Elucidating Targetable Genetic Vulnerabilities in Relapsed/Refractory Diffuse Large B-cell Lymphoma

by

Audrey Li

B.S. Computer Science and Engineering, MIT, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Audrey Li Department of Electrical Engineering and Computer Science May 10, 2024
Certified by:	Gad Getz Professor of Pathology, Thesis Supervisor
Certified by:	Chris Love Professor of Chemical Engineering, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Elucidating Targetable Genetic Vulnerabilities in Relapsed/Refractory Diffuse Large B-cell Lymphoma

by

Audrey Li

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Diffuse large B-cell lymphoma (DLBCL), the most prevalent form of non-Hodgkin lymphoma is marked by significant heterogeneity in its morphology, genetic irregularities, and clinical behavior. Current prognostic tools, including the International Prognostic Index and cell-of-origin transcriptional classifications such as germinal center B-cell-like and activated B-cell-like, do not adequately reflect DLBCLs complex nature. Front-line standard of care treatment predominantly consists of a regimen with cyclophosphamide, doxorubicin, prednisone, rituximab, and vincristine (R-CHOP); however, the relapse rate remains high, underscoring the need for improved diagnostic and therapeutic methods.

In this comprehensive analysis, we investigated the genetic substructure of DLBCL in both newly diagnosed and relapsed/refractory cases, focusing on genetic abnormalities pertinent to relapsed settings and the immune microenvironment's influence on therapy response. Our findings revealed significant enrichment of specific genetic clusters, notably clusters 2 and cluster 5, which are associated with an inferior prognosis and high relapse rates following R-CHOP therapy. These clusters were characterized by distinct genetic alterations, including prevalent mutations in *TP*53, *BCL*2, and *MYD*88. The results of this study suggest that integrating detailed genetic profiling into the clinical management of DLBCL could significantly refine therapeutic approaches, tailoring them to the unique genetic backdrop of each patient's disease. This approach promises to enhance the precision of prognostic assessments and the efficacy of subsequent therapeutic interventions, paving the way for personalized medicine in the treatment of DLBCL.

Thesis supervisor: Gad Getz

Title: Professor of Pathology

Thesis supervisor: Chris Love

Title: Professor of Chemical Engineering

Acknowledgments

I would like to express my profound gratitude to several individuals, lab groups, and institutions whose expertise and support were indispensable in the completion of this research project.

I wish to express my deepest appreciation to Dr. Gad Getz, Dr. Chip Stewart, and the Getz Lab for their tremendous mentoring, guidance, and support throughout the entirety of this project. Drs. Getz and Stewart's mentorship not only shaped the course of this research but also contributed significantly to my professional growth and understanding of genomic science. I am also immensely grateful to Dr. Margaret Shipp, Dr. Eleonora Calabretta, and the Shipp Lab for their invaluable clinical insights and annotations that enriched our understanding of the data. Eleonora's meticulous efforts in creating detailed graphs and tables provided clarity and allowed us to present our data in a meaningful and impactful manner. I would also like to thank Dr. Chris Love for being a co-supervisor on my thesis. His support and expertise are crucial in the course of my thesis work.

The data in this study was made possible through a collaboration between multiple esteemed institutions, including Dana-Farber Cancer Institute, Brigham and Women's Hospital, Massachusetts General Hospital, University of Rochester Medical Center, the City of Hope Cancer Center, and the Broad Institute. I am thankful for the clinically annotated cohort of formalin-fixed paraffin-embedded (FFPE) tissue samples provided by these institutions.

The contributions of all these individuals and institutions were crucial to the success of this research, and I am deeply thankful for their support and collaboration. Their collective expertise and dedication have left a lasting impact on this study and on my personal and professional development.

Contents

List of Figures

Chapter 1

Introduction

Diffuse large B-cell lymphoma (DLBCL) is recognized as the predominant form of non-Hodgkin lymphoma (NHL), making up 30% to 40% of 545,000 yearly NHL diagnoses globally [1]. DLBCL exhibits a high degree of heterogeneity, evident through its diverse morphology, immunohistochemical profiles, genetic irregularities, and clinical characteristics. Traditional prognostic methods, such as the clinical International Prognostic Index (IPI) or the transcriptional cell-of-origin (germinal center B-cell-like [GCB] and activated B-cell-like [ABC] classifications), fall short in reflecting the intricate nature of this illness [2].

DLBCL patients are commonly treated with a regimen that combines an anti-CD20 monoclonal antibody, rituximab, with a combination chemotherapy regimen, cyclophosphamide, adriamycin, vincristine and prednisone (R-CHOP). While this treatment approach leads to a cure in 60-70% of cases, there remains a significant portion of patients who relapse [3]. Addressing relapsed DLBCL continues to be a critical medical need, underscoring the necessity for advancements in therapeutic strategies, where novel therapeutic agents are primarily assessed within the context of relapsed or refractory disease. Unfortunately, the current approach of testing novel targeted agents in patients defined by clinical prognostic categories, or tumor transcriptional subgroups has led to multiple negative phase III trials [4].

Currently, the medical field lacks established, data-driven guidelines for selecting traditional and emerging treatments for DLBCL, often applying these therapies to a broad patient base or groups identified only by clinical prognostic indicators or gene expression categories. This work aims to address these pain points with recently identified genetic markers specific to DLBCL to facilitate a more tailored and effective approach to choosing initial and secondary treatments for affected individuals. We will characterize the genetic substructure of both newly diagnosed and relapsed/refractory (R/R) DLBCL and analyze paired diagnostic and R/R biopsy specimens from a large clinically annotated patient cohort. More specifically, the work aims to: 1) identify specific genetic abnormalities and complementary pathways for targeted therapy in relapsed settings; and 2) identify potential genetic bases of immune

evasion that may inform our use of immune-based therapies.

1 Background

With DLBCL's varied clinical and biological profile, while roughly 65% of DLBCL patients can achieve remission with primary induction treatments, the rest face poor prospects, accentuating the urgency for novel therapeutic options [3]. The standard practice of administering new targeted therapies to a general patient population or those identified by clinical prognostic factors or transcriptional signatures, such as germinal center B-cell (GCB) and activated B cell (ABC), are often made without the benefit of genetic data [5], [6]. Furthermore, this approach has seen various promising agents, such as obinutuzumab, bortezomib, ibrutinib, and lenalidomide, not meet their therapeutic endpoints in randomized clinical trials [4]. The need to incorporate genetic insights into treatment selection remains a pivotal aspect of improving outcomes for this patient population.

R/R DLBCL generally carries a poor prognosis, yet the outcomes vary significantly among patients. Three distinct categories of recurrence have been identified, each with its own expected prognosis. The first category of primary refractory patients fails to respond to initial R-CHOP treatment and is unlikely to benefit from any additional current therapies. The second category includes those patients who relapse within 12 months of treatment; these individuals typically do not achieve long-lasting responses to salvage therapy. Lastly, the third category encompasses patients who relapse more than 12 months after treatment. This group tends to have a more favorable outlook, with the disease responding to chemotherapy, leading to a cure rate of approximately 30% when treated with standard chemo-immunotherapy [7].

Through analyzing recurrent mutations, somatic copy number alterations (SCNAs), and structural variants (SVs) in newly diagnosed DLBCLs, five distinct DLBCL categories (i.e. clusters) were identified as marked by unique genetic patterns [8]. We observed significant differences in PFS among patients with these genetically distinct transcriptional types, with notably higher relapse rates in Cluster 2 (C2) and Cluster 5 (C5) tumors and unfavorable trajectory in Cluster 3 (C3) tumors. Of interest, the previously described cell-of-origin (COO) transcriptional subclassification was insufficient to identify these poor-prognosis DLBCLs. A complete assessment, encompassing mutations, SCNAs, and SVs, is necessary to better understand the substructure and prognostic differences in DLBCL, highlighting the need for a broad-based evaluation method.

2 Related Work

Clinically, for patients with primary refractory DLBCL, the odds of a positive response to any salvage treatment, including intense myeloablative therapy and autologous stem cell transplantation (ASCT), are low. In a similar vein, for those patients whose disease re-emerges

within six to 12 months post-induction therapy, the likelihood of a successful outcome with existing salvage treatments is notably diminished. Patients with DLBCL recurrence occurring more than twelve months following induction therapy have a 30% chance of achieving lasting remission with the current salvage therapies, such as ASCT [9].

Recent research has identified common mutations, somatic copy number alterations (SCNAs), and structural variants (SVs) in newly diagnosed DLBCL, leading to the identification of five genetic distinct subgroups (Clusters 1-5, [C1-5]) [8]. These subgroups included: 1) a high-risk ABC-enriched DLBCL category with prevalent 18q arm *BCL2* gene copy gains and $MYD88^{L265P}$ and *CD79B* mutations (C5); 2) a lower-risk ABC-enriched DLBCL group with genetic changes similar to those in marginal zone lymphomas and modifications in immune escape mechanisms (C1); 3) a high-risk GCB-enriched DLBCL subset with *BCL2* structural variants and mutations, *PTEN* mutations and/or copy losses and changes in epigenetic modifying enzymes (C3); 4) another lower-risk GCB-enriched DLBCL group with changes in the *JAK/STAT* and *BRAF* pathways and various histones (C4); and 5) a group that is not enriched in GCB or ABC tumors with frequent loss of *TP53* function, $9p21.3/CDKN2A$ deletion, and associated genomic instability (C2). These genetic subsets have been associated with differing patient outcomes following R-CHOP, with groups C2, C3, and C5 showing significantly poorer progression-free survival (PFS). The findings from our study indicate a higher propensity for relapse in DLBCLs categorized within clusters 2, 3, and 5 after initial treatment with R-CHOP therapy [8]. These DLBCLs, identified as high-risk due to their molecular characteristics, exhibit specific genetic vulnerabilities that can be targeted therapeutically, indicating the potential for tailored combination treatments.

Among the different types of DLBCL relapses, the central nervous system (CNS) relapse is a rare but devastating event, affecting approximately 5% of patients and leading to survival that often spans only a few months [10]. The clinical CNS International Prognostic Index (CNS-IPI) is the most commonly employed method for identifying high-risk patients, yet it has a low positive predictive value. Additionally, patients with the ABC or non-GCB transcriptional subtypes of DLBCL are found to be at an increased risk of CNS relapse. Despite stratifying patients based on the CNS-IPI for prophylactic treatment, the administration of intravenous or intrathecal chemotherapy has not been effective in preventing CNS relapses [11].

In a recent study conducted by Hilton et al., 129 patients with multiple biopsies were taken for genetic assessment, including transformed lymphoma, accounting for approximately 25% of cases [12]. Biopsies were examined using a combination of break-apart FISH for MYC, BCL2, and/or BCL6 rearrangements, along with NanoString DLBCL90 for COO [13] and dark zone signature (DZsig) classification [14], and either whole-genome (WGS) or whole-exome sequencing (WES). In cases of primary refractory disease, the observed pattern of tumor evolution indicated that innate chemoresistance was present from the time of diagnosis and that the composition of mutations remained relatively unchanged during treatment. For

late relapses, the analysis pointed to the existence of persistent clonal precursor cell (CPC) populations that are responsible for multiple DLBCL manifestations over time, although the genetics-based classification remained the same, indicating that early CPC mutations play a significant role in the biology of the disease. However, the study has limitations, including the absence of copy number alteration (CNA) analysis and a lack of detailed mutational patterns associated with the type of relapse and CNS relapses [12].

Recently, a variety of targeted and immunotherapy-based medications have received approval for clinical application in R/R DLBCL, yielding promising outcomes. These novel agents are currently being evaluated in a broad patient pool or within subsets delineated by clinical prognostic factors or tumor gene expression profiles. However, there is a pressing need for reliable predictive instruments that can accurately determine the most suitable salvage therapy tailored for individual patients, ensuring optimized treatment efficacy [4].

Our previous analyses revealed the need to define recurrent mutations, SCNAs, and SVs to fully understand the substructure and prognostic differences in DLBCL, underscoring the importance of a broad-based evaluation method. In this study, we analyze the comprehensive genetic signatures and associated subtypes of relapsed DLBCL to develop a rational, personalized treatment approach for patients in need.

Chapter 2

Methods

Our work has two central aims: 1) define the genetic signatures of R/R DLBCL and assess differences with newly diagnosed DLBCLs; and 2) identify potential targetable vulnerabilities associated with specific comprehensive genetic signatures.

Concerning these two objectives, we analyzed the genetic signatures of R/R DLBCLs. In addition, through investigating the CNS DLBCL relapse patients, we hoped to discover whether they are more likely to be identified with the C5 molecular signature, which has been associated with extranodal tropism. We also assessed whether patients with primary refractory, early relapsed, or late relapsed DLBCL have additional defining genetic abnormalities and associated therapeutically-targetable vulnerabilities.

More specifically, we conducted the following steps in the process of developing a robust predictive tool to deliver the most appropriate salvage therapy to each patient:

1. Whole exome sequencing, DNA quality control, and analysis of quality control outputs.
2. Select samples to be included in the prediction pipeline based on tumor purity and contamination level.
3. Mutation detecting and discovery of cancer genes (MutSig2CV analysis).
4. Copy number variation (CNV) detection and analysis, including GISTIC2.0 analyses.
5. Structural variant (SV) detection and analysis.
6. Data analysis, including the application of the molecular classifier with consensus clustering and interpretation to answer the clinical questions.

