

UNIQUE CHROMOSOME ABERRATIONS DISTINGUISH DIFFUSE LARGE B-
CELL LYMPHOMA AND BURKITT LYMPHOMA: AN ANALYSIS OF
CYTOGENETICS AND PREDICTOR MODEL CLASSIFIERS

By

Rolando García

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics
School of Health Related Professions
Rutgers, the State University of New Jersey

May 2015



Final Dissertation Defense Approval Form

UNIQUE CHROMOSOME ABERRATIONS DISTINGUISH DIFFUSE LARGE B-CELL
LYMPHOMA AND BURKITT LYMPHOMA: AN ANALYSIS OF CYTOGENETICS
AND PREDICTOR MODEL CLASSIFIERS

BY

Rolando García

Dissertation Committee

Shankar Srinivasan PhD – committee chair

Frederick Coffman PhD – committee member

Shashi Mehta PhD – committee member

Approved by the Dissertation Committee

_____	Date _____
_____	Date _____
_____	Date _____

TABLE OF CONTENTS

I. Introduction	1
1.1 Introduction	1
1.2 Background Information	2
1.3 Statement of Problem	9
1.4 Objectives and Goals of the Research	15
1.5 Research Hypothesis	16
II. Literature Review	19
2.1 Epidemiology	18
2.2 Morphological and Phenotypic Features of BL and DLBCL	20
2.3 Cytogenetic Aberrations in DLBCL and BL	34
2.4 Chromosome Imbalances by Array Comparative Genomic Hybridization in DLBCL subtypes and BL	37
2.5 Gene Expression Profiles, Micro RNA, and Single Nucleotide Polymorphisms in DLBCL and BL	40
2.6 Classification of Non Hodgkin Lymphoma by Bioinformatics Methods: Logistic Regression, Unsupervised Cluster Analysis, and Artificial Neural Networks	48
III. Methods	55
3.1 Overview	52
3.2 CyDAS: an Online Bioinformatics Tool-kit for Chromosome Analysis	56

3.3	Analysis of the Mitelman Dataset in CyDAS.....	58
3.4	Building a Reliable Set of RCAs for Distinguishing DLBCL (MYC+ and MYC-) vs. BL.....	60
3.5	Cluster Predictive Models of DLBCL and BL.....	63
3.6	Logistic Regression Model of DLBCL and BL.....	65
3.7	Neural Networks and SVM models for the Classification of DLBCL and BL.....	66
3.8	Discrimination (AUC and specificity) Studies of Predictive Models.....	68
3.9	Internal validation Techniques.....	69
3.10	External Validation: an Extended Literature Review of Array CGH in DLBCL Subtypes vs. BL and Compared with RCAs.....	69
IV.	Results.....	71
4.1	Genetic Complexity of DLBCL vs. BL and CyDAS analysis	71
4.2	RCAs between MYC+ DLBCL and BL.....	73
4.3	RCAs Differences between MYC+ and MYC- DLBCL.....	73
4.4	RCAs Differences between DLBCL (MYC+ MYC-) vs. BL.....	75
4.5	Differences between DLBCL (<i>MYC+</i> and <i>MYC-</i>) vs. BL in a Multivariable Context	78
4.6	Cluster Analysis of DLBCL and BL based on RCAs	79
4.7	Discrimination Analysis: Area under the Curve and Specificity...	89
4.8	A Core Set of RCAs (Primary RCAs)	90

4.9	Additional RCAs Identified DLBCL vs. BL.....	92
4.10	RCAs Guidelines for Distinguishing DLBCL from BL.....	98
4.11	Correlation of RCAs and Copy Number Aberrations in DLBCL and BL.....	100
4.12	CNVs Differences between DLBCL and BL: a CGH Dataset Analysis.....	101
V.	Discussion.....	106
5.1	Overview.....	106
5.2	A Complex Chromosome Complement Defines DLBCL from BL.	107
5.3	RCAs Delineate DLBCL (MYC+ and MYC-) vs. BL.....	109
5.4	Reliable RCAs in the Published Literature.....	110
5.5	RCAs Define Molecular Subtypes of DLBCL.....	112
5.6	RCAs Applied to Supervise Predictor Models Show High Specificity.....	114
5.7	Surrogate markers associated with RCAs may Prove Helpful in Distinguishing DLBCL from BL	116
5.8	Extended Literature Review and Copy Number Variations Correlates with Cytogenetic Data.....	117
VI.	Summary and Conclusion	
6.1	Conclusion.....	122
6.2	Summary.....	126
6.3	Future Research.....	128

Summary of Contributions.....	130
References.....	131

ABSTRACT

Background: Under current classification of lymphoid neoplasms, majority of lymphomas can be reliably classified; however, overlapping features between diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL) with or without *MYC* gene rearrangement are problematic to diagnose.

Purpose: The aim of this study was to identify recurrent chromosome abnormalities to distinguish these entities and to test for their specificities using predictor models.

Dataset and Methods: The Study involved the analysis of publicly available information and institutional cases. Two distinct datasets were used to build (n = 338) and test (n=177) predictor models. An independent group t-test performed with the Statistical Analysis Software (SAS) was used to assess the differences in the number of aberrations between groups. The Fisher exact test was then used to determine correlations between RCAs and the two entities. A p-value less than .05 was considered significant. Discrimination of models was determined by the ROC curve. All analyses were performed using R; only SAS was used for a logistic regression model. Subsequent supervised models were constructed (n = 515) and copy number variations analysis (n = 249) was conducted for validity.

Results: RCAs associated with DLBCL: +X, 1qL, 1p36L, +2, -2, +3, -4, +7, -8, 9qL, +12, 14qL, 15qL, 16qL, +16, 17pL, +18, 19pL and 22qL. Specificity of models was 85-100%. In terms of the area under the curve (AUC) of the ROC curve, predictor classifiers were classified as excellent models (0.9 - 0.93). Only the LR was below 0.9. When datasets were combined, additional RCAs were identified (6pG, 6qL, +5, +11, -

10/-15, -10/-14, 1qG and 13qL), with latter two describing BL. Subsequent analysis by an artificial neural network model showed a specificity of 95-100%. In terms of validity, findings from an extended array CGH review showed a number of RCAs correlated with copy number aberrations. Moreover, an analysis of CNVs revealed similar results.

Conclusion: Our findings revealed unique RCAs that suggest distinct biological activities between DLBCL and BL, these RCAs may be used to augment diagnostic accuracy and help clinicians better manage these patients. In terms of predictor classifiers, ANN models outperformed all others classifiers.

Acknowledgements

First, I would like to dedicate this work to my wife, Patricia García and to all patients suffering from diffuse large B-cell lymphoma and Burkitt lymphoma – may this information provided here serve to advance patient care. I would also like to thank everyone involved in my graduate formation throughout my studies at Rutgers University - School of Health Related Professions. Special thanks to Dr. Elaine Keohane for her patience, time invested on my papers, and her valuable comments (“distill the most important information”). I would also like to thank all my members of my doctoral committee for their valuable input and comments, Dr. Shankar Srinivasan, Dr. Frederick Coffman and Dr. Mehta Shasi. Especial thanks to Dr. Srinivasan who always had a smile in his e-mails and most importantly that guided me throughout the dissertation phases – many and many thanks, and in a more playful note that kept my marriage intact. In addition, I would also like to thank Dr. Prasad Koduru, cytogenetic director at UT Southwestern, who has taught me to illustrate and better explain my research results.

To my family, who has always supported my educational endeavors even in the first grade – thanks mom (Irene Mendoza), and my special thanks to my wife, who is a constant champion for the things I do, not to mention my personal public relation supporter – I love you. Also, a special thanks to my nephew-in law (Sebastian Yepes) for delivering the dissertation to the doctoral committee under adverse weather conditions. Moreover, thanks to all our friends who put up with our absences throughout the PhD studies. Lastly, I cannot probably count the ways how our God has been there with us

supporting our journey of life every step of the way – May God continue to bless every one of us, inspire us to help and to teach us to become more Christ-like.

LIST OF TABLES

Table 1-1: Distinguishing Features of DLBCL and BL.....	13
Table 2-1: Immunophenotypic and Genetic Characteristics of DLBCL and BL...	30
Table 2-2: Cytogenetic Profile of DLBCL and BL.....	36
Table 4-1: Quantitative Representation of Aberrations at Chromosome 1p by CyDAS.....	72
Table 4-2: Morphological Groups for <i>MYC</i> ⁺ DLBCL and BL.....	73
Table 4-3: RCAs Differences between <i>MYC</i> ⁺ and <i>MYC</i> ⁻ DLBCL.....	75
Table 4-4: RCAs Differences between characteristic groups of DLBCL vs. BL.....	76
Table 4-5: Cytogenetic Findings in ISCN format of DLBCL and BL samples.....	79
Table 4-6: Additional RCAs Associated with DLBCL (<i>MYC</i> ⁺ and <i>MYC</i> ⁻) and BL.....	92
Table 4-7: Classification of RCAs based on Molecular Subtypes of DLBCL and BL.....	94
Table 4-8: Genomic Imbalances in DLBCL and BL and Correlated RCAs.....	100
Table 4-9: Total Number of Cases and Subgroups of DLBCL and BL.....	102
Table 4-10: CNVs Differences between DLBCL and BL.....	102

Table 4-11: Significant CNVs Differences between Distinct DLBCL Subtypes	
and BL.....	104
Table 5-1: Features of DLBCL (GCB and ABC subtypes).....	113

LIST OF FIGURES

Figure 1-1: A multiple input model of the perceptron.....	7
Figure 1-2: Morphological characteristics of DLBCL and BL.....	12
Figure 2-1: Chromosome ideogram of gains and losses by array CGH in DLBCL and BL.....	39
Figure 3-1: Qualitative and quantitative aberrations generated by CyDAS.....	57
Figure 3-2: Screenshot of the Mitelman Case Quick Searcher.....	58
Figure 3-3: Screenshot of the CyDAS interphase analysis of datasets for gains and Losses.....	59
Figure 3-4: Datasets and analysis work-flow.....	61
Figure 3-5: Typical column dendrogram of a hierarchical cluster.....	63
Figure 3-6: Schematic representation of an ANN model.....	67
Figure 4-1: Distribution of gains and losses in DLBCL and BL identified by CyDAS.....	72
Figure 4-2: Ideogram representation of characteristic gains and losses in DLBCL vs. BL.....	78
Figure 4-3: Hierarchical cluster heat map.....	87

Figure 4-4: Hierarchical cluster with p-values	88
Figure 4-5: ANN classification of the test dataset.....	89
Figure 4-6: Receiver operating characteristic (ROC) curve of prediction models...	90
Figure 4-7: Hierarchical cluster dendrogram with p-values of RCAs.....	91
Figure 4-8: Correlational matrix of the total number of RCAs.....	93
Figure 4-9: ROC curves of supervised models.....	96
Figure 4-10: Odds ratio estimates of RCAs implicated in DLBCL.....	97
Figure 4-11: Discrimination and calibration of logistic regression model.....	98
Figure 4-12: Proposed RCA Algorithm to better classified DLBCL and BL	99
Figure 5-1: Schematic of the Hans algorithm.....	114
Figure 5-2. PI3K pathway in BL.....	119

LIST OF EQUATIONS

3-1: Specificity = True Negative / True Negative + false positive..... 62

3-2: $p = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k} / 1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k}$ 66

LIST OF ABBREVIATIONS

ABC.....	Activated B-cell
ANNs.....	Artificial neural networks
AUC.....	Area under the curve
DLBCL.....	Diffuse large B-cell
lymphoma	
BL.....	Burkitt lymphoma
CNAs.....	Copy number aberrations
CNVs.....	Copy number variations
CGH.....	Comparative genomic
hybridization	
GCB.....	Germinal center B-cell
HCs.....	Hierarchical clusters
NHL.....	Non-Hodgkin lymphoma
ISCN.....	International system for
	Human cytogenetic
	Nomenclature
KI-67.....	Proliferation index
LG.....	Logistic regression
q arm.....	Long arm of a chromosome
p arm.....	Short arm of a chromosome
R.....	R statistical package
RCAs.....	Recurrent cytogenetic
aberrations	
ROC.....	Receiver operating curve

SAS.....	Statistical analysis software
SVM.....	Support vector machine

Chapter I

I. Introduction

1.1 Introduction

Under the current World Health Organization (WHO) Classification scheme of hematolymphoid neoplasms, diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL) fall under the umbrella of mature B-cell non-Hodgkin (NHL) lymphomas. Distinguishing between these two entities is critical since prognosis and treatment greatly differs. BL is not cured with existing therapies used for DLBCL and requires more aggressive protocols. Therefore, an accurate diagnosis is critical to avoid the over-treatment or under-treatment of these patients. Although establishing a diagnosis for DLBCL and BL may appear straightforward, differential diagnosis between the two remains problematic in daily clinical practice.¹ This is due to the overlapping morphological, immunophenotypic and cytogenetic features between these two entities.¹⁻³ In fact, the Non-Hodgkin's Lymphoma Classification project identified just 53% agreement between experienced hematopathologist when distinguishing BL, atypical BL and DLBCL.⁴ To further support this view, two recent DNA microarray studies have shown that nearly 6% of histological DLBCL actually represent the molecular BL subtype.^{5,6} In view of these challenges, the 2008 WHO classification established the new provisional category, B-cell lymphoma, unclassifiable, with features intermediate between DLBCL and BL (I/DLBCL-BL).⁷ This category only serves to bring together B-cell lymphomas that share both BL or DLBCL features but failed a diagnosis of BL or DLBCL. Within this new classification, the WHO also outlined BL by a simple *MYC* gene positive karyotype compared to more complex karyotypes for I-DLBCL/BL and

DLBCL. However, this may not be entirely correct since molecular studies have shown an increase in genetic complexity in the molecular BL subtype.^{5,8} Chromosome structural rearrangements are also now included in this classification system, for example, genomic gains of the long arm of chromosome 3 (3q), short arm of chromosome 9 (9p) and 18q are classified as activated B-cell DLBCL (ABC-DLBCL) and chromosome translocations between chromosome 14 and 18 [t(14;18)] and 12 q gains belong to the germinal center GC-DLBCL subset, while t(14;18) now rules out BL.⁷ Despite such efforts to properly assign BL from DLBCL, additional genetic criteria are clearly needed.

1.2 Background Information

To better distinguish BL from DLBCL, this current study will use an integrative approach analysis of cytogenetic data, gene expression profiles and bioinformatic tools to augment diagnostic accuracy. In the following paragraphs, we provide background information on the various technologies that will be used in this study.

Cytogenetics as a Diagnostic Tool for Hematological Disorders

Cytogenetics is a subspecialty of pathology that studies chromosome structure and variation in cancer and non-cancer states. Normally, an individual carries a total of 22 homologous chromosomes in addition to a sex chromosome pair (XY for males and XX for females) for a total of 46 chromosomes. Each chromosome has a distinctive band pattern when trypsin is applied to chromosomes and stain with Wright stain (G-banded chromosomes). Given such distinct banding pattern, each chromosome can be readily identified under light microscopy; it is this banding pattern that serves to describe structural variations in chromosomes and also serves as the basis for chromosome

nomenclature (the International System for Human Cytogenetic Nomenclature or ISCN). Because of the advances in cytogenetic techniques, it is nowadays an integral part of the diagnosis, prognosis and therapy management for many hematological malignancies.

Molecular Cytogenetics: Comparative Genomic Hybridization (CGH)

Introduced in the early 90's, classical CGH was the first molecular cytogenetic technique that offered a comprehensive assessment of the entire genome. Both classical and array CGH allow for the detection of DNA copy number changes providing a map of the regions in the genome that are gain or lost. Nowadays, array CGH has but completely replaced classical CGH; however, for the purpose of this research study and because there is valuable data available for analysis on many disease states, including DLBCL and BL, we will describe classical CGH in more detail. In classical CGH, differentially labeled DNA (test DNA in green and control in red) are hybridized to chromosome metaphase spreads. A copy number difference between a test and control sample is shown as a green and red fluorescence ratio difference on metaphase spreads. This fluorescence ratio between test and control DNA is carried out by an imaging system. In the case of genomic regions that are gained or amplified, an increased in green fluorescence ratio is detected, whereas regions of the chromosomes that are lost produced a decreased fluorescence ratio. The advantage of this technique is the wide area screening of the genome for copy number changes and the amount of DNA needed (a few nanograms of DNA), while its disadvantage is resolution. Genomic gains or losses smaller than 5-10 MB are not detected.⁹ In the case of array CGH, this limitation is overcome.

Bioinformatics

In the last decade, with the completion of the human genome project in 2003 and the rapidly evolving microarray-based technologies, an explosion of biological and biomedical data has become available. It is this vast information in emerging fields such as genomics, transcriptomics and proteonomics that has precipitated the rise of bioinformatics, an information science that is concerned with the application of computer techniques and technologies that gather, manipulate classified, store, retrieve, visualize, predict and analyze data. Bioinformatics is an interdisciplinary science that interphase biology, computer science, statistics and mathematics all merged into a single discipline. Similarly, the National Institute of Health (NIH) working definition of bioinformatics is the development, research tools and application of computational applications to expand the use of biological and medical data, but also to store, acquire, organize, archive, analyze or to visualize these data. The impact of bioinformatics in the biomedical field is not only due to its ability to handle large volumes of data, but also in the ability to predict, evaluate and interpret clinical findings. Thus, the goal of bioinformatics is to uncover and find patterns in this wealth of information and gain a better understanding of the genome and biological systems. For the purpose of this study, bioinformatics is essential to uncover distinguishing genetic features between DLBCL and BL from large datasets, and also to predict between these two entities.

Logistic Regression, Unsupervised and Supervised Prediction Models

Predictive models are widely used in data classification tasks throughout the literature. The following sections offer a brief summary of the various models that will be used in this study.

In a logistic regression model, the relationship between categorical dichotomous outcome variables is modeled. The model predicts an outcome based on a series of predictor variables, in our case, a set of recurrent cytogenetic aberrations that predict BL or DLBCL. In addition, it also provides information on how important independent variables relate to the outcome and generates an odds ratio for this relationship. Once a model has been selected, a test of goodness-of-fit is performed; this is usually the Hosmer-Lemeshow test that determines the prediction ability of the model or “calibration”.¹⁰ This test generates a chi-square statistic, where a small value describes a better performance of the model. Another calibration tool to assess the quality of LR models is calibration plots. It compares the similarity between two approximations of a probability, in other words, it plots fitted vs. actual values. A second feature that is used to test the accuracy of the model also referred to “discrimination” or how well the model distinguishes between the two different tumor types is the receiver-operating characteristic (ROC) curve. It is the most commonly used tool to evaluate discrimination¹¹ and it is most often used with diagnostic tests based on specificity and sensitivity. Finally, and perhaps more importantly, is the validation of the model. This is normally carried out with a test dataset, that is, a different set of data that was used to build the model.

Another statistical method used in classification tasks is the analysis of clusters. Clusters analysis is a form of unsupervised learning, that is, clusters are partitioned and derived from no previous knowledge about the classification of objects. There is a large number of clustering analysis methods; however, the most popular method is the k-means partitioning algorithm.¹² Partitioning methods are helpful when a set of clusters are

desired.¹³ For numerical and binary datasets, the k-means algorithm has a mean vector that finds the nearest cluster for an object, while at the same time reducing object dissimilarity by a distance metric, for example, a Euclidian distance. K-means repetitively designates objects to the nearest cluster revising the mean with every repetition cycle. The repetition goes on until a certain threshold specified by the user is reached, that is, pending a specified number of objects changing clusters.¹² There are limitations to k-means, such as processing speed, parameters that can be designated by the user, responsiveness to outliers and initial means, as well as shaped clusters. Various other algorithms have been developed to address some of these shortcomings. One alternative to using the k-means and that addresses outliers is the k-medoids algorithm or PAM (partition around medoids). In this instance, the k-medoids assigns the mean to the object that is closest to the center and reduces distances among the different objects within a cluster.¹⁴

Supervised Models: Artificial neural networks-ANNs and Support Vector Machine-SVM

Unlike unsupervised methods that derive clusters with no previous knowledge of classification, supervised models classify by learning experience, that is, models are trained and build a predictive model that is subsequently tested against a test dataset. One such example of a supervised model is artificial neural networks (ANNs). ANN models are supervised learning methods that are highly precise in the assessment of outcome prediction. These models also generate near-perfect classifications and are able to depict causality between input and output data using IF-THEN rules.¹⁵⁻¹⁷ These networks achieve this by constantly adjusting weighted connections to minimize errors in

matching input to the output datum. In this context, ANNs have the capacity to learn by example by way of highly interconnected processing units referred to as a perceptron that is analogous to a biological neuron. Indeed, it is the perceptron that serves as the basic framework for neural networks and the processing information parallels that of the brain. Figure 1-1 illustrates the basic structure and function of the perceptron.

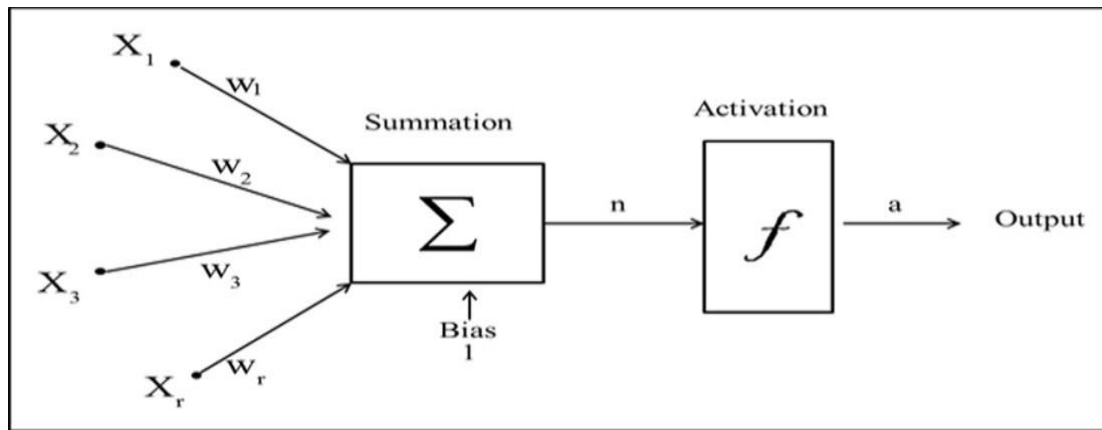


Figure 1-1. A multiple input model of the perceptron. Single inputs $X_1, X_2, X_3, \dots, X_r$ are weighted suitable elements $W_1, W_2, W_3, \dots, W_r$. The summation of the weighted input and the bias equal to 1 represents the total input (n). Thereafter, it continues into to an activation function (f) that consequently generates the neuron output (a). Figure was adapted from Hanrahan et al.¹⁸

Neural networks generally use a sigmoid function due to the fact that it possess better learning capability and improved precision.¹⁹ However, in recent years investigators have explored other types of functions such as the radial basis functions (RBFs).¹⁸ Given the functionality of a single perceptron, it is limited by a given weight and threshold value; however, when multiple neurons are interconnected in layers

moving forward to a single output the power of neural networks is truly appreciated. In fact, multilayered perceptrons (MLP) are well suited for classifications, pattern recognition, predictions and decision support systems. Another type of supervised learning method that is used in biological systems for classification tasks is a support vector machine (SVM) model. This approach produces a lower number of classification inaccuracies compared to other methods.²⁰ First introduced in 1995 by Cortes and Bapnik²¹, it maps input vectors to high dimensional planes (hyper-planes) instead of linear maps; it then uses this hyper-plane to construct a linear decision border between two different classes or groups. In this regard, with the existence of one hyper-plane, there exist an infinite number of hyper-planes. The optimal feature space or hyper-plane that is used by the model to separate between groups is selected based on the closest point of each group to the hyper-plane, but at the same time as measured distance, usually a Euclidian distance that is as farthest away from this optimal feature space as possible.²⁰

Internal and External Validation Techniques

The objective of any predictive model is to deliver an effective outcome prediction for future patients, in our case, new patients with a suspected differential diagnosis of DLBCL and BL. Thus, the dataset and analysis to construct prediction models is essentially only of interest to learn for future events. Therefore, validation is an important element of the process when constructing diagnostic classifiers or prediction models. Here, two types of validity are considered, internal and external validity. In terms of internal validity, it is the reproducibility of the model, that is, different settings are performed within the same underlying population and then results from these settings are then evaluated to determine the validity of the model. For the purpose of this study,

the following are used: split-sample validation (samples are divided into two different groups), cross validation (assessment of model performance is carried out by assigning a random part of the underlying population consecutively, usually in deciles or a 10 cross validation that functions as validation component) and boot-strap validation (the process of sampling with replacement from an underlying population, this process introduces a random element and normally 100 to 200 bootstraps are required to achieve stable results). With regards to external validation or the general applicability of the model to future patients or new datasets, method transportability (assessment of the model using a different method or approach) is used in the present study.

1.3 Statement of Problem

Because of the diagnostic challenges due to the overlapping morphological, immunophenotypic and cytogenetic features of BL and DLBCL and since therapy greatly differs between the two tumor types, it is important to investigate possible markers that may be used to better define these entities. But first, an outline of some the features currently established in clinical practice for these neoplasms is useful to cover at this time.

DLBCL is the most common form of NHL accounting for 30-40% of adult cases.²² It is a heterogeneous group of neoplasms that are characterized by the presence of large number of neoplastic B-cells of varied size and morphological variants, primarily consisting of immunoblastic, anaplastic and centroblastic morphology variants.²³ It exhibits a diverse clinical presentation, varied response to treatment and expresses distinct molecular profiles.²² By gene expression profiling, three molecular subgroups of DLBCL can be reliably identified that have both a distinct patho-biology and prognosis.

These are the germinal center B-cell like DLBCL (GCB) with a more favorable prognosis, the activated B-cell DLBCL (ABC) characterized by a poor clinical outcome and the primary mediastinal B-cell lymphoma distinguished by a favorable prognosis.²⁴ Apart from the molecular subgroups, a large number of DLBCL subtypes are now well established, the majority of these neoplasms fall under the classification of DLBCL not otherwise specified (DLBCL, NOS). In this group, there are no characteristic genetic aberrations; however, cases exhibit complex karyotypes and some of the recurrent aberrations that are observed include *BCL-6* translocations to the immunoglobulin heavy chain gene (*IGH*) and non-*IGH* genes with a frequency of 25-40%, *BCL-2* translocations to *IGH* observed in 15-20% and *MYC* (8q24) translocations involving *IGH* and non-*IGH* genes in up to 16% of the cases, a rare occurrence in DLBCL.²⁵⁻³⁰ The most frequent *MYC* translocations involve the juxtaposition of *MYC* to the *IGH* locus at 14q32 [t(8;14)(q24;q320)]. Other frequent *MYC* variants consist of the t(8;22)(q24;q11) and t(2;8)(p11;q24) translocations resulting in the juxtaposition of *MYC* to the immunoglobulin light chain lamda (*IGL*) and *IG* Kappa (*IGK*) genes respectively.³¹ In addition to these previously discussed genetic aberrations, other genetic lesions may also include the presence of *MYC* in combination with *BCL-2* or *BCL-6* gene rearrangements, so-called double hit lymphomas that can also occur in rare occasions. In terms of karyotypes, DLBCL may contain a *MYC* positive complex karyotype (6 or more chromosome aberrations in addition to a *MYC* rearrangement) or it may include a *MYC* simple karyotype.⁷ Morphologically, as described earlier, only large cells are common in this entity. The proliferation index, measured by the Ki67 score i.e., an antigen found in growing and dividing cells, is below 90%, but it can be measured above 90% in rare

occasions, while the bcl-2 protein expression is negative to strongly positive. By Immunophenotype, DLBCL express CD20, may express CD10 in 25-50% and CD5 appears in 10% of the cases.^{7,32} Briefly, a bcl-2 positive, ki-67 less than 90%, rare *MYC* rearrangements, *BCL-2* and *BCL-6* rearrangements favors a diagnosis of DLBCL.¹

In contrast, Burkitt lymphoma generally presents in extranodal sites e.g., abdominal masses, jaw, bone marrow or in the peripheral blood in the form of acute lymphoblastic leukemia. In this regard, it is unlike DLBCL that is localized to both nodal and extranodal sites. BL has three clinical variants: endemic form occurring in Africa, sporadic present throughout the world (1-2% in the US) and the immunocompromised BL seen generally in AIDS patients.^{7,33}

The gender ratio is also different between the two entities; in BL there is a male predominance compared to DLBCL that shows no gender preference. In contrast to DLBCL, the genetic hallmarks for BL, although non-specific, are *MYC* rearrangements, primarily t(8;14) and variant translocations t(8;22) and less often t(2;8) translocation. Although most of BL cases do exhibit *MYC* rearrangements, approximately 10% of cases lack a detectable *MYC* rearrangement by fluorescent in situ hybridization (FISH) analysis.^{2,5} In this case, other genetic techniques are available that may be used to detect cryptic *MYC* rearrangements. Presently, studies are ongoing to determine whether or not *MYC* negative BL cases really do exist. Unlike DLBCL, non-*IG/MYC* rearrangements, *BCL-2* with no *MYC* rearrangements, *BCL-6* with no *MYC* rearrangements and double hit lymphomas are absent in BL.⁷ In terms of karyotypes, *MYC* simple karyotypes are generally a characteristic finding in BL, but *MYC* complex karyotypes may appear in rare occasions. Morphologically, unlike a diffuse infiltrate of large B-cell seen in DLBCL,

BL is characterized by monomorphic small to medium sized B-cells with a “starry sky” appearance (see Figure 1-2), a high proliferation index greater than 95% measured by Ki-67 staining, positive for CD10, CD20, bcl-6 and absence of bcl-2 expression.⁷ Lastly, we should also point out; DLBCL accounts for 25,000 newly diagnosed cases in the US in contrast to only 2,000 of BL³⁴.

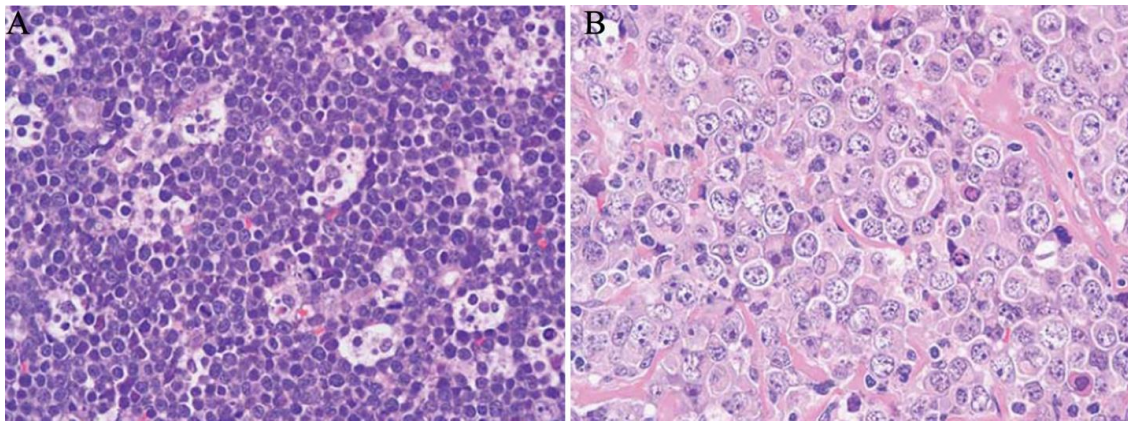


Figure 1-2 . Morphological characteristics of DLBCL and BL. Panel A shows a characteristic BL morphology with small to medium sized monophorphous cells, whereas panel B illustrates a large pleomorphic cell infiltrate more typical of DLBCL. Figure courtesy of Seegmiller and colleagues.³⁵

Previously, slight morphological deviation from strict guidelines for BL but with immunophenotype and molecular characteristics of BL were formerly termed “Burkitt-like lymphoma” or “atypical Burkitt”; however, molecular studies have showed these variants to have a molecular profile comparable to classical BL, and this term is no longer applicable.^{5,36} By gene expression profiles, BL shows a distinct molecular signature from DLBCL; however, cases were reported with an intermediate signature.^{5,6} With respect to treatment, these patients require aggressive chemotherapeutic regimens due to the high

tendency of central nervous system involvement. Whereas a low chemotherapeutic regimen consisting of cyclophosphamide, doxorubicin, vincristine, prednisone plus the anti B-cell monoclonal antibody CD20 rituximab (R-CHOP) is given to patients with DLBCL, a more aggressive treatment that requires prophylactic intrathecal chemotherapy for BL is needed.^{7,22} In short, *BCL-2* negative, a Ki-67 of greater than 95% and simple karyotypes (no more than two chromosome aberrations) involving *IG/MYC* rearrangements favors a diagnosis of BL.^{1,35} Table 1-1 summarizes distinguishing features between DLBCL and BL.

Table 1-1. Distinguishing Features of DLBCL and BL

Characteristic	BL	DLBCL
Gender	Prevalence in males	Gender prevalence
Infiltrate	Extranodal	Nodal and extranodal
Morphology	Small to medium cells	Large cells
	Monomorphic	Pleomorphic
Proliferation (Ki-67)	Greater than 90%	Common < 90%, rare <90%
Immunophenotype	CD10+,bcl-6+,bcl-2-2+/-	CD10+,bcl-6+,bcl-
<i>MYC</i> rearrangement	Yes (90% or more)	Infrequent (up to 16%)

<i>IG/MYC</i>	Yes	Infrequent
Non- <i>IG/MYC</i>	No	Infrequent
Simple karyotype/ <i>MYC</i> +	Yes	Infrequent
Complex karyotype/ <i>MYC</i> +	Infrequent	Possible when <i>MYC</i> present
<i>BCL-2/MYC</i> -	No	Occasional
<i>BCL-6/MYC</i> -	No	Occasional
Double hit	No	Infrequent

In this context, differential diagnosis between DLBCL and BL may appear non-problematic, that is, findings of *BCL-2* negative, Ki-67 greater than 95%, *IG-MYC* rearrangement with a simple cytogenetic karyotype and monomorphic B-cells with a “starry-sky” appearance favors a diagnosis of BL, whereas a *bcl-2* positive, Ki-67 less than 90%, *MYC* negative rearrangements, *BCL-6* and *BCL-2* rearrangements favors DLBCL. Nevertheless, overlapping clinical, morphological and genetic features between DLBCL and BL with or without *MYC* gene rearrangement still remain problematic to diagnose in daily clinical practice. Undoubtedly, new diagnostic markers are needed to better distinguish between these two entities. Therefore, a wide array of technologies ranging from conventional cytogenetics, classical CGH, application of bioinformatic tools, and the use of gene expression profiles may shed some new light into the genetic make-up and chromosome imbalances of DLBCL and BL, augmenting diagnostic accuracy and resulting in development of more precise therapeutic modalities.

