The genetic landscape of relapsed-refractory DLBCL

by

Christopher Rushton

Bachelor of Science, University of Victoria, 2015

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2022

Declaration of Committee

Name:	**Christopher Rushton**
Degree:	**Doctor of Philosophy**
Title:	**The genetic landscape of relapsed-refractory DLBCL**
Committee:	**Chair:** Valentin Jaumouillé Assistant Professor, Molecular Biology and Biochemistry
	Ryan D. Morin Supervisor Associate Professor, Molecular Biology and Biochemistry
	Fiona Brinkman Committee Member Professor, Molecular Biology and Biochemistry
	David W. Scott Committee Member Associate Professor, Medicine University of British Columbia
	Jonathan Choy Examiner Professor, Molecular Biology and Biochemistry
	Björn Chapuy External Examiner Associate Professor, Hematology and Oncology University of Göttingen

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin Lymphoma, with ~4,400 Canadians diagnosed annually. DLBCL patients are treated with a standard frontline immunochemotherapy (R-CHOP) comprised of several chemotherapeutics and the monoclonal antibody rituximab, which binds to the B-cell surface marker CD20. While R-CHOP is generally effective for DLBCL, for patients where frontline treatment fails (relapsed-refractory DLBCL, rrDLBCL), prognosis is poor with a median survival time of 6 months. Although numerous rrDLBCL salvage therapies have been developed, their efficacies have been limited partially due to the heterogeneity of DLBCL and an incomplete understanding of the genomic features associated with relapse disease. We hypothesize that the genomic landscape of rrDLBCL will be distinct from diagnostic DLBCL and contain recurrent mutations contributing to both treatment failure and advanced disease. To this end, we performed ultra-deep targeted sequencing of 63 genes (CAPP-Seq) on 135 rrDLBCL liquid biopsies and found six genes significantly enriched for mutations at relapse. *KMT2D* and *TP53* were mutated in half of all rrDLBCL samples, with *TP53* enriched for dominant negative hotspot mutations. 8% of rrDLBCL cases harbour *MS4A1* mutations, which encodes CD20. Mutations in *MS4A1* were clonally selected following treatment and *in vitro* attenuated the binding of rituximab and other anti-CD20 antibodies, including those undergoing clinical trials. To expand upon these findings, we performed whole exome sequencing on 155 rrDLBCL samples and found mutations in *TET2* and *TMEM30A* significantly depleted at relapse which, in conjunction with enrichment of mutations in *KMT2D* and *CREBBP*, suggest broad epigenetic changes in rrDLBCL. Using additional copy number information from 77 rrDLBCL liquid biopsies (n=222) we observed a high burden of copy number variants in rrDLBCL and novel recurrent deletions of RNA regulators *HNRNPU* and *HNRNPD*, the MHC Class I regulator *IRF2*, and recurrent gains involving the B-cell proliferation factors *IKZF3* and *TCF3*, representing candidate therapeutic targets. 13 regions were significantly enriched for events in rrDLBCL, including these novel events and others regulating apoptosis (*TP53, PTEN, BCL2*) and proliferative signaling (*MIR17HG, BCL6*). Overall, we have further characterized the genetics of rrDLBCL and identified mechanisms of treatment resistance and possible therapeutic targets.

Keywords: **Diffuse large B-cell lymphoma; relapsed DLBCL; rituximab; treatment resistance; biomarkers; DNA sequencing**

This thesis is dedicated to my grandfather, Earnest Ernie Rushton, who passed away in 2003 after a long battle with lung cancer, and Liz Chavez, research technician at BC cancer, who contributed significantly to this project and passed away in October 2021 abruptly from cancer.

This thesis is also dedicated to all patients who participated in this project, whose samples will continue to improve cancer research and patient care.

Acknowledgements

Over the course of my degree and these projects, I have had the opportunity to meet and work with an exemplary group of individuals from the Morin lab, BC Cancer, SFU, UBC, and other universities across Canada. First, I must acknowledge and thank the unique members of the Morin lab, both current and former, as without their support these projects and my degree would not have been possible. Specifically, I would like to thank Quratulain Qureshi for her good sense of humour and seemingly excessive levels of sass, Manuela Cruz for being an easy person to talk to with a plethora of fried chicken recommendations, and Bruno Grande, whose R and figure-creating expertise (termed “Brunifying”) set such a high bar in the lab that no individual could hope to reach it. I must also thank Sarah “broomit” Arthur who worked with me on the first relapse paper, and was an invaluable resource both inside and outside of the lab, and Nicole Thomas for being an excellent R resource as well as a fantastic coffee and beer companion when the stress of graduate school was too much to bear. Of course, one must also acknowledge Miguel Alcaide, who not only was an extremely productive wet lab member despite his advanced age and was responsible for sequencing the majority of samples within this project.

I have also had the pleasure of working with outstanding and incredibly knowledgeable group of individuals from BC Cancer and McGill university. Nathalie Johnson and one of her graduate students, Ryan Rys, provided hundreds of samples for this project as well as associated metadata and are continuing to collect and sequence samples from rrDLBCL patients. It is with a rather heavy heart that I am unable to continue on with these projects and analyze their samples, as I have no doubt they will prove invaluable to lymphoma research. I would like to acknowledge the significant contributions of Liz Chavez, a former technician at BC cancer, who processed and sequenced hundreds of relapse liquid biopsies used in the second portion of this project and taught me wet-lab considerations for DNA sequencing, and whom I am indebted to. Unfortunately, Liz Chavez passed away suddenly from cancer in fall of 2021. It is hoped that through the contributions of this project and others, we can improve the outcomes of all cancer patients to avoid the heartache that comes with such a loss.

I was also exceptionally fortunate to have the support of extremely experienced and helpful committee members. David Scott not only provided extensive lymphoma

knowledge but also provided helpful patient-focused insight into the implications of this research and other projects. Fiona Brinkman was essential in navigating the more student-focused portions of grad school, and being a TA for her MBB342 course was an invaluable if stressful experience. Furthermore, the contributions of the MBB graduate caucus and the various social and academic events they organize, as well as the ever-helpful MBB staff members, particularly Mimi Fourine and Christine Beauchamp, were essential in my development over the past several years.

Of course, undertaking graduate school and a PhD specifically is a very intensive and exhausting process, which would have not been possible without the support of my family and friends. I must thank my aunt and uncle, Ken and Marilyn Rushton, for their eternal patience with my inconsistent and insane working hours (especially while thesis writing) while I was renting their basement suite, and for providing much-appreciated transportation to SFU every time our lab server failed, which happened more frequently than one would hope. Above all I will thank my parents, Keith Rushton and Susan Pauley, for guidance and support from the start of graduate school up to the final stages, and I apologize for not returning as many of their calls as I intended.

And finally, above all else, I must wholeheartedly thank my mentor, guide, arch nemesis, and senior supervisor Ryan Morin. When I started graduate school in September of 2016, I had next to no bioinformatics experience other than BLAST searches, two courses of basic Java coding, and a mediocre B+ GPA. For reasons that I still do not understand today, he decided to take me on as a graduate student in his lab. For that I am eternally grateful as I can not imagine my life without this opportunity or cancer genomics research. While he may have subjected me to analyze cohorts of questionable quality (which will remain anonymous) resulting in a significantly receded hairline, he has always been a truthful, generally laid-back, and extremely helpful resource to the lab and others.

Table of Contents

Declaration of Committee ii
Ethics Statement iii
Abstract iv
Dedication vi
Acknowledgements vii
Table of Contents ix
List of Tables xii
List of Figures xiii
List of Acronyms xv
Glossary xvii

Chapter 1. Introduction 1
1.1. Cancer development and genetics 1
1.1.1. Driver and passenger mutations 1
1.1.2. Oncogenes and tumour suppressors 2
1.1.3. Tumour development and heterogeneity 3
1.1.4. Hallmarks of cancer and tumour formation 5
1.2. Lymphoma and DLBCL 7
1.2.1. Lymphoma subtypes, and non-Hodgkin lymphoma 8
1.2.2. Molecular classification of DLBCL, Cell of Origin (COO) 9
1.2.3. Treatment of frontline DLBCL 10
1.2.4. Treatment of relapsed-refractory DLBCL 10
1.3. Illumina DNA sequencing 12
1.4. Illumina sequencing approaches 15
1.4.1. Whole genome sequencing 16
1.4.2. Whole exome sequencing 17
1.4.3. Custom captures 17
1.5. Analysis of Illumina sequencing data 18
1.5.1. Read alignment and duplicate marking 19
1.5.2. Quality control 20
1.5.3. Simple somatic mutation detection 21
1.5.4. Copy number variant detection 22
1.5.5. Structural variant detection 24
1.6. Biopsies and cell-free DNA 25
1.6.1. Tumour tissue biopsies 25
1.6.2. Whole blood and buffy coat 26
1.6.3. Blood plasma and cell-free DNA 26
1.6.4. Characteristics of cfDNA and ctDNA 27
1.6.5. Applications of liquid biopsies 30
1.7. Genetics of DLBCL 31
1.7.1. Genetics of diagnostic DLBCL 31
1.7.2. Genetics of molecular subgroups 32

1.7.3. Genetic subgroups 33
1.7.4. Genetics of rrDLBCL 34
1.8. Research Aims and Outline 34

Chapter 2. Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma 35
2.1. Abstract 36
2.2. Introduction 36
2.3. Methods 38
2.3.1. Targeted sequencing and mutational analysis of rrDLBCLs 38
2.3.2. Meta-analysis of untreated DLBCLs 39
2.3.3. Identifying genes associated with rrDLBCL 40
2.3.4. Evaluation of MS4A1 protein expression and antibody reactivity 41
2.3.5. PT255 exome sequencing and single-cell analysis 42
2.4. Supplemental Methods 42
2.4.1. rrDLBCL sample collection 42
2.4.2. Blood processing and DNA extraction 43
2.4.3. Library construction and targeted enrichment 43
2.4.4. Sequence alignment and somatic variant calling 44
2.4.5. Analysis of untreated DLBCL cohort 45
2.4.6. Quality control and validation of untreated cohorts and mutation frequency 45
2.4.7. Genetic subgroupings of rrDLBCL cases 46
2.4.8. Survival analysis in untreated DLBCL 47
2.4.9. Site-directed mutagenesis of *MS4A1* 47
2.4.10. Immunohistochemistry of tissue sections and cell lines 47
2.4.11. Immunoblotting of cells expressing wild-type or mutant CD20 48
2.4.12. Single Cell Analysis of PT255 49
2.5. Results 50
2.5.1. Enrichment of mutations in rrDLBCL 50
2.5.2. Recurrent clonal selection following rituximab-based therapy 54
2.5.3. *KMT2D* and *TP53* mutations are poor prognostic markers in diagnostic DLBCL 57
2.5.4. Differential representation of EZB and MCD subgroups in rrDLBCL 60
2.5.5. Mutations in *MS4A1* attenuate rituximab binding 60
2.5.6. *MS4A1*-harbouring subclones drive rapid treatment resistance 64
2.6. Discussion 68

Chapter 3. Recurrent copy number alterations contribute to a distinct genetic landscape in rrDLBCL 71
3.1. Abstract 71
3.2. Introduction 72
3.3. Methods 75
3.3.1. rrDLBCL sample collection 75
3.3.2. Sample processing, library preparation, and sequencing 75
3.3.3. Read alignment and somatic variant calling 76

3.3.4. Copy number calling 77
3.3.5. Diagnostic cohort 78
3.3.6. Mutation frequency comparison 78
3.3.7. Mutual exclusivity analysis and clustering 79
3.4. Results 79
3.4.1. The exome-wide mutation landscape of rrDLBCL 79
3.4.2. Mutations with prognostic potential in rrDLBCL 81
3.4.3. Recurrent CNVs inform on the biology of rrDLBCL 82
3.4.4. Recurrent CNV drivers and novel events are enriched in rrDLBCL 85
3.5. Discussion 90

Chapter 4. General Discussion 93
4.1. Summary of research findings 93
4.1.1. Mutations in *KMT2D* and *TP53* dominate the landscape of rrDLBCL 93
4.1.2. Mutations in *MS4A1* directly contribute to treatment resistance 94
4.1.3. Recurrent copy number alterations contribute to a unique landscape in rrDLBCL tumours 95
4.2. Implications of research 96
4.2.1. The genetics of rrDLBCL are generally similar to diagnostic DLBCL 96
4.2.2. rrDLBCL is genetically heterogeneous, and there is no single mechanism of R-CHOP resistance 97
4.2.3. Candidate therapeutic targets 97
4.2.4. Mechanisms of treatment resistance 98
4.3. Ongoing work and future directions 99
4.3.1. Serial sampling of rrDLBCL cases 99
4.3.2. The epigenetic and transcriptomic landscape of rrDLBCL 100
4.3.3. Mechanism and impact of *MS4A1* mutations in DLBCL 101
4.3.4. Contribution of non-coding events 101
4.4. Closing perspective 102

References 103

Appendix A. Supplementary Data file associated with Chapter 2 131

Appendix B. Supplementary Data File Associated with Chapter 3 133

List of Tables

Table 1-1-1. Comparison of Illumina sequencing approaches 16
Table 1-2. A brief set of QC issues encountered during Illumina library preparation and sequencing, and their downstream effects on the sequencing data 21
Table 2-1. Patients and samples used in clonal evolution analysis 41

List of Figures

Figure 1-1. Example of tumour heterogeneity, where a tumour is comprised of multiple subclones (colours), and one of these subclones is resistant to treatment (red subclone). 4
Figure 1-2. Hallmarks of cancer. All the phenotypes required by a normal cell to become fully malignant. 6
Figure 1-3. Overview of Illumina sequencing, including (A) library preparation, and (B) Sequencing by synthesis 14
Figure 1-4. A general workflow for analysis of Illumina sequencing data 19
Figure 1-5. Example copy number variant detection using (A) read depth, and (B) B-allele frequency. 23
Figure 1-6. An example structural variant (translocation), and how Illumina sequencing reads will appear when mapped to the reference genome. 25
Figure 1-7. Overview and key differences between (A) tissue biopsies, and (B) liquid biopsies 29
Figure 1-8. Leveraging Unique Molecular Identifiers to perform error correction following Illumina sequencing 30
Figure 2-1. Variant calling workflows for rrDLBCL cases (left), and Untreated DLBCL exome cases (right). 40
Figure 2-2. Mutation landscape of lymphoma-related genes in 135 rrDLBCL cases. 51
Figure 2-3. Differentially mutated genes between rrDLBCL and untreated DLBCL. 52
Figure 2-4. Mutation patterns in genes enriched for mutations within the population of rrDLBCLs. 53
Figure 2-5. Clonal evolution patterns of regression and selection in rrDLBCL. 56
Figure 2-6. Prognostic potential of *KMT2D* and *TP53* mutations in untreated DLBCL. 57
Figure 2-7. Prognostic potential of *TP53* and *KMT2D* truncating mutations in untreated DLBCL. 58
Figure 2-8. Prognostic association of *KMT2D* and/or *TP53* mutations on DLBCL. 59
Figure 2-9. Distribution and functional impact of *MS4A1* mutations in rrDLBCL. 63
Figure 2-10. Comparison of CD20 binding of CHO-S cells transfected with plasmids containing either wild-type or mutant (Y86H, Y86C and L66R) MS4A1. 64
Figure 2-11. Plasma and single-cell sequencing of multiple time points in a DLBCL patient (PT255). 67
Figure 3-1. Cohorts and analysis workflows used for the rrDLBCL CNV cohort, SNV cohort, and merged cohorts. 77
Figure 3-2. Landscape of simple somatic mutations in rrDLBCL, exome-wide, and events significantly differentially perturbed in rrDLBCL. 80
Figure 3-3. The mutational landscape of rrDLBCL, broken down by molecular subgroups. 81
Figure 3-4. Significantly (p_{adj} <0.1) differentially mutated genes between diagnostic DLBCL (blue bars) and rrDLBCL (red bars) within the GCB molecular subgroup 82

Figure 3-5. The landscape of copy number variants across rrDLBCL. 84
Figure 3-6. Landscape of copy number variants across rrDLBCL, subset to ABC (top) and GCB (bottom) cases. 85
Figure 3-7. Significantly differentially perturbed events between diagnostic and rrDLBCL. 87
Figure 3-8. Comparison of copy number events between diagnostic DLBCL (B, D) and rrDLBCL (A,C), specifically within the ABC (A,B) and GCB (C,D) molecular subgroups. 88
Figure 3-9. Patterns of recurrent CNVs across rrDLBCL samples. 89

List of Acronyms

ABC	Activated B-Cell DLBCL
aSHM	Aberrant somatic hypermutation
BAF	B-allele frequency (relevant for SNPs)
CAPP-Seq	Cancer personalized profiling by deep sequencing
CAR-T	Chimeric antigen receptor T-cell therapy
CCF	Cancer cell fraction
CCTG	Canadian Cancer trials group
cfDNA	Cell-free DNA
CHIP-Seq	Chromatin immunoprecipitation sequencing
CNV	Copy number variant
COO	Cell of origin
CTC	Circulating tumour cell
ctDNA	Circulating tumour DNA
DHIT	Double-hit DLBCL
DLBCL	Diffuse large B-Cell lymphoma
dsDNA	Double-stranded DNA
FFPE	Formalin-fixed paraffin embedded
FL	Follicular lymphoma
Gb	Gigabase
GCB	Germinal centre B-cell DLBCL
HDACI	Histone deacetylase inhibitor
HL	Hodgkin lymphoma
IHC	Immunohistochemistry
ILSG	International Lymphoma Study Group
Indel	Small insertion or deletion
IPI	International Prognostic Index
Kb	Kilobases
lpWGS	Low-pass WGS (lpWGS)
mAb	Monoclonal antibody
MAF	Minor allele frequency (relevant for SNPs) (Not to be confused with the Mutation Annotation File format)
Mb	Megabase
MCL	Mantle cell lymphoma

MHC	Major Histocompatibility complex
miRNA	Micro RNA
MRD	Minimal residual disease
NCI	National Cancer Institute
NHL	Non-Hodgkin lymphoma
NK cell	Natural Killer cell
NMF	Non-negative matrix factorization
ORR	Overall response rate (clinical trial)
OS	Overall survival
PCR	Polymerase chain reaction
PFS	Progression-free survival
PR	Partial response
QC	Quality Control
RNA-Seq	RNA Sequencing
ROS	Reactive Oxygen Species
rrDLBCL	Relapsed-refractory DLBCL
SFU	Simon Fraser University
SHM	Somatic hypermutation
SNP	Single nucleotide polymorphism (germline mutation)
SNV	Single nucleotide variant
ssDNA	Single stranded DNA
SSM	Simple somatic mutation
SV	Structural variant
tFL	Transformed Follicular lymphoma
TMA	Tissue Microarray
UBC	University of British Columbia
UNC	Unclassified DLBCL (molecular subgroups)
UTR	Untranslated region (of a gene)
VAF	Variant allele frequency (relevant for SSMs)
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization
WT	Wild-type

Glossary

Adapter dimers	Two Illumina adapters which have ligated directly to each other, without any input DNA
Angiogenesis	Inducing the growth of new blood vessels to support tumour growth
Apoptosis	Programmed cell death
Apoptotic bodies	Portions of a cell's membrane, organelles, and DNA, generated after a cell has undergone apoptosis
B-allele frequency	The state of a heterozygous SNP in a tumour sample (i.e. 1/1 (normal heterozygous), 2/1 (gain of one allele), 0/1 (loss of heterozygosity)
Bridge amplification	When a DNA fragment bound to a flowcell "bends over" and hybridizes to a complimentary oligonucleotide on the flowcell
Buffy coat	A blood fraction comprised of immune cells and platelets
Capture efficiency	Following sequencing, the proportion of reads in a with fall within the desired regions (capture space)
Capture space	The regions of the human genome which are targeted for sequencing
Chromatin	A nucleosome bound to DNA
Circulating tumour cells	Complete tumour cells which have detached from the main tumour and are circulating in the bloodstream
Cluster (sequencing)	A group of DNA fragments bound to the flowcell which all share the same sequence and are physically close to one another in a group
Coding region	A region of the genome encoding a protein sequence
Complete response	A cancer patient for whom treatment is successful, and there are no (or very minor) indication of a tumour
Constitutional DNA	A sample or sequencing data from healthy tissue, usually from the same individual as a tumour sample
Convergent evolution	When two organisms/cells acquire a shared phenotype independently
Coverage	The number of times a specific genomic loci has been sequenced
Cytoband	A named region of a chromosome, named from staining and inspecting the banding patterns of chromosomes during cell division
Discordant read pair	In paired-end sequencing, when one read maps to one genomic locus and the other maps to a distant genomic locus. Indicative of a structural variant

DLBCL90	A gene expression-based assay for classifying FFPE tissue biopsies from DLBCL patients into COO subgroups, identify molecular PMBCL, and double-hit signature cases.
DNA adduct	DNA which is covalently crosslinked to a protein or chemical. Usually damages the DNA
Domain	A portion of a protein with a specific structure and function
Double-hit DLBCL	DLBCL cases harbouring a translocation of *MYC* and one of *BCL2* or *BCL6*, placing them under control of a constitutively activated enhancer.
Driver mutation	A mutation which provides the cell with a selective advantage
Duplicates	Read pairs which originate from the same DNA molecule. See PCR and Optical Duplicates
Euchromatin	“Unpacked” chromatin, transcriptionally accessible
Family (read)	A set of reads which all originate from the same starting (parental) DNA molecule. See duplicates
Fresh Frozen	A tissue biopsy which has been snap frozen in liquid nitrogen and preserved at very low temperatures
GC Bias	Regions of the genome with high AT or GC content which tend to display reduced sequencing coverage (due to PCR amplification biases)
Genome equivalents	The number of “complete” human genome sequences within a sample
Germline variant	A variant shared by all cells in an individual, usually inherited from an individual’s parents
Hallmarks of cancer	A list of all phenotypes a cell must acquire to become fully cancerous
Haploinsufficient tumour suppressor	A tumour suppressor gene where both copies of the gene must remain functional for it to act as a tumour suppressor
Heterochromatin	“Packed”/condensed chromatin, transcriptionally repressed
High molecular weight DNA	Long fragments of cfDNA, originating from necrotic cells
Hot spot (gene)	A region/locus of a gene which is mutated recurrently across many samples
Lane (sequencing)	An individual channel within a flowcell
Library (DNA)	A set of input DNA fragments with sequencing adapters ligated to each end

Linker DNA	The DNA located between two chromatin subunits
Loss of heterozygosity	When a heterozygous SNP in the normal sample is homozygous in the tumour sample
Loss-of-function mutation	A mutation which prevents the protein from performing its normal physiological function, usually through complete inactivation
Lymph2Cx	A gene expression-based assay for classifying FFPE tissue biopsies from DLBCL patients into COO subgroups
Matched normal	A sample/sequencing data from healthy tissue, from the same individual as a tumour sample
Missense mutation	A coding mutation which leads to the change of a single amino acid
Molecular barcode	A short (<14 bp) sequence ligated to the ends of a DNA molecule. Used to identify duplicates
Multiplexing	When multiple samples are sequenced on the same sequencing run
Mutation	A change in an organism's DNA
Non-coding region	A region of the genome which does not directly encode a protein
Nonsense mutation	A coding mutation which truncates the resulting protein
Nucleosome	An octamer of 8 histone proteins, used to bind and condense DNA
Off-target reads	In capture-based sequencing, reads which map to loci outside the desired capture space
Optical Duplicates	Multiple read pairs which originate from the same starting DNA molecule when the sequencer mistakenly identifies a single cluster on the flowcell as multiple clusters
Paired-end sequencing	Sequencing a DNA fragment from both ends
Partial response	A cancer patient for whom therapy is initially successful with notable reduction in tumour size, but the tumour persists, and treatment fails
Passenger mutation	A mutation which does not affect the fitness of the cell
PCR Duplicates	Multiple read pairs which originate from the same starting DNA molecule as a result of the duplication of that molecule via PCR
Plasma	A blood fraction containing proteins, DNA, salts, lipids, and water
Positive selection	A mutation which conveys a phenotypic and selective advantage to a cell/organism, and is selected for in a population

Read (sequencing)	A single DNA sequence output by the sequencing machine
Read depth	The number of reads within a given genomic window (commonly referred to as a bin)
Read mapping	Determine which fragment of an organism's genome a read or read pair represents, usually be comparing the read sequence to a fully assembled version of the reference genome
Reference genome	A "complete" and existing sequence of an organism's genome. Usually used during short read alignment (see read mapping)
Sequencing by synthesis	Where a reversibly terminated DNA nucleoside is integrated into a growing DNA strand during DNA replication, the base is read, and then the terminator is removed, enabling subsequent bases to be sequenced
Shotgun sequencing	Breaking up input DNA into multiple fragments, and sequencing the smaller fragments in parallel
Simple somatic mutation	Single nucleotide variants and small insertions/deletions
Somatic variant	A variant specific to cancerous cells, not found in the constitutional DNA
Splice site mutation	A mutation affecting the splice donor or acceptor sequence of a protein
Split read	A read where one portion is aligned/originates from one locus and the other portion is aligned/originates from a (normally) distant genomic locus. Indicative of structural variant
Stable disease	A cancer patient who is treated, but the tumour shows minimal or no response to therapy
Structural variant	Large-scale genomic alterations which re-organize portions of the organism's genome
Subclone	A population of cells in the tumour sharing the same phenotype
Telomeres	Sacrificial DNA sequences on the ends of human chromosomes, which protect the chromosome during cell division
Tissue biopsy	To physically excise and remove a portion of the tumour
Transmembrane domain	A protein domain which crosses a cellular membrane. Usually contains external hydrophobic residues
Triple-hit DLBCL	DLBCL cases harbouring translocations of *MYC, BCL2*, and *BCL6*, placing them under control of a constitutively active enhancer (see Double-hit DLBCL)

Tumour cell contamination	The presence of tumour cells in a normal/healthy sample
Tumour heterogeneity	Cells within the same tumour harbouring distinct mutation profiles, acquiring unique mutations
Tumour microenvironment	Tumour cells and otherwise normal cells which interact with and support the tumour mass
Tumour purity	The proportion of tumour/malignant cells in a tumour sample
Tumour suppressor	A gene which inhibits tumour growth when fully functional
Tumour-normal pair	Sequencing both a tumour sample and a sample of healthy cells from the same patient
Unmatched normal	A healthy/constitutional sample from a patient other than the tumour sample
Warburg effect	A cancer cell will prioritize glycolysis/lactate pathways for energy production, over oxidative phosphorylation
Whole genome sequencing	The process of sequencing the entire human genome

Chapter 1.

Introduction

1.1. Cancer development and genetics

Fundamentally, cancer is a disease that arises from normal human cells which acquire a set of somatic mutations over many rounds of cell division. Under normal conditions, cells are placed under extensive regulatory control to divide only when necessary (for example, during wound healing[1,2] or human development[3]). Some mutations can bypass these regulatory systems, allowing a cell to achieve uncontrolled cell division or evasion of apoptosis or other anti-cancer defenses within the body. With a sufficient combination of such mutations, clonal populations of cells harbouring such mutations can eventually become cancerous.

1.1.1. Driver and passenger mutations

Cells obtain somatic mutations through errors in DNA replication[4] or through DNA-damaging processes such as environmental exposures (carcinogens)[5–7]. The resulting mutations can exhibit various downstream effects depending on the type of mutation and where they occur. The human genome in diploid cells comprises approximately 6.4 gigabases [Gb] with only small portions directly encoding proteins[8]. While the remaining non-coding portion of the genome does encoding functional regions (for example, regulating gene expression, regulatory RNAs) our ability to predict the function of mutations in these regions is relatively limited[9,10]. As most mutational processes introduce mutations at semi-arbitrary locations[11], most somatic mutations acquired by a cell are so-called “passenger” events, as they have no effect on the fitness of that cell relative to its normal counterpart. Those mutations with a functional impact (for example, those that alter the function of a protein) can either benefit a cancer cell or can be detrimental. By chance, most will be detrimental to an important cellular process and would result in a fitness disadvantage, thereby undergoing negative selection. However, in rare cases mutations will benefit cellular fitness and provide the cell with a competitive advantage relative to other cells in the same tissue, and these would

undergo positive selection. Such events are termed driver mutations and are key to tumour development and cancer progression. Importantly, although these mutations are beneficial to the cell in the context of Darwinian evolution, such behaviour can be detrimental to the multicellular organism. Tumours generally acquire few driver mutations but harbour orders of magnitude more passenger mutations[12]. It should be noted that the functional relevance of mutations is dependent, in part, on the biological context such as cell differentiation state, microenvironment or environmental factors. As a result, a “passenger” mutation may act as a driver in another context. For instance, selective pressure may only exist temporarily such as during exposure to therapies. Driver mutations generally contribute to the fitness of cells by modifying either the function or dosage of protein-coding genes. While these generally manifest as protein-coding changes, driver copy number alterations leading to increases or decreases in gene dosage, as well as other regulatory mutations that affect protein abundance, have also been observed [13–15].

1.1.2. Oncogenes and tumour suppressors

Cancer-associated genes are commonly categorized as either tumour suppressor genes or oncogenes and tend to display distinct patterns of mutations in cancerous cells. If a gene encodes a protein that negatively regulates cell growth and division (for example, by preventing the cell from dividing when in close contact with other cells) and is predominantly affected by loss-of-function mutations, it is generally considered a tumour suppressor gene[16]. Typically, mutations in tumour suppressor genes introduce a premature stop codon in the protein (nonsense mutation); change the reading frame (insertion or deletion mutation, indel); or prevent the resulting transcript from being properly spliced (splice site mutation). Because diploid cells normally have two copies of autosomal genes, tumour cells typically acquire driver mutations perturbing both copies of a tumour suppressor gene. Tumour suppressor genes with haploinsufficiency been described but are relatively rare[17,18]. Furthermore, certain loss-of-function mutations can have a dominant negative effect, disrupting protein function even when a functional copy of the gene remains[19]. These events tend to occur in proteins which form protein complexes with multiple subunits, where a non-functional molecule of the protein can affect the function of complex. Mutations displaying this

phenotype tend to disrupt a key amino acid (missense mutation), typically those involved in the enzyme's active site.

In contrast to tumour suppressors, the activity of proteins encoded by oncogenes drive cellular growth or inhibit apoptosis. Because randomly acquired mutations are usually deleterious, mutations within oncogenes tend to be localized to specific amino acids or domains that promote sustained activity. These can include mutations which disrupt inhibitory domains, modify post translocation modifications, or enhance protein activity directly[20,21] . Regulatory mutations that directly increase the expression of oncogenes are also common, particularly in lymphomas. For instance, genomic translocations which place anti-apoptotic genes (such as *BCL2*) or pro-proliferation genes (such as *MYC*) under control of a constitutively active enhancer in B-cells results in significantly increased and constitutive expression of these genes.

1.1.3. Tumour development and heterogeneity

Cancers represent a population of cells arising from a single founding cell through the stepwise acquisition of mutations while under selective pressure. When a cell acquires a malignant mutation, it continues to grow and divide until the cell encounters another checkpoint which inhibits growth and division. Benign (or premalignant) populations can persist until a subsequent daughter cell acquires an additional mutation that enables it to avoid this new restriction, enabling further growth and division until another proliferative block is encountered. This process continues until a fully formed malignant tumour develops[22], able to grow without restriction and metastasize to different sites in the body. Solid tumours commonly recruit supporting cells from the immune system such as regulatory T-Cells and modulated dendritic cells to evade immune destruction[23,24], as well as stromal cells such as fibroblasts, which further promote immune suppression, produce oncogenic growth factors and induces angiogenesis, and supports metastases[25]. However, even an early tumour is comprised of hundreds of thousands or even millions of cells[26], and different cells in the tumour can acquire different mutations[27] (Figure 1-1). This tumour heterogeneity has profound biological and clinical implications. From a biological perspective, these subclonal populations of cells (termed "subclones") compete amongst themselves for limiting supplies of nutrients and oxygen in their microenvironment. If a subclone harbours a beneficial mutation, it can outcompete and eliminate other subclones in a classic

selective sweep[28]. In rare cases, subclones may actually cooperate with one another, acquiring synergistic phenotypes which further advance tumour development and metastases[29,30]. Tumour heterogeneity can also arise from differences in the environmental conditions of different tumour niches. For instance, the center of a tumour is generally deficient of oxygen and nutrients, whereas cells near the tumour's periphery tend to have abundant nutrients and oxygen but are exposed to the body's immune system, thereby having distinct selective pressures within the same population.

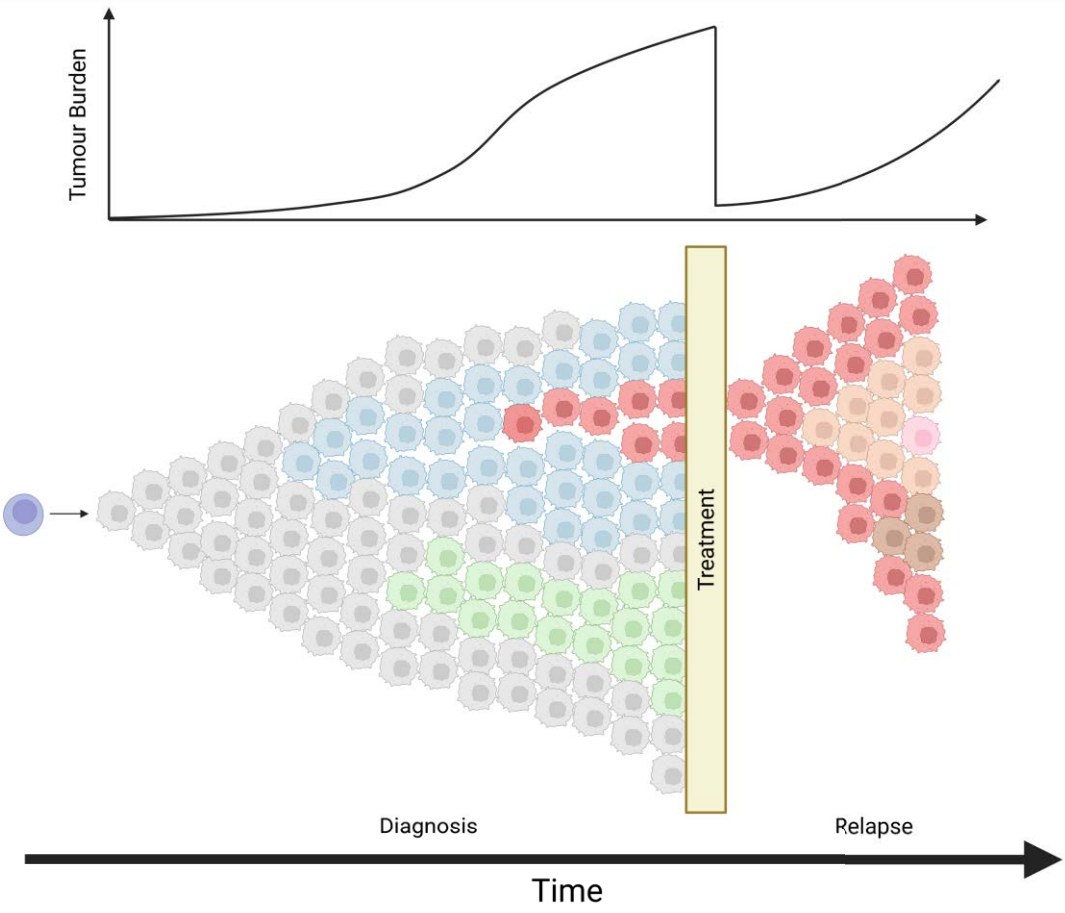


Figure 1-1. Example of tumour heterogeneity, where a tumour is comprised of multiple subclones (colours), and one of these subclones is resistant to treatment (red subclone). This subclone subsequently expands and becomes the dominant clone.

Tumour heterogeneity and the presence of multiple genetically and phenotypically distinct subclones is also relevant in the context of treatment. As most therapeutics target proteins which tumour cells depend upon[31] or features required by rapidly growing cells[32,33], this treatment constitutes a strong selective pressure on the tumour. Subclonal populations may harbour mutations that render them resistant to the

chosen therapy and, upon exposure, this subclone can persist while the bulk of the tumour cells are killed, which can be perceived as an initial response to treatment. With a reduction of competing cells, this resistant subclone can theoretically thrive among abundant resources, allowing it to generate a tumour now entirely resistant to therapy. Tumour heterogeneity and selection of a resistant subclone are strongly associated with treatment failure[34–36].

1.1.4. Hallmarks of cancer and tumour formation

Numerous biological processes and checkpoints act on a cell to prevent it from undergoing unregulated cell division. To become fully malignant, a normal cell must acquire several phenotypes to avoid these restrictions, with the comprehensive list of the phenotypes required termed the “hallmarks of cancer” and reviewed by Hanahan and Weinberg[37,38] (Figure 1-2). In brief, there are ten major attributes a cell must acquire to become malignant. Cell division is tightly regulated by sustained negative feedback loops and anti-proliferative signals which prevent cell division, and expression of pro-proliferative signals is tightly controlled. Cells which acquire mutations which enable constitutive proliferation signaling (for instance, via the Ras signaling[39]) and ignore anti-proliferative signals (for instance, by blocking the TFG-β signaling pathway[40]) enable them to grow and divide without the usual external stimulation.

Each time a cell undergoes cell division, the entire genome of the cell must be replicated. As DNA polymerase requires a primer to begin replication, every subsequent round of cell division shortens the genome[41]. Human genomes are protected by telomeres, sacrificial DNA sequences on the end of each chromosome which are shorted during cell division[42]. When telomeres become too short (i.e. if the cell has undergone too many rounds of cell division), cell division either ceases, or genomic damage occurs[43]. To avoid this biological limitation, malignant cells must acquire “replicative immortality” by reactivating the telomerase *TERT*[44] which lengthens the telomere sequences.

As malignant cells continue to acquire driver mutations and the tumour continues to increase in size, it eventually outstrips its blood supply and encounters a deficit of oxygen and nutrients. To account for this, a tumour tends to both promote the growth of new blood vessels (inducing angiogenesis)[45] to further supply the tumour, and modifies

the cell's underlying energetics and metabolism. Normal mammalian cells generally produce energy from glucose and oxygen through oxidative phosphorylation, although some cell types leverage anaerobic glycolysis to produce energy under hypoxic conditions[46]. The latter process generates significant less energy than oxidative phosphorylation[47], but has the additional effect of generating numerous metabolites used by cellular components[48,49]. Even with functioning mitochondria and when abundant oxygen is present, malignant cells tend to prioritize anerobic glycolysis over oxidative phosphorylation, as these additional metabolites are required for continued cell division.

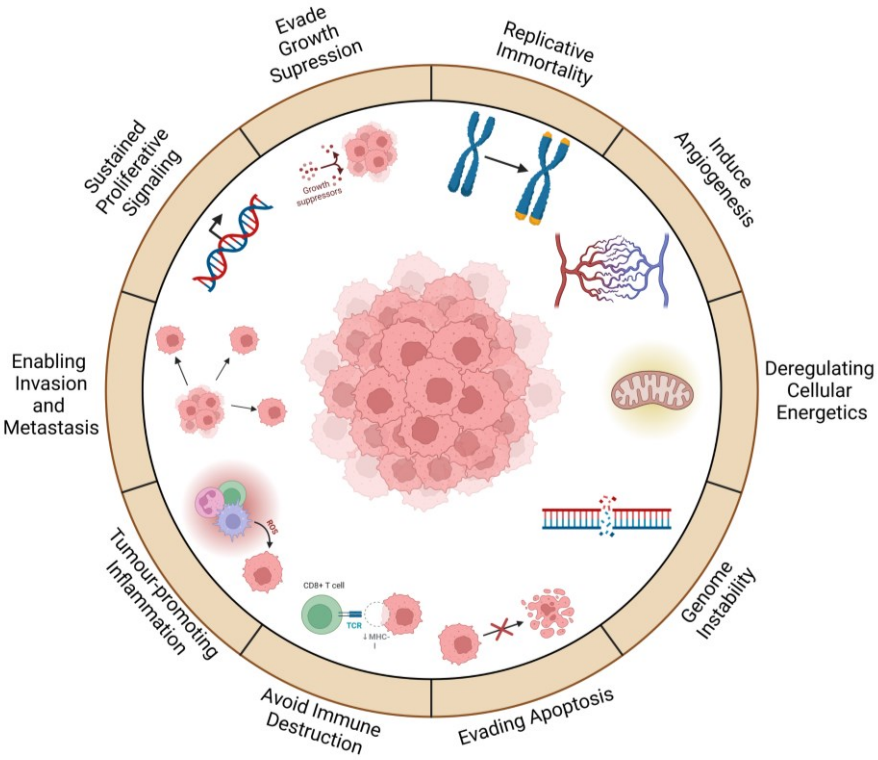


Figure 1-2. Hallmarks of cancer. All the phenotypes required by a normal cell to become fully malignant. Adapted from Hanahan and Weinberg (2011).

Tumour cells require multiple driver mutations to become fully malignant[12], but human cells are host to extensive DNA repair[50] and carcinogen-processing pathways[51] which ensure a low mutation rate[52]. As a higher mutation rate increases the likelihood of acquiring a beneficial driver mutation, cancer cells frequently inactivate pathways responsible for DNA repair and genomic stability[53] and become more suspectable to mutagens and errors during DNA replication[54]. Many pathways and proteins involved in

DNA repair and maintenance also act as genomic guardians, inducing apoptosis if catastrophic DNA damage occurs[55]. Apoptosis can also be induced when constitutive proliferative signaling occurs as a further safeguard against tumour formation[56]. To avoid this, many apoptotic pathways are perturbed by loss-of-function driver mutations in malignant cells. Chief among these is the master DNA damage and apoptotic regulator, and "guardian of the genome" *TP53*, which is mutated or otherwise inactivated in half of all human cancers[57].

