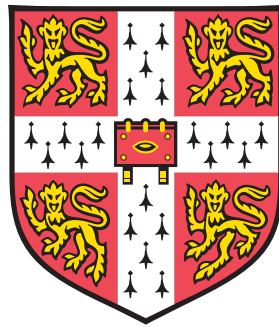


Analysing the B-cell repertoire:
Investigating B-cell population dynamics in
health and disease.

University of Cambridge
Jesus College



A thesis submitted for the degree of
Doctor of Philosophy

Rachael Bashford-Rogers

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

August 2014

Declaration

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

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

Rachael Bashford-Rogers
August 2014.

Abstract

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

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

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

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Paul Kellam and Prof. Allan Bradley for giving me the opportunity to carry out this project and for all their invaluable advice, support and encouragement. Many thanks also to my thesis committee, Dr Brian Huntly, and my post-doctoral supervisor, Dr Anne Palser, for their critical and constructive assessment of my work. In particular I thank Anne and Paul for their day-to-day guidance and manuscript proofreading. I would like to thank Dr George Vassiliou for his continual advice and guidance, as well as providing a wealth of fruitful collaborations. I thank the Wellcome Trust for my PhD studentship, and Jesus College and the Society for General Microbiology (SGM) for funding conference travel.

I extend my gratitude my collaborators, particularly Dr George Follows, Dr Danny Douek, Dr Mike Hubank, Dr Saad Idris, Dr Joanna Baxter, Dr. Clare Hodkinson, Dr Katerina Nicolaou and Dr Paul Costeas, with whom this work was made possible. A special thanks goes to the rest of the Kellam lab for countless productive discussions and constructive criticisms throughout the PhD programme. I also thank the Wellcome Trust Sanger Institute sequencing teams, and in particular Dr David Harris, for their technical expertise for generating the sequencing data. I further thank the Cambridge Cancer Trials Centre, and the patients and staff of Addenbrooke's Haematology Translational Research Laboratory.

On a personal note, I want to express my biggest gratitude to my wonderful parents and brother who have always encouraged me to pursue my goals and on whose help I could always count. In particular, I thank my family for their care, support, and guidance throughout my whole education. I am also very grateful to all my trusted friends who in various ways offered their support and encouragement during the period of my studies. A special thanks to Daniela Robles, Abigail Perrin and Michelle Wareham for their trusted friendships and our many tea breaks.

This does not give the extent of gratitude to all the people who have helped me through the PhD journey. Thank you all.

Rachael Bashford-Rogers

Wellcome Trust Sanger Institute, August 2014.

Contents

Chapter 1	1
1. Introduction	1
1.1. STRUCTURE OF THE ADAPTIVE IMMUNE SYSTEM.....	1
1.1.1. Structure of antibodies.....	1
1.1.2. Antibody isotypes.....	4
1.1.3. Generation of antibody diversity.....	7
1.1.4. B-cell development.....	7
1.1.4.1. Immunoglobulin gene rearrangements.....	7
1.1.4.2. B-cell receptor editing and allelic exclusion	14
1.1.5. B-cell response to antigens.....	15
1.1.6. Class switch recombination.....	17
1.1.7. B-cell memory responses.....	20
1.1.7.1. Generating T-cell dependent antigen immunological memory.....	20
1.1.7.2. Generating T-cell independent antigen immunological memory.....	21
1.1.7.3. Immunological memory recall.....	22
1.2. MEASURING B-CELL POPULATION STRUCTURE.....	24
1.2.1. Low-throughput B-cell receptor analyses.....	24
1.2.2. High-throughput B-cell receptor analyses	26
1.2.3. B-cell receptor repertoires.....	30
1.2.3.1. B-cell repertoires in model species.....	30
1.2.3.2. Diversity of the immune repertoire.....	32
1.2.3.3. Immune repertoire variation with age.....	35
1.2.3.4. B-cell repertoire responses to vaccines and natural infections.....	37
1.2.3.5. In vivo B-cell evolutionary processes.....	39
1.3. CHRONIC LYMPHOCYTIC LEUKAEMIA (CLL).....	43
1.3.1. Aetiology and epidemiology.....	43
1.3.2. Biology, pathogenesis and diagnosis of CLL	43
1.3.3. Monoclonal B lymphocytosis as a possible pre-leukemic phase.....	45
1.3.4. Disease staging in CLL.....	46
1.3.5. Prognostic markers in CLL	48
1.3.6. Current treatments for CLL.....	51
1.3.7. B-cell receptors in CLL.....	55