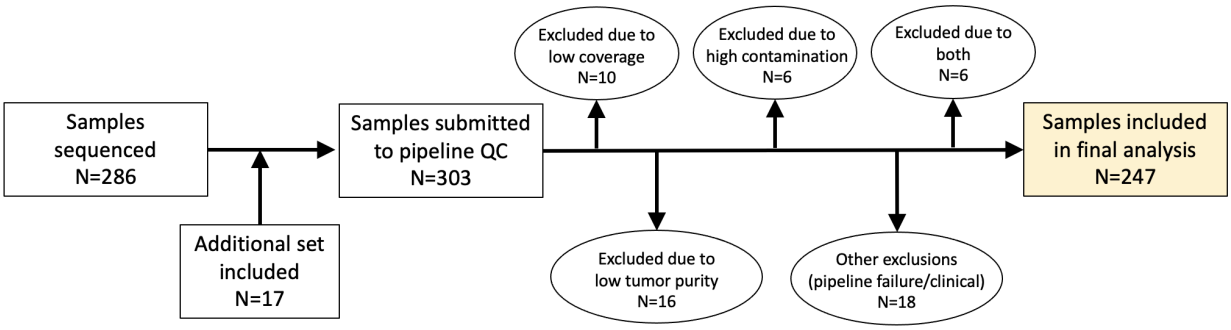

Figure 2.1: Flow chart detailing sample processing for exome+ sequencing

3 Data

As a collaboration effort between Dana-Farber Cancer Institute, Brigham and Women's Hospital, Massachusetts General Hospital, University of Rochester Medical Center, and City of Hope Cancer Center, a clinically annotated cohort of formalin-fixed paraffin-embedded (FFPE) tissue samples from patients with R/R DLBCL was assembled for exome+ sequencing (i.e., sequencing of the whole-exome and a set of genomic loci tailored for detection of SVs in lymphomas). As shown from the flow chart (Figure 2.1), the dataset accounts for a total of 286 cases originally sequenced. Another set of previously sequenced cohorts was also added to augment the dataset. The entire set of 303 samples then passed through QC. Through the QC process, exclusions were made based on specific criteria: low coverage (10 samples excluded), high contamination (6 samples excluded), and a combination of both factors (6 samples excluded). Further exclusions were made for low tumor purity, which accounted for 16 samples, and additional reasons, such as pipeline failure or clinical factors, led to 18 more exclusions. Following these rigorous quality checks, a total of 247 samples are deemed suitable for the final analysis, providing a robust dataset for assessing genomic landscapes and potentially guiding therapeutic strategies for R/R DLBCL.

From the 247 samples included in the final analysis, a subset was established comprising 217 cases specifically identified as the R/R DLBCL cohort, and the rest of the 30 samples were deemed as the CAR-T cohort. Within the R/R DLBCL cohort, the study further categorized the samples: 45 consisted of pairs of diagnostic and first relapse biopsies. Of these, 34 were systemic relapses, and 11 were central nervous system (CNS) relapses. These paired biopsies help in understanding the extent to which potential baseline genetic alterations determine therapy response, monitoring clonal and subclonal evolution of pre-existing genetic alterations, and detecting the occurrence of new genetic defects. Additionally, there were 18 biopsies from cases diagnosed as primary refractory to treatment. The dataset also included 130 biopsies from first relapses, which encompassed both paired and unpaired samples with respect to the initial diagnostic specimens. These relapse biopsies broke down further into

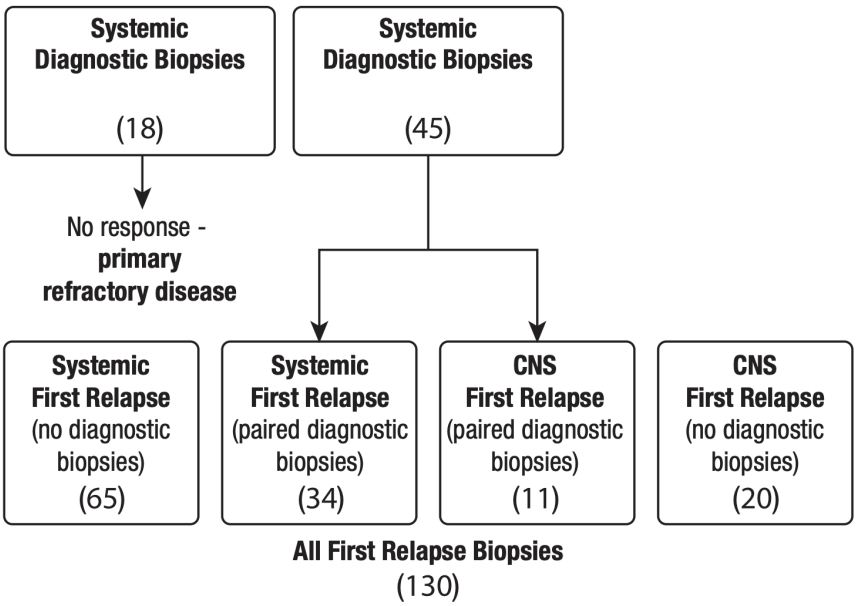

Figure 2.2: Diagram detailing the division of some subsets of samples within our study.

99 systemic relapses and 31 CNS relapses. Moreover, the relapses were temporally classified, with 65 characterized as early relapses occurring within 12 months of diagnosis and another 65 identified as late relapses happening after 12 months. This detailed stratification underscores the complexity of R/R DLBCL and the potential for this dataset to contribute to a deeper understanding of disease progression and resistance.

For reference, we enumerate the data breakdown into subsets below.

1. All cases in the "relapsed DLBCL" series (n=217)
2. All cases with paired diagnostic and first relapse specimens
    2.1. All diagnostic specimens (with paired first relapse samples) (n=45)
    2.2. All first relapse specimens (with paired diagnostic samples) (n=45)
    2.3. Diagnostic specimens (with paired systemic first relapse samples) (n=34)
    2.4. Systemic first relapse specimens (with paired diagnostic samples) (n=34)
    2.5. Diagnostic specimens (with paired CNS first relapse) (n=11)
    2.6. CNS first relapse (with paired diagnostic samples) (n=11)
3. Primary refractory samples
    3.1. Diagnostic samples (n=18)
    3.2. Diagnostic samples (with paired first progression specimens) (n=4)

3.3. First progression specimens (all with paired diagnostic samples) (n=4)

4. All first relapse samples (systemic and CNS first relapse samples with and without a paired diagnostic specimen) (n=128)
5. All first systemic relapse samples (includes those with and without a paired diagnostic specimen) (n=99)
6. All CNS first relapse samples (includes those with and without a paired diagnostic specimen) (n=31)
7. All first relapse samples within 12 months (excludes primary refractory diagnostic specimens [Subset 3.1]) (n=65)
8. All first relapse samples within 12 months (includes primary refractory diagnostic specimens [Subset 3.1]) (n=83)
9. All first relapse samples after 12 months (n=65)

4 Methods

4.1 Exome sequencing, alignment, and DNA quality control

To comprehensively analyze genetic defects, whole exome sequencing (WES) and DNA quality control were performed using an extended bait set that captures SVs (including translocations). The WES targets (n= 204,360 targets covering 34.8Mb of exons) were augmented by 422 targets covering 866 kB designed to capture known SV breakpoints in non-Hodgkins Lymphoma. We processed the sequence data with the Broad Institute's Genomic Platform exome alignment pipeline. For each sample, the pipeline produced a single Binary Alignment Map (BAM) file by combining data from multiple libraries and flow cell runs. The output is also coupled with the corresponding BAM index file and Picard metrics files.

4.2 Significance analysis of recurrent SCNAs using GISTIC

We identified arm-level and focal peaks of recurrent copy number alterations using GISTIC2.0 with segmented copy-number data generated by HapASeg (Priebe et al., manuscript in preparation), as input [15]. Events with a q-value of less than 0.1 were reported to be significant.

4.3 Contamination

The amount of cross-individual contamination for tumor pairs was estimated by the ContEst tool from the Broad Institute's Genome Analysis Toolkit [16].

4.4 Mutation detection

Somatic SNVs, small insertions, and deletions (indels) were identified using MuTect1.0 and Strelka pipelines, respectively [17], [18]. We used a suitable panel of normal (PoN)from unrelated healthy controls that met quality control standards to 1) remove common germline events and 2) remove sequencing artifacts. Mutation annotation was carried out using the Oncotator tool [19].

4.5 Artifact filtering

OxoG is an artifact signature that arises due to oxidative damage to guanine in the library preparation phase. This damage prompts guanine to pair with adenine instead of cytosine incorrectly, leading to a G>T mutation in the observed sequence. These artifacts are strand-specific, in contrast to somatic mutations that affect both DNA strands. This distinct strand bias is a key indicator to differentiate between true genetic events and OxoG artifacts. Additionally, the cohort under study exhibits single nucleotide errors linked to using Formalin-Fixed Paraffin-Embedded (FFPE) samples. The formaldehyde in these samples triggers cytosine deamination, resulting in C>T mutations. These are akin to mutations seen with aging but exhibit the same strand bias as OxoG. This similarity in strand bias enables the application of established algorithms for orientation bias detection, previously applied to FFPE samples, to identify these artifact mutations [20].

4.6 Copy number analysis

We used two methods to determine the allele-specific copy numbers for the samples. One is Allelic Capseg [21], and more recently, HapASeg has been developed to address some Allelic CapSeg limitations. HapASeg emphasizes the measurement of allelic imbalance of phased heterozygous germline variants. It has been shown to improve copy number estimates from noisy whole genome sequencing data without requiring a PoN (Priebe et al., manuscript in preparation).

4.7 Estimation of purity, ploidy, and cancer cell fraction (CCF) using ABSOLUTE

The ABSOLUTE algorithm was used to obtain purity, ploidy, and CCF estimates for mutations and copy numbers [22]. ABSOLUTE requires manual review by trained reviewers. Due to the prevalence of heterozygous germline variants in the tumor-only samples as inputs to ABSOLUTE, the process for determining solutions was more reliant on the copy ratio profile than mutation allele frequency distributions.

4.8 Germline somatic log-odds filter for tumor-only samples: Tonly2

For every genetic variation that passed all preceding filters, SNV or indel, key factors such as cancer cell fractions (CCFs) of somatic events, tumor purity, ploidy, and copy number variations (CNVs), were analyzed. These parameters were also used in determining the log ratio of the probability that its allele fraction is consistent with the allele fraction modeled for a hypothetical germline event and the probability it is consistent with a modeled somatic event. SNVs that are more likely to be germline variants are removed by Tonly2 (Chu et al., manuscript in preparation).

4.9 Significance analysis of recurrently mutated genes: MutSig2CV

We applied the MutSig2CV algorithm to discover the genes that exhibited significant mutations, with those having a q-value below 0.1 being classified as significant [23]. Significant genes were identified in different data subsets (including paired tumor/normal and tumor-only subsets), suggesting that any germline mutations remaining after this pipeline are most likely randomly distributed throughout the genome and unlikely to affect the significantly mutated genes detected by MutSig2CV.

4.10 Consensus clustering of genetic alterations

Generation of the gene-by-sample matrix

We used a reference set of data of 699 DLBCL samples combined from Chapuy et al. (2018) [8] and Schmitz et al. (2018) [24] (referred to as the DLBclass cohort hereafter).

Creation of the gene-by-sample matrix involved compiling previously identified significantly mutated genes (identified by MutSig2CV and the clustering and visualization of mutations in protein structures (CLUMPS) [25] with a Benjamini-Hochberg false discovery rate (FDR) q-value ≤ 0.1 and a frequency $\geq 3\%$) from the DLBclass cohort, significant somatic copy number alterations (SCNAs identified by GISTIC2.0 with a q-value 0.1 and a frequency $\geq 3\%$), and chromosomal rearrangements (frequency $\geq 3\%$) from Chapuy et al. (2018) into a comprehensive matrix (encoding non-synonymous mutations as 2, synonymous mutations as 1, no mutation as 0, high-grade copy ratio (CN) gain [CN $\geq$ 3.7 copies] as 2, low-grade CN gain [3.7 copies $\geq$ CN $\geq$ 2.2 copies] as 1, CN neutral as 0, low-grade CN loss [1.1 $\leq$ CN $\leq$ 1.6 copies] as 1, high-grade CN loss [CN $\leq$ 1.1 copies] as 2, presence of chromosomal rearrangement as 3, absence as 0, and not assessed chromosomal rearrangements as N/A) [8].

Non-negative matrix factorization consensus clustering

In an effort to reliably determine tumors possessing common genetic characteristics, we employed a modified non-negative matrix consensus clustering approach as shown in Chapuy et al. (2018) [8]. This process involved submitting the gene-by-sample matrix, which included

data on SNVs, CNVS, and SVs, to the NMF consensus clustering algorithm without subjecting the matrix to normalization. The best cluster number was identified to be "$k = 5$". Furthermore, we analyzed marker genes linked to each cluster through a Fisher test, and the p-values obtained were then adjusted for the FDR.

