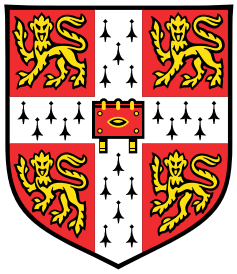
Machine Learning for Precision Oncology

Genomic Classification and Analysis of Diffuse Large B Cell Lymphoma



Camilo Ruiz

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
MPhil in Biological Sciences

King's College September 2017

I would like to dedicate this thesis to my family …

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 20,000 words excluding appendices, bibliography, footnotes, tables, and equations.

Camilo Ruiz
September 2017

Contributions

This project was carried out in collaboration with the Haematological Malignancy Research Network, the Leeds Teaching Hospital, and the Sanger Institute. Patients were diagnosed and samples were collected at St James' Hospital in Leeds (Dr. Sharon Barrons, Dr. Catherine Cargo, Dr. Cathy Burton, and Dr. Jan Taylor). Clinical outcome data was collected and processed at the University of York (Dr. Eve Roman, Dr. Alexandra Smith, Dr. Russell Patmore, Dr. Simon Crouch, and Dan Painter). DNA sequencing was conducted by the sequencing core at the Wellcome Trust Sanger Institute. Genome assembly was conducted by the Sanger Institute's Cancer and Somatic Mutation Group's IT team led by Adam Butler and Jon Teague (i.e. from raw sequencing files to initial VCF files). Finally, I conducted all analysis downstream of genome assembly. I constructed a computational pipeline to identify driver mutations within the set of sequencing variants. I then manually curated these variants in collaboration with Dr. Peter Campbell and Dr. Philip Beer, a consulting Haematopathologist at Leeds Teaching Hospital. Subsequently, I used these drivers to conduct a genomic landscape analysis of the B-NHLs within our cohort. Finally, I used Bayesian Dirichlet processes to generate a preliminary classification of the disease. The project was advised by Dr. Peter Campbell at the Sanger Institute. Additionally, my thesis committee included Dr. Peter Campbell, Dr. Philip Beer, Dr. Moritz Gerstung of the European Bioinformatics Institute, and Dr. Brian Tom of the MRC Biostatistics Unit at the University of Cambridge.

Acknowledgements

To my advisor, Professor Peter Campbell: Thank you for mentoring me and guiding me on this journey. Thank you for trusting me with this project and pushing me to succeed.

To the members of my thesis committee Professor Peter Campbell, Professor Moritz Gerstung, Professor Brian Tom, and Dr. Philip Beer: Thank you for challenging me and providing guidance along this project. I appreciate your continual support and insight in helping me avoid mistakes and guiding me towards new insights.

To our collaborators Dr. Sharon Barrans, Dr. Catherine Cargo, Dr. Simon Crouch, Dr. Cathy Burton, Dr. Russell Patmore, Dr. Alexandra Smith, and Dr. Jan Taylor: Thank you for providing the clinical samples that made this study possible. Also thank you for helping integrate some of the preliminary clinical data and offering broad advice on the project direction.

To the Campbell Lab, especially Grace Collord, Dr. Daniel Leongamornlert, and Dr. Francesco Maura: Thank you for your endless help. You helped me overcome a great deal of challenges, whether graphing in R or interpreting the clinical output. A special thanks to Grace in particular for providing the graphics code necessary to visualize gene-level mutational profiles.

To the Gates Cambridge Trust: Thank you for giving me the opportunity to study at Cambridge as a Gates Scholar. I am indebted to the Trust and hope to use what I have learned to improve the world.

To my friends and family: Thank you for your constant support and love, in both the challenging and fruitful times.

Abstract

The most common lymphoma in adults, Diffuse Large B Cell Lymphoma (DLBCL) accounts for 30-35% of all non-Hodgkin lymphoma (NHL) cases. Even though DLBCL is curable in advanced stages, up to one-third of patients will not achieve a cure with their initial therapy. Today, there is no effective way to predict which patients will or will not be cured by first-line chemotherapeutic treatment. Patients who are not initially cured relapse, develop chemoresistance, and ultimately die of their disease.

Current classification and prognostication schemes do not account for much of the genetic and molecular heterogeneity of DLBCL. Indeed, the gold standard WHO classification uses clinical data, morphology, phenotype, cytogenetics, and molecular characteristics to demarcate DLBCL subtypes. However, it does not incorporate many of the genetic lesions that both cause DLBCL and make it heterogeneous. As a result, the most common WHO subtype of DLBCL – DLBCL, not otherwise specified (DLBCL NOS)–likely encapsulates multiple disease subtypes for which conventional diagnostic approaches have not yet yielded clear methods of discrimination.

The prognostication and treatment guidelines for DLBCL are similarly uniform, again not reflecting the heterogeneity inherent to DLBCL. The gold standard clinical prognostic tool, the Revised International Prognostic Index (R-IPI), sorts patients into three risk groups based on factors such as age and whether their lactate dehydrogenase level is elevated. None of the R-IPI factors, however, accounts for the genetic basis of DLBCL and cannot therefore incorporate prognostic information from genetic variability between patients within the same risk group. Virtually all DLBCL patients receive the same first-line therapy, R-CHOP, despite the probability that the genetic and biological heterogeneity will result in heterogeneous response to the potential treatments available. Up to one third of patients will not be cured by R-CHOP and their prognosis suffers significantly in the case of relapse.

