.28); four weeks after nectin4-blind CDV infection (2016.12.27); eight weeks after nectin4-blind CDV infection (2017.01.23); and four weeks after influenza challenge (2017.03.06 or 2017.03.20).

For animals infected with nectin4-blind CDV, whole blood was collected once weekly after infection in Na2-EDTA tubes for virus titration. Total leukocyte counts were obtained as described above. Blood was centrifuged for 15 min at 3,000 rpm, and the plasma was harvested. The remaining cells were mixed with five volumes of Red-Blood Cell (RBC) Lysis Buffer based on the initial volume of blood (≈2 mL) and incubated for 10–15 min at room temperature. After RBC lysis, the cells were centrifuged and washed once with PBS, and then resuspended in DMEM containing 5% FBS and 1% L-glutamine. Titrations were performed by limiting dilution in 96-well plates and titers are expressed as TCID\textsubscript{50}/mL.

### 2.6.3 Influenza MX10 challenge

Ferrets were infected intranasally with 5 × 10\textsuperscript{5} TCID\textsubscript{50} of the virulent 2009 pandemic H1N1 influenza virus strain MX10. Nasal washes were performed once daily, and a clinical examination was performed twice daily for the first four days after infection, and once on day 7.

### 2.6.4 Titration of total antibodies

Determination of antibody titers was performed using immunoperoxidase monolayer assay (IPMA). MDCK cells were seeded into 96-well plates at a density of 1 × 10\textsuperscript{6} cells per plate. The cells were then infected with 10\textsuperscript{2} TCID\textsubscript{50} per well of pandemic influenza H1N1pdm09.
(A/California/07/2009) in DMEM without FBS. After 48 hrs incubation at 37°C, media were removed, and the cells were washed once with 150 µL per well of PBS diluted 1/3 in water. The wash was removed and the cells were air-dried for 10-15 min at room temperature. Cells were fixed by heated at 65°C for 8 hrs. Ferret plasma samples were initially diluted at 1:50 in PBS, and then titrated on fixed infected MDCK cells in duplicate in 2-fold dilution steps until 1:6400. Plasma dilutions were incubated at room temperature for 2 hrs. The plasma dilutions were then removed and the plates were washed once with 150 µL PBS per well for 15 min at room temperature. Ferret antibodies bound to influenza-infected cells were detected by incubation with a rabbit goat anti-ferret secondary antibody coupled to horseradish peroxidase (HRP) at a dilution of 1:750 in 50 µL PBS per well for 1 hr at room temperature. The secondary antibody was then removed and the plates were washed once with 150 µL PBS per well at 15 min at room temperature. Assays were developed by staining with 3-amino-9-ethyl-carbazole (AEC) at room temperature. Reactions were stopped by removal of the substrate, followed by the addition of 100 µL of water per well. Assays were then read the next day.

2.6.5 Determination of neutralizing antibody titers

Ferret plasma samples were diluted in 50 µL per well of serum-free DMEM in 96-well plates as described above, starting at a dilution of 1:10 and continuing until 1:1280. Plasma dilutions were mixed with 50 µL per well of serum-free DMEM containing $10^2$ TCID$_{50}$ of influenza virus. The plasma-virus mixtures were incubated at 37°C for 20 min to allow complexes to form, after which they were added to MDCK cells. Neutralisation assays were incubated at 37°C for 48 hrs, and then fixed and stain by IPMA as described above. The primary antibody for neutralisation assays was a polyclonal ferret anti-H1N1pdm09 immune serum at a dilution of 1:500 in 50 µL PBS per well, followed by a goat anti-ferret-HRP secondary antibody at a dilution of 1:750.
3 Results

3.1 Effect of measles on lymphocyte population structure

Measles-associated lymphopenia in the majority of measles cases is resolved several days after disappearance of rash [279]. Since the measles cases used in this study were sampled on average 45 days after rash, full recovery of their B and T cell counts is expected to have occurred. To test if any effects of lymphopenia are observed at this late timepoint, the changes in lymphocyte population structure before and after measles were characterised using FACS sorting. An average of 392,658 live PBMCs were sorted for each paired sample using the gating strategy explained in ‘Methods’ Section (Figure IV.3). The frequency of each cellular compartment as a proportion of total sorted cells was monitored across timepoints.

The CD3$^{+}$CD45Ra$^{+}$ T naïve population was present at highest frequency across samples (mean frequency – 54%) while CD19$^{+}$CD27$^{+}$ was the least common lymphocyte subset constituting on average ≈3.9% of the lymphocytes sorted from a given PBMC sample. The adult cohort exhibited significant difference in the relative contribution of CD19$^{+}$CD27$^{-}$ B cells and the two T cells compartments compared to the measles cases (p-value<0.001) (Figure IV.2). This is consistent with a previously described patterns of age-associated changes in lymphocyte composition [368]. A significant change in cell frequency across timepoints was not observed for the measles-infected or any of the control cohorts suggesting a highly stable lymphocyte composition in uninfected and vaccinated individuals and a recovery of B and T cell counts post lymphopenia in measles cases (Figure IV.3).
3 Results

Figure IV.3: Lymphocyte frequencies across sampling timepoints. Cell frequencies as percentage of total sorted cells per sample of sorted B- and T-lymphocytes from peripheral blood mononuclear cells (PBMCs) across sample cohorts: measles (yellow), uninfected controls (blue), vaccinated controls (green). Timepoints 1 and 2 represent PBMC samples derived before and after measles virus infection or vaccination respectively. Significance for individual samples across timepoints (shown in yellow) – Wilcoxon signed rank test, paired, (p-value=0.03 for CD19\(^{+}\)CD27\(^{-}\) and CD19\(^{+}\)CD27\(^{+}\)). "***" – p-value<0.001, "**" – p-value<0.01 (not significant after correction for multiple testing).

3.2 Changes in B cell repertoire diversity following measles

As expected measles-infected individuals showed a complete recovery of the cellular composition of the B cell compartment ≈45 days following measles. RNA from the sorted B cell
Characterisation of measles-associated immune suppression using B cell receptor sequencing

populations was then used for BCR receptor sequencing to investigate if the reconstitution of B cell frequencies post infection results also in recovery of BCR diversity. B cell repertoires from all samples were derived using the isotype-resolved BCR sequencing approach described in Chapter III. All samples were sequenced at equimolar concentrations to avoid high variation in sequencing depth (For a full list of QC metrics for each sample, see Table A.1). No significant differences (p-value>0.05) were observed in the number of joined reads and number of unique BCR sequences across measles infected cases and uninfected controls and across timepoints.

3.3 Dynamics of the naïve B cell compartment following measles virus infection

3.3.1 Incomplete reconstitution of naïve cell compartment in a subset of measles cases

Antigen inexperienced (naïve) B cells express CD150 and are also depleted during measles-associated lymphopenia [279]. Reconstitution of this B cell subset in peripheral blood requires re-population from progenitor cells in the bone marrow via homeostatic proliferation. Therefore, if the antigen-independent stage of B cell development in bone marrow is not impaired, the recovery of the naïve B cell counts should result in a recovery of the BCR diversity in this population. To test this, BCR repertoires were derived from the CD19+ CD27− B cell subset before and after MeV infection. The rate of somatic mutation in variable genes and the isotype identity were used to select all unmutated IgD+ and/or IgM+ BCRs which represent the true naïve B cells, unexposed to antigenic stimulation [233].

The frequencies V gene usage profiles were highly correlated across measles and control samples (mean $R^2 = 0.8$; mean p-value=$1.06 \times 10^{-6}$) consistent with the expected high overlap of genetic composition in the antigen inexperienced repertoire across individuals. (Figure IV.4.T1). Following measles infection, two individuals (B-35, B-42) exhibited a dramatic shift in the genetic composition of their naïve compartment leading to a complete loss of correlation in V gene profile with the rest of the measles cases as well as seronegative
and vaccinated controls (mean co-cluster p-value - 0.0648) (Figure IV.4.T2). The rest of the measles-infected individuals exhibited successful recovery of their naïve compartment and maintained high similarity of their V gene profiles (mean co-cluster p-value without B-35, B-42 - $1.401057 \times 10^{-08}$, $R^2 = 0.8$).

**Figure IV.4: Correlation of V gene frequencies of B naïve cells across individuals.**
Correlation between V gene frequencies (as percentage of the total sequenced repertoire for each individual) across MeV infected, uninfected and vaccinated control individuals before (T1) and after (T2) measles infection/vaccination (or across timepoints for uninfected controls). Pie charts represent the correlation coefficient as a proportion of total correlation of 1. Colour legend corresponds to the values of Pearson’s ranked correlation. Blank squares represent sample pairs with non-significant correlation (p-value $>0.05$/number V genes). Coloured rectangles on the right of each correlation plot represent the different cohort groups: measles cases (yellow), uninfected controls (blue), vaccinated controls (green).

Impairment of B cell reconstitution could result either in reduced diversity of the pre-immune B cell repertoire or skewed representation of individual V genes. To assess the specific genetic signature underlying the observed shift in genetic composition of two of the infected measles cases, I compared the variable gene composition of the naïve repertoires of all individuals before and after infection (Figure IV.5). Consistent with previous reports of healthy adult repertoires [369, 370], IGHV4-34 and IGHV3-23 genes dominated the naïve repertoires of both children and adult groups before infection with average frequency of 12%.
and 10% respectively. These variable gene profiles suggest genetic structure of the B-naïve compartment of all individuals reflects the expected profile of a healthy immune status.

**Figure IV.5: Variable gene frequencies in the B naïve compartment across individuals.**
Variable gene frequencies (as percentage of the total BCR repertoire) of IgD and/or IgM unmutated B19⁺ B27⁻ B cell populations of measles and control individuals across the two sampling timepoints (T1 and T2). V genes are ordered according to their location along the IGH locus (5’-3’ top to bottom). Coloured rectangles under each heatmap represent the different cohort groups: measles cases (yellow), uninfected controls (blue), vaccinated controls (green). The colour key represents the z-score calculated for each column by subtracting the mean of all V gene frequencies per individual from each individual V gene frequency data point and then dividing those points by the standard deviation of all points.