As tumour cells acquire non-synonymous mutations, both driver and passenger, they generate mutant proteins not normally present in the body. Almost all human cell types present protein peptides on their cellular surface in the Major Histocompatibility Complex (MHC) Class 1 complex[58]. These presented peptides are recognized by CD8+ T-Cells which induce destruction of cells presenting foreign proteins (for instance, those infected by viruses or cancerous cells). While downregulation of MHC Class 1 is common in cancer cells to evade immune destruction[59,60], Natural Killer (NK) cells destroy cells which don't present MHC Class 1[61] which is common in foreign invaders such as bacteria. To avoid further immune destruction, tumours generally recruit cells which supress the normal immune response, such as regulatory T-cells[62] and those which promote inflammation of the tumour site[23,24]. This inflammation has the further effect of immune suppression through the release of specific cytokines by pro-inflammatory cells and by producing factors associated with cell growth and angiogenesis[63]. It should be noted that due to the generally random nature of driver mutations, these hallmarks may not be acquired in a specific order, and a single driver event can provide multiple hallmarks.

1.2. Lymphoma and DLBCL

Lymphomas are a type of cancer which arise from cells involved in the immune system, namely lymphocytes, and form a tumour within the lymphatic system. Lymphomas are the 5th most common form of cancer within Canada, with an estimated 12,150 Canadians diagnosed with lymphoma each year[64]. Despite its high incidence, the rate of lymphoma mortality in Canada has notably declined over the past two decades, reflecting improvements in disease classification and treatment[64]. However, lymphoma remains the 8th deadliest cancer in Canada, especially among children, accounting for a

an estimated 3.5% of all cancer-associated deaths. There remains significant room to improve in both disease management and treatment.

1.2.1. Lymphoma subtypes, and non-Hodgkin lymphoma

At the broadest level, lymphomas can be divided into three major subgroups: Hodgkin Lymphoma, mature T- and NK-cell neoplasms, and B-cell neoplasms. Hodgkin lymphoma, a cancer arising from B-Cells and characterized by the presence of Reid-Sternberg cells in the tumour, is generally rare, accounting for 0.4% of all cancer cases diagnosed worldwide[65] and only 10% of lymphoma cases[66]. Of the remaining non-Hodgkin lymphomas (NHLs), mature T- and NK-Cell neoplasms include lymphomas derived from T-lymphocytes and natural killer cells which have migrated out of the thymus, and account for 10-12% of all lymphoid malignancies[67,68]. The remaining NHLs largely arise from B-Cells[69]. B-cell lymphomas are extremely diverse, with 46 distinct recognized subtypes[70] but can broadly be divided into low-grade/indolent and high-grade/aggressive B-Cell lymphoma. Low-grade lymphomas generally grow slowly, and as such patients with indolent lymphomas generally have superior outcomes compared to those with aggressive lymphomas. However, while indolent lymphomas are initially responsive to therapies[71] these tumours generally persist following treatment and gradually becomes treatment resistant. The most common subtype of indolent lymphoma (and the second most common type of NHL) is follicular lymphoma (FL), which is characterized by a hallmark translocation which places the anti-apoptotic and proto-oncogene *BCL2* under the control of an immunoglobulin enhancer[72] and arise from B-cells within follicles of the germinal center[73]. While patients with follicular lymphoma (FL) is generally not fatal in-and-of-itself, with 5-year overall survival (OS) of 92%[73], FL can undergo histological transformation to more aggressive forms of lymphoma.

In contrast to the slow growth and long outcomes of indolent lymphomas, high-grade/aggressive lymphomas are characterized by rapidly growing and generally aggressive tumours and inferior patient outcomes if left untreated. Compared to the persistence of indolent lymphomas, high-grade lymphomas are generally responsive to therapy. These include Burkitt lymphoma, characterized by translocations placing the proto-oncogene *MYC* under control of the immunoglobulin enhancer, and Mantle cell lymphoma, which arise from cells in the mantle zone of the germinal center and is characterized by translocations of the cell cycle regulator *CCND1*[74,75].

The most common type of NHL is termed diffuse large B-cell lymphoma (DLBCL), representing 40% of all NHL cases. DLBCL is so named due to the morphological characterization of the disease, where large lymphoid cells diffuse and displace normal lymphoid tissue. While distinct subtypes of DLBCL exist corresponding to associated viral infection (HHV8+ DLBCL[76] and EBV+ DLBCL[77]), most cases are classified as DLBCL, not otherwise specified (NOS).

1.2.2. Molecular classification of DLBCL, Cell of Origin (COO)

In 2000, Alizadeh *et al* discovered a gene expression signature which divided DLBCL cases into two molecular subgroups[78]. The first, termed Germinal Center B-Cell DLBCL (GCB), showed elevated expression of genes expressed by B-Cells within the germinal center of lymphoid tissues, and thus thought to arise from germinal center B-cells. The second, termed Activated B-Cell (ABC) DLBCL, shows expression of genes associated with plasma cells[79] and NF-κB signaling[80], and thus thought to arise from B-Cells which are in the process of leaving the germinal center and differentiating into plasmablasts. ABC-DLBCL cases shows significantly inferior outcomes when treated with chemotherapy compared to GCB-DLBCL[78], and thus these molecular subtypes are a predictive marker of treatment outcome. The prevalence of ABC and GCB-DLBCL varies significantly with geographical location, with GCB-DLBCL more common than ABC-DLBCL in North America and many European countries, while ABC-DLBCL is more prevalent in Asian and Pacific nations[81].

Initially, molecular classification of DLBCL samples into these prognostic subgroups was met with difficulty, as gene expression profiling required fresh frozen tissue biopsies. Tumour biopsies are generally stored via formalin fixation, which fragments RNA and renders gene expression profiling difficult[82]. To avoid this issue and assign tumours into molecular subgroups using FFPE tissue, several groups developed Immunohistochemical methods of assigning cell of origin (COO) via the presence or absence of specific proteins within cancerous cells[83–85]. Chief among these is the Hans algorithm, which classifies samples into GCB or non-GCB based on the presence or absence of CD10, BCL6, and MUM1[83]. However, these classification approaches had notable limitations, such as poor inter-site reproducibility[86] and the binary classification of any sample into GCB or ABC subgroups with no intermediate/unclassified group for samples not expressing any associated signature. To address these limitations, Scott *et*

al.[87] developed the Lymph2Cx assay, which calculates a COO-confidence score using the expression of 20 genes from FFPE tissue[88] using a digital gene expression technique called Nanostring[89]. Lymph2Cx showed high reproducibility between labs and recapitulated an Unclassified group for samples presenting neither gene expression signature. Lymph2Cx has subsequently been expanded to classify DLBCL into other morphological entities[90,91], with the most recent iteration being termed DLBCL90.

1.2.3. Treatment of frontline DLBCL

Until 2002, patients diagnosed with DLBCL were treated in the frontline setting with a chemotherapy combination known as CHOP. CHOP consists of cyclophosphamide, doxorubicin, vincristine, and prednisone, which cumulatively inhibit DNA synthesis, mitosis, and modulate the tumour microenvironment[32,92–96]. This regimen was standard-of-care for DLBCL until 2002, when Coiffier *et al*.[97] showed significant improvements in elderly patients suffering from DLBCL via the addition of the monoclonal antibody rituximab. This antibody interacts with the cell surface marker CD20, which is present on all mature B-Cells[98]. Rituximab + CHOP (R-CHOP) has been standard-of-care for all DLBCL despite DLBCL's heterogeneity, with 5-year OS of 55-65%[99–101]. Treatment failure and relapse following R-CHOP tend to occur within two years of treatment[102], with patients disease-free after two years displaying comparable overall survival to that of the general population[103]. While numerous studies have attempted to improve upon R-CHOP via the addition of novel agents, none have shown significantly improved patient outcomes, and thus most variants to date have attempted to reduce toxicity while maintaining overall response rate[104–107].

1.2.4. Treatment of relapsed-refractory DLBCL

Although R-CHOP is effective for 60-70% of DLBCL cases, for patients where R-CHOP is ineffective and relapsed disease develops (relapsed-refractory DLBCL, rrDLBCL), prognosis is generally poor. This is especially true for cases which are refractory to frontline therapy (relapse within 12 months), with salvage therapy response rates of 13-23%, and median overall survival of 6.3 months[108]. In 2014, the Canadian cancer trials group (CCTG) established Rituximab plus GDP (R-GDP) as a standard salvage therapy option for rrDLBCL[109]. This is comprised of gemcitabine, a DNA nucleoside analog which incorporates itself into a growing DNA strand and prevents

further elongation[110], dexamethasone, a corticosteroid[111], and cisplatin, which generates DNA-DNA and DNA-protein crosslinks preventing cell division and inducing apoptosis[33]. While R-GDP displayed an overall response rate (ORR) of 45.1%, only 13% of patients displayed a complete response to therapy, and the majority of cases relapse[109].

To address the poor performance of current rrDLBCL therapies, a plethora of salvage therapies are currently undergoing investigation for rrDLBCL. Many of these are molecularly targeted agents perturbing key components for lymphoma survival. For instance, tazemetostat is an *EZH2* inhibitor which has shown promise for rrDLBCL cases harbouring *EZH2* activating mutations[112]. Ibrutinib is an inhibitor of the protein BTK, which is critical for B-cell receptor-mediated NF-κB signaling. Given the limited efficacy of mono-agent therapies for rrDLBCL, many subsequent studies have combined multiple targeted agents, both with and without anti-CD20 antibodies and chemotherapy. A phase II study ibrutinib, rituximab, and lenalidomide, which inhibits the B-cell regulators IKZF1 and IKZF3, showed an ORR of 65% and a CR rate of 41% in ABC-rrDLBCL cases[113]. VIPOR is a phase II trial combining ibrutinib, prednisone, obinutuzumab (anti-CD20 monoclonal antibody), lenalidomide, and venetoclax (an inhibitor of the anti-apoptotic protein BCL2), with an ORR of 86% and CR of 68% in relapsed DLBCL cases, and 52% and 29% in refractory DLBCL[114]. Additional therapies leveraging bivalent antibodies (BITE) which bring together malignant cells and T-Cells resulting in T-Cell activation have also been explored. Glofitamab is a bivalent anti-CD20/CD3 antibody which, in a phase I trial of 171 rrDLBCL cases, showed an ORR of 53.8% and CR of 36.8%[115].

Currently, the most promising rrDLBCL treatment uses a patient's own T-Cells and genetically modifies them to produce a chimeric antigen receptor (CAR) containing both a domain specific for a feature expressed by tumour cells (ex. CD20), and a domain that activates the T-cell[116–118]. These modified T-cells are then provided to the patient intravenously and, upon recognizing tumour antigens, further expand and proliferate. Not only do these CAR-T cells directly destroy malignant cells, but they can circulate and destroy metastatic tumour, and persist as memory cells[118]. This therapy, called CAR-T, has been explored extensively for lymphoid malignancies[117,119,120], including rrDLBCL[121–123], with promising efficacy. For instance, a phase 1 trial exploring an anti-CD19 CAR-T in 93 rrDLBCL cases showed an ORR of 52% and CR of 40%[123]. This is especially promising given the enrollment criteria for CAR-T clinical trials generally require patients

to have undergone not only frontline therapy (R-CHOP), but also multiple rounds of salvage therapy. However, as CAR-T cell therapy is personalized by case, it is extremely expensive, with CAR-T production alone costing $450,000 for every patient[124]. Thus, CAR-T therapy is commonly used as a last resort when all other treatment options are unviable. While several groups are attempting to reduce the cost of CAR-T therapy, more cost-effective therapies are required for widespread use.

1.3. Illumina DNA sequencing

To determine the sequence of a given strand of DNA, several methods of DNA sequencing have been developed, with the continuing goals of lowering cost, increasing throughput, and increasing the length of DNA that can be sequenced. Modern sequencing techniques enable tumour samples to be sequenced comprehensively and cost-effectively, allowing driver and novel mutations in tumour genomes to be uncovered. Currently, the most common method of DNA sequencing is Illumina sequencing, which leverages a sequencing by synthesis approach to determine the sequence of small DNA fragments.

To sequence a sample using Illumina sequencing,[125] input DNA is first fragmented to ~150-600bp fragments[125] (fragmenting),and the ends of resulting DNA fragments are repaired to create blunt-end molecules (end repair). An additional adenosine nucleotide is then added to the 3’ end of the dsDNA (A-tailing) which enables additional fixed dsDNA sequence (adapters) to be ligated to both ends of each fragment (ligation). These adapters contain several important sequences for downstream steps. The end product of this is commonly referred to as a DNA library (Figure 1-3A).

Several optional steps are commonly performed on a library prior to sequencing, tailored for the downstream application and type of input. One common clean-up step selects for DNA fragments within a given size range (size selection), removing adapter dimers (i.e. adapters which have ligated to each other without any source DNA between them) and very short DNA fragments. PCR amplification using adapter-specific primers is also common to select for DNA fragments ligated to adapters, although this introduces PCR-amplification biases and is generally avoided for samples with sufficient input DNA[126,127]. For DNA extracted from formalin-fixed, paraffin embedded (FFPE) tissue blocks, library preparation and PCR often introduces C/G -> T/A DNA damage

artifacts[128] through deamination of cytosine bases to uracil[129]. Thus, some protocols include treatment with a uracil DNA glycosylase to remove uracil bases prior to PCR amplification[130,131]. For studies focused on specific regions of the genome, hybridization-capture (hybrid-capture) of DNA fragments corresponding to given regions of the genome is also performed (explained in more detail in 1.4.2).

To obtain the sequence of a DNA library, the library is first passed over a glass slide (termed a flowcell) which harbour bound oligonucleotides complimentary to the adapter sequences (termed the P5 and P7 sequences, on the very end of the adapters). When passed over the flowcell, the DNA library hybridizes to the flowcell, and the complimentary strand is synthesized to generate a DNA fragment physically bound to the flowcell. The original (and physically not attached) template is then washed away. As the opposite end of this DNA fragment contains the adapter and P5/P7 sequence, which is complimentary to the flowcell oligonucleotides, these synthesized DNA fragments tend to bend over and hybridize to another bound oligonucleotide, which is then synthesized again to yield a second DNA molecule with an identical sequence (minus any PCR errors). This process is termed bridge amplification and is repeated numerous times to form a cluster of DNA molecules on the flowcell close together with an “identical” DNA sequence (Figure 1-3B).

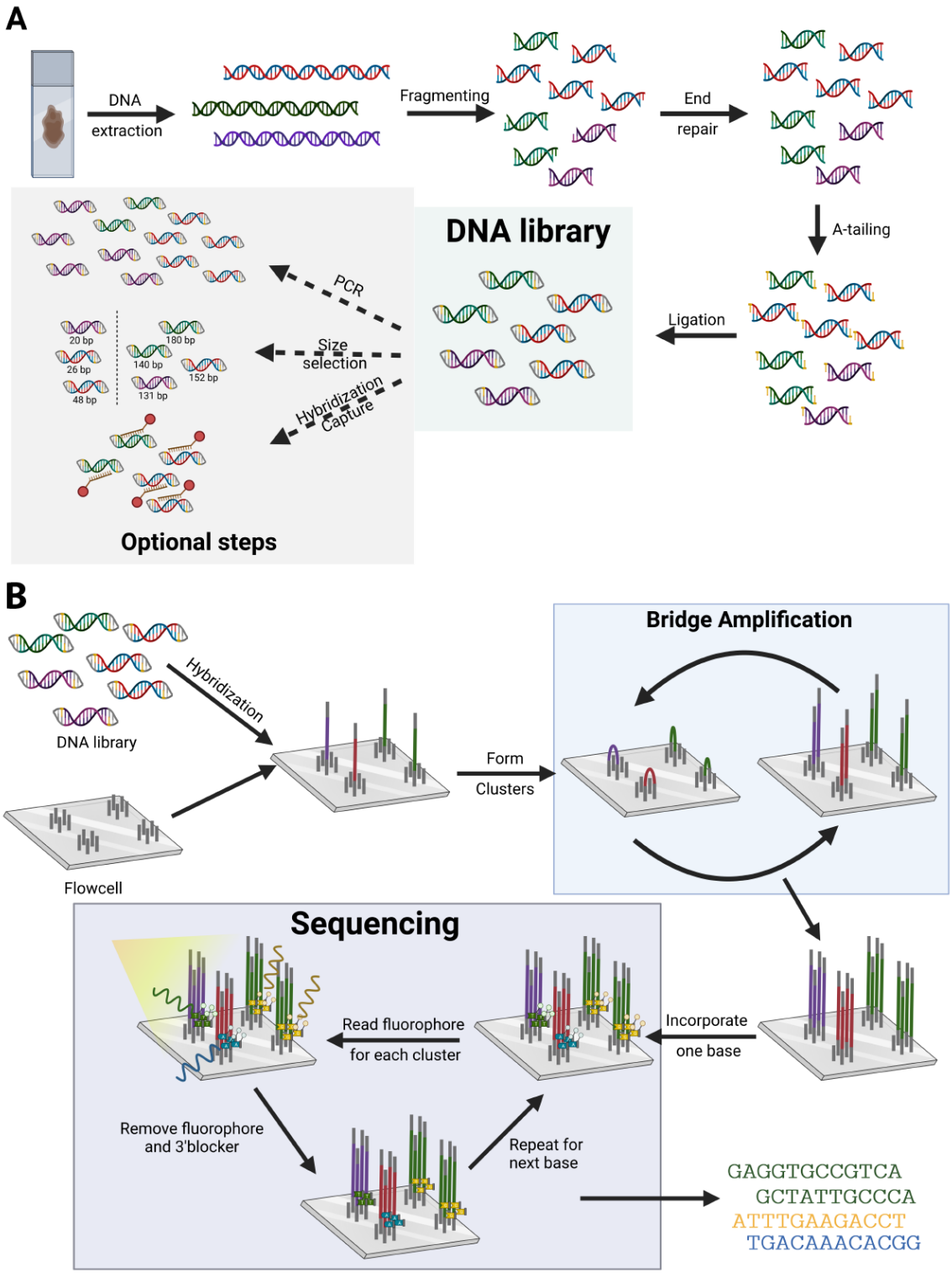


Figure 1-3. Overview of Illumina sequencing, including (A) library preparation, and (B) Sequencing by synthesis

To sequence these individual clusters, and thus obtain a sequence for the original DNA fragment, a DNA polymerase and four types of modified nucleotides are passed over the flowcell. These nucleotides harbour both a reversible terminator bound to the 3'hydroxyl group and a fluorophore/linker combination, with a different colour fluorophore corresponding to each base. A single nucleotide is then incorporated into each growing daughter strand and replication ceases. The leftover nucleotides and DNA polymerase are then washed away, and the fluorescence of each cluster is read by the sequencing machine. To sequence the next base, the reversible terminator and fluorophore are removed, and the above process is repeated for the next base. This process is repeated until the desired number of bases (commonly referred to as read length) is obtained.

Illumina sequencing processes several advantages compared to Sanger sequencing, due to the shotgun sequencing approach where smaller DNA fragments are sequenced in parallel compared to a single larger molecule. While this approach drastically improves sequencing throughput and reduces costs by orders of magnitude[132] while maintaining sequencing accuracy, it comes with notable disadvantages, mainly the limited size of DNA fragments which can be sequenced. While improvements to Illumina sequencing and the advent of paired-end sequencing (where both ends of the DNA fragment are sequenced) have improved fragment size[133], the ~500 base pair limit of paired-end sequencing remains a major limitation for many downstream applications, which new sequencing approaches such as nanopore sequencing have attempted to address.

1.4. Illumina sequencing approaches

Illumina sequencing is currently the most common sequencing technique for human genomic studies. As different studies have different objectives and sources of DNA, variants of Illumina sequencing have been devised to balance sequencing depth (i.e. the number of times a single base was sequenced [fold-coverage]), sequence breadth (i.e. how many bases were sequenced [capture regions]) and cost. In cancer genomics studies, it is common both to sequence a tumour sample and a sample prepared from healthy cells from the same individual (commonly referred to as a tumour-normal pair) to distinguish germline and somatic variants[134]

1.4.1. Whole genome sequencing

The original method of sequencing involves sequencing all DNA extracted from a given source. In the case of human cells, this would consist of the entire human nuclear genome (hence whole genome sequencing [WGS]), mitochondrial genome, and any viral sequences integrated into host genome. The resulting sequencing reads thus cover all coding and non-coding regions, enabling comprehensive downstream analysis and identification of driver and passenger mutations genome-wide. However, given the large size of the human genome, the number of reads required to sequence the whole genome are relatively high. Samples which undergo WGS are subsequently limited in sequencing depth, with 80x fold-coverage common for tumour samples and 40x fold-coverage common for matched normals/constitutional samples[135]. Despite this limited depth, WGS still costs around 5500 CAD for such a tumour normal pair[135], with billions of read pairs required per sample.

One alternative method of WGS while reducing the associated costs is to sequence the entire genome extremely shallowly (usually <=1x average fold-coverage). While the resulting low-pass WGS (lpWGS) data is coverage-sparse and unusable for identifying somatic SNVs or *de-novo* identification of germline single nucleotide polymorphisms (SNPs), it has shown utility in identifying cases harbouring previously known SNPs[136], as well as identifying somatic and germline copy number variants (CNVs)[137–139], although the limited resolution prohibits the detection of relatively focal copy number events.

Table 1-1-1. Comparison of Illumina sequencing approaches

Sequencing type	What is it?	Benefits	Limitations
Whole genome sequencing (WGS)	Sequence the entire human genome	- Sequencing data available for all coding and non-coding regions in the genome - Enables comprehensive downstream analyses	- Relatively expensive - Issues with repetitive regions (Telomeres, centromeres etc.)
Whole exome sequencing (WES)	Sequence only protein-coding regions	- Significantly cheaper than WGS	- Limited information on non-coding regions - Introduces capture biases in targeted regions
Custom capture	Sequence only regions of interest	- Increased sequencing target flexibility	- Limited information on regions not sequenced

		- Decreased sequencing costs compared to WES	- Introduces capture biases in targeted regions - Substantially higher off-target sequencing rate - Custom probes are generally expensive

1.4.2. Whole exome sequencing

In the vast majority of cases, proteins are the effector molecules of the genome, but protein-coding sequences comprise a minority of the human genome[8]. Thus, many sequencing studies only sequence protein-coding DNA (i.e. the "exome"). This is generally achieved using a hybridization-capture approach[140,141], where DNA baits complimentary to the sequences of interest are hybridized with the DNA library. The baits are covalently bound to a magnetic bead[142], and when this mixture is passed across a magnet, only the baits (and complimentary DNA) are bound. The DNA is then eluted from the beads and sequenced. While whole exome sequencing (WES) is significantly cheaper than WGS[143,144] despite generally higher sequencing coverage (~100x), WES comes with many limitations. Notwithstanding the obvious limitation of not sequencing non-coding regions, WES introduces biases in sequencing coverage due to differences in probe binding affinity for its target template[145], as well as biases due to the GC content of the sequenced regions[146]. It should also be noted that, while many commercial exome kits exist, not all exomes are created equal. Some kits specifically including baits for the untranslated regions (UTRs) flanking protein-coding genes, and differences in the design of each bait will notably influence the inter-probe coverage bias.

1.4.3. Custom captures

For studies interested in only a handful of genes or regions, custom captures can further reduce sequence costs and/or increase sequencing depth compared to WES. This requires a custom set of capture probes to be designed for the regions of interest, with numerous commercial solutions available (albeit at a non-trivial cost). Compared to WES, custom captures allow for substantially greater flexibility, allowing any genomic region of interest (coding or non-coding) to be sequenced. However, in addition to all the

limitations of WES, whose limitations are often exacerbated due to the smaller capture regions, the capture efficiency (i.e. the ratio of reads falling within the capture regions compared to non-captured regions) is often lower than WES due to the smaller capture space. For ultra-deep sequencing applications, a double-capture, where two rounds of hybridization-capture are performed, can be used to improve capture efficiency[147].

1.5. Analysis of Illumina sequencing data

Following sequencing, Illumina sequencers produce one (single-end sequencing) or two (paired-end sequencing, most common) output file(s) containing the basecalls and associated quality scores for each cluster (henceforth read). While an assortment of bioinformatics workflows are available for human genomic studies, the following sections will briefly outline a general workflow for most cancer genomics applications[148]: This includes mapping reads against a reference genome, identifying and flagging duplicate reads, and downstream quality control. This is followed by identification of single nucleotide variants (SNVs) and small insertions/deletions (indels), identification of copy number variants (CNVs) and finally identification of large-scale structural variants (SVs) (Figure 1-4).

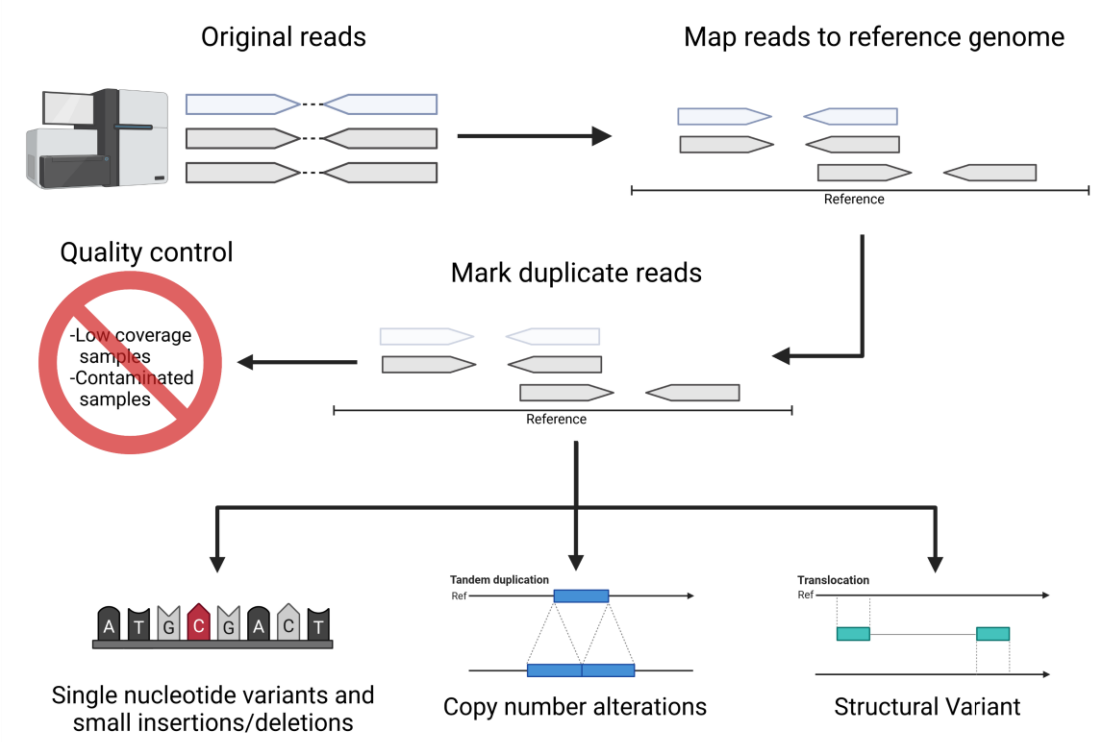


Figure 1-4. A general workflow for analysis of Illumina sequencing data

1.5.1. Read alignment and duplicate marking

As shotgun-based sequencing approaches fragment the host genome in small segments (<250bp), the first step of any analysis is to determine which fragments of DNA represent which portion of the genome. While *de novo* assembly, where overlapping reads are stitched together to form a longer contig, is possible with Illumina sequencing, the numerous interspersed repetitive sequences across the human genome[149] and the short length of Illumina reads render even a partial genome assembly extremely difficult[150,151]. An alternative solution is to compare the sequence of each read against an existing fully assembled version of the genome and determine where that read (or read pair) originates based on sequence similarity. This process is referred to as read mapping, and utilizes existing human genome assemblies (GRCh37, GRCh38, T2T-CHM13) (referred to as a reference genome) generated using longer read sequencing technologies. A plethora of tools exist for read mapping, with the Burrows-Wheeler aligner (BWA)[152] and minimap2[153] commonly used.

During the process of Illumina library preparation and sequencing, some DNA molecules may be sequenced multiple times. These duplicate reads originate from 1)

Multiple PCR copies of a DNA fragment binding to the flowcell and being sequenced (PCR duplicates, most common), or 2) A single cluster on the flowcell being erroneously recognized as multiple clusters (optical duplicates, usually rare). To avoid overrepresentation of duplicated sequences in downstream analysis, duplicates are flagged and ignored, so that a single sequencing read (pair) corresponds to a single DNA molecule in the original library. While numerous tools for marking duplicate reads exist[154], the Picard toolkit's MarkDuplicates tool is extensively used for duplicate removal.

1.5.2. Quality control

Countless issues can affect the quality of Illumina sequencing data (Table 1-2); however, the two most common quality control (QC) issues encountered are samples prepared from formalin-fixed, paraffin-embedded (FFPE) tissue blocks, and sequencing of samples with limited input DNA. Fixation of tissue within paraffin blocks has been performed by pathologists and clinicians since the 19th century, as it allows tissues to be stored almost indefinitely at room temperature with minimal degregation[155]. However, formalin fixation reduces the amount of DNA that can be sequenced[156] and induces C/G -> T/A transitions which may erroneously be interpreted as real mutations[128]. Samples with limited amounts of input DNA require several rounds of PCR to generate sufficient input for sequencing; thus the duplicate rate of these samples tends to be higher, resulting in reduced effective sequencing coverage[157].

To identify samples with reduced coverage, higher error rates, contamination, or other QC issues, numerous software packages have been developed operating on unaligned (ex. FASTQC) or aligned (ex. Picard toolkit, Qualimap2[159], HTQC[160]) sequencing data. Samples with coverage below a given cut-off, those with an extremely high background error rate, or extensive contamination are generally excluded from further downstream analysis.

Table 1-2. A brief set of QC issues encountered during Illumina library preparation and sequencing, and their downstream effects on the sequencing data

Sequencing Issue	Downstream effects
Limited input DNA	Low sequencing coverage High duplicate rate
Poor capture efficiency	Low coverage of on-target regions Increased number of off-target reads
Formalin-Fixed, Paraffin-embedded tissue source	Reduced sequencing coverage DNA damage artifacts
Sample contamination (ex. Bacterial contamination)	Large number of unmapped reads (to human genome) Reduced genome sequencing coverage
Over-fragmented library	Shorter read lengths Increased rate of read mismapping Reduced sequencing coverage Higher duplicate rate C/G->A/T DNA damage[158]
Excessive PCR cycles	Higher duplicate rate Increased GC coverage bias

1.5.3. Simple somatic mutation detection

Almost all Illumina sequencing projects aim to identify single nucleotide variants (SNVs) and small insertions and deletions (indels) (cumulatively simple somatic mutations, SSMs) within a sample of interest. At the basic level, a somatic variant caller will cycle through every position in the reference genome (or capture regions, if specified), and obtain all sequencing reads (and the corresponding base) overlapping that position. These bases are then compared to the reference base to find any support for a variant. If such support exists, a corresponding confidence score is assigned based upon the number of read supporting the alternate allele, base quality scores, mapping quality scores, and other features dependent on the variant caller in question. If a matched normal is also provided, support for a variant at the genomic locus in question is also evaluated in the normal, to distinguish germline variants (supported in the matched normal) from somatic variants (no support in the normal). Note that, while this process of distinguishing somatic from germline variants appears trivial, it is not unusual for constitutional samples (typically sourced from peripheral blood[161]) to contain some level of tumour cells[162,163].

A plethora of somatic variant callers have been developed[162,164–170], but benchmarking studies[171,172] show that Strelka2[165] and MuTect2[164] outperform other

variant callers for most use cases. However, it must be strongly emphasized that all somatic variant callers have their own strengths and weaknesses, and the end user must evaluate how these relate to the features of their dataset in question. In our experience, Strelka2 notably outperforms MuTect2 in tumour-normal pairs with high levels of tumour cell contamination in the normal, while Strelka2 calls excessive numbers of false positive variants in samples prepared from FFPE tissues. MuTect2 performs de-novo assembly of candidate insertions and deletions (indels) to accurate determine how many reads support such an event, while the Strelka2 pipeline allows exceptionally large indels (>80bp) to be called. In some use cases (for instance, ultra-deep targeted sequencing approaches[173]), neither MuTect2 nor Strelka2 may be suitable. Generating a consensus list of variants based on multiple variant callers does address the weakness of any one caller and is utilized in practice[40,174,175], but this may also sacrifice the strengths of any single tool.

1.5.4. Copy number variant detection

There are three primary methods of identifying copy number variants (CNVs) from Illumina sequencing data: read depth, B-allele frequency, and breakpoint detection. In a read depth approach, the genome is broken up into windows or bins (usually several hundred base pairs in length), and the number of reads in these bins are counted, adjusted for GC content, and compared to expected number of reads in a bin for a copy-neutral segment. If fewer reads are present than expected, this corresponds to a deletion, while additional reads correspond to a gain or high-level amplification of that region (Figure 1-5A). The expected number of reads can be calculated from a matched normal (if available), or a panel of unrelated normal samples, although the latter approach is generally less accurate especially for targeted sequencing approaches. In a B-allele frequency approach, heterozygous germline SNPs from a matched normal are identified, and the B-allele frequency (i.e. ratio of reads supporting the reference and alternate alleles) of this SNP is calculated in the tumour sample (Figure 1-5B). As a heterozygous SNP is expected to have a 50/50 ratio of reads supporting each allele, a deviation from this ratio indicates additional of fewer copies of a given allele are present. For instance, if 66% of reads support allele A of a SNP and 33% support allele B (3:2 ratio), this indicates that allele A has been duplicated, and there are 3 copies of the corresponding locus. In a contrasting example, if 100% of reads support allele A while

0% support allele B, this indicates there has been a loss-of-heterozygosity event at this locus. This indicates that all copies of this locus represent allele A, either because only a single copy of the locus remains (deletion, CN=1), or a reciprocal translocation resulted in the loss of allele B (copy number state >1). In breakpoint detection, read pairs overlapping genomic positions corresponding to the boundaries of a copy number event are identified (described in 1.5.5).

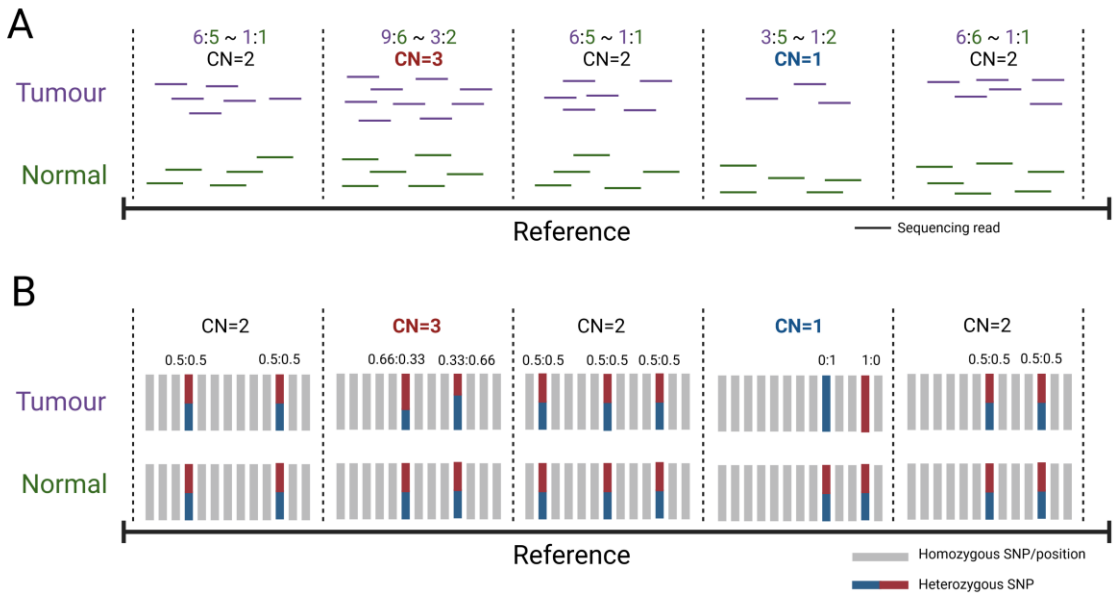


Figure 1-5. Example copy number variant detection using (A) read depth, and (B) B-allele frequency. Note that this example utilizes a matched normal

In contrast to the identification of SNVs, identifying CNVs is relatively difficult due to technical and sample-specific biases. For instance, GC bias results in lower coverage of regions with high AT or GC content[176,177], and thus the expected number of reads in a given bin will vary based on GC content. Furthermore, as most tumour samples are impure (i.e. contaminated by normal cells), the signal of somatic CNVs will be diluted by DNA from healthy diploid cells, and CNVs become increasingly difficult to detect in samples with low tumour content[178]. These features affect both WGS and WES data and are generally accounted for by most modern tools. The majority of tools also require a corresponding matched normal to be sequenced, both to assist in correcting these biases and to enable heterozygous SNPs to be called and leveraged to identify CNVs.

Detection of CNVs from capture-based sequencing is generally more difficult than for WGS data, both because the coverage of off-target regions is almost negligible,

and because coverage of on-target regions can vary significantly due to differences in the affinity of a given bait for its target DNA fragments. Almost all capture-compatible CNV callers require a matched normal[179,180], or at the bare minimum a set of unmatched normals[181] which have undergone an identical sequencing workflow to correct for such biases. CNVs within off-target regions are also extremely difficult to detect in capture-based sequencing data, although tools leveraging off-target reads show promise in addressing these limitations[180,181]. Samples which have undergone lpWGS further require custom software and extremely large bin sizes (on the order of 250kb-1mb) to account for the sparsity of the corresponding sequencing data[182,183]. Due to these complicating factors and the diversity of datasets, a wide plethora of CNV callers are utilized in practice. In general, Battenberg[184] is a solid choice for WGS data, Sequenza[179] can be used for WES data, cnvkit[181] generally performs well for custom-captures and samples without a matched normal, and ichorCNA[182] is a good choice for lpWGS data.

1.5.5. Structural variant detection

Structural variants are large-scale genomic alterations which re-organize portions of the genome. As these events generally occur in repetitive sequences[185] and given the short length sequencing reads, structural variants (SVs) are generally difficult to detect in Illumina sequencing data, and are almost impossible to comprehensively identify from WES. SV callers attempt to identify SVs using two types of reads: split reads and discordant read pairs (Figure 1-6)[186]. Sequencing reads overlapping a genomic breakpoint will comprise DNA from two distinct (and often distant) genomic loci. Thus, when these reads are mapped to the reference genome, a portion of the read will map to one portion of the genome while the other maps to the other end of the breakpoint. These are termed split reads due to the split alignment. In paired-end sequencing data, it is possible for one read to map to one end of the breakpoint completely (i.e. not split) while the other maps completely to the other end of the breakpoint (and a distant genomic locus), and are termed discordant read pairs. SV callers identify candidate breakpoints using these types of reads, then generate a directed acyclic graph of the breakpoint and supporting reads[187,188]. While this has been used to identify SVs in practice[161], in reality SV calling performance is highly variable, with recall ranging from 11[189]-70%[190] and relatively high false positive rates[191], especially in FFPE samples[192].

Thus, sequencing approaches with longer reads (for instance, Nanopore sequencing) are generally preferred for SV identification.

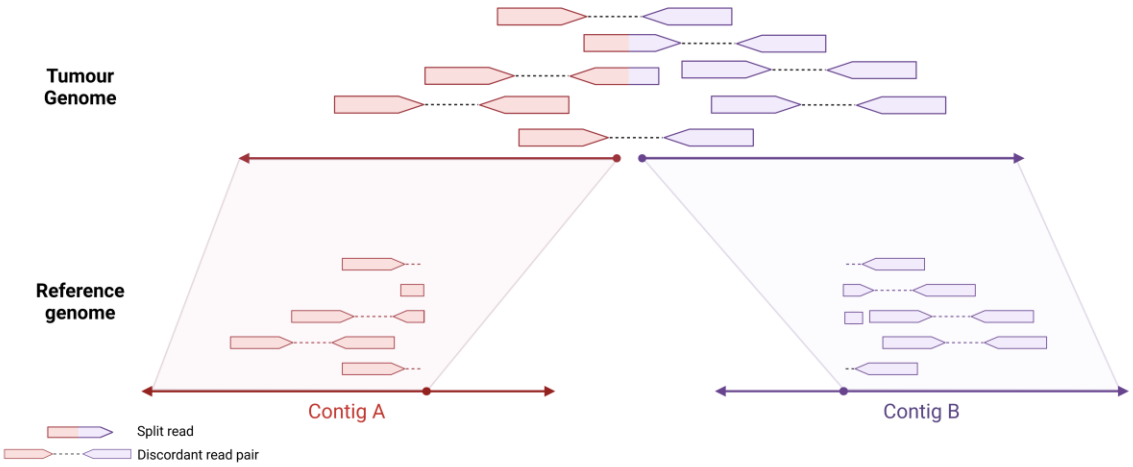


Figure 1-6. An example structural variant (translocation), and how Illumina sequencing reads will appear when mapped to the reference genome. Split reads and discordant read pairs (paired-end sequencing only) can be used to detect structural variants

1.6. Biopsies and cell-free DNA

In cancer genomic studies, a source of both tumour cells and normal cells are required to sequence both tumour DNA and the corresponding constitutional DNA. There are several candidate sources of both tumour and constitutional samples.

1.6.1. Tumour tissue biopsies

The traditional method of obtaining a tumour sample is to physically excise a portion of the tumour. These tissue biopsies are commonly used by pathologists to inspect the morphology of tumour cells, the tissue itself, as well as associated supporting cells, but can be used as a source of tumour material for genomics, transcriptomics, and proteomic analyses. Tissue biopsies are generally stored using two approaches; fresh-frozen, where the sample is snap frozen in liquid nitrogen and stored at -80^{o}C, or formalin fixation, paraffin embedded (FFPE), where a tissue sample is fixed in formalin and embedded in a paraffin block. While FFPE tissue biopsies can be stored at room temperature for extended periods of time without substantial degradation, formalin

fixation induces DNA and RNA damage which complicate downstream analysis[128,156], and thus fresh-frozen biopsies are strongly preferred for genomic studies.

While tissue biopsies are extensively collected and used in cancer genomics, they come with several limitations. First, collection of a physical biopsy may be impossible in some cancer types due to the physical location of the tumour, such as tumours within the central nervous system[193]. Second, collection of a tissue biopsy is a surgical procedure, and is both expensive and prone to complications[194] which can vary depending on the tumour type and location[195,196]. These costs and risks generally inhibit multiple biopsies from being collected from a single patient. Furthermore, the amount of material collected may limit downstream applications, and insufficient tumour material is collected in as many as 30% of tissue biopsies[197].