4.11 Data analysis and interpretation

The data analysis process was separated into two steps. Initially, DLBCL classification based on the Chapuy et al. (2018) classification logic was applied to the relapsed DLBCL samples [8].

After performing consensus clustering, additional genetic anomalies were identified and examined for their clinical relevance. In particular, through the comparative analysis of diagnostic and relapsed biopsy pairs, we uncovered new defining genetic lesions or existing sub-clones emerging from treatment-specific selective pressure, which could shed light on resistance to first-line treatments, timing of relapse, and response to salvage treatment. The subset of patients with primary refractory DLBCL is especially critical for identifying genetic changes in those with the least likelihood of cure. Furthermore, studying the genetically annotated cohort treated with CAR-T cell therapy could reveal molecular mechanisms of immune evasion that undermine the efficacy of novel, approved immunotherapeutic drugs.

4.12 Validation of the molecular classifier in R/R DLBCL

We developed a molecular classifier based on the recurrent genetic alterations in newly diagnosed DLBCLs to identify C1-C5 tumors (manuscript in preparation) prospectively. The efficacy of our new molecular classifier is evaluated to determine its predictive capacity for the relapse setting. This classifier will be refined and optimized, as needed, to include genetic variations and newly recognized genetic alterations linked to treatment resistance.

Chapter 3

Results

5 Quality control results

5.1 Sample filtering

Sample filtering was performed based on two criteria: coverage and contamination. For coverage, samples with $<$ 10x sequencing coverage cannot provide valuable information; tumor purity is needed to determine the quality of information arising from samples with coverage between 10x and 50x. Samples with contamination, defined as the fraction of DNA from someone other than the patient, have a higher risk of reporting germline variants in the contaminating DNA as somatic, and hence, mutations consistent with the contamination level are removed. As a result, tumors with $>$ 5% contamination have a significant limit on sensitivity to mutations with low variant allele fractions (VAF), which are typically subclonal mutations. The median contamination for this cohort was 0.15% (interquartile range 0.036-0.5%). A total of 14 samples were omitted due to the above concerns, with most having a coverage lower than 35x and/or contamination greater than 10%).

5.2 Target coverage and insert size

As Figure 3.1a shows, the cohort has a large range of target coverage with a median coverage of 198x. Roughly 85% of samples have coverage $>$ 100x. Due to the nature of using FFPE samples, the insert size, also known as the fragment length or number of base pairs (bp) between read pair ends, is shortened to a median of around 120 bp (Figure 3.1b); in particular, we note 17 of the 286 cases have insert size $<$ 100 bp. Short fragments yield less sensitivity for detecting SVs. It also results in analysis limitations arising from possible mismappings and adapter leakage into the sequenced reads. Longer fragments tend to improve the alignments.

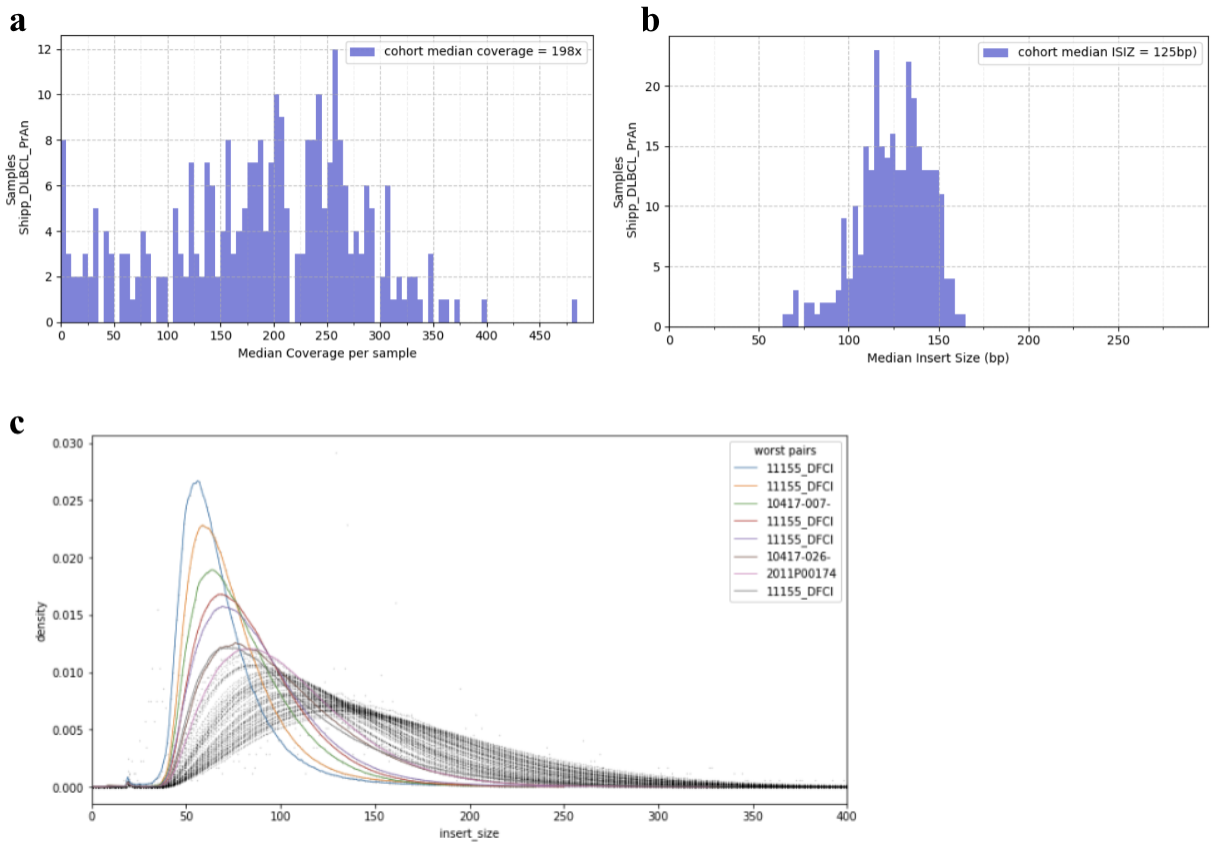

Figure 3.1

(a) Distribution of median coverage values of the samples in the cohort with the specific coverage level on the horizontal axis and number of corresponding samples on the vertical axis.

(b) Distribution of median insert size in base pairs of the samples in the cohort with insert size on the horizontal axis and number of corresponding samples on the vertical axis.

(c) Graph showing insert size of the samples corresponding to the density of each, with samples having the shortest insert sizes colored according to the panel on the right.

5.3 OxoG damage

The OxoG Q score is defined as the PHRED quality score for the difference in "PRO" and "CON" pair orientation error rates. A Q score of less than 30 indicates that the damage from OxoG is a dominant mode of mutation artifact, which corresponds to an artifactual excess of C>A false mutations on one strand and no G>T variants on the other strand. Figure 3.2a shows that the cohort is centered around a Q score of 40, with a dribble down to 30. We then manually investigated the BAM files of those samples with a Q score of around 30 in the Integrative Genomics Viewer (IGV), but no obvious OxoG damage was observed [26].

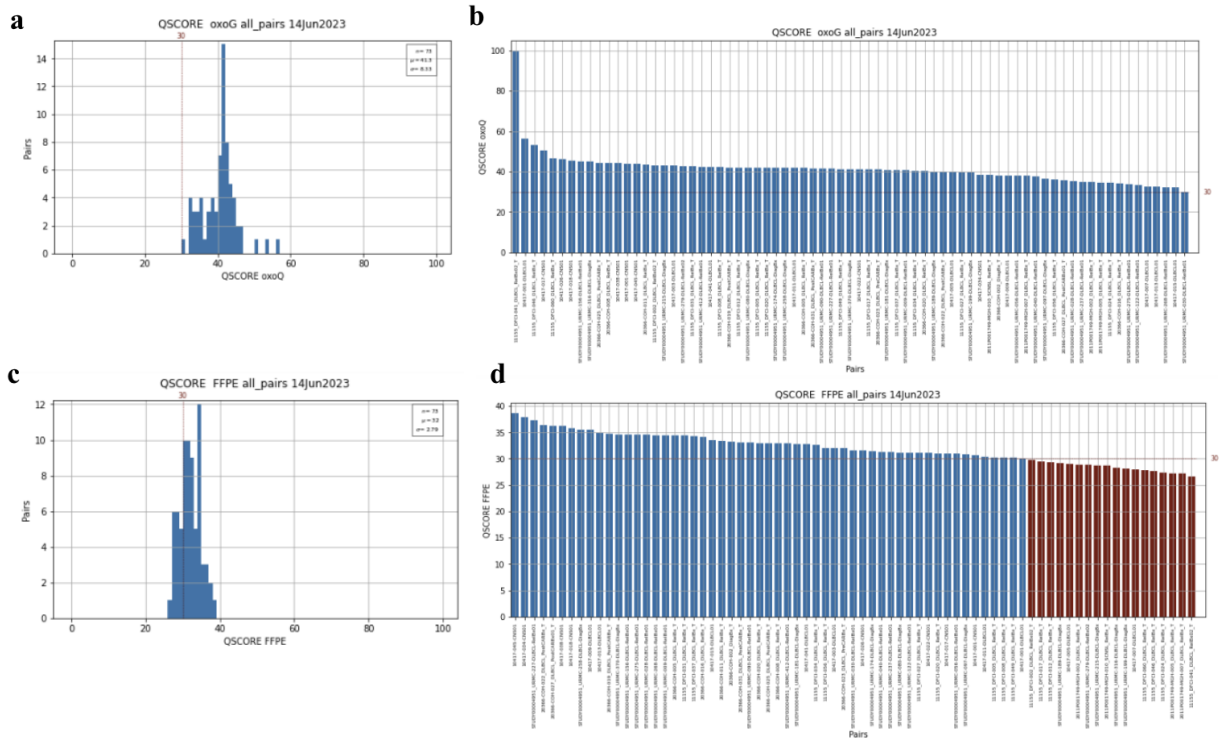

Figure 3.2

(a) OxoG Q score distribution of the samples in the cohort with the score on the horizontal axis and number of corresponding samples on the vertical axis.

(b) Detailed OxoG Q scores bar graph for each sample.

(c) FFPE damage Q score distribution of the samples in the cohort with the score on the horizontal axis and number of corresponding samples on the vertical axis.

(d) Detailed FFPE damage Q scores bar graph for each sample.

5.4 FFPE damage

We then calculated the deamination Q scores, which is defined in the same convention as the Oxo Q scores (the PHRED quality score for the difference in "PRO" and "CON" pair orientation error rates). A Q score of lower than 30 indicates significant FFPE damage, corresponding to an artifactual excess of C to T false mutations on one strand and no G to A variants on the other strand. Here, the cohort is centered on Q of around 32 (Figure 3.2c), with 18 BAMs slightly below 30. Through manual reviews and investigations through IGV, we again confirm that no FFPE damage was observed.

5.5 Purity and tumor content

We scanned all samples in the cohort for allele imbalance at sites of common heterozygous sites through the HapASeg algorithm. All samples with >100x mean target coverage show

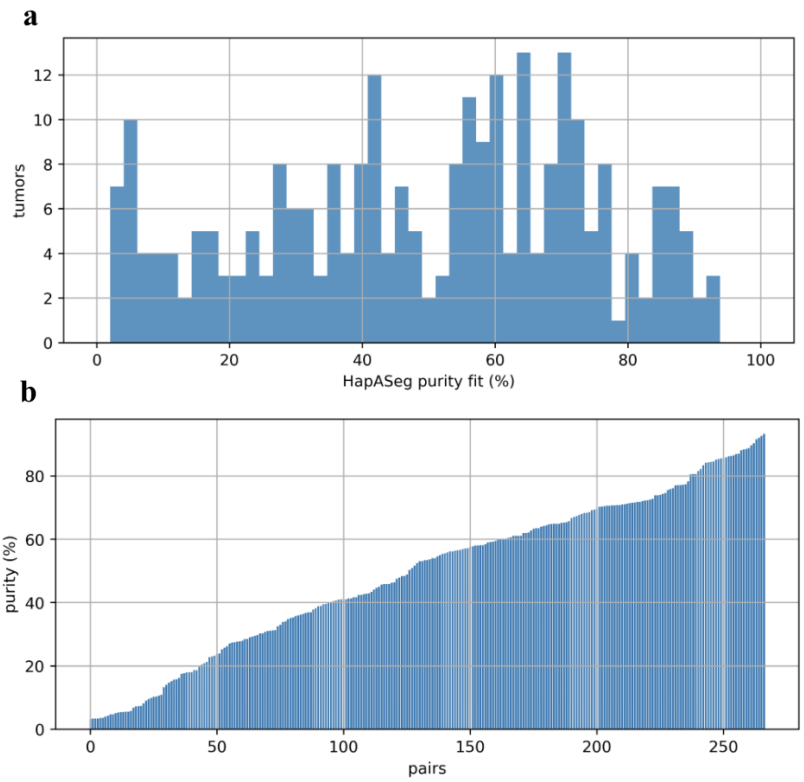

Figure 3.3

(a) Percentage of purity distribution calculated from the ABSOLUTE algorithm.