The healthy profile was successfully recovered after measles for the majority of infected individuals with the exception of two individuals: B-35 and B-42. Immune reconstitution of the naïve compartment was incomplete in these individuals which was evidenced by a collapse in BCR diversity with loss of ≈80% of V gene families. As a result, the naïve B cells of these two individuals had lost the V gene signature characteristic of a healthy pre-immune status and exhibited an overexpression of single V genes (IGHV1-2 and IGHV2-5 for B35 and B42 respectively). The naïve compartment of individual B-35 was highly clonal.
with IGHV1-2 constituting 56% of the post measles repertoire and the previously dominant IGHV4-34 and IGHV3-23 genes now reduced to $\approx 0.04\%$ of naïve repertoire. For individual B-42 expression of genes IGHV4-34 and IGHV3-23 was undetectable after infection and the B- naïve repertoire was dominated by IGHV2-5 (18%), IGHV3-9 (15%), IGHV1-8 (15%) and IGHV3-11 (14%). Interestingly, the observed V gene reconstitution pattern in both individuals was biased toward overexpression of genes proximal to the V-DJ recombination site. This suggests a spatial effect of V gene diversity generation following depletion which is consistent with previously reported preference for recombination of V genes closer to the DJ regions during B cell development in the bone marrow [371].

This dramatic loss in diversity could not be explained by any technical limitations in processing of these two samples as they exhibited higher than the median input cell counts compared to the rest of the individuals (98,748 and 65,600 cells compared to median values of 36,518 and 58,868 for B35 and B42 respectively) and had higher number of derived unique BCR sequences than 7 other individuals with normal V gene profile (For detailed information of sample QC metrics, See Table A.1). The disease manifestation of individual B-35 has reported as “severe” by the parents but both individuals generated similar levels of anti-measles IgG antibodies (See “Methods”). The individuals were not related and lived in two different villages.

### 3.3.2 Restricted V gene usage as a source of reduced diversity in the naïve compartment

Increased clonality in the antigen-inexperienced repertoire can result either from aberrant VDJ recombination during B cell development leading to the production of a small number of VDJ combinations, or from an abnormal expansion of a single BCR clone as observed in B cell malignancy. To test if the observed collapse in diversity in the subset of measles cases is driven by an overrepresentation of a single BCR clone or from restriction of produced VDJ clones, I compared the frequencies of individual VJ combinations before and after infection to test if there was a clonal expansion of a single VJ gene combination. Uninfected
Characterisation of measles-associated immune suppression using B cell receptor sequencing

individuals and vaccinated controls exhibited highly stable VJ gene frequency profile with correlation coefficient ($R^2$) of 0.94 and 0.96 respectively (Figure IV.6). By contrast, measles cases exhibited a shift in VJ gene profile leading to a reduced correlation across timepoints ($R^2 = 0.6$). The observed increase in frequency of a subset of the VJ gene combinations after measles was primarily driven by the contraction of the naïve repertoire of individuals B-35 and B-42) which increased the relative contribution of these gene combinations to the total repertoire. The naïve compartment of individual B-35 was restricted to only 30 VJ gene combinations with IGHV1-2|IGHJ4 BCRs constituting $\approx 25\%$ of the repertoire post measles compared to 3.6% before infection. Similarly, individual B-42 had only 54 VJ gene combinations with the top 5 most common V genes (IGHV2-5, IGHV3-9, IGHV1-8, IGHV3-11, IGHV3-7) all recombined with IGHJ4 and constituting 36% of the naïve B cell repertoire. The average frequency of the most common VJ genes in control groups was 7.4% for uninfected and 5% for vaccinated individuals.

**Figure IV.6: Correlation of VJ gene frequencies of B naïve cells across timepoints.**
Correlation of VJ gene frequencies across timepoints for uninfected controls and measles cases. VJ gene frequency is shown as % of total repertoire per sample. Lower panels represent $\log_{10}$ -transformed values of frequency fold change for individual VJ gene combinations. Colours of each dot correspond to a specific sample as shown in the colour legends. $R^2$ values represent the results of Spearman ranked coefficient testing. Only the VJ gene combinations for individuals B-35 and B-42 are coloured in the 'Measles' panel to highlight their effect on the reduced correlation across timepoints. The VJ gene combinations of all remaining measles cases are shown in grey.
The overrepresentation of IGHV1-2|IGHJ4 BCR in subject B-35 was not driven by a clonal expansion of a single BCR sequence, but was represented by 650 distinct BCRs with 23 different CDR3 lengths varying between 7 and 31 amino acids (mean - 17 amino acids). By comparison, the pre-measles repertoire of the same individual this VJ gene combination was represented only by 155 BCR sequences distributed over 15 amino acid lengths but exhibited the same average amino acid length (17 amino acids). Network analysis [269] on BCR clonal identity for both individuals B-35 and B-42 further confirmed the lack of expansion of a single BCR and the presence of distinct BCR clones, despite the use of a limited number of VJ gene combinations (Figure IV.7). Together, these suggest that the high clonality observed in the naïve repertoires of individuals B-35 and B-42 does not a result of an overrepresentation of a single clone but reflects a restriction in the used V gene pool.

Figure IV.7: BCR diversity of naïve repertoires of individuals B-35 and B-42. Network properties [269] of CD19^+CD27^- B cell receptor repertoires of individuals B-35 and B-42 post measles. Each dot represents a unique BCR sequence. Each vertex connects two BCRs with 1 nt difference anywhere along the BCR sequence. The percentage of vertices shown represents how many of the total number of sequences are shown in the diagram after removal of the singleton reads for clarity.
3.4 Dynamics of the memory compartment B cell repertoire following measles virus infection

After characterisation of the effect of measles infection on the naïve compartment, I performed similar analyses on the B memory compartment to investigate the effect of measles infection on previously acquired immunity. Unlike the naïve repertoire which requires efficient bone marrow function for reconstitution, the memory repertoire is shaped by the exposure history of an individual and represents a combination of clonal expansion events reflecting previous responses to different antigens. Therefore, the inter-individual variation in repertoire diversity is expected to be higher than in the naïve compartment.

3.5 Changes in isotype composition of the B memory compartment post measles

Measles infection leads to development of measles-specific IgM and IgG antibodies which confer serological protection from re-infection. Such seroconversion was observed for all measles cases in the studied cohort (See Table IV.1). Differential expression of IgG subclasses has been reported in the convalescent phase of infection compared to late serological response, which is dominated by IgG1 and IgG4 subclasses [372]. Therefore, I first tested if expansion in IgM of IgG isotypes is also observed in the B memory compartment of measles-infected individuals. Isotype profiles were calculated for each individual based on the frequency of BCRs with a given isotype. The most common isotype classes before measles were IgM and IgA1 with average frequencies of 36% and 32% of the BCR repertoires across individuals. Significant increase in frequency was observed only for IgG1 subclass (p-value: 0.03, Wilcoxon signed-rank test) after measles, but not in control individuals. The frequencies of the rest of the antibody classes remained consistent across timepoints, suggesting a lack of substantial changes in isotype composition of the B memory compartment following measles.
3 Results

![Graph](image)

**Figure IV.8: Changes in isotype composition of B memory cells post measles**
The frequency of individual isotypes was calculated as percentage of reads with the respective isotype per sample. Uninfected and vaccinated controls are grouped together in the "Controls" panel. "*"-denotes p-value < 0.05, Wilcoxon signed-rank test

### 3.6 Convergence of variable gene usage across individuals post measles

Next, I tested the effect of measles on the variable gene usage across individuals. The B memory repertoires of all individuals before measles were dominated by 5 most highly expressed V genes (IGHV3-23, IGHV1-18, IGHV1-2, IGHV4-34, IGH4-59) genes which together constituted between 27% and 54% of the V gene repertoires across samples (Figure IV.9). The degree of clonality of V gene usage varied across individuals with maximum V gene frequency between 9% and 21% of the BCR repertoire. The highest clonality of V gene usage was observed for individual B-25 where IGHV1-2 gene constituted 21% of the B memory repertoire before infection. Interestingly, the IGHV4-34 gene which was previously reported to undergo negative selection in the memory compartment of healthy adults [373], had an average frequency of ≈6% across individuals in this cohort and is the most highly expressed gene in 3 of the children reaching maximum frequency of 18% (B-27) before infection. Measles infection did not result in significant shifts in V gene usage in any of
the individuals (p-value > 0.05, Wilcoxon signed rank test). Despite the development of high clonal V gene repertoire in the naïve compartment, individuals B-35 and B-42 did not exhibit the same clonality in their memory pool. Post-measles repertoires of all individuals were dominated by the same top 5 V genes (IGHV3-23, IGHV1-18, IGHV4-34, IGHV1-2, IGHV4-59) with similar average repertoire contribution across samples (between 29% and 52% of repertoire). The maximum V gene frequency per sample varied between 7% and 18% of the repertoire across individuals.

Figure IV.9: Variable gene frequencies in the B memory compartment across individuals.
Variable gene frequencies (as percentage of total BCR repertoire) of B19\(^+\) B27\(^+\) B cell populations across measles and control individuals across the two sampling timepoints (T1 and T2). V genes are ordered according to their location along the IGH locus (5'–3' top to bottom). Coloured rectangles under each heatmap represent the different cohort groups: measles cases (yellow), uninfected controls (blue), vaccinated controls (green). The colour key represents the z-score calculated for each column by subtracting the mean of all V gene frequencies per individual from each individual V gene frequency data point and then dividing those points by the standard deviation of all points.
3.6.1 Convergence of variable gene usage across individuals post measles

Next, I compared the degree of correlation of V gene frequencies across samples from the measles-infected and control cohorts. The mean correlation coefficient across samples before infection was $R^2 = 0.66$ with 25 pairs of individuals exhibiting non-significant correlation of their V gene profiles (p-value > 0.001) (??). The mean co-cluster correlation between all samples was also non-significant after correction for multiple testing across all V genes (p-value=0.0012), suggesting a high individual-specific variation of V gene frequencies in resting B memory pool.

![Figure IV.10: Correlation of V gene frequencies of B memory cells across samples.](image)

Correlation between V gene frequencies (as a percentage of total BCR repertoire per individual) across measles-infected and uninfected control individuals before (T1) and after (T2) measles infection (or across timepoints for uninfected controls). Colour legend corresponds to the values of Spearman ranked correlation. Pie chats represent the correlation coefficient as a proportion of total correlation of 1. Blank squares represent sample pairs with correlation below the significance threshold. (p-value > 0.05/(Number of V genes for each time point).