1.4.	B-CELL ACUTE LYMPHOBLASTIC LEUKAEMIA.....	57
1.4.1.	<i>Aetiology and epidemiology.....</i>	57
1.4.2.	<i>Biology, pathogenesis and diagnosis of ALL.....</i>	57
1.4.3.	<i>Prognostic markers in ALL.....</i>	58
1.4.4.	<i>Current treatments for ALL.....</i>	59
1.4.5.	<i>Monitoring minimal residual disease in ALL.....</i>	60
1.5.	AIMS AND HYPOTHESES	64
Chapter 2		65
2.	Materials and methods	65
2.1.	SAMPLES.....	65
2.2.	B-CELL METHODS.....	67
2.2.1.	RT-PCR.....	67
2.2.2.	<i>RNA capture for sequencing BCR repertoires</i>	69
2.2.3.	<i>5' Rapid amplification of cDNA ends (5'RACE) of B-cell receptors.....</i>	69
2.2.4.	<i>Sequencing methods.....</i>	69
2.2.5.	<i>Per-base error quantification</i>	70
2.2.6.	<i>Reference-based V-D-J assignment</i>	70
2.2.7.	<i>Network assembly and analysis</i>	70
2.2.8.	<i>Diversity measure calculations.....</i>	72
2.2.9.	<i>Estimation of cluster sizes due to sequencing error.....</i>	72
2.2.10.	<i>Phylogenetic analysis of BCR sequences</i>	73
2.2.11.	<i>Linear discriminant analysis of BCR repertoire parameters.....</i>	73
Chapter 3		74
3.	Developing computational methods for assessing B-cell receptor populations from next-generation sequencing.....	74
3.1.	INTRODUCTION.....	74
3.2.	RESULTS.....	75
3.2.1.	<i>Next-generation sequencing of IgH variable genes.....</i>	75
3.2.2.	<i>Next-generation sequencing error rate.....</i>	79
3.2.3.	<i>Percentage of identical BCR reads between samples.....</i>	83
3.2.4.	<i>Limitations of V-D-J gene classification.....</i>	85
3.2.5.	<i>BCR sequences organise into networks based on sequence diversity.....</i>	87
3.2.6.	<i>Population measures capture network and sample diversity.....</i>	93
3.2.7.	<i>Network property sensitivity to sequencing depth and edge lengths.....</i>	101
3.2.8.	<i>Minimal effect of sequencing errors on network properties.....</i>	104

3.2.9.	<i>BCR repertoire network parameters relate to CLL development.....</i>	106
3.2.10.	<i>Following malignant B-cell clonal dynamics by BCR sequencing.....</i>	109
3.2.11.	<i>Phylogenetic analysis of B-cell clones.....</i>	118
3.3.	CONCLUSIONS.....	123
Chapter 4	126
4.	Comparison of BCR amplification and sequencing methods.....	126
4.1.	INTRODUCTION.....	126
4.2.	RESULTS.....	126
4.2.1.	<i>Generation of BCR sequencing datasets for comparative studies.....</i>	126
4.2.2.	<i>Theoretical framework for sampling and sequencing BCR repertoires.....</i>	130
4.2.3.	<i>Sequencing depth requirement.....</i>	138
4.2.4.	<i>Assessing the stochasticity of sampling B-cell repertoires.....</i>	141
4.2.5.	<i>Comparison between independent primer sets.....</i>	148
4.2.6.	<i>Assessing differences between sequencing methods.....</i>	152
4.2.7.	<i>Assessing different RNA-capture and amplification methods.....</i>	156
4.2.8.	<i>Effect of amplicon length.....</i>	160
4.2.9.	<i>RNA versus DNA: which is best for BCR sequencing?.....</i>	163
4.3.	CONCLUSIONS.....	166
Chapter 5	168
5.	Minimal residual disease in B-acute lymphoblastic leukaemia.....	168
5.1.	INTRODUCTION.....	168
5.2.	RESULTS.....	168
5.2.1.	<i>BCR sequencing of longitudinal samples from B-ALL patients.....</i>	168
5.2.2.	<i>Comparison of ALL and CLL repertoires.....</i>	171
5.2.3.	<i>BCR sequencing sensitivity to detect B-ALL clones.....</i>	174
5.2.4.	<i>Detecting B-ALL BCRs in clinical samples.....</i>	179
5.2.5.	<i>Detecting B-ALL BCRs in RNA and DNA.....</i>	187
5.2.6.	<i>Distinguishing between B-ALL and healthy samples.....</i>	189
5.2.7.	<i>ALL Relapse: a case study of CSF relapse.....</i>	194
2.1.	CONCLUSIONS.....	201
Chapter 6	203
6.	Overall summary and future work.....	203
6.1.	OVERALL SUMMARY.....	203
6.2.	FUTURE WORK.....	206

