Fall 12-18-2015

The Role of Genetic Alterations in Tumor Initiation, Progression and Transformation

Weiwei Zhang
University of Nebraska Medical Center

THE ROLE OF GENETIC ALTERATIONS IN TUMOR INITIATION, PROGRESSION AND TRANSFORMATION

by

Weiwei Zhang

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Pathology & Microbiology Graduate Program

Under the Supervision of Professors Wing C. Chan and Kai Fu

University of Nebraska Medical Center
Omaha, Nebraska

November, 2015

Supervisory Committee:

Dhundy Bastola, Ph.D. James D. Eudy, Ph.D.

Shelley D. Smith, Ph.D.

ACKNOWLEDGEMENTS

It has been a great journey. I have been blessed by many kind people in both my personal and professional life.

First and foremost, I would like to thank my mentor, Dr. Wing C. Chan for his dedicated supervision and support throughout the research of this project. Dr. Chan displays an outstanding eye as a scientist with an unprecedented respect towards individuals. He gave a lot of advice, helpful in my research and life. I am grateful that I still get enough attention from him in spite of his busy schedule even though I am far from California.

I would also like to thank Dr. Kai Fu for being my co-mentor and taking care of me when my lab moved to City of Hope. I am greatly appreciative of his support and the trust he has provided for me.

I would like to thank my exceptional supervisory committee, Dr. Shelley Smith, Dr. James Eudy, and Dr. Dhundy Bastola, for their valuable time and suggestions. It would be impossible for me to complete my doctoral dissertation without them. I also want to thank Dr. Robert Boissy who served on my committee at the earlier period of my study.

I have been closely working with three people during my Ph.D. period. First, I would like to thank Dr. McKeithan who put extreme effort into the follicular lymphoma project and brought many good suggestions for my thesis. He is a great scientist and, more importantly, a great person. Our previous graduate student Dr. Himabindu Ramachandrareddy complimented him, saying “You know more than Google”. I could not agree more with that statement. Dr. McKeithan always brings up tough critical questions, hurtful to hear but so true that one can’t even argue. I am always surprised by how knowledgeable he is, how many details he has paid

attention to, and the beauty of the figures he designs. I am grateful that I have an opportunity to work with Dr. McKeithan, learn from him, and be friends with him.

The second person I would like to thank is Dr. Alyssa Bouska. Alyssa is a senior research associate in our lab, with whom I intensely interact daily. Alyssa contributed wet works in my thesis. She is trained as a biologist who is obviously good at performing experiments, but she has started picking up some analysis which is very impressive.

And the third person is Dr. Qiang Gong. Qiang is a bioinformaticist who introduced many analysis methods to me especially at the very beginning of my research. It is very nice of him to share his experience at such an early stage so that I can avoid a lot of unnecessary mistakes. Qiang and I have closely worked together on multiple projects including the follicular lymphoma project. He brought up many good points and suggestions about my thesis. I am pleased to discuss and solve problems with him.

Thanks to Dr. Chan, we have excellent research environment in the lab. I would like to thank all my lab members, Dr. Javeed Iqbal, Chao Wang, Dr. Can Kucuk, Bei Jiang, Joseph Rohr, Zhongfeng Liu, Cynthia Lachel, and Xiaozhou Hu. I also want to especially thank Wanqin Xie and Lin Huang from Zhixin Zhang's lab, and Chao Wang from our lab. They helped me survived my first Ph.D. year when I was taking the four toughest graduate biological courses at UNMC with a very limited biology background.

There is a saying in China, even the best cook needs high good ingredients. In my case, I have the University of Nebraska Medical Center, which has offered the best facilities, services, and opportunities as well as a group of people from their Graduate Program who are willing to help. I have the Department of Pathology and Microbiology which provide precious samples and unlimited support. I have the Holland Computing Center, its amazing employees.

Lastly, but of special importance, I would like to thank my parents, Pei Zhang and Aiyin Wang, who never had a chance to go to high school but were so determined to believe that the United States is the right place for my education that they provided all they could afford to support me. I still don't understand what encouraged them to make such a big decision but I would have achieved nothing without their unconditional love. I would also like to thank my husband, Daniel Bailey, who happens to be a great programmer, loves mathematics, and provides me my audience. Daniel has enlighten me with many great suggestions when I shared my analysis ideas with him, comforted and supported me when I was stressed, and showed his love and encouragement when it came to important decisions. People say the secret of marriage is communication. I like to joke about it by saying "communicate by what? Scripts or equations?".

All in all, I am a lucky person. I love what I am doing, and enjoyed every bit of my Ph.D. life; the good, the bad, happiness, and frustration.

THE ROLE OF GENETIC ALTERATIONS IN TUMOR INITIATION, PROGRESSION AND TRANSFORMATION

Weiwei Zhang, Ph.D.

University of Nebraska Medical Center, 2015

Supervisor: Wing C. Chan, M.D.

Follicular lymphoma (FL) is the second most common lymphoma in the United States. Although it is generally an indolent lymphoma, FL is not curable, and, in about 30% of patients, the FL undergoes transformation into an aggressive lymphoma (tFL) with marked worsening of prognosis. To identify mutations preferentially present in tFL, we performed whole exome sequencing (WES) on paired FL and tFL arising in the same patients and developed a mutational analysis pipeline. After we identified potentially important genes that have been found to be mutated in our paired FL and tFL study, we constructed a custom capture platform including these genes as well as other genes known to be mutated in B-cell lymphomas. We were able to use this focused sequencing platform to analyze additional samples at greater sequencing depth. Clonal architecture and evolution can be readily identified; however, the DNA samples were fragmented using restriction enzymes, which compromised duplicate analysis. We developed a new approach with a statistical model to solve the problems. Samples from uninvolved tissue of the same patients are commonly used to distinguish germline variants from somatic mutations; however, the germline DNA was often not available for our samples. , We designed a filtering based method to limit the number of germline variants that would be mistakenly called somatic mutations and validated this approach using a dataset with paired normal samples. We also introduced a novel idea based on machine learning to predict somatic mutations from paired FL and tFL samples without healthy tissue. Five machine learning algorithms were tested in

datasets with known somatic mutations, and their performance was evaluated by statistical measures. The results indicated somatic mutations can be reliably predicted. In order to provide complementary information, we integrated our mutation data with copy number abnormality data and found genes more frequently mutated in tFL cases. The recurrently mutated genes are often involved in epigenetic regulation, the JAK-STAT or the NF-κB pathway, immune surveillance, and cell cycle regulation, or are transcription factors involved in B cell development. As no entirely tFL specific mutations are found, the transformation event needs to cooperate with pre-existing alterations and future studies will focus on identifying cooperative mutations for FL transformation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABC: Activated B-cell like

aCGH: Array comparative genomic hybridization

APCs: antigen presenting cells

AUC: Area under curve

GC: Germinal center

BLL: Burkitt-like Lymphoma

BWA: Burrows-Wheeler Aligner

CBS: Circular binary segmentation

CGH: Comparative genomic hybridization

CHIP-seq: Chromatin immunoprecipitation sequencing

CN: Copy number

CNA: Copy number abnormality

CNV: Copy Number Variation

CSMs: Candidate somatic mutations

D: Diversity

dbSNP: The single nucleotide polymorphism database

DLBCL: Diffuse Large B cell Lympphoma

FFPE: Formalin-fixed paraffin-embedded

FISH: Fluorescent in situ hybridization

FL: Follicular Lymphoma

GATK: Genome Analysis Toolkit

GCB: Germinal center B-cell like

GEP: Gene expression profiling

HL: Hodgkin Lymphoma

Ig: Immunoglobulin

IgD: Immunoglobulin D

IgH: Immunoglobulin heavy chain

IgL: immunoglobulin light chain

IgM: Immunoglobulin M

J: Joining

KM: Kaplan-Meier

LLMPP: Lymphoma/Leukemia Molecular Profiling Project

MeDIP-seq: Methylated DNA immunoprecipitation sequencing

ncRNAs: Non-coding RNAs

MCR: Minimal common region

NHL: non-Hodgkin lymphoma

NNET: Neural networks

OS: Overall survival

Post GC: Postgerminal center

PRC2: Polycomb repressive complex 2

PTCL: Peripheral T cell lymphoma

RAG1: Recombination-activating gene 1

RAG2: Recombination-activating gene 2

rCNAs: Recurring Copy Number Variation

REAL: Revised European-American Lymphoma

RF: Random forest

ROC: Receiver operating characteristic

RP: Recursive partitioning

RRBS: Reduced representation bisulfite sequencing

RSSs: Recombination signal sequences

SAS: Statistical analysis system

SNP: Single nucleotide polymorphism

SVM: support vector machine

TdT: Terminal ldeoxynucleotidyl transferase

TET2: Tet methylcytosine dioxygenase 2

TF: Transcription factor

tFL: Transformed Follicular Lymphoma

TR: Tree

UNMC: University of Nebraska Medical Center

UTR: Untranslated regions

V: Variable

VAF: Variant-allele frequency

WES: Whole Exome Sequencing

WGBS: Whole genome bisulfite sequencing

WGS: Whole Genome Sequencing

WTSS: Whole transcriptome shotgunsequencing

CHAPTER I

INTRODUCTION

A. Lymphoma biology

Lymphomas are tumors derived from lymphatic cells. Lymphomas have many subtypes, with the two main categories being Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL). NHL contains approximately 90% of lymphomas, and the majority of NHL is derived from B-lymphocytes at various stages of B cell differentiation (Table 1-1). A small percentage is derived from T-lymphocytes, and rarely from natural killer (NK) cells and dendritic cells[1]. The tumor behavior varies significantly, from indolent, e.g. Follicular Lymphoma (FL), to aggressive, e.g. Diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL). According to Revised European-American Lymphoma (REAL), DLBCL (31%) and FL (22%) are the two most common lymphoid neoplasms in the United States and Western Europe[2].

Mature B-cell neoplasms
Chronic lymphocytic leukemia/small lymphocytic lymphoma
B-cell prolymphocytic leukemia
Splenic marginal zone lymphoma
Hairy cell leukemia
Splenic lymphoma/leukemia, unclassifiable
Lymphoplasmacytic lymphoma
Heavy chain diseases
Plasma cell myeloma
Solitary plasmacytoma of bone
Extraosseous plasmacytoma
Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue
Nodal marginal zone lymphoma
Follicular lymphoma
Primary cutaneous follicle centre lymphoma
Mantle cell lymphoma
Diffuse large B-cell lymphoma, NOS
Diffuse large B-cell lymphoma associated with chronic inflammation
Lymphomatoid granulomatosis
Primary mediastinal (thymic) large B-cell lymphoma
Intravascular large B-cell lymphoma
ALK-positive large B-cell lymphoma
Plasmablastic lymphoma
Large B-cell lymphoma arising in HHV8-associated multicentric Castleman disease
Primary effusion lymphoma
Burkitt lymphoma
B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and Burkitt lymphoma
B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Hodgkin lymphoma

Table 1-1. WHO classification of mature B-cell neoplasms[1].

A.1. Structure of the lymphoid system

The lymphoid system plays a vital role in defense against pathogens from viruses to parasites. It is comprised of lymphoid tissue and recirculating lymphocytes.

Lymphoid tissue has two major forms, central (primary) lymphoid tissue and peripheral (secondary) lymphoid tissue (Figure 1-1). The bone marrow and thymus are the two organs of the central lymphoid system. They are the sites of B and T cell maturation respectively. The matured B and T cells express different antigen receptors and migrate into the peripheral lymphoid tissue to defend against pathogens invading the body. Cellular and humoral immune responses take place in the peripheral lymphoid tissue, which includes spleen, bone marrow, tonsil, mucosa-associated lymphoid tissues and lymph nodes. The spleen responds predominantly to blood-borne antigens. The bone marrow is both a central and a peripheral lymphoid organ because it gives rise to B and NK cells. The tonsil and mucosa-associated lymphoid tissues react to antigens entering via the surface mucosal barriers. Lymph nodes mount immune responses to antigens or antigen presenting cells (APCs) entering through the afferent lymph vessel. Lymph nodes are major sites of B, T and other immune cells and provide sites for interaction of the lymphocytes, APCs and other cells. Lymph nodes are composed of three major areas, the cortex, paracortex and medulla (Figure 1-2). The cortex contains primarily B cells, some of which aggregate in primary follicles. When an immune response occurs, the follicles develop a central area, with large proliferating cells, termed a germinal center (GC). The paracortex contains primarily T cells and many antigen presenting cells. The medulla contains both T and B cells and many plasma cells and scavenger phagocytic cells. Lymphocytes and antigens enter the cortex through the afferent lymphatic vessel and filter down through the paracortex and into medulla before leaving the lymph node via the efferent lymphatic vessel and moving on.

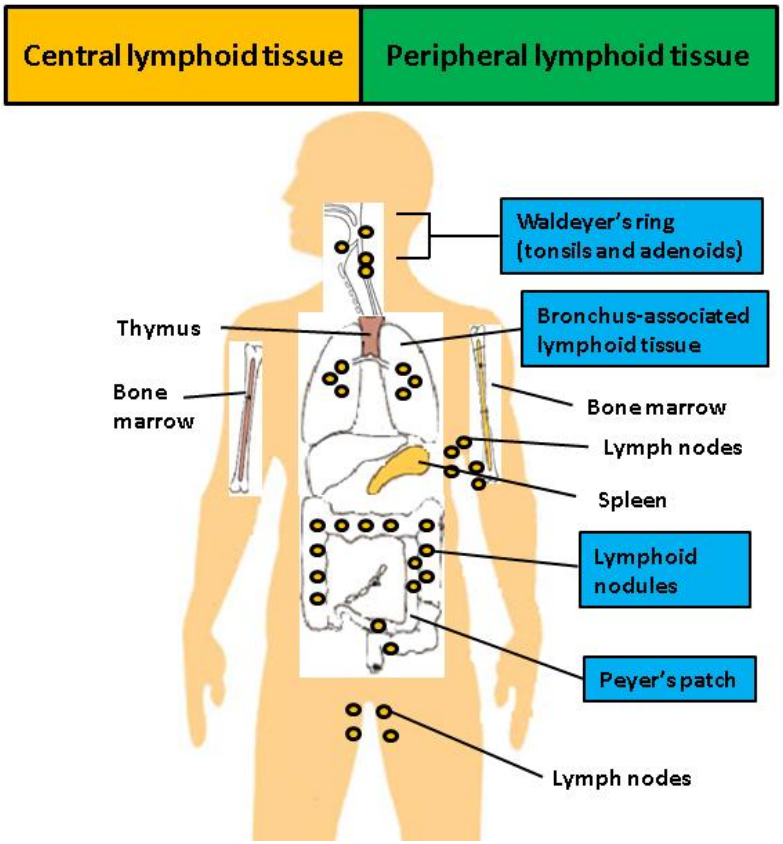

Figure 1-1. Major lymphoid organs and tissue.

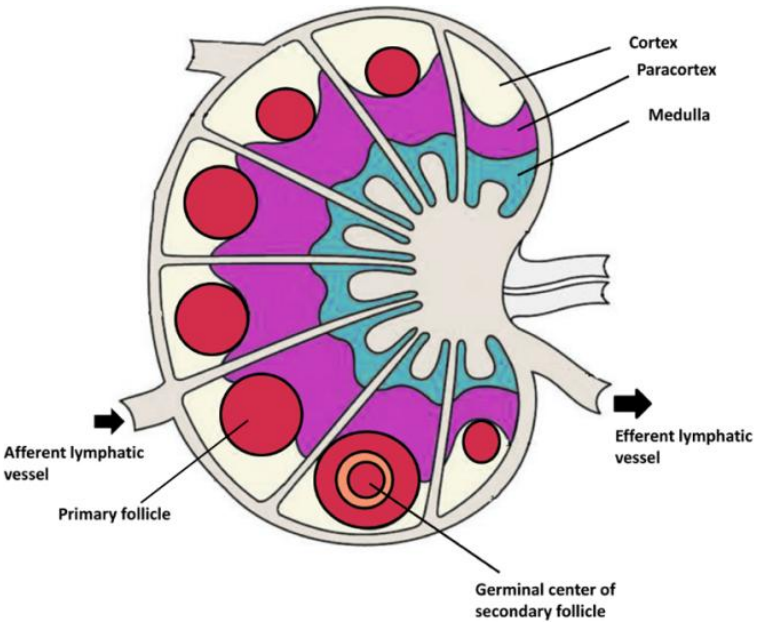

Figure 1-2. Schematic structure of the lymph node.

A.2. B cell differentiation

B cell differentiation in the central and peripheral lymphoid tissues involves epigenetic and gene expression changes. In general, B cell differentiation and migration occur through multiple stages, associated with changes in the immunoglobulin (Ig) gene loci, expression of cytokine and receptors and Ig protein (Table 1-2, Figure 1-3). A cluster of genes encoding the Ig heavy chain is located on chromosome 14. It includes a series of variable (V) genes, joining (J) genes, diversity (D) genes and C genes. Clusters of genes encoding the B cell receptor light chains are located on chromosome 2 and chromosome 22, respectively. Each cluster includes a series of V genes, J genes and one or more C genes, and only one of the light chains is used to construct the Ig molecule. DNA rearrangement generates a diverse array of antigen-specific molecules by shuffling, cutting and recombining individual V, D, and J segments into a heavy chain and similar recombination of individual V and J segments for the light chain immunoglobulin molecule. Two heavy chain and two light chain molecules transcribed from the rearranged IgH and IgL (Ig kappa and Ig lambda) of a cell will be assembled into a full immunoglobulin molecule to attack one particular antigen without attacking the body itself during the B cell differentiation[3]. The first stage of differentiation starts when the hematopoietic stem cells in the bone marrow receive signals from bone marrow stromal cells and begin B cell development. The second stage is the CD34+ progenitor B cells express some B cell characteristic markers and initiate IgH rearrangement. The Ig gene recombination is initiated by the recombination-activating gene 1 (RAG1)-recombination-activating gene 2 (RAG2) protein complex (Figure 1-4). At the third stage, the progenitor B cells differentiate into precursor B cells. The early precursor B cells involve the recombination of the D and J segments of the IgH. The late precursor B cells involve the rearrangement of the V segment to previous recombined DJ segments. This stage results in the complete rearrangement of the IgH gene. IgL rearrangement takes place only after the IgH gene

arrangement. It occurs in a similar manner to the rearrangement of IgH. In contrast to the IgH, the IgL does not possess D segment. Therefore, it only takes one step to form a V J segments. At the fourth stage, the precursor B cells differentiate into immature B cells. The light chain arrangement results in the expression of an IgM protein on the cell surface. At this stage, the immature B cells haven't encountered any antigen yet and they are also unable to initiate an immune response to foreign antigens. At the fifth stage, the immature B cells leave the bone marrow and migrate to peripheral lymphoid tissues. With the expression of IgM and IgD together, the immature B cells give rise to mature naïve B cells. The mature naïve B cells have the capability of responding to antigen. At the sixth stage, some of the mature naïve B cells encounter antigens and transform to extrafollicular B blasts, and later to short-lived plasma cells or primed B cells. Other mature naïve B cells form primary follicles. The primed B cells migrate into primary follicles with cognate T-cells and differentiate into centroblasts (CBs) with rapid proliferation. In this stage, somatic hypermutation occurs and provides additional variation that may improve the antibody responses to antigens as advantageous mutations lead to increasing affinity to antigen. The whole process is called affinity maturation. CBs may move to the light zone of the GC and become centrocytes (CC). CC with advantageous mutations survive and differentiate into long-lived plasma cells or memory B cells. At the seventh stage, the memory B cells with IgM protein migrate to marginal zone. CC can undergo further DNA rearrangement to change the class of the Ig through isotype switching. Long-lived plasma cells with IgG, IgA and IgE predominantly stay in the bone marrow. As a result of B cell development, naïve mature B cells, GC B cells, memory B cells and long-lived plasma cells are the four major forms of mature B cells. FL s are at the developmental stage of GC B cells.

	B-cells	Immunoglobulin Genes	Somatic Mutations	Ig Protein	Corresponding Lymphoma	
Foreign antigen independent	Stem cell	Germline	None	None	B-LBL/ALL	Bone marrow
	Pro-B-cell	Germline	None	None		
	Pre-B-cell	IgH-rearrangements μ chain (Cytoplasm)	None	Igμ		
	Immature B-cell	IgL/IgH-rearrangements IgM (Membrane)	None	IgM (membrane)		
Foreign antigen dependent	Mature native B-cell	IgH/L rearrangements IgM and IgD (Membrane)	None	IgM/IgD	B-CLL, MCL	Peripheral lymphoid tissue
	Germinal center (Centroblasts and Centrocytes)	IgH/L-rearrangements Class switch	Introduction of somatic mutations	Ig (minimal or absent)	BL, FL, LPHL, DLBCL	
	Memory B-cell	IgH/L rearrangements	Somatic mutations	IgM	MZL, B-CLL	
Terminal differentiation	Plasma cell	IgH/L rearrangements	Somatic mutations	IgG>IgA>IgD	Plasmacytoma/myeloma	

Table 1-2. B cell development and the corresponding lymphoma[2].

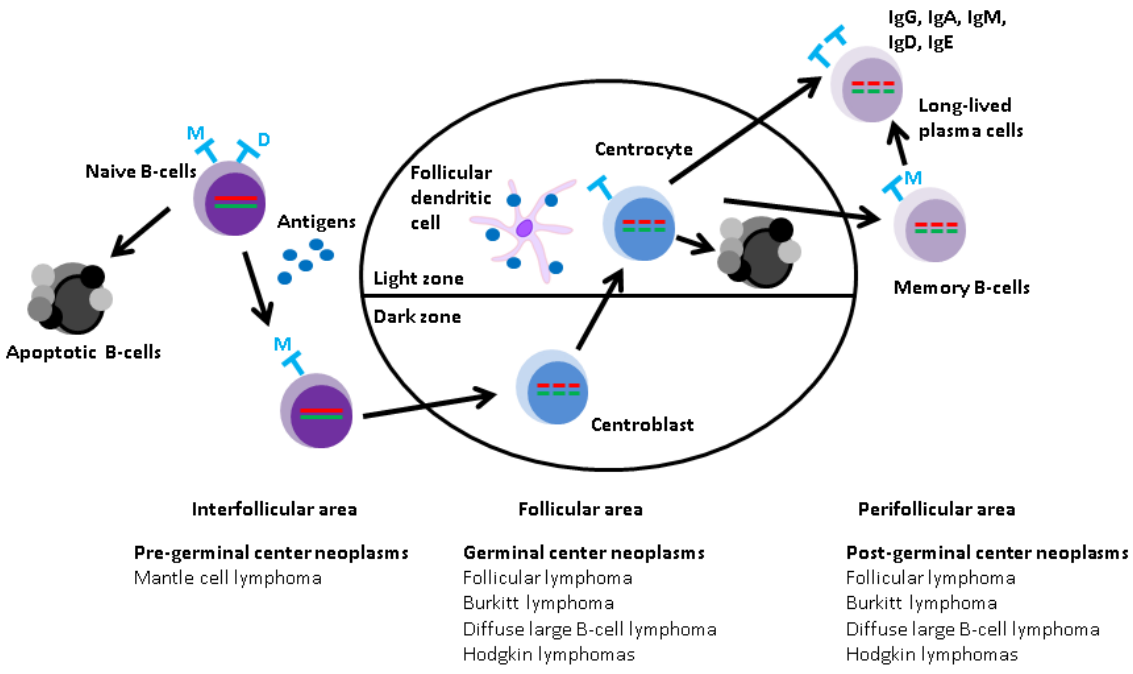

Figure 1-3. Events in B cell development[2].

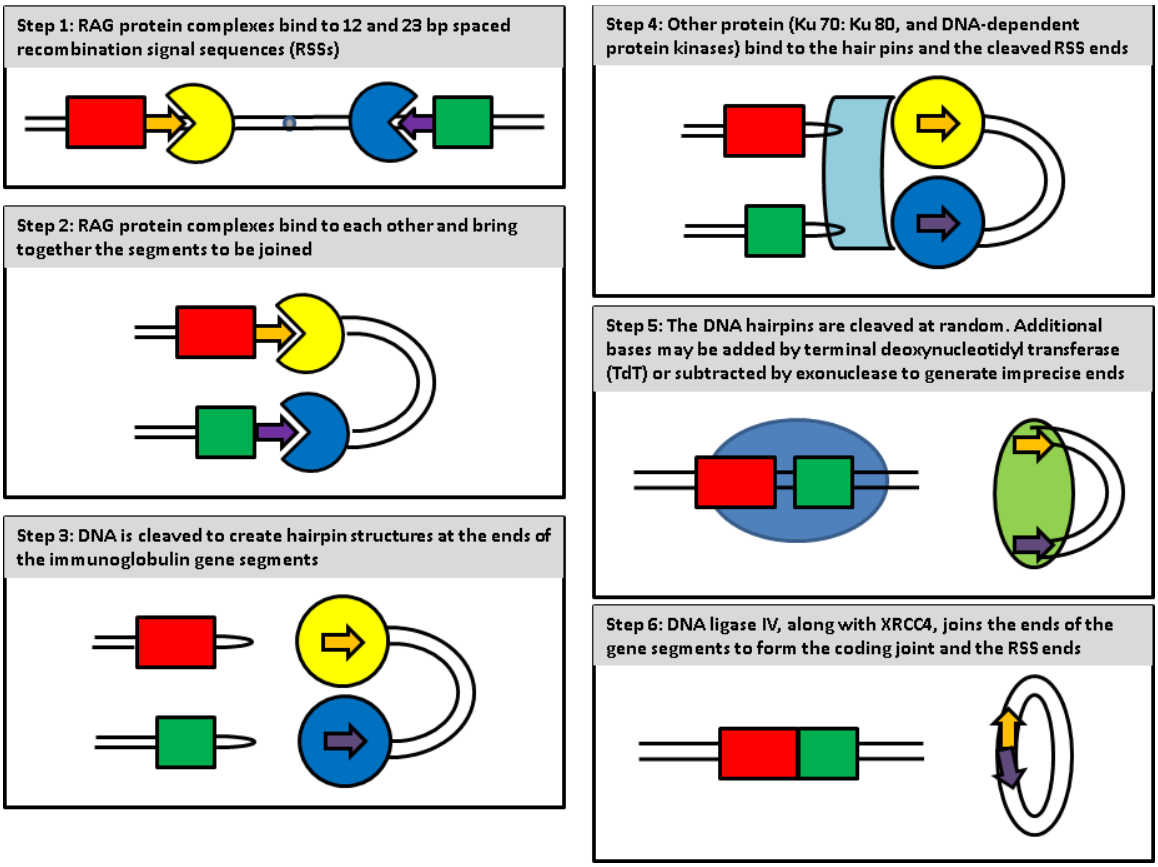

Figure 1-4. RAG protein involvement in the rearrangement of immunoglobulin gene segments[4]. RAG protein complexes are illustrated as light purple and light orange colored domes even though all RAG protein complexes are identical. This complex breaks double-stranded DNA between the B cell receptor coding segments and recombination signal sequences during the process of VDJ recombination. A ubiquitously expressed set of non-homologous end joining proteins repairs and joins the cleaved DNA ends afterwards.

A.3. FL

FL, the second most common histologic subtype of NHL in North America, arises from the proliferation of malignant germinal center B cells (Table 1-2). Despite the fact that FL is generally slow-growing, it is still not curable and it becomes less sensitive to chemotherapy when relapse happens. Up to 90% of FL cases have a translocation between chromosomes 14 and 18 leading to the overexpression of BCL2 in germinal center B cells (Figure 1-4). The BCL2 gene is normally found on chromosome 18q21, and the translocation moves the gene near to the immunoglobulin heavy chain enhancer element on chromosome 14. t(14;18) is the classical

cytogenetic aberration in FL, but there are other translocations of chromosome 18, such as t(2;18) or t(18;22), which juxtapose BCL2 to the loci of the IgL chain (κ or λ), which also result in BCL2 overexpression[5,6]. However, the translocation alone is insufficient for the establishment of FL but it provides a survival advantage for the cell in the germinal center microenvironment, where the cell can accumulate additional abnormalities. Clearly, these secondary events are necessary to contribute to disease progression.

A.4. Transformed follicular lymphoma (tFL)

FL can transform into an aggressive lymphoma with poor prognosis. There are two main types of tFL, DLBCL and Burkitt-like lymphoma (BLL). DLBCL is the most frequent tFL; gene expression profiling studies have demonstrated that DLBCL consists of two major distinct molecular subtypes, germinal center B-cell like (GCB) DLBCL and activated B-cell like (ABC) DLBCL. GCB DLBCL arises from normal germinal center B cells whereas ABC DLBCL arises from postgerminal center (Post GC) B cells and has significant lower overall survival rate than GCB DLBCL. BLL is currently known as unclassifiable B cell lymphoma with features that are between DLBCL and BL.

In most cases, tFL emerges from the FL clone by acquiring additional abnormalities that give it a growth advantage and make it more aggressive and able to grow without the GC microenvironment. In a few instances, however, the DLBCL appears to arise from the emergence of an unrelated second lymphoma. It is very important to identify the transformation because it represents a change in the biology of the disease and in the patient's clinical course. After FL transformation, the patient has a more rapidly progressing disease and short survival, commonly less than 2 years. Occasionally, tFL is found at the presentation of the lymphoma and this type of

tFL may be biologically distinct from the usual type that arise later in the course of a FL and has better prognosis[7].

B. Genetic abnormalities

B.1. Genetic abnormalities that contribute to the development of FL

The t(14;18) is considered as the first hit in FL development (Figure 1-5). It causes dysregulation of tumor cell apoptosis, but it is not sufficient to result in clinical disease. Therefore lymphomagenesis requires a number of additional genetic and cytogenetic events. These later events change the biological and clinical behavior of the clone, and finally generating the FL. Aberrant BCL2 overexpression allows the survival of the abnormal GCB cells in the GC microenvironment where continuous somatic hypermutation or recombination activity in class switch increase the instability of the genome and probability of second hits that promote the formation of FL. A number of secondary chromosomal alterations have been reported, for example, the most common alterations are the partial trisomies of chromosomes 1q, 7, 8 and 18q, and 1p and 6q deletion. Deletion in 6q usually follows by the deletion of 1q[5]. Some genetic abnormalities are associated with late disease or transformation[8,9]. In rare case, the histologic progression of FL involves *MYC* rearrangements[10].

BCL6 translocation is also common in FL and can occur in cases with or without the classical BCL2 rearrangement. Constitutive overexpression of BCL6 may be an important mechanism is the pathogenesis of FL[11,12].

With next generation sequencing (NGS), many mutations have also been recently identified in multiple studies (Figure 1-5), for example, *KMT2D (MLL2*), *EZH2*, and *CREBBP* are often mutated in early FL[8,13-16].

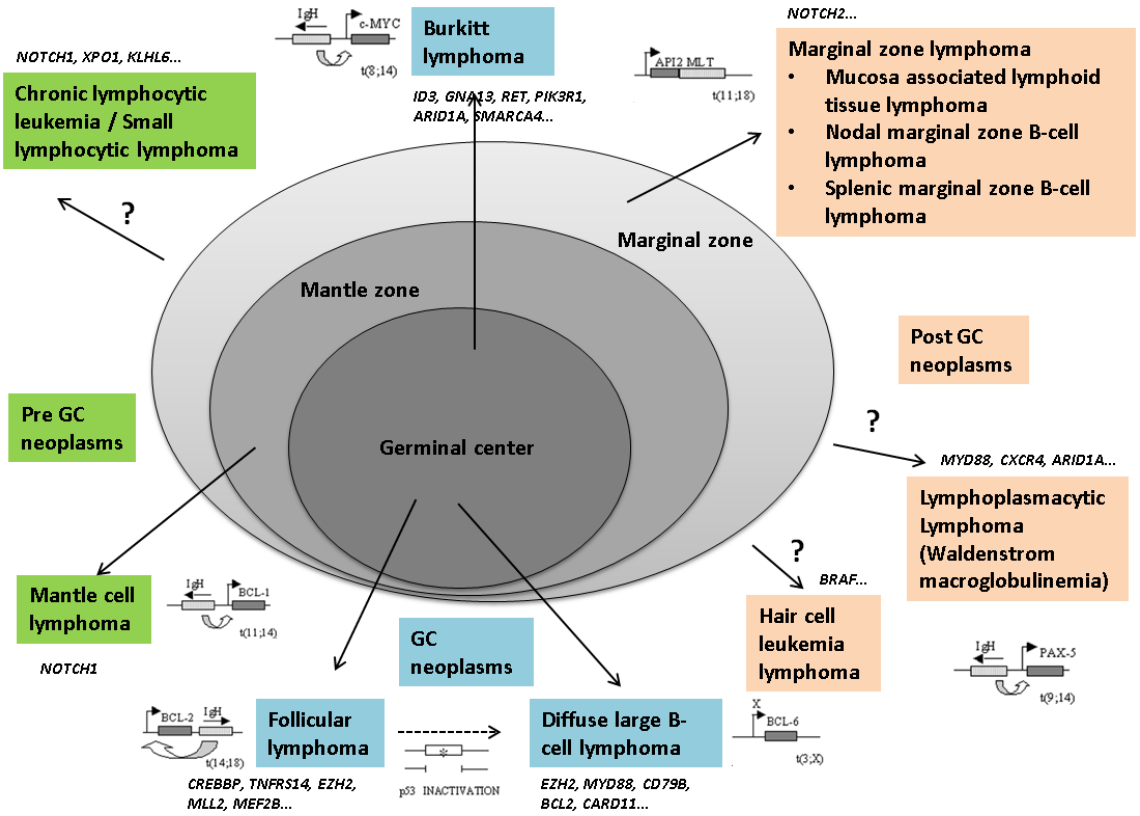

Figure 1-5. Model of B cell NHL histogenesis and pathogenesis[2].

B.2. Genetic and cytogenetic abnormalities that drive transformation

DLBCL is considered de novo if it develops in the absence of a precursor malignancy such as FL. In contrast, when a person is diagnosed with DLBCL but he previously or concurrently had FL diagnosis, then it may represent a tFL. If the DLBCL and the FL share a common precursor, they should share many of the copy number abnormalities (CNAs) and mutations, thus supporting the idea that they are clonally related and not independent tumors. FL is usually indolent, but 40% of patients will develop tFL with poor prognosis. There must be secondary genetic and cytogenetic abnormalities that drive the transition from an indolent to an aggressive stage. Identification of these abnormalities that drive transformation is critical for a better understanding of the progression and evolution of tFL.

To date, studies of the genetic abnormalities in de novo DLBCL have revealed a large number of findings, for example: chromosome alterations affect 3q27 recurrently in DLBCL. This region includes BCL6, a transcriptional repressor that belongs to the family of transcription factors containing zinc-fingers and is required for GC formation and B cell immune response. 25% of GCB DLBCL cases have chromosomal rearrangements of *BCL2*. It is possible that some such cases may actually represent transformation of a clinically inapparent FL.

Current concepts describe the transformation as the result of heterogeneous genetic and cytogenetic abnormalities. The genetic abnormalities in transformation have been represented in four models[8]. The four models are grouped into 2 categories, linear evolution from FL and divergent evolution from a preclinical progenitor. In fact, transformation is all linear. The divergent model is a sampling artifact because the tumor sampled does not contain the direct precursor of the transformed tumor. However, the precursor clone for the transformed tumor may be a minor subclone at the time of sampling and it is very informative to identify these subclones in the study of clonal evolution in transformation.

Multiple studies suggest several discrete mechanisms are involved in driving transformation from FL. These mechanisms include *TP53* mutations[9] and deletions, inactivation of p16 and dysregulation of *MYC*[17].

DLBCL molecular subtypes have been reported to have different oncogenic pathways. SPIB gain or amplification, deletion of CDKN2A tumor suppressor locus and trisomy 3 happen more commonly in ABC DLBCL whereas amplification of mir-17-92 and loss of the tumor suppressor PTEN are only recurrent in GCB DLBCL[18]. Similarly, mutations affecting the NF-kB pathways are much more prevalent in the ABC subtype. It would be highly interesting to compare the copy number variation (CNV) and mutation profiles of tFL with these DLCBL subtypes.

C. CNV analysis

CNV refers to gain or loss of chromosomal regions in the genome. CNV regions vary from kilobase range change to gain or loss of entire chromosomes. Most CNV is stable and heritable. CNV can cause disease, affect gene expression level and change phenotype. CNV analysis can be used to detect chromosomal abnormalities that contribute to lymphoma.

C.1. CNV general technology

Techniques to identify cytogenetic aberrations have changed drastically in the past decade. Fluorescent in situ hybridization (FISH), comparative genomic hybridization (CGH) and array comparative genomic hybridization (aCGH) are three technologies that have been widely used in molecular cytogenetics. FISH is based on the capability of a fluorescent–labeled single stranded DNA to hybridize to its complementary DNA sequence[19]. FISH is often used to identify chromosomal rearrangement including translocations, inversions, etc. but CNVs can also be detected. CGH is designed for CNV detection. It can efficiently compare two genomic DNA samples by competitive hybridization to normal metaphases[19]. Normal genomic DNA serves as the standard for comparison of the test DNA. Compared to standard FISH, which can only test one or several genes at a time, CGH is able to detect gains or losses genome wide. aCGH utilizes the same principles as traditional CGH but has tremendously higher resolution. Hybridizing is performed on a high density DNA array format and able to detect small CNVs in the genome[20].

C.2. Single-nucleotide polymorphism array analysis procedure

In our previous study, DNA from FL and tFL tumors was hybridized to high-resolution GeneChip Human Mapping 250K Nsp single-nucleotide polymorphism (SNP) arrays. In contrast to aCGH, which uses competitive hybridization of fragmented tumor DNA and control DNA labeled with different fluorophores to a microarray platform to detect CNAs, SNP array contains

oligonucleotide probes that interrogate both copy number (CN) and SNP sites. Therefore, SNP arrays are able to detect both DNA CN and SNP-based genotypes at submegabase resolution, including loss of heterozygosity (LOH), and uniparental disomy[21].

A circular binary segmentation algorithm[22] was used to analyze the resulting data. This algorithm segmented chromosomes into similar log2 ratios and connected the change point to the locations of regions with aberrant DNA copy numbers. Therefore, it identified regions of CN gain or loss. Minimal common regions (MCRs) were determined for recurrent CNAs (rCNAs) in all samples.

D. NGS analysis

NGS, also known as high throughput sequencing or massively parallel sequencing, has dramatically changed the way scientists extract genetic information from biological systems (Table 1-3). These technologies help us to develop great insight into the abnormalities affecting the genome, transcriptome and epigenome through DNA sequencing of genomic DNA, cDNA and bisulfite treated DNA with cheaper cost. NGS allows us to discover point mutations as well as structural alterations such as indels, inversion and translocations that contribute to diseases. Sequencing of the transcriptome provides us the capability to identify changes of gene expression, alternative splicing, gene fusions, mutations and non-coding RNA species. Sequencing after Bisulfite treatment can be used to determine the global cytosine methylation status of DNA.

DNA Level	RNA Level	Epigenetic Level
Whole genome sequencing (WGS) • Discover the genetic variations in a genome-wide range	Whole transcriptome shotgun sequencing (WTSS) ▪ Exam differential gene expression ▪ Discover novel genes	Whole genome bisulfite sequencing (WGBS) ▪ Whole genome-wide level ▪ High accuracy and resolution
Whole exome sequencing (WES) ▪ Discover the causative, susceptibility loci ▪ More efficient	Small RNA sequencing ▪ Exam miRNA gene expression ▪ Gene regulatory networks and miRNA target study	Methylated DNA immunoprecipitation sequencing (MeDIP -seq) ▪ Based on immunoprecipitation for methylated DNA enrichment ▪ Whole genome-wide and cost effective
Target Region sequencing ▪ Discover novel variants or validate the candidate variants in the target region	Non-coding RNA sequencing ▪ Identify novel non-coding RNA ▪ Discover disease-specific biomarkers	Reduced representation bisulfite sequencing (RRBS) ▪ Promoter regions with substantial genome coverage ▪ Based on enzyme digestion and bisulfite treatment
Single cell sequencing ▪ Genetic variation research at single cell level ▪ Explore cancer cells evolution during tumor progression	Cell line sequencing ▪ Obtain a clear and comprehensive genetic patters ▪ Obtain mutation information of high accuracy	Chromatin immunoprecipitation sequencing (ChIP-seq) ▪ Genome-wide protein-DNA interaction studies ▪ High resolution

Table 1-3. Multilevel high throughput NGS[23].

D.1. NGS: general technology

There are multiple platforms for performing NGS and the Illumina platform is one of the most commonly used technologies. It has been reported to have high sensitivity and reliability in DNA mutation detection. It is a sequencing method based on engineered polymerases and reversible terminator bases. Under the catalysis of the polymerase, the DNA templates are copied base by base using the deoxyribonucleotide triphosphates which are fluorescently-labeled and reversibly terminated. During each cycle, the fluorescence signal is captured by a build-in camera at the point of incorporation and the nucleotides are identified. After that, the fluorescence label and the blocking group are removed allowing the addition of the next bases. The critical difference NGS extends this process across millions of DNA fragments in a massively parallel fashion instead of sequencing a single DNA fragment.

D.2 Whole exome sequencing (WES) analysis procedure

WGS sequences all DNA sequences in an organism's genome, and has more uniform quality for the identification of variants as well as large insertions and deletions and other structural alterations. On the other hand, WES is an efficient strategy to selectively sequence the exome. The exome include all regions in genes that are translated into protein including also splice junctions and some regions important in regulating transcription. Since the exome constitutes only about 1% of the human genome, WES is more cost efficient and less computationally intensive than WGS. Sequencing depth can generally be higher.

E. Custom capture panel

To obtain mutations preferentially present in FL and tFL, we performed WES on paired FL and tFL arising in the same patients and developed a mutational analysis pipeline. Even if the cost of sequencing decreases a lot, it is still expensive to sequence a large number of samples. It is also time-consuming to analyze the huge amount of data, especially since the mutated genes that contribute to the disease are limited in number. Therefore, once we identified potentially important genes that are mutated from WES studies, a focused sequencing platform to analyze genes of interest can be constructed. This platform can analyze a large number of samples at greater sequencing depth and less cost to identify transformation associated mutations and their possible cooperation in the transformation process.

F. Application of genetic abnormality analysis in cancer research

Gains and losses of CN can range in size from thousands to millions of bases. For large CNAs, it can be difficult to determine which genes contribute to cancer. It is possible that multiple genes in these regions are contributing. Examining GEP and mutations may help to identify the target genes in these regions.

NGS provides a great opportunity to exam the genetic mutations at a genome-wide level. We are able to go through the mutations in coding regions and splice sites, and have a view of the mutational landscape of FL and tFL, and figure out the alterations that cause transformation. We incorporate CNV data into mutation studies so that the combination of the two approaches can provide a comprehensive view of genetic abnormality. The integrated data help us to identify the abnormalities associated with transformation of FL and characterize the progression of FL.

