Characterizing Genetic Drivers of Lymphoma through High-Throughput Sequencing

by

Jenny Zhang

University Program in Genetics and Genomics
Duke University

Date:_______________________

Approved:

___________________________

Sandeep Dave, Supervisor

___________________________

Jen-Tsan Ashley Chi

___________________________

Alexander J. Hartemink

___________________________

Micah A. Luftig

Dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy in the
University Program in Genetics and Genomics
in the Graduate School
of Duke University

2016
ABSTRACT

Characterizing Genetic Drivers of Lymphoma through High-Throughput Sequencing

by

Jenny Zhang

University Program in Genetics and Genomics
Duke University

Date:_______________________
Approved:

___________________________
Sandeep Dave, Supervisor

___________________________
Jen-Tsan Ashley Chi

___________________________
Alexander J. Hartemink

___________________________
Micah A. Luftig

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University Program in Genetics and Genomics in the Graduate School of Duke University

2016
Abstract

The advent of next-generation sequencing, now nearing a decade in age, has enabled, among other capabilities, measurement of genome-wide sequence features at unprecedented scale and resolution.

In this dissertation, I describe work to understand the genetic underpinnings of non-Hodgkin’s lymphoma through exploration of the epigenetics of its cell of origin, initial characterization and interpretation of driver mutations, and finally, a larger-scale, population-level study that incorporates mutation interpretation with clinical outcome.

In the first research chapter, I describe genomic characteristics of lymphomas through the lens of their cells of origin. Just as many other cancers, such as breast cancer or lung cancer, are categorized based on their cell of origin, lymphoma subtypes can be examined through the context of their normal B Cells of origin, Naïve, Germinal Center, and post-Germinal Center. By applying integrative analysis of the epigenetics of normal B Cells of origin through chromatin-immunoprecipitation sequencing, we find that differences in normal B Cell subtypes are reflected in the mutational landscapes of the cancers that arise from them, namely Mantle Cell, Burkitt, and Diffuse Large B-Cell Lymphoma.

In the next research chapter, I describe our first endeavor into understanding the genetic heterogeneity of Diffuse Large B Cell Lymphoma, the most common form of
non-Hodgkin’s lymphoma, which affects 100,000 patients in the world. Through whole-genome sequencing of 1 case as well as whole-exome sequencing of 94 cases, we characterize the most recurrent genetic features of DLBCL and lay the groundwork for a larger study.

In the last research chapter, I describe work to characterize and interpret the whole exomes of 1001 cases of DLBCL in the largest single-cancer study to date. This highly-powered study enabled sub-gene, gene-level, and gene-network level understanding of driver mutations within DLBCL. Moreover, matched genomic and clinical data enabled the connection of these driver mutations to clinical features such as treatment response or overall survival. As sequencing costs continue to drop, whole-exome sequencing will become a routine clinical assay, and another diagnostic dimension in addition to existing methods such as histology. However, to unlock the full utility of sequencing data, we must be able to interpret it. This study undertakes a first step in developing the understanding necessary to uncover the genomic signals of DLBCL hidden within its exomes. However, beyond the scope of this one disease, the experimental and analytical methods can be readily applied to other cancer sequencing studies.

Thus, this dissertation leverages next-generation sequencing analysis to understand the genetic underpinnings of lymphoma, both by examining its normal cells
of origin as well as through a large-scale study to sensitively identify recurrently mutated genes and their relationship to clinical outcome.
Dedication

To my parents.
# Contents

Abstract ........................................................................................................................................ iv

List of Figures .................................................................................................................................. xii

Acknowledgements .......................................................................................................................... xviii

1. Introduction ................................................................................................................................... 20
   1.1 Background of next-generation sequencing ........................................................................... 20
   1.2 Application of techniques to B Cell Lymphomas ................................................................. 21
   1.3 Distinction of Subgroups of DLBCL by Gene Expression Profiling ........................................ 22

2. The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells ........................................................................ 24
   2.1 Introduction ............................................................................................................................ 24
   2.2 Results .................................................................................................................................... 26
      2.1.1 Exome sequencing reveals recurrent genetic alterations in mantle cell lymphoma ........ 26
      2.1.2 Defining the landscape of gene mutations in mantle cell lymphoma ............................ 27
      2.1.3 Genetic differences between mantle cell lymphoma and other lymphomas .......... 30
      2.1.4 Epigenetic profiling of normal B cells and its relationship to gene expression .......... 32
      2.1.5 Role of B-cell differentiation stage in mantle cell lymphoma ..................................... 35
      2.1.6 Epigenetic profiling of normal B cells and its relationship to mutations in lymphoma ... 36
   2.3 Discussion ............................................................................................................................... 38
   2.4 Methods ................................................................................................................................. 42
      2.4.1 MCL sample acquisition and processing ................................................................. 42
4. Integrative genomic and clinical analysis of 1001 diffuse large B cell lymphoma (DLBCL) patients reveals genetic drivers and risk groups .......................................................... 86

4.1 Introduction ............................................................................................................. 86

4.2 Results .................................................................................................................... 89

4.2.1 Exome sequencing overview ............................................................................. 89

4.2.1 Interesting clinical features ............................................................................... 89

4.2.2 Gene-level mutational patterns identify oncogenes vs. tumor suppressors .... 90

4.2.2.1 The mutational pattern of oncogenes ............................................................. 90

4.2.2.2 The mutational pattern of tumor suppressors.................................................. 95
4.2.3 Integrating mutational patterns and copy number alterations to identify recurrently altered genes in DLBCL ................................................................. 97

4.2.4 ABC/GCB logistic regression mutational classifier ........................................... 99

4.2.5 Identifying recurrently mutated pathways and their significance .................... 100

4.2.6 Association of gene expression, mutation, and copy number with DLBCL survival .................................................................................................................. 105

4.3 Discussion .............................................................................................................. 107

4.4 Methods ............................................................................................................... 107

4.4.1 Exome sequencing of DNA from FFPE samples ................................................. 108

4.4.1.1 Introduction to FFPE samples and motivation ............................................... 108

4.4.1.2 Parallel analysis of frozen and FFPE DLBCL samples from the same individual indicates high variant and mutation overlap ................................................. 109

4.4.1.3 Mutational spectra are identical between FFPE and Frozen cases ............ 111

4.4.1.4 Sequence depth and coverage ........................................................................ 112

4.4.1.5 Gene discovery using FFPE cases ................................................................. 113

4.4.1.6 Conclusions .................................................................................................. 115

4.4.2 High-throughput Exome Library Preparation .................................................... 115

5. Conclusions and future directions ........................................................................... 116

References ............................................................................................................... 119

Biography .................................................................................................................. 128
List of Figures

Figure 1: (a) Each column represents a patient and each row represents a gene. Mutations are color-coded with yellow for missense mutation, purple for frameshift mutation, red for nonsense mutation, and orange for in-frame insertion or deletion. (b) The bar graph indicates frequency of variants found by gene across all samples, subdivided by not-synonymous (blue) and synonymous (grey) mutations. .................................. 29

Figure 2: The network indicates functional groupings of the genes mutated in MCL. Nodes represent significantly mutated genes that are also a part of a significant functional group. Edges connect nodes that belong to the same functional gene set. Colored ovals identify the gene sets to which these nodes belong. .............................................. 30

Figure 3: The bar graph depicts the proportion of mutated cases that belong to each lymphoma type for mantle cell lymphoma (MCL), Burkitt lymphoma (BL), GCB diffuse large B cell lymphoma (DLBCL), and ABC diffuse large B cell lymphoma............................................ 31

Figure 4: Epigenetic profiles of H3K4me1, H3K4me3, H3K27 me3, H3Ac, H3K36me3, and PolII are shown in 50-bp read resolution from 2kb upstream to 10 kb downstream of all annotated transcription start sites. transcription start sites. ................................................................. 33

Figure 5: Epigenetic profiles of H3K4me1, H3K4me3, H3Ac, H3K36me3 and H3K27me3 around the transcription start sites (TSS) are shown for the top 10% most expressed, bottom 10% expressed, and all genes for naïve and germinal center B cells. ........................... 34

Figure 6: The heat map depicts the chromatin signal at gene level for H3K36me3, H3ac, and H3K4me1 open chromatin markers in genes with the most significant open chromatin differences between normal and germinal center B cells. On the rightmost heatmap, gene expression in mantle cell lymphoma samples (64 cases) and Burkitt lymphoma (23 cases) is indicated for the same genes. Orange indicates higher chromatin signal or higher gene expression; blue indicates lower................................................................. 36

Figure 7: The box plots on the left illustrate difference in open chromatin score for genes differentially mutated between BL and MCL. On the horizontal axis, MCL Mutated Genes is defined as genes with a significantly higher mutation rate in MCL compared to BL, and the reverse for BL Mutated Genes. The box plots on the right indicate expression fold change between normal germinal center and naïve B cells for the same two sets of genes. The size of the individual data points is scaled to the number of mutated cases in the corresponding disease. ........................................................................................................ 38
Figure 8: Results from sequencing a lymphoma genome. (A) The Circos diagram [58] summarizing the somatically acquired genetic variants in a diffuse large B cell lymphoma (DLBCL) genome. The outermost ring depicts the chromosome ideogram oriented clockwise, pter-qter. The next ring indicates copy number alterations in the DLCBL genome. The next 3 rings indicate somatically acquired mutations in intergenic regions, potential regulatory regions, and the exome respectively. (B) The pie chart depicts the relative number of somatically acquired mutations in the DLBCL genome, which can be classified by their genomic location as intergenic, intronic, potential regulatory or transcribed regions (left panel). The right panel depicts the breakdown of different mutation types observed in the transcribed regions. (C) The histogram depicts the mutation profile of DLBCL. The proportion of mutations in each of the 6 mutational classes is shown. Transitions represent the majority of the somatically acquired mutations (P<10^{-6}).

Figure 9: (A) The bar graph depicts the coverage achieved in each of the cases. The black line indicates our targeted level of 30-fold coverage. (B) This bar graph depicts the overlap between the variants identified by exome sequencing and multiplex PCR followed by deep sequencing (Raindance) in 179 genes in 8 cases, as well as exome sequencing and SNP arrays in 43 cases.

Figure 10: (A) The plot indicates the average number of additional sequence variants detected in the exomes as a function of adding each additional case. (B,C) The plots indicate the cumulative estimated number of (B) rare (C) somatically acquired exome variants discovered as a function of sample size (N=1 through N=34).

Figure 11: (A) The pie chart indicates the relative distribution of missense, nonsense, frameshift, and synonymous base alterations in the entire dataset. (B) The pie chart indicates the relative distribution of missense, nonsense, frameshift, and synonymous base alterations in the 322 DLBCL cancer genes. (C) The histogram shows the relative distribution of different mutation classes in the 322 DLBCL cancer genes. The difference between the rates of transitions and transversions was highly statistically significant (P<10^{-6}). (D) The plot shows the relative sizes of all insertions/deletions in the entire dataset. (E) The plot shows the relative sizes of all insertions/deletions among the 322 DLBCL cancer gene variants.

Figure 12: The heat map indicates the pattern of mutations of the 322 DLBCL cancer genes in 73 primary DLBCLs and 21 DLBCL cell lines. Each column represents a patient or cell line and each row represents a DLBCL cancer gene. Mutation types are indicated in the legend.
Figure 13: The frequency of the 322 DLBCL cancer genes are indicated in the graph in descending order.

Figure 14: The bar graph shows the frequency of the 12 genes that were found to be differentially mutated in the activated B cell like DLBCL subgroup and the germinal center B cell like DLBCL subgroup.

Figure 15: The figure depicts the relative distribution of genetic mutations by gene ontology categories. The spans of the arcs indicate the relative numbers of different genes annotated with respect to gene ontology [69] terms. Representative genes in each group are shown along-side each arc. Terms that match the described hallmarks of cancer(3) are marked with an asterisk.

Figure 16: (A) Deep sequencing reads identify a novel mutation in PIK3CD in a DLBCL patient. Sequencing reads matching the genome perfectly are shown in gray. The two samples differ only in a single nucleotide that is G in the tumor, but T (i.e. identical to reference genome) in the matched normal. The data were displayed using the Integrated Genomics Viewer(2). (B) Chromatograms display the results from Sanger sequencing in the same case. The sequenced bases demonstrate a T→G alteration in the tumor but not the matched normal.

Figure 17: Distribution of mutations occurring in the PI3 kinase pathway related genes: PIK3R1, PIK3CD and MTOR. Each blue diamond marks an individual mutation. 11 separate events occurred in these 3 genes.

Figure 18: (A) A molecular model of PIK3CD based on threading through the PIK3CG crystal structure. The location of the Asn 972 Lys (N972K) mutation identified in our sequencing data is highlighted within the C-lobe of the catalytic domain. (B) The Western blot depicts the alteration in phosphorylated AKT expression initially and 3 hours after the withdrawal of IL3. The relative expression of phospho-AKT (S473), with total AKT and beta-actin loading controls, are shown. (C) The relative change in PI3 kinase activity is depicted as a function of altered phospho-AKT protein expression normalized to beta-actin expression in 3 separate experiments. The p-values were computed using a student’s t-test comparing the altered PI3 kinase activity in the cells before and after IL3 withdrawal. (+ indicates IL3 exposure, - indicates IL3 withdrawal). (D) The IC50s for 21 cell lines treated with a PI3 kinase inhibitor are shown. The 3 cell lines with mutated MTOR have approximately 5-fold lower IC50s than the 18 cell lines with wild-type MTOR (P=0.005).
Figure 19: Overlaps in Genes Discovered in Multiple Cancer Studies. The Venn Diagrams depict the comparison of gene mutations from the four DLBCL studies. The number in parentheses indicates the number of genes identified in each study. The gene lists were: Morin et al (Table S1, genes with confirmed somatic cases), Lohr et al(23) (Table 1, top 58 genes), and Pasqualucci et al (Table S3 and Figure 4, validated somatic genes).

Figure 20: Copy number alterations in the DLBCL whole-genome compared to its matched control. Copy number changes are depicted on log2scale.

Figure 21: Summary of DLBCL variant and cancer gene identification from 73 primary tumor samples.

Figure 22: Power calculations by Lawrence et al (1) indicate that 500 paired cases of DLBCL need to be sequenced to sensitively detect genes recurrently mutated in 5% of DLBCL.