1.4 Objectives and Goals of the Research

The overall goal of this project is to identify the most relevant cytogenetic markers associated with each tumor type and subsequently apply a set of bioinformatic algorithms (i.e., logistic regression, unsupervised hierarchical and k-means partition clustering, as well as artificial neural networks and support vector machine models) to determine the reliability of these markers and identify predictor models that may be used to augment diagnostic accuracy between the two histological tumor types. It should be noted here, that the principal requirement in this study is to determine discrimination (i.e., the ability to distinguish between two groups), this is usually determined by specificity and the area under the curve (AUC) using the ROC curve, and calibration and sensitivity are less required in this context. Moreover, to further validate these cytogenetic markers and to gain an insight into the molecular aspects of DLBCL and BL, copy number variations (in terms of chromosome gains and losses) using CGH data information and gene expression profiles will also be explored. To this end and to determine the most significant markers associated with each tumor type, we will collect cytogenetics and CGH datasets from various publicly available databases. Next, we will use the Pearson correlation to associate RCAs to a specific tumor type, determine specificities of all significant RCAs, correlate the different datasets, apply predictor models to analyze these RCAs and subsequently obtained a set of genetic markers that are highly predictive of tumor type. ROC curves will evaluate discrimination of predictor models. To the best of our knowledge, this is the first attempt to define DLBCL vs. BL from chromosome architecture down to the molecular level, and to use a cytogenetic approach to reliably

distinguish between the two diagnostic tumors types. This study will aim to answer the following questions:

- 1) What are the RCAs that predict DLCBL from BL?
- 2) What is the specificity of these RCAs?
- 3) With respect to the analysis of classical CGH, what are the specific gains and losses among the various diagnostic categories?
- 4) What are the correlations between RCAs and CNVs?
- 5) After a comprehensive review of array CGH of DLBCL subtypes and BL, what are the chromosomes imbalances distinguishing the various DLBCL subtype from BL?
- 6) What predictive models, logistic regression, supervised or unsupervised perform better?
- 7) After the application of the various bioinformatics algorithms, what are the top ranked markers predictive of DLBC?

1.5 Research Hypotheses

The following outlines the various hypotheses that will be tested to determine whether or not unique diagnostic cytogenetic markers may be used to distinguish between DLBCL vs. BL. Here, the null hypothesis is designated as H_0 and the alternative hypothesis is defined as H_1 . The first hypothesis tests whether there is a difference in the total number of chromosome aberrations (both numerical and structural) between the two different groups. Hypothesis 2, 3, 4, 6 and 7 tests whether there is a difference in the frequency of either RCAs or CNVs accordingly, while hypothesis number 5 tests

whether individual RCAs determine the dependent variable or outcome (DLBCL or BL) in a multivariable context. Hypothesis number 6 tests whether there is a difference in chromosome imbalances as measured by array CGH between DLBCL subtypes and BL. Of note, this hypothesis is based on the review of the literature and is an expanded analysis of a previously published report³⁷ by our group in 2012. All hypotheses are in line with the overall aim of this study.

- 1) Is the average total number of chromosome aberrations in DLBCL greater in number than BL?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 2) Is there a difference in the frequency of different RCAs between *MYC*⁺ DLBCL vs. BL?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 3) Is there are a difference in the frequency of different RCAs between *MYC*⁺ DLBCL vs. *MYC*⁻ DLBCL?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 4) Is there a difference in the frequency of different RCAs between DLBCL (*MYC*⁺ and *MYC*⁻) vs. BL?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 5) Is there a difference between DLBCL and BL with respect to RCAs within a multivariable context?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 6) Is there a difference in chromosome imbalances (measured by array CGH) in DLBCL subtypes compared to BL (only those chromosome aberrations found in array CGH with a frequency of 20% or more were considered of relevance)?

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

- 7) Is there a difference in the frequency of individual CNVs (e.g. gain of 18q, gain of 3) between the two different histological tumor types? (proposed for further research)

Hypothesis 1: $H_0 = H_1$

Alternative hypothesis: $H_0 > H_1$

Chapter II

II. Literature Review

2.1 Epidemiology

Three distinct variants of Burkitt lymphoma are recognized based on geographical location, age groups and risk factors. These consist of the endemic BL, sporadic BL and BL associated with immunodeficiency, primarily due to the human immunodeficiency virus (HIV).⁷ The endemic variant for the most part is confined to equatorial parts of the African continent. It is characterized by affecting facial bones in children with an incidence at 2 to 9 years of age, and has been correlated with the distribution of malaria in Africa.^{38,39} The sporadic variant occurs outside of Africa affecting for the most part the abdominal viscera and bone marrow. It is seen in children and adults with no age prevalence.³⁸ The incidence is relatively low in both the US and Europe at only 1-2% of all non-Hodgkin lymphomas.⁷ The third variant of BL associated with HIV (immunodeficiency associated) has no geographical or age prevalence. All three variants typically involve extranodal sites and are at risk to disseminate into the central nervous system.⁴⁰

In contrast, DLBCL is a heterogeneous group of lymphomas accounting for 30-40% of all adult non-Hodgkin lymphoma cases, with most cases falling under the DLBCL not otherwise specified.^{7,25} The median onset of disease is in the 7th decade, but it can also present in children and young adults. The incidence is more prevalent in males than that of females and disease presentation may involve nodal or extranodal sites with approximately 40% of the cases confined to extranodal sites.^{41,42}

2.2 Morphological and Phenotypic Features of BL and DLBCL

Morphologically, BL are small to medium sized cells with a diffuse infiltrate of repetitive tumor cells arranged in sheets, round nucleus, clumped chromatin and with various visible nucleoli. The tumor cells have a high mitotic index, on average greater than 95% measured by the immunohistochemical expression of MIB-1 (a monoclonal antibody that binds to Ki-67, an antigen associated with the proliferation of cells and used as a index to measure the proliferation of tumor cells)⁴³, and have a characteristic “starry-sky” appearance due to intermingle of phagocytic cells. In contrast, DLBCL are larger and more pleomorphic to those tumor cells encountered of BL. However, as mentioned in previous sections, there exists a continuum of morphological features that overlap between the two entities and as a result further ancillary studies are required.

By immunophenotype, BL is positive for CD79a, CD20 and PAX5.²⁵ However, in most of the cases, BL has a classical signature for CD10 expression, bcl-6, and negative for bcl-2 expression.⁴⁴ Besides these markers, BL also can express CD43, CD38, TCL1 but it is negative for TdT, CD138, CD44, CD23, CD5 and CD34.⁴⁵ The Epstein Barr virus (EBV) varies across BL variants.²⁵ In contrast, as a group DLBCL can express bcl-6, bcl-2, CD45, CD45RA, CD79a, PAX5, IRF4/MUM-1, CD22, CD19 and CD20.⁷ Expression of CD30 is associated with the anaplastic variant but can also be expressed variably in other DLBCL subtypes. Approximately 10% of DLBCL will express CD5^{7,25} and CD10+ characteristic of BL is also prevalent in a subset of DLBCL (30-60%). The EBV is also present in a subset of DLBCL and in plasmablastic lymphoma CD38 and CD138 are expressed.⁷ In the sections that follow, we describe

some of the major studies that have included protein expression (immunohistochemical staining, IHC), immunophenotypic investigations and the use of limited genetic components to aid in the differential diagnosis of DLBCL vs. BL.

In 2001, the Southwest Oncology Group⁴⁶ outlined phenotypic differences between Burkitt like lymphoma, (BLL, n=13) and DLBCL (n= 55) by IHC staining and limited use of FISH to assess *MYC* rearrangements. In this study, a broad antibody panel to distinguish between the two diagnostic entities was used. When BL was compared against DLBCL, the Ki-67 showed a proliferation index higher than DLBCL, a higher expression for CD10 and oncoprotein p53, a lower expression for bcl-2, and a constant lower expression for cell adhesion molecules. The CD44 was uniformly lower in BLL as well. For the presence of the EBV, the study reported 30% of the cases tested positive for BL, while 25 of the cases tested for DLBCL were negative. In terms of *MYC* rearrangements, four of 5 BL tested positive for the t(8;14) translocation or *MYC* rearrangement and none of the three cases analyzed for DLBCL showed a *MYC* rearrangement. Although this study is in agreement with the current literature, the scope of genetic testing performed by this group was extremely limited.

Unlike the previous study that reported no *MYC* rearrangements in DLBCL, Nakamura and colleagues⁴⁷ in 2002 compared both immunophenotypic and genetic characteristics between BL and *MYC*+ DLBCL (n = 10, n = 5 each). The study population included 5 pediatric cases with BL and remaining ten patients were adults. The IHC studies showed BL positive for CD20 (100%), CD10 (90%), bcl-2 (10%), and bcl-6 (90%), all BL cases were negative for the EBV virus. In contrast, *MYC*+ DLBCL were positive CD20 (100%), CD10 (40%), bcl-2 (75%), bcl-6 (100%), and none of this

cases tested positive for EBV. The Ki-67 proliferation index was significantly higher in BL than *MYC*+ DLBCL. The authors of this study concluded that a high proliferation index of nearly 100% and a monotonous proliferation of neoplastic cells should suspect BL, while CD10+ and bcl-2 negative should further help in distinguishing between BL and *MYC*+ DLBCL. Although the study reported chromosomal abnormalities for 6 of 10 BL and 3 of 5 DLBCL cases, there was no direct comparison between the two in terms of chromosome aberrations.

Following with the immunophenotypic features between the two diagnostic entities, Frost and colleagues in 2004⁴³ explored the expression of 6 proteins that included CD10, bcl-6, *MYC*, bcl-2, CD138 and MYB-1 in 55 classical BL and DLBCL pediatric cases (n=33, BL; n=20, DLBCL). To assess the usefulness of this IHC panel, the study tested four additional pediatric samples that could not be accurately classified as DLBCL or BL. This study showed a significant difference in staining pattern between the two entities, mainly a strong staining of *MYC*, absence of bcl-2 and a 100% protein expression of MYB-1 predicted BL. There was no difference in CD10, CD138 or bcl-6 protein expression. Based on these findings, the study classified the four intermediate cases as BL. Although these results are part in agreement with the literature (*MYC* staining is not part of the standard work-up for BL or DLBCL)¹, one limitation is the small numbers of cases analyzed and even much smaller were the cases used to assess the utility of *MYC*, absence of bcl-2 and MIB-1. In retrospect, it is unlikely that staining for *MYC* can serve as sensitive diagnostic marker because there was substantial overlap in *MYC* staining for both BL and DLBCL as the author rightly pointed out.

In line with this previous work, Gormley and colleagues⁴⁸ in 2005 used a panel of eight protein expression markers of germinal center (GC) and activated B-cell (ABC) to distinguish DLBCL vs. BL. The GC and ABC markers in this study included bcl-6, CD10, cyclin H and MUM-1, CD138, PAK1, CD44 and bcl-2 respectively. A total of 39 DLBCL and 16 BL cases were evaluated retrospectively. Based on morphology, cases were further subdivided into centroblastic and immunoblastic for DLBCL and classical and variant form for BL. To support a morphological distinction between a variant form of BL and DLBCL, the study relied on a high proliferation index or Ki-67 score and detection of a *MYC* rearrangement by FISH analysis. However, a drawback to establishing this distinction is that *MYC* rearrangements with a high Ki-67 score (more than 95%) can also be present in DLBCL⁴⁹. Therefore, it is uncertain whether these characteristics may be completely reliable in distinguishing between a variant form of BL and DLBCL.⁴⁸ Despite these shortcomings and using a hierarchical clustering, the study reported two major clusters based on protein expression patterns that significantly correlated with a morphological diagnosis. For example, the germinal center markers CD-10 and bcl-6, as expected, stained stronger in BL than the DLBCL cluster; however; cyclin H a GC marker surprisingly did not stain BL but was more prevalent in DLBCL. The bcl-2 marker was not uniquely expressed on both of the clusters. Both the CD44 and MUM-1 were absent from BL and positive in the DLBCL cluster, while PAK-1 and CD138 showed no distinction between the two diagnostic groups. Based on hierarchical cluster analysis, the sensitivity was only reported at 81% and specificity at 87% for a diagnosis of BL. Even though these findings were able to show some distinction between the two diagnostic entities based on a multi protein expression panel (GC and ABC

markers), sensitivity and specificity results, in our view, remain low. What is more and as previously stated, the criteria used for the separation between a variant form of BL and DLBCL are limited based on the Ki-65 score and FISH analysis of *MYC*.

In a similar fashion, Haralambieva and associates² analyzed protein expression markers in 74 adult DLBCL and 10 pediatric BL cases (Ki-67, CD-10, bcl-2, bcl-6), this study also incorporated genetic markers by FISH analysis, mainly to detect *BCL-6*, *BCL-2* and *MYC* rearrangements. The study proceeded with two algorithms (algorithm A, a review process by four hematopathologist that assessed morphology, IHC, age, site of involvement and algorithm B that included a Ki-67 >90%, bcl6+, CD10+, bcl-2 negative, *MYC*+, *BCL-2* negative and *BCL-6* negative for a diagnosis of BL) to make the distinction between the two diagnostic entities. All 10 reference pediatric BL cases were accurately diagnosed by the two algorithms. Of the 74 adult cases, 21 and 52 cases were diagnosed as BL and DLBCL by algorithm A respectively, while 23 and 51 were categorized as BL and DLBCL each with algorithm B. However, we should point out that an independent consensus by all four pathologists only reached 24% in BL and 69% in DLBCL. In addition, only 55% of all cases independently reviewed and finalized were uniformly classified by morphological and cytochemical features. A higher consensus was only reached when all four hematopathologist discussed the cases concurrently. Interestingly, one observation made by the study was a more homogeneous clinical presentation of BL by algorithm B (site of involvement and gender were more uniform in BL). Indeed, it is now well established that BL most often presents with extranodal site involvement.⁷ Moreover, the study concluded that 29% of BL cases would have been incorrectly diagnosed using just morphological and cytochemical measures based on the

criteria of algorithm B. Therefore, subsequent investigations relied on a multi-parameter approach that would include molecular characteristics to accurately diagnose BL from DLBCL.

Following this latter findings and to improve the selection criteria for establishing an accurate diagnosis, Cogliatti and colleagues⁴⁹ in 2006 reported a differential diagnostic algorithm based on morphology, immunophenotype and genetic features. For a diagnosis of BL, the group used a classical morphology, CD10+, bcl-6+, bcl-2-, Ki-67 >95%, *MYC* rearrangement and absence of a t(14;18) translocation. Using these criteria, the group divided a total number of 39 cases into four distinct groups: 1) DCI (n=5), classical BL, all criteria were present in this group, 2) DCII (n=11), atypical BL with deviation of two criteria, 3) DCIII (n=9), *MYC*+ DLBCL with two or more deviations of the criteria selected and 4) DCIV (n=14), highly proliferative B-cell lymphoma with morphology intermediate between atypical BL and DLBCL. A Ki-67 of > 95% was more prevalent in both classical and atypical BL. In terms of immunophenotypic findings, the study concluded the cell phenotype to be sensitive but not specific for a diagnosis of BL. Of the four groups listed by the study, classical BL had a homogeneous morphological, immunophenotypic and genetic characteristics. In light of these newly proposed diagnostic algorithm and based on standard therapeutic protocols at the time, the study concluded a total of 53.3% of patients were considered undertreated and 38.1% of these patients were deemed over treated. Subsequent stratification of tumors by this classifier exhibiting a typical BL morphology confirmed five of six cases with a classical BL that is, CD10+, CD20+, BCL-2 negative and bcl-6+. Eleven of the 39 cases that showed variation of immunophenotype, were pleomorphic, harbored a *MYC* rearrangement and

with evidence of a t(14;18) translocation were classified as atypical BL. Nine additional cases were categorized as DLBCL. These cases showed a morphology consistent with DLBCL, variable immunophenotype, *MYC* rearrangement and at times with a t(14;18) translocation. The remaining 14 cases were classified as B-cell lymphomas with intermediate features between DLBCL and BL. Despite this effort of including morphological, immunophenotypic and genetic components in a new classifier, there still remained cases that were not adequately classified and were placed into this mix-bag of DLBCL and BL. In brief and as previously mentioned, immunophenotype is sensitive to BL but not specific and broader genetic classifiers than just *MYC* and t(14;18) are clearly needed to further stratify these cases.

Unlike the two previous studies that reported a need to use genetic markers to augment distinction between BL and DLBCL, Chuang et al. in 2007⁵⁰, reported a comparative study where histopathology and IHC, primarily CD10, bcl-2, bcl-6 and Ki-67 could confidently distinguish between the two groups. This study compared the results from 28 BL and 16 DLBCL with a high proliferation index with or without the characteristic starry sky appearance (DLBCL-HPSS) that were diagnosed by histopathology, IHC, FISH results for the *MYC* rearrangement and by presence of the Epstein Barr virus m-RNA (EBER) in neoplastic cells. The study reported a significant difference in CD10, bcl-2, Ki-67 and combined expression of CD10/bcl-2 negative/bcl-6 + in BL vs. DLBCL-HPSS ($p < .001$). Alternatively, the MUM-1 expression was more prevalent in DLBCL-HPSS compared to BL ($p < .001$). With respect to EBER and *MYC* rearrangements, 7 of 28 (25%) in BL vs. 0 of 16 (0%) in DLBCL were positive for EBER ($p = .037$) and 26 of 27 (96%) in BL vs. 1 of 15 (7%) showed a *MYC* rearrangement by

FISH analysis ($p < .001$). Given these results, the study concluded that CD10, bcl-2, bcl-6, Ki-67 greater than 95% and with an added benefit from a positive result from EBER, a confident distinction between BL and DLBCL could be made in most cases. While these results are persuasive, the study population is small and a larger cohort is needed to substantiate findings. Also and as previously mentioned, a measurement of Ki-67 $> 95\%$ may not be entirely specific for BL since *MYC*⁺ DLBCL may have similar expression measurements.⁴⁹ And, even though immunophenotypic markers listed here may distinguish between the two entities in general, there still a number of cases that may pose a challenge, particularly *MYC*⁺ DLBCL with a high Ki-67 as just mentioned or a subset of DLBCL that are characterized by the expression of CD10+.

To properly distinguish between BL and a subset of CD10+ DLBCL (n=13, n = 10, respectively), a study in 2010 using flow cytometric immuno-phenotyping (FCI) with novel antibodies CD18, CD44, CD54, CD20 and CD43 showed statistically significant differences between these two entities with respect to CD44 and CD54 expression⁵¹ and concluded that CD44 is an excellent marker to properly differentiate DLBCL from BL. Likewise, McGowan and colleagues⁵² using a multi-parameter scoring system of frequently used antibodies in FIC reported significant differences in CD10+ DLBCL and BL. This study assessed the mean florescent intensities (MFI) for CD10, CD20, CD38, CD79B, CD71 and CD43 in the two tumor types. The study revealed MFI for CD10, CD43, CD71, percentage of neoplastic cells with CD71 and forward side scatter (FSC) significantly associated with BL, whereas CD79b was statistically significant in DLBCL. Limitations of this study are the small study group. Therefore, further investigations with a larger cohort are required to confirm these findings.

Moving from a scoring system based on MIF of routinely used antibodies, Naresh and colleagues⁵³ in 2011 explored an IHC and FISH scoring system that was used in three phases for the differential diagnosis of BL and DLBCL. For Phase 1, it consisted of scoring cases based on morphology and immuno-stains (0-3 points), CD10 (1 if positive) and bcl-2 (2 points if negative and 1 point for positive). A cumulative score of 5-6 suggested BL, 3-4 points did not exclude BL and less than 3 points excluded BL. For the second phase, Ki-67 (2 points > 95%, 1-90-95%), CD38 (1 if positive) and CD44 (1 if negative) were added for a total of 6 point variables. Cumulative scores for this second phase were as followed: greater than 8 points indicated a diagnosis of BL, 6-7 points BL was not excluded and less than 6 points excluded BL. For the final phase, a FISH analysis for a *MYC-IGH* rearrangement, *BCL-2* and *BCL-6* rearrangements were added to the scoring system. Final cumulative scores of greater than 8 points suggested a diagnosis of BL, while 6-7 points did not exclude BL. This system was assessed on 252 patients and a final diagnosis of BL, not BL or BL not excluded (BLNE) were assigned. In total, 4.8% of cases were not resolved by the algorithm. Of the cases that were diagnosed as not-BL (n=116), a total of 103 cases were categorized as DLBCL and six cases were diagnosed as U-DLBCL/BL. Of special interest, of the cases diagnosed as BL, 5% showed bcl-2 expression, 2.5% were negative for CD10 expression, and in 9.3% of the cases the Ki-67 was <95%. Additionally, in 47 of the cases a CD38-/CD44+ immunophenotype was restricted to the not-BL group and, five (10%) among 48 cases diagnosed as BL, a *MYC/IGH* rearrangement was not detected. Whether this is due to technical issues, *MYC* activation through micro-RNA expression or whether true *MYC*-BL really exists still remains an unresolved matter. In brief, additional cases to validate

this algorithm are needed. Moreover, although the proposed algorithm successfully classifies the great majority of cases, in our view additional work is clearly needed to appropriately classify all unresolved cases. In this context, we believe additional genetic markers may be needed to fully discriminate between DLBCL and BL. Also, although limited cases were found with a deviation in BL phenotype, that is, bcl-2+, CD10- and Ki-67 < 95%, more work is needed to fully characterize this subgroup.

In the same year that Naresh et al.⁵³ proposed a three-phase algorithm, Lu and colleagues⁵⁴ studied 97 pediatric cases with a diagnosis of BL (n=81), DLBCL (n=8) and BL/DLBCL(n=8) using immunohistochemistry, Epstein-Barr virus (EBER) in situ hybridization and FISH. This study reported no difference in CD10 and bcl-6 among all three categories, expression of bcl-2 and MUM-1 was significantly higher in DLBCL and BL/DLBCL groups compared to BL, Ki-67 > 90% was indicative of BL, EBER was significantly higher in BL and extra copies of *BCL-6* was more prevalent in both DLBCL and BL/DLBCL. Although most of these findings are in agreement with the previous literature, extra copies of *BCL-6* and MUM-1 need further validation given the fact that only 8 cases of DLBCL were compared against BL. Here, we should point out that that expression of MUM-1 does not exclude a BL diagnosis⁵⁴. Of interest, this study concluded that intermediate cases (BL/DLBCL) resemble DLBCL and may belong to a subset of DLBCL, particularly given the fact that there were no differences in immunophenotypic features between these two entities.

Taken together and based on our initial review of morphological, immunophenotypic and limited use of genetic markers, we conclude that both protein expression patterns and immuno-phenotypic characteristics are sensitive but not specific

to BL, and therefore subsequent ancillary studies are needed. Table 2-1 summarizes this initial review.

Table 2-1. Immunophenotypic and Genetic Characteristics Evaluated in BL and DLBCL

Reference	Marker	BL (%)	DLBCL(%)	p-Value
Nakamura ⁴⁷	CD10	90	40	NA
	CD20	100	100	NA
	bcl-2	10	75	NA
	bcl-6	100	100	NA
	EBV	0	0	NA
	Ki-67	98.1	66.3	.0001
SOG ⁴⁶	Ki-67	88	53	NA
	CD10	85	27	NA
	p53	54	16	NA
	bcl-2	15	53	NA
	Lack of any CAM	92	27	NA
	CD44	8	86	NA
	<i>MYC</i> R	80	0	NA
	EVB	30	0	NA

Frost 2004 ⁴³	bcl-2 (no stain)	88	37	.005
	Ki-67(100%)	82	0	.0001
	myc stain (4+)	76	25	.0001
	CD10 (4+)	91	100	
NS				
	bcl-6	70	75	NS
Gormley 2005 ⁴⁸	bcl-2	rare cases	some cases	NA
	Ki-67	99	56	.0001
	CD44	Primarily -	Strong +	NA
	MUM-1	Primarily -	Strong +	NA
	CD10	+++++	++	NS
	bcl-6	+++++	++	NS
	Cyclin H	Absent	+++++	NA
Cogliati 2006 ⁴⁹	CD10	80	67	NA
	bcl-2(no stain)	80	56	NA
	bcl-6+	100	78	NA
	Ki-67 >95%	100	89	NA
	<i>MYC</i> +(genotype)	100	100	NA
	t(14;18)-	100	89	NA

Haralambieva 2005 ²	CD10+	23/23	33/51	NA
	Ki-67	95-99	50-99	NA
	bcl-6+	23/23	40/46	NA
	bcl-2+	0	37/51	NA
	CD10+,bcl-6+, bcl-2 neg	23/23	0	NA
	<i>MYC</i> + (genotype)	23/23	7/51	NA
	<i>BCL</i> -2+ (genotype)	0	9/51	NA
	<i>BCL</i> -6+ (genotype)	0	11/50	NA
Chuang 2007 ⁵⁰	CD10+	100	6	<0.01
	bcl-2	11	69	<0.01
	bcl-6	96	88	.543
	CD10+/bcl-2 neg	86	6	<.001
	bcl-6+			
	MUM-1	18	94	<0.01
	KI-67>95	96	13	<0.01
Schniederjan 2010 ⁵¹	CD44+	2/13	9/10	.001
	CD54+	100	220	.01

McGowen 2012 ⁵²	MFI of CD10+ CD43,CD71,CD79b, FSC (BL > DLBCL)	NA	NA	<.05
Naresh 2011 ⁵³	CD38-/CD44+	Negative	Positive	NA
	bcl-2	5	72	NA
	CD10-	2.5	67	NA
	Ki-67 < 95%	9.3	94	NA
	<i>MYC</i> R	90	5	NA
Lu 2011 ⁵⁴	CD10	91	75	0.396
	bcl-6	86	63	0.209
	bcl-2	3	50	.001
	MUM-1	17	63	.012
	Ki-67 (mean)	93	83	.006
	<i>MYC</i> + (genotype)	98	37.5	0
	<i>BCL-6</i> extra copies	0	37.5	0

Key: SOW, South Oncology Group; CAM, cell adhesion molecules; R, rearrangement; EBV, Epstein-Barr virus.

2.3 Cytogenetic Aberrations in DLBCL and BL

Although there are a number of direct comparisons between DLBCL and BL in terms of immunophenotypic features, the cytogenetic profile for the two entities is much more limited and a direct comparison between the two is narrow in scope. In the next paragraphs we explore existing reports that outline some of the cytogenetic findings for both DLBCL and BL.

In 2004, a study by Au and colleagues⁵⁵ evaluated clinico-pathological features and cytogenetics for Non-Hodgkin lymphoma (NHL) with 8q24/*MYC* aberrations. Of the total number of NHL cases reviewed by this study, a total of 45 were BL and 15 were diagnosed as DLBCL. The study reported a fewer number of secondary aberrations, that is, simple karyotypes in BL compared to the rest of the cases ($p < 0.001$). A similar finding was also reported by our group in 2010.³⁵ Therefore, a lack of genetic complexity may be suggestive of BL.⁵⁶ This is further supported by an array CGH study that reported more chromosome imbalances in non-molecular BL, as well as a higher frequency of non-immunoglobulin heavy chain (non-*IG*) –*MYC* rearrangements and *IG-MYC* rearrangements in the context of complex karyotypes.⁵ However, we now know this is not entirely true since a subset of BL may carry an *IG-MYC* rearrangement in combination with complex karyotypes (i.e., with higher genetic complexity).⁵ Despite a detailed cytogenetic outline of cases provided by this study, the authors failed to make a direct comparison of secondary aberrations with 8q24/ *MYC* rearrangements between the two entities of interest (this is especially true when BL appears with a higher genetic complexity). In our view, a unique pattern of secondary aberrations can be used to further distinguish between BL and DLBCL. In fact, following our initial work in 2010,

a preliminary analysis of RCAs in 26 well defined BL and DLBCL cases (n = 23, n = 3 each), showed loss of 1p32-p36, 2p11-p25, -4, 6q21-q27, -10, 13q loss, 15q loss, and 16p13 rearrangements significantly associated with DLBCL (unpublished data).

Following these initial studies, Boerma and colleagues⁵⁷ used the Mitelman database of chromosome aberrations to outline the cytogenetic profile of BL and non-BL lymphomas. This study included a total of 538 BL and 327 non-BL cases. An additional 108 cases of BL were included without an apparent *MYC* translocation. Of the total number of BL cases identified, only 481 cases met the criteria of a core group of BL (*IG-MYC* rearrangement with no *BCL-2*, *BCL-6* or *CCND1* rearrangements). Consistent with previous reported observations, the core BL group revealed a low genetic complexity score. In terms of age groups in the BL group, adults carry a higher frequency of variant *MYC* rearrangements and a higher incidence of chromosome 8q gains. However, the genetic complexity for both adults and pediatrics was similar. This is corroborated by earlier studies that suggest that BL in both adult and pediatric patients is the same tumor neoplasm.^{5,6} In the core BL group, the most recurrent (>4%) aberrations were gains of chromosome 1q, 7, 12, and losses to 6q, 13q32-34 and 17p. When comparing differences between the core BL and non-BL groups, the study reported a higher genetic complexity in non-BL cases ($p < 0.0001$), while genetic gains of chromosome 7, 11, 12, 18 and X and losses of 4q, 6q13-27, 9, 10p, 15, 17p and 17q were more prevalent in non-BL. Importantly, these observations remained the same when the non-BL group was limited to a morphological diagnosis of DLBCL. Among the other reported observations included: BL without *MYC* rearrangement revealed genetic differences from the core BL group (an ongoing discussion persists on whether or not a *MYC*- BL really exist), a non-

IG-MYC rearrangement is uncommon and differs significantly from the core BL group, and double hit lymphomas (*MYC* rearrangement with concurrent *BCL-2*, *BCL-6* or *CCND1* rearrangements) must not be classified as BL. Even though this latter study was able to define a core BL group and show distinct cytogenetics differences between the core group and DLBCL, additional work is needed to confirm these findings. Also, sensitivity and specificity analysis were lacking. Moreover, it is of interest to see whether a cytogenetic algorithm can be implemented based on recurrent chromosome aberrations to see how well it can classify DLBCL from BL. Table 2-2 outlines the cytogenetic profile of DLBCL and BL.

Table 2-2. Cytogenetic profile of DLBCL and BL

Reference	BL	non-BL/DLBCL
Seegmiller 2010 ³⁵	Simple karyotype (<2 aberrations)	Complex karyotype
Hummel 2006 ⁵	<i>IG-MYC</i> rearrangement	non- <i>IG/MYC</i> rearrangement
	<i>IG-MYC</i> + simple karyotype	<i>IG-MYC</i> + complex karyotype
Boerma 2009 ⁵⁷	1q G, +7,+12, 6q L, 13q L, 17p L	+7, +11, +12, +18, +X, 4q L, 6q13-27L, -9, -10, -15, 17 p L, 13qL, 17q L
Garcia 2011		1p32-p36 L, 2p L, -4, 6qL
(Unpublish data)		-10, 13q L, -15, 16p13 R

Key: G, gain, L, loss, R, rearrangement.

2.4 Chromosome Imbalances by Array Comparative Genomic Hybridization in DLBCL subtypes and BL

To address one of our research questions (hypothesis number 6), that is, whether distinct DLBCL subtypes carry unique chromosome imbalances compared to BL, our group initially detailed copy number aberrations (CNAs) by array CGH that are distinctive of well-defined WHO DLBCL subtypes.³⁷ Following this initial review, this section will expand on this work and explore CNAs in BL. Of note, only those aberrations by array CGH that were present in 20% or more in the various DLBCL subtypes and BL were included in this review. In addition, an analysis of classical CGH is here proposed for further research.

Following our initial publication of array CGH in distinct DLBCL subtypes, here we expand this published review to include BL. Garcia and colleagues⁵⁸ in 2003 analyzed 46 BL patients using classical CGH. In this study, the most frequent aberrations included gains of 12q (26%), 22q (20%) and Xq (22%). In disagreement with these previous findings, Barth and associates⁵⁹ reported five of seven (29%) BL cases with loss of chromosome 12q. However, gains of 1q (43%) and loss of 17p (29%) were aligned with the results reported by Garcia et al. Later in 2008, Salaverria and colleagues⁸ analyzed a total of 51 patients with a molecular signature of BL (26 with classical BL features and eight cases with histomorphological features of DLBCL referred to discrepant BL). Aberrations detected for both these groups included gains of 1q, 7q,

8q24, 13q11-q13, 13q31-q32, 13q33, but only 1q gain was reported with an incidence greater than 20% in the classical BL group. With respect to the discrepant BL group, loss of 13q14 was only present in this group and had a significant difference in terms of the aberrations just listed above compared to the classical BL group ($p < 0.05$). Of interest, some of the cases within this category expressed both *BCL-2* mRNA as well as bcl-2 protein and although these cases cluster within BL, a lower level of BL germinal center B-cell expression genes was evident in a previous study.⁶ In addition to outlining some of the chromosome imbalances in BL, this study also compared chromosome imbalances between BL ($n = 51$) and 161 previously reported cases of both ABC and GCB molecular subtypes of DLBCL. The analysis showed losses of chromosome 11q24-q25 more prevalent in BL compared to DLBCL ($p = 0.03$), whereas gains of chromosome 3, 18 and losses of 6q16-q27 were found more frequently in the ABC molecular subtype of DLBCL ($p < 0.05$). No differences were noted between BL and the GCB subgroup of DLBCL. Unlike the previous reports just listed above, Toujani and colleagues⁶⁰ in 2009 used an array CGH platform to analyze 13 BL tumors and 15 BL cell lines. This study reported gains at chromosome 1q (44%), 13q (26%) and 7q (22%). Likewise, a year later, Scholtysik and colleagues⁶¹ reported the highest recurrent genomic gains (> 7 cases) occurring at 1q31, 3q27, 6q15, 11q24.3, and 13q31.3 in 39 cases of BL analyzed. The most frequent genomic losses (> 6 cases) were 3q13, 17p13, 19q13 and Xp22. Figure 2-1 accounts a summary of copy number alterations in both DLBCL subtypes (based on our earlier review of DLBCL subtypes) and BL. The following are abbreviations used for the different DLBCL subtypes: PMBL, Primary mediastinal large B-cell lymphoma; Bone, Primary large B-cell lymphoma of bone; CNS, DLBCL of the

central nervous system; Leg type, Primary cutaneous large B-cell lymphoma leg type; T/HR, T-cell/histiocyte-rich B-cell lymphoma (no aCGH data available); CD5+ , De novo CD5 large B-cell lymphoma; PAL, Pyothorax-associated lymphoma; PL, Plasmablastic lymphoma; PEL, primary effusion lymphoma.

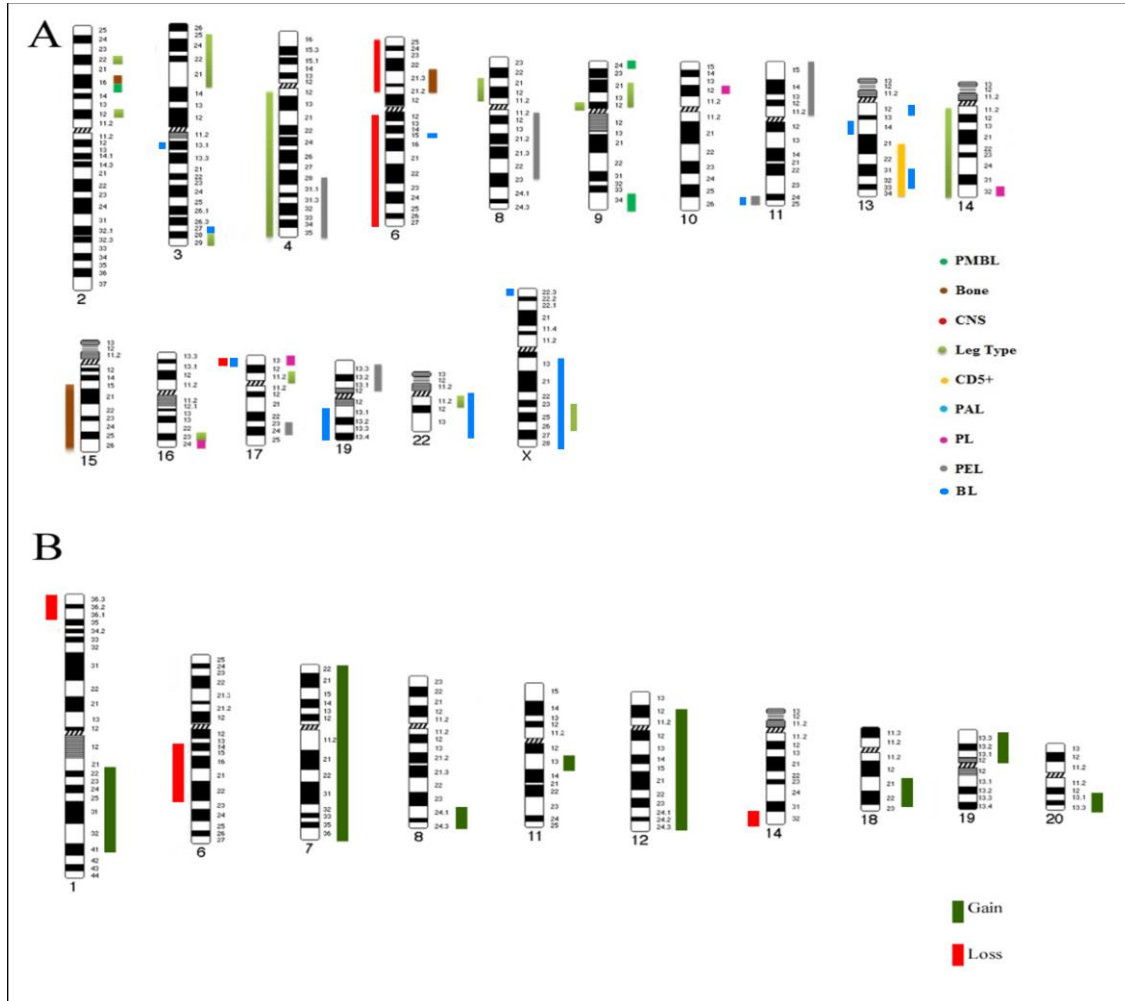


Figure 2-1. Chromosome ideogram of gains and losses by array CGH in DLBCL and BL. Panel A shows an ideogram spectrum for unique gains and losses for DLBCL subtypes and BL. Chromosome band region gains are provided on the right hand side of each chromosome, while genetic losses are shown on the left hand side of each chromosome

ideogram. Each category is represented by a different color coded bar as illustrated in the schematic. Panel B illustrates common CNA for all DLBCL subtypes. Commonalities were highlighted if three or more subtypes had a CNA at that particular band region.