1.6.2. Whole blood and buffy coat

A blood sample can be fractionated into three major components via centrifugation. Blood plasma comprises the largest fraction by volume and contains lightweight components such as cell-free proteins, DNA, lipids, salts, and small macromolecules. The middle portion, called the buffy coat, is comprised of leukocytes and platelets, and represents ~1% of the blood fraction. The remaining fraction is comprised almost exclusively of red blood cells, which grant it its characteristic red colour. The buffy coat is particularly useful as it contains whole cells which can be used as a source of constitutional DNA[161]. However, this fraction can also contain malignant cells which have detached from the main tumour and are circulating in the bloodstream. These are termed circulating tumour cells (CTCs) and can contaminate the buffy coat with tumour DNA during bulk sequencing, contaminating the constitutional sample (tumour cell contamination)[162,163]. Although generally rare, CTCs can also be isolated and used as a source of tumour DNA and specialized downstream applications, such as single cell sequencing[198].

1.6.3. Blood plasma and cell-free DNA

When a cell undergoes apoptosis, the genome and organelles are disassembled and packed into a portion of the cellular membrane (membrane blebbing)[199,200]. The resulting apoptotic bodies are released into the bloodstream and circulate for a brief time

until they are cleared by macrophages[201]. While this extra-cellular DNA, termed cell-free DNA (cfDNA), has a relatively short half life of 1-2 hours[202], it has profound utility in cancer therapy. For instance, cfDNA levels tend to be elevated in patients with cancer, and is correlated with tumour size[203] and disease stage[204]. However, cfDNA levels vary significantly between individuals and are also elevated by other factors such as injury or trama[205], and thus caution must be applied when comparing cfDNA levels between individuals.

While the majority of cfDNA in cancer patients originates from healthy cells which have undergone apoptosis, a proportion of cfDNA originates from tumour cells and is termed circulating tumour DNA (ctDNA). Not only are ctDNA levels (the fraction of cfDNA which originates from tumour cells) correlated with tumour stage and burden[206] but represent a candidate source of tumour genetic material. Thus, a liquid biopsy (typically a blood sample, but other types of liquid biopsies exist[207]) can be collected from a patient, and the cfDNA extracted and used to detect somatic mutations.

1.6.4. Characteristics of cfDNA and ctDNA

The majority of cfDNA and ctDNA originate from cells which have undergone apoptosis, although necrosis and other mechanisms also contribute to cfDNA[208]. The human genome is compressed into chromosomes by chromatin subunits, comprised of an octamer of histones proteins and 146 bases of DNA wrapped around this histone complex[209,210]. Post-translational modification of individual histones regulate DNA accessibility (how “packed up” the DNA is) and is a central component of epigenetic regulation of gene expression[211]. These chromatin subunits are connected by linker DNA not bound to a nucleosome and significantly more accessible. Thus, when a cell undergoes apoptosis and produces DNA endonucleases[212], these endonucleases preferentially cleave the more accessible linker DNA and the DNA of actively transcribed genes (unpacked from the associated nucleosome, euchromatin). The resulting cfDNA is fragmented into small fragments with a mean size of 166bp[213], although supplementary peaks corresponding the DNA of multiple nucleosomes are also observed (2 nucleosomes = ~330bp, 3 nucleosomes = ~490bp), albeit at a reduced frequency. Longer DNA fragments (commonly referred to as high molecular weight DNA) are also observed (>1kb), usually originating from cells which have undergone necrosis. ctDNA fragments tend to be shorter than cfDNA, with a mean fragment size of 144bp[214],

although a significant number of ctDNA fragments shorter than 100bp are also observed. This short fragmentation of cfDNA impairs the utility of longer-read sequencing technologies such as nanopore sequencing. It should also be strongly emphasized that Illumina sequencing libraries prepared from cfDNA should not be fragmented prior to library construction, as additional fragmentation will substantially reduce the amount of usable DNA (DNA fragments that are too short will fail size selection) and will dilute the sample with high molecular weight DNA which tends to originate from healthy cells. The regular fragmentation pattern of cfDNA and ctDNA can also be used to determine the epigenetic state and expression of a gene, as unpacked euchromatin will be more accessible to DNA endonucleases and will show a random fragmentation pattern with lower coverage[215].

In a traditional tumour tissue biopsy, the majority of cells originate from the tumour, with minor representation of normal cells within the tumour microenvironment. Thus, tissue biopsies tend to have high tumour purity (Figure 1-7A). In contrast, the majority of cfDNA in liquid biopsies originates from healthy cells, even in highly advanced cancer cases, and thus a minority of DNA sequenced originates from malignant cells[216] (Figure 1-7B). While WES and WGS is possible for liquid biopsies with extremely high ctDNA levels (>30%), due to the limited sequencing depth of these approaches, specialized sequencing techniques are needed to detect somatic events in samples with lower ctDNA levels. One technique developed by Newman *et al*.[173] leverages a custom capture panel targeting regions recurrently mutated in the disease of interest. These regions are then sequenced extremely deeply (~ 10,000x coverage) to enable the detection of SSMs in samples with extremely low ctDNA levels (0.01%). While this approach (called CAPP-Seq) has been used extensively[217], sequencing to this depth has several limitations. First, the capture regions must be relatively small to limit the sequencing bandwidth (and thus cost) required. Second, some sample lack sufficient input DNA to reach the desired sequencing depth. If a liquid biopsy only contains DNA representative of 1,000 cells (termed genome equivalents), a given region of the genome can only be sequence to that depth. Third, Illumina sequencing has a background error rate of 0.1-0.5% (depending on the sequencer)[218], and thus it is difficult to distinguish real somatic variants from sequencing errors at variant allele frequencies (VAF) below 0.1% VAF.

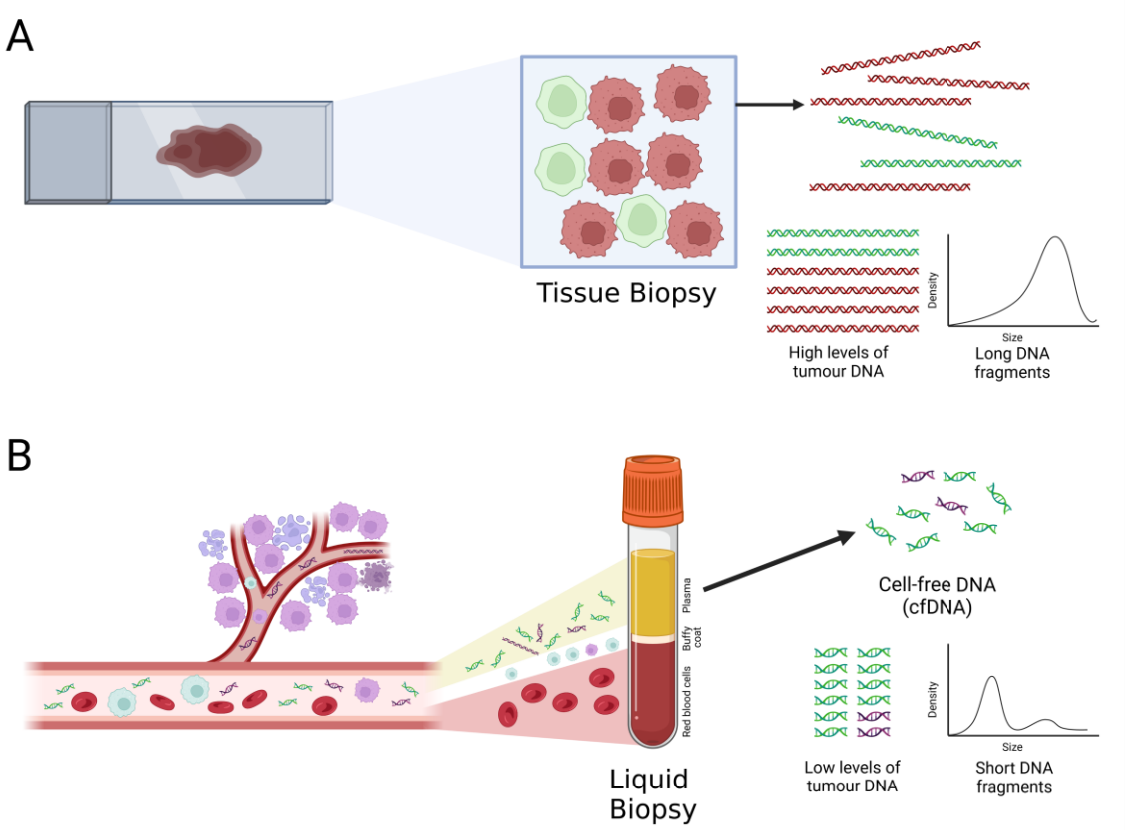


Figure 1-7. Overview and key differences between (A) tissue biopsies, and (B) liquid biopsies

One method of detecting SSMs below the error rate of Illumina sequencing machines is to leverage the PCR and optical duplicates generated during library preparation and sequencing to perform error correction of the resulting read pairs. As all duplicates of a DNA molecule should share an identical sequence, if differences exist, they must result from errors introduced during DNA replication or sequencing. Molecular barcodes are short semi-random DNA sequences ligated to the ends of each DNA fragment during library construction (also referred to as Unique Molecular Identifiers, UMIs) (Figure 1-8) [219]. After sequencing, one can identify all reads which originate from the same parental DNA molecule (termed family) as the family will all share the same UMI sequence (sequencing errors in the barcode notwithstanding) and map to the same place in the reference genome. The DNA sequence between family members can then be compared and collapsed to correct errors. While this approach requires a relatively high duplicate rate (generally undesirable in most sequencing applications as it reduces effective coverage and/or increases sequencing costs), it can reduce sequencing error

rate to 0.0001%[219], enabling confident detection of somatic mutations even in samples with extremely low ctDNA.

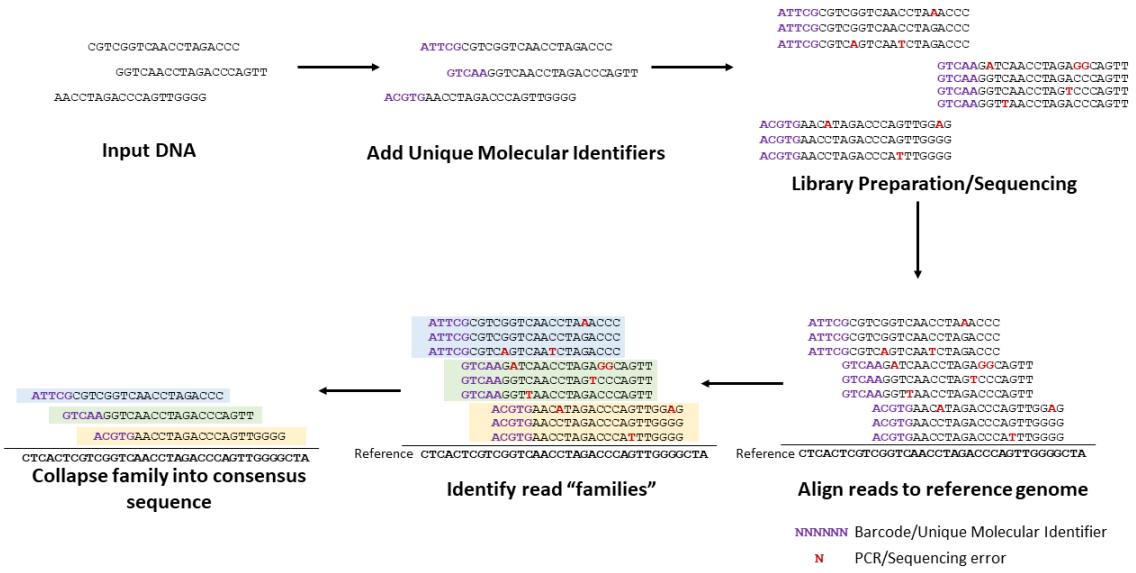


Figure 1-8. Leveraging Unique Molecular Identifiers to perform error correction following Illumina sequencing

1.6.5. Applications of liquid biopsies

Liquid biopsies have numerous advantages compared to traditional biopsies, notably their lower costs, minimal invasiveness, and increased accessibility compared to traditional tumour biopsies[208]. Not only can liquid biopsies be collected from cases where traditional tumour biopsies are unviable, but their superior accessibility and lower costs allow for serial liquid biopsies to be collected from the same patient. This can be applied following treatment to evaluate the tumour and evaluate treatment response[220,221]. Furthermore, if treatment is partially effective but the tumour persists (minimal residual disease, MRD), serial liquid biopsies can detect MRD with higher sensitivity than other approaches[204]. The relatively low cost and accessibility of liquid biopsies also enables a population to be screened for tumour-specific biomarkers and mutations[222]. However, liquid biopsies come with several limitations. For instance, not all cancer patients have detectable ctDNA, even following ultra-deep sequencing in highly advanced cases, and thus the utility of liquid biopsies in these cases is extremely limited. As RNA is relatively unstable compared to DNA[223], cell-free RNA has an even shorter half-life in the

bloodstream than cfDNA, and samples must be rapidly stored to minimize RNA degradation prior to RNA sequencing[224]. Several studies have shown that miRNAs and circular RNAs (which are more stable than mRNA) can act as tumour biomarkers in liquid biopsies[225]. Due to the lack of intact cells, single-cell sequencing is largely unviable from liquid biopsies, although sequencing of circulating tumour cells (CTCs) extracted from the buffy coat of whole blood can be used. Finally, the fragmented nature of cfDNA, while it has its own utility[215], largely impairs the use of long-read sequencing approaches, and thus detection of structural variants is extremely difficult.

1.7. Genetics of DLBCL

One of the overarching goals of cancer genomics is to comprehensively identify genes recurrently mutated in a given type of cancer, and the resulting effects of those mutations on the cell and tumour. These driver genes can then be explored as candidate therapeutic targets. Numerous large-scale sequencing studies and a plethora of smaller studies have attempted to comprehensively identify driver genes and mutations in DLBCL.

1.7.1. Genetics of diagnostic DLBCL

Several large scale genomic studies, utilizing WES[226–228] and WGS[15] of DLBCL FF and FFPE biopsies, have explored the landscape of somatic mutations in diagnostic DLBCL. The most commonly mutated gene in diagnostic DLBCL is the lysine methyltransferase *KMT2D*[229] and the master tumour suppressor gene *TP53*, which both act as tumour suppressors and acquire loss-of-function mutations. The NF-κB signaling component *MYD88* is also recurrently mutated, and acquires an activating hotspot mutation (Leu265Pro) leading to constitutively active NF-κB signaling[20,230] and promoting B-cell survival. Mutations in the Histone H1 (linker histone) gene *HIST1H1E* leads to broad epigenetic dysregulation and results in increased expression of normally repressed genes via the formation of euchromatin[231]. *CD79B* encodes Igβ, a component of the B-cell co-receptor which mediates B-cell receptor signaling. Mutations within *CD79B* tend to occur in the ITAM domain, specifically Tyr196, and prevent active *CD79B* from responding to inhibitory signaling circuitry[232]. Further recurrently mutated genes include the histone acetyltransferase *CREBBP*, acquiring loss-of-function mutations

(specifically within the acetyltransferase domain)[233] which impair histone H3 acetylation and subsequent transcriptional activation of target genes, and *CARD11*, which accumulates mutations in a coil-coil domain which further enhance NF-κB signaling[234].

Copy number alterations also contribute significantly to the genetic landscape of DLBCL[235,236]. Arm-level or whole-chromosome gains of chromosome 7 are relatively common, along with recurrent deletions of the q arm of chromosome 6 and the p arm of chromosome 17. More specifically, recurrent deletions are observed perturbing the master tumour suppressor gene *TP53* (cytoband 17p13.1), the MHC Class 1 component *B2M* (15q21.1), the cell cycle regulators *CDKN2A*/*CDKN2B* (9p21.3), *PTEN* (10q23.31) and *RB1* (13q14.2), and *TNRFSF14* (1p36.32), whose deletion leads to recruitment of T-cells and subsequent production of pro-inflammatory cytokines[237]. Recurrent copy number gains are observed affecting *REL* (2p16.1, generally extremely focal and high-level), encoding c-REL which is a component of the NF-κB complex and required for canonical NF-κB signaling[238], *BCL6* (3q27.3), which prevents B-cell differentiation into plasma cells[239] and impairs DNA repair and apoptotic pathways[240,241], *MYC* (8q24), whose amplification promotes B-cell proliferation, survival, invasion, and deregulates cellular energetics[242], and the master anti-apoptotic factor *BCL2* (18q21.33). Further recurrent gains are observed affecting the micro-RNA cluster *MIR17HG* (13q31.3), which promotes cell cycle progression and proliferation[243], and *MDM2* (12q15), which binds and inactivates *TP53*[244]. It should also be noted that *BCL2*, *BCL6*, and *MYC* commonly experience chromosomal translocations which place these genes adjacent to and under control of a constitutively activated enhancer leading to significantly expression[245].

1.7.2. Genetics of molecular subgroups

As DLBCL molecular subgroups are associated with distinct morphological stages of B-cell development, with corresponding unique transcriptional profiles[78], each subgroup acquires a unique pattern of driver mutations. ABC-DLBCL is characterized by constitutively active NF-κB signaling, and as such ABC-DLBCL tend to acquire driver mutations which enable and further enhance NF-κB signaling. These include activating mutations of *CD79B, MYD88*, and the NF-κB transcription factor complex regulator *NFKBIZ*[15]. In contrast, GCB-DLBCL commonly acquires mutations perturbing epigenetic regulators, with mutations in *CREBBP* and *HIST1H1E*, activating mutations in the

histone methyltransferase *EZH2*[246] and transcription factor *MEF2B* (leading to increased expression of *BCL6*)[247] enriched in GCB-DLBCL. This distinct pattern of genomic alterations also extends to copy number events, with gains of *BCL2*, *BCL6*, and *MIR17HG* significantly enriched in ABC-DLBCL, while gains of *REL* and deletions of *PTEN, FAS*, *B2M*, and *TNRFSF14*, as well as rearrangements involving *BCL2*, are enriched in GCB-DLBCL.

1.7.3. Genetic subgroups

Several groups have recently uncovered additional DLBCL subgroups harbouring shared genetic features, and have attempted to classify DLBCL cases into genetic subgroups with prognostic and therapeutic implications[227,228,248,249]. Chief among these is LymphGen[249], which classifies DLBCL tumours into six genetic subgroups using SSMs (mainly coding mutations, but including some non-coding regions), CNVs, and SVs. These genetic subgroups overlap existing molecular subgroups to some extent, with GCB samples generally classified into either EZB or ST2 subgroups. EZB is named for a high frequency of mutations in *EZH2* and translocations of *BCL2* and characterized by mutations in numerous epigenetic modifiers (*CREBBP, KMT2D, EP300, EZH2*). ST2 is dominated by mutations in *SGK1* and *TET2* and associated with constitutive PI3K and JAK/STAT signaling. DLBCL cases transformed from follicular lymphoma tend to be classified within the EZB subgroup. ABC-DLBCL cases tend to be classified as either MCD, or BN2. The MCD subgroup dominated by hotspot mutations in *MYD88* and *CD79B* and is characterized by constitutively active NF-κB signaling, while BN2 is named for high frequency of mutations in *NOTCH2* and translocations involving *BCL6*, and harbouring mutations in BCR-dependent NF-κB signaling pathways. Two additional genetic subgroups, A53 and N1, are not strongly associated with either molecular subgroup. A53 cases are characterized by loss of *TP53* and a high burden of copy number alterations, while the N1 subgroup is defined by gain-of-function mutations perturbing *NOTCH1.* These genetic subgroups have prognostic significance, with MCD and N1 cases associated with inferior outcomes across DLBCL tumours, while on a molecular subgroup basis, MCD and A53 represent cases with inferior prognosis within ABC-DLBCL, and EZB and A53 represent a subset of GCB-DLBCL with inferior outcomes.

1.7.4. Genetics of rrDLBCL

As treatment exerts a strong selective pressure on the tumour and individual cells, one would expect that rrDLBCL tumours would be enriched for mutations which contribute to treatment failure. To this end, several studies have attempted to explore the landscape of rrDLBCL to identify genetic features underpinning treatment resistance. Unfortunately, many of these studies have been limited by small sample sizes, as tissue biopsies are generally not collected upon relapse. Previously, the largest genomic study of rrDLBCL consisted of WES on 47 rrDLBCL samples[250], and observed a high frequency of mutations perturbing *EZH2*, *CREBBP*, and *MYD88*. An additional study performing exome sequencing on 38 rrDLBCL cases reported an enrichment of mutations affecting *TP53*, *KMT2C*, *FOXO1*, *STAT6*, *MYC*, and *CCND3*[251]. Several studies have identified recurrent CNVs which may contribute to immune evasion, through recurrent deletions and mutations of genes encoding MHC-Class 1 components, and *B2M*[252–254], with functional loss of MHC Class 1 on tumour cells. However, mutations directly implicated with treatment resistance have currently not been uncovered, with most studies to date limited to a dozen samples.

1.8. Research Aims and Outline

During this project, we aimed to explore and characterize the genomic landscape of relapsed-refractory DLBCL. If mutations perturbing a gene contribute to treatment resistance (and thus provide a selective advantage) one would expect these mutations to be enriched and prevalent at relapse. Thus, by characterizing the landscape and repertoire of mutations in rrDLBCL, we hope to identify events prevalent at diagnosis but further enriched at relapse, representing candidate biomarkers of treatment failure, as well as initially rare events which are selected following treatment, representing mutations acquired following the selective pressure of therapy. We also aim to identify recurrent events in rrDLBCL which could act as therapeutic targets.

This thesis is comprised of an introductory chapter, two data chapters outlining research to this end, and a final overview discussion chapter.

Chapter 2. Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma

This chapter has previously been published as a research paper[255]. Christopher K. Rushton*, Sarah E. Arthur*, Miguel Alcaide, Matthew Cheung, Aixiang Jiang, Krysta M. Coyle, Kirstie L. S. Cleary, Nicole Thomas, Laura K. Hilton, Neil Michaud, Scott Daigle, Jordan Davidson, Kevin Bushell, Stephen Yu, Ryan N. Rys, Michael Jain, Lois Shepherd, Marco A. Marra, John Kuruvilla, Michael Crump, Koren Mann, Sarit Assouline, Joseph M. Connors, Christian Steidl, Mark S. Cragg, David W. Scott, Nathalie A. Johnson†, and Ryan D. Morin†. 2020. "Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma." *Blood Advances* 4(13): 2886-2898. *Contributed Equally. †Contributed Equally.

Contribution: R.D.M., D.W.S., and N.A.J. conceptualized the study; R.D.M. provided the methodology; C.K.R. provided the software; C.K.R., S.E.A., M.A., A.J., K.M.C., N.T., L.K.H., K.L.S.C., and M.S.C. provided the formal analysis; C.K.R., S.E.A., M.A., R.N.R., K.M.C., and A.J. led the investigation; S.E.A., M.A., M. Cheung, N.M., S.D., J.D., R.N.R., K.B., S.Y., M.J., L.S., J.K., M. Crump, M.S.C., K.M., S.A., and N.A.J. provided the resources; C.K.R., M.A., and N.A.J. curated the data; C.K.R., S.E.A., M.A., R.D.M., D.W.S., and N.A.J. contributed to writing of the original draft:; C.K.R., S.E.A., and R.D.M. provided the visualization; S.E.A. and C.K.R. contributed to the writing, review, and editing; R.D.M., N.A.J., and D.W.S. provided supervision; and R.D.M., D.W.S., N.A.J., J.M.C., M.A.M., and C.S. contributed to acquiring funding.

The authors acknowledge the support of the patients and their families, who have consented to participate in lymphoma tissue banking within the Banque de Cellules Leucémiques du Québec at the Jewish General Hospital (JGH). The authors thank the Jewish General Hospital Foundation and the Cole Foundation for supporting the Lymphoma Cell Bank at the JGH.

This work was supported by Terry Fox New Investigator Award 1021 and Program Project grants 1043 and 1061. R.D.M. holds a Canadian Institutes of Health Research (CIHR) New Investigator Award and a Scholar of Michael Smith Foundation for Health Research. The authors gratefully acknowledge funding support from the Terry Fox Research Institute (J.M.C.), Genome Canada (J.M.C. and C.S.), Genome British

Columbia (J.M.C.), CIHR (J.M.C., R.D.M., and N.A.J.), and the British Columbia Cancer Foundation (J.M.C. and R.D.M.). N.A.J. is also supported by funds from Canadian Cancer Society Research Institute (705478).

2.1. Abstract

DLBCL patients are typically treated with immunochemotherapy containing rituximab (rituximab, cyclophosphamide, oncovin, and prednisone [R-CHOP]); however, prognosis is extremely poor if R-CHOP fails. To identify genetic mechanisms contributing to primary or acquired R-CHOP resistance, we performed target-panel sequencing of 135 relapsed/refractory DLBCLs (rrDLBCLs), primarily comprising circulating tumor DNA from patients on clinical trials. Comparison with a metacohort of 1670 diagnostic DLBCLs identified 6 genes significantly enriched for mutations upon relapse. *TP53* and *KMT2D* were mutated in the majority of rrDLBCLs, and these mutations remained clonally persistent throughout treatment in paired diagnostic-relapse samples, suggesting a role in primary treatment resistance. Nonsense and missense mutations affecting *MS4A1*, which encodes CD20, are exceedingly rare in diagnostic samples but show recurrent patterns of clonal expansion following rituximab-based therapy. *MS4A1* missense mutations within the transmembrane domains lead to loss of CD20 in vitro, and patient tumors harbouring these mutations lacked CD20 protein expression. In a time-series from a patient treated with multiple rounds of therapy, tumor heterogeneity and minor *MS4A1*-harbouring subclones contributed to rapid disease recurrence, with *MS4A1* mutations as founding events for these subclones. *TP53* and *KMT2D* mutation status, in combination with other prognostic factors, may be used to identify high-risk patients prior to R-CHOP for posttreatment monitoring. Using liquid biopsies, we show the potential to identify tumors with loss of CD20 surface expression stemming from *MS4A1* mutations. Implementation of non-invasive assays to detect such features of acquired treatment resistance may allow timely transition to more effective treatment regimens

2.2. Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin Lymphoma (NHL), representing 30-40% of cases diagnosed in North America. DLBCL

can arise de-novo or through histologic transformation from indolent lymphoid malignancies, most commonly transformed follicular lymphoma ("tFL"). Patients diagnosed with DLBCL are generally treated with a standard immunochemotherapy regimen comprising four chemotherapeutic agents and the anti-CD20 monoclonal antibody rituximab (R-CHOP), which is curative for 60-70% of DLBCL cases[103,256]. However, for patients with DLBCL that which is refractory to frontline treatment and those who experience subsequent relapse (relapsed/refractory DLBCL, "rrDLBCL"), outcomes are extremely poor, with a 2-year overall survival of 20-40%[108,257]. While numerous treatments are under investigation to improve both frontline and salvage therapy, the success of these new therapies has been limited. The advancement of therapeutics in the relapse setting has likely been encumbered by our limited understanding of the genetic and molecular features that underlie innate and acquired resistance to R-CHOP. Identifying such mechanisms may reveal additional treatment options and lead to biomarkers allowing patients to be paired with appropriate treatments.

Whereas the genomic landscape of diagnostic DLBCL is well understood, the genomic and molecular features of both rrDLBCL and DLBCLs that arise through histologic transformation remains elusive due to the difficulties in obtaining tumor tissue from relapsed patients. Early studies exploring the genomic landscape of rrDLBCL identified several candidate genes enriched for mutations among rrDLBCL cases, including *TP53, STAT6, FOXO1, SOCS1, and PIM1*[251,258,259]. Indeed, mutations in some of these may be prognostic at diagnosis (e.g. *FOXO1*[260] and *TP53*[261,262]), whereas others may reflect a more diverse representation of DLBCLs beyond those arising de novo including tFLs. However, previous studies of rrDLBCL have been limited by small sample sizes, with the largest single cohort comprising 47 cases[250]. In addition to comparing mutation prevalence between untreated DLBCL and rrDLBCL, some studies compared the clonal population structure and mutation burden between paired diagnostic and relapse samples[250,263]. Such analyses nominated additional candidate genes whose mutation could contribute to treatment resistance, such as BCL2 and CREBBP[250], but these results have not been independently confirmed. The genetic heterogeneity of DLBCL warrants a more comprehensive study of rrDLBCL to definitively identify genes associated with relapse and to characterize the role of these mutations for resistance to components of R-CHOP.

Although cell-free DNA is commonly used for non-invasive quantitative monitoring of disease burden[264–266], with sufficient levels of circulating tumor DNA (ctDNA), liquid biopsies can also provide a source of tumor genetic material allowing broad genetic characterization of tumors[208,267,268]. In DLBCL, mutations found within ctDNA reflect somatic mutations irrespective of anatomical biases, providing opportunity for comprehensive exploration of tumor genetics and heterogeneity[269]. This can be accomplished using a single time point[270] but is more powerful when applied to serial samples as the variant allele frequency can reveal clonal dynamics and thus putative resistance mechanisms[271].

To more thoroughly survey the genetic mechanisms of R-CHOP resistance in DLBCL, we explored the genetics of rrDLBCL in 135 cases relying on a combination of tumor tissue and plasma-derived ctDNA collected after relapse. By comparing the mutational profiles of these cases to a large cohort of untreated DLBCLs, we identified 6 genes significantly enriched for mutations. Many of these genes are commonly mutated in untreated DLBCL, notably *KMT2D* and *TP53,* and remain clonally stable over the course of therapy. Another of these genes, *MS4A1*, encodes the B-cell surface marker CD20 and is the target of rituximab. *MS4A1* missense mutations are restricted to transmembrane domains and inhibit binding of both rituximab and other anti-CD20 antibodies. These finding have the potential to identify patients at a high risk of R-CHOP failure prior to frontline treatment and those with tumors likely to be resistant to rituximab-based secondary therapies and other CD20-targeted immunotherapies.

2.3. Methods

2.3.1. Targeted sequencing and mutational analysis of rrDLBCLs

This study included samples from 135 patients with rrDLBCL with 117 of these comprising plasma collected within 3 clinical trials or the general patient population treated in Quebec (Supplemental Table S1, S2, and S3, Appendix A). This study was reviewed and approved by the Research Ethics Boards of the University of British Columbia-BC Cancer and the Jewish General Hospital (18-030), in accordance with the Declaration of Helsinki. Plasmas were collected and processed as previously described[268,272] and detailed in Section 2.4.2. The remaining 18 cases represent tissue biopsies previously described by our group[251]. With the exception of these 18 cases with

existing exome data, all samples were subjected to library construction using custom adaptors with unique molecule identifiers. Libraries were enriched by hybridization-capture using a custom set of LockDown oligonucleotides targeting the exons of 63 genes (Supplemental Table S4, Appendix A). The genes on this panel represent well-established DLBCL genes from previous publications and included *MS4A1* based on preliminary exome and genome data from PT255 and the 18 rrDLBCL exomes. Following enrichment, all libraries were multiplexed and sequenced using Illumina chemistry using 125- or 150-bp paired reads on either MiSeq or HiSeq2500 instruments. After alignment, reads were collapsed into consensus sequences using in-house pipeline that leverages unique molecule identifier information. Single nucleotide variants and small insertions and deletions (henceforth simple somatic mutations) were identified with Strelka2[165] with custom post-filtration steps to remove artifacts (Section 2.4.4; Supplemental Table S5, Appendix A).

2.3.2. Meta-analysis of untreated DLBCLs

To obtain a cohort representative of diagnostic DLBCLs, we compiled exome data from three previously published cohorts[226–228] and a cohort of paired tumor/normal genomes[15], amounting to 1670 cases termed the “untreated” cohort, because all biopsies were obtained prior to treatment. As matched constitutional samples were not available for the majority of these exome cases, and because the supplied variant calls were generated using diverse pipelines, we reprocessed all exomes through a standardized variant calling workflow for unpaired tumor samples, including filtering of common and rare germline variants (see section 2.4.5, Figure 2-1).

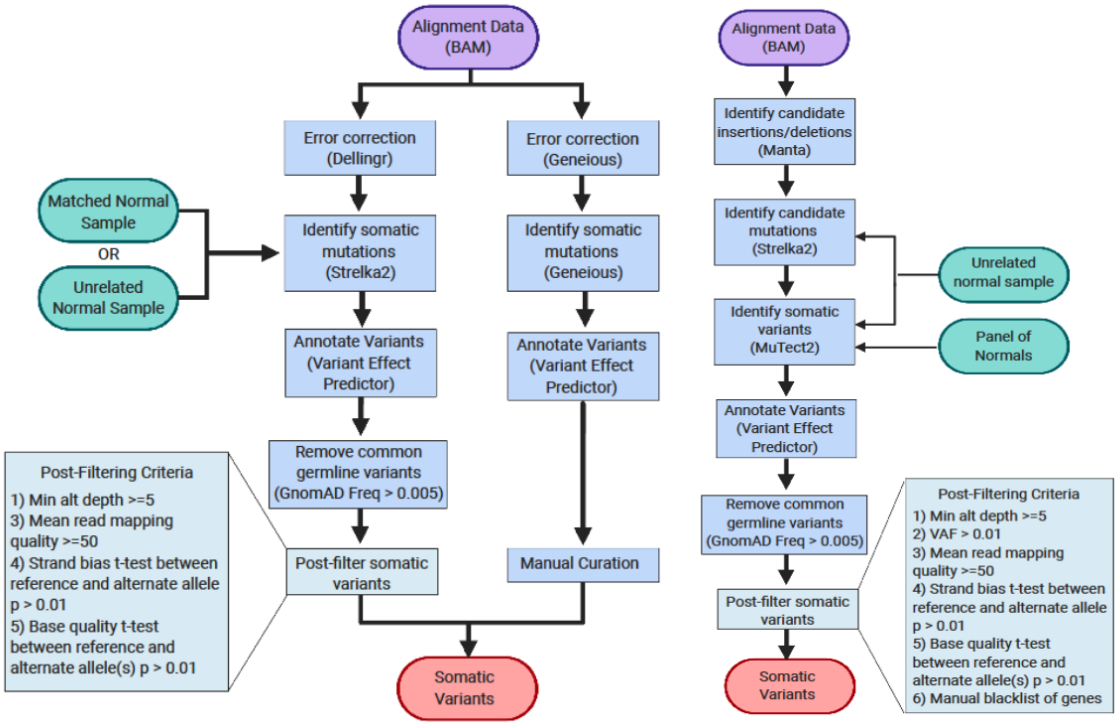


Figure 2-1. Variant calling workflows for rrDLBCL cases (left), and Untreated DLBCL exome cases (right).

2.3.3. Identifying genes associated with rrDLBCL

. We identified mutations and hotspots associated with rrDLBCL using two complementary approaches. First, we compared the gene and hotspot mutation frequency between rrDLBCL and untreated DLBCL to identify genes enriched for mutations in rrDLBCL. Mutation hotspots considered here are listed in Supplemental Table S6, Appendix A. The mutation frequency of all genes in our panel was compared between the rrDLBCL cohort and the untreated DLBCL cohort, as well as an additional diagnostic cohort[248], using Fisher's exact test and Benjamini/Hochberg false discovery rate threshold of 0.1 (Supplemental Table S7 and S8, Appendix A). Second, leveraging the paired samples representing time points prior to and following treatment (Table 2-1), we compared the tumor genomic landscape between time points to identify genes that recurrently showed evidence of clonal selection. Mutations were classified based on the log ratio of the cancer cell fraction (CCF) between the two time points, where log-fold change (T2 CCF)/(T1 CCF) > 1.0 indicates a mutation underwent clonal expansion following treatment, log-fold change < 1.0 indicated the mutation was depleted following treatment, with all other values considered stable. We also explored the prevalence of

different genetic subgroups using the LymphGen classifier (See 2.4.7; Supplemental Table S2 and S9, Appendix A)[249], and the prognostic implications of clonal and highly recurrent events in rrDLBCL (See 2.4.8 and Supplemental Table S10, Appendix A)

Table 2-1. Patients and samples used in clonal evolution analysis

Cohort	Clinical Trial Identifier	Cell-of-origin and DHITSig assigned using	Diagnostic Tumour Source	Relapse Tumour Source	Diagnostic tumour biopsy + relapse plasma pairs
LY.17	NCT02436707	DLBCL90	Tumour Biopsy	cfDNA	10
Epizyme	NCT01897571	DLBCL90	Tumour Biopsy	cfDNA	47
Total	-				57

2.3.4. Evaluation of MS4A1 protein expression and antibody reactivity

Suspension-adapted Chinese hamster ovary (CHO-S) cells (Life Technologies) were cultured in FreestyleCHO media supplemented with 8 mM glutamine (Gibco). Cells were maintained between 0.3 and 1.5 x 10^6 cells/mL in a humidified shaking incubator at 37°C in 8% CO2. CHO-S cells (107) were transfected with 10 mg of plasmid DNA containing MS4A1 wild-type (WT) or mutant constructs. Details of mutagenesis are included in 2.4.9 and Supplemental Table S11, Appendix A. For each transfection, efficiency was determined using a positive control (green fluorescent protein [GFP]) to demonstrate that cells were permissive for transfections. For all experiments, WT CD20 transfections were performed in parallel with mutants. Transfected cells (1.5 x 10^5) were opsonized with 1.5 mg of unlabeled anti-CD20 antibody for 30 minutes at 4°C. Unbound antibody was washed twice in 2 mL wash buffer (phosphate-buffered saline containing 1% bovine serum albumin and 10 mM sodium azide) and centrifuged at 400g for 5 minutes and resuspended in ;150 mL of wash buffer. Primary antibody was detected with 0.2 mg/mL of anti-human immunoglobulin G (IgG)–phycoerythrin or anti-mouse IgG-phycoerythrin polyclonal antibodies (Stratech) and stained for 30 minutes at 4°C. Cells were washed in 2 mL of wash buffer before acquiring on a FACSCalibur fluorescence-activated cell sorter. Flow cytometry data were analyzed in FCSExpress v.3 (De Novo software, Pasadena, CA). rituximab (human [h]IgG1), ofatumumab (hIgG1), tositumomab (B1, murine [m]IgG1), and obinutuzumab (non-glycomodified hIgG1 type II relative of obinutuzumab) or an isotype control (mIgG1 or hIgG1) were used to stain the cells. Immunoblotting was performed largely as reported previously[273]. In brief, 5 x 10^6

cells were lysed in radioimmunoprecipitation assay buffer with 20 mg separated on a 10% Bis-Tris gel. CD20 expression was assessed using rabbit anti-CD20 clone EP459Y (Abcam) alongside an HRP-conjugated anti-rabbit secondary antibody (NA9340, Sigma) detected using a ChemiDoc-It Imaging System. Full details of immunoblotting are included in 2.4.11.

2.3.5. PT255 exome sequencing and single-cell analysis

We performed exome sequencing on a single relapsed case (PT255) representing 3 time points: (1) the diagnostic biopsy (diagnosis, D); (2) cell-free DNA (cfDNA) collected following second relapse (relapse 2, R2/P1); and (3) cfDNA collected following third relapse (relapse 3, R3/P5). Somatic variants, copy number alterations, and clonal population structure were analyzed as described above. Somatic coding variants were chosen from this bulk tumor and plasma exome sequencing to represent different clones at varying time points, with the Fluidigm Access Array used for multiplexing amplicon sequencing of selected variants in PT255 plasma samples and circulating tumor single cells from selected time points following relapse (Supplemental Table S12 and S13, Appendix A).

2.4. Supplemental Methods

2.4.1. rrDLBCL sample collection

251 patient samples were collected from three clinical trials exploring candidate treatment options for patients with relapsed-refractory DLBCL: LY.17 (NCT02436707), Obinituzumab-GDP [OZM073] (NCT02750670), Epizyme (NCT01897571), a retrospective cohort of rrDLBCL patients (Montreal), and a previously published rrDLBCL cohort (QCROC-2) (Supplementary Table S1 and S2, Appendix A). The LY.17, OZM073, Epizyme, and the Montreal cohorts consisted exclusively of blood samples, while the QCROC2 cohort consisted of tumor tissue biopsies. All patients were previously treated with R-CHOP, and in many cases were treated with several additional salvage therapies. COO and DHITsig were assigned for each sample using the DLBCL90 Nanostring assay[91] except for the Montreal cohort, which was evaluated by Immunohistochemistry[83] and FISH. Tumors identified as transformed from other lymphoid malignancies or clinically identified as PMBCL were analyzed separately. This

study was reviewed and approved by the Research Ethics Boards of the University of British Columbia-BC Cancer and the Jewish General Hospital (18-030), in accordance with the Declaration of Helsinki. For two of the clinical trials (LY.17 and Epizyme), diagnostic tumor tissue biopsies were provided prior to therapy for 57 patients (Table 2-1 and Supplemental Table S3, Appendix A). Both *de novo* DLBCL cases and cases transformed from other lymphoid malignancies were included in this cohort.

2.4.2. Blood processing and DNA extraction

Blood samples from DLBCL patients were either immediately centrifuged following collection or preserved in Streck Cell-free DNA BCT® blood collection tubes (Streck, La Vista, NE, USA) and processed within 1-2 weeks to separate plasma from cells. Plasma aliquots were kept at –80 °C for extraction at a later date. We used the MagMAX Cell-free DNA isolation kit (ThermoFisher Scientific, Waltham MA, USA) or the QIAamp® circulating nucleic acid kit (Qiagen, Hilden, Germany) to isolate cell-free DNA from 0.5-4 mL of plasma. Total DNA yields were estimated using a Qubit fluorometer (ThermoFisher Scientific).

Formalin-Fixed Paraffin-Embedded (FFPE) tissue slides and blocks corresponding to diagnostic biopsies were manually processed and DNA was extracted using the MagMAX™ FFPE DNA/RNA Ultra Kit (ThermoFisher Scientific). Constitutional DNA samples for each patient were extracted from buffy coats using the FFPE AllPREP kit (Qiagen) (LY.17. tumor biopsies and constitutional DNA samples) or DNeasy Blood & Tissue kit (Qiagen) (Epizyme tumor biopsies and constitutional DNA samples).