Figure 23: IPI clustering in 365 patients compared to complete separation by IPI in 1001 patients.

Figure 24: EZH2 and MYD88 exhibiting classical mutational patterns of activating mutations. Each panel shows the coding regions, with key domains marked along the x-axis. The y-axis indicates the number of DLBCL cases affected. Along each gene, mutations are indicated; recurrent mutations are scaled based on the number of cases. The dominating feature within each gene is a mutation peak at a single amino acid, indicating strong selection for that alteration in the activation of that particular gene.

Figure 25: Single amino acid modifications are sufficient to risk-stratify DLBCL. As shown on the left, EZH2 Y646F/N mutations are indicative of a less aggressive form of DLBCL with better overall survival, whereas MYD88 L265P separates a subtype of DLBCL with poorer overall survival.

Figure 26: Comparison of mutational patterns in PIK3CD in 94 cases of DLBCL vs. 1001 cases. The R38C hotspot was missed completely in the smaller study. In the Kaplan-Meier plot on the right, we also see that these point mutations appear to stratify a subtype of the disease with better overall survival.

Figure 27: Examples of genes with recurrent inactivating mutations in DLBCL. For each gene, the union of mutations detected in 1001 DLBCL cases is show using the cbioportal lollipop tool. The height of each mutation indicates the number of observations. Red
dots indicate stopgain or frameshift mutations, and green dots indicate missense mutations...

Figure 28: Genes with recurrent activating alterations (left) and loss-of-function alterations (right) in 1001 DLBCL.

Figure 29: Genes most differentially mutated in ABC DLBCL vs. GCB DLBCL are shown above. The cases on the left and right were categorized by gene expression to be ABC and GCB DLBCL respectively. The upper genes are enriched for mutation in ABC, whereas the lower genes are enriched in GCB. These mutational patterns indicated that it would be possible to classify DLBCL into ABC and GCB subgroups using mutations instead of gene expression.

Figure 30: Clustering of genes recurrently mutated in 1001 DLBCL. Top: overall heatmap of gene clustering based on pairwise Jaccard distance. Below: A zoomed view of gene blocks at the bottom, to highlight some genes. Note the tight clustering of HRAS/KRAS/NRAS, as well as that of the PI3K family.

Figure 31: Recurrently mutated pathways in DLBCL. In the center is depicted a Cytoscape gene network of recurrently mutated genes in DLBCL, grouped and color-coded by pathway. The size of each gene is scaled to the number of samples observed to be mutated in DLBCL. For example, some of the most frequently mutated genes include MYD88 and KMT2D/MLL2. Surrounding the network are mutational heatmaps color-coded by mutational type; yellow indicates missense, green stopgain, blue copy number loss, and red copy number gain.

Figure 32: Features that stratify DLBCL by hazard ratio. On the left are forest plots of features, and on the right are Kaplan-Meier curves of select examples. (A) Gene expression-based features. (B) Gene-level mutational and copy number features. (C) combinations of features that result in additive differences in overall survival.

Figure 33: A rare overlap heatmap is shown for 11 FFPE-frozen pairs, ordered by pair. It is evident that pairs are significantly more similar to one another (off-diagonal blocks), than to cases from other patients. The values mapped on the diagonal indicate perfect overlap of each sample with itself.

Figure 34: The distribution of fraction of variants overlapping between FFPE and frozen cases from different individuals (blue) is compared to that of cases from the same individual (orange) via box plots.
Figure 35: The mutational spectra is compared between FFPE and Frozen sequencing data from the same individual. ................................................................. 112

Figure 36: Percent of region sequenced at depth of at least 20 as a function of number of uniquely mapped reads. The fraction of the exome sequenced at depth of 20 or greater is plotted against the number of reads uniquely mapped for FFPE (blue) and frozen (orange). The overall fitted curve is indistinguishable between frozen and FFPE samples. Thus, once number of reads is fixed, the depth and coverage are identical for FFPE and frozen samples. ........................................................................................................ 113

Figure 37: Number of recurrently mutated genes overlapping in 4 studies: 1. Ours, solely discovered with FFPE cases, 2. 109 confirmed recurrently somatically mutated by Morin et al, 3. 58 significantly mutated genes by Lohr et al, and 4. 39 genes with more than 1 mutation as found by Pasqualucci et al, figure 4. Blue numbers indicate study overlaps with genes discovered in our study with muTect events >2 (1643), and orange numbers indicate overlaps with our study, of genes with at least 1 muTect event (3420 genes). 114
Acknowledgements

I would like to thank my advisor, Sandeep Dave, for his encouragement, support and trust throughout this journey. He has provided both the intellectual freedom to come up with ideas, as well as the dedication to follow them through, by generously offering his own ideas, expertise, research resources, and all-around dedication. It is his continual multi-faceted support through the expected ups and downs of research that enabled these projects to come to completion.

My committee members Jen-Tsan Ashley Chi, Alex Hartemink, and Micah Luftig have always been easily approachable and challenged me to think about the research problems in a different way, by offering ideas and questions drawn from their diverse expertise and backgrounds.

The Dave lab has been a wonderful home I look forward to seeing every day. Members generously offer their expertise, and well as a healthy dose of humor. In particular, my graduate school sisters Andrea Moffitt and Adrienne Greenough have been excellent sounding boards, always available for a coffee or garden walk where we discuss research and life. Cassandra Love has contributed extensively to every project described in this dissertation, and without her mind and skills they would have been impossible. Anupama Reddy is a role model and mentor.

The UPGG program is a friendly and diverse department, with excellent faculty and classmates. In particular, one of the most rewarding experiences has been watching
my peers grow over our years spent together in the program, initially encouraging each other as we navigated the course requirements, and subsequently as we each navigated our PhD research.

Longtime Durham friends: we have grown together from fresh-out-of-college kids to adults who have acquired spouses, houses, kids, and a good number of Duke and/or UNC degrees. You’ve always been available for trustworthy personal and academic advice, or just for unwinding via a Washington Duke run or delicious meal.

Finally, my family has offered both scientific and non-scientific feedback, and their continual support and love. I am infinitely lucky to have them.
1. Introduction

1.1 Background of next-generation sequencing

Next-generation sequencing enables the genome-wide interrogation of DNA, RNA, and protein-DNA interactions at an affordable cost. It offers two levels of information: quantitative values, in the form of density of reads mapping to particular genomic locations, and sequence information, via examination of the sequences of those reads. Whereas array-based techniques require us to design probes based on our hypotheses as to what will be present in the sample, next-generation sequencing requires no such pre-determined assumptions. With the advent of exome-seq, RNA-seq, chip-seq, and whole-genome sequencing at continually declining costs, the proverbial lamplight defining the boundaries of what is visible has widened into a floodlight. No longer are we required to pick a handful of genes to study; we can simply study all annotated genes. Moreover, as costs have fallen, we have been afforded the ability to profile a large sample size rather than agonize over case or two is the most interesting to study.

The challenge has shifted from generating the data to understanding it. Similar to our experiences when microarray technology was first developed, analysis of high-throughput data requires careful quality control, estimation of signal and noise, reproducible analyses, and achieving enough power in the context of multiple hypothesis testing.
Next-generation sequencing is being applied to increasing numbers of cancer types; The Cancer Genome Atlas (TCGA) has sequenced 38 disease types, including gastrointestinal, breast, skin, and brain cancers. Other groups independent of TCGA, including ours, have also sequenced and analyzed additional cancers. Even sequencing projects of normal individuals, or individuals with non-cancer abnormalities, provide useful insight for cancer studies because they can serve as controls. The ultimate goal of sequencing cancers is to separate the driver mutations, or cumulative genomic changes that cause a normal cell to become cancerous, from the passenger mutations, which are random and believed to have no direct bearing on cancer development. These driver mutations help us to understand how cancer arises, and can inform diagnosis and individualized treatment.

The most popular approaches thus far for identifying driver mutations have focused on gene-level analysis. Typically, genes somatically mutated at frequencies significantly higher than the background mutation rate are identified as implicated in the disease. Known pathways and gene sets are then applied after the gene list has been determined to try to group the genes in a meaningful manner and draw overarching conclusions about cell functions that have been disrupted in the cancer.

1.2 Application of techniques to B Cell Lymphomas

B Cell lymphomas comprise roughly 75% of all lymphomas, with the remainder consisting of T Cell as well as Hodgkin’s lymphomas. Under the umbrella of B Cell
lymphomas are a myriad of diseases arising from different stages of normal B Cell differentiation.

We have applied here genomic analytical techniques primarily to Mantle Cell Lymphoma and Diffuse Large B Cell lymphoma (DLBCL) cases to identify and validate driver mutations. Previous in-depth study of the disease in the laboratory also has helped us to develop a special understanding of DLBCL, both from the clinical and biological/genetic perspective, upon which we can build this novel work.

From a clinical perspective, genetic study of DLBCL is important because DLBCL is the most common form of non-Hodgkin lymphoma, with an annual incidence of 25,000 in the United States. It causes nearly 10,000 deaths per year. New therapeutics have been extensively explored in DLBCL with over a 100 clinical trials in the past several decades. In spite of this intense interest, the chemotherapy regimens used to treat DLBCL have changed little. Less than half of patients with this disease are cured after treatment with chemotherapy and rituximab. An important reason for the failure of many clinical trials in DLBCL may be that the disease is treated as a single entity, even though it is known to be molecularly heterogeneous.

1.3 Distinction of Subgroups of DLBCL by Gene Expression Profiling

Gene expression profiling of patients with DLBCL demonstrated that the tumors comprised at least two distinct diseases with different cells of origin, distinct cytogenetic differences and different response rates to anthracycline-based chemotherapy regimens.
One subgroup of DLBCL, termed germinal center B cell-like (GCB) DLBCL, was found to share many characteristics of normal germinal center B cells including the expression of genes such as BCL6 and CD10. On the other hand, another subgroup termed activated B cell-like (ABC) DLBCL, expressed genes that are downstream of the Nuclear factor kappa B (NF-kB) signaling pathway, including BCL2, Pim-1 kinase and IR4. While ABC and GCB DLBCL each constituted about 40% of the total cases, nearly 20% of the cases were found to have variable levels of expression of the distinction genes and were deemed unclassified.
2. The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells

This research chapter is based on a research article published by Jenny Zhang, Dereje Jima, Andrea Moffitt, and Qingquan Liu, et al, in the journal Blood in 2014(4).

2.1 Introduction

Mantle cell lymphoma is an uncommon form of non-Hodgkin’s lymphoma characterized by generally high relapse rate and mortality. Although translocation of \( \text{CCND1} \) is a defining feature of the disease, the role of collaborating somatic mutations that contribute to mantle cell lymphoma remains unknown (5-7).

Cancers have traditionally been classified based on their anatomic tissue of origin. Hematopoietic malignancies (leukemias and lymphomas) in particular have been extensively subclassified beyond tissue of origin, based upon the specific point in the B cell differentiation stage from which they are thought to arise. Thus, the study of these tumors enables the direct comparison between the normal stages of cell differentiation and tumors that are thought to arise from them.

As we better understand the genetic abnormalities underlying tumors, it is unclear whether the traditional methods of classifying tumors based on their tissue of origin remain useful. For instance, it has been proposed that tumors could be reclassified based on their particular mutational profiles rather than their cell of origin(8). While some genes appear to be mutated frequently in a large number of cancers (e.g. TP53), a
number of oncogenes have been found to be relatively specific to certain cancers. The lineage specificity of mutations has not been fully defined.

A number of genes related to B cell differentiation are recurrently mutated in lymphomas (e.g. BCL6(9), EZH2(10)). Expression of mutant EZH2 hampers germinal center (GC) differentiation and drives aberrant proliferation(11). However, the relationship between the differentiation stage and the development of specific mutations in cancers remains to be defined. While early efforts have attempted to characterize the association of chromatin structure to genetic alterations, they have been limited to aggregate alterations at the megabase scale(12), or rearrangement breakpoints(13), or known risk variants(14). Work comprehensively defining the chromatin state of more than one normal cell type and direct association of gene mutations in tumors that arise from those normal cells has been lacking.

The ENCODE project(15) has elucidated the chromatin structure of a number of cell types, but a similar definition of primary human mature B cells has been lacking. A particular advantage of studying mature B cells is that their relationship to leukemias and lymphomas is far better defined compared to other cell types (e.g. those corresponding to solid tumors). In this study, we sought to better understand the role epigenetic alterations that determine chromatin structure play in the gene expression and mutational profiles of B cell tumors.
We performed exome sequencing in 56 cases of mantle cell lymphoma to broadly identify the mutational landscape of the disease. We noted that the genetic profiles of mantle cell lymphomas were distinct from other aggressive lymphomas. We then defined the chromatin structure of the normal counterpart B cells (naïve and germinal center B cells respectively) for mantle cell lymphoma and Burkitt lymphoma by profiling their epigenetic markers using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for markers H3K4me1, H3K4me3, H3Ac, H3K36me3, H3K27me3 and PolII. We found that the somatic mutational profiles of mantle cell lymphoma and Burkitt lymphomas overlapped strongly with areas of open chromatin in their normal counterpart B cells, identifying the epigenetically determined chromatin structure in normal B cells as a potential determinant in the acquisition of somatic mutations.

Thus, our data define the broad genomic landscape of mutations in mantle cell lymphoma and point to an interplay between epigenetic states in normal cells and the development of genetic alterations that lead to cancer.

2.2 Results

2.1.1 Exome sequencing reveals recurrent genetic alterations in mantle cell lymphoma

We performed exome sequencing on 56 mantle cell lymphoma tumors, 28 with paired normal tissue (i.e. total 84 exomes). DNA was extracted from these tissues and paired end sequencing libraries were constructed, followed by exome-enrichment using a solution-based capture approach available commercially through Agilent. On average,
we achieved 102-fold coverage of the targeted exonic regions. Over 90% of the targeted exons were covered at an average depth exceeding 10-fold.

Sequencing reads were mapped to the reference human genome, and high-quality mismatches were classified as synonymous and nonsynonymous variants. Nonsynonymous variants were further subclassified as missense, nonsense and small insertions/deletions (indels). Missense variants comprised the highest proportion of alterations (48.4%), whereas synonymous and nonsense variants and indels comprised 46.8%, 0.7% and 4% respectively.