References	209
Appendix A	240
<i>Published works</i>	<i>240</i>

List of Figures

FIGURE 1.1. REPRESENTATIVE STRUCTURE OF AN ANTIBODY.....	3
FIGURE 1.2. STAGES OF B-CELL MATURATION.	10
FIGURE 1.3. ARRANGEMENT OF THE HUMAN IGH GENE LOCUS ON CHROMOSOME 14.	11
FIGURE 1.4. PHYLOGENETIC SEQUENCE RELATIONSHIPS BETWEEN THE HUMAN A) IGHV AND B) IGHD GENES.	12
FIGURE 1.5. STAGES OF IMMUNOGLOBULIN GENE REARRANGEMENT.	13
FIGURE 1.6. MECHANISM OF CLASS-SWITCH RECOMBINATION.	19
FIGURE 1.7. FEATURES OF PRIMARY AND SECONDARY RESPONSE.	23
FIGURE 1.8. DIFFERENT IGH RNA SEQUENCING METHODS.....	28
FIGURE 1.9. ALIGNMENT OF HUMAN IGHV AND J GENES WITH BIOMED-2 PRIMER ANNEALING LOCATIONS.	29
FIGURE 1.10. SCHEMATIC DIAGRAM OF THE DIFFERENT TYPES OF BCR REPERTOIRE.	34
FIGURE 1.11. LINEAGE TREE CONSTRUCTED BY IGTREE.....	40
FIGURE 1.12. MAXIMUM PARSIMONY TREES OF B-CLONES.	42
FIGURE 2.1. SEQUENCING OF B-CELL RECEPTOR REPERTOIRES.	67
FIGURE 2.2. OUTLINE OF NETWORK GENERATION METHOD.	71
FIGURE 3.1. SEQUENCING OF B-CELL RECEPTOR REPERTOIRES.	84
FIGURE 3.2. PERCENTAGE OF REFERENCE SEQUENCES MATCHED TO 454 READS.....	86
FIGURE 3.3. GENERATION OF B-CELL RECEPTOR SEQUENCE NETWORKS.....	88
FIGURE 3.4. B-CELL RECEPTOR REPERTOIRES FROM DIFFERENT SAMPLES.	90
FIGURE 3.5. DISTRIBUTION OF MUTATIONS BETWEEN CONNECTED VERTEX SEQUENCES.....	92
FIGURE 3.6. MEASURES DIFFERENTIATING BETWEEN B-CELL RECEPTOR POPULATIONS.	95
FIGURE 3.7. COMPARISON OF DIVERSITIES FROM FR1 AND FR2 PRIMER SETS.	96
FIGURE 3.8. B-CELL RECEPTORS NETWORKS FOR FR1 AND FR2 PRIMER AMPLIFIED HEALTHY DONORS.	97
FIGURE 3.9. MEASURES DIFFERENTIATING BETWEEN B-CELL RECEPTOR DOMINANT CLUSTERS.	99
FIGURE 3.10. COMPARISON OF CLUSTER 1 AND CLUSTER 2 SEQUENCES FOR CLL PATIENT 5.	100
FIGURE 3.11. VARIATION OF BCR POPULATION MEASURES WITH SAMPLING DEPTH.....	102
FIGURE 3.12. NETWORK STRUCTURE VARIATION WITH EDGE LENGTH.....	103
FIGURE 3.13. ASSESSMENT OF ERROR IN BCR NETWORKS.....	105
FIGURE 3.14. VARIATION OF B-CELL RECEPTOR POPULATIONS.	107
FIGURE 3.15. BCR DIVERSITY VARIATION WITH TIME SINCE CLL DIAGNOSIS.	108
FIGURE 3.16. TREATMENT TIMES AND WHITE BLOOD CELL COUNT OVER TIME FOR TEMPORAL CLL SAMPLES.	111
FIGURE 3.17. DYNAMICS OF CLL BCR REPERTOIRES AND WHITE BLOOD CELL COUNTS.	115
FIGURE 3.18. DYNAMICS OF CLL BCR REPERTOIRES PROPERTIES.	117
FIGURE 3.19. UNROOTED MAXIMUM PARSIMONY TREES OF THE MALIGNANT CLL CLUSTERS.	122
FIGURE 4.1. SIMULATION DISTRIBUTIONS.	133
FIGURE 4.2. PERCENTAGES OF BCR SEQUENCES SHARED BETWEEN REPEATED SAMPLES.	134
FIGURE 4.3. EXPERIMENTAL DESIGN FOR ASSESSING BCR SEQUENCING REPRODUCIBILITY.	137