(b) Detailed purity percentage of all the pairs ranked from lowest to highest.

clear heterozygous site shifts. We applied HapASeg to infer the segmented allele-specific copy-number ratios. These segmentation data served as the input to ABSOLUTE, which estimated the purity and ploidy of each tumor/normal pair. Lower coverage samples yielded poor data, and the purity and ploidy could not be ascertained. The HapASeg algorithm presented a much-improved method for detecting heterozygous variant allele frequencies shifted from 50% and hence improved the segmentation data. Overall, the median purity was around 53.4% (Figure 3.3).

6 Significantly mutated driver genes

We discovered significantly altered driver genes through WES analysis of the 217 non-CAR-T samples that passed through the QC pipeline. To analyze all 217 samples for potential cancer-causing genes, we utilized state-of-the-art novel computational techniques to separate inherited variants and other discrepancies in samples that only had tumor data.

We divided the data into several subsets described above, then applied the MutSig2CV tool [23] to subsets 2.1, 2.2, and 4. In particular, we aimed to compare the results between

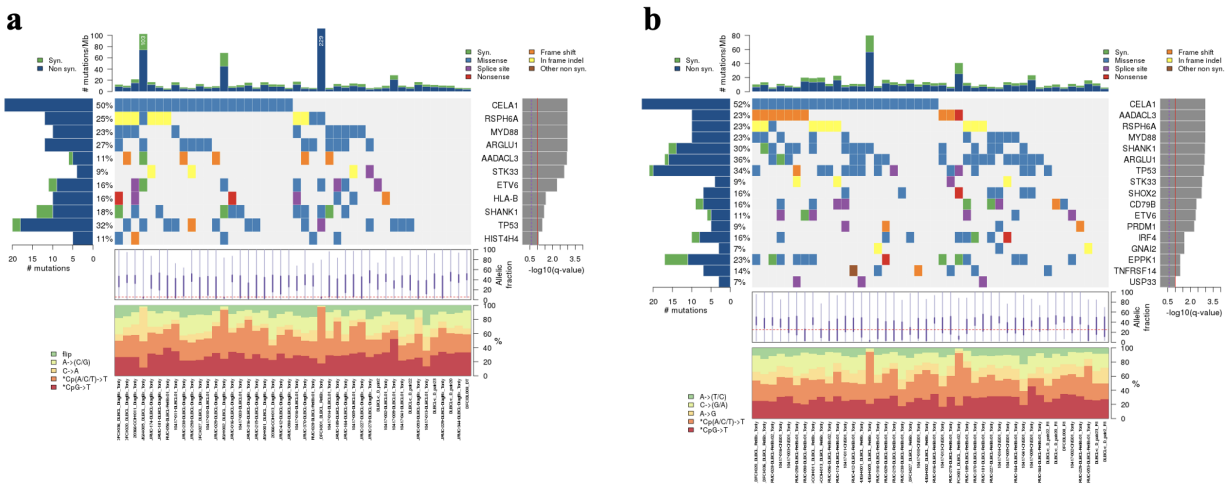

Figure 3.4

(a) Frequently mutated genes were discovered in all diagnostic specimens with paired first relapse samples (subset 2.1). The analysis included the number and frequency of these recurrent mutations displayed on the left side. In the center, there is a gene-sample matrix presenting these frequently mutated genes, which are color-coded according to mutation type and organized by their significance (determined by MutSig2CV q value, shown on the right). The overall mutation density across the entire cohort is depicted at the top, while the allelic fraction of mutations is illustrated at the bottom.

(b) Similarly, frequently mutated genes were discovered in first relapse specimens with paired diagnostic samples (subset 2.2).

the paired diagnostic and relapsed specimens (subsets 2.1 and 2.2, respectively) and comprehensively analyze the drivers discovered in all first relapse samples, including systemic and CNS first relapse samples with and without a paired diagnostic specimen (subset 4).

6.1 Paired diagnostic and relapsed specimens

For the diagnostic paired subset, we discovered 11 candidate cancer genes with a significant discovery threshold (q value < 0.1; Figure 3.4a). The list includes known mutational drivers such as the tumor suppressor *TP*53; key cell growth factor regulator *ETV*6; the immunomodulatory pathway component *HLA-B*; and NF-kB signaling pathways members such as *MYD*88.

For the relapsed paired subset, we discovered 17 significantly mutated cancer driver genes (Figure 3.4b). This number is higher than the number of significantly mutated genes in the paired diagnostic subset, which is consistent with our hypothesis that the patients acquire more mutation events during relapse. Similar to the diagnostic subset, the list includes *MYD*88, *TP*53, and *ETV*6. Furthermore, other than the genes that passed the threshold for significance (q value < 0.1), we noticed several other previously reported mutational drivers,

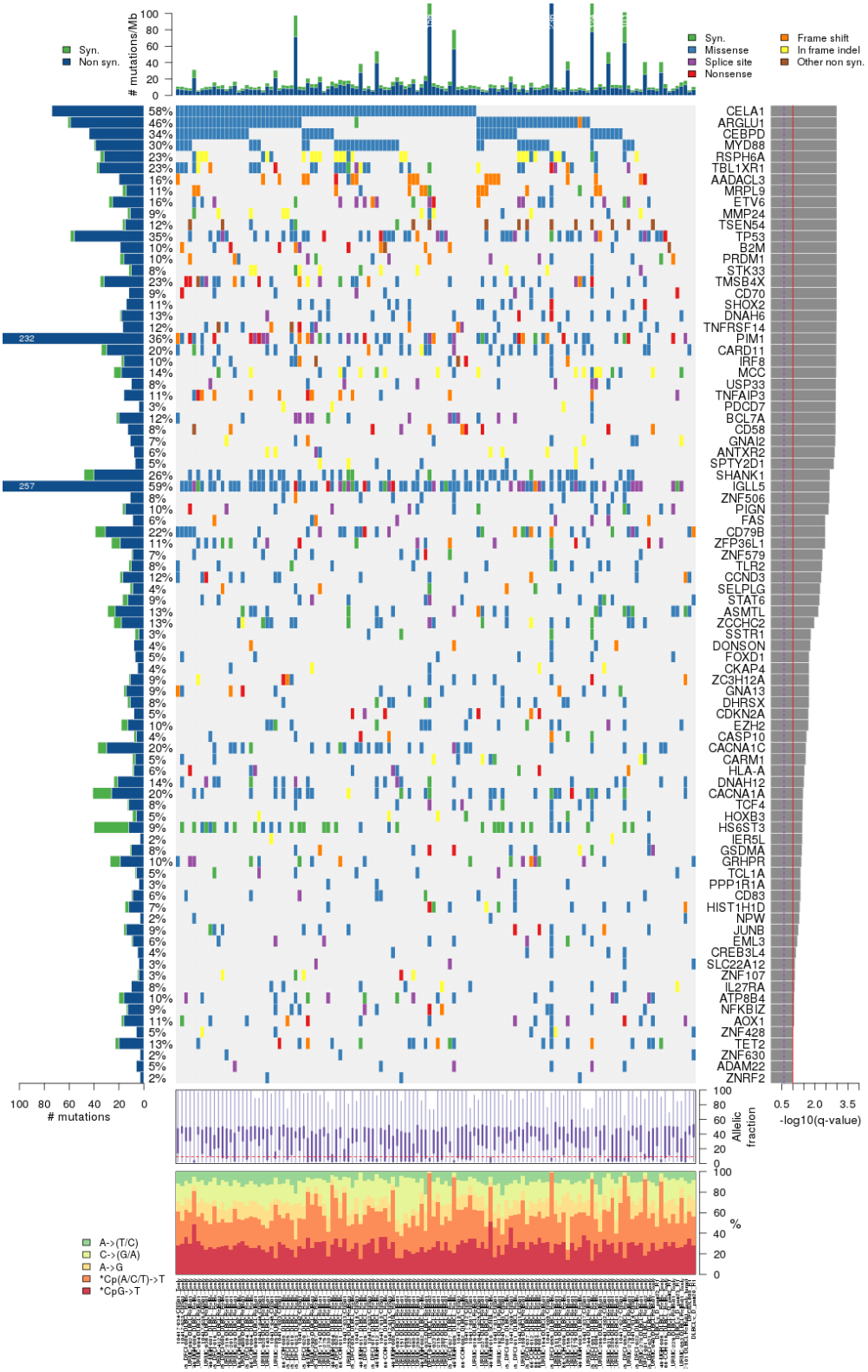

Figure 3.5: Frequently mutated genes were discovered in all first relapse specimens (subset 4).

including NF-kB signaling pathway member *MYD*88; B-cell receptor (BCR) signaling component *CD*79*B*; oncogenic transcription factor *IRF*4; and proto-oncogene *GNAI*2.

6.2 All first relapsed specimens

In the MutSig2CV analysis of the 128 samples part of subset 4 (all first relapsed samples with and without a paired diagnostic specimen), we again noticed similar drivers as in previously reported mutational drivers in addition to the ones mentioned in subsets 2.1 and 2.2. Notably, we obtained the components of the BCR pathway *CD*79*B*; Toll-like receptor (TLR) *MYD*88; JAK/STAT pathways *STAT*6, *PIM*1; immunomodulatory pathway *CD*70, *B*2*M*, *CD*58, *HLA-A*, *CD*83; and NF-kB signaling pathways *FAS*, *BCL*7*A*, *TNFAIP*3. We also noted additional NF-kB modifiers *CARD*11 and *NFKBIZ*.

*TP*53 mutations were observed in 35% of the cases, which exceeds the previously documented frequency of approximately 20% at the time of diagnosis [8], indicating a possible evolution of genetic profiles in DLBCLs. Additionally, *CCND*3 mutations were identified in 13% of cases. This mutation is associated with the development of chemo-resistance [27] and is not typically identified as a driver gene upon initial diagnosis.

Furthermore, *STAT*6 and *NFKBIZ* showed mutations in 9% of the cases each. These genes, like *CCND*3, were also implicated as potential contributors to chemo-resistance in the findings of Morin et al. (2016) [27], and are not classified as driver genes at diagnosis, suggesting a potential role in disease progression or treatment response.

The frequencies of the remaining mutated genes were found to be comparable to those observed in historical diagnostic cases of DLBCLs, indicating a consistent mutation landscape over time for these genes.

6.3 Comparison with previously studied DLBCL cohort

Paired diagnostic and relapsed samples (subset 2.1 and 2.2)

Frequency scatter plots and quantile-quantile (Q-Q) plots were generated to draw comparisons between our DLBCL cohort and the reference DLBCL cohort from Chapuy et al. (2018) [8]. We used the list of 98 candidate cancer genes (CCGs) from the Chapuy cohort as a benchmark for frequency comparison. Figure 3.6a illustrates a scatter plot where each CCG was plotted with its mutation frequency in the Chapuy cohort against ours. A Pearson correlation coefficient of 0.70 indicates a relatively strong positive correlation between the two datasets, suggesting that many mutation frequencies are conserved across both cohorts. However, *IGLL*5 is identified as an outlier, where its frequencies in our cohort are identified to be much higher compared to the Chapuy cohort.

Figure 3.6c shows the Q-Q plot of the p-values from our gene mutation frequency analysis

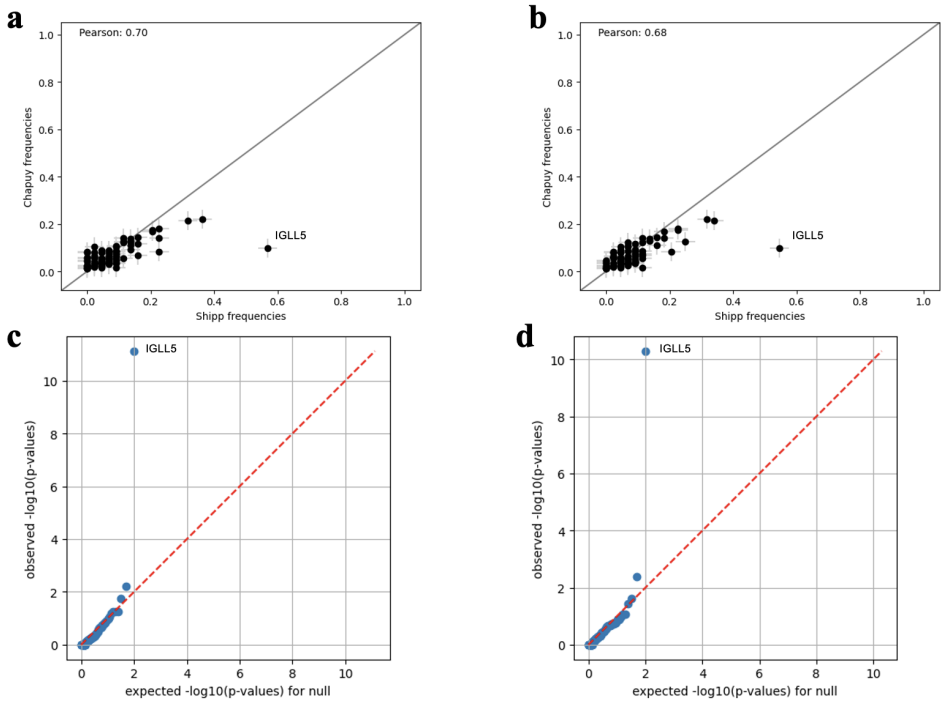

Figure 3.6

(a) Scatter plot of CCGs frequencies in subset 2.1 with the Chapuy cohort.