We verified the accuracy of identifying genetic variants through deep sequencing by performing Sanger sequencing in the same cases for 53 distinct variants. As in our previously described work with hundreds of variants (16, 17), we found over 90% agreement between the two methods. These results confirmed the accuracy of our methods for high throughput sequencing and identifying genetic variants.

2.1.2 Defining the landscape of gene mutations in mantle cell lymphoma

We initially examined the exomes of the mantle cell lymphoma cases with paired germline DNA. We identified somatic mutations affecting 537 genes in at least one tumor/germline pair among these cases. Transitions comprised the majority of the somatic variants (P<10⁻³, chi-squared test). On average, we observed 32 somatic alterations per sample (range 1-61), a rate similar to that which we observed in Burkitt lymphoma cases (16), and approximately half the rate we observed in DLBCL (17).
Upon inclusion of an additional 28 validation samples in our analysis, we identified 1189 rare variants corresponding to these 537 somatically mutated genes. These variants were not present in publicly available data from normal controls including dbSNP135(18), the 1000 Genomes Project(19), and available exome sequencing data from healthy individuals without lymphoma(20, 21).

We further required each of the identified genes to have three somatically acquired and/or non synonymous rare events predicted to be functionally significant. These methods, detailed in the Supplement(4), allowed us to identify 37 recurrently mutated genes in mantle cell lymphoma. These genes are depicted in Figure 1a and listed in Supplementary Table 2 (4). The individual variants are listed in Supplementary Table 3 (4). The relative frequency of synonymous and nonsynonymous variants for each gene is shown in Figure 1b. 12 of the 37 genes have been previously implicated in cancer(22).
Figure 1: (a) Each column represents a patient and each row represents a gene. Mutations are color-coded with yellow for missense mutation, purple for frameshift mutation, red for nonsense mutation, and orange for in-frame insertion or deletion. (b) The bar graph indicates frequency of variants found by gene across all samples, subdivided by not-synonymous (blue) and synonymous (grey) mutations.

The most frequently mutated genes in mantle cell lymphoma were ATM (41.9%), CCND1 (14%), MLL2 (19.6%) and TP53 (18.6%). Other frequently mutated genes included known oncogenes and tumor suppressor genes such as RB1, SMARCA4, APC, NOTCH1, and UBR5. Our data also implicated a number of genes not previously associated with mantle cell lymphoma, including POT1, FAT4 and ROBO2. Silencing mutations (frameshift and nonsense mutations) comprised a sizeable fraction of the genetic events in ATM, MLL2, MLL3, RB1 and ROBO2, suggesting that those alterations result in a loss of function in mantle cell lymphoma.
Pathway and gene set analysis of the significantly mutated genes reveals a crucial role for many of these genes in cell cycle regulation (SMC1A, POT1, RB1), cell adhesion (FAT4, DLC1, CDH8), development (ROBO2, ANK2, CTNNA2), and chromatin modification (SMARCA4, MLL2, MLL3, WHSC1). Figure 2 illustrates the links between mutated genes that belong to the same functional gene sets.

![Network Diagram](https://via.placeholder.com/150)

**Figure 2:** The network indicates functional groupings of the genes mutated in MCL. Nodes represent significantly mutated genes that are also a part of a significant functional group. Edges connect nodes that belong to the same functional gene set. Colored ovals identify the gene sets to which these nodes belong.

### 2.1.3 Genetic differences between mantle cell lymphoma and other lymphomas

In order to better understand the genetic differences between common non-Hodgkin lymphomas, we analyzed exome sequencing data that we(16, 17) and others(9,
23-26) have generated from Burkitt lymphoma and diffuse large B cell lymphoma (DLBCL). We identified all genes that were mutated at a frequency of 10% or higher in at least one tumor type and differentially mutated among at least one of the lymphoma types (P<0.05, Fisher’s exact test). We plotted the relative frequencies of these genes for mantle cell lymphoma, Burkitt lymphoma and the molecular subgroups of DLBCL in Figure 3.

![Figure 3: The bar graph depicts the proportion of mutated cases that belong to each lymphoma type for mantle cell lymphoma (MCL), Burkitt lymphoma (BL), GCB diffuse large B cell lymphoma (DLBCL), and ABC diffuse large B cell lymphoma.](image)

We found a number of genes that were predominantly mutated in each disease. Mutations in \textit{ATM}, \textit{CCND1}, and \textit{RB1} occurred mostly in mantle cell lymphoma. Mutations in \textit{ID3} and \textit{MYC} occurred predominantly in Burkitt lymphoma. Mutations in \textit{PIM1}, \textit{BCL2} and \textit{CREBBP} occurred mostly in DLBCLs. A number of genes had overlapping patterns of mutations between two or more of the diseases, including \textit{TP53}, \textit{GNA13}, \textit{ARID1A} and \textit{SMARCA4}.
We further examined the association between individual genes that were recurrently mutated in these diseases (Supplement Figure 2) (4). We found that ATM and RB1 mutations tended to co-occur in the same cases, reflecting their involvement in the DNA damage response and its importance in mantle cell lymphoma.

2.1.4 Epigenetic profiling of normal B cells and its relationship to gene expression

Gene expression and replication must be preceded by conformational changes in the chromatin to an open state (27). We sought to define the relationship between the epigenetically determined chromatin state of primary mature B cells and the lymphomas that are thought to arise from them.

We performed FACS-sorting of normal naïve B cells, germinal center B cells and memory B cells from otherwise normal individuals undergoing tonsillectomy (>95% purity verified by flow cytometry in each case)(28). We profiled the chromatin structure and epigenetic state of the normal B cells through ChIP-seq on 6 different markers: H3K4me1, H3K4me3, H3Ac, H3K27me3, H4K36me3, and PolII. These markers collectively identify a comprehensive epigenetic portrait of these primary human cells. A few of these markers have been published previously in B cells, and the identified regions overlap strongly with our data(11).

We plotted (Figure 4) the aggregate density for each of the six markers associated with gene expression over a range from 2kb upstream to 10 kb downstream of all CCDS-annotated transcription start sites (TSS). Consistent with work in other cell types (27, 29),
we found that H3K4me1 and H3Ac exhibited strong peaks before and after the TSS, 
PolII and H3K4me3 peaked directly at the TSS, and H3K36me3 exhibited elevated levels 
through the gene body (Figure 4), reflecting transcriptionally active genes. H3K4me3 
levels declined rapidly over the gene body, consistent with the notion that it marks 
 promoter sites. The repressive marker H3K27me3 also was elevated before and after the 
 TSS, an indication of the transcriptionally inactive genes.

Figure 4: Epigenetic profiles of H3K4me1, H3K4me3, H3K27 me3, H3Ac, 
H3K36me3, and PolII are shown in 50-bp read resolution from 2kb upstream to 10 kb 
downstream of all annotated transcription start sites. transcription start sites.

We further plotted the chromatin epigenetic marker profiles around the 
transcription start sites for both highly and lowly expressed genes for each cell type(30). 
We observed similar patterns in both cell types (Figure 5). The genes that were 
expressed the most highly in these B cells were highly associated (P<10^−6, Pearson
correlation test and Kolmogorov-Smirnov test) with the epigenetic alterations that identify open chromatin (H3K4Me1, H3K4Me3, H3Ac, H3K36Me3). Conversely, the epigenetic alteration associated with closed chromatin (H3K27Me3) was found to be highly associated with repressed gene expression. The markers H3K4Me1, H3Ac and H3K36Me3 were most strongly associated with the gene body whereas H3K4Me3 was most strongly associated with promoter regions, as has been described previously(27).

Figure 5: Epigenetic profiles of H3K4me1, H3K4me3, H3Ac, H3K36me3 and H3K27me3 around the transcription start sites (TSS) are shown for the top 10% most expressed, bottom 10% expressed, and all genes for naïve and germinal center B cells.
2.1.5 Role of B-cell differentiation stage in mantle cell lymphoma

Although B cell differentiation state is the basis for classifying lymphomas, the extent to which it is related to the molecular profiles of mantle cell lymphoma and other lymphomas has not been fully defined. In order to investigate this further, we examined the relationship between differentially activated chromatin marks between cell types and gene expression in the tumors of corresponding cell-of-origin (Figure 6). Based on the observed patterns of distribution of chromatin marks and the body of the gene, we defined open chromatin in all gene regions as a combination of marks corresponding to H3K36me3, H3ac, and H3K4me1. Differential chromatin levels for each gene were calculated by computing the difference of total number of aligned ChIP-seq reads from naïve and germinal center B cells that fell within the gene region, normalized by the length of the gene and the sequencing library size, similar to methods described previously (Supplement Table 4)(4, 27).

We examined the gene expression profiles of mantle cell lymphomas(31) and Burkitt lymphomas(32) performed on the same microarray platform. We observed a striking overlap (Figure 6) between the epigenetic profiles defining open chromatin in naïve and germinal center cells and the gene expression profiles that distinguish mantle cell lymphoma and Burkitt lymphoma, an overlap that was highly statistically significant (P<10⁻⁶, chi-squared test), and likely reflected the similarity of these tumors to their supposed cells of origin.
Thus, we concluded that the epigenetically defined regions of open chromatin of their cells of origin are highly related to gene expression profiles of mantle cell lymphoma and Burkitt lymphoma.

2.1.6 Epigenetic profiling of normal B cells and its relationship to mutations in lymphoma

We next examined the association between the mutational patterns observed in mantle cell lymphoma and Burkitt lymphoma and these chromatin marks in their cells of origin, naïve B cells and germinal center B cells. We defined an open chromatin score
for each gene using a sum of the markers for open chromatin on the gene body
(H3K36Me3, H3Ac, and H3K4Me1; Figure 6, Supplement Appendix 4 (4)).

Using these epigenetic markers, differences in open chromatin between naïve B
cells and germinal center B cells were computed for genes that were differentially
mutated in mantle cell lymphoma and Burkitt lymphoma. We plotted the relative state
of open chromatin in naïve and germinal center B cells (Figure 7; individual markers
shown in Supplement Appendix 4 (4)) in these genes. We found that difference in gene
mutation frequency between mantle cell lymphoma and Burkitt lymphoma is highly
associated with differences in open chromatin in their corresponding cells of origin
(P<0.001, chi-squared test). Differences in mutation frequency were not associated with
expression differences of these genes in the normal B cells (Figure 7). In addition, those
genes mutated in both BL and MCL without significant differences in mutation rate did
not show significant differences in open chromatin between the corresponding normal B
cell types.
2.3 Discussion

Our data identifies the genetic heterogeneity underlying mantle cell lymphoma and implicates a number of novel genes in the development of the disease. Our study overlapped significantly with recently published studies of 29 tumor exomes (33) and eighteen transcriptomes (6, 29) (e.g. ATM, CCND1, TP53, WHSC1, MLL2, NOTCH1, UBR5) and targeted sequencing studies (e.g. BIRC3(34), ATM and TP53(35)). Our work
implicates other mutated genes in the mantle cell lymphoma including RB1, POT1, ROBO2, SMARCA4, and MLL3. Similar to other lymphomas(17), our data indicate a striking heterogeneity underlying mantle cell lymphomas, with relatively few genes mutated in more than 10% of the cases. While our data indicate a paucity of activating mutations in oncogenes that can be readily targeted therapeutically, a functional analysis of gene mutations implicates cell cycle progression, cell adhesion, and signal transduction as broad oncogenic processes that could be targeted.

If the acquisition of somatic mutations in tumors was completely stochastic, then one might expect that virtually every oncogene and tumor suppressor gene to be involved in most cancers. However, most cancers show a degree of specificity in their mutational spectrum. Our work identifies B cell stage as a critical determinant of acquired somatic alterations in lymphomas. “Lineage addiction” has been previously noted in the context of individual genes(36). Both BCL6(37) and EZH2(11) have already been defined as important regulators of B cell differentiation and oncogenes in germinal center B cell-derived tumors(10, 38). Our study is among the first to define the extent to which B cell stage is associated with mutations in lymphomas.

The state of differentiation of each cell type ultimately resides in the chromatin structure of the cells. The epigenetic modifications of histones enable the coiling and uncoiling of DNA that is a prerequisite for DNA repair, replication and transcription. The ENCODE project(15) has defined the chromatin structure of a number of cell types
determining gene regulatory regions including promoters, enhancers and insulators. To our knowledge, our data provide the first comprehensive description of these regions in human naïve, germinal center (dark zone) and memory B cells.

The normal cells of origin of hematopoietic malignancies are better defined than those of most other tumor types. While our recognition of cells of origin is necessarily limited by our evolving knowledge of states of cellular differentiation that exist within every lineage, the mature B cell types (germinal center dark zone cells and naïve cells) nevertheless capture the broad phenotypes of Burkitt lymphoma(39) and mantle cell lymphomas respectively. Mantle cell lymphomas arise from a subset of CD5-positive naïve cells, while a small proportion of mantle cells appear to be of post-germinal center origin. Nevertheless, the vast majority of mantle cell lymphomas largely maintain the phenotype of naïve B cells. This is highlighted by the strong overlap between the gene expression profiles of naïve B cells and mantle cell lymphomas.

Chromatin structure has been shown to be a potential determinant of local mutation rates in different cancers(12, 14). One recent study examined, at megabase scale, the relationship between relative mutational rates across the genome and corresponding chromatin structure(12). It found that genomic regions of higher SNV density associate with more closed chromatin.

Our study is, to our knowledge, the first to elucidate the relationship between differential gene-level mutation rates between two cancer types (MCL and BL) and the
difference between chromatin states of their corresponding normal cells of origin (naïve and germinal center B cells). Our data indicate that at the gene level, the epigenetically defined differences in chromatin structure between Naïve and Germinal Center B cells is independently associated with both expression profiles and mutational profiles of B cell lymphomas derived from those cell types.

Mutation of chromatin modifying genes has been found to be a recurrent feature of many cancers. In lymphomas, EZH2 mutations have been shown to modify H3K27me3, a process that might influence lymphoid transformation in germinal center derived B cells. The role of transcription-associated mutagenesis has been described in prokaryotes(40) and yeast(41). While it is possible that similar mechanisms underlie cancers, the association between chromatin structure and mutation rates is not explained by transcription alone. There was no relationship between differential gene expression and mutation, suggesting that other unknown mechanisms underlie the association between mutations and chromatin structure.