G. Overview of this dissertation

The motivation of the study is to identify genetic changes that would provide important insight into the mechanisms of FL tumorigenesis and transformation; however, there are a number of obstacles. One obstacle is the shortage of matching normal samples in FL research. The normal sample is currently the only way to confidently filter out all germline variants but it is difficult for researchers to collect normal samples in the FL field. Another difficulty is the limited number of cryopreserved or fresh tumor samples available for analysis. There are also computational challenges to overcome such as removal of passenger mutations and duplicate analysis.

In this study, we applied a variety of statistical methods to improve the performance of the analysis. We introduced a binomial distribution based statistical model to estimate duplicate ratio in custom capture panel that uses restriction enzymes to capture the genes of interest. Instead of applying typical log-rank in survival analysis, advanced test statistics with different weights were provided based on abnormalities that occur in the early or late stage of FL development. We developed a variety of mutation detection pipelines for different NGS platforms and sample combinations. We applied a two-step filtering method to identify somatic

mutations based on the biological features of FL to provide potential driver mutations that contribute to the disease. We also proposed a novel computational model using the machine learning concept to distinguish the germline variants from somatic mutations without corresponding normal samples in sequencing analysis.

The rest of this dissertation is organized as follows. In chapter 2, we introduce the pipelines on which the methods were built and evaluated. We also describe the samples to which our methods were applied and the datasets in which our methods were validated. In chapter 3, we report the initial results and validation results of the methods applied in the analysis. In chapter 4, we associate the CN abnormalities with our mutation data, integrate our mutation data with other two datasets, combine the studies with other biological analysis, and generate more comprehensive understanding of FL initiation and progression. In chapter 5, we describe the future study in identifying epigenetic alterations, contributions of mutations in regulatory region and non-coding regions, functional study of the individual mutated genes in FL and transformation, extending our current pipelines to FFPE samples, and improving features applied in machine learning models. In chapter 6, we summarize how the pipelines and approaches designed in this thesis significantly improve the data analysis, provide reliable results, and detect novel abnormalities in FL and tFL research.

CHAPTER II

METHOD

A. WES and data analysis

A.1. WES dataset

We obtained 12 paired FL and tFL frozen tissue samples from the same patients before and after transformation. The samples were provided by the University of Nebraska Medical Center (UNMC), Lymphoma/Leukemia Molecular Profiling Project (LLMPP), or Aarhus University Hospital. This study was approved by the UNMC institutional review board.

A.1.1. Sample and patient materials

The time between the FL diagnosis and transformation varied from 1 year to 9 years. The tFL samples were all diagnosed as DLBCL by a panel of LLMPP hematopathologists. Our previous gene expression profiling (GEP) analysis had classified the tFL samples as activated B-cell (ABC)-like, unclassifiable (UC), or germinal center B-cell (GCB)-like tFL. The clonal relationship between the biopsies was confirmed by comparing the genetic profile based on clinical data, SNP array and/or sequencing to confirm that there was close similarity between pairs. When BCL2 status was unknown, BCL2 rearrangement was assessed by PCR to determine if the frequency of t(14;18) positivity expected for an FL dataset was observed in the tFL dataset. We did PCR with primer that prime in 3 places where the break point often occurs, major breakpoint region, minor cluster region and intermediate cluster region (Table 2-1).

Case ID	FL biopsy	tFL biopsy	Exome Sequencing	Paired Case	FL biopsy date	tFL biopsy date	t(14;18) status (tFL)	subtype
Case 1	x	x	x	x	9/28/2004	1/20/2006	pos (MBR)	GCB-like
Case 2	x	x	x	x	12/10/2003	9/24/2007	pos (MBR)	GCB-like
Case 3	x	x	x	x	4/6/1993	4/29/1994	unknown (MCR,MBR,ICR neg)	ABC-like
Case 4	x	x	x	x	1/20/1988	5/24/1991	unknown (MCR,MBR,ICR neg)	GCB-like
Case 5	x	x	x	x	4/28/1989	2/14/1990	pos (MCR)	GCB-like
Case 6	x	x	x	x	1/28/2000	11/12/1997	positive (clinical data)	GCB-like
Case 8	x	x	x	x	6/6/2002	11/1/2004	pos (MCR)	ABC-like
Case 9	x	x	x	x	5/23/1995	10/27/2004	unknown (MCR,MBR,ICR neg)	ABC-like
Case 10	x	x	x	x	1/20/1996	4/28/1999	pos (MCR)	ABC-like
Case 11	x	x	x	x	1997	1998	unknown (MCR,MBR,ICR neg)	ABC-like
Case 12	x	x	x	x	1992	1993	positive (der18 aCGH)	GCB-like
Case 22	x	x	x	x	9/12/1994	10/13/1995	pos (MBR)	UC

Table 2-1. Sample and patient information for WES.

A.1.2. Library preparation

Library preparation was performed according to the manufacturer's protocol using Illumina's TruSeq DNA sample prep kits. Exome capture was performed according to the manufacturer's protocol using either Illumina's TruSeq exome enrichment kit (11 paired samples) or Agilent's SureSelect exome enrichment kit (1 paired sample). 1 µg of DNA per sample was used for capture, and the equivalent of 1 to 3 samples was sequenced per lane on an Illumina HiSeq2000 or HiSeq2500 sequencer.

To prepare the samples using the Illumina's TruSeq exome enrichment kit, DNA was sheared into 100-300 base-pair long fragments using a Covaris sonicator. After fragmentation, ends were repaired, A-overhangs were added at the 3'-end of the DNA fragment, and adaptors were ligated to both ends of the DNA fragments. These DNA fragments were then denatured into single-stranded DNA and hybridized to biotin-labeled DNA probes (Truseq) or biotin-labeled RNA

probes (SureSelect) specific for the targeted regions. After enrichment using streptavidin beads, the enriched DNA fragments were eluted from the solution. After amplification, the DNA molecules were ready for cluster generation and subsequent sequencing.

A.2. Pipeline for calling variants in WES

As shown in Figure 2-1, the pipeline is as follows: After sequencing, the raw read quality was first evaluated by FastQC (v0.10.1), and then mapped to human reference genome hg19 using the Burrows-Wheeler Aligner (BWA) (0.7.5a-r405). Genome Analysis Toolkit (GATK) (v3.1-1-g07a4bf8) was used for local realignment and base quality recalibration. Duplicate marking was done with Picard (v1.115). Variant calling was performed with VarScan2 (v2.3.6). Basic filtering, including base quality, total read depth, read depth supporting the reference sequence, read depth supporting the variant sequence, forward read depth supporting variant and reverse read depth supporting variant, was applied to filter out unreliable variants. The variants were then annotated using ANNOVAR[24]. We filtered out known germline variants (single-nucleotide polymorphisms [SNPs] and indel polymorphisms) based on dbSNP138 and only selected mutations that were nonsynonymous, stop gain, stop loss, insertion or deletion in the coding region or splice sites. Since we have paired FL and tFL samples, we categorized our mutations into three types: FL-unique mutations indicated mutations only detected in the FL sample not the paired tFL sample; tFL-unique mutations indicated mutations only detected in the tFL sample not the paired FL sample; and shared mutations indicated mutations detected in both paired FL and tFL samples.

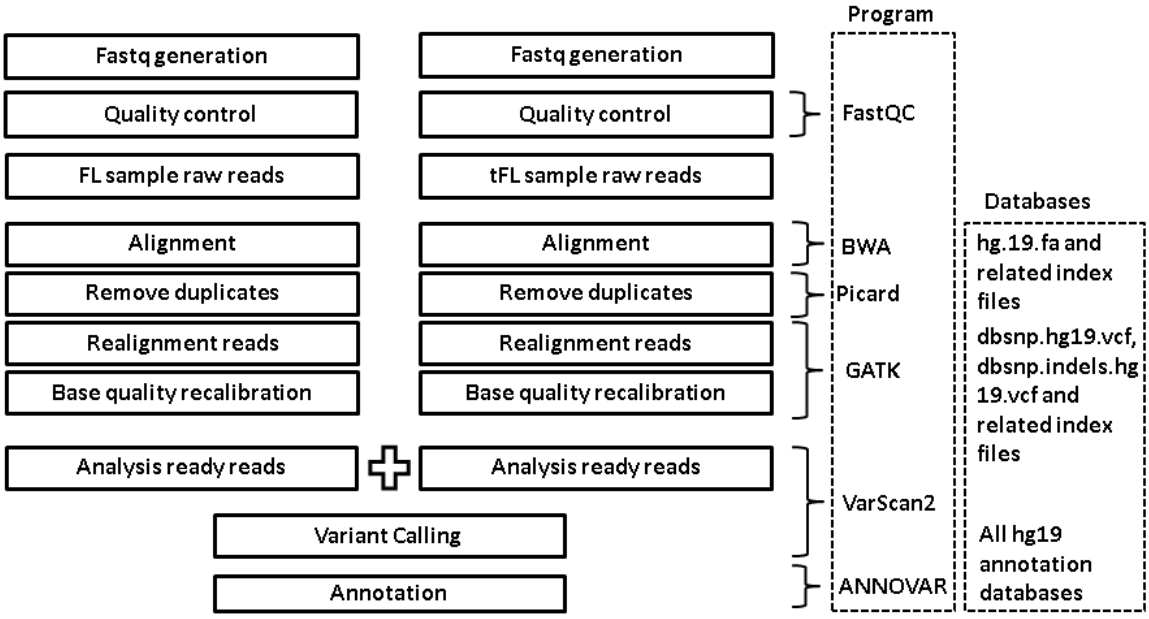

Figure 2-1. Pipeline designed for WES with paired sample.

A.3. WES performance analysis

In order to evaluate the performance of WES, we applied sequence performance analysis (Figure 2-2) to check the total number of reads used for the alignment, the number of reads that were successfully mapped to the reference, the number of aligned reads that were mated with their paired reads, the fraction of duplicates in the sequenced reads, the regions in the reference that were aligned by the reads, the depth of each position in the aligned region, and the coverage and depth of coding regions and splice sites.

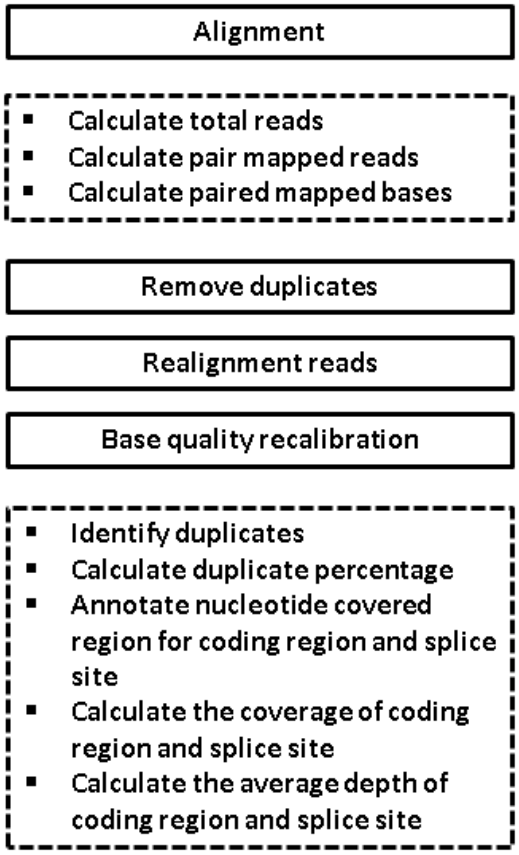

Figure 2-2. Pipeline designed for WES performance evaluation.

A.4. Pipeline validation

Sanger sequencing is an accurate analysis method for variants validation when the variant frequency presents at 20% or greater in the samples. We applied Sanger sequencing to validate 50 variants to confirm the reliability of the WES pipeline.

B. Custom capture panel sequencing and data analysis

B.1. Custom capture panel dataset

Samples included in the custom capture panel pipeline were all frozen tissues including 7 pairs of FL and tFL samples, 4 triplets with a tFL and two FL samples, and 15 single tFL samples. The 7 pairs of FL and tFL samples were obtained at FL diagnosis and at later transformation

diagnosis.3 out of these 7 pairs of FL and tFL samples were performed by WES earlier and resequenced for the performance comparison. The 4 sample triplets were obtained along with the disease progression. The samples were provided by the UNMC, the LLMPP, or Aarhus University Hospital. This study was approved by the UNMC institutional review board.

B.1.1 Sample and patient materials

The time between the FL diagnosis and transformation varied from less than 1 year to 7 years. The tFL samples were diagnosed as DLBCL by a panel of LLMPP hematopathologists. Our earlier GEP analysis had assigned the tFL samples as ABC-like, UC and GCB-like. When BCL2 status was unknown, BCL2 rearrangement was assessed by PCR to determine if the frequency of t(14;18) positivity expected for an FL dataset was observed in the tFL dataset. We did PCR with primer that prime in 3 places where the break point often occurs, major breakpoint region, minor cluster region and intermediate cluster region (Table 2-2).

Case ID	FL 1 biopsy	FL 2 biopsy	tFL biopsy	Custom panel	Paired case	FL 1 biopsy date	FL 2 biopsy date	tFL biopsy date	t(14;18) status (tFL)	Sub type
Case3	X		X	X	X	4/6/1993		4/29/1994	unknown (MCR,MBR,ICR neg)	ABC-like
Case5	X		X	X	X	4/28/1989		2/14/1990	pos (MCR)	GCB-like
Case10	X		X	X	X	1/20/1996		4/28/1999	pos (MCR)	ABC-like
Case15			X	X					pos (ICR)	GCB-like
Case16			X	X					pos (ICR)	GCB-like
Case18			X	X					unknown (MCR,MBR,ICR neg)	GCB-like
Case19			X	X					positive (clinical info)	ABC-like
Case20			X	X					pos (MBR)	GCB-like
Case21			X	X					unknown (MCR,MBR,ICR neg)	ABC-like
Case23	X	X	X	X	X	11/16/1992	2/9/1995	1/26/1998	pos (MBR)	ABC-like
Case24	X	X	X	X	X	1993	1993	2000	pos (MBR)	GCB-like
Case25	X	X	X	X	X	1990	1990	1993	positive (der18 aCGH)	GCB-like
Case26	X	X	X	X	X	1994	1995	1995	pos (MBR)	GCB-like
Case27	X		X	X	X	1994		1999	unknown (MCR,MBR,ICR neg)	GCB-like
Case28	X		X	X	X	1988		1994	positive (LLMPP sheet-MBR)	GCB-like
Case29	X		X	X	X	9/14/1994		11/18/1994	unknown (MCR,MBR,ICR neg)	UC
Case31	X		X	X	X	1/21/2008		7/17/2012	unknown (MCR,MBR,ICR neg)	UC
Case32			X	X					pos (MCR)	ABC-like
Case33			X	X					unknown (MCR,MBR,ICR neg)	unknown
Case34			X	X					pos (MBR)	GCB-like
Case35			X	X					pos (MBR)	ABC-like
Case36			X	X					pos (MBR)	GCB-like
Case37			X	X					positive (der18 aCGH)	GCB-like
Case38			X	X					unknown (MCR,MBR,ICR neg)	GCB-like
Case39			X	X					unknown (MCR,MBR,ICR neg)	GCB-like
Case40			X	X		4/6/1993		4/29/1994	positive for BCL2(FISH)	GCB-like

Table 2-2. Sample and patient information for custom panel sequencing.

B.1.2. Custom capture panel design

The potentially interesting genes that were included in the custom capture panel were selected based on the preliminary analysis of some of our WES analysis and previously published lymphoma sequencing studies. The custom capture panel allows us to sequence genes of interest in a large number of samples at greater sequencing depth. We expected the custom capture panel to allow us to evaluate clonal architecture and evolution. The criteria for selection included:

1. B-cell expressed genes[25] that were recurrently mutated in the 11 initial WES cases,

2. Classic cancer genes (http://cancer.sanger.ac.uk/cancergenome/projects/classic),

3. Genes frequently mutated in lymphomas according to the Catalog of Somatic Mutations in Cancer (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) or

4. Genes recurrently mutated in previously published B-cell lymphoma sequencing studies[8,26-29]. 496 genes potentially involved in the FL and tFL disease process were included in the custom panel (Table 2-3).

ABCA7	4	CDT1	3	FCGBP	3	JAK3	2	NPR2	3	SGK1	3
ABCB1	3	CELSR1	3	FGFR1	1	KANSL2	3	NRAS	3	SHKBP1	3
ABL1	4	CENPE	3	FGFR4	3	KDM2B	3	NTN3	3	SHPRH	3
ACACB	3	CEP350	3	FLNC	3	KIAA0100	3	NUMA1	3	SIDT2	4
ACSF3	4	CHD2	3	FLYWCH1	1	KIAA0430	3	NUP98	4	SIN3A	3
ACSS1	3	CHD3	3	FNDC3B	3	KIAA1109	3	NVL	3	SMAD4	2
ACTB	3	CHD8	4	FOLH1	3	KIAA1211	3	ODF2	3	SMARCA4	2
ACTG1	3	CHTF18	3	FOXO1	5	KIAA1551	3	OFD1	2	SMARCB1	2
ACVR1B	2	CIC	3	FOXO4	3	KIF1B	1	OGT	4	SMG7	3
ADAMTS18	3	CIITA	5	FRG1B	1	KIF4A	3	P2RY8	4	SNRNP200	3
ADRBK1	3	CLASP1	3	FRYL	3	KLF2	2	PABPC1	1	SNX19	3
AFF1	4	CLEC16A	3	FTCD	3	KLHL6	4	PACS1	3	SOCS1	4
AFF4	3	CLSTN3	3	FTH1	2	KMT2A	5	PALD1	3	SPEN	4
AKAP13	3	CLUH	3	FUBP1	5	KMT2C	6	PASD1	2	SPG11	3
AKAP8	2	CNOT1	5	GAK	3	KMT2D	7	PASK	3	SPTBN1	3
AKAP9	4	CNOT3	3	GAS7	3	KMT2E	3	PAX5	2	SPTBN5	5
AKT1	1	COL7A1	3	GATA2	1	KNTC1	3	PAXIP1	3	SRSF2	1
ALMS1	3	CREB3L2	3	GATA3	1	KPNA5	3	PBRM1	6	STAT3	5
ANKLE2	3	CREBBP	9	GCN1L1	4	KRAS	4	PCBP1	1	STAT6	4
ANKRD12	3	CRTC1	3	GGA1	3	KTN1	3	PDCD11	1	STK11	1
ANKRD17	3	CRTC2	0	GIGYF2	4	LAMP1	3	PDE4DIP	3	STK4	4
ANXA1	3	CSNK1D	3	GNA11	2	LCP1	3	PDS5B	4	SYK	4
APC	5	CSRP2BP	3	GNA13	4	LILRB1	3	PGAP2	1	SYNE1	8
ARHGAP30	3	CTBP2	1	GNAS	1	LRBA	4	PHF3	3	TAF1	4
ARHGEF12	3	CTCF	3	GOLGA3	3	LRCH4	3	PHF6	3	TAGLN	3
ARHGEF2	4	CTNNA1	1	GOLGA4	3	LRP10	4	PHKA2	3	TARSL2	3
ARID1A	4	CTNNB1	3	GPR82	3	LRRC7	2	PHRF1	3	TBL1XR1	4
ARID1B	3	CUL7	4	GRB2	2	LRRK1	5	PICALM	3	TBP	2
ARRDC2	3	CXCR5	3	GSE1	3	MACF1	3	PIK3CA	3	TCF3	4
ASPM	3	CYLD	4	GTPBP8	3	MALT1	3	PIK3R1	5	TDG	2
ASXL1	4	CYP1A2	1	GTSE1	2	MAP2K4	1	PIM1	6	TET2	4
ATAD3B	3	DARS2	3	GYS1	3	MAPK1	1	PKD1	3	TIGAR	3

ATF7IP	3	DAXX	1	HCK	3	MBTPS1	3	PLCG2	4	TINF2	2
ATM	4	DAZAP1	3	HDAC7	3	MCTP2	3	PLEKHA5	3	TMC8	3
ATP10A	3	DDX3X	4	HEATR5B	3	MDN1	4	PMS1	3	TMEM30A	5
ATRX	5	DDX56	4	HELZ	3	MED12	4	POU2F2	1	TNFAIP2	3
AXIN1	1	DENND5A	3	HERC1	2	MED26	3	PPP2R1A	2	TNFAIP3	6
B2M	6	DEPDC5	3	HIST1H1B	3	MEF2B	5	PPP2R3C	3	TNFRSF14	6
BAP1	1	DGKZ	3	HIST1H1C	5	MEN1	2	PPP2R5A	1	TNK2	3
BAZ2A	3	DHX15	4	HIST1H1D	7	MET	2	PRDM1	3	TOPBP1	3
BBS10	3	DIAPH1	1	HIST1H1E	4	METTL9	3	PRDM15	2	TP53	9
BBX	3	DIP2A	3	HIST1H2AC	4	MGA	3	PRKAR1A	1	TPR	3
BCL10	7	DMD	4	HIST1H2AG	3	MIA3	3	PRKDC	5	TRAF3	4
BCL11A	5	DMXL1	3	HIST1H2BC	5	MICAL3	3	PRMT6	3	TRIP11	4
BCL2	7	DNM2	5	HIST1H2BG	2	MKI67	3	PTCH1	3	TSC2	4
BCL2L10	1	DNMBP	3	HIST1H3B	1	MLH1	2	PTEN	6	TSC22D1	2
BCL6	5	DNMT3A	2	HIST1H3H	3	MLLT10	5	PTPN11	2	TSPAN32	1
BCLAF1	1	DOCK11	3	HIVEP2	3	MON2	3	PTPRH	3	TTC27	3
BCOR	4	DOT1L	3	HIVEP3	3	MS4A1	3	PUM1	3	U2AF1	1
BCR	5	DPF2	3	HLA-A	2	MSH2	3	PXDN	3	U2AF2	4
BIRC6	4	DPYD	3	HLA-B	2	MSH6	4	RAPGEF1	4	UBAP2	4
BLM	3	DTX1	3	HLA-C	2	MST1	4	RAPGEF2	3	UBC	3
BOD1L1	3	DYNC1H1	3	HLA-DMB	2	MTG2	3	RASGEF1A	1	UBE2A	2
BRAF	3	DZIP3	3	HMGB1	1	MTR	3	RB1	4	UBR4	4
BRCA1	3	EBF1	4	HNF1A	1	MUC4	4	RBM15	3	UBR5	3
BRCA2	5	EDEM3	3	HPS3	3	MUM1	3	RBM39	3	UPF1	3
BRD2	3	EGFR	4	HPS5	3	MYBL2	3	RBMX	3	UPF2	3
BRD4	4	EGLN1	3	HRAS	2	MYC	5	REV1	3	USP10	3
BRD8	3	EIF4A2	3	HSPA8	3	MYD88	7	REV3L	1	USP19	3
BTAF1	3	ELP2	2	HTT	3	MYH11	3	RFTN1	5	USP34	3
BTBD3	3	EML4	4	HUWE1	4	MYH9	3	RFX7	4	VHL	2
BTG1	4	EP300	7	HVCN1	3	MYO18A	2	RFXAP	3	VPS13C	3
BTG2	4	EP400	3	ICE1	3	MYO1G	2	RGS12	3	WDR76	4
BUB1B	4	EPHA7	3	ID3	1	MYRIP	3	RHOH	1	WDR90	1
CALR	1	ERAP1	1	IDH1	4	NBAS	3	RLTPR	3	WHAMM	1
CAMTA1	4	ERICH1	3	IDH2	1	NBEAL2	3	RNF103	3	WHSC1L1	3
CARD11	8	ERMARD	3	IFI16	3	NCOR1	3	RNF213	7	WRN	3
CARS	4	ETS1	2	IGF1R	3	NCOR2	2	RNF40	3	XPO1	4
CASC5	4	EWSR1	3	IGFN1	2	NF1	4	ROBO1	4	YY1AP1	2
CBL	2	EXOC4	3	IGLL5	2	NF2	1	ROCK2	3	ZC3H18	3
CCDC94	3	EXOSC6	1	IKZF1	2	NFE2L2	2	RPN2	3	ZMYM3	4
CCND3	6	EZH2	7	IKZF3	5	NFKB2	4	RRP1B	3	ZNF142	3
CCNH	3	FAM186A	1	IL7R	2	NFKBIA	3	RTTN	3	ZNF500	3

CD22	3	FANCA	4	INO80	4	NIN	4	RUNX1	3	ZNF521	6
CD36	7	FANCD2	3	INPP5D	3	NISCH	3	S1PR2	3	ZNF600	3
CD58	5	FANCE	3	INTS10	3	NKAP	3	SAMD9	3	ZNF608	3
CD70	3	FANCF	3	IQGAP1	3	NOA1	1	SCAPER	3	ZNF708	3
CD74	5	FAS	4	IRF4	4	NOC2L	3	SEC23IP	3	ZNF830	1
CD79B	7	FASN	4	IRF8	4	NONO	3	SENP6	4	ZNF85	1
CDAN1	3	FBLN2	3	ITPKB	3	NOTCH1	6	SETD2	3	ZRSR2	1
CDC73	2	FBXO11	3	ITPR2	4	NOTCH2	3	SETDB1	4	ZWILCH	1
CDH1	2	FBXO31	2	JAK1	2	NPIPB15	1	SF3A2	3		
CDKN2A	3	FBXW7	6	JAK2	2	NPM1	1	SF3B1	5		

Table 2-3. Genes selected in custom capture. Note: The number indicates the recurrence of each gene was reported by different resources.

B.1.3. Library preparation

Agilent's Haloplex custom gene enrichment panel was designed using Agilent's SureDesign software. 200 ng of DNA per sample was used for capture, and 10-12 samples were sequenced per lane.

To prepare the samples, we first used restriction enzymes to fragment the DNA samples. After DNA denaturation, the biotinylated HaloPlex probes guided the targeted fragments to form circular DNA molecules. These biotinylated HaloPlex probes were specifically designed to hybridize to both ends of the fragmented DNA samples. Sample barcodes were incorporated in the DNA sample at the same time. Afterwards, magnetic streptavidin beads were used to retrieve perfectly hybridized DNA fragments. The circular DNA molecules were then closed by ligation and amplified by PCR. After amplification, the DNA molecules were ready for sequencing.

B.2. Pipeline for calling variants in custom capture panel sequencing

Since we had cases with either one, two, or three samples, the pipelines were designed slightly differently (Figure 2-3, Figure 2-4 and Figure 2-5). In general, after sequencing, the raw read quality was first evaluated by FastQC (v0.10.1), and then the raw reads were mapped to human reference genome hg19 using BWA (0.7.5a-r405). GATK (v3.1-1-g07a4bf8) was used for

local realignment and base quality recalibration. Variant calling was performed with VarScan2 (v2.3.6). Basic filtering was performed as described above for the WES samples. The variants were then annotated using ANNOVAR[24]. We filtered out known germline variants (SNPs and indel polymorphisms) based on dbSNP138 and only selected mutations that were nonsynonymous, stop gain, stop loss, insertion or deletion in the coding region or splice sites. Since we had sample pairs and triplets, we categorized our mutations into three types: FL-unique; tFL-unique; and shared mutations as described above for the WES samples.

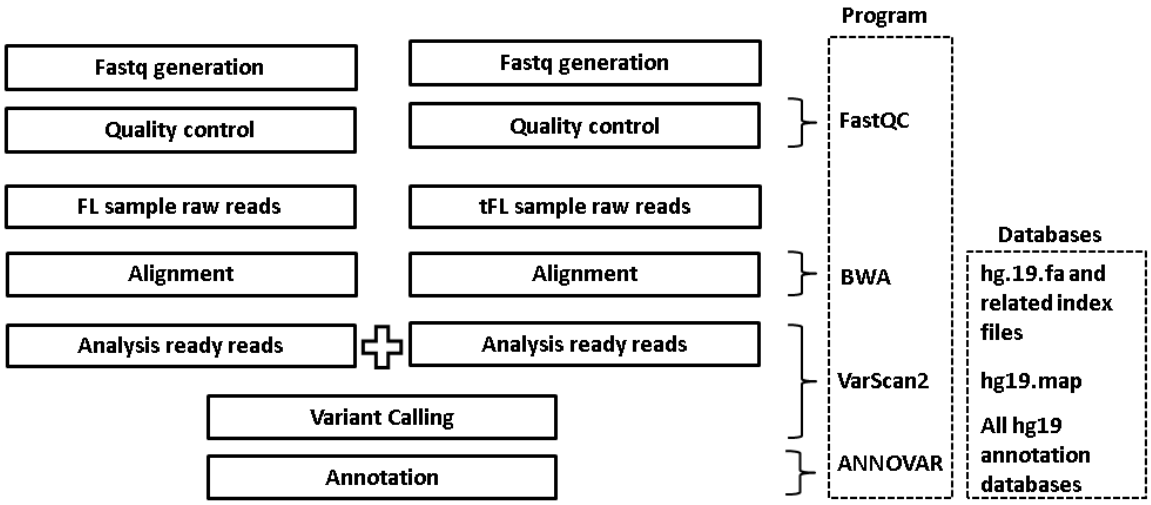

Figure 2-3. Pipeline designed for custom capture panel with paired sample.

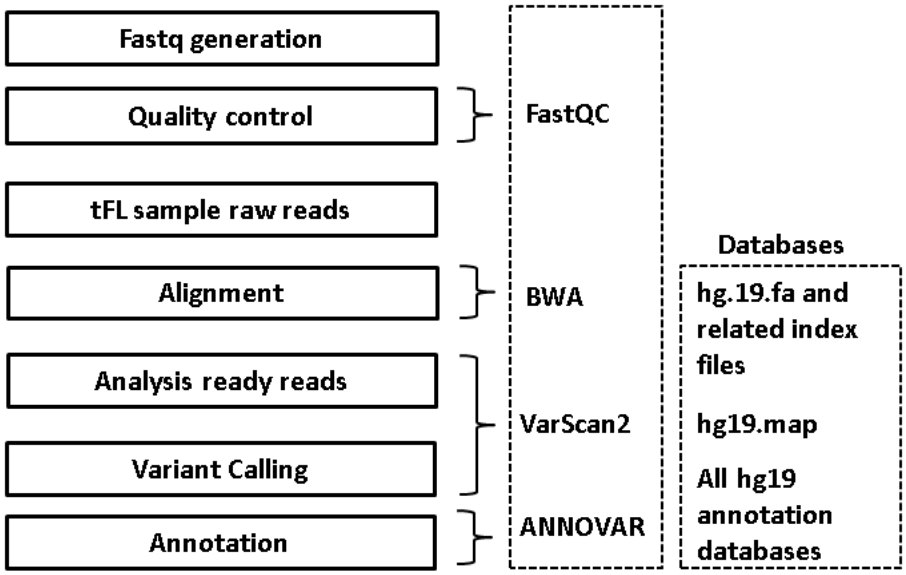

Figure 2-4. Pipeline designed for custom capture panel with single sample.

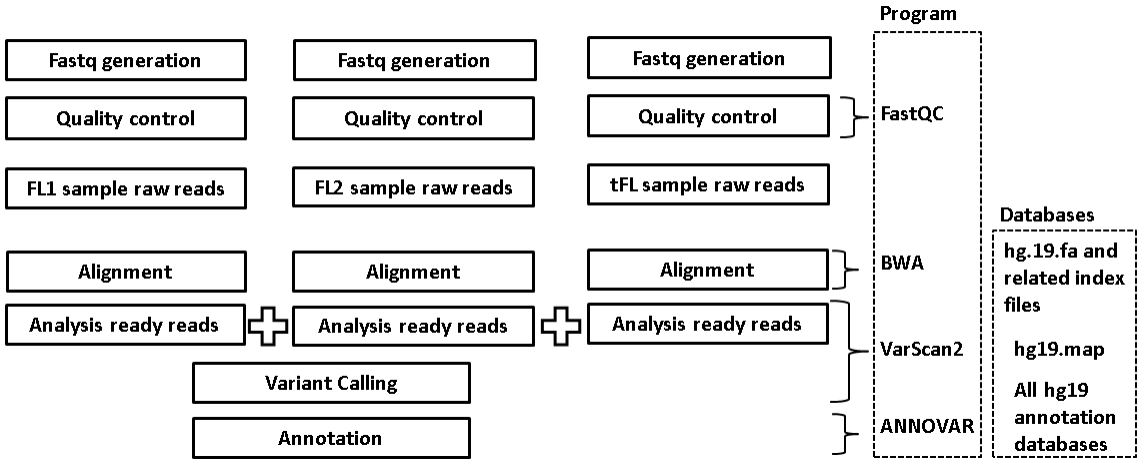

Figure 2-5. Pipeline designed for custom capture panel with tripled sample.

B.3. Custom capture panel performance analysis

We applied custom capture panel performance analysis (Figure 2-6) to check the total number of reads used for alignment, the number of reads that were mapped to the reference and the number of aligned reads that were mated with their paired reads.

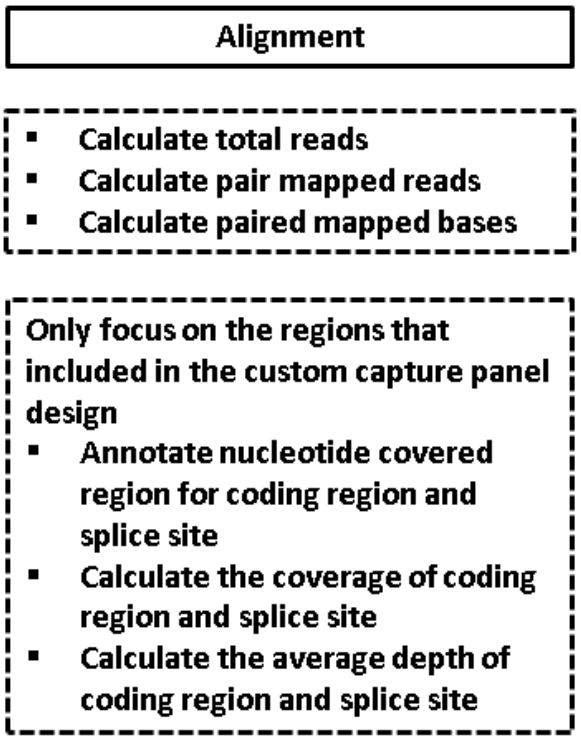

Figure 2-6. Custom capture panel evaluation pipeline.

B.4. Pipeline validation

We selected 3 samples that were previously sequenced by WES and applied custom capture panel sequencing to them again to evaluate the performance of the custom capture panel.

We made two mutation lists from the 3 samples: one for WES, and the other one for custom capture panel sequencing. The mutations were chosen in the regions covered by both WES and custom capture panel sequencing. The overlap of the two mutation lists, the variant frequency, mutation depth, and overall coverage were compared in the 3 samples. The previous Sanger

sequencing results that we did in the 3 paired overlap samples were used again to validate the mutations detection accuracy in the custom capture panel sequencing. KARPAS422 is a cell line with known mutations. We sequenced the cell line by custom capture panel and checked if it detected the known mutations in the cell line.

C. Binomial distribution model for estimating the duplicate ratio

The restriction enzyme based custom capture panels select genes at exactly same start and end positions. Because of this, the well-known duplicates marking program Picard cannot identify duplicates. This duplicate generation is due to the amplification by PCR causing multiple reads to be derived from a single initial molecule. We designed a binomial distribution statistical model to estimate the duplicate ratio (the average number of reads per initial molecule). The model is based on the variant frequency of heterozygous germline variants being approximately 50% and the p-values from the statistical test for each heterozygous germline variant should be uniformly distributed.

C.1. Data preprocess

After basic filtering, the variants have balanced reads supporting alternative alleles and good base quality. We then used dbSNP to annotate the heterozygous germline variants by using our own scripts instead of annotation programs as we were only interested in heterozygous germline variants. We also checked the distribution of the variant frequency, total read number and the number of reads supporting the alternative allele to avoid bias or outliers that might potentially affect the estimation.

C.2. Estimating the duplicate ratio

For each sample, we generated the p-values for each heterozygous germline variant based on the binomial test with 50% probability of success. We only used the p-values between 0.1

and 0.9. The reason that we did not use p-values <0.1 is that sequencing errors introduced many apparent variants with low variant frequencies that fortuitously matched dbSNP, the p-values close to zero were overrepresented and rejected the null hypothesis (probability to get reads supporting one allele is 50%). The reason we didn't use p-values >0.9 is the majority of total reads are assigned equally in heterozygous germline variants so that the p-values were close to 1 and did not reject the null hypothesis. We then selected an alpha value which makes the slope of the regression model fitting the frequencies of the p-values to the p-value close to 0. The alpha value is the estimated PCR-duplicate ratio that can adjust the slope of the linear regression model to the uniformly distributed pattern.

C.3. Model validation

We simulated a dataset based on the real total number of reads in our samples and assigned a duplicate ratio equal to 1, 2 and 4 to confirm the patterns of the p-values from the binomial test. If there are no duplicates, most of the hypotheses will not be rejected; the majority of the total reads are assigned equally; therefore, there should be a big peak on the far right of the histogram of p-values; and the rest of the p-values are uniformly distributed. If there are PCR-generated duplicates, most of the hypotheses will be rejected; a majority of total reads are assigned unequally; and p-values close to zero should be overrepresented. As there is a bigger PCR-duplicate ratio, the peak close to zero is expected to be more exaggerated. We also did 10,000,000 simulations based on a real number of reads from all of our samples with duplicate ratio equal to 1 to confirm the binomial distribution model. We should see the trend be flat except the p-values close to 1, which confirms that the p-values from the binomial test will be uniformly distributed if there are no duplicates.

D. Filtering method for removing germline variants

The variants called by either the WES or custom capture panel sequencing pipelines include two groups, germline variants and somatic mutations. Germline variants are heritable variations that are present in the germ cells and can pass to all cells in the body of the progeny. In contrast, somatic mutations are not inherited from a parent and cannot be transmitted to offspring. The conventional mutation analysis requires paired normal and tumor samples to distinguish germline variants from somatic mutations. The general idea is that the variants that are detected in both normal and tumor samples are germline variants, and conversely, the variants that are only found in the tumor sample are considered somatic mutations. We were interested in detecting somatic mutations that drove transformation, but our samples lacked corresponding normal samples to filter out germline variants that were absent from dbSNP, which are predominantly SNPs present at very low population frequency (private SNPs). A filtering based method was used to filter out germline variants and keep the somatic mutations. The filtering method is based on the biological understanding of FL. The mutations that drive transformation or contribute to FL often: occur recurrently; associate with critical CNAs and oncogenes; have an impact on amino acid structure and function of protein. Because FL is a B cell lymphoma, the mutated genes have to be expressed in B cell to be able to play roles in FL development. Datasets from reliable organizations provide comprehensive information to help filter out false positive mutations. As examples, there are common errors that may occur in sequencing technology or variants that are reported in dbSNP.

D.1. Databases collection

Multiple databases were collected or constructed to help sort out germline variants and somatic mutations. The databases are listed as following:

1. List of genes expressed in B cell. Genes that were expressed in B cells based on our previous GEP analysis and whole transcriptome sequencing data from lymphoma cell[25]. Genes were considered not significantly expressed in B-cells if their maximum FPKM value was less than 1 of 5 naïve B-cell samples, 4 germinal center B-cell samples, and 5 DLBCL samples that include both GCB and ABC subtypes. For a few genes not included in this dataset, our Affymetrix dataset of normal B-cell and DLBCL samples was used to determine expression and a maximum log2 value of 8.5 was set as the threshold of expression;

2. Sanger cancer gene list (http://cancer.sanger.ac.uk/cancergenome/projects/classic);

3. Cosmic cancer gene list (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/);

4. List of apparent recurrent mutations in 13 normal samples. Mutations that were detected in normal samples were considered as private SNPs or sequencing artifacts;

5. List of mutated genes from previous published lymphoma sequencing studies[8,26-29];

6. dbSNP138;

7. Annotation databases from ANNOVAR (http://annovar.openbioinformatics.org/en/);

8. Paralog gene list (http://massgenomics.org/2013/06/ngs-false-positives.html). The paralog gene list is used to avoid 2 types of false positive variants. One is from genes that are physically large so that they tend to accumulate a lot of mutations. The other is from the misalignment to a paralog, in which case the reads supporting the variant allele were originally from another part of the genome.

D.2. Data preprocessing

Each database we used in the filtering method has its own specific format. Some databases include extremely comprehensive formats. We extracted the information and parsed it to a

simple version so that we can easily apply it to future interaction. We also preprocessed the format of our variant list in order to connect it to various databases.

D.3. Germline variant filtering

To obtain a confident somatic variant list, a filtering process was applied. This pipeline generated a mutation list with two tiers of confidence and one rescued tier. First, variants were excluded if they met specified criteria from the following:

1. Existed in the dbSNP database (dbSNP138) with same orientation;

2. Synonymous mutations;

3. Found in a set of 13 unrelated normal samples as SNPs or artifacts;

4. Shared by samples from six or more cases or 50% of total samples unless they are from a gene known to have frequent mutations at the site (such as EZH2);

5. Found other variants in closely neighboring positions;

6. Mutations in intergenic regions, introns, and 5’ or 3’ untranslated regions (UTR);

7. Genes previously reported as frequently showing false positive mutations (http://massgenomics.org/2013/06/ngs-false-positives.html).

Second, the remaining variants were considered as candidate somatic mutations (CSMs). Mutations were included in our list only if they met specified criteria from the following:

1. Not being found in all samples from the same patient and cannot be explained by copy number (CN) abnormality;

2. Previously identified in a B-cell lymphoma genome;

3. Being a truncating mutation;

4. Previously identified in a non-B-cell-lymphoma cancer genome;

5. Identified in a gene from the Cancer Gene census list;

6. Having a reduced variant frequency which cannot be explained by CN abnormality.

Based on these 6 criteria above, the reported variants were classified into two tiers. Mutations satisfying one of the first three conditions, or two of the other three conditions were classified as high-confidence Tier 1 mutations; mutations that met only condition 6 were classified as less confident Tier 2 mutations. Because we used very strict filtering criteria to try to ensure that we did not include germline variants (private SNPs) in our mutation data, we may have filtered out some true mutations. To avoid underestimation of prevalence of important genes, other CSMs in B-cell expressed, recurrently mutated genes assigned to Tier 1 and 2 lists were rescued and designated Tier R. All the other mutations were considered as likely germline variants, and were thus excluded from the following analysis.

D.4. Filtering method validation

To validate the performance of the pipeline for filtering out germline variants, we applied the pipeline to a published FL sequencing dataset. This dataset has FL, tFL and normal samples from the same patient so that we can use the normal samples for filtering. 6 WGS samples with matched FL, tFL and normal samples were downloaded from a published paper[9]. We generated two somatic mutation lists. One somatic mutation list was generated using our WES pipeline (Figure 2-1) and did not utilize the matched normal data. The other somatic mutation list was generated using the standard somatic calling pipeline VarScan2 taking into account the normal data. This somatic mutation list did not include germline variants. We then compared the two somatic mutation lists at each criterion and validated the filtering method.

