1. Introduction

1.1. Classifying cancer, a deeply heterogeneous disease

Cancer is an extremely heterogeneous disease, showing distinct clinical and biological manifestations between cancer types, within subtypes, and even between patients with the same subtype. Such heterogeneity results from the pathogenesis of cancer: as somatic mutations accumulate over time, in a myriad of genes and tissues, a variety of pathways are dysregulated leading to cell proliferation. Patients of the same cancer type may carry distinct causative mutations. Indeed, different tumour cells within a patient may also carry distinct causative mutations. Overall, the myriad combinations of genetic mutations targeting distinct genes, cells, and tissues generate different clinical courses, survival likelihoods, and treatment responses between patients.

To deal with such heterogeneity, classification schemes have been developed. By grouping patients according to common characteristics, broad patterns emerge with patients sorted according to common prognoses and responses to treatments. Historically, such classification has relied on histological, morphological, and immunohistochemical examination of the patient's tumour cells. Such an approach, however, is lacking in a few respects. First, different cancer types have been shown to share similar histological, morphological, and immunohistochemical characteristics in spite of having distinct genetic causes and treatment responses. As a result, traditional classification systems often fail to resolve categories at a high enough level precisely because they do not incorporate the causative genetic changes leading to disease. Second, resulting classes are often difficult to interpret in the context of the pathways distinguishing diseases, making translation to therapy more challenging. Indeed, a distinct morphological profile does not immediately suggest a new therapeutic target. Thus, even when a new class is demarcated, it is often challenging to directly improve its clinical course. Finally, the clinical insights of some distinct classes have struggled with widespread relevance and reproducibility. For example, DLBCL was traditionally classified according to centroblastic, immunoblastic, and anaplastic subtypes with distinct clinical courses. Such clinical differences, however, have struggled with reproducibility. Additionally, the morphological subtype with the worst clinical course (anaplastic) has shown to occur in only 7.4% of cases, making widespread clinical relevance poor[1].

With the advent of more readily available patient samples and cheap sequencing, classification schemes have been shifting toward resolving cancer on the basis of molecular and genetic differences. Throughout, blood cancers have led the way. Indeed, Chronic Myeloid Leukemia began with morphological characterization[2–4] which then gave way to the Philadelphia Chromosome and the BCR-ABL mutation as the primary classification characteristics[5]. Acute Myeloid Leukaemia then followed with the first identification of a specific genetic subtype: Acute Promyelocytic Leukaemia[6–9]. Both of these categories of disease, defined by their canonical genetic lesion, now have specific targeted therapies against this genetic change, radically improving treatment outcomes for those patients. In solid tumours, Ewing's Sarcoma was defined by a t(11;22) translocation[10]; breast cancer became defined by *ERBB2*[11,12]; and non small cell lung cancers are increasingly defined by specific kinase mutations[13].

Broadly, genetic and molecular classification approaches share a series of advantages over traditional approaches. First, these classifications rely on the causative genetic and molecular changes that underlie cancer. As a result, they are more likely to be clinically relevant, durable, and reproducible. Even as treatments change, for example, the underlying genetic structure of cancers are likely to remain the same. Second, genetic classifications group patients on the basis of pathways rather than morphology, leading to improved biological insights. By extracting the unique pathways that distinguish patient groups, the pathogenesis of distinct cancers become clearer. Finally, genetic classifications can improve clinical prognostication and suggest therapeutic targets. Targeted therapies inhibiting a specific gene that defines a genetic class can be reserved exclusively for patients of that class, improving treatment selection. Similarly, when a new patient class emerges that is resistant to traditional therapies, the pathway dysregulations allowing such resistance can be examined and new target combinations can be suggested.

1.2. A purely genetic classification for DLBCL

While an effective classification scheme could benefit all cancers, it could especially benefit DLBCL. Compared to other cancers, DLBCL exhibits a higher degree of genetic heterogeneity since it derives from Germinal Centre B cells which often have unstable genomes. Additionally, an effective classification could immediately help clinical outcomes. 30% of DLBCL patients today are not cured by R-CHOP, the front line chemotherapeutic treatment. These patients subsequently relapse upon which their prognosis suffers significantly. At present, there is no way to pre-emptively identify these patients in spite of the fact that they likely exhibit genomic differences that prevent effective R-CHOP treatment. A classification system that identifies these patients would enable physicians to move them toward more aggressive clinical regimens such as stem cell transplantation or experimental therapies. It could also help develop more targeted clinical trial protocols, in which only those patients likely to relapse are recruited.

