6. Discussion

Here, we have provided the largest sequencing study on B-NHLs to date, proposing a novel genetics-based classification and profiling the mutational landscapes of FL, BL, and DLBCL with greater resolution than previously described.

6.1. Genomic Landscape and Gene Level Analysis

Our genomic landscape analysis for DLBCL NOS, FL, and BL was largely consistent with literature expectations but provided additional resolution due to the size and depth of our study. DLBCL NOS, FL, and BL all exhibited classic long tail distributions although DLBCL NOS in particular showed the greatest heterogeneity: the most recurrently mutated genes in DLBCL NOS accounted for a lower fraction of the overall mutations than those in FL and BL. Such a result was consistent with our later classification finding in which Class 5 contained 85 distinguishing genes all rarely mutated, indicating high heterogeneity. Because of the scope of our study, we also identified a variety of novel driver mutations, some rare, occurring across the 292 genes in our study.

Additionally, our landscape analysis found a small number of genes that showed a high mutation frequency across DLBCL NOS, FL, and BL (i.e. *KMT2D, CREBBP, TNFRSF14, TP53, SOCS1, B2M, ARID1A, CCND3, TNFAIP3, IRF8*). These mutational similarities initially pointed to the need for similar pathway dysregulations for B-NHLs to progress. By contrast, however, the only gene that was commonly mutated across all classes in our classification analysis was *TP53*. The difference in these results demonstrates that genetic classification can more accurately distinguish classes than histology; and importantly, can resolve pathway differences that demarcate patients into classes that have consistent pathway mutations that are largely non-overlapping.

Our mutation analysis demonstrated that patients, regardless of B-NHL condition, generally have 3-4 driver mutations. This insight, combined with the later classification description of co-mutation within classes, shows that multiple pathways tend to be dysregulated within B-NHLs and DLBCL. As a result, oncogenic cooperation may be occurring to, for example, increase proliferation while also evading the immune system. The presence of multiple driver mutations increases the complexity of pathogenesis and also classification. Rather than single genes demarcating novel classes, combinations of genetic mutations distinguish patients. As a result, far more possibilities exist and heterogeneity similarly increases.

At the gene level, we found genes broadly falling into oncogenic and tumour suppressor mutation profiles as expected and identified the presence of expected mutational processes such as aberrant somatic hypermutation. More interestingly, we identified clusters of disrupting mutations in specific gene domains that we suspect caused gains in function and thus allowed oncogenic activity. The specific mechanisms and mutations had not, to our knowledge, been previously reported for B-NHLs. For example, we observed a high proportion of frameshift and missense mutations in the death domain of the *FAS* gene, which generally initiates a caspase cascade leading to apoptosis. We suspect the inactivation of the *FAS* domain improves tumour cell survival. Similarly, we found a high number of frameshift and nonsense mutations in the SMAD/FHA domain of *IRF8* which we suspect could cause a gain in function that prevents apoptosis. In *SGK1*, we found a series of essential splice site mutations affecting a single exon, causing a likely gain in function and flagging that exon's importance in *SGK1* regulation. None of the above mechanisms, to our knowledge, had been previously reported in the context of DLBCL or B-NHLs.

6.2. Classification

Our classification system resolved seven distinct categories of B-NHLs, successfully separating FL, BL, and DLBCL while simultaneously highlighting the inherent heterogeneity of DLBCL. Compared to the WHO classification, we demonstrated significant heterogeneity and potential for further resolution within given subtypes. Indeed, patients marked as DLBCL NOS patients by the WHO classification were present in all seven classes identified here, indicating the necessity for further resolution.

We cannot directly compare our work to the cell of origin classification due to the absence of gene expression data from our dataset, however, Class 2 shared genetic characteristics largely consistent with ABC-DLBCL. The future addition of gene expression data to our study will allow us to directly compare our classification with the cell of origin classification. Crucially, we will be able to answer whether or not cell of origin can be distinguished on the basis of genetic mutations alone. If so, our approach could become an important surrogate for gene expression profiling as a way of determining cell of origin, which has already shown clinical relevance with the ABC-DLBCL group responding differently to targeted treatments than the GCB-DLBCL group.