FIGURE 4.4. BCR SAMPLING PROBABILITIES.	140
FIGURE 4.5. GENE-USAGE FREQUENCY CORRELATIONS BETWEEN SEQUENCING REPEATS.	142
FIGURE 4.6. BCR CLONALITY MEASURES CORRELATIONS BETWEEN SEQUENCING REPEATS.	143
FIGURE 4.7. GENE-USAGE FREQUENCY CORRELATIONS BETWEEN RT-PCR REPEATS.	145
FIGURE 4.8. BCR CLONALITY MEASURES CORRELATIONS BETWEEN RT-PCR REPEATS.	146
FIGURE 4.9. INDIVIDUAL BCR FREQUENCY CORRELATIONS BETWEEN RT-PCR REPEATS.	147
FIGURE 4.10. ASSESSING THE REPRODUCIBILITY OF SAMPLES AMPLIFIED BY THE FR1 AND FR2 PRIMER SETS.	149
FIGURE 4.11. GENE-USAGE FREQUENCY CORRELATION BETWEEN FR1 AND FR2 PRIMER SETS.	150
FIGURE 4.12. COMPARISON OF BCR SEQUENCING NETWORKS BETWEEN FR1 AND FR2 PRIMER SETS.	151
FIGURE 4.13. COMPARING DIFFERENT BCR SEQUENCING METHODS.	154
FIGURE 4.14. INDIVIDUAL BCR FREQUENCY CORRELATIONS BETWEEN DIFFERENT SEQUENCING METHODS.	155
FIGURE 4.15. COMPARING DIFFERENT BCR AMPLIFICATION METHODS.	158
FIGURE 4.16. INDIVIDUAL BCR FREQUENCY CORRELATIONS BETWEEN DIFFERENT AMPLIFICATION METHODS.	159
FIGURE 4.17. VARIATION OF DIVERSITY MEASURES WITH READ-LENGTH.	161
FIGURE 4.18. ALIGNMENT OF RNA CAPTURE READS TO BCR SEQUENCE.	162
FIGURE 4.19. COMPARISON OF RNA AND DNA REPERTOIRES.	165
FIGURE 5.1. COMPARING THE B-CELL REPERTOIRE IN B-ALL WITH CLL.	173
FIGURE 5.2. BCR SEQUENCING SENSITIVITY.	177
FIGURE 5.3. B-ALL BCR POPULATIONS.	181
FIGURE 5.4. BI-CLONAL B-CELL EXPANSION IN B-ALL PATIENT 859.	186
FIGURE 5.5. DETECTION OF B-ALL BCR SEQUENCES IN RNA AND DNA SAMPLES.	188
FIGURE 5.6. DISTINGUISHING BETWEEN B-ALL AND HEALTHY B-CELL POPULATIONS.	193
FIGURE 5.7. PHYLOGENETICS OF B-ALL CSF RELAPSE.	197
FIGURE 5.8. POTENTIAL MECHANISMS OF GENERATING RELAPSE B-ALL B-CELL POPULATIONS.	198