E. Machine learning model for removing germline variants

Predictive modeling is a concept of building a trained model that learns certain features from a training dataset and then applying the trained model to make predictions from other datasets. Typically, there are two main categories of predictive modeling. One is able to predict continuous outcomes and is based on the relationships between attributes and trends. This is called a regression model. The other one is capable of predicting discrete outcomes and is based on grouping the attributes. This is called pattern classification which can be divided into two types: supervised and unsupervised learning. Supervised learning uses the known discrete outcomes in the training dataset to train the model. In unsupervised learning the outcomes in the training dataset are not known, therefore, the program tries to find unknown information from the attributes in the dataset. In our study, we used supervised learning modeling to predict germline variants in pairs of FL and tFL samples without normal samples.

E.1. Dataset preprocessing

We collected two published sequencing datasets[9] that have FL, tFL and normal samples from the same patients so that we can use one as the training dataset and the other one for the testing dataset. We first applied a standard somatic calling pipeline using VarScan2 and generated variants. The variants were classified into germline variant and somatic mutation groups by comparing with their corresponding normal samples. We then annotated the variants with well-known prediction scores by ANNOVAR[24]. The variants were variant call format, which packs information in couple columns. We parsed the information into a more organized and convenient format for the models to do the training and testing. The parsed format separated the packed information into individual columns, reorganized the columns, and only kept the information we need for next steps.

E.2. Model training

We applied 5 robust machine learning algorithms to our training dataset. These were SVM, recursive partitioning (RP), Breiman and Cutler's random forests (RF) for classification and regression, classification and regression tree (TR) and feed-forward neural networks (NNET). We used R (statistical language) package *e1071*, *rpart*, *randomForest*, *tree*, and *nnet* for each of the respective models listed. The features that we used for training were the variant frequency of the FL sample, variant frequency of the corresponding tFL samples, dbSNP, and SIFT score. SIFT score is a score that can predict amino acid changes that affect protein function; the smaller the score is, the more deleterious the variant. We applied two sets of features. One set is called a basic feature set; it includes variant frequency for FL, variant frequency for tFL, and dbSNP. All our variants have the information for the basic feature set. Therefore, all of them were included in the training dataset. The other set is called a complex feature set, this set added the SIFT score as the extra feature. SIFT fails to provide a score for a very small number of the variants in the datasets. The machine learning algorithms do not work with missing data; therefore, we only used the mutations that have SIFT score information for our complex feature set. We trained all the 5 models with two feature sets in Okosun's dataset and tested them in Pasqualucci's dataset with statistical measures individually. ROCR[30] is another R package for evaluating the performance of classifiers. We used it to evaluate the performance of our models. We used 5 different R packages, each of them has its own design. Therefore, when we applied the statistical measures to the tested models, we made 5 different scripts to retrieve information required by the calculation.

E.3. Model selection and validation

We applied the 5 different models to the two published datasets. The statistical values, sensitivity, specificity, area under the receiver operating characteristic (ROC) curve (AUC), and false discover rate (FDR), were calculated for each model based on its performance in the testing dataset. We also applied the trained models to one of our own sequencing samples. These samples had the best coverage and depth. As we mentioned earlier, our own sequencing data does not have corresponding normal samples to remove germline variants, but we applied a filtering based method to effectively remove most of the germline variants. Therefore, the somatic mutations predicted by the machine learning models should have decent overlap with the results generated by our filtering based method. The best machine model was selected by considering all the statistical results.

F. Data integration for downstream analysis

To have a more comprehensive view of the genomic alterations and to use the CNV data to assist the analysis of mutation data, we integrated our mutation data with our previous corresponding CNV data and other two sequencing data to investigate the genetic landscape of the disease.

F.1. Integration of mutation and CNA datasets

Integration of the mutation data with CNA data can provide complementary information for us to examine their association with FL transformation. In our CNV analysis, a circular binary segmentation algorithm was applied to segment chromosomes into regions of similar CN log2 ratio. It translated noisy intensity measurements into regions of equal CN log2 ratios and connected the change point to the locations of regions with aberrant DNA CNs and thereby identified regions of CN gain and loss. Each chromosome in each sample was divided into many

segments with similar CN. These segment files were used to assign the CN for each mutation in our analysis. We also identified mutated genes that were included in the regions of each rCNA for data interaction.

F.2. Integration of mutation datasets

In order to enhance the investigation of the genetic events that drive the transition from FL to tFL, we integrated other two published mutation datasets and identified recurrent mutations.

For each dataset, we first retrieved the mutations with their original sample name, mutation positions, variants, mutation types, and gene names. Then we classified each mutation as FL-unique, tFL-unique or shared. All the 3 datasets were then combined into a comprehensive matrix. In this matrix, the rows included all the gene names that had been detected in any of the 3 datasets, the columns included all the paired samples that had been used for mutation detection. Additionally, 5 columns were added: FL-unique mutation frequency, tFL-unique mutation frequency, shared mutation frequency, genes expressed in B cell, and genes selected in our custom capture panel. For each sample and each gene, mutation types were annotated; some of them had more than one type. We classified them as genes with only shared mutation, genes with only FL-unique mutation, genes with both FL-unique and shared mutations, genes with tFL-unique mutation, genes with both shared and tFL-unique mutations, genes with both FL-unique and tFL-unique mutation, genes with FL-unique, shared and tFL-unique mutations, and genes with CNV. We tallied the frequencies for sorting the mutations conveniently. We also identified mutations that were involved in potentially important domains and pathways from the combined dataset. We also identified mutations that were detected in FL samples at low levels but had increased variant frequency in the tFL for clonal analysis.

G. Survival analysis

Most standard statistical models require a normal distribution and cannot be applied in survival analysis because the time to the event occurrence is rarely normally distributed and we often lack follow-up information on the patients. For an example of the latter, logistic regression can study how risk factors are associated with disease or affect the time to the disease, but patients may drop out of the study or fail to develop a disease before the end of the study. All these realities make logistic regression unfit for survival analysis. The Kaplan-Meier (KM) method uses information from the dropout patients, rather than simply throwing it away, to estimate the survival probability at a given time.

The log-rank test is a widely used hypothesis test in survival analysis. It is applied to compare the survival distributions of two groups of patients. One group includes patients with a factor, and the other group includes patients without the factor. After the test, we can come up with a conclusion about whether the two groups (with and without the factor) have identical survival functions. The log-rank test works well when the factor has the exact same weight or influence at each observed event time. The majority of the situations in the biological field are, however such that the characteristics of the factors do not always have exactly the same impact during the disease development. For example, loss on 9p21.3 occurring early in FL development[31] has more influence on patients' early overall survival. If we use a log-rank test, we will easily miss the biological abnormality information that can improve the analysis. Therefore, in our study, we designed a survival analysis that associated an understanding of how the abnormalities contribute to FL development and provided a more accurate and comprehensive test.

Statistical Analysis System (SAS), a professional software suite for statistical analysis that is flexible to select the test that matches the biological property, is convenient to navigate to different analysis parts, and facilitates the display and retrieval of integrated information, was applied to our survival analysis.

G.1. Dataset description

The samples that were used for the survival analysis were obtained from the LLMPP or the UNMC Pathology/Oncology Database. This study was approved by the UNMC institutional review board and conducted in accordance with the declaration of Helsinki. Diagnoses were confirmed by a panel of LLMPP hematopathologists.

We originally performed CN analysis using the high-resolution GeneChip Human Mapping 250K Nsp SNP array (Affymetrix) on 225 FL and 84 tFL samples. After preliminary analysis, 198 FL and 79 tFL samples had sufficient quality and were kept for rCNA identification. The method for rCNA identification was described in a previous published paper[31]. In general, the raw CEL files were imported into Genotyping Console 4.1 software (Affymetrix) generating SNP genotypes and probe-intensity log2 ratios (relative to 48 normal controls provided by Affymetrix). A DNAcopy R package from Bioconductor was applied to segment probe values and estimate CN. The recurrent abnormalities, represented as rCNAs, were then identified.

We had clinical data for 149 FL cases and 21 tFL cases. These tFL samples were defined as DLBCL that occurred in patients diagnosed with FL. Since we had a very small number of tFL with clinical data, we only applied the survival analysis to FL cases.

G.2. Applying KM method

In survival analysis, the time variable was survival time, and the event was death. Censoring is the key analytical problem in survival analysis; it occurs when we have the individual true

survival time interval, but it is unknown most of time. Most survival time intervals are right-censored, because the survival time has been cut off before the true survival time, in other words, the observed survival time is shorter than the true survival time. Our goal is to use the right-censored observed survival time in FL patients to suggest the true survival time. The KM method was applied to estimate the survival probability at a given time. We made use of a risk set at a given time. By risk set, we mean the information we have on the individuals who have survived at that given time. We also used the information we have on censored people up to the time of censorship instead of discarding the information. The KM survival curves were plotted to interpret the survival data.

We examine CNVs as factors that potentially influence the patients' survival time. We used them to divide the clinical data into two groups for future survival comparison. For each group, we applied the KM method to the FL clinical data. The number of censored patients, deceased patients and living patients were all calculated at the time point that patients either dropped or deceased in the study. A survival curve was generated by the KM method for each group for future survival curves comparison.

G.3. Applying log-rank test

The log-rank test is the most popular method for testing whether two KM curves are statistically equivalent. The null hypothesis is that there is no overall difference between the two survival curves. In our study, the null hypothesis was that there is no overall difference between the group with the abnormality and the group without the abnormality. Under this null hypothesis, the log-rank test used a chi-square distribution to test a statistic with one degree of freedom. The categories of the outcomes were defined by each of the ordered failure times for the entire set of data and the statistics used the categories of the outcomes to calculate the

observed cell counts versus expected cell. The expected cell count (Equation 2-1 and Equation 2-2) was the proportion of the number of patients in one group out of the total patients in both compared groups at risk at time j (Equation 2-3) multiplied by the total number of deaths at time j in both groups (Equation 2-4). The log-rank test statistic (Equation 2-5) divided the square of the sum of the observed counts (m_{ij}) minus expected counts (e_{ij}) over all failure times for either one of the two comparing groups by the variance of the sum of the observed counts minus expected counts over all failure times for either one of the two comparing groups. A p-value for the log-rank test was determined from the chi-square distribution tables. We used the p-value to conclude whether or not this abnormality affects the survival.

Equation 2-1: $e_{1j} = \left(\frac{n_{1j}}{n_{1j}+n_{2j}}\right) \times (m_{1j} + m_{2j})$ expected cell count for group 1 at jth failure

Equation 2-2: $e_{2j} = \left(\frac{n_{2j}}{n_{1j}+n_{2j}}\right) \times (m_{1j} + m_{2j})$ expected cell count for group 2 at jth failure

Equation 2-3: $n_{ij}/(n_{1j} + n_{2j})$ proportion of patients at jth failure, i=1, 2 represents the group number

Equation 2-4: $(m_{1\ j} + m_{2\ j})$ the total number of deaths at jth failure, 1 and 2 represent the group 1 and group 2

Equation 2-5: $\frac{(\sum_j (m_{ij}-e_{ij}))^2}{var(\sum_j (m_{ij}-e_{ij}))}$ *i*=1, 2 represents the group number, *j* is the order of failure time

G.4. Applying alternatives to the log-rank test

To test the null hypothesis that two survival curves are statistically equivalent, there are several alternatives to the log-rank test as shown in Table 2-4. The main difference in the methods is the weights at the jth failure time in test statistic (Equation 2-6). The log-rank test uses the summed observed minus expected failures in each group to form the test statistic. This

simple sum gives the same weight to each failure time when combing observed minus expected failures in each group. In contrast, the alternative tests apply different weights at the jth failure time. Take the Wilcoxon test as an example. It places more emphasis on the information at the beginning of the survival curve where the number at risk is large allowing early failures to receive more weight than later failures. This type of weighting is often used if a factor related to survival has stronger effect in the earlier phase and tends to be less effective over time. Therefore, if the CN abnormality occurs in the early stage of FL, the Wilcoxon test will be a better choice than the log-rank test. Similarly, the Tarone-Ware test weights the early failure time more heavily, but the weight is between log-rank and Wilcoxon. Prentice[32] illustrated that the Tarone-Ware test is always superior to either the log-rank test or Wilcoxon test. The Peto test weights the jth failure time by the survival estimate$\hat{s}(t_j)$; it gives more weight for the earlier survival time as well. The Fleming-Harrington test has the most flexibility in terms of the choice of weights in the test. We can choose the values for parameters p and q. For example, if p=1 and q=0, then $w(t) = \hat{s}(t_{j-1})^P[1 - \hat{s}(t_{j-1})]^0 = \hat{s}((t_{j-1})$ giving more weight for the earlier survival times when $\hat{s}(t_{j-1})$ is close to one. In contrast, if p=0 and q=1, then $w(t) = \hat{s}(t_{j-1})^0[1 - \hat{s}\big(t_{j-1}\big)]^1 = 1 - \hat{s}((t_{j-1})$, I giving more weight for the later survival times when $\hat{s}(t_{j-1})$. If p=0 and q=0, then $w(t) = \hat{s}(t_{j-1})^0[1 - \hat{s}(t_{j-1})]^0 = 1$; thus, it reduces to the log-rank test. Selecting the appropriate weighted test according to the property of the abnormality can provide more accurate analysis.

Test Statistic	$w(t_j)$
Log-rank	1
Wilcoxon	n_j
Tarone-Ware	$\sqrt{n_j}$
Peto	$\tilde{s}(t_j)$
Flemington-Harrington	$\hat{s}(t_{j-1})^P[1-\hat{s}(t_{j-1})]^q$

Table 2-4. Weights in alternative test statistics.

Equation 2-6: test statistic= $\frac{(\sum_j w(t_j)(m_{ij}-e_{ij}))^2}{var(\sum_j w(t_j)(m_{ij}-e_{ij}))}$ i=1,2 represents the group number, j is the order of failure time, $w(t_j)$ weight at jth failure time.

From our previous study, we had the average number of abnormalities associated with each rCNAs calculated to model the progression of rCNAs with FL development[31]. We can estimate the temporal order of the rCNA in the disease from this calculation. According to the published study, samples with abnormalities that occur in the early stage of FL are expected to have generally a low number of abnormalities, whereas samples with abnormalities that occur late in the disease are expected to have a high number of abnormalities. Therefore, early rCNAs would be expected to have a lower average number of other abnormalities and we should select a test statistic that puts more weight for the earlier survival time. In contrast, test statistics that put more weight on the later survival time could be applied to rCNAs that tended to occur late.

G.5. Applying survival analysis by SAS

The SAS programming language is specifically designed for statistical analysis. It can also retrieve and manipulate data from various sources. We applied the KM method and the alternative tests by using SAS and generated a comprehensive result including survival estimates

in different failure times with the survival standard error, number of failure patients and number of risk patients, summary statistics for our time variable, a summary of censoring, survival curves with confident intervals, marked time and survival probability of censored patients, and multiple weighted alternative tests.

CHAPTER III

INITIAL RESULTS AND DISCUSSION

A. WES analysis

A.1. Introduction

To identify the genetic changes that drive the transformation and contribute to its biological and clinical behavior, we sequenced 12 pairs of FL and tFL cases and applied a WES pipeline specifically designed for paired FL and tFL samples to identify mutations in the dataset. We then evaluated the performance of the WES by calculating the mapping condition, percentage of duplicates, coverage, and depth in annotated coding regions and splice sites of the samples. To confirm the reliability and capability of the WES mutation detection pipeline, Sanger sequencing was applied to validate the accuracy of the variant identification.

A.2. Sequencing performance

There were 12 pairs of FL and tFL samples sequenced by WES. In general, the WES had very good performance (Table 3-1). Most of the reads were pair mapped (range: 96.23%-98.85%, SD: 0.69%, Figure 3-1). The average percentage of duplicates was 34.41% (range: 19.44%-56.91%, SD: 11.57%, Figure 3-2). The average depth of bases at coding regions was 90 (range: 30-199, SD: 37, Figure 3-3). The average depth of bases at splice sites was 73 (range: 26-180, SD: 32, Figure 3-6). The average coverage of coding regions was 94.44% (range: 90.96%-99.67%, SD: 2.09%, Figure 3-4). The average coverage of coding region with at least 10 reads was 87.64% (range: 75.71%-94.85%, SD: 5.34%, Figure 3-4). The average coverage of splice sites was 96.81% (range: 94.50%-99.85%, SD: 1.36%, Figure 3-4).

Cases	Total Reads	Duplicates	%Duplicates	Pair-Mapped Reads	Pair-Mapped Bases
FL-1	237,590,222	74,826,008	31.49%	233,505,162	23,584,021,362
tFL-1	266,693,964	71,921,285	26.97%	256,652,884	25,921,941,284
FL-2	167,722,410	95,443,258	56.91%	163,370,368	16,500,407,168
tFL-2	151,505,060	81,911,778	54.07%	147,515,422	14,899,057,622
FL-3	173,523,760	71,057,500	40.95%	170,494,962	17,219,991,162
tFL-3	102,426,766	48,077,593	46.94%	100,197,678	10,119,965,478
FL-4	78,691,960	38,993,802	49.55%	76,703,946	7,747,098,546
tFL-4	160,164,134	68,724,439	42.91%	157,400,100	15,897,410,100
FL-5	95,405,906	42,166,753	44.20%	93,430,749	9,436,505,649
tFL-5	112,526,146	57,537,049	51.13%	110,389,686	11,149,358,286
FL-6	140,051,736	49,327,829	35.22%	137,578,938	13,895,472,738
tFL-6	76,965,874	26,150,440	33.98%	74,417,132	7,516,130,332
FL-8	138,771,418	45,034,843	32.45%	136,867,478	13,823,615,278
tFL-8	161,157,660	48,874,483	30.33%	158,777,402	16,036,517,602
FL-9	153,123,582	30,802,959	20.12%	151,108,590	15,261,967,590
tFL-9	145,388,704	28,812,740	19.82%	143,571,681	14,500,739,781
FL-10	145,153,806	29,265,025	20.16%	143,360,954	14,479,456,354
tFL-10	106,939,220	22,798,673	21.32%	105,289,995	10,634,289,495
FL-11	141,973,378	27,596,351	19.44%	140,231,461	14,163,377,561
tFL-11	125,551,362	26,238,424	20.90%	123,869,734	12,510,843,134
FL-12	172,894,910	54,648,179	31.61%	170,898,776	17,260,776,376
tFL-12	182,211,848	58,432,593	32.07%	180,015,195	18,181,534,695
FL-22	144,067,832	45,448,865	31.55%	141,669,627	14,308,632,327
tFL-22	144,306,904	45,991,619	31.87%	142,075,691	14,349,644,791
max	266,693,964	95,443,258	56.91%	256,652,884	25,921,941,284
min	76,965,874	22,798,673	19.44%	74,417,132	7,516,130,332
mean	146,867,023	49,586,770	34.41%	144,141,400	14,558,281,446
median	144,730,355	47,034,606	31.97%	142,718,323	14,414,550,573
SD	43560873	19772323	11.57%	42352468	4277599295
Cases	**CDS covered**	**CDS 10X covered**	**CDS average DP**	**Splice covered**	**Splice average DP**
FL-1	98.74%	94.85%	199	99.31%	180
tFL-1	99.67%	94.04%	155	99.85%	138
FL-2	93.30%	82.78%	79	95.55%	65
tFL-2	93.58%	83.47%	75	96.03%	61
FL-3	93.80%	88.25%	92	96.70%	77

tFL-3	92.40%	81.74%	45	95.45%	39
FL-4	90.96%	75.71%	30	94.71%	26
tFL-4	93.74%	86.92%	75	96.34%	64
FL-5	92.11%	79.83%	44	95.26%	37
tFL-5	93.22%	83.81%	47	95.95%	41
FL-6	93.30%	86.50%	86	95.93%	73
tFL-6	91.40%	78.70%	43	94.50%	37
FL-8	92.73%	84.63%	86	95.78%	69
tFL-8	92.67%	84.88%	96	95.80%	76
FL-9	95.67%	92.11%	112	97.75%	87
tFL-9	95.44%	91.36%	95	97.67%	74
FL-10	95.79%	92.38%	123	97.84%	95
tFL-10	95.38%	90.89%	83	97.31%	64
FL-11	95.72%	92.11%	109	97.80%	86
tFL-11	95.70%	91.37%	92	97.70%	73
FL-12	95.59%	92.25%	107	97.64%	78
tFL-12	95.26%	92.10%	108	97.52%	79
FL-22	95.09%	91.35%	95	97.46%	70
tFL-22	95.21%	91.30%	94	97.51%	70
max	99.67%	94.85%	199	99.85%	180
min	90.96%	75.71%	30	94.50%	26
mean	94.44%	87.64%	90	96.81%	73
median	94.44%	89.57%	92	97.00%	72
SD	2.09%	5.34%	37	1.36%	32

Table 3-1. General statistical information of 12 pairs of FL and tFL WES sequenced samples.

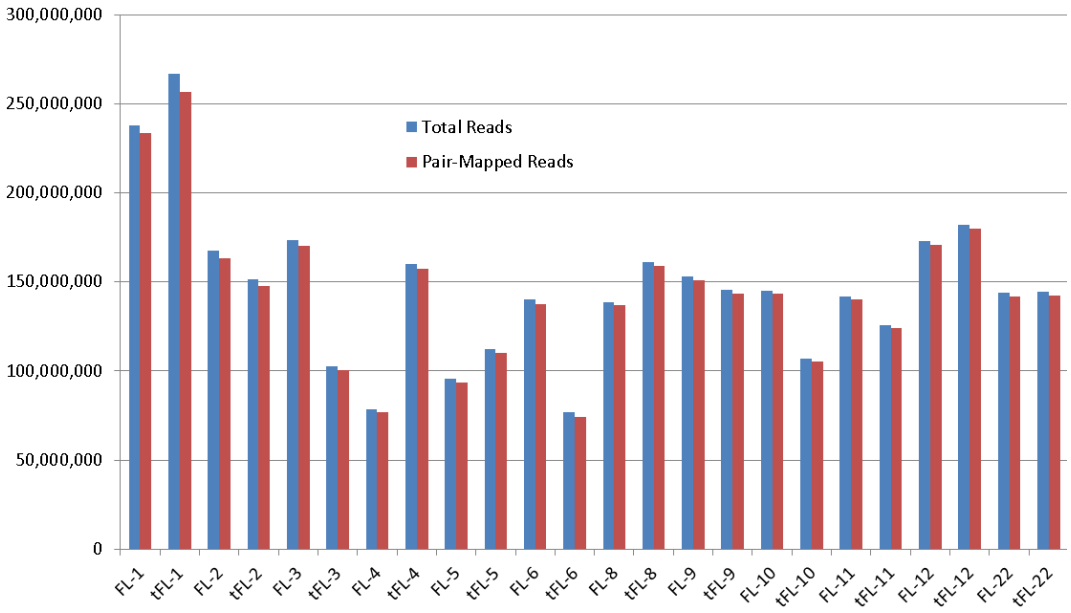

Figure 3-1. Numbers of reads in 12 pairs of FL and tFL WES sequenced samples.

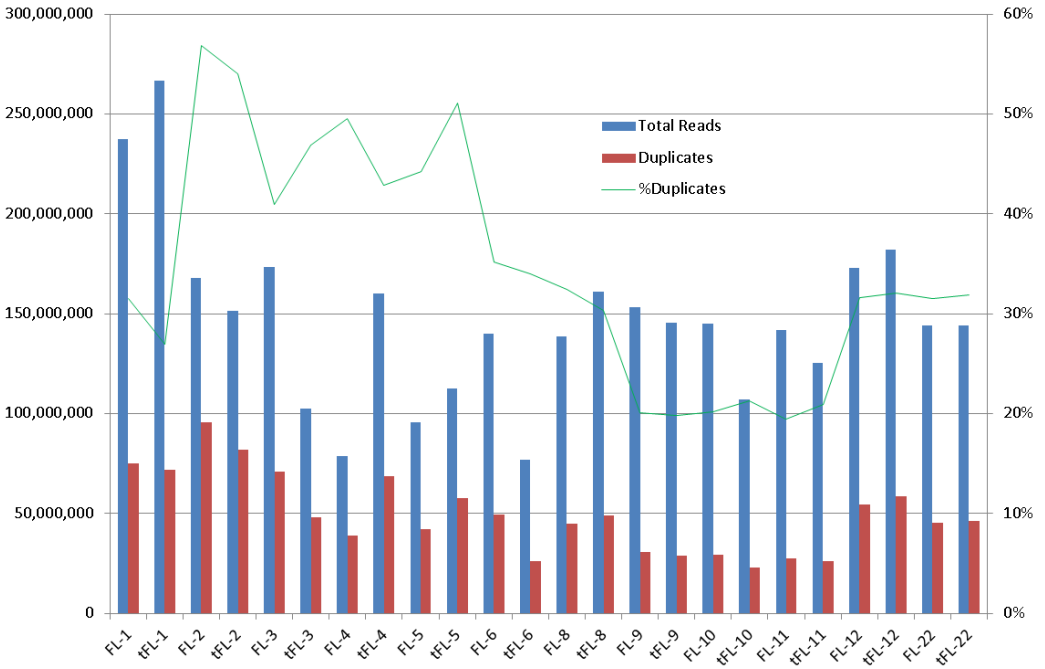

Figure 3-2. Percentage of duplicates in 12 pairs of FL and tFL WES sequenced samples.

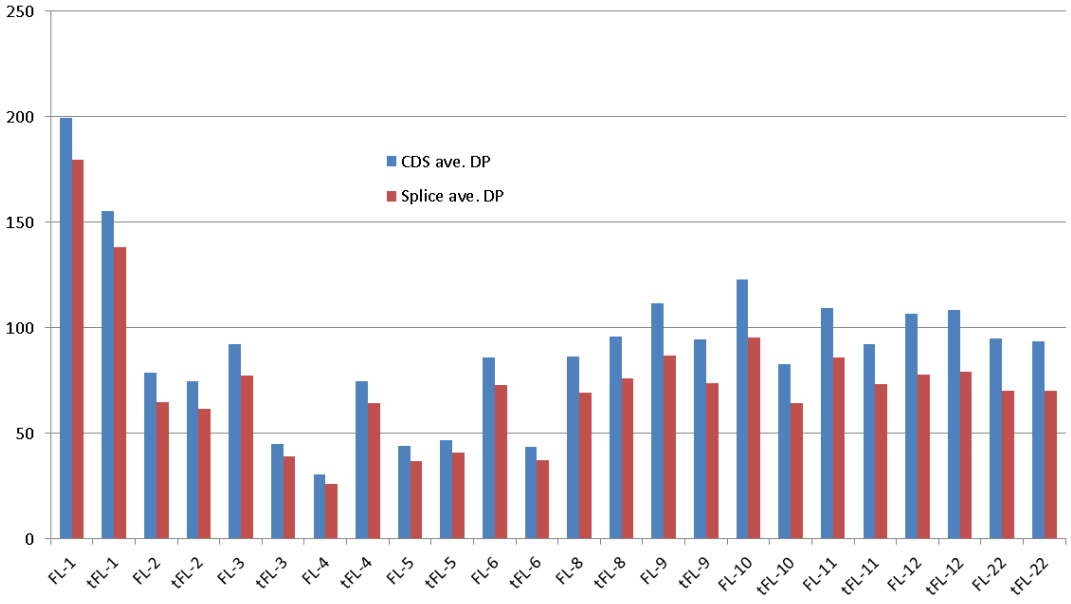

Figure 3-3. Read depth at coding regions and splice sites for 12 pairs of FL and tFL WES sequenced samples.

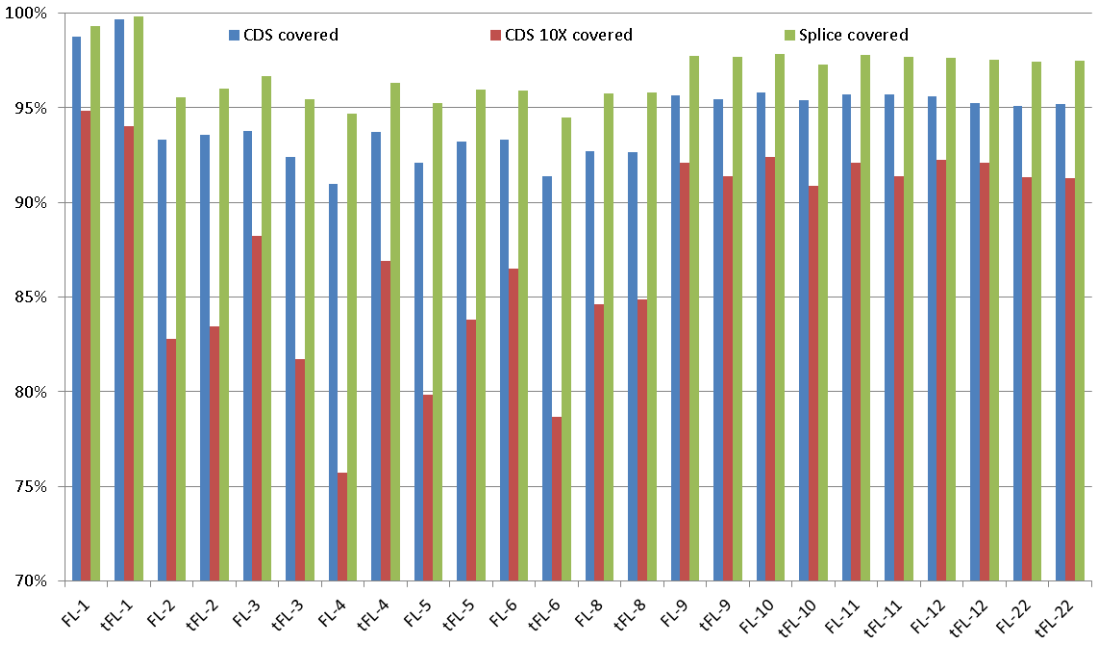

Figure 3-4. Coverage of coding regions and splice sites in pairs of FL and tFL WES sequenced samples.

A.3. Validation of the WES mutation detection pipeline

Sanger sequencing was applied to evaluate the capability of variant identification in the WES pipeline. 50 variants were selected from a confident variant list in Table 3-2. 47 out of 47 variants were confirmed, and the primers in the other 3 variants did not work.

Gene	aa. change	Chr	Pos	Ref	Var	Case	Type	Confirmed by sanger?
ADCY9	C->S	chr16	4164141	A	T	5-tFL	snp	YES
ANKRD12	S->T	chr18	9255822	T	A	5-tFL	snp	YES
CLSTN3	A->V	chr12	7287964	C	T	5-tFL	snp	YES
HMGN5	T->R	chrX	80371830	G	C	5-tFL	snp	YES
IL4R	Y->N	chr16	27363936	T	A	5-tFL	snp	YES
LTBP1	P->R	chr2	33623527	C	G	5-tFL	snp	YES
OR13H1	-	chrX	130678922	G	-A	5-tFL	indel	YES
PTGES	V->M	chr9	132502006	C	T	5-tFL	snp	YES
ROBO1	L->F	chr3	78987869	T	A	5-tFL	snp	YES
SDK1	D->A	chr7	4011182	A	C	5-tFL	snp	YES
TP53	D->V	chr17	7577598	T	A	5-tFL	snp	YES
TP53	M->V	chr17	7577545	T	C	5-tFL	snp	YES
ZNF302	V->L	chr19	35175831	G	C	5-tFL	snp	YES
ZNF594	K->R	chr17	5085165	T	C	5-tFL	snp	YES
CDH23	N->H	chr10	73544792	A	C	6-tFL	snp	YES
FUBP1	K->E	chr1	78435630	T	C	6-tFL	snp	YES
GRIP1	D->H	chr12	66786109	C	G	6-tFL	snp	YES
MYST1	H->Q	chr16	31142184	C	G	6-tFL	snp	YES
SLC7A14	W->L	chr3	170218933	C	A	6-tFL	snp	YES
SPEN	-	chr1	16256410	C	+A	6-tFL	indel	YES
TATDN3	Q->P	chr1	212977991	A	C	6-tFL	snp	YES
TNFAIP3	Q->E	chr6	138200146	C	G	6-tFL	snp	YES
ABCA12	-	chr2	215901791	T	A	9-tFL	snp	YES
AKAP13	-	chr15	86123376	A	-CCACAG	9-tFL	indel	YES
BCL9L	G->D	chr11	118779065	C	T	9-tFL	snp	YES
CDH7	T->M	chr18	63547760	C	T	9-tFL	snp	YES
LOC643677	Q->H	chr13	103393996	C	A	9-tFL	snp	YES
PGLYRP3	V->A	chr1	153274966	A	G	9-tFL	snp	YES
PIK3R1	P->T	chr5	67576825	C	A	9-tFL	snp	YES
SLITRK5	D->E	chr13	88330325	C	A	9-tFL	snp	YES
SPATA20	G->R	chr17	48628075	G	C	9-tFL	snp	YES
WNT3A	V->G	chr1	228210442	T	G	9-tFL	snp	YES
TBP	-	chr6	170871046	A	-	7-	indel	YES

					CAGCAG CAG	tFL*		
RRAGC	I->F	chr1	39322649	A	G	6-FL	snp	YES
RRAGC	I->F	chr1	39322649	A	G	6-tFL	snp	YES
RRAGC	I->F	chr1	39322697	T	A	8-FL	snp	YES
RRAGC	I->F	chr1	39322697	T	A	8-tFL	snp	YES
TPR	R->Q	chr1	186291702	C	T	4-tFL	snp	YES
SLC35F5	I->T	chr2	114493380	A	G	4-tFL	snp	YES
SLC35F5	A->T	chr2	114493426	C	T	4-tFL	snp	YES
MIR142	V->A	chr17	56408621	A	G	4-tFL	snp	YES
MIR142	V->A	chr17	56408621	A	G	8-FL	snp	YES
MIR142	V->A	chr17	56408621	A	G	8-tFL	snp	YES
MIR142	S->G	chr17	56408625	T	C	9-tFL	snp	YES
MIR142	V->A	chr17	56408657	A	G	11-FL	snp	YES
MIR142	V->A	chr17	56408657	A	G	11-tFL	snp	YES
TMSB4X	-	chrX	12994363	A	G	4-tFL	snp	YES
DDX3X	R->S	chrX	41205842	C	A	6-tFL	snp	primer failed
MGAT4B	S->T	chr5	179226060	C	G	6-tFL	snp	primer failed
MYO16	-	chr13	109704824	G	T	9-tFL	snp	primer failed

Table 3-2. Sanger sequencing validation in of 50 variants found in WES sequenced samples.
Note: * sample removed from the final analysis.

A.4. Mutations identified by WES

We identified a total of 1191 mutations (in 666 different genes) in the 12 paired WES dataset (Table 3-3). 838 mutations (in 363 different genes) were detected in both paired FL and tFL samples (shared mutations). 446 out of the 838 mutations were in genes expressed in B cells (in 187 different genes). 114 were in the confident final list (28 different genes). 90 (89 different genes) were only detected in FL samples (FL-unique mutations). 51 out of the 90 mutations were in genes expressed in B cells (50 different genes). 13 were in the confident final list (12 different genes). 263 (249 different genes) were only detected in tFL samples (tFL-unique mutations). 135 out of the 263 mutations were in genes expressed in B cells (121 different

genes), 31 were in the confident final list (23 different genes). The recurrent mutations identified in our dataset were highly concordant with previously published sequencing reports, affirming the effectiveness of our analysis pipeline.

Type	Sequence type	Total	T1+T2+TR +express	T1	T2	TR	T1+express +recurrent	T1+T2+TR+ express+re current
share	WES	838	446	178	252	16	98	114
FL-unique	WES	90	51	46	5	0	13	13
tFL-unique	WES	263	135	119	15	1	27	31
	sum	**1191**	**632**	**343**	**272**	**17**	**138**	**158**

Table 3-3. Number of mutations detected by WES in different mutation types and filtering.

A.5. Discussion

We were particularly interested in mutations that contributed to transformation. Therefore, we separated the mutations into FL-unique mutations, tFL-unique mutations, and shared mutations so that we can assess if some of them occurred more frequently in one of the groups. If the mutations occur more often in the tFL-unique group, we considered them to be more likely to be related to transformation. Similarly, if the mutations occur more often in the FL-unique group, we considered them to be more likely contributing to FL development. When we had sequenced a decent number of samples, we noticed that several mutations of a gene can be found in one sample. For example, BCL2 mutations were detected as a FL-unique mutation and also a shared mutation in the exactly same sample but at different positions. This could indicate mutations of both alleles of BCL2 or there were subclones harboring different mutations. These mutations were very carefully noted for further integration.

B. Custom capture panel analysis

B.1. Introduction

We sequenced 7 pairs of FL and tFL samples, 4 triples of two FL samples and a tFL sample, and 15 single tFL samples, and applied various pipelines to analyze the sequencing data as described above. To confirm the reliability and capability of the custom capture panel mutation detection, we compared the variants detected using the custom capture panel with the variants detected by WES on three paired samples analyzed by both approaches. We also checked the known mutations in a sequenced cell line. After the mutation detection pipelines and custom capture panel were proved to be effective, a simulation was applied to the data from the reliable variants to confirm the approach for estimating the duplicate ratio (the average number of reads per initial molecule). We then estimated the duplicate ratio in samples sequenced by custom capture panel.

B.2. Sequencing performance

A total of 43 samples (3 paired samples were also sequenced by WES) were sequenced using the custom capture panel. In general, the custom capture panel had very good performance (Table 3-4). Most of the reads were pair mapped (range: 82.41%-91.21%, SD: 1.94%, Figure 3-5). The average depth at coding regions was 982 (range: 491-1515, SD: 224, Figure 3-6) and the average depth at splice sites was 849 (range: 426-1317, SD: 194, Figure 3-6), both of which are much deeper than WES. The average coverage at coding regions was 95.31% (range: 88.25%-99.77%, SD: 1.45%, Figure 3-7), and the average coverage at splice sites was 89.60% (range: 83.16%-90.13%, SD: 1.01%, Figure 3-7).

Cases	Total Reads	Pair-Mapped Reads	Pair-Mapped Bases	%Pair-Mapped Reads	CDS ave DP	Splice ave DP	CDS cov	Splice cov
FL-3	32,115,446	29,290,422	2,958,332,622	91.20 %	786	683	95.43 %	89.82 %
tFL-3	31,611,696	28,753,286	2,904,081,886	90.96 %	757	655	95.43 %	89.91 %
FL-5	39,545,852	35,684,534	3,604,137,934	90.24 %	1237	830	99.77 %	90.06 %
tFL-5	42,809,742	38,735,308	3,912,266,108	90.48 %	975	853	95.76 %	90.13 %
tFL-15	33,735,752	30,695,784	3,100,274,184	90.99 %	814	708	95.56 %	89.94 %
tFL-16	40,419,724	36,422,198	3,678,641,998	90.11 %	930	813	95.66 %	90.09 %
tFL-17	38,825,300	34,737,906	3,508,528,506	89.47 %	932	839	88.25 %	83.16 %
tFL-18	33,919,492	30,732,778	3,104,010,578	90.61 %	843	724	95.37 %	89.92 %
tFL-19	37,800,608	34,477,870	3,482,264,870	91.21 %	994	813	99.70 %	89.75 %
tFL-20	42,563,636	38,736,302	3,912,366,502	91.01 %	1048	915	95.35 %	89.86 %
tFL-21	31,746,862	28,752,256	2,903,977,856	90.57 %	773	680	95.34 %	89.87 %
KARPAS-422	40,802,670	36,187,510	3,654,938,510	88.69 %	952	841	95.50 %	89.92 %
FL-23-1	38,999,680	35,230,592	3,558,289,792	90.34 %	1000	837	95.52 %	89.90 %
FL-23-2	40,960,006	36,782,770	3,715,059,770	89.80 %	1017	858	95.33 %	89.83 %
tFL-23	36,783,674	33,217,988	3,355,016,788	90.31 %	941	799	95.37 %	89.86 %
FL-24-1	25,211,878	22,258,334	2,248,091,734	88.29 %	603	536	95.00 %	89.66 %
FL-24-2	36,757,574	32,288,728	3,261,161,528	87.84 %	873	778	95.18 %	89.73 %
tFL-24	42,562,004	37,436,676	3,781,104,276	87.96 %	1012	885	95.40 %	89.92 %
FL-25-1	32,177,034	28,231,026	2,851,333,626	87.74 %	784	660	95.21 %	89.70 %
FL-25-2	29,521,614	24,330,136	2,457,343,736	82.41 %	605	529	95.59 %	89.99 %
tFL-25	39,425,936	32,595,856	3,292,181,456	82.68 %	750	667	95.70 %	89.97 %
FL-26-1	30,333,532	26,278,374	2,654,115,774	86.63	708	605	95.23	89.82

				%			%	%
FL-26-2	32,761,944	28,522,064	2,880,728,464	87.06 %	821	690	95.29 %	89.79 %
tFL-26	37,275,886	32,372,154	3,269,587,554	86.84 %	880	753	95.42 %	89.94 %
FL-27	47,299,092	42,250,582	4,267,308,782	89.33 %	1146	1021	95.25 %	89.59 %
tFL-27	51,565,566	46,014,100	4,647,424,100	89.23 %	1214	1086	95.24 %	89.65 %
FL-28	44,351,228	40,003,606	4,040,364,206	90.20 %	1049	930	95.09 %	89.54 %
tFL-28	46,206,520	40,649,858	4,105,635,658	87.97 %	1131	986	95.21 %	89.58 %
FL-29	48,257,298	42,504,248	4,292,929,048	88.08 %	1216	1044	95.13 %	89.63 %
tFL-29	28,674,836	25,401,852	2,565,587,052	88.59 %	700	613	94.99 %	89.48 %
FL-30	51,330,246	45,304,466	4,575,751,066	88.26 %	1266	1109	95.18 %	89.57 %
tFL-30	40,658,460	36,030,640	3,639,094,640	88.62 %	1003	869	95.06 %	89.44 %
FL-31	33,138,820	29,408,576	2,970,266,176	88.74 %	818	709	95.03 %	89.52 %
tFL-31	20,070,002	17,797,300	1,797,527,300	88.68 %	491	426	94.85 %	89.38 %
tFL-32	54,421,832	46,090,264	4,655,116,664	84.69 %	1301	1119	95.00 %	89.53 %
tFL-33	39,634,324	35,351,892	3,570,541,092	89.20 %	978	848	95.01 %	89.59 %
tFL-34	53,733,670	48,043,530	4,852,396,530	89.41 %	1319	1159	95.22 %	89.61 %
tFL-35	52,127,122	46,398,420	4,686,240,420	89.01 %	1314	1134	95.21 %	89.72 %
tFL-36	61,553,502	54,997,112	5,554,708,312	89.35 %	1515	1317	95.19 %	89.69 %
tFL-37	48,163,280	43,010,422	4,344,052,622	89.30 %	1192	1026	95.19 %	89.71 %
tFL-38	50,147,304	45,078,500	4,552,928,500	89.89 %	1208	1061	95.25 %	89.68 %
tFL-39	43,741,788	38,832,258	3,922,058,058	88.78 %	1056	946	95.09 %	89.57 %
tFL-40	52,922,576	47,175,804	4,764,756,204	89.14 %	1281	1142	95.18 %	89.72 %
max	61,553,502	54,997,112	5,554,708,312	91.21 %	1515	1317	99.77 %	90.13 %
min	20,070,002	17,797,300	1,797,527,300	82.41 %	491	426	88.25 %	83.16 %

mean	40,384,312	35,883,170	3,624,200,156	88.84 %	982	849	95.31 %	89.60 %
median	39,939,521	35,724,863	3,608,211,163	89.17 %	979	840	95.23 %	89.72 %
SD	8,677,316	7,751,512	782,902,664	1.94%	224	194	1.45 %	1.01%

Table 3-4. General statistical information for custom capture panel sequenced samples.