2.5 Gene Expression Profiles, Micro RNA and Single Nucleotide Polymorphisms in DLBCL and BL

The use of gene expression profiles (GEP) has facilitated to quantitate the expression of a large number of genes in a single experiment. With the development of the “lymphochip”⁶², a complementary DNA microarray for lymphoid disorders, it has been possible to initially characterize various B-cell malignancies including DLBCL at the molecular level. Using hierarchical clustering of differentially expressed genes in DLBCL, Alizadeh and associates⁶³ initially identified three distinct clusters. Following this initial classification of DLBCL, Rosenwald and colleagues⁶² in 2002 analyzed an expanded number of cases (n = 274) and using the same hierarchical clustering method validated the existing three distinct expression patterns. The first two cluster groups were designated as the germinal center B-cell like DLBCL (GCB-DLBCL) and the activated B-cell like DLBCL (ABC-DLBCL). The GCB-DLBCL expressed genes that characterize the germinal center B-cell molecular signature, while the ABC-DLBCL had the molecular signature of activated B-cells. A third group, termed primary mediastinal DLBCL could not be classified into the GCB or ABC molecular subtypes. Both GCB and ABC molecular subgroups are not identified by morphological characteristics and are variable in terms of immunophenotype by immunohistochemistry.⁶⁴ The GCB subtype expressed numerous markers of GC differentiation that included CD10, CD38, and *BCL-*

6. Furthermore, two characteristic genetic events in DLBCL, the t(14;18) translocation and c-rel gene amplification were observed exclusively in the GCB subtype.⁶² Alternatively, the ABC molecular subtype highly expressed genes similar to those that are normally seen in activated B-cells including *IRF4*, responsible for the proliferation of B-cells in response to antigen activation.⁶⁵ Therefore, overexpression of *IRF4* may in turn contribute to the abnormal proliferation of neoplastic cells. In addition, other unique genetic characteristics limited to this group include the highly expressed *BCL-2*, an anti-apoptotic gene,⁶³ and the constitutive activation of the nuclear factor kB (NF-kB) pathway.⁶⁶ In terms of clinical outcome, the ABC subtype has a lower overall survival and event free survival than the GCB subtype.

In 2006, two major GEP studies carried out by Dave et al. and the other by Hummel and associates,^{5,6} investigated the gene expression of BL in an attempt to distinguish it from DLBCL at the molecular level. In the first study by Dave and colleagues⁶, two sets of cases were analyzed. For the first set, a total of 71 cases originally diagnosed as classic BL (n = 25), atypical BL (n = 20), DLBCL (n = 20) and high grade lymphoma, not otherwise specified (n = 6) based on morphologic, phenotypic characteristics and cytogenetics were included for GEP analysis. These specimens were later reclassified by GEP. The second set of cases consisted of 223 previously investigated specimens with a diagnosis of DLBCL. Nine additional cases were high grade DLBCL with a Ki-67 proliferation index of almost 100%. These specimens were subsequently divided into the three major molecular subtypes (GCB, ABC, and primary mediastinal DLBCL). Only one of the cases from this group was reclassified as a BL. This study used a custom DNA microarray of 2524 distinctive genes that are

differentially expressed in non-Hodgkin lymphomas. The Affymetrix U133 Plus 2.0 array was also used to profile a subset of specimens. Initially, the group developed a gene expression classifier that separated BL from DLBCL. Cases were separated into two groups based on the classifier, one with a detectable molecular signature of *MYC* target genes, characteristic of BL, and a second group lacking this signature. Thereafter, the molecular signature of *MYC* target genes was directly compared with the GCB, ABC and primary mediastinal DLBCL. During this process and only after a 4 pairwise comparison that supported BL, was a diagnosis of molecular BL assigned. In terms of biological mechanisms between the two entities, BL had a high expression of *MYC* target genes as anticipated, these are a subset of genes that are normally expressed in germinal center B-cells, whereas the major histocompatibility complex (MHC) I set of genes and the NF- κ B target genes were more prevalent in DLBCL. These latter set of genes included genes such as *STAT3*, *BCL-3*, *NFKBIA*, *CD44*, *FLIP*, and *BCL2A1*. Moreover, both BL and DLBCL with *MYC* rearrangement/ t(8;14) translocation were unmistakably distinguishable from each other with respect to expression pattern. Of the total number of specimens analyzed and based on gene expression profile, eight cases that were pathologically characterized (pathology was based on morphological findings, immunophenotype and cytogenetics) as DLBCL were later reclassified as BL. These cases were referred as discrepant BL. Interestingly, four of these cases expressed bcl-2. Thus, based on these findings, a small number of cases cannot be properly diagnosed with routine clinical diagnostic tools.

In the same year, Hummel and associates⁵ used a molecular profile approach to define BL. In this study, 220 specimens were included for analysis (BL, atypical BL and

DLBCL). The study initially designed a BL core group expression profile that subsequently identified an expanded group of lymphoma cases. The BL core group consisted of the following characteristics: a Ki-67 score greater than 95%, CD5-, bcl-2 negative, CD10+, bcl-6+ and CD20+. By applying the core group expression profile to a training set of 105 lymphomas, the group identified 58 genes that represented a molecular signature of BL (mBL). Each case was then given a mBL score (0-1) based on gene expression similarity between the BL core group and the various tumor specimens. A score greater than .95 was considered a diagnosis of mBL, while a score less than .05 were characterized as a non-mBL. The rest of the cases were assigned intermediate. This process classified 22% (n = 36 + 8 cases that comprised the initial core group) of the specimens with an expression signature of mBL, 20% as intermediate, and 58% of cases as non-mBL. Of the 36 specimens with a mBL signature, 21 were classified as atypical BL due to variant morphology or immunophenotype. Importantly, 11 of 36 mBL showed typical morphological features of DLBCL. In terms of immunophenotype, all mBL express CD10 and bcl6. Similar to the previous study, a few number of mBL cases (n = 7) expressed bcl-2. The study was also able to categorize two distinct cytogenetic groups based on expression. A *MYC-IGH* rearrangement with less than 6 chromosome aberrations (*MYC* simple) frequently related to the mBL group, and *MYC* complex tumor specimens more in line with an intermediate signature.

Both of these studies help to improve the distinction of BL and DLBCL, for example, both studies agree that a number of cases with a unique morphological feature of DLBCL are nonetheless consistent with an expression profile that is closest to BL. Thus, in a sense these studies have expanded the continuum of BL to tumor specimens that are

otherwise categorize as DLBCL by conventional morphological features. Also, we should highlight that a small number of BL specimens in both studies showed expression of bcl-2. And lastly, we should point out that there was a 100% agreement between classical cases of BL and the molecular signature of these cases. Following these two major GEP studies, Deffenbacher and colleagues⁶⁷ in 2012 expanded the use of genomic profiling to include both GEP and micro RNAs (miRNA, small non coding RNA that binds to a target mRNA inhibiting the translation of a given protein) profiles to distinguish between pediatric and adult non-Hodgkin lymphomas, mainly BL and DLBCL. In this study, GEP and miRNA clearly distinguish between pediatric BL (n = 57) and DLBCL (n= 13). In the pediatric DLBCL category, the GCB molecular subtype was more prevalent than the ABC subtype (6:1). Moreover, when pediatric and adult cases for both BL and DLBCL were compared in a supervised hierarchical clustering model, two distinct clusters were apparent with minor differences between the two, suggestive of similarities between pediatric and adult tumor cells at the molecular level. However, array CGH observations identified some unique copy number aberrations in pediatric DLBCL cases (-4p, -19q and +16p), indicative of distinct underlying biology. With respect to miRNAs, there were 35 differentially expressed miRNAs between BL and DLBCL. As expected, tumors with a diagnosis of BL had a miRNA profile regulated by *MYC*, including miR-17-92. A second clustered was identified on chromosome X, mainly composed of miR18b, miR20b and miR106a. In DLBCL, a subset of cases expressed similar patterns of BL, suggesting *MYC* deregulation in these cases. In terms of age group, expression patterns for pediatric and adult DLBCL revealed no significant differences and only miR-9 was elevated in the adult group compared to the pediatric BL

group. In short, both GEP and miRNA unambiguously distinguished pediatric BL from DLBCL, but in terms of age groups, it appears that both pediatric and adult DLBCL may differ in their underlying oncogenic mechanisms.

Lastly, we should mention a recent publication by Walther and associates⁶⁸ in 2013 that reported the aberrant expression of lymphocyte enhancer factor 1 (LEF1) in sporadic BL.

Similar to the just described study, Lenze and associates⁶⁹ used miRNA profiles to distinguish subtypes of BL, mainly sporadic, endemic and human immunodeficiency virus related-BL (HIV-BL) from DLBCL. The group of this study generated a profile of miRNAs for endemic (n = 18), sporadic (n = 31), HIV-BL (n = 15) and DLBCL (n = 86) to see whether miRNA expression could be used to distinguish across the various entities. A total of 38 miRNAs were differentially expressed between BL and DLBCL, primarily downregulation of MiR-155, suppression of *MYC* associated miRNAs (has-miRs-23a, -146b, -30d, -29b, -26a and -23b) and reduce levels of NF-kB (hsa-miRs-221, -222, -146a, and 146b) associated miRNAs in BL. In this context, we should point out that miRNA-146a, -146b and -155 are activated by the NF-kB signaling pathway, and therefore such findings highlights the importance of this pathogenetic pathway in distinguishing DLBCL from BL. This is further supported by previously described GEP studies. For the BL subtypes, only 6 miRNAs were differentially expressed between the sporadic and endemic form of BL, while no miRNA differences between the HIV-BL and the endemic variant were detected. Using an unsupervised cluster analysis of miRNAs, the study showed a clear distinction between BL and DLBCL with no apparent subdivisions of BL, indicating that the three BL variants belong to a single entity. Although this study further

supports NF- κ B as a distinctive pathway and clearly distinguished between the two entities by cluster analysis, two DLBCL were misclassified and 23 cases showed overlapping miRNA expression between BL and DLBCL, reminiscent of tumor samples with intermediate gene expression profiles.

In a similar fashion and exploring miRNA molecular signatures, Di Lisio and associates⁷⁰ in 2012 analyzed a number tumor samples that included 12 BL and 29 DLBCL. A direct comparison and confirmation by quantitative real-time polymerase chain reaction showed a total of 19 differentially expressed miRNAs, confirming previously described findings by Lenze and associates that miR-155, -29b and -146a showed reduced expression in BL. In addition, miR-17-3p, -595 and -663 were downregulated in DLBCL in this study. Using the SOTA algorithm for unsupervised clustering of the 19 differentially expressed miRNAs, a total of 66 out 71 cases were correctly classified (93% accuracy). Interestingly, some of the BL tumor samples of this study did not harbor a *MYC* rearrangement; however, analysis of mRNA and protein expression patterns of *MYC* showed no significant difference between *MYC*⁺ and *MYC*⁻ BL. Furthermore, these cases fail to cluster based on *MYC* expression level or *MYC* status. These findings may support observations that a subset of cases with classical BL morphology may not carry a *MYC* rearrangement and other molecular mechanisms may be involved in *MYC* deregulation. Therefore, further investigations are needed to determine whether *MYC*⁻ BL really exists as previously stated in this literature review.

Similarly, single nucleotide polymorphisms (SNPs), another high resolution microarray based technology that detects genome wide numerical abnormalities, specific amplifications, low DNA copy number changes, and loss of heterozygosity with no loss

or gain of DNA, also referred to somatic uniparental dysomy (UPD) or loss of heterozygosity (LOH) has been used to identify chromosome imbalances in non-Hodgkin lymphomas. In the previously described study of Deffenbacher and associates⁷¹, the group also used a SNP array to detect copy number aberrations (CNAs) in DLBCL; however, there was no direct comparison between BL and DLBCL. In this analysis, 18 well defined pediatric DLBCL (11 GCB, 2 ABC, 2 primary mediastinal DLBCL and 3 non-classifiable DLBCL) were included in the analysis. Cases with a diagnosis of primary mediastinal DLBCL carry gains of chromosome 2p. Other previously observed aberrations in DLBCL^{72,73} were also observed by SNP analysis, mainly loss of the long arm of chromosome 6, -4p14, +7, +12, 17p-, 17q+, -19q13.32, and LOH at 1q, 2, 6p, 9p, as well as 19p. In a recent report by Scholtysik and associates⁷⁴ in 2012, frequent deletion at 19p encompassing *TNFSF7* and *TNFSF9* were both found in both BL and DLBCL. Despite the great potential of SNP arrays in detecting chromosome imbalances, there is limited work in this area and further investigations are needed, at least in directly comparing BL and DLBCL.

Even though both gene expression and miRNAs profiles can reliably distinguish between the two diagnostic entities, these are complex techniques and not currently available in most clinical diagnostic laboratories. Therefore, more conventional techniques such as immunohistochemistry and cytogenetics, normally included in a lymphoma work-up, may serve as surrogate markers for expression profiles e.g., bcl-2 protein overexpression measured by immunohistochemical methods may be used instead of *BCL-2* gene expression. Indeed, Soldini and associates⁷⁵ recently used a differential protein expression algorithm based on gene expression profiles to distinguish between

BL and DLBCL. In this study, CSE1L and ID3 protein over expression, earlier identified by Dave and Hummel^{5,6} by GEP, were associated with BL and STAT-3 with DLBCL ($p < 0.001$). Moreover, recurrent cytogenetic aberrations may also be used as surrogate markers for gene expression similar to the example just listed above. One such example is gain of chromosome 11 that correlates with CD44 (located at 11p13) overexpression, a distinct marker of the NF- κ B molecular signaling pathway and characteristic of DLBCL.^{6,76} In terms of SNPs, more research is needed and similar to the drawbacks of both GEP and miRNAs; it is not routinely available for clinical diagnostics.

2.6 Classification of Non Hodgkin Lymphoma by Bioinformatics Methods: Logistic Regression, Cluster Analysis, and Artificial Neural Networks

Even though logistic regression (LR) models are found all through the literature, to the best of our knowledge, there are no studies applying LR classifiers specifically to non- Hodgkin lymphoma based on relevant cytogenetic markers. Nonetheless, there is a long list of LR classifiers that use a combination of biomarkers to predict a wide spectrum of malignancies including non- Hodgkin lymphomas.^{77,78} Unlike LR classifiers, unsupervised cluster analyses used to distinguish BL from DLBCL have been developed. To some extent we have already described some of these classifiers in previous sections. For example, Gormley and associates⁴⁸ used an unsupervised hierarchical clustering method and applied 8 GCB and ABC markers to 13 BL and 5 DLBCL tumor samples with a sensitivity of 81% and specificity of 87%. Likewise and using a similar approach, Alizadeh and associates⁶³ successfully identify three distinct molecular clusters in DLBCL (GCB, ABC and primary mediastinal DLBCL categories)

and Dave et al⁶ used a classifier based on GEP to correctly categorize 25 confirmed cases of BL. Similarly, both Lenze⁶⁹ and Di Lisio⁷⁰ used unsupervised clustering applied to miRNAs to appropriately classify BL from DLBCL. In the case of Di Lisio et al. 93% of the cases submitted for cluster analyses were correctly classified by this method. Alternatively, within the last decade, studies have reported using supervised methods for classification purposes. In the paragraphs that follow, we summarize the potential usefulness of artificial neural networks (ANNs), a supervised pattern recognition method, to classify non Hodgkin lymphoma.

In 2003, O'Neill and colleagues¹⁶ explored a 4026 gene panel network of 40 DLBCL patients. This study was designed to predict clinical outcome on individual patients and to distinguish DLBCL from other donors including other different types of lymphomas. For predicting a diagnosis of DLBCL, this study initially trained 10 networks with 96 donors based on 4096 gene set. This generated a subgroup of 292 genes with three errors produced. Thereafter, the gene set of 292 was treated in two different ways. Both treatments resulted in 19 genes predicting DLBCL with two errors produced. The gene set of 19 was also used in a follow-up study of 46 donors. Here eleven networks were trained. In this subsequent run, the networks produced a total of one error or 98% correct. This study shows how networks can be trained to establish a perfect prognosis (up to ten years) and near-perfect classification of DLBCL. However, it should be noted that the sample population used in this study may not be totally representative of a larger data set, and for this reason, much larger studies are needed to validate these results.

In the same year that O'Neill and associates published their results, Ando et al.¹⁷ used a fuzzy neural network (FNN), an advanced ANN model, on gene expression data that selected a combination of genes to predict the classification and outcome of DLBCL patients after anthracycline-based therapy. This particular network allowed high accuracy (91%) with classification of DLBCL patients into 5 different groups.

Using a similar methods, Liu and colleagues¹⁵ used a combinational feature selection coupled with ensemble neural network to enhance the classification of leukemia, lung cancer, prostate cancer, ovarian cancer, colon cancer and DLBCL (n = 47 samples based on a 4026 gene set). Here, the focal point is on the latter. In this study, a combinational approach for selecting the most useful gene sets for classification was used. Presently, a number of selection methods (correlation coefficient, signal to noise ratio, mutual information and Euclidean distance)⁷⁹⁻⁸¹ are available for determining top-ranked genes associated with a particular disease. However, Li and associates⁸² concluded that low ranked-genes are also important in selecting genes that may produce a more accurate classification. Based on this, several studies have proposed improved selection of gene sets.^{81,83-86} In this present study, Liu and colleagues selected a combination of methods to obtain better classification scheme. These included a wilcon's ranksum test⁸⁷ to obtain the top-ranked gene sets, a statistical analysis termed PCA⁸⁵ to obtain key components for a good approximation and a clustering method along with a t-test to obtained the top-ranked genes for the clusters.^{81,87} Once the list of genes was selected, the study used ensemble multiple networks to optimize classification. Combining classification results generated by ensemble networks, Liu et al.¹⁵ were able to obtain more accurate results compared to previous classifiers for the same datasets.

Only ovarian classification showed a lower predictive accuracy. Some limitations of the study worth noting include the following: the study failed to describe in detailed the set of genes useful in classifying the various neoplasms, it did not mention the possible role of genes selected and how these may induce disease, the method used is more labor intensive, and it should also be noted that the number of networks forming the model at 100 members in this study is not necessarily the best model.

In a more recent study in 2012 by Cui and associates⁸⁸, the group classified DLBCL (n = 40) into GCB and ABC molecular subtypes by using an ensemble neural network applied to 1277 miRNAs. A multi-layer feed forward network was used for this purpose and a sensitivity analysis was performed to evaluate the importance of input variables, also termed profile method.⁸⁹ This method was improved by calculating the root mean square error (RMSE) of the neural network. The ratio of the initial RMSE vs. the changed RMSE of the network was then assessed, the bigger this ratio the more important the input variable. Based on this, the authors selected 5 highly important miRNAs that better classify the various DLBCL data sets, among these included the NF-kB target miR-146, previously described by Lenze and Di Lisio.^{69,70} Nonetheless, the neural network model used here was not able to accurately predict all samples using the 5 selected miRNAs (e.g. of 10 samples processed with 5 selected miRNAs, one ABC subtype was misclassified as GCB).

To date, there are a limited number of reports of ANNs to classify DLBCL. However, ANN predictive models are plentiful in the scientific literature; for example, a high prevalence of published publications on prostate and cervix cancer, as well as publications on rare diseases abound.⁹⁰ In a systematic review of neural networks,

Lisboa et al. reported a total of 396 hits for neural networks and cancer from 1994 to 2006.⁹⁰ And, perhaps the most influential work on neural networks and its uses on classification of cancer was the study published by Khan et al. in 2001.⁸³ Here, Khan and colleagues demonstrated the potential use of ANNs in classifying diagnostic disease entities; this particular study was able to classify four distinct categories of round blue cells based on microarray data using a dataset of 6000 genes. Likewise, the study by O'Neill and associates¹⁶ was the first publish manuscript to demonstrate a 100% accuracy in the prognosis of large B-cell lymphomas (n = 40), while reducing unwanted noise level in complex datasets. Similarly, Liu et al. reported a higher predictive accuracy using a combinational selection approach coupled with ensemble networks to classify a number of cancers including DLBCL. Since then and coupled with our assessment of four major studies reviewed here, ANN-based predictive models have established a favorable outlook in developing new classification and predictive models in cancer studies; although with the caveat that ANNs have not eclipsed conventional statistical methods currently in place today, and conflicting results of prominent cancer studies reported in the literature⁹⁰ have also added concerns about the use ANN-based models.

Like any other method, ANN-based predictive models have limitations. First, the time necessary to train ANNs may take a considerable time to finalize; and if hidden layers are increased, so does the time is increased. Because of this, only a few hidden layers are used to train ANNs. Another potential drawback is over-fitting i.e., training data committed to memory causes the network to perform less efficiently on prospective cases. Bias in error rates are also another concern. An added concern is their inability to explain how the network reached a solution or commonly referred to “the black box” of

ANN modeling. And, finally, the input data must closely represent the specific scenario under study. If this is not case, results are then invalidated.⁹¹ Notwithstanding these shortcomings, some of the limitations just outlined here can be addressed by staging the following frameworks: a method that can explain the response of networks such as sensitivity analysis or rule extraction, as well as log-odds ratio should be set up as part of the analysis process⁹⁰, error rates can be minimized by cross-validation methods such as Monte Carlo sampling⁹² and over-fitting can be addressed by re-sampling (this allows for performance estimates on new datasets to confidently avoid over-fitting) or by weight decay, a method that keeps weight values low since over-fitting produces large weight values.⁹² Collectively, re-sampling, weight decay and cross validation techniques are one of the most common types of regularization that should be applied during the training of data to optimize predictive performance of ANNs.⁹¹ Once these common practices are established, ANN-based models provide many advantages. For one is their ability to make sense of complicated datasets over conventional statistical techniques, datasets need not conform to normally distributed sample populations, they can manage complex datasets with non-linear relationships and they can also handle noisy information. What is more, ANN models can generalize, that is, their ability to make sense of information that is dissimilar to that provided by the training dataset, and therefore in this respect have the potential application for diagnostic purposes.⁹¹ In terms of performance when compared to LR models or other statistical methods, ANN models have advantages when it comes to forecasting dichotomous outcome and are excellent in diagnostic support. In addition, multiple learning algorithms may be integrated into ANN models compared to only one in LR.⁹³ Thus, at least in theory ANN models should outperform LR.

Nonetheless, a comparative study from 2002 that surveyed the performance of LR vs. ANN in 72 papers from the literature reported that ANN models only marginally outperformed LR.⁹⁴ In this same study, the authors also reported overall better performance from both ANN and LR than that of other statistical tools, mainly those of decision trees and k -nearest neighbors. Only support vector machines performed comparable to ANN and LR. In more recent studies, ANN has outperformed LR or in some instances has shown comparable performance.^{74,95-101} In brief, ANN-based models provide a powerful classification tool that may use to predict BL from DLBCL.

Chapter III

III. Methods

3.1 Overview

To generate unique RCAs that may be used to highly predict DLBCL from BL, a combination of bioinformatic and statistical tools were applied to the following datasets: cytogenetics, copy number variations measured by comparative genomic hybridization, selected tumor samples comprised from the published literature and a small number of institutional cases. These datasets were retrieved from the Mitelman Database of Chromosome Aberrations in Cancer (available at <http://cgap.nci.nih.gov/Chromosomes/Mitelman>), SKY-M FISH & CGH Database at the National Cancer Institute (available at <http://www.ncbi.nlm.nih.gov/sky/>). A total of 338 cases with cytogenetic data (DLBCL, n = 254; BL, n = 84) were queried from the Mitelman database to build a diagnostic model for DLBCL and BL based on RCAs. Initially, cytogenetic data was collected into three groups. These included the following: 1) *MYC*⁺ DLBCL vs. BL, 2) *MYC*⁺ DLBCL vs. *MYC*⁻ DLBCL and 3) a combined dataset of *MYC*⁺ and *MYC*⁻ DLBCL vs. BL. Cytogenetic data was initially analyzed with CyDAS, a bioinformatics tool-kit, that analyzes cytogenetic data in ISCN format. To validate results from CyDAS and to determine differences in the number of aberrations (numerical and structural) between comparative groups, a manual curation of the data was performed for statistically significant RCAs generated from CyDAS; an independent group t-test was performed to determine differences in the number of aberrations. The specificity of these RCAs was also calculated. To test for the reliability

of RCAs and develop a set of predictive models, a distinct cytogenetic dataset of 177 cases (DLBCL, n = 117; BL, n = 60) collected from the literature coupled with a small number of institutional cases was applied to cluster analysis (hierarchical clustering (HC) heat map, HC with p-values, k-means partition using the PAM algorithm), neural networks and logistic regression. Discrimination analysis (area under the receiver operator characteristic curve – ROC) to determine specificity and sensitivity and to assess the functionality of these predictive models was performed. To address internal validity, a number of techniques were used, mainly split sample validation, boot-strap resampling, 10-fold cross validation and well-established RCAs were used as internal controls [e.g., (14;18) translocation or BCL-2 rearrangements and +12 seen in GCB-DLBC and +3/+18 prevalent in the ABC-DLBCL molecular subtypes]. For external validation, method transportability was considered. In this regard, findings of this study were compared and correlated with an extended analysis of array CGH and a copy number variation analysis (n = 259).

3.2 CyDAS: An Online bioinformatics Toolkit for Chromosome Analysis

CyDAS, a bioinformatics tool-kit capable of cytogenetic data analysis in ISCN format, available at <http://www.cydas.org> allows statistical analysis in cancer cytogenetics. This program decodes the genomic aberrations contained in a karyotype into numerical value and assigns to individual chromosome band regions. CyDAS allows for a “birds-eye view” of the different aberrations (gains, losses and rearrangements) for every chromosome band region and display data alongside a chromosome ideogram and in a table format. The resulting image illustrates numeric imbalances with red coloring

for genomic losses and green coloring representing genomic gains (See Figure 3-1). Cytogenetic data selected from the Mitelman database was up loaded to CyDAS for analysis.

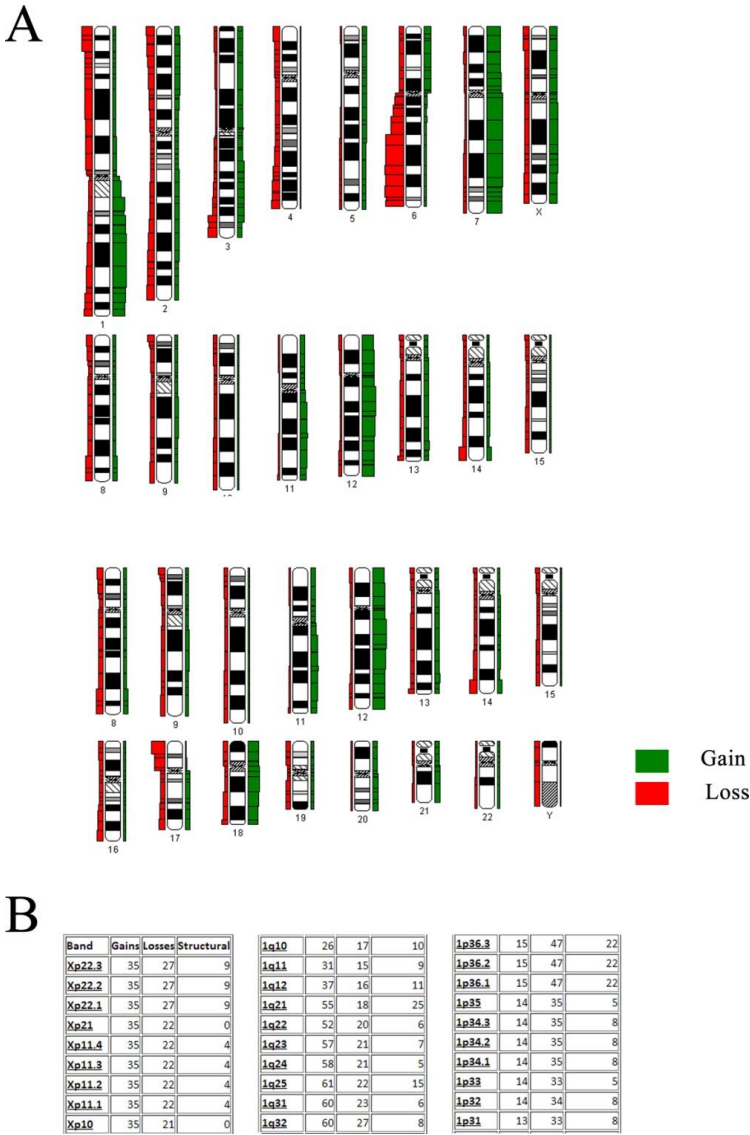


Figure 3-1. Qualitative and quantitative aberrations generated by CyDAS. Panel A illustrates RCAs in chromosome ideogram format, whereas panel B shows a quantitative representation of RCAs in a table format. In this example, the quantitative tables only show gains, losses and structural rearrangements for chromosomes Xp, 1p and 1q.

3.3 Analysis of the Mitelman Dataset in CyDAS

CyDAS graphically analyses chromosome aberration datasets downloaded from the Mitelman Database. Using the case quick searcher (see Figure 3-2), a text file is downloaded from the Mitelman database with a complete dataset specified by a query.

The screenshot shows a web form for searching the Mitelman database. It includes several sections with radio buttons and text input fields, as well as two expandable lists.


- Sole Abnormality:** Radio buttons for ☒ No and ☐ Yes.
- Abnormality:** Radio buttons for ☒ And and ☐ Or, followed by a text input field containing "t(8;14)".
- Breakpoint:** Radio buttons for ☒ And and ☐ Or, followed by an empty text input field.
- Topography:** A list box with the following items: Adrenal, Anus, Bladder, Blood vessel, and Bone and soft tissues (all sites). To the left of the list is the text "Expand Topography List:".
- Morphology:** A list box with the following items: Dermatofibrosarcoma protuberans/Bednar tumor/giant cell fibroblastoma, Desmoid-type fibromatosis, Desmoplastic small round cell tumor, Diffuse large B-cell lymphoma (highlighted in blue), and Diffuse-type giant cell tumor. To the left of the list is the text "Expand Morphology List:". The entire list box is enclosed in a yellow border.
- Special Morphology:** An empty text input field.
- At the bottom, there are two buttons: "Submit" and "Reset Form", separated by the word "or".

Figure 3-2. Screenshot of the Mitelman Case Quick Searcher. The case quick searcher allows the user to query samples with a specified abnormality, breakpoint, topography and morphology.

In this search example, a translocation between chromosomes 8 and 14 [t(8;14)(q24;q32)] with a diffuse large B-cell (DLBCL) morphology was selected. Once the query form is submitted, the total numbers of specified cases are displayed in Mitelman format. A text

file of all cases in Mitelman format is subsequently uploaded into CyDAS (see Figure 3-3).

File:

DLBCL MITELMAN.txt 

Format:

☒ Mitelman ☐ Custom Format (Banding Analysis) ☐ Custom Format (CGH)

For Custom Format Only:

Separator: ☒ Tabulator ☐ Pipe ("|") ☐ Blank (" ")

☒ Has Header Line

MapViewer:

☒ NCBI ☐ Ensemble

Banding Resolution:

☐ 2 Digits ☒ 400 Bands ☐ 550 Bands

Number of cases:

The maximum number of cases is limited for technical reasons (duration of calculation) to the default value given below. You may use an even lower value.

Maximum Cases:

Total Cases:

Valid Cases:

Other drawing parameters:

If no cutoff is desired, use "0".

CutOff Structural

CutOff Numeric

Figure 3-3. Screenshot of the CyDAS interphase analysis of datasets for gains and losses. The text file uploaded from the Mitelman database and used in the CyDAS online application is highlighted at the red arrow. The format for the file is in Mitelman format and chromosome banding resolution is set at 400.

The online analysis system from CyDAS (analysis of datasets for gains and losses) was used to process the dataset. CyDAS determines the number of gains and losses of all chromosome band regions. These results are displayed in columns alongside a chromosome ideogram and a quantitative representation is also generated in a table format as previously indicated.

From the image generated in CyDAS, a clear overrepresentation of gains and losses is then visualized. For more quantitative results, the table generated by CyDAS provides specific number of events i.e., gains and losses or RCAs, for all chromosome band regions. Such collection of both qualitative and quantitative RCAs may be repeated for different neoplastic states represented in the Mitelman Database. For this project, two datasets for DLBCL and BL were generated from the Mitelman Database and analyzed with CyDAS to obtain a “bird’s eye view” of all RCAs in DLBCL and BL. Moreover, an independent group t-test, performed with the Statistical Analysis Software – SAS (Cary, North Carolina), was used to assess the differences in the number of aberrations between *MYC*+ DLBCL vs. BL and *MYC*+ vs. *MYC*- DLBCL.

3.4 Building a Reliable Set of RCAs for Distinguishing DLBCL (*MYC*+ and *MYC*-) vs. BL

This study involved the analysis of the publicly available information from the Mitelman Database of Chromosome Aberrations in Cancer (available at <http://cgap.nci.nih.gov/Chromosomes/Mitelman>), and other relevant publications accessed through the PubMed search. The Mitelman Database was queried using the key words – B-cell lymphoma, diffuse large B-cell lymphoma (DLBCL), Burkitt lymphoma (BL), t(8;14), t(8;22), t(2;8), 8q24. Lymphoma in pediatric population (patients <20 years of age), tumors in which karyotype was not fully described or did not conform to ISCN nomenclature were excluded. Since a simple karyotype (pseudo-diploid karyotype) was considered as indicative of BL these cases were also excluded from the analysis. Data from these searches were processed as indicated in Fig. 3-4.