Overall, DLBCL shows a high heterogeneity compared to other cancers. Unlike similar genetic classification schemes, such as that for AML, DLBCL presented a category with a larger number of rarely mutated genes (Class 6). The separation of these rarely

mutated genes into their own class rather than their presence within other classes points to the increased heterogeneity of DLBCL compared to other cancers. Indeed, the large number of potential driver lesions that can cause cancer within this category point to the potential for pathogenesis in a variety of different ways. Each likely follow distinct mechanisms and effective resolution of this class would require substantially higher sample sizes in order to create additional subcategories. Such heterogeneity reinforces the distinct clinical responses to treatments and the need for classification to resolve such differences.

Our classification approach additionally demonstrated its ability to resolve patients who had likely transformed. The first example was the identification of Class 1 patients, a class with hallmark mutations for FL, that were diagnosed by our clinicians as having DLBCL. Since the transformation of FL into DLBCL is well documented, such a result was expected and consistent with the literature. More surprising, however, was the fact that Class 5, consisting primarily of DLBCL and BCL, Int. patients demonstrated hallmark mutations of SMZL, likely corresponding to patients that had transformed from SMZL. Crucially, only a genetic classification approach of this sort – not histology alone – could identify the root disease from which DLBCL had transformed. Biologically, our result reinforces the possibility of SMZL to transform into DLBCL, which had been previously reported but rarely[185]. Clinically, it could suggest that Class 5 Patients have a distinct pathogenesis and thus may respond differently to novel treatments compared to other DLBCL subtypes.

Overall for aggressive diseases such as DLBCL which often transform from indolent cancers, the ability to distinguish the original genetic mutations that led to cancer could substantially affect patient outcomes. We expect our approach, therefore, to generalize across other cancers, identify additional indolent diseases and their transformation pathways, and flag patients which may respond more effectively to distinct regimens.

Our classification is based on causal genetic changes, and as a result, is likely to be durable, reproducible, and clinically relevant. We note that while treatments and clinical practices may change over time, improving the survival of DLBCL and B-NHL patients, the underlying genomic changes causing B-NHLs will remain consistent. Therefore, our classification represents fundamentally different pathogenesis mechanisms inherent to DLBCL and captures lasting biological information. With the addition of translocation, copy number, and gene expression data in a follow up study, this classification will additionally gain resolution, accuracy, reproducibility, and clinical relevance.

6.3. Comparison to Recent Large Scale DLBCL Genomics Study

Recently, Reddy *et al.* published an integrative analysis of 1,001 DLBCL samples that complements the results of this manuscript[15]. Whole exome sequencing, transcriptomics, copy number analysis, and FISH tests were conducted . Additionally, 400 of the samples had paired normals. In comparison, our study conducted targeted sequencing, transcriptomics, copy number analysis, and FISH tests on 962 DLBCL samples without paired normals. The targeted sequencing has been completed, and the outcomes of the remaining analyses are being processed by collaborators. The complementarities between our studies enable synergies to refine genomic analysis and classification of DLBCL.

First, Reddy *et al.*'s genomic analysis is generally consistent with this work. The genes in our study with the highest number of driver mutations were generally consistent with Reddy *et al.*'s list of frequently mutated genes with a few exceptions discussed in Section 4.1.2. A few other notable differences exist. Reddy *et al.* conducted whole exome sequencing rather than targeted sequencing of genes. Whole Exome Sequencing allows Reddy *et al.* to identify driver genes with previously unreported pattern of mutations, something not possible through our targeted study. Indeed, a few of the 150 genes identified as drivers are not present within our bait set (*DUSP2, ZNF608*, and *BIRC6*) and we thus do not report variants in these genes. Conversely, our targeted sequencing study also uncovered genes and specific mutations not present in Reddy et al.'s study. For example, we found splicing errors in *SGK1* which were not reported by Reddy et al.'s work. Therefore, we see these studies as complementary. A meta-analysis involving both sets of variants would prove helpful to fully understanding the genomic changes underlying DLBCL.