Measles infection lead to removal of a large proportion of intra-individual variation and to increased sharing of the V gene structure of the memory repertoire ($R^2 = 0.74$; p-value = 1.7998 \times 10^{-5}) V gene profile was shared not only between measles cases but also exhibited increased correlation with uninfected controls (mean $R^2 = 0.76$ after measles
compared to 0.67 before measles). Individuals who were poorly correlated with B-55 and B-99 seronegative samples before infection (B-05, B-25, B-27, B-71 B-83, B-88) now exhibited significant convergence of V gene profile (p-value<0.001). Individual B-95 which was sampled earliest after measles (day 24, compared to median sampling point of 41 days post infection) showed the lowest convergence to the rest of the individuals post infection with average $R^2 = 0.58$. Similar shift in repertoire overlap was observed on the level of VJ gene combinations where measles infection lead to significant correlation in expression profiles across individuals (p-value= $6.05e^{-13}$).

3.7 Loss of clonally expanded B cell populations as a source of convergence between memory repertoires post measles

The convergence in memory repertoires across individuals was reflected in the significant change in VJ gene profile of 5 of the measles cases after infection (p-values: B-25 – 0.02, B-27 – 0.01, B-83 – 1.5 \times 10^{-5}, B-95 – 7.6 \times 10^{-5}; Wilcoxon signed-rank test). The shift in V gene profile was largely driven by a depletion of the most commonly expressed V gene before infection (between 0.7 and 5-fold reduction across individuals) (Figure IV.11). The observed depletion was not focused on specific V gene families and affected highly expressed genes, irrespective of their gene identity. More than 5-fold depletion of the highly expressed V genes was also observed in individual B-95, but the most frequent V gene family V1-18 was not affected. High V gene loss in individuals B-25 and B-42 was also observed for V genes present at low frequency before infection, likely representing a limitation of the sampling of rare clones.
Figure IV.11: Fold reduction in V gene frequency post measles.
Fold change of V gene frequency in B memory repertoires of individuals with a significant shift in repertoire convergence and VJ gene profile post measles. The x axes of each panel represent the frequency of individual V genes before infection as percentage from total V gene repertoire per sample. The y axes represent the fold change (Frequency before/Frequency after) in frequency of the respective V genes.

The high abundance of IGHV1-2 gene in individual B-25 was reduced 7-fold and this V gene constituted only 3% of the post-measles repertoire. Similar effect was observed for the highly expressed V4-34 gene in individual B-27 which was reduced to ≈3% post infection. For both individuals, the genes that were dramatically depleted were associated with clonal expansion events before measles and were represented by a single expanded clonotype (defined as VJ gene combination plus CDR3 amino acid length) - IGHV1-2|IGHJ4|CDR3-13aa for individual B-25 and IGHV4-34|IGHJ6|CDR3-16aa, suggesting that loss of individual-
specific clonally expanded populations is the likely reason for shift in VJ gene profiles and increased repertoire convergence across individuals post measles.

3.8 Clonotype reconstruction post measles

Removal of clonal expansion events representing individual’s specific history of antigen exposure would lead to a form of immune re-setting and reduced inter-individual variation. This was demonstrated by the increase in convergence across individuals observed on variable gene level (Figure IV.9) and appeared to result largely from loss of individual clonal expansions (Figure IV.11). Therefore, I next tested if changes in frequency of individual clonotypes (defined as V+J gene identity, plus CDR3 amino acid length) are indeed the reason for the increased overlap in V gene usage across individuals. Clonotype composition of uninfected and vaccinated controls remained highly stable across timepoints and without significant changes in frequency of any clonotype (p-value > 0.05). By contrast, measles cases showed significant change (p-values < 0.05) in frequency of 18 clonotypes associated with 15 different V genes (Figure IV.12). The shift in clonotype abundance was largely driven by clonotype loss after measles (16/18 clonotypes) with increase in frequency of only clonotypes IGHV3-23|IGHJ4|CDR3-18 and IGHV3-73|IGHJ4|CDR3-16. Interestingly, clonotypes with the same VJ gene and CDR3 length composition have previously been reported for specificity to influenza hemagglutinin suggesting of a possible depletion of influenza-specific memory clones. However, direct demonstration of the antigen-specificity of the depleted clones cannot be achieved with the available data.
3 Results

Figure IV.12: Restructuring of the B memory clonotype composition post measles.
B cell clonotypes exhibiting significant change in frequency post measles. Frequency (as proportion of total clonotype repertoire) before (1.PRE) and after (2.POST) measles was compared using Wilcoxon ranked test. Correction for multiple testing was not performed. p-values for each clonotype are shown in the table. Previous association of a clonotype with influenza specificity is designated with ‘Y’.

<table>
<thead>
<tr>
<th>Clones</th>
<th>P-value</th>
<th>Influenza Abs</th>
</tr>
</thead>
<tbody>
<tr>
<td>VH1-18, IGH6.20</td>
<td>0.00068</td>
<td>Y</td>
</tr>
<tr>
<td>VH1-21, IGH6.20</td>
<td>0.00064</td>
<td>Y</td>
</tr>
<tr>
<td>VH2-16, IGH6.20</td>
<td>0.00076</td>
<td>Y</td>
</tr>
<tr>
<td>VH2-25, IGH6.20</td>
<td>0.00074</td>
<td>Y</td>
</tr>
<tr>
<td>VH1-6, IGH6.13</td>
<td>0.00165</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-18, IGH6.18</td>
<td>0.00050</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-18, IGH6.16</td>
<td>0.00056</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-16, IGH6.16</td>
<td>0.00186</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-11, IGH6.11</td>
<td>0.00011</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-39, IGH6.16</td>
<td>0.00176</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-29, IGH6.16</td>
<td>0.00015</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-39, IGH6.16</td>
<td>0.00157</td>
<td>Y</td>
</tr>
<tr>
<td>VH6-1, IGH6.20</td>
<td>0.00183</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-9, IGH6.13</td>
<td>0.00090</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-9, IGH6.17</td>
<td>0.00042</td>
<td>Y</td>
</tr>
<tr>
<td>VH1-4, IGH6.16</td>
<td>0.00067</td>
<td>Y</td>
</tr>
<tr>
<td>VH4-39, IGH6.13</td>
<td>0.0013</td>
<td>Y</td>
</tr>
</tbody>
</table>

3.9 Maintenance of antigen-specific clones in the B cell memory compartment post measles

Since BCR sequences corresponding to the same clonotypes do not always reflect identical antigen specificity, I focused further on B clones with the same V, D, J gene annotation and identical CDR3 amino acid sequence as a way to monitor the maintenance of antigen-specific clones present before infection. I followed the degree of repertoire overlap for each individual and the changes in relative contribution of the overlapping clones to the overall repertoire. Considering the consistent sampling strategy and sensitivity of repertoire capture across timepoints, any changes in relative frequency of the shared clones should be a result of restructuring in the memory compartment affecting the relative contribution of a given clonotype to the total diversity of the BCR repertoire.

Seronegative and vaccinated controls maintained highly correlated frequencies of shared repertoires across timepoints ($R^2 = 0.76$) with the exception of individual ‘FLU6’ where the
Characterisation of measles-associated immune suppression using B cell receptor sequencing

B cell clonotypes present before vaccination had increased 3-fold 40 days post vaccination. This was an effect driven by a single clonally expanded clonotype (V6-1|D3-3|J4) which constituted ≈10 percent of the total BCR repertoire, likely represent a recent recall response (Figure IV.13). Interestingly, this individual also had an abundance of non-lymphocyte, non-granulocyte population during FACS sorting (data not shown) which could represent mast cells and an ongoing allergic/autoimmune response. No clinical information is available to confirm this.

Measles cases exhibited lower clonotype overlap across timepoints compared to control cohorts (median 21 clonotypes for measles compared to 53 for controls) and low correlation in frequency of shared clonotypes ($R^2 = 0.35$). The majority of individuals had a reduction in relative abundance of clones present before infection, likely due to a loss of these clones and repopulation of B memory pool with new clones or clones present at low frequency before infection.
Figure IV.13: Clonal dynamics in the B memory compartment across timepoints.
Changes in frequency of BCR clonotypes present across sampling timepoints for seronegative and vaccinated controls and measles virus infected cases. Clonotypes were defined as BCR sequences with the same V, D, J gene assignment and identical CDR3 amino acid sequence. Upper panels represent all individuals from each group; lower panel has FLU6 and B-35 removed to show the lower frequency clonotypes. The size of the dots represents the total number of shared clonotypes for any given sample across timepoints. The colours represent the respective control groups or the parent reported severity of measles disease.

Indeed, individual B-35 which had the highest degree of memory clonotype overlap across timepoints (142 clones) had up to 4-fold reduction in frequency of >80% of overlapping clones and increased representation of clones present at low frequency before infection (Figure IV.14). Interestingly, the small subset of clones that were increased in frequency post infection, corresponded to V genes proximal to V-DJ recombination site (V1-2, V6-1, V3-7,
Characterisation of measles-associated immune suppression using B cell receptor sequencing

V4-4). The most clonotypes that were maintained before and after infection in this individual were associated with V1-2 gene expression, which is consistent with the overrepresentation of this gene in the naïve compartment of the same individual.

![Clonal dynamics in the B memory compartment of individual B35.](image)

Figure IV.14: Clonal dynamics in the B memory compartment of individual B35.

Frequency of overlapping BCR clonotypes before (T1, blue) and after (T2, orange) measles infection in individual B35. Clonotypes were defined as BCR sequences with the same V, D, J gene assignment and identical CDR3 amino acid sequence. Each point represents a unique BCR clone. The V gene annotation represents the V gene identity of the respective clones. Clone frequency is represented as percentage of total BCR repertoire; \( \log_{10} \) - transformed values are shown.

3.10 Immunological amnesia in a ferret model of measles

The characterisation of the genetic composition of B cell naïve and memory compartments using BCR sequencing outlined two possible modes of immune restructuring following measles: restriction in the diversity of the naïve compartment which could lead to limited ability to build novel immune responses and depletion of previously available memory clones which acts as immune re-setting and loss of individual-specific diversity.

To address how these changes in B memory and naïve compartments can contribute to the phenotype of immune suppression and susceptibility to secondary infections, I used a ferret model of measles in collaboration with Dr. Veronika von Messling and Dr Bevan Sawatsky at Paul Ehrlich Institute, Germany. Morbilliviruses share considerable structural
and functional similarities. Even though disease severity varies among the respective host species, the underlying pathogenesis and the clinical signs are comparable [374]. The canine distemper virus (CDV) causes severe and often lethal disease in dogs and ferrets and has been demonstrated as a suitable model for characterisation of morbillivirus pathogenesis and associated immune-suppression [375, 376].