(b) Scatter plot of CCGs frequencies in subset 2.2 with the Chapuy cohort.

(c) Quantile-Quantile (Q-Q) plot comparing the distribution of observed p-values from the mutation frequency analysis of cohort 2.1 against the expected distribution under the null hypothesis.

(d) Q-Q plot comparing the distribution of observed p-values from the mutation frequency analysis of subset 2.2 against the expected distribution under the null hypothesis.

compared to the expected distribution under the null hypothesis between cohorts using the Fisher Exact test. The alignment of the p-values roughly falls along the red diagonal line, showing that p-values do not demonstrate significant deviation from the null model. Nevertheless, *IGLL5* remains an outlier, as identified at the top left of figure 3.6c.

The relapsed cohort follows a similar trend for both the frequency plot and the Q-Q plot, and the Pearson correlation coefficient is calculated to be 0.68 (Figures 3.6b, 3.6d).

All First Relapse Samples (Subset 4)

Subset 4 includes all first relapsed samples (systemic and CNS first relapse samples with and without a paired diagnostic specimen). The scatter plot displays a Pearson correlation coefficient of 0.73, slightly higher than previously reported, suggesting a consistently strong relationship between mutation frequencies in our cohort and the Chapuy cohort (Figure 3.7a).

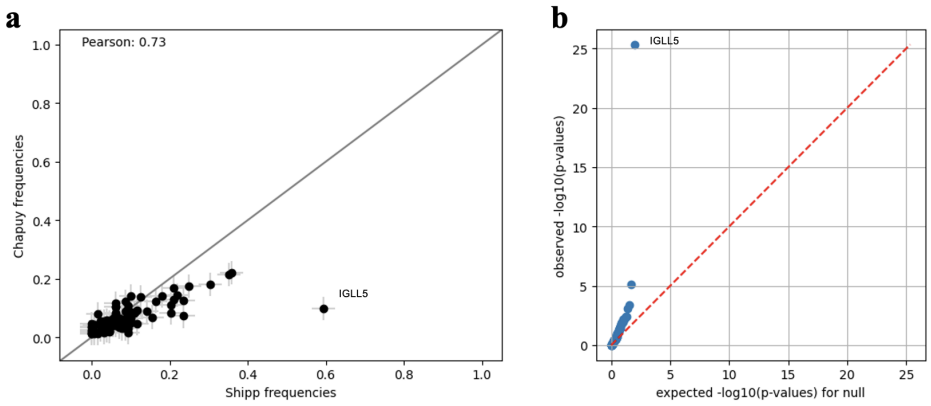

Figure 3.7

(a) Scatter plot of CCGs frequencies in subset 4 with the Chapuy cohort.

(b) Q-Q plot comparing the distribution of observed p-values from the mutation frequency analysis of subset 4 against the expected distribution under the null hypothesis.

The Q-Q plot for subset 4 (Figure 3.7b) illustrates an upward shift from the diagonal line when plotting the observed p-values against the expected distribution under the null hypothesis. This deviation, especially at the higher significance end of the p-value spectrum (i.e., high $-\log_{10}$(p-values)), is more pronounced compared to earlier observations. This observation illustrates that the relapsed samples accumulated more mutations compared to the baseline DLBCL cohort.

7 Significantly Recurrent SCNAs

Next, we separated the samples into primary and relapsed sample groups then identified significantly recurrent Somatic Copy Number Alterations (SCNAs) with the GISTIC2.0 program based on each group's data (Figure 3.8), which are then each separated into arm-level events and focal events for amplification (or gains) and deletion (or losses). The events correlate highly with the previously studied DLBCL cohorts [8].

8 DLBCL Molecular Clustering

We applied non-negative matrix factorization (NMF) consensus clustering to a list of 163 identified genetic driver alterations, dividing the samples in our study into five robust subsets of tumors (clusters) with discrete genetic signatures (hereafter referred to as coordinate genetic signatures; C1–C5).

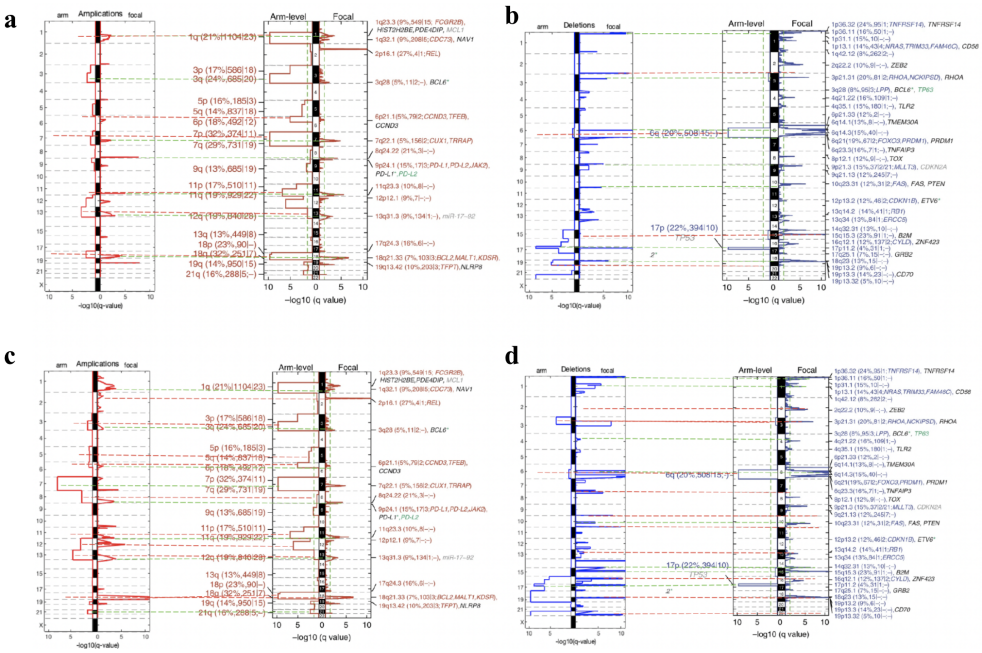

Figure 3.8

(a) GISTIC2.0-defined recurrent copy number amplifications for primary DLBCL samples with arm-level events (left) and focal events (right), as compared to previous studies of DLBCL cohorts [8]. Chromosomes are shown on the vertical axis. The green line denotes agreements with known previous cohorts, while the red line represents disagreements.

(b) GISTIC2.0-defined recurrent copy number deletions for primary DLBCL samples with arm-level events (left) and focal events (right), as compared to previous studies of DLBCL cohorts [8].

(c) GISTIC2.0-defined recurrent copy number amplifications for relapsed DLBCL samples with arm-level events (left) and focal events (right), as compared to previous studies of DLBCL cohorts [8].

(d) GISTIC2.0-defined recurrent copy number deletions for relapsed DLBCL samples with arm-level events (left) and focal events (right), as compared to previous studies of DLBCL cohorts [8].

8.1 Clustering Distribution

We obtained a large portion of C2 and C5 tumors among all the specimens. More notably, when we narrowed down the predictions based on the confidence level given by the classifier, which included roughly 70.5% of the samples, we noticed that C2 and C5 also have the highest portion of confidence over 0.7, representing more accurate predictions (Figure 3.9a). A similar clustering distribution is observed for subset 4 (all first relapse samples), where the samples are dominated by C2 and C5 DLBCLs (Figure 3.9b). These findings are notable because newly diagnosed C2 and C5 DLBCLs are more likely to relapse following initial induction chemotherapy [8].

We further classified the first systemic relapse samples (subset 5) and the first CNS relapse samples (subset 6) separately. Both of these subsets have a minimal number of samples from

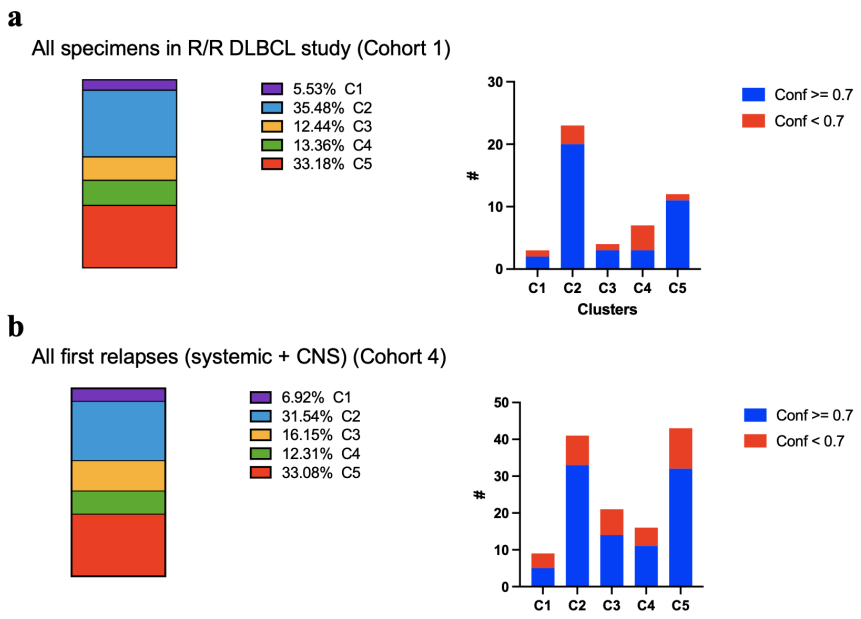

Figure 3.9

(a) The DLBCL classifier using the non-negative matrix factorization (NMF) consensus clustering designed by Chapuy et al. (2018) [8] applied to all specimens (subset 1) in the study with percentages of each cluster illustrated in the bar graph (left). Stacked bar graph of counts of predictions with confidence of over 0.7 and under 0.7 broken down by each cluster (right).

(b) The DLBCL classifier applied to all first relapse specimens (subset 4) in the study with percentages of each cluster illustrated in the bar graph (left). Stacked bar graph of counts of predictions with confidence of over 0.7 and under 0.7 broken down by each cluster (right).

subset 1, which is notable because newly diagnosed C1 DLBCLs are less likely to relapse following induction therapy [8]. While subset 5 represents enrichment in both C2 and C5, the first CNS relapse subset is significantly dominated by C5 tumors. This is of interest because newly diagnosed primary CNS lymphomas have a genetic signature very similar to that of C5 tumors [8], [28], [29]. We then broke down each of these two subsets into early ($<=$ 12 months) and late relapses ($>$ 12 months); we saw a much more apparent divide in the concentration of the C2 and C5 tumors in both subdivided subsets. In both subset 5 and subset 6, the late relapses are enriched in C5 tumors, while the early relapses have a more even division between the different clusters (Figure 3.10). Nevertheless, subset 5's early relapse samples are still primarily dominated by C2 tumors.

8.2 Subset Comparisons and Analyses

Allele Frequency Comparisons

Subset 4 We first compared all first relapse samples (subset 4) with the DLBclass subset of newly diagnosed DLBCLs. We noticed more driver alterations in the first relapse

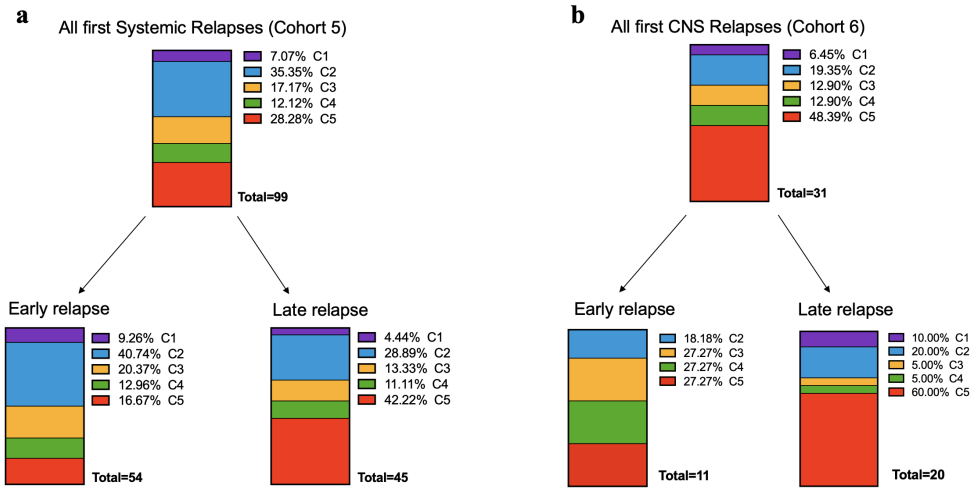

Figure 3.10

(a) The DLBCL classifier applied to all systemic relapse specimens (subset 5) in the study with percentages of each cluster illustrated in the bar graph (top). The cohort is further broken down into early and late relapsed patients, with the clustering assignment of the specimens illustrated in bar graphs (bottom).