2.4.3. Library construction and targeted enrichment

ctDNA libraries were constructed using Illumina-compatible adapters carrying unique molecule identifiers (UMI) as described previously[219,274]. Equimolar amounts of libraries were pooled and enriched using a custom panel of 63 lymphoma-related genes (Supplementary Table S4, Appendix B) comprised of xGen® lockdown probes and pre-designed gene capture pools (Integrated DNA Technologies, Coralville, IA, USA). Enriched libraries were analysed on Illumina instruments (MiSeq or NextSeq) using 150 bp paired-end sequencing chemistry.

Tumor and constitutional DNA samples were incorporated into genomic DNA libraries by using the QIAseq FX DNA Library Kit (Qiagen). Similarly, we generated equimolar pools of tumor and normal DNA libraries and enriched them on coding regions using the xGen® Exome Research Panel (Integrated DNA Technologies, Coralville, IA, USA).

2.4.4. Sequence alignment and somatic variant calling

Raw sequencing reads were mapped to the human reference genome GRCh38 using BWA mem[152]. For libraries constructed with custom adapters containing UMIs[219], aligned reads were processed using Dellingr[274] (https://github.com/morinlab/Dellingr), which trims the UMIs and leverages them to identify duplicate reads following alignment. In this process duplicate reads are combined into a consensus sequence, thereby performing error-correction and removal of redundant bases from overlapping read pairs. For samples sequenced without UMIs (tumor biopsies), duplicate reads were flagged using Picard (http://broadinstitute.github.io/picard/), and soft-clipping of redundant bases from overlapping read pairs was performed using bamtools (https://github.com/pezmaster31/bamtools). Quality control was performed using Qualimap2[159] , and samples with insufficient coverage of the capture space were excluded from downstream analyses.

Simple somatic mutations (SSMs) were identified using Strelka2[165], providing the candidate small insertions and deletions predicted by Manta[188] (Figure 2-1). In cases where no constitutional DNA from the individual was available, we used sequence data from constitutional DNA representing a single (unrelated) patient in place of a matched normal and subsequently removed variants with a minor allele frequency above 0.005 in any population, as specified in gnomAD[275]. As Strelka2 was noted to systematically under-call variants strongly supported by our molecular barcodes due to low variant allele fraction, we supplemented Strelka2 outputs using a previously described Geneious workflow which directly leverages the molecular barcodes[219]. Variant annotation was performed using Variant Effect Predictor[276], specifically vcf2maf (https://github.com/mskcc/vcf2maf). Variant calls were post-filtered to remove those with: 1) Less than 5 reads supporting the alternate allele, 2) A mean read mapping quality of less than 50, 3) Read mapping strand bias $p<0.01$, as determined using Fisher's exact test, and 4) Base quality bias $p<0.01$, as determined using Student's t-test on all reads

aligned at this position. Supplemental Table S4, Appendix A contains somatic variant calls retained for all downstream analyses. For cases with multiple time points, exome sequencing was performed on the diagnostic tumor biopsy and constitutional DNA, which also allowed the detection of copy number alterations using Sequenza[179]. To minimize noise, the exome data was pre-filtered to remove any variants not observed in the ExAC database for the purposes of B-allele frequency calculation[275]. Clonal population structure was derived using PyClone[277] using copy number information from the matched tumor, where available. As we were unable to obtain copy number alterations from our plasma samples due to low ctDNA levels and the small capture space, we used the same copy number information for both the plasma and tissue biopsy. In situations where no copy number profile was available for a given patient, we specified a default diploid profile for both time points.

2.4.5. Analysis of untreated DLBCL cohort

Using the original sequence alignments obtained (Schmitz: NCI NCICCR-DLBCL, Reddy: EGA EGAS00001002606, Chapuy: dbGAP phs000450.v3.p1), candidate somatic variants were first identified using the unfiltered variant calls generated from Strelka2[165], using small insertion and deletions identified from Manta[188]. Candidate unfiltered variant positions were then converted into BED format and provided to MuTect2[148], which was run in unpaired mode using a panel of normals generated from 58 unrelated normal WGS samples. Further candidate germline variants were removed by filtering out any variant with a population allele frequency of >0.005 in gnomAD[275], and using the same post-filtering criteria as the rrDLBCL cohort (described in 2.4.4), along with a variant allele frequency filter of 1%. As matched normal genomes were available for the Arthur genome cohort, we elected to use the original variant calls provided. Variant calls generated from Reddy, Chapuy, and Arthur cohorts were lifted over to GRCh38 using Crossmap[278].

2.4.6. Quality control and validation of untreated cohorts and mutation frequency

Due to variable coverage and analytical approaches used for the untreated DLBCL exome data sets, all exomes were assessed for sufficient coverage at each locus in our panel. This was accomplished using the GATK CollectCallableLoci tool[148],

using the following criteria to consider a position “callable”: minimum depth of 30, maximum fraction of reads with a low mapping quality of 40%, with reads required to have a minimum base quality of 10 and minimum mapping quality of 40 to be counted. For each sample, any genes whereby less than 50% of the gene was considered “callable” was excluded from consideration, unless a mutation had been called in that gene. By counting mutations even when they occurred in genes flagged as “uncallable”, we artificially increased the mutation prevalence in the untreated cohort, and thus biased ourselves against identifying genes enriched for mutations. We subsequently compared the mutation frequency of all genes and select hotspots (Supplemental Table S6, Appendix A) between the untreated and rrDLBCL cohort using a Fisher’s exact test and Benjamin/Hochberg false discovery rate correction, with a p_{adj} <0.1 considered significant (Supplemental Table S7, Appendix A). To ensure our differentially mutated genes were not enriched for mutations simply due to increased sequencing depth in our rrDLBCL cohort, and thus ability to detect mutations, we further compared the mutation frequency between our rrDLBCL cohort and an additional diagnostic cohort comprised of ultra-deep (500x) sequencing of 293 genes[248] (Supplemental Table S8, Appendix A).

2.4.7. Genetic subgroupings of rrDLBCL cases

We inferred genetic subgroup labels for each rrDLBCL tumor using the LymphGen classifier[249]. As our rrDLBCL cohort is primarily comprised of liquid biopsies with limited information of copy-number alterations and structural variants, we assigned genetic subgroupings using only simple somatic mutation data. To compare the genetic subgroup prevalence against our untreated DLBCL cohort, we restricted our analyses to the cohort of cases described by Schmitz *et al*[228], and selected only SNV features in genes sequenced in our rrDLBCL cohort, and disregarded copy-number alterations and structural alterations to ensure the results were comparable. Notably, this restricted set of features precluded cases from being assigned to A53 and reduced the sensitivity of the BN2 and ST2 subgroups. The prevalence of each subgroup was compared using a chi-squared test and Benjamini/Hochberg false discovery rate correction, with a p_{adj} < 0.1 considered significant (Supplemental Table S2 and S9, Appendix A).

2.4.8. Survival analysis in untreated DLBCL

The prognostic association of each differentially mutated gene or mutation hotspot was individually evaluated using overall survival (OS) time from 1670 samples in the untreated DLBCL cohort. For each exome, if less than 50% of a given gene was considered callable and no mutation was detected in that gene (see 2.4.6), the mutation status was considered missing data for that patient (NA) and was excluded from analysis. This strict criterion addresses the variable sequence coverage of some genes across different cohorts. Univariate survival analysis was performed using the survminer package (V 0.4.6, R version 3.5.3) using the Kaplan-Meier method[279]. Cox proportional hazard models[280] were fit using the survival package (V2.3.8, R version 3.5.3) (Supplemental Table S10, Appendix A). We fit additional Cox models incorporating International Prognostic Index (IPI), COO, and the source cohort for each case. Additional models were fit using *KMT2D* truncating mutations to evaluate the effects of truncating mutations on patient outcomes.

2.4.9. Site-directed mutagenesis of *MS4A1*

Site-directed mutagenesis (SDM) to generate *MS4A1* missense mutations was performed using NEB Q5-SDM (protocol E0554) kit. Bespoke non-overlapping primers were produced by LifeTechnologies (Supplemental Table S11, Appendix A). 5µL of the mutation reaction was used to transform C2571 cells (NEB) by heat shock transformation as recommended by NEB Q5-SDM kit. Plasmids were propagated in 10 to 100 mL bacterial cultures and purified using QIAGEN mini/maxiprep kits. All clones were sequenced at Source BioScience and sequences aligned using DNALasergene SeqManPro programme to confirm the mutagenesis and the absence of additional mutations.

2.4.10. Immunohistochemistry of tissue sections and cell lines

Immunohistochemistry was performed at the Segal Cancer Centre Research Pathology Facility (Jewish General Hospital). Tumor biopsies from consented patients were collected, formalin fixed, and paraffin embedded (FFPE) by clinical pathology at the Jewish General Hospital. FFPE blocks were cut at 4 µm and H&E stained to confirm tissue morphology. Separate tissue microarray blocks (TMAs) were constructed for

diagnostic and relapse samples by selecting representative tissue sections on the FFPE block, removing cores from the donor block and placing them into the corresponding TMA block. Cell line controls were added to the TMA blocks via similar preparation as tumor biopsies. Briefly, cell lines were expanded in standard culture conditions until confluence. Cells were pelleted, washed, and fixed in 10% neutral buffered formalin for 30 minutes at room temperature. Afterwards, the cell pellets were washed and placed in 70% ethanol for subsequent paraffin embedding. FFPE cell line cores were taken and placed into the corresponding TMAs, as previously mentioned.

Tissue samples were cut at 4-µm, placed on SuperFrost/Plus slides (TOMO, VWR) and dried overnight at 37°C, before IHC processing. Slides were then loaded onto the Discovery XT Autostainer (Ventana Medical System). Solutions used for automated immunohistochemistry were obtained from Ventana Medical System unless otherwise specified. Slides underwent d e- paraffinization, heat-induced epitope retrieval (CC1 prediluted solution Ref: 950-124, standard protocol). Double immunostaining for CD20 and PAX5 was sequentially performed online using a heat protocol. Mouse monoclonal anti-CD20 (Clone L26, Roche) was prediluted and auto- applied for 32 min at 37°C, followed by a detection kit (Omnimap anti-Mouse HRP Ref: 760-4310 and ChromoMap-DAB Ref: 760-159). Slides were washed with warm soapy water, followed by reaction buffer (Ref: 950-300) and loaded for a subsequent immunostaining with a rabbit monoclonal anti-PAX5 (Clone SP34, Roche) antibody. This was prediluted, and auto-applied for 32min at 37°C, then followed by the appropriate detection kit (OmniMap anti-Rabbit HRP Ref: 760-4311 and the Discovery Purple Kit, Ref: 760-229). A negative control was performed through omission of the primary antibody. Slides were counterstained with Hematoxylin for 8 minutes, blued with Bluing Reagent for 4 minutes, removed from the autostainer, washed in warm soapy water, dehydrated through graded alcohols, cleared in xylene, and mounted with Eukitt Mounting Medium (Eukitt, Fluka Analytical). Sections were analyzed by conventional light microscopy or scanned at 40X using the Aperio AT Turbo Scanner (Leica Biosystems).

2.4.11. Immunoblotting of cells expressing wild-type or mutant CD20

$5x10^6$ cells from each transfection were lysed in 30µL RIPA buffer (25 mM Tris-HCl, 150 mM NaCl, 1% Nonidet P-40, 1% sodium deoxycholate, 0.1% SDS containing

protease inhibitor mixture [Sigma], 50 mM NaF, and 0.2 mM Na3VO3). Whole cell lysate protein concentration was measured using a BioRad Protein Assay with known concentrations of Bovine Serum Albumin used as a standard curve. 20µg of lysate was loaded onto a 10% Bis-Tris gel and separated under a constant voltage. The proteins were transferred onto a nitrocellulose membrane using the iBlot (ThermoFisher), blocked with 5% w/v milk solution (Marvel) and probed with rabbit anti-CD20 (clone EP459Y, Abcam) overnight at 4°C. Unbound antibody was removed by washing three times in TBS-Tween, before bound antibody was detected with HRP-conjugated anti-Rabbit secondary antibody (NA9340, Sigma) at room temperature for 1 hr. The membrane was washed three times in TBS-Tween before ECL substrate (ThermoFisher) was added and detected using a ChemiDoc-It Imaging System. Equivalent loading was confirmed by blotting with rabbit anti-tubulin (Clone#2144, Cell Signalling Technology).

2.4.12. Single Cell Analysis of PT255

To isolate single cells from patient cell population (PT255), surface antigens were used to detect and sort the desired cells using flow cytometry (BD FACSAria Fusion, BD Biosciences, San Jose, California, USA). To stain and sort the cells, 10^6 cells were thawed, washed, and resuspended in warm PBS and stained with LIVE/DEAD Fixable Aqua Dead Cell Stain (ThermoFisher Scientific, Grand Island, New York, USA) to assess cell viability. Cells were washed twice with cold PBS and Fc were blocked with 20% serum for 15 minutes. Cells were washed with staining buffer [PBS supplemented with 2% foetal bovine serum (FBS; Wisent, St- Bruno, Québec, Canada)] and labeled (or not, for negative controls) with CD3 (1 µl; clone SK7; BD Biosciences), CD19 (1 µl; clone H1B19; BD Biosciences) and CD20 (5 µl; clone 2H7; BD Biosciences) antibodies for 1 hour on ice in the dark. Cells were then washed twice with flow cytometry buffer, resuspend in 100% FBS, and analyzed by flow cytometry. The gates for positive-staining cells were determined by comparison with unstained cells. Cells that were either CD3(-)CD19(+)CD20(-) or CD3(-)CD19(+)CD20(+) were single cell sorted in a 96 well plate for further genetic analysis.

Primers were designed for these and germline variants using Primer3 (Supplemental Table S12, Appendix A). All forward primers were tailed with the sequence CGCTCTTCCGATCTCTGNNNN, and all reverse primers were tailed with the sequence TGCTCTTCCGATCTGACNNNN for use in downstream sequencing. 48

single-plex PCRs were performed using the Fluidigm AccessArray as per manufacturer's protocols in Appendix G Section 4: Amplicon Tagging on the Access Array IFC, with some modifications described below. Pre-amplification was performed on isolated single cells and plasma DNA using Platinum Multiplex Master Mix, adjusting manufacturer's protocols for a 5μL reaction. A 48-plex primer mix (1μM each primer) was used. Thermal cycling conditions were 56°C(10min), 94°C(11min), 30X[94(30sec), 60°C(30sec), 72°C(30sec)], 4°C(hold). Barcoding was performed using primers with sequences: forward, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCT; reverse, CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCG (with the region denoted as XXXXXX reserved for sample indexes). Products were pooled for sequencing using the Post-PCR Amplicon Purification and Quantitation protocol and sequenced on an Illumina MiSeq.

Sequencing reads were aligned against the human reference genome hg19 using BWA mem[152]. A custom python script was used to calculate variant allele fractions and coverage at each of the 48 variant loci (Supplemental Table S13, Appendix A). Samples were analyzed in duplicate, and the average variant allele fraction was used for analysis. False positive variants were removed as well as loci where primers failed to give a total of 29 variants used for analysis.

2.5. Results

2.5.1. Enrichment of mutations in rrDLBCL

The pattern of mutations observed in rrDLBCL largely resembles that of untreated DLBCL (Figure 2-2). As this survey was focused on the genetic landscape following relapse, we searched for genes enriched for mutations after treatment failure. Such mutations are expected to represent either features of primary treatment resistance or examples of mutations subjected to clonal expansion under the selective pressures exerted by therapy. This analysis revealed 6 genes enriched for mutations: *KMT2D, TP53, CREBBP, FOXO1, NFKBIE*, and *MS4A1*, with another two genes depleted for mutations in rrDLBCL (Figure 2-3).

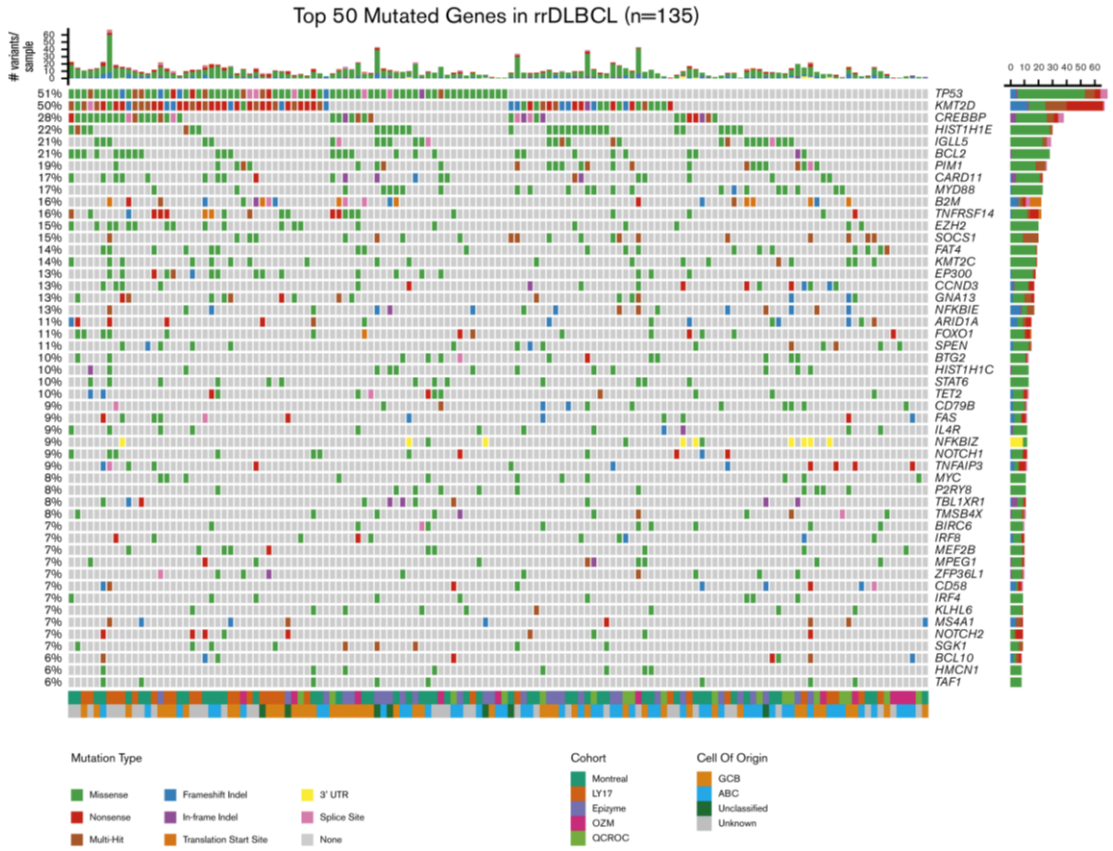


Figure 2-2. Mutation landscape of lymphoma-related genes in 135 rrDLBCL cases. Exonic mutations affecting the top 50 most recurrently mutated genes in our cohort of 135 rrDLBCL samples representing 5 different cohorts (Section 2.3.1). The inferred effect of each mutation is indicated by colour. Noncoding mutations are suppressed with the exception of *NFKBIZ*, which includes 3′ UTR mutations that have been previously described as driver mutations. The 2 covariate tracks on the bottom show COO information (where available) and the source cohort for each sample. Bar plots above and to the right of the plot indicate number of mutations per patient and number of patients with a mutation in that gene, respectively. Although the mutation landscape closely resembles untreated DLBCL, there are some notable differences. For example, approximately half of all rrDLBCLs harbored mutations in either *TP53* (51%) or the histone methyltransferase *KMT2D* (50%) with 31% of cases harbouring mutations in both genes.

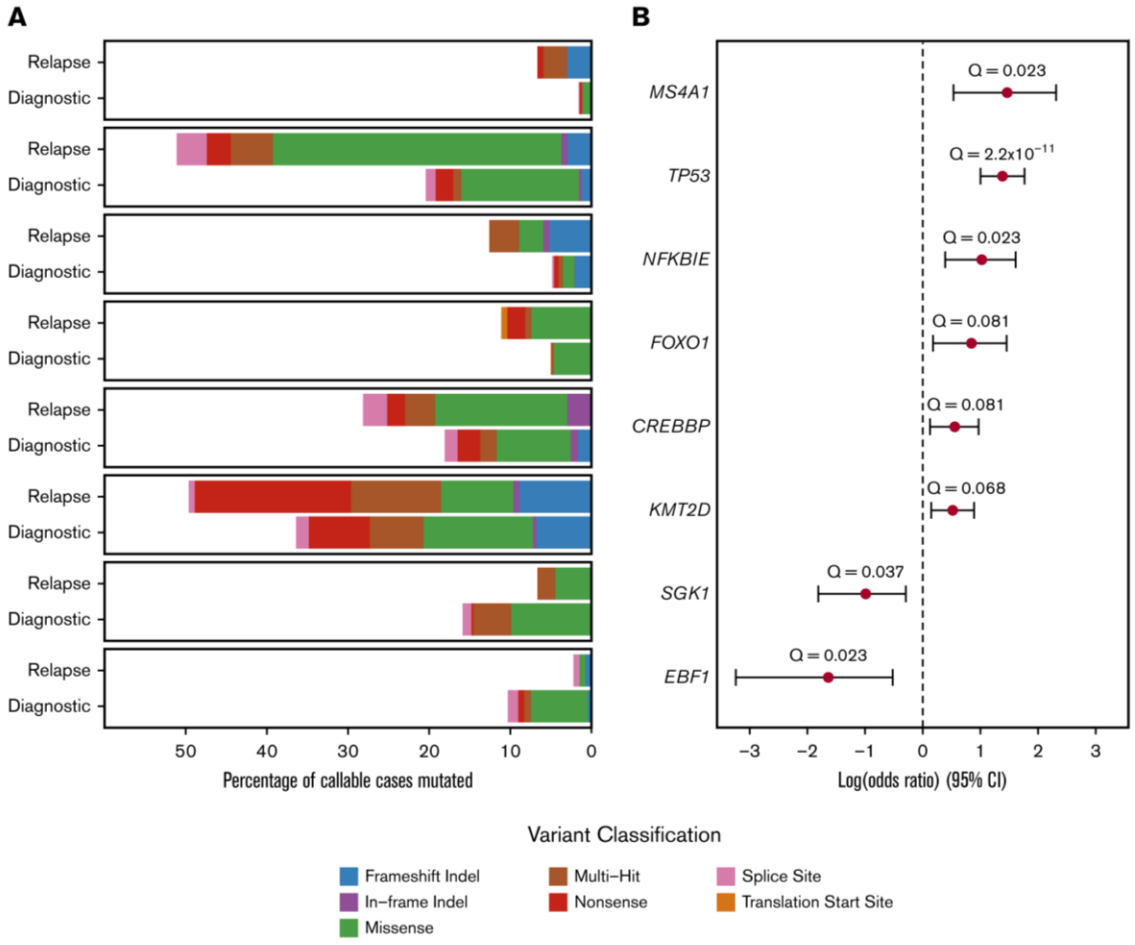


Figure 2-3. Differentially mutated genes between rrDLBCL and untreated DLBCL. (A) Mutation type and frequency of each differentially mutated gene in the untreated and rrDLBCL cohorts, using a significance threshold of 0.1 following false discovery rate correction. Untreated cases with insufficient coverage (not callable) in the gene of interest were not counted in the denominator for that gene (see 2.4.6). (B) Forest plot showing the odds ratio for all differentially mutated genes, as determined by the Fisher's exact test, for all differentially mutated genes (Supplemental Table S7, Appendix A). CI, confidence interval.

The lysine methyltransferase *KMT2D* is a tumor suppressor in DLBCL and follicular lymphoma (FL)[229] and was mutated in half of all rrDLBCLs (Figure 2-3A). In addition to a significant increase in *KMT2D* mutations in rrDLBCL relative to untreated DLBCL (q = 0.0678; OR, 1.68), loss-of-function mutations were further enriched at relapse (55/135 rrDLBCLs [40.7%] vs 304/1314 untreated [23.1%], q = 1.7×10^{-5}). Similar to the pattern in untreated cases, truncating mutations were observed across the length of the protein (Figure 2-4A) and tend to occur before the N-terminal SET domain, which catalyzes H3K4 methylation.

Figure 2-4. Mutation patterns in genes enriched for mutations within the population of rrDLBCLs. Lollipop plots displaying the mutations discovered in the 6 genes (*KMT2D* [A], *TP53* [B], *CREBBP* [C], *FOXO1* [D], *NFKBIE* [E], *MS4A1* [F]) found to be significantly enriched for mutations at relapse compared with untreated DLBCL. Mutations in rrDLBCL are displayed above each gene, and mutations in the untreated cohort are displayed below each gene. The number of mutated cases and percentage of cases with mutations in that gene are shown beside each gene (red: relapse; blue: untreated). The size of a lollipop and vertical displacement represent the number of patients with nonsilent mutations observed at that position. Note that lollipops were scaled down in the untreated cohort, and thus, the size of a lollipop cannot be directly compared between the untreated and relapse cohorts. Relevant protein domains are displayed for genes with differing mutation patterns within these domains. There is a general enrichment for recurrent mutations in the untreated cohort, most pronounced in *KMT2D*. These are attributed to rare germline variants that we were unable to filter due to their absence in any database of common variants.

The majority (51%) of rrDLBCLs harbored a *TP53* mutation ($q = 2.25 \times 10^{-11}$; OR 3.99). In contrast to *KMT2D*, mutations were predominately missense and affected the DNA-binding L1-sheet-helix domain (Figure 2-4B). We observed recurrent mutations affecting known TP53 hotspots, including Arg175, Arg248, and Arg273, which either bind directly to DNA or coordinate DNA binding[281]. The Arg248 residue, which directly binds

to the minor groove of DNA, was the only hotspot significantly enriched for mutations in rrDLBCL (q = 0.0807; OR 3.29).

The rrDLBCLs were also enriched for mutations affecting each of *CREBBP* (q = 0.0807; OR 1.74), *NFKBIE* (q = 0.0232; OR 2.78), and *FOXO1* (q = 0.087; OR 2.33). The majority of *CREBBP* missense mutations affected the acetyltransferase domain (28/48 mutations, 58%) (Figure 2-4C), with the remainder predominantly causing truncation. Mutations in *FOXO1* could broadly be defined into 2 classes: those that disrupt the forkhead DNA-binding domain and those that disrupt *FOXO1* phosphorylation (Figure 2-4D). The latter include mutations-targeted Tyr24, adjacent residues, or the canonical start codon, which both affect the regulation of FOXO1 nuclear localization[260]. We also observed recurrent frameshift deletions affecting Tyr254 in *NFKBIE*, a negative regulator of NF-κB signaling (Figure 2-4E). Although mutations in *NFKBIE* were significantly enriched in rrDLBCL, the prevalence of mutations affecting the Tyr254 hotspot was not significantly higher in this cohort.

MS4A1 exhibited the strongest enrichment for mutations in rrDLBCL (q = 0.023; OR 4.32). *MS4A1* encodes CD20, the B-lymphocyte antigen and target of rituximab and several other therapeutic monoclonal antibodies (mAbs). Although truncating mutations were observed across the length of *MS4A1*, a striking number of missense mutations were also observed (Figure 2-4F). None of these are predicted to directly affect residues comprising the rituximab epitope nor the epitopes of other mAbs. Instead, the recurrent missense variants were predicted to affect the transmembrane domains of the small loop, including 3 examples of a Tyr86 mutation (2 Tyr86Cys, 1 Tyr86His). Outside of rrDLBCL, mutations affecting this residue appear to be exceedingly rare as they were absent from the entire untreated DLBCL cohort and only appear in a single tumor in COSMIC[282].

2.5.2. Recurrent clonal selection following rituximab-based therapy

To further explore genetic mechanisms that contribute to treatment resistance, we inferred the clonal structure and dynamics in the 57 patients with serially collected samples representing time points prior to and following rituximab-containing treatment regimens, including *de novo* DLBCL, tFL, and other B-cell lymphomas. For each set of samples, we inferred the CCF of each mutation detected in pre-treatment tumor tissue

biopsies and post-treatment plasma samples. We then compared individual CCFs between paired samples and categorized mutations as enriched (clonal expansion), depleted (clonal regression), or stable at relapse within that tumor (Figure 2-5A). Figure 2-5B-I shows representative time series for mutations of interest. Overall, coding mutations affecting *TP53* (Figure 2-5B,D-I) and *KMT2D* (Figure 2-5D,F-G,I) tended to be stable following therapy, including all examples of *TP53* Arg248 mutations and *KMT2D* loss-of-function mutations (Figure 2-5D,F,I). We observed numerous examples showing clonal expansion of a single *KMT2D* mutation and clonal depletion of a separate *KMT2D* mutation, suggesting a persistent selective advantage for *KMT2D* loss (Figure 2-5G). Mutations affecting *CREBBP* and *NFKBIE*, including *NFKBIE* Tyr254, were similarly stable prior to and following treatment in most patients. Taken together, mutations in these genes appear to generally represent a component of the founding clone.

In contrast to these genes, *MS4A1* mutations exhibited a consistent trend toward clonal expansion in patients following rituximab-containing therapy (Figure 2-5B-E). This includes several cases inferred to harbor multiple subclonal populations with distinct *MS4A1* mutations, with each exhibiting clonal expansion (Figure 2-5D). *MS4A1* mutations were consistently undetectable in diagnostic tissue and appear to result from consistent positive selection under the pressure of R-CHOP and other therapies. This trend along with their predominance in rrDLBCL relative to untreated DLBCLs indicates a role of these mutations in contributing to acquired treatment resistance during exposure to rituximab-containing therapy.

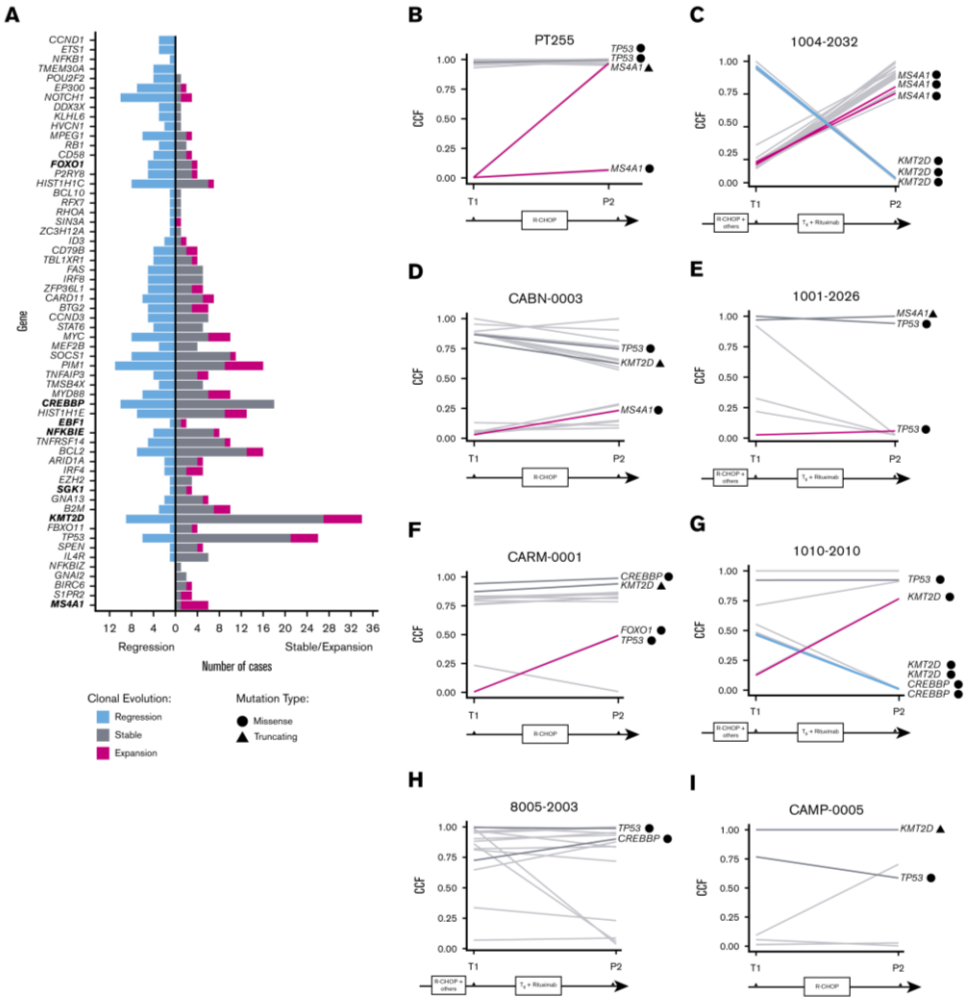
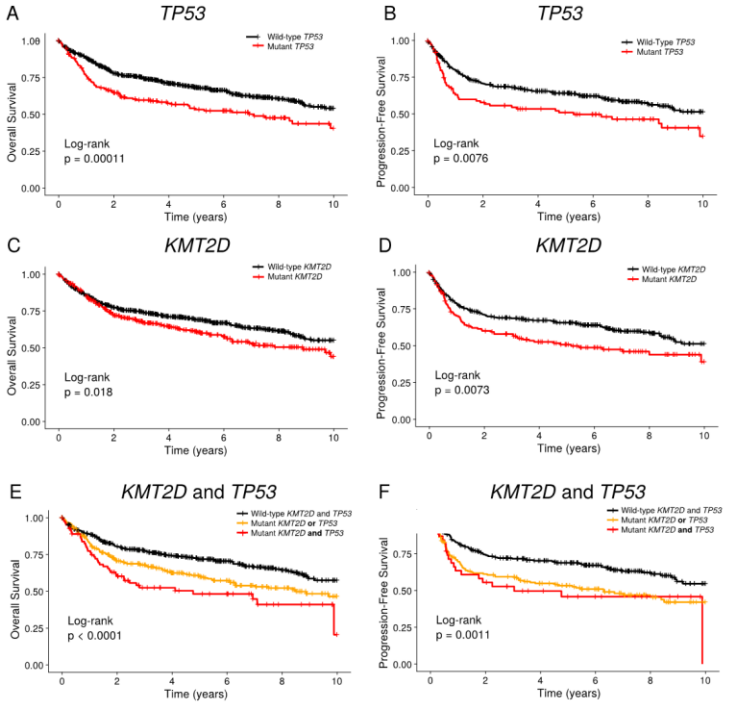


Figure 2-5. Clonal evolution patterns of regression and selection in rrDLBCL. (A) Number of cases with a mutation that regressed (blue), expanded (pink), or remained stable (gray) in a given gene following therapy. Time points are from a tumor biopsy before treatment (T1) and a plasma sample after treatment (P2). Genes in clusters that predominately undergo clonal expansion treatment are near the bottom, and genes in clusters depleted following treatment are near the top. (B-I) Clonal evolution plots for several patients following therapy, using a pretreatment tumor tissue biopsy and a posttreatment plasma sample. Each line represents a single coding mutation, and the relative CCF of each mutation before and after therapy is used to flag mutations that undergo clonal expansion (pink), depletion (blue), or remain stable (gray). The eight genes differentially mutated are labeled and highlighted, and the mutation type is indicated by the adjacent symbol.

2.5.3. *KMT2D* and *TP53* mutations are poor prognostic markers in diagnostic DLBCL

The clonal and stable nature of many rrDLBCL-associated mutations implies they may contribute to innate treatment resistance. In this scenario, such mutations should intuitively be associated with inferior outcomes. To assess this, we searched for associations between non- silent mutations in each of these genes and overall survival (OS) and progression-free survival (PFS) in our untreated cohort. Mutations in each of *TP53* and *KMT2D* were individually associated with shorter OS and PFS in diagnostic DLBCL (Figure 2-6A-D).



G

Variable	Hazard Ratio	95% CI	p value
Overall Survival			
KMT2D	1.34	[1.06;1.69]	0.0148
TP53	1.59	[1.24;2.05]	<0.001
Overall Survival			
KMT2D Truncating	1.34	[1.03;1.74]	0.03
TP53	1.58	[1.22;2.04]	<0.001
Progression-Free Survival			
KMT2D	1.41	[1.06;1.87]	0.0167
TP53	1.47	[1.08;2.00]	0.0151
Progression-Free Survival			
KMT2D Truncating	1.26	[0.93;1.72]	0.1361
TP53	1.49	[1.09;2.02]	0.0118

Figure 2-6. Prognostic potential of *KMT2D* and *TP53* mutations in untreated DLBCL. (A-F) Kaplan-Meier survival curves showing differences in overall survival (OS) (A,C,E) and progression-free survival (PFS) (B,D,F) in untreated DLBCL cases harbouring *TP53* mutations (A,B), *KMT2D* mutations (C,D), or both (E-F), using our cohort of 1670 untreated DLBCL cases. Cases with insufficient coverage (not callable) in the gene of interest were excluded from analysis (see 2.4.8). All cases were censored at 10 years. (G) Cox proportional hazard models for OS and PFS in untreated DLBCL. All 6 genes enriched for mutations in rrDLBCL were initially included in each model, but only *KMT2D* and *TP53* mutations remained significant following feature selection. All cases were censored at 10 years.

Whereas mutations in *TP53* have previously been characterized as a poor prognostic marker in DLBCL[261], *KMT2D* has not previously been independently associated with inferior outcomes. Cases harbouring both a *KMT2D* and *TP53* mutation displayed shorter OS (Figure 2-6E) and PFS (Figure 2-6F), a trend recently observed in mantle-cell lymphoma[283]. This extends to K*MT2D* truncating mutations, which are associated with inferior OS alone (Figure 2-7A-B), and in conjunction with *TP53* mutations (Figure 2-7C-D). In contrast, *MS4A1* mutations were not significantly associated with patient outcomes, a finding that we attribute to the low mutation prevalence in untreated DLBCL.

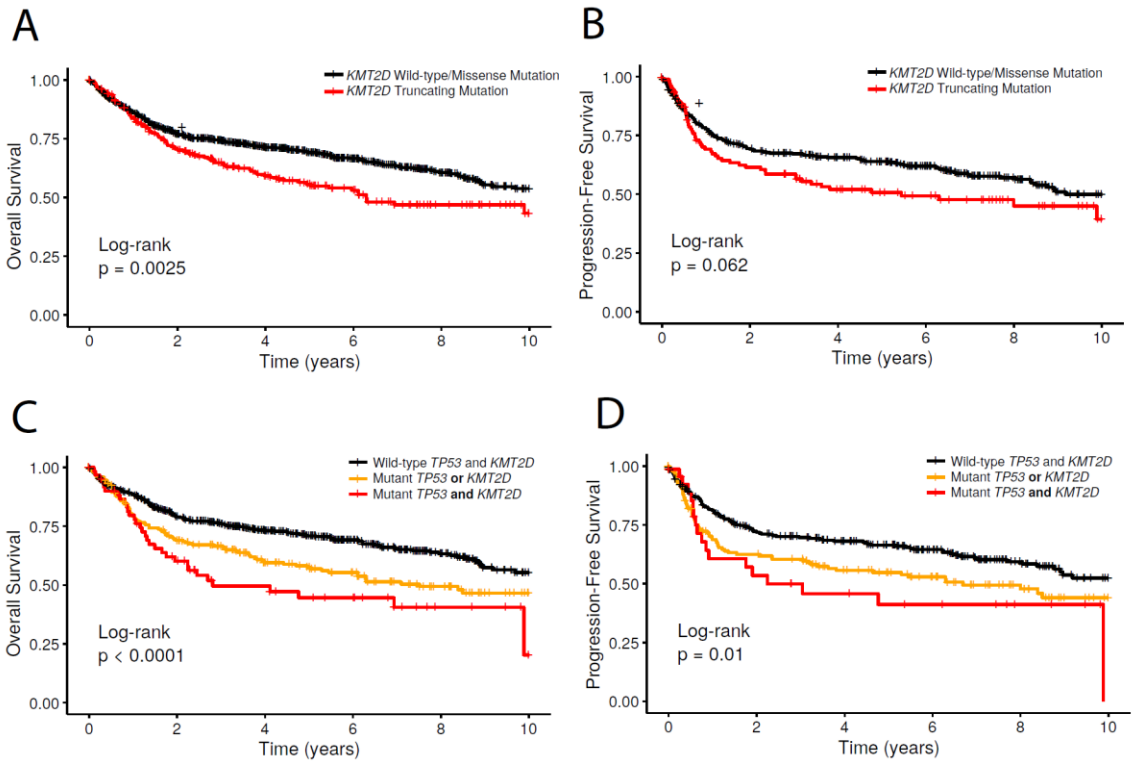


Figure 2-7. Prognostic potential of *TP53* and *KMT2D* truncating mutations in untreated DLBCL. Kaplan-Meier survival curves showing differences in overall survival (A,C) and progression-free survival (B,D) in untreated DLBCL cases harbouring *KMT2D* truncating mutations (A,B), or *KMT2D* truncating mutations and *TP53* mutations (C-D), using our cohort of 1670 untreated DLBCL cases. Cases with insufficient coverage (not callable) in the gene of interest were excluded from analysis (see 2.4.8). All cases were censored at 10 years for plotting.