We designed a study to test the effect of lymphotropic challenge on vaccination-induced immunity to influenza and the changes in susceptibility to secondary challenge following virus-associated lymphopenia.

3.10.1 Effect of a lymphotropic challenge on the serological immunity to influenza virus

The effect of lymphotropic challenge on previously acquired immunity was first tested by measuring influenza-neutralising antibody titers generated after vaccination with live-attenuated (LAIV) and tetravalent inactivated influenza vaccines (TIIV). LAIV vaccine generated the highest titer of influenza-specific antibodies which were maintained 12 weeks post vaccination (Figure IV.15). Animals subjected to lymphotropic challenge with canine distemper virus (CDV) following had a significant reduction in antibody titers compared to unchallenged controls for both vaccine group. Since serological immunity is maintained by antibody-secreting plasma cells, this reduction in serum titers suggests an effect of CDV on the plasma cell pool. Upon challenge with Influenza A (MX10 strain), irrespective of the CDV challenge, TIIV vaccinated groups managed to generate similar levels of influenza-neutralising antibody titers higher than unvaccinated controls. By contrast, in the LAIV vaccinated group, CDV infection dramatically impaired the ability of the ferrets to generate protective serological immunity to influenza and neutralising antibody titers remained lower than for the unvaccinated controls.
Characterisation of measles-associated immune suppression using B cell receptor sequencing

Figure IV.15: Effect of a lymphotropic challenge on vaccine-induced serological immunity to influenza virus.

a. Dynamics of Influenza A-neutralising antibody titers following single dose of live-attenuated (LAIV) or two-doses of trivalent inactivated (TIIV) influenza vaccine with or without lymphotropic challenge with canine distemper virus (CDV). b. Timepoints of TIIV prime and boost vaccination are at week 0 and week 4, respectively. The single dose of LAIV is given at Week 4. CDV Challenge for both vaccine groups is at week 8 followed by Influenza A challenge (MX10 strain) at Week 16. Measurements of MeV neutralising antibodies were taken at each timepoint. The ‘No vaccine’ group represents unvaccinated controls subject only to a final challenge with influenza virus without CDV infection or vaccination. Significance of the difference in antibody titers was calculated using Mann-Witney test for each timepoint comparing the value derived for the vaccine+CDV vs. the non-CDV control for each vaccine. ‘***’ - p-value <0.001; ‘**’-p-value < 0.01; ‘*’-p-value < 0.05
3.10.2 Effect of lymphotropic challenge on severity of secondary infections

The effect of depletion of serological memory on the susceptibility to secondary infection was assessed by monitoring the severity of clinical symptoms following influenza challenge in vaccinated animals. Severity was assessed as a combined clinical score of body temperature and weight loss following challenge. The highest severity was detected at 36h post infection in CDV infected TIIV vaccine group. The LAIV vaccine provided significantly reduced severity of clinical manifestation at the same timepoint. However, this protective effect was lost in after CDV infection and ferrets experienced similarly severe disease as unvaccinated controls. (Figure IV.16).

*Figure IV.16: Effect of a lymphotropic challenge on the severity of secondary infection.*
Severity of clinical symptoms of influenza infection represented as a combined clinical score of weight loss and fever. The clinical score was measured twice a day for seven days after influenza challenge. ‘No vaccine’ group represents unvaccinated controls subject only to final challenge with influenza virus without CDV infection or vaccination. Significance of the difference in clinical score was calculated using Mann-Witney test for each timepoint comparing the value derived for the vaccine+CDV vs. the non-CDV control for each vaccine. ‘**’-p-value < 0.01;
4 Discussion

In this chapter I used the isotype-resolved BCR repertoire sequencing to investigate the immunological mechanisms underlying measles-associated immune suppression and tested the hypothesis of immunological amnesia as a source of increased rate of secondary infections post measles.

Immunological amnesia associated with loss of previously acquired B memory clones was tested directly by following the changes in BCR diversity of B memory compartment of measles-infected individuals. Immunological memory is created by establishment of antibody-producing long-lived plasma cell in bone marrow and generation of B and T memory cells maintained as a clonal expansion in the peripheral blood that can be activated upon re-exposure with the same antigen.

The depletion of individual-specific expanded B memory clones and increased convergence in B memory repertoires across individuals post measles suggests an effect of lymphotropic challenge on private clonal expansion events, reflecting the specific exposure history of an individual. The specific loss of highly expanded clones can be explained by observed preferential MeV replication in activated B/T memory cells [287] and their likely availability in peripheral blood. Significant loss of clonotypes associated with responses to other pathogens further supports the effect of measles on previously acquired immunity. The clonal composition of the memory B cell compartment showed reduced overlap with the pre-measles repertoire compared to controls. The lower relative contribution of pre-measles memory clones to the post-infection repertoire suggests a degree of immune re-setting and re-population of the B memory pool. The effect of re-setting is confirmed by the increased overlap in repertoires diversity between measles cases and uninfected and vaccinated controls. Estimation of the relative contribution of loss of previous memory vs repopulation of the memory pool with measles-specific clones would require antigen-specific information and cannot be fully resolved with the available data.
In addition to the effect observed in the B memory pool, the diversity of the naïve compartment was also impaired in a subset of the infected individuals. The B cell naïve compartment is continuously re-populated with bone marrow-derived B cells via a negative loop mechanism of regulation by mature B cells [377]. This ensures a highly stable pool of naïve B cells with a broad range of potential antigen specificities encoded in the diversity of the recombined BCR genes. The observed severe contraction of immune diversity in the naïve compartment of a subset of measles-infected individuals suggests an effect of measles on bone marrow reconstitution of B cell diversity. According to the macaque model of measles, MeV replicates successfully in bone marrow of infected animals, likely affecting plasma cells and haemopoietic stem cells [287, 279]. In vitro MeV infection is successfully established in both CD34$^+$ CD150$^-$ and CD34$^+$ CD150$^+$ hematopoietic stem and progenitor cell (HS/PCs) subsets and virus can be transmitted between stem cells and stromal cells or through infected mature B and T cells [378]. The ability to exchange virus between different cell types populating the bone marrow niche suggests that the bone marrow is a possible environment for sustained virus replication. MeV infection does not affect the viability and colony forming capacity of HS/PCs but impairs their short-term ability to reconstitute the bone marrow niche in NOD/SCID IL-2R$^c-$ mice resulting in reduced efficiency of engraftment [378]. This early MeV-induced impairment in the mouse model does not result in significant changes in counts of haemopoietic CD45$^+$ cells (or CD14$^+$ granulocyte or CD19$^+$ B lymphocyte populations) measured 13 weeks post engraftment but suggests an effect of MeV on the population of early HS/PCs.

Consistent with this mouse model of bone marrow impairment, the observed contraction in naïve cell diversity in measles infected individuals was not associated with reduced B-naïve cell counts but resulted from restricted usage of variable genes with a pronounced effect of V gene topology on the detected expression levels. This bias toward recombination site proximal genes reflects previous reports of V gene selection during recombination which is dependent on the chromatin accessibility of the IGH locus and thus selects for V genes proximal to the V-DJ recombination site [379, 380]. Together, the observed pattern of limited diversity and biased reconstitution of the naïve compartment suggest a model
of impairment of the bone marrow niche which does not affect the overall capacity for homeostatic proliferation but creates gaps in available early progenitor cells, thus limiting the breadth of possible immune reactivities. Depending on the developmental stage of the affected progenitor cells, this would either lead to delayed reconstitution of the breadth of the naïve compartment or permanent gaps in HSC population leading to compromised long-term maintenance of homeostatic proliferation upon subsequent infections.

Such long-term effect on the ability to successfully maintain the peripheral B cell compartment is observed in patients treated with the anti-CD20 antibody rituximab which causes lymphopenia similar to the one observed during measles. Despite the increased number of naïve B cells in bone marrow and peripheral blood, patients have significant reduction in CD27$^+$ B memory cells more than 2 years after treatment [381, 382]. This suggests that transient impairment of the bone marrow niche can lead to long-term effects on the renewal of B cell memory and generation of immune reactivity to new antigenic exposure [383]. The observed pattern of skewed reconstitution in B cell diversity suggests that the success of novel immune response would depend on the nature of the antigen and the ability to be recognised by V-DJ recombination proximal genes, as they would be the only available source of naïve B cell clones. Increased abundance of such clones in memory compartment was indeed observed for individual B-35 despite the overall depletion of memory clones available before measles (Figure IV.12) confirming the biased reconstitution of immune diversity post measles.

In addition to the restricted ability to reconstitute the naïve compartment, impairment of the bone marrow niche during measles infection can compromise the function of antigen-specific plasma cells responsible for serological immunity. The effect of a lymphotropic infection on serological memory was demonstrated using the ferret model. Indeed, CDV infection lead to significantly reduced titers of influenza neutralising antibodies in both vaccine groups thus removing the protective effect of vaccination. Since the half-life of serum IgG is 15-30 days [384], the maintenance of influenza-specific antibody titers 8 weeks post vaccination is reliant on successful generation of a pool of short or long-lived plasma
cells in the bone marrow [385]. Therefore, the reduction in antibody titers following CDV suggests an effect of lymphotropic infection on plasma cell pool established post vaccination. This is an alternative mechanism for establishment of immunological amnesia to the originally proposed loss of peripheral B cell memory. The loss of serological immunity would result in a restricted ability for early neutralisation of previously encountered pathogens and likely an increased severity of clinical symptoms (depending on the pathogenicity of the pathogen). Such increase in severity of disease was observed for both CDV infected vaccine groups following influenza challenge and lead to more severe clinical score even compared to influenza naive ferrets.

While serological memory is important for the early control of infection, the generation of B memory cells is required for successful recall responses. In the ferret model of vaccination with LAIV, CDV lead to both a reduction in protective anti-influenza antibodies, but also to inability to generate a memory recall response following re-exposure to the virus. Whether CDV infection causes loss of vaccine-generated B memory clones or the lack of a recall response stems from an impaired B cell activation could not be established from the serological and clinical assays and would require genetic characterisation of lymphocyte dynamics in the ferret model as shown in humans.

