3. Methods

Two phases composed this study. First, driver variants were extracted from genetic calls and merged with relevant clinical data. Second, this joint clinical and genetic data was harnessed to create a novel genetic classification of DLBCL.

3.1. Dataset

3.1.1. Patient Cohort

Patient samples came from the Haematological Malignancy Research Network (HMRN), a UK population-based registry whose methods have been previously described[131,132]. In short, fresh frozen or formalin-fixed, paraffin-embedded (FFPE) tissue samples were collected from 1607 lymphoma patients over 15 years. All samples collected were diagnostic biopsies. DNA was subsequently extracted for sequencing. Patient characteristics are available in Figure 1b.

Since patient samples were collected over 15 years, around 90% of curatively treated patients received rituximab. At the time of this manuscript, the information relating which patients did and did not receive rituximab was not yet processed and transferred to us by our collaborators. This will primarily affect the survival analysis at the end of this study, which is marked as being preliminary and will be heavily revised in future versions of this work. The 11% of DLBCL NOS patients marked as “not treated” were treated with palliative intent.

3.1.2. Library Preparation and Sequencing

Genetic sequencing targeted the exon region of 292 genes, specific SNPs in noncoding regions to facilitate copy number analysis, and known hot spot mutations outside of exon regions. Custom RNA baits were designed according to manufacturer guidelines (Agilent). Genomic DNA (125uL, 40ng/uL) was fragmented and prepared for Illumina DNA library sequencing via a Bravo automated liquid handler. Prepared samples were then indexed to a unique DNA barcode with 6 cycles of PCR. Next, the Agilent SureSelect protocol was used to prepare and hybridize 16 equimolar pools of libraries to custom RNA baits. RNA baits were designed to target the exons of 292 genes implicated in lymphomas and myeloid cancers. Additionally, baits targeted a series of SNPs in non-coding regions to allow later extraction of copy number changes. Finally, an Illumina HiSeq machine with a 75-base pair paired-end protocol was used to sequence enriched pools of 96 cases.

3.1.3. Clinical Data

The following clinical data was collected for all patients: sex; age at diagnosis, WHO Diagnostic Group: Diffuse large B-cell lymphoma, Follicular lymphoma, Burkitt lymphoma, B-cell lymphoma (intermediate between DLBCL and classical HL; Diagnostic Subtype ICDO3: Diffuse large B-cell lymphoma (NOS); Follicular lymphoma, Burkitt lymphoma, Intravascular large B-cell lymphoma, Follicular lymphoma: large cell, Plasmablastic large B-cell lymphoma, T-cell/histiocyte-rich large B-cell lymphoma, B-cell lymphoma (intermediate between DLBCL and classical HL); overall survival: days since pathology report; survival status; and treatment: treated, not known, watch and wait. Additional clinical variables were also collected and are currently being processed by our collaborators.

3.2. Genetic Data Preparation

3.2.1. Sequencing Alignment

To align raw sequencing data to the human genome (NCBI Build 37), the BWA algorithm[133] was used. The coverage depth at each base-pair position was determined utilizing Bedtools® v2.15.0[134]. Sequencing was performed to an average target depth of 500x reads per base, although there was inevitably patient-to-patient and gene-to-gene variation around this target.

3.2.2. Variant Calling

DLBCL includes a spectrum of genetic mutations including indels, complex rearrangements, and point mutations. We utilized two approaches to call relevant variants. First, point mutations were called using a modified version of the CaVEMan [135]algorithm with a single cord blood sample designated as the normal (Cancer Variants through Expectation Maximisation, https://github.com/cancerit/CaVEMan). CaVEMan calls variants by comparing sequencing data from each tumour sample with a designated normal sample and then calculating the likelihood of a mutation at each base-pair position locus. Thereby, CaVEMan identifies point mutations. Second, indel mutations were called using a modified version of the Pindel algorithm[136]. Third, Samtools mpileup was utilized to specifically identify mutations in known hotspot regions[137] like the *TERT* promoter. Finally, we manually reviewed all remaining variants using a genome browser (Gbrowse®)[138].

3.2.3. Variant Filtering

After calling the full set of variants, we removed off-target variants and variants that were suspected errors. These variants were removed based on (1) their presence in an off-target region, (2) a set of standard CaVEMan filters, (3) a set of standard Pindel filters, (4) a manually implemented set of additional filters, and (5) manual review.

First, we removed unmapped reads, PCR duplicates, and variants in off-target regions. Off-target variants were removed using Bedtools v2.15.0[134].