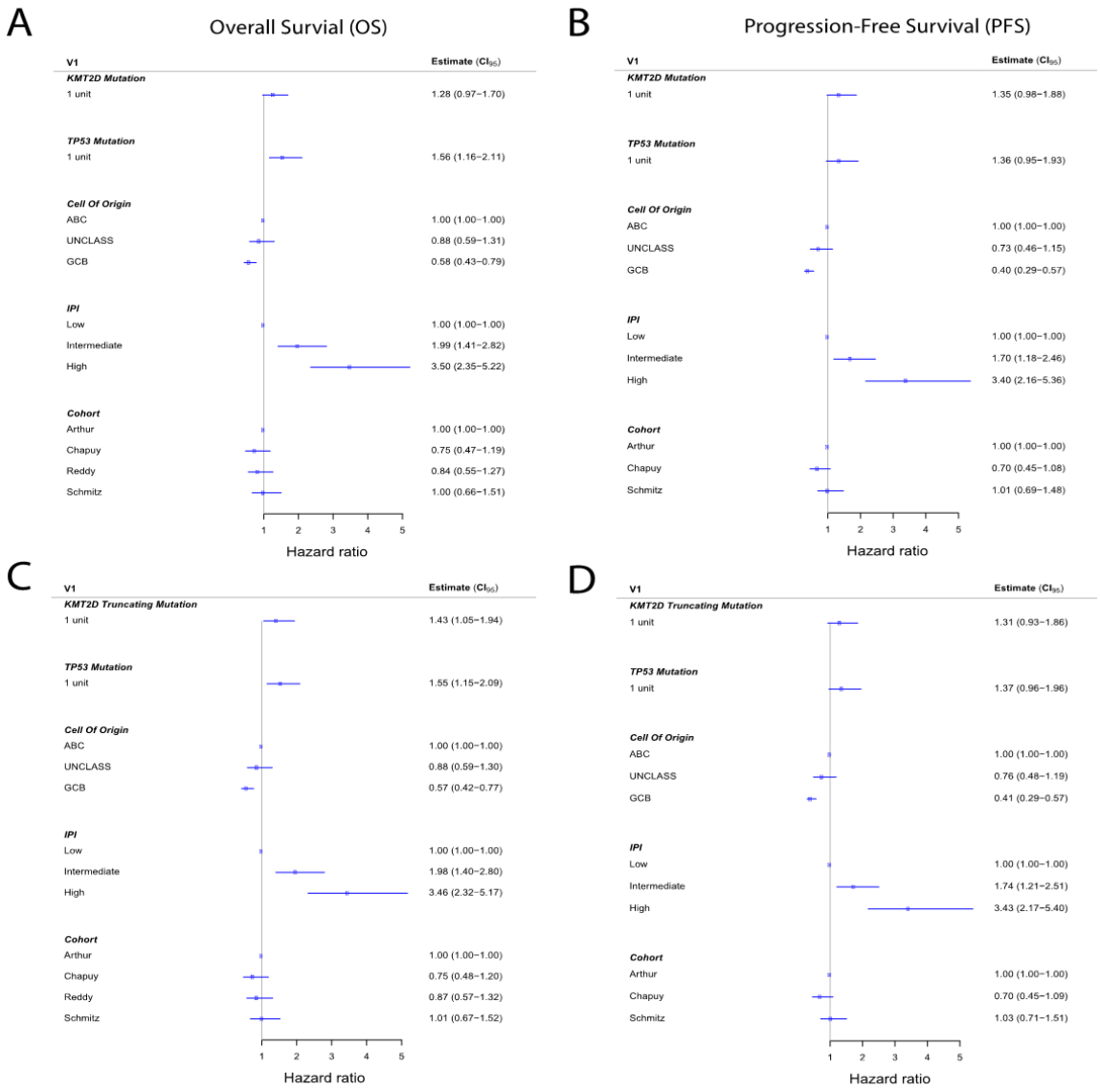


Figure 2-8. Prognostic association of *KMT2D* and/or *TP53* mutations on DLBCL. These forest plots summarize the Cox proportional hazard models for patient OS (A,C) and PFS (B,D) within the untreated DLBCL cohort. The features explored include *TP53* mutations, *KMT2D* mutations (A,B) or *KMT2D* truncating mutations (C,D), as well as IPI and COO classification. Cases lacking IPI, COO, or considered "uncallable" within *KMT2D* or *TP53* were excluded from analysis (see 2.4.8). PFS information was not provided for samples from the Reddy cohort[284].

We next tested showed that *KMT2D* and *TP53* mutations were significantly associated with inferior PFS and OS in a multivariate setting (Figure 2-6G), whereas *KMT2D* truncating mutations were only associated with inferior OS. In a model incorporating previously described prognostic features of DLBCL, including COO and International Prognostic Index classification, *KMT2D* and *TP53* mutation status remained

independently associated with shorter survival (Figure 2-8, Supplemental Table S10, Appendix A). Due to the high frequency of *TP53* and *KMT2D* mutations in untreated DLBCL, and because these mutations tend to remain stable following therapy, these each represent strong candidates for prognostic biomarkers of eventual treatment failure on R-CHOP.

2.5.4. Differential representation of EZB and MCD subgroups in rrDLBCL

As DLBCL is a genetically heterogeneous disease, recent studies have attempted to classify DLBCL tumors into subgroups based upon shared genetic features with therapeutic implications[227,228,248,249]. To explore this in rrDLBCL, we assigned rrDLBCL tumors into genetic subgroups, which enabled classification of 49% (66/135) of cases (Supplemental Table S2, Appendix A). Cases were most commonly classified as EZB in rrDLBCL (45/135 cases, 33%). In contrast, the prevalence of BN2 (5/135, 3.7%) and MCD (6/135, 4.4%) tumors was lower than previously reported[249]. The majority of ABC-DLBCL tumors (31/47, 66%) were not assigned to any genetic subgroup. Given our limited feature set, we compared the prevalence of cases in each genetic subgroup to a subset of the untreated DLBCL cohort (see 2.4.7). EZB tumors were significantly over-represented among our rrDLBCL cohort (33% vs 17%, q=0.00023) whereas MCD tumours were significantly less prevalent (4.4% vs 14%, q=0.014) (Supplemental Table S9, Appendix A). An enrichment of EZB tumors was not particularly surprising given the inferior prognosis of some of these cases, particularly those described as EZB-M+. Given the inferior outcomes of MCD tumours in ABC-DLBCL[78], a reduced representation of these cases in rrDLBCL was unexpected. As we anticipate additional methods for performing genetic classification of DLBCL, further exploration of the distribution of each genetic subgroup in rrDLBCL is clearly warranted.

2.5.5. Mutations in *MS4A1* attenuate rituximab binding

We next explored the functional effects of *MS4A1* mutations and their potential role in promoting rituximab resistance. We transfected a CD20− cell line with wild-type (WT) or mutant CD20 constructs representing common *MS4A1* missense mutations observed in patients (Figure 2-9A). We showed that all 3 *MS4A1* mutants had significantly decreased binding of rituximab or other anti-CD20 antibodies, including

tositumomab (B1), ofatumumab, and obinutuzumab derivatives by flow cytometry, with two mutants (Y86C and L66R) showing a complete absence of binding (Figure 2-9B; Figure 2-10). Consistent with the other mutation tested, cells with ectopic expression of Tyr86Cys were not recognized by any of the anti-CD20 antibodies. In contrast, cells expressing Tyr86His were recognized by all four antibodies, albeit at a significantly reduced amount. Because this assay requires expression on the plasma membrane, we next explored whether the mutations affected the expression of CD20 within the cell using immunoblotting with a CD20 antibody that binds within the cytoplasmic domain. Consistent with the result from flow cytometry, an immunoblot of cell lysates showed reduced CD20 protein with the Y86H mutant and no visible expression with the other two mutants (Figure 2-9C). We separately performed CD20 staining on cell lines derived from tumors naturally harbouring *MS4A1* mutations. Both Gly98Arg and Tyr86His cell lines were negative for CD20 staining by immunohistochemistry using L26, another mAb recognizing the C-terminal cytoplasmic region of CD20 (Figure 2-9D). Taken together, we conclude that *MS4A1* missense mutations can directly contribute to rituximab resistance by reducing CD20 expression and/or stability.

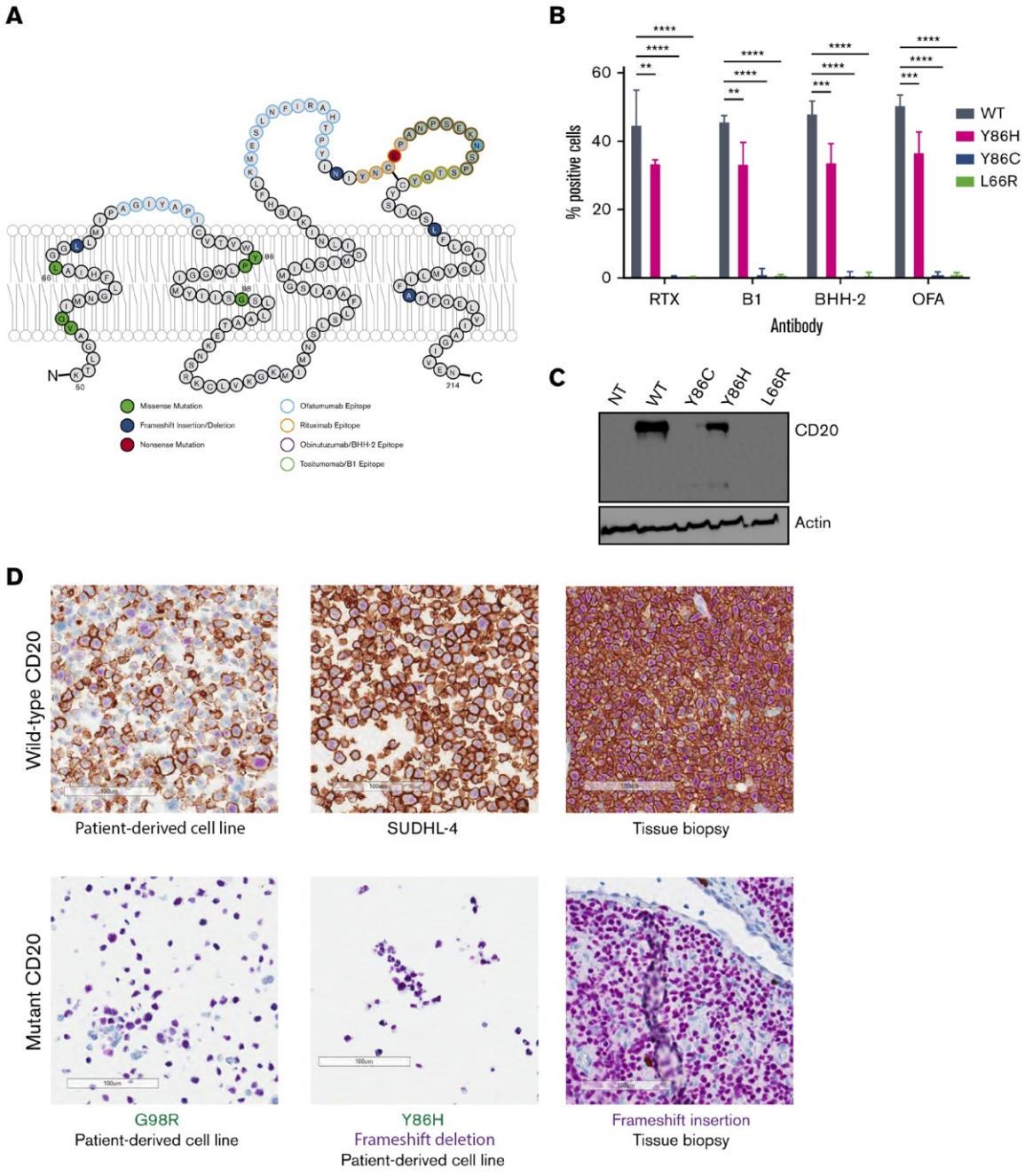

A
Missense Mutation
Frameshift Insertion/Deletion
Nonsense Mutation
Ofatumumab Epitope
Rituximab Epitope
Obinutuzumab/BHH-2 Epitope
Tositumomab/B1 Epitope
B
% positive cells
WT
Y86H
Y86C
L66R
RTX
B1
BHH-2
OFA
Antibody
C
NT
WT
Y86C
Y86H
L66R
CD20
Actin
D
Wild-type CD20
Patient-derived cell line
SUDHL-4
Tissue biopsy
Mutant CD20
G98R
Patient-derived cell line
Y86H
Frameshift deletion
Patient-derived cell line
Frameshift insertion
Tissue biopsy

Figure 2-9. Distribution and functional impact of *MS4A1* mutations in rrDLBCL. (A) Topology of *MS4A1* transmembrane domains and extracellular loops, as annotated by Uniprot and elsewhere[285]. *MS4A1* mutations observed in the rrDLBCL cohort have been labeled, along with the predicted binding epitope of 4 different CD20 mAbs. (B) Comparison of antibody binding between CHO-S cells transfected with plasmids expressing either WT CD20 or 1 of 3 mutants (Tyr86His, Tyr86Cys, and Leu66Arg) for 4 different CD20 antibodies: rituximab (RTX), tositumomab (B1), obinutuzumab (BHH-2), or ofatumumab (OFA). The percent of positively stained cells was compared between mutants within each antibody (adjusted P values from 2-way analysis of variance of 3 replicates: *P > .1, **P > .01, ***P > .001, ****P > .0001). See also Figure 2-10. (C) Representative western blot (of 2 independent experiments performed) showing CD20 expression of CHO-S cells transfected with WT or mutant CD20 (Y86H, Y86C, and L66R) and a nontransfected (NT) control. (D) Immunohistochemistry of CD20 in a cell line and tumor tissue biopsy harbouring WT CD20 as well as 2 patient-derived cell lines harbouring G98R (PT255), and Y86H along with a frameshift mutation, respectively. CD20 is stained red using the L26 CD20 antibody and B-cell nuclei were stained purple using a Pax5 antibody, visualized at ×20 original magnification.

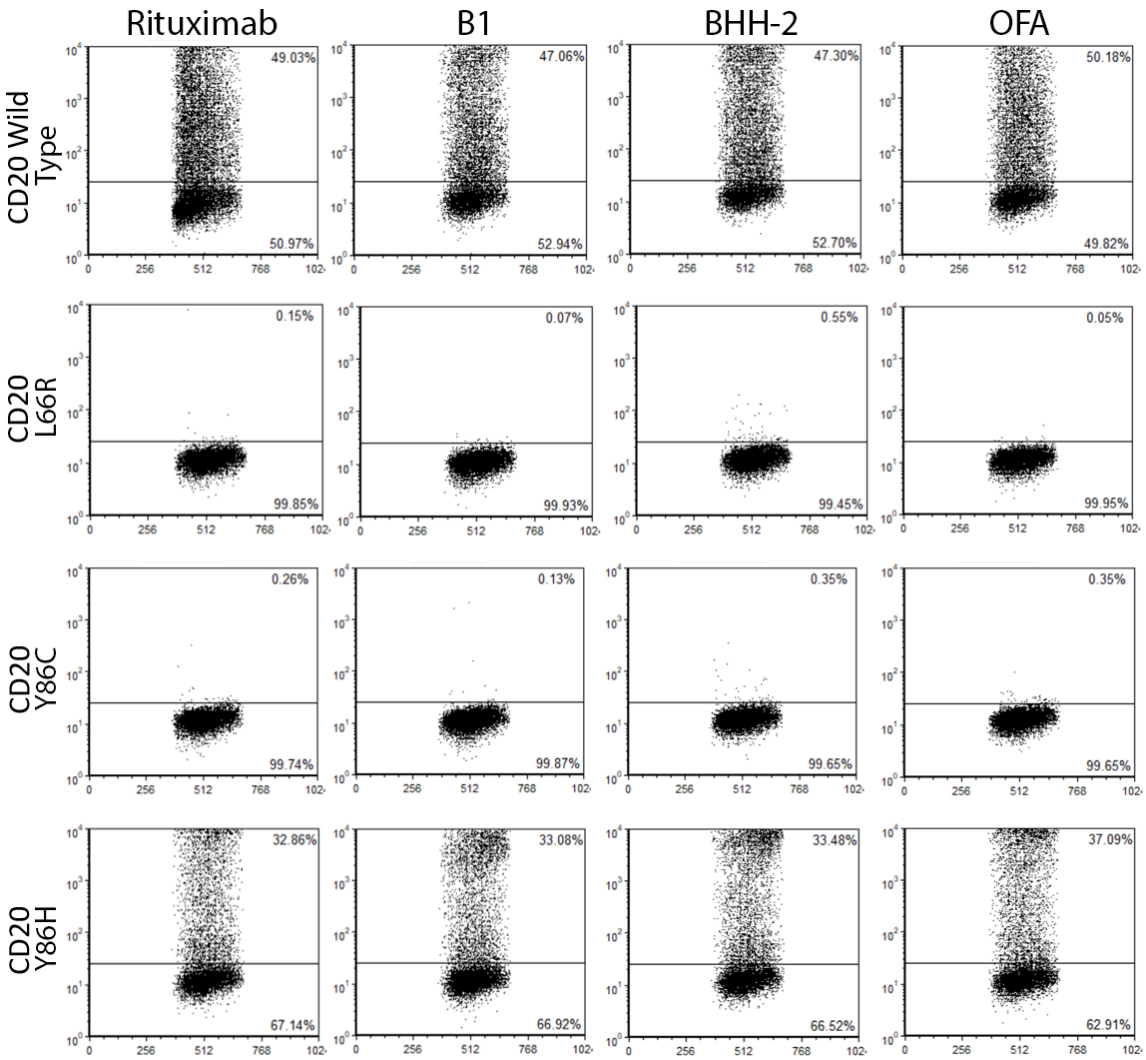


Figure 2-10. Comparison of CD20 binding of CHO-S cells transfected with plasmids containing either wild-type or mutant (Y86H, Y86C and L66R) MS4A1. CHO-S cells were transfected with plasmids and then assessed 24 hours later for CD20 surface expression using different CD20 antibodies: rituximab, tositumomab (B1), obinutuzimab (BHH-2) or ofatumumab (OFA) followed by detection with a PE-labelled $F(ab)_2$ secondary antibody. Representative FSC x FL1 plots shown, indicating cut-offs for % positive cells (upper portion).

2.5.6. *MS4A1*-harbouring subclones drive rapid treatment resistance

To further explore how multiple rounds of therapy can influence clonal structure in a *MS4A1*-mutant patient (PT255), we followed the progression of a patient with chemorefractory aggressive high-grade B-cell lymphoma using multiple complementary approaches (Figure 2-11). We initially performed exome sequencing on 3 time points beginning with the untreated diagnostic tumor biopsy (diagnosis, D), followed by cfDNA

collected after failure of both R-CHOP and subsequent high-dose chemotherapy (relapse 2, R2/P1) and a second cfDNA sample following additional rounds of chemotherapy including prednisone (relapse 3, R3/P5) (Figure 2-11A). We identified several distinct subclonal populations in these samples (Figure 2-11B) and selected mutations representative of each population: trunk (clonal), R2-associated (high prevalence at R2), and R3-associated (high prevalence at R3) for validation. We measured the variant allele frequency for these representative mutations in each time point and additional cfDNA samples and circulating tumor cells from blood collected between R2 and R3.

This analysis revealed a heterogenous clonal structure consisting of distinct subclones at each relapse (Figure 2-11C). Following R-CHOP and high-dose chemotherapy (R2), we observed emergence of a population containing an *MS4A1* truncating mutation and a missense mutation within the kinase domain of *DDR2*, which has been described in lung cancer and may confer susceptibility to dasatinib[286]. This R2-associated subclone was undetectable at R3 and was replaced by a distinct population harbouring a *MS4A1* missense mutation (Gly98Arg) and a truncating mutation affecting *NR3C1*, which encodes the glucocorticoid receptor and could contribute to resistance against steroids such as prednisone[287]. Although some mutations present in this later population were detectable at low levels following deep sequencing at R2, the extent of clonal expansion was striking given that <3 weeks elapsed between R2 and R3. In particular, this subclone exhibited rapid clonal expansion in the 8 days separating samples P3 and P4. Taken together, these data show that rapid changes in clonal structure can contribute to treatment resistance in DLBCL.

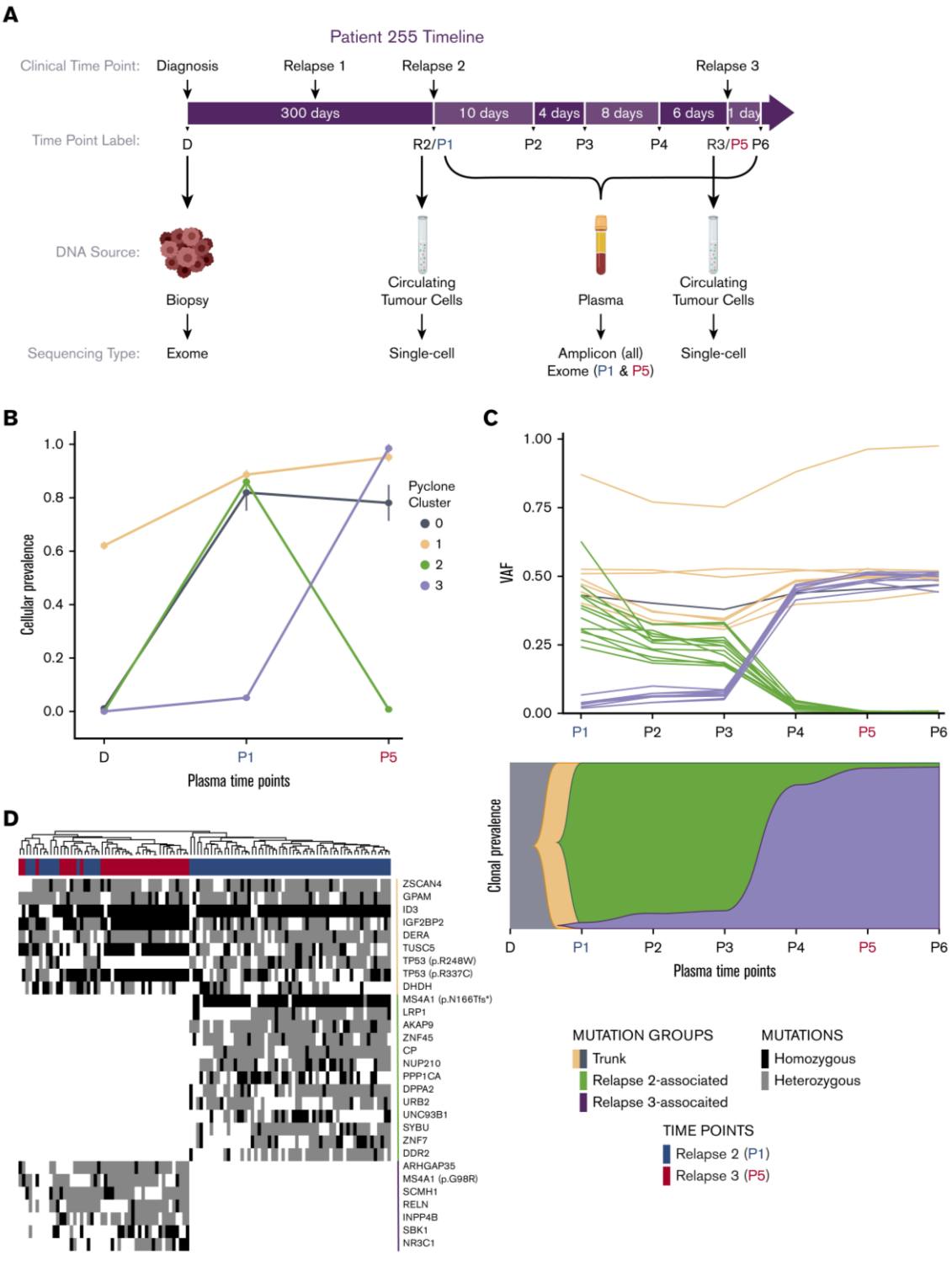

A
Patient 255 Timeline
Clinical Time Point:
Diagnosis
Relapse 1
Relapse 2
Relapse 3
300 days
10 days
4 days
8 days
6 days
1 day
Time Point Label:
D
R2/P1
P2
P3
P4
R3/P5
P6
DNA Source:
Biopsy
Circulating Tumour Cells
Plasma
Circulating Tumour Cells
Sequencing Type:
Exome
Single-cell
Amplicon (all)
Exome (P1 & P5)
Single-cell
B
Cellular prevalence
Plasma time points
Pyclone Cluster
0
1
2
3
C
VAF
Clonal prevalence
Plasma time points
D
ZSCAN4
GPAM
ID3
IGF2BP2
DERA
TUSC5
TP53 (p.R248W)
TP53 (p.R337C)
DHDH
MS4A1 (p.N166Tfs*)
LRP1
AKAP9
ZNF45
CP
NUP210
PPP1CA
DPPA2
URB2
UNC93B1
SYBU
ZNF7
DDR2
ARHGAP35
MS4A1 (p.G98R)
SCMH1
RELN
INPP4B
SBK1
NR3C1
MUTATION GROUPS
Trunk
Relapse 2-associated
Relapse 3-associated
MUTATIONS
Homozygous
Heterozygous
TIME POINTS
Relapse 2 (P1)
Relapse 3 (P5)

Figure 2-11. Plasma and single-cell sequencing of multiple time points in a DLBCL patient (PT255). (A) Timeline of events for PT255. Clinical time point shows the timing of diagnosis (D) and relapses (R2, relapse 2; and R3, relapse 3) relative to blood sample collection (P1 to P6). Bulk tumor DNA was separately obtained from a biopsy at diagnosis, circulating tumor cells extracted at R2 and R3, and cfDNA extracted from plasma samples P1 to P6 after R2. Varying types of sequencing was performed on DNA from each time point, as summarized below. (B) Results from running PyClone[277] on exome sequencing of DNA obtained from diagnosis, R2(P1) and R3(P5). Clusters 0 and 1 contain trunk mutations seen at both P1 and P5; cluster 2 contains R2-specific mutations, and cluster 3 contains mutations that were subclonal at R2 and clonal at R3. (C) Amplicon sequencing of a subset of mutations found in the clusters in panel B from all 6 plasma time points reveal a more complete but similar evolution of the tumor as inferred from bulk sequence analysis in panel B. Below shows the suspected proportion of the tumor made up of each clone at individual time points. (D) Single-cell amplicon sequencing of circulating tumor cells taken at R2 and R3 revealed 2 distinct populations of cells containing mutations specific to each of R2 and R3. Genes are ordered by group and by frequency of mutation detected (top to bottom), suggesting a relative order of mutation acquisition.

Given the emergence of subclones with different genetic features, we next sought to validate the clonal dynamics observed in R2 and R3 using single-cell sequencing. We determined mutation status and ploidy for the same set of mutations in a total of 74 isolated single cells from R2 and 35 isolated cells from R3. This confirmed that the R2-associated and R3-associated subclones exist in mutually exclusive subpopulations and confirmed the subpopulation at R2 representing the dominant clone found at R3 (Figure 2-11D). This also revealed genetic features that could not be inferred from bulk sequencing alone, such as a 17p deletion affecting *TP53* and *TUSC5* in the R3-associated clone and *MS4A1* loss of heterozygosity in the R2-associated clone. The *MS4A1* missense (R3-associated) and frameshift (R2-associated) mutations were detected in the majority of cells from each time point and thus were interpreted to represent early events in the foundation and development of these individual subclones. The vast majority of cells (> 99%) were negative for cell surface expression of CD20, consistent with each of these *MS4A1* mutations causing loss of CD20 expression. As rituximab can persist for weeks following treatment[288], these *MS4A1* mutations likely provided the founder cells with a strong selective advantage, resulting in the independent emergence of multiple resistant subclones.

2.6. Discussion

By comparing the genetic landscapes of untreated and rrDLBCL, we highlight the potential role of two DLBCL-associated genes, *KMT2D* and *TP53*, as contributors to primary treatment resistance. Mutations affecting these genes were enriched in rrDLBCL and were typically clonal in matched pretreatment samples (Figure 2-5). *TP53* mutations are known to be associated with inferior patient outcomes in DLBCL[261,262], shown to be enriched for mutations in rrDLBCL[259], and can contribute to resistance against chemotherapeutics, which induce DNA damage[289,290]. For instance, mutations affecting Arg248, a residue critical in DNA-binding that was enriched for mutations in our rrDLBCL cohort, can increase expression of cytochrome P450, which promotes resistance against a diverse range of chemotherapeutics in vitro and leads to inferior patient outcomes[291]. Given the high prevalence of *TP53* and *KMT2D* mutations in untreated DLBCL (20.7% and 36.9%, respectively), and their clonal prevalence and stability, mutations affecting these genes likely contribute to lymphomagenesis and primary refractory disease. Indeed, *KMT2D* mutations have been described as early drivers in DLBCL and FL[292] and contribute to increased cell survival and proliferation[293]. Although loss of H3K4me3 methylation results in transcriptional repression of numerous tumor suppressor genes, the contribution toward treatment resistance remains to be elucidated. *KMT2D* is commonly mutated in DLBCL overall, and these mutations may be enriched in the C3 genetic subgroup, which is associated with inferior prognosis within GCB-DLBCL[227]. Here, *KMT2D* mutations were associated with inferior prognosis in our untreated cohort regardless of COO or IPI (Figure 2-8). Together with genetic features such as double-hit/triple-hit, COO, and the DHITsig expression signature[91], *KMT2D* and *TP53* mutations may facilitate the identification of high-risk patients for alternative treatments.

One barrier that has limited the genetic exploration of rrDLBCL is the lack of tissue biopsies, which are not routinely collected upon relapse. With sufficient levels of ctDNA, liquid biopsies have been shown to accurately reflect that mutational landscape of both the primary tumor and the distal sites[269]. Collection of posttreatment liquid biopsies is gaining adoption as it can noninvasively inform on treatment response[270,272,294] and, as demonstrated herein, affords the opportunity for serial sampling such that clonal dynamics can be inferred within the context of treatment resistance. As some patients in this study exhibited rapid changes in population

structure (Figure 2-11), noninvasive methods will be required to allow prospective detection of resistance-associated mutations.

In this study, mutations in *MS4A1* recurrently exhibited clonal expansion following rituximab-based therapy (Figure 2-5). Single-cell analysis of a case harbouring two mutually exclusive MS4A1-containing subclones revealed that these mutations were acquired after exposure to R-CHOP and became founder events for the multiple subclones that occurred at both relapses (Figure 2-11). Curiously, many *MS4A1* mutations were not predicted to truncate the protein, and these missense variants did not directly affect the rituximab binding epitope. Prior work utilizing Sanger sequencing and a smaller rrDLBCL cohort also found limited evidence for MS4A1 mutations within the rituximab epitope[295], leaving the phenomenon of reduced CD20 expression unexplained. We explored the influence of these mutations on anti-CD20 antibody interactions and found that common missense mutations attenuated mAb recognition (Figure 2-9B), largely as a result of reduced expression, with patient-derived cell lines harbouring these mutations appearing on CD20− (Figure 2-9D). Although the underlying mechanism of CD20 loss stemming from these transmembrane domain missense mutations remains unresolved, the most likely explanation is that they impair correct protein folding and subsequent stable expression, rather than simply destroying the antibody epitope(s), as five different mAbs were unable to detect expression, including one targeting the cytoplasmic domain in the C terminus. Given the low mutation frequency and low clonal prevalence of *MS4A1* mutations prior to therapy, we hypothesize that these mutations provide limited (if any) fitness advantage until the tumor is exposed to anti-CD20 antibodies. Furthermore, this suggests additional unidentified mechanisms by which tumor cells inhibit CD20 surface expression, possibly through other genetic or epigenetic mechanisms. These findings reinforce the necessity of evaluating tissue biopsies following relapse for CD20 expression in trials including immunotherapy targeting this protein. These CD20− non-Hodgkin lymphoma cases are known to have poor outcomes with available therapies[296] and thus represent a population in need of alternative therapies.

In summary, we have identified six genes that are significantly enriched for mutations in rrDLBCL: *KMT2D, TP53, CREBBP, NFKBIE, FOXO1*, and *MS4A1*. The enrichment of *KMT2D* mutations in the rrDLBCL population and its association with inferior outcome suggests distinct biology or natural history of these DLBCLs, as *KMT2D*

and *CREBBP* mutations are among the most common genetic feature of FL. One explanation for our observations is that a substantial proportion of de novo DLBCLs result from occult transformation from FL. Further evidence supporting this possibility has recently been gleaned through the genetic analysis of DLBCLs with *MYC* and *BCL2* translocations[297]. In contrast to these early mutations, *MS4A1* mutations are rare in untreated DLBCL and were generally undetectable prior to therapy. Our data indicate that these mutations directly contribute to rituximab resistance, resulting in rapid clonal selection and expansion in the presence of rituximab-containing therapy. Furthermore, our single-cell data highlight the significant clonal heterogeneity of rrDLBCL, and the contribution of *MS4A1* mutations toward rapid treatment resistance. The recurrent loss of CD20 expression in the rrDLBCL population may have profound implications given the widespread use of rituximab and the ongoing targeting of CD20 with additional mAbs and more modern forms of immunotherapy.

Chapter 3.
Recurrent copy number alterations contribute to a distinct genetic landscape in rrDLBCL

This chapter has been prepared as a manuscript draft. Christopher K. Rushton, Ryan Rys, Elizabeth Chavez, Laura Hilton, Miguel Alcaide, Kostia Dreval, Matthew Cheung, Manuela Cruz, Krysta Coyle, Barbara Meissner, Susana Ben-Neriah, Neil Michaud, Scott Daigle, Jordan Davidson, Jasper Wong, Michael Jain, Lois Shepherd, Marco Marra, John Kurivilla, Michael Crump, Koren Mann, Sarit Assouline, Joseph M. Connors, Christian Steidl, David W. Scott, Nathalie A. Johnson, and Ryan D. Morin

Contributions: R.D.M., D.W.S., and N.A.J. conceptualized the study; C.K.R., R.D.M., R.R., and N.J. provided the methodology; C.K.R. provided the software; C.K.R., M.A., K.D., K.C. and L.K.H. provided the formal analysis; C.K.R. and R.R. led the investigation; R.R., E.C., L.H., M.A., K.D., M. Cheung., M.C., K.C., B.M., S.B., N.M., S.D., J.D., J.W., M.J., L.S., M.M., J.K., M.C., K.M., S.A., provided the resources; C.K.R., R.R., and N.A.J. curated the data; C.K.R., and R.R. contributed to writing of the original draft:; C.K.R. provided the visualization; C.K.R., and R.R. contributed to the writing, review, and editing; R.D.M., N.A.J., and D.W.S. provided supervision; and R.D.M., D.W.S., N.A.J., J.M.C., M.M., and C.S. contributed to acquiring funding.

3.1. Abstract

Patients with diffuse large B-cell lymphoma (DLBCL) are generally treated with immunochemotherapy (R-CHOP), but for the 30-40% of patients who relapse or who have treatment-refractory disease (rrDLBCL), prognosis is generally poor. While numerous novel therapies have been developed for rrDLBCL patients with promising efficacy, there remain a notable portion of patients where these treatments are unavailable or ineffective. To explore genetic features which contribute to rrDLBCL biology and identify mechanisms of treatment resistance or actionable events, we performed a combination of whole genome sequencing (WGS) and whole exome sequencing (WES) on 107 plasma (liquid) and tumour (tissue) biopsies collected following R-CHOP and combined this with rrDLBCL sequencing data from previous studies for a total of 155 cases with exome sequencing data. Following somatic variant

calling exome-wide, we identified four genes significantly enriched for mutations at relapse (*KMT2D*, *TP53*, *STAT6*, and *MYC*), and three (*TMEM30A*, *TET2*, and *BCL10*) significantly depleted for events. Based on mutational profiles, EZB was the most predominant subgroup in our cohort (26.3% of cases), while the majority (54.5%) of ABC-rrDLBCLs were unclassified via LymphGen. We further bolstered our cohort with low-pass WGS data from 67 rrDLBCL liquid biopsies, for a total of 222 cases with genome-wide copy number data. Analysis of these data identified 13 regions significantly enriched for CNVs in rrDLBCL, including well described lymphoma drivers (*TP53, PTEN, STAT6*, *MIR17HG*), as well as several novel rrDLBCL CNVs. These include recurrent deletions of MHC Class I regulator *IRF2*, RNA splicing regulator *HNRNPD*, and gains of genes involved in B-cell maturation and differentiation, including *IKZF3* and *TCF3*. The reduced representation of *TET2* mutations and low prevalence of ST2 cases could imply that such cases are also less likely to relapse on standard therapy. The frequent observation of deletions affecting *HNRNPD* points to an under-appreciated role of RNA-binding proteins in DLBCL relapse, while deletions of *IRF2* could contribute to the propensity of rrDLBCL to evade destruction by the immune system.

3.2. Introduction

Non-Hodgkin lymphoma (NHL) is the 6th most common form of cancer in Canada, with an estimated 11,400 new cases diagnosed each year[298]. 60% of NHL cases classified are classified diffuse large B-cell lymphoma (DLBCL)[299] which is characterized by its genetic, phenotypic, and clinical heterogeneity[300]. DLBCL can arise *de novo* or through histological transformation from other lymphoid malignancies, most commonly follicular lymphoma[301,302]. DLBCL patients are generally treated with a frontline immunochemotherapy (R-CHOP)[256], which is effectively curative for 60-70% of cases[102,103]. For the subset of cases where frontline treatment fails (relapsed-refractory DLBCL, rrDLBCL), patient outcomes are generally poor, especially cases refractory to frontline therapy[108]. A plethora of salvage therapies are under investigation to improve rrDLBCL treatment, including numerous targeted therapies[113,114,303–306], bi-specific antibodies[307,308], and CAR-T cell therapy[123,309]. While many of these experimental therapies show promise, notably CAR-T cell therapy, patient long-term outcomes remain heterogeneous and unacceptably poor. To improve the outcomes of patients requiring

salvage therapy, the genetic features of rrDLBCL must be characterized to identify recurrent genetic aberrations that may lead to novel therapeutic strategies.

A popular approach to stratifying cases in hopes of overcoming the heterogeneity characteristic of DLBCL is to group cases that harbour similar molecular or genetic features reflecting shared biology. In the cell-of-origin (COO) system, DLBCL can be divided into two molecular subgroups based on gene expression patterns: activated B-cell like (ABC), characterized by constitutive NF-kB signaling, and germinal center-like (GCB), with ABC-DLBCL displaying inferior outcomes following R-CHOP[310]. Recently, several groups have identified genetic subgroups in DLBCL with prognostic and therapeutic implications[227,228,248,249]. In contrast, the LymphGen system assigns cases into six genetic subgroups based on single nucleotide variants and small insertions and deletions (cumulatively "simple somatic mutations", SSMs), copy number variants (CNVs) and structural variants (SVs), with MCD and EZB subgroups representing a subset of ABC and GCB cases, respectively, with inferior outcomes[249]. While ~60% of diagnostic DLBCL tumours are assigned into a genetic subgroup via the LymphGen algorithm, with a high prevalence of MCD, BN2, and EZB cases, the prevalence and frequency of these genetic subgroups in rrDLBCL has not been established.

Several large-scale studies have characterized the landscape of somatic alterations in DLBCL, mostly in the context of the COO subgroups[15,226–228,236]. GCB-DLBCL is dominated by recurrent mutations perturbing epigenetic modifiers, including *GNA13*[311], *EZH2*[312], *CREBBP*, and *EP300*, as well as recurrent copy number gains of *REL* and deletions of *TNFRSF14*, *B2M*, *PTEN*, and *FAS*[236]. ABC-DLBCL is characterized by constitutively active NF-kB signaling, with characteristic hotspot mutations affecting the NF-kB and JAK-STAT signaling regulator *MYD88*[20,313], *CD79B*[232], *CARD11*, *PRDM1*, *PIM1*, and *NFKBIZ*[15]. Furthermore, ABC-DLBCLs tend to display recurrent amplifications of *BCL2* and *BCL6*[236]. As described in the previous chapter, mutations affecting the lysine methyltransferase *KMT2D*[293] and the master apoptotic regulator *TP53* are common in both subtypes and may represent prognostic markers of poor prognosis in patients treated with R-CHOP[227,314,315].

As the selective pressure of treatment provides a selective pressure that benefits cells with natural resistance, it follows that through genomic analysis of rrDLBCL we

should be capable of delineating genetic events are enriched at relapse, thereby implicating them in treatment resistance. Mutations associated with escape from immune surveillance have been reported at a high frequency in rrDLBCL, including recurrent deletions and SSMs perturbing *HLA-A*, *HLA-B*, *HLA-C*, *CD70*, *CD58*, and *B2M*, which collectively contribute to loss of functional MHC class I protein[252–254]. A significant enrichment of SSMs perturbing NF-kB signaling components *MYD88*, *PIM1*, and *CD79B* as well as enrichment of mutations perturbing *KMT2D* and *STAT6* have also been observed at relapse[263,316,251]. Furthermore, recurrent deletions of 6q22, gains of 13q21 (*MIR17HG*), 2p14 (*REL*) and chromosome 7 have been also described, albeit anecdotal trends not statistically significant when compared to diagnostic DLBCL[263,316]. Some of these events have therapeutic implications for patients receiving standard of care. For instance, gains of *BCL2* and loss of *TP53* are common in rrDLBCL and are associated with resistance against various chemotherapeutics, including components of R-CHOP[317]. Though many of these require further exploration and may not directly lead to alternative therapeutics (e.g. *TP53*), this information can have general utility for determining the risk of relapse on standard therapy. In some cases, alternative therapies may be warranted. For example, the reliance of a tumour on *BCL2* could be exploited with therapeutics that disrupt BCL2 activity via the BH3 domain. Our group previously compared the mutation frequency of 63 lymphoma-associated genes between 135 rrDLBCL cases to diagnostic DLBCL and found six genes significantly enriched for events at relapse, including *MS4A1*, whose mutations attenuated anti-CD20 antibody binding *in vitro*[255]. However, these previous studies have been generally limited by small sample sizes and a restricted view of genetic events, with only a handful of studies performing exome-wide comparisons incorporating both SSMs and copy number alterations (CNVs) and lacking statistical power, instead opting to compare genetic features between paired diagnostic-relapse cases[252,253]. Given the genetic heterogeneity of DLBCL and the high prevalence of *TP53* alterations at relapse, indicating genomic instability, a broad, exome-wide analysis of rrDLBCL including CNVs is needed to fully resolve the complex biology underlying treatment resistance in DLBCL, especially within the context of genetic subgroups.

To explore the landscape of SSMs and CNVs in rrDLBCL, we performed a combination of whole exome sequencing (WES), whole genome sequencing (WGS), and ultra-low pass WGS (lpWGS) on tissue and liquid biopsies from 247 rrDLBCL cases.

Through an exome-wide analysis, we report five genes significantly enriched for mutations in rrDLBCL, and two genes (*TMEM30A, TET2*) significantly depleted for mutations at relapse. Furthermore, we observed a high burden of CNVs in rrDLBCL, with many recurrent events encompassing well described lymphoma drivers. 13 of these recurrent events were enriched for events at relapse, including several novel relapse-specific events encompassing genes involved in RNA regulation, antigen presentation, and B-cell proliferation.