However, taken together, the results of the human and the ferret studies provide complementary evidence for two likely mechanism of immune-suppression associated with measles: 1) loss of previously generated B memory cells which leads to reduced breadth of possible recall responses 2) impairment of the bone marrow niche which can lead to a loss of plasma-cell derived serological immunity and in some cases to restricted naïve B cell pool which limits the reactivity to novel antigens.
Chapter V

Discussion and future directions
Discussion and future directions

The multi-stage process of generation of BCR diversity during B cell development makes the production of two B cells with identical CDR3 regions a virtually impossible event in the lifetime of an individual [386]. Therefore, with the aid of BCR sequencing, the BCR molecule can be used as a unique identifier of each cell that can be detected and tracked during the evolution of an adaptive immune response. The ability to use the genetic composition of the immunoglobulin receptor genes as a measure of B cell population size and activation status has revolutionised the field of immunology. Less than ten years since its first application, immune repertoire sequencing has facilitated a multitude of investigations ranging from basic biology of the immune system to understanding of infection [387], vaccination [131, 212, 388], malignancy [206, 205] and immune-mediated diseases [389, 390]. With the continuous advances in the available sequencing technologies and platforms for integration of BCR sequencing with other methods for immune cell phenotyping, the potential for further application of genetic profiling of B cell responses in research and clinical settings is enormous.

In this dissertation, I developed and optimised a methodology for a reliable capture of the diversity of B cell populations using high-throughput sequencing and then applied this approach to the study of immune mechanisms underlying measles-associated immune suppression. In the following sections I will discuss the importance of my findings in the context of some recent advances and outstanding challenges in the field of BCR sequencing.

1 Cross-validation of experimental approaches and analysis tools for BCR sequencing

In Chapter II, I investigated one of the most substantial technical challenges for the use of BCR sequencing – the presence of technical noise associated with library preparation and sequencing that can bias the interpretation of the derived measures of BCR diversity. The work on comparing different BCR amplification strategies with incorporation of molecular barcodes provided a demonstration of the magnitude of PCR error and biased template
1 Cross-validation of experimental approaches and analysis tools for BCR sequencing

capture that can substantially limit the utility of BCR sequencing. This highlighted the importance of a careful selection of a library preparation strategy for reliable capture of the clonal composition of B cell populations. The presence of method-specific biases was recently demonstrated for T cell receptor sequencing after comparison of multiple published protocols with commercially available platforms [391]. Such systematic comparison is lacking for BCR sequencing, but is likely to have similar sources of technical biases as demonstrated by my work on comparing barcode incorporation strategies. The requirement for reproducibility of repertoire capture is important in the context of longitudinal studies aiming to capture immunological changes over time in response to immune perturbation. The ability to re-capture a substantial proportion of BCR clones present at low frequency before perturbation is crucial for the distinction between background variation in BCR diversity and clonal dynamics in response to infection and vaccination. B cell clones of clinical importance are also often present at very low frequency even after infection and vaccination (e.g. broadly neutralising antibodies [392, 393]) so the combination of highly sensitive and reproducible capture of B cell diversity is key to their identification.

More work on cross-validation of different experimental strategies and analysis tools is required to enable the combination of datasets and meta-analysis studies which would improve the power to detect common immunological patterns across individuals. Important steps in this direction are made with the increasing number of dedicated software tools for characterisation of different aspects of BCR repertoire diversity [394, 154, 395, 396]. This aids the generation of more reproducible datasets and an easy validation of the performed analyses. Active efforts in this direction are made by the Adaptive Immune Receptor Repertoires (AIRR) community which unites multiple researchers in the field of immune repertoire sequencing who work together towards data standardisation, maintenance of public data repositories and setting up of minimal standards for data generation. 

1 http://airr.imacs.sfu.ca
Discussion and future directions

2 Development of novel methods for phylogenetic inference

The rapid increase in applications of B cell receptor sequencing in different research contexts uncovered limitations in currently available statistical and phylogenetic frameworks to study highly diverse B cell populations. The high sequence complexity of BCR genes with both genetic recombination and somatic mutation presents a challenge for phylogenetic inference models. A recently performed comparison of methods for phylogenetic reconstruction of B cell clonal lineages demonstrated that the majority of currently used methods fail to reliably recapitulate the evolution of BCR repertoires as they assume independent mutation of neighbouring nucleotides (which is not the case for BCRs [397]) and do not account for the smaller timescales of antibody evolution [156] Therefore, reconstructing the trajectories of clonal evolution of B cell populations is still an unresolved problem and a subject of continuous methods development in attempt to provide a more reliable inference framework for the evolution of B cell responses from immune repertoire data.

A phylogenetic model taking into account the specific characteristics of antibody evolution has only recently been described [398]. The validation of its accuracy is difficult to achieve from peripheral blood samples which contain a mixture of clonal lineages from different germinal centre reactions with varying degrees of convergence. In an attempt to overcome this inherent biological complexity, a recent collaboration between the laboratories of Gabriel Victora and Erick Matsen IV used data from germinal centre-derived single B cells to build a new phylogenetic model for B cell lineage reconstruction (pre-print available on arxiv: ²). Unlike previous lineage tree models [399, 400], this approach takes into account the abundance of each clone from a clonal lineage as a way to infer directionality of the mutational process and demonstrates an improved ability to uncover the true phylogenetic relationships between B cells undergoing affinity maturation in germinal centres.

In light of this work, the use of barcoded primers for correction of technical noise has an added advantage to other error-correction approaches because it ensures not only the accuracy of detection of somatic hypermutations (SHMs) but also provides a means to use unique molecular identifiers for quantification of clonal frequencies. The role of clone abundance for phylogenetic inferences highlights the importance of sufficient biological and technical sampling of the total B cell pool. The phylogenetic reconstruction of an accurate clonal lineage will be difficult if sampling depth is poor because clones appearing earlier in a clonal expansion event will be underrepresented or missing in the overall detected BCR sequences and thus not accounted for when constructing a phylogeny. Such underrepresentation of the early stages of a clonal evolution is particularly challenging when sampling peripheral blood where B cell population is derived from multiple germinal centre reactions. In this context, distinguishing between B cell clones that have evolved from a single clonal lineage or result from a convergent evolution of multiple germinal centre reactions in response to the same antigen is difficult to achieve without sampling the early stages of clonal evolution.

The skewed sampling of clonally expanded B cell populations largely stems from the biology of B cell evolution in response to an antigen where selected B cell clones expand clonally while early low-affinity ones are gradually lost. Therefore, this biological undersampling will always constitute a bias in the captured repertoire diversity. To this end, the use of sequencing strategies with improved sensitivity (as demonstrated for the optimised 3’MPLX method) will be essential to improve the technical sampling of diverse repertoires and overcome at least in part the burden of limited biological sampling of B cell pools. The use of methods for the generation of isotype-resolved BCR sequencing data as demonstrated in Chapter III of this dissertation can provide a means to validate novel methods of lineage reconstruction. Due to the fixed directionality of the class-switch process, tree branches predicting transitions between classes that are impossible given the location of the constant region genes on the IGH locus would be indicative of undersampled B cell ancestors and can be used as a measure of the accuracy of the inferred phylogeny. Combined with better biological sampling of cells undergoing affinity maturation (e.g. of germinal centre B cell derived via direct lymph node aspirates [401]) the study of somatic mutation patterns and
isotype composition of B cell repertoires has the potential to provide better insights into the true evolutionary processed leading to the generation of an antigen-specific responses.

3 Deriving predictive patterns of immune function using BCR sequencing

Historically, the generation of genetic diversity in B cell populations has been considered a stochastic process independent of genetic background [402–404]. Contrary to this notion, by characterising the evolution of isotype-specific responses in Chapter IV, I identified an important effect of the genetic composition of the B cell receptor on the class-switch fate of a B cell. This finding suggests an interaction between the variable gene background of an immunoglobulin molecule and its effector functions. This work is in line with recent findings of a high degree of genetic predetermination of diversity in variable genes [182] and demonstrates that the isotype composition of BCR repertoires could also be dependent on the genetic background of a B cell population. The conservation of B cell genetic substructure across individuals opens new avenues of research based on modelling the interaction between genetic predetermination and stochastic variation in the development of antibody repertoires [405]. Indeed, a theoretical framework of the evolution of B cell diversity predicts a high degree of repertoire predetermination as a necessary evolutionary adaptation for lymphocyte populations [406] and suggests a likely functional role as a first line of natural antibodies with high degree of cross-reactivity across pathogens [407]. The ability to derive common features of naïve B cell repertoire is a possible path towards the design of immunogens that would target shared BCR clones and thus ensure lower heterogeneity in responses across individuals, which is essential in vaccine development. This approach has been demonstrated in the design of an HIV immunogen that promotes the generation of broadly neutralising antibodies by selective targeting of specific germline genes [408, 409]. Since the naïve B cell populations have highly conserved V gene repertoire [410, 402], eliciting neutralising antibodies in a large proportion of the population could
3 Deriving predictive patterns of immune function using BCR sequencing

potentially be achieved via the use of computationally designed immunogens targeting highly represented germlines and public clones [411]. Despite being a potentially powerful tool for diagnostic and therapeutic purposes, the use of BCR repertoire data for identification of common immunological signatures and prediction of immune responses requires further characterisation of the genetic complexity of the IGH locus and robust statistical frameworks to model the high redundancy of human B cell repertoire (explained below).

3.1 Incomplete characterisation of germline diversity of human IGH loci

The lack of complete annotation of human V(D)J alleles limits the use of the higher degree of conservation in naïve repertoires to derive predictive signatures of B cell function. Several computational tools for variable gene prediction have been recently described and predict a substantially greater germline diversity than currently annotated [412, 413]. The highly repetitive nature of human immunoglobulin loci [414] makes sequence assembly from short read data extremely difficult as the degree of sequence dissimilarity between duplicate blocks is similar to the one observed between alleles of the same locus. Thus, the currently available sequences of IGH in the most recent assembly of the human genome constitutes an incomplete reconstruction of the IGHV cluster [415] and many of the identified alternate haplotypes represent variants that are present at higher frequencies than those found in the reference genome [416, 417]. The advances in long-read technologies are likely to provide a better understanding of the allelic variation of human IGH loci, but their high error rate in highly repetitive regions is still a major limitation for accurate annotation of germline genes [418, 419].