Our study identifies the epigenetically defined chromatin structure of normal B cells as a common denominator of both expression profiles and mutational profiles of B cell lymphomas. This work thus provides an important starting point for understanding the genetic diversity of mantle cell lymphomas and the interplay between genetic and epigenetic alterations in cancer.
2.4 Methods

2.4.1 MCL sample acquisition and processing

Mantle cell lymphoma (MCL) tumors (N=56) and normal tissue (N=28) were obtained from the institutions that constitute the Hematologic Malignancies Research Consortium (HMRC) (30). Genomic DNA was extracted as described previously (30). Patient tumor and normal samples were collected according to IRB guidelines. This study was conducted in accordance with the Declaration of Helsinki.

2.4.2 Exome capture and library preparation

Genomic DNA was sheared to 250 bp using the Covaris S2 platform. Exome capture libraries were prepared as described previously (17).

2.4.3 B cell isolation

Tonsils were obtained according to IRB guidelines from patients undergoing routine tonsillectomy at a North Carolina hospital. Mononuclear tonsillar B cells were prepared and stained as described previously (28). Stained cells were FACs sorted into 10 million each of naive B cells (CD19+IgD+CD27–CD38+), GC B cells (CD19+IgD–CD38++), and memory B cells (CD19+IgD–CD27+CD38dim) for each chIP antibody.

2.4.4 Chromatin immunoprecipitation (ChIP) and sequencing library preparation

Flow-sorted B cells were cross-linked with 1% formaldehyde, pelleted, lysed, and sonicated to a median size around 200 bp. After sonication, the nuclear extract was washed and incubated overnight at 4 ºC with antibodies coupled to magnetic beads.
(Invitrogen, 112.03D): anti-H3K4me3 (Millipore, 07-473), anti-H3K4me1 (Abcam, ab-8895), anti-H3K27me3 (Millipore, 07-449), anti-H3Ac (Millipore, 06-599), anti-PolII (Santa Cruz, sc-899x) and anti-H3K36me3 (Abcam, ab-9050). Isotype-matched IgG control antibodies were used in parallel as negative control. The beads were washed, and the immunoprecipitated complex was eluted and reverse cross-linked at 65°C overnight. The samples were subsequently digested and purified. Sequencing libraries were prepared using the Illumina Genomic DNA Sample Prep kit (FC-102-1001) according to the provider’s protocol.

2.4.5 Exome data processing and analysis

Exome Sequence Alignment, Variant Calling, and Annotation

Reads in fastq format(42) were pre-processed with GATK(43) to remove Illumina adapter sequences and Phred-scaled base qualities of 10 and below as described previously(16). After GATK processing, reads were mapped to hg18 using Burrows-Wheeler Aligner (BWA(31)), followed by Novoalign V2.06.09 (Novocraft.com). Remaining unmapped reads were clipped to 35bp to remove adapter matches in the 3′ end of the read, and re-aligned with BWA.

All alignments were output as BAM files(44) and merged using Picard (http://picard.sourceforge.net). The data will be available on dbGaP. PCR/optical duplicates were marked by Picard. Base quality recalibration was performed using GATK. To improve accuracy and quality of the calls, localized indel realignments were
performed using GATK(43). Regions that should be realigned are determined by the GATK Realigner Target Creator. Single nucleotide variant (SNV) and indel variants were called using SAMtools(45). SAMtools mpileup was run on all cases and output to a single VCF file. Each variant was required to have an instance of genotype quality greater than 30 and read depth greater than 5. Individual SNVs and indels were annotated with gene names and predicted function using Annovar (http://wannovar.usc.edu/)(46).

**Identification of somatically acquired mutations using muTect**

muTect(47) was used to identify somatic mutations in the 28 MCL tumor-normal pairs. BAM files were generated as described above. muTect version 1.1.1 was executed on each tumor-normal pair.

**Control exomes**

In addition to NHLBI 6500 exomes (http://evs.gs.washington.edu/EVS/), 1000 Genome Project(19), and dbSNP135(18) frequencies, we also filtered our mantle cell lymphoma variants against 189 non-cancer exomes generated in-house and 256 publicly available exomes that we re-processed(20, 21). This allowed us to reduce the number of variants observed that are simply part of normal human variation.

**Coverage**
SAMtools(45) flagstat was used to compute the number and percent of reads that mapped to the genome. Both depth and breadth of coverage for each exome were computed using BEDTools(48).

Sanger Sequence Validation

DNA regions of interest were amplified using primers designed to target exonic regions containing the variant, as described previously(49). The amplified fragments were verified by agarose gel, and reactions were purified with ExoSAP-IT using manufacturer instructions (Affymetrix, 78250).

Significantly differentially mutated genes

Differentially mutated genes were selected for further analysis by Fisher's exact test between the number of mutated and unmutated cases in Burkitt lymphoma and mantle cell lymphoma (p<0.05).

Gene pathway analysis

Using GATHER Gene Ontology annotation (http://gather.genome.duke.edu/) and GSEA MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp), enrichment of this set of 37 MCL mutated genes was measured in known gene lists. Enrichment was considered to be significant at an FDR q-value less than 0.05.

2.4.6 ChIP-seq data processing and analysis

Alignment and peak calling
The B cell ChIP-seq and mock samples were aligned to the reference human genome (hg19) using Bowtie(50). Significantly (P<0.0001, FDR<5%) enriched peak regions were detected using the MACS software(51) using the following parameters (--nomodel, --shiftsize=100, --bw=250, --mfold=10,30, -w -S, -g hs, --pvalue=0.0001). The data will be available on GEO.

Characterizing ChIP-seq reads in gene regions

For each pair of ChIP marker and B cell sample (Naïve and GC), a score was calculated for each gene in a similar fashion to RNA-seq RPKM. Specifically, for each gene region (5 kb padded on either side), the number of aligned reads to fall in that gene region were counted using coverageBed(48). This number is divided by the length of the gene region in base pairs, and again by the total number of aligned reads in that particular sample, to adjust for coverage differences. The scores were multiplied by 1 billion to get to a reasonable number range and log2 transformed. Transcription start and stop sites for each gene were extracted from RefSeq. Gene region BED files, padded by 5kb on either end of the transcription start and stop sites, were generated in hg19 using BEDTools. Differences between Naïve and GC were calculated by subtracting the GC score from the Naïve score for corresponding genes and markers.

Open chromatin calculation
The open chromatin score for each gene is defined as the sum of H3K4me1, H3Ac, and H3K36me3 individual gene scores. Differences between Naïve and GC are calculated by subtracting one cell type chromatin score from the other.

2.4.7 Microarray data analysis

Microarray data from public sources (BL, MCL, B cell)

All gene expression microarray data was generated on the Affymetrix Human Genome U133 Plus 2.0 platform. Burkitt lymphoma (BL) gene expression microarray CEL files were published previously(32). Mantle cell lymphoma (MCL) gene expression microarray CEL files were downloaded from Gene Expression Omnibus (GEO), accession number GSE21452(31). Germinal center and naïve B cell gene expression microarray was published previously(28), available under GEO accession number GSE12366.

Processing microarray data and selecting significant genes

BL, MCL, and B cell expression CEL files were normalized together using RMA(52). Differentially expressed genes between naïve and germinal center B cells were selected using a t-test p-value less than 0.05, absolute fold change greater than 1, and mean expression greater than 5.5.
3. The Genetic Heterogeneity of Diffuse Large B Cell Lymphoma

This research chapter is based on a research article published by Jenny Zhang and Vladimir Grubor, et. al, in the journal PNAS in 2013(17).

3.1 Introduction

Diffuse large B cell lymphoma (DLBCL) is the most common form of lymphoma in adults(53). While nearly half the patients can be cured with standard regimens, the majority of relapsed patients succumb. Thus, there is an urgent need to identify the genetic underpinnings of the disease and to identify novel treatment strategies. Gene expression profiling(54, 55) has uncovered distinct molecular signatures for DLBCL subtypes that have unique biology and prognoses. High-throughput sequencing has provided rich opportunities for the comprehensive identification of the genetic causes of cancer(56-58). While exhaustive portraits of individual cancer genomes are emerging, the degree to which these genomes represent the disease is unclear.

We have generated a detailed analysis of a DLBCL genome by sequencing a primary human tumor and paired normal tissue. We further characterized the genetic diversity of DLBCL by sequencing the exomes of 73 DLBCL primary tumors (34 with matched normal DNA) and 21 DLBCL cell lines for comparative purposes. This in-depth sequencing identified 322 DLBCL cancer genes that were recurrently mutated in DLBCLs. We also experimentally validated the effects of genetic alteration of PIK3CD, a novel oncogene that we identified in DLBCL. Our work provides one of the largest
genetic portraits yet of human DLBCLs, and offers new insights into the molecular heterogeneity of the disease, especially in the context of other recently published studies in DLBCL(9, 24).

3.2 Results

3.2.1 Sequencing of a lymphoma genome uncovers the spectrum of somatic variation in DLBCL

Lymphoma biopsy tissue and unaffected bone marrow were obtained from the same patient. Using the Illumina platform, we generated at total of 171 gigabases (GB) of 100 base-pair (bp) paired-end sequences from the tumor and matched normal genomes corresponding to an average per-base sequencing coverage of 37-fold and 20-fold respectively.

We identified 23,214 somatic sequence alterations occurring throughout the lymphoma genome, summarized in Figures 8 and SI Appendix (Table S1) (17).

Transitions accounted for about 60% of these events (Figure 8C), similar to patterns observed in a number of other malignancies(49, 59, 60) and suggests that the majority of these DLBCL mutations arise from stochastic endogenous processes rather than environmental exposures, for example in the context of tobacco exposure and lung cancer(58).

Known oncogenes(61) found to be somatically mutated in this DLBCL patient included ARID1A, SETD2, CARD11, and PIK3R1. Of these genes, only CARD11(62) has been previously experimentally identified as an oncogene in DLBCL. We also identified
structural genetic alterations using approaches described previously (63, 64). In all, we identified 7 deletions and 3 amplifications. Known oncogenes that were implicated by these copy number alterations include PTEN (chr 10) and CDKN2A (chr 9).
Figure 8: Results from sequencing a lymphoma genome. (A) The Circos diagram [58] summarizing the somatically acquired genetic variants in a diffuse large B cell lymphoma (DLBCL) genome. The outermost ring depicts the chromosome ideogram oriented clockwise, pter-qter. The next ring indicates copy number alterations in the DLCBL genome. The next 3 rings indicate somatically acquired mutations in intergenic regions, potential regulatory regions, and the exome respectively. (B) The pie chart depicts the relative number of somatically acquired mutations in the DLBCL genome, which can be classified by their genomic location as intergenic, intronic, potential regulatory or transcribed regions (left panel). The right panel depicts the breakdown of different mutation types observed in the transcribed regions. (C) The histogram depicts the mutation profile of DLBCL. The proportion of mutations in each of the 6 mutational classes is shown. Transitions represent the majority of the somatically acquired mutations (P<10^-6).
3.2.2 Exome Sequencing defines the spectrum of coding region mutations in DLBCL

To identify recurrently mutated genes in DLBCLs, we obtained a total of 73 cases of primary human samples. We divided the primary human cases into a discovery set (N=34) and a prevalence set (N=39). For each of the discovery set cases of primary DLBCLs, we also sequenced paired normal tissue. In addition, we sequenced the exomes of 21 DLBCL cell lines that are widely used to model the disease.

We performed whole-exome sequencing for all of these DLBCL and paired normal cases using the Agilent solution-based system of exon capture, which targets the NCBI Consensus CDS database (CCDS)(65). We generated over 500GB of mappable sequence data and generated sequence data for 94% (median) of the targeted bases in each sample. Our average exome coverage was 47-fold (median 42.5-fold) per targeted base (Figure 9A). In all, we identified 121,589 distinct novel variants in these cases.

3.2.3 Validation of Genetic Variant Identification

To verify our methods for exome capture and bioinformatics analyses, we performed exome sequencing on a single Hapmap sample (NA12762) previously published(19). We found over 99% concordance with the published data. We also used three different approaches for further validation.

First, we performed high-throughput, multiplex PCR in microdroplets (Raindance technologies(66)) and sequencing to over 100-fold coverage for 179 genes (SI
Appendix (Table S5)) in 8 cases. Over 99% of the variants were identical (Figure 9B, SI Appendix (Table S6)) (17). The 100-fold coverage did not result in a significant increase in variant discovery, confirming our estimates our coverage of the exome was adequate for the identification of variants in most instances.

Second, we genotyped 43 of these cases using an Illumina SNP array comprising over 733,000 probes. We found excellent concordance (94.2%) between the microarray calls and the exome sequencing analysis in that sample (Figure 9B). Finally, we performed PCR amplification and Sanger sequencing for 25 variants corresponding to 24 genes in 50 cases. Again, we found excellent concordance (93.5%) of the exome capture calls with the calls generated by conventional Sanger sequencing (SI Appendix (Figure S2))(17). Taken together, these results indicate that our methods for exome-enrichment, sequencing and bioinformatic analysis produce robust results.
We empirically explored the power of our study to identify novel genetic variants by plotting the expected number of new variants that would be discovered from each additional case of DLBCL. We found a progressively smaller number of new variants by sequencing each additional sample (Figure 10A). By N=25, the number of novel variants contributed by each additional sample fell to less than 1% of the total (Figure 10A). These values corresponded well with a regression model for exponential increase ($R^2 = 0.8768, P<10^{-6}$).

However, the vast majority of these variants depicted in Figure 10A were common variants, and present in our normal controls. The number of rare variants (<1% frequency in the general population) discovered per additional sample remained
relatively constant and increased linearly with each additional case (Figure 10B).

Similarly, the number of somatic variants identified in our discovery set (i.e. variants absent from the corresponding paired normal tissue) also increased linearly with the addition of each individual pair of tumor-normal sets (Figure 10C). The number of individual genes implicated by sequencing additional samples showed a similar linear increase as a function of the number of cases. Since cancer arises predominantly from such somatically acquired rare variants, these findings have implications for the number of samples needed to comprehensively characterize a heterogeneous disease like DLBCL, as discussed below.