3.3. Methods

3.3.1. rrDLBCL sample collection

Tissue or liquid biopsies were collected from 199 patients from a combination of three clinical trials (LY.17 [NCT02436707], Obinituzumab-GDP [OZM073][NCT02750670], and Epizyme [NCT01897571]) or from the routine patient population from Quebec (Montreal) or British Columbia, Canada (LSARP) (Supplemental Table S1, Appendix B). All cases were treated with R-CHOP or R-CHOP equivalent, and biopsies were collected following failure of R-CHOP and, in some cases, additional salvage therapies. For patients with a tissue biopsy available, cell-of-origin (COO) was assigned using the DLBCL90 Nanostring[91] assay except for the Montreal cohort, which was assigned using the Hans algorithm[83]. Note that for many liquid biopsies, the COO was assigned at diagnosis. This study was reviewed and approved by the Research Ethics Boards of the University of British Columbia-BC Cancer agency and the Jewish General Hospital (18-030) in accordance with the Declaration of Helsinki. This cohort was additionally augmented with rrDLBCL samples from two previously published cohorts: 20 rrDLBCL samples from Schmitz et al[228]. and 28 rrDLBCL samples from the QC2 trial[251].

3.3.2. Sample processing, library preparation, and sequencing

Formalin-fixed paraffin-embedded (FFPE) tissue samples and matched constitutional DNA were extracted as previously described[255]. DNA libraries were prepared using QIAseq FX DNA Library Kit (Qiagen). Blood and plasma samples from LY.17, OZM073, Epizyme, and Montreal cohorts were extracted as previously described[173,255]. DNA libraries from liquid biopsies were prepared using either custom in-

house Unique Molecular Identifiers (UMIs) or using the xGen ctDNA kit (IDT technologies). For samples which have undergone whole exome sequencing, target enrichment and capture was performed using the XGen exome research panel V1 (IDT technologies). All sequencing was performed on an Illumina HiSeq2500 or HiSeqX using 150bp or 125bp paired-end chemistry. Samples were sequenced to a target depth 80x coverage (whole genome sequencing), 150x coverage (whole exome sequencing), or 0.1-0.3x coverage (ultra-low pass WGS, lpWGS).

3.3.3. Read alignment and somatic variant calling

Sequencing reads passing Illumina's chastity filter were aligned against the human reference genome GRCh37 (WGS data) or GRCh38 (WES and lpWGS data) using bwa mem[152]. For exome samples prepared using UMIs, family identification and error correction was performed using a custom in-house pipeline as previously described[219]. For samples prepared from tissue biopsies and lpWGS data, duplicate reads were identified and flagged using Picard MarkDuplicates[318]. Quality control was performed using Qualimap2[159], the Picard toolkit (CollectWGSMetrics and CollectHsMetrics), and samtools[154].

Simple somatic mutations and small insertions/deletions (SSMs) were identified using a consensus of four variant callers: Strelka2[165], MuTect2[164], SAGE, and LoFreq[169] (Figure 3-1). All callers were run in paired mode, using an unmatched normal from a different patient if a matched normal from the same patient was unavailable. All tools were run with the appropriate settings for each sequencing type, disabling depth filters and providing the exome capture space for samples which underwent WES. Candidate single nucleotide variants and small insertions and deletions (cumulatively simple somatic mutations, SSMs) from each tool were compared using Starfish[319], and SSMs called by at least three tools were considered real. LoFreq does not detect indels, thus requiring indels to have been identified by each of Strelka2, MuTect2, and SAGE. For both pipelines, variants were annotated using vcf2maf (https://github.com/mskcc/vcf2maf), using Variant Effect Predictor[276], and post-filtered to remove variants with a GnomAD[275] population allele frequency >0.01 in any population, and to remove recurrent variants which were observed in >20% of unpaired samples. Final somatic variant calls from WGS data were converted to hg38 genomic coordinates

using Crossmap[278] (Supplemental Table S2, Appendix B). Visualizations and quality control of final SSM calls were generated using the R packages maftools[320] and ggplot2.

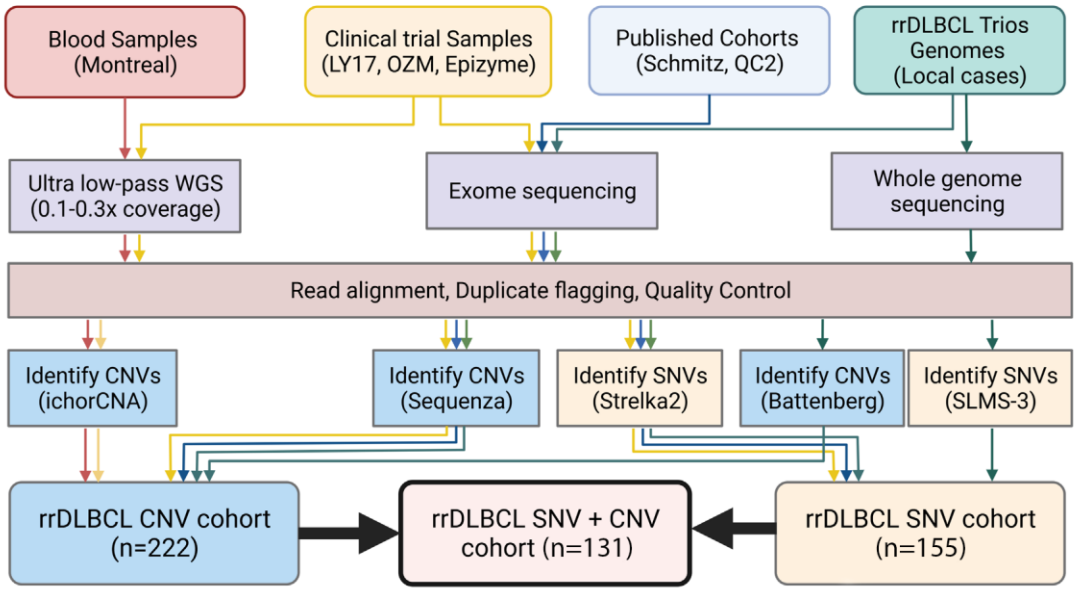


Figure 3-1. Cohorts and analysis workflows used for the rrDLBCL CNV cohort, SNV cohort, and merged cohorts

3.3.4. Copy number calling

Copy number variants (CNVs) were identified using three approaches. For lpWGS samples, CNVs were identified using ichorCNA[182] and HMMCopy's readCounter using 500kb bins and a custom panel of normals generated from 69 samples with no detectable ctDNA. All CNV profiles were manually inspected and ichorCNA was re-run with fixed purity/ploidy priors for samples where the default ichorCNA solution was deemed inaccurate. For paired WES samples, CNVs were identified using Sequenza[179], pre-filtering bins to remove candidate SNP positions not observed in GnomAD[275]. For paired WGS samples, CNVs were identified using Battenberg[184], and CNVs were covered to hg38 coordinates using liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver) (Supplemental Table S3, Appendix B). Regions recurrent perturbed were identified using GISTIC2[321], and were manually inspected using Integrative Genome Viewer[322] to filter peaks stemming from systematic mapping artifacts, and merge adjacent peaks (Supplemental Table S4, Appendix B). The genetic subgroup of each sample was

predicted via the LymphGen algorithm[249], using a custom python script (https://github.com/ckrushton/LGenIC) to format CNV, SSMs, and SV calls (where available) to generate the prerequisite input files. Visualizations were generated using the R packages GenVisR[323], GenomeTornadoPlot[324], and ggplot2.

3.3.5. Diagnostic cohort

Aligned reads, copy number calls, and associated metadata were downloaded from Schmitz *et al*[228] and grouped with a previously published WGS cohort by our group[15]. SSMs were identified using the pipeline listed above, using the same unmatched normal from the rrDLBCL cohort for the Schmitz cohort (as no normal was available for any Schmitz cases). CNVs from the WGS data were identified using Battenberg[184]. Given the difficulty in calling CNVs from unmatched sequencing data which has undergone hybridization capture, we elected to use the array-based CNV calls provided by Schmitz *et al*. We further subset the Schmitz cases to 1) remove 20 rrDLBCL samples treated with Ibrutinib, 2) select for samples with both SSM and CNV data, and 3) Balance the proportion of ABC, GCB, and unclassified cases from our combined diagnostic cohort with the proportion observed in rrDLBCL (final n=467). Survival analysis was performed R package Survminer (Version 0.4.9, R version 4.1.3), excluding any cases with incomplete information for any of the variables considered.

3.3.6. Mutation frequency comparison

To compare the frequency of SSMs between the diagnostic and relapse cohort, we combined the SSM calls between our diagnostic and relapsed cohort and identified 58 genes with evidence of positive selection via OncoDriveFML[325] (Q<0.01) (Supplemental Table S5, Appendix B). We supplemented this with three additional genes, *FOXO1, MS4A1*, and *MYC*, each having previously been observed enriched for mutations in rrDLBCL or identified in preliminary analyses. The frequency of mutations in these genes was compared between diagnostic and rrDLBCL cohorts using a Fisher's exact test and Benjamini/Hochberg false discovery correction, with Q<0.1 classified as significantly differentially mutated (Supplemental Table S6, Appendix B). We excluded genes observed exclusively mutated in one cohort but not the other as these were observed to represent recurrent germline events upon manual inspection. To compare

CNVs prevalence between the two cohorts, we first normalized all CNV profiles to a diploid state, and the copy number state at the center of each GISTIC peak was assigned for each sample using a custom python script. The prevalence of gains (CN state > 2) and deletions (CN state <2) was then compared for each amplification and deletion peak respectively using a fisher's exact test and Benjamini/Hochberg false discovery correction in a custom R script, classifying events with Q<0.1 as significantly enriched/depleted for gains/amplifications.

3.3.7. Mutual exclusivity analysis and clustering

All rrDLBCL samples with CNV information (n=222) were annotated with the copy number state of each recurrently gained/deleted region identified by GISTIC analysis of these data[321] following ploidy normalization using a custom python script. The co-occurrence/mutual exclusivity of each recurrent CNV was assigned using maftools somaticInteractions[320]. For clustering, the copy number status of CNVs was converted into a binary matrix, and two sets of GISTIC peaks (gains of 1q21.3/1q25.2 and deletions of 6q23.3/6q16.3) were merged into a single "meta-peak" due to their close genomic proximity and high concordance of events, with events affecting either sub-peak resulting in the meta-peak being assigned "perturbed". Clustering was performed using the R package NMF (Version 0.2.4, R version 4.1.3), obtaining four clusters using after 1000 iterations using the "Brunet" algorithm[326].

3.4. Results

3.4.1. The exome-wide mutation landscape of rrDLBCL

To explore the landscape of SSMs in our rrDLBCL cohort, we performed a combination of WES and WGS on 155 rrDLBCL samples and identified coding somatic variants exome-wide. Overall, the landscape of somatic mutations in rrDLBCL exome-wide is broadly similar to diagnostic DLBCL (Figure 3-2A), with a high burden of mutations perturbing *KMT2D* (mutated in 41% of rrDLBCL cases), *TP53* (32%), *CREBBP* (25%), *MYD88* (19%), and *PIM1* (15%). Many of these mutation patterns are associated with one of the two COO groups, Specifically, mutations in *MYD88* and *CD79B* significantly enriched in ABC-rrDLBCL and mutations perturbing *SOCS1*, *CREBBP*, and *EZH2* significantly more abundant in GCB-rrDLBCL (Figure 3-3).

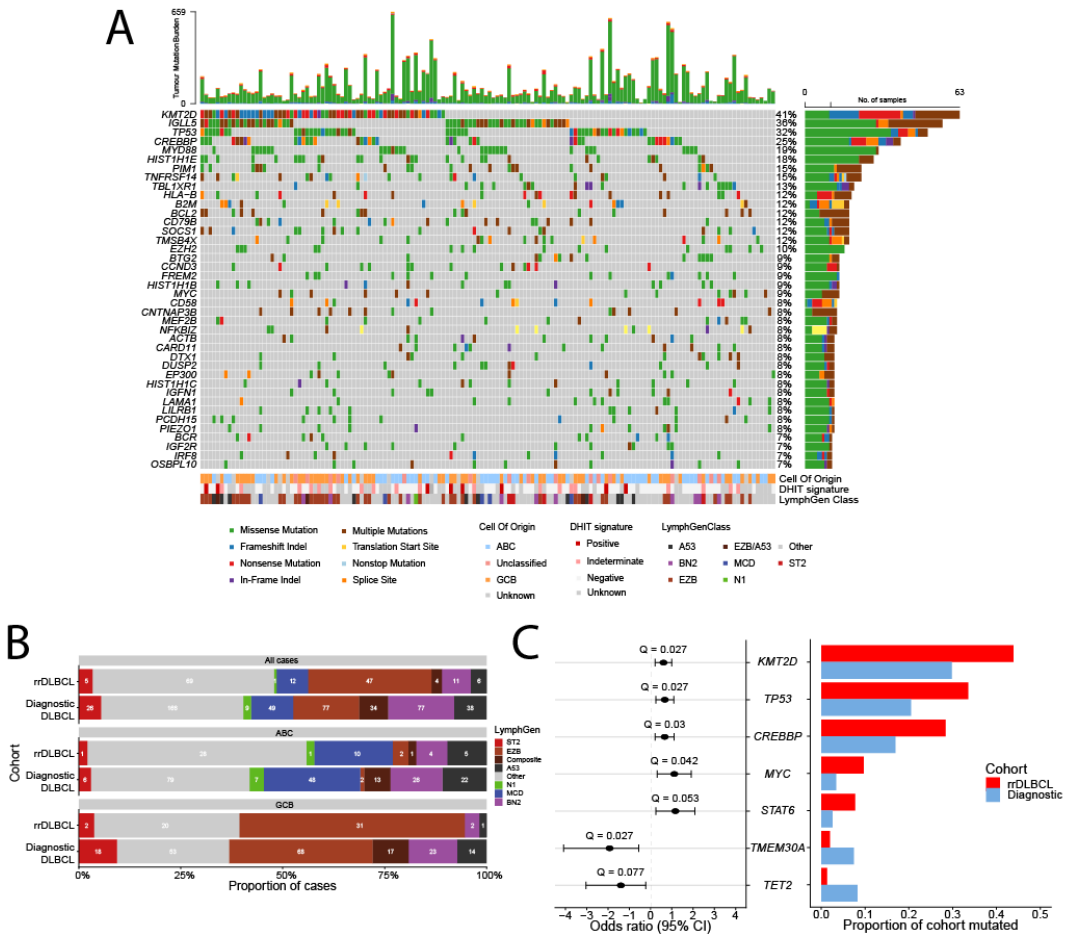


Figure 3-2. Landscape of simple somatic mutations in rrDLBCL, exome-wide, and events significantly differentially perturbed in rrDLBCL. (A). Mutation heatmap displaying the top 40 recurrently mutated genes in rrDLBCL exome-wide, across all 155 rrDLBCL samples with exome data (x-axis). Box colours correspond to the mutation type(s) observed in that sample. Additional covariate tracks are also provided. (B). Distribution of genetic subgroups in the rrDLBCL cohort compared to the diagnostic cohort, broken down by molecular subgroups. (C) Forest plot (left) and bar plot (right) summarizing the odds ratio and mutation frequency, respectively, of genes significantly ($p_{adj} < 0.1$) enriched or depleted for mutations in rrDLBCL (red) compared to diagnostic DLBCL (blue).

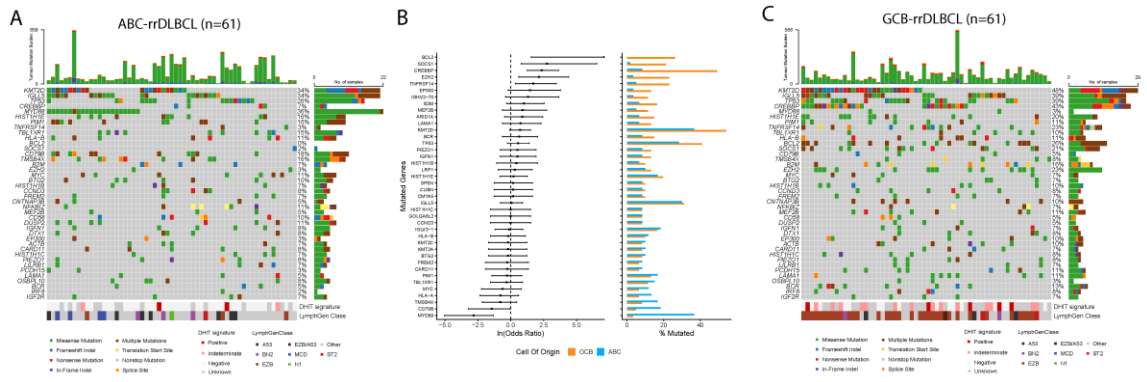


Figure 3-3. The mutational landscape of rrDLBCL, broken down by molecular subgroups. The mutation frequency and prevalence of ABC-rrDLBCL(A), and GCB-rrDLBCL (C), along with a forest plot (B) showing the overall frequency of mutations between the two genetic subgroups.

We next sought to explore the distribution of genetic subgroups within our rrDLBCL cohort using the LymphGen classifier (Figure 3-2B). EZB was the most prevalent genetic subgroup in our cohort (30.3% of cases and 55.4% of GCB-rrDLBCLs), and the prevalence of EZB cases was significantly higher (p_{adj} = 0.00163) in our rrDLBCL cohort compared to diagnostic DLBCL (n=467). This enrichment likely reflects the presence of rrDLBCL cases transformed from follicular lymphoma. In contrast, BN2 cases were significantly under-represented among rrDLBCL, (p_{adj} = 0.00472), consistent with the superior prognosis of BN2 cases treated with R-CHOP[249]. Curiously, despite the poor prognosis of MCD cases following R-CHOP, we did not observe an enrichment of MCD rrDLBCL cases in our cohort, with the majority (53.8%) of ABC-rrDLBCLs genetically unclassified via LymphGen.

3.4.2. Mutations with prognostic potential in rrDLBCL

We next focused on identifying the individual genes enriched or depleted for mutations in rrDLBCL as they might represent candidate biomarkers of poor or good treatment outcomes, respectively. We assembled a candidate gene list comprised of recurrently mutated genes that displayed mutation patterns consistent with positive selection (Supplemental Table S5, Appendix B) and genes previously associated with rrDLBCL. The prevalence of mutations in these genes were compared to a representative diagnostic cohort comprised of previously published exome and genome cases. Seven genes were significantly differentially mutated between diagnostic DLBCL and rrDLBCL with 5 having more mutations in the latter (Figure 3-2C). Of the genes

enriched for mutations in rrDLBCL, four (*TP53, KMT2D, CREBBP, STAT6*) were previously described in studies from our lab and elsewhere[255,316]. Three of these genes remain significantly enriched in rrDLBCL when a GCB-specific comparison was performed (*CREBBP*, *STAT6* and *TP53*) (Figure 3-4). In contrast, no genes found to be enriched for mutations among ABC rrDLBCLs. Mutations in *MYC*, though enriched in rrDLBCL, are almost exclusively missense mutations and most overlap AID recognition motifs (18/32, 56%, Supplemental Table S2, Appendix B).

Only two genes were significantly depleted for mutations in rrDLBCLs, namely *TET2* and *TMEM30A*. Mutations in either of these were exceptionally rare in rrDLBCL (1.3% and 2.0% of cases, respectively). While mutations in *MS4A1* showed evidence of positive selection in rrDLBCL (Supplemental Table S5, Appendix B), this comparison did not show them to be significantly enriched in rrDLBCL. This can be attributed, in part, to the low prevalence of *MS4A1* mutations in both cohorts (mutated in 4.5% rrDLBCL and 0% of diagnostic DLBCL), which may limit our statistical power to detect a difference.

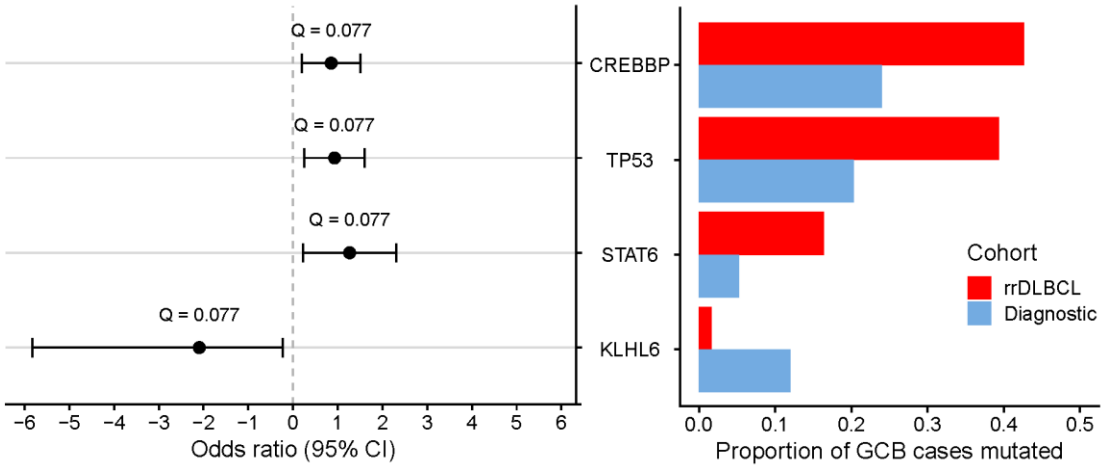


Figure 3-4. Significantly differentially mutated genes between diagnostic DLBCL (blue bars) and rrDLBCL (red bars) within the GCB molecular subgroup. The left panel shows a forest plot with the odds ratio of each significant gene, as determined using a Fisher's exact test. The right panel is a bar plot showing the proportion of each cohort harbouring a mutation in the respective gene.

3.4.3. Recurrent CNVs inform on the biology of rrDLBCL

Given the low number of genes we found differentially mutated in rrDLBCL, we next sought to explore the landscape of recurrent copy number events to determine the

interplay between CNV and SSM in rrDLBCL biology. We augmented our cases having exome or WGS data with lpWGS from additional samples, yielding a total of 222 rrDLBCLs with CNV calls. Overall, rrDLBCL tumours are burdened by high levels of CNVs (Figure 3-5), with a high frequency of arm-level or whole-chromosome gains involving chromosome 7 (47.7% of cases), gains of 18q23 (encompassing *BCL2*, 44.1%), as well as deletions of 6q16.3 (43.2%) and 6q23.3 (*TNFAIP3*, 45.0%), and 17p13.1 (37.4%). We observed recurrent events perturbing well-described lymphoma drivers, including gains of *REL* (2p16.1), *BCL6*, (3q29), *MIR17HG* (13q31.3) and *BCL2* (18q22.1), and deletions of *CDKN2A/CDKN2B* (9p21.3), *PTEN* (10q23.31), *RB1* (13q14.2), and *B2M* (15q15.1).

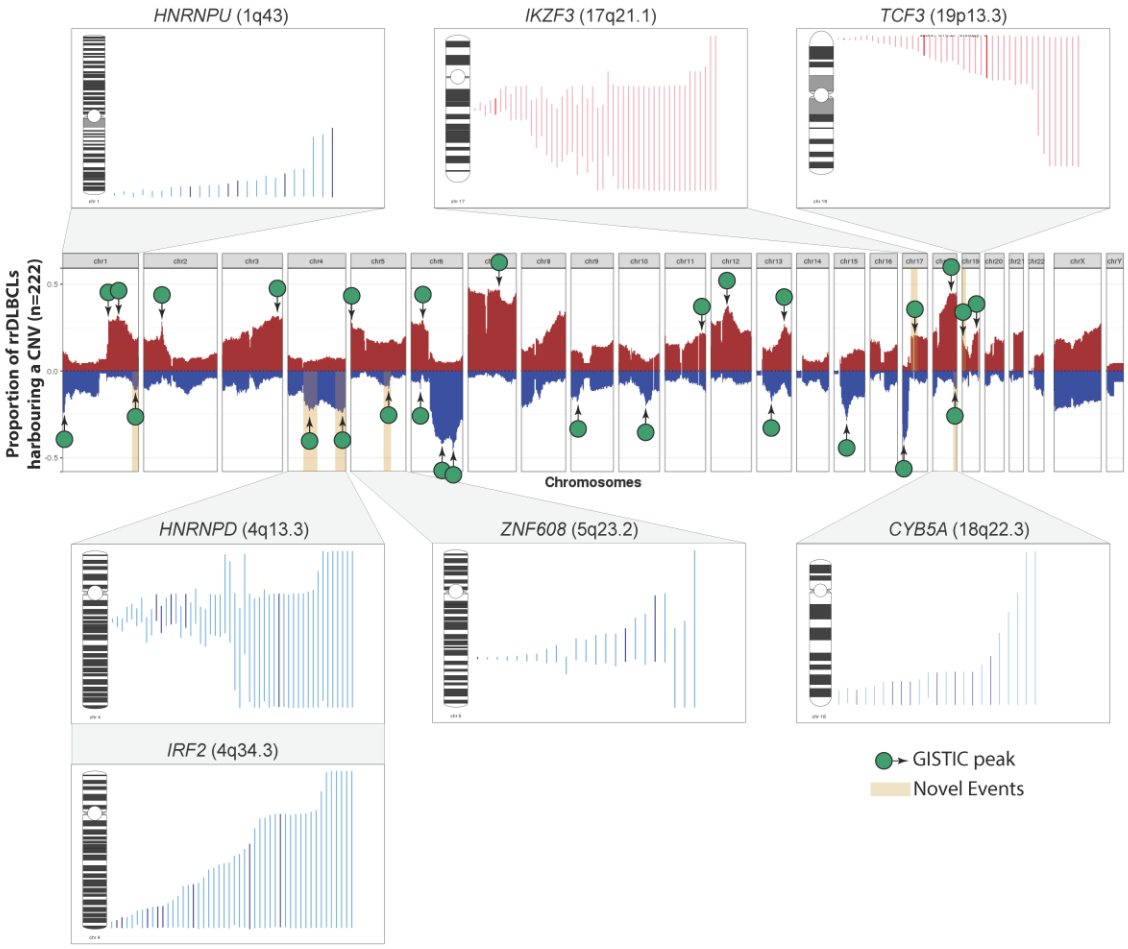


Figure 3-5. The landscape of copy number variants across rrDLBCL. The proportion of samples harbouring a copy number gain (red) or deletion (blue) at each genomic locus are indicated in the copy number frequency plot, with significantly recurrently perturbed regions as identified by GISTIC labelled. Recurrent CNVs which have not previously been described in DLBCL are labelled, with tornado plots representing the suspected target and copy number segments overlapping that region.

Among the recurrent CNVs we identified several novel events within our cohort that have not previously been described in DLBCL (Figure 3-5). We observed recurrent deletions of 1q43 and 4q13.3 centered on the RNA splicing regulators *HNRNPU* and *HNRNPD*, respectively, with the deletion of 1q43 observed almost exclusively in ABC cases (Figure 3-6). The deletion of 4q34.3, centered on the MHC Class I regulator *IRF2,* is another prevalent event that has not been described. Deletions of 18q22.3, while rare in rrDLBCL, have been previously reported in pancreatic cancer, in a study that nominated *CYB5A*[327] *as* the target of this event. In that study, loss of this locus resulted in decreased autophagy in malignant cells. In our cohort, this event was almost

exclusively observed in GCB-rrDLBCL cases (Figure 3-6). Two newly identified regions subject to recurrent gains in rrDLBCL were 17q21.1 and 19p13.3, with the gain of 17q21.1 commonly perturbed in GCB-rrDLBCL cases and encompassing the B-cell maturation regulator *IKZF3*/Aiolos. 19p13.3 includes the B-cell differentiation regulator *TCF3* and copy number gains involving this locus were significantly more prevalent in ABC-rrDLBCL cases.

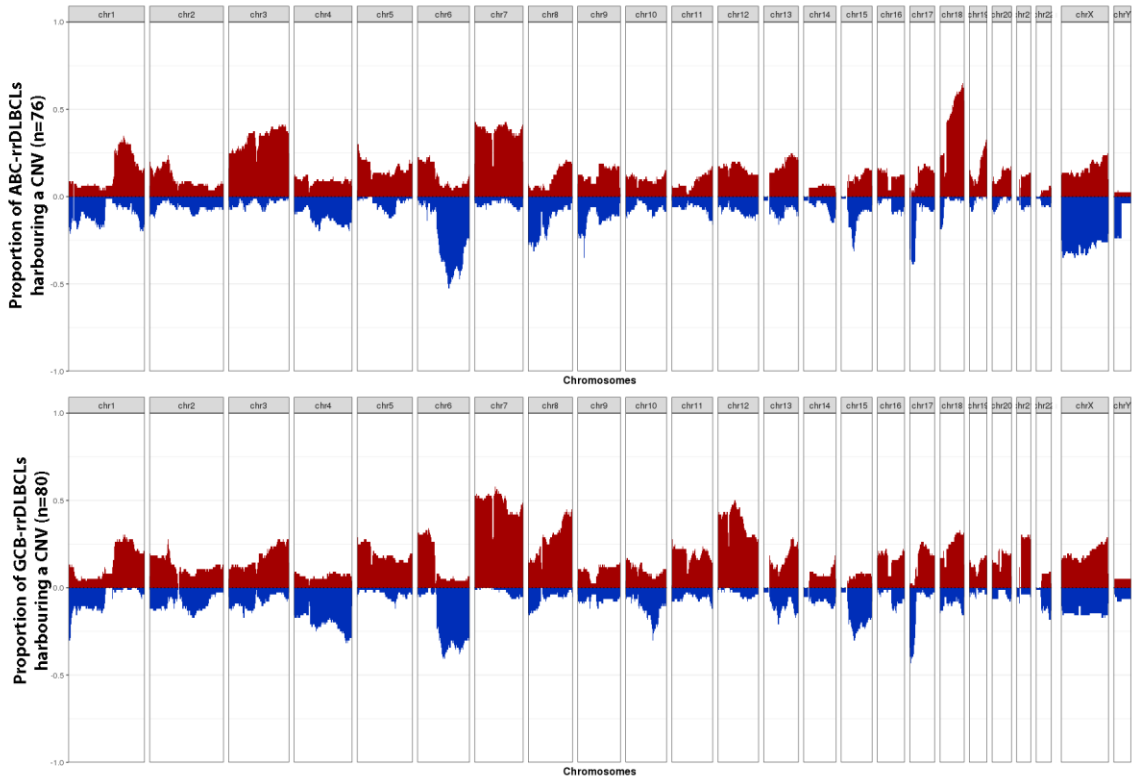


Figure 3-6. Landscape of copy number variants across rrDLBCL, subset to ABC (top) and GCB (bottom) cases. The proportion of cases harbouring copy number gains (red) or deletions (blue) are indicated at each genomic locus

3.4.4. Recurrent CNV drivers and novel events are enriched in rrDLBCL

As rrDLBCL cases harbour a high burden and repertoire of copy number events, we next sought to compare the frequency of these recurrent events to our diagnostic cohorts. Of the 29 regions recurrently perturbed in rrDLBCL, 13 were found to be significantly enriched at relapse, with one event (gains of 5p13.33) significantly depleted (Figure 3-7). This includes four of the novel recurrent events in rrDLBCL

(4q13.3/*HNRNPD* deletions, 4q34.3/*IRF2* deletions, 17q21.1/*IKZF3* gains, and 19p13.3/*TCF3* gains), with 19p13.3 gains displaying the most significant enrichment. Gains of 19p13.3 and 17q21.1 were also enriched in a subtype-specific comparison of ABC cases (Figure 3-8). Gains involving the *STAT6* locus (12q13.3) were enriched at relapse, consistent with the enrichment of SSMs. Of the remaining loci, *PTEN* and *TP53* are both deleted at a high frequency in rrDLBCL, and gains of *MIR17HG* and *BCL2* are associated with increased NF-kB signaling and cell survival. Deletions of *TP53* (17p13), while prevalent in both ABC and GCB-rrDLBCL, are notable enriched in GCB cases (Figure 3-8). We further attempted to compare the prevalence of SSMs and CNVs using rrDLBCL cases with SSM and CNV information (n=131). While 10 genes were significantly differentially perturbed, representing a combination of differentially perturbed genes in the SSM and CNV only comparison, all events failed FDR correction (Q>0.1).

Given the high frequency and enrichment of recurrent CNVs in rrDLBCL, we next sought to establish patterns and identify groups of CNVs which might indicate shared biological modules. After comparing all recurrent CNVs across our rrDLBCL cohort (Figure 3-9), gains of 18q22.1/*BCL2* and deletions of 18q22.3/*CYB5A* were significantly mutually exclusive, consistent with the ABC/GCB pattern observed for these events. Deletions of 17p13.2/*TP53* significantly co-occurred with deletions of well-established tumour suppressors 1p36.32/*TNFAIP3*, 10q23.31/*PTEN*, and 15q15.1/*B2M*, as well as several novel rrDLBCL events (deletions of 4q13.3, 4q34.3, and gains of 17q21.1), suggesting that loss of *TP53* enable a cellular phenotype which enable the acquisition of these events. Gains of 3q29/*BCL6* also significantly co-occurred with gains of 13q31.3/*MIR17HG* and 18q22.1/*BCL2*. Finally, gains of the 7q22.1 locus, overlapping *CDK6* and *CDK14*, also co-occurred with gains of the 12q13.3 locus which, in addition to *STAT6*, overlaps the cell cycle regulators *CDK2* and *CDK4*, suggesting a subset of cases with significantly enhanced cell-cycle progression.

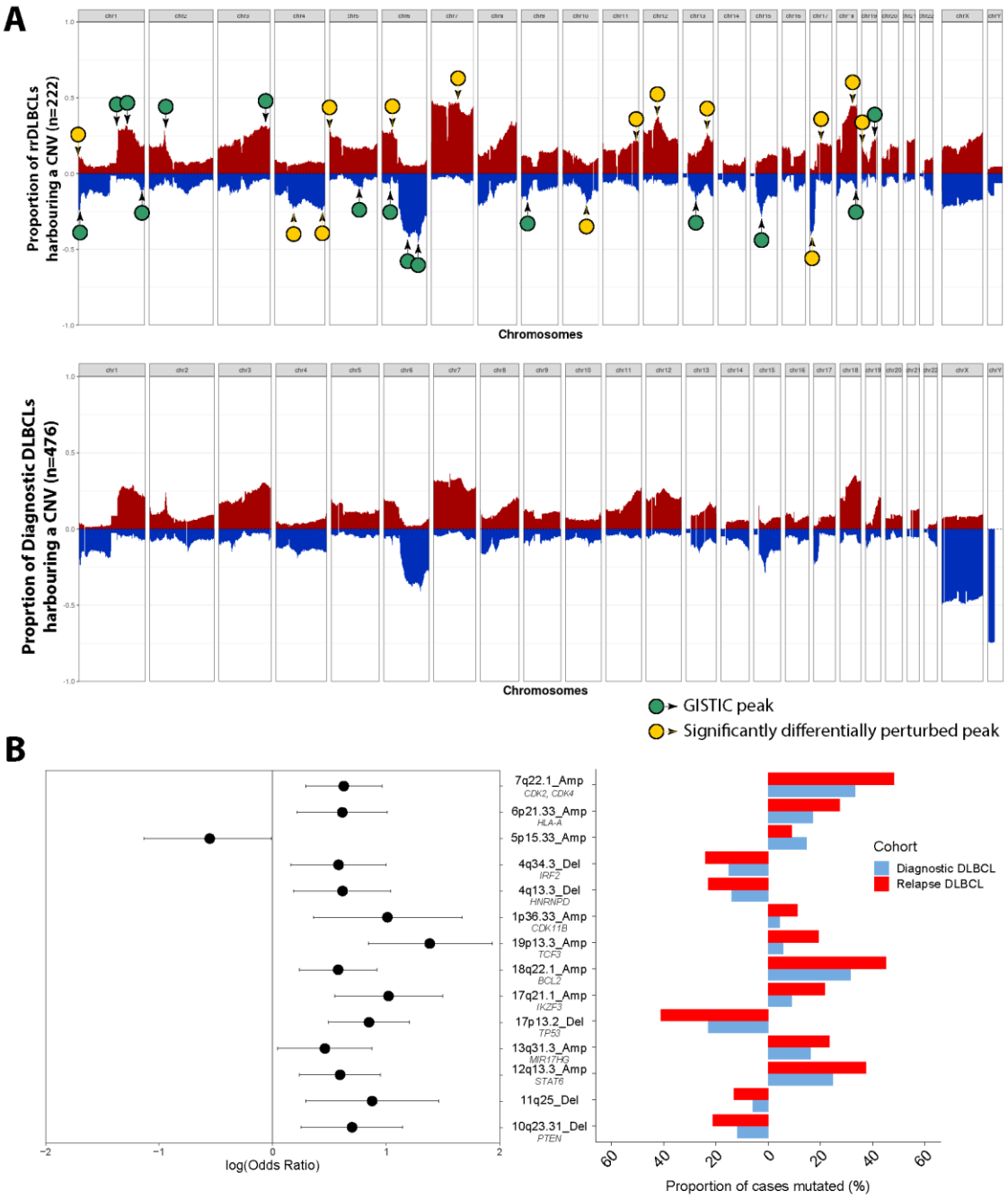


Figure 3-7. Significantly differentially perturbed events between diagnostic and rrDLBCL. (A) Landscape of copy number variants between rrDLBCL (top) and diagnostic DLBCL (bottom), with GISTIC peaks and significantly differentially perturbed GISTIC peaks indicated by the green and yellow bubbles, respectively. (B) Forest plot and bar plot summarizing recurrent CNVs which are significantly ($p_{adj} < 0.1$) differentially perturbed between diagnostic (blue) and rrDLBCL (red) and their associated frequency in each cohort

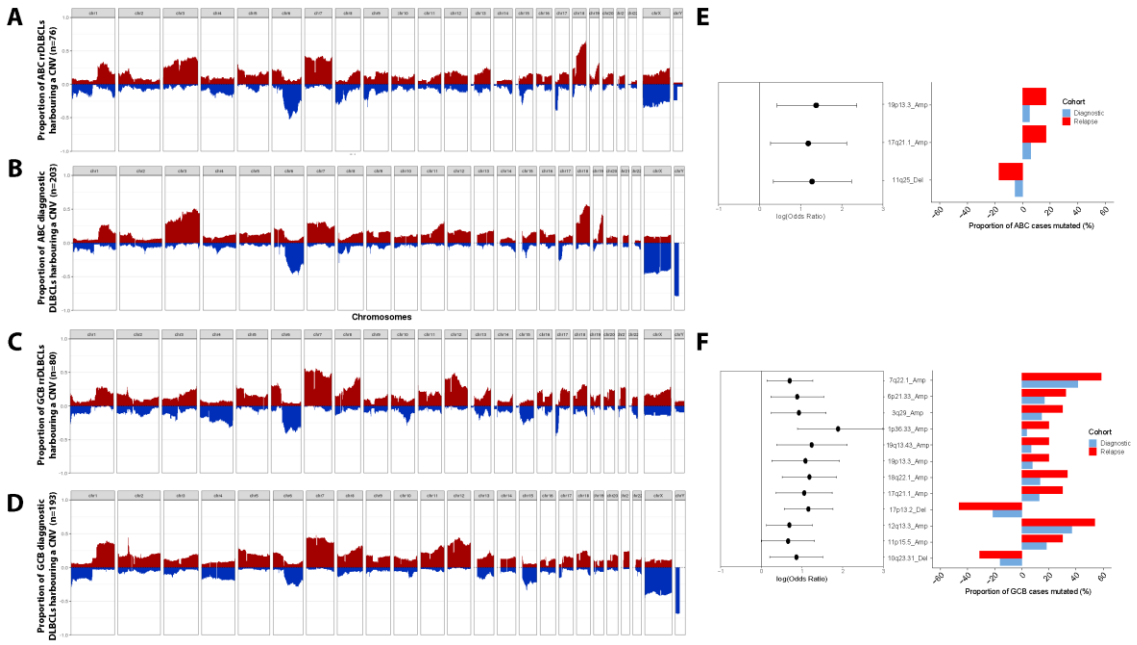


Figure 3-8. Comparison of copy number events between diagnostic DLBCL (B, D) and rrDLBCL (A,C), specifically within the ABC (A,B) and GCB (C,D) molecular subgroups. The frequency of copy number gains (red) and deletions (blue) at each locus are indicated genome-wide. Forest plots and copy number frequency bar plot for each significantly differentially mutated region within ABC (E) and GCB (F) specific comparisons.