3.2 Limited models for inferring antigen-specificity from BCR data

While the germline BCR repertoires show high degrees of genetic pre-determination, the degree of conservation of immune signatures is substantially reduced in the antigen-experienced
Discussion and future directions

B cell populations [212, 420]. Models for association of signatures of expressed genes in immune cells with an outcome of immune perturbation are a common theme in the field of systems immunology [421–423]. Yet, appropriate statistical frameworks that can derive predictive patterns from BCR repertoire data are currently missing. The development of such methods is hindered by the highly redundant and cross-reactive nature of adaptive immune responses where one antigen can be recognised by multiple immune receptor molecules [424] and one receptor can recognise multiple antigens (millions in the case of T cell receptors [425]). This feature is an important aspect of the evolution of adaptive immunity as a non-cross-reactive repertoire would require an extremely high number of individual lymphocytes to cover the full spectrum of necessary antigen specificities [426]. However, this substantially reduces the predictive power of sequencing as individuals with different clonal composition of their BCR repertoires may be protected against the same infections. To this end, performing immune repertoire sequencing on larger cohorts of individuals and following how their responses to specific immune perturbations will be essential to overcome this effect of the heterogeneity of antigen-experienced B cell repertoires. Deriving more data on the genotype-phenotype relationship between BCR sequence and adaptive immune function will enable the development of better models for the use immune receptor sequencing data in a predictive fashion.

4 BCR sequencing for discovery of novel immune phenotypes

4.1 Providing evidence for immune dysfunction

The ability to characterise the composition of B cell populations by sequencing their immune receptor genes provides a means to study the cellular heterogeneity of adaptive immune responses at a level undetectable with other molecular assays In this way, the genetic
characterisation of B cell diversity in response to immune perturbation has the potential to uncover novel mechanisms of immune (dys)function.

In Chapter IV, I used the optimised methodology for isotype-resolved BCR sequencing from Chapter III to address a long-standing immunological paradox associated with measles virus infection: the prolonged state of immune-suppression despite recovery of normal cell counts. The longitudinal sampling approach and the availability of control cohorts enabled the calculation of baseline BCR diversity for each individual and comparison of the magnitude of repertoire changes after infection to the background variation observed across multiple time points in uninfected individuals. The presence of control samples is a commonly overlooked aspect of BCR repertoire characterisation in the context of infection and vaccination. Yet, given the highly dynamic nature of adaptive immune response, the inclusion of controls is essential to identify specific immune phenotypes associated with the specific infections.

As predicted by the hypothesis of ‘immunological amnesia’ [366, 279], measles infection resulted in depletion of previously expanded B memory pools and reduction in inter-individual diversity reflecting previous exposure history. This change in B cell composition could be caused by specific loss of previously available B memory cells or can reflect a repopulation of peripheral memory compartment with measles-specific clones, thus reducing the relative frequency of the previously abundant clones and making them more difficult to detect. Neither of these scenarios can be demonstrated directly from the BCR sequencing data due to the limited biological sampling that can be achieved by sampling peripheral blood. However, the profound B cell depletion during measles-associated lymphopenia [279] and the observed prolonged period of immune suppression [270] make the possibility of simple recirculation of B memory populations an unlikely scenario as it would not explain the increased susceptibility to secondary infections. Direct testing of BCR specificity from B memory cells before and after infection would be required to demonstrate that the population of B memory cells post infection is primarily measles-specific and change in diversity of the memory pool reflects a change in antigen-specificity.
Discussion and future directions

Even with the addition of antigen-specific information, the BCR data alone cannot assess completely the functional consequences of a depleted B memory pool (due to the limited predictive power, as described above). Therefore, the empirical validation of the observed restructuring in immune memory pools by the use of a ferret model was valuable for investigating if measles-associated changes in B cell diversity lead to an increased susceptibility to previously encountered pathogens. Such functional steps are an important aspect of any study where genetic or transcriptional variation is used as a predictor of distinct physiological phenotypes. Supporting findings from BCR sequencing data with molecular assays that can confirm changes in immune reactivity are required to improve the predictive power of immune repertoire analysis. The functional follow-up of sequence profiling of the immune cell diversity is often challenging as the breadth of immune responses is too high-dimensional and not always easy to access empirically, especially in humans. Therefore, characterising the diversity of immune receptor repertoires in other species that can be readily used as animal models might provide an opportunity to study in greater detail the relationship between immune receptor diversity and immune phenotypes and build better analysis tools to infer this relationship in humans. In the context of the ferret model, which is missing a complete annotation of IGH locus, I have used a variable gene prediction tool [427] to develop and test a primer set for amplification of ferret heavy chain genes which can be used for a follow-up work in confirming the immunological consequences of measles evidenced by human data.

4.2 Uncovering novel mechanisms of immune dysfunction

In addition to confirming the hypothesis of depletion of immune memory, the use of BCR repertoire sequencing uncovered an unexpected effect of measles infection on the diversity of the naïve compartment. Since the naïve compartment is populated by B cells derived from bone marrow, the naïve cells pool should be recovered by normal lymphocyte homeostasis [428]. Therefore, the highly biased variable gene composition of the repertoires of two of measles cases is indicative of incomplete reconstitution of naïve mature B cells in peripheral
4 BCR sequencing for discovery of novel immune phenotypes

blood. Although, this phenotype is observed in only two of the 20 tested cases, it provides a possible explanation for the presence of secondary complications in subset of individuals without any changes in the virus genotype (in the US between 10-30% of measles cases result in hospitalisations and pneumonia is the most common severe complication with incidence on 1:20 individuals) [429, 430]. The long-term consequences of a restricted breadth of the naïve compartment are difficult to ascertain from the available data. Whether the detected depletion constitutes a transient effect which would be eventually resolved, requires further follow-up of the two affected individuals. The ferret model of infection could not provide additional evidence for impaired ability to generate immune responses to novel antigens, as the animals were previously exposed to influenza through vaccination. An experiment to test this hypothesis is currently being performed.

The effect of measles virus infection on increased incidence of non-measles disease suggests a ≈3-year window of detectable increase in susceptibility to other pathogens [270]. The work presented in this dissertation demonstrates two possible mechanisms to explain this observation: 1) the long-term consequences of measles disease could be caused by depleted immune memory and/or 2) compromised ability to respond to new antigens (as a result of the restricted naïve B cell pool). The relative strengths of these effects are likely to be dependent on the overall pathogen burden in a country where measles outbreaks occur, which would determine the likelihood of exposure to other pathogens. A population-based study to explore the association between measles and overall infectious disease mortality in countries with higher incidence of infectious disease (as compared to US, UK and Denmark investigated by Mina et al [270]) is a possible approach to test the effect of immune suppression in the context of high pathogen burden.

Importantly, neither of these two immunological mechanisms of measles-associated immunosuppression would have been possible to uncover using conventional molecular assays, as both the naïve and the memory compartments recover their normal cell counts post measles. This demonstrates the utility of a BCR sequencing to identify aspects of adaptive
Discussion and future directions

immune (dys)function previously underappreciated due to limitations in available assays and analysis tools.

5 Final remarks

Immune repertoire sequencing is revolutionising the field of immunology by providing a robust method for genetic monitoring of adaptive immune responses. While the technology is being applied to an increasing range of research and clinical settings, several aspects of the methods used for data acquisition and analysis remain unresolved. Standardisation of laboratory techniques and analysis pipelines is lacking which limits the direct comparability of results. The lack of dedicated repositories for BCR sequences with associated metadata hinders efficient data sharing across studies. The substantial size of the obtained sequencing datasets complicates the use of any currently available tools for data visualisation and phylogenetic analysis. Despite the recognised potential for utility in diagnostic and clinical settings, the practical application of BCR sequencing will require a well-documented and validated relationship between observed immune dynamics and specific disease/response phenotypes. Despite all these limitations, the work presented in this dissertation demonstrates the ability of BCR repertoire sequencing to address aspects of immune function that have been subject to decades of extensive work without conclusive results. This gives a strong motivation for a continuous development and extended application of this technology in a wider range of research and clinical contexts where it can improve our understanding of the immune system. In much the same way our immune system is in a constant process of evolution and adaptation to its environment, the tools we use to explore its functionality must similarly evolve.
References


References


References


References


169
References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


[413] D. Gadala-Maria, G. Yaari, M. Uduman, and S. H. Kleinstein. Automated analysis of high-throughput b-cell sequencing data reveals a high frequency of novel...