In this study, we propose a novel classification scheme for B-NHLs and DLBCL based purely on genetic changes. By conducting targeted deep sequencing of 1607 B-NHL patients and subsequently classifying these patients on the basis of genetics alone, we: (1) identify novel mutation patterns such as the aberrant splicing of an exon in *SGK1*, (2) produce the first ever purely genetic classification of B-NHLs broadly and DLBCL in particular, (3) unlock previously unknown patterns of co-mutation which shed light on unique pathogenesis mechanisms, (4) identify novel subclasses of DLBCL, including one with hallmark SMZL mutations, revealing new insights regarding DLBCL pathogenesis, and (5) set the stage for a follow up clinical study examining the unique lesions that give 30% of DLBCL patients poor R-CHOP responses[14], thus shedding light on the critical clinical question of DLBCL.

Our study occurs in three main stages (Figure 1a). First, we identify driver mutations in 292 genes implicated in lymphoid and myeloid malignancies across 1607 patients. Second, we conduct mutational analysis at the landscape level and at the gene-level for DLBCL, FL, and BL – the primary B-NHLs included in our study. Finally, we utilize Bayesian Dirichlet Processes – a machine learning classification approach – to classify our samples on the basis of genetics alone.

Our study draws its effectiveness from its depth and size. We sequence 1607 total patients spread across a range of B-NHL subtypes, with the largest patient populations for DLBCL and FL (Figure 1b). Our study is one of only two studies of such scope[15] and is roughly 10X larger than all other previous DLBCL and B-NHL genetic sequencing studies, allowing us to consider more B-NHL subtypes. Additionally, our targeted sequencing approach allows us to sequence at greater depth, thus identifying rarer and clinically useful variants previously missed. Combined, such scope and scale finally allows us to use Bayesian Dirichlet Processes – a machine learning approach that can effectively delineate co-mutation patterns with a sufficiently large dataset. While we apply this approach to DLBCL and B-NHLs in this study, the broad methodology should hold equally for other cancers. As a result, we see this as a foundational study for a new paradigm in cancer classification. Additionally, upon further work which will incorporate gene expression data, copy number changes, and translocation data, we will be able to (1) compare our classification robustly with the cell of origin classification based on gene-expression profiling, potentially providing a surrogate and (2) present the most integrative classification scheme to date.