3.2.4 Patterns of Exome Variation and Identification of DLBCL Cancer Genes

We began analysis by aligning the sequencing reads to the genome and determining the distribution of our mutations. 53.8% of the variants were missense, and

Figure 10: (A) The plot indicates the average number of additional sequence variants detected in the exomes as a function of adding each additional case. (B,C) The plots indicate the cumulative estimated number of (B) rare (C) somatically acquired exome variants discovered as a function of sample size (N=1 through N=34).
nonsense, frameshift, and synonymous variants comprised 1.1%, 2.4%, and 42.7% of the total number of variants respectively (Figure 11A). These overall patterns of genetic variation in the DLBCL exomes are quite similar to what we would expect in the variation of normal exomes. We eliminated common genetic variants that occurred in our set of normal controls identified from dbSNP, the 1000 genomes project, and 256 recently published exomes from otherwise healthy individuals.

Our methods for identifying candidate DLBCL cancer genes are detailed in the SI Appendix. We developed a statistical model for comparing the characteristics of the somatically mutated genes in our discovery set of 34 tumor-normal pairs to previously validated cancer genes. We modeled several variables including frequency of non-synonymous variation in the gene, frequency of somatic mutation, gene size, rate of non-synonymous variation in healthy controls, and the predicted effect of the genetic alteration on the encoded protein. DLBCL cancer genes were identified as those that had a score distribution most similar to those of the previously validated cancer genes (P<10⁻⁶).

We identified 322 candidate DLBCL cancer genes as recurrently somatically mutated in DLBCL. The majority of the 52 known cancer-related genes and the remaining 270 genes have not been previously identified as having a role in lymphomas. Among these 322 DLBCL cancer genes, we identified a total of 1418 variants in these cases (Figure 11B, Dataset S4) (17). There was a higher proportion of each category of
non-synonymous mutations including missense (66.4%), nonsense (2.4%) and frameshift (2.5%), and fewer synonymous mutations (28.7%) in these genes compared to the patterns observed in the entire exome.

Once again, we observed a predominance of transitions (P<10^{-6}, chi-squared test, Figure 11C). Overall, the sizes of the insertions and deletions in these DLBCLs preserved reading frames, with peaks observed at insertion/deletion sizes of 3, 6, 9 and so on (Figure 11D). However, in the 322 DLBCL cancer genes, we noted a significant depletion of indel sizes that were multiples of three (Figure 11E, P<0.01, chi-squared test).
3.2.5 Identification of protein coding sequence variation in DLBCL

The mutational patterns of these 322 DLBCL cancer genes in the 73 primary lymphomas, as well as 21 cell lines, are depicted in Figure 12. The median number of DLBCL cancer gene alterations per patient was 16 (mean 17).
Figure 12: The heat map indicates the pattern of mutations of the 322 DLBCL cancer genes in 73 primary DLBCLs and 21 DLBCL cell lines. Each column represents a patient or cell line and each row represents a DLBCL cancer gene. Mutation types are indicated in the legend.

Figure 13 shows the frequency of the DLBCL cancer genes, which followed a classic “long tail” distribution. Our data identify a number of known cancer-related genes in DLBCL that have been previously reported and include TP53(68), MYD88(69), PIM1(70), CARD11(62) and BCL6(70). Our data also implicate a number of new cancer-related genes that were not previously linked to DLBCL, including PIK3R1, ARID1A, MTOR, and IDH1. These data indicate that the spectrum of mutations and genes involved in lymphomas, and potentially other cancers, may be much larger than has been previously appreciated.
3.2.6 Gene expression-based subgroups of DLBCL demonstrate distinct mutation patterns

To better understand potential subgroup-related differences in observed patterns of DLBCL mutations, we performed gene expression profiling using Affymetrix microarrays to distinguish ABC DLBCLs (N=29) and GCB DLBCLs (N=35). We found 12 genes (Figure 14) with a frequency of at least 10% in each subgroup that were differentially mutated between the two groups (P<0.05, Fisher’s exact test). Genes that were more frequently mutated in ABC DLBCLs included MYD88, KLHL14, CD79B and SIGLEC10, whereas GNA13, BCL2, and EZH2 were more frequently mutated in GCB DLBCL. Of these, we also found GNA13 and EZH2 to be recurrently mutated in Burkitt lymphoma(16), another tumor derived from germinal center B cells.
3.2.7 Functional Categorization of Recurrently Mutated Genes in DLBCL

Twelve gene ontologies accounted for over half of the DLBCL cancer genes (N=203, Figure 15). Biological processes comprising signal transduction (e.g. PIK3CD, PDGFRA) and chromatin modification (e.g. MLL3, SETD2) were most commonly implicated as DLBCL cancer genes (SI Appendix, Figure S3 (17)). A number of these biological processes have been directly implicated as hallmarks and enabling characteristics of cancer(3). Thus a number of these DLBCL cancer genes have directly discernible roles that impact the growth and development of tumors.
The role of signal transduction pathways in tumors is of particular interest because they may be therapeutic targets for small molecule inhibitors. Signaling pathways (71) including JAK-STAT, ubiquitin, WNT, NF-KB, Notch, and PI3 kinase signaling were recurrently mutated in DLBCLs, although mutations in each signaling pathway occurred in only a minority of these cases. Many of these pathways have not been conclusively implicated in lymphomas and might be potential therapeutic targets in DLBCL subsets defined by mutations in them, a notion that we further explored experimentally.

### 3.2.8 Determination of genes enriched with AICDA-related mutations

We further examined the potential role of AICDA (AID) in the acquisition of these somatic mutations in our DLBCL cases. For the 322 DLBCL cancer genes, we...
determined the number of acquired mutations where the reference sequence was “C” and of those, the fraction that fell into WRCY motifs (and the reverse complement), which are associated with AICDA activity(72).

We performed a Fisher’s exact test to determine the significance of enrichment for mutations in the WRCY motifs compared to the background rate. We found significant enrichment (P<0.05) of WRCY-motifs in PIM1, BTG1 and CD79B, suggesting that AICDA is a significant contributor to the somatic alterations in these genes. Gene alterations in PIM1, BTG1 and CD79A have not been described in most solid tumors, suggesting AICDA-related alterations are a lymphoma-specific mechanism, similar to their described role in B cell biology.

**3.2.9 PIK3CD is a Novel Oncogene, and PI3 Kinase Inhibition is a Potential Therapeutic Approach in DLBCL**

Deregulation of the PI3 kinase pathway is a common feature of many cancers(73). We observed 3 separate cases with mutations in the PIK3CD gene, which is not thought to be an oncogene. A single point mutation T → G (Figure 16A), confirmed by Sanger sequencing (Figure 16B) was identified in the catalytic domain of the PIK3CD gene, which altered the encoded amino acid from one with an uncharged side chain (asparagine) to one with a positively charged side chain (lysine).
We found mutations in 2 additional known oncogenes in the PI3-kinase pathway, PIK3R1 and MTOR, pointing to deregulation of the PI3-kinase pathway as an important oncogenic mechanism in DLBCLs. Other key members of the pathway with known oncogenic roles, including PTEN, FOXO3, and GSK3, were not mutated in our cases. Similar to patterns observed previously in PIK3CA (73), the mutations in PIK3CD, PIK3R1 and MTOR appear to spread across multiple locations of the gene (Figure 17) rather than clustering in a single hotspot.

Figure 16: (A) Deep sequencing reads identify a novel mutation in PIK3CD in a DLBCL patient. Sequencing reads matching the genome perfectly are shown in gray. The two samples differ only in a single nucleotide that is G in the tumor, but T (i.e. identical to reference genome) in the matched normal. The data were displayed using the Integrated Genomics Viewer (2). (B) Chromatograms display the results from Sanger sequencing in the same case. The sequenced bases demonstrate a T→G alteration in the tumor but not the matched normal.
We modeled the PIK3CD protein structure based on that of its paralog PIK3CG, which has been determined through crystallography (74). The identified mutation lies in the catalytic domain in the predicted structure of the protein (Figure 18A). We over-expressed the wild type and the mutant PIK3CD constructs in the FL5.12 lymphoma cell line that is well-characterized for its IL3-dependent PI3 kinase signaling (75). In these cells, withdrawal of IL3 is associated with measurable decrease in PI3 kinase signaling and decreased phosphorylated AKT, which is directly downstream of PI3 kinase and provides pro-proliferative signals. In cells expressing the wild-type form of PIK3CD, we found that withdrawal of IL3 was associated with a measurable decrease in phosphorylated AKT S473 (Figure 18B). There was no measurable decrease in the phosphorylated AKT in cells expressing the mutant form of PIK3CD, suggesting that the mutation had an activating effect (Figure 18B). These observations were confirmed in three experimental replicates, all of which showed that IL3 withdrawal was associated with significant down-regulation in phosphorylated AKT in cells expressing wild-type PIK3CD (P=0.04), but not in cells expressing mutant PIK3CD (P=0.44, Figure 18C). ELISA

Figure 17: Distribution of mutations occurring in the PI3 kinase pathway related genes: PIK3R1, PIK3CD and MTOR. Each blue diamond marks an individual mutation. 11 separate events occurred in these 3 genes.
experiments also demonstrated similar patterns of PI3 kinase activation in cells expressing mutant PIK3CD (SI Appendix (Figure S8) (17)).

Among PIK3CD, PIK3R1 and MTOR, only MTOR was found to be mutated in multiple cell lines (in addition to patient cases). We investigated the effects of a small molecule inhibitor of PI3kinase, BKM120 (Novartis), on the viability of 21 DLBCL cell lines (Figure 18D). The 3 cell lines with MTOR mutations had, on average, a 5-fold higher sensitivity to PI3 kinase inhibition than those 18 cell lines that did not harbor these mutations (P=0.005, Wilcoxon rank test). These results strongly suggest that the presence of mutations in MTOR is associated with sensitivity to PI3 kinase inhibition.
3.2.10 Comparison to other Genetic Studies Reveals the Striking Genetic Heterogeneity of DLBCL

Shortly after the completion of our study in June 2011, and during revisions, three separate studies exploring the genetics of DLBCL using similar methodologies and deep sequencing were published(9, 23, 24). Multiple studies applying these methodologies in the same cancer-type have generally been lacking thus far and these
publications provided an unusual opportunity for testing the overlapping mutations identified by the different studies.

We constructed Venn diagrams depicting the overlap between genes identified in the three studies and our study as shown in Figure 19. We initially assumed that our study comprising 73 primary DLBCL cases would be sufficient to identify the vast majority of recurrent genetic alterations in the disease. Surprisingly, we noted relatively modest overlaps of roughly 10-20% between the 4 different studies. Even genes that overlapped between different studies often varied. Similar patterns were observed when we simply compared the overlap between all somatically mutated genes in these studies. The overlap of more frequently mutated genes was still incomplete; when we limited the analysis to those 17 genes that were mutated in over 10% of the cases reported by Lohr et al(23), the overlap between the different studies including ours approached 70% (Figure S6). Even in that scenario, different studies had overlap with different genes. The remaining 30% of the genes mutated in at least 10% of the cases in that study were not detected by any of the remaining three studies. The overlap is still lower for genes with fewer mutation events. These observations suggest that there is considerable
genetic heterogeneity in the disease that contributes the observed patterns of disparate mutations.

Figure 19: Overlaps in Genes Discovered in Multiple Cancer Studies. The Venn Diagrams depict the comparison of gene mutations from the four DLBCL studies. The number in parentheses indicates the number of genes identified in each study. The gene lists were: Morin et al (Table S1, genes with confirmed somatic cases), Lohr et al (Table 1, top 58 genes), and Pasqualucci et al (Table S3 and Figure 4, validated somatic genes).

Although genes that do not overlap among studies might signify some false positives, our analysis indicates that a number of validated oncogenes and tumor suppressor genes were identified in just one study. Examples include NOTCH1(76) and CD74(76), BCL10, IRF4, MALT1, and TET2(9), BCR and ETV6(23), and PIK3R1, MTOR,
KIT, PDGFRA and ARID1A in our study. The number of identified cancer genes appeared to rise linearly as a function of the size of the study, further indicating that the differences between the individual studies arise from the inherent genetic heterogeneity of the DLBCL tumors, an effect that we also observed in other cancers (Figure S7).

### 3.3 Discussion

Through whole genome sequencing and whole-exome sequencing, we have identified the spectrum of sequence variation that occurs in DLBCL. Our data suggest that the majority of genetic variants in DLBCL are stochastically acquired. In all, we identified a total of 322 candidate DLBCL cancer genes that have recurrent somatic mutations in patients with DLBCL. We have identified a role for a number of known and novel oncogenes. Many of these genes, including ARID1A, KIT and IDH1 have not been previously implicated in DLBCL by any previous study.

A central observation of our study is the striking genetic heterogeneity that underlies a relatively common cancer. As we have demonstrated, there is a relatively low overlap between four different studies that explore the genetics of DLBCL. While differences in methodology, the diversity of patient populations, and number of patients might contribute to this low overlap, our data indicate that major driver of the low overlap is the inherent genetic heterogeneity of the disease. Consistent with this observed heterogeneity, we demonstrate that the number of rare variants and somatic
mutations rises linearly with the increased number of cases, suggesting that continued sequencing of tumors will implicate new variants and new cancer-related genes.

Gene expression profiling has previously revealed aspects of the heterogeneity of the disease, particularly with regard to cell of origin of DLBCLs (54). Our data indicate that recurrent mutations in 12 genes were clearly enriched between ABC DLBCLs and GCB DLBCLs. Thus, the two DLBCL subgroups share the mutational patterns of many more genes, suggesting that shared mechanisms underlie their biology. The striking genetic heterogeneity observed in the disease as a whole is also recapitulated in these subgroups.

The recognition of recurrent mutations in the gene coding regions in the disease is an important early step towards understanding its biology and potential therapeutic possibilities. Somatic mutations have previously been observed in multiple genes in the NF-κB pathway, including TNFAIP3 (A20) and CARD11 (77). Both of these genes were found to have somatic cases in our study and at least one of the recently published DLBCL studies (9, 23, 24). Histone-modifying genes, such as MLL2 and MEF2B, were found to be frequently mutated in DLBCL (32% and 11.4%, respectively) (9). While MLL2 was not included in our exome capture library, which was designed using build 36 of the human genome, we also observed somatic mutations in MEF2B and CREBBP, an acetyltransferase gene reported previously (76). MLL3, which forms complexes with MLL2, was the most frequently mutated gene in our cases. Our data also implicate
AICDA-related mutations as a major mechanism underlying genetic mutations in the genes PIM1, BTG1 and CD79B. These observations highlight the diverse biological mechanisms underlying the observed genetic diversity.