Appendix A

Supplementary tables
Table A.1: Quality control of BCR sequencing reads for measles cases and control samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B05_20POST</td>
<td>519259</td>
<td>373409</td>
<td>168213</td>
<td>83009</td>
<td>19433</td>
<td>7948</td>
<td>Measles</td>
</tr>
<tr>
<td>B05_20PRE</td>
<td>480671</td>
<td>349275</td>
<td>171021</td>
<td>99660</td>
<td>13419</td>
<td>4910</td>
<td>Measles</td>
</tr>
<tr>
<td>B05_27POST</td>
<td>462171</td>
<td>351676</td>
<td>141778</td>
<td>65367</td>
<td>13445</td>
<td>5481</td>
<td>Measles</td>
</tr>
<tr>
<td>B05_27PRE</td>
<td>531000</td>
<td>354878</td>
<td>209343</td>
<td>116791</td>
<td>31443</td>
<td>13078</td>
<td>Measles</td>
</tr>
<tr>
<td>B07_20POST</td>
<td>497636</td>
<td>357221</td>
<td>181904</td>
<td>89907</td>
<td>17480</td>
<td>7037</td>
<td>Measles</td>
</tr>
<tr>
<td>B07_20PRE</td>
<td>531861</td>
<td>335552</td>
<td>255679</td>
<td>175588</td>
<td>49191</td>
<td>19494</td>
<td>Measles</td>
</tr>
<tr>
<td>B07_27POST</td>
<td>567522</td>
<td>422507</td>
<td>260137</td>
<td>150266</td>
<td>23330</td>
<td>7249</td>
<td>Measles</td>
</tr>
<tr>
<td>B07_27PRE</td>
<td>390798</td>
<td>276531</td>
<td>153204</td>
<td>75285</td>
<td>9245</td>
<td>3129</td>
<td>Measles</td>
</tr>
<tr>
<td>B08_20POST</td>
<td>481791</td>
<td>320418</td>
<td>210815</td>
<td>129147</td>
<td>37582</td>
<td>14032</td>
<td>Measles</td>
</tr>
<tr>
<td>B08_20PRE</td>
<td>546458</td>
<td>363367</td>
<td>254282</td>
<td>160527</td>
<td>47032</td>
<td>21013</td>
<td>Measles</td>
</tr>
<tr>
<td>B08_27POST</td>
<td>288679</td>
<td>237283</td>
<td>134933</td>
<td>73463</td>
<td>17516</td>
<td>4731</td>
<td>Measles</td>
</tr>
<tr>
<td>B08_27PRE</td>
<td>554233</td>
<td>447426</td>
<td>170526</td>
<td>87221</td>
<td>19859</td>
<td>6377</td>
<td>Measles</td>
</tr>
<tr>
<td>B09_20POST</td>
<td>498916</td>
<td>362630</td>
<td>195758</td>
<td>110825</td>
<td>22458</td>
<td>9047</td>
<td>Measles</td>
</tr>
<tr>
<td>B09_20PRE</td>
<td>509524</td>
<td>355913</td>
<td>213570</td>
<td>129469</td>
<td>23167</td>
<td>9423</td>
<td>Measles</td>
</tr>
<tr>
<td>B09_27POST</td>
<td>542600</td>
<td>376014</td>
<td>248151</td>
<td>150949</td>
<td>34045</td>
<td>8884</td>
<td>Measles</td>
</tr>
<tr>
<td>B09_27PRE</td>
<td>535048</td>
<td>351147</td>
<td>242085</td>
<td>150397</td>
<td>44477</td>
<td>12982</td>
<td>Measles</td>
</tr>
<tr>
<td>B12_20POST</td>
<td>521120</td>
<td>336050</td>
<td>231660</td>
<td>146511</td>
<td>31928</td>
<td>13028</td>
<td>Measles</td>
</tr>
<tr>
<td>B12_20PRE</td>
<td>470219</td>
<td>332674</td>
<td>184209</td>
<td>88316</td>
<td>20366</td>
<td>8689</td>
<td>Measles</td>
</tr>
<tr>
<td>B12_27POST</td>
<td>486612</td>
<td>372313</td>
<td>203466</td>
<td>125828</td>
<td>13234</td>
<td>3465</td>
<td>Measles</td>
</tr>
<tr>
<td>B12_27PRE</td>
<td>494468</td>
<td>352687</td>
<td>213642</td>
<td>87213</td>
<td>11136</td>
<td>3258</td>
<td>Measles</td>
</tr>
<tr>
<td>B13_20POST</td>
<td>566313</td>
<td>338020</td>
<td>276016</td>
<td>210326</td>
<td>74121</td>
<td>31579</td>
<td>Measles</td>
</tr>
</tbody>
</table>
Table A.1: Quality control of BCR sequencing reads (continued)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B13_20PRE</td>
<td>538212</td>
<td>393338</td>
<td>236161</td>
<td>126496</td>
<td>44756</td>
<td>20438</td>
<td>Measles</td>
</tr>
<tr>
<td>B13_27POST</td>
<td>626147</td>
<td>402774</td>
<td>204201</td>
<td>100160</td>
<td>42079</td>
<td>20041</td>
<td>Measles</td>
</tr>
<tr>
<td>B13_27PRE</td>
<td>614458</td>
<td>393624</td>
<td>281150</td>
<td>202400</td>
<td>49843</td>
<td>20041</td>
<td>Measles</td>
</tr>
<tr>
<td>B16_20POST</td>
<td>549736</td>
<td>339216</td>
<td>195900</td>
<td>116795</td>
<td>28258</td>
<td>12261</td>
<td>Measles</td>
</tr>
<tr>
<td>B16_27POST</td>
<td>645843</td>
<td>459311</td>
<td>311793</td>
<td>136155</td>
<td>25913</td>
<td>7526</td>
<td>Measles</td>
</tr>
<tr>
<td>B16_27PRE</td>
<td>553343</td>
<td>371519</td>
<td>234920</td>
<td>118648</td>
<td>17449</td>
<td>4627</td>
<td>Measles</td>
</tr>
<tr>
<td>B25_20POST</td>
<td>604137</td>
<td>349057</td>
<td>268714</td>
<td>187978</td>
<td>60194</td>
<td>27138</td>
<td>Measles</td>
</tr>
<tr>
<td>B25_20PRE</td>
<td>548744</td>
<td>343109</td>
<td>216252</td>
<td>134794</td>
<td>37216</td>
<td>14721</td>
<td>Measles</td>
</tr>
<tr>
<td>B25_27POST</td>
<td>497011</td>
<td>360635</td>
<td>254434</td>
<td>84262</td>
<td>30837</td>
<td>10438</td>
<td>Measles</td>
</tr>
<tr>
<td>B25_27PRE</td>
<td>553597</td>
<td>412988</td>
<td>229881</td>
<td>120875</td>
<td>10449</td>
<td>3288</td>
<td>Measles</td>
</tr>
<tr>
<td>B27_20POST</td>
<td>563600</td>
<td>362887</td>
<td>206891</td>
<td>104239</td>
<td>21021</td>
<td>8760</td>
<td>Measles</td>
</tr>
<tr>
<td>B27_20PRE</td>
<td>636837</td>
<td>372283</td>
<td>277660</td>
<td>218748</td>
<td>85388</td>
<td>36230</td>
<td>Measles</td>
</tr>
<tr>
<td>B27_27POST</td>
<td>677864</td>
<td>473320</td>
<td>292336</td>
<td>196830</td>
<td>33482</td>
<td>10720</td>
<td>Measles</td>
</tr>
<tr>
<td>B27_27PRE</td>
<td>624701</td>
<td>403232</td>
<td>251712</td>
<td>174718</td>
<td>74291</td>
<td>25033</td>
<td>Measles</td>
</tr>
<tr>
<td>B35_20POST</td>
<td>180804</td>
<td>137194</td>
<td>71537</td>
<td>21213</td>
<td>8923</td>
<td>5155</td>
<td>Measles</td>
</tr>
<tr>
<td>B35_20PRE</td>
<td>622825</td>
<td>338808</td>
<td>269415</td>
<td>184590</td>
<td>52549</td>
<td>20682</td>
<td>Measles</td>
</tr>
<tr>
<td>B35_27POST</td>
<td>807343</td>
<td>554751</td>
<td>394420</td>
<td>158318</td>
<td>55933</td>
<td>22591</td>
<td>Measles</td>
</tr>
<tr>
<td>B35_27PRE</td>
<td>838674</td>
<td>523621</td>
<td>341126</td>
<td>170113</td>
<td>41757</td>
<td>13685</td>
<td>Measles</td>
</tr>
<tr>
<td>B42_20POST</td>
<td>197762</td>
<td>115820</td>
<td>50144</td>
<td>26053</td>
<td>10131</td>
<td>5753</td>
<td>Measles</td>
</tr>
<tr>
<td>B42_20PRE</td>
<td>471586</td>
<td>298526</td>
<td>166926</td>
<td>79795</td>
<td>36272</td>
<td>21803</td>
<td>Measles</td>
</tr>
<tr>
<td>B42_27POST</td>
<td>443301</td>
<td>351210</td>
<td>205277</td>
<td>125058</td>
<td>25705</td>
<td>10674</td>
<td>Measles</td>
</tr>
</tbody>
</table>
Table A.1: Quality control of BCR sequencing reads (continued)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B42_27PRE</td>
<td>536742</td>
<td>362059</td>
<td>219508</td>
<td>110146</td>
<td>13781</td>
<td>4689</td>
<td>Measles</td>
</tr>
<tr>
<td>B54_20POST</td>
<td>593224</td>
<td>367810</td>
<td>229638</td>
<td>129474</td>
<td>37907</td>
<td>18424</td>
<td>Measles</td>
</tr>
<tr>
<td>B54_20PRE</td>
<td>704886</td>
<td>425105</td>
<td>317765</td>
<td>190763</td>
<td>83967</td>
<td>48077</td>
<td>Measles</td>
</tr>
<tr>
<td>B54_27POST</td>
<td>496658</td>
<td>323064</td>
<td>210428</td>
<td>131894</td>
<td>26269</td>
<td>9600</td>
<td>Measles</td>
</tr>
<tr>
<td>B54_27PRE</td>
<td>759102</td>
<td>493082</td>
<td>314916</td>
<td>242201</td>
<td>68205</td>
<td>24100</td>
<td>Measles</td>
</tr>
<tr>
<td>B55_20POST</td>
<td>426301</td>
<td>301756</td>
<td>142826</td>
<td>84294</td>
<td>38704</td>
<td>20551</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B55_20PRE</td>
<td>507643</td>
<td>347753</td>
<td>178770</td>
<td>118690</td>
<td>22016</td>
<td>9308</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B55_27POST</td>
<td>470631</td>
<td>361935</td>
<td>190581</td>
<td>127928</td>
<td>12464</td>
<td>4162</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B55_27PRE</td>
<td>405255</td>
<td>292101</td>
<td>178451</td>
<td>127568</td>
<td>36546</td>
<td>9573</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B56_20POST</td>
<td>465116</td>
<td>320746</td>
<td>177335</td>
<td>108616</td>
<td>31260</td>
<td>13634</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B56_20PRE</td>
<td>422497</td>
<td>297600</td>
<td>137963</td>
<td>85480</td>
<td>19617</td>
<td>8396</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B56_27POST</td>
<td>734530</td>
<td>486315</td>
<td>326870</td>
<td>228017</td>
<td>71499</td>
<td>19668</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B56_27PRE</td>
<td>428690</td>
<td>301756</td>
<td>168825</td>
<td>106476</td>
<td>16751</td>
<td>5199</td>
<td>Uninfected</td>
</tr>
<tr>
<td>B71_20POST</td>
<td>374191</td>
<td>239120</td>
<td>131878</td>
<td>75943</td>
<td>24859</td>
<td>12150</td>
<td>Measles</td>
</tr>
<tr>
<td>B71_20PRE</td>
<td>461301</td>
<td>261855</td>
<td>207058</td>
<td>153217</td>
<td>51997</td>
<td>22441</td>
<td>Measles</td>
</tr>
<tr>
<td>B71_27POST</td>
<td>548913</td>
<td>359109</td>
<td>228524</td>
<td>155804</td>
<td>71218</td>
<td>28880</td>
<td>Measles</td>
</tr>
<tr>
<td>B71_27PRE</td>
<td>423338</td>
<td>278584</td>