List of Tables

TABLE 1.1. PROPERTIES OF IMMUNOGLOBULIN ISOTYPES.	5
TABLE 1.2. SUMMARY OF VACCINE STUDIES BASED ON LOW-RESOLUTION B-CELL REPERTOIRE CHARACTERISATION.	25
TABLE 1.3. SUMMARY OF STUDIES OF B-CELL REPERTOIRES IN MODEL SPECIES.	31
TABLE 1.4. NUMBER OF POTENTIAL HUMAN BCR GENE SEGMENT COMBINATIONS.	32
TABLE 1.5. SUMMARY OF STUDIES OF B-CELL REPERTOIRES FROM HEALTHY INDIVIDUALS.	33
TABLE 1.6. SUMMARY OF STUDIES OF IMMUNE REPERTOIRE VARIATION WITH AGE.	36
TABLE 1.7. SUMMARY OF STUDIES OF ANTIGEN-SPECIFIC ANTIBODY REPERTOIRES.	37
TABLE 1.8. SUMMARY OF STUDIES OF B-CELL REPERTOIRES FROM VACCINATIONS.	38
TABLE 1.9. RAI STAGE MEDIAN SURVIVAL.	46
TABLE 1.10. BINET STAGE MEDIAN SURVIVAL.	47
TABLE 1.11. GENOMIC MARKERS IN CLL ASSOCIATED WITH PROGNOSIS.	48
TABLE 1.12. GENOMIC AND CELL-BASED PROGNOSTIC FACTORS IN ALL.	59
TABLE 1.13. THE MAIN CLINICAL ASSAYS USED TO MONITOR MRD IN ACUTE LYMPHOBLASTIC LEUKAEMIA.	63
TABLE 2.1. TABLE OF SAMPLES USED.	66
TABLE 2.1. HUMAN B-CELL RECEPTOR PCR PRIMERS.	68
TABLE 3.1. PATIENT SAMPLE INFORMATION.	76
TABLE 3.2. SAMPLE INFORMATION AND NUMBER OF SEQUENCING READS.	77
TABLE 3.3. SAMPLE INFORMATION AND NUMBER OF SEQUENCING READS FROM THE BOYD ET AL. DATASET.	78
TABLE 3.4. SAMPLE INFORMATION AND NUMBER OF SEQUENCING READS FOR CONTROL GENES.	80
TABLE 3.5. ESTIMATED AVERAGE PER-BASE 454 ERROR FREQUENCIES BY TYPE.	81
TABLE 3.6. ESTIMATED AVERAGE PER-BASE MISEQ ERROR FREQUENCIES.	82
TABLE 3.7. FILTERED BCR DEPTHS FOR TEMPORAL CLL PATIENT SAMPLES.	109
TABLE 4.1. SAMPLES USED IN THIS STUDY FOR EACH AMPLIFICATION METHOD.	128
TABLE 4.2. MEAN AND STANDARD DEVIATION OF READ DEPTHS PER SAMPLE.	129
TABLE 4.3. MEAN DIVERSITY MEASURES FOR EACH SAMPLE TYPE.	129
TABLE 4.4. ESTIMATION OF NUMBER AND PERCENTAGE OF SAMPLED PERIPHERAL BLOOD B-CELLS.	131
TABLE 4.5. B-CELL SAMPLING SIMULATION PARAMETERS.	132
TABLE 4.6. TECHNICAL INFORMATION OF THE NEXT-GENERATION SEQUENCING PLATFORMS USED IN THIS STUDY.	153
TABLE 5.1. B-ALL PATIENT SAMPLE INFORMATION.	170
TABLE 5.2. FALSE POSITIVE RATE FOR DETECTING B-ALL MRD.	178
TABLE 5.3. CORRELATIONS BETWEEN THE PERCENTAGE OF B-ALL BCRs MATCHED AND QPCR LEVELS.	182
TABLE 5.4. PERCENTAGES OF B-ALL CLONOTYPIC BCR SEQUENCES IN REPEATED SAMPLES.	183
TABLE 5.5. TABLE OF THE PROPERTIES OF THE LARGEST TWO CLUSTERS IN PATIENT 859.	187
TABLE 5.6. DETECTION OF B-ALL CELLS IN PATIENT 859.	194
TABLE 5.7. PROBABILITIES OF BCR REPERTOIRE OVERLAP BETWEEN DAY 0 BM AND DAY 567 CSF SAMPLES.	200

Nomenclature

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