The genetic heterogeneity of DLBCLs and other cancers implies that no matter what recurrently altered gene or pathway is considered, only a minority of patients is likely to be affected. For that subgroup of patients whose tumors harbor a growing number of recognized genetic lesions that can be targeted therapeutically, the recognition of such alterations can make a crucial difference in their management. A number of genetic mutations we identified, including those in PIK3CD, KIT, and PDGFRA, suggest therapeutic possibilities in the affected patients. Our data suggest that such targeted therapeutic approaches in patients will need to be combined with carefully selected assays for those genetic lesions to better understand their role in response to targeted therapies. Our data also have major implications for how we model cancers and the need to ascertain whether extant mouse and other models recapitulate the primary disease. Thus, our study sheds new light on the genetic heterogeneity of lymphomas, as well as cancers in general, and underscores the need for individualized approaches for treating patients.

3.4 Methods
3.4.1 Sample acquisition and processing

Archival lymphoma tumors (N=73) and normal tissue (N=34) from 73 patients were obtained from the institutions that constitute the Hematologic Malignancies Research Consortium (HMRC) (30). These cases were anonymized, shipped to Duke University, and processed in accordance with a protocol approved by the Institutional Review Board at Duke University. RNA and genomic DNA were extracted from these 73 cases in addition to 21 DLBCL cell lines using column-based methods described previously(30).

3.4.2 Whole genome sequencing

3.4.2.1 Library Preparation

Whole genome sequencing libraries were prepared using methods described in the “Sample Preparation” section of the Agilent SureSelect protocol (pre-capture portion). Genomic DNA was sheared to 500 bp using Covaris settings: duty cycle-10%, intensity-5, frequency-200 cycles/burst, duration-135s, waterbath temperature-4°C, and quantified by BioAnalyzer (Agilent) using the DNA1000 chip. Then, it was end-repaired, A-tailed, and ligated to Illumina paired-end adapters at a ratio of 2µl per µg of DNA as quantified by BioAnalyzer. The ligated library was amplified for 6 cycles using Illumina PE PCR primers and 2x Phusion HF Master Mix. Post-PCR, the library was purified and assayed on BioAnalyzer to determine size and concentration. Libraries were diluted to 5 pM for Illumina clustering and paired-end sequenced over 9 days.
3.4.2.2 Sequence Alignment

Raw reads in fastq format were masked for Illumina adapter sequences, barcodes, and Phred-scaled base qualities of 10 and less using GATK(43). All the alignments were output as BAM files and merged using Picard (http://picard.sourceforge.net). PCR/optical duplicates were marked with Picard, and base quality recalibration and localized Indel realignments were performed using GATK (43). Read alignments were visualized with Integrative Genomics Viewer(2).

SAMtools mpileup with settings “-C50 -m3 -F0.0002” was run for the samples concurrently and output to a VCF file. Individual SNVs and Indels were annotated with gene names and predicted function using SequenceVariantAnalyzer(78), dbSNP130, HapMap v3 allele frequencies, 1000 Genome Project pilot 1 allele frequencies and CCDS Gene IDs using BEDTools, AWK, and custom Python scripts. Predictions for phenotypic severity of variants were determined using mutation assessor(79) (www.mutationassessor.org).

3.4.2.3 Structural Variant Calling and Annotation

The discovery of structural genetic aberrations was not the primary objective of this study. There are major limitations imposed by the short-read format in the detection of such variants and the methods for the detection of such variants are still evolving. We nevertheless surveyed our whole genome sequencing data using established methods to identify copy number variation and structural rearrangements.
3.4.2.4 Copy Number Variation

We identified copy number variants by using an approach similar to that described previously (63). In order to define the alterations in copy number throughout the genome, we began by segmenting the genome into non-overlapping intervals of 200 KB each. Each of these intervals represented an individual bin for identifying segmentation through a Hidden Markov Model. The copy number calculations were computed based on the model.

Briefly, we computed the total number of sequencing reads mapping to each 200 KB interval in both the tumor and the normal genomes from the same patient. The per-interval counts were median-centered for each sample, and then a ratio was computed of the number of reads mapping to the tumor and normal and log2-transformed. The copy number comparison between the whole-genome DLBCL and its matching normal is depicted in 200 kb intervals below (Figure 20). The y-axis is the log2 ratio between median-centered values for the DLBCL and matching normal. We identified 7 deletions and 3 amplifications. Known oncogenes that were implicated by these copy number alterations include PTEN (chromosome 10) and P16/CDKN2A (chromosome 9). We also identified small deletions in the immunoglobulin heavy chain and light chain loci that correspond to somatic rearrangement of these genes.
Figure 20: Copy number alterations in the DLBCL whole-genome compared to its matched control. Copy number changes are depicted on log2scale.
3.4.3 Identification of DLBCL Cancer Genes

The overall schema for identifying DLBCL cancer genes is summarized below:

- **All Variants in DLBCL:** 121,589
  - Select variants that are
    1. Present in tumor, but not present in paired normal
    2. Not a common source of genetic variation
    3. Annotated as frameshift, missense, nonsense, stop gained or stop lost in discovery set

- **Discovery Set:** 2589 Variants (2140 Genes)
  - Select variants that are
    1. Not a common source of genetic variation
    2. Annotated as frameshift, missense, nonsense, stop gained or stop lost in discovery or validation set

- **Discovery + Validation Set:** 4928 Variants (2140 Genes)
  - Select genes that have
    1. Recurrent mutations in discovery and validation sets
    2. Low rate of non synonymous variation in normal controls
    3. Variants predicted to significantly alter amino acid function

- **Discovery + Validation Set:** 1418 Variants (322 Genes)

**Figure 21: Summary of DLBCL variant and cancer gene identification from 73 primary tumor samples.**

Genes mutated in DLBCL were identified by analyzing the 73 primary tumor samples. The initial set of DLBCL mutations were determined from the 34 DLBCL primary tumors with paired normal samples, which constituted the discovery set. Data from cell lines were not used in this analysis.
For each of these cases, we identified mutations that were present in tumor but absent from the paired normal cases (somatically mutated). We eliminated common genetic variants by excluding those that occurred in the general population as identified from the following sources: dbSNP (18), publicly available data (pilot 1) from the 1000 genomes project (19), 256 recently published exomes from otherwise healthy individuals (20, 21, 67), one additional hapmap exome that we sequenced in this study, and those found to have a minor allele frequency of greater than 1% in the 6500 exome dataset from the NHLBI Exome Sequencing Project.

We identified 5884 variants that were somatically mutated in at least one of these 34 tumor-normal pairs. From this list, we identified 2589 variants that represented frameshift, nonsense, missense, or loss of a stop codon changes, corresponding to 2140 genes. These 2140 genes were examined in all primary DLBCL cases and found to have 4928 frameshift, nonsense, missense or loss of a stop codon variants. Among these variants were also 125 variants from 58 genes that were identified as potential mutational hotspots, occurring 4 or more times in DLBCLs (and none in controls). We estimated the functional impact of each of these 4928 variants on the encoded protein using a program that outputs a functional index score (described below). We also tallied the number of variants by gene and noted whether the gene had been previously annotated as a cancer gene in the COSMIC database (80). Finally, we estimated the rate of nonsynonymous variation in these genes in normal controls. We limited this analysis
to previously sequenced 257 exomes from otherwise healthy individuals because these cases have similar exonic coverage as our DLBCLs, and were processed using methods identical to those used to characterize the DLBCL exomes.

We generated a statistical model for genes likely to be drivers. It takes into account 4 features: gene size, background nonsynonymous mutation rates in normal samples, somatically acquired events, and the rate of these events in carriers. Given that mutations are rare and the number of genes is high relative to the number of samples, standard regression techniques do not apply. Also chi-squared tests of independence, or other similar tests, for each individual gene, besides having the obvious problem of multiple testing, would never account for important mutants that occur with very low frequency but have other important characteristics. After filtering for genes in which we observed a minimum of 1 somatic event and 1 additional rare event from the same class or presence in the COSMIC database, we ranked genes based on their distance from known cancer genes.

We calculated the distance of a gene from a pool of known cancer genes based on the 4 variables listed above. Let \( \mu \) and \( \Sigma \) be the mean vector and covariance matrix for the population of known cancer genes, from which we calculate cancer gene candidate \( j \)'s distance. We use the well-known Mahalanobis distance \( D=(g_j-\mu)^T \Sigma^{-1}(g_j-\mu) \) where \( \Sigma = \frac{n_1\Sigma_1+n_2\Sigma_2}{n_1+n_2-2} \).
This distance and has several desirable properties: unlike the Euclidean metric, which can only deal with circular forms and is a special case of \( D \) where \( \Sigma = I \), the identity matrix. Here, because \( \mu \) and \( \Sigma \) are unknown, we estimate them from the same mean and sample covariance matrix for the population. In the special case where we assume that the population is multivariate normal, \( D \) is drawn from the chi-squared distribution with \( p \) degrees of freedom. Therefore we can test whether each individual gene belongs to the population of cancer genes or not, based on this assumption.

We compared two populations: known cancer genes from the literature and our candidate novel cancer genes. The p-value measuring the level of distance is calculated from the F-distribution \( F(d_1,d_2) \) with \( d_1 \) and \( d_2 \) degrees of freedom respectively:

\[
D \sim F(p, n - p - 1)
\]

Genes closest in distribution (P<10^{-6}) to known cancer genes were identified as DLBCL cancer genes. Previously annotated cancer genes were required to have 1 involved case, while genes that were not previously annotated as cancer genes were required to have a minimum of two involved primary tumor DLBCL cases. Using these statistics, we found that 90% of the known cancer genes and the newly identified DLBCL cancer genes had at least one variant with a functional index of 0.9 or higher, and a rate of non-synonymous variants of less than one per case in the unmatched controls.
Using these criteria, we identified a total of 426 genes that were recurrently mutated. Excluding those for which more than two-thirds of the variants also were found by the NHLBI Exome Sequencing Project, 322 genes remained. 52 genes within this list were previously annotated as cancer genes.

3.4.4 Biological Validation

3.4.4.1 Cell Culture

As described previously(28), Lymphoma cell lines were cultured with RPMI1640 media supplemented with 10% v/v Fetal Bovine Serum (FBS) and 1% v/v Penicillin/Streptomycin supplied at 10,000U penicillin and 10 mg streptomycin/ml (BJAB, Farage, Karpas422, Pfeiffer, RL, SCI 1, SKI, Toledo, U2932, WSU_NHL, HT), or RPMI1640 supplemented with 15% v/v FBS and 1 % v/v P/S (RCK8, SUDH4, SUDHL7), Iscove's modified dulbecco's medium (IMDM) supplemented with 20% v/v human plasma and 1% P/S (RCK8, SUDHL4, SUDHL7), or alpha-mem media supplemented with 10% FBS and 1% P/S (TMD8). Cells were grown in a 5% CO2 environment at 37°C.

3.4.4.2 Cell Viability Assays

The effect the PI3K inhibitor drug BKM120 (Novartis) on the 22 lymphoma cell lines sequenced in this study was assayed on 96-well format using MTT viability assays. 40,000 cells were grown in drug at 10 concentrations: 25µM, and 9 serial 1:2 dilutions down to 0.05 µM. For each condition tested, there were 5 technical replicates. Two controls were included: media only and drug-free cells. If the IC50 was not reached
within this drug concentration range, the experiment was repeated with an additional
drug concentration of 50µM. After 48 hours, 15µl MTT was added to each well, and the
plate was incubated at 37°C for 4 hours. 100µl of MTT detergent solution was added,
and color was developed in the dark at room temperature overnight, after which 570 nm
absorbance was measured. After normalization to control wells, the IC50 values were
calculated (Table S9).

3.4.4.3 PI3 Kinase Dependence

To model the three dimensional structure of the PI3KCD protein, we
threaded the protein sequence (NP_005017) through the crystal structure of the PI3KCG
protein (PDB:1HE8) using the PHYRE algorithm(81). Given the very strong homology
through the core structural domains, the output was of high quality. Thus, the catalytic,
Ras binding, and C2 domains were easily discernible and properly oriented in the
PI3KCD model as displayed in Figure 18D. The PyMol program (DeLano Scientific) was
used for positioning of somatically mutated residues observed in our sequencing studies
and rendering of the model figure.

FL5.12 myristoylated Akt (myrAkt) cells obtained from Jeffrey Rathmell’s
laboratory were cultured in RPMI (Invitrogen, 11875-093) supplemented with 10% FCS,
500 pg/mL recombinant mouse IL3 (rmIL3), 2 mM L-glutamine, 10 mM HEPES, 1% v/v
Penicillin/Streptomycin (Invitrogen, 15140-122) and 0.1% v/v βME (Invitrogen, 21985-023). Cells were split down to a density of 50k/ml daily.
The PIK3CD shuttle clone (Genecopoeia, GC-M0163-CF) open reading frame was inserted into the pEF-DEST51 plasmid (Invitrogen, 12285-011) using Gateway cloning (Invitrogen, 12538-120). The point mutation was created by site-directed mutagenesis (Stratagene, 200521). Wild-type and mutant PIK3CD plasmid insert sequences were confirmed over the entire length of the ORF by Sanger sequencing.

For western blot analysis, 1.5 million FL5.12 myrAkt cells were transfected with 2µg wild-type or mutant by Amaxa (Nucleofector V, program G-016), concurrent with addition of doxycycline at 1µg/ml in the media to induce myrAkt expression. At 18 hours post-transfection, each transfection was split in half and washed twice in Phosphate-buffered saline. The control cells were re-suspended in normal FL5.12 growth media, whereas the remainder was re-suspended in media lacking IL3. P-Akt was measured by Western blots 3 hours later to compare the cells in which IL3 was replaced to those in which it was withdrawn.

Cells for PI3K activity ELISA were transfected in a similar manner, but with 50 M cells transfected in 5 batches of 10M cells by Amaxa. Cells were also subject to IL3 withdrawal 18H post-transfection and harvested 3 hours later.

