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

List of Figures

List of Tables

Nomenclature

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