Second, Reddy *et al.* take a distinct approach to DLBCL classification. Reddy *et al.* classify patients on the basis of gene expression patterns. As a result, they can identify functional signatures based on gene expression such as the Monti Host Response signature. Conversely, our study classifies DLBCL on the basis of genetic lesions. Therefore, we can identify patterns at the genetic level such as our Class 5 which is suspected to contain SMZL patients. Ultimately, future work could seek to simultaneously incorporate both gene expression patterns and genetic lesions as the basis for classification. Therefore, both types of findings could be drawn out from the clusters. Note that this is distinct from Reddy *et al.*'s work which first generated a classification based on gene expression and subsequently identified the genetic alterations associated with each cluster.

In spite of the distinct approach to classification, some commonalities were observed. Namely, our genetic classification identified *MYD88* and *CDKN2A* as defining Class 2 which we suspect to be primarily composed of ABC-DLBCL. Both of these genes had more

genomic alterations in the ABC-DLBCL expression cluster of Reddy *et al*.'s work than in other clusters.

Third, Reddy *et al.*'s study also conducted a functional CRISPR screen and created a prognostication model with implications for our study. First, the CRISPR screen only identified 35 of the 150 driver genes Reddy *et al.* had initially flagged as having functional relevant to DLBCL cell lines. This result reinforces the need to biologically validate the driver variants we have discovered. Second, Reddy *et al.* created a prognostication model that outperformed the R-IPI by using only genetic and molecular features. The prognostication first enumerates all combinations involving up to 4 distinct genetic and molecular features and affecting at least 20 patients. These 313 combinatorial features are then fed into an Elastic regression. We hope to make two improvements when developing a similar prognostication model for our dataset. First, we hope to use a more robust feature selection method such as bootstrapping or stepwise regression. Second, we hope to include additional clinical characteristics into the set of regression features. Indeed prior work for AML[143] has shown that clinical variables often have even more predictive power than genetics[143,186]. A regression model incorporating both may provide more accurate classification.

Finally, the union of these works could provide validation for both studies. Comparison of genomic variants could validate pipelines and drivers in both studies. Testing whether Reddy *et al.*'s cohort classifies into similar genetic clusters as ours could validate our genetic classification. Finally, testing Reddy *et al*.'s prognostication tool on our cohort could validate its generalizability.

6.4. Future Work

While the aforementioned project describes the genetic landscape and provides a genetic classification of various B-NHL malignancies, substantial additional potential exists.

6.4.1. Incorporating Copy Number Analysis, Gene Expression, and Translocation Data

First, the incorporation of copy number analysis, gene expression, and translocation information will add to both the pathogenesis insights derived from this project as well as the resolution of classification. Crucially, both copy number amplifications/deletions and translocations are well known to affect progression of B-NHLs while also providing subtype-differentiating lesions. Current work is underway implementing a custom algorithm to extract copy number from this targeted, unmatched dataset. Similarly, translocation data from collaborators is currently being processed and will be added. Once incorporated, our study

will be one of the two largest and most complete genetic analyses of B-NHL, and DLBCL in particular, ever conducted, thereby enabling new insights regarding causality, molecular progression, and differentiating feature of each disease. Moreover because specific copy number and translocation changes are known to predominantly present in specific subtypes (i.e. *MYC* translocation in BL), the incorporation of such data will draw sharper divisions between classes of our classification and potentially define entirely new classes.