RIPA Lysis buffer (1 × phosphate-buffered saline [PBS], 1% Nonidet P-40, 0.5% sodium deoxycholate, 0.1% SDS, 10 mM phenylmethylsulfonyl fluoride, 1 µg/mL aprotinin, and 100 mM sodium orthovanadate) was added to 750,000 cells and incubated on ice for 30 minutes. The mixture was spun down and the supernatant was transferred...
to a new tube as the whole cell extract. A total of 20 µg of cell lysate was separated on a 4 – 18 % Tris-Bis NuPAGE gel (Invitrogen) and transferred using the iBlot transfer device (Invitrogen) program 2 for 6 minutes. The blots were probed using 1:1000 rabbit-phospho-AKT (Cell Signaling Technologies, #4060), 1:2000 rabbit totalAKT (Cell Signaling Technologies, #9272) and 1:1000 mouse-anti-β-actin (Santa Cruz Biotechnologies, sc-47778) overnight at 4°C. The antibodies were detected using 1:10,000 goat-anti-rabbit or 1:10,000 goat-anti-mouse horse radish peroxidase conjugated antibodies (Santa Cruz Biotechnologies). Western Blotting Luminol Reagent (Santa Cruz Biotechnologies) was used to visualize the bands corresponding to each antibody.

After 3 hours of IL3 withdrawal, cells were washed in PBS and lysed in 80 µl lysis buffer consisting of 10 mM Tris pH 7.4, 150 mM NaCl, 1% Triton X-100, 1% deoxicholic acid, 0.1% SDS, and 5 mM EDTA supplemented with 1% each of protease inhibitor cocktail (Sigma cat#P-8340) Serine/Threonine phosphatase inhibitor cocktail (Sigma cat #P-2850), Tyrosine phosphatase inhibitor (Sigma cat #P-5726) and PMSF (100mM stock). The lysate was vortexed 10s, incubated on ice for 10 minutes, vortexed again, and then incubated on ice again. Then, it was sonicated using Covaris settings: duty cycle:5%, intensity:4, Cycles/burst:200 for 2 1-minute pulses. The lysate was centrifuged at 4°C for 10 minutes at 16,000 rcf, and the supernatant was transferred to a fresh tube.
100µg of protein was used per well for PI3K activity measurement by ELISA (Echelon Biosciences part number K-1000s) per manufacturer instructions.
4. Integrative genomic and clinical analysis of 1001 diffuse large B cell lymphoma (DLBCL) patients reveals genetic drivers and risk groups

4.1 Introduction

As described in the previous chapter, the largest single genetic study of DLBCL involved 94 cases (17). Analysis indicated that the overlap in concordant mutations in four separate DLBCL studies (17) was relatively low because the vast majority of recurrently mutated genes affect fewer than 10% of cases, so that sub-100 case studies are incapable of comprehensively detecting most such genes. Based on power calculations by Lawrence et al, for cancers with a mutation rate of DLBCL, 1000 cases need to be sequenced.
To comprehensively define the genetic underpinnings of DLBCL, we embarked on the largest single-cancer exome sequencing study to date. Using custom in-house high-throughput methods, we performed whole exome sequencing on over 1000 DLBCL, which is an order of magnitude larger than existing DLBCL studies. Matched normal DNA was available for half of the cases and enabled the measurement of somatic mutations. The cases originated from institutions all over the world and encompassed all stages of the disease. While previous studies were only powered to detect mutations

Figure 22: Power calculations by Lawrence et al (1) indicate that 500 paired cases of DLBCL need to be sequenced to sensitively detect genes recurrently mutated in 5% of DLBCL.

http://www.tumorportal.org/figure/5
present in at least 15% of samples, this 1000-sample study is powered for detection of mutations in fewer than 5% of cases.

Significant and creative method development was required to complete this endeavor, to address the sheer sample size, as well as the sample source. First of all, in order to perform whole-exome sequencing on 1001 cases, new methods would be necessary because simply applying existing methods 10 times over would require a duration 10–times longer, as well as cost 10-fold more. Secondly, the development of methods to exome-sequence DNA from Formalin-Fixed Paraffin-embedded samples widened the set of biopsies eligible for genetic analysis. The adoption of sequencing from FFPE rather than frozen DNA as used previously required careful comparative analysis of exome sequencing of FFPE and Frozen DNA from the same lymphoma cases.

The method development to increase the throughput of exome sequence and expand analysis to FFPE DNA is detailed in the methods section of this chapter.

The other significant difference, besides scale, in this study compared to the one described in chapter 3, is the availability of clinical data. Many clinical features of DLBCL have been previously well-defined, but this study is unprecedented in scale and enabled the identification of both clinical and genetic features that correlate with survival outcome.
4.2 Results

4.2.1 Exome sequencing overview

Whole-exome sequencing at an average depth of 50x was applied to 1001 cases of DLBCL, as well as paired normal DNA in half of those cases. All 1001 cases also included detailed clinical features including treatment response, patient age at diagnosis, patient demographics, recurrence, as well as disease-free and overall survival.

Raw fastq data were processed in a similar way as described previously (4, 17), to detect deviations from the reference genome in the form of single-nucleotide variants (SNVs), small insertion-deletions (indels), and copy number alterations.

4.2.1 Interesting clinical features

The International Prognostic Index is a clinical score used to rank lymphoma patients based on In 2007, it was suggested that the International Prognostic Index should be revised in the post-Rituximab treatment era because the 6 stages were observed to cluster into three groups (82). In fact, what we see is that in the current study of 1001 cases all treated with Rituximab, the original IPI classification still separates the cases by risk. Thus, with adequate sample size, the original IPI score still indicates outcome in patients, as shown in the figure below.
4.2.2 Gene-level mutational patterns identify oncogenes vs. tumor suppressors

Recurrent mutations were observed at three genomic scales of increasing size: the single amino acid, the gene, and the gene network.

With 1001 cases at our disposal, it was possible to gauge the spatial distribution of mutations along genes in order to detect enriched regions of mutation, even in some cases to the precision of a single base.

4.2.2.1 The mutational pattern of oncogenes

For example, the most striking single-base signal was a T-C transversion at hg19 genomic location chr3: 38182641 that results in the activating MYD88 L265P mutation, which has previously been reported in Waldenstrom’s macroglobulemia (83). This exact
point mutation independently occurred in an astounding 143 (14%) of cases, and was the strongest proto-oncogene in the study. It was clearly not the result of the population background of the cohort, as the variant was determined to be absent in the matched normal DNA of those individuals, and also exceedingly rare in the general population. Thus MYD88 L265P is the strongest single driver mutation in DLBCL. However, it is not the only such mutation; another strong example is the EZH2 Y646F/N missense mutation, which was found in 65 (6.5%) of cases. EZH2 is a histone methyltransferase that effects tri-methylation of H3K27, a repressive chromatin modification. The mutational patterns of EZH2 and MYD88 visualized using the cbioportal tool are shown below.
In fact, these hotspot mutations were such strong indicators of the nature of the DLBCL that cases could be risk-stratified based on their presence alone. Similar to the use of markers such as Estrogen/Progesterone/HER2 receptors routinely measured on all breast cancer diagnoses, one could imagine strong and specific genetic markers such as EZH2 Y646F/N and MYD88L265P to be routinely assayed from DLBCL biopsies, to inform disease subtype, treatment, as well as prognosis. The figure below shows corresponding Kaplan-Meier plots for DLBCL cases separated solely based on status of these corresponding mutations.

Figure 24: EZH2 and MYD88 exhibiting classical mutational patterns of activating mutations. Each panel shows the coding regions, with key domains marked along the x-axis. The y-axis indicates the number of DLBCL cases affected. Along each gene, mutations are indicated; recurrent mutations are scaled based on the number of cases. The dominating feature within each gene is a mutation peak at a single amino acid, indicating strong selection for that alteration in the activation of that particular gene.
The pattern of a particular recurrently mutated gene, PIK3CD, highlights the power of sequencing 1001 cases of DLBCL compared to fewer than 100. In that smaller, 94-case study detailed in the previous chapter, PIK3CD was identified as an oncogene based on the observation of 3 mutations in the catalytic domain. However, upon revisiting PIK3CD mutations, this time in 1001 cases, we noticed another strong hotspot at the R38C location. With observations in 25 patients, R38C became a glaringly obvious hotspot, but because it only represents a frequency of 2.5%, our previous study was not powered to detect it nor to classify it as a hotspot mutation.

The exact same corresponding amino acid changes have previously been observed to be activating in the alpha subunit of PI3K, PIK3CA, thus lending more

Figure 25: Single amino acid modifications are sufficient to risk-stratify DLBCL. As shown on the left, EZH2 Y646F/N mutations are indicative of a less aggressive form of DLBCL with better overall survival, whereas MYD88 L265P separates a subtype of DLBCL with poorer overall survival.
confidence to the hypothesis that they are also activating in our observations in PIK3CD. Moreover, they are tantalizing therapeutic targets for PI3 Kinase inhibition, for which several drugs already exist. Once again, as was seen with EZH2 and MYD88, the PIK3CD hotspot mutations alone were sufficient to risk-stratify the disease.

Figure 26: Comparison of mutational patterns in PIK3CD in 94 cases of DLBCL vs. 1001 cases. The R38C hotspot was missed completely in the smaller study. In the Kaplan-Meier plot on the right, we also see that these point mutations appear to stratify a subtype of the disease with better overall survival.

Thus, increasing the sample size by an order of magnitude not only increases the sensitivity to detect genes recurrently mutated in 5% or fewer cases, but also to detect the important locations of importance within a particular recurrently mutated gene.

To summarize, additional genes exhibited a hotspot-like mutational pattern in a similar manner to MYD88, EZH2, and PIK3CD, and this mutation pattern as a strong indicator of proto-oncogene activating mutations. The observed hotspot pattern is quite logical, as there are limited ways in which to hyper-activate a gene, and thus a disease
that selects for gene activation through sequence modification can only do so through limited avenues, with very specific modifications in the amino acid sequence.

4.2.2.2 The mutational pattern of tumor suppressors

The previous section detailed how repeated mutational observations within the same gene enabled the detection of oncogenes. Likewise, repeated mutation observations provide the spatial data to detect tumor suppressor genes, but the pattern is different from that of oncogenes. The hallmark mutational pattern of a tumor suppressor gene is an abundance of stopgain mutations, usually in the earlier portion of the gene, as well as frameshift mutations. The intuition is easy to understand; whereas there are very few ways to turbocharge a gene, there are infinite ways to disable it, through stopgain mutations that truncate the amino acid polymer, frameshift mutations that garble all subsequent protein structure, or strategically placed missense mutations that substitute a small amino acid for a bulky one, or a neutral one for a charged one, that wreak havoc on active pockets or protein folding.

Several such examples are shown below in the figure for genes NFKBIA, NFKBIE, and SPEN. As we can see in the figure, each gene is littered with stopgain mutations throughout, and the frequency of mutations is more evenly distributed, rather than concentrated at a specific amino acid. These genes also make sense in the context of lymphoma. NFKBIA and NFKBIA are inhibitors of the NF-kB pathway, which is well-known to be activated through over-expression in ABC-DLBCL, and also through
activation via MYD88 L265P as described in the previous section. One can imagine that MYD88 L265P is equivalent to pushing down the gas pedal on the NFkB pathway, whereas disabling NFKBIA or NFKBIE through loss-of-function mutations is similar to cutting the brake lines, all to the same effect.

SPEN is introduced in this study as a novel candidate tumor suppressor gene in DLBCL. It has previously been reported to be recurrently mutated, with inactivating events, in marginal zone lymphoma (84). The normal function of SPEN is to repress NOTCH signaling and thus negatively regulate B cell differentiation into MZ B cells. SPEN was one of the strongest previously unknown genes found to harbor abundance of inactivating mutations, and it supports the validity of our gene discovery methods that SPEN also turns out to be a gene of functional significance in B cell development that has already been reported in another lymphoma type.
4.2.3 Integrating mutational patterns and copy number alterations to identify recurrently altered genes in DLBCL

As described in the previous section, oncogenically-mutated genes tend to have focal amino acid modifications (MYD88; PIK3CD; EZH2; etc.), whereas a tumor-suppressor-mutated gene has a mutational pattern abundant in stopgains, frameshifts, and missense mutations throughout the gene (NFKBIA, NFKBIE, SPEN). In conjunction with gene-level mutational patterns, copy number alterations can also indicate whether

Figure 27: Examples of genes with recurrent inactivating mutations in DLBCL. For each gene, the union of mutations detected in 1001 DLBCL cases is show using the cbioportal lollipop tool. The height of each mutation indicates the number of observations. Red dots indicate stopgain or frameshift mutations, and green dots indicate missense mutations.
a gene has a tumor-suppressor or oncogene role. Tumor-suppressor genes are expected to exhibit loss-of-function through mutations or copy number loss, and oncogenes are expected to be activated via primarily focal missense mutations or copy number gain. Using copy number calls as determined by the exome sequencing data, we were able to compare each gene’s mutational profile to its copy number profile, and develop a gene-level “oncogene score” and “tumor suppressor score”. The top genes are depicted in the figure below.

![Figure 28: Genes with recurrent activating alterations (left) and loss-of-function alterations (right) in 1001 DLBCL.](image)

Many key genes categorize as expected, such as for example previously identified oncogenes BCL2 and MYC, or tumor suppressor TP53. One surprising gene that is identified as activating is NOTCH2, which exhibits many nonsense mutations, which would typically be expected to result in loss of protein function. However, on
closer examination, one notices that the nonsense mutations all result in the deletion of
the regulatory PEST domain of the gene, resulting in increased stability of the protein
product.

4.2.4 ABC/GCB logistic regression mutational classifier

As described in chapters 2 and 3, DLBCL can be classified based on microarray
gene expression into two subtypes, Activated-like DLBCL, and Germinal-center-like
DLBCL. These names come from the supposed cells of origin of their respective
subtypes. The subtypes also are of clinical importance, because ABC-DLBCL patients
typically have poorer prognoses compared to GCB-DLBCL.

As described in section 3.2.6, it was apparent that these subtypes of DLBCL
exhibit not only gene expression differences, but mutational differences as well (Figure
3). From 35 cases of GCB-DLBCL and 29 cases of ABC-DLBCL, 12 genes were found to
be differentially mutated, with 8 more frequently mutated in ABC-DLBCL, and 4 more
frequently mutated in GCB-DLBCL (Figure 14). Now, with an order of magnitude more
cases available, we wondered if we could develop a novel classifier of ABC and GCB
subtypes of lymphoma, using not gene expression data, but mutational data.

Using logistic regression trained on ABC/GCB labels based on gene expression,
we created a mutation-based predictor for ABC and GCB subtypes that achieved 80%
cross-validation accuracy, which suggests exome sequencing data could be utilized in
the absence of gene expression data to classify ABC and GCB subtypes. A few of the most significantly mutated genes are shown in the mutational heatmap below.