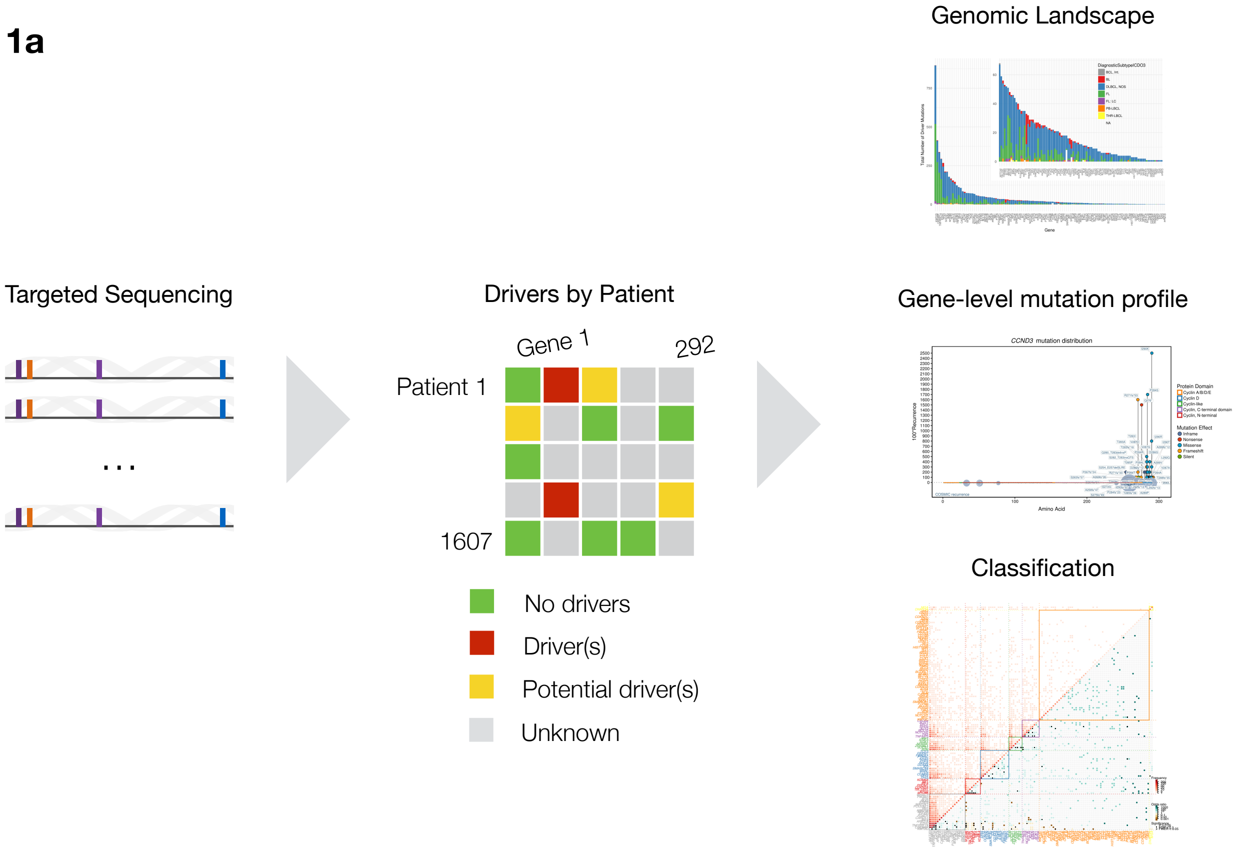

Figure 1 Overview of Study. (a) Process Overview. Targeted Sequencing of 292 genes was conducted on 1607 lymphoma samples. Subsequently, variants were called, filtered into somatic mutations, and annotated as drivers or passengers. Finally, three analyses were conducted investigating the genomic landscape of B-NHLs, examining the mutation profiles of crucial lymphoma genes, and creating the first ever purely genetic classification of B-NHLs and DLBCL in particular. **(b)** Patient Cohort Overview.

1b

	Overall	Sex		Age	OS	Treatment				Survival Status	
	n (% total)	F (% subtype)	M (% subtype)	Median (Range)	Median (Range)	Not known (% subtype)	Not treated (% subtype)	Treated (% subtype)	Watch and wait (% subtype)	0 (% subtype)	1 (% subtype)
Total	1607 (100%)	771 (47.9%)	832 (51.7%)	66 (3, 98)	2112 (-19, 4655)	69 (4.2%)	125 (7.7%)	1185 (73.7%)	224 (13.9%)	871 (54.2%)	732 (45.5%)
Diagnostic Group WHO											
DLBCL-HL intermediate	15 (0.9%)	6 (40%)	9 (60%)	54 (11, 81)	2155 (15, 2533)	0 (0%)	0 (0%)	15 (100%)	0 (0%)	12 (80%)	3 (20%)
Burkitt Lymphoma	39 (2.4%)	6 (15.3%)	33 (84.6%)	38 (3, 86)	2023 (-1, 4623)	1 (2.5%)	7 (17.9%)	31 (79.4%)	0 (0%)	24 (61.5%)	15 (38.4%)
Diffuse large B-cell lymphoma	962 (59.8%)	445 (46.2%)	517 (53.7%)	69 (8, 98)	1819 (-19, 4655)	46 (4.7%)	114 (11.8%)	802 (83.3%)	0 (0%)	439 (45.6%)	523 (54.3%)
Follicular lymphoma	587 (36.5%)	314 (53.4%)	273 (46.5%)	64 (20, 98)	2493 (1, 4655)	22 (3.7%)	4 (0.6%)	337 (57.4%)	224 (38.1%)	396 (67.4%)	191 (32.5%)
Diagnostic Subtype ICDO3											
DLBCL-HL intermediate	15 (0.9%)	6 (40%)	9 (60%)	54 (11, 81)	2155 (15, 2533)	0 (0%)	0 (0%)	15 (100%)	0 (0%)	12 (80%)	3 (20%)
Burkitt lymphoma	39 (2.4%)	6 (15.3%)	33 (84.6%)	38 (3, 86)	2023 (-1, 4263)	1 (2.5%)	7 (17.9%)	31 (79.4%)	0 (0%)	24 (61.5%)	15 (38.4%)
Diffuse large B-cell lymphoma, NOS	925 (57.5%)	430 (46.4%)	495 (53.5%)	69 (8, 98)	1824 (-19, 4655)	44 (4.7%)	107 (11.5%)	774 (83.6%)	0 (0%)	422 (45.6%)	503 (54.3%)
Follicular lymphoma	566 (35.2%)	305 (53.8%)	261 (46.1%)	64 (20, 98)	2477 (1, 4655)	22 (3.8%)	4 (0.7%)	318 (56.1%)	222 (39.2%)	378 (66.7%)	188 (33.2%)
Follicular lymphoma: large cell	21 (1.3%)	9 (42.8%)	12 (57.1%)	57 (37, 84)	3313 (88, 4589)	0 (0%)	0 (0%)	19 (90.4%)	2 (9.5%)	18 (85.7%)	3 (14.2%)
Intravascular large B-cell lymphoma	1 (0%)	0 (0%)	1 (100%)	71 (71, 71)	2232 (2232, 2232)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	1 (100%)	0 (0%)
Plasmablastic large B-cell lymphoma	14 (0.8%)	5 (35.7%)	9 (64.2%)	71 (18, 95)	426.5 (2, 3379)	1 (7.1%)	4 (28.5%)	9 (64.2%)	0 (0%)	3 (21.4%)	11 (78.5%)
T-cell/histiocyte-rich large B-cell lymphoma	22 (1.3%)	10 (45.4%)	12 (54.5%)	65.5 (30, 89)	1836 (7, 2782)	1 (4.5%)	3 (13.6%)	18 (81.8%)	0 (0%)	13 (59%)	9 (40.9%)