(b) The DLBCL classifier applied to all CNS relapse specimens (subset 6) in the study with percentages of each cluster illustrated in the bar graph (top). The subset is further broken down into early and late relapsed patients, with the clustering assignment of the specimens illustrated in bar graphs (bottom).

samples (subset 4) versus the newly diagnosed DLBCLs (DLBclass). Several genes have higher frequencies in subset 4 versus diagnostic samples in a statistically significant manner (alpha value of 0.05). These include genes that are known targets of aberrant somatic hypermutation *IGLL*5, *BCL*2; tumor suppressor *TP*53; chromatin modifiers *KMT*2*D*; the Wnt pathway components *TBL*1*XR*1; DNA demethylation regulator *TET*2; B cell transcription factor *IRF*4; RNA spliceosome complex *SF*3*B*1; and common lymphoid protein-encoding gene *POU*2*AF*1. As first relapse samples (subset 4) are enriched in C2 and C5 (Figure 3.9b), we observed many of the signature genes in C2 and C5 are more frequently altered in relapsed DLBCLs. In particular, *BCL*2 copy gains and $MYD88^{L265P}$ mutations are major components of the C5 signature, and *TP*53 mutations are typically detected by C2 tumors [8]. One major outlier we noticed that had a much lower mutation frequency is *BCL*6, which is known to be a driver of C1 tumors [8]. These findings are noteworthy because newly diagnosed C1 DLBCLs are less likely to relapse following induction therapy, and C1 DLBCLs are less common in our relapsed DLBCL subset (compared with newly diagnosed DLBCLs) (Figure 3.9b).

The enrichment of CNAs at specific chromosomal locations indicates regions that harbor genes with potential oncogenic or tumor suppressor functions. The loss or gain of these genes can contribute to genomic instability—a hallmark of cancer—and affect the effectiveness of

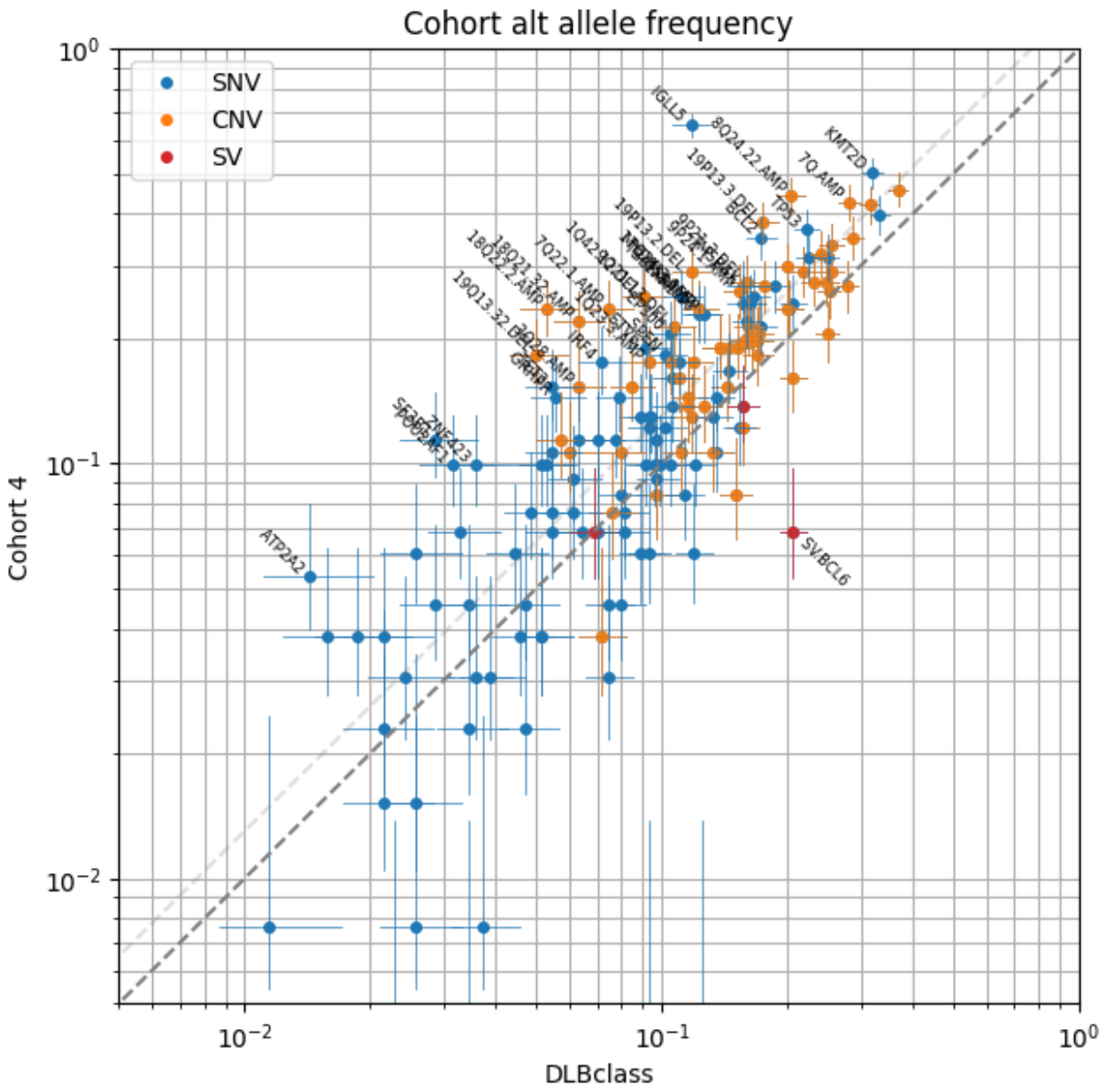

Figure 3.11: Alternate allele frequency comparisons between subset 4 and DLBclass, with blue dots indicating SNVs, orange dots indicating CNVs, and red dots indicating SVs. The mutation frequency differences deemed statistically significant under an alpha value of 0.05 are labeled in black.

therapy. In particular, subset 4 harbors amplification of $8q24.22$; deletions at $19p13.3$ involving *CD*70, which may influence tumor immunity; amplification in $9p24.1$, involving *PD-L*1, which is a known mechanism of immune evasion; amplifications at $7q22.1$ targeting *CDK*6, which influences cell cycle regulation; and deletions at $19p13.2$, involving *SMARCA*4, which may disrupt chromatin remodeling and gene expression. Furthermore, deletions at $1q42.12$ and amplifications at $18q21.32$ are associated with an unfavorable outcome in patients with

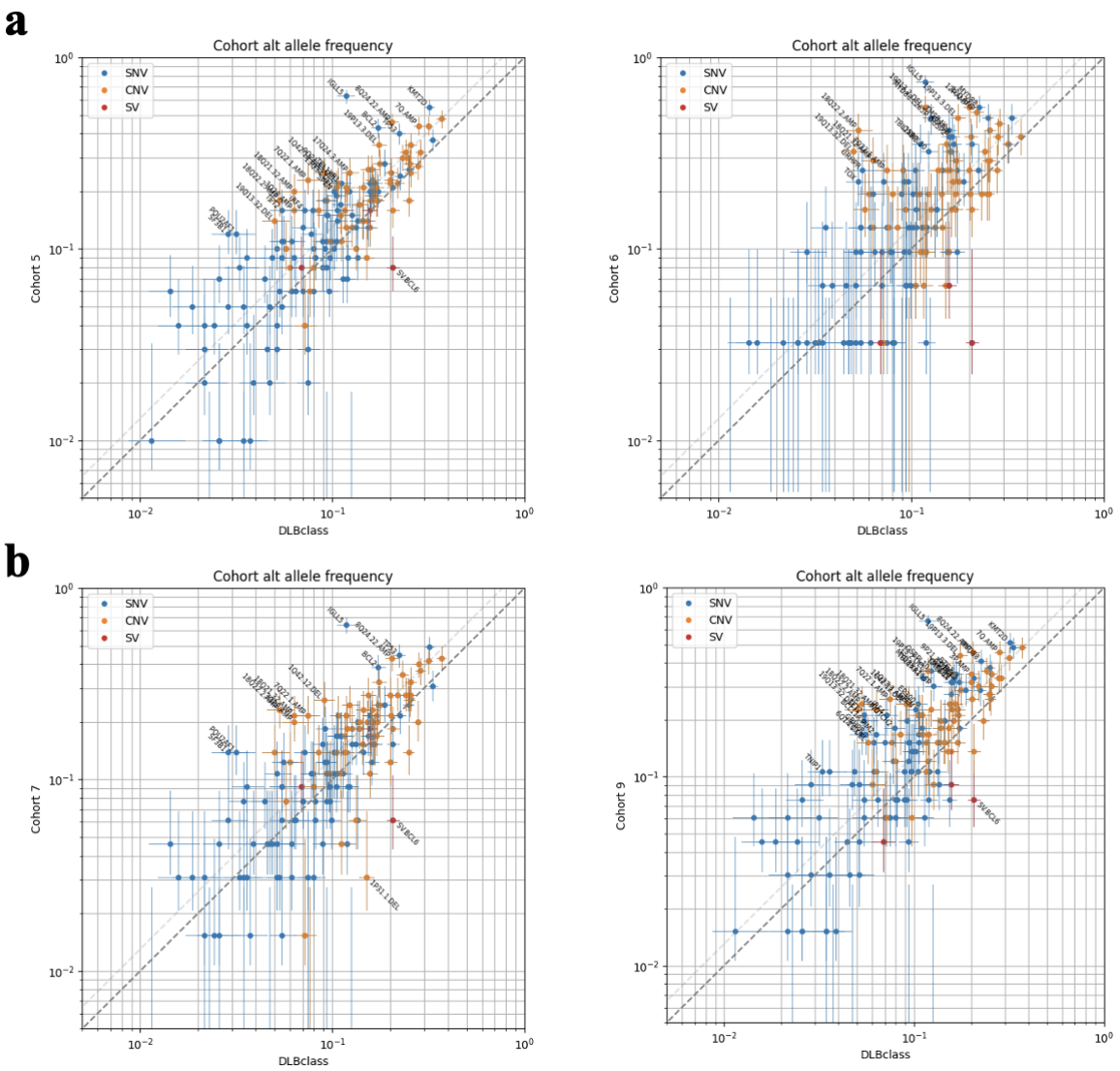

Figure 3.12: Alternate allele frequency comparisons between subsets 5 and 6 with DLBclass (a) and subsets 7 and 9 with DLBclass (b), with blue dots indicating SNVs, orange dots indicating CNVs, and red dots indicating SVs. The mutation frequency differences deemed statistically significant under an alpha value of 0.05 are labeled in black.

newly diagnosed DLBCL [8].

Subset 5 and 6 We next analyzed the frequencies of mutations in subsets 5 and 6 (all first systemic relapse samples and all first CNS relapse samples, respectively) as compared to newly diagnosed DLBCLs (DLBclass) (Figure 3.12). First systemic relapse DLBCLs (subset 5) show a higher frequency of mutations in both C2 and C5 tumor drivers, including *TP*53, *BCL*2, *KMT*2*D*, *TET*2, *SF*3*B*1, and *POU*2*AF*1, confirming our clustering observations (Figure 3.10). These mutations are associated with various aspects of cellular regulation, such as apoptosis (*BCL*2), DNA damage response (*TP*53), epigenetic modifications

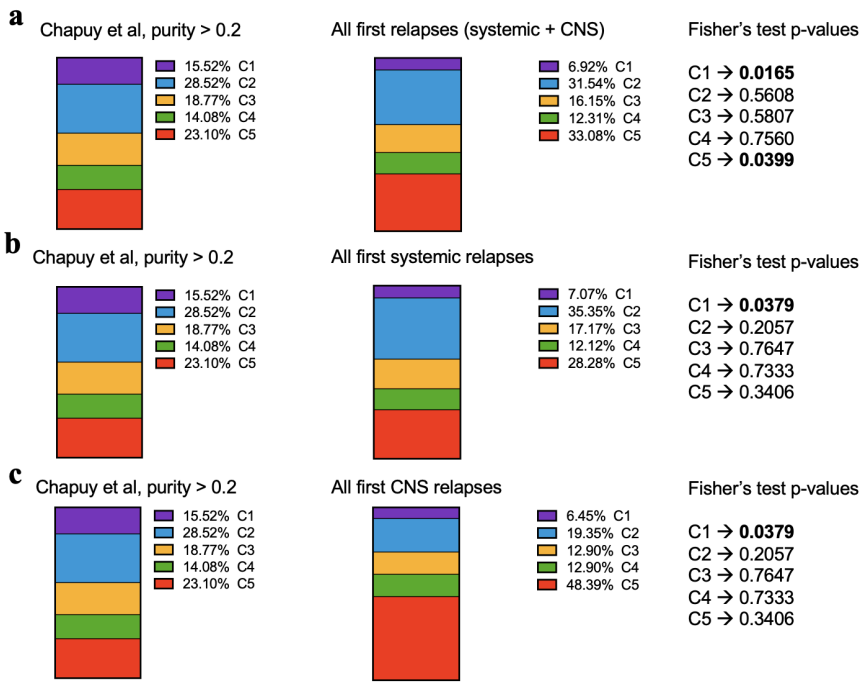

Figure 3.13: Stacked bar graphs representing the inputs to the Fisher test for testing cluster distributions of subsets 4(a), 5(b), and 6(c) as compared to the Chapuy et al. cohort (2018) [8] with the p-values listed to the right (bolded values show p values below the significance threshold).