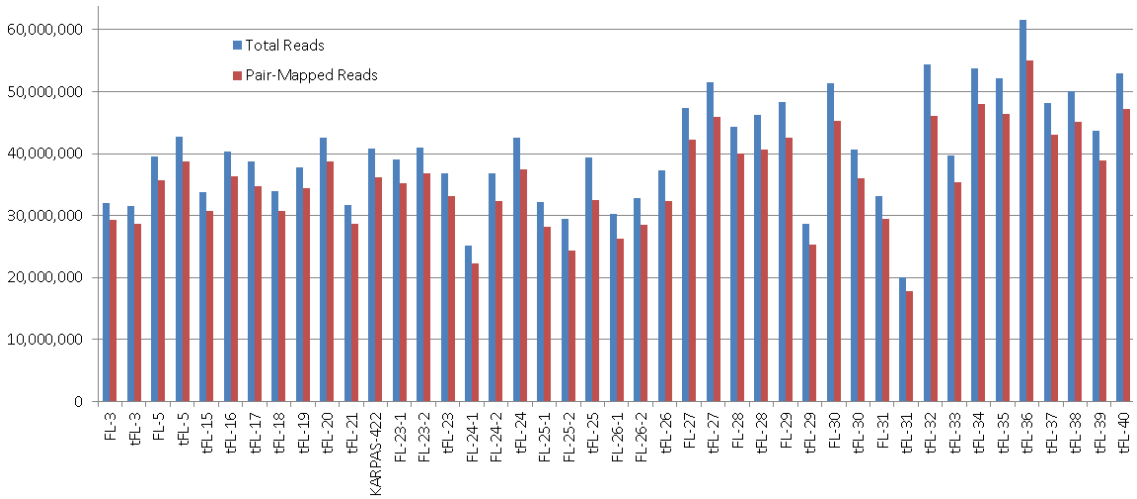

Figure 3-5. Numbers of reads in 43 FL/tFL custom capture panel sequenced samples.

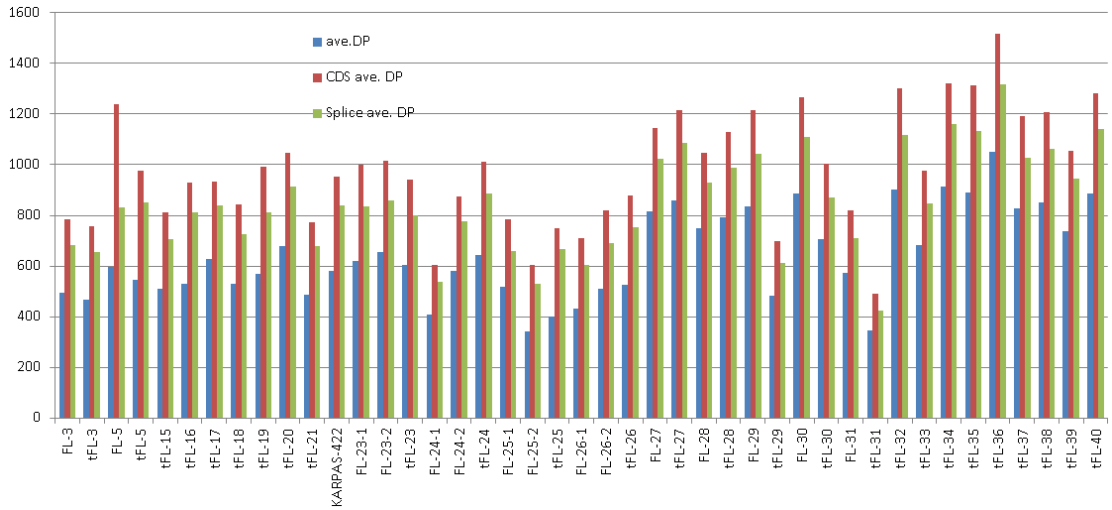

Figure 3-6. Depth of coverage for coding regions and splice sites in 43 FL/tFL custom capture panel sequenced samples.

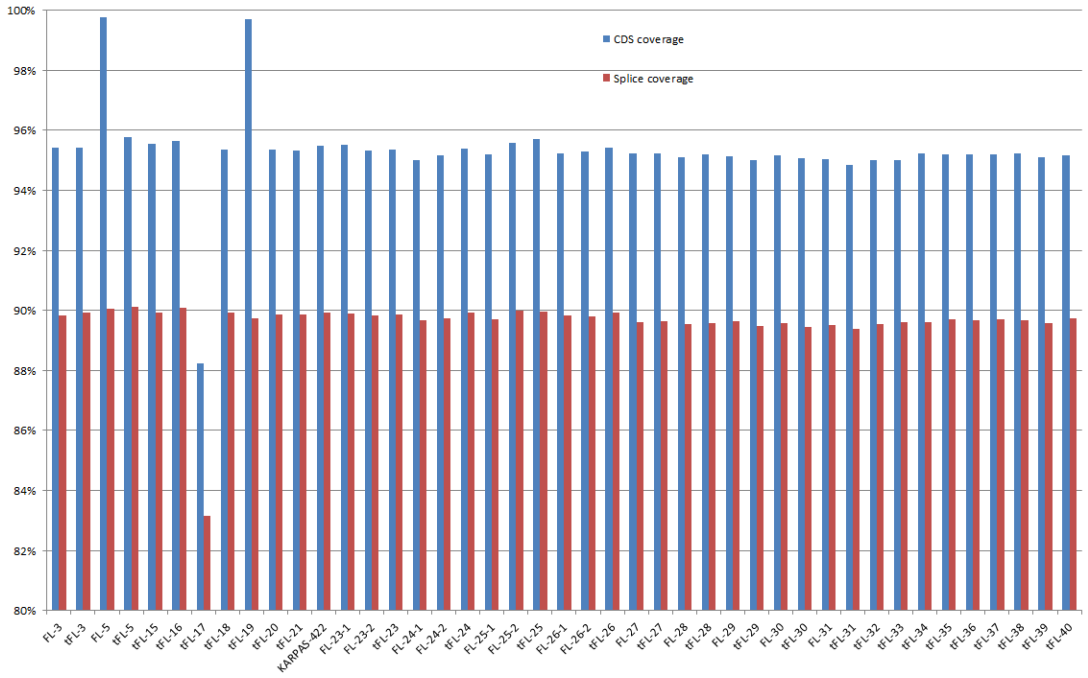

Figure 3-7. Coverage of coding regions and splice sites in 43 FL/tFL custom capture panel sequenced samples.

B.3. Custom capture panel mutation detection pipelines validation

We sequenced 3 of the paired cases for which we had WES data. After that we determined the number of overlapping variants with sufficient depth in both platforms. We confirmed that more than 98% of the variants that were detected by WES were successfully detected in the custom capture panel in all 3 paired cases (Table 3-5). We also reported all the known mutations in KARPAS-422. We then calculated the depth for mutations (mutations in Tier 1 and 2 in Table 3-6) found in both platforms and compared their variant frequency. We confirmed that the mutation depth in custom capture panel was much deeper than WES and the majority of variant frequencies between the two platforms were similar (Table 3-6).

Case	Variants Found in WES	Sufficient Custom Depth	Validated	Validate Rate
FL-3	109,803	2,788	2,742	98.40%
tFL-3	98,423	2,522	2,485	98.50%
FL-5	96,326	2,544	2,492	98.00%
tFL-5	99,221	2,789	2,742	98.30%
FL-10	112,456	2,898	2,845	98.20%
tFL-10	104,922	2,703	2,650	98.00%
Total	**621,151**	**16,244**	**15,956**	

Table 3-5. Variant overlap for 3 paired cases sequenced by both WES and custom capture panel. Sufficient custom depth indicates that the mutations found in WES also had sufficient depth in custom capture panel for comparison.

Case	Chromosome| Position |Altered-Base	Gene	WES data			Custom-seq data		
			Var Read Depth	Total Read Depth	Var Freq	Var Read Depth	Total Read Depth	Var Freq
FL-10	chr14|23345905|C	LRP10	40	96	42%	547	1765	31%
FL-10	chr1|85736474|T	BCL10	125	193	65%	17	245	7%
FL-3	chr1|110882568|G	RBM15	12	21	57%	35	160	22%
FL-5	chrX|41200829|C	DDX3X	11	31	36%	35	81	43%
tFL-10	chr4|40245479|A	RHOH	72	146	49%	1740	3550	49%
tFL-10	chr16|3786704|G	CREBBP	24	29	83%	347	456	76%
tFL-10	chr6|157528317|A	ARID1B	46	89	52%	628	1653	38%
tFL-10	chr3|187449624|A	BCL6	20	58	35%	440	1047	42%
tFL-10	chr12|49433524|-CT	KMT2D(MLL2)	50	62	81%	203	244	83%
tFL-10	chr17|80209332|A	CSNK1D	57	153	37%	522	1213	43%
tFL-10	chr13|31037742|A	HMGB1	57	136	42%	44	295	15%
tFL-10	chr21|44521518|A	U2AF1	38	102	37%	28	172	16%
tFL-3	chr17|62007128|T	CD79B	6	30	20%	447	2236	20%
tFL-3	chr4|3134324|A	HTT	15	61	25%	60	334	18%
tFL-3	chr15|42035019|A	MGA	21	39	54%	101	146	69%
tFL-5	chr16|3807881|A	CREBBP	18	34	53%	658	1646	40%
tFL-5	chr12|49438694|T	KMT2D(MLL2)	17	32	53%	683	1484	46%
tFL-5	chr12|49418731|A	KMT2D(MLL2)	17	43	40%	546	1162	47%
tFL-5	chr7|148508728|T	EZH2	20	43	47%	317	857	37%
tFL-5	chr17|7577598|A	TP53	14	25	56%	3	28	11%
tFL-5	chr12|7287964|T	CLSTN3	26	60	43%	1021	2430	42%
tFL-5	chr16|3808033|G	CREBBP	6	18	33%	451	1074	42%

tFL-5	chr16|3807917|C	CREBBP	25	47	53%	999	2379	42%
tFL-5	chr12|57493137|C	STAT6	50	57	88%	654	909	72%
tFL-5	chr17|78269544|C	RNF213	13	26	50%	297	874	34%
tFL-5	chr16|3808046|G	CREBBP	7	16	44%	451	1074	42%
tFL-5	chr16|85936784|G	IRF8	18	32	56%	328	820	40%
tFL-5	chrX|39923127|G	BCOR	9	19	47%	259	740	35%
tFL-5	chr3|78987869|A	ROBO1	24	52	46%	609	1449	42%
tFL-5	chr16|3808030|C	CREBBP	6	19	32%	451	1074	42%
tFL-5	chr12|57493818|C	STAT6	51	60	85%	836	1114	75%
tFL-5	chr17|7577545|C	TP53	14	37	38%	5	6	83%

Table 3-6. Comparison of depth in genes found mutated by WES and custom gene sequencing.

B.4. Mutations identified by the custom capture panel

We identified a total of 484 mutations (134 different genes) in the custom capture panel dataset (Table 3-7). 169 were detected in single samples (74 different genes) and 80 were in the confident final list (42 different genes). 156 (in 61 different genes) were detected in both paired FL and tFL samples (shared mutations) and 66 were in the confident final list (27 different genes). 19 (in 17 different genes) were only detected in FL samples (FL-unique mutations) and 8 were in the confident final list (8 different genes). 140 (in 84 different genes) were only detected in tFL samples (tFL-unique mutations) and 68 were in the confident final list (40 different genes).

Type	Sequence type	Total	T1+T2+TR+express	T1+express	T2+express	TR+express	T1+express+recurrent	T1+T2+TR+express+recurrent
single	custom	169	157	91	40	26	80	129
share	custom	156	143	94	40	9	66	95
FL-unique	custom	19	17	14	2	1	8	10
tFL-unique	custom	140	130	98	27	5	68	88
	sum	**484**	**447**	**297**	**109**	**41**	**222**	**322**

Table 3-7. Number of mutations detected by custom capture panel in different mutation types and filtering.

B.5. Validation of a binomial distribution model

Because most of our tumor samples have a considerable admixture of normal cells, the true variant frequency for somatic mutations is expected to be less than 0.5, but the frequency for germline variants should equal 0.5. Thus, we wanted to be able to use a proportion test to determine the likelihood that a fraction of variant counts was significantly different from 0.5. We performed the proportion test on the data from the custom platform on heterozygous germline variants (SNPs). For almost all heterozygous germline variants, approximately 50% of reads should support each allele. Using dbSNP, a compendium of germline variants we can extract heterozygous SNPs from the variants called using the mutation analysis pipeline. Unfortunately, when we performed the proportion test on these known SNPs, many more than expected showed a significant difference from 0.5. One possible explanation is that multiple reads were derived from a single initial molecule due to the PCR step; the average number of reads per initial molecule was designated “the duplicate ratio.” If there are no duplicates (duplicate ratio =1), p-values from the binomial test for each heterozygous germline variant will be uniformly distributed, and the slope of the regression model fitting the frequencies of the p-

values to the p-value will be close to 0; however, if there are PCR-generated duplicate reads, p-values close to zero will be overrepresented. We designed a binomial distribution statistical model to estimate the duplicate ratio. We estimated the duplicate ratio based on the slope of the linear regression model of p-values generated from binomial tests and frequencies of the p-values.

To validate the model, we extracted all the heterozygous germline variants from the mutation detection pipelines and did the duplicate ratio simulation based on the extracted real total reads to display the patterns of p-values and frequencies of the p-values with and without duplicates. Figure 3-8 displayed the patterns of no duplicates, duplicate ratio equal to 2 and duplicate ratio equal to 4. The number of total reads applied in Figure 3-8 was randomly selected from the real total reads of all the samples (range, 100-1200). According to the rationale for the binomial distribution model, if there are no duplicates, most of the null hypotheses, the probability that a read supports one allele is 0.5, will not be rejected, and the majority of total reads are assigned equally. That is why there was a big peak on the far right, whereas the rest of p-values were uniformly distributed in Figure 3-8 (No duplicates). If there are PCR-generated duplicates, most of the hypotheses will be rejected, implying that at the majority of sites, the total reads are assigned unequally, and p-values close to zero will be overrepresented. That is why there was a big peak on the left as displayed in Figure 3-8 (duplicate*2 and duplicate*4). As the PCR-duplicate ratio increases, the peak close to zero becomes more exaggerated; therefore, when the PCR-duplicate ratio was 2, the frequency of p-values close to 0 was a little bit over 1500 and it increased to over 3000 when the PCR-duplicate ratio reached 4 in Figure 3-8. From the results of the 10,000,000 simulations we did based on a real number of reads from all of our samples with duplicate ratio equal to 1, we can see the trend of the histogram was flat except the p-values close to 1 in Figure 3-9 as we expected,

which confirmed that the p-values from binomial test will be uniformly distributed if there are no duplicates.

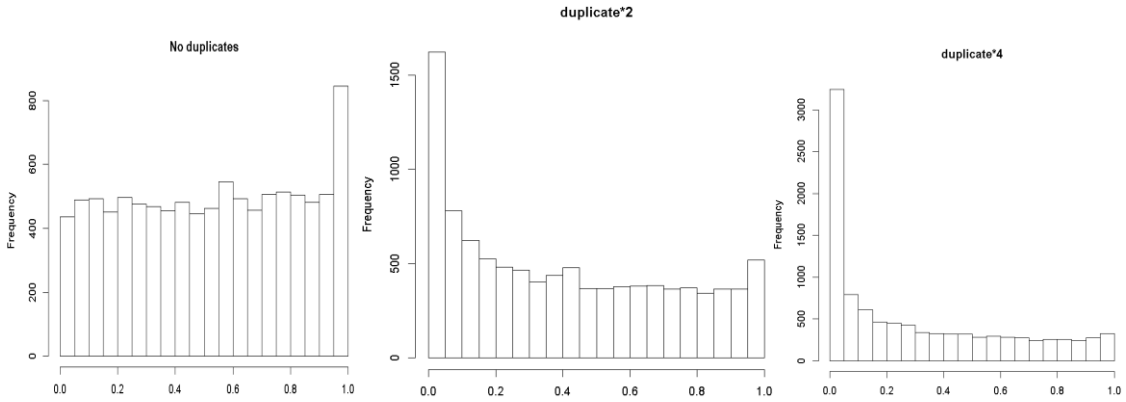

Figure 3-8. Patterns of different duplicate ratio.

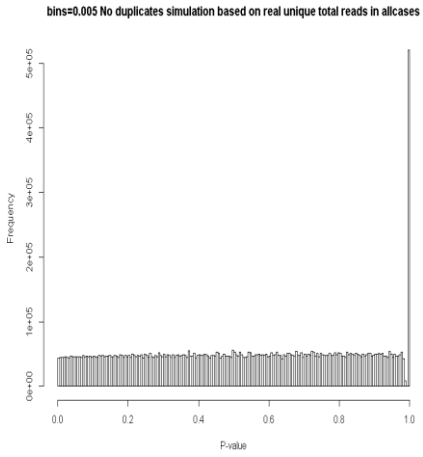

Figure 3-9. Pattern when the duplicate ratio is equal to 1.

B.6. Duplicate ratio estimation

We applied the Binomial distribution model and estimated the duplicate ratio for all the samples sequenced using the custom capture panel (Table 3-8). The average duplicate ratio was 3.3. We then adjusted the reads in the 3 paired cases for which we had WES data and custom capture panel with predicted duplicate ratio. With a few exceptions, the read depth in the data from the custom capture panel was still much deeper than from WES in genes found mutated in both platforms (Table 3-9).

Case	Duplicate ratio	Case	Duplicate ratio	Case	Duplicate ratio
Case15	2.76	**Case25FL2**	2.28	**Case31tFL**	2.36
Case16	2.98	**Case25tFL**	2.62	**Case32**	3.8
Case17	3	**Case26FL1**	2.92	**Case33**	3.8
Case18	2.2	**Case26FL2**	2.62	**Case34**	2.96
Case19	2.26	**Case26tFL**	1.75	**Case35**	2.7
Case20	2.879	**Case27FL**	12	**Case38**	2.96
Case21	2.92	**Case27tFL**	2.82	**Case39**	3.2
Case23FL1	2.78	**Case28FL**	4.8	**Case3FLcustom**	2.24
Case23FL2	2.7	**Case28tFL**	6.7	**Case3tFLcustom**	2.4
Case23tFL	2.3	**Case29FL**	3.26	**Case40**	13
Case24FL1	2.28	**Case29tFL**	2	**Case5FLcustom**	2.74
Case24FL2	2.76	**Case10FLcustom**	2.94	**Case5tFLcustom**	2.56
Case24tFL	2.24	**Case10tFLcustom**	2.7	**CaseKARPAS422**	2.9
Case25FL1	2.46	**Case31FL**	2.56		

Table 3-8. Estimated duplicate ratio in all samples sequenced by custom capture panels.

Case	Chromosome|Position|Altered-Base	WES			Custom			Adjusted custom	
		Var Read Depth	Total Read Depth	Var Freq	Var Read Depth	Total Read Depth	Var Freq	Var Read Depth	Total Read Depth
FL-10	chr14|23345905|C	40	96	42%	547	1765	31%	186	600
FL-10	chr1|85736474|T	125	193	65%	17	245	7%	6	83
FL-3	chr1|110882568|G	12	21	57%	35	160	22%	16	71
FL-5	chrX|41200829|C	11	31	36%	35	81	43%	13	30
tFL-10	chr4|40245479|A	72	146	49%	1740	3550	49%	644	1315
tFL-10	chr16|3786704|G	24	29	83%	347	456	76%	129	169
tFL-10	chr6|157528317|A	46	89	52%	628	1653	38%	233	612
tFL-10	chr3|187449624|A	20	58	35%	440	1047	42%	163	388
tFL-10	chr12|49433524|-CT	50	62	81%	203	244	83%	75	90
tFL-10	chr17|80209332|A	57	153	37%	522	1213	43%	193	449
tFL-10	chr13|31037742|A	57	136	42%	44	295	15%	16	109
tFL-10	chr21|44521518|A	38	102	37%	28	172	16%	10	64
tFL-3	chr17|62007128|T	6	30	20%	447	2236	20%	186	932
tFL-3	chr4|3134324|A	15	61	25%	60	334	18%	25	139
tFL-3	chr15|42035019|A	21	39	54%	101	146	69%	42	61
tFL-5	chr16|3807881|A	18	34	53%	658	1646	40%	257	643
tFL-5	chr12|49438694|T	17	32	53%	683	1484	46%	267	580
tFL-5	chr12|49418731|A	17	43	40%	546	1162	47%	213	454
tFL-5	chr7|148508728|T	20	43	47%	317	857	37%	124	335
tFL-5	chr17|7577598|A	14	25	56%	3	28	11%	1	11
tFL-5	chr12|7287964|T	26	60	43%	1021	2430	42%	399	949

tFL-5	chr16|3808033|G	6	18	33%	451	1074	42%	176	420
tFL-5	chr16|3807917|C	25	47	53%	999	2379	42%	390	929
tFL-5	chr12|57493137|C	50	57	88%	654	909	72%	255	355
tFL-5	chr17|78269544|C	13	26	50%	297	874	34%	116	341
tFL-5	chr16|3808046|G	7	16	44%	451	1074	42%	176	420
tFL-5	chr16|85936784|G	18	32	56%	328	820	40%	128	320
tFL-5	chrX|39923127|G	9	19	47%	259	740	35%	101	289
tFL-5	chr3|78987869|A	24	52	46%	609	1449	42%	238	566
tFL-5	chr16|3808030|C	6	19	32%	451	1074	42%	176	420
tFL-5	chr12|57493818|C	51	60	85%	836	1114	75%	327	435
tFL-5	chr17|7577545|C	14	37	38%	5	6	83%	2	2

Table 3-9. Comparison of read depth in genes found mutated by WES and custom gene sequencing after duplicate ratio adjustment.

B.7. Discussion

We observed the depth in 2 mutations detected by custom capture was not as good as WES and the 2 mutations were very close to each other. These errors are probably due to the poor design of that specific region. The average duplicate ratio was 3.3, but there are four samples with a much larger estimated duplicate ratio than others. It is probably because the tumor genome in these four samples has CNAs that have a large fraction of the dbSNP SNPs, which are truly not equal to 0.5, these dbSNP SNPS would interfere with the approach used.

C. Somatic mutation filtering

C.1. Introduction

In order to remove germline variants from our mutation list, we applied different layers of filtering with a number of reliable datasets. We also validated our somatic mutation filters in samples with corresponding normal samples to ensure that our filtering method can exclude most germline variants from our list of likely somatic mutations. We set up two tiers of mutations, listed according to stringency.

C.2. Positive and negative filtering validation

We applied the filtering to a dataset that had paired FL and tFL samples with normal samples from the same patients. In Table 3-10, we listed the number of germline variants and somatic mutations in different layers of filtering. In negative filtering, we first removed synonymous mutations and mutations detected in intergenic and intronic regions, and 5' and 3' UTRs. The total number of mutations after this step was 51,685. Then dbSNP filtered out a majority of variants and left 4655 variants including 4038 germline variants and 617 somatic mutations. Mutations that are very close to each other were considered false positives and were removed; this step left 1390 mutations in the list including 877 germline variants and 513 somatic mutations. The paralog gene database removed 141 apparent mutations and left 1249 mutations for the next step. Mutations found in other unrelated normal samples were considered private SNPs or artifacts. This resulted in 1183 mutations remaining in the study including 709 germline variants and 474 somatic mutations. After this step, we checked the recurrence of these remaining mutations and removed the ones that had high frequency. In general, negative filtering removed 51199 variants and kept 943 variants with 485 germline variants and 458 mutations. The proportion of germline variants decreased in Figure 3-10. The

positive filtering filtered out 457 mutations and kept 486 mutations including 98 germline variants and 388 somatic mutations. Since mutations can fall in different categories among the positive filtering criteria, we made Table 3-11, Table 3-12 and Table 3-13 to give a specific look at the distribution of germline variants and somatic mutations, we also made Figure 3-11 to display the portion of germline variants and somatic mutations in different categories. The FDR of the filter set is 20.2%.

	Total	**Germline**	**Somatic**
non-protein-changing	51685	-	-
dbSNP	4655	4038	617
Adjacent	1390	877	513
paralog genes	1249	758	491
Normal samples	1183	709	474
Too Many Positive	943	485	458
Positive Filtering	486	98	388

Table 3-10. Evaluation of the somatic filters performance of the negative and positive filtering.

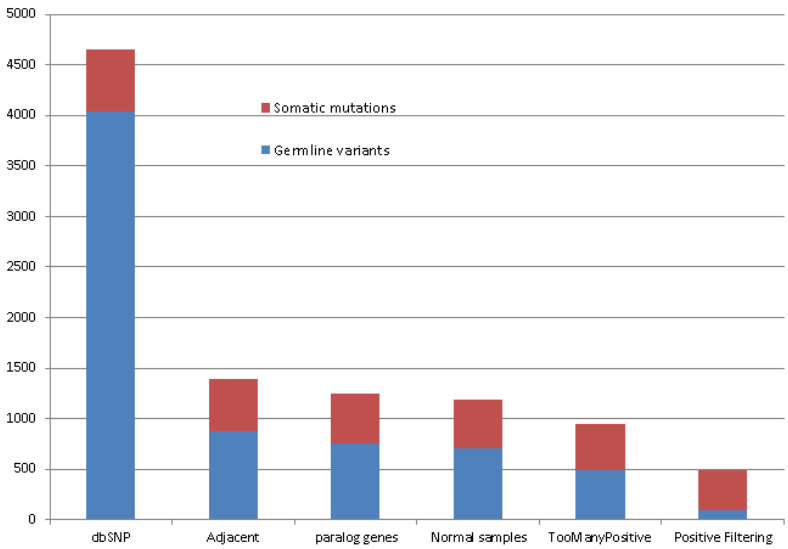

Figure 3-10. Proportions of germline variants and somatic mutations at different stages of negative and positive filtering.

Categories 388 true somatic mutations identified by	
Intra-Case-Specific	30
Intra-Case-Specific, Cancer-Gene	1
Intra-Case-Specific, Loss-Of-Function, Reduced-VAF	5
Intra-Case-Specific, Loss-Of-Function, Reduced-VAF, Cancer-Gene	1
Intra-Case-Specific, Reduced-VAF	107
Intra-Case-Specific, Reduced-VAF, Cancer-Gene	5
Intra-Case-Specific, Reported-Site	1
Intra-Case-Specific, Reported-Site, Cancer-Gene	2
Intra-Case-Specific, Reported-Site, Loss-Of-Function, Cancer-Gene	1
Intra-Case-Specific, Reported-Site, Reduced-VAF	4
Loss-Of-Function, Cancer-Gene	3
Loss-Of-Function, Reduced-VAF	31
Loss-Of-Function, Reduced-VAF, Cancer-Gene	4
Reduced-VAF	145
Reduced-VAF, Cancer-Gene	15
Reported-Site, Cancer-Gene	1
Reported-Site, Loss-Of-Function	4
Reported-Site, Loss-Of-Function, Cancer-Gene	1
Reported-Site, Loss-Of-Function, Reduced-VAF	4
Reported-Site, Loss-Of-Function, Reduced-VAF, Cancer-Gene	1
Reported-Site, Reduced-VAF	16
Reported-Site, Reduced-VAF, Cancer-Gene	6

Table 3-11. True positive somatic mutations in positive filtering.

Categories 98 false mutations belonged to	
Intra-Case-Specific, Reported-Site, Reduced-VAF	1
Loss-Of-Function, Reduced-VAF	4
Reduced-VAF	76
Reduced-VAF, Cancer-Gene	2
Reported-Site, Cancer-Gene	2
Reported-Site, Reduced-VAF	12
Reported-Site, Reduced-VAF, Cancer-Gene	1

Table 3-12. False positive somatic mutations in positive filtering.

	Germline	Somatic
Intra-Case-Specific	1	157
Reported-Site	16	41
Reduced-VAF	96	344
Loss-Of-Function	4	55
Cancer-Gene	5	41

Table 3-13. Evaluation of the performance of the somatic filters in positive filtering.

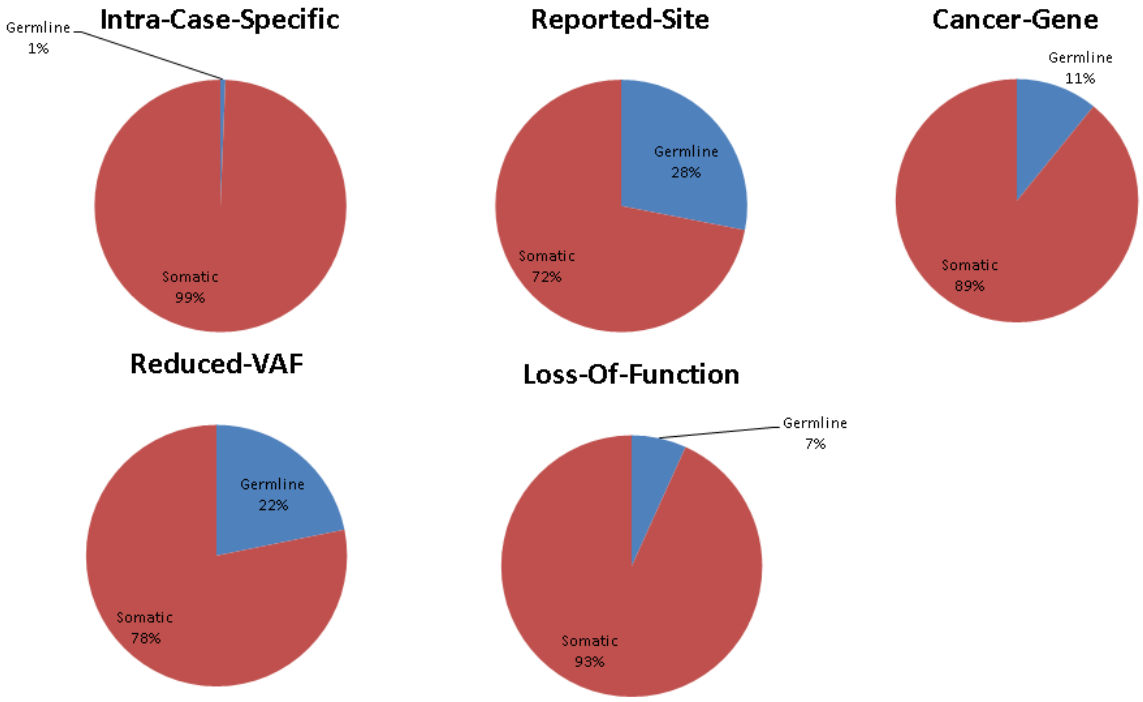

Figure 3-11. Proportions of germline and somatic variants among different categories of positive filtering.

C.3. Discussion

We applied very strict filters to avoid mistakenly including germline variants in our reliable somatic mutation list. These strict rules might be too strict and let some real somatic mutations slip away. For example, some real somatic mutations might not fall into negative filtering but also not have enough strong positive signals to pass the positive filter either. For such mutations, we added one more rescue rule, that is, the variants that did not pass the positive filter but

were in genes recurrently mutated in other samples were considered potential somatic mutations and were not removed.

D. Somatic mutation prediction by machine learning methods

D.1. Introduction

We applied various machine learning methods to two published datasets[8,9] (Pasqualucci dataset and Okosun dataset) to provide straightforward predictions of somatic mutations instead of filtering by layers. Sensitivity, specificity, FDR (Table 3-14) and AUC were used to evaluate the performance. We also applied the trained models to our dataset and checked the consistence with our filtering based method. RF turned out to have had the best performance in general.

		"Gold" standard		
		Negative	Positive	
Test outcome	Negative	TN	FN	
	Positive	FP	TP	FDR=FP/(TP+FP)
		Specificity=TN/(TN+FP)	Sensitivity=TP/(TP+FN)	

Table 3-14. 2X2 contingency table to calculate sensitivity, specificity and FDR. TN indicates true negative, FN indicates false negative, FP indicates false positive, TP indicates true positive, and FDR indicates false discover rate.

D.2. Machine learning validation

We downloaded two published datasets which have pairs of FL and tFL with corresponding normal samples. We used one dataset as training data, and the other as testing data with two sets of different features to predict somatic mutations. We assigned one set of features with basic information and the other set with more complex information. We calculated 4 statistical

measures to evaluate the performance of the models. Sensitivity measures the proportion of successfully detected true somatic mutations out of all true somatic mutations. Specificity measures the proportion of successfully detected true germline variants out of all true germline variants. FDR measures the proportion of false predicted somatic mutations out of all predicted somatic mutations. AUC measures the area under the ROC curve. 0.90 to 1 is considered to be excellent; 0.8 to 0.9 is considered to be good; 0.7 to 0.8 is considered to be fair; 0.6 to 0.7 is considered to be poor; and 0.5 to 0.6 is considered failure. In general, all the models had very decent statistical measures (Table 3-15, Table 3-16): the specificity and AUC were excellent. The sensitivity and the FDR varied among the different models. Complex features were slightly better than basic features. Taking all of the measures into consideration, RF had the best performance; it had much lower FDR and very similar in sensitivity to the others. We also applied the trained models to our own dataset and checked the overlap with the filtering based method. The majority of the predicted somatic mutations were detected by positive and negative filtering method, which confirmed the machine learning model to be very robust.

machine learning models	sensitivity	specificity	AUC	FDR	TN	FN	FP	TP	prediction in our sample (overlap with filtering method)
SVM	0.8443114	0.99837076	0.94872	0.15315	31252	52	51	282	55(42)
rp	0.8473054	0.99830687	0.98378	0.15774	31250	51	53	283	56(44)
rf	0.8353293	0.99869022	0.99722	0.12813	31262	55	41	279	50(38)
tree	0.8473054	0.99766796	0.92249	0.20506	31230	51	73	283	61(46)
nnet	0.8742515	0.99763601	0.99577	0.20219	31229	42	74	292	65(49)

Table 3-15. Performance of machine learning models with basic features. SVM indicates support vector machine, rp indicates recursive partitioning, rf indicates random forest, nnet indicates neural networks, AUC indicates area under curve, FDR indicates false discover rate, TN indicates true negative, FN indicates false negative, FP indicates false positive, and TP indicates true positive.

machine learning models	sensitivity	specificity	AUC	FDR	TN	FN	FP	TP	prediction in our sample (overlap with filtering method)
SVM	0.809677	0.998735	0.954981	0.128472	29221	59	37	251	49*(37)
rp	0.880645	0.997437	0.982434	0.215517	29183	37	75	273	62(45)
rf	0.854839	0.998291	0.99739	0.15873	29208	45	50	265	54(42)
tree	0.883871	0.997881	0.940876	0.184524	29196	36	62	274	63(47)
nnet	0.870968	0.998394	0.976341	0.148265	29211	40	47	270	50(40)

Table 3-16. Performance of machine models with complex features. SVM indicates support vector machine, rp indicates recursive partitioning, rf indicates random forest, nnet indicates neural networks, AUC indicates area under curve, FDR indicates false discover rate, TN indicates true negative, FN indicates false negative, FP indicates false positive, and TP indicates true positive. Note: * one spot was removed because the training dataset does not have it covered.

D.3. Discussion

SIFT score is the feature we added in the set of complex features. It predicts amino acid changes that affect protein function[33]. If a variant is a somatic mutation that contributes to the disease, SIFT should be more likely to predict it to be deleterious. Conversely, germline variants that exist in a normal sample are more likely to be functionally neutral with a benign prediction from SIFT. Therefore, SIFT score should be able to provide very useful information to improve the training models. From the results, we did see the models that were trained with complex features performed better than those trained only with basic features. SIFT does not assign a score to all the variants; therefore, we had to filter out those that do not have a SIFT score in the training data. This limited the size of the training data, which may explain why the complex feature did not improve the statistical measures a lot. We do not have normal samples in our own dataset, but our somatic mutation filtering is very reliable. The overlap of the two independent methods is strong evidence indicating that the machine learning model has the capability to predict somatic mutations. It also confirms our filtering method.

E. CNA and patient outcome

E.1. Introduction

FL is indolent with about 8 to 10 years survival time. However, tFL progresses more rapidly with shorter survival, commonly less than two years. Therefore it is very important for us to accurately identify abnormalities that are associated with survival so that we can monitor the patients more closely and provide better treatment plans. Compared to standard survival analysis, our advanced survival analysis provided alternative tests that were able to put weight in different stages of the survival curve and chose the most accurate test based on the character of each individual rCNA.

E.2. Survival analysis for patients with different rCNAs

In our previous study, we used a CNV analysis pipeline on high-resolution SNP array processed samples to identify abnormalities that contribute to FL transformation. For each rCNA, the average number of rCNAs present in cases with that rCNAs was calculated to model the progression of rCNAs in FL evolution. Abnormalities that occur in the early stage of FL are expected to have higher frequency and be found preferentially in tumors that have a lower average number of other abnormalities, whereas abnormalities that occur late in the disease are expected to be found preferentially in tumors that have a high average number of other abnormalities.

In our survival analysis, we first inferred the temporal order of the rCNAs based on the average abnormality numbers in Figure 3-12[31]. For example, loss on 9p21.3 (rCNA402) and 6q13-q15 (rCNA341) occurred early in FL development. We then put the weight in different stages of the survival curve based on the biological characters of the rCNAs to test the difference between the two overall survival curves.

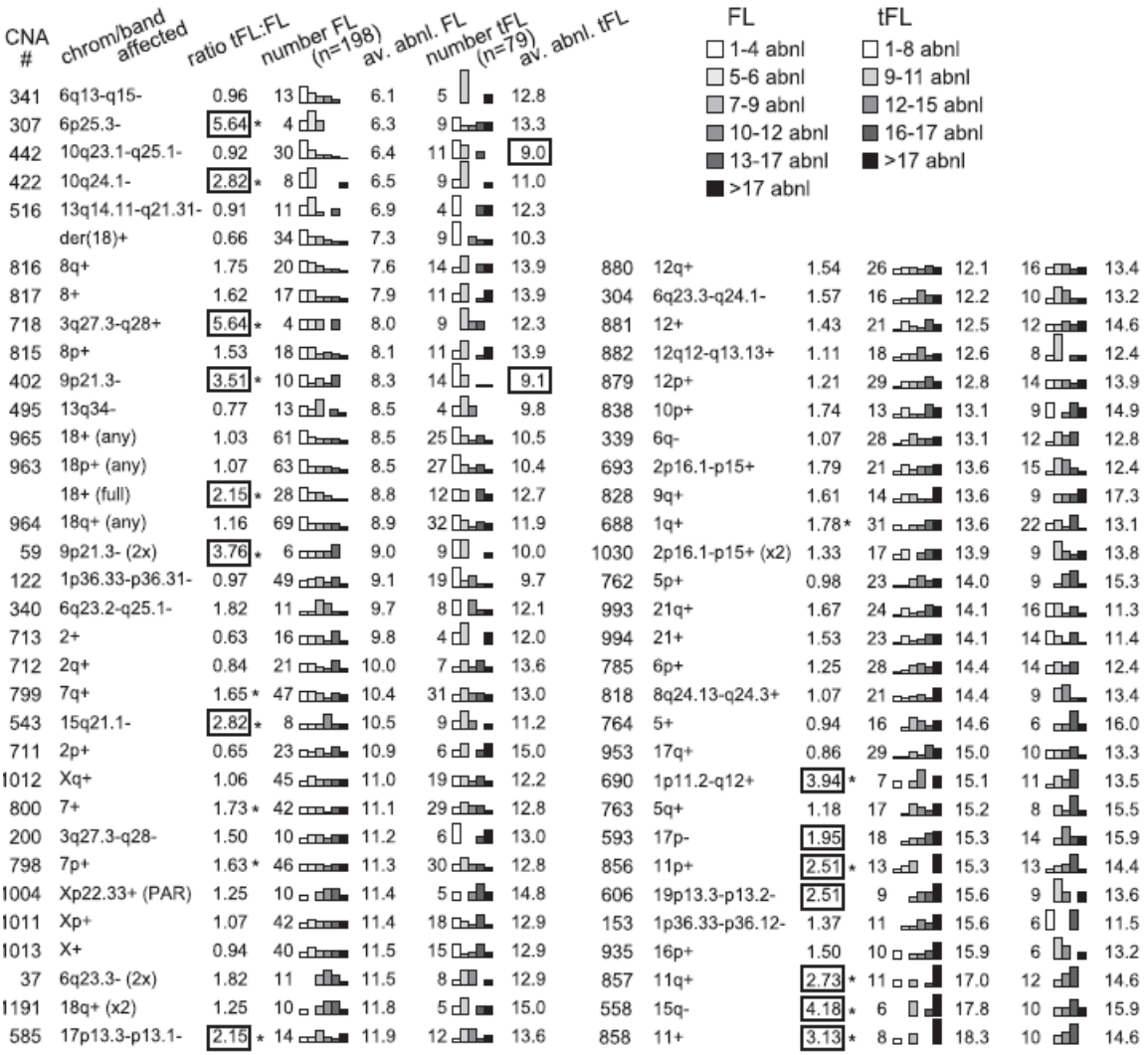

Figure 3-12. Number of abnormalities associated with rCNAs[31]. The av.abnl indicates the average numbers of abnormalities with a specific CNA.