In this study, we propose a novel, purely genomic classification for DLBCL and other B-cell non-Hodgkin lymphoma (B-NHLs) that incorporates the genetic heterogeneity inherent to the disease. By analysing the genetic lesions of 1607 B-NHL patients over 15 years and then performing a machine-learning based clustering, we identify seven distinct classes with characteristic genetic lesions and patterns of co-mutation. These classes aptly distinguish Follicular Lymphoma (FL) and Burkitt Lymphoma (BL) samples from DLBCL samples while simultaneously resolving the heterogeneity of DLBCL. Class 5, for example, shows hallmark mutations of Splenic Marginal Zone Lymphoma (*NOTCH2, BCL10, SPEN*),

suggesting these DLBCL patients represent transformed lymphomas. Such a conclusion could not have been drawn from histology alone and importantly, suggests these patients may respond differently to novel therapies compared to other DLBCL subtypes. We also present a genomic landscape analysis more complete and powerful than prior work since our study is nearly 10X larger than the largest prior B-NHL genetics study. We present mutation profiles at the gene level for nearly 200 genes implicated in lymphoma, identifying previously unreported mutations such as the aberrant splicing of a single exon in *SGK1*. Future work adding copy number, gene expression, and translocation data will enhance the robustness and resolution of our classification scheme and landscape analysis.

Table of Contents

List of Figures

in the coding sequence, (2) protein domains from UniProtKB, and (3) bubbles. Bottom half of plots show bubbles sized according to the number of mutations found in COSMIC. **(a)** Tumour suppressor genes exhibit disrupting mutations spread throughout the coding sequence of the gene. *ARID1A* is shown as a representative example. **(b)** Highly recurrent missense mutations may disrupt a key residue. *SOCS1* is shown as a representative example. **(c, d)** *TNFRSF14* and *BTG2* exhibited recurrent nonsense, frameshift, and nonstop mutations**.**

Figure 7 Gene-level analysis demonstrates known and novel oncogene hot spots. (a) Oncogenes exhibit missense hot spots. *XPO1* is shown as a representative example. **(b)** We additionally identified novel hotspots in known oncogenes. *CARD11* is shown as a representative example. **(c)** We created the mutational profile for *STAT3*, a known but uncharacterized oncogene.

Figure 8 Gene-level analysis shows the potential for genes to serve as both tumour suppressors and oncogenes. *TP53* is shown as a representative example.

Figure 9 Gene-level analysis shows patterns of aberrant somatic hypermutation. *B2M* is shown as a representative example.

Figure 10 Gene-level analysis reveals disrupting mutations clustered in highly specific domains. (a) *BCL10*, **(b)** *IRF8*, **(c)** *FAS*, **(d)** *ARID1B*, **(e)** *NOTCH1*, **(f)** *NOTCH2*, **(g)** *KLF2*, **(h)** *TCF3*, **(i)** *SMARCB1*.

Figure 11 Co-mutation and mutual exclusivity patterns generate eight distinct classes in FL, BL, and DLBCL. Lower triangle depicts pairwise association between lesions in genetic classes. The colour of each tile corresponds to the odds ratio for each pair, with brown representing mutual exclusivity and blue indicating co-mutation. Odds ratios are computed by observed co-mutation rates compared to expected co-mutation based on each lesion's gene frequency. Coloured tiles represent significant relationships ($p < 0.05$), asterisks show significant family wise error rates (FWER < 0.05), boxes show false discovery rates < 0.1 (FDR < 0.1). Upper triangle depicts absolute occurrences of co-mutation for each pair, coloured on a gradient.

Figure 12 Each class shows a distinct mutational signature profile. (a) Number of driver mutations across all classes, coloured by proposed class assignment for patient with that mutation. **(b-i)** Mutational signature of each class. Numbers next to class show number and fraction of patients assigned to that class. Each bar shows the median posterior probability of a given lesion with error bars corresponding to the 2.5 and 97.5 quantiles.

Figure 13 Classes show distinct subtype compositions and survival outlooks. (a, b) Patient assignment to WHO diagnostic groups or subtypes compared to patient assignment to proposed classes. **(c)** Kaplan-Meier plot for proposed classes.

Nomenclature

ABC-DLBCL: Activated B Cell-Like Diffuse Large B-Cell Lymphoma

BCL, Int.: B-cell lymphoma, intermediate between DLBCL and classical HL

BL: Burkitt Lymphoma

B-NHL: B Cell non-Hodgkin Lymphoma

DLBCL: Diffuse Large B Cell Lymphoma

FL: Follicular Lymphoma

FL-LC: Follicular lymphoma, large cell

GC: Germinal Centre

GCB-DLBCL: Germinal Centre B Cell-Like Diffuse Large B-Cell Lymphoma

GZL: B-cell lymphoma, intermediate between DLBCL and classical HL

IV-LBCL: Intravascular large B-cell lymphoma

PB-LBCL: Plasmablastic large B-cell lymphoma

SMZL: Splenic marginal zone lymphoma

THR-LBCL: T-cell/histiocyte-rich large B-cell lymphoma