(*KMT2D* and *TET2*), RNA splicing (*SF3B1*), and lymphoid development (*POU2AF1*). On the other hand, subset 6, with nearly 50% samples identified as C5 DLBCLs, exhibits a different mutational spectrum. It mostly has mutations in *MYD88*, with $MYD88^{L265P}$ in particular, and increased mutations in *TBL1XR1*, corresponding to C5 tumor signatures.

Both subsets have an enrichment of specific CNAs. Both subsets demonstrated deletions at $19p13.3$ (involving *CD70*, the *CD27* partner modulator of T-cell co-stimulations) and amplification at $18q21.32$ (involving *MALT1*), which is associated with R-CHOP induction therapy [8].

The differential patterns of genetic alterations in systemic and CNS relapse subsets may reflect distinct evolutionary paths of DLBCL under therapeutic pressure and different microenvironmental influences at the relapse sites. These findings underscore the potential for site-specific therapy and the need for targeted surveillance strategies based on the mutational profile at diagnosis.

Subset 7 and 9 We further compared subsets 7 and 9 to analyze the potential differences in the genetic signatures of early and late first relapsed DLBCL (Figure 3.12). In early relapsed DLBCLs (subset 7), there is a higher frequency of mutations in *IGLL5*, *TP53*, *BCL2*, *SF3B1*, and *POU2AF1* compared to the DLBclass baseline samples; these gene alterations are characteristic of both C2 and C5 tumors. Deletions at $1q42.12$ and amplifications at

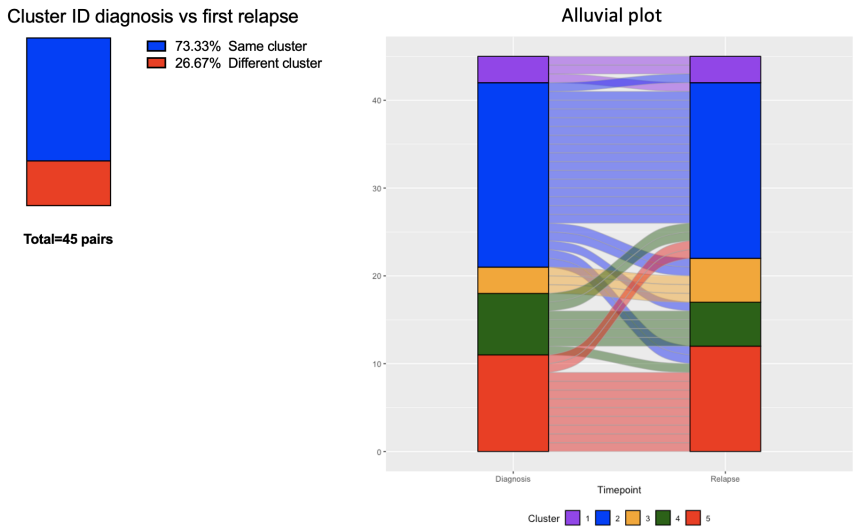

Figure 3.14: Alluvial plot of clustering results of paired diagnostic and relapsed samples (n=45).

18q21.32, both associated with inferior outcomes in newly diagnosed DLBCLs treated with standard induction therapy (R-CHOP), also appear here, underscoring their possible role in early relapse.

DLBCLs with later relapse (> 12 months, subset 9) exhibit increased mutations in *IGLL*5, *TET*2, *CD*79*B*, *TBL*1*XR*1, *BTG*2, *OSBPL*10, *BCL*11*A*, and *IRF*4. The inclusion of *TET*2 points towards an epigenetic component that could be more influential in the later stages of relapse, affecting DNA methylation patterns over time. Mutations in *CD*79*B*, involved in BCR signaling, and *IRF*4, important for immune response regulation, could indicate adaptations to evade immune surveillance over a longer period. *BTG*2 is a tumor suppressor gene, and its mutation may eventually lead to unchecked cell growth, contributing to late relapse.

The CNAs, similar to those observed in the early relapse subset, include amplifications at 8q24.22 and deletions at 19p13.3. However, DLBCLs with late relapse (> 12 months, subset 9) uniquely exhibit deletions at 19p13.2, affecting *SMARCA*4, which has roles in chromatin remodeling.

Fisher Test Comparison

We conducted a Fisher's test with Benjamini-Hochberg correction (FDR $q < 0.1$) test to compare the clustering results of the high-purity samples from the Chapuy cohort with subsets 4, 5, and 6 from our study (Figure 3.13). Setting the significance threshold to 0.05, we reject the null hypothesis for C1 tumor distribution between subset 4 (all first relapses) and Chapuy (p-value of 0.0078), as well as between subset 5 and Chapuy (p-value of 0.0185), again confirming our observation that there are few C1 tumors in our cohorts. Moreover,

we also reject the null hypothesis comparing C2 DLBCLs from subset 5 (p-value of 0.0342) and C5 DLBCLs from subset 6 (p-value of 0.0101), demonstrating the enrichment in C2 and C5 tumors, respectively, in first systemic relapses and CNS relapses (Figure 3.13). Taken together, these data suggest that "good-risk" C1 DLBCLs are less commonly represented in relapsed DLBCLs, C2 DLBCLs are more frequent in systemic relapsed DLBCLs, and C5 DLBCLs are more common in CNS relapsed disease.

Alluvial Plot Analysis

We use alluvial plots to visualize paired patient sample changes over time changes in cluster designation in paired diagnostic and relapsed DLBCLs. When comparing paired diagnostic and relapsed tumor biopsies (subsets 2.1 to 2.2), we found that 33 of the paired biopsies cases had the same designations, while 12 cases were different (Figure 3.14). Specifically, five cases shifted to C2 at relapse and acquired additional $TP53$ alterations (mutations or $17p$ deletions), which are important signatures of C2 tumors as discussed above, then acquired the other during relapse in two of five cases, resulting in bi-allelic hits.

Chapter 4

Discussion

9 Findings and Clinical Interpretations

The analysis of diffuse large B-cell lymphoma (DLBCL) biopsies from diagnostic and relapse samples has revealed distinct patterns of cluster enrichment that may hold the key to understanding disease progression and resistance mechanisms. In a broader analysis of all the first relapsed DLBCL cases (subset 4), there is a significant enrichment in the high-risk C2 and C5 tumors, with reduced frequencies of good risk in C1 DLBCLs. These tumors have high frequencies in mutations associated with therapy resistance, including $TP53$ (36%), $BCL2$ (35%), and $KMT2D$ (50%). This confirms the aggressive nature of these genetic aberrations and their role in the persistence and resurgence of the disease. Further, significant drivers include C5-defining genes such as $TBL1XR1$(25%), $IRF4$ (17%), $SF3B1$ (11%), $POU2AF1$ (14%), and $GRHPR$ (9%). This pattern indicates that the features defining C2 and C5, potentially including gene expression profiles, mutations, or micro-environmental interactions, are crucial for relapse pathophysiology.

Enriched CNAs like $1q42.12$ deletion and $18q21.32$ amplification, previously linked to PFS in previous studies [8], are also prevalent in first relapsed cases (subset 4). This reiterates the impact of these genetic changes on disease progression and treatment outcomes. The high frequency of the $19p13.3$ deletion (38%), which results in $CD70$ copy number loss, is striking compared to its presence in diagnostic biopsies (17%).

Given the important role of $CD70/CD27$ in T-cell co-stimulation, this is a potential genetic basis of immune evasion in relapsed DLBCL. More precisely, CNS relapse samples (subset 6) show a marked enrichment in C5, pointing to distinct biological behaviors or treatment responses in these cases. Conversely, systemic relapses (subset 5) have a slight predominance of C2, which may reflect different driving forces in systemic DLBCLs compared to CNS disease. Consistent with classifier assignments, systemic relapses are enriched in $TP53$ mutations, whereas CNS relapses are enriched in $MYD88$ mutations (both $MYD88^{L265P}$ and

others), as well as other C5-defining drivers (*TBL1XR1*, *TMSB4X*, *GRHPR*).

Our data further indicates a temporal component to cluster enrichment, with early relapses predominantly falling into C2, while late relapses tend to fall into C5. This temporal distinction persists even when systemic and CNS relapses are analyzed separately, highlighting the consistency of the association between cluster identity and relapse timing.

Interestingly, within the paired diagnostic and relapse subsets, the majority of cases (73%) retain their original cluster identity upon relapse. This stability implies that the primary factors defining cluster identity are fundamental to the biology of the DLBCL and may persist through the course of the disease. However, changes to C2, with associated *TP53* alterations, might be expected in relapsed DLBCL.

In summary, the cluster analysis of DLBCL diagnostic and relapse samples has unveiled significant patterns of enrichment that could inform future research and treatment strategies. The strong association between specific clusters and relapse patterns, particularly the enrichment of C5 in CNS relapses and C2 in early relapses, emphasizes the potential for cluster-specific approaches in DLBCL treatments.

Our data provide a granular view of the genetic alterations driving DLBCL relapse and underscore the critical need for personalized treatment strategies. These strategies should be responsive to the genetic subtypes, the kinetics of relapse, and the role of the immune system, with emphasis on novel agents that target these specific genetic vulnerabilities. The persistence of certain genetic profiles from diagnosis to relapse further highlights the potential of early genetic profiling in guiding long-term disease management and improving patient outcomes.

10 Future Work

10.1 Limitations

Despite applying numerous tools to filter out germline mutations and artifacts, including Tonly2, BLAT realignment filter, and the MAFPonFilter, we still noticed that the list of significant genes produced by MutSig2CV tool contains many artifacts. We will continue to investigate the characteristics of artifacts and design a tool to effectively filter them from our pipeline.

10.2 Recurrent genetic alterations

Our research aims to characterize further recurrent genetic alterations in DLBCL as initially described in significant studies such as Chapuy et al. (2018) [8]. For subset 1, we plan to scrutinize the recurrently mutated genes, chromosomal rearrangements involving *BCL2*, *BCL6*, and *MYC*, as well as other chromosomal rearrangements utilizing the latest SV

pipeline. Additionally, we intend to examine recurrent SCNAs at both the arm-level and focal areas to provide a comprehensive genetic landscape of these malignancies.

10.3 NMF classification

To delve deeper into the genetic structure of DLBCL, NMF classification will be applied to various subsets. This includes primary refractory diagnostic samples (subset 3.1), all first relapse samples (subset 4), all systemic first relapse samples (subset 5), and all CNS first relapse samples (subset 6). Through this classification, we aim to uncover any additional genetic substructures that might influence the prognosis and therapeutic responses in these distinct groups.

10.4 Comparison of genetic alterations across subsets

A comparative analysis of recurrent genetic alterations will be conducted across several pairs of cohorts. These include all paired newly diagnosed and relapse samples (subsets 2.1 versus 2.2), diagnostic specimens with paired first systemic relapse samples (subsets 2.3 versus 2.4), and diagnostic specimens with paired first CNS relapse samples (subsets 2.5 versus 2.6). Special attention will be given to identifying new genetic alterations that emerge at relapse. Comparisons will also be made between first systemic and first CNS relapse samples, as well as between primary refractory diagnostic samples and early versus late first relapse samples, to identify distinct genetic alterations that correlate with different relapse timings.

10.5 Mutational signature analysis

Given the artifact limitation, as we discussed above, we will work to eliminate and filter out the artifacts further to improve the mutational signature analysis. The mutational signature analysis will compare all diagnostic specimens (subset 2.1) versus all first relapse samples (subset 4). This analysis will focus on identifying new mutational signatures associated with alkylator therapy at relapse, and any new genetic alterations driven by such therapies.

10.6 Clonal evolution studies

We will further conduct clonal evaluation studies between the subsets with the Phylogic N-Dimensional with Timing (PhylogicNDT) tool, which can statistically model phylogenetic and evolutionary trajectories based on mutation and copy-number data representing samples taken at single or multiple time points [30]. This study will encompass comparisons between all diagnostic specimens with paired first relapse specimens (subset 2.1) and all first relapse specimens with paired diagnostic samples (subset 2.2). We will also explore clonal dynamics in patients with paired diagnostic primary refractory specimens (subset 3.1) and first progression (subset 3.2) samples. A specific focus will be placed on selected patients

in the CAR-T cohort who experience multiple relapses to understand the clonal shifts and selection pressures imposed by advanced therapies.

These studies are pivotal in advancing our understanding of DLBCL at the genetic level. By elucidating the complex interplay of genetic alterations across various stages of the disease, we can improve diagnostic precision, predict therapeutic outcomes, and ultimately enhance the management and treatment strategies for patients with DLBCL.

Chapter 5

Conclusion

11 Conclusion

DLBCL, the most common type of non-Hodgkin lymphoma, presents significant clinical challenges due to its heterogeneous nature, both in terms of its biological underpinnings and its response to treatment. This study aims to dissect the complex genetic landscape of DLBCL, particularly focusing on the recurrent genetic alterations and their association with treatment outcomes in R/R DLBCL cases. Through extensive genetic analyses, we identified specific genetic substructures and vulnerabilities that contribute to the disease's prognosis and therapy resistance.