Second, the incorporation of gene expression data in particular will allow us to compare our classification to the cell-of-origin classification currently leading the literature. By providing additional differentiating information (i.e. genetic mutations, copy number changes, and translocations), our dataset will be able to refine the cell-of-origin categories currently based purely on gene expression. Importantly, our study may also be able to define whether ABC and GCB DLBCL are indeed distinct entities or whether information inherent to genetic mutations rather than gene expression provide more convincing differentiation among DLBCL subtypes. Finally, by adding additional genetic information to the samples classified via the cell-of-origin classification, our analysis will provide mechanistic insight into the pathogenesis of GCB-DLBCL in particular whose pathogenesis is presently unknown[19].

6.4.2. Survival Analysis for Classification

Only a preliminary survival analysis was conducted to understand the distinct clinical courses of the identified classes within this study. A full survival analysis would additionally correct here for age, date of diagnosis, centre, treatment, and a variety of other variables. Such corrections are especially critical because our study incorporates samples taken over 15 years. The introduction of CHOP and subsequently R-CHOP therefore occurred within the time window of our study and the substantially improved outcomes for patients receiving these treatments versus previous ones must be accounted for. Similarly, improvement in general clinical treatment must also be accounted for.

Such a survival analysis could generate crucial clinical insights. By distinguishing which subclasses of DLBCL and the other B-NHL malignancies presented here both (1) exhibit the worst clinical course and (2) are the least likely to respond to treatment, we may be able to identify the subset of patients which should be moved toward more aggressive treatments such as stem cell transplantations and considered for experimental therapies. Moreover, by specifically conducting this analysis on the subset of DLBCL patients that respond poorly to an R-CHOP regimen vs. those that respond well to an R-CHOP regimen,

we will hopefully be able to delineate the causative genetic and molecular differences that prevent cure in 30% of DLBCL cases. If we are able to sufficiently distinguish these patients, additional studies could then fully characterize their distinct pathogenesis, leading to suggestions for new treatments and therapies that will help them. Additionally, such a survival analysis could be coupled with a survival analysis for the specific genetic lesions that are most deleterious. By identifying such lesions, both within given classes and across all classes, we would be able to more effectively identify the patients with the most aggressive clinical course and subsequently shift them onto more intensive therapies and potentially experimental clinical trials.

6.4.3. Validation of M7-FLIPI Prognostication Tool for FL

Our dataset could validate the M7-FLIPI prognostication tool for FL. M7-FLIPI seeks to risk stratify FL patients receiving first-line immunochemotherapy by considering their mutations in seven genes (*EZH2, ARID1A, MEF2B, EP300, CREBBP,* and *CARD11*), their Follicular Lymphoma International Prognostic Index (FLIPI), and their Eastern Cooperative Oncology Group performance status (ECOG)[187,188]. Our dataset contains 337 FL patients which were treated and 222 which were placed under a "watch and wait" regimen (Figure 1b). All samples were diagnostic biopsies, and all of these FL patients have the relevant genetic, clinical, and survival data required to utilize the M7-FLIPI prognostication tool. Once the appropriate clinical data is processed to subset treated patients based on the treatments they receive, we believe our dataset will be sufficiently large to validate the M7-FLIPI prognostication tool.

6.4.4. Prediction of Treatment Outcomes Based on Genetics

Finally, future work will focus on providing a machine learning based approach to improve the prognostication of DLBCL patients. The gold standard clinical prognostic tool, the Revised International Prognostic Index (R-IPI), sorts patients into three risk groups based on factors such as age and whether their lactate dehydrogenase level is elevated.[14] None of the R-IPI factors, however, account for the genetic basis of DLBCL and cannot therefore incorporate prognostic information from genetic variability between patients within the same risk group. Virtually all DLBCL patients receive the same first-line therapy, R-CHOP, despite the probability that the genetic and biological heterogeneity will result in heterogeneous response to the potential treatments available.[189] By utilizing a machine learning based approach that considers all possible lesions as well as clinical variables, we

may be able to more effectively predict which patients are likely to respond well to R-CHOP and which are not. If such an identification is possible, the patients at greater risk may be moved toward more aggressive treatments or experimental therapies.