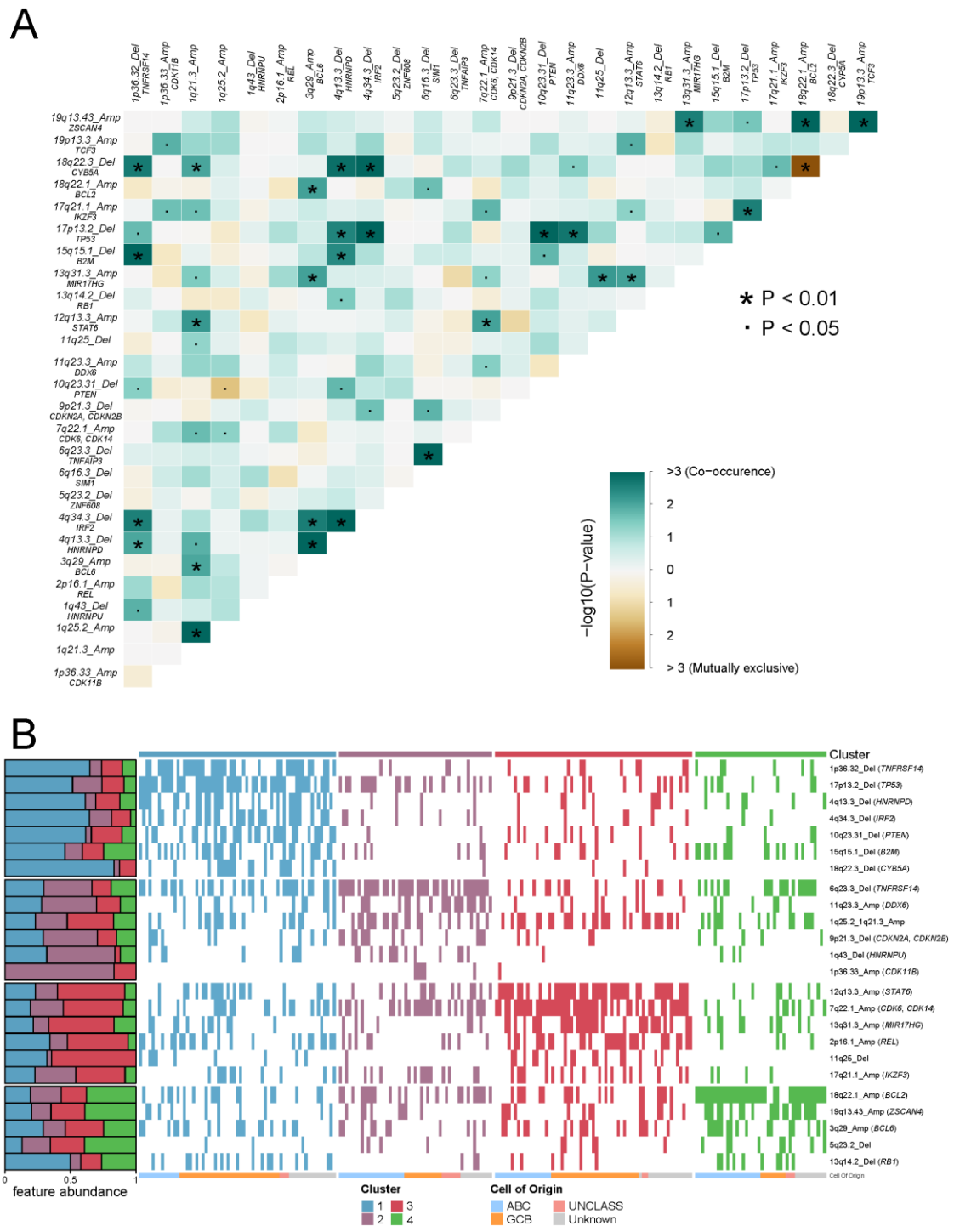


Figure 3-9. Patterns of recurrent CNVs across rrDLBCL samples. (A) Somatic Interactions plot showing recurrent CNVs which co-occur (green) or are mutually exclusive (brown) across rrDLBCL samples. Significant interactions are also shown (. and *). (B) NMF clustering results of recurrent CNVs across rrDLBCL samples. Each cluster is represented by a unique colour, and the proportion of events represented by samples in each cluster is reflected in the stacked bar chart on the leftmost side.

3.5. Discussion

Despite numerous candidate salvage therapies available for rrDLBCL, patient outcomes generally remain poor, stemming from the genetic and molecular heterogeneity of DLBCL and rrDLBCL. Through an exome-wide examination of SSMs and CNVs in rrDLBCL, we uncovered a high burden of CNVs and recurrent, enriched events perturbing well described lymphoma drivers. These enriched events include several novel events encompassing genes involved in antigen presentation (*IRF2*), B-cell proliferation and growth (*IKZF3*, *TCF3*), and RNA splicing (*HNRNPU*, *HNRNPD*).

Initially, we explored the landscape of SSMs exome-wide in our 155 rrDLBCL cases with WES or WGS data. Many genes recurrently mutated at high prevalence (*KMT2D, TP53, CREBBP,* and *MYD88*), and those enriched for events at relapse (*KMT2D, TP53, CREBBP, STAT6*) have been associated with rrDLBCL by our group and others[251,253,255,263]. We observed two genes*, TMEM30A* and *TET2*, significantly depleted for mutations at relapse. Homozygous loss of function of *TMEM30A* have been described in DLBCL and are associated with improved prognosis following R-CHOP due to elevated macrophage infiltration[328], consistent with the depletion of mutations observed in rrDLBCL. *TET2* is a hydroxymethyltransferase which initiates demethylation of 5' methylcytosine, a common mark of epigenetic silencing, via conversion to 5'hydroxymethylcytosine, leading to transcriptional activation of the target genes. Loss of *TET2* has been shown to impair B-Cells from existing the germinal center, thus inhibiting plasma cell differentiation, and is an early driver in lymphomagenesis[329–332]. Curiously, not only were SSMs perturbing *TET2* significantly depleted in rrDLBCL, but the *TET2* locus (4q24) appeared to be retained (copy-neutral) despite the high prevalence of deletions at adjacent loci 4q13.3(*HNRNPD*) and 4q34.3(*IRF2*). Many genes perturbed by *TET2* loss of function are also perturbed by disruption of *CREBBP* induced H3K27 acetylation[329], which is enriched for mutations in rrDLBCL. In conjunction with the high frequency of loss-of-function mutations in *KMT2D,* this suggests that epigenetic dysregulation may contribute significantly to the overall biology of rrDLBCL. As DNA methyltransferase inhibitors which promote hypomethylation are currently being evaluation for rrDLBCL[333], functional *TET2* may reduce the effectiveness of this therapeutic approach.

In contrast to the limited number of genes differentially mutated by SSMs in rrDLBCL, the copy number landscape of rrDLBCL was distinct, with an overall higher burden of CNVs and numerous recurrent events, many of which were enriched upon relapse. This included the 12q13.3 locus, overlapping *STAT6*, consistent with the enrichment of SSMs observed in rrDLBCL. The minimal common region also encompasses the B-cell maturation transcription factor *IKZF4* and the cell cycle regulators *CDK2* and *CDK4,* which are responsible for G1-S transition[334]. Gains of the 7q22.1 locus are also enriched in rrDLBCL and encompass the cell cycle regulators *CDK6* and *CDK14*, with this recurrent gain significantly co-occurring with gains of 12q13.3. In conjunction with recurrent deletions of 9p21.3*/CDKN2A* and 13q14.2/*RB1* further suggests that constitutive proliferation contributes to rrDLBCL biology, with a high frequency of 18q22.1/*BCL2* gains in ABC-rrDLBCL and deletions of 17p13.2/*TP53* reducing apoptosis.

Within our rrDLBCL cohort, we identified several recurrent CNVs not previously described in lymphomas. Recurrent deletions of 4q34.3/*IRF2* were observed at high prevalence in rrDLBCL, and loss of *IRF2* has been shown to impair MHC-Class I antigen presentation and upregulate expression of PD-L1, contributing to immune evasion[335]. Previous rrDLBCL studies have observed recurrent deletions of *HLA-A* and *B2M*, which have been associated with immune evasion[252,253]. Given the sequence similarities between *HLA-A*, *HLA-B*, and *HLA-C* and the difficulty in mapping reads to a single HLA sequence using Illumina sequencing, we were unable to evaluate the copy number state *of HLA-A* in our study; however, *B2M* was recurrently deleted in our rrDLBCL cohort.

We additionally identified two novel recurrent amplification peaks in rrDLBCL, centered on 17q21.1/*IKZF3* and 19p13.3/*TCF3*, which were observed and enriched in both ABC and GCB-rrDLBCL. *IKZF3* encodes the lymphoid-specific transcription factor Aiolos and is crucial for B-cell development, with inhibition of Aiolos in DLBCL cell lines leading to reduced proliferative signaling and promotion of T-cell activation[336]. As 25% of rrDLBCLs harbour an amplification of *IKZF3*, this represents a candidate therapeutic target, with two existing small molecule inhibitors, avadomide and lenalidomide, leading to Aiolos degradation[336,337] and promising efficacy for a subset of rrDLBCL cases[303,338–340]. We also did not observe any cases harbouring *IKZF3*-G152A mutations, conveying avadomide resistance in our rrDLBCL cohort. The transcription factor *TCF3* is also essential for B-cell development and lineage commitment[341]. Fusions involving *TCF3*

have been observed in pediatric lymphomas[342,343], and recurrent mutations previously have been described in Burkitt lymphoma[344–346]. Notably, upregulation of *TCF3* has been observed in prostate cancer and decrease sensitivity to doxorubicin, and could contribute to resistance against R-CHOP [347].

One of the most notable observations in our rrDLBCL cohort were novel deletions of the RNA splicing regulators 1q43/*HNRNPU* and 4q13.3/*HNRNPD*. *HNRNPD* and *HNRNPU* recognize and bind AU-rich elements of mRNA, thus regulating expression of target genes. Curiously *HNRNPU* and *HNRNPD* both target and stabilize MYC[348–350], and decreased expression of these genes are associated with decreased MYC protein abundance[349]. *HNRNPU* expression in bladder cancer has been associated with resistance against platinum based chemotherapies, with knockdowns of *HNRNPU* increasing cisplatin sensitivity[351]. While rare, these recurrent deletions of 1q43 may identify a subset of cases with promising outcomes following platinum-based chemotherapy. Given the high recurrence of CNVs and SSMs perturbing epigenetic (*KMT2D, CREBBP, TET2*) and transcriptomic regulators (*HNRPD, HNRNPU*), further studies exploring the transcriptomic and epigenetic landscape of rrDLBCL are needed to fully understand the biology of rrDLBCL and identify additional mechanisms of treatment resistance.

Chapter 4. General Discussion

4.1. Summary of research findings

Despite the utility of R-CHOP as a standard frontline therapy for DLBCL, outcomes for rrDLBCL patients remain generally poor. While several promising salvage therapies for rrDLBCL have recently been explored (notably CAR-T cell therapy and bi-specific antibodies), along with numerous molecularly targeted agents and inhibitors, outcomes for rrDLBCL are heterogeneous, and a notable proportion of cases will fail to respond to salvage therrapies. As genetic events which contribute to therapeutic resistance will undergo positive selection in the context of therapy, these events are expected to be enriched in relapsed tumours. Thus, we collected and sequenced liquid and tissue biopsies from rrDLBCL cases to identify such genetic events and discovered novel relapse-specific mutations, candidate biomarkers of R-CHOP failure, and mutations directly implicated in R-CHOP resistance.

4.1.1. Mutations in *KMT2D* and *TP53* dominate the landscape of rrDLBCL

At the start of this project, tissue biopsies were generally not collected from DLBCL patients upon relapse, and thus genetic characterization of rrDLBCL tended to be limited to small-scale studies where such tissue biopsies were available. To avoid these limitations, we collected liquid biopsies from rrDLBCL patients enrolled in three clinical trials: LY17, a multi-arm clinical trial testing several salvage therapies for rrDLBCL, OZM073, examining obinutuzumab-GDP as a salvage therapy in place of the standard R-GDP, and Epizyme, exploring the *EZH2* inhibitor tazemetostat. Additional liquid biopsies were collected from the routine patient population in Montreal, Quebec. As liquid biopsies generally harbour low levels of ctDNA, we performed CAPP-Seq of 63 genes both recurrently mutated in lymphomas as well as candidate drivers of treatment resistance.

Through this comparatively large (135 cases) rrDLBCL study, we observed a high prevalence of mutations perturbing the master tumour suppressor gene *TP53* and the lysine methyltransferase *KMT2D*, which were mutated in 50% of rrDLBCL cases (Figure 2-2) and enriched at relapse. The high prevalence of *TP53* mutations

corresponds with the poor prognostic potential of such mutations both in DLBCL and other cancers [57,262,281]. *KMT2D* mutations were also extremely prevalent in rrDLBCL, and although mutations in *KMT2D* alone had not been implicated as a prognostic marker for DLBCL, *KMT2D* mutations have been associated with genetic subgroups which display inferior outcomes following R-CHOP[227]. *KMT2D* has been described as a haploinsufficient tumour suppressor gene, with loss of *KMT2D* enhancing B-cell proliferation and dysregulating pathways involved in cell cycle regulation and apoptosis[292]. The high prevalence of *KMT2D* mutations in this initial rrDLBCL cohort may also reflect the inclusion of cases transformed from other lymphoid malignancies, as *KMT2D* mutations are common in FL[292]. In our diagnostic cohort, mutations in *KMT2D* and *TP53* were associated with inferior PFS and OS in the context of R-CHOP, further supporting the enrichment of these mutations at relapse. As mutations perturbing these genes tend to be clonal (Figure 2-5) and persist following salvage therapies, they represent candidate predictive biomarkers of both R-CHOP failure and failure of salvage therapies. Unfortunately, as we lacked the clinical outcomes of patients enrolled on these trials at the time of this study, the prognostic potential of these mutations in the context of these salvage therapies could not be evaluated.

4.1.2. Mutations in *MS4A1* directly contribute to treatment resistance

In this initial rrDLBCL cohort, we also observed an enrichment of mutations perturbing *MS4A1*, which encodes the B-cell surface marker CD20 and the target of both rituximab and other anti-CD20 monoclonal antibodies. While frameshift mutations were observed in *MS4A1* (localized to the large loop [Figure 2-9]), we primarily observed missense mutations affecting the small loop of CD20 and neighbouring transmembrane domains, generating hydrophilic residues in these domains. These mutations prevented affected cells from being recognized and bound by all anti-CD20 mAbs tested, and thus represent a direct, acquired mechanism of treatment resistance. This is clinically relevant both in the frontline setting and at relapse, where an initially CD20+ tumour can lose CD20. As CD20 status is not routinely re-evaluated during salvage therapies, these mutations could indicate that “CD20+” tumours may indeed be intrinsically resistant to many salvage therapy regimens.

4.1.3. Recurrent copy number alterations contribute to a unique landscape in rrDLBCL tumours

While mutations in *MS4A1* represent a novel, acquired, and clinically relevant mechanism of resistance against anti-CD20 mAbs, *MS4A1* mutations are generally rare in rrDLBCL (7% of cases), and other genomic features which convey resistance remain undiscovered. Our initial study harboured two major limitations: First, we were restricted to a small subset of genes which had largely been implicated in DLBCL previously and thus our ability to discover novel mechanisms of resistance was limited. Second, although large scale copy-number alterations contribute significantly to the genomic landscape of DLBCL, we were unable to evaluate their contribution to relapse biology given the limitations of CAPP-Seq. To address these limitations, we performed a follow-up study (Chapter 3) exploring the landscape of SSMs and CNVs in rrDLBCL exome-wide using a combination of WES and WGS data from both tissue and liquid biopsies (n=155).

The landscape of SSMs in rrDLBCL found in our exome cohort (Figure 3-3) was comparable to that found using out CAPP-Seq cohort (Figure 2-2). However, the prevalence of mutations perturbing *KMT2D* (41% vs 51%) and *TP53* (32% vs 50%) varied substantially between the two cohorts. As the CAPP-Seq cohort was primarily comprised of clinical trial samples, which tended to represent more advanced cases (failed frontline and multiple salvage therapies), these samples may have represented more aggressive, advanced, and resistant cases than the exome cohort, which was primarily comprised of samples collected from routine patient care following R-CHOP only. These clinical trials may have also excluded cases with extremely poor prognosis given their enrollment criteria.

After selecting for genes which showed evidence of positive selection across both DLBCL and rrDLBCL and comparing mutation frequency across these genes, we discovered five genes significantly enriched for mutations in rrDLBCL, four of which (*KMT2D, TP53, CREBBP, STAT6*) were found in our previous analysis. We also observed an enrichment of mutations perturbing *MYC*, which are indicative of *MYC* translocations. Mutations in *TMEM30A* were exceedingly rare and significantly depleted in rrDLBCL (mutated in 1.9% of cases), which is supported by its status as a good prognostic marker following R-CHOP therapy[328]. Mutations in the DNA demethylation

initiator *TET2* (1.3% of cases) were equally rare, suggesting broad methylome changes in rrDLBCL.

While only a handful of genes were differentially perturbed between diagnostic and rrDLBCL, we observed a plethora of recurrent CNVs in rrDLBCL enriched at relapse (Figure 3-7). Many of these enriched events perturbed genes involved in B-cell proliferation (*MIR17HG*), cell cycle regulation (*PTEN, CDK11B*), apoptosis (*BCL2, TP53*) and JAK/STAT signaling (*STAT6*). The enrichment of gains involving the *STAT6* locus and deletions of *TP53* support the enrichment of SSMs observed perturbing these genes. Through this analysis, we also discovered novel events which have not previously been associated DLBCL. The novel deletions of the RNA regulators *HNRNPD* and *HNRNPU*, in conjunction with the depletion of *TET2* mutations and enrichment of *KMT2D* and *CREBBP* mutations at relapse, further suggests that epigenetic dysregulation notably contribute to relapse biology. Recurrent deletions of *IRF2* have been shown to downregulate MHC Class I-mediated antigen presentation, thus contributing towards immune evasion[335]. *IKZF3* and *TCF3* have been shown to promote B-cell proliferation[344,352,353], and thus gains on these genes may enhance their oncogenic potential. Knockdowns of *IKZF3* and *TCF3* or downstream components have been associated with increased apoptosis in Burkitt Lymphoma (BL)[352] and DLBCL[337], and thus represent candidate therapeutic targets. However, these genes are only gained in ~20% of rrDLBCL tumours, and thus cases harbouring such events would need to be identified prior to therapy.

4.2. Implications of research

4.2.1. The genetics of rrDLBCL are generally similar to diagnostic DLBCL

Through an exome-wide analysis of 155 rrDLBCL cases, we have cataloged the repertoire of SSMs and CNVs in rrDLBCL and discovered that the landscape of events in rrDLBCL is generally comparable to that of diagnostic DLBCL with mutations perturbing common DLBCL drivers retained and persisting following treatment. While we uncovered several novel rrDLBCL specific events (namely mutations in *MS4A1*, deletions of *HNRNPD, HNRNPU, IRF2*, and gains involving *IKZF3* and *TCF3*), these events tended to occur intermittently across our rrDLBCL cohort. Thus, a relapsed

DLBCL tumour is not genetically distinct from its pre-treatment treatment tumour, but instead reflects and extends the original tumours genetics and biology. This is specifically relevant in the context of genetic DLBCL subgroups, which may selectively constrain the repertoire of mutations a tumour could acquire.

4.2.2. rrDLBCL is genetically heterogeneous, and there is no single mechanism of R-CHOP resistance

DLBCL tumours are characterized by their genetic heterogeneity[226], and rrDLBCL tumours reflect and share this heterogeneity. Through an exome-wide analysis of rrDLBCL tissue and liquid biopsies, we did not observe novel, prevalent SSMs exclusive to rrDLBCL. If a single set of genes or pathway were responsible for R-CHOP resistance, one would expect mutations or events within this pathway to be highly prevalent in relapse tumours. Given the heterogeneity of rrDLBCL, and the differing mutational constraints and repertoires of different molecular and genetic subgroups, the spectrum of mutations a given tumour can acquire is exceptionally diverse. Further mutational processes, such as aSHM[354], can generate passenger mutations which are initially rare but convey a selective advantage following therapy. Indeed, mutations in *MS4A1* (although not associated with aSHM) appeared to act as passenger mutations prior to therapy, given the low prevalence of such events. Mutations perturbing *MS4A1* were also more frequently observed in samples collected following multiple types of salvage therapy. Given the genetic constraints of each molecular and genetic subgroup in DLBCL, and the wide repertoire of possible mutations, it is possible that resistance-associated mutations will be unique for different subgroups. It is also likely that the dysregulation of epigenetic (*KMT2D, CREBBP*, *TET2*), and transcriptomic (*HNRNPD* and *HNRNPU*) factors contribute to treatment resistance through epigenetic and/or transcriptomic disregulation, and thus further research exploring these avenues is needed.

4.2.3. Candidate therapeutic targets

While the genetic landscape of rrDLBCL is generally similar to that of diagnostic DLBCL, we did observe several events enriched at relapse which are possible therapeutic targets. For instance, we observed a high prevalence and enrichment of gains perturbing *BCL2* (42% of rrDLBCL), especially in ABC-rrDLBCL. While BCL2

inhibitors such as venetoclax have shown promising efficacy in relapsed-refractory chronic lymphocytic leukemia[355] and MCL[356], venetoclax has generally performed poorly in patients with FL or DLBCL[356] (with a ORR of 18% in rrDLBCL), despite high *BCL2* expression in these malignancies. We also observed novel recurrent gains perturbing *IKZF3* and *TCF3* in rrDLBCL. *IKZF3*, encoding Aiolos, is a key lymphoid maturation factor[353] which transcriptionally represses interferon-stimulated genes (ISGs)[336] which normally induce apoptosis upon interferon stimulation. Lenalidomide and avadomide promote degradation of Aiolos, and have been shown to induce apoptosis in DLBCL cell lines[357]. Aiolos has also been shown to interact with histone deacetylases and transcriptionally repress target genes[357], compounding the loss-of-function mutations observed in *CREBBP* and enriched in rrDLBCL. While histone deacetylase inhibitors (HDACIs) alone have shown poor efficacy in treating rrDLBCL, with an ORR of 10%[358,359], combinatorial therapies combining lenalidomide and HDACIs have shown improved responses in cell lines resistant to lenalidomide monotherapy[357]. Furthermore, *TCF3*, encoding the transcription factor E2A, and the JAK/STAT signaling regulator *STAT6* also represent candidate targets of therapeutic potential.

4.2.4. Mechanisms of treatment resistance

Although mutated in a minority of rrDLBCL cases, we discovered that mutations perturbing *MS4A1* were enriched at relapse, and these mutations impaired CD20 from being presented on the surface of lymphoma cells (Figure 2-9B). Thus, these mutations, both truncating and missense (Figure 2-9A) act as a direct mechanism of resistance against rituximab and other anti-CD20 mAbs. Given that many rrDLBCL salvage therapies include anti-CD20 monoclonal antibodies, cases harbouring *MS4A1* mutations should have CD20 expression re-evaluated prior to salvage therapy, although treatment options for CD20- DLBCL cases are generally limited. We would also expect *MS4A1* mutations to act as a predictive biomarker in the context of therapy; however, due to the limited number of cases with *MS4A1* mutations, this could not be robustly evaluated.

4.3. Ongoing work and future directions

4.3.1. Serial sampling of rrDLBCL cases

In Chapter 2 and Chapter 3, we sequenced liquid biopsies collected from the routine patient population in Montreal, Quebec. While preliminary CAPP-Seq data was incorporated in Chapter 2, and a sizable number of cases with lpWGS data (77) were incorporated in Chapter 3, we have subsequently collected a total of 588 liquid biopsies from 252 patients with rrDLBCL. These liquid biopsies are collected at diagnosis, during and following frontline and salvage therapies, and also includes patients treated with both standard salvage therapy (R-GDP) and experimental salvage therapies such as tazemetostat, anti-CD19-CD3 BITE, and CAR-T cell therapy. We have performed lpWGS on all 588 samples and detected ctDNA in 157 cases (26.7%) despite the limited sensitivity of lpWGS (>=7% ctDNA required). To improve on this sensitivity and to identify SSMs, we are currently performing CAPP-Seq on all samples with an expanded gene panel incorporating both described lymphoma drivers, candidate genes associated with treatment resistance (for instance, CD19), and regions frequently affected by aSHM, with a theoretical sensitivity of 1% ctDNA. As of October 2022, CAPP-Seq using this panel has been completed on 403/588 samples. This cohort is extremely heterogeneous in terms of time points, treatments performed, and patient characteristics. As such, we are planning to compare the overall repertoire of mutations across rrDLBCL samples, compare mutations between time points to look for examples of clonal evolution, and group cases and samples with similar treatment regimens (i.e., before and following R-CHOP, before and following R-GDP) to identify recurring events selected by these treatments. For candidate cases where the tumour is initially response to therapy but the patient later relapses, we are planning to sequence interim time points extremely deeply (>10,000x coverage) with the aim of MRD detection. We will also evaluate ctDNA levels from samples collected before and immediately following salvage therapy to evaluate ctDNA as a predictor of treatment outcomes. Samples are also being collected and sequenced from DLBCL and NHL cases treated with CAR-T therapy in British Columbia, and these samples will be analyzed in a similar manner.

4.3.2. The epigenetic and transcriptomic landscape of rrDLBCL

Through an exploration of the genomic features of rrDLBCL, we observed an enrichment of events perturbing genes involved in histone methylation (*KMT2D*) and acetylation (*CREBBP*), as well as RNA regulation (*TET2*). Due to the rapid acquisition of treatment resistance observed in some rrDLBCL cases (for instance, PT255, Section 2.5.6) and the prevalence of these events, epigenetic and transcriptomic regulation may contribute significantly to therapeutic resistance. Our group has recently performed RNA sequencing on 79 relapse tissue biopsies from 72 rrDLBCL patients. Through a preliminary analysis of expression patterns and a differential expression analysis to identify genes dysregulated in rrDLBCL, we discovered that transcriptomic differences between the diagnostic and relapse tissue biopsy are generally minor and heterogeneous between cases, further supporting the similarities in the mutational landscape observed following DNA sequencing. Indeed, when incorporating diagnostic-relapse pairs the transcriptome of a rrDLBCL tumour appears to be most similar to the corresponding diagnostic tumour, and not other relapse tumour biopsies, further suggesting that features associated with treatment failure are either intrinsic to the tumour itself or are minor additions to the biology of the disease. Due to the genetic and molecular diversity of rrDLBCL, a larger transcriptomic cohort of samples is needed.

Given the high prevalence of *KMT2D*, *CREBBP*, and *TET2* mutations in rrDLBCL, the resulting epigenetic effects of these events must be examined. A previous study exploring DNA methylation patterns in 13 diagnostic-relapse rrDLBCL pairs found that, while methylation of rrDLBCL tumours generally reflected their diagnostic counterparts, there are patterns of convergent evolution of rrDLBCL tumours following R-CHOP, and increased hypomethylation of promoters involved in TGF-β signaling[360]. Further large-scale studies investigating methylation patterns across rrDLBCL pairs are needed, both to identify other patterns of DNA methylation, and find patterns specific to molecular and genetic subgroups. These would necessitate diagnostic-relapse-normal “trios”, and such events could be determined using long read sequencing technologies (nanopore sequencing)[361,362] and modified Illumina sequencing approaches such as bisulfite sequencing[363] . While bisulfite sequencing has been performed on liquid biopsies[364], alternative cfDNA-specific sequencing approaches such as cfMeDIP-seq[365], and techniques correlating cfDNA fragmentation patterns with actively transcribed genes[215] can infer methylomic and transcriptomic patterns in liquid biopsies. Patterns of

histone post-translational modification (H3K27Me3 etc.), and the DNA associated with such events could be devised using CHIP-Seq[366]. These epigenetic marks could then be compared between diagnostic and relapse samples to identify relapse-specific epigenetic patterns. These patterns could further be stratified by the presence or absence of mutations in key epigenetic modifiers such as *KMT2D, CREBBP,* and *TET2* to evaluate the downstream effect of these mutations.

4.3.3. Mechanism and impact of *MS4A1* mutations in DLBCL

In Chapter 2.5.3, we observed that mutations in *MS4A1* attenuate the binding of both rituximab and other anti-CD20 monoclonal antibodies. Curiously, frameshift events were generally restricted to the large loop of CD20, while missense mutations were restricted to the small loop and transmembrane domains (Figure 2-9A). We also did not observe focal deletions or other CNVs perturbing the *MS4A1* in our rrDLBCL cohort. As deletion of the *MS4A1* locus would prevent antibody binding, the retention of the CD20 locus and pattern of mutations observed in rrDLBCL is curious and suggests retention of the *MS4A1* locus is selectively advantageous in DLBCL. Further work comparing the phenotypes of DLBCL cells harbouring wild-type *MS4A1*, *MS4A1* with missense mutations withing the small loop region, and those lacking *MS4A1* locus entirely could provide further insight into the effect of *MS4A1* mutations and the selective pressures on the *MS4A1* locus.

4.3.4. Contribution of non-coding events

While the vast majority of DLBCL and rrDLBCL studies have focused on evaluating the repertoire of coding mutations across DLBCL (including those outlined here), the role and contribution of non-coding drivers has been largely underexplored. Non-coding drivers have been observed in DLBCL, including aSHM disrupting superenhancers[367] and non-coding regions of specific genes (such as the NF-KB signaling component *NFKBIZ*[15]). Our group has performed WGS of diagnostic-relapse-normal “trios” for rrDLBCL cases, and we will investigate the contribution of such events.

4.4. Closing perspective

Through the two large-scale genomic studies of rrDLBCL described in Chapter 2 and Chapter 3, we have sought to explore and characterize the landscape of genomic features in rrDLBCL, and how such events compare to diagnostic DLBCL. Through this approach, we have discovered a direct mechanism of treatment resistance (mutations in *MS4A1*), biomarkers of treatment failure (mutations in *KMT2D* and *TP53*), and candidate therapeutic targets (*IKZF3* and *TCF3*). This project also highlights the utility of liquid biopsies in cancer genomics research, enabling serial samples to be easily collected from patients undergoing treatment, and the application of low-cost sequencing approaches (lpWGS and CAPP-Seq) to identify genetic features of rrDLBCL. These techniques can further be implemented to screen for candidate therapeutic targets prior to treatment, monitoring treatment response, and monitoring for the emergence of resistance mutations following therapy. Through additional genomic, epigenomic, and transcriptomic studies of rrDLBCL, it is hoped that additional mechanisms of treatment resistance and biomarkers of treatment failure can be uncovered.

References

1. Schaffer, C. J. & Nanney, L. B. Cell Biology of Wound Healing. in *International Review of Cytology* (ed. Jeon, K. W.) vol. 169 151–181 (Academic Press, 1996).

2. Schreml, S., Szeimies, R.-M., Prantl, L., Landthaler, M. & Babilas, P. Wound healing in the 21st century. *J. Am. Acad. Dermatol.* **63**, 866–881 (2010).

3. Nurse, P. A Long Twentieth Century of the Cell Cycle and Beyond. *Cell* **100**, 71–78 (2000).

4. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res.* **34**, 2311–2321 (1974).

5. Frenkel, K. Carcinogen-mediated oxidant formation and oxidative DNA damage. *Pharmacol. Ther.* **53**, 127–166 (1992).

6. Miller, E. C. & Miller, J. A. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer* **47**, 2327–2345 (1981).

7. Miller, J. A. & Miller, E. C. The Concept of Reactive Electrophilic Metabolites in Chemical Carcinogenesis: Recent Results with Aromatic Amines, Safrole, and Aflatoxin B1. in *Biological Reactive Intermediates: Formation, Toxicity, and Inactivation* (eds. Jollow, D. J. et al.) 6–24 (Springer US, 1977). doi:10.1007/978-1-4613-4124-6_2.

8. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74 (2012).

9. Biémont, C. & Vieira, C. Junk DNA as an evolutionary force. *Nature* **443**, 521–524 (2006).

10. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

11. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

12. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

13. Houldsworth, J. *et al.* Relationship between REL amplification, REL function, and clinical and biologic features in diffuse large B-cell lymphomas. *Blood* **103**, 1862–1868 (2004).

14. Lenz, G. *et al.* Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci.* **105**, 13520–13525 (2008).

15. Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).

16. Weinberg, R. A. Oncogenes and tumor suppressor genes. *CA. Cancer J. Clin.* **44**, 160–170 (1994).

17. Mao, J.-H. *et al.* Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. *Nature* **432**, 775–779 (2004).

18. Park, B.-J. *et al.* The Haploinsufficient Tumor Suppressor p18 Upregulates p53 via Interactions with ATM/ATR. *Cell* **120**, 209–221 (2005).

19. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* **23**, 2330–2338 (2004).

20. Ngo, V. N. *et al.* Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119 (2011).

21. Kim, Y. *et al.* CD79B and MYD88 mutations in diffuse large B-cell lymphoma. *Hum Pathol* **45**, 556–564 (2014).

22. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

23. Balkwill, F. & Mantovani, A. Inflammation and cancer: back to Virchow? *The Lancet* **357**, 539–545 (2001).

24. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).

25. Kalluri, R. The biology and function of fibroblasts in cancer. *Nat. Rev. Cancer* **16**, 582–598 (2016).

26. Del Monte, U. Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle* **8**, 505–506 (2009).

27. Heppner, G. H. Tumor Heterogeneity. *Cancer Res.* **44**, 2259–2265 (1984).

28. Tomlinson, I. & Bodmer, W. Selection, the mutation rate and cancer: Ensuring that the tail does not wag the dog. *Nat. Med.* **5**, 11–12 (1999).

29. Calbo, J. *et al.* A Functional Role for Tumor Cell Heterogeneity in a Mouse Model of Small Cell Lung Cancer. *Cancer Cell* **19**, 244–256 (2011).

30. Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. *Nat. Rev. Cancer* **15**, 473–483 (2015).

31. Onrust, S. V., Lamb, H. M. & Barman Balfour, J. A. Rituximab. *Drugs* **58**, 79–88 (1999).

32. Fornari, F. A., Randolph, J. K., Yalowich, J. C., Ritke, M. K. & Gewirtz, D. A. Interference by doxorubicin with DNA unwinding in MCF-7 breast tumor cells. *Mol. Pharmacol.* **45**, 649–656 (1994).

33. Siddik, Z. H. Cisplatin: mode of cytotoxic action and molecular basis of resistance. *Oncogene* **22**, 7265–7279 (2003).

34. Almendro, V. *et al.* Inference of Tumor Evolution during Chemotherapy by Computational Modeling and In Situ Analysis of Genetic and Phenotypic Cellular Diversity. *Cell Rep.* **6**, 514–527 (2014).

35. Turke, A. B. *et al.* Preexistence and Clonal Selection of MET Amplification in EGFR Mutant NSCLC. *Cancer Cell* **17**, 77–88 (2010).

36. Shah, N. P. *et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2**, 117–125 (2002).

37. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).

38. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

39. Chang, E. H., Furth, M. E., Scolnick, E. M. & Lowy, D. R. Tumorigenic transformation of mammalian cells induced by a normal human gene homologous to the oncogene of Harvey murine sarcoma virus. *Nature* **297**, 479–483 (1982).

40. Robbins, Y. *et al.* Dual PD-L1 and TGF-b blockade in patients with recurrent respiratory papillomatosis. *J. Immunother. Cancer* **9**, e003113 (2021).

41. Olovnikov, A. Principle of marginotomy in template synthesis of polynucleotides. in *Dokl. Akad. Nauk. SSSR* vol. 201 1496–1499 (1971).

42. Blackburn, E. H. & Gall, J. G. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.* **120**, 33–53 (1978).

43. Lundblad, V. & Szostak, J. W. A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell* **57**, 633–643 (1989).

44. Kyo, S. *et al.* Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT). *Nucleic Acids Res.* **28**, 669–677 (2000).

45. Folkman, J. Tumor Angiogenesis: Therapeutic Implications. *N. Engl. J. Med.* **285**, 1182–1186 (1971).

46. Hargreaves, M. & Spriet, L. L. Skeletal muscle energy metabolism during exercise. *Nat. Metab.* **2**, 817–828 (2020).

47. Sahlin, K., Tonkonogi, M. & Söderlund, K. Energy supply and muscle fatigue in humans. *Acta Physiol. Scand.* **162**, 261–266 (1998).

48. Mazurek, S. & Eigenbrodt, E. The tumor metabolome. *Anticancer Res.* **23**, 1149–1154 (2003).

49. Fantin, V. R., St-Pierre, J. & Leder, P. Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell* **9**, 425–434 (2006).

50. Lindahl, T. & Wood, R. D. Quality Control by DNA Repair. *Science* **286**, 1897–1905 (1999).

51. Guengerich, F. P. Cytochrome P450 and Chemical Toxicology. *Chem. Res. Toxicol.* **21**, 70–83 (2008).

52. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci.* **107**, 961–968 (2010).

53. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071–1078 (2009).

54. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability — an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).

55. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *science* **316**, 1160–1166 (2007).

56. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432**, 307–315 (2004).

57. Olivier, M., Hollstein, M. & Hainaut, P. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008 (2010).

58. Kloetzel, P. M. Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII. *Nat. Immunol.* **5**, 661–669 (2004).

59. Angell, T. E., Lechner, M. G., Jang, J. K., LoPresti, J. S. & Epstein, A. L. MHC Class I Loss Is a Frequent Mechanism of Immune Escape in Papillary Thyroid Cancer That Is Reversed by Interferon and Selumetinib Treatment In Vitro. *Clin. Cancer Res.* **20**, 6034–6044 (2014).

60. Gettinger, S. *et al.* Impaired HLA Class I Antigen Processing and Presentation as a Mechanism of Acquired Resistance to Immune Checkpoint Inhibitors in Lung Cancer. *Cancer Discov.* **7**, 1420–1435 (2017).

61. Kärre, K., Ljunggren, H. G., Piontek, G. & Kiessling, R. Selective rejection of H–2-deficient lymphoma variants suggests alternative immune defence strategy. *Nature* **319**, 675–678 (1986).

62. Sakaguchi, S., Yamaguchi, T., Nomura, T. & Ono, M. Regulatory T Cells and Immune Tolerance. *Cell* **133**, 775–787 (2008).

63. Schoppmann, S. F. *et al.* Tumor-Associated Macrophages Express Lymphatic Endothelial Growth Factors and Are Related to Peritumoral Lymphangiogenesis. *Am. J. Pathol.* **161**, 947–956 (2002).

64. Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian cancer Statistics 2021. (2021).

65. Zhou, L. *et al.* Global, regional, and national burden of Hodgkin lymphoma from 1990 to 2017: estimates from the 2017 Global Burden of Disease study. *J. Hematol. Oncol.J Hematol Oncol* **12**, 107 (2019).

66. Shankland, K. R., Armitage, J. O. & Hancock, B. W. Non-Hodgkin lymphoma. *The Lancet* **380**, 848–857 (2012).

67. Dearden, C. E. *et al.* Guidelines for the management of mature T-cell and NK-cell neoplasms (excluding cutaneous T-cell lymphoma). *Br. J. Haematol.* **153**, 451–485 (2011).

68. Liang, X. & Graham, D. K. Natural killer cell neoplasms. *Cancer* **112**, 1425–1436 (2008).

69. Smith, A. *et al.* Lymphoma incidence, survival and prevalence 2004–2014: sub-type analyses from the UK's Haematological Malignancy Research Network. *Br. J. Cancer* **112**, 1575–1584 (2015).

70. SH, S. *et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*.

71. Fisher, R. I. *et al.* New treatment options have changed the survival of patients with follicular lymphoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 8447–8452 (2005).

72. de Jong, D. *et al.* Activation of the c-myc Oncogene in a Precursor-B-Cell Blast Crisis of Follicular Lymphoma, Presenting as Composite Lymphoma. *N. Engl. J. Med.* **318**, 1373–1378 (1988).

73. Carbone, A. *et al.* Follicular lymphoma. *Nat. Rev. Dis. Primer* **5**, 1–20 (2019).

74. Cheah, C. Y., Seymour, J. F. & Wang, M. L. Mantle Cell Lymphoma. *J. Clin. Oncol.* **34**, 1256–1269 (2016).

75. Cortelazzo, S., Ponzoni, M., Ferreri, A. J. M. & Dreyling, M. Mantle cell lymphoma. *Crit. Rev. Oncol. Hematol.* **82**, 78–101 (2012).

76. Calabrò, M. L. & Sarid, R. Human Herpesvirus 8 and Lymphoproliferative Disorders. *Mediterr. J. Hematol. Infect. Dis.* **10**, e2018061 (2018).

77. Oyama, T. *et al.* Senile EBV+ B-cell lymphoproliferative disorders: a clinicopathologic study of 22 patients. *Am. J. Surg. Pathol.* **27**, 16–26 (2003).

78. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

79. Wright, G. *et al.* A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9991–9996 (2003).

80. Davis, R. E., Brown, K. D., Siebenlist, U. & Staudt, L. M. Constitutive Nuclear Factor κB Activity Is Required for Survival of Activated B Cell–like Diffuse Large B Cell Lymphoma Cells. *J. Exp. Med.* **194**, 1861–1874 (2001).

81. Nowakowski, G. S. *et al.* Variable global distribution of cell-of-origin from the ROBUST phase III study in diffuse large B-cell lymphoma. *Haematologica* **105**, e72–e75 (2020).

82. von Ahlfen, S., Missel, A., Bendrat, K. & Schlumpberger, M. Determinants of RNA Quality from FFPE Samples. *PLoS ONE* **2**, e1261 (2007).

83. Hans, C. P. *et al.* Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103**, 275–282 (2004).

84. Choi, W. W. L. *et al.* A New Immunostain Algorithm Classifies Diffuse Large B-Cell Lymphoma into Molecular Subtypes with High Accuracy. *Clin Cancer Res* **15**, 5494–5502 (2009).

85. Colomo, L. *et al.* Clinical impact of the differentiation profile assessed by immunophenotyping in patients with diffuse large B-cell lymphoma. *Blood* **101**, 78–84 (2003).

86. de Jong, D. *et al.* Immunohistochemical Prognostic Markers in Diffuse Large B-Cell Lymphoma: Validation of Tissue Microarray As a Prerequisite for Broad Clinical Applications—A Study From the Lunenburg Lymphoma Biomarker Consortium. *J. Clin. Oncol.* **25**, 805–812 (2007).

87. Scott, D. W. *et al.* Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin Determined by Digital Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue Biopsies. *J. Clin. Oncol.* **33**, 2848–2856 (2015).

88. Scott, D. W. *et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214–1217 (2014).

89. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with colour-coded probe pairs. *Nat. Biotechnol.* **26**, 317–325 (2008).

90. Mottok, A. *et al.* Molecular classification of primary mediastinal large B-cell lymphoma using routinely available tissue specimens. *Blood* **132**, 2401–2405 (2018).

91. Ennishi, D. *et al.* Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **37**, 190–201 (2019).

92. Skladanowski, A. & Konopa, J. Adriamycin and daunomycin induce programmed cell death (apoptosis) in tumour cells. *Biochem. Pharmacol.* **46**, 375–382 (1993).

93. William Lown, J., Sim, S.-K., Majumdar, K. C. & Chang, R.-Y. Strand scission of DNA by bound adriamycin and daunorubicin in the presence of reducing agents. *Biochem. Biophys. Res. Commun.* **76**, 705–710 (1977).

94. Phillips, D. R., White, R. J. & Cullinane, C. DNA sequence-specific adducts of adriamycin and mitomycin C. *FEBS Lett.* **246**, 233–240 (1989).

95. Na, G. C. & Timasheff, S. N. Interaction of vinblastine with calf brain tubulin: multiple equilibria. *Biochemistry* **25**, 6214–6222 (1986).