Our analysis of mutations, CNAs, and SVs in biopsies from R/R DLBCL patients underscores a significant enrichment in clusters C2 and C5, which are associated with poor prognoses and high relapse rates. These clusters exhibited distinct genetic profiles that were strongly correlated with differential responses to the standard R-CHOP therapy. For instance, mutations commonly associated with therapy resistance such as *TP*53 and *BCL*2, were predominantly found in these high-risk clusters. This enrichment suggests that patients within these clusters are less likely to respond to conventional therapies and may benefit from alternative strategies that target these specific genetic alterations. Our findings also point towards a nuanced understanding of DLBCL, where the timing of relapse, the location of relapse (CNS versus systemic), and the primary treatment response (refractory versus sensitive) are all interwoven with specific molecular and cellular characteristics of the lymphoma clusters.

The integration of comprehensive genetic profiling into clinical practice could revolutionize the management of DLBCL. By understanding the specific genetic alterations that drive poor outcomes, clinicians can tailor treatments to target these vulnerabilities. For instance, the addition of novel therapeutic agents that specifically inhibit the pathways altered in C2 and C5 could potentially improve response rates and survival outcomes for patients with high-risk DLBCL.

Future research should focus on eliminating artifacts from the list of significantly mutated genes detected by MutSig2CV and validating the findings in larger cohorts and integrating real-time genetic profiling into clinical trials. Artifacts could arise due to systematic false positive mutations in particular genes or, alternatively, incorrect background mutation frequency estimated by MutSig2CV. The development of a robust molecular classifier to predict relapse and tailor treatments could significantly enhance the precision medicine approach in DLBCL. Additionally, further studies are needed to explore the mechanisms of immune evasion and resistance in R/R DLBCL, which could lead to the development of more effective combination therapies that include immunomodulatory agents.

In conclusion, this study provides a critical foundation for the personalized treatment of DLBCL, highlighting the importance of genetic substructures in determining the disease course and treatment response. By leveraging these insights, the next generation of DLBCL treatment could see significantly improved outcomes, turning a once uniformly fatal diagnosis into a condition that is as individually manageable as it is genetically diverse.

References

[1] A. Sethi, A. Tandon, H. Mishra, and I. Singh, "Diffuse large b-cell lymphoma: An immunohistochemical approach to diagnosis.," *Journal of oral and maxillofacial pathology : JOMFP*, vol. 23, no. 2, pp. 284–288, Aug. 2019, Place: India, ISSN: 0973-029X 1998-393X. DOI: 10.4103/jomfp.JOMFP_294_18.

[2] International Non-Hodgkin's Lymphoma Prognostic Factors Project, "A predictive model for aggressive non-hodgkin's lymphoma," *The New England Journal of Medicine*, vol. 329, no. 14, pp. 987–994, Sep. 30, 1993, ISSN: 0028-4793. DOI: 10.1056/NEJM199309303291402.

[3] B. Coiffier, E. Lepage, J. Briere, *et al.*, "CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-b-cell lymphoma," *The New England Journal of Medicine*, vol. 346, no. 4, pp. 235–242, Jan. 24, 2002, ISSN: 1533-4406. DOI: 10.1056/NEJMoa011795.

[4] L. H. Sehn and G. Salles, "Diffuse large b-cell lymphoma," *New England Journal of Medicine*, vol. 384, no. 9, pp. 842–858, Mar. 4, 2021, Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMra2027612, ISSN: 0028-4793. DOI: 10.1056/NEJMra2027612. [Online]. Available: https://doi.org/10.1056/NEJMra2027612 (visited on 11/12/2023).

[5] A. A. Alizadeh, M. B. Eisen, R. E. Davis, *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, Feb. 2000, Number: 6769 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/35000501. [Online]. Available: https://www.nature.com/articles/35000501 (visited on 11/12/2023).

[6] A. Rosenwald, G. Wright, W. C. Chan, *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma," *The New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, Jun. 20, 2002, ISSN: 1533-4406. DOI: 10.1056/NEJMoa012914.

[7] A. Goy, "Succeeding in breaking the r-chop ceiling in dlbcl: Learning from negative trials," *Journal of Clinical Oncology*, vol. 35, no. 31, pp. 3519–3522, 2017, PMID: 28926287. DOI: 10.1200/JCO.2017.74.7360. eprint: https://doi.org/10.1200/JCO.2017.74.7360. [Online]. Available: https://doi.org/10.1200/JCO.2017.74.7360.

[8] B. Chapuy, C. Stewart, A. J. Dunford, *et al.*, "Molecular subtypes of diffuse large b cell lymphoma are associated with distinct pathogenic mechanisms and outcomes," *Nature Medicine*, vol. 24, no. 5, pp. 679–690, May 1, 2018, ISSN: 1546-170X. DOI: 10.1038/s41591-018-0016-8. [Online]. Available: https://doi.org/10.1038/s41591-018-0016-8.

[9] M. Crump, S. S. Neelapu, U. Farooq, *et al.*, "Outcomes in refractory diffuse large b-cell lymphoma: Results from the international SCHOLAR-1 study," *Blood*, vol. 130, no. 16, pp. 1800–1808, Oct. 2017, _eprint: https://ashpublications.org/blood/article-pdf/130/16/1800/1402936/blood769620.pdf, ISSN: 0006-4971. DOI: 10.1182/blood-2017-03-769620. [Online]. Available: https://doi.org/10.1182/blood-2017-03-769620.

[10] S. H. Bernstein, J. M. Unger, M. LeBlanc, J. Friedberg, T. P. Miller, and R. I. Fisher, "Natural history of CNS relapse in patients with aggressive non-hodgkin's lymphoma: A 20-year follow-up analysis of SWOG 8516—the southwest oncology group," *Journal of Clinical Oncology*, vol. 27, no. 1, pp. 114–119, Jan. 1, 2009, ISSN: 0732-183X. DOI: 10.1200/JCO.2008.16.8021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4879698/ (visited on 11/12/2023).

[11] N. Schmitz, S. Zeynalova, M. Nickelsen, *et al.*, "CNS international prognostic index: A risk model for CNS relapse in patients with diffuse large b-cell lymphoma treated with r-CHOP," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 34, no. 26, pp. 3150–3156, Sep. 10, 2016, ISSN: 1527-7755. DOI: 10.1200/JCO.2015.65.6520.

[12] L. K. Hilton, H. S. Ngu, B. Collinge, *et al.*, "Relapse timing is associated with distinct evolutionary dynamics in diffuse large b-cell lymphoma," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 41, no. 25, pp. 4164–4177, Sep. 1, 2023, ISSN: 1527-7755. DOI: 10.1200/JCO.23.00570.

[13] H. Nguyen, A. Perry, P. Skrabek, *et al.*, "Validation of the double-hit gene expression signature (DLBCL90) in an independent cohort of patients with diffuse large b-cell lymphoma of germinal center origin," *The Journal of Molecular Diagnostics*, vol. 23, no. 5, pp. 658–664, May 1, 2021, Publisher: Elsevier, ISSN: 1525-1578. DOI: 10.1016/j.jmoldx.2021.02.005. [Online]. Available: https://www.jmdjournal.org/article/S1525-1578(21)00043-X/fulltext (visited on 12/01/2023).

[14] B. Ylstra and D. de Jong, "DNA or RNA? Classification of B-cell lymphomas," *Blood*, vol. 141, no. 20, pp. 2413–2415, May 2023, ISSN: 0006-4971. DOI: 10.1182/blood.2022018741. eprint: https://ashpublications.org/blood/article-pdf/141/20/2413/2069140/blood_bld-2022-018741-c-main.pdf. [Online]. Available: https://doi.org/10.1182/blood.2022018741.

[15] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biology*, vol. 12, no. 4, R41, Apr. 28, 2011, ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-4-r41. [Online]. Available: https://doi.org/10.1186/gb-2011-12-4-r41 (visited on 05/11/2024).

[16] K. Cibulskis, A. McKenna, T. Fennell, E. Banks, M. DePristo, and G. Getz, "ContEst: estimating cross-contamination of human samples in next-generation sequencing data," *Bioinformatics*, vol. 27, no. 18, pp. 2601–2602, Jul. 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr446. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/18/2601/48864972/bioinformatics_27_18_2601.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr446.

[17] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nature Biotechnology*, vol. 31, no. 3, pp. 213–219, Mar. 2013, Number: 3 Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.2514. [Online]. Available: https://www.nature.com/articles/nbt.2514 (visited on 12/02/2023).

[18] S. Kim, K. Scheffler, A. L. Halpern, *et al.*, "Strelka2: Fast and accurate calling of germline and somatic variants," *Nature Methods*, vol. 15, no. 8, pp. 591–594, Aug. 2018, Number: 8 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0051-x. [Online]. Available: https://www.nature.com/articles/s41592-018-0051-x (visited on 12/02/2023).

[19] A. H. Ramos, L. Lichtenstein, M. Gupta, M. S. Lawrence, T. J. Pugh, G. Saksena, M. Meyerson, and G. Getz, "Oncotator: Cancer variant annotation tool.," *Human mutation*, vol. 36, no. 4, E2423–2429, Apr. 2015, Place: United States, ISSN: 1098-1004 1059-7794. DOI: 10.1002/humu.22771.

[20] M. Costello, T. J. Pugh, T. J. Fennell, *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 Research*, vol. 41, no. 6, e67–e67, Jan. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gks1443. eprint: https://academic.oup.com/nar/article-pdf/41/6/e67/25340074/gks1443.pdf. [Online]. Available: https://doi.org/10.1093/nar/gks1443.

[21] S. Carter, M. Meyerson, and G. Getz, "Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping," *Nature Precedings*, pp. 1–1, Oct. 5, 2011, Publisher: Nature Publishing Group, ISSN: 1756-0357. DOI: 10.1038/npre.2011.6494.1. [Online]. Available: https://www.nature.com/articles/npre.2011.6494.1 (visited on 12/02/2023).

[22] S. L. Carter, K. Cibulskis, E. Helman, *et al.*, "Absolute quantification of somatic DNA alterations in human cancer," *Nature Biotechnology*, vol. 30, no. 5, pp. 413–421, May 2012, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.2203. [Online]. Available: https://www.nature.com/articles/nbt.2203 (visited on 05/09/2024).

[23] M. S. Lawrence, P. Stojanov, P. Polak, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jul. 1, 2013, ISSN: 1476-4687. DOI: 10.1038/nature12213. [Online]. Available: https://doi.org/10.1038/nature12213.

[24] R. Schmitz, G. W. Wright, D. W. Huang, *et al.*, "Genetics and pathogenesis of diffuse large b-cell lymphoma," *New England Journal of Medicine*, vol. 378, no. 15, pp. 1396–1407, 2018. DOI: 10.1056/NEJMoa1801445. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1801445. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1801445.

[25] A. Kamburov, M. S. Lawrence, P. Polak, I. Leshchiner, K. Lage, T. R. Golub, E. S. Lander, and G. Getz, "Comprehensive assessment of cancer missense mutation clustering in protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 40, E5486–5495, Oct. 6, 2015, ISSN: 1091-6490. DOI: 10.1073/pnas.1516373112.

[26] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature Biotechnology*, vol. 29, no. 1, pp. 24–26, Jan. 2011, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/nbt.1754. [Online]. Available: https://www.nature.com/articles/nbt.1754 (visited on 05/11/2024).

[27] R. D. Morin, S. Assouline, M. Alcaide, *et al.*, "Genetic landscapes of relapsed and refractory diffuse large b-cell lymphomas," *Clinical Cancer Research*, vol. 22, no. 9, pp. 2290–2300, May 1, 2016, ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-15-2123. [Online]. Available: https://doi.org/10.1158/1078-0432.CCR-15-2123 (visited on 04/21/2024).

[28] B. Chapuy, M. G. M. Roemer, C. Stewart, *et al.*, "Targetable genetic features of primary testicular and primary central nervous system lymphomas," *Blood*, vol. 127, no. 7, pp. 869–881, Feb. 2016, ISSN: 0006-4971. DOI: 10.1182/blood-2015-10-673236. eprint: https://ashpublications.org/blood/article-pdf/127/7/869/1353834/869.pdf. [Online]. Available: https://doi.org/10.1182/blood-2015-10-673236.

[29] B. Chapuy, H. Cheng, A. Watahiki, *et al.*, "Diffuse large b-cell lymphoma patient-derived xenograft models capture the molecular and biological heterogeneity of the disease," *Blood*, vol. 127, no. 18, pp. 2203–2213, May 2016, ISSN: 0006-4971. DOI: 10.1182/blood-2015-09-672352. eprint: https://ashpublications.org/blood/article-pdf/127/18/2203/1392993/2203.pdf. [Online]. Available: https://doi.org/10.1182/blood-2015-09-672352.

[30] I. Leshchiner, E. A. Mroz, J. Cha, *et al.*, "Inferring early genetic progression in cancers with unobtainable premalignant disease," *Nature Cancer*, vol. 4, no. 4, pp. 550–563, Apr. 2023, Publisher: Nature Publishing Group, ISSN: 2662-1347. DOI: 10.1038/s43018-023-00533-y. [Online]. Available: https://www.nature.com/articles/s43018-023-00533-y (visited on 05/09/2024).