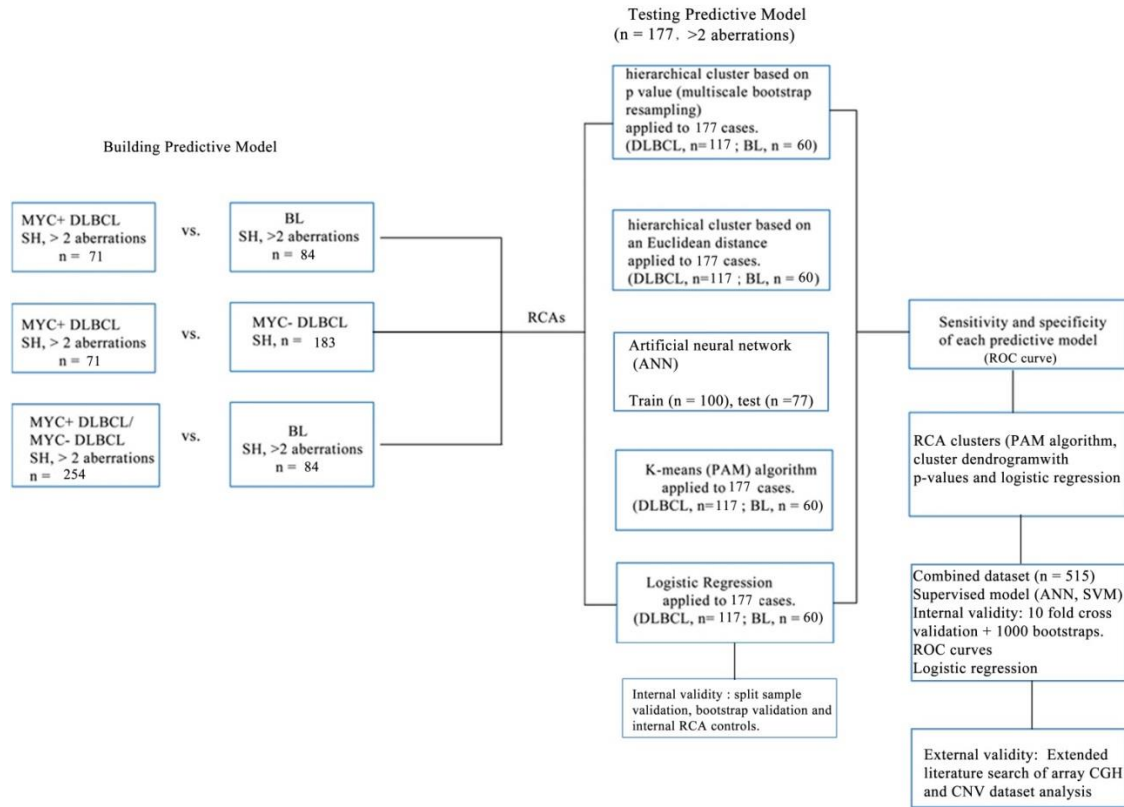


Figure. 3-4. Datasets and analysis work-flow. Key: SH, single hit.

Cytogenetic data from these tumors was initially classified into three analysis groups – MYC+ DLBCL vs BL, MYC+ DLBCL vs. MYC- DLBCL, and DLBCL (MYC+ and MYC-) vs. BL. All three groups were evaluated for effect size and statistical power using the R statistical package.

The first comparative group consisted of 71 DLBCL and 84 BL with MYC rearrangement. The second comparative group consisted of 183 MYC- DLBCL was evaluated and compared to that identified in the 71 MYC+ DLBCL. To further evaluate RCA differences between MYC+ vs. MYC- DLBCL, a larger dataset of sample karyotypes not restricted to more than two aberrations (MYC+, n = 104; MYC-, n = 202) was also analyzed. The third comparative group consisted of 254 cases of DLBCL

(MYC+ and MYC-) and 84 cases of BL. Lastly a fourth comparative group was identified to evaluate whether double hit lymphomas with *MYC*+ and *BCL*-2+ rearrangements (n = 55) are close in chromosome architecture to SH DLBCL (n = 306), we compared all these groups using the Fisher Exact Test, a statistical test that measures the strength of association between two variables in a 2 x 2 contingency table.

Cytogenetic information from the online karyotype of these selected cases was uploaded into CyDAS, a bioinformatics program developed to analyze the structural and numerical abnormalities of tumor karyotypes (Hiller et al.,¹⁰² available at <http://www.cydass.org>). From these results, the differences between the two comparative groups were tabulated and then confirmed by a manual analysis of all karyotypes included in the analysis for accuracy. Fisher exact test was then used to determine the strength of association of different aberrations between the two comparative groups. To provide further evidence between statistically correlated RCAs in the two distinct tumor types (DLBCL and BL), specificity of all RCAs was determined. Indeed, the higher the specificity, a measurement that rules in disease, the higher the diagnostic probability of a given RCA. Although we should highlight that a clue with 100% specificity, is not essentially present in every patient with the disease.¹⁰³ The equation for specificity is shown below.

$$\text{Specificity} = \text{True Negative} / \text{True Negative} + \text{false positive} \quad (3-1)$$

3.5 Cluster Predictive Models of DLBCL and BL.

To further assess the performance of RCAs in predicting DLBCL vs. BL and to build a predictive model, we applied two unsupervised hierarchical clusters-HCs (a HC heat map based on Euclidian distance and one based on p-values) to a fourth and distinct dataset of DLBCL (*MYC*⁺ and *MYC*⁻, *n* = 116) and BL (*n* = 60). Heat maps use colors instead of numbers to represent data, for example, in a dichotomous table of 0 and 1 representing the presence and absence of RCAs, the low and high values in the heat map are set to blue and bright red respectively. One advantage of heat maps is the ability to visualize clustered data that are not otherwise apparent in a large number table. Normally heat maps are combined with hierarchical clustering to arrange similar items or cluster items based on a comparable distance, usually a Euclidian distance. The result is a hierarchical cluster that is presented in a heat map as a dendrogram, a tree like structure (see Figure 3-5 for a column dendrogram).

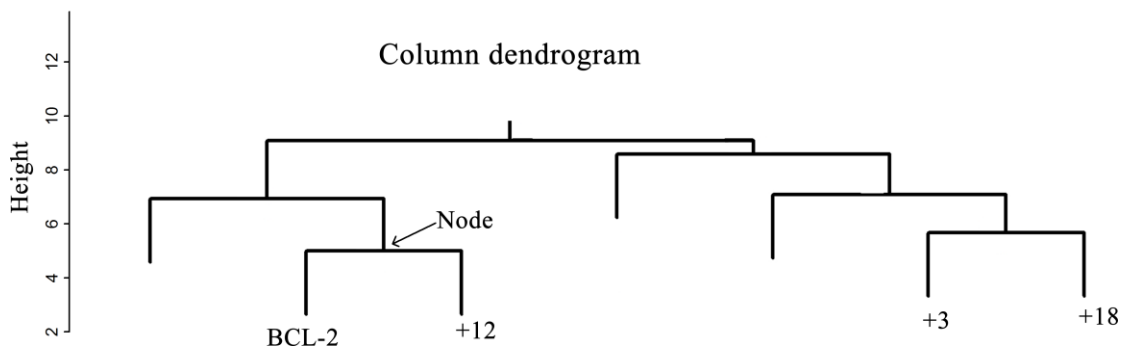


Figure 3-5. Typical column dendrogram of a hierarchical cluster. The dendrogram illustrates the distance or similarity between columns (e.g., variables BCL-2, +12 are closer to each, whereas +3, +18 form a separate and distinct cluster) and the node each of the variables belong base on the clustering calculation.

In the case of a HC based on p-values, it evaluates the uncertainty in the hierarchical cluster analysis by calculating a p-value for each of the clusters in the dendrogram. In this analysis, two types of p-values are generated, approximately unbiased (AU) and the bootstrap probability (BP). Of these, the AU type p-value determined by multiscale bootstrap has a better estimate than the BP p-value, which is generated by a normal bootstrap resampling technique.¹⁰⁴ Misclassified cases generated by the two HCs were assessed with a K-means partition using the PAM algorithm¹⁴, a logistic regression analysis and a RCA cluster based on p-values. With respect to a K-means partition cluster, it has a mean vector that finds the nearest cluster for an object, while at the same time reducing object dissimilarity by a distance metric, usually a Euclidian distance. There are limitations to k-means, such as processing speed, parameters that can be designated by the user, responsiveness to outliers and initial means, as well as shaped clusters. Various other algorithms have been developed to address some of these shortcomings. One alternative to using the k-means and that addresses outliers is the k-medoids algorithm or PAM (partition around medoids). In this instance, the k-medoids assigns the mean to the object that is closest to the center and reduces distances among the different objects within a cluster.¹⁴ To test the internal validity and functionality of the predictor model, we included *BCL-2* rearrangements. Although *BCL-2* rearrangements exclude a BL diagnosis, these cases were included as an internal control. The reason here is that those cluster members farther away from BL, in this instance *BCL-2* rearrangements, should reside closest to a DLBCL cluster membership in a predictive model. The 177 cases applied to HCs (heat map and p-value), PAM algorithm and logistic regression consisted of 12 institutional cases of

DLBCL histology previously reported³⁵, an additional 7 BL cases not published and 157 cases from the publish literature. The statistical R package (The R Core Team, Vienna, Austria) using the gplots, pvclust and cluster libraries were used for all types of analyses described above.

3.6 Logistic Regression Model of DLBCL and BL

The test-dataset of 177 cases previously used for the clustering analyses and containing greater than 2 chromosome aberrations was applied to a logistic regression analysis. Findings from the logistic regression were taken into account when assessing misclassified cases using hierarchical clustering techniques. Logistic regression is a method that is used to evaluate the impact of a given predictor (independent variable or RCAs in this project) on a categorical dependent variable (in our case BL or DLBCL). It is used to predict categorical outcomes when two or more categories are present.¹⁰⁵ In our case, a logistic regression analysis allows to evaluate how well a set of RCAs explains BL or DLBCL. It also calculates the prediction ability or calibration of the model by performing a test of goodness of fit, normally the Hosmer-Lemeshow test. It also describes the comparative importance of the individual RCAs and determines specificity and sensitivity of the model (concordance statistic or c statistic).¹⁰⁵ In logistic regression, odds ratios are used instead of probabilities. The odds ratio approach gives a measurement of the effect of independent variables (RCAs) on the dependent variable (BL or DLBCL), for example, a 7.5 odds ratio of a given RCA (e.g., gain of chromosome 6p) associated in DLBCL is interpreted as the odds of a patient with DLBCL having a gain of chromosome 6p is 7.5 more than that of a patient with BL. The logistic regression equation for prediction is listed below:

$$p = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k} / 1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k} \quad (3-2)$$

Here, p represents the probability of the event occurring (outcome); e designates the base of the natural logarithm or exponent function, $X_1 \dots X_k$ designates k independent variables, β_0 is a constant and β_i is the coefficient of the independent variable X_i . In this project, a multivariable analysis (logistic regression) with stepwise selection (stepwise independent variable selection) was then used to determine the relationship between the two diagnostic tumor types (dependent variable) and the various reliable predictor RCAs (independent variables).

3.7 Neural Networks and SVM models for the Classification of DLBCL and BL

Data from 177 sample karyotypes with DLBCL ($n = 117$) and BL ($n = 60$) were randomly divided into training and testing groups (100 cases were used to train the model and 77 cases were used to test the model). The total numbers of cases comprised of 177 cases used previously to develop the predictive model based on clustering and logistic regression. Every sample karyotype of the dataset was described by an input of 21 RCAs (1 RCA, BCL-2, served as internal control), four hidden layers, including one bias node along with one output node (See Figure 3-6).

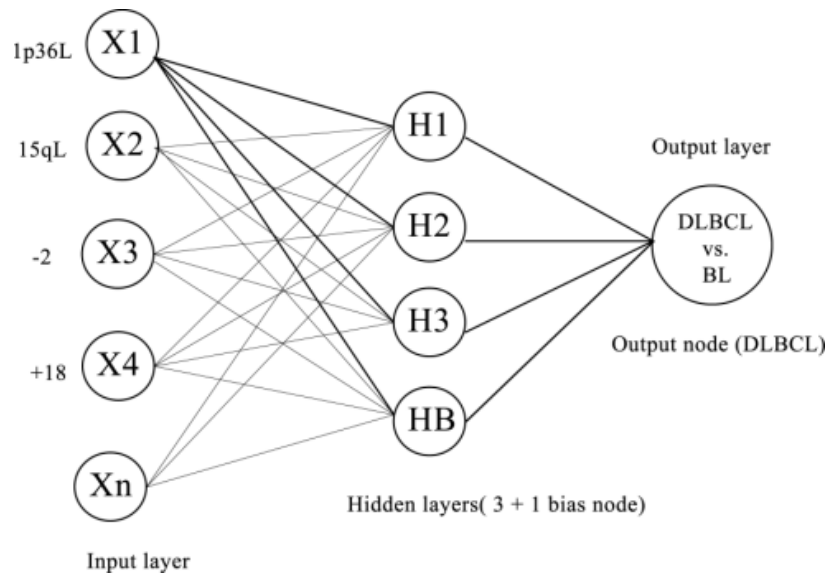


Figure 3-6. Schematic representation of an ANN model. The model contains 21 input nodes (X1,X2,Xn), three nodes within a single hidden layer, one bias node and one single output node representing DLBCL.

The logistic sigmoid activation function was used for each of the hidden and output nodes. An additional ANN model was processed with only one hidden layer. The ANN was processed by R using the neuralnet and nnet packages. The training phase of the ANN model, described as the optimum parameters (e.g., weights and biases) of the model to accurately predict a patient with DLBCL or BL, was carried out by the globally convergent algorithm based on back propagation (BP). For the training phase, approximately 2/3 ($n = 100$) of the total cases were used for this purpose. In the course of this phase, the ANN model is provided with input data, that is, RCAs of sample karyotypes and the correct answer is also assigned to the model. During this process, the network adjusts the connections weights to diminish observed inaccuracies in matching both inputs and outputs. Neural network connections are fixed once the network has

correctly classified all training dataset. The ANN model was then tested using the remaining 1/3 of the total dataset (n= 70) to evaluate the accuracy of the model.

In order to further evaluate reliable RCAs in the context of a larger sample population, a SVM and ANN models were constructed of all tumor samples with cytogenetic information. The total samples (n = 515) included the dataset used to identify reliable RCAs (n =334) plus a second dataset used to test predictor models (n = 177). The statistical R package using the e1071 library was used to develop the SVM model.

3.8 Discrimination (AUC and specificity) Studies of Predictive Models

Traditionally, to assess how well a model performs, both specificity and sensitivity are determined by calculating the area under the receiver operator characteristic curve (ROC)¹⁰⁶, also referred to the c statistic or discrimination (this is the principal component of a predictive model if a high risk group is to be identified – in our case, DLBCL vs. BL). An acceptable model has a c statistics above 0.7, over 0.8 is considered a good model and anything above 0.9 is considered an excellent model. In general, discrimination or the c statistic is preferable when a diagnostic test is required.⁷³ Thus, for the purpose of this present study and the primary requirement required to accurately diagnose one entity from the other, it is the discrimination of the model (how well the model can distinguish between DLBCL vs. BL) that will take precedence. To determine discrimination (i.e., ROC curves, specificities, sensitivities and the c statistic) for all predictor models, SAS using a logistic regression and the statistical R package (ROCR and Epi libraries) were used in this study.

3.9 Internal validation Techniques

To address internal validation in this study, a number of techniques were used. Among these methods used included a split-sample validation (one dataset used to construct reliable RCAs while a different dataset was used to evaluate the model), 10-cross validation (10 different deciles of the underlying sample population serve as validation) Bootstrap validation (repetitive re-sampling from an underlying population and each used as an original dataset for analysis) used in the hierarchical cluster with p-values and internal RCA controls (e.g., initially included in the study BCL-2, +3/+18 and later added 3q27 rearrangements all indicative of a tumor phenotype consistent of DLBCL).

3.10 External Validation: an Extended Literature Review of Array CGH in DLBCL Subtypes vs. BL and Compared with RCAs

To address external validity by method transportability (i.e., collection of data and analyzed with an alternative method), we performed an extended literature review of array CGH. In 2012, our group initially carried out a literature review of copy number aberrations performed by array CGH in DLBCL subtypes and reported unique copy number aberrations between distinct DLBCL subtypes. In this present study, we expanded this initial analysis and compared the different DLBCL subtypes vs. BL to determine correlations between RCAs and copy number variations by array CGH. Only those copy number aberrations that were listed in the literature with a frequency of 20% or greater were considered significant. Moreover, we should highlight the analysis of copy number variations (using a classical CGH dataset), as well as a dataset of gene expression profiles that were used to further support our cytogenetic findings.

In terms of the copy number variations-CNV analysis and to identify unique CNVs between the two distinct entities, a total of 249 (n=78, BL; n=171, DLBCL) cases were retrieved from the SKY-M FISH & CGH database at the National Cancer Institute (available at <http://www.ncbi.nlm.nih.gov/sky/>) and the Progenetix Database, a database of copy number abnormalities in cancer available at <http://www.progenetix.org>. All cases included in this analysis were previously analyzed by classical comparative genomic hybridization (CGH). The total number of DLBCL cases included five distinct categories, mainly ABC, GCB, NOS, transformed DLBCL and relapse DLBCL. For the BL subgroups, four distinct subtypes were identified. These included: atypical BL, classic BL, discrepant BL and BL-NOS. The two major diagnostic groups were grouped together and analyzed (DLBCL vs. BL) using the Fisher Exact Test to assess CNVs differences between the two diagnostic groups. A logistic regression analysis to evaluate the impact of individual CNVs on the two tumor types was then performed. The statistical software package SAS was used for the logistic regression. A p-value of $< .05$ was considered significant of all RCAs and CNVs data analysis.

Chapter IV

IV. Results

4.1 Genetic Complexity of DLBCL vs. BL and CyDAS analysis

Research question 1: Is the average total number of chromosome aberrations in *MYC*+ DLBCL greater in number than BL? Initial analysis of all aberrations (numerical and structural) identified a total of 899 observations or 12.6 / tumor in *MYC*+ DLBCL, and a total of 814 observations or 9.7 / tumor in BL; this difference was significant by an independent group t-test ($p = .003$).

An analysis of structural and numerical analysis using the CyDAS also revealed significant differences in the number of chromosome aberrations. Among these, gain of chromosomes X, 2, 3, 5, 7, 11, 12, 18 and 21, and loss of chromosomes X, 1p, 2, 3q, 4, 6q, 8, 9p, 9q, 10, 14, 15q, 16, 17 were more prevalent in DLBCL than in BL (see Fig. 4-1). A quantitative representation of the number of aberrations registered by CyDAS at 1p21-36 is illustrated in Table 4-1. In this instance, there were a total of 67 losses in DLBCL and 9 losses in BL, and the difference was significant ($p=0.001$); this suggests 1p36 loss is associated with a DLBCL histology rather than a histology of BL.

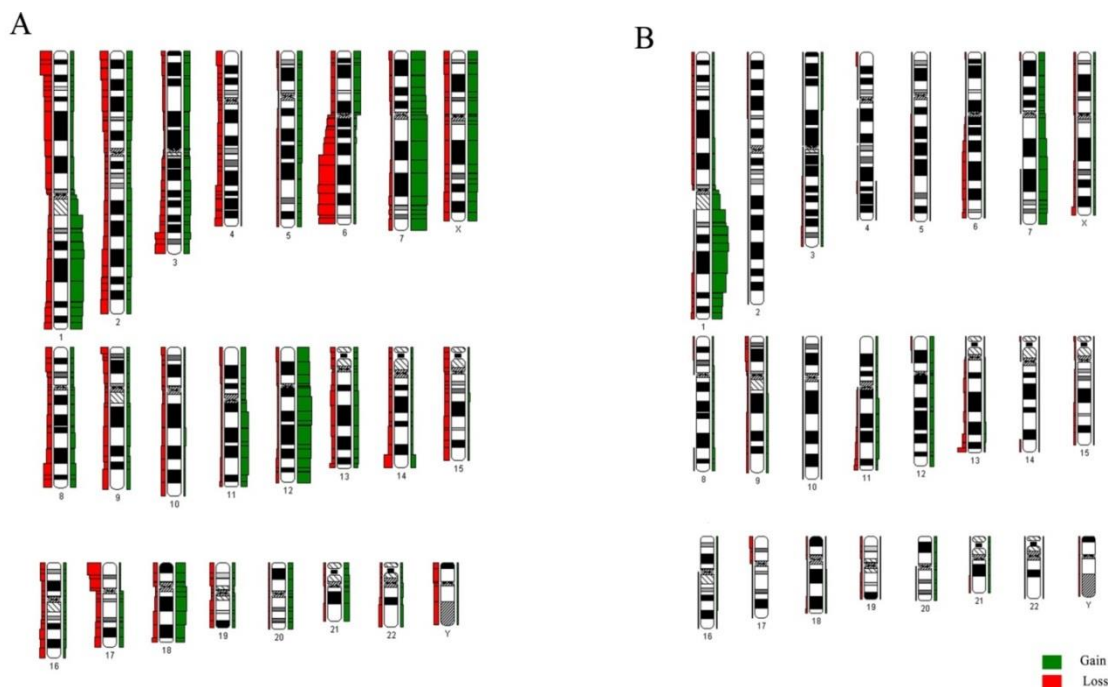


Figure 4-1. Distribution of gains and losses in DLBCL and BL identified by CyDAS. A. DLBCL(MYC+ and MYC-), B. BL histology.

Table 4-1. Quantitative representation of aberrations at chromosome 1p by CyDAS

	G	L	R
<u>1p36.3</u>	20	67	27
<u>1p36.2</u>	20	67	27
<u>1p36.1</u>	20	67	27
<u>1p35</u>	19	48	5
<u>1p34.3</u>	19	48	10
<u>1p34.2</u>	19	48	10
<u>1p34.1</u>	19	48	10
<u>1p33</u>	19	44	5
<u>1p32</u>	19	45	10
<u>1p31</u>	18	43	9
<u>1p22</u>	18	42	16
<u>1p21</u>	21	37	10

	G	L	R
<u>1p36.3</u>	7	9	4
<u>1p36.2</u>	7	9	4
<u>1p36.1</u>	7	9	4
<u>1p35</u>	7	9	3
<u>1p34.3</u>	7	9	4
<u>1p34.2</u>	7	9	4
<u>1p34.1</u>	7	9	4
<u>1p33</u>	7	9	3
<u>1p32</u>	7	9	3
<u>1p31</u>	7	10	4
<u>1p22</u>	9	10	6
<u>1p21</u>	9	10	3

DLBCL (MYC+ and MYC-)

BL

Key: G, gains; L, losses; R, rearrangements.

4.2 RCAs between MYC+ DLBCL and BL

Research question 2: Are there differences in the frequency of different RCAs between *MYC*+ DLBCL vs. BL? Since in tumors with clonal evolution CyDAS may have registered a chromosome aberration at a specific band more than once, the frequency of tumors with all statistically significant aberrations was manually verified. Initial manual analysis was performed on 71 *MYC*+ DLBCL and 84 BL cases. The type of *MYC* translocation and frequencies of tumors in each cytogenetic class is presented in Table 4-2.

Table 4-2. Morphological Groups for *MYC*+ DLBCL and BL

Group	t(8;14)	t(8;22)	t(2;8)	non- <i>IG/MYC</i>	Total
DLBCL	51	6	6	6	71
BL	63	13	8	0	84

Key: BL, Burkitt lymphoma; DLBCL, diffuse large B-cell lymphoma.

A manual curation of RCAs of this comparative group revealed +X, +7, 15q loss, +16, 17p loss, and +18 were significantly associated with *MYC*+ DLBCL than BL (Table 4-4).

4.3 RCAs differences between MYC+ and MYC- DLBCL.

Research question 3: Are there differences in the frequency of different RCAs between *MYC*+ DLBCL vs. *MYC*- DLBCL? In order to verify whether the significant RCAs identified in the above analysis were specific to *MYC*+ DLBCL, or they represent

DLBCL in general (*MYC*⁺ and *MYC*⁻ DLBCL as a single entity), chromosome abnormalities of 183 *MYC*⁻ DLBCL were evaluated and compared along the same lines as *MYC*⁺ DLBCL and BL. In this instance, a total of 2417 aberrations or an average of 13.2 / tumor was identified in this group. The pattern of aberrations and their frequency (13.2/tumor vs. 12.6/tumor, $p = 0.59$) in *MYC*⁻ DLBCL was similar to that seen in *MYC*⁺ DLBCL. The aberrations that showed significant positive association with *MYC*⁻ DLBCL were 1p36L, 1qL, 19pL and monosomy 8. Therefore, presence of these aberrations in a DLBCL tumor may suggest the absence of a *MYC* rearrangement. In contrast, +7 and 15qL were associated with the *MYC*⁺ group. However, in a larger dataset of *MYC*⁺ (n= 104) vs. *MYC*⁻ DLBCL (n = 202) where the number of aberrations was not restricted to more than two per karyotype , only 1qL, 16qL and monosomy 8 remained significant, and +7 and 15qL were not statistically significant.

In terms of 1p36 L in this larger dataset, the *MYC*⁻ group showed a higher prevalence for this aberration, although non-significant compared to the *MYC*⁺ DLBCL group ($p = .07$, data not shown). The above results indicate only marginal differences between the *MYC*⁺ and *MYC*⁻ groups. Therefore in general, these RCAs may be useful in distinguishing between DLBCL and BL. Moreover, those aberrations associated with the *MYC*⁻ group may characterize the ABC molecular subtype of DLBCL since presence of *MYC* rearrangements in DLBCL are more associated with the germinal subtype of DLBCL. Table 4-3 illustrates differences between *MYC*⁺ and *MYC*⁻ DLBCL.

Table 4-3. RCAs Differences between *MYC*+ and *MYC*- DLBCL

RCA Value	MYC+ DLBCL (n=71) SH (> 2 aberrations)	MYC- DLBCL (n=183) SH (> 2 aberrations)	p-
1p36 L	10	57	.04
1qL	6	48	.01
+7	26	36	.03
15qL	21	25	.02
19pL	5	24	.01
16qL	4	29	.05
-8	1	24	.002

Key: L, loss; G, gain.

4.4 Differences between DLBCL (*MYC*+ and *MYC*-) vs. BL.

Research question 4: Are there differences in the frequency of different RCAs between DLBCL (*MYC*+ and *MYC*-) vs. BL? Since the difference in the occurrence of different RCAs between *MYC*+ and *MYC*- DLBCLs was only marginal, we combined the data from all DLBCL tumors and compared it with that from BL. The combined data confirmed the previously identified RCAs by CyDAS and identified an additional two RCAs (1qL and 22qL) significantly associated with DLBCL compared to BL (see Table 4-4). With regards to 9qL, and 19pL, these RCAs were not significant in comparing DLBCL with BL; however, they remained significant when only the *MYC*- DLBCL group was compared with BL ($p = .04$ for 9qL and 19pL). In the case of +16, it remained significant in the *MYC*+ DLBCL vs. BL. Figure 4-2 is a visual representation of Table 4-4 illustrating all the RCAs that may be used to distinguish between DLBCL and BL.

Table 4-4. RCAs Differences between characteristic groups of DLBCL vs. BL

RCA	DLBCL SH (> 2 aberrations)	BL (n=84) SH (> 2 aberrations)	p-Value	Specificity
-----	-------------------------------	-----------------------------------	---------	-------------

DLBCL (*MYC*+ and *MYC*-, n= 254) vs. BL

+X	36	4	.03	95.6
1p36 L	67	6	.001	93.4
1qL	48	5	.004	94.4
+2	14	0	.026	100
-2	22	0	.003	100
+3	29	3	.05	96.6
-4	29	4	.015	95.6
9qL	42	6	.07	93.4
+12	52	7	.03	92.4
14qL	50	5	.009	94.4
15qL	46	4	.006	95.6
17pL	57	8	.02	91.4
+18	43	5	.03	94.4
22qL	20	0	.006	100
-8	25	1	.01	98.8

***MYC*+ DLBCL (n = 71) vs. BL**

+16	4	0	.04	100
+X	11	4	.05	95.6
+7	27	15	.036	85.5
15qL	29	3	.0004	96.7
+16	4	0	.04	100
17pL	16	8	.05	92.6
+18	13	5	.04	94.4

***MYC*- DLBCL (n = 183) vs. BL**

+X	25	4	.05	95.6
1p36L	57	6	.0002	93.4
1qL	48	5	.0007	94.4
+2	11	0	.04	100
-2	19	0	.001	100
+3	22	3	.04	96.7
-4	25	4	.05	95.6
-8	24	1	.002	98.8
9qL	33	6	.04	93.4
+12	39	7	.03	92.4
14qL	45	5	.002	94.4
15qL	25	4	.05	95.6
16qL	29	5	.05	94.4
17pL	41	8	.04	92.6
+18	30	5	.05	94.4
19pL	23	3	.04	96.7
del(22)q	17	0	.004	100

Key: L, loss; G, gain. Specificities for 9qL and 20 are better reflective of *MYC*-DLBCL vs. BL only.

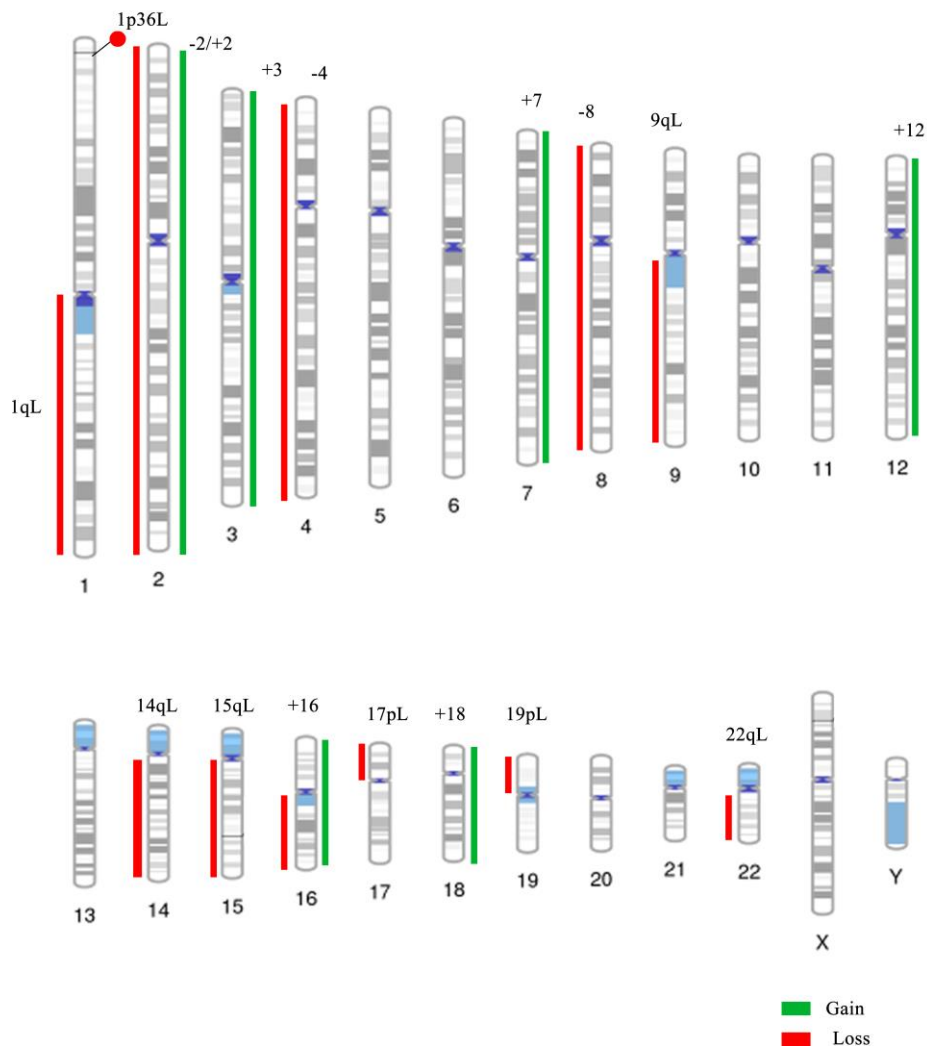


Figure 4-2. Ideogram representation of characteristic gains and losses in DLBCL vs. BL. Ideogram shows a visual representation of gains and losses in DLBCL compared to BL.

4.5 Differences between DLBCL (*MYC*⁺ and *MYC*⁻) vs. BL in a Multivariable Context.

Research question 5: Are there differences between DLBCL and BL with respect to RCAs within a multivariable context? A logistic regression model obtained a discrimination of 85%; this was determined by the concordance statistic or c statistic

(discriminative ability of the model, a measurement of specificity and sensitivity) generated by the logistic regression model. RCAs that remained independent predictors of DLBCL in this latter model included +18, +X, 17pL, 15qL, 1p36L,+3, -4, -8, +16, BCL-2, +2, and -2, with the seven last RCAs predicting DLBCL perfectly.

4.6 Cluster Analysis of DLBCL and BL based on RCAs

To assess the performance of RCAs in predicting DLBCL vs. BL, we applied three unsupervised clustering methods (HC based on Euclidean distance, HC with p-values and a K-means partition analysis using the PAM algorithm), a logistic regression model and a supervised model using artificial neural networks to a total of 177 cases (DLBCL, n=117; BL, n=60). The complete cytogenetic characteristics of all cases (n = 224) that were initially selected are found in Table 4-5. Of note, only those cases with more than 2 aberrations were used for this initial analysis (n = 177).