Our survival analysis generated pdf format output with convenient bookmarks (Figure 3-13). The results included survival estimation at different time points with the corresponding number of failures, the number of remaining patients in the study and number of censored patients, p-values for all alternative tests, survival curves for the two compared groups (with and without the abnormality) with 95% confidence intervals and censored patients at different time points within the survival curves (Figure 3-14). For example, we had 5 patients who had rCNA402 (Table 3-18). 4 of them died at different time points (0.734, 0.767, 1.792 and 9.949). One of

them was lost to follow-up at 5.8940. The probability that a patient is alive past time=0 is 100%; the probability that a patient is alive past time=0.734 (268 days) is 80%; and the probability that a patient is alive past time=0.7670 (280 days) is 60%, when it comes to time=9.949 (3631 days), the probability that a patient is alive is 0%. If we do not take the biological property into consideration and just use the log-rank test, the log-rank statistic is 5.5652 and the corresponding p-value is 0.0183 (Table 3-19) indicating that the null hypothesis (all survival curves are the same) should be rejected. We can therefore conclude that groups with and without rCNA402 have significantly different KM survival curves and that rCNA402 is associated with poor outcome in FL. However, according to the calculated average numbers of abnormalities (Figure 3-12), rCNA402 occurs quite early (8.3 in Figure 3-12) in the FL development; therefore, the Wilcoxon test is a much more appropriate test for rCNA402, which puts more weight on the earlier survival curve, and the p-value (0.0080) from Wilcoxon test shows a much more significant difference between groups with and without rCNA402 (Table 3-19) than the p-value (0.0183) from log-rank test. We are also 95% confident that the entire survivor function fell within upper curve and lower curve as figure 3-3 shows. We applied the survival analysis to other rCNAs that appear to occur in the early stage of FL as the previous study mentioned, and showed that rCNA304 (p-value<0.0001 by Tarone-Ware test), rCNA341 (p-value =0.05 by Tarone-Ware test), rCNA1013 (p-value =0.0457 by Tarone-Ware test) and rCNA818 (p-value=0.0469 by Tarone-Ware test) are associated with poor survival for the patients (Figure 3-14 and Table 3-19).

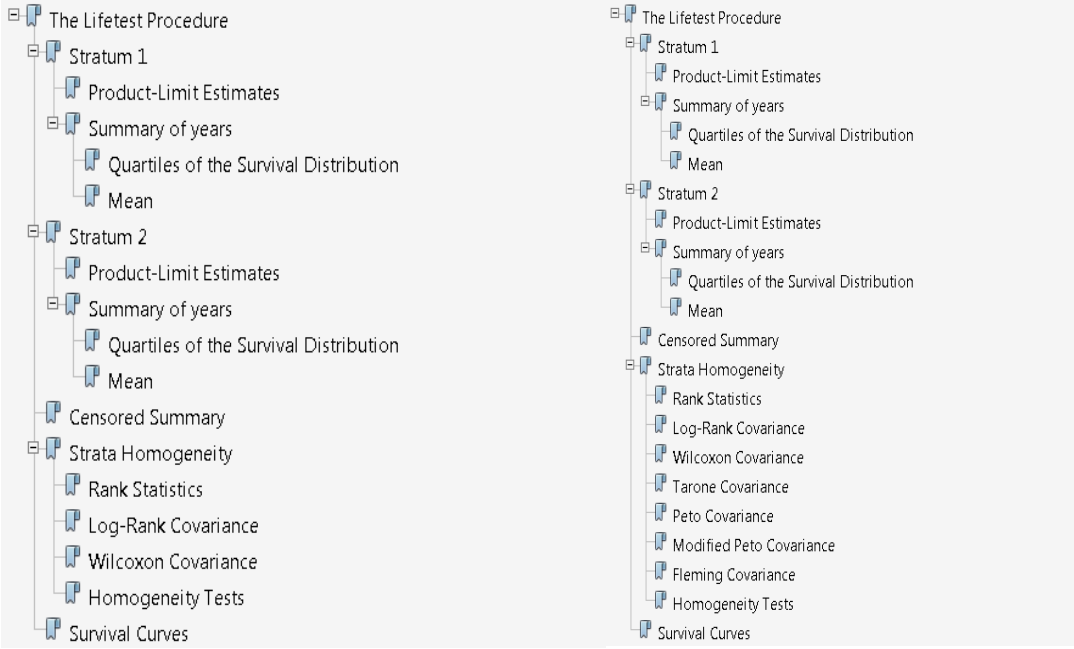

Figure 3-13. pdf format output with bookmark.

The LIFETEST Procedure

Stratum 1: group = abnorm

Product-Limit Survival Estimates						
years		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000		1.0000	0	0	0	5
0.7340		0.8000	0.2000	0.1789	1	4
0.7670		0.6000	0.4000	0.2191	2	3
1.7920		0.4000	0.6000	0.2191	3	2
5.8940	*	.	.	.	3	1
9.9490		0	1.0000	.	4	0

Note: The marked survival times are censored observations.

Table 3-17. Survival estimation at different time points in rCNA402.

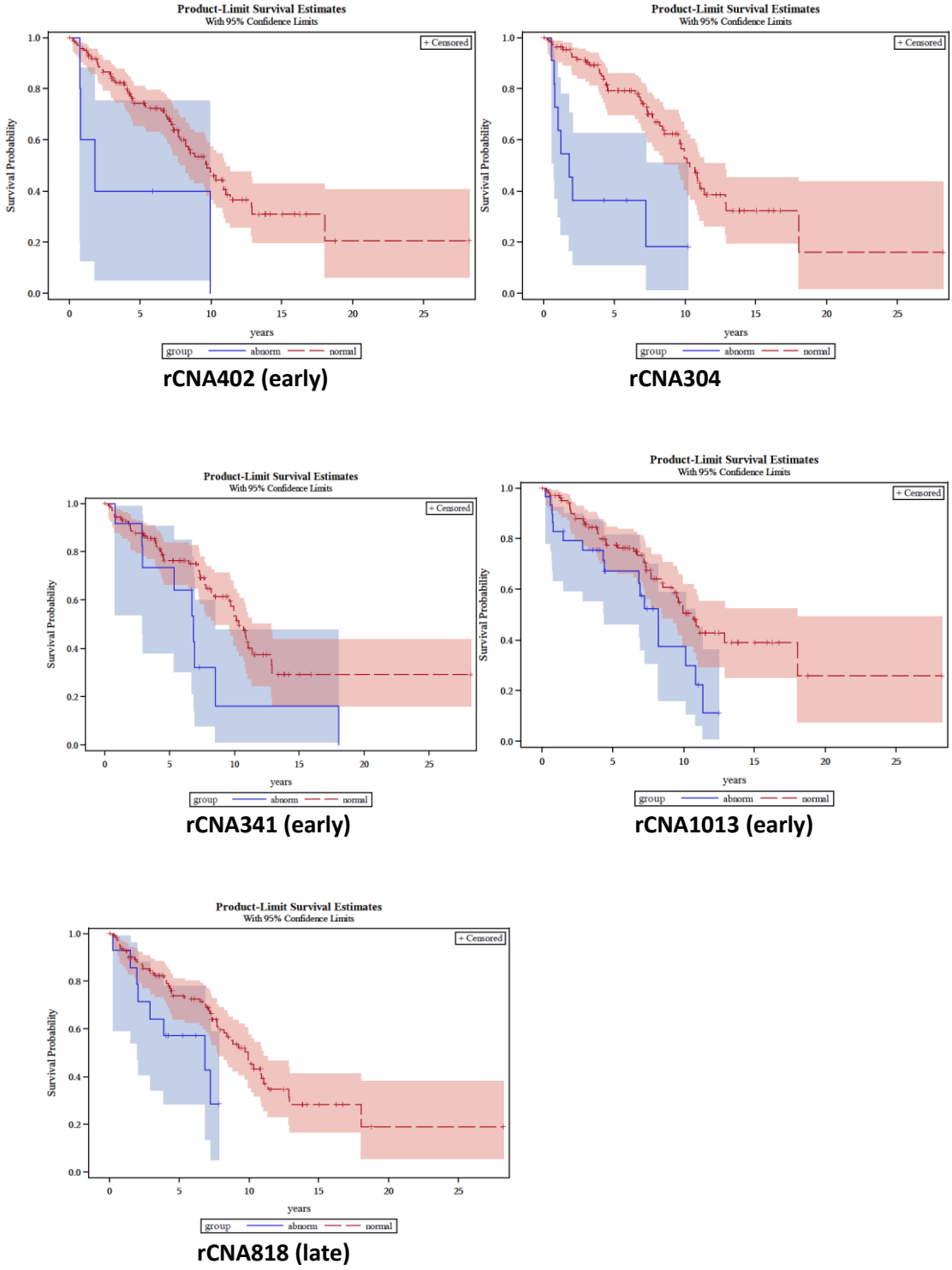

Figure 3-14. Survival curves of rCNA402, rCNA304, rCNA341, rCNA1013 and rCNA818.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	5.5652	1	0.0183
Wilcoxon	7.0233	1	0.0080
Tarone	6.3001	1	0.0121
Peto	6.2719	1	0.0123
Modified Peto	6.2756	1	0.0122
Fleming(1)	6.2684	1	0.0123

rCNA402 (early)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	17.0212	1	<.0001
Wilcoxon	23.6447	1	<.0001
Tarone	21.0568	1	<.0001
Peto	20.7670	1	<.0001
Modified Peto	20.8619	1	<.0001
Fleming(1)	20.7133	1	<.0001

rCNA304

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.4802	1	0.0343
Wilcoxon	2.8682	1	0.0903
Tarone	3.7752	1	0.0520
Peto	3.8982	1	0.0483
Modified Peto	3.8298	1	0.0503
Fleming(1)	3.9813	1	0.0460

rCNA341 (early)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.7466	1	0.0294
Wilcoxon	3.5357	1	0.0601
Tarone	3.9940	1	0.0457
Peto	4.1397	1	0.0419
Modified Peto	4.1038	1	0.0428
Fleming(1)	4.1613	1	0.0414

rCNA1013 (early)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	4.2893	1	0.0384
Wilcoxon	3.5605	1	0.0592
Tarone	3.9504	1	0.0469
Peto	3.9415	1	0.0471
Modified Peto	3.9327	1	0.0474
Fleming(1)	3.9458	1	0.0470

rCNA818 (late)

Table 3-18. Tests in rCNA402, rCNA304, rCNA341, rCNA1013 and rCNA818.

E.3. Discussion

We collected a decent number of clinical samples but some of the abnormality groups had a very small number of patients. It could cause a large standard error or a situation where one or two patients' failure makes a noticeable difference between the groups being compared.

There are multiple tests for abnormalities occurring in the early stage of the disease, and each of them is slightly different. Compared to the log-rank test, the Wilcoxon test puts the most weight in the earlier phase of the disease. The Tarone-Ware test is more moderate, and the weight is between a log-rank test and the Wilcoxon test on the earlier phase. One study group proved that the Tarone-Ware test is better than both the log-rank test and Wilcoxon test, so if the rCNAs do not indicate that abnormalities occur in a very early stage, we chose Tarone-ware test instead of other tests. We can also take all the p-values from the tests into consideration. The Fleming-Harrington test can be applied to both abnormalities occurring in the early or late stage depending on the parameters. It is a very flexible method, but there are a large number of combinations for the parameters. How the parameters are assigned can also make a difference.

CHAPTER IV

DOWNSTREAM ANALYSIS AND DISCUSSION

A. Somatic mutation integration

We integrated somatic mutations that were identified in WES and custom capture sequencing using strict positive and negative filtering and selected recurrently mutated genes in the combined WES and custom capture panel dataset. 50 recurrently mutated genes were identified in the combined dataset. The genes most frequently mutated in tFL included *KMT2D (MLL2)*, *CREBBP*, *BCL2*, *EZH2*, *MEF2B*, and *TNFRSF14* as noted in Figure 4-1. We also identified recurrent shared mutations, FL-unique mutations and tFL-unique mutations from WES and custom capture sequencing each individually for FL transformation and subclone analysis (Table 4-1 and Appendix A). Mutations detected in single tFL samples (without a corresponding FL sample) also contributed to FL transformation analysis (Table 4-1 and Appendix A).

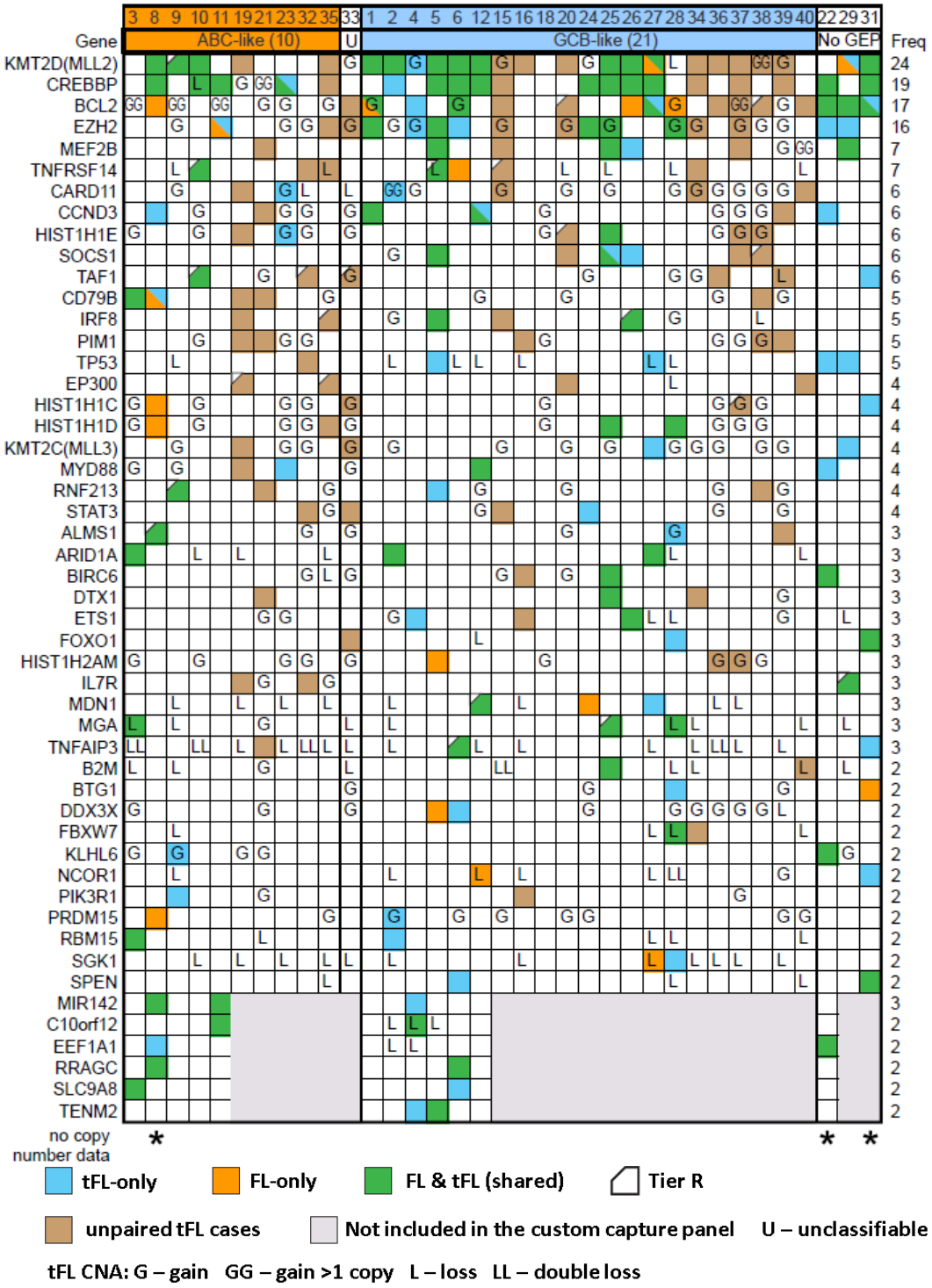

Figure 4-1. Genes found to be recurrently mutated in tFLs. Only genes with at least one Tier 1 mutation found in a tFL/FL pair were selected. Its distribution in other situations is also shown with different colors representing different statuses of the individual cases. Blocks with 1 color indicate that one mutation was observed. Similarly, blocks with 2 colors indicate that more than one status was observed. Genes are also noted by copy number gain or loss in the tFL cases. Copy number data are not available for case 8, 22, and 31. Tier R mutations were "rescued" from the stringent filtering criteria associated with Tier 1 and Tier 2 mutations.

Class	T1+T2+TR	Genes
FL-unique	2	KMT2D(MLL2), BCL2
tFL-unique	10	KMT2D(MLL2), CREBBP, BCL2, EZH2, CARD11, CCND3, SOCS1, TP53, KMT2C(MLL3), MYD88
shared	16	KMT2D(MLL2), CREBBP, BCL2, EZH2, MEF2B, TNFRSF14, CCND3, SOCS1, IRF8, HIST1H1D, ARID1A, BIRC6, MGA, MIR142, C10orf12, RRAGC
single	22	KMT2D(MLL2), CREBBP, BCL2, EZH2, MEF2B, TNFRSF14, CARD11, CCND3, HIST1H1E, SOCS1, TAF1, CD79B, IRF8, PIM1, EP300, HIST1H1C, KMT2C(MLL3), RNF213, STAT3, DTX1, HIST1H2AM, IL7R

Table 4-1. Recurrent somatic mutations identified in WES and custom capture sequencing. Genes from Tier R are labeled in green.

B. Somatic mutations within regions of rCNAs

Using our previous CNA data, we identified the CN for the recurrent mutations (Appendix B) and listed the mutations within rCNAs in FL and tFL samples (Appendix C). The integration of mutation data with CNA data provided complementary information to investigate the role of genetic alterations in tumor initiation and progression. In Figure 4-1, genes with CNA were denoted as G (CN=3, CN gain), L (CN=1, CN loss), GG (CN>3, CN amplification) and LL (CN<1, double loss).

TNFRSF14 is a likely driver gene in one of the most frequent rCNAs in FL and tFL (loss of 1p36.33-p36.31, rCNA122 in Table 4-2). It was mutated in 20% (7/35) of cases. One of the mutations was a FL-unique mutation in a paired case (chr1|2493172|A in FL-6). This case also had a CN loss (CN=1) affecting the gene. Two of the mutations were shared mutations (chr1|2493112|C in case 5 and chr1|2493111|A in case 10). One of the two shared mutations was associated with a CN loss (case 5, CN=1). The rest of the mutations were all found in

unpaired single tFL cases. One of the single tFL cases (chr1|2492063|-C in tFL-35) had a CN loss (CN=1) affecting TNFRSF14. Thus, homozygous loss of wild-type *TNFRSF14* is a relatively frequent event in FL and tFL.

CARD11 is an important oncogenic gene with frequent CN gain. It was mutated in 17% (6/35) of cases. Two of the mutations were tFL-unique mutations (chr7|2979495|G in tFL-2 and chr7|2979501|G in tFL-23). Both of the tFL-unique mutations had a CN amplification (CN=4 in tFL-2 and CN=3 in tFL-23) affecting the locus. The other cases with mutations were all unpaired single tFL cases. Two of the single tFL cases (chr7|2984163|T in tFL-15, chr7|2979466|G in tFL-15, and chr7|2977614|A in tFL-34) had CN gain (CN=3) affecting *CARD11*. The *CARD11* locus is affected by gains in 24% of FL samples and 39% of tFL samples in our previous studies. All the evidence suggests that *CARD11* may be activated by both CN gain and mutation, and coordinate the activation of the NF-κB pathway.

HIST1H1E is another important oncogenic gene with frequent copy number gain[34]. It was mutated in 17% (6/35) of cases. Two out of three paired cases had tFL-unique mutations (chr6|26156911|A in tFL-20 and chr6|26156797|T in tFL-23). One of the two tFL-unique mutations also had a CN gain (CN=3). The other 3 cases of the 6 were all unpaired single tFL cases. Two (chr6|26156947|G in tFL-37 and chr6|26157271|G in tFL-38) out of the three single tFL cases also had CN gain (CN=3). The evidence suggests that *HIST1H1E* may be affected by both CN gain and mutation.

EZH2 is a known mutated gene in FL and tFL. It inhibits genes responsible for suppressing tumor development, and blocking EZH2 activity may slow tumor growth. It was mutated in 46% (16/35) of cases. Most of the mutations were the two highly recurrent mutations (chr7|148508727|A and chr7|148508728|T) reported by other groups as well. Two mutations

were found in the same tFL case (chr7|148508745|C and chr7|148508763|A in tFL-34). One mutation was a tFL-unique mutation (chr7|148506437|A in tFL-11). These mutations were often shared FL/tFL mutations or found in single tFL samples with CN gains affecting the *EZH2* gene. The *EZH2* locus is involved in an rCNA that was affected by gains in 24% of FL samples and 39% of tFL samples in our previous studies (rCNA799 in Table 4-2). The CN gain and mutations in *EZH2* could enhance histone methylation at H3K27 and turn off genes inhibiting cell proliferation and differentiation.

Our previous study identified several recurrent losses where the likely driver genes were identified. In our previous study, we identified a region of homozygous loss on 6q that occurred in 10% of tFLs and included only *TNFAIP3* (rCNA37 in Table 4-2). Forty three percent (15/35) of our samples had copy losses including this gene (Figure 4-1). Three mutations were found in TNFAIP3: one unpaired tFL mutation (tFL-21, chr6|138200194|G), one tFL-unique mutation (chr6|138192455|T in tFL-31), and a shared mutation (chr6|138200146|G in case 6) with a copy loss in its FL sample (Table 4-2). All the evidence suggests that *TNFAIP3* may be inactivated by both CN loss and mutation. A larger heterozygous rCNA on 6q occurring in 10% of tFLs (rCNA340 in Table 4-2) includes 102 genes including *SGK1*, and a recurrent mutation of the kinase *SGK1* (chr6|134495706|G in FL-27 and chr6|134495724|C in tFL-28) occurred in two paired samples. Thirty four percent (12/35) of our samples had copy losses of this gene (Figure 4-1). *CREBBP* is the second most frequently mutated gene in our study. Mutation occurred in 54% (19/35) of our cases. Most of the mutations in *CREBBP* were shared FL/tFL mutations; the rest of the mutations were detected in single, unpaired tFL samples. One case had a shared *CREBBP* mutation and a CNL affecting *CREBBP* in the tFL. A small loss on chr16 (in 5% of tFLs in rCNA564 in Table 4-2) encompasses *CREBBP* and 7 other genes, and is likely driven by *CREBBP*.

TP53 was recurrently mutated in tFLs with 14% (5/35) of cases in our sequencing study. A 17p loss occurs in 9% of FLs and 18% of tFLs (rCNA593 in Table 4-2). One case had heterozygous CN loss affecting *TP53* in both FL and tFL samples in our previous study, but the mutations were only detected in tFL samples. We had one mutation in an unpaired tFL sample (chr17|7578286|G in tFL-32) and four tFL-unique mutations (chr17|7577545|C and chr17|7577598|A in tFL-5, chr17|7577097|A in tFL-22, chr17|7578265|G in tFL-27, chr17|7576571|C and chr17|7577093|G in tFL-29). One tFL-unique mutation had a CN loss (CN=1) affecting *TP53*. It is probable that the *TP53* expression is reduced by CN loss earlier in the disease, and the mutations occur later, as they were only found in tFLs.

rCNA	CNA band	Type	Freq in FLs	Freq in tFLs	Numbers of gene in rCNA	Mutated Genes in Our Cases
122	1p36.33-p36.31-	Loss	25%	24%	51	TNFRSF14,
343	6p22.2-p21.33-	Loss	1%	1%	173	HIST1H3A, HIST1H3B, HIST1H1C, HIST1H1E, HIST1H2BG, HIST1H1D, PRSS16, ZNF184, HIST1H2AM, ZKSCAN3, PPP1R10,
799	7q+	Amp or Gain	24%	39%	739	CD36, CACNA2D1, PCLO, ZNF804B, AKAP9, CYP51A1, RELN, CCDC136, PLXNA4, EZH2, MLL3,
1191	18q+ (x2)	Amp	5%	6%	242	ESCO1, DSC1, EPG5, LOXHD1, CTIF, WDR7, PHLPP1, BCL2, CCDC102B, NETO1,
37	6q23.3- (x2)	dbLoss	6%	10%	1	TNFAIP3,
340	6q23.2-q25.1-	Loss	6%	10%	102	SGK1, TNFAIP3,
564	16p13.3-	Loss	3%	5%	8	CREBBP,
593	17p-	Loss	9%	18%	431	ZNF594, TP53, PER1, ZNF18, CDRT1, NCOR1, C17orf51,

Table 4-2. Mutations identified in rCNAs. rCNAs refer to the abnormalities which were identified and described in Bouska et al, 2013. Red color indicates mutated genes.

C. Somatic mutations acquired during transformation of FL

We analyzed paired samples to identify genes that are likely to contribute to transformation. The genes that are mutated preferentially in the tFL samples are considered likely drivers of the indolent FL into aggressive tFL. *TP53* (Table 4-3) and *USH2A* were the only two recurrent tFL-unique mutations that were found exclusively in our 12 paired WES samples. *USH2A* is probably not expressed in B cells based on our previous GEP analysis. To identify more somatic mutations acquired during transformation, we expanded the analysis to FL and tFL sample pairs sequenced using the custom capture panel and detected 2 additional genes that were recurrently mutated only in tFL samples. These two genes were *CARD11* and *KMT2C (MLL3)*. *CARD11* was mutated in 17% (6/35) of cases and also affected by CN gain (Table 4-3). *KMT2C (MLL3)* was mutated in 11% (4/35) of cases (Table 4-3). Two of the mutations were tFL-unique mutations (chr7|151873463|A in tFL-27 and chr7|152027794|-C in tFL-29). The other mutations were all found in unpaired single tFL cases. One single tFL sample had two mutations (chr7|151860157|G and chr7|151868408|T in tFL-19); the other single tFL sample had a mutation (chr7|151891609|G in tFL-33) and also had a CN gain (CN=3) affecting *KMT2C (MLL3)*. We also identified a set of genes with tFL-unique mutations in 2 or more cases, including *EZH2*, *CCND3* (Table 4-3), and *MYD88* (Table 4-3). These mutations were not exclusive to tFL, and some of the mutations are actually known to occur early in the development of FL, but it is possible they can also be late mutations that cooperate with other mutations in transformation or mutations that are present in the subclones that transformed later. Comparing each gene to all the others, three of the 50 genes in Figure 4-1 showed a significant difference in the fraction of tFL-mutated paired cases that had a tFL-unique mutation. TP53 showed a significant increase ($p= 0.024$, two-sided Fisher's exact test), consistent with its importance in transformation. In

contrast, MLL2 and CREBBP showed a significant decrease (p=0.045 each), consistent with mutations in these genes being usually early events in FL.

10 out of 20 paired cases had *EZH2* mutations, all of which occurred in the SET domain. In 4 of the cases (tFL-4, tFL-6, tFL-22 and tFL-29), the mutations were only present in the tFL sample. In 5 cases (cases 1, 5, 24, 25 and 28), the mutation was shared with the FL biopsy, and in 1 case (case 11) the mutation was detected at very low VAF (<4%) in the FL biopsy and became clonal in the tFL biopsy.

Sample ID	Chr|Pos|VarAllele	Gene	MutType	Var Freq	CN
tFL-5	chr17|7577545|C	TP53	nonsynonymous	0.378	2
tFL-5-cus	chr17|7577545|C	TP53	nonsynonymous	0.833	NA
tFL-5	chr17|7577598|A	TP53	nonsynonymous	0.56	2
tFL-5-cus	chr17|7577598|A	TP53	nonsynonymous	0.107	NA
tFL-22	chr17|7577097|A	TP53	nonsynonymous	0.176	NA
tFL-27	chr17|7578265|G	TP53	nonsynonymous	0.63	1
tFL-29	chr17|7576571|C	TP53	stopgain	0.469	2
tFL-29	chr17|7577093|G	TP53	nonsynonymous	0.511	2
tFL-32	chr17|7578286|G	TP53	nonsynonymous	0.371	2
tFL-2	chr7|2979495|G	CARD11	nonsynonymous	0.303	4
tFL-15	chr7|2984163|T	CARD11	nonsynonymous	0.47	3
tFL-15	chr7|2979466|C	CARD11	nonsynonymous	0.46	3
tFL-19	chr7|2977613|T	CARD11	nonsynonymous	0.32	2
tFL-19	chr7|2979486|G	CARD11	nonsynonymous	0.4	2
tFL-23	chr7|2979501|G	CARD11	nonsynonymous	0.13	3
tFL-34	chr7|2977614|A	CARD11	nonsynonymous	0.28	3
tFL-40	chr7|2985468|T	CARD11	nonsynonymous	0.41	2
tFL-19	chr7|151860157|G	KMT2C(MLL3)	nonsynonymous	0.25	2
tFL-19	chr7|151868408|T	KMT2C(MLL3)	nonsynonymous	0.341	2
tFL-27	chr7|151873463|A	KMT2C(MLL3)	nonsynonymous	0.39	2
tFL-29	chr7|152027794|-C	KMT2C(MLL3)	frameshift_deletion	0.34	2
tFL-33	chr7|151891609|G	KMT2C(MLL3)	nonsynonymous	0.37	3
tFL-12	chr6|41903688|G	CCND3	nonsynonymous	0.437	2
FL-12	chr6|41903688|T	CCND3	nonsynonymous	0.31	2
tFL-21	chr6|41903731|A	CCND3	stopgain	0.099	2
tFL-22	chr6|41903755|A	CCND3	stopgain	0.25	NA

tFL-39	chr6|41903710|G	CCND3	nonsynonymous	0.44	2
FL-12	chr3|38182641|C	MYD88	missense	0.401	2
tFL-12	chr3|38182641|C	MYD88	missense	0.833	2
tFL-19	chr3|38182641|C	MYD88	missense	0.81	2
tFL-22	chr3|38182025|T	MYD88	nonsynonymous	0.143	NA
tFL-23	chr3|38182032|G	MYD88	nonsynonymous	0.19	2

Table 4-3. Mutations in TP53, CARD11, KMT2C (MLL3), CCND3, and MYD88.

D. Genes mutated in ABC-like vs. GCB-like lymphomas

We classified tFL as ABC-like, UC, or GCB-like based on the gene expression signatures using in our previous study. The classification information was available for 32 out of 35 samples. Unexpectedly, a substantial number of tFL cases are classified as the ABC type instead of GCB type. *CD79B* was mutated in 14% of cases (5/35), more frequently in ABC-like (4/10=40%) than in GCB-like tFL (1/21=5%) (p=0.0274, Fisher's exact test in Table 4-4), consistent with findings in de novo DLBCL[35]. Interestingly, the mutations were clustered in the immunoreceptor tyrosine-based activation motif (ITAM) domain.

The NF-κB pathway plays an important role of ABC-like tFL. Therefore, we investigated the mutations that can activate this pathway in our data. *CARD11* (6/35), *MYD88* (4/35, 1 is a shared mutation), *TNFAIP3* (3/35), and *BCL10* (1/35) were found mutated (Table 4-5). *CARD11* is affected by CN gains in FL (24%) and tFL (39%) (Table 4-6), and it also had CN gain in two sequenced samples which had tFL-unique mutations (CN=3 in tFL-2 and tFL-23). *TNFAIP3* is affected by CN losses in FL (26%) and tFL (34%) (Table 4-6). *CARD11* mutations (Table 4-5) occurred within or adjacent to the coiled-coil domain (CCD) in our case. These mutations likely disrupt binding of the CCD to the inhibitory domain (ID), allowing *CARD11* to assume its active conformation even without phosphorylation, and consequently activating NF-κB in the absence of BCR engagement[36,37]. *BCL10* mutation (R58Q) occurred in its CARD domain near the acidic

patch that binds to the basic patch of the *CARD 11* CARD domain, and three acidic residues (E50, E53, and E54) were identified as important for binding[38]. The R58Q mutation substitutes a glutamine for the basic arginine, which could increase binding to *CARD11* by increasing the net negative charge. The *MYD88* was mutated in 11% (4/35, 1 case doesn't have GEP data) of cases (Table 4-5), more commonly in the ABC-like (2/10) than in the GCB-like tFL (1/21). All of the mutations were in the Toll/interleukin-1 homology domain (TIR), including two cases with the most frequent mutation, L265P[39].

BCL2, KMT2D (*MLL2)*, *EZH2* and *SOCS1* were mutated more frequently in GCB-like tFL (Table 4-4) in our study. Of note, *BCL2* and *SOCS1* were only mutated in GCB-like tFLs, consistent with what has been reported in DLBCL[8,40].

Gene	all mut freq (n=35)	FL mut freq (n=20)	tFL mut freq (n=35)	tFL only mut freq (n=20)*	ABC tFL mut freq (n=10)	GCB tFL mut freq (n=21)	p-values of ABC vs GCB tFL mutations
BCL2	49%	45%	40%	5%	0%	48%	0.012
CD79B	14%	10%	14%	0%	40%	5%	0.027
IL7R	9%	5%	9%	0%	20%	0%	0.097
HTT	6%	10%	6%	0%	20%	0%	0.097
KMT2D(MLL2)	69%	65%	69%	5%	50%	81%	0.105
EZH2	46%	30%	46%	20%	20%	52%	0.129
SOCS1	17%	10%	17%	5%	0%	29%	0.141
MYD88	11%	5%	11%	10%	20%	5%	0.237
KLHL6	6%	5%	6%	5%	10%	0%	0.323
AKAP13	3%	0%	3%	5%	10%	0%	0.323
ANXA1	3%	0%	3%	5%	10%	0%	0.323
ARID1B	3%	5%	3%	0%	10%	0%	0.323
BCL10	3%	5%	3%	0%	10%	0%	0.323
BCL11A	3%	0%	3%	5%	10%	0%	0.323
BCL6	3%	5%	3%	0%	10%	0%	0.323
CD58	3%	0%	3%	0%	10%	0%	0.323
CSNK1D	3%	0%	3%	5%	10%	0%	0.323
FTH1	3%	0%	3%	0%	10%	0%	0.323
HELZ	3%	5%	3%	0%	10%	0%	0.323
HMGB1	3%	5%	3%	0%	10%	0%	0.323
HPS5	3%	5%	3%	0%	10%	0%	0.323

MIA3	3%	5%	3%	0%	10%	0%	0.323
NOTCH2	3%	0%	3%	0%	10%	0%	0.323
PHF6	3%	5%	3%	0%	10%	0%	0.323
PRDM1	3%	0%	3%	0%	10%	0%	0.323
RAPGEF2	3%	0%	3%	5%	10%	0%	0.323
RHOH	3%	5%	3%	0%	10%	0%	0.323
SETD2	3%	5%	3%	0%	10%	0%	0.323
TBL1XR1	3%	0%	3%	5%	10%	0%	0.323
TET2	3%	0%	3%	0%	10%	0%	0.323
TINF2	3%	0%	3%	5%	10%	0%	0.323
U2AF1	3%	5%	3%	0%	10%	0%	0.323
UBAP2	3%	0%	3%	5%	10%	0%	0.323
ZNF142	3%	0%	3%	0%	10%	0%	0.323
TNFRSF14	20%	15%	17%	0%	30%	14%	0.358
DMD	9%	5%	9%	0%	0%	14%	0.533
ETS1	9%	5%	9%	5%	0%	14%	0.533
TAF1	17%	5%	17%	5%	20%	10%	0.577
EP300	11%	0%	11%	0%	20%	10%	0.577
RNF213	11%	5%	11%	5%	20%	10%	0.577
MEF2B	20%	15%	20%	5%	10%	24%	0.634

Table 4-4. P-values of ABC vs GCB tFL mutations. *Only calculated the frequencies in paired cases.

Sample ID	Chr|Pos|VarAllele	Gene	CN
FL-10	chr1|85736474|T	BCL10	2
FL-10-cus	chr1|85736474|T	BCL10	2
tFL-10-cus	chr1|85736474|T	BCL10	4
tFL-2	chr7|2979495|G	CARD11	4
tFL-15	chr7|2984163|T	CARD11	3
tFL-15	chr7|2979466|C	CARD11	3
tFL-19	chr7|2977613|T	CARD11	2
tFL-19	chr7|2979486|G	CARD11	2
tFL-23	chr7|2979501|G	CARD11	3
tFL-34	chr7|2977614|A	CARD11	3
tFL-40	chr7|2985468|T	CARD11	2
FL-3	chr17|62007128|T	CD79B	2
FL-3-cus	chr17|62007128|T	CD79B	NA
tFL-3	chr17|62007128|T	CD79B	2
tFL-3-cus	chr17|62007128|T	CD79B	NA
tFL-8	chr17|62006836|T	CD79B	NA
FL-8	chr17|62007129|T	CD79B	NA
tFL-19	chr17|62006798|C	CD79B	2
tFL-21	chr17|62006799|C	CD79B	2
tFL-38	chr17|62007480|C	CD79B	2
FL-12	chr3|38182641|C	MYD88	2
tFL-12	chr3|38182641|C	MYD88	2
tFL-19	chr3|38182641|C	MYD88	2
tFL-22	chr3|38182025|T	MYD88	NA
tFL-23	chr3|38182032|G	MYD88	2
FL-6	chr6|138200146|G	TNFAIP3	0
tFL-6	chr6|138200146|G	TNFAIP3	2
tFL-21	chr6|138200194|G	TNFAIP3	2
tFL-31	chr6|138192455|T	TNFAIP3	NA

Table 4-5. Genes mutated in NF-κB pathway. The first Sample ID column indicates the samples in which the mutations were detected, the Chromosome| Position |Altered-Base column indicates the coordinates of the mutations, the Gene column indicates the genes of the mutations, the CN column indicates the CN estimated in this gene from our previous study, NA is noted if we don't have the CN information.

Gene	Samples	% Loss/double Loss	% Double Loss	% Gain/Amplification	% Amplification
CARD11	FL (n=198)	0.000	0.000	0.242	0.030
	tFL (n=79)	0.025	0.000	0.392	0.089
TNFAIP3	FL (n=198)	0.263	0.056	0.000	0.000
	tFL (n=79)	0.342	0.101	0.000	0.000

Table 4-6. Frequency of CNAs affecting *CARD11* and T*NFAIP3* based on previously published data.

E. Mutations affecting miRNA

MicroRNAs are short non-coding RNAs that are involved in post-transcriptional regulation of gene expression. They affect both the stability and translation of mRNAs. The seed sequence is an essential region for the binding of the miRNA to the target mRNA. There are 422 microRNAs targeted by illumina's TruSeq exome enrichment kit. We used the BED file provided by Illumina, which has details on all of the target microRNAs regions, and calculated the coverage and depth of all the microRNAs in all our WES samples to confirm the microRNA capture performance. The average coverage and depth for all microRNAs in all samples is 96.70% and 80 respectively.

The only micro-RNA found to be mutated was *miR-142*, listed in Table 4-7. *MiR-142* is a hematopoietic-specific micro RNA precursor whose 3p and 5p arms are both functional and expressed at similar levels[41]. *MiR-142* was mutated in 3 of 12 tFL cases, two of which were shared with the corresponding FL and one was a tFL-unique mutation. Interestingly, all the mutations were located in the seed sequences (nucleotides 2-8) in Figure 4-2. The mutation (chr17|56408621|G) that was identified in 2 cases and affects miR-142-3p, was also detected in DLBCL by another group[42]. The other shared mutation affecting miR-142-5p is a new finding.

SampleID	Chr|Pos|VarAllele	Gene	VarFreq
tFL-4	chr17|56408621|G	MIR142	0.19
FL-8	chr17|56408621|G	MIR142	0.448
tFL-8	chr17|56408621|G	MIR142	0.471
FL-11	chr17|56408657|G	MIR142	0.086
tFL-11	chr17|56408657|G	MIR142	0.207

Table 4-7. Mutations in miR-142. The first column (Sample ID) indicates the samples in which the mutations were detected, the second column (Chromosome| Position |Altered-Base) indicates the coordinates of the mutations, the third column (Gene) indicates the genes affected, the fourth column (Var Freq) indicates the variant frequencies of the mutations.

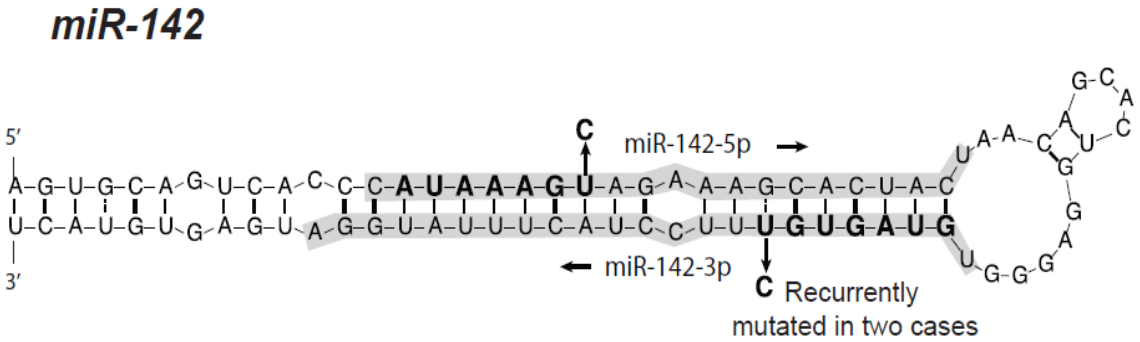

Figure 4-2. Domains/regions affected by mutations for miR-142.

F. Recurrently mutated genes in 3 datasets

To increase our analytical power and to gain a more comprehensive view of the genes that likely drive transformation, we combined our data with two published datasets for further analysis (Figure 4-3). We identified additional genes that tend to be associated with transformation, including *MYC, EBF1, IRF4, RPN1, SOCS1, SYNE1, SGK1, PIM1, EP300, BMP7, ETS1, SARDH, TAF1, FBXO11 and HIST1H1E* summarized in Table 4-8 below.

MYC had a tFL-unique mutation in only 1 of our samples but 4 samples in other datasets. *EBF1* had 4 tFL-unique mutations (2 FL-unique mutations) in 4 samples in other datasets. *IRF4* had 3 tFL-unique mutations in 3 samples in another dataset. *RPN1* had a tFL-unique mutation in only 1 of our samples and 2 samples in another dataset. *SOCS1* had tFL-unique mutations in 2 of our samples (2 shared mutations) and 2 samples in other datasets. *SYNE1* had a tFL-unique

mutation in only 1 of our samples (1 FL-unique mutation) and 2 samples in another dataset. *SGK1* had a tFL-unique mutation in only 1 of our samples (1 FL-unique mutation) and 2 samples in another dataset. *PIM1* had 4 tFL-unique mutations in 4 samples in another dataset (2 shared mutations). *EP300* had 2 tFL-unique mutations in 2 samples in another dataset (1 shared mutaiton). *BMP7* had 2 tFL-unique mutations in 2 samples in another dataset (1 shared mutaiton). *ETS1* and *TAF1* both had a tFL-unique mutation in 1 of our samples (1 shared mutation) and 1 sample in another dataset. *SARDH* had 2 tFL-unique mutations in 2 samples in another dataset. *FBX011* had 2 tFL-unique mutations 2 samples in other datasets. *HIST1H1E* had 1 tFL-unique mutation in 1 of our samples (1 shared mutation) and 3 samples in other datasets.