4.2.5 Identifying recurrently mutated pathways and their significance

The 407 most recurrently mutated genes in 1kDLBCL were identified based on enrichment of somatic mutations, lack of population variability, and predicted function of mutation.
In order to understand the pathways enriched in DLBCL, we utilized Reactome (www.reactome.org), a curated pathway database. However, within reactome, a particular gene may appear in multiple pathways, and any given pair of genes that are not particularly tightly related may appear in the same pathway because said pathway contains over a thousand genes. To create a consensus pathway similarity metric, the co-occurrence of all pairs of genes was computed, the logic being that if a given pair of genes very frequently appears in the same pathways, they are likely to be more closely
interacting. After computing the gene similarity matrix, genes were clustered based on similarity in pathway membership (Ward-distance), as shown below.

Figure 30: Clustering of genes recurrently mutated in 1001 DLBCL. Top: overall heatmap of gene clustering based on pairwise Jaccard distance. Below: A zoomed view of gene blocks at the bottom, to highlight some genes. Note the tight clustering of HRAS/KRAS/NRAS, as well as that of the PI3K family.
Guided by the clustering of recurrently mutated genes and collective information from the Reactome pathway database, as well as the GeneCards database, we developed a pathway-centric view of genes mutated in DLBCL, as depicted in the network figure below. Thus, with 1001 cases, it is possible to group DLBCL cases by mutational cases not only by direct gene overlap, but also by pathway.

We observe pathways widely mutated in all malignancies (apoptosis, N = 389; cell cycle, N = 208), as well as those known to be important in B-Cell biology (NFkB; N = 454, Interferon signaling; N = 176), and even those pathways that were initially discovered to be important in other malignancies (RAS signaling; N = 116, PI3K; N = 188). Within each subnetwork, gene alterations occurred in consistent patterns; in NFkB, positive regulators of the pathway such as MYD88 and CARD11 were activated, whereas negative regulators such as TNFAIP3, NFKBIE, and NFKBIA were inactivated.

These pathways demonstrate the overarching mutational patterns of DLBCL in the context of general hallmarks of cancer, cell of origin B-cell biology, and therapeutic targets via drug repurposing through pathway inhibitors first applied in diseases other than lymphoma.
Figure 31: Recurrently mutated pathways in DLBCL. In the center is depicted a Cytoscape gene network of recurrently mutated genes in DLBCL, grouped and color-coded by pathway. The size of each gene is scaled to the number of samples observed to be mutated in DLBCL. For example, some of the most frequently mutated genes include MYD88 and KMT2D/MLL2. Surrounding the network are mutational heatmaps color-coded by mutational type; yellow indicates missense, green stopgain, blue copy number loss, and red copy number gain.
4.2.6 Association of gene expression, mutation, and copy number with DLBCL survival

We examined the associations between the mutations and clinical outcome in all 1001 patients. All survival analyses were conducted using nearly equally split training and validation sets, corrected for multiple comparisons with significance of \( P<0.01 \) in the validation set (Figure 32). The cell of origin classification was strongly associated with survival in our cases and was independent of MYC and BCL2 co-expression, which was separately associated with survival (Figure 32A). Figure 32B shows hazard ratios for select genes, as well as associated Kaplan-Meier survival curves for a subset of those genes. We further identified combinations of different genetic and expression features that point to context dependence for survival associations (Figure 32C). For instance, mutations in KLHL14 were associated with a particularly poor prognosis in ABC DLBCL, while CREBBP mutations in ABC DLBCL patients were associated with better prognosis than average GCB DLBCLs. Mutations in EZH2 and CD70 were associated with a highly favorable prognosis within the GCB DLBCL subgroup. TP53 mutations were found to be prognostic only in the presence of MLL2 mutations and high BCL2 expression. Importantly, these risk groups are mutually exclusive and inform clinical outcome significantly better than existing metrics.
Figure 32: Features that stratify DLBCL by hazard ratio. On the left are forest plots of features, and on the right are Kaplan-Meier curves of select examples. (A) Gene expression-based features. (B) Gene-level mutational and copy number features. (C) combinations of features that result in additive differences in overall survival.
4.3 Discussion

In this largest single-cancer whole exome sequencing study to date, we have comprehensively and sensitively identified genetic alterations in DLBCL and used them to detect underlying gene network patterns that clinically stratify the disease. These findings advance our understanding of the heterogeneous genetic alterations of DLBCL and inform therapeutic targets to utilize in personalized medicine approaches to treating these patients. As large-scale gene expression microarray studies have enabled advances in understanding of DLBCL 15 years ago, this exome sequencing analysis of 1001 cases today adds an entirely new dimension to our understanding, through the axis of mutational data.

As sequencing costs continue to drop, the cost of whole-exome sequencing will converge on that of other routine medical tests, and the cost will instead shift to the analysis of that data. Our study takes the first step in that direction, by deeply exploring the mutational heterogeneity in DLBCL and the relationship between those mutations and clinical outcome, as well as ABC/GCB subtype.

4.4 Methods

New methods had to be developed for this large-scale DLBCL exome project in two areas: FFPE exome sequencing, and decreasing the costs while increasing the throughput of exome library preparation.
4.4.1 Exome sequencing of DNA from FFPE samples

4.4.1.1 Introduction to FFPE samples and motivation

The most common and cheapest way to archive tumor tissue is Formalin-Fixation and paraffin-embedding (FFPE). In 2007, it was estimated that more than a billion tissue samples were stored around the world, mostly as FFPE. The ability to tap into these rich archives rather than restrict exome sequence analysis to fresh frozen biopsies drastically expands the sample size and statistical power of sequence studies. However, FFPE preparation requires nucleic acid to be cross-linked and as a result fragmented into pieces of several hundred base pairs. Moreover, FFPE may cause artifactual changes, predominantly C:G → T:A (85).

To analyze the effect of FFPE on DLBCL mutation discovery, we performed side-by-side parallel whole-exome sequence analysis of 12 sets of paired FFPE and frozen tumors, and evaluated our overall mutation discovery sensitivity compared to previous DLBCL studies utilizing fresh-frozen biopsies.

The analysis show that when extracted and sequenced properly, data from FFPE and frozen samples of the same individual are equal in mutation discovery sensitivity, and that a dataset of DLBCL mutations derived solely from FFPE biopsies is equally or more powerful than previous datasets derived from frozen biopsies.
4.4.1.2 Parallel analysis of frozen and FFPE DLBCL samples from the same individual indicates high variant and mutation overlap

Paired FFPE and Frozen biopsies from the same DLBCL case were prepared and whole exome sequenced identically, and then compared to one another. To apply the comparison, population variants were removed from analysis to remove the component contributed by common ethnicity, and to enrich for those variants and tumor mutations that are specific to a particular individual. After removal of the population component, a similarity metric between a sample and all other samples was computed as the number of population-rare variants shared as fraction of the total number of variants in the sample, with 0 equal to no overlap, and 1 equaling complete overlap. As can be observed in the figures below, there is a significant rare variant similarity between FFPE and Frozen DLBCL biopsies from the same individual (yellow/bright blue, high overlap), compared to unrelated samples (dark blue, poor overlap). A perfect overlap is observed when a sample is compared to itself. This similarity metric analysis indicates that sequencing of FFPE tissue preserves variants and mutations specific to that DLBCL case, and that FFPE preparation is not generating false variants that cause the FFPE biopsy to differ from the frozen biopsy.
Figure 33: A rare overlap heatmap is shown for 11 FFPE-frozen pairs, ordered by pair. It is evident that pairs are significantly more similar to one another (off-diagonal blocks), than to cases from other patients. The values mapped on the diagonal indicate perfect overlap of each sample with itself.

Figure 34: The distribution of fraction of variants overlapping between FFPE and frozen cases from different individuals (blue) is compared to that of cases from the same individual (orange) via box plots.
4.4.1.3 Mutational spectra are identical between FFPE and Frozen cases

An additional concern of FFPE preparation is the introduction of an FFPE artifactual single-nucleotide, typically C:G > T:A. Equivalent mutations were collapsed to 6 possibilities: C:G > A:T, A:T > T:A, C:G > G:C, A:T > G:C, A:T > C:G, and C:G > T:A. The distribution of these 6 possible mutations was compared between FFPE and frozen cases. We find that the mutational spectra are indistinguishable between the two groups, suggesting that the C:G > T:A artifact is not a significant contributor to false variant calls.

An additional concern of FFPE preparation is the introduction of an FFPE artifactual single-nucleotide, typically C:G > T:A. Equivalent mutations were collapsed to 6 possibilities: C:G > A:T, A:T > T:A, C:G > G:C, A:T > G:C, A:T > C:G, and C:G > T:A. The distribution of these 6 possible mutations was compared between FFPE and frozen cases. We find that the mutational spectra are indistinguishable between the two groups, suggesting that the C:G > T:A artifact is not a significant contributor to false variant calls, as shown in the stacked bargraph below.
4.4.1.4 Sequence depth and coverage

An additional concern with respect to FFPE samples is that the preservation method may cause some regions of the genome to be sequenced unevenly. To measure the depth and coverage, we analyzed FFPE exon coverage as a function of input reads. If a particular sample has poorly sequenced regions, then the coverage throughout the genome will be uneven and cannot corrected by increasing the overall depth. On the other hand, if the genome is being sequenced evenly, then the fraction of the exome region with depth above threshold will become asymptotically close to 1 as the number of sequence reads increases.

Figure 35: The mutational spectra is compared between FFPE and Frozen sequencing data from the same individual.
4.4.1.5 Gene discovery using FFPE cases

The gene discovery sensitivity of FFPE DLBCL cases can be evaluated by whether it confirms genes previously identified in other studies. We expect the genes discovered to far exceed previous numbers recorded because our study size is 1000, and power calculations suggest that we are powered to detect genes mutated in as low as 3% of DLBCL, whereas previous sub-100 – sample studies were only powered to comprehensively detect genes mutated in at least 30% of cases. The overlap is not 100% because 1. gene discovery bioinformatics methods differ from study to study, 2. the patient cohort may differ, especially for the smaller studies, and 3. smaller studies will still, by chance, detect some of the rarely mutated DLBCL genes not found in this larger study.

Figure 36: Percent of region sequenced at depth of at least 20 as a function of number of uniquely mapped reads. The fraction of the exome sequenced at depth of 20 or greater is plotted against the number of reads uniquely mapped for FFPE (blue) and frozen (orange). The overall fitted curve is indistinguishable between frozen and FFPE samples. Thus, once number of reads is fixed, the depth and coverage are identical for FFPE and frozen samples.
As shown above, with the exception of 2 genes, our 1kDLBCL FFPE DLBCL study re-discovers all genes detected by 2 or more previous studies, suggesting that non-overlapping genes occur because of slight differences in methods and patient cohorts of prior studies.
4.4.1.6 Conclusions

By evaluating variant overlap, mutational spectra, coverage, and gene overlap with previous studies, we show that FFPE cases are equivalent to frozen cases for mutation detection. FFPE exome sequencing of tumors drastically expands the number of cases available to sequencing.

4.4.2 High-throughput Exome Library Preparation

To increase the throughput of exome library preparation, a custom, in-house 96-well barcode system was designed based on the Illumina sequencing system whereby a separate read is used to identify the barcode of a particular library. This enabled the libraries to be exome-captured in the same pool, thus saving costs on baits as well as increasing the throughput of the preparation process. Extensive method development was required to maximize the capture efficiency, design barcode sequences with maximal edit distance yet optimal composition for sequencing, and balance the relative abundance of libraries such that in multiplexed format they were sequenced as evenly as possible.
5. Conclusions and future directions

We have described here a journey towards understanding the driver mutations of lymphoma, first by looking back and understanding the epigenetics of their normal cells of origin, then by characterizing the full breadth of genomic heterogeneity in the most common non-Hodgkin’s lymphoma, DLBCL, with an eye towards its connection to the clinical outcome of the patient.

Future directions include biological validation of candidate mutations in cell line and mouse models, exploration of intratumoral heterogeneity, drug repurposing for personalized medicine, and analysis of germline variants.

Experimentally, these studies have yielded high-confidence driver mutations; however the underlying mechanisms and specific biological context require experimental exploration via introduction of mutations to cell line and mouse models, and perturbation of the involved pathways.

Some of the most interesting future directions involve exploring the genetic intratumoral heterogeneity within individual lymphoma cases, rather than between cases. The primary focus of this dissertation has been to characterize and understand lymphoma at a population level, or how it differs from patient to patient. However, within each case, there is a population of subclones that undergo selection due to treatment and competition for resources from the host, and this is a currently untapped avenue for further exploration. Comprehensive characterization of subclones requires
sequencing at higher depth, perhaps an order of magnitude more than was applied here, as well as sophisticated analytical tools. Although the cost of sequencing has dropped rapidly, it is still prohibitively expensive to sequence 1001 cases at extremely high (500-1000x) depth. A suitable compromise would be to select cases that have the most interesting subclonality structure based on existing data, for example based on the abundance of many low allelic-fraction variants.

As alluded to in Chapter 4 in the discussion of biological pathways, we discovered some recurrently mutated genes that are clinically actionable using drugs typically applied to other malignancies. Those cases, even if they affect a small fraction of DLBCL patients, are excellent candidates for clinical trials that utilize drug repurposing. It would be an improvement to the current first-line therapy for DLBCL, which is not tailored to the mutational profile of the tumor. Moreover, such trials would augment the understanding of drug effectiveness on the same actionable genetic target in different disease and mutational contexts.

A final avenue for future directions involves consideration of germline variants. The current studies were focused on somatically acquired mutations, but with increasing power and higher-resolution understanding of normal human variability, it may be possible to understand lymphomas in the context of germline background with respect to susceptibility and overall prognosis.
Thus, we hope the work laid out here will serve as a starting point for further applications at the bench, bedside, and computer.
References


52. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high


identifies robust subtypes including one characterized by host inflammatory

56. Ley TJ, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid

57. Chapman MA, et al. (2011) Initial genome sequencing and analysis of multiple


183.


Biography

Jenny was born in Maryland in 1984. She graduated from MIT in 2006 with a B.S. in Physics. Following a one-year post-baccalaureate IRTA at the National Institutes of Health and three years of research at Duke University, she entered the Duke University Program in Genetics and Genomics in 2010.

Publications