Table 4-5. Cytogenetic Description in ISCN format of DLBCL and BL

ID	DX	Cytogenetics
Institutional Cases		
1	DLBCL	43,XX,del(1)(q42q44),der(2)t(2;8)(p13;q24),der(3)t(2;3)(p13;q27),-4,add(5)(q33),add(8)(q24),add(9)(q22),-10,add(10)(q24),-15,add(17)(p13)
2	DLBCL	80~85<4n>-X,-Y,del(1)(p32p36.1),-2,ins(2;?)(q31;?)x2,-4,-4,-5,add(5)(p15),add(6)(q13),add(6)(q23),ins(6;?)(q23;?),add(7)(q11.2),t(8;14)(q24;q32)x2,add(9)(q22),add(9)(q22),-10,-10,-11,-12,-13,-15,+16,add(16)(p11.2),add(16)(p13.3),-18,-18,-20,-21,-22,-22,+7~13mar[cp5]/75~87,idem,add(19)(p13.3)[cp10]
3	DLBCL	46,XX,t(1;17;1)(q23;p11.2;q42),der(5;15)(p10;q10)[1 cell]/43~45,XX,-3,add(5)(q13),add(7)(q22),-15,-17,add(22)(q13),+mar1,+1~3mar[cp6 cells]/43~45,XX,-3,add(5)(q13),add(6)(p21.1),add(7)(q22),-15,-17,add(22)(q13),+mar1,+1~3mar[cp8 cells]
4	DLBCL	49~50,XY,+X,add(1)(q21),der(6)t(6;6)(q21;p12),t(8;14)(q24.1;q32),+11,der(13)t(2;13)(q21;q34),der(15)t(1;15)(q12;q22),+17[cp2]
5	DLBCL	48,XY,+X,+3,add(7)(q32),add(10)(p11.2),der(15;21)(q10;q10),i(17)(q10),+18[9]/46,XY[9]
6	DLBCL	38~59,XY,+X,-2,add(3)(q27),+add(6)(p15),+7,+i(8)(q10),dup(10)(q22.3q11.2),+11,+12,13,del(14)(q24q32),+15,add(18)(q12.2),+20,-21,+mar1,+mar2,+mar3,+mar4
7	DLBCL	48,X,-Y,+X,del(6)(q13q21),+7,add(8)(p21),add(14)(q32),add(16)(q22),+der(?)t(?;1)(?;q12)[7 cells]/46,XY[13 cells]
8	DLBCL	50,X,-Y,-2,add(6)(q15),+del(6)(q13q21),+7,+8,add(10)(q22),+11,+12,t(14;18)(q32;q21),-15,der(15)t(15;15)(p11.2;q15),+18,+mar[1 cell]/46,XY[19 cells]
9	DLBCL	53,X,-Y,+X,der(4)add(4)(p16)dup(4)(q25q27),+5,add(12)(p13),+add(12)(p13),-15,add(16)(p13),+add(17)(q23),add(19)(q13.3),+21,+21,+21,+mar1,+mar2[7 cells]/46,XY[10 cells]

10	DLBCL	45,XY,t(2;11)(q23;q12),add(3)(p12),-9,-9,add(14)(q32),add(17)(p11.2),+mar[9]/46,XY[11]
11	DLBCL	48~91<4n>XXY,+X,-4,+der(6)t(1;6)(q21;q13),del(9)(q22),add(12)(p11.2),del(12)(p13),add(13)(p11.2)x2,-14,add(14)(q32),der(16)t(12;16)(q13;p13.3)x2,-17,dup(18)(q21q22)x2,-19,+20,+1~2rs
12	DLBCL	46,XX,add(6)(q11)[1]/46,XX,idem,+12[1]/91,idemx2,i(X)(p10)x2,-8,+12,-15[cp5]/92,idemx2,i(X)(p10)x2,+add(6)(q11),-8,+12,-15[cp3]/46,XX[7]
13	BL	46,XY,t(8;14)(q24.1;q32) [20]
14	BL	46,XX,t(8;14)(q24.1;q32) [20]
15	BL	46,XY,t(8;14)(q24.1;q32) [5]
16	BL	46,XY,t(8;22)(q24.1;q11.2) [20]
17	BL	46,XY,t(8;14)(q24.1;q32) [16]
18	BL	46,XX,t(8;14)(q24.1;q32) [19]
19	BL	46,XY,t(8;14)(q24.1;q32) [20]
20	BL	46,XY,der(6)t(1;6)(q12;q13),t(8;14)(q24.1;q32) [20]
21	BL	47,XY,t(8;14)(q24.1;q32),+12 [20]
22	BL	46,XY,t(8;14)(q24.1;q32),+20[12]/ 46,XY,t(8;14)(q24.1;q32),add(17)(p11.2)[4]
23	BL	46,XY,del(4)(q27q31),t(8;14)(q24.1;q32) [10]
24	BL	46,XY,dup(1)(q?12q?24),i(7)(q10),t(8;14)(q24.1;q32) [8]
25	BL	46,XY,der(2)del(2)(q?13q21)del(2)(q23q31),t(8;22)(q24.1;q11.2),del(17)(p11.2)
26	BL	47,XY,t(8;14)(q24.1;q32),add(18)(q21),+19
27	BL	46,XY,t(8;14)(q24.1;q32), add(13)(q34), del(17)(p11.2)
28	BL	47,+X,idi(X)(p11.2)x2,t(8;14)(q24.1;q32)[20]
29	BL	47,XY,+del(7)(p15p22),t(8;14)(q24.1;q32) [10]/47,idem,ins(1;1)(q25;q31.2q21)[3]
30	BL	47,XX,t(8;14)(q24.1;q32),+12[15]/46,idem,-X,[3]/47,idem,-X,+mar[2]
31	BL	47,XY,+8,der(8)t(8;14)(q24.1;q32),t(8;14)(q24.1;q32),dup(11)(q13q23),add(13)(q22)[19]
32	BL	46,XY,t(8;14)(q24.1;q32) [2 cells]/46,idem,t(2;21)(p21;q22) [15 cells]/45,idem,-Y,t(4;17)(q21;p11.2),del(6)(q21q23) [4]
33	BL	46, XY,del(2)(p21p23),t(8;14)(q24.1;q32)[3]/47,idem,- del(2)(p21p23),+7[1]/ 47, idem,del(2)(p21p23),+der(7)del(7)(p13p15)inv(7)(p22q11.2)[15]
34	BL	47, XY,t(2;8)(p12;q24.1),add(3)(q27),t(12;22)(p13;q11.2),+add(12)(q24.1), add(14)(q32),der(14)t(1;14)(q12;p11.2)add(14)(q32), add(15)(q24) [15]/47,idem,add(2)(q31),-add(3)(q27),+add(3)(q27) [4]
35	BL	46,XX,t(8;14)(q24;q32),add(13)(p11.2)[20]
36	BL	46,XX,t(8;14)(q24.1;q32),der(18)t(1;18)(q12;p11.32) [18 cell],46,X,add(X)(p22.1),t(8;14)(q24.1;q32),der(18)t(1;18)t(1;18)(q12;p11.32) [cp2 cells]

Reference ^{47,107-109}

37	BL	46,XX,t(8;14)(q24;q32),add(11)(q23),add(13)(q34)
38	BL	46,XX,del(6)(q13q23),t(8;14)(q24;q32),add(11)(q23),add(13)(q34)
39	BL	46,XY,ins(1;?)(q21;?),t(8;14)(q24;q32),add(11)(q23),add(17)(p11)
40	BL	46,XY,add(13)(q),t(8;14)(q24;q32),add(17)(p)
41	BL	46,XY,t(8;14)(q24;q32),-13,+der(13)t(13;?)(q26;?)
42	BL	46,XY,dup(1)(q21q32),t(8;14)(q24;q32)
43	BL	46,XY,t(8;14)(q24;q32),+add(1)(q),del(15)(q)
44	DLBCL	51,XX,add(1)(q?4),+3,+3,+7,t(8;14)(q24;q32),+10,+18
45	DLBCL	48,XY,+X,add(2)(p?2),add(3)(p?5),add(3)(q2?),add(6q?1),+add(7)(p1?),-8,add(11)(q23),-13,der(14)t(8;14)(q24;q32),add(15)(q2?),add(16)(p1?),-21,+4mar

Reference ⁵⁵

46	DLBCL	46,XY,t(8;14)(q24;q32),add(11)(q23),der(16)t(13;16) (q22;q24),der(21)t(1;21)(p10;p13)[6]/47,XY,+i(1)(q10),t(8;14)(q24;q32),der(16)t(13;16)(q22;q24)[6]/48,XY,+X,+i(1)(q10),t(8;14)(q24;q32),der(16)t(13;16)(q22;q24)[13]
47	DLBCL	46,XY,+der(1)t(1;7)(q32;q22)del(1)(p11pter),ins(3)(p21), -4,t(8;14)(q24;q32),add(13)(q32),add(21)(p11)[22]/45,idem,t(8;14)(14;22)(q24;q32p11;q11),-22[3]
48	DLBCL	47,XY,t(1;12)(p31;p11),add(8)(q24),add(11)(q25)[13]/47,idem,add(10)(q26),-12,+mar[2]
49	DLBCL	77 - 84,XXX,-X,add(1)p21,dic(1;1)(p36;p13),+2,-5,+7,t(8;14)(q24;q32)x2,-9,-10,-12,-13x2, der(14)t(8;14)(q24;q32)x2,-15,-16,i(21)(q10),-22,+3 - 4mar[cp6]
50	DLBCL	48,X,-X,-4,add(6)(q21),+7,t(8;9)(q24;p13),+der(8)t(8;9)(q24;p13),+10,+12,del(14)(q22),add(18)(q12),-22,+mar[9]
51	DLBCL	86 - 97,XXYY,+X,-1,t(1;4)(q21;p15),del(3q),add(3)(q12),-4, add(5)(q34)x2,del(6)(q13),+del(6)(q13),add(8)(q24),add(10)(q26),+1 - 5 mars[cp15]
52	DLBCL	54 - 58,XY,+del(1)(p22),+2,-3,-4,+7,del(8)(q24qter),-9,+add(12)(p11),-14,+15,+15,del(17)(q25), +7-10mars,+2dm[cp10]
53	DLBCL	46,X,der(X)t(X;8)(q22;q24.3),del(2)(p13p21),der(3)del(3)(q21q26)t(3;14)(q27;q32),add(4)(p11),add(6)(q11),der(8)t(X;8)(q22;q22),del(10)(p11.1p11.2),der(14)t(3;14)(q27;q32),-21,+mar[25]

54	DLBCL	47,X,-Y,add(1)(p36),dup(2)(p15p23),-3,-del(4)(q25q31),+7,t(8;14)(q24;q32),-9,del(12)(p12),add(13)(q14),-15,-16x2,-17,+18,+22x2,+3mars,+ring[cp8]
55	DLBCL	37-42,X,-X,-2,-8,-12,-19,del(19)(p13),-20,-21,-22[cp5]/48,XX,+del(3)(q27),add(8)(q24),-9,-17,+18,-22,+3mars[1]
56	DLBCL	47,XY,+der(?)t(1;?)(q24;?),t(3;14)(q27;q32),del(8)(q24),add(13)(q14),t(18;22)(q21;q11)[2]
57	BL	46,X,add(X)(q11),inv dup(1)(q12q32), t(8;14)(q24;q32), t(13;3)(3;22;7)(q14;q27p21;q11;q34)[9]
58	BL	46,X,del(X)(p11),t(2;8)(p11;q24),del(6)(q13)[7]/46,idem,der(7)t(1;7)(q11;q11)[11]
59	BL	46,XY,inv dup(1)(q12q32),add(7)(q32),t(8;14)(q24;q32),add(13)(q14)[9]/50,XY,+1,add(7)(q32),+add(7)(q32),-10,+12,t(8;14)(q24;q32),+19,+mar[cp3]
60	BL	48,XY,t(8;14)(q24;q32),der(13)(q32),+20,+mar [15]

Reference ⁵¹

61	BL	48,XX,+7,t(8;14)(q24;32),+13
62	BL	46,XX,t(8;14)(q24;q32),inv(9)c+12
63	BL	43-46,XY,add(1)(q21),t(8;14)(q24;q32),+1-2mar
64	BL	44,XY,add(6)(p21),t(8;14)(q24;q32),add(13)(q34),add(14)(p11),-15,-22
65	BL	48-49,XY,t(8;14)(q24;q32),+9,add(9)(q22),add(13)(q34),add(18)(23),+1-2mar
66	BL	47,XY,t(8;14)(q24;q32),+12,der(14)
67	DLBCL	89-90,XXYY,+X,+X,add(1)(p33),add(1)(p21),add(7)(p22)x2,+12,t(14;18)(q32;q21),del(20)(q11.2)x2,+3
68	DLBCL	44,X,-Y,+9,der(9;13)(q10;q10),-15
69	DLBCL	74-77,XXX,add(3)(p12),-10,-10,-10,add(13)(p11.2),add(16)(p13.3),+20,+20,+20,+7,-17mar,4-6dmin[cp]
70	DLBCL	50,XX,add(1)(q44),+2,+7,del(10)(q24),t(14;18)(q23;q21),del(15)(22),+16,+21[cp]
71	DLBCL	62-90,XXY,-Y,-4,-14

Reference ¹¹⁰

73	DLBCL	48,XY,+Y,+18,-19,+mar
74	DLBCL	47,XX,der(1)dup(1)(q21q23)dup(1)(q25q42),+3,del(6)(q13q23),-11,+der(14)t(11;14)(q13;p13),-15,+18,add(21)(p11)/47,idem,del(4)(q21q25),add(5)(q13)/47,idem,del(4),add(5),+7,+15
75	DLBCL	60,XY,-X,+Y,-1,-2,+i(3)(q10),-4,-6,-8,-9,-10,-13,-15,-17,+18,der(19)t(11;19)(q13;p13),-21
76	DLBCL	49,XY,add(6)(q11),+add(7)(q32),+16,add(17)(q25),+18,der(19)t(11;19)(q13;p13)add(19)(q13)
77	DLBCL	46,XY,add(2)(q31),add(3)(q12),t(3;8)(p21;q11),der(7)t(7;11)(p22;q13),add(16)(q22)/47,idem,+add(3)
78	DLBCL	48,XY,+3,t(4;7)(q31;p22),del(9)(q22q32),+der(12)t(12;18)(q13;q11)/47,idem,der(10)t(10;11)(p13;q13),-12,/48,idem,del(6)(q21q23),+7,der(22)t(1;22)(q21;q13)

Reference ¹¹¹

84	DLBCL	49-56,XY,+1,del(3)(p12p14),+5,t(14;18)(q32;q21),+i(21)(q10),+1-7mar
85	DLBCL	47,XY,t(12;22)(q13;q11),t(14;18)(q32;q21),+mar/47,idem,+der(1)t(1;4)(p13;q12),-4

Reference ¹¹²

87	DLBCL	39-48,XY,+X,del(6)(q23),+9,der(19)t(2;19)(p14;q13)
88	DLBCL	47-51,X,del(X)(q22-24),+3,der(9)t(8;9)(q13;p21-22),+16,+der(16)t(X;16)(q22-24;q13),+18,der(19)t(5;19)(q14;q13)
89	DLBCL	41-48,XY,+X,del(1)(p12),+der(1)t(1;22)(q23;q12),+5,der(10)t(10;14)(q24;q32),+12,+13
90	DLBCL	39-48,XY,+X,del(6)(q23),+9,der(19)t(2;19)(p14;q13)
91	DLBCL	47-51,X,del(X)(q22-24),+3,der(9)t(8;9)(q13;p21-22),+16,+der(16)t(X;16)(q22-24;q13),+18,der(19)t(5;19)(q14;q13)
92	DLBCL	45-46,X,del(X)(q26),del(1),del(5),der(7)t(1;7)(q32;q36),der(17)t(5;17)(p13;p13),der(22)t(1;22)(q22-25;q13)
93	DLBCL	37-48,XX,+der(X)t(X;1)(p11;q21),+5,der(6)t(6;19)(q25-27;?),+7,+i(10)(q10),+11,+del(12)(p10)
94	DLBCL	44-47,XY,der(2)t(2;11)(q37;q23),del(3)(p21p23-25),+der(3)t(3;14)(q27;q32),der(9)t(9;14)(p13;q32),-15,del(15)(q15q22),der(17)t(11;17)(q13;q24),+18,-21
95	DLBCL	40-46,XY,der(1)del(1)(p22)del(1)(q21-22),-4,del(6)(q2?1q2?5),-11,del(11)(p13),der(11)t(1;11)(p22;q24)t(11;18)(p15;q21)ins(12;1)(q15;q44q25-32),-19
96	DLBCL	44-49,X,-X,-1,-2,del(2)(p11),+der(3)t(1;3)(p21-22;q21),-4,der(6)del(6)(p12)del(6)(q13),+der(6)t(1;6)(q21;q27),+der(8)t(8;12)(p11;q15)t(8;12)(q13;q13),+der(9)t(9;14)(q34;q22),del(12)(q13),+13,+16,i(17)(q10),+18,+18,-22
97	DLBCL	82-95,XXYY,-1,del(3)(p21p23-25),der(3)t(3;7)(p21;q22),+der(3)t(3;13)(?p21;?q22),-4
98	DLBCL	47-50,XY,+del(3)(?p13),+del(3)(p21p23-25),der(14)t(1;14)(p11-13;q32)+ider(18)(q10)del(18)(q21)

Reference ¹¹³

99	DLBCL	46-47,X,-Y,del(1)(q41),ins(1;?)(q21;?),+?3,add(6)(q13),+7,add(14)(q32),-19,+r
100	DLBCL	48,X,+X,-Y,add(3)(q2?5),del(5)(q14q23),add(6)(q12),add(8)(q21),+12,?del(18)(q12q21),add(22)(q13),+mar
101	DLBCL	101-104,YY,+Y,add(X)(q2?1)x2,+2,+2,add(2)(p1?3)x4,+6,+7,add(7)(q22)x3,-8,-8,+10,+10,+14,-15,+16,+19,+20,+21,+der(?)t(?;2)(?;p12),+der(?)t(?;8)(?;q11)x2
102	DLBCL	57,XY,+X,+5,?del(6)(q1?q2?),+add(7)(p11),dic(9;12)(p2?1;q22),+10,+ins(11;?)(q21;?),+12,add(13)(q22),-14,add(14)(q32),add(18)(q21),+20,+21,+der(?)t(?;1)(?;q12),+r,+3mar
103	DLBCL	75-99,XXYY,+Y,der(1)t(1;11)(q32;q13)x2,-2,-2,-5,-5,add(6)(q?15),add(6)(q?21),add(7)(q22)x2,i(7)(q10),-8,add(8)(q24)x2,-9,-9,-9,-9,-10,-11,-11,del(11)(q13q21),-12,-13,-14,-15,-16,-16,-17,-17,-17,-17,dup(18)(q12q23)x2,-19,-19,-20,-22,+der(?)t(?;5)(?;q13)x2,+der(?)t(?;17)(?;q21)x2,+13mar
104	DLBCL	45-49,X,-Y,del(2)(p11p13),+add(3)(q2?6),del(4)(q2?2),del(5)(q14q23),del(6)(p24),del(6)(q15q16),+add(7)(q31),del(9)(p21p22),add(10)(p1?2),dup(11)(q25q22),r(12)(p13q24),i(17)(p10),i(17)(q10),add(19)(p13),+20,+21
105	DLBCL	45-50,dup(X)(p21p22),-Y,del(6)(q2?3),del(6)(q15),+8,t(11;14)(p11;q11),+12,+13,del(14)(q31),+18
106	DLBCL	86-89,XXYY,+X,add(1)(p11),del(2)(p11),-3,-4,del(6)(q15q21)x3,+7,t(7;19)(q11;q13),+8,-10,-11,-12,-13,-13,-15,-15,-16,-17,-17,-18,-19,add(19)(p13),der(19)t(7;19)(q11;q13),-20,-22,+2mar
107	DLBCL	92-115,XXYY,+X,add(3)(q27-29),-6,-6,-8,-8,-9,-10,-10,+15,-16,-17,-19,-19,-22,+13-23mar

Reference ¹¹⁴

108	BL	46,XX,del(6)(q14q22),t(8;14)(q24;q32)
109	BL	46,XY,t(8;14)(q24;q32)
110	BL	47,XY,t(2;8)(p11;q24),+7/47,XY,t(2;8),+16/48,XY,t(2;8),+7,+16
111	BL	46,XX,dup(1)(p34p36),dup(1)(q12q32),der(14)t(8;14)(q24;q32)
112	BL	46,XX,t(8;14)(q24;q32),der(13)t(1;13)(q21;q32)dup(13)(q31q32)/46,idem,dup(1)(q21q25)
113	BL	46,XX,t(8;14)(q24;q32)
114	BL	45,XY,t(1;11)(p36;q13),del(4)(q13q22),dup(7)(q21q36),t(8;14)(q24;q32),-15,del(17)(p11)
115	BL	46,XY,dup(7)(q21q35),t(8;14)(q24;q32)
116	BL	46,XX,t(1;13)(q21;q22),t(8;14)(q24;q32),inv(9)(p11q12)c/46,XX,t(8;14),inv(9)c,t(13;21)(q13;q22)/46,XX,t(6;16)(q15;q24),t(8;14),inv(9)c/46,XX,der(4)t(4;7)(p11;q11),t(8;14),inv(9)c,der(13)dup(13)(q22q11)del(13)(q22)/46,XX,add(2)(p14),t(8;14),inv(9)c,add(13)(q32)
117	BL	46,XY,t(8;14)(q24;q32),dup(13)(q31q34)
118	BL	47,XY,+X,t(8;14)(q24;q32)/48,idem,+7/48,idem,+1,der(1;2)(q10;q10)/52,idem,+i(1)(q10),+4,+6,+11/48,idem,+der(1;7)(q10;p10)
119	BL	47,XY,t(8;22)(q24;q11),+19
120	BL	42-44,X,dic(X;?)(q28;?),add(4)(p16),-6,t(8;14)(q24;q32),der(9)t(9;11)(p13;q13),dup(9)(q21q34),del(11)(q12),del(15)(q12),add(17)(p11),dup(18)(q12q23),-19,-22,+2mar
121	BL	46,XY,t(8;14)(q24;q32)
122	BL	46,XY,t(8;14)(q24;q32)
123	BL	46,XY,t(8;14)(q24;q32)
124	BL	46,XY,t(8;14)(q24;q32)/46,idem,dup(1)(q21q32)/46,idem,-4,+mar
125	BL	46,XY,t(8;14)(q24;q32)
126	BL	46,XY,t(8;14)(q24;q32)
127	BL	46,XX,del(6)(q15q21),t(8;14)(q24;q32),+10,der(13)t(1;13)(q21;q34)
128	BL	46,XY,del(6)(q12q25),t(8;14)(q24;q32),del(11)(q14),del(17)(p11)/46,XY,t(8;14),dup(13)(q31q33),del(17)/46,XY,t(8;14),t(9;11)(q12;q23),del(17)
129	BL	46,XY,dup(1)(q12q25),t(8;14)(q24;q32),add(17)(p13)
130	BL	46,X,-Y,der(6)t(6;10)(p21;q24),add(8)(q24),del(10)(q24),+13,ins(13;12)(q14;?)x2,der(14)t(8;14)(q24;q32),der(14)t(14;14)(q32;q?),t(15;16)(q15;q24),del(18)(q?),der(20)t(6;20)(p21;q13)/46,idem,der(8)t(8;13)(p23;q21)ins(8;12)(p23;?)
131	BL	46,XY,t(8;14)(q24;q32)
132	BL	47,XY,dup(1)(q21q42),?del(6)(q16q25),+7,t(8;14)(q24;q32),add(11)(q25)
133	BL	46,XY,t(8;14)(q24;q32),add(13)(q22)
134	BL	48,XX,+7,der(8)t(8;14)(q24;q32),+12,del(14)(q24),der(14)t(14;14)(q24;q32)t(8;14)/48,idem,-X,+r
135	BL	46,XY,t(8;14)(q24;q32)
136	BL	46,XY,dup(1)(p12p36),t(8;14)(q24;q32)
137	BL	46,XY,del(2)(p16p25),dup(12)(p12p13),t(8;14)(q24;q32),del(13)(q33),dup(13)(q31q33)
138	BL	46,XY,dup(1)(q12q31),t(8;14)(q24;q32)
139	BL	46,XY,del(6)(q21q23),t(8;14)(q24;q32)

Reference ¹¹⁵

140	BL	46,XY,t(1;22)(q11;q11),t(8;14)(q24;q32)/46,XY,dup(1)(q11q44),t(8;14)(q24;q32)
141	BL	46,XY,der(8)t(1;8)(q23;p21)t(8;14)(q24;q32),der(14)t(14;21)(q11;q22),der(21)t(8;14;21)(q24;q11q32;q22)
142	BL	46,XX,t(8;14)(q24;q32)/46,idem,del(2)(p11),ins(5;2)(p14;p11p?)
143	BL	46,XY,dup(1)(q31q44),add(6)(p25),t(8;14)(q24;q32)
144	BL	46,XY,t(8;14)(q24;q32),der(11)inv(11)(q21q23)hsr(11)(q23),der(16)hsr(16)(q?)t(7;16)(q22;q24)/46,idem,dup(1)(q21q25),-der(16)
145	BL	46,XY,der(2)t(2;14)(p12;q32),+7,dup(11)(q13q24),der(14)t(2;14)(p12;q32),der(14)t(1;14)(q12 or p11;p11)t(8;14)(q24;q32),-21,der(22)t(21;22)(q11;p11)/47,idem,+X,der(18)t(1;18)(q21;q2?2)

146	BL	46,XY,+i(7)(p10),t(8;14)(q24;q32),der(13)t(1;13)(q22;q32)
147	BL	46,XY,dup(1)(q21q41),del(2)(q22-24q37),t(8;14;20)(q24;q32;q11)/46,idem,-dup(1),der(1)t(1;4)(q?32;q?)/46,idem,-dup(1),der(Y)t(Y;1)(q12;?p3?)
148	BL	46,XY,dup(1)(q12q31),t(8;14)(q24;q32)/46,idem,der(13)t(7;13)(q21;q34)
149	BL	46,XX,der(6)t(6;7)(p23;?q22),t(8;14)(q24;q32)/46,XX,t(8;14),der(18)t(7;18)(q21;q22)
150	BL	46,Y,t(X;3)(q2?;p21),t(5;12)(q22;q12),der(8)t(8;14)(q24;q32),der(9)t(1;9)(?;p2?),del(13)(q12),der(14)t(13;14)(?q21;p11)t(8;14)(q24;q32)
151	BL	46,XY,der(1)t(1;1;6)(?q31;?q12;q?21),der(1)t(1;7)(q44;?),del(3)(q21),der(5)t(1;5)(?;q35),del(6)(q2?1),der(6)t(6;7)(q13-21;?),t(8;14)(q24;q32),der(13)t(1;13;9)(?;?q34;?)
152	BL	47,X,t(X;14)(p11;q32),del(6)(q13q24),inv(6)(p23q13),t(8;22)(q24;q11),+9
153	BL	47,Y,t(X;1)(p22;q21),trp(1)(q21q32),t(8;14)(q24;q32),der(19)t(2;19)
154	BL	54,XY,+X,+Y,del(1)(q11),+7,t(8;14)(q24;q32),del(9)(q11),+?der(9)t(9;11)(q12;?),del(13)(q?),+del(13)(q?),+18,der(21)t(1;21)(?;p11),der(22)t(1;22)(q?;q11),+der(22)t(1;22),+2mar/54,XY,+X,+Y,dup(1)(q?),+7,t(8;14),del(13)(q?),+del(13)(q?),+18,+der(19)t(1;19)(?;q10),der(21),+mar
155	BL	46,XY,del(1)(q32),t(8;14)(q24;q32),der(11)t(1;11)(q32;q2?3),dup(13)(q2?q?)
156	BL	48,XY,+r(X),+r(1)(p3?2;q12),del(6)(q16q23),der(7)?inv(7)(q11q36)?trp(7)(p22?q36),t(8;14)(q24;q32)
157	BL	47,XX,+der(1)t(1;16)(p13;?p11),del(6)(q15q24),t(8;22)(q24;q11)
158	BL	46,XY,dup(1)(q12q25),t(8;14)(q24;q32),der(13)t(13;16)(q33;?)
159	BL	46,XY,dup(1)(q12q31),del(4)(q26),dup(6)(q23q26),t(8;14)(q24;q32),der(13)dup(13)(q?q?)t(4;13)(q2?6;q?)/46,idem,der(15)t(1;15)(q1?2;p1?),der(11)t(1;11)(q?;q2?3)
160	BL	46,X,der(Y)t(X;Y)(q21;q12),der(1)t(1;9)(p34;p24)t(1;8)(q21;q24),t(2;6)(p1?1;q14-16),der(6)del(6)(q22q25)t(6;14)(p2?q32),del(8)(q24) or add(8)(q24),der(9)t(1;9)(p34;p24),+12,der(14)t(1;14)(q21;q32),t(14;19)(q32;?p13)/46,idem,t(12;17)(q24;q21)
161	BL	46,XY,t(8;22)(q24;q11),del(9)(q11q22),t(10;14)(q22;q32)/45,idem,-Y
162	BL	46,X,der(Y)t(Y;1)(q12;q11),r(6),t(8;14)(q24;q32)

Institutional Cases (cont.)

163	BL	49,XX,add(1)(q25),t(8;14)(q24.1;q32),add(10)(q22),add(17)(p11.2),+22,+mar1,+mar2 [2 cells]/49,XX,add(1)(q25),der(1)t(1;1)(p13;q21)ins(1;?)(p13;?),t(8;14)(q24.1;q32),add(10)(q22),der(13)t(1;13)(p13;q32),add(17)(p11.2),+22,+mar1,+mar2 [3 cells]46,XX [15 cells]
164	BL	46,XY,del(2)(p21p23),t(8;14)(q24.1;q32)[3]/47,idem,-del(2)(p21p23),+7[1]/47,idem,-del(2)(p21p23),+der(7)del(7)(p13p15)inv(7)(p22q11.2)[15]/46,XY[1]
165	BL	46,XY,add(3)(p11.2),t(8;14)(q24.1;q32),der(13)t(3;13)(p21;q32) [12 cells]/46,idem,del(13)(q12q14) [3 cells]
166	BL	46,XY,t(8;14)(q24.1;q32)[20]
167	BL	46,XX,t(8;14)(q24.1;q32)[cp5]
168	BL	46~47,XY,t(8;14)(q24.1;q32)[cp5]/47,idem,+add(1)(p22)[6]/46,idem,dup(1)(q12q32)[3]/46,idem,der(18)t(1;18)(q21;q23)[3]
169	BL	46,XY,t(8;14)(q24.1;q32)[15]

Reference ¹¹⁶

170	DLBCL	51,XX,+3,del(6)(q13q21),+7,del(9)(p22p24),del(9)(q13q32),+12,+i(18)(q10)+add(22)(p11)
-----	-------	---

Reference ³⁴

171	DLBCL	45,XY,del(1)(p32),-2,-5,-8,-9,t(10;11)(q24;p15),add(12)(q24),-17,del(18)(q21),+4mar
172	DLBCL	84,XXX,-X,add(1)(p13)x2,-2,-3,del(6)(q13)x2,-10,-10,+12,-13,-13,-14,-14,add(14)(q32),-15,-17,-17,del(18)(q21)x2,-19,-20,-22,+6mar
173	DLBCL	45,XX,-1,dic(1;11)(q11;q23),-3,add(6)(q13),+der(9)t(9;11)(q34;q13),der(11),-14,add(19)(q13),der(21)t(14;21)(q11;p13),+mar
174	DLBCL	89,XX,del(X)(q22)x2,-1,-4,-5,-6,-6,add(17)(p11),-22,+3mar
175	DLBCL	89,XXXX,add(2)(q10),add(3)(q27),-6,-8,-9,+10,-11-19,-21,-22,+3mar/88,idem,-4/88,idem,-9
176	DLBCL	48,XX,add(4)(q21),+9,+11
177	DLBCL	49,XY,t(1;7)(p13;q32),+add(3)(q11),+9,add(10)(q22),+18,add(19)(q13)

Reference ¹¹⁷

178	DLBCL	46,XX,del(2)(p?),add(6)(q?),del(6)(q?),del(7)(q?),add(9)(p?),add(11)(q?),add(12)(p?),del(13)(q?),der(14)t(6;14)(p21;q32),del(20)(p?)
179	DLBCL	49,XY,dup(1)(q21q32),add(2)(p13),+3,-6,t(6;14)(p21;q32),+der(8)del(8)(p21)t(8;21)(q22;?),add(9)(p24),add(10)(p11),del(11)(q13),+12,+18,-19,del(19)(p13),der(21)t(8;21),+mar
180	DLBCL	48,XX,+18,+18,-19,+r
181	DLBCL	??,X?,add(6)(q?),add(14)(q?)
182	DLBCL	50,XX,+X,t(12;22)(p13;q11),+16,+18,+19
183	DLBCL	44,X,-Y,add(2)(q32),del(4)(q?),der(14)t(8;14)(q24;q32),-15,add(17)(p11),i(21)(q10)

Reference ¹¹⁸

184	DLBCL	49,X,-Y,+X,+2,del(3)(q13q21),add(4)(p16),add(7)(p22),inv(8)(p21q11)
185	DLBCL	47,XY,-2,inv(3)(q?),del(5)(q?),-6,add(6)(p?),add(7)(p?),del(8)(p?),del(11)(q?),add(12)(q?),-21,+r,+3mar
186	DLBCL	44-47,X,-Y,add(1)(p21),t(2;9)(q11;p24),t(3;6)(q27;p12),add(5)(p15),-7,-8,+9,del(11)(p14),add(17)(p13),del(17)(q23),+add(18)(q23),+21,-der(?)t(?;1)(?;p21)
187	DLBCL	48,X,-Y,add(1)(p34),der(1)(q25q44),del(8)(p22),inv(10)(q22q24),dup(14)(q24q32),+18,+i(18)(q10),add(19)(p13),+mar
188	DLBCL	47,XX,del(3)(q21),add(7)(p22),del(8)(p22),add(9)(p24),inv(14)(?q11q32),+18,-18,add(19)(p13)/47,idem,add(17)(p13)

Reference¹¹⁰

189	DLBCL	47,XX,+X,t(1;11)(p36;q13),der(6)(p11p21)add(6)(q21),i(18)(q10)
190	DLBCL	47,XX,add(6)(q15),der(8)t(8;8)(p21;q13),der(15)t(15;21)(q26;q11),add(19)(q13),+mar
191	DLBCL	49,XY,add(1)(p13),del(1)(p13),+der(1)t(1;1)(q23;p13),+der(2;11)(q10;q10),inv(4)(p15q21),+add(6)(q15),del(8)(p21),+9,t(13;18)(q22;q23),-15,der(22)t(7;22)(q11;p11)/95,idem,-2
192	DLBCL	89,XXXX,-1,-2,-2,add(2)(p13),+3,-4,add(4)(p12),-5,-6,del(7)(q32),der(7)t(3;7)(p13;q32),der(8)t(8;8)(p21;q13)x2,+11,der(12)t(12;21)(q24;q11)x2,add(13)(q32),+14,-15,-16,-17,+18,+18,-19,-22,+3mar
193	DLBCL	45,XY,der(2)t(2;8)(p23;q13),add(8)(p21),add(9)(p22),t(9;14)(p13;q32),del(10)(p11),-17/45,idem,i(21)(q10)/89,idemx2
194	DLBCL	45,XX,del(1)(p13p22),der(3;17)t(3;17)(q27;p11)del(3)(p21p23),-4,add(8)(p21),der(9)t(4;9)(p12;p24),der(11)t(4;11)(p12;q23),der(11)del(11)(q23)inv(11)(q11q23),add(14)(p11),t(15;22)(p11;q11),-17,+20
195	DLBCL	49,XY,del(1)(p34),t(3;14)(q27;q32),t(3;6;7)(p24;p25;p15),der(6)t(1;6)(p34;q21),add(8)(p21),dup(8)(p23p21),+10,+11,+21
196	DLBCL	49,XX,+X,dup(1)(q31q42),+der(3)t(3;3)(p13;q12),add(7)(q22),del(8)(p21),add(17)(q25),add(19)(p13)
197	DLBCL	46,X,-Y,t(3;3)(q11;q27),+6,add(6)(q21x2,r(18)/46,idem,add(7)(p22),add(12)(p13)
198	DLBCL	86,XX,-Y,-Y,-1,dup(1)(q21q25)x2,t(3;11)(q13;q13),-4,t(4;9)(q31;q22)x2,-8,-14,-15,t(17;18)(p11;p11)x2,+18,der(19)t(19;22)(p13;q11)t(6;19)(p21;q13),der(22)t(19;22)(p13;q11)/84,idem,-6,-10
199	DLBCL	46,X,-X,+18,der(19)r(19;?)p13q13;/46,idem,t(3;4)(p25;q31)/46,idem,der(4)t(3;4)/47,idem,der(4)t(3;4),+18
200	DLBCL	52,XX,+3,+12,+16,+18,+18,-19,+r/51,idem,-21