96. Cronstein, B. N., Kimmel, S. C., Levin, R. I., Martiniuk, F. & Weissmann, G. A mechanism for the antiinflammatory effects of corticosteroids: the glucocorticoid receptor regulates leukocyte adhesion to endothelial cells and expression of endothelial-leukocyte adhesion molecule 1 and intercellular adhesion molecule 1. *Proc. Natl. Acad. Sci.* **89**, 9991–9995 (1992).

97. Coiffier, B. *et al.* CHOP Chemotherapy plus Rituximab Compared with CHOP Alone in Elderly Patients with Diffuse Large-B-Cell Lymphoma. *N. Engl. J. Med.* **346**, 235–242 (2002).

98. Nadler, L. M. *et al.* A unique cell surface antigen identifying lymphoid malignancies of B cell origin. https://www.jci.org/articles/view/110005/scanned-page/134 (1981) doi:10.1172/JCI110005.

99. Sehn, L. H. *et al.* Introduction of Combined CHOP Plus Rituximab Therapy Dramatically Improved Outcome of Diffuse Large B-Cell Lymphoma in British Columbia. *J. Clin. Oncol.* **23**, 5027–5033 (2005).

100. Feugier, P. *et al.* Long-Term Results of the R-CHOP Study in the Treatment of Elderly Patients With Diffuse Large B-Cell Lymphoma: A Study by the Groupe d'Etude des Lymphomes de l'Adulte. *J. Clin. Oncol.* **23**, 4117–4126 (2005).

101. Sant, M. *et al.* Survival for haematological malignancies in Europe between 1997 and 2008 by region and age: results of EUROCARE-5, a population-based study. *Lancet Oncol.* **15**, 931–942 (2014).

102. Coiffier, B. *et al.* Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d'Etudes des Lymphomes de l'Adulte. *Blood* **116**, 2040–2045 (2010).

103. Maurer, M. J. *et al.* Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32**, 1066–1073 (2014).

104. Hainsworth, J. D. *et al.* A randomized, phase 2 study of R-CHOP plus enzastaurin vs R-CHOP in patients with intermediate- or high-risk diffuse large B-cell lymphoma. *Leuk. Lymphoma* **57**, 216–218 (2016).

105. Moccia, A. A. *et al.* Long-term outcomes of R-CEOP show curative potential in patients with DLBCL and a contraindication to anthracyclines. *Blood Adv.* **5**, 1483–1489 (2021).

106. Lue, J. K. & O'Connor, O. A. A perspective on improving the R-CHOP regimen: from Mega-CHOP to ROBUST R-CHOP, the PHOENIX is yet to rise. *Lancet Haematol.* **7**, e838–e850 (2020).

107. Mian, M. *et al.* R-CHOP versus R-COMP: Are They Really Equally Effective? *Clin. Oncol.* **26**, 648–652 (2014).

108. Crump, M. *et al.* Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. *Blood* **130**, 1800–1808 (2017).

109. Crump, M. *et al.* Randomized Comparison of Gemcitabine, Dexamethasone, and Cisplatin Versus Dexamethasone, Cytarabine, and Cisplatin Chemotherapy Before Autologous Stem-Cell Transplantation for Relapsed and Refractory Aggressive Lymphomas: NCIC-CTG LY.12. *J. Clin. Oncol.* **32**, 3490–3496 (2014).

110. Huang, P., Chubb, S., Hertel, L. W., Grindey, G. B. & Plunkett, W. Action of 2′,2′-Difluorodeoxycytidine on DNA Synthesis1. *Cancer Res.* **51**, 6110–6117 (1991).

111. Coates, T. *et al.* The mechanism of action of the antiinflammatory agents dexamethasone and Auranofin in human polymorphonuclear leukocytes. *Blood* **62**, 1070–1077 (1983).

112. Ribrag, V. *et al.* Interim Results from an Ongoing Phase 2 Multicenter Study of Tazemetostat, an EZH2 Inhibitor, in Patients with Relapsed or Refractory (R/R) Diffuse Large B-Cell Lymphoma (DLBCL). *Blood* **132**, 4196 (2018).

113. Goy, A. *et al.* Ibrutinib plus lenalidomide and rituximab has promising activity in relapsed/refractory non–germinal center B-cell–like DLBCL. *Blood* **134**, 1024–1036 (2019).

114. Melani, C. Phase 1b/2 Study of Vipor (Venetoclax, Ibrutinib, Prednisone, Obinutuzumab, and Lenalidomide) in Relapsed/Refractory B-Cell Lymphoma: Safety, Efficacy and Molecular Analysis. in (ASH, 2020).

115. Hutchings, M. *et al.* Glofitamab, a Novel, Bivalent CD20-Targeting T-Cell–Engaging Bispecific Antibody, Induces Durable Complete Remissions in Relapsed or Refractory B-Cell Lymphoma: A Phase I Trial. *J. Clin. Oncol.* **39**, 1959–1970 (2021).

116. Irving, B. A. & Weiss, A. The cytoplasmic domain of the T cell receptor ζ chain is sufficient to couple to receptor-associated signal transduction pathways. *Cell* **64**, 891–901 (1991).

117. Kalos, M. *et al.* T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci. Transl. Med.* **3**, 95ra73-95ra73 (2011).

118. Gill, S., Maus, M. V. & Porter, D. L. Chimeric antigen receptor T cell therapy: 25years in the making. *Blood Rev.* **30**, 157–167 (2016).

119. Lee, D. W. *et al.* T cells expressing CD19 chimeric antigen receptors for acute lymphoblastic leukaemia in children and young adults: a phase 1 dose-escalation trial. *The Lancet* **385**, 517–528 (2015).

120. Kochenderfer, J. N. *et al.* B-cell depletion and remissions of malignancy along with cytokine-associated toxicity in a clinical trial of anti-CD19 chimeric-antigen-receptor–transduced T cells. *Blood* **119**, 2709–2720 (2012).

121. Osborne, W. *et al.* Phase I Alexander study of AUTO3, the first CD19/22 dual targeting CAR T cell therapy, with pembrolizumab in patients with relapsed/refractory (r/r) DLBCL. *J. Clin. Oncol.* **38**, 8001–8001 (2020).

122. Neelapu, S. S. *et al.* Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma. *N. Engl. J. Med.* **377**, 2531–2544 (2017).

123. Schuster, S. J. *et al.* Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* **380**, 45–56 (2019).

124. Lyman, G. H., Nguyen, A., Snyder, S., Gitlin, M. & Chung, K. C. Economic Evaluation of Chimeric Antigen Receptor T-Cell Therapy by Site of Care Among Patients With Relapsed or Refractory Large B-Cell Lymphoma. *JAMA Netw. Open* **3**, e202072 (2020).

125. Head, S. R. *et al.* Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* **56**, 61-passim (2014).

126. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

127. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *https://doi.org/10.2144/000113809* https://www.future-science.com/doi/10.2144/000113809 (2018) doi:10.2144/000113809.

128. Williams, C. *et al.* A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens. *Am. J. Pathol.* **155**, 1467–1471 (1999).

129. Do, H. & Dobrovic, A. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. *Oncotarget* **3**, 546–558 (2012).

130. Berra, C. M. *et al.* Use of uracil-DNA glycosylase enzyme to reduce DNA-related artifacts from formalin-fixed and paraffin-embedded tissues in diagnostic routine. *Appl. Cancer Res.* **39**, 7 (2019).

131. Serizawa, M. *et al.* The efficacy of uracil DNA glycosylase pretreatment in amplicon-based massively parallel sequencing with DNA extracted from archived formalin-fixed paraffin-embedded esophageal cancer tissues. *Cancer Genet.* **208**, 415–427 (2015).

132. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

133. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).

134. Mandelker, D. & Ceyhan-Birsoy, O. Evolving Significance of Tumor-Normal Sequencing in Cancer Care. *Trends Cancer* **6**, 31–39 (2020).

135. Schwarze, K. *et al.* The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet. Med.* **22**, 85–94 (2020).

136. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics* **22**, 197 (2021).

137. Chen, X. *et al.* Low-pass Whole-genome Sequencing of Circulating Cell-free DNA Demonstrates Dynamic Changes in Genomic Copy Number in a Squamous Lung Cancer Clinical Cohort. *Clin. Cancer Res.* **25**, 2254–2263 (2019).

138. Wang, H. *et al.* Low-pass genome sequencing versus chromosomal microarray analysis: implementation in prenatal diagnosis. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 500–510 (2020).

139. Chau, M. H. K. *et al.* Low-pass genome sequencing: a validated method in clinical cytogenetics. *Hum. Genet.* **139**, 1403–1415 (2020).

140. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).

141. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).

142. Parham, N. J. *et al.* Specific magnetic bead based capture of genomic DNA from clinical samples: application to the detection of group B streptococci in vaginal/anal swabs. *Clin. Chem.* **53**, 1570–1576 (2007).

143. Gordon, L. G. *et al.* Estimating the costs of genomic sequencing in cancer control. *BMC Health Serv. Res.* **20**, 492 (2020).

144. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130 (2018).

145. Barbitoff, Y. A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci. Rep.* **10**, 2057 (2020).

146. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).

147. Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* **12**, 423–425 (2015).

148. Genomics in the Cloud [Book]. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/.

149. Liehr, T. Repetitive Elements in Humans. *Int. J. Mol. Sci.* **22**, 2072 (2021).

150. Paszkiewicz, K. & Studholme, D. J. De novo assembly of short sequence reads. *Brief. Bioinform.* **11**, 457–472 (2010).

151. Liao, X. *et al.* Current challenges and solutions of de novo assembly. *Quant. Biol.* **7**, 90–109 (2019).

152. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

153. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

154. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

155. Fox, C. H., Johnson, F. B., Whiting, J. & Roller, P. P. Formaldehyde fixation. *J. Histochem. Cytochem.* **33**, 845–853 (1985).

156. Do, H. & Dobrovic, A. Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clin. Chem.* **61**, 64–71 (2015).

157. McNulty, S. N., Mann, P. R., Robinson, J. A., Duncavage, E. J. & Pfeifer, J. D. Impact of Reducing DNA Input on Next-Generation Sequencing Library Complexity and Variant Detection. *J. Mol. Diagn.* **22**, 720–727 (2020).

158. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).

159. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

160. Yang, X. *et al.* HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* **14**, 1–4 (2013).

161. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

162. Taylor-Weiner, A. *et al.* DeTiN : Overcoming Tumor in Normal Contamination. *Nat. Methods* **15**, 531–534 (2018).

163. Stieglitz, E. *et al.* The genomic landscape of juvenile myelomonocytic leukemia. *Nat. Genet.* **47**, 1326–1333 (2015).

164. Benjamin, D. *et al. Calling Somatic SNVs and Indels with Mutect*. https://doi.org/10.1101/861054 (2019) doi:10.1101/861054.

165. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).

166. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28**, 1811–1817 (2012).

167. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

168. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).

169. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

170. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

171. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE* **11**, e0151664 (2016).

172. Cai, L., Yuan, W., Zhang, Z., He, L. & Chou, K.-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.* **6**, 36540 (2016).

173. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).

174. Sievers, C. *et al.* Comprehensive multiomic characterization of human papillomavirus-driven recurrent respiratory papillomatosis reveals distinct molecular subtypes. *Commun. Biol.* **4**, 1–11 (2021).

175. Wang, M. *et al.* SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Sci. Rep.* **10**, 12898 (2020).

176. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).

177. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).

178. Chen, Y.-C. *et al.* Comprehensive Assessment of Somatic Copy Number Variation Calling Using Next-Generation Sequencing Data. 2021.02.18.431906 Preprint at https://doi.org/10.1101/2021.02.18.431906 (2021).

179. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **26**, 64–70 (2015).

180. Kuilman, T. *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* **16**, 49 (2015).

181. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.* **12**, e1004873 (2016).

182. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1–13 (2017).

183. Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J. & Menten, B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res.* **47**, 1605–1614 (2019).

184. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

185. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).

186. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).

187. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).

188. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma. Oxf. Engl.* **32**, 1220–1222 (2016).

189. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).

190. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

191. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718 (2012).

192. Haile, S. *et al.* Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res.* **47**, e12 (2019).

193. Escudero, L., Martínez-Ricarte, F. & Seoane, J. ctDNA-Based Liquid Biopsy of Cerebrospinal Fluid in Brain Cancer. *Cancers* **13**, 1989 (2021).

194. Kelly, R. J. *et al.* Complications and Economic Burden Associated With Obtaining Tissue for Diagnosis and Molecular Analysis in Patients With Non–Small-Cell Lung Cancer in the United States. *J. Oncol. Pract.* **15**, e717–e727 (2019).

195. Posielski, N. M. *et al.* Risk Factors for Complications and Nondiagnostic Results following 1,155 Consecutive Percutaneous Core Renal Mass Biopsies. *J. Urol.* **201**, 1080–1087 (2019).

196. Zhang, Y., Shi, L., Simoff, M. J., J Wagner, O. & Lavin, J. Biopsy frequency and complications among lung cancer patients in the United States. *Lung Cancer Manag.* **9**, LMT40 (2020).

197. VanderLaan, P. A. *et al.* Success and failure rates of tumor genotyping techniques in routine pathological samples with non-small-cell lung cancer. *Lung Cancer* **84**, 39–44 (2014).

198. Lin, D. *et al.* Circulating tumor cells: biology and clinical significance. *Signal Transduct. Target. Ther.* **6**, 1–24 (2021).

199. Coleman, M. L. *et al.* Membrane blebbing during apoptosis results from caspase-mediated activation of ROCK I. *Nat. Cell Biol.* **3**, 339–345 (2001).

200. Atkin-Smith, G. K. *et al.* A novel mechanism of generating extracellular vesicles during apoptosis via a beads-on-a-string membrane structure. *Nat. Commun.* **6**, 7439 (2015).

201. Erwig, L.-P. & Henson, P. M. Clearance of apoptotic cells by phagocytes. *Cell Death Differ.* **15**, 243–250 (2008).

202. Khier, S. & Lohan, L. Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Future Sci. OA* **4**, FSO295 (2018).

203. Thierry, A. R. *et al.* Origin and quantification of circulating DNA in mice with human colourectal cancer xenografts. *Nucleic Acids Res.* **38**, 6159–6175 (2010).

204. Bettegowda, C. *et al.* Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **6**, 224ra24-224ra24 (2014).

205. Wang, H.-C. *et al.* Serial Plasma Deoxyribonucleic Acid Levels as Predictors of Outcome in Acute Traumatic Brain Injury. *J. Neurotrauma* **31**, 1039–1045 (2014).

206. Strijker, M. *et al.* Circulating tumor DNA quantity is related to tumor volume and both predict survival in metastatic pancreatic ductal adenocarcinoma. *Int. J. Cancer* **146**, 1445–1456 (2020).

207. Escudero, L. *et al.* Circulating tumour DNA from the cerebrospinal fluid allows the characterisation and monitoring of medulloblastoma. *Nat. Commun.* **11**, 5376 (2020).

208. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).

209. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA: Chromatin structure is based on a repeating unit of eight histone molecules and about 200 DNA base pairs. *Science* **184**, 868–871 (1974).

210. Noll, M. & Kornberg, R. D. Action of micrococcal nuclease on chromatin and the location of histone H1. *J. Mol. Biol.* **109**, 393–404 (1977).

211. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).

212. Arends, M. J., Morris, R. G. & Wyllie, A. H. Apoptosis. The role of the endonuclease. *Am. J. Pathol.* **136**, 593–608 (1990).

213. Lo, Y. M. D. *et al.* Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Sci. Transl. Med.* **2**, 61ra91-61ra91 (2010).

214. Underhill, H. R. *et al.* Fragment Length of Circulating Tumor DNA. *PLoS Genet.* **12**, e1006162 (2016).

215. Esfahani, M. S. *et al.* Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597 (2022).

216. Osumi, H., Shinozaki, E., Yamaguchi, K. & Zembutsu, H. Early change in circulating tumor DNA as a potential predictor of response to chemotherapy in patients with metastatic colourectal cancer. *Sci. Rep.* **9**, 17358 (2019).

217. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).

218. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

219. Alcaide, M. *et al.* Targeted error-suppressed quantification of circulating tumor DNA using semi-degenerate barcoded adapters and biotinylated baits. *Sci. Rep.* **7**, 1–19 (2017).

220. Honoré, N., Galot, R., van Marcke, C., Limaye, N. & Machiels, J.-P. Liquid Biopsy to Detect Minimal Residual Disease: Methodology and Impact. *Cancers* **13**, 5364 (2021).

221. Khan, K. H. *et al.* Longitudinal Liquid Biopsy and Mathematical Modeling of Clonal Evolution Forecast Time to Treatment Failure in the PROSPECT-C Phase II Colourectal Cancer Clinical Trial. *Cancer Discov.* **8**, 1270–1285 (2018).

222. Sozzi, G. *et al.* Clinical Utility of a Plasma-Based miRNA Signature Classifier Within Computed Tomography Lung Cancer Screening: A Correlative MILD Trial Study. *J. Clin. Oncol.* **32**, 768–773 (2014).

223. Lesnik, E. A. & Freier, S. M. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. *Biochemistry* **34**, 10807–10815 (1995).

224. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).

225. Wen, G., Zhou, T. & Gu, W. The potential of using blood circular RNA as liquid biopsy biomarker for human diseases. *Protein Cell* **12**, 911–946 (2021).

226. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494.e15 (2017).

227. Chapuy, B. *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).

228. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* **378**, 1396–1407 (2018).

229. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011).

230. Avbelj, M. *et al.* Activation of lymphoma-associated MyD88 mutations via allostery-induced TIR-domain oligomerization. *Blood* **124**, 3896–3904 (2014).

231. Yusufova, N. *et al.* Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature* **589**, 299–305 (2021).

232. Davis, R. E. *et al.* Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463**, 88–92 (2010).

233. Pasqualucci, L. *et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471**, 189–195 (2011).

234. Lenz, G. *et al.* Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319**, 1676–1679 (2008).

235. Karube, K. *et al.* Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* **32**, 675–684 (2018).

236. Sebastián, E. *et al.* High-resolution copy number analysis of paired normal-tumor samples from diffuse large B cell lymphoma. *Ann. Hematol.* **95**, 253–262 (2016).

237. Kotsiou, E. *et al.* TNFRSF14 aberrations in follicular lymphoma increase clinically significant allogeneic T-cell responses. *Blood* **128**, 72–81 (2016).

238. Oeckinghaus, A. & Ghosh, S. The NF-κB Family of Transcription Factors and Its Regulation. *Cold Spring Harb. Perspect. Biol.* **1**, a000034 (2009).

239. Ye, B. H. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nat. Genet.* **16**, 161–170 (1997).

240. Phan, R. T. & Dalla-Favera, R. The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature* **432**, 635–639 (2004).

241. Ranuncolo, S. M., Polo, J. M. & Melnick, A. BCL6 represses CHEK1 and suppresses DNA damage pathways in normal and malignant B-cells. *Blood Cells. Mol. Dis.* **41**, 95–99 (2008).

242. Dang, C. V. *et al.* The c-Myc target gene network. *Semin. Cancer Biol.* **16**, 253–264 (2006).

243. Benhamou, D. *et al.* The c-Myc/miR17-92/PTEN Axis Tunes PI3K Activity to Control Expression of Recombination Activating Genes in Early B Cell Development. *Front. Immunol.* **9**, 2715 (2018).

244. Watanabe, T. *et al.* The MDM2 Oncogene Overexpression in Chronic Lymphocytic Leukemia and Low-Grade Lymphoma of B-Cell Origin. *Blood* **84**, 3158–3165 (1994).

245. Tsujimoto, Y., Finger, L. R., Yunis, J., Nowell, P. C. & Croce, C. M. Cloning of the Chromosome Breakpoint of Neoplastic B Cells with the t(14;18) Chromosome Translocation. *Science* **226**, 1097–1099 (1984).

246. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42**, 181–185 (2010).

247. Ying, C. Y. *et al.* MEF2B mutations lead to deregulated expression of the oncogene BCL6 in diffuse large B cell lymphoma. *Nat. Immunol.* **14**, 1084–1092 (2013).

248. Lacy, S. E. *et al.* Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report. *Blood* doi:10.1182/blood.2019003535.

249. Wright, G. W. *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568.e14 (2020).

250. Greenawalt, D. M. *et al.* Comparative analysis of primary versus relapse/refractory DLBCL identifies shifts in mutation spectrum. *Oncotarget* **8**, 99237–99244 (2017).

251. Morin, R. D. *et al.* Genetic Landscapes of Relapsed and Refractory Diffuse Large B-Cell Lymphomas. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **22**, 2290–2300 (2016).

252. Nesic, M. *et al.* The mutational profile of immune surveillance genes in diagnostic and refractory/relapsed DLBCLs. *BMC Cancer* **21**, 829 (2021).

253. Wise, J. F. *et al.* Mutational dynamics and immune evasion in diffuse large B-cell lymphoma explored in a relapse-enriched patient series. *Blood Adv.* **4**, 1859–1866 (2020).

254. Jiang, Y. *et al.* Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. *Genome Biol.* **15**, 432 (2014).

255. Rushton, C. K. *et al.* Genetic and evolutionary patterns of treatment resistance in relapsed B-cell lymphoma. *Blood Adv.* **4**, 2886–2898 (2020).

256. Mounier, N. *et al.* Rituximab plus CHOP (R-CHOP) overcomes bcl-2—associated resistance to chemotherapy in elderly patients with diffuse large B-cell lymphoma (DLBCL). *Blood* **101**, 4279–4284 (2003).

257. Rovira, J. *et al.* Prognosis of patients with diffuse large B cell lymphoma not reaching complete response or relapsing after frontline chemotherapy or immunochemotherapy. *Ann. Hematol.* **94**, 803–812 (2015).

258. Nijland, M. *et al.* Mutational Evolution in Relapsed Diffuse Large B-Cell Lymphoma. *Cancers* **10**, 459 (2018).

259. Mareschal, S. *et al.* Whole exome sequencing of relapsed/refractory patients expands the repertoire of somatic mutations in diffuse large B-cell lymphoma. *Genes. Chromosomes Cancer* **55**, 251–267 (2016).

260. Trinh, D. L. *et al.* Analysis of FOXO1 mutations in diffuse large B-cell lymphoma. *Blood* **121**, 3666–3674 (2013).

261. Xu-Monette, Z. Y. *et al.* Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. *Blood* **120**, 3986–3996 (2012).

262. Zenz, T. *et al.* TP53 mutation and survival in aggressive B cell lymphoma. *Int. J. Cancer* **141**, 1381–1388 (2017).

263. Melchardt, T. *et al.* Clonal evolution in relapsed and refractory diffuse large B-cell lymphoma is characterized by high dynamics of subclones. *Oncotarget* **7**, 51494–51502 (2016).

264. Chan, K. C. A. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin. Chem.* **59**, 211–224 (2013).

265. Cheng, H. *et al.* Analysis of ctDNA to predict prognosis and monitor treatment responses in metastatic pancreatic cancer patients. *Int. J. Cancer* **140**, 2344–2350 (2017).

266. Kurtz, D. M. *et al.* Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **36**, 2845–2853 (2018).

267. Thierry, A. R., El Messaoudi, S., Gahan, P. B., Anker, P. & Stroun, M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev.* **35**, 347–376 (2016).

268. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).

269. Rossi, D. *et al.* Diffuse large B-cell lymphoma genotyping on the liquid biopsy. *Blood* **129**, 1947–1957 (2017).

270. Chabon, J. J. *et al.* Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nat. Commun.* **7**, 11815 (2016).

271. Murtaza, M. *et al.* Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.* **6**, 8760 (2015).

272. Assouline, S. E. *et al.* Phase 2 study of panobinostat with or without rituximab in relapsed diffuse large B-cell lymphoma. *Blood* **128**, 185–194 (2016).

273. Cleary, K. L. S., Chan, H. T. C., James, S., Glennie, M. J. & Cragg, M. S. Antibody distance from the cell membrane regulates antibody effector mechanisms. *J. Immunol. Baltim. Md 1950* **198**, 3999–4011 (2017).

274. Alcaide, M., Rushton, C. & Morin, R. D. Ultrasensitive Detection of Circulating Tumor DNA in Lymphoma via Targeted Hybridization Capture and Deep Sequencing of Barcoded Libraries. *Methods Mol. Biol. Clifton NJ* **1956**, 383–435 (2019).

275. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

276. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

277. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

278. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

279. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

280. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972).

281. Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death Differ.* **25**, 154–160 (2018).

282. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

283. Ferrero, S. *et al.* KMT2D mutations and TP53 disruptions are poor prognostic biomarkers in mantle cell lymphoma receiving high-dose therapy: a FIL study. *Haematologica* (2019) doi:10.3324/haematol.2018.214056.

284. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B-Cell Lymphoma. *Cell* **171**, 481-494.e15 (2017).

285. Teeling, J. L. *et al.* The Biological Activity of Human CD20 Monoclonal Antibodies Is Linked to Unique Epitopes on CD20. *J. Immunol.* **177**, 362–371 (2006).

286. Xu, C. *et al.* NSCLC Driven by DDR2 Mutation Is Sensitive to Dasatinib and JQ1 Combination Therapy. *Mol. Cancer Ther.* **14**, 2382–2389 (2015).

287. Haowen, X. *et al.* Haploinsufficiency for NR3C1 Drives Glucocorticoid Resistance By Inactivation the PI3K/AKT/GSK3β/Bim Pathway in Adult Acute Lymphoblastic Leukemia. *Blood* **132**, 1328 (2018).

288. Regazzi, M. *et al.* Pharmacokinetic Behavior of Rituximab: A Study of Different Schedules of Administration for Heterogeneous Clinical Settings. *Ther. Drug Monit.* **27**, 785–792 (2005).

289. Sun, Y., Xia, P., Zhang, H., Liu, B. & Shi, Y. P53 is required for Doxorubicin-induced apoptosis via the TGF-beta signaling pathway in osteosarcoma-derived cells. *Am. J. Cancer Res.* **6**, 114–125 (2016).

290. Ungerleider, N. A. *et al.* Breast cancer survival predicted by TP53 mutation status differs markedly depending on treatment. *Breast Cancer Res. BCR* **20**, 115 (2018).

291. Xu, J. *et al.* Unequal prognostic potentials of p53 gain-of-function mutations in human cancers associate with drug-metabolizing activity. *Cell Death Dis.* **5**, e1108 (2014).

292. Zhang, J. *et al.* Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat. Med.* **21**, 1190–1198 (2015).

293. Ortega-Molina, A. *et al.* The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nat. Med.* **21**, 1199–1208 (2015).

294. Roschewski, M., Staudt, L. M. & Wilson, W. H. Dynamic monitoring of circulating tumor DNA in non-Hodgkin lymphoma. *Blood* **127**, 3127–3132 (2016).

295. Johnson, N. A. *et al.* CD20 mutations involving the rituximab epitope are rare in diffuse large B-cell lymphomas and are not a significant cause of R-CHOP failure. *Haematologica* **94**, 423–427 (2009).

296. Qunaj, L., Castillo, J. J. & Olszewski, A. J. Survival of patients with CD20-negative variants of large B-cell lymphoma: an analysis of the National Cancer Data Base. *Leuk. Lymphoma* **59**, 1375–1383 (2018).

297. Cucco, F. *et al.* Distinct genetic changes reveal evolutionary history and heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia* (2019) doi:10.1038/s41375-019-0691-6.

298. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2022. *CMAJ* **194**, E601–E607 (2022).

299. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375–2390 (2016).

300. Sehn, L. H. & Gascoyne, R. D. Diffuse large B-cell lymphoma: optimizing outcome in the context of clinical and biologic heterogeneity. *Blood* **125**, 22–32 (2015).

301. Al-Tourah, A. J. *et al.* Population-Based Analysis of Incidence and Outcome of Transformed Non-Hodgkin's Lymphoma. *J. Clin. Oncol.* **26**, 5165–5169 (2008).

302. Lossos, I. S. & Gascoyne, R. D. Transformation of follicular lymphoma. *Best Pract. Res. Clin. Haematol.* **24**, 147–163 (2011).

303. Carpio, C. *et al.* Avadomide monotherapy in relapsed/refractory DLBCL: safety, efficacy, and a predictive gene classifier. *Blood* **135**, 996–1007 (2020).

304. Eyre, T. A. *et al.* A phase II study to assess the safety and efficacy of the dual mTORC1/2 inhibitor vistusertib in relapsed, refractory DLBCL. *Hematol. Oncol.* **37**, 352–359 (2019).

305. Ansell, S. M. *et al.* Nivolumab for Relapsed/Refractory Diffuse Large B-Cell Lymphoma in Patients Ineligible for or Having Failed Autologous Transplantation: A Single-Arm, Phase II Study. *J. Clin. Oncol.* **37**, 481–489 (2019).

306. Hawkes, E. A. *et al.* Avelumab in Combination Regimens for Relapsed/Refractory DLBCL: Results from the Phase Ib JAVELIN DLBCL Study. *Target. Oncol.* **16**, 761–771 (2021).

307. Viardot, A. *et al.* Phase 2 study of the bispecific T-cell engager (BiTE) antibody blinatumomab in relapsed/refractory diffuse large B-cell lymphoma. *Blood* **127**, 1410–1416 (2016).

308. Coyle, L. *et al.* Open-Label, phase 2 study of blinatumomab as second salvage therapy in adults with relapsed/refractory aggressive B-cell non-Hodgkin lymphoma. *Leuk. Lymphoma* **61**, 2103–2112 (2020).

309. Locke, F. L. *et al.* Long-term safety and activity of axicabtagene ciloleucel in refractory large B-cell lymphoma (ZUMA-1): a single-arm, multicentre, phase 1–2 trial. *Lancet Oncol.* **20**, 31–42 (2019).

310. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

311. Morin, R. D. *et al.* Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).

312. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42**, 181–185 (2010).

313. Lee, J.-H., Jeong, H., Choi, J.-W., Oh, H. & Kim, Y.-S. Clinicopathologic significance of MYD88 L265P mutation in diffuse large B-cell lymphoma: a meta-analysis. *Sci. Rep.* **7**, 1785 (2017).

314. Stefancikova, L. *et al.* Prognostic impact of p53 aberrations for R-CHOP-treated patients with diffuse large B-cell lymphoma. *Int. J. Oncol.* **39**, 1413–1420 (2011).

315. Xu-Monette, Z. Y. *et al.* Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. *Blood* **120**, 3986–3996 (2012).

316. Juskevicius, D. *et al.* Distinct genetic evolution patterns of relapsing diffuse large B-cell lymphoma revealed by genome-wide copy number aberration and targeted sequencing analysis. *Leukemia* **30**, 2385–2395 (2016).

317. Sartorius, U. A. & Krammer, P. H. Upregulation of Bcl-2 is involved in the mediation of chemotherapy resistance in human small cell lung cancer cell lines. *Int. J. Cancer* **97**, 584–592 (2002).

318. *Picard Toolkit*. (Broad Institute, 2022).

319. Chen, Z. *et al.* Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.* **10**, 3501 (2020).

320. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).

321. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

322. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

323. Skidmore, Z. L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012–3014 (2016).

324. Hong, C., Thiele, R. & Feuerbach, L. GenomeTornadoPlot: a novel R package for CNV visualization and focality analysis. *Bioinformatics* **38**, 2036–2038 (2022).

325. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 1–13 (2016).

326. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**, 4164–4169 (2004).

327. Giovannetti, E. *et al.* Role of CYB5A in Pancreatic Cancer Prognosis and Autophagy Modulation. *JNCI J. Natl. Cancer Inst.* **106**, djt346 (2013).

328. Ennishi, D. *et al.* TMEM30A loss-of-function mutations drive lymphomagenesis and confer therapeutically exploitable vulnerability in B-cell lymphoma. *Nat. Med.* (2020) doi:10.1038/s41591-020-0757-z.

329. Dominguez, P. M. *et al.* TET2 Deficiency Causes Germinal Center Hyperplasia, Impairs Plasma Cell Differentiation, and Promotes B-cell Lymphomagenesis. *Cancer Discov.* **8**, 1632–1653 (2018).

330. Quivoron, C. *et al.* TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Is a Recurrent Event during Human Lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).

331. Rosikiewicz, W. *et al.* TET2 deficiency reprograms the germinal center B cell epigenome and silences genes linked to lymphomagenesis. *Sci. Adv.* **6**, eaay5872 (2020).

332. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).

333. Zhang, X., Hu, J., Chen, Q., Zhang, M. & Young, K. H. Decitabine Can Improve the Efficacy of Second-Line Chemotherapy in Relapsed or Refractory Diffuse Llarge B Cell Lymphoma. *Blood* **134**, 5221 (2019).

334. Aleem, E., Kiyokawa, H. & Kaldis, P. Cdc2–cyclin E complexes regulate the G1/S phase transition. *Nat. Cell Biol.* **7**, 831–836 (2005).

335. Kriegsman, B. A. *et al.* Frequent Loss of IRF2 in Cancers Leads to Immune Evasion through Decreased MHC Class I Antigen Presentation and Increased PD-L1 Expression. *J. Immunol.* **203**, 1999–2010 (2019).

336. Hagner, P. R. *et al.* CC-122, a pleiotropic pathway modifier, mimics an interferon response and has antitumor activity in DLBCL. *Blood* **126**, 779–789 (2015).

337. Krönke, J. *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301–305 (2014).

338. Michot, J.-M. *et al.* Avadomide plus obinutuzumab in patients with relapsed or refractory B-cell non-Hodgkin lymphoma (CC-122-NHL-001): a multicentre, dose escalation and expansion phase 1 study. *Lancet Haematol.* **7**, e649–e659 (2020).

339. Ribrag, V. *et al.* Phase Ib study of combinations of avadomide (CC-122), CC-223, CC-292, and rituximab in patients with relapsed/refractory diffuse large B-cell lymphoma. *eJHaem* **3**, 139–153 (2022).

340. Salles, G. *et al.* Tafasitamab plus lenalidomide in relapsed or refractory diffuse large B-cell lymphoma (L-MIND): a multicentre, prospective, single-arm, phase 2 study. *Lancet Oncol.* **21**, 978–988 (2020).

341. Kee, B. L., Quong, M. W. & Murre, C. E2A proteins: essential regulators at multiple stages of B-cell development. *Immunol. Rev.* **175**, 138–149 (2000).

342. Rowsey, R. A. *et al.* Characterization of TCF3 rearrangements in pediatric B-lymphoblastic leukemia/lymphoma by mate-pair sequencing (MPseq) identifies complex genomic rearrangements and a novel TCF3/TEF gene fusion. *Blood Cancer J.* **9**, 1–8 (2019).

343. Kubota-Tanaka, M. *et al.* B-lymphoblastic lymphoma with TCF3-PBX1 fusion gene. *Haematologica* **104**, e35–e37 (2019).

344. Yamazaki, T., Liu, L., Conlon, E. G. & Manley, J. L. Burkitt lymphoma-related TCF3 mutations alter TCF3 alternative splicing by disrupting hnRNPH1 binding. *RNA Biol.* **17**, 1383–1390 (2020).

345. Schmitz, R. *et al.* Burkitt Lymphoma Pathogenesis and Therapeutic Targets from Structural and Functional Genomics. *Nature* **490**, 116–120 (2012).

346. Grande, B. M. *et al.* Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood* **133**, 1313–1324 (2019).

347. Patel, D. & Chaudhary, J. Increased expression of bHLH transcription factor E2A (TCF3) in prostate cancer promotes proliferation and confers resistance to doxorubicin induced apoptosis. *Biochem. Biophys. Res. Commun.* **422**, 146–151 (2012).

348. Yugami, M., Kabe, Y., Yamaguchi, Y., Wada, T. & Handa, H. hnRNP-U enhances the expression of specific genes by stabilizing mRNA. *FEBS Lett.* **581**, 1–7 (2007).

349. Liao, B., Hu, Y. & Brewer, G. Competitive binding of AUF1 and TIAR to MYC mRNA controls its translation. *Nat. Struct. Mol. Biol.* **14**, 511–518 (2007).

350. Zhang, B. *et al.* The splicing regulatory factor hnRNPU is a novel transcriptional target of c-Myc in hepatocellular carcinoma. *FEBS Lett.* **595**, 68–84 (2021).

351. Shi, Z. *et al.* Targeting HNRNPU to overcome cisplatin resistance in bladder cancer. *Mol. Cancer* **21**, 37 (2022).

352. Wilke, A. C. *et al.* SHMT2 inhibition disrupts the TCF3 transcriptional survival program in Burkitt lymphoma. *Blood* **139**, 538–553 (2022).

353. Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *EMBO J.* **16**, 2004–2013 (1997).

354. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).

355. Roberts, A. W. *et al.* Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **374**, 311–322 (2016).

356. Davids, M. S. *et al.* Phase I First-in-Human Study of Venetoclax in Patients With Relapsed or Refractory Non-Hodgkin Lymphoma. *J. Clin. Oncol.* **35**, 826–833 (2017).

357. Hagner, P. R. *et al.* Interactome of Aiolos/Ikaros Reveals Combination Rationale of Cereblon Modulators with HDAC Inhibitors in DLBCL. *Clin. Cancer Res.* **28**, 3367–3377 (2022).

358. Puvvada, S. D. *et al.* A phase II study of belinostat (PXD101) in relapsed and refractory aggressive B-cell lymphomas: SWOG S0520. *Leuk. Lymphoma* **57**, 2359–2369 (2016).

359. Crump, M. *et al.* Phase II trial of oral vorinostat (suberoylanilide hydroxamic acid) in relapsed diffuse large-B-cell lymphoma. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **19**, 964–969 (2008).

360. Pan, H. *et al.* Epigenomic evolution in diffuse large B-cell lymphomas. *Nat. Commun.* **6**, 6921 (2015).

361. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

362. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).

363. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).

364. Legendre, C. *et al.* Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clin. Epigenetics* **7**, 100 (2015).

365. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).

366. Collas, P. The Current State of Chromatin Immunoprecipitation. *Mol. Biotechnol.* **45**, 87–100 (2010).

367. Bal, E. *et al.* Super-enhancer hypermutation alters oncogene expression in B cell lymphoma. *Nature* **607**, 808–815 (2022).

Appendix A.
Supplementary Data File associated with Chapter 2

Description

The accompanying Excel spreadsheet hosts several sets of additional results corresponding to Chapter 2, with each table corresponding to a separate tab in the spreadsheet.

List of supplemental tables

Table S1. Overview of relapsed-refractory DLBCL cohort and patient information. A sample with at least one detectable somatic mutation is considered having detectable tumour DNA.

Table S2. Patient-specific metadata and breakdown of rrDLBCL samples. This includes cell-of-origin classification, LymphGen classification, and if a constitutional sample was available for that patient.

Table S3. Summary of samples with a source of tumor DNA at diagnosis and upon relapse.

Table S4. Gene panel capture space of 63 lymphoma-associated genes. All coordinates are relative to GRCh38. Note that additional probes were included for non-coding regions and specific exons of additional genes not included in this analysis due to variable coverage in exome data.

Table S5. rrDLBCL somatic variant calls, in MAF format. Note that this includes variant calls outside the capture space specified in Table S4.

Table S6. List of mutation hotspots/regions within the capture space examined for mutation enrichment. All coordinates are relative to GRCh38.

Table S7. Differentially mutated genes between the rrDLBCL cohort and untreated cohorts. Comparison was performed using a fisher's exact test and

Benjamin-Hochberg false discovery rate threshold, with all genes harbouring a Q value below 0.1 considered significant. Odds ratios were scaled using $\log_e$. See 2.3.1 and 2.3.2 for a description of the untreated and relapse cohorts. This is summarized in Figure 2-3.

Table S8. Differentially mutated genes between the rrDLBCL cohort and diagnostic DLBCL cases listed by Lacy *et al*. Comparison was performed using a fisher's exact test and Benjamin-Hochberg false discovery rate threshold, with all genes harbouring a Q value below 0.1 considered significant. Odds ratios were scaled using $\log_e$. See methods for a description of the relapse cohort.

Table S9: Summary of genetic subgroups within the rrDLBCL cohort using the Wright classifier, and their prevalence compared to the untreated cohort.

Table S10. Cox proportional hazard models for *KM2D* and *TP53* mutations, in the context of other prognostic covariates. Feature importance for patient OS (A,C) and PFS (B,D) within our untreated DLBCL cohort, examining *TP53* mutation status, *KMT2D* mutation status (A,B) or *KMT2D* truncating mutation status (C,D) along with the International Prognostic Index (IPI) stage and COO subgroup. Cases lacking IPI and COO information, or with insufficient coverage in *KMT2D* or *TP53* were excluded from analysis. Cohort was included as a feature to ensure cases from a given cohort did not display inferior outcomes. PFS information was not provided for samples from the Reddy cohort.

Table S11. Primer sequences used to introduce mutations within *MS4A1*.

Table S12. Primers used for Access Array amplicon sequencing of PT255. Forward (F) and reverse (R) primers for each gene used in amplicon sequencing experiments. Primers were tailed with Illumina sequence adapters.

Table S13. Variants used for PT255 analysis from amplicon sequencing results.

Filename:
Rushton_AppendixA.xlsx

Appendix B.
Supplementary Data File Associated with Chapter 3

Description

The accompanying Excel spreadsheet hosts several sets of additional results corresponding to Chapter 3, with each table corresponding to a separate tab in the spreadsheet.

List of supplemental tables

Table S1. Overview of rrDLBCL cohort, including patient characteristics and sequencing type. Molecular and genetic subgroup labels are also included.

Table S2. Exome-wide somatic variant calls of rrDLBCL samples with WES or WGS data, in MAF format. For WGS samples, this will include non-coding variants outside the traditional “exome” capture space. Coordinates are relative to the hg38 reference genome

Table S3. Somatic copy number segments from all rrDLBCL samples (WES, WGS, lpWGS). Copy number state is specified as log(2) rations. Segments are relative to the hg38 reference genome. Note some samples have somatic variant calls but lack CNV calls due to low sample quality.

Table S4. Regions significantly recurrently perturbed by CNVs within the rrDLBCL cohort, as identified using GISTIC2.

Table S5. Genes whose mutations show evidence of positive selection across the rrDLBCL SNV cohort and diagnostic cohorts, as determined via OncoDriveFML.

Table S6. Genes significantly definitely perturbed for mutations between diagnostic and rrDLBCL cohorts. Only genes with evidence of positive selection were considered for analysis.

Filename:
Rushton_AppendixB.xlsx