Overall, the recurrent mutations (Figure 4-3) identified in our dataset were highly concordant (92%) with the other two datasets.

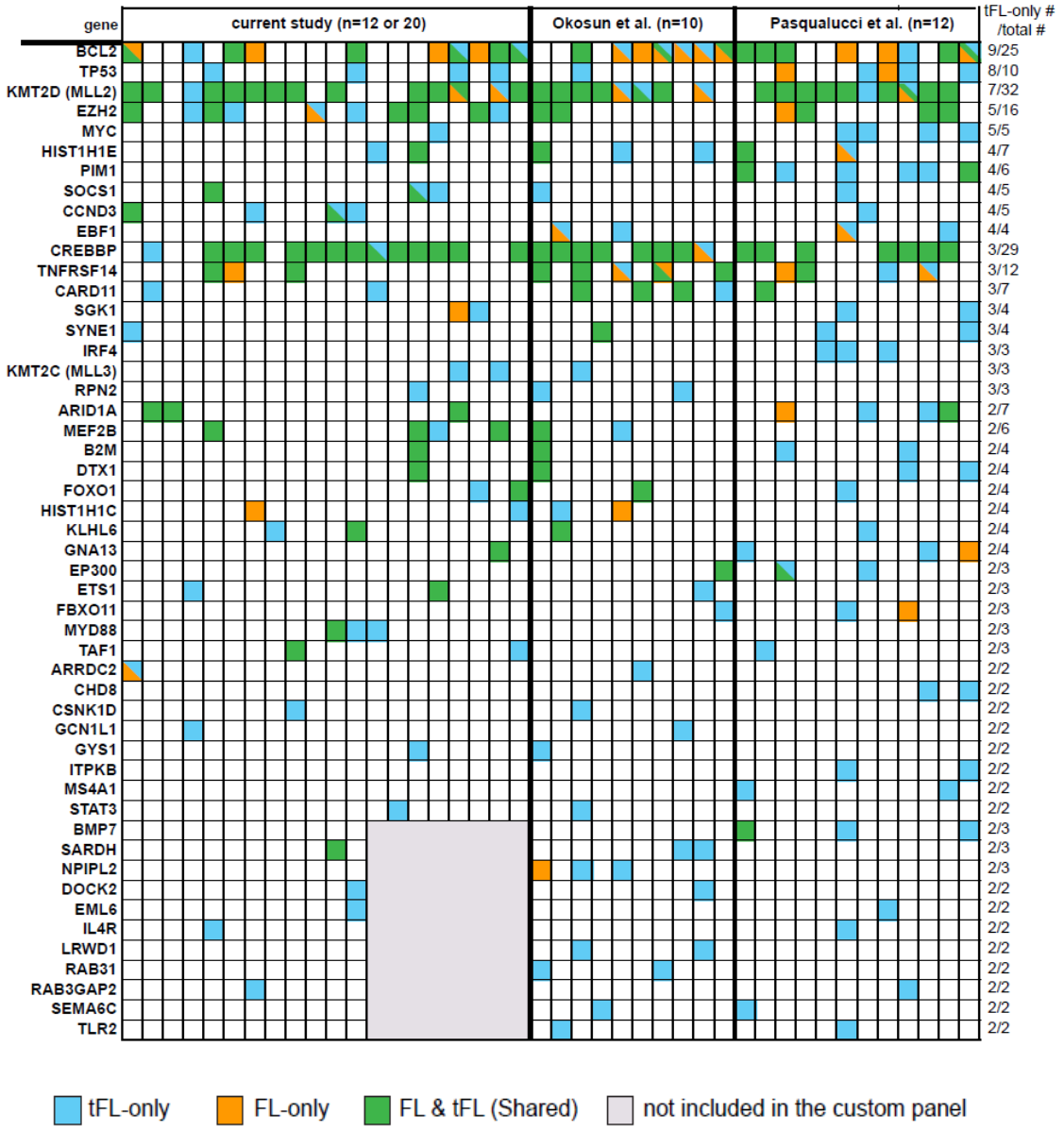

Figure 4-3. tFL-unique mutated genes found in more than two cases in 3 combined datasets. The data was summarized from a combine set of 42 FL and tFL paired samples. Only genes expressed in B cells are shown. Genes in bold were selected in custom capture panel. The color of each block represents the mutation type of the corresponding genes and cases.

Gene	Current study			Okosun et al.			Pasqualucci et al.			Mutations in 1 dataset	Mutations in 3 datasets
	FL-unique	tFL-unique	shared	FL-unique	tFL-unique	shared	FL-unique	tFL-unique	shared		
MYC	0	1	0	0	0	0	0	4	0	1	5
EBF1	0	0	0	1	2	0	1	2	0	0	6
IRF4	0	0	0	0	0	0	0	3	0	0	3
RPN1	0	1	0	0	2	0	0	0	0	1	3
SOCS1	0	2	2	0	1	0	0	1	0	4	6
SYNE1	0	1	0	0	0	1	0	2	0	1	4
SGK1	1	1	0	0	0	0	0	2	0	2	4
PIM1	0	0	0	0	0	0	0	4	2	0	6
EP300	0	0	0	0	0	1	0	2	1	0	4
BMP7	0	0	0	0	0	0	0	2	1	0	3
ETS1	0	1	1	0	1	0	0	0	0	2	3
SARDH	0	0	0	0	0	0	0	2	1	0	3
TAF1	0	1	1	0	0	0	0	1	0	2	3
FBXO11	0	0	0	0	1	0	1	1	0	0	3
HIST1H1E	0	1	1	0	2	1	1	1	1	2	8

Table 4-8. Additional genes that tend to be associated with transformation in combined datasets.

G. Pathway analysis

In addition to the NF-κB pathway described in ABC-like DLBCL, our previous CN analysis study revealed that the rCNAs affect in B-cell transcription factors, cell cycle regulation, and immune surveillance pathways in the transformation of FL.

Similarly, we found that mutations also commonly target genes involved in these same pathways (Table 4-9). For example, several regions with small deletions or amplifications were likely driven by B-cell transcription factors. Recurrent mutations affecting B-cell transcription factors were also identified, including mutation of MEF2B. Four of the cases harbored previously described *MEF2B* mutations[43] that block association with the co-repressor CABIN1, increasing transcriptional activity (3 cases, D83V; 1 case E73K; Figure 4-4).

We applied David Bioinformatics Resources 6.7 to identify pathways significantly enriched for mutations in our tFL WES mutations. Table 4-9 shows BIOCARTA and KEGG pathways

enriched in our list of mutations. The top 3 enriched pathways involved B-cell receptor signaling, IL-7 signaling, and JAK-STAT activation.

The Switch/Sucrose non-fermentable (SWI/SNF) is a nucleosome remodeling complex involved in chromatin remodeling. It is capable of altering the position of nucleosomes along DNA. In tFL, SWI/SNF members appear to be targets of both mutation and copy loss. *ARID1A* is mutated in 9% of tFL cases (3/35 in Table 4-10) and 15% of tFL cases (5/32) have a CNA affecting *ARID1A* including a small recurrent CN loss that encompassed *ARID1A* at 5% frequency (Table 4-11). Additionally, 1/35 tFLs had an *ARID1B* mutation and 1/12 cases had *ARID4B* and *ARID5B* mutation (Table 4-10). A small, rare, recurrent CNA affected *ARID1B*, but almost 23% of tFLs had a larger loss on chr6 that included *ARID1B* (Table 4-11). In addition to SWI/SNF family members, many genes involved in chromatin organization and modification were mutated in our cases (Figure 4-3), as was also observed by others. *EZH2* and *MLL3* mutations were described above.

We also identified frequent CNAs and mutations affecting genes that regulate B-cell migration and AKT/mTOR pathway activation. This pathway has recently been shown to be mutated in DLBCL and cell lines[26,27,44,45] . Figure 4-5 depicts the pathway with S1P interacting with its receptors. S1PR2 (a G-protein-coupled receptor) signals through GNA13 (G-protein), which interacts with ARHGEF1 (a RHO guanine-nucleotide exchange factor) and RHOA. Interruption of this pathway promotes migration of B-cells out of the GC and activation of AKT. All three genes were mutated in our dataset (Table 4-12). The effect of S1P signaling through S1PR1 is the opposite of S1PR2 signaling, and CD69 interacts with and inhibits S1PR1. No mutation in *CD69* or *S1PR1* was detected by us or others, but analysis of our CNA data indicated CNL involving *CD69* in 7.6% of tFLs (Table 4-11), which could partially reduce the normal inhibitory influence on S1PR1. Indeed, copy loss affecting *S1PR2*, *GNA13*, *ARHGEF1*, *P2RY*8, and/or *CXCR4* loci occurs more frequently in tFL compared to FL (21.5% vs 7.5% in Table 4-13).

There was a single case that had 3 separate FL-tFL shared mutations in *ARHGEF1*. A second case had a shared mutation in *GNA13*. Additionally, the Pasqualucci dataset identified *GNA13* mutations in 3 cases (2 tFL-unique mutations and 1 FL-unique mutation in Figure 4-3).

AKT pathway activation can be enhanced by concurrently activating mutations affecting the mTOR pathway. Two cases harbored a shared FL-tFL mutation affecting conserved residues within the switch 1 region of *RRAGC* (Table 4-12), and other two studies identified mutations in 3 cases (2 FL-unique mutations and 1 shared mutation in Table 4-14) affecting the same region. Additional genes that affect the PI3K/AKT/mTOR pathway were mutated in our dataset including *TSC2, PIK3R1,* and *PTEN* (Table 4-12).

Category	**Term**	**PValue**	**Genes**	**FDR (%)**
KEGG	hsa04662:B cell receptor signaling pathway	0.003698	CARD11, CD19, FCGR2B, LILRB3, PPP3R1, CD79B, PIK3R1	4.112991
BIOCARTA	h_il7Pathway:IL-7 Signal Transduction	0.011012	NMI, BCL2, CREBBP, IL2RG	12.26147
KEGG	hsa04630:Jak-STAT signaling pathway	0.012082	STAT6, IL2RA, CCND3, IL4R, CREBBP, IL2RG, PRL, PIK3R1, STAT2	12.87195
BIOCARTA	h_pmlPathway:Regulation of transcriptional activity by PML	0.013431	CREBBP, TP53, RARA, RB1	14.76345
BIOCARTA	h_rarrxrPathway:Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription in carcinoma cells	0.013431	NCOA2, RARA, NCOR2, POLR2A	14.76345
BIOCARTA	h_telPathway:Telomeres, Telomerase, Cellular Aging, and Immortality	0.02238	BCL2, TP53, RB1, POLR2A	23.46293
KEGG	hsa05222:Small cell lung cancer	0.026054	BCL2, TP53, RB1, PTEN, PIK3R1, TRAF3	25.86317

BIOCARTA	h_pcafpathway:The information-processing pathway at the IFN-beta enhancer	0.02906	HMGB1, CREBBP, POLR2A	29.41808
KEGG	hsa05340:Primary immunodeficiency	0.032244	CD19, TAP1, IL2RG, RFXAP	31.03271
KEGG	hsa05215:Prostate cancer	0.032409	BCL2, CREBBP, TP53, RB1, PTEN, PIK3R1	31.16593
KEGG	hsa05214:Glioma	0.036878	TP53, CAMK2B, RB1, PTEN, PIK3R1	34.68509
BIOCARTA	h_carm-erPathway:CARM1 and Regulation of the Estrogen Receptor	0.053205	CREBBP, SPEN, NCOR2, POLR2A	47.58013
BIOCARTA	h_il4Pathway:IL 4 signaling pathway	0.053409	STAT6, IL4R, IL2RG	47.71387
BIOCARTA	h_egfr_smrtePathway: Map Kinase Inactivation of SMRT Corepressor	0.053409	MAP3K1, RARA, NCOR2	47.71387

Table 4-9. Mutated genes classified in KEGG/BIOCARTA pathways.

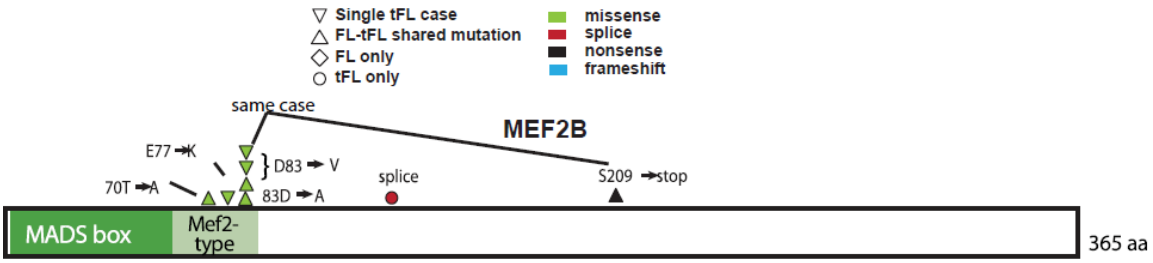

Figure 4-4. Domains/regions affected by mutations for MEF2B.

Sample ID	Chr|Pos|VarAllele	Gene	MutType	Var Freq	CN
FL-2	chr1|27106915|T	ARID1A	stopgain	0.055	NA
tFL-2	chr1|27106915|T	ARID1A	stopgain	0.232	2
FL-3-cus	chr1|27023162|G	ARID1A	nonsynonymous	0.073	NA
tFL-3-cus	chr1|27023162|G	ARID1A	nonsynonymous	0.042	NA
FL-27	chr1|27089478|T	ARID1A	stopgain_	0.141	2
tFL-27	chr1|27089478|T	ARID1A	stopgain	0.451	2
FL-6	chr10|63759897|A	ARID5B	nonsynonymous	0.519	2
tFL-6	chr10|63759897|A	ARID5B	nonsynonymous	0.222	2
FL-6	chr10|63759918|G	ARID5B	nonsynonymous	0.494	2
tFL-6	chr10|63759918|G	ARID5B	nonsynonymous	0.258	2
FL-10	chr6|157528317|A	ARID1B	stopgain	0.413	2
FL-10-cus	chr6|157528317|A	ARID1B	stopgain	0.34	2
tFL-10	chr6|157528317|A	ARID1B	stopgain	0.517	2
tFL-10-cus	chr6|157528317|A	ARID1B	stopgain	0.38	2
FL-22	chr1|235420509|C	ARID4B	nonsynonymous	0.357	NA
tFL-22	chr1|235420509|C	ARID4B	nonsynonymous	0.27	NA

Table 4-10. Mutations in ARID1A, ARID1B, ARID4B and ARID5B.

Gene	Samples	% Loss/double Loss	% Double Loss	% Gain/Amplification	% Amplification
ARID1A	FL (n=198)	0.096	0.000	0.005	0.000
	tFL (n=79)	0.152	0.000	0.000	0.000
ARID1B	FL (n=198)	0.162	0.000	0.000	0.000
	tFL (n=79)	0.228	0.013	0.000	0.000
CD69	FL (n=198)	0.015	0.005	0.157	0.025
	tFL (n=79)	0.076	0.000	0.152	0.013

Table 4-11. Frequency of CNAs affecting *ARID1A*, *ARID1B* and *CD69* based on previously published data.

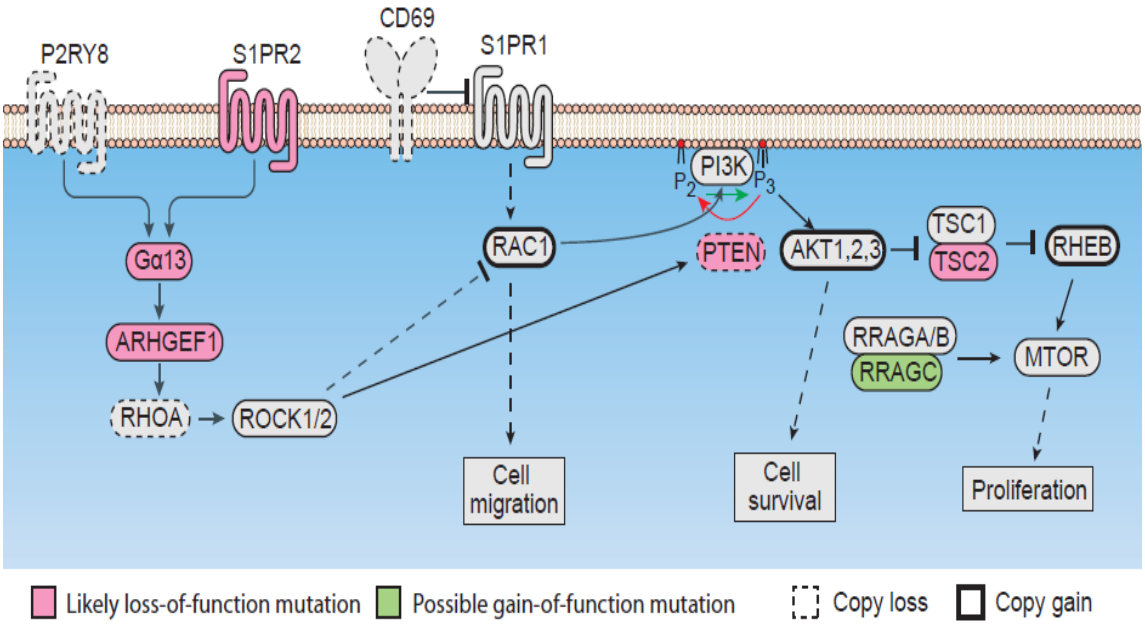

Figure 4-5. Abnormalities of the S1PR1 and S1PR2 pathway are associated with FL transformation. Black arrows and bar-headed lines indicate activation or inhibition, respectively; dotted lines indicate an indirect effect. Different color and shape of border lines were used to mark the types of mutations or copy changes observed in our case series (n=35), within which, 80% of the cases carry at least one of the genetic abnormalities.

Sample ID	Chr|Pos|VarAllele	Gene	MutType	Var Freq	CN
tFL-22	chr19|10334805|T	S1PR2	stopgain	0.223	NA
FL-22	chr19|42398308|A	ARHGEF1	nonsynonymous	0.31	NA
tFL-22	chr19|42398308|A	ARHGEF1	nonsynonymous	0.25	NA
FL-22	chr19|42398565|+A	ARHGEF1	frameshift_insertion	0.353	NA
tFL-22	chr19|42398565|+A	ARHGEF1	frameshift_insertion	0.202	NA
FL-22	chr19|42406962|A	ARHGEF1	nonsynonymous	0.3	NA
tFL-22	chr19|42406962|A	ARHGEF1	nonsynonymous	0.333	NA
FL-29	chr17|63010412|T	GNA13	nonsynonymous	0.199	2
tFL-29	chr17|63010412|T	GNA13	nonsynonymous	0.43	2
FL-6	chr1|39322649|G	RRAGC	nonsynonymous	0.278	2
tFL-6	chr1|39322649|G	RRAGC	nonsynonymous	0.234	2
FL-8	chr1|39322697|A	RRAGC	nonsynonymous	0.34	NA
tFL-8	chr1|39322697|A	RRAGC	nonsynonymous	0.446	NA
tFL-4	chr10|89624275|T	PTEN	stopgain	0.333	1
tFL-24	chr16|2104347|A	TSC2	nonsynonymous	0.23	2
tFL-9	chr5|67576825|A	PIK3R1	nonsynonymous	0.318	2
tFL-16	chr5|67591106|G	PIK3R1	nonsynonymous	0.1	2

Table 4-12. Mutations in S1PR2 pathway and PI3K/AKT/mTOR pathway genes.

Samples	% Loss/double Loss	% Gain/Amplification
FL (n=198)	0.076	0.359
tFL (n=79)	0.215	0.405

Table 4-13. Frequency of CNAs affecting any of the following genes:GNA13, ARHGEF1, P2RY8,S1PR2, and/or CXCR4 on previously published data.

Sample	Gene Symbol	Chromosome	Genomic position	ref|variant	Mutation type	Data resource
S4_FL	RRAGC	chr1	39322646	C|T	missense	Oko
S4_tFL	RRAGC	chr1	39322646	C|T	missense	Oko
S7_FL3	RRAGC	chr1	39322723	G|T	missense	Oko
23	RRAGC	chr1	39322649	A|G	missense	Pas

Table 4-14. Mutations in RRAGC in other sequencing studies. The first column (Sample ID) indicates the samples that the mutations were detected, the second column (Gene Symbol) indicates the genes of the mutations, the third column (Chromosome) indicates the chromosome of the mutations, the fourth column (Genomic position) indicates the position of the mutations, the fifth column (ref|variant) indicates the reference and variant of the mutations, the sixth column (Mutation type) indicates the mutation types, the seventh column (Data resource) indicates the datasets of the mutations were detected.

H. Domains and regions affected by mutations

B-cell receptor (BCR) signaling complex includes both CD79 and surface immunoglobulin. CD79 generates a signal when BCR recognize antigens. *CD79B* is one of the two distinct chains of CD79 and it has an ITAM, which is a conserved sequence of 4 amino acids. The ITAM plays an important role in signal transduction. We detected 6 mutations in *CD79B*. One mutation was in the IG-like V-type domain. Three mutations were between the transmembrane domain and ITAM domain. 2 tFL-unique mutations were within ITAM domain. The other two studies also found 2 shared mutations in the ITAM domain. Figure 4-6 shows all the mutations.

After antigen stimulation in normal B cells, the BCR sends signal to activate the NF-κB pathway, which among other effects promotes survival. *CARD11* is a signaling scaffold protein that forms a complex with BCL10 and MALT1, which leads to K63-ubiquitination of MALT1 and eventually to the activation of the IKK complex, which activates the NF-κB pathway. We detected 8 mutations in *CARD11* (Table 4-3). 2 of them were tFL-unique mutations, the rest 6 were found in single tFL samples. The other two studies also found 4 shared mutations and 1 tFL-unique mutation (Figure 4-3). 9 out of 12 mutations were within the coiled-coil domain, 3 of

them were within the caspase activation and recruitment domain (CARD), and no mutations were found in the inhibitory domain, PDZ domain, SRC homology 3 domain (SH3) or guanylate kinase-like domain (Figure 4-7).

RRAGC is a protein that is encoded by the *RRAGC* gene. This protein is a monomeric guanine nucleotide-binding protein and forms a heterodimer with RRAGA or RRAGB. It is primarily localized in the cytoplasm. When RRAGC binds GTP or GDP, it acts as a switch for downstream pathways. As we mentioned in the previous section, we detected 2 shared mutations in our study (Table 4-12). 3 shared mutations and 1 FL-unique mutations were detected in the other two datasets (Table 4-14). All of the mutations were within P-loop containing nucleoside triphosphate hydrolase domain as depicted in Figure 4-8.

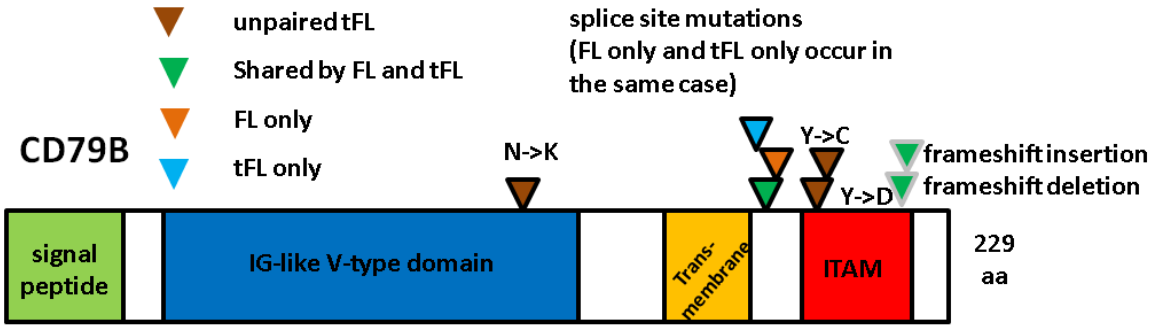

Figure 4-6. Domains/regions affected by mutations for CD79B. The mutations from our case series have black outline, while those from the other two published datasets have gray outline.

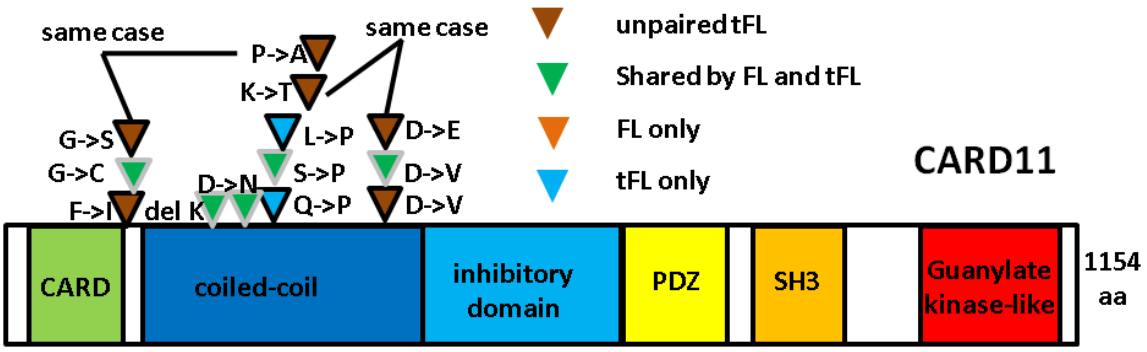

Figure 4-7. Domains/regions affected by mutations for CARD11. The mutations from our case series have black outline, while those from the other two published datasets have gray outline.

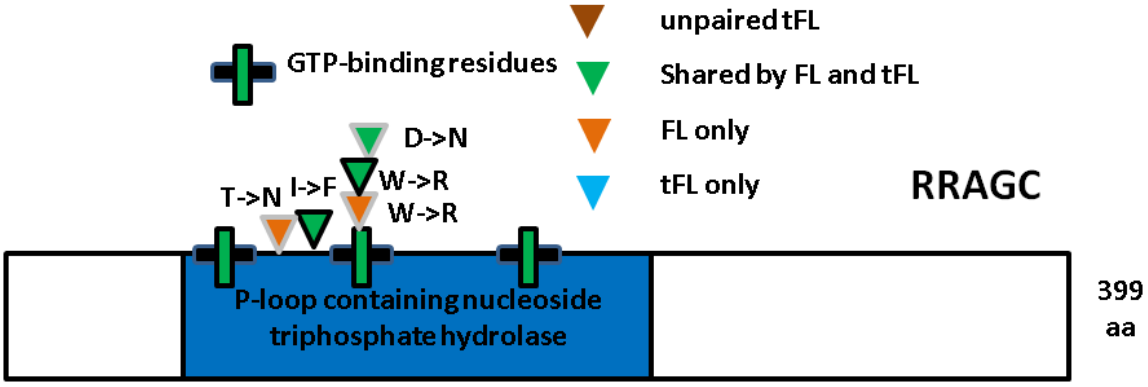

Figure 4-8. Domains/regions affected by mutations for RRAGC. Sites of GTP binding are noted. The mutations from our case series have black outline, while those from the other two published datasets have gray outline.

I. Subclonal mutations

Mutations present in small subclones in FL and which later became dominant clones in corresponding tFL may help to drive the transformation process. We investigated all our FL and tFL paired samples and revealed a number of genes in which the VAF increased in the tFL (Table 4-15). Several of the genes that were identified as subclonal in the FL such as *CARD11*, *CD79B*, *EZH2*, *FOXO1*, *HIST1H1E*, and *MYD88* are recurrently mutated in our dataset and likely important for disease progression. *S1PR2*, which is involved in germinal center B-cell

confinement[46] was also detected at low VAF in the FL sample but emerged as clonal in the transformed sample.

Sample ID	# Variant Reads	# Total Reads	%VAF	% Tumor content	% Adjusted VAF	CN	chr|pos|var	Gene
FL-8	1	144	0.69%	88%	0.79%	NA	chr2|191064806|G	C2orf88
tFL-8	42	146	28.77%	79%	36.51%	NA		
FL-8	1	116	0.86%	88%	0.98%	NA	chr1|220379264|G	RAB3GAP2
tFL-8	50	174	28.74%	79%	36.47%	NA		
FL-23-1	6	614	1.00%	84%	1.19%	3	chr7|2979501|G	CARD11
tFL-23	108	829	13.00%	26%	49.43%	3		
FL-23-2	7	711	0.99%	84%	1.18%	NA	chr7|2979501|G	CARD11
tFL-23	108	829	13.00%	26%	49.43%	3		
FL-23-1	12	1194	0.99%	84%	1.18%	3	chr6|26156797|T	HIST1H1E
tFL-23	116	1052	10.99%	26%	41.79%	3		
FL-23-2	11	1161	0.99%	84%	1.18%	NA	chr6|26156797|T	HIST1H1E
tFL-23	116	1052	10.99%	26%	41.79%	3		
tFL-10	2	180	1.11%	90%	1.24%	5	chr1|85736474|T	BCL10
FL-10	125	193	64.77%	81%	79.67%	NA		
tFL-24	9	910	0.98%	77%	1.27%	2	chr5|138260327|C	CTNNA1
FL-24-1	111	464	24.03%	92%	26.09%	2		
tFL-24	9	910	0.98%	77%	1.27%	2	chr5|138260327|C	CTNNA1
FL-24-2	128	532	23.99%	92%	26.05%	NA		
tFL-9	1	87	1.15%	84%	1.37%	1	chr22|24530363|G	CABIN1
FL-9	23	150	15.33%	24%	64.14%	2		
FL-8	2	156	1.28%	88%	1.46%	NA	chr6|121562671|A	TBC1D32
tFL-8	43	152	28.29%	79%	35.90%	NA		
FL-8	1	74	1.35%	88%	1.54%	NA	chr5|96507052|C	RIOK2
tFL-8	18	64	28.13%	79%	35.70%	NA		
FL-11	1	131	0.76%	49%	1.54%	2	chr7|148506437|A	EZH2
tFL-11	18	95	18.95%	27%	69.41%	2		
FL-22	1	109	0.92%	58%	1.59%	NA	chr1|233515277|G	KIAA1804
tFL-22	15	90	16.67%	43%	39.13%	NA		
FL-11	1	121	0.83%	49%	1.69%	2	chr5|79354080|A	THBS4
tFL-11	30	99	30.30%	27%	110.99%	2		
FL-22	2	150	1.33%	58%	2.29%	NA	chr19|10334805|T	S1PR2
tFL-22	29	130	22.31%	43%	52.37%	NA		
FL-28	2	216	0.97%	36%	2.68%	NA	chr13|41240279|A	FOXO1

tFL-28	27	182	14.98%	37%	40.38%	2		
FL-22	1	64	1.56%	58%	2.69%	NA	chr3|38182025|T	MYD88
tFL-22	7	49	14.29%	43%	33.54%	NA		
FL-11	1	70	1.43%	49%	2.91%	2	chr1|186036995|T	HMCN1
tFL-11	16	78	20.51%	27%	75.13%	2		
FL-11	1	69	1.45%	49%	2.95%	2	chr1|178861391|T	RALGPS2
tFL-11	8	73	10.96%	27%	40.15%	2		
FL-11	2	119	1.68%	49%	3.41%	2	chrX|53631710|C	HUWE1
tFL-11	16	85	18.82%	27%	68.94%	1		
FL-2	1	106	0.94%	26%	3.56%	NA	chr2|97370071|C	FER1L5
tFL-2	24	95	25.26%	32%	78.45%	2		
FL-9	1	113	0.88%	24%	3.68%	4	chr18|47801358|T	MBD1
tFL-9	56	223	25.11%	84%	30.00%	6		
FL-3	2	167	1.20%	32%	3.79%	2	chr8|139732980|C	COL22A1
tFL-3	8	65	12.31%	43%	28.56%	2		
FL-2	1	85	1.18%	26%	4.47%	NA	chr4|90035447|T	TIGD2
tFL-2	6	59	10.17%	32%	31.58%	1		
FL-11	2	85	2.35%	49%	4.78%	3	chr18|66721328|A	CCDC102B
tFL-11	12	109	11.01%	27%	40.33%	7		
tFL-31	3	108	3.14%	58%	5.41%	NA	chr18|60985880|A	BCL2
FL-31	27	151	17.88%	32%	55.53%	NA		
tFL-22	2	85	2.35%	43%	5.52%	NA	chr13|111372118|A	ING1
FL-22	13	77	16.88%	58%	29.10%	NA		
FL-11	4	110	3.64%	49%	7.40%	2	chr13|39262875|T	FREM2
tFL-11	21	85	24.71%	27%	90.51%	2		
FL-2	3	127	2.36%	26%	8.94%	NA	chr7|117431643|T	CTTNBP2
tFL-2	14	114	12.28%	32%	38.14%	4		
FL-3	2	53	3.77%	32%	11.89%	2	chr19|38834308|T	CATSPERG
tFL-3	5	19	26.32%	43%	61.07%	2		
FL-2	6	191	3.14%	26%	11.89%	NA	chr2|210745783|G	UNC80
tFL-2	31	147	21.09%	32%	65.50%	2		
FL-3	3	67	4.48%	32%	14.13%	2	chr17|62007128|T	CD79B
tFL-3	6	30	20.00%	43%	46.40%	2		
FL-2	2	53	3.77%	26%	14.28%	NA	chr10|53458823|A	CSTF2T
tFL-2	11	55	20.00%	32%	62.11%	0		

Table 4-15: Mutations present in subclones with increase VAF in tFL. Red indicates subclones, green indicates major clones.

J. Discussion

The goal of our study is to understand which genetic events drive the indolent FL to aggressive tFL, so that we can identify which patients are in a high-risk group, determine the best treatment plan, and target the promising abnormalities for therapies. We performed WES on 12 paired FL and tFL samples and used a custom capture panel on 23 additional cases with deep sequencing. We identified recurrently mutated genes and pathways from the sequencing study, and integrated CNA information from our previous study.

Our integrated data strongly suggests that the activation of NF-κB pathway is critically important for the transformation of some FL cases. We have identified ABC-like and GCB-like tFL as likely to be associated with certain genetic abnormalities. For example, CARD11, MYD88, and TNFAIP3 are more frequently mutated in de novo ABC-DLBCL and are associated with NF-κB pathway activation. They are expected to have a similar influence on tFL. The corresponding CNA data indicated that some of the mutations cooperate with CNAs to generate homozygous alterations that amplify their functional consequences. Other studies reported that the ABC-like and GCB-like divergence may be present already at the FL stage; however, there is no evidence that the ABC-like FL is more likely to undergo transformation or have worse survival but the number of cases studied is low.

Our pathway analysis found that the commonly targeted genes in transformation encode B-cell transcription factors and proteins involved in cell cycle regulation, immune surveillance, JAK-STAT activation, and the p53 pathway. These pathways highly overlap with the pathways we previously identified in CNAs.

One of the characteristics of tFL is the loss of GC confinement. We detected mutations or CN losses in S1PR2, GNA13, ARHGEF1, P2RY8, CXCR4, and CD69. These genes are involved in the

S1PR1 and S1PR2 pathway, which is associated with FL transformation. The inactivation of the S1PR2 pathway or activation of the S1PR1 pathway promotes migration of B-cells out of the GC and activation of the AKT pathway. This may be a critical event in the change from a follicular to a diffuse state in the process of transformation.

The activation of the AKT pathway likely provides a survival signal for the B-cells that move out of the GC and hence lose the supportive microenvironment of the GC. The concomitant activation of the mTOR pathway may enhance the effect of AKT activation. We identified mutations in several genes that are involved in mTOR pathway activation. RRAGC heterodimerizes with RRAGA or RRAGB and can recruit mTORC1 through RAPTOR to the lysosomal membrane, where mTORC1 can be activated by RHEB. Activity of the heterodimer is higher when GDP is bound to RRAGC. The switch 1 region of the RRAGC yeast ortholog Gtr2p undergoes a conformation change depending upon GTP or GDP binding; thus, mutations of this region may affect RRAGC activity and hence mTORC1 activation. TSC2 is a GTPase-activating protein (GAP) that inhibits mTORC1 activation by promoting the conversion of RHEB-GTP to RHEB-GDP. Additional genes that were mutated in PI3K/AKT/mTOR signaling pathway included PIK3R1, INPP5F and PTEN..

There are several mutated genes involved in the JAK-STAT pathway. The mutations in SOCS1 (Figure 4-9) and STAT6[47] have been reported by other studies. STAT3 is another mutated gene that has not been discussed. We identified 4 mutations in STAT3, they all affected amino acids that are resolved within the STAT3 crystal structure. The K658N substitution within the SH2 domain is identical to an activating mutation previously identified by other studies. According to the gain of function characteristic of this mutation, we would expect the tumor to shift toward ABC-tFL and our GEP data indicate that this case is unclassifiable. Another K340T substitution may act in an opposite way: it likely represents a loss of function mutation and may interfere

with binding to the target sequence. Because the mutant protein retains the ability to dimerize, it likely will have a dominant-negative effect. This mutation was found in a GCB-like tFL as it may act to block further differentiation.

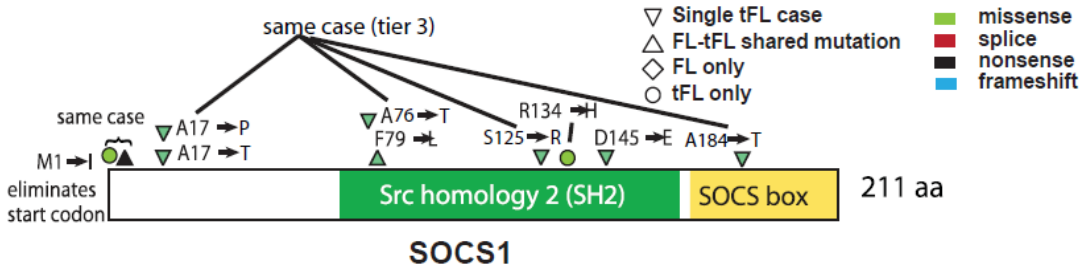

Figure 4-9. Domains/regions affected by mutations for SOCS1.

Chromatin structure deregulation and disruption is important in the pathogenesis of FL and many mutations affecting chromatin modifiers have been identified. SWI/SNF family members, which are involved in chromatin remodeling, are often mutated in a variety of cancers. In our study, ARID1A and ARID1B were recurrently mutated and also identified in CN loss in tFL. Additional genes were mutated that affect chromatin organization and modification. For example, mutations in CREBBP and MLL2, which were highly recurrent, are likely to occur as early events in FL. in contrast. EZH2, another frequently mutated gene, which is mutated almost twice as frequently in tFL compared to FL, may tend to appear later in the disease and may promote transformation. There is evidence of a positive feedback loop between EZH2 and MYC, via a micro-RNA network, and increase in MYC was largely confined to tFLs. The availability of specific inhibitors for EZH2 makes it a promising target for tFL treatment.

MiR-142 mutation is a unique finding in the study. The mutations were detected within the seed sequence in 3 out of 12 tFL samples. We applied TargetScan 5.2 to mir-142 to identify

possible targets; it showed differences in the predicted targets between mutant and WT. All the mutations are expected to alter complementarity to target genes. CYLD, a negative regulator of NF-κB, was listed as a top predicted target for mutant mir-142-5p by TargetScan, and this mutation is a novel finding. One shared mutation affecting mir-142-3p is exactly the same as the one reported by another group[42]. They found that this mutation results in both gain and loss of function. Novel targets sites for the mutant miR in the *ZEB2* 3' UTR may lead to its down-regulation.

There were only a few mutations appearing uniquely in tFLs. Therefore, the transformation likely occurs due to a combination of several genetic changes that cooperate together to push the FL to the transformation. A larger dataset is required to determine which combinations are important in transformation.

CHAPTER V

FUTURE WORK

When an immune response occurs, B cells migrate within follicles and develop a germinal center with large proliferating cells. This event initiates massive clonal expansion, somatic hypermutation and class switch recombination. These events increase genomic instability and predispose GCB-cells to the development of lymphoma. Our sequencing study has demonstrated the genetic abnormalities that contribute to FL and tFL. While cancer has been viewed as the result of progressive accumulation of genetic and epigenetic abnormalities, it is known that B cell differentiation is regulated through the expression of transcription factors along with epigenetic modulation in the germinal center. In our study, we have identified multiple mutated genes that may contribute to epigenetic alterations. For example, in *EZH2* (encoding a histone-lysine N-methyltransferase), the mutation usually occurs at only one position and is a gain-of-function mutation that enhances histone methylation at histone H3K27, an inhibitory mark. Mutation of *MLL2* (lymphoid leukemia 2) was detected as loss–of-function and decreases histone methylation at histone H3K4, an activating mark. Certain genes can be turned off by both enhanced histone H3K27 methylation and decreased H3K4 methylation, including genes inhibiting cell proliferation and differentiation. TET2 (Tet methylcytosine dioxygenase 2) can convert 5-methylcytosine to 5-hydroxymethylcytosine and lead to eventually to DNA demethylation. All the information suggests that aberrant DNA methylation occurs simultaneously with genetic abnormalities in FL and tFL development. Aberrant DNA methylation in the right context of the genes, can lead to gene silencing, e.g. hypermethylation of CpG islands of DNA sequence in tumor cells could result in the silencing of tumor suppressors. Within the context of chromatin, gene activation or inactivation is highly dependent on the methylation in the tail lysine residues of histone proteins. Therefore, there is an interaction between DNA methylation and histone modification. These modifications follow different chemical reactions and have different groups of enzymes involved, but changes in one may

impact changes in the other. Histone modifications are actually more complex and include acetylation, phosphorylation, and ubiquitylation. In future studies, we can apply reduced representation bisulfite sequencing (RRBS) analysis to investigate genome-scale alterations in DNA methylation. We can also apply chromatin immunoprecipitation sequencing (ChIP-seq) for global analysis of histone modifications. ChIP-sequencing could help identify the genes affected by the mutation in histone modifying genes. By correlating the results with our previous gene expression data, we can identify genes that are down-regulated by aberrant DNA methylation or chromatin modification and the pathways involved. If we integrate our gene expression data, CNV data, epigenetic and somatic mutation data, we should be able to provide an illuminating insight into the genome, transcriptome, and epigenome of FL and tFL, demonstrate connections and interaction among them, and identify the mechanisms relevant to FL and tFL development.