Reference¹¹⁵

201	DLBCL	46,XY,der(1)t(1;2)(p13;?),del(2)(q12q31),der(2)t(1;2)(p11;p11)ins(2;1)(q3?2;q?),der(8)t(8;14)(q24;q32),del(9)(q12),der(13)t(3;13)(q24q25;p11)trp(13)(q31q34),der(14)t(2;1;14;8)(?;?;q32;q24),der(16)dup(16)(q11q22)t(9;16)(q22;q22),del(17),der(21)t(17;21)(p11;p11)
202	DLBCL	X,der(X)t(X;4)(p21;q21),der(X)t(X;10;4)(p21;?;q12),der(3)t(3;18)(q11;q21),dup(3)(q13q24),del(4)(q12),der(4)t(X;4)(?;p16),t(8;22)(q24;q11),der(9)t(3;9)(p13;p21),t(9;20)(q10;p10),der(10)t(3;10)(?;q11),dup(11)(q23q13),der(11)dup(11)(q23q13)t(1;11)(?;q25),der(11)dup(11)(q23q13)t(3;11)(?;q25),der(11)dup(11)(q23q13)t(7;11)(?;q25),dup(11)(q23q13),trp(13)(q22q32),der(14)t(X;14)(q13;q24),der(15)t(1;15)(q11;p11),der(16)t(1;16)(p2?;q1?) ,der(17)t(X;17)(q21;p11),der(17)t(X;17;6)(p11;p11q25;?),der(18)t(4;18)(?;q23),der(18)t(6;18)(?;q23),dup(18)(q11q23),der(22)t(16;22)(?;q1?3)
203	DLBCL	46,XY,-3,+7,t(8;14)(q24;q32),der(17)t(3;17)(q23;p11),hsr(18)(q?)
204	DLBCL	46,XY,dup(7)(q35q11),t(8;14)(q24;q32)/46,XY,der(7)ins(7;7)(p21;q36q21),t(8;14)(q24;q32)
205	DLBCL	45,X,-Y,i(1)(q10),+del(5)(q2?),-6,t(8;14)(q24;q32),der(15)t(7;15)(?;p11),der(21)t(7;21)(?;q22),der(22)t(6;22)(?;q13)/45,idem,der(17)t(17;20)(p11;?q11)/45,idem,del(17)(q2?1),der(17)t(17;20)
206	DLBCL	45,X,-Y,i(1)(q10),+del(5)(q2?),-6,t(8;14)(q24;q32),der(15)t(7;15)(?;p11),der(21)t(7;21)(?;q22),der(22)t(6;22)(?;q13)/45,idem,der(17)t(17;20)(p11;?q11)/45,idem,del(17)(q2?1),der(17)t(17;20)
207	DLBCL	46,XY,t(8;14)(q24;q32),del(13)(q22q31)/47,idem,+20
208	DLBCL	40-53,XY,+X,-Y,del(Y)(q11),der(1)?dup(1)(q?)t(1;18)(q;q?12),-3,-4,der(4)t(4;18)(q1?3;?),der(5)t(3;5)(?p21;q3?1),?del(6)(q?),der(6)t(4;6)(?;p21),der(7)?t(7;10;22),t(8;14)(q24;q32),der(8)t(8;18)(p12-21;q12-21),-10,del(11)(q14),der(11)t(11;13)(q?14;q?21),der(11)t(11;17)(q13;?p11),der(12)t(1;12)(?;p13),ider(13)del(13)(q?14),der(13)dup(13)(?q21q34)t(13;18)(q34;?),der(13)t(7;13)(?q32;q34),der(14)t(7;14)(?q11;p?11),-15,-16,der(17)t(3;5)(q21?3;q3?)t(3;15)(q2?9;?)t(15;17)(?;p1?2),+18,+18,der(18)t(4;18)(?;p11),+der(18),i(18)(q10),+19,del(19)(q1?3),der(19)t(7;19)(q22;q13),der(19)t(10;19)(q21;p13 or q13),ins(20;18)(p11;?),-21,-22,der(22)t(Y;22)(q11;p1?)
209	DLBCL	46,XY,dup(7)(q?),t(8;14)(q24;q32),der(14)t(7;14)(?;q3?2)
210	DLBCL	49,XX,+X,der(1)t(1;14)(q;q?32),+7,der(8)t(8;14)(q24;q32)t(1;14)(q?21;q32)t(1;8)(q?42;q24),+13,der(14)t(8;14)t(1;8)(q?32;q1?)del(1)(q?32q?42)t(1;8)(q?42;q24)/48-49,idem,-20,der(22)t(14;22)
211	DLBCL	47,X,-X,trp(1)(q11q21),t(2;10)(q12-14;q21),t(5;9)(q11;q31),+7,+der(7)del(7)(q31q35)t(1;7)(q11;p?21),+der(7)t(7;8;13)(p13;?q24;?),t(8;14)(q24;q32),del(8)(q1?),t(9;10)(q22;p11),del(13)(q14q22),del(17)(p12)
212	DLBCL	85,XXXX,der(1)t(1;4)(p11;q2?1),+der(1)t(1;4),-3,-4,der(5)t(X;5)(?;q2?),del(7)(q22q32)x2,+der(7)t(7;10),-8,der(8)t(6;8),t(8;14)(q24;q32)x2,-9,-10,-14,-15,-15,-16,der(17)t(16;17)(?p10;q10)x2,-18,-18,+r(15)x2
213	DLBCL	51-54,XX,+der(X)t(X;2)(p22;?),+Y,t(1;16)(q21;q12),-3,-4,+der(6)t(1;6)(q21;q13),+7,t(8;14)(q24;q32),-14,t(14;18)

214	DLBCL	(q32;q21),del(15)(q1?),+del(15)(q1?),+der(16)t(1;16),+17,+18,del(18)(q?),+?del(18)(q?),+20,+21,-22
215	DLBCL	52,XY,+X,+3,+7,der(8)t(8;18)(q2?2;q21),t(8;14)(q24;q32),t(9;11)(p13;p13),+der(14)t(8;14),+18,+18
216	DLBCL	48,XX,der(1)t(1;15)(p13;?),+7,t(8;14)(q24;q32),r(11),+13,-15,+18
217	DLBCL	48,XY,+der(1)t(1;2)(p21;p1?),+7,t(8;14)(q24;q32)
218	DLBCL	48,XX,+t(3;18)(p1?3;q11),+der(7)t(6;7)(p21;q36),t(8;14)(q24;q32),del(15)(q15q24)
		45,XY,der(5)t(5;15)(q11;?),del(13)(q31q33),der(13;20)(q10;q10),del(17)(p11p12),der(20;22)(p10;q10)/
		82,YYY,-
		X,der(X)t(X;2;3)(?q21;?;p12),der(X)t(X;3)(p11-21;?),der(1)hsr(1)t(1;1)(p?22;q?21),der(2)t(2;5)(?;q11),
		-3,der(3)t(3;4),-4,der(4)t(2;4;5),?del(5)(q31q33),t(8;22)(q24;q11),der(8)t(8;22)(q24;q11),der(9)t(X;9;9;2)
219	DLBCL	(?q13;?p13;q22;?),-11,-12,-13,-14,-15,-16,-18,-19,+21,-22
220	DLBCL	47,XX,+idic(1)(p12),t(8;14)(q24;q32),del(10)(p14),der(14)t(8;14)
		46,XX,der(2)t(2;6)(p?14-16;q2?1),t(2;8)(p12;q24),t(4;14)(p12;q2?),der(6)t(6;7;?16;9)(q2?1;?;?;?p?),
		der(9)t(2;9)(p?;p2?2)
221	DLBCL	46,XX,t(8;14)(q24;q32)/46,idem,der(1)t(1;2)(q?31;?),der(2)t(2;17),der(3)t(2;3)(?;q2?),der(3)t(3;13)(p2?5;?),
		der(3)t(3;17)(q2?6;p?11),der(4)t(4;13)(q?31;q2?),+5,+del(5)(q22),der(6)t(2;6)(?;q2?),+12,dup(13)(q?),der(17)t(
		3;17)(q2?6;p?11),der(17)t(2;17)(?;q2?)
222	DLBCL	46,XX,der(1)del(1)(p12)dup(1)(q21q31),del(6)(q11),der(6)inv(6)(p21q11)t(6;7)(q11;q32),der(7)t(6;7)
		(q11;q32),der(8)t(1;8)(p3?;q24),t(8;14)(q24;q32),der(14)t(12;14)(q21;q32)/46,idem,der(12)t(6;12)
223	DLBCL	50,XY,+X,-Y,+5,t(8;14)(q24;q32),+12,+19,+20
224	DLBCL	46,XY,t(8;14)(q24;q32),dup(11)(q13q23),der(13)t(1;13)(q21;q34)/46,idem,-der(13),ins(13;1)(q34;?)/45,idem,-
		der(13),ins(13;1),-22
225	DLBCL	44-49,XY,-2,del(7)(q11q21),+ider(7)del(7)(q11q21),t(8;22)(q24;q11),+19,-21/23-26,XY,+5,der(8)t(8;22),+10,-
		15,+18,+21
226	DLBCL	46,XY,dup(1)(q12q32),t(8;14)(q24;q32),del(17)(p11)

Reference ⁵⁶

227	BCLU	46,XX,t(4;7)(q13;q21),t(8;14)(q24.1;q32)[6]/46,XX[14]
228	BCLU	47,XY,del(3)(p12p21),+del(3)(q12q29),t(8;22)(q24;q11.2),del(11)(q22q23),add(14)(q32),-19,+mar[20]
229	BCLU	47,XX,der(1)t(1;8)(p36.3;q22),der(8)t(8;14)(q24.1;q23)t(8;t(14;18))(q24.1;q22),der(14)t(8;14),
		der(18)t(14;18)(q32;q21),+mar[20]
230	BCLU	46,XX,der(1)dic(1;1)(q10;q31)t(1;5)(q21;q33),der(6)t(6;8)(q21;q22.1),del(8)(q13q24.1),
		der(11)dup(11)(q13q23)t(8;11)(q22.1;q25),der(13)t(6;13)(p12;q34),add(15)(p13),der(18)t(1;18)(p36.1;q23)[16]
		/
		46,XX[4]
231	BCLU	47,XY,der(1)t(1;3)(p36.3;p21),t(1;9)(p22;p13),+der(8)t(8;8)(p22;q22),del(9)(p22p24),add(13)(q22),
		der(14)add(14)(p11.2)t(14;18)(q32;q21),inv(15)(p11.2q15),-16,der(18)t(14;18)(q32;q21),+mar[20]

Reference ³⁵

232	BCLU	48,XX,+der(X)t(X;1)(q26;q21),t(2;8)(p12;q24.1),+der(8)t(2;8)(p12;q24.1),der(11)t(11;14)(p15;q32), add(14)(q32)
233	BCLU	48,XX,dup(11)(q13q23),+12,ins(14;8)(q32;q24.1q24.1),+20,ish ins(14;8)(MYC+;IGH+;MYC+;IGH-)[17]/48,idem,
		add(9)(q34)[cp3]
234	BCLU	47,XY,der(3)t(3;14)(q27;q32),der(8)t(8;14)(q24.1;q32)t(3;14)(q27;q32),del(10)(q23q23),der(14)t(8;14)(q24.1;q32),
		+der(17)t(1;17)(q12;p13)[3]/48,idem,+X [9 cells]/50,idem,+X,+7,+add(14)(q32) [2]/50,idem,+X,+add(14)(q32),+20

Key: DLBCL, diffuse large B-cell lymphoma; BL, Burkitt lymphoma; BCLU,

Unclassifiable with features intermediate between DLBCL and BL.

The HC heat map (see Figure 4-3) generated twelve large clusters when the dendrogram was cut at a specified level. Of these, eleven clusters represented DLBCL, while only one was limited to BL (cluster 12). As expected and in terms of genetic complexity, BL in cluster 12 was characterized by a low number of RCAs (1-2 for the most part),

whereas the remaining of the clusters contained a more complex chromosome complement (more than 3 RCAs). The Total number of misclassified cases by this model included 15 cases (DLBCL, 5; BL, 10). In terms of the HC dendrogram with p-values (Figure 4-4) and the K-means partition using the PAM algorithm (data not shown), twelve cases were misclassified by the HC with p-values (DLBCL, 1; BL, 11) and a total of 10 cases (DLBCL, 3; BL, 7) by the PAM algorithm. In terms of the number of clusters generated for each of the latter models, a total of 21 clusters (existence of cluster < .05) were generated by the HC with p-values, while 75 clusters were pre-designated to the PAM algorithm. In terms of the logistic regression model and as previously indicated, it obtained an ability to discriminate between the two histological tumor types of 85%. In addition and based on cytogenetic findings from this current study, six of eight cases of unclassifiable with features intermediate between DLBCL and BL – BCLU (Table 4-5, cases 227-234) were re-classified as DLBCL and two were assigned as BL (case 227 and 230). Case number 227 (46,XX,t(4;7)(q13;q21),t(8;14)(q24.1;q32)[6]) lacks any identifiable RCA associated with DLBCL and contains less than two chromosome alterations (i.e., a simple karyotype) besides the *MYC* translocation [i.e., t(8;14)]. This latter case may be best described as a *MYC* + BL lacking typical BL morphological and phenotypic features (given its classification of BCLU in the literature). In terms of the second re-classified case of BL (46,XX,der(1)dic(1;1)(q10;q31)t(1;5)(q21;q33),der(6)t(6;8)(q21;q22.1),del(8)(q13q24.1),der(11)dup(11)(q13q23)t(8;11)(q22.1;q25),der(13)t(6;13)(p12;q34),add(15)(p13),der(18)t(1;18)(p36.1;q23)[16]), it may seemed more problematic to classify. First, it lacks a detectable *MYC* translocation and contains a more complex karyotype. However and

based on a previous cytogenetic study¹¹⁹ where duplication of chromosome 11 was a frequent finding in *MYC*- BL, we believe this case is best classified as a *MYC*- BL with typical BL morphology and phenotypic features. Moreover, loss of the long arm (q) of chromosome 13, present in this sample's karyotype [der(13)t(6;13)(p12;q34)], has been also identified as recurrent aberration in BL⁵⁷.

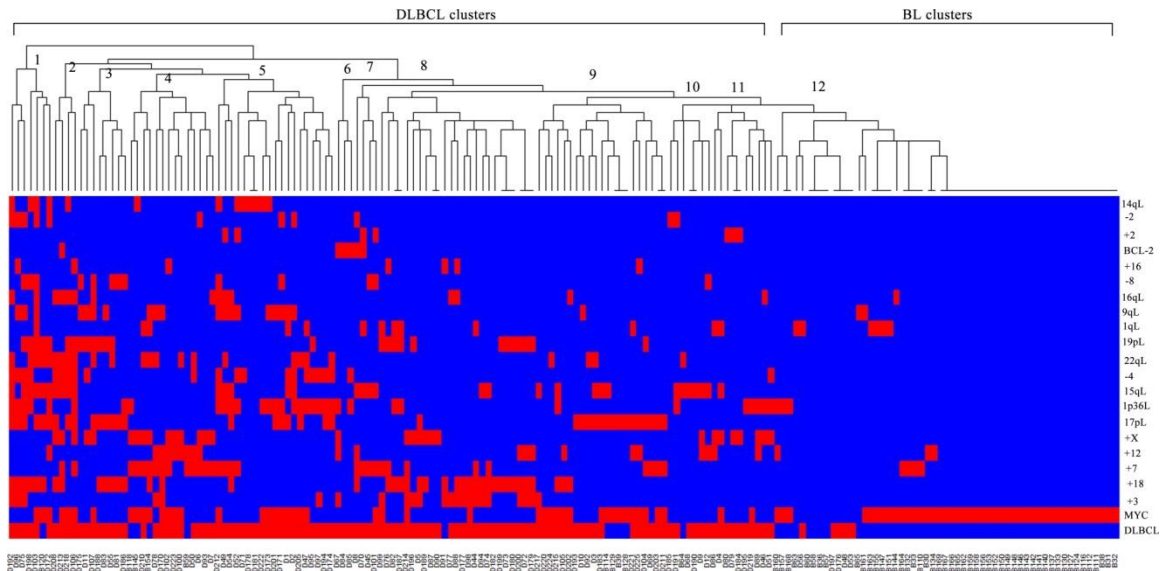


Figure 4-3. Hierarchical cluster heat map. Heat map modeled on 21 RCAs and applied to 117 DLBCL and 60 BL cases. The red and blue staining areas represent presence or absence of individual RCAs for each of the diagnostic tumor samples.

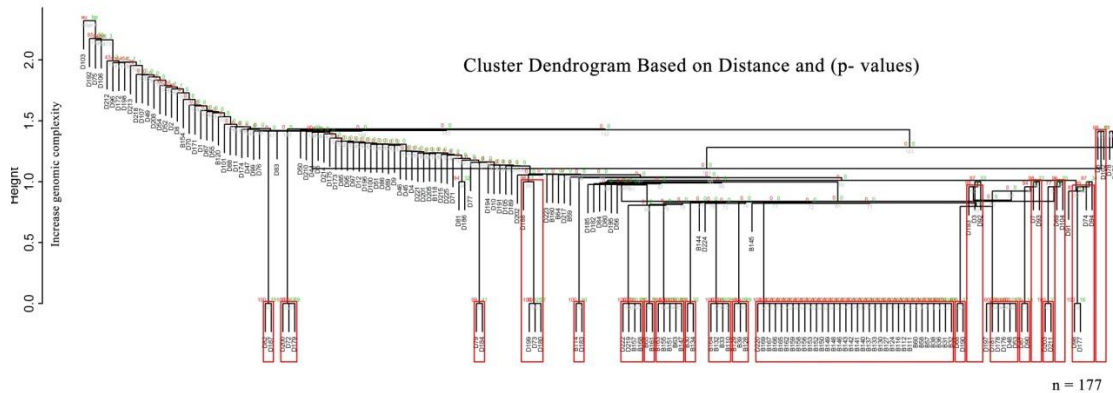


Figure 4-4. Hierarchical cluster with p-values applied to 177 cases and based on 21 RCAs. Existence of a cluster is ($p < .05$) drawn in a red rectangle. Cases placed at a higher height distance reflect an increase in genomic complexity based on the number of RCAs present.

With regards to the supervised model (ANNs), a total of seven cases were misclassified by this model (DLBCL, 6; BL, 1) (see Figure 4-5). The learning and training groups as indicated previously were 100 and 77 (DLBCL, 50; BL, 27) each. A score index less than 0.2 was considered a diagnosis of BL, this based on the threshold level or optimal cut point between DLBCL and BL obtained from the ROC curve (see Figure 4-6). ANN models containing more than one hidden layer were also constructed; however, results were comparable to the one hidden layer ANN model (data not shown).

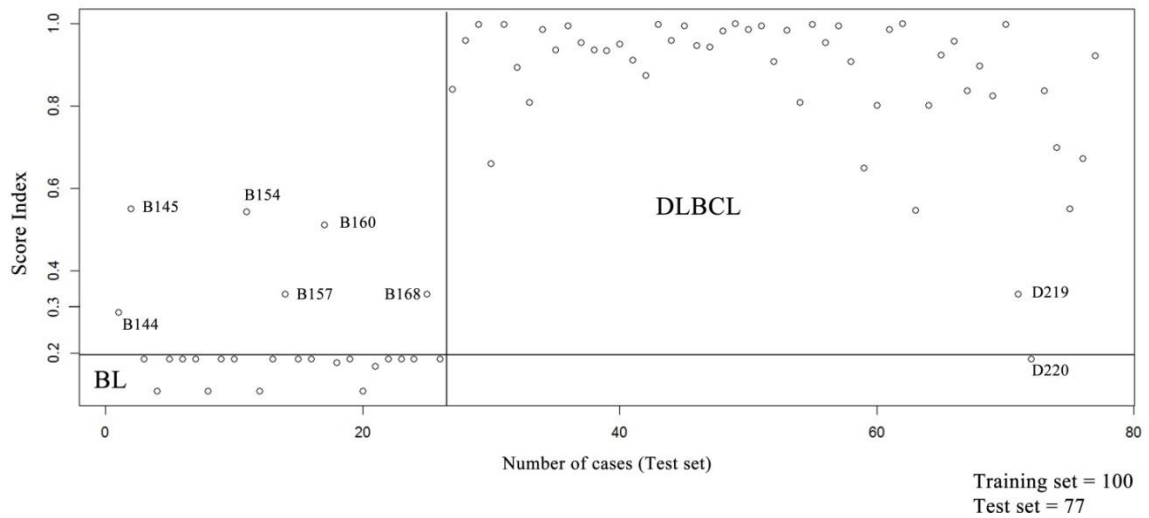


Figure 4-5. ANN classification of the test dataset. Classification of 77 cases by the ANN model based on 21 RCAs (1-26, BL; 27-77, DLBCL) by an artificial neural network containing one hidden layer. A score index closer to 1 is more certain of DLBCL. The threshold limit between DLBCL and BL was set at 0.2, this to allow for the small number of BL cases with a higher genomic complexity (molecular BL subtype).

4.7 Discrimination Analysis: Area under the Curve and Specificity

To assess the discrimination ability or performance measure of each of the models, in terms of specificity, sensitivity and the area under the curve - AUC, an ROC curve was constructed. Findings from this analysis showed the k-means partition using the PAM algorithm with the highest AUC (0.93). However, it was the ANN model with the highest specificity at 92.3 (Figure 4-6). When comparing cluster models, including the ANN model, with a logistic regression, all cluster models outperformed the logistic regression (0.9 or higher vs. 0.85).

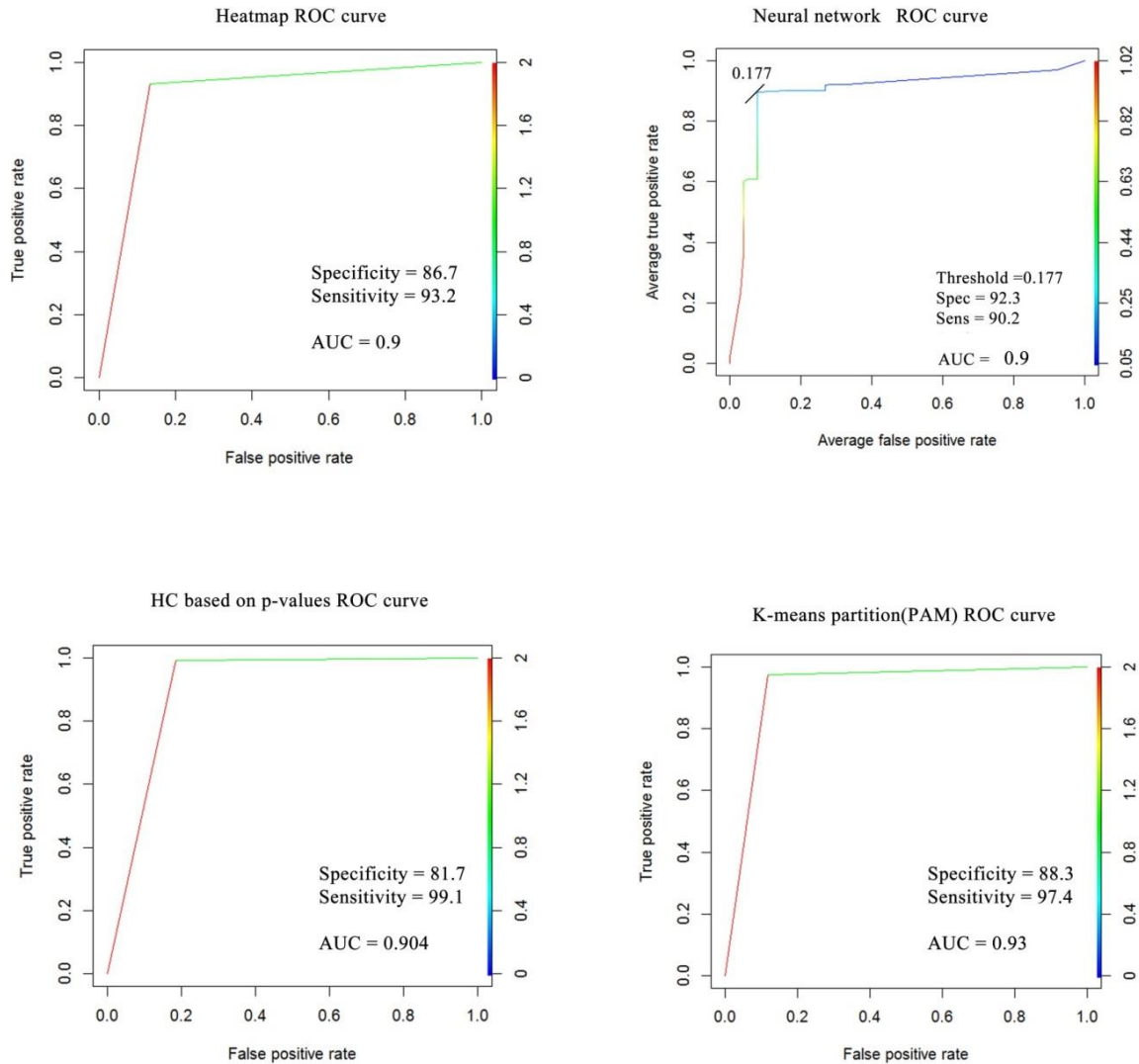


Figure 4-6. Receiver operating characteristic (ROC) curve of prediction models. Performance of each of the predictor models was assessed with the ROC curve. Each model was analyzed with 177 sample cases and 21 RCAs.

4.8 A Core Set of RCAs (Primary RCAs)

To further evaluate RCAs, a closer inspection of all RCAs applied to the four predictive models was further explored with a HC p-value (Figure 16), PAM algorithm and logistic regression. The goal in this instance was to develop a primary set of RCAs

that are more certain of DLBCL. Since *BCL-2*, our internal control, resides closer to DLBCL than BL, those RCAs clustered around *BCL-2* suggest a DLBCL phenotype. In the case of a HC p-value dendrogram of RCAs, +2 and +16 clustered around *BCL-2*, while -8, 14qL and -2 were found in nearby cluster in both the HC p-value dendrogram and PAM algorithm. Therefore, these RCAs seem to suggest closeness to DLBCL. Indeed, all but one (14qL) predicted DLBCL with 100% accuracy in the test dataset of 177 cases. Moreover, a closer inspection of the larger clusters (cluster 1 and cluster 2 vs. cluster 3, $p < .0001$) of the dendrogram showed these set of RCAs belonging to the GC-DLBCL molecular subtype (these RCAs clearly clustered around characteristic GC-DLBCL markers, *BCL-2* and +12). Likewise, RCAs that clustered around +18, a well-established marker for the ABC-DLBCL molecular subtype, were indicative of a DLBCL phenotype. In this regard, +3, 15qL, 1p36L and 17pL (all found in cluster 3 and independent predictors of DLBCL by the logistic regression model) suggest an ABC-DLBCL molecular subtype. Thus for a primary set of RCAs that are highly predictive of DLBCL, we propose the following: +2, +16, -8, 14qL, -2, +3, +18, 15qL, 1p36L and 17pL. Of note, since 1qL along with 16qL and 9qL are considered close to DLBCL, but seem to belong to a different cluster from *BCL-2* (Figure 4-7), these RCAs are here considered as secondary.

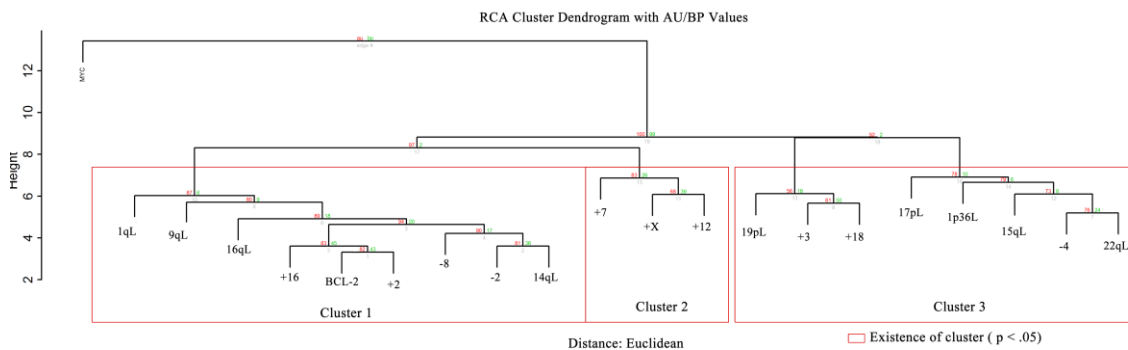


Figure 4-7. Hierarchical cluster dendrogram with p-values of RCAs. Dendrogram illustrates two larger clusters of RCAs (cluster 1 + cluster 2 vs. cluster 3). First two clusters are characteristic of GC-DLBCL and cluster 3 of ABC-DLBCL.

4.9 Additional RCAs Identified DLBCL and BL

Moreover, when combining the two datasets together (n = 515), that is the testing dataset used to evaluate predictor models, combined with the initial dataset to identify RCAs, additional chromosome aberrations (n = 8) were identified. In this analysis, gain of the short arm of chromosome 6 (6p Gain), 6qL, +5, +11 and concurrent losses of chromosome 10 and 15 (-10/-15), as well as -10/-14 were found associated with DLBCL. On the other hand, 13q loss and gain of the long arm of chromosome 1 were more prevalent in BL (Table 4-6). Of note, an additional internal control (3q27 rearrangements exclusively present in DLBCL) was also added to the list of reliable RCAs. Therefore, a total of 31 RCAs were identified as unique markers that may be use to distinguish between the two entities. To further explore the relationship between RCAs, we constructed a correlational matrix of all RCAs (Figure 4-8).

Table 4-6. Additional RCAs Associated with DLBCL (*MYC*⁺ and *MYC*⁻) and BL

RCA	Morphology	p-value
6p gain	DLBCL	.03
6qL	DLBCL	.0001
-10	DLBCL	.02
-10/-14	DLBCL	.04
-10/-15	DLBCL	.005
1qG	BL	.0001

+11
+5
13qL

DLBCL
DLBCL
BL

.047
.02
.0006

Key: L, loss; G, gain

	DLBCL	6qL	-10	-10/-14	1qG	3q27	13qL	6pG	+5	+11	-10/-15	+7	16qL	+X	1p36L	1qL	+3	9qL	+12	15qL	17pL	+18	19pL	14qL	+2	-2	-4	-8	22qL	BCL-2	MYC	+16			
DLBCL	1.0000	0.1883	0.1059	0.0402	0.2849	0.3119	-0.1024	0.0999	0.1089	0.2685	0.1247	0.1188	0.1402	0.1125	0.1276	0.0268	0.1838	0.1383	0.1287	0.1776	0.1809	0.2038	0.1480	0.1717	0.1293	0.1548	0.1772	0.1876	0.1687	0.2070	-0.0920	0.0540			
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
6qL	0.1883	1.0000	0.1746	0.1042	0.3077	0.3307	-0.1003	-0.0254	-0.2371	0.2727	0.1547	0.2192	0.1863	0.2059	0.1057	0.0279	-0.0122	0.2814	0.2412	0.1611	0.1767	-0.0049	0.1048	0.1150	-0.0547	0.2339	0.1641	0.1102	0.1929	-0.0363	-0.1100	0.0306			
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
-10	0.1059	0.1746	1.0000	0.8381	0.2051	0.1010	0.1488	-0.0238	0.0182	0.3120	0.8648	0.3841	0.2857	0.1975	0.2482	0.0879	-0.0234	0.3874	-0.0351	0.2777	0.1848	-0.0038	0.1810	0.2030	0.0818	0.3824	0.2800	0.2849	0.2370	0.0448	-0.1281	0.0143			
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
-10/-14	0.0402	0.1042	0.8381	1.0000	0.0347	-0.0452	0.0787	-0.0434	0.0114	-0.2412	0.8472	0.3749	0.3068	-0.0214	0.2774	0.0009	-0.0201	0.1847	0.2072	0.2324	0.1478	-0.0811	0.2247	0.3911	-0.0344	0.1702	0.1218	0.1816	-0.0243	-0.1162	0.0300				
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
1qG	-0.2849	-0.3077	-0.2051	-0.0347	1.0000	-0.0710	0.0697	-0.0440	-0.0697	-0.0041	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0048	-0.0738	0.0387	-0.1892	0.0362				
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
3q27	0.3119	0.3307	0.1010	0.0347	-0.0710	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362				
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
13qL	-0.1024	-0.1003	-0.1488	-0.0238	0.0697	-0.0732	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362			
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
6pG	0.2685	0.2727	0.3120	0.3749	0.0142	0.2338	0.0772	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+5	0.1247	-0.2371	0.2727	0.1547	-0.0041	0.2192	-0.0142	0.0697	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+11	0.2685	0.2727	0.3120	0.3749	0.0142	0.2338	0.0772	0.0142	0.2338	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
-10/-15	0.1247	-0.2371	0.2727	0.1547	-0.0041	0.2192	-0.0142	0.0697	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+7	0.1863	0.2059	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
16qL	0.1402	0.1863	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+X	0.1125	0.2059	0.2727	0.3314	0.0347	0.2059	0.0772	0.0142	0.2338	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
1p36L	0.2075	0.1987	0.2402	0.2075	0.0210	0.2749	0.0210	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749	0.0210	0.2749		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
1qL	0.0268	0.2059	0.2727	0.3314	0.0347	0.2059	0.0772	0.0142	0.2338	1.0000	-0.0440	0.0697	-0.0732	0.0142	0.2338	0.0772	0.0142	0.0696	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+3	0.1863	0.2059	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
9qL	0.1863	0.2059	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
+12	0.1863	0.2059	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92	92			
15qL	0.1863	0.2059	0.2857	0.3068	0.0201	0.2774	-0.0214	0.0697	0.0621	0.0871	-0.0647	-0.0431	0.0472	0.1474	0.0268	-0.0802	-0.1041	0.0241	-0.0268	-0.0048	-0.0738	0.0387	0.0387	-0.1892	0.0362	-0.1162	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300		
	92	92	92	92	92																														

morphological entities: +18 and 19p loss (19pL has a strong correlation with -8 and 1p36L, markers of ABC-DLBCL in this study) for the ABC-DLBCL, *BCL2* and +12 for the GC-DLBCL molecular subtype and 1qG for BL.

Table 4-7. Classification of RCAs based on Subtypes of DLBCL and BL

RCA	Associated RCAs	p-value
ABC-DLBCL		
+18	+3	.0001
	+7	.0005
	19pL	.009
	9qL	.04
19pL	6qL	.0006
	-8	.0001
	16qL	.0001
	1p36L	.0002
	9qL	.01
	15qL	.0001
	14qL	.0001
	-4	.0001
	17pL	.0001
	-10/-15	.0001
GCB-DLBCL		
BCL-2	+12, +11	.0001
	+7, +2	.0008
	+5	.004
	+X	.007

	-4	.0009
	-2	.04
DLBCL		
3q27	+7	.005
	16qL	.0001
	19pL	.02
	14qL	.0001
DLBCL		
Non-ABC/GCB		
+16	16qL	.04
BL		
MYC	IqG	.0001
	13qL	.04

Additional RCAs (n = 8, plus 3q27 rearrangements) in combination with the previously identified RCAs were subsequently used in two supervised models (SVM and ANN). A discrimination analysis of these two models was also performed (Figure 4-9). Of note, sample partition with 10 fold cross validation and bootstrap of 1000 iterations was used for internal control on both supervised models.

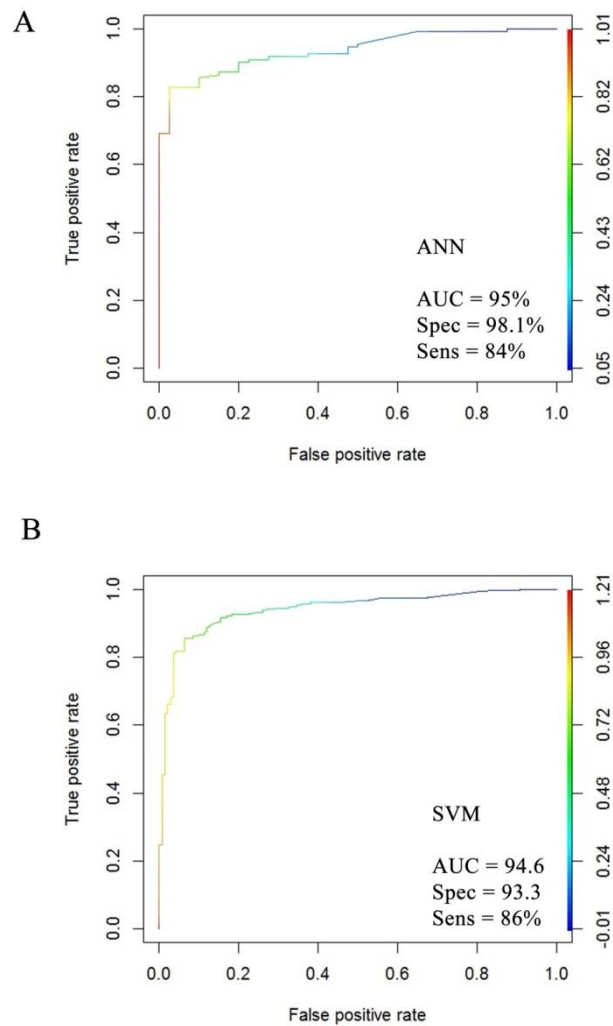


Figure 4-9. ROC curves of supervised models (Panel A, ANN with one hidden layer; B, SVM). Results are the averages of a 10 fold cross validation and 1000 boot-straps. The ROC curves illustrated here are the closest representation of the averages obtained from a 10-cross validation and 1000 boot-straps.