<td>159932</td>
<td>87737</td>
<td>24118</td>
<td>9550</td>
<td>Measles</td>
</tr>
<tr>
<td>B75_20POST</td>
<td>338582</td>
<td>213081</td>
<td>123182</td>
<td>71983</td>
<td>27468</td>
<td>15319</td>
<td>Measles</td>
</tr>
<tr>
<td>B75_20PRE</td>
<td>381659</td>
<td>250765</td>
<td>96675</td>
<td>47156</td>
<td>16006</td>
<td>8188</td>
<td>Measles</td>
</tr>
<tr>
<td>B75_27POST</td>
<td>438663</td>
<td>294090</td>
<td>160717</td>
<td>94055</td>
<td>23621</td>
<td>10151</td>
<td>Measles</td>
</tr>
<tr>
<td>B75_27PRE</td>
<td>507251</td>
<td>340867</td>
<td>211638</td>
<td>98046</td>
<td>32948</td>
<td>13663</td>
<td>Measles</td>
</tr>
</tbody>
</table>
Table A.1: Quality control of BCR sequencing reads (continued)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B78_20POST</td>
<td>455289</td>
<td>300888</td>
<td>143131</td>
<td>65416</td>
<td>17867</td>
<td>8495</td>
<td>Measles</td>
</tr>
<tr>
<td>B78_20PRE</td>
<td>350636</td>
<td>228578</td>
<td>106624</td>
<td>51288</td>
<td>15098</td>
<td>7350</td>
<td>Measles</td>
</tr>
<tr>
<td>B78_27POST</td>
<td>452218</td>
<td>318438</td>
<td>173472</td>
<td>96929</td>
<td>9894</td>
<td>3860</td>
<td>Measles</td>
</tr>
<tr>
<td>B78_27PRE</td>
<td>412123</td>
<td>279356</td>
<td>165985</td>
<td>96428</td>
<td>16236</td>
<td>6212</td>
<td>Measles</td>
</tr>
<tr>
<td>B83_20POST</td>
<td>471649</td>
<td>272404</td>
<td>194719</td>
<td>144426</td>
<td>77128</td>
<td>52202</td>
<td>Measles</td>
</tr>
<tr>
<td>B83_20PRE</td>
<td>452048</td>
<td>282518</td>
<td>156740</td>
<td>90342</td>
<td>36909</td>
<td>20491</td>
<td>Measles</td>
</tr>
<tr>
<td>B83_27POST</td>
<td>524794</td>
<td>358589</td>
<td>214726</td>
<td>144475</td>
<td>44372</td>
<td>15977</td>
<td>Measles</td>
</tr>
<tr>
<td>B83_27PRE</td>
<td>413670</td>
<td>269130</td>
<td>166649</td>
<td>101541</td>
<td>5616</td>
<td>2077</td>
<td>Measles</td>
</tr>
<tr>
<td>B88_20POST</td>
<td>496938</td>
<td>290323</td>
<td>221261</td>
<td>170061</td>
<td>102203</td>
<td>72682</td>
<td>Measles</td>
</tr>
<tr>
<td>B88_20PRE</td>
<td>505924</td>
<td>316212</td>
<td>208671</td>
<td>130020</td>
<td>50624</td>
<td>25629</td>
<td>Measles</td>
</tr>
<tr>
<td>B88_27POST</td>
<td>423424</td>
<td>258959</td>
<td>164091</td>
<td>128571</td>
<td>54046</td>
<td>21717</td>
<td>Measles</td>
</tr>
<tr>
<td>B88_27PRE</td>
<td>485383</td>
<td>322736</td>
<td>195319</td>
<td>130498</td>
<td>16396</td>
<td>6128</td>
<td>Measles</td>
</tr>
<tr>
<td>B90_20POST</td>
<td>382175</td>
<td>249936</td>
<td>132881</td>
<td>70258</td>
<td>29511</td>
<td>17184</td>
<td>Measles</td>
</tr>
<tr>
<td>B90_20PRE</td>
<td>383129</td>
<td>266636</td>
<td>84560</td>
<td>35870</td>
<td>11131</td>
<td>5538</td>
<td>Measles</td>
</tr>
<tr>
<td>B90_27POST</td>
<td>501048</td>
<td>340077</td>
<td>212752</td>
<td>139339</td>
<td>37419</td>
<td>13704</td>
<td>Measles</td>
</tr>
<tr>
<td>B90_27PRE</td>
<td>492561</td>
<td>345860</td>
<td>202187</td>
<td>92550</td>
<td>19718</td>
<td>7162</td>
<td>Measles</td>
</tr>
<tr>
<td>B94_20POST</td>
<td>761435</td>
<td>439584</td>
<td>329174</td>
<td>247874</td>
<td>84049</td>
<td>39314</td>
<td>Measles</td>
</tr>
<tr>
<td>B94_20PRE</td>
<td>670684</td>
<td>414138</td>
<td>294481</td>
<td>190441</td>
<td>74203</td>
<td>39850</td>
<td>Measles</td>
</tr>
<tr>
<td>B94_27POST</td>
<td>809868</td>
<td>534750</td>
<td>323008</td>
<td>241243</td>
<td>87065</td>
<td>32120</td>
<td>Measles</td>
</tr>
<tr>
<td>B94_27PRE</td>
<td>666895</td>
<td>426812</td>
<td>268989</td>
<td>163326</td>
<td>31776</td>
<td>12087</td>
<td>Measles</td>
</tr>
<tr>
<td>B95_20POST</td>
<td>674950</td>
<td>426538</td>
<td>272355</td>
<td>152538</td>
<td>37694</td>
<td>16324</td>
<td>Measles</td>
</tr>
</tbody>
</table>
Table A.1: Quality control of BCR sequencing reads (continued)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>B95_20PRE</td>
<td>594278</td>
<td>377186</td>
<td>231434</td>
<td>145622</td>
<td>32570</td>
<td>13428</td>
<td>Measles</td>
</tr>
<tr>
<td>B95_27POST</td>
<td>552731</td>
<td>405915</td>
<td>232146</td>
<td>179879</td>
<td>7510</td>
<td>2133</td>
<td>Measles</td>
</tr>
<tr>
<td>B95_27PRE</td>
<td>514919</td>
<td>31663</td>
<td>22956</td>
<td>149135</td>
<td>2133</td>
<td>6088</td>
<td>Measles</td>
</tr>
<tr>
<td>B99_20POST</td>
<td>554399</td>
<td>394502</td>
<td>238354</td>
<td>128181</td>
<td>15114</td>
<td>5112</td>
<td>Measles</td>
</tr>
<tr>
<td>B99_27POST</td>
<td>440851</td>
<td>371400</td>
<td>172389</td>
<td>110653</td>
<td>8468</td>
<td>19252</td>
<td>Measles</td>
</tr>
<tr>
<td>B99_27PRE</td>
<td>475246</td>
<td>347631</td>
<td>193899</td>
<td>110653</td>
<td>8468</td>
<td>19252</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU1_20POST</td>
<td>505333</td>
<td>253652</td>
<td>128399</td>
<td>10964</td>
<td>8468</td>
<td>19252</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU1_27POST</td>
<td>449980</td>
<td>232994</td>
<td>143255</td>
<td>63302</td>
<td>6023</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU1_27PRE</td>
<td>487986</td>
<td>289901</td>
<td>159187</td>
<td>70226</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU2_20POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU2_27POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU3_20POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU3_27POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU4_20POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU4_27POST</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
<tr>
<td>FLU4_27PRE</td>
<td>487342</td>
<td>238530</td>
<td>182363</td>
<td>127264</td>
<td>7228</td>
<td>7052</td>
<td>Measles</td>
</tr>
</tbody>
</table>
Table A.1: Quality control of BCR sequencing reads (continued)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw reads (F)</th>
<th>Raw reads (R)</th>
<th>Joined reads</th>
<th>Reads with BCs</th>
<th>Unique BCs</th>
<th>Unique seq</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLU4_27POST</td>
<td>502782</td>
<td>283935</td>
<td>159330</td>
<td>97993</td>
<td>30919</td>
<td>11854</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU4_27PRE</td>
<td>473520</td>
<td>257157</td>
<td>149511</td>
<td>93181</td>
<td>24394</td>
<td>10335</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU5_20POST</td>
<td>413587</td>
<td>193399</td>
<td>133084</td>
<td>81830</td>
<td>28394</td>
<td>14645</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU5_20PRE</td>
<td>484326</td>
<td>234489</td>
<td>165428</td>
<td>122972</td>
<td>47382</td>
<td>24930</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU5_27POST</td>
<td>514294</td>
<td>303691</td>
<td>167335</td>
<td>79649</td>
<td>25770</td>
<td>10337</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU5_27PRE</td>
<td>478643</td>
<td>299723</td>
<td>170920</td>
<td>92571</td>
<td>37625</td>
<td>16146</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU6_20POST</td>
<td>424323</td>
<td>216519</td>
<td>127300</td>
<td>61838</td>
<td>11860</td>
<td>5452</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU6_20PRE</td>
<td>451143</td>
<td>228109</td>
<td>153701</td>
<td>70147</td>
<td>15910</td>
<td>7469</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU6_27POST</td>
<td>542261</td>
<td>319220</td>
<td>185952</td>
<td>110450</td>
<td>53337</td>
<td>24532</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU6_27PRE</td>
<td>473430</td>
<td>309143</td>
<td>175627</td>
<td>81604</td>
<td>26256</td>
<td>13876</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU7_20POST</td>
<td>496782</td>
<td>245816</td>
<td>162951</td>
<td>97654</td>
<td>29012</td>
<td>14333</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU7_20PRE</td>
<td>718484</td>
<td>350415</td>
<td>252568</td>
<td>182316</td>
<td>72297</td>
<td>37407</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU7_27POST</td>
<td>474501</td>
<td>239929</td>
<td>137944</td>
<td>69808</td>
<td>9534</td>
<td>4142</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU7_27PRE</td>
<td>397644</td>
<td>221918</td>
<td>133702</td>
<td>90581</td>
<td>30361</td>
<td>14406</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU8_20POST</td>
<td>441642</td>
<td>222479</td>
<td>158507</td>
<td>103657</td>
<td>55471</td>
<td>37138</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU8_20PRE</td>
<td>508829</td>
<td>263649</td>
<td>184468</td>
<td>100793</td>
<td>48118</td>
<td>32733</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU8_27POST</td>
<td>550448</td>
<td>324701</td>
<td>173482</td>
<td>102422</td>
<td>36920</td>
<td>14233</td>
<td>Vaccine</td>
</tr>
<tr>
<td>FLU8_27PRE</td>
<td>508171</td>
<td>311592</td>
<td>181744</td>
<td>58926</td>
<td>17350</td>
<td>8343</td>
<td>Vaccine</td>
</tr>
</tbody>
</table>