NGS has provided an opportunity to fully describe the spectrum of mutations that contribute to diseases. The majority of sequencing studies are focused on somatic mutations that are located in coding regions of the genome and on distinguishing driver mutations from passenger mutations among these somatic mutations. It is a very efficient strategy. In our study, we first concentrated on protein-coding regions, and then selected recurrently mutated genes for further analysis. However, we and others have noticed that most somatic mutations are actually located in non-coding regions which are often excluded from WES analysis, but may still play a role in the disease process. For example, mutation of the regulatory region of the TERT gene has been found in malignant melanoma and may be important in its pathogenesis. RNA has traditionally been considered as a messenger between DNA and protein but, recent studies indicate that RNA is involved in the regulation of genome organization and gene expression. Additionally, mRNA splice sites, UTR regulation elements, promoters, transcription factor (TF)

binding sites, enhancers and noncoding RNAs (ncRNAs) could be functionally important in the non-coding regions. Among the non-coding regions, the majority of the genomes are actually transcribed into ncRNAs that include several families of small RNAs such as the microRNA family and the Piwi family of RNAs. More interestingly, there are long ncRNAs (lncRNAs) that appear to control various levels of gene expression and potentially are involved in human disease. It has been reported that lncRNA functions as the interface between DNA and specific chromatin remodeling activities. For example, the expression level of the lncRNA *HOTAIR* is used to predict metastasis and survival in breast cancer[48]. The increasing expression of the same lncRNA in epithelial cancer cells has also been reported to induce Polycomb repressive complex 2 (PRC2) and lead to alter histone H3K27 methylation and increase cancer invasiveness and metastasis[48]. As another example, lncRNAs recruit Polycomb group (PcG) complexes to target genes and regulate the activity of PcG protein[49]. lncRNA regulation is likely dependent on direct interactions with PcG proteins such as EZH2, SUZ12, and CBX . *MALAT1* is another ncRNA whose expression is also associated with metastasis and affects survival in lung cancer[50]. All these studies suggest that lncRNAs participate in epigenome alterations and that they may play a vital role in disease. According to the ample evidence for important roles of ncRNA, we can extend our current study to non-coding regions especially ncRNAs and explore the abnormalities that might contribute to FL and tFL.

WTS can be used for the analysis of ncRNA alone but a global analysis of non-coding regions would require WGS and the availability of corresponding normal DNA is critical. We will also have to develop a pipeline that is specifically designed for mutation identification in non-coding regions and for the analysis of large structural alteration. Instead of detecting nonsynonymous variants, we would focus on variants within ncRNAs, enhancers, and mRNA promoters and TF binding site. It has been reported that variants identified in TF binding site are related to cancer

progression[51,52]. We can also look into is how mutations in regulatory regions of the genome might play a role in the FL and tFL development. For example, the structural alteration and mutations in the first exon-intron region of *BCL6* are common in B-cell lymphoma[53]. This less than 3 kb region has been shown to contain negative regulatory elements, and the first intron is found to be commonly deleted in B-cell lymphoma as a result of chromosomal rearrangements. Our current study is concentrated on recurrently mutated genes in coding regions, according to the above evidence, we can extend our study to mutations in regulatory regions that can downregulate transcription of important genes.

miRNA s are another potential target for investigation. Variants in miRNA sequence could alter binding specificity, therefore leading to the alterations of expression and translation of target mRNA[54]. We have found mutations in only one miRNA but a more extensive study is needed to assess the importance of miRNA mutation in FL/tFL. Similarly, mutations of critical miRNA binding sites on important genes should also be evaluated.

Chromatin open or active regions are considered most likely to contain key regulatory elements[55]. Therefore, these regions can be potential targets for investigation as well. Since there may not be clear and uniform information about these regions, particularly in the cell type of interest, the challenge during pipeline design can be that the variants within these regions are more difficult to annotate and interpret than the variants within amino acid coding regions. The other challenge might be in locating databases to use. Since non-coding regions are not as well-studied as coding regions, we probably will need to collect different databases from various resources and reformat them into a convenient format to apply to the pipeline.

KMT2D(MLL2), *CREBBP*, *BCL2*, *EZH2*, *MEF2B*, *TNFRSF14*, *CARD11* and *CCND3* are the top recurrent genes in our study. Although none of them is unique to the transformation, the

functional study of these mutants could provide insight on how they contribute to the tumor development.

In our sequencing study, the very limited number of frozen or fresh tissue samples has been a limitation. However, there are much larger numbers of formalin-fixed paraffin-embedded (FFPE) samples in tissue banks, and they are often well-characterized with histological, immunophenotypical and follow-up clinical data. We have set up robust pipelines for mutation detection in frozen/fresh samples and for reliable downstream analysis. We would like to apply them to FFPE samples. As it is commonly known that the DNA may be highly fragmented in FFPE samples and errors can introduced into sequencing due to formalin fixation, the challenge will be to improve the quality of DNA from FFPET and to remove the artifacts introduced by formalin fixation.

For the experimental part, there are many available methods designed specifically for FFPE samples. For example, the QIAamp DNA FFPE Tissue Kit uses special lysis conditions to overcome inhibitory effects caused by formalin crosslinking of nucleic acids while releasing DNA from tissue sections. Another possible method for FFPE samples would be water-soluble bifunctional catalysts, which can enhance the removal and decrosslinking of adducts from RNA and DNA bases[56]. Also, our laboratory has developed a modified Qiagen protocol for extracting DNA from FFPE samples resulting in higher yield of high quality DNA.

For the analysis part, we would like to add additional filtering steps to remove artifacts in our current pipelines. Because there are many publications[57,58] about common artifacts in FFPE samples, we should be able to apply their methods or adjust their methods to estimate the threshold VAF at which a variant can be trusted not to be artifactual. Our current pipeline will be robust enough to detect the majority of real variants. Our main focus will be on how to detect

true low frequency variants. We assume the artifacts tend to occur at certain locations instead of randomly[57]. In order to detect the hidden patterns, we will investigate the difference between the reliable artifacts and variants around their neighboring region, for example, the proportion of certain sequences, and the frequency of same variants. Statistical methods will be applied to measure the differences.

In our machine learning prediction study, the SIFT score in the complex feature set improved the training model and provided noticeably better prediction. However, SIFT score cannot predict every single variant (SIFT can predict most of the variants and have the best AUC score compared to other related programs[59]), therefore, we sacrificed some of our training dataset, which might impair the performance of the training model. Also, we cannot predict the variants that do not have a SIFT score in the testing dataset or real dataset. To address this limitation, we would like to assign comparable scores to the variants that SIFT score is unable to predict based on its own theory. The rationale[33] applied to SIFT score prediction is based on whether an amino acid substitution affects the protein structure. If the substituted amino acid doesn't have a similar property, it tends to be predicted as deleterious. For example, if a position located in the regions contain only hydrophobic amino acids, the SIFT score would assume this position can only contain amino acids with a hydrophobic character. Therefore, if the amino acid changes to any other one with hydrophilic character, this change would tend to be predicted as deleterious. In contrast, if this amino acid changes to any other one with a hydrophobic character, the mutation would tend to be predicted as tolerated. SIFT score also presumes that the important amino acids are highly conserved, so if the amino acid changes at a very conserved position, the position would tend to be predicted as deleterious. We would like to apply the similar idea in assigning comparable scores to the variants that cannot be predicted by their SIFT score. For example, if the variant is a nonsense mutation and it results in a

truncated, incomplete, nonfunctional protein product, the consequence is similar or even worse than if the amino acid substitution occurs in a conserved region, and it would be assigned as deleterious. Conversely, if the variant is a missense mutation and the amino acid changes to another amino acid with a similar character, it would be assigned as benign. This concept was also taken into consideration in our filtering method to remove germline variants.

In this machine learning prediction study, we selected the features that are strongly associated with the characters of somatic mutations to ensure the model gets trained with the most useful information. We would like to add more features to improve the model in the future. There are more features annotated by ANNOVAR that we can include in the model. The challenge is that not all of the features provide useful information for somatic mutation prediction; instead, some of them might even degrade the model. The solution is that when we consider adding a feature, we first understand the theory behind the feature, the connection between the new feature and somatic mutation, then the distribution of the new feature to avoid biasing the data, and most importantly, we will test the new features in the model to confirm the performance. We would like to use forward selection, which tests the addition of each new feature that improves the model the most and repeat this process until none improve the model.

We have identified recurrently mutated genes and important pathways that are affected by mutations and CNAs but there are no mutations or CNAs that were completely unique to tFL. It is quite clear that the transformation is the result of the transformation event in combination of certain existing genetic events. We need to have a larger dataset to figure out the genetic combinations that can cause transformation. We are going to use our custom capture panel and the pipelines we applied to it to extend the study.

CHAPTER VI

CONCLUSIONS

In our study, we designed multiple reliable pipelines to identify mutations from samples sequenced by WES and by a custom capture approach and evaluated the sequencing performance in each sample.

All the pipelines include the following: read quality control for all raw reads from the sequencers, read mapping against reference genome (alignment), mapping performance evaluation, duplicate ratio adjustment, variants calling, variant quality evaluation, variant depth evaluation, gene-related annotation, and database-related annotation. Each of these pipelines was constructed for use on particular sequencing types (WES vs. custom capture) and sample types (single vs. pairs vs. triplets) to make sure it takes full consideration of its corresponding features and expectations. All the variant calling pipelines were validated by Sanger sequencing to confirm their performance. Samples performed by two different sequencing platforms were carefully investigated to validate and compare their performance.

In custom capture panel sequencing, the method we employed used restriction enzymes to fragment DNA samples. A novel, statistically-based method was applied to estimate the duplicate ratio in samples with identical sequence reads that are generated by restriction enzymes. Simulation was performed for evaluation. This duplicate ratio estimation method enhances the statistical accuracy of variant frequency estimates and their changes from FL to tFL during clonal evolution.

An innovative two-way filtering based method was applied to retrieve somatic mutations from samples without corresponding normal samples. It is ideal to have healthy tissue from the same patient so that we can exclude germline variants from somatic mutations, but such tissue is not always available. Our layer by layer filtering approach was designed based on the biological character of the disease and the mutations and employed multiple reliable databases

to select reliable somatic mutations that may contribute to the disease. We classified the variants into two tiers of confidence. To validate our method, we applied it to a dataset which has corresponding normal samples without using the data from the normal samples. Then we compared the filtered results with the true somatic mutations identified from the full set of data. The validation proves our method is reliable and can filter out most of germline variants with 20.2% FDR.

We are the first to introduce a machine learning approach in predicting somatic mutations from samples without corresponding normal samples. Five different robust machine learning models were trained on one dataset (training dataset) with known germline variants. Two different sets of features were applied during training. Then the trained models were applied to another dataset (testing dataset) also with corresponding normal samples. The prediction performance was evaluated by a set of statistical measures based on the comparison of predicted results and true results. Random forest turned out to be the best model for germline variants prediction based on a comprehensive consideration.

We also designed an advanced survival analysis for CNV data. Compared to commonly used analysis, this advanced approach put different emphasis on survival curves based on the biological understanding of each CNA and brings more statistical power to the comparison test. The comprehensive output provides more flexibility for reviewers to retrieve information.

By integrating the information on mutation and CNAs, we have identified recurrently mutated genes and important pathways that may contribute to FL and to its transformation. We have found recurrent mutations of miR-142, which is a novel finding in FL and tFL studies. We also detected a number of mutations that appear to be more prevalent in tFL. The genes most frequently mutated in tFL included *TP53, KMT2D* (*MLL2*), *CREBBP*, *EZH2*, *BCL2*, *miR-142*, and

MEF2B. Many recurrently mutated genes are involved in important pathways, such as epigenetic regulation, the JAK-STAT or the NF-κB pathways, immune surveillance, and cell cycle regulation. Some other recurrently mutated genes are transcription factors involved in B-cell development. An especially interesting pathway is the S1P-activated pathway, which likely regulates lymphoma cell migration and survival outside of follicles during transformation. We detected mutations and CNAs along this pathway. We found few genes that were mutated only in tFL samples; therefore, it must be a combination of different genetic and epigenetic events that cooperate to drive an indolent FL to an aggressive tFL. A larger dataset and statistical analysis will be required to sort out the necessary combination of abnormalities required for transformation. The custom capture panel and analytical and experimental methods we have developed for frozen as well as FFPET will allow us to study a large set of patient samples to further our understanding of cooperativity of mutants and mechanisms of transformation.

CHAPTER VII

REFERENCES

1. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 WHO classification of lymphoid neoplasms and beyond: Evolving concepts and practical applications. *Blood*. 2011;117(19):5019-5032.

2. Harris NL, Stein H, Coupland SE, et al. New approaches to lymphoma diagnosis. *Hematology Am Soc Hematol Educ Program*. 2001:194-220.

3. Doan T. Immunology. In: *Lippincott's illustrated reviews; variation: Lippincott's illustrated reviews.* 2nd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013:99-107.

4. Lieber MR, Ma Y, Pannicke U, Schwarz K. Mechanism and regulation of human non-homologous DNA end-joining. *Nature reviews Molecular cell biology*. 2003;4(9):712-720.

5. Ott G, Rosenwald A. Molecular pathogenesis of follicular lymphoma. *Haematologica*. 2008;93(12):1773-1776.

6. Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. *J Clin Invest*. 2012;122(10):3424-3431.

7. Casulo C, Burack WR, Friedberg JW. Transformed follicular non-hodgkin lymphoma. *Blood*. 2015;125(1):40-47.

8. Pasqualucci L, Khiabanian H, Fangazio M, et al. Genetics of follicular lymphoma transformation. *Cell reports*. 2014;6(1):130-140.

9. Okosun J, Bödör C, Wang J, et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat Genet*. 2014;46(2):176-181.

10. Yano T, Jaffe ES, Longo DL, Raffeld M. MYC rearrangements in histologically progressed follicular lymphomas. *Blood*. 1992;80(3):758-767.

11. Horsman DE, Okamoto I, Ludkovski O, et al. Follicular lymphoma lacking the t (14; 18)(q32; q21): Identification of two disease subtypes. *Br J Haematol*. 2003;120(3):424-433.

12. Gu K, Fu K, Jain S, et al. T (14; 18)-negative follicular lymphomas are associated with a high frequency of BCL6 rearrangement at the alternative breakpoint region. *Modern Pathology*. 2009;22(9):1251-1257.

13. Jardin F. Next generation sequencing and the management of diffuse large B-cell lymphoma: From whole exome analysis to targeted therapy. *Discovery medicine*. 2014;18(97):51-65.

14. Green MR, Gentles AJ, Nair RV, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood*. 2013;121(9):1604-1611.

15. Bodor C, Grossmann V, Popov N, et al. EZH2 mutations are frequent and represent an early event in follicular lymphoma. *Blood*. 2013;122(18):3165-3168.

16. Li H, Kaminski MS, Li Y, et al. Mutations in linker histone genes HIST1H1 B, C, D, and E; OCT2 (POU2F2); IRF8; and ARID1A underlying the pathogenesis of follicular lymphoma. *Blood*. 2014;123(10):1487-1498.

17. Davies AJ, Rosenwald A, Wright G, et al. Transformation of follicular lymphoma to diffuse large b-cell lymphoma proceeds by distinct oncogenic mechanisms. *Br J Haematol*. 2007;136(2):286-293.

18. Lenz G, Wright GW, Emre NC, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci U S A*. 2008;105(36):13520-13525.

19. Tönnies H. Modern molecular cytogenetic techniques in genetic diagnostics. *Trends Mol Med*. 2002;8(6):246-250.

20. Stankiewicz P, Beaudet AL. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr Opin Genet Dev*. 2007;17(3):182-192.

21. Heinrichs S, Li C, Look AT. SNP array analysis in hematologic malignancies: Avoiding false discoveries. *Blood*. 2010;115(21):4157-4161.

22. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557-572.

23. Xu F, Wang Q, Zhang F, et al. Impact of next-generation sequencing (NGS) technology on cardiovascular disease research. *Cardiovascular diagnosis and therapy*. 2012;2(2):138-146.

24. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

25. Béguelin W, Popovic R, Teater M, et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer cell*. 2013;23(5):677-692.

26. Morin RD, Mendez-Lago M, Mungall AJ, et al. Frequent mutation of histone-modifying genes in non-hodgkin lymphoma. *Nature*. 2011;476(7360):298-303.

27. Lohr JG, Stojanov P, Lawrence MS, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A*. 2012;109(10):3879-3884.

28. Schmitz R, Young RM, Ceribelli M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*. 2012;490(7418):116-120.

29. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*. 2012;44(1):47-52.

30. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-3941.

31. Bouska A, McKeithan TW, Deffenbacher KE, et al. Genome-wide copy-number analyses reveal genomic abnormalities involved in transformation of follicular lymphoma. *Blood*. 2014;123(11):1681-1690.

32. Prentice RL. Linear rank tests with right censored data. *Biometrika*. 1978;65(1):167-179.

33. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.

34. Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, Bose R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics*. 2015;31(22):3561-3568.

35. Davis RE, Ngo VN, Lenz G, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature*. 2010;463(7277):88-92.

36. Lamason RL, McCully RR, Lew SM, Pomerantz JL. Oncogenic CARD11 mutations induce hyperactive signaling by disrupting autoinhibition by the PKC-responsive inhibitory domain. *Biochemistry (N Y)*. 2010;49(38):8240-8250.

37. Lenz G, Davis RE, Ngo VN, et al. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science*. 2008;319(5870):1676-1679.

38. Li S, Yang X, Shao J, Shen Y. Structural insights into the assembly of CARMA1 and BCL10. *PLoS One*. 2012;7(8):e42775.

39. Ngo VN, Young RM, Schmitz R, et al. Oncogenically active MYD88 mutations in human lymphoma. *Nature*. 2011;470(7332):115-119.

40. Schif B, Lennerz JK, Kohler CW, et al. SOCS1 mutation subtypes predict divergent outcomes in diffuse large B-cell lymphoma (DLBCL) patients. *Oncotarget*. 2013;4(1):35-47.

41. Landgraf P, Rusu M, Sheridan R, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*. 2007;129(7):1401-1414.

42. Kwanhian W, Lenze D, Alles J, et al. MicroRNA-142 is mutated in about 20% of diffuse large b-cell lymphoma. *Cancer medicine*. 2012;1(2):141-155.

43. Ying CY, Dominguez-Sola D, Fabi M, et al. MEF2B mutations lead to deregulated expression of the oncogene BCL6 in diffuse large B cell lymphoma. *Nat Immunol*. 2013;14(10):1084-1092.

44. Muppidi JR, Schmitz R, Green JA, et al. Loss of signalling via G [agr] 13 in germinal centre B-cell-derived lymphoma. *Nature*. 2014;516(7530):254-258.

45. Morin RD, Mungall K, Pleasance E, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood*. 2013;122(7):1256-1265.

46. Green JA, Suzuki K, Cho B, et al. The sphingosine 1-phosphate receptor S1P2 maintains the homeostasis of germinal center B cells and promotes niche confinement. *Nat Immunol*. 2011;12(7):672-680.

47. Yildiz M, Li H, Bernard D, et al. Activating STAT6 mutations in follicular lymphoma. *Blood*. 2015;125(4):668-679.

48. Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. 2010;464(7291):1071-1076.

49. Yap KL, Li S, Muñoz-Cabello AM, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell*. 2010;38(5):662-674.

50. Ji P, Diederichs S, Wang W, et al. MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*. 2003;22(39):8031-8041.

51. Jiang J, Jia P, Shen B, Zhao Z. Top associated SNPs in prostate cancer are significantly enriched in cis-expression quantitative trait loci and at transcription factor binding sites. *Oncotarget*. 2014;5(15):6168-6177.

52. Lin VC, Huang C, Lee Y, et al. Genetic variations in TP53 binding sites are predictors of clinical outcomes in prostate cancer patients. *Arch Toxicol*. 2014;88(4):901-911.

53. Kikuchi M, Miki T, Kumagai T, et al. Identification of negative regulatory regions within the first exon and intron of the BCL6 gene. *Oncogene*. 2000;19(42):4941-4945.

54. Gopalakrishnan C, Kamaraj B, Purohit R. Mutations in microRNA binding sites of CEP genes involved in cancer. *Cell Biochem Biophys*. 2014;70(3):1933-1942.

55. He Y, Carrillo JA, Luo J, et al. Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. *Frontiers in genetics*. 2014;5.

56. Karmakar S, Harcourt EM, Hewings DS, et al. Organocatalytic removal of formaldehyde adducts from RNA and DNA bases. *Nature chemistry*. 2015.

57. Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res*. 2013;41(6):e67.

58. Yost SE, Smith EN, Schwab RB, et al. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res*. 2012;40(14):e107.

59. Hicks S, Wheeler DA, Plon SE, Kimmel M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum Mutat*. 2011;32(6):661-668.

APPENDICES

APPEDIX A: RECURRENT SOMATIC MUTATIONS IDENTIFIED IN WES AND CUSTOM CAPTURE PANEL SEQUENCING

Sample ID	Chromosome| Position |Altered-Base	Gene	Mutation Type	Var Freq	Class	Platform
FL-2	chr1|27106915|T	ARID1A	stopgain	6%	shared	WES
tFL-2	chr1|27106915|T	ARID1A	stopgain	23%	shared	WES
FL-1	chr18|60985508|T	BCL2	missense	19%	shared	WES
tFL-1	chr18|60985508|T	BCL2	missense	35%	shared	WES
FL-22	chr18|60985814|C	BCL2	missense	12%	shared	WES
tFL-22	chr18|60985814|C	BCL2	missense	17%	shared	WES
FL-6	chr18|60985889|A	BCL2	missense	27%	shared	WES
tFL-6	chr18|60985889|A	BCL2	missense	27%	shared	WES
FL-22	chr2|32693596|T	BIRC6	stopgain	29%	shared	WES
tFL-22	chr2|32693596|T	BIRC6	stopgain	30%	shared	WES
FL-6	chr17|65908884|-ACT	BPTF	non-frameshift deletion	35%	shared	WES
tFL-6	chr17|65908884|-ACT	BPTF	non-frameshift deletion	40%	shared	WES
FL-4	chr10|98742885|T	C10orf12	stopgain	5%	shared	WES
tFL-4	chr10|98742885|T	C10orf12	stopgain	49%	shared	WES
FL-11	chr10|98743591|A	C10orf12	stopgain	8%	shared	WES
tFL-11	chr10|98743591|A	C10orf12	stopgain	26%	shared	WES
FL-1	chr6|41903707|A	CCND3	missense	46%	shared	WES
tFL-1	chr6|41903707|A	CCND3	missense	50%	shared	WES
FL-3	chr17|62007128|T	CD79B	splicing	5%	shared	WES
tFL-3	chr17|62007128|T	CD79B	splicing	20%	shared	WES
FL-10	chr16|3786704|G	CREBBP	missense	48%	shared	WES
tFL-10	chr16|3786704|G	CREBBP	missense	83%	shared	WES
FL-22	chr16|3786706|G	CREBBP	missense	32%	shared	WES
tFL-22	chr16|3786706|G	CREBBP	missense	26%	shared	WES
FL-6	chr16|3786739|G	CREBBP	missense	31%	shared	WES
tFL-6	chr16|3786739|G	CREBBP	missense	33%	shared	WES
FL-12	chr16|3788605|G	CREBBP	missense	41%	shared	WES
tFL-12	chr16|3788605|G	CREBBP	missense	51%	shared	WES
FL-11	chr16|3788618|A	CREBBP	missense	20%	shared	WES
tFL-11	chr16|3788618|A	CREBBP	missense	37%	shared	WES
FL-12	chr16|3795311|-A	CREBBP	frameshift deletion	36%	shared	WES
tFL-12	chr16|3795311|-A	CREBBP	frameshift deletion	49%	shared	WES

FL-5	chr16|3807881|A	CREBBP	stopgain	23%	shared	WES
tFL-5	chr16|3807881|A	CREBBP	stopgain	53%	shared	WES
FL-5	chr16|3807917|C	CREBBP	missense	21%	shared	WES
tFL-5	chr16|3807917|C	CREBBP	missense	53%	shared	WES
FL-5	chr16|3808030|C	CREBBP	missense	43%	shared	WES
tFL-5	chr16|3808030|C	CREBBP	missense	32%	shared	WES
FL-5	chr16|3808033|G	CREBBP	missense	41%	shared	WES
tFL-5	chr16|3808033|G	CREBBP	missense	33%	shared	WES
FL-5	chr16|3808046|G	CREBBP	missense	33%	shared	WES
tFL-5	chr16|3808046|G	CREBBP	missense	44%	shared	WES
FL-8	chr16|3828175|C	CREBBP	stopgain	45%	shared	WES
tFL-8	chr16|3828175|C	CREBBP	stopgain	39%	shared	WES
FL-8	chr16|3860683|T	CREBBP	missense	51%	shared	WES
tFL-8	chr16|3860683|T	CREBBP	missense	40%	shared	WES
FL-1	chrX|31854912|C	DMD	missense	23%	shared	WES
tFL-1	chrX|31854912|C	DMD	missense	40%	shared	WES
FL-6	chr1|184692957|T	EDEM3	missense	33%	shared	WES
tFL-6	chr1|184692957|T	EDEM3	missense	24%	shared	WES
FL-22	chr6|74229068|A	EEF1A1	missense	29%	shared	WES
tFL-22	chr6|74229068|A	EEF1A1	missense	26%	shared	WES
tFL-4	chr7|148508727|A	EZH2	missense	23%	shared	WES
tFL-6	chr7|148508727|A	EZH2	missense	44%	shared	WES
FL-1	chr7|148508728|T	EZH2	missense	20%	shared	WES
tFL-1	chr7|148508728|T	EZH2	missense	42%	shared	WES
FL-5	chr7|148508728|T	EZH2	missense	26%	shared	WES
tFL-5	chr7|148508728|T	EZH2	missense	47%	shared	WES
FL-9	chr11|18327713|T	HPS5	missense	10%	shared	WES
tFL-9	chr11|18327713|T	HPS5	missense	37%	shared	WES
FL-3	chr4|3134324|A	HTT	missense	14%	shared	WES
tFL-3	chr4|3134324|A	HTT	missense	25%	shared	WES
FL-5	chr16|85936784|G	IRF8	missense	32%	shared	WES
tFL-5	chr16|85936784|G	IRF8	missense	56%	shared	WES
FL-22	chr3|183210327|C	KLHL6	missense	39%	shared	WES
tFL-22	chr3|183210327|C	KLHL6	missense	13%	shared	WES
FL-5	chr12|49418731|A	KMT2D(MLL2)	splicing	32%	shared	WES
tFL-5	chr12|49418731|A	KMT2D(MLL2)	splicing	40%	shared	WES
FL-12	chr12|49426001|-C	KMT2D(MLL2)	frameshift deletion	44%	shared	WES
tFL-12	chr12|49426001|-C	KMT2D(MLL2)	frameshift deletion	45%	shared	WES

FL-6	chr12|49427255|A	KMT2D(MLL2)	stopgain	52%	shared	WES
tFL-6	chr12|49427255|A	KMT2D(MLL2)	stopgain	36%	shared	WES
FL-2	chr12|49431070|C	KMT2D(MLL2)	missense	6%	shared	WES
tFL-2	chr12|49431070|C	KMT2D(MLL2)	missense	27%	shared	WES
FL-8	chr12|49431800|-G	KMT2D(MLL2)	frameshift deletion	55%	shared	WES
tFL-8	chr12|49431800|-G	KMT2D(MLL2)	frameshift deletion	40%	shared	WES
FL-6	chr12|49432396|A	KMT2D(MLL2)	stopgain	37%	shared	WES
tFL-6	chr12|49432396|A	KMT2D(MLL2)	stopgain	31%	shared	WES
FL-1	chr12|49433220|A	KMT2D(MLL2)	stopgain	56%	shared	WES
tFL-1	chr12|49433220|A	KMT2D(MLL2)	stopgain	95%	shared	WES
FL-10	chr12|49433524|-CT	KMT2D(MLL2)	frameshift deletion	76%	shared	WES
tFL-10	chr12|49433524|-CT	KMT2D(MLL2)	frameshift deletion	81%	shared	WES
FL-5	chr12|49438694|T	KMT2D(MLL2)	stopgain	49%	shared	WES
tFL-5	chr12|49438694|T	KMT2D(MLL2)	stopgain	53%	shared	WES
FL-3	chr15|42035019|A	MGA	missense	59%	shared	WES
tFL-3	chr15|42035019|A	MGA	missense	54%	shared	WES
FL-8	chr17|56408621|G	MIR142	-	45%	shared	WES
tFL-8	chr17|56408621|G	MIR142	-	47%	shared	WES
FL-11	chr17|56408657|G	MIR142	-	9%	shared	WES
tFL-11	chr17|56408657|G	MIR142	-	21%	shared	WES
FL-12	chr3|38182641|C	MYD88	missense	40%	shared	WES
tFL-12	chr3|38182641|C	MYD88	missense	83%	shared	WES
FL-6	chr1|39322649|G	RRAGC	missense	28%	shared	WES
tFL-6	chr1|39322649|G	RRAGC	missense	23%	shared	WES
FL-8	chr1|39322697|A	RRAGC	missense	34%	shared	WES
tFL-8	chr1|39322697|A	RRAGC	missense	45%	shared	WES
FL-3	chr20|48500446|A	SLC9A8	missense	24%	shared	WES
tFL-3	chr20|48500446|A	SLC9A8	missense	22%	shared	WES
FL-1	chr18|60985760|T	BCL2	missense	5%	FL-unique	WES
FL-8	chr18|60985880|A	BCL2	missense	24%	FL-unique	WES
FL-12	chr6|41903688|T	CCND3	missense	31%	FL-unique	WES
FL-8	chr17|62007129|T	CD79B	splicing	45%	FL-unique	WES
FL-5	chrX|41200829|C	DDX3X	missense	36%	FL-unique	WES
FL-11	chr7|148508728|T	EZH2	missense	3%	FL-unique	WES
FL-8	chr6|26056530|C	HIST1H1C	missense	34%	FL-unique	WES

FL-8	chr6|26234591|G	HIST1H1D	missense	46%	FL-unique	WES
FL-12	chr17|16012098|T	NCOR1	splicing	56%	FL-unique	WES
FL-8	chr21|43256276|C	PRDM15	missense	31%	FL-unique	WES
FL-3	chr1|110882568|G	RBM15	missense	57%	FL-unique	WES
FL-12	chr13|37394096|G	RFXAP	splicing	49%	FL-unique	WES
FL-6	chr1|2493172|A	TNFRSF14	stopgain	67%	FL-unique	WES
tFL-4	chr18|60985405|G	BCL2	missense	24%	tFL-unique	WES
tFL-4	chr18|60985411|C	BCL2	missense	18%	tFL-unique	WES
tFL-2	chr7|2979495|G	CARD11	missense	30%	tFL-unique	WES
tFL-12	chr6|41903688|G	CCND3	missense	44%	tFL-unique	WES
tFL-8	chr6|41903710|G	CCND3	missense	13%	tFL-unique	WES
tFL-22	chr6|41903755|A	CCND3	stopgain	25%	tFL-unique	WES
tFL-8	chr17|62006836|T	CD79B	splicing	52%	tFL-unique	WES
tFL-2	chr16|3786206|C	CREBBP	splicing	33%	tFL-unique	WES
tFL-6	chrX|41205842|A	DDX3X	missense	43%	tFL-unique	WES
tFL-8	chr6|74228924|T	EEF1A1	missense	37%	tFL-unique	WES
tFL-4	chr11|128391823|A	ETS1	missense	41%	tFL-unique	WES
tFL-11	chr7|148506437|A	EZH2	missense	19%	tFL-unique	WES
tFL-22	chr7|148508728|T	EZH2	missense	30%	tFL-unique	WES
tFL-9	chr3|183273154|C	KLHL6	stopgain	37%	tFL-unique	WES
tFL-4	chr12|49435906|-T	KMT2D(MLL2)	frameshift deletion	21%	tFL-unique	WES
tFL-4	chr12|49443750|-T	KMT2D(MLL2)	frameshift deletion	49%	tFL-unique	WES
tFL-22	chr3|38182025|T	MYD88	missense	14%	tFL-unique	WES
tFL-9	chr5|67576825|A	PIK3R1	missense	32%	tFL-unique	WES
tFL-2	chr21|43256634|C	PRDM15	missense	11%	tFL-unique	WES

tFL-2	chr1|110883003|T	RBM15	missense	15%	tFL-unique	WES
tFL-5	chr17|78269544|C	RNF213	missense	50%	tFL-unique	WES
tFL-6	chr20|48467300|-T	SLC9A8	frameshift deletion	21%	tFL-unique	WES
tFL-6	chr1|16256410|+A	SPEN	frameshift insertion	37%	tFL-unique	WES
tFL-1	chr6|152456309|C	SYNE1	missense	50%	tFL-unique	WES
tFL-22	chr17|7577097|A	TP53	missense	18%	tFL-unique	WES
tFL-5	chr17|7577545|C	TP53	missense	38%	tFL-unique	WES
tFL-5	chr17|7577598|A	TP53	missense	56%	tFL-unique	WES
tFL-15	chr7|2984163|T	CARD11	nonsynonymous	47%	single	custom
tFL-15	chr7|148508728|G	EZH2	nonsynonymous	46%	single	custom
tFL-15	chr12|49433790|-G	KMT2D(MLL2)	frameshift_deletion	25%	single	custom
tFL-15	chr16|3788617|T	CREBBP	nonsynonymous	33%	single	custom
tFL-15	chr16|85942671|C	IRF8	nonsynonymous	18%	single	custom
tFL-15	chr17|40475318|G	STAT3	nonsynonymous	29%	single	custom
tFL-15	chr18|60985724|C	BCL2	nonsynonymous	22%	single	custom
tFL-15	chr19|11143994|A	SMARCA4	nonsynonymous	30%	single	custom
tFL-15	chr19|19260045|A	MEF2B	nonsynonymous	30%	single	custom
tFL-16	chr5|67591106|G	PIK3R1	nonsynonymous	10%	single	custom
tFL-16	chr6|37138763|G	PIM1	nonsynonymous	28%	single	custom
tFL-16	chr12|49427447|A	KMT2D(MLL2)	stopgain	39%	single	custom
tFL-19	chr3|38182641|C	MYD88	missense	81%	single	custom
tFL-19	chr5|35857093|A	IL7R	nonsynonymous	23%	single	custom
tFL-19	chr6|37138423|C	PIM1	nonsynonymous	36%	single	custom
tFL-19	chr7|2977613|T	CARD11	nonsynonymous	32%	single	custom
tFL-19	chr12|49425693|-C	KMT2D(MLL2)	frameshift_deletion	54%	single	custom
tFL-19	chr17|62006798|C	CD79B	nonsynonymous	44%	single	custom
tFL-20	chr7|148508727|A	EZH2	nonsynonymous	19%	single	custom
tFL-20	chr12|49415846|A	KMT2D(MLL2)	stopgain	44%	single	custom
tFL-20	chr12|49435698|G	KMT2D(MLL2)	splicing	43%	single	custom
tFL-20	chr16|11349287|T	SOCS1	nonsynonymous	38%	single	custom
tFL-20	chr16|89857858|T	FANCA	nonsynonymous	33%	single	custom
tFL-20	chr22|41556727|A	EP300	splicing	49%	single	custom
tFL-31	chr1|16265861|G	SPEN	nonsynonymous	37%	single	custom
tFL-31	chr6|26056305|C	HIST1H1C	nonsynonymous	30%	single	custom

tFL-31	chr6|138192455|T	TNFAIP3	stopgain	39%	single	custom
tFL-31	chr12|49446429|T	KMT2D(MLL2)	stopgain	29%	single	custom
tFL-31	chr13|41240211|A	FOXO1	nonsynonymous	27%	single	custom
tFL-31	chr16|3788618|A	CREBBP	nonsynonymous	38%	single	custom
tFL-31	chr17|16046921|C	NCOR1	nonsynonymous	31%	single	custom
tFL-31	chr18|60985760|T	BCL2	nonsynonymous	6%	single	custom
tFL-31	chr18|60985880|A	BCL2	nonsynonymous	3%	single	custom
tFL-31	chrX|70613191|A	TAF1	nonsynonymous	58%	single	custom
tFL-32	chr1|2492152|G	TNFRSF14	nonsynonymous	11%	single	custom
tFL-32	chr5|35867526|G	IL7R	nonsynonymous	56%	single	custom
tFL-32	chr17|7578286|G	TP53	nonsynonymous	37%	single	custom
tFL-32	chr17|40477033|C	STAT3	nonsynonymous	36%	single	custom
tFL-33	chr7|148508728|T	EZH2	nonsynonymous	20%	single	custom
tFL-33	chr17|40474427|G	STAT3	nonsynonymous	20%	single	custom
tFL-33	chr18|60985301|G	BCL2	nonsynonymous	25%	single	custom
tFL-33	chr18|60985626|T	BCL2	nonsynonymous	19%	single	custom
tFL-34	chr1|2488138|A	TNFRSF14	stopgain	43%	single	custom
tFL-34	chr4|153244046|G	FBXW7	nonsynonymous	22%	single	custom
tFL-34	chr7|2977614|A	CARD11	nonsynonymous	28%	single	custom
tFL-34	chr7|148508727|A	EZH2	nonsynonymous	14%	single	custom
tFL-34	chr7|148508745|C	EZH2	nonsynonymous	14%	single	custom
tFL-34	chr7|148508763|A	EZH2	nonsynonymous	17%	single	custom
tFL-34	chr12|49415647|T	KMT2D(MLL2)	stopgain	17%	single	custom
tFL-34	chr12|49425098|A	KMT2D(MLL2)	stopgain	25%	single	custom
tFL-34	chr16|3790421|T	CREBBP	nonsynonymous	20%	single	custom
tFL-35	chr1|2492063|-C	TNFRSF14	frameshift_deletion	78%	single	custom
tFL-35	chr6|26234923|T	HIST1H1D	nonsynonymous	42%	single	custom
tFL-35	chr7|148508727|G	EZH2	nonsynonymous	36%	single	custom
tFL-35	chr12|49432738|A	KMT2D(MLL2)	stopgain	87%	single	custom
tFL-35	chr16|3790421|T	CREBBP	nonsynonymous	86%	single	custom
tFL-36	chr6|27860754|C	HIST1H2AM	stopgain	15%	single	custom
tFL-36	chr12|49425446|-CT	KMT2D(MLL2)	frameshift_deletion	43%	single	custom
tFL-36	chr18|60985896|T	BCL2	nonsynonymous	22%	single	custom
tFL-37	chr7|148508728|T	EZH2	nonsynonymous	27%	single	custom
tFL-37	chr12|49431667|A	KMT2D(MLL2)	stopgain	47%	single	custom
tFL-37	chr12|49434187|+T	KMT2D(MLL2)	frameshift_insertion	39%	single	custom
tFL-37	chr16|3820888|+GTGCA	CREBBP	frameshift_insertion	45%	single	custom
tFL-37	chr16|11348901|C	SOCS1	nonsynonymous	24%	single	custom

tFL-37	chr18|60985286|T	BCL2	nonsynonymous	24%	single	custom
tFL-37	chr18|60985884|C	BCL2	nonsynonymous	32%	single	custom
tFL-37	chr19|19260045|A	MEF2B	nonsynonymous	53%	single	custom
tFL-38	chr6|37138355|T	PIM1	nonsynonymous	38%	single	custom
tFL-38	chr12|49424074|T	KMT2D(MLL2)	nonsynonymous	16%	single	custom
tFL-38	chr12|49435294|A	KMT2D(MLL2)	stopgain	76%	single	custom
tFL-38	chr17|62007480|C	CD79B	nonsynonymous	36%	single	custom
tFL-39	chr6|37139111|T	PIM1	nonsynonymous	27%	single	custom
tFL-39	chr6|41903710|G	CCND3	nonsynonymous	44%	single	custom
tFL-39	chr12|49435479|A	KMT2D(MLL2)	stopgain	34%	single	custom
tFL-39	chr16|3788657|G	CREBBP	nonsynonymous	31%	single	custom
tFL-40	chr1|155920754|A	ARHGEF2	stopgain	49%	single	custom
tFL-40	chr7|2985468|T	CARD11	nonsynonymous	41%	single	custom
tFL-40	chr15|45003764|A	B2M	stopgain	60%	single	custom
tFL-40	chr18|60795973|A	BCL2	nonsynonymous	37%	single	custom
tFL-40	chr22|41566522|G	EP300	nonsynonymous	20%	single	custom
FL-3-cus	chr1|27023162|G	ARID1A	nonsynonymous	7%	shared	custom
FL-3-cus	chr1|110882568|G	RBM15	nonsynonymous	16%	shared	custom
FL-3-cus	chr17|62007128|T	CD79B	splicing	14%	shared	custom
FL-5-cus	chr7|148508728|T	EZH2	nonsynonymous	32%	shared	custom
FL-5-cus	chr12|49418731|A	KMT2D(MLL2)	splicing	40%	shared	custom
FL-5-cus	chr12|49438694|T	KMT2D(MLL2)	stopgain	34%	shared	custom
FL-5-cus	chr16|3807881|A	CREBBP	stopgain	37%	shared	custom
FL-5-cus	chr16|3807917|C	CREBBP	nonsynonymous	39%	shared	custom
FL-5-cus	chr16|3808030|C	CREBBP	nonsynonymous	38%	shared	custom
FL-5-cus	chr16|3808033|G	CREBBP	nonsynonymous	38%	shared	custom
FL-5-cus	chr16|3808046|G	CREBBP	nonsynonymous	38%	shared	custom
FL-5-cus	chr16|11349099|C	SOCS1	nonsynonymous	36%	shared	custom
FL-5-cus	chr16|85936784|G	IRF8	nonsynonymous	38%	shared	custom
FL-30	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	54%	shared	custom
FL-30	chr16|3786704|G	CREBBP	nonsynonymous	35%	shared	custom