Our results indicate higher performance of the ANN model compared to a SVM diagnostic classifier (specificity of the ANN model ranged from 95.1-100%, while that of the SVM model fluctuated from 92.3-94%). To answer research question 5 (is there a difference between DLBCL and BL with respect to RCAs within a multivariable

context), a logistic regression model was also performed to explore the impact of each of the RCAs on DLBCL (see Figure 4-10). Under this model, 6qL, 1qG, +5, +X, 1p36L, +3, 17pL, +18, 14qL and -4 remained independent predictors of DLBCL. Of note, *BCL2* and 3q27 rearrangements, as expected for our internal controls, as well as -10/-15 predicted DLBCL every single time (specificity and sensitivity of 100%). Also, we should note that 1q gain (odds ratio of 0.21) is model here for DLBCL and not BL. To obtain the risk ratio of 1q gain in BL, simply inverting the risk ratio of 0.21 of DLBCL shows the risk ratio of the other group, in this case that of BL ($1/0.21 = 4.76$)¹²⁰.

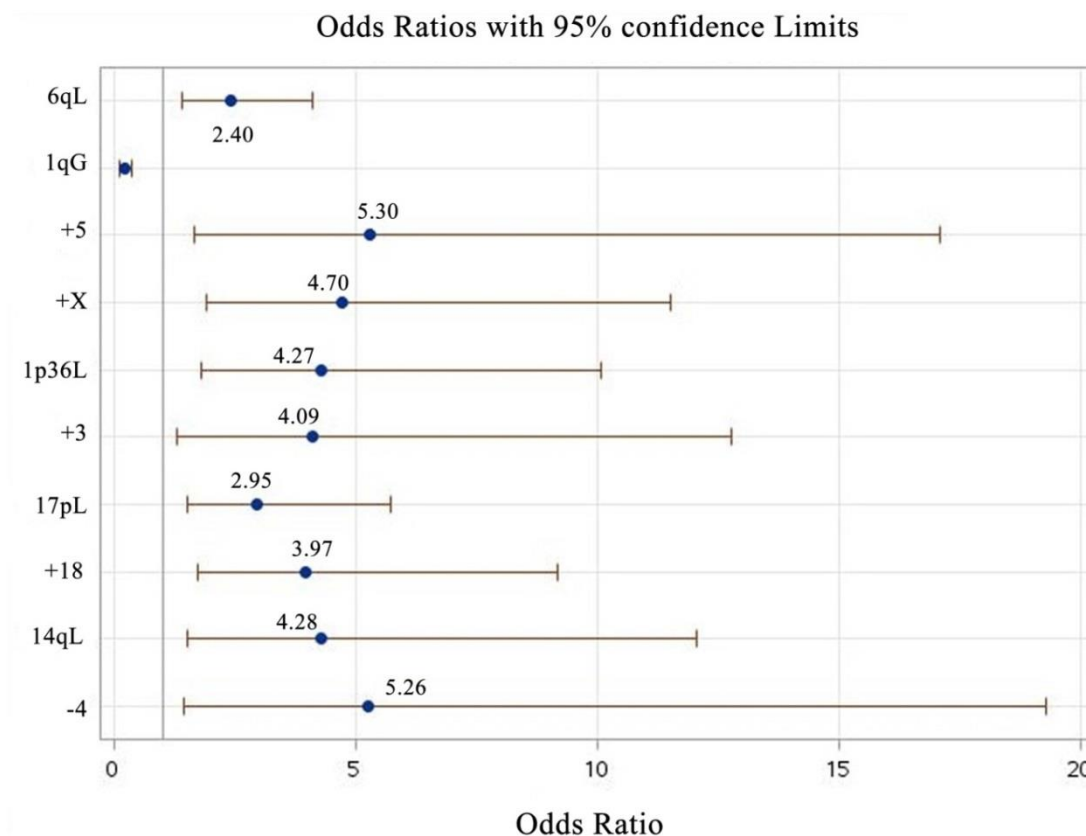


Figure 4-10. Odds ratio estimates of RCAs implicated in DLBCL. Schematic illustrates independent predictors of DLBCL and their respective odds ratio.

To determine discrimination and calibration, a c statistics and the Hosmer-Lemeshow goodness of fit - HL test were defined (Figure 4-11). A c statistic of 0.83 described the logistic regression as a good model (discrimination) and the Hosmer-Lemeshow Goodness-of-fit test defined how well the model performed (calibration). In this instance, a p-value > .05 signified the model fits the data well ($p > 0.9161$).

A Discrimination (c statistic)			
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.3	Somers' D	0.661
Percent Discordant	14.2	Gamma	0.699
Percent Tied	5.5	Tau-a	0.269
Pairs	50978	c	0.830

B Calibration		
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
3.2725	8	0.9161

Figure 4-11. Discrimination and calibration of logistic regression model. Panel A illustrates the c statistics result, while Panel B shows results from the Hosmer-Lemeshow Goodness-of-fit test.

4.10 RCAs Guidelines for Distinguishing DLBCL from BL

Given these observations, we propose the following set of guidelines to assess DLBCL from BL: 1) Evaluate chromosome complement based on the number of cytogenetic aberrations, 2) presence of primary cytogenetic aberrations (i.e., those RCAs that are associated with DLBCL with a p value < .01 by the correlational matrix or by previous analyses such as a logistic regression model). These primary RCAs include

3q27 rearrangements, *BCL2*, 6qL, -10/-15, +5, +7, 16qL, +X, 1p36L, +3, 9qL, +12, 15qL, 17pL, +18, 19pL, 14qL, +2, -2, -4, -8, and 22qL and 3) presence of secondary RCAs (i.e., those associated with DLBCL with a p-value < .05 by the correlational matrix). For the latter, these included: -10, -10/-14, 6pG and +11. Presence of one primary RCA suggests DLBCL, while presence of one or more secondary RCAs in the context of 7 or more cytogenetic aberrations suggests DLBCL.^{5,6} Figure 4-12 illustrates a proposed cytogenetics-based algorithm constructed on RCA findings.

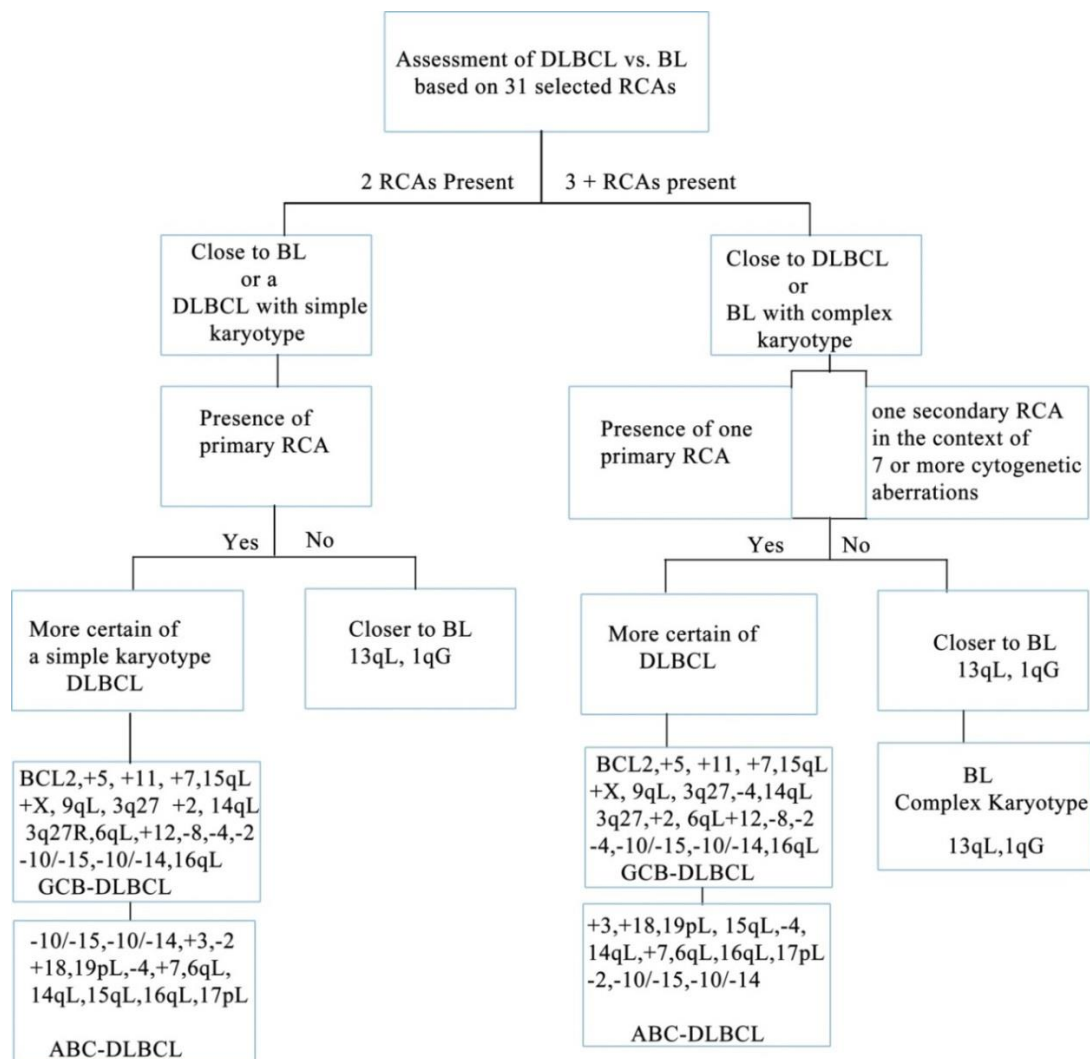


Figure 4-12. Proposed RCA algorithm for classification of DLBCL and BL.

4.11 Correlation of RCAs and Copy Number Aberrations in DLBCL and BL

Research question 6: Is there a difference in chromosome imbalances (measured by array CGH) in DLBCL subtypes compared to BL? To address the general usefulness (external validity) of our findings and to answer research question 6, the present study considered copy number aberrations by array CGH (a literature review) between DLBCL subtypes and BL. Of note and as indicated previously, only those copy number aberrations with a frequency of more than 20% were considered of relevance. Findings from this review revealed a number of RCAs correlated with copy number aberrations, mainly +3, +18, +16, +2, -8, -4, 15qL, 1p36L, +7, +12, 9pL, 14qL, 6pG, +11 and 13qL, while +X and 17p were prevalent in both tumor types. RCAs not correlated with the literature review of array CGH were -2, 16qL, 19pL, 22qL, 1qL, +5 and -10/-15 (see Table 4-8).

Table 4-8. Genomic Imbalances in DLBCL and BL and Correlated RCAs

Copy Number Imbalance	DLBCL	BL	RCAs
1p36L	+	-	1p36L
2pG	+	-	+2
3pG	+	-	+3
4qL/4qG	+	-	-4
6pG / 6pL / 6qL	+	-	6pG
8pL / 8qG	+	-	-8
9pL/ 9pG / 9qL	+	-	9qL

10pG	+	-	NA
11pG	+	-	+11
14qL/ 14q32 G	+	-	14qL
15qL	+	-	15qL
16q23-q24 L	+	-	16qL
17pG / 17qG	+	-	NA
19pG	+	-	NA
3q13.1L / 3q27G	-	+	NA
6q15G	-	+	NA
13q14L / 13q12G	-	+	13qL
19qL	-	+	NA
XpL	-	+	NA

Key: p, short arm of the chromosome; q, long arm of the chromosome; L, loss; G, gain; NA, not applicable.

4.12 CNVs Differences between DLBCL and BL: a CGH Dataset Analysis

With respect to the CNV analysis, Table 4-9 highlights the total number of cases along with sub-type classifications of the two distinct groups, while Table 4-10 shows significant CNVs associated with DLBCL and BL.

Table 4-9. Total Number of Cases and Subgroups of DLBCL and BL

Morphological Group	Total
DLBCL NOS	24
ABC DLBCL	63
GCB DLBCL	64
Transformed DLBCL	9
Relapse DLBCL	11
DLBCL Total	171
Classic BL	17
Atypical BL	12
BL NOS	41
Discrepant BL	8
BL Total	78

Key: NOS, not otherwise specified; Discrepant, molecular signature of BL

Table 4-10. CNVs Differences between DLBCL and BL

CNV p-Value	Gain/Loss	DLBCL/BL n/(%)	Specificity	DX
X cen-qter .004	G	17 (9.9)/0	100	DLBCL
X cen-pter .001	G	19 (11.1)/0	100	DLBCL
3 cen-qter .0008	G	22 (12.9)/0	100	DLBCL
4p15-pter .003	G	0/5 (6.4)	100	BL
5 cen-pter .032	G	18 (10.5)/2 (2.6)	98.7	DLBCL
6 cen-pter .002	G	18 (10.5)/0	100	DLBCL

6 cen-qter .04	L	15 (8.8)/1(1.3)	99.4	DLBCL
8q24-qter .02	G	3 (1.8)/7 (9.0)	96.0	BL
13q31-q32 .01	G	0/4 (5.1)	100	BL
17p13-pter .004	L	16 (9.4)/0	100	DLBCL
17 cen-pter <.0001	G	0/10 (12.8)	100	BL
18 cen-qter .0003	G	24 (14)/0	100	DLBCL

Key: cen, centromere; pter, p terminal region of chromosome; qter, q terminal region of chromosome; DX, diagnostic.

A multivariable analysis (logistic regression) of all significant CNVs in BL revealed gains of 8q24 remained an independent predictor of BL (odds ratio of 7.3) and gains of 4p15-pter, 13q31-32, and 17 cen-pter predicted BL. In contrast, gains of Xcen-qter, Xcen-pter, 3 cen-qter, 6cen-pter, 18 cen-qter and 17p13-pter loss predicted DLBCL every single time, whereas 5 cen-pter gain (odds ratio of 2.9) remained an independent predictor (data not shown). Genomic gains of 6p, X, 3, 5, 18 and loss of 17p in DLBCL was further supported by our previous RCA analysis. All CNVs associated with DLBCL correlated perfectly with previous RCA findings. The overall correlation between CVN and RCA data was 71.4%. With respect to any significant differences found in DLBCL subtypes vs. BL, a number of CNVs were identified associated with both ABC and GCB subtypes when compared to BL (see Table 4-11).

Table 4-11. Significant CNVs Differences between Distinct DLBCL Subtypes and BL

CNV Value	G/L	n / (%)	Specificity	Dx	p-
GCB vs. BL					
17 cen-pter .005	G	0/10(12.8)	100	BL	
17 cen-qter .02	G	1 (1.6)/10(12.8)	98.6	BL	
17p13-pter .02	L	5 (7.8)/0	100	DLBCL	
18 cen-qter .02	G	5 (7.8)/0	100	DLBCL	
ABC vs. BL					
1p34-pter .02	L	4 (6.25)/0	100	DLBCL	
1q25-qter .02	L	4/(6.25)/0	100	DLBCL	
3cen-pter .004	G	13(20.6)/3(3.8)	94	DLBCL	
3cen-qter <.0001	G	18(28.5)/0	100	DLBCL	
4p15-pter .038	G	0/5 (6.4)	100	BL	
6cen-pter .005	G	6(9.5)/0	100	DLBCL	
6cen-qter .006	L	8(12.7)/1(1.3)	98.2	DLBCL	
8q24 .038	G	0/5(6.4)	100	BL	
13q14-q22 .038	L	0/5(6.4)	100	BL	
17cen-pter .003	G	0/10(12.8%)	100	BL	

17p13-pter .002	L	7(11.1)/0	100	DLBCL
18cen-pter .001	G	9(14.3)/2(2.6)	96.4	DLBCL
18cen-qter <.0001	G	14(22.2)/0	100	DLBCL
Xcen-pter .0002	G	10(15.9)/0	100	DLBCL
Xcen-qter .001	G	8(12.7)/0	100	DLBCL

Key: cen, centromere; pter, p terminal region of chromosome; qter, q terminal region of chromosome; DX, diagnostic of tumor type.

Chapter V

V. Discussion

5.1 Overview

Only a limited number of studies have examined and systematic compared the chromosome architecture on both DLBCL and BL. In this present study, we explored RCAs in an effort to better define between DLBCL and BL. We associated data and outlined a set of diagnostic models (considered excellent models based on an AUC of 0.9 or greater) that better classified DLBCL from BL. Our findings, as expected, indicated a higher frequency in the total number of structural and numerical cytogenetic aberrations in DLBCL compared to BL. Also, unique recurring cytogenetic aberrations were identified between the two histological tumor types that suggest distinct molecular pathogenesis. Independent comparisons of both *MYC*⁺ and *MYC*⁻ DLBCL, vs. BL also showed unique RCAs that may be used to properly distinguish between these groups. Here, our analysis showed only marginal differences between the *MYC*⁺ and *MYC*⁻ DLBCL groups, allowing for these latter two groups as a whole to serve as a comparative group vs. BL. In a similar manner, we also defined a core set of RCAs that are closely associated with molecular subtypes of DLBCL. Knowledge of the presence or absence of *MYC* in DLBCL, as well as presence of RCAs that are closer in distance to specific molecular subtypes of DLBCL may be useful in further risk stratification of these patients. Moreover, we were able to distinguish between the two histological tumors types (DLBCL vs. BL) with high levels of confidence by using both unsupervised and supervised predictor models applied to RCAs; however, our findings showed artificial

neural networks superior diagnostic classifiers compared to all other models in terms of specificity (95-100% obtained from the ROC curve). Lastly, our results also revealed specific copy number variations and gene expression signatures that correlated with RCA data.

The diagnostic criteria of BL are well established; nonetheless, there remain cases that show atypical morphology, immuno-phenotype and genetic features and thus distinguishing between the two neoplastic entities, that is DLBCL and BL, remains problematic. Likewise, DLBCL in some instances may present with morphological features reminiscent of BL, such as a starry sky appearance that is so characteristic of BL. In addition and as previously mentioned in one study, six percent of histological DLBCL represent the molecular BL subtype⁵. In adults, the accurate classification of these two entities has crucial clinical importance since treatment and prognosis relies on their correct classification. Despite numerous studies in the area to better classify these two entities, additional clinical markers are clearly needed.

5.2 A Complex Chromosome Complement Defines DLBCL from BL

Our studies indicate DLBCL contains a higher level of genomic complexity than BL (greater than 2 RCAs). Indeed, even after a total of 21 reliable RCAs were applied to all sample karyotypes (second dataset used to evaluate predictor models), in general, BL retained no more than two RCAs. This was mostly evident in both the heat-map (cluster 12 belonging to BL only showed 1-2 RCAs) and the ANN model (most BL cases were homogeneous with a simple karyotype and found at a threshold level just below 0.2). Therefore, we conclude that a simple karyotype with no more than 2 RCAs is, for the

most part, characteristic of a diagnosis of Burkitt lymphoma. Indeed and in agreement with the published literature, BL is characterized by a simple karyotype. In 2009, Boerma and colleagues⁵⁷ reported BL with a low genomic complexity score when compared to other B-cell non-Hodgkin lymphomas with a *MYC* rearrangement, including DLBCL. In a similar manner, our group³⁵ in 2010 reported cases containing less than two additional chromosome aberrations associated with a diagnosis of BL and with an overall better outcome. Likewise in 2010 and using a microarray genomic hybridization method, Capello and co-workers¹²¹ reported HIV positive BL with a lower number of genomic lesions when compared to HIV positive DLBCL ($p < 0.032$). Subsequent studies by Havelange et al.¹¹⁵ in 2013 using both conventional karyotyping and fluorescence in situ hybridization (FISH), showed similar results. In this latter study, the chromosome complement of BL was less complex compared to other *MYC*+ B-cell lymphomas. Therefore, a lack of complex genetic aberrations in combination with a *MYC* rearrangement, especially a translocation of an immunoglobulin partner gene with *MYC* (*IG/MYC*), is more suggestive of BL. Nevertheless and as previously mentioned, previous molecular studies have extended the spectrum of BL to include cases with morphological features of DLBCL and expression of BCL-2^{5,6} and containing more complex genetic aberrations⁸. Thus, additional genomic characteristics other than a *MYC* translocation coupled with a simple karyotype are needed to clearly distinguish a ‘true’ BL from DLBCL. In this regard, the following sections explore RCAs that may be used to better classify a ‘true’ BL from DLBCL.

5.3 RCAs Delineate DLBCL (MYC+ and MYC-) vs. BL

With respect to the frequency of RCAs in both the initial and combined dataset analysis of the two tumor types, our data showed unique cytogenetic aberrations present in DLBCL compared to BL. These findings are in line with a previous study in 2008 by Boerma and colleagues⁵⁷. Using a more systematic approach, we corroborated the latter study and identified 4q loss (-4), 6q loss, +18, +7, +11, +12, +X, 17pL and 15qL in *MYC* + DLBCL. Thus, further support for the presence of these RCAs in DLBCL. Moreover and in contrast with this earlier study, our analysis extended the number of RCAs in both tumor types and identified 13qL and 1q gain significantly more prevalent in BL, while gains of chromosomes 2, 3, 5, 6p and 16, as well as losses to chromosome 1p36, 1q, 2, 8, 14q, 16q, 19p, 20q and 22q were significantly associated with DLBCL. Discrepancy between our current findings and the study of Boerma and colleagues may be explained by the fact that the latter study only included *MYC*+ DLBCL. Indeed, 1p36L, 1qL, +2, -2, +3, -8, 14qL and 22qL, were all implicated with the *MYC*- DLBCL group in the present study, while gain of 6p was also more prevalent in the *MYC*- DLBCL group (data not shown). Because both *MYC*+ and *MYC*- DLBCL may pose a challenge in distinguishing between BL, and given the fact that up to 10% of cases with a *MYC* rearrangement can go undetected by conventional FISH techniques due to the great variation in *MYC* breakpoints (i.e., a *MYC*- DLBCL may in fact represent a *MYC*+ DLBCL), both *MYC*+ and *MYC*- DLBCL were included in this current study. In terms of practical implications of these findings and given the fact that chromosome analysis is a routine work-up of all suspected lymphomas, presence of these RCAs, easily identifiable in a sample karyotype and readily available in clinical settings, can with a high degree of

confidence distinguish between DLBCL and BL, thus reducing the number of patients that are incorrectly treated and assessing a more accurate prognosis of these patients.

5.4 Reliable RCAs in the Published Literature

To the best of our knowledge, these set of RCAs are here first reported to highly predict DLBCL from BL. Indeed, these non-random RCAs preferentially seen in DLBCL compared to BL seem to suggest a role in the pathogenesis of disease. For example, deletions at 1p36 normally detected in DLBCL may be important in lymphomagenesis. The tumor suppressor gene (TSG) p73 that maps to 1p36 was commonly deleted in 27% of DLBCL cases by fluorescence in situ hybridization¹²². Further support for 1p36 loss in DLBCL compared to BL, comes from a protein overexpression study of ID3 (located at 1p36) that was correlated with a diagnosis of BL vs. DLBCL ($P < .001$) in a recent study⁷⁵. Moreover, a review of array CGH revealed loss of 1p36 as a common occurrence in DLBCL subtypes³⁷. In this latter instance, loss of 1p36 was commonly deleted in 20% or more of the cases. Subsequent survey of the literature assessing for copy number variations by array CGH of BL and later compared to previously reported findings on DLBCL by our group in 2012³⁷ showed no distinct deletions at 1p36 for BL (Figure 2-1). Similarly and further supported in our analysis, RCAs in a multivariable context also identified 1p36L as an independent predictor of DLBCL (odds ratio of 4.27). Of interest, both +3 and +18 along with 1p36 loss remained relatively close in distance in terms of the k-means partition analysis using the PAM algorithm, as well as the RCA cluster dendrogram based on Euclidian distance and p-values. This may suggest that 1p36L is preferentially found in the ABC-DLBCL

molecular subtype of DLBCL (or *MYC*- DLBCL). Indeed, there was a significant difference in 1p36L between *MYC*+ and *MYC*- DLBCL; however, it should be mentioned that in a larger dataset, 1p36L was only more prevalent in a *MYC*- DLBCL group ($p = .07$, data not shown). In terms of the remaining RCAs listed here, all appear in agreement with the published literature. A comparative genomic hybridization study ($n = 52$) of DLBCL tumors, Berglund et al.¹²³ reported chromosomal losses to 6q23-q terminal (20%) and gains were preferentially observed in Xq25-q26 (43%), 3q24-q25 (11%), 12 centromere-q14 (20%), 7 (11%) and 18q12-21. In terms of chromosome loss to the long arm (q) of chromosome 6, it is cited in the published literature as characteristic feature of the ABC-DLBCL molecular subtype. Here, our findings were also in line with the literature and found 6q loss significantly more prevalent in DLBCL vs. BL when we combined the two datasets ($n = 515$). We should also note that 6q loss remained significant even when the number of RCAs for analysis was not restricted to more than two RCAs for sample karyotypes (data not shown). In terms of loss of chromosome 2, 4 and 9q, Chen and associates¹²⁴ reported losses to chromosomes 2p25.2-25.3 and 4p15.32-q35.2, as well as 1p31.1-p36.33 and 15q11.2-q26.3 using CGH in 64 DLBCL cases, while Deffenbacher et al.⁷¹ reported losses to 1p36.22-p36.32 (25%), 1p35.1 (30%) and 9q22.31 (20%) using CGH in DLBCL. Similarly, Oudejans and colleagues⁷⁶ described losses to 4q12 and 15q11.2-q21.3. With regards to gains of chromosome 16 found more prevalent in DLBCL compared to BL in this current study, previous work by Takeuchi and associates¹²⁵ support our findings. In this latter study, gains of 16p13.3, 16q24.3, 16q22.1 and losses to 4q12 were identified in DLBCL. In these same study, top clones in the ABC-DLBCL subset included gains to 3p, 3q and 18q21. Moreover and when

considering gain of chromosome 6p, Kasugai and colleagues¹²⁶ published gains of chromosome 6p21 leading to the overexpression of *CCND3* in DLBCL. Lastly and in terms of DLBCL, we should mention the loss of the long arm of chromosome 20. This latter chromosome abnormality was found more prevalent in DLBCL when the number of RCAs was not restricted to more than two RCAs in the comparative analysis of the two groups (data not shown). Our findings of 20q loss in DLBCL are supported by a new diagnostic algorithm by Soldini and co-workers⁷⁵. Here, the *CSE1L* gene mapped at 20q was over-expressed in BL compared to DLBCL. Of note, loss of 20q was not part of our diagnostic models. When considering BL, we should highlight duplication of the long arm of chromosome 1 [dup(1)(q)] and 13q loss were identified as frequent aberrations in recent studies of BL with complex karyotypes^{115,127}. Although some of the chromosome alterations described in this present study have been previously identified more prevalent in DLBCL than BL, mainly those identified by Boerma and associates where they directly compared between the two tumor types, or just simply other RCAs that are described specifically in either DLBCL or BL (e.g., 13qL and dup (1)q in BL or loss of chromosome 2, 4, 9q in DLBCL), we should highlight other RCAs, to the best of our knowledge, appear new reports in the literature. Among these include: +5, 1q loss, -8, 14qL, 19pL and 22qL.

5.5 RCAs Define Molecular Subtypes of DLBCL

DLBCL, as previously stated, contains three distinct molecular subtypes that originate from distinct B-cell differentiation stages (cell of origin, COO) that correlate with gene expression profiling: the germinal center B-cell like (GCB) – DLBCL, Activated B-cell-like (ABC)- DLBCL and the primary mediastinal large B-cell lymphoma (PMBCL). The

COO based gene expression signature also correlates with prognosis, with the ABC-DLBCL having an inferior prognosis. Table 5-1 outlines some of the characteristic features between the GCB and ABC subtypes.

Table 5-1. Features of DLBCL (GCB and ABC subtypes)

	GCB	ABC
Frequency	~50%	~45%
5-year OS	~60%	~30-40
Cytogenetics	t(14;18), <i>MYC</i> , 3q27	3q27, +3, 6qL
Protein markers	CD10, BCL6, LM02, GCET1,2	MUM-1, CCND2, FOXP1, BCL2

Because Gene expression profiling of COO is expensive, requires fresh tissue, not readily available in clinical practice and results laborious to perform, at present immunohistochemical (IHC) algorithms are used as surrogate markers of COO-DLBCL (GCB-DLBCL and non-GCB-DLBCL). Among these include the Hans, Choi, Visco and Young, and Tally algorithms. Below find an illustration of the Hans algorithm. It has approximately an 80% concordance with gene expression profiling classification of GCB and non-GCB DLBCL. Since the introduction of the Hans algorithm, others have improved concordance classification of gene expression profiling by introducing additional IHC stains. One example is the Choi algorithm (Figure 5-1) that reported a 93% concordance with gene expression by adding FOXP1¹²⁸.

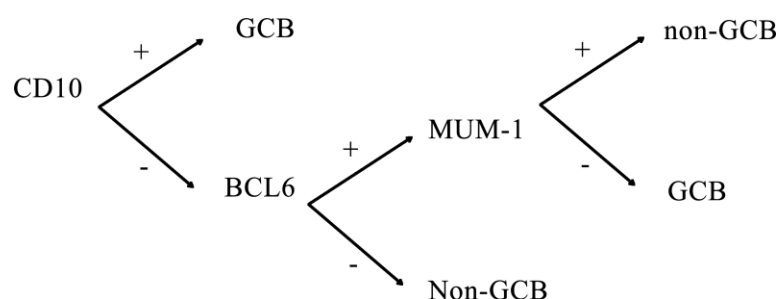


Figure 5-1. Schematic of the Hans algorithm. The Hans algorithm is used for the classification of germinal center and non-germinal center DLBCL.

In this regard and in the context of a cytogenetic evaluation of lymphomas in a clinical setting, our findings show that RCAs may also be used as surrogate markers of COO that correlates with gene expression, while at the same time provide an accurate classification of DLBCL and BL. In this terms, presence of +5, +11, +X, +2, 9qL and -8 may be added to the list of already well-established RCAs [+12 and (14;18) translocation] used for detecting neoplastic cells derived from the germinal center (i.e., GCB-DLBCL). Likewise, +3/+18, 6qL along with 1p36L, 19pL and 16qL based on our findings are more in line with the ABC-DLBCL molecular subtype. Knowledge of those RCAs that clearly outline COO-DLBCL, as well as knowledge of the *MYC* status in DLBCL can further help risk stratified these patients.

5.6 RCAs Applied to Supervise Predictor Models Show High Specificity

Taken together and with the use of a reliable set of RCAs along with the application of unsupervised and supervised predictor models, especially that of supervised models, our findings showed high levels of discrimination for the ANN model (95-100% specificity) and supervised SVM model (92-94% specificity) , when distinguishing between DLBCL and BL. Of note, since the primary need was to

distinguish one neoplasm from the other (i.e., ‘rule-in’ DLBCL), the primary requirement was to evaluate discrimination, and more specifically the degree of specificity of predictor models using the ROC curve. In this context and to the best of our knowledge, this is the first reported set of predictor models based on chromosome alterations that resolve DLBCL from BL with high specificity. Moreover, the RCA classifier established here outperforms previously reported classifiers in terms of specificity and is comparable to a very recent molecular targeted gene expression classifier just published in January of 2015. This latter study obtained a specificity of 91-100 for BL¹²⁹. In terms of diagnostic classifiers for DLBCL and BL, it should be noted that only limited studies have evaluated such predictor models. For example, Gormley and associates⁴⁸ using a hierarchical cluster based model on 8 ABC-DLBCL and GC-DLBCL markers reported a specificity of BL of 87% and based on the data provided by the authors a lower specificity of DLBCL. Although the predictor models presented in this current study predicted tumor samples with a high level of certainty; nonetheless, it should be noted that one of the limitations of these models may be their inability to resolve simple karyotypes found in DLBCL (e.g., a single *MYC* rearrangement with no additional aberrations or cases with 1 or 2 RCAs that are not otherwise identified as significant in this study). Moreover and because it is expected of predictor models not to classify with a hundred percent accuracy, mainly due underlying markers that may not prove easily detectable as significant, other ancillary studies are needed to resolve these cases. In fact, one other limitation of this study was that only cytogenetic markers were evaluated, and clearly a more comprehensive set of clinical biomarkers are warranted for evaluation. Despite these shortcomings, predictor models applied to a set of reliable RCAs, mainly ANN and

SVM supervised models provide a high level of confidence when distinguishing between DLBCL and BL.

5.7 Other Surrogate Biomarkers associated with RCAs may Prove Helpful in Distinguishing DLBCL from BL

Based on our findings and in terms of the utility of these set of recurrent cytogenetic markers to distinguish between DLBCL and BL, other biomarkers associated with the presence or absence of these RCAs may also be used to distinguish between the two neoplasms. For example, -4 and +11 correlate to a protein expression level of CD38-/CD44+ (CD38 resides at 4p15 and CD44 at chromosome 11p13) in non-BL⁵³. Moreover, and as indicated previously by an earlier study⁵¹, expression of CD44 makes a reliable candidate to immuno-phenotypically discriminate between DLBCL and BL. Therefore, use of CD38 and CD44 in a lymphoma work-up for differential diagnosis of these neoplasms is highly recommended. Likewise and in a similar manner, these set of RCAs described herein may also be used to better define provisional cases that are assigned to the temporary container of B-cell lymphoma, unclassifiable, with features between DLBCL and BL (BCLU). In this case, primary antibodies directed against the overexpression of SOX11, SRY sex determining region Y-box 11 (present at chromosome 2p25 and highly expressed in both BL and BCLU)¹³⁰ and detected by immunohistochemistry-IHC, may better assign these cases as a B-cell lymphoma closer to BL in this BL-DLBCL continuum since monosomy of chromosome is more in agreement with a DLBCL tumor phenotype. Thus, detection of SOX11 by IHC supported by a correlation of morphological, immunophenotypic and cytogenetic findings may indicate the use of a more aggressive chemotherapeutic regimen amenable to BL in border line cases. Other biomarkers that are not fully documented in the literature in

terms of distinguishing DLBCL from BL that may be useful are CD83 (located at chromosome 6p and most likely over-expressed in DLBCL that may be detected by IHC or immunophenotypic methods) and CD200 (consistent with trisomy of chromosome 3). In short and for future studies, these biomarkers should be evaluated in a large cohort of patients to assess their utility in a clinical setting.