Second, variants were removed based on the CaVEMan filters below:

1. DTH: Less than 1/3 of mutant alleles were >= 25 base quality
2. RP: Coverage was less than 8 and no mutant alleles were found in the first 2/3 of a read (shifted 0.08 from the start and extended 0.08 more than 2/3 of the read length)

3. MN: More than 0.03 of mutant alleles that were >= 15 base quality found in the matched normal
4. PT: mutant alleles all on one direction of read (1 read allowed on opposite strand) and in second half of the read. Second half of read contains the motif GGC[AT]G in sequenced orientation and the mean base quality of all bases after the motif was less than 20
5. MQ: Mean mapping quality of the mutant allele was <21
6. SR: Position falls within a simple repeat using the supplied bed file
7. CR: Position falls within a centromeric repeat using the supplied bed file
8. PH: Mutant reads were on one strand (permitted proportion on other strand: 0.04) and mean mutant base quality was less than 21
9. TL: More than 10 percent of reads covering this position contained an indel according to mapping
10. SRP: More than 80 percent of reads contain the mutant allele at the same read position
11. HSD: Position falls within a high sequencing depth region using the supplied bed file
12. AN: Position could not be annotated against a transcript using the supplied bed file
13. VUM: Position has >= 3 mutant alleles present in at least 1 percent unmatched normal samples in the unmatched VCF
14. SE: Coverage is >= 10 on each strand but mutant allele is only present on one strand
15. MNP: Tumour sample mutant allele proportion – normal sample mutant allele proportion < 0.2

Third, indel variants were removed based on the filters built into Pindel[136].

Fourth, we removed additional variants based on manual filters. To remove variants within the error limits of CaVEMan and Pindel, we removed: variants with a read depth less than 10, variants with less than 3 reads, and variants with a variant allele fraction less than 0.05. To remove variants due to polymerase slippage in homopolymeric regions of the genome, we removed variants with a repeat length greater than 4 that also occurred in over 10% of individuals. Finally, we removed variants with insufficient read depth (<10). For reference, the average read depth across our study was ~500x reads per base.

3.2.4. Driver Identification

In order to identify driver mutations within the set of mutant calls, we first removed suspected germline polymorphisms. Next, we executed a pipeline for automated driver annotation. Finally and crucially, we reviewed all annotations manually before marking variants as Drivers, Passengers, or Variants of Unknown Significance.

First, we removed suspected germline polymorphisms by annotating the variants according to their population frequency in ExAC non-TCGA v0.3[139]. Any variants with a population frequency in ExAC non-TCGA > 0.001 were considered likely germline polymorphisms. While ExAC non-TCGA is contaminated with some relatively common somatic driver mutations, we reduced the risk of mistakenly removing common drivers by keeping a whitelist of common driver mutations and also examining suspected somatic mutations during the manual review step.

Next, we executed a pipeline for automated driver annotation. In order to be considered a driver, a variant must:

1. Not have a Vagrent [140]annotated mutation effect of the following type: THREE_PRIME_UTR, FIVE_PRIME_UTR, FIVE_PRIME_FLANK, THREE_PRIME_FLANK, INTRONIC, SPLICE_REGION, SILENT.
2. While also fulfilling any of the four conditions below:
    a. In a whitelist of well known driver mutations;
    b. Recurrence in COSMIC v82[141] > 3;
    c. Recurrence in COSMIC subsetted to hametopoetic and lymphoid diseases > 3;
    d. Likely to be a driver mutation based on it's effect and presence in a known tumour suppressor gene or known oncogene. As an example, a truncating mutation in a tumour suppressor gene would be considered a likely driver via this process.

Finally, a manual review process triaged suspected passengers and suspected passengers into two final categories of drivers and passengers. Beyond manually reviewing all annotations, this step was particularly important for removing missense variants that were only recurrent because of Somatic Hyper Mutation and not otherwise expected to be drivers.

3.3. Classification

3.3.1. Classification Techniques

To separate DLBCL patients into maximal, non-overlapping clusters, we utilized Bayesian Dirichlet Processes[142]. Bayesian Dirichlet Processes utilize a mixture model with an infinite prior distribution for the proportion and number of clusters. A Markov chain Monte Carlo method is then used learn the number, proportion, and assignments of the clusters. Analysis relied on the R package https://github.com/nicolaroberts/hdp which implements the non-hierarchical Dirichlet process we used. To fit the data, we used 100,000 burn-in iterations and 20,000 samples at 60 iterations between samples. After fitting the data, we merged clusters more than 5% similar on a cosine similarity metric and requested that only 99% of the data require explanation. Relevant code was adapted from a prior AML study by Papaemmanuil et al.[143]

3.3.2. Statistical Analysis

R version 3.3.3 was used for all statistical analysis and visualization.