FL-23-1	chr16|3789578|A	CREBBP	splicing	40%	shared	custom
FL-23-2	chr16|3789578|A	CREBBP	splicing	18%	shared	custom
FL-24-1	chr6|90402481|G	MDN1	nonsynonymous	23%	shared	custom
FL-24-1	chr7|148508728|G	EZH2	nonsynonymous	40%	shared	custom
FL-24-2	chr7|148508728|G	EZH2	nonsynonymous	38%	shared	custom
FL-24-1	chr16|3788618|A	CREBBP	nonsynonymous	76%	shared	custom
FL-24-2	chr16|3788618|A	CREBBP	nonsynonymous	79%	shared	custom
tFL-24	chr16|3788618|A	CREBBP	nonsynonymous	66%	shared	custom
FL-25-1	chr7|148508727|A	EZH2	nonsynonymous	18%	shared	custom
FL-25-2	chr7|148508727|A	EZH2	nonsynonymous	10%	shared	custom
tFL-25	chr7|148508727|A	EZH2	nonsynonymous	10%	shared	custom
FL-25-1	chr12|49428448|A	KMT2D(MLL2)	stopgain	61%	shared	custom
FL-25-2	chr12|49428448|A	KMT2D(MLL2)	stopgain	66%	shared	custom
FL-25-1	chr12|113496211|G	DTX1	nonsynonymous	27%	shared	custom
FL-25-2	chr12|113496211|G	DTX1	nonsynonymous	31%	shared	custom
FL-25-1	chr15|45003779|C	B2M	nonsynonymous	17%	shared	custom
FL-25-2	chr15|45003779|C	B2M	nonsynonymous	26%	shared	custom
FL-25-1	chr16|3900873|A	CREBBP	stopgain	21%	shared	custom
FL-25-2	chr16|3900873|A	CREBBP	stopgain	31%	shared	custom
FL-25-1	chr16|11349320|A	SOCS1	stopgain	23%	shared	custom
FL-25-2	chr16|11349320|A	SOCS1	stopgain	33%	shared	custom
FL-25-1	chr19|19257600|C	MEF2B	stopgain	28%	shared	custom
FL-25-2	chr19|19257600|C	MEF2B	stopgain	29%	shared	custom
FL-25-1	chr19|19260045|A	MEF2B	nonsynonymous	33%	shared	custom
FL-25-2	chr19|19260045|A	MEF2B	nonsynonymous	35%	shared	custom
FL-26-1	chr11|128332397|C	ETS1	stopgain	27%	shared	custom
FL-26-2	chr11|128332397|C	ETS1	stopgain	22%	shared	custom
FL-26-1	chr12|49426598|A	KMT2D(MLL2)	stopgain	23%	shared	custom
FL-26-2	chr12|49426598|A	KMT2D(MLL2)	stopgain	24%	shared	custom
FL-26-1	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	53%	shared	custom
FL-26-2	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	47%	shared	custom
FL-26-1	chr16|3781333|T	CREBBP	nonsynonymous	31%	shared	custom
FL-26-2	chr16|3781333|T	CREBBP	nonsynonymous	34%	shared	custom
FL-26-1	chr16|3831211|+C	CREBBP	frameshift_insertion	23%	shared	custom
FL-26-1	chr18|60985883|T	BCL2	nonsynonymous	26%	shared	custom
FL-27	chr1|27089478|T	ARID1A	stopgain	14%	shared	custom
FL-27	chr12|49424062|T	KMT2D(MLL2)	splicing	23%	shared	custom
FL-27	chr16|3788618|A	CREBBP	nonsynonymous	21%	shared	custom

tFL-27	chr16|3788618|A	CREBBP	nonsynonymous	79%	shared	custom
FL-27	chr18|60985760|T	BCL2	nonsynonymous	26%	shared	custom
tFL-27	chr18|60985760|T	BCL2	nonsynonymous	41%	shared	custom
FL-28	chr4|153332538|A	FBXW7	nonsynonymous	14%	shared	custom
FL-28	chr7|148508727|A	EZH2	nonsynonymous	12%	shared	custom
tFL-28	chr7|148508727|A	EZH2	nonsynonymous	18%	shared	custom
FL-29	chr18|60985529|G	BCL2	nonsynonymous	13%	shared	custom
FL-29	chr19|19260045|G	MEF2B	nonsynonymous	62%	shared	custom
FL-31	chr1|16265861|G	SPEN	nonsynonymous	20%	shared	custom
FL-31	chr12|49446429|T	KMT2D(MLL2)	stopgain	19%	shared	custom
FL-31	chr13|41240211|A	FOXO1	nonsynonymous	18%	shared	custom
FL-31	chr16|3788618|A	CREBBP	nonsynonymous	14%	shared	custom
FL-31	chr18|60985880|A	BCL2	nonsynonymous	18%	shared	custom
FL-5-cus	chrX|41200829|C	DDX3X	nonsynonymous	43%	FL-unique	custom
FL-24-2	chr6|90402481|G	MDN1	nonsynonymous	22%	FL-unique	custom
FL-26-2	chr18|60985883|T	BCL2	nonsynonymous	25%	FL-unique	custom
FL-27	chr6|134495706|G	SGK1	nonsynonymous	15%	FL-unique	custom
FL-27	chr12|49427261|A	KMT2D(MLL2)	stopgain	14%	FL-unique	custom
FL-28	chr18|60985815|G	BCL2	nonsynonymous	18%	FL-unique	custom
FL-29	chr12|49418462|C	KMT2D(MLL2)	stopgain	24%	FL-unique	custom
FL-31	chr12|92538182|T	BTG1	nonsynonymous	11%	FL-unique	custom
tFL-3-cus	chr1|27023162|G	ARID1A	nonsynonymous	4%	tFL-unique	custom
tFL-3-cus	chr1|110882568|G	RBM15	nonsynonymous	22%	tFL-unique	custom
tFL-3-cus	chr17|62007128|T	CD79B	splicing	20%	tFL-unique	custom
tFL-5-cus	chr7|148508728|T	EZH2	nonsynonymous	37%	tFL-unique	custom
tFL-5-cus	chr12|49418731|A	KMT2D(MLL2)	splicing	47%	tFL-unique	custom
tFL-5-cus	chr12|49438694|T	KMT2D(MLL2)	stopgain	46%	tFL-unique	custom
tFL-5-cus	chr16|3807881|A	CREBBP	stopgain	40%	tFL-unique	custom
tFL-5-cus	chr16|3807917|C	CREBBP	nonsynonymous	42%	tFL-unique	custom
tFL-5-	chr16|3808030|C	CREBBP	nonsynonymous	42%	tFL-	custom

cus					unique	
tFL-5-cus	chr16|3808033|G	CREBBP	nonsynonymous	42%	tFL-unique	custom
tFL-5-cus	chr16|3808046|G	CREBBP	nonsynonymous	42%	tFL-unique	custom
tFL-5-cus	chr16|11349099|C	SOCS1	nonsynonymous	40%	tFL-unique	custom
tFL-5-cus	chr16|85936784|G	IRF8	nonsynonymous	40%	tFL-unique	custom
tFL-5-cus	chr17|7577545|C	TP53	nonsynonymous	83%	tFL-unique	custom
tFL-5-cus	chr17|7577598|A	TP53	nonsynonymous	11%	tFL-unique	custom
tFL-5-cus	chr17|78269544|C	RNF213	nonsynonymous	34%	tFL-unique	custom
tFL-30	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	83%	tFL-unique	custom
tFL-30	chr16|3786704|G	CREBBP	nonsynonymous	76%	tFL-unique	custom
tFL-23	chr3|38182032|G	MYD88	nonsynonymous	19%	tFL-unique	custom
tFL-23	chr7|2979501|G	CARD11	nonsynonymous	13%	tFL-unique	custom
tFL-23	chr16|3786704|T	CREBBP	nonsynonymous	15%	tFL-unique	custom
tFL-23	chr16|3789578|A	CREBBP	splicing	12%	tFL-unique	custom
tFL-24	chr7|148508728|G	EZH2	nonsynonymous	33%	tFL-unique	custom
tFL-24	chr17|40485721|G	STAT3	nonsynonymous	31%	tFL-unique	custom
tFL-25	chr12|49428448|A	KMT2D(MLL2)	stopgain	53%	tFL-unique	custom
tFL-25	chr12|113496211|G	DTX1	nonsynonymous	26%	tFL-unique	custom
tFL-25	chr15|45003779|C	B2M	nonsynonymous	23%	tFL-unique	custom
tFL-25	chr16|3900873|A	CREBBP	stopgain	25%	tFL-unique	custom
tFL-25	chr16|11349320|A	SOCS1	stopgain	26%	tFL-unique	custom
tFL-25	chr16|11349333|A	SOCS1	nonsynonymous	26%	tFL-unique	custom
tFL-25	chr19|19257600|C	MEF2B	stopgain	23%	tFL-unique	custom
tFL-25	chr19|19260045|A	MEF2B	nonsynonymous	21%	tFL-unique	custom
tFL-26	chr11|128332397|C	ETS1	stopgain	37%	tFL-unique	custom
tFL-26	chr12|49426598|A	KMT2D(MLL2)	stopgain	42%	tFL-	custom

					unique	
tFL-26	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	47%	tFL-unique	custom
tFL-26	chr16|3781333|T	CREBBP	nonsynonymous	40%	tFL-unique	custom
tFL-26	chr16|3831211|+C	CREBBP	frameshift_insertion	30%	tFL-unique	custom
tFL-26	chr16|11348935|T	SOCS1	nonsynonymous	42%	tFL-unique	custom
tFL-26	chr19|19257993|A	MEF2B	splicing	81%	tFL-unique	custom
tFL-27	chr1|27089478|T	ARID1A	stopgain	45%	tFL-unique	custom
tFL-27	chr6|90453359|T	MDN1	stopgain	36%	tFL-unique	custom
tFL-27	chr7|151873463|A	KMT2C(MLL3)	nonsynonymous	39%	tFL-unique	custom
tFL-27	chr12|49424062|T	KMT2D(MLL2)	splicing	32%	tFL-unique	custom
tFL-27	chr17|7578265|G	TP53	nonsynonymous	63%	tFL-unique	custom
tFL-27	chr18|60985722|T	BCL2	nonsynonymous	39%	tFL-unique	custom
tFL-28	chr2|73677838|G	ALMS1	nonsynonymous	17%	tFL-unique	custom
tFL-28	chr4|153332538|A	FBXW7	nonsynonymous	15%	tFL-unique	custom
tFL-28	chr6|134495724|C	SGK1	nonsynonymous	13%	tFL-unique	custom
tFL-28	chr12|92539226|C	BTG1	nonsynonymous	12%	tFL-unique	custom
tFL-28	chr13|41240279|A	FOXO1	nonsynonymous	15%	tFL-unique	custom
tFL-29	chr7|148508727|A	EZH2	nonsynonymous	28%	tFL-unique	custom
tFL-29	chr7|152027794|-C	KMT2C(MLL3)	frameshift_deletion	34%	tFL-unique	custom
tFL-29	chr12|49420082|A	KMT2D(MLL2)	nonsynonymous	50%	tFL-unique	custom
tFL-29	chr12|49420132|A	KMT2D(MLL2)	nonsynonymous	41%	tFL-unique	custom
tFL-29	chr17|7576571|C	TP53	stopgain	47%	tFL-unique	custom
tFL-29	chr17|7577093|G	TP53	nonsynonymous	51%	tFL-unique	custom
tFL-29	chr18|60985529|G	BCL2	nonsynonymous	33%	tFL-unique	custom
tFL-29	chr19|19260045|G	MEF2B	nonsynonymous	97%	tFL-unique	custom
tFL-31	chr1|16265861|G	SPEN	nonsynonymous	37%	tFL-	custom

					unique	
tFL-31	chr6|26056305|C	HIST1H1C	nonsynonymous	30%	tFL-unique	custom
tFL-31	chr6|138192455|T	TNFAIP3	stopgain	39%	tFL-unique	custom
tFL-31	chr12|49446429|T	KMT2D(MLL2)	stopgain	29%	tFL-unique	custom
tFL-31	chr13|41240211|A	FOXO1	nonsynonymous	27%	tFL-unique	custom
tFL-31	chr16|3788618|A	CREBBP	nonsynonymous	38%	tFL-unique	custom
tFL-31	chr17|16046921|C	NCOR1	nonsynonymous	31%	tFL-unique	custom
tFL-31	chr18|60985760|T	BCL2	nonsynonymous	6%	tFL-unique	custom
tFL-31	chr18|60985880|A	BCL2	nonsynonymous	3%	tFL-unique	custom
tFL-31	chrX|70613191|A	TAF1	nonsynonymous	58%	tFL-unique	custom

The first column (Sample ID) indicates the samples in which the mutations were detected. The second column (Chromosome| Position |Altered-Base) indicates the coordinates of the mutations. The third column indicates the genes altered by the mutations. The fourth column indicates the mutation type. The fifth column (Var Freq) indicates the variant frequencies of the mutations. The sixth column indicates the classes of the mutations. The last column indicates the platform used to detect the mutations.

APPENDIX B: INTEGRATED MUTATION AND CN INFORMATIN FOR TNFRSF14, CARD11, HIST1H1E, EZH2, KMT2D (MLL2), BCL2, TNFAIP3, SGK1, CREBBP, and TP53

Sample ID	Chr|Pos|VarAllele	Gene	MutType	Var Freq	CN
FL-5	chr1|2493112|C	TNFRSF14	nonsynonymous	0.667	NA
FL-5-cus	chr1|2493112|C	TNFRSF14	nonsynonymous	0.81	NA
tFL-5	chr1|2493112|C	TNFRSF14	nonsynonymous	1	1
tFL-5-cus	chr1|2493112|C	TNFRSF14	nonsynonymous	0.881	1
FL-6	chr1|2493172|A	TNFRSF14	stopgain	0.667	1
FL-10	chr1|2493111|A	TNFRSF14	splicing	0.409	2
FL-10-cus	chr1|2493111|A	TNFRSF14	splicing	0.48	2
tFL-10	chr1|2493111|A	TNFRSF14	splicing	0.474	2
tFL-10-cus	chr1|2493111|A	TNFRSF14	splicing	0.55	2
tFL-15	chr1|2489802|A	TNFRSF14	nonsynonymous	0.629	2
tFL-32	chr1|2492152|G	TNFRSF14	nonsynonymous	0.11	2
tFL-34	chr1|2488138|A	TNFRSF14	stopgain	0.431	2
tFL-35	chr1|2492063|-C	TNFRSF14	frameshift_deletion	0.78	1
tFL-2	chr7|2979495|G	CARD11	nonsynonymous	0.303	4
tFL-15	chr7|2984163|T	CARD11	nonsynonymous	0.47	3
tFL-15	chr7|2979466|C	CARD11	nonsynonymous	0.46	3
tFL-19	chr7|2977613|T	CARD11	nonsynonymous	0.32	2
tFL-19	chr7|2979486|G	CARD11	nonsynonymous	0.4	2
tFL-23	chr7|2979501|G	CARD11	nonsynonymous	0.13	3
tFL-34	chr7|2977614|A	CARD11	nonsynonymous	0.28	3
tFL-40	chr7|2985468|T	CARD11	nonsynonymous	0.41	2
tFL-19	chr6|26156958|T	HIST1H1E	nonsynonymous	0.33	2
tFL-19	chr6|26156976|C	HIST1H1E	nonsynonymous	0.32	2
tFL-20	chr6|26156911|A	HIST1H1E	nonsynonymous	0.41	2
tFL-23	chr6|26156797|T	HIST1H1E	nonsynonymous	0.11	3
FL-25-1	chr6|26156787|C	HIST1H1E	nonsynonymous	0.2	2
FL-25-2	chr6|26156787|C	HIST1H1E	nonsynonymous	0.301	NA
tFL-25	chr6|26156787|C	HIST1H1E	nonsynonymous	0.24	2
tFL-37	chr6|26156947|G	HIST1H1E	nonsynonymous	0.22	3
tFL-38	chr6|26157271|G	HIST1H1E	nonsynonymous	0.3	3
FL-1	chr7|148508728|T	EZH2	nonsynonymous	0.204	2
tFL-1	chr7|148508728|T	EZH2	nonsynonymous	0.422	2
tFL-4	chr7|148508727|A	EZH2	nonsynonymous	0.228	3
FL-5	chr7|148508728|T	EZH2	nonsynonymous	0.262	NA
FL-5-cus	chr7|148508728|T	EZH2	nonsynonymous	0.321	NA
tFL-5	chr7|148508728|T	EZH2	nonsynonymous	0.465	2

tFL-5-cus	chr7|148508728|T	EZH2	nonsynonymous	0.37	NA
tFL-6	chr7|148508727|A	EZH2	nonsynonymous	0.441	2
tFL-11	chr7|148506437|A	EZH2	nonsynonymous	0.189	2
FL-11	chr7|148508728|T	EZH2	nonsynonymous	0.034	2
tFL-15	chr7|148508728|G	EZH2	nonsynonymous	0.46	3
tFL-20	chr7|148508727|A	EZH2	nonsynonymous	0.19	3
tFL-22	chr7|148508728|T	EZH2	nonsynonymous	0.3	NA
FL-24-1	chr7|148508728|G	EZH2	nonsynonymous	0.4	2
FL-24-2	chr7|148508728|G	EZH2	nonsynonymous	0.38	NA
tFL-24	chr7|148508728|G	EZH2	nonsynonymous	0.331	2
FL-25-1	chr7|148508727|A	EZH2	nonsynonymous	0.18	3
FL-25-2	chr7|148508727|A	EZH2	nonsynonymous	0.1	NA
tFL-25	chr7|148508727|A	EZH2	nonsynonymous	0.1	3
FL-28	chr7|148508727|A	EZH2	nonsynonymous	0.12	NA
tFL-28	chr7|148508727|A	EZH2	nonsynonymous	0.18	3
tFL-29	chr7|148508727|A	EZH2	nonsynonymous	0.28	2
tFL-33	chr7|148508728|T	EZH2	nonsynonymous	0.2	3
tFL-34	chr7|148508727|A	EZH2	nonsynonymous	0.14	3
tFL-34	chr7|148508745|C	EZH2	nonsynonymous	0.14	3
tFL-34	chr7|148508763|A	EZH2	nonsynonymous	0.17	3
tFL-35	chr7|148508727|G	EZH2	nonsynonymous	0.36	2
tFL-37	chr7|148508728|T	EZH2	nonsynonymous	0.27	3
FL-1	chr12|49433220|A	KMT2D(MLL2)	stopgain	0.559	2
tFL-1	chr12|49433220|A	KMT2D(MLL2)	stopgain	0.951	2
FL-2	chr12|49431070|C	KMT2D(MLL2)	nonsynonymous	0.062	NA
tFL-2	chr12|49431070|C	KMT2D(MLL2)	nonsynonymous	0.271	2
tFL-4	chr12|49435906|-T	KMT2D(MLL2)	frameshift_deletion	0.21	3
tFL-4	chr12|49443750|-T	KMT2D(MLL2)	frameshift_deletion	0.493	3
FL-5	chr12|49418731|A	KMT2D(MLL2)	splicing	0.317	NA
FL-5-cus	chr12|49418731|A	KMT2D(MLL2)	splicing	0.4	NA
tFL-5	chr12|49418731|A	KMT2D(MLL2)	splicing	0.395	2
tFL-5-cus	chr12|49418731|A	KMT2D(MLL2)	splicing	0.47	NA
FL-5	chr12|49438694|T	KMT2D(MLL2)	stopgain	0.486	NA
FL-5-cus	chr12|49438694|T	KMT2D(MLL2)	stopgain	0.34	NA
tFL-5	chr12|49438694|T	KMT2D(MLL2)	stopgain	0.531	2
tFL-5-cus	chr12|49438694|T	KMT2D(MLL2)	stopgain	0.46	NA
FL-6	chr12|49427255|A	KMT2D(MLL2)	stopgain	0.521	2
tFL-6	chr12|49427255|A	KMT2D(MLL2)	stopgain	0.364	2
FL-6	chr12|49432396|A	KMT2D(MLL2)	stopgain	0.367	2
tFL-6	chr12|49432396|A	KMT2D(MLL2)	stopgain	0.308	2
FL-8	chr12|49431800|-G	KMT2D(MLL2)	frameshift_deletion	0.545	NA

tFL-8	chr12|49431800|-G	KMT2D(MLL2)	frameshift_deletion	0.395	NA
FL-9	chr12|49433087|T	KMT2D(MLL2)	nonsynonymous	0.504	2
tFL-9	chr12|49433087|T	KMT2D(MLL2)	nonsynonymous	0.427	2
FL-10	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	0.764	2
FL-30	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	0.54	2
tFL-10	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	0.806	2
tFL-30	chr12|49433524|-CT	KMT2D(MLL2)	frameshift_deletion	0.832	2
FL-12	chr12|49426001|-C	KMT2D(MLL2)	frameshift_deletion	0.438	2
tFL-12	chr12|49426001|-C	KMT2D(MLL2)	frameshift_deletion	0.449	2
tFL-15	chr12|49433790|-G	KMT2D(MLL2)	frameshift_deletion	0.25	3
tFL-16	chr12|49427447|A	KMT2D(MLL2)	stopgain	0.39	2
tFL-19	chr12|49425693|-C	KMT2D(MLL2)	frameshift_deletion	0.54	2
tFL-20	chr12|49415846|A	KMT2D(MLL2)	stopgain	0.44	2
tFL-20	chr12|49435698|G	KMT2D(MLL2)	splicing	0.43	2
FL-25-1	chr12|49428448|A	KMT2D(MLL2)	stopgain	0.61	2
FL-25-2	chr12|49428448|A	KMT2D(MLL2)	stopgain	0.66	NA
tFL-25	chr12|49428448|A	KMT2D(MLL2)	stopgain	0.53	2
FL-26-1	chr12|49426598|A	KMT2D(MLL2)	stopgain	0.231	NA
FL-26-2	chr12|49426598|A	KMT2D(MLL2)	stopgain	0.239	3
tFL-26	chr12|49426598|A	KMT2D(MLL2)	stopgain	0.421	2
FL-26-1	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	0.531	NA
FL-26-2	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	0.47	3
tFL-26	chr12|49433902|-AG	KMT2D(MLL2)	frameshift_deletion	0.47	2
FL-27	chr12|49424062|T	KMT2D(MLL2)	splicing	0.23	2
tFL-27	chr12|49424062|T	KMT2D(MLL2)	splicing	0.32	2
FL-27	chr12|49427261|A	KMT2D(MLL2)	stopgain	0.14	2
FL-29	chr12|49418462|C	KMT2D(MLL2)	stopgain	0.24	2
tFL-29	chr12|49420082|A	KMT2D(MLL2)	nonsynonymous	0.501	2
tFL-29	chr12|49420132|A	KMT2D(MLL2)	nonsynonymous	0.409	2
FL-31	chr12|49446429|T	KMT2D(MLL2)	stopgain	0.19	NA
tFL-31	chr12|49446429|T	KMT2D(MLL2)	stopgain	0.29	NA
tFL-34	chr12|49415647|T	KMT2D(MLL2)	stopgain	0.17	2
tFL-34	chr12|49425098|A	KMT2D(MLL2)	stopgain	0.25	2
tFL-35	chr12|49432738|A	KMT2D(MLL2)	stopgain	0.87	2
tFL-36	chr12|49425446|-CT	KMT2D(MLL2)	frameshift_deletion	0.43	2
tFL-37	chr12|49431667|A	KMT2D(MLL2)	stopgain	0.47	2
tFL-37	chr12|49434187|+T	KMT2D(MLL2)	frameshift_insertion	0.39	2
tFL-38	chr12|49424074|T	KMT2D(MLL2)	nonsynonymous	0.16	4
tFL-38	chr12|49435294|A	KMT2D(MLL2)	stopgain	0.76	4
tFL-39	chr12|49435479|A	KMT2D(MLL2)	stopgain	0.34	3
tFL-39	chr12|49435750|A	KMT2D(MLL2)	nonsynonymous	0.51	3

tFL-39	chr12|49435765|C	KMT2D(MLL2)	nonsynonymous	0.51	3
FL-1	chr18|60985508|T	BCL2	nonsynonymous	0.192	3
tFL-1	chr18|60985508|T	BCL2	nonsynonymous	0.353	3
FL-1	chr18|60985760|T	BCL2	nonsynonymous	0.05	3
tFL-4	chr18|60985405|G	BCL2	nonsynonymous	0.237	2
tFL-4	chr18|60985411|C	BCL2	nonsynonymous	0.179	2
FL-6	chr18|60985889|A	BCL2	nonsynonymous	0.267	2
tFL-6	chr18|60985889|A	BCL2	nonsynonymous	0.271	2
FL-8	chr18|60985880|A	BCL2	nonsynonymous	0.244	NA
tFL-15	chr18|60985724|C	BCL2	nonsynonymous	0.219	2
tFL-20	chr18|60985443|G	BCL2	nonsynonymous	0.42	2
FL-22	chr18|60985814|C	BCL2	nonsynonymous	0.121	NA
tFL-22	chr18|60985814|C	BCL2	nonsynonymous	0.169	NA
FL-26-1	chr18|60985883|T	BCL2	nonsynonymous	0.261	NA
FL-26-2	chr18|60985883|T	BCL2	nonsynonymous	0.251	2
tFL-27	chr18|60985722|T	BCL2	nonsynonymous	0.388	2
FL-27	chr18|60985760|T	BCL2	nonsynonymous	0.259	2
tFL-27	chr18|60985760|T	BCL2	nonsynonymous	0.411	2
FL-28	chr18|60985815|G	BCL2	nonsynonymous	0.181	NA
FL-29	chr18|60985529|G	BCL2	nonsynonymous	0.13	2
tFL-29	chr18|60985529|G	BCL2	nonsynonymous	0.331	2
tFL-31	chr18|60985760|T	BCL2	nonsynonymous	0.06	NA
FL-31	chr18|60985880|A	BCL2	nonsynonymous	0.179	NA
tFL-31	chr18|60985880|A	BCL2	nonsynonymous	0.031	NA
tFL-33	chr18|60985301|G	BCL2	nonsynonymous	0.25	2
tFL-33	chr18|60985626|T	BCL2	nonsynonymous	0.19	2
tFL-36	chr18|60985896|T	BCL2	nonsynonymous	0.22	2
tFL-37	chr18|60985286|T	BCL2	nonsynonymous	0.239	4
tFL-37	chr18|60985884|C	BCL2	nonsynonymous	0.32	4
tFL-38	chr18|60985308|T	BCL2	nonsynonymous	0.459	2
tFL-40	chr18|60795973|A	BCL2	nonsynonymous	0.37	2
tFL-40	chr18|60985798|T	BCL2	nonsynonymous	0.42	2
FL-6	chr6|138200146|G	TNFAIP3	nonsynonymous	0.471	0
tFL-6	chr6|138200146|G	TNFAIP3	nonsynonymous	0.542	2
tFL-21	chr6|138200194|G	TNFAIP3	nonsynonymous	0.141	2
tFL-31	chr6|138192455|T	TNFAIP3	stopgain	0.391	NA
FL-27	chr6|134495706|G	SGK1	nonsynonymous	0.15	1
tFL-28	chr6|134495724|C	SGK1	nonsynonymous	0.129	2
tFL-2	chr16|3786206|C	CREBBP	splicing	0.33	2
FL-5	chr16|3807881|A	CREBBP	stopgain	0.233	NA
FL-5-cus	chr16|3807881|A	CREBBP	stopgain	0.37	NA

tFL-5	chr16|3807881|A	CREBBP	stopgain	0.529	2
tFL-5-cus	chr16|3807881|A	CREBBP	stopgain	0.4	NA
FL-5	chr16|3807917|C	CREBBP	nonsynonymous	0.213	NA
FL-5-cus	chr16|3807917|C	CREBBP	nonsynonymous	0.39	NA
tFL-5	chr16|3807917|C	CREBBP	nonsynonymous	0.532	2
tFL-5-cus	chr16|3807917|C	CREBBP	nonsynonymous	0.42	NA
FL-5	chr16|3808030|C	CREBBP	nonsynonymous	0.433	NA
FL-5-cus	chr16|3808030|C	CREBBP	nonsynonymous	0.38	NA
tFL-5	chr16|3808030|C	CREBBP	nonsynonymous	0.316	2
tFL-5-cus	chr16|3808030|C	CREBBP	nonsynonymous	0.42	NA
FL-5	chr16|3808033|G	CREBBP	nonsynonymous	0.414	NA
FL-5-cus	chr16|3808033|G	CREBBP	nonsynonymous	0.38	NA
tFL-5	chr16|3808033|G	CREBBP	nonsynonymous	0.333	2
tFL-5-cus	chr16|3808033|G	CREBBP	nonsynonymous	0.42	NA
FL-5	chr16|3808046|G	CREBBP	nonsynonymous	0.333	NA
FL-5-cus	chr16|3808046|G	CREBBP	nonsynonymous	0.38	NA
tFL-5	chr16|3808046|G	CREBBP	nonsynonymous	0.438	2
tFL-5-cus	chr16|3808046|G	CREBBP	nonsynonymous	0.42	NA
FL-6	chr16|3786739|G	CREBBP	nonsynonymous	0.306	2
tFL-6	chr16|3786739|G	CREBBP	nonsynonymous	0.333	2
FL-8	chr16|3828175|C	CREBBP	stopgain	0.449	NA
tFL-8	chr16|3828175|C	CREBBP	stopgain	0.393	NA
FL-8	chr16|3860683|T	CREBBP	nonsynonymous	0.512	NA
tFL-8	chr16|3860683|T	CREBBP	nonsynonymous	0.396	NA
FL-10	chr16|3786704|G	CREBBP	nonsynonymous	0.481	2
FL-30	chr16|3786704|G	CREBBP	nonsynonymous	0.35	2
tFL-10	chr16|3786704|G	CREBBP	nonsynonymous	0.828	1
tFL-30	chr16|3786704|G	CREBBP	nonsynonymous	0.761	1
FL-11	chr16|3788618|A	CREBBP	nonsynonymous	0.197	2
tFL-11	chr16|3788618|A	CREBBP	nonsynonymous	0.373	2
FL-12	chr16|3788605|G	CREBBP	nonsynonymous	0.413	2
tFL-12	chr16|3788605|G	CREBBP	nonsynonymous	0.51	2
FL-12	chr16|3795311|-A	CREBBP	frameshift_deletion	0.355	2
tFL-12	chr16|3795311|-A	CREBBP	frameshift_deletion	0.49	2
tFL-15	chr16|3788617|T	CREBBP	nonsynonymous	0.328	2
FL-22	chr16|3786706|G	CREBBP	nonsynonymous	0.317	NA
tFL-22	chr16|3786706|G	CREBBP	nonsynonymous	0.262	NA
tFL-23	chr16|3786704|T	CREBBP	nonsynonymous	0.15	2
FL-23-1	chr16|3789578|A	CREBBP	splicing	0.4	2
FL-23-2	chr16|3789578|A	CREBBP	splicing	0.18	NA
tFL-23	chr16|3789578|A	CREBBP	splicing	0.12	2

FL-24-1	chr16|3788618|A	CREBBP	nonsynonymous	0.761	2
FL-24-2	chr16|3788618|A	CREBBP	nonsynonymous	0.789	NA
tFL-24	chr16|3788618|A	CREBBP	nonsynonymous	0.661	2
FL-25-1	chr16|3900873|A	CREBBP	stopgain	0.21	2
FL-25-2	chr16|3900873|A	CREBBP	stopgain	0.31	NA
tFL-25	chr16|3900873|A	CREBBP	stopgain	0.25	2
FL-26-1	chr16|3781333|T	CREBBP	nonsynonymous	0.31	NA
FL-26-2	chr16|3781333|T	CREBBP	nonsynonymous	0.34	2
tFL-26	chr16|3781333|T	CREBBP	nonsynonymous	0.4	2
FL-26-1	chr16|3831211|+C	CREBBP	frameshift_insertion	0.23	NA
tFL-26	chr16|3831211|+C	CREBBP	frameshift_insertion	0.3	2
FL-27	chr16|3788618|A	CREBBP	nonsynonymous	0.209	2
tFL-27	chr16|3788618|A	CREBBP	nonsynonymous	0.791	2
FL-31	chr16|3788618|A	CREBBP	nonsynonymous	0.141	NA
tFL-31	chr16|3788618|A	CREBBP	nonsynonymous	0.377	NA
tFL-34	chr16|3790421|T	CREBBP	nonsynonymous	0.2	2
tFL-35	chr16|3790421|T	CREBBP	nonsynonymous	0.86	2
tFL-37	chr16|3820888|+GTGCA	CREBBP	frameshift_insertion	0.45	2
tFL-39	chr16|3788657|G	CREBBP	nonsynonymous	0.311	2
tFL-5	chr17|7577545|C	TP53	nonsynonymous	0.378	2
tFL-5-cus	chr17|7577545|C	TP53	nonsynonymous	0.833	NA
tFL-5	chr17|7577598|A	TP53	nonsynonymous	0.56	2
tFL-5-cus	chr17|7577598|A	TP53	nonsynonymous	0.107	NA
tFL-22	chr17|7577097|A	TP53	nonsynonymous	0.176	NA
tFL-27	chr17|7578265|G	TP53	nonsynonymous	0.63	1
tFL-29	chr17|7576571|C	TP53	stopgain	0.469	2
tFL-29	chr17|7577093|G	TP53	nonsynonymous	0.511	2
tFL-32	chr17|7578286|G	TP53	nonsynonymous	0.371	2

The first column (Sample ID) indicates the sample in which the mutations were detected. The second column (Chromosome| Position |Altered-Base) indicates the coordinates of the mutation. The third column indicates the gene altered by the mutation. The fourth column indicates the mutation type. The fifth column indicates the variant frequency of the mutation. The sixth column indicates the CN estimated for this gene from our previous study; note that all CN>3 are shown as 4. NA is noted if we do not have the CN information.

APPENDIX C: MUTATIONS IDENTIFIED IN REGIONS OF COPY NUMBER ABNROMALITIES

rCNA	CNAband	Frequency in FLs	Frequency in FLs	Recurrently mutated genes in our cases
964	18q+	35%	41%	BCL2
965	18+	31%	32%	BCL2
122	1p36.33-p36.31-	25%	24%	TNFRSF14
799	7q+	24%	39%	EZH2, MLL3
798	7p+	23%	38%	CARD11
1012	Xq+	23%	24%	TAF1
1011	Xp+	21%	23%	DMD, DDX3X, HUWE1
800	7+	21%	37%	CARD11, EZH2, MLL3
1013	X+	20%	19%	DMD, DDX3X, HUWE1, TAF1
688	1q+	16%	28%	ARHGEF2, EDEM3
442	10q23.1-q25.1-	15%	14%	C10orf12
953	17q+	15%	13%	STAT3, MIR142, CD79B, BPTF, RNF213
339	6q-	14%	15%	EEF1A1, MDN1, PRDM1, SGK1, TNFAIP3, SYNE1
785	6p+	14%	18%	HIST1H1C, HIST1H1E, HIST1H1D, HIST1H2AM, PIM1, CCND3, CUL7
880	12q+	13%	20%	MLL2, BTG1, DTX1
993	21q+	12%	20%	PRDM15
711	2p+	12%	8%	BIRC6, SPTBN1, ALMS1
762	5p+	12%	11%	IL7R
994	21+	12%	18%	PRDM15
881	12+	11%	15%	MLL2, BTG1, DTX1
593	17p-	9%	18%	TP53, NCOR1
882	12q12-q13.13+	9%	10%	MLL2
763	5q+	9%	10%	PIK3R1
304	6q23.3-q24.1-	8%	13%	TNFAIP3
713	2+	8%	5%	BIRC6, SPTBN1, ALMS1
764	5+	8%	8%	IL7R, PIK3R1
585	17p13.3-p13.1-	7%	15%	TP53
341	6q13-q15-	7%	6%	EEF1A1, MDN1
856	11p+	7%	16%	HPS5
37	6q23.3- (x2)	6%	10%	TNFAIP3
340	6q23.2-q25.1-	6%	10%	SGK1, TNFAIP3
153	1p36.33-p36.12-	6%	8%	TNFRSF14, SPEN
857	11q+	6%	15%	ETS1
1191	18q+ (x2)	5%	6%	BCL2

935	16p+	5%	8%	CREBBP, SOCS1
124	1p36.11-p35.3-	5%	5%	ARID1A
936	16q+	5%	9%	IRF8, FANCA
989	20q+	5%	8%	TM9SF4, SLC9A8
990	20+	5%	8%	TM9SF4, SLC9A8
422	10q24.1-	4%	11%	C10orf12
543	15q21.1-	4%	11%	B2M
444	11q24.3-	4%	3%	ETS1
858	11+	4%	13%	HPS5, ETS1
937	16+	4%	6%	CREBBP, SOCS1, IRF8, FANCA
974	19p+	4%	3%	DAZAP1, SMARCA4, MEF2B
342	6q16.3-q22.33-	4%	4%	PRDM1
1087	7q+ (x2)	4%	1%	EZH2, MLL3
839	10q+	4%	4%	C10orf12
840	10+	4%	4%	C10orf12
976	19+	4%	3%	DAZAP1, SMARCA4, MEF2B
714	2p16.3-p14+	3%	5%	SPTBN1
769	6p21.2-p21.1+	3%	1%	CCND3, CUL7
1014	Xp21.1-p11.1+	3%	5%	DDX3X, HUWE1
1078	6p+ (x2)	3%	1%	HIST1H1C, HIST1H1E, HIST1H1D, HIST1H2AM, PIM1, CCND3, CUL7
1227	Xq+ (x2)	3%	4%	TAF1
558	15q-	3%	13%	MGA, B2M
665	Xp-	3%	4%	DMD, DDX3X, HUWE1
666	Xq-	3%	5%	TAF1
733	3q+	3%	10%	KLHL6
564	16p13.3-	3%	5%	CREBBP
787	7p22.3-p21.3+	3%	8%	CARD11
966	18q21.2-q21.33+	3%	4%	BCL2
1002	22q+	3%	1%	IGLL5, EP300
1027	1q+ (x2)	3%	3%	ARHGEF2, EDEM3
1086	7p+ (x2)	3%	6%	CARD11
1192	18+ (x2)	3%	0%	BCL2
1226	Xp+ (x2)	3%	5%	DMD, DDX3X, HUWE1
605	18q21.33-q23-	3%	4%	BCL2
667	X-	3%	4%	DMD, DDX3X, HUWE1, TAF1
732	3p+	3%	8%	MYD88
734	3+	3%	6%	MYD88, KLHL6
952	17p+	3%	3%	TP53, NCOR1
954	17+	3%	3%	TP53, NCOR1, STAT3, MIR142, CD79B, BPTF, RNF213

312	6q21-	2%	5%	PRDM1
1088	7+ (x2)	2%	1%	CARD11, EZH2, MLL3
1143	12q+ (x2)	2%	0%	MLL2, BTG1, DTX1
1228	X+ (x2)	2%	3%	DMD, DDX3X, HUWE1, TAF1
515	13q-	2%	3%	RFXAP, FOXO1
687	1p+	2%	1%	TNFRSF14, SPEN, ARID1A, RRAGC, RBM15
689	1+	2%	1%	TNFRSF14
735	3q26.1-q29+	2%	3%	KLHL6
898	13q+	2%	5%	RFXAP, FOXO1
1042	2p21-p14+ (x2)	2%	1%	SPTBN1
1122	11p+ (x2)	2%	1%	HPS5
1144	12+ (x2)	2%	0%	MLL2, BTG1, DTX1
1203	21q+ (x2)	2%	1%	PRDM15
1204	21+ (x2)	2%	0%	PRDM15
561	15q11.2-q21.1-	2%	0%	MGA, B2M
583	16q-	2%	4%	IRF8, FANCA
668	Xp22.33-p21.1-	2%	1%	DMD
801	7p22.3-p21.1+	2%	1%	CARD11
919	15q+	2%	5%	MGA, B2M
946	17p12-p11.2+	2%	1%	NCOR1
968	19p13.3+	2%	3%	DAZAP1
1065	5p+ (x2)	1%	3%	IL7R
1075	6p21.32-p12.2+ (x2)	1%	0%	PIM1, CCND3, CUL7
1182	17p12-p11.2+ (x2)	1%	0%	NCOR1
119	Xp- (x2)	1%	0%	DMD, DDX3X, HUWE1
1215	Xp11.22-q11.1+ (x2)	1%	3%	HUWE1
1223	Xp11.4+ (x2)	1%	0%	DDX3X
1229	Xp11.23-q11.1+ (x2)	1%	1%	HUWE1
152	1p-	1%	5%	TNFRSF14, SPEN, ARID1A, RRAGC, RBM15
223	3q27.1-	1%	1%	KLHL6
269	4p-	1%	9%	HTT
343	6p22.2-p21.33-	1%	1%	HIST1H1C, HIST1H1E, HIST1H1D, HIST1H2AM
470	11q23.3-q25-	1%	3%	ETS1
581	16q24.1-q24.3-	1%	0%	FANCA
653	22q-	1%	1%	IGLL5, EP300
765	5p11-q13.3+	1%	0%	PIK3R1

854	11q24.3+	1%	0%	ETS1
872	12q13.12+	1%	3%	MLL2
955	17q22-q24.1+	1%	1%	MIR142, CD79B
1007	Xp21.2-p21.1+	1%	3%	DMD
1022	1p36.33-p36.32+ (x2)	1%	1%	TNFRSF14
1066	5q+ (x2)	1%	0%	PIK3R1
1067	5+ (x2)	1%	0%	IL7R, PIK3R1
1082	7p22.3-p22.1+ (x2)	1%	1%	CARD11
1089	7p22.3-p21.1+ (x2)	1%	3%	CARD11
1138	12q13.12-q13.13+ (x2)	1%	1%	MLL2
1173	16p+ (x2)	1%	0%	CREBBP, SOCS1
1174	16q+ (x2)	1%	0%	IRF8, FANCA
1175	16+ (x2)	1%	0%	CREBBP, SOCS1, IRF8, FANCA
1185	17q+ (x2)	1%	0%	STAT3, MIR142, CD79B, BPTF, RNF213
120	Xq- (x2)	1%	0%	TAF1
121	X- (x2)	1%	0%	DMD, DDX3X, HUWE1, TAF1
154	1p21.1-p12-	1%	5%	RBM15
155	1q21.1-q23.3-	1%	1%	ARHGEF2
236	3p-	1%	1%	MYD88
270	4q-	1%	8%	FBXW7
271	4-	1%	8%	HTT, FBXW7
332	6p21.1-	1%	1%	CCND3
333	6p21.1-p12.3-	1%	1%	CUL7
469	11q-	1%	0%	ETS1
493	12q23.3-q24.31-	1%	1%	DTX1
557	15q15.1-q15.3-	1%	1%	MGA
562	15q13.1-q22.2-	1%	1%	MGA, B2M
582	16p-	1%	0%	CREBBP, SOCS1
619	19p-	1%	1%	DAZAP1, SMARCA4, MEF2B
621	19-	1%	0%	DAZAP1, SMARCA4, MEF2B
951	17q22-q24.2+	1%	1%	CD79B
985	20q13.13+	1%	1%	SLC9A8
100	17p- (x2)	0%	1%	TP53, NCOR1
1051	3p+ (x2)	0%	1%	MYD88
1052	3q+ (x2)	0%	1%	KLHL6
1053	3+ (x2)	0%	1%	MYD88, KLHL6
1079	6p25.3-p21.33+ (x2)	0%	3%	HIST1H1C, HIST1H1E, HIST1H1D, HIST1H2AM

108	19p- (x2)	0%	1%	DAZAP1, SMARCA4, MEF2B
193	2p-	0%	1%	BIRC6, SPTBN1, ALMS1
237	3q-	0%	3%	KLHL6
238	3-	0%	1%	MYD88, KLHL6
274	4p16.3-p15.2-	0%	3%	HTT
441	10q-	0%	1%	C10orf12
594	17q-	0%	3%	STAT3, MIR142, CD79B, BPTF, RNF213
595	17-	0%	3%	TP53, NCOR1, STAT3, MIR142, CD79B, BPTF, RNF213
637	20q-	0%	1%	TM9SF4, SLC9A8
752	4p+	0%	1%	HTT
753	4q+	0%	1%	FBXW7
754	4+	0%	1%	HTT, FBXW7
805	7q33-q36.3+	0%	3%	EZH2, MLL3
895	13q13.2-q13.3+	0%	3%	RFXAP