5.8 Extended Literature Review and Copy Number Variations

We correlated a number of CNVs to cytogenetic data. In this analysis, all CNVs significantly associated with DLBCL correlated with the RCA data, while gains of 4p15-pter, 8q24 and 17cen-pter uniquely distinguished BL from DLBCL in this present CNV analysis. However, we should point out that gain of 8q24 has been implicated in a number of DLBCL subtypes³⁷. In terms of 1p36 loss in DLBCL, the CNV analysis showed a significant difference between the ABC DLBCL subtype and BL in terms of 1p36 loss ($p < .02$), and gene expression of *ID3* (inhibitor of DNA binding 3, dominant negative helix-loop-helix protein) located at 1p36 was significantly under-expressed in DLBCL compared to BL. Further evidence of ID3 in BL comes from a recent study of protein overexpression where it was correlated with a diagnosis of BL ($P < .001$)⁷⁵. Moreover, an extended literature review of array CGH revealed loss of 1p36 as a common occurrence in DLBCL subtypes³⁷. In this latter instance, loss of 1p36 was commonly deleted in 20% or more of the cases. Subsequent survey of the literature assessing CNV by array CGH of BL revealed no distinct deletions at 1p36 for BL. In addition and in a multivariable context, 1p36 loss remained an independent predictor of DLBCL vs. BL (odds ratio of 4.27). Deletions of 1p36 have been implicated in Non-Hodgkin lymphoma. In one study, Stoffel and associates¹²² reported deletions to p73, a

tumor suppressor and key in lymphomagenesis that is mapped to 1p36 and related to p53, in 25% of follicular lymphoma and 27% of DLBCL cases. Immunohistochemical analysis of these cases revealed a high Ki-67, indicating a high proliferation rate. Other correlated RCAs and CNVs worth mentioning include: gain of 3cen-qter that was significantly associated with the ABC subtype compared to the atypical, classic and BL-NOS groups ($p < .0001$) and +12 highly correlated with DLBCL ($p < .0001$). Furthermore, and of interest, both +3 and +18 along with 1p36 loss, all associated with an ABC subtype DLBCL remain relatively close in distance in terms of the k-means partition analysis and a RCA cluster dendogram construct, further supporting the presence of 1p36 loss in the ABC-DLBCL molecular subtype. Of interest and in regards to BL, genes identified in a biological pathway that are directly or indirectly involved with *MYC* and that have been implicated in a BL diagnosis to drive neoplastic progression include *CSEIL* (mapped at 20q13), *TCF3*(19p13) and *ID3* (1p36) overexpression^{75,131}. This suggests these genes are under expressed in DLBCL. Indeed, loss of 1p36, 19p and 20q were found more prevalent in DLBCL in our cytogenetic analysis. . Of note, we should mention that loss of chromosome 20q was also significantly associated with DLBCL when compared to BL. This difference was notable when the number of RCAs present in sample karyotypes were not restricted to more than 2 chromosome alterations when performing the comparative analysis between groups (data not shown). Therefore and in a biological context, *ID3* and *TCF3* overexpression results in VPREB3 (pre-B lymphocyte protein 3) increase expression, resulting in the activation of the pro-survival PI3K kinase pathway (phosphatidylinositol-3-OH). In this pathway, *MYC* controls the direct expression of *EBF1* and that of *TCF3*

indirectly (through a negative feedback loop) that in turn activates the expression of VPREB3, resulting in a deregulated PI3K signaling pathway and subsequently supporting survival of BL cells¹³¹ (see Figure 5-2).

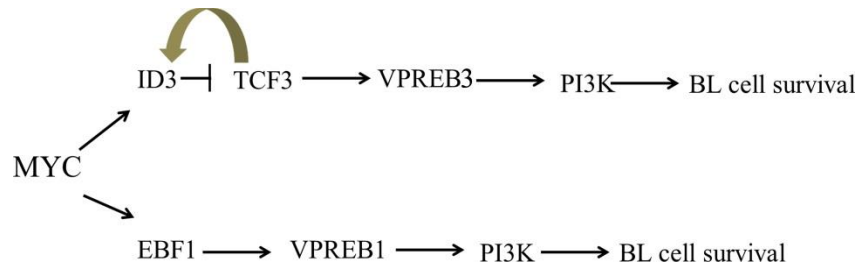


Figure 5-2. PI3K pathway in BL. Involvement of *MYC*, *ID3*, *TCF3*, VPREB1 in the phosphatidylinositol-3-OH kinase pathway in BL. Schematic adapted from Soldini et al.¹³¹

This pathway activation has been shown to play a key role in BL and concurrent expression of *ID3*/ *VPREB3* is highly specific of BL¹¹³. Based on Soldini and associates¹³¹, *ID3*-/ *VPREB3*- expression were absent in a BL group (n = 22). These observations suggest that *ID3*-/ *VPREB3*- or a single positivity of either of these two tumor markers excludes a diagnosis of BL. Of note, *ID3*-/ *VPREB3*- was also present in border line cases (BCLU, 19.2%) and DLBCL (37.2%). Regarding DLBCL vs. BL and in terms of RCAs, these latter findings are in agreement with current observations of 1p36 (*ID3*-) and 19p (*TCF3*-) loss in DLBCL, and the gene expression of this neoplasm (i.e., both *ID3* and *TCF3* are under expressed in DLBCL, while overexpression of these two genes is observed in BL). Nonetheless, it should be mention that VPREB3 expression is also found in border line cases (BCLU) and in 82% of DLBCL with a *MYC* translocation¹¹⁴. Thus, presence of VPREB3+ expression in *MYC*+ DLBCL further

supports our observations of 1p36 and 19p loss in *MYC*-DLBCL, or for that matter; presence of these RCAs are more in line with the ABC-DLBCL molecular subtype (the GCB-DLBCL are characterized by *MYC* translocations). In addition, in BCLU cases (n = 26) a double positive of ID3+/VPREB3+ was present in 100% and in contrast with the findings from Rodig and associates¹¹⁴ 87.5% of cases lacking a *MYC* translocation were observed with ID3+/VPREB3+¹³¹. Therefore, we suggest that a chromosome complement containing 1p36 and 19p loss excludes BCLU and is more in line with a DLBCL tumor. Moreover and in regards with reports of ID3+/VPREB3+ in *MYC*-DLBCL, it should be noted that other molecular mechanisms other than *MYC* may be responsible for ID3+/VPREB3+ modulation. Hence and based on this information, we propose concurrent anomalies of 1p36 and 19p loss may be used to exclude both BL and BCLU. Other notable gene worth mentioning in DLBCL is *CD44*. Mapped at chromosome 11p13 and correlated with trisomy of chromosome 11 with our cytogenetic data, is involved in driving cell cycle progression and replication of neoplastic cells by acting on *CCND2*, and this latter gene subsequently acting on *CDK5R1* to activate the G1/S transition of the cell cycle (this confirmed by a biological pathway generated from previously reported differentially expressed genes of DLBCL vs. BL – data not shown)^{5,6}. Of interest, *CDK5R1* (cyclin-dependent kinase 5 regulatory subunit 1) at chromosome 17q11.2, is linked to the G1/S transition in the mitotic cycle by *CCND2* as mentioned earlier and regulated by both micro RNA-103 and micro RNA-107 and, both differentially expressed in BL and DLBCL (fold change of -3.99 and -4.00 in BL vs. DLBCL each)^{69,132} suggesting *CDK5R1* plays a role in the molecular pathogenesis of DLBCL. Moreover, correlational studies between copy number aberrations by

comparative genomic hybridization and microarray expression profiles have shown an association between losses of chromosome 10, 14 and *CDK5RI* overexpression¹³³. Indeed, these observations are in agreement with our cytogenetic data (both -10/-14 are closely associated with a diagnosis of DLBCL). Another correlation of interest is deletion of the long arm (q) of chromosome 13, more prevalent in BL than DLBCL (p = .0006). In this regard, Onnis and co-workers reported a trend of down regulation of micro RNA, miR-9-1 (mapped at 13q), in a subset of BL cases that lacked a *MYC* translocation¹³⁴. Therefore, it may be argued that cases lacking a detectable *MYC* translocation with 13q deletion reminiscent of BL by morphology and/or immunophenotype may indeed represent true *MYC*- BL; however, further research is needed in this area.

Chapter VI

VI. Summary and Conclusion

6.1 Conclusion

In light of the prognostic and therapeutic consequences, distinguishing between DLBCL and BL is critically important. Therefore, a correct diagnosis requires a set of highly reliable tools that can discriminate between the two histological tumor types. Presently, morphological basis, as well as immunophenotypic and cytogenetics features are used in a clinical setting to properly diagnose DLBCL from BL. Nonetheless, as mentioned previously, overlapping features between the two entities still pose a diagnostic challenge, and thus additional criteria is needed that can be easily implemented in a clinical setting (at present molecular profiling is complex, laborious and not an routine diagnostic tool in a clinical laboratory). In this study, cytogenetic information from the Mitelman database was collected to identify reliable RCAs that may be used to distinguish between the two distinct neoplasms. Since chromosome analysis is part of a lymphoma work-up, a diagnostic classifier based on RCAs is highly desirable. To answer our first research question-RQ1 (is the average number of chromosome aberrations in DLBCL greater than in BL?), an independent t-test showed DLBCL with a higher number of aberrations than BL ($p < .003$), these results are supported by previous studies (BL is characterized by a simple karyotype³⁵). Our initial analysis also revealed a total of 20 RCAs that were significantly associated with DLBCL. These included the following: +3, +18, +7, +12, +X, -8, +16, -4, +2, -2, loss of 1p36, 17p, 15q, 22q, 9q, 19p, 1q, 16q and 14q. Of these and to answer RQ2 (is the number of RCAs different in *MYC*+ vs. BL), +X, +7, 15q loss, +16, 17 p loss and +18 were more prevalent in a *MYC*+

DLBCL vs. BL ($p < .05$). To further assess DLBCL vs. BL in more general terms, that is both *MYC*⁺ and *MYC*⁻ DLBCL vs. BL and to address RQ3 (is the number of RCAs different between *MYC*⁺ DLBCL from *MYC*⁻ DLBCL), our analysis showed marginal differences between the two groups. From these, we conclude that both *MYC*⁺ and *MYC*⁻ DLBCL may be used as comparative group vs. BL, and at the same time provide information regarding the *MYC* status of DLBCL to further risk stratify these patients. We should also point out that cases with *MYC*⁻ DLBCL and BL may also be difficult to properly diagnose and thus, knowledge of RCAs unique to *MYC*⁻ DLBCL are as well important to assess. In this context and when *MYC*⁺ and *MYC*⁻ DLBCL were combined and compared with BL - RQ4 (is the number of RCAs different from *MYC*⁺ and *MYC*⁻ DLBCL different from BL); our data showed 20 unique RCAs associated with DLBCL. Subsequent analysis in a multivariable context- RQ5 (is there a difference between DLBCL and BL with respect to RCAs when multiple variable are included), revealed RCAs that remained independent predictors of outcome – in this cases DLBCL, mainly +18, +X, 17p loss, 15q loss, 1p36 loss, +3, -4, +16, +2 and -2. To evaluate these RCAs with a different methodology and to explore RQ6 (is there a difference in chromosome imbalances by array CGH between DLBCL and BL), we performed an extended literature review of array CGH between the two entities. Comparing the percentages of cases reported in the literature and setting an arbitrary difference of 20% or more as significant, this current study identified a number of genomic gains and losses that correlated with RCAs. For example, 1p36 loss, +2, +3, -4, 6p gain, -8, 9q loss, +11, 14q loss, 15q loss, 16q loss and 13q loss were all correlated. Likewise, when we assess copy number variations -RQ7 (is the number of CNVs different between DLBCL and BL), we

identified similar results. Of interest, 1p36 loss, 1q loss, 6p gain, 6q loss, 17p loss, +18 and +3 were associated with the ABC molecular subtype of DLBCL vs. BL.

Next, to further investigate RCAs and to assess the clinical utility of cytogenetic markers, a number of diagnostic models were constructed. For this analysis, a distinct cytogenetic dataset of 177 cases from the published literature, as well as a set of institutional cases were collected and tested. These sample karyotypes were subsequently applied to a set of predictor models and reliable RCAs. Among these models included: unsupervised models that included hierarchical clusters including one based on Euclidean distance and a second one based on both Euclidean distance and a p-value, k-means partition using the PAM algorithm, logistic regression and supervised models that included both neural networks and a SVM model. The ability of models to discriminate between the two B-cell non-Hodgkin lymphomas was assessed with the receiver operating curve-ROC, here the specificity was considered as the major requirement to 'rule-in' DLBCL or BL. In these terms, the higher the specificity, the better the discrimination ability of the model. In a similar manner, the area under the curve-AUC for the ROC curve or the concordance statistic – c for the logistic regression model were calculated to evaluate the clinical usefulness of the various models. According to some, a c statistics (comparable to the AUC) of 0.7 is acceptable, 0.8 is described as a good model and over 0.9 is considered an excellent model. With respect to both unsupervised and supervised models, the clustered heat-map based on Euclidean distance showed the following: 86.7, specificity; 93.2, sensitivity and an AUC of 0.9. The cluster based on both Euclidian distance and a p-value had a specificity of 81.7 and a sensitivity of 99.1 with an AUC of 0.9. For the K-means partition using the PAM algorithm with a pre-

designation of 75 clusters, it showed a specificity of 88.3 and a sensitivity of 97.3 with an AUC of 0.93. In terms of the logistic regression, a c statistics revealed a discrimination of 85% and RCAs that remained independent predictors of DLBCL included +18, +3, -4, -8, +16, +2, -2 and loss of 17p, 15q and 1p36. For the ANN model, the ROC curve showed a specificity of 92.3 and a sensitivity of 90.2, while the AUC measured 0.9. Thus, our findings revealed ANN models is a superior diagnostic classifier and in general, unsupervised and supervised models were both deemed 'good' to 'excellent' models respectively. Moreover, to explore the relationship between RCAs, a cluster dendrogram of RCAs based on Euclidean distance and a p-value, along with a PAM algorithm and logistic regression were constructed. This analysis showed both distinct RCAs between the two molecular subtypes of DLBCL (activated B-cell DLBCL, +3, +18, loss of 1p36, 17p and 19p loss; germinal center DLBCL, +X, +12, +2, +16, 9q loss and -8) and a primary set of 12 RCAs that were more closely associated with DLBCL, mainly +3, +18, +16, +2, -8, -2, +X, -4, 15q, 1p36, 16q and 19p. These primary set of RCAs were then used to construct an algorithm that may be used to evaluate patient's sample karyotypes between DLBCL vs. BL. Of note, to determine RCAs between the two molecular subtypes of DLBCL, three internal controls were used for this purpose: *BCL-2* for the GC-DLBCL and both +3 and +18 for the ABC-DLBCL. Lastly, 5 additional RCAs, 6p gain, +11, +5 and -10/-15, -10/-14 were defined more prevalent in DLBCL ($p < .05$), while loss of 13q loss and 1q gain were found associated with BL. These were obtained when the two cytogenetic datasets were combined (dataset used to identify RCAs + dataset used to test predictor models). A correlational matrix of all RCAs was then constructed; findings from this analysis further supported our previously

proposed algorithm to assess sample karyotypes and further delineated DLBCL vs. BL, as well as molecular subtypes of DLBCL. In this instance, +5 and +11 were added to the list of RCAs that are characteristic of the GCB-DLBCL. Moreover, a logistic regression that included all cytogenetic data identified 6q loss, +5, +X, 1p36 loss, +3, 17p loss, +18, 14q loss and -4 as independent predictors of DLBCL. Subsequently, a support vector machine - SVM and ANN model of all cases was generated and plotted against a ROC curve. Here, a specificity ranging from 92-94% and an AUC of 0.93 was obtained for the SVM model, while 95-100% specificity was determined for the ANN model with an AUC of 0.90. Therefore, presence of these RCAs (18 previously identified plus an additional 6 RCAs obtained from the combining of datasets) increased specificity and showed a high reliability of these RCAs when distinguishing between the two distinct neoplasms. In brief, our findings indicate that a diagnostic classifier based on RCAs can be used with great level of confidence to consistently diagnose between the two tumor types. In addition and in terms of internal validity, a 10 cross validation and 1000 bootstraps obtained stable results for both SVM and ANN models. Likewise, correlations of RCAs to copy number variations (both classical CGH and a literature review of array CGH) showed generalization (external validity) of these results.

6.2 Summary

In summary, this is the first comprehensive cytogenetic analysis that tries to systematically establish unique cytogenetic markers that can be used to reliably distinguish between DLBCL vs. BL in a clinical setting. We described certain genomic aberrations that were significantly more common in one than the other and that may display distinct biological activities and/or pathways. Based on these current findings,

we proposed a set of RCAs that can help guide clinicians to better manage each patient and provide them with the optimal treatment. With this in context, both unsupervised and supervised predictor models applied to RCAs were explored to correctly classify between the two distinct neoplasms. Of these, ANN models obtained the highest levels of specificity. Moreover, our findings add knowledge to unique cytogenetic markers that may be used as surrogate markers to identify molecular subtypes of DLBCL (cell of origin-COO classification in DLBCL). Likewise, unique RCAs were also identified that excluded a *MYC* rearrangement in DLBCL. This added information (i.e., COO and *MYC* status) provides further risk stratification of DLBCL patients. In addition, we identified distinct RCAs that may serve to re-classify BCLU cases, so called border line cases (e.g., loss of chromosome 2 resulting in *SOX11* under-expression precludes a classification of BL or BCLU). Similarly, we re-classify BCLU cases accordingly and provide further evidence of chromosome 13 q loss in BL and possibly an association with duplication of the long arm of chromosome 11 in *MYC*- BL (See Table 3-5 case number 230, the latter RCA, i.e., duplication of 11q was previously outlined in a cytogenetic study¹¹⁹). Although, we provide a detailed outline of the cytogenetic profiles of DLBCL and BL (evaluated for internal and external validity), a major limitation of this study is the lack of additional biomarkers that may be applied to a diagnostic classifier (ANN model) or a bioinformatics algorithm that may be used to distinguish between the two tumor types. We should also point out that DLBCL with no RCAs present or those with a simple karyotype (containing random or non-random RCAs not identified in this study) are not detected by our classifiers.

6.3 Future Research

Moving forward and to gain a better understanding of those cases that fall into the provisional container of B-cell lymphoma, unclassifiable, with features intermediate between BL and DLBC-BCLU, further research is needed to characterize these cases at the molecular level, as well as assigned these findings to specific RCAs (limited or no comprehensive data is available on BCLU cases with both molecular and cytogenetics data). Similarly, added research is needed to better characterize atypical cases of BL that are CD10-, BCL2 weak + and MUM1 weak + at the molecular and cytogenetic level. Likewise, rare cases (~ 3%) that are characterize as molecular BL otherwise similar to BL but lack a detectable *MYC* rearrangement needs further research. Briefly, although we provide a detailed classification of both DLBCL and BL in terms of cytogenetic markers, the ability to enhance prediction models and better predict border line cases requires a broad spectrum of ancillary markers in a large independent cohort study. This will lead to better classification systems that can consistently distinguish between difficult cases of DLBCL and BL and decrease the number of patients that are under-treated or over-treated. As more detailed information is gained from these additional investigations, more precise therapeutic modalities can also be developed that can benefit these patients.

Summary of Contributions

All Figures and Tables	page 1-119	RG
CyDAS analysis	page 71	RG
t-test analysis	page 71	RG
Identification of reliable RCAs	page 71	RG
Cytogenetic analysis in ISCN format	page 71-78	RG
All cluster analyses	page 79-89,96	RG
Cross validation / Bootstrapping	page 96	RG
ROC analysis	page 90,96	RG
Correlation matrix	page 93-95	RG
All logistic regression analyses	page 97-98	RG
Extended array CGH analysis	page 100-101	RG
CNV analysis	page 101-105	RG

Key: RG, Rolando Garcia.

References

1. Bellan C, Stefano L, Giulia de F, Rogena EA, Lorenzo L. Burkitt lymphoma versus diffuse large B-cell lymphoma: a practical approach. *Hematol Oncol*. Jun 2010;28(2):53-56.
2. Haralambieva E, Boerma EJ, van Imhoff GW, et al. *Clinical, immunophenotypic, and genetic analysis of adult lymphomas with morphologic features of Burkitt lymphoma*: Am J Surg Pathol 2005.
3. McClure RF, Remstein ED, Macon WR, et al. Adult B-cell lymphomas with burkitt-like morphology are phenotypically and genotypically heterogeneous with aggressive clinical behavior. *Am J Surg Pathol*. Dec 2005;29(12):1652-1660.
4. A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. The Non-Hodgkin's Lymphoma Classification Project. *Blood*. 1997;89(11):3909-3918.
5. Hummel M, Bentink S, Berger H, et al. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med*. Jun 8 2006;354(23):2419-2430.
6. Dave SS, Fu K, Wright GW, et al. Molecular diagnosis of Burkitt's lymphoma. *N Engl J Med*. Jun 8 2006;354(23):2431-2442.
7. Swerdlow SH, Campo, E., Harris, N.L., Jaffe, E.S., Pileri, S.A., Stein, H., Thiele, J., Vardiman, J.W. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 2008.
8. Salaverria I, Zettl A, Bea S, et al. Chromosomal alterations detected by comparative genomic hybridization in subgroups of gene expression-defined Burkitt's lymphoma. *Haematologica*. Sep 2008;93(9):1327-1334.
9. Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP. Genome screening by comparative genomic hybridization. *Trends Genet*. Oct 1997;13(10):405-409.
10. Lemeshow S, Hosmer DW, Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. Jan 1982;115(1):92-106.
11. Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. *Ann Thorac Surg*. Mar 2003;75(3):753-757.
12. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform*. May 2009;10(3):297-314.
13. Chopra P, Kang J, Yang J, Cho H, Kim HS, Lee MG. Microarray data mining using landmark gene-guided clustering. *BMC Bioinformatics*. 2008;9:92.
14. Kaufmann L RP. Clustering by means of medoids. In: Dodge Y, (ed). *Statistical Data Analysis Based on the L 1 Norm and Related Methods*. Elsevier Science,. 1987:405-416.
15. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC Bioinformatics*. Sep 27 2004;5:136.
16. O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics*. Apr 10 2003;4:13.

17. Ando T, Suguro M, Kobayashi T, Seto M, Honda H. Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. *Cancer Sci.* Oct 2003;94(10):906-913.
18. Hanrahan G. Computational neural networks driving complex analytical problem solving. *Anal Chem.* Jun 1 2010;82(11):4307-4313.
19. Schalkoff RJ, ed. A. Artificial Neural Networks. McGraw-Hill: New York. 1997.
20. Byvatov E, Schneider G. Support vector machine applications in bioinformatics. *Appl Bioinformatics.* 2003;2(2):67-77.
21. Cortes CV, V. support vector networks. *Machine Learning.* 1995;20:273-297.
22. Coiffier B. Diffuse large cell lymphoma. *Curr Opin Oncol.* Sep 2001;13(5):325-334.
23. Rosenwald A, Ott G. Burkitt lymphoma versus diffuse large B-cell lymphoma. *Ann Oncol.* Jun 2008;19 Suppl 4:iv67-69.
24. Frick M, Dorken B, Lenz G. New insights into the biology of molecular subtypes of diffuse large B-cell lymphoma and Burkitt lymphoma. *Best Pract Res Clin Haematol.* Mar 2012;25(1):3-12.
25. de Leval L, Hasserjian RP. Diffuse large B-cell lymphomas and burkitt lymphoma. *Hematol Oncol Clin North Am.* Aug 2009;23(4):791-827.
26. Barrans S, Crouch S, Smith A, et al. Rearrangement of MYC is associated with poor prognosis in patients with diffuse large B-cell lymphoma treated in the era of rituximab. *J Clin Oncol.* Jul 10 2010;28(20):3360-3365.
27. Savage KJ, Johnson NA, Ben-Neriah S, et al. MYC gene rearrangements are associated with a poor prognosis in diffuse large B-cell lymphoma patients treated with R-CHOP chemotherapy. *Blood.* Oct 22 2009;114(17):3533-3537.
28. Obermann EC, Csato M, Dirnhofer S, Tzankov A. Aberrations of the MYC gene in unselected cases of diffuse large B-cell lymphoma are rare and unpredictable by morphological or immunohistochemical assessment. *J Clin Pathol.* Aug 2009;62(8):754-756.
29. Copie-Bergman C, Gaulard P, Leroy K, et al. Immuno-fluorescence in situ hybridization index predicts survival in patients with diffuse large B-cell lymphoma treated with R-CHOP: a GELA study. *J Clin Oncol.* Nov 20 2009;27(33):5573-5579.
30. Yoon SO, Jeon YK, Paik JH, et al. MYC translocation and an increased copy number predict poor prognosis in adult diffuse large B-cell lymphoma (DLBCL), especially in germinal centre-like B cell (GCB) type. *Histopathology.* Aug 2008;53(2):205-217.
31. Neri A, Barriga F, Knowles DM, Magrath IT, Dalla-Favera R. Different regions of the immunoglobulin heavy-chain locus are involved in chromosomal translocations in distinct pathogenetic forms of Burkitt lymphoma. *Proc Natl Acad Sci U S A.* Apr 1988;85(8):2748-2752.
32. Tagawa H, Suguro M, Tsuzuki S, et al. Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. *Blood.* Sep 1 2005;106(5):1770-1777.
33. Raphael M, Gentilhomme O, Tulliez M, Byron PA, Diebold J. Histopathologic features of high-grade non-Hodgkin's lymphomas in acquired immunodeficiency

- syndrome. The French Study Group of Pathology for Human Immunodeficiency Virus-Associated Tumors. *Arch Pathol Lab Med*. Jan 1991;115(1):15-20.
34. Dave SS. Genomic stratification for the treatment of lymphomas. *Hematology Am Soc Hematol Educ Program*. 2013;2013:331-334.
 35. Seegmiller AC, Garcia R, Huang R, Maleki A, Karandikar NJ, Chen W. Simple karyotype and bcl-6 expression predict a diagnosis of Burkitt lymphoma and better survival in IG-MYC rearranged high-grade B-cell lymphomas. *Mod Pathol*. Jul 2010;23(7):909-920.
 36. Hasserjian RP, Ott G, Elenitoba-Johnson KS, Balague-Ponz O, de Jong D, de Leval L. Commentary on the WHO classification of tumors of lymphoid tissues (2008): "Gray zone" lymphomas overlapping with Burkitt lymphoma or classical Hodgkin lymphoma. *J Hematop*. Jul 2009;2(2):89-95.
 37. Tirado CA, Chen W, Garcia R, Kohlman KA, Rao N. Genomic profiling using array comparative genomic hybridization define distinct subtypes of diffuse large B-cell lymphoma: a review of the literature. *J Hematol Oncol*. 2012;5:54.
 38. Orem J, Mbidde EK, Lambert B, de Sanjose S, Weiderpass E. Burkitt's lymphoma in Africa, a review of the epidemiology and etiology. *Afr Health Sci*. Sep 2007;7(3):166-175.
 39. Facer CA, Playfair JH. Malaria, Epstein-Barr virus, and the genesis of lymphomas. *Adv Cancer Res*. 1989;53:33-72.
 40. Ziegler JL, Bluming AZ, Morrow RH, Fass L, Carbone PP. Central nervous system involvement in Burkitt's lymphoma. *Blood*. Dec 1970;36(6):718-728.
 41. Armitage JO, Weisenburger DD. New approach to classifying non-Hodgkin's lymphomas: clinical features of the major histologic subtypes. Non-Hodgkin's Lymphoma Classification Project. *J Clin Oncol*. Aug 1998;16(8):2780-2795.
 42. Harris NL, Jaffe ES, Stein H, et al. A revised European-American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. *Blood*. Sep 1 1994;84(5):1361-1392.
 43. Frost M, Newell J, Lones MA, Tripp SR, Cairo MS, Perkins SL. Comparative immunohistochemical analysis of pediatric Burkitt lymphoma and diffuse large B-cell lymphoma. *Am J Clin Pathol*. Mar 2004;121(3):384-392.
 44. Dogan A, Bagdi E, Munson P, Isaacson PG. CD10 and BCL-6 expression in paraffin sections of normal lymphoid tissue and B-cell lymphomas. *Am J Surg Pathol*. Jun 2000;24(6):846-852.
 45. Rodig SJ, Vergilio JA, Shahsafaei A, Dorfman DM. Characteristic expression patterns of TCL1, CD38, and CD44 identify aggressive lymphomas harboring a MYC translocation. *Am J Surg Pathol*. Jan 2008;32(1):113-122.
 46. Braziel RM, Arber DA, Slovak ML, et al. The Burkitt-like lymphomas: a Southwest Oncology Group study delineating phenotypic, genotypic, and clinical features. *Blood*. Jun 15 2001;97(12):3713-3720.
 47. Nakamura N, Nakamine H, Tamaru J, et al. The distinction between Burkitt lymphoma and diffuse large B-Cell lymphoma with c-myc rearrangement. *Mod Pathol*. Jul 2002;15(7):771-776.
 48. Gormley RP, Madan R, Dulau AE, et al. Germinal center and activated b-cell profiles separate Burkitt lymphoma and diffuse large B-cell lymphoma in AIDS and non-AIDS cases. *Am J Clin Pathol*. Nov 2005;124(5):790-798.

49. Cogliatti SB, Novak U, Henz S, Schmid U, Moller P, Barth TF. Diagnosis of Burkitt lymphoma in due time: a practical approach. *Br J Haematol.* Aug 2006;134(3):294-301.
50. Chuang SS, Ye H, Du MQ, et al. Histopathology and immunohistochemistry in distinguishing Burkitt lymphoma from diffuse large B-cell lymphoma with very high proliferation index and with or without a starry-sky pattern: a comparative study with EBER and FISH. *Am J Clin Pathol.* Oct 2007;128(4):558-564.
51. Schniederjan SD, Li S, Saxe DF, et al. A novel flow cytometric antibody panel for distinguishing Burkitt lymphoma from CD10+ diffuse large B-cell lymphoma. *Am J Clin Pathol.* May 2010;133(5):718-726.
52. McGowan P, Nelles N, Wimmer J, et al. Differentiating between Burkitt lymphoma and CD10+ diffuse large B-cell lymphoma: the role of commonly used flow cytometry cell markers and the application of a multiparameter scoring system. *Am J Clin Pathol.* Apr 2012;137(4):665-670.
53. Naresh KN, Hazem A.H Ibrahim, Stefano Lazzi, Patricia Rince, Monica Onorati, Maria R. Ambrosio et al., Diagnosis of Burkitt Lymphoma Using an Algorithmic Approach-Applicable in Both Resource-poor and Resource-rich Countries. *British Journal of Haematology.* 2011;154:770-776.
54. Lu B, Zhou C, Yang W, et al. Morphological, immunophenotypic and molecular characterization of mature aggressive B-cell lymphomas in Chinese pediatric patients. *Leuk Lymphoma.* Dec 2011;52(12):2356-2364.
55. Au WY, Horsman DE, Gascoyne RD, Viswanatha DS, Klasa RJ, Connors JM. The spectrum of lymphoma with 8q24 aberrations: a clinical, pathological and cytogenetic study of 87 consecutive cases. *Leuk Lymphoma.* Mar 2004;45(3):519-528.
56. Perry AM, Crockett D, Dave BJ, et al. B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and burkitt lymphoma: study of 39 cases. *Br J Haematol.* Jul 2013;162(1):40-49.
57. Boerma EG, Siebert R, Kluin PM, Baudis M. Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of today's knowledge. *Leukemia.* Feb 2009;23(2):225-234.
58. Garcia JL, Hernandez JM, Gutierrez NC, et al. Abnormalities on 1q and 7q are associated with poor outcome in sporadic Burkitt's lymphoma. A cytogenetic and comparative genomic hybridization study. *Leukemia.* Oct 2003;17(10):2016-2024.
59. Barth TF, Muller S, Pawlita M, et al. Homogeneous immunophenotype and paucity of secondary genomic aberrations are distinctive features of endemic but not of sporadic Burkitt's lymphoma and diffuse large B-cell lymphoma with MYC rearrangement. *J Pathol.* Aug 2004;203(4):940-945.
60. Toujani S, Dessen P, Ithzar N, et al. High resolution genome-wide analysis of chromosomal alterations in Burkitt's lymphoma. *PLoS One.* 2009;4(9):e7089.
61. Scholtysik R, Kreuz M, Klapper W, et al. Detection of genomic aberrations in molecularly defined Burkitt's lymphoma by array-based, high resolution, single nucleotide polymorphism analysis. *Haematologica.* Dec 2010;95(12):2047-2055.

62. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*. Jun 20 2002;346(25):1937-1947.
63. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. Feb 3 2000;403(6769):503-511.
64. Hans CP, Weisenburger DD, Greiner TC, et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*. Jan 1 2004;103(1):275-282.
65. Mittrucker HW, Matsuyama T, Grossman A, et al. Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science*. Jan 24 1997;275(5299):540-543.
66. Rosenwald A, Staudt LM. Gene expression profiling of diffuse large B-cell lymphoma. *Leuk Lymphoma*. 2003;44 Suppl 3:S41-47.
67. Deffenbacher KE, Iqbal J, Sanger W, et al. Molecular distinctions between pediatric and adult mature B-cell non-Hodgkin lymphomas identified through genomic profiling. *Blood*. Apr 19 2012;119(16):3757-3766.
68. Walther N, Ulrich A, Vockerodt M, et al. Aberrant lymphocyte enhancer-binding factor 1 expression is characteristic for sporadic Burkitt's lymphoma. *Am J Pathol*. Apr 2013;182(4):1092-1098.
69. Lenze D, Leoncini L, Hummel M, et al. The different epidemiologic subtypes of Burkitt lymphoma share a homogenous micro RNA profile distinct from diffuse large B-cell lymphoma. *Leukemia*. Dec 2011;25(12):1869-1876.
70. Di Lisio L, Sanchez-Beato M, Gomez-Lopez G, et al. MicroRNA signatures in B-cell lymphomas. *Blood Cancer J*. Feb 2012;2(2):e57.
71. Deffenbacher KE, Iqbal J, Liu Z, Fu K, Chan WC. Recurrent chromosomal alterations in molecularly classified AIDS-related lymphomas: an integrated analysis of DNA copy number and gene expression. *J Acquir Immune Defic Syndr*. May 1 2010;54(1):18-26.
72. 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*. Sep 9 2008;105(36):13520-13525.
73. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology*. Oct 2003;229(1):3-8.
74. Scholtysik R, Nagel I, Kreuz M, et al. Recurrent deletions of the TNFSF7 and TNFSF9 genes in 19p13.3 in diffuse large B-cell and Burkitt lymphomas. *Int J Cancer*. Sep 1 2012;131(5):E830-835.
75. Soldini D, Montagna C, Schuffler P, et al. A new diagnostic algorithm for Burkitt and diffuse large B-cell lymphomas based on the expression of CSE1L and STAT3 and on MYC rearrangement predicts outcome. *Ann Oncol*. Sep 11 2012.
76. Oudejans JJ, van Wieringen WN, Smeets SJ, et al. Identification of genes putatively involved in the pathogenesis of diffuse large B-cell lymphomas by integrative genomics. *Genes Chromosomes Cancer*. Mar 2009;48(3):250-260.
77. Lombardi C, Tassi GF, Pizzocolo G, Donato F. Clinical significance of a multiple biomarker assay in patients with lung cancer. A study with logistic regression analysis. *Chest*. Mar 1990;97(3):639-644.

78. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*. Aug 2002;48(8):1296-1304.
79. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. Dec 8 1998;95(25):14863-14868.
80. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. Oct 15 1999;286(5439):531-537.
81. Jaeger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. *Pac Symp Biocomput*. 2003:53-64.
82. Li J, Liu H, Ng SK, Wong L. Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics*. Oct 2003;19 Suppl 2:ii93-102.
83. Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. Jun 2001;7(6):673-679.
84. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*. Aug 29 2000;97(18):10101-10106.
85. Speed T. Statistical analysis of gene expression microarray data. *CRC Press*. 2003:190-197.
86. Wall ME, Dyck PA, Brettin TS. SVDMAN--singular value decomposition analysis of microarray data. *Bioinformatics*. Jun 2001;17(6):566-568.
87. Devore J. Probability and statistics for engineering and the sciences. 4th edition. *Ducbury Press*. 1995.
88. Xingran Cui C-WL MFA, Quan Liu, Jiann-Shing Shieh. Diffuse large B-cell lymphoma classification using linguistic analysis and ensembled artificial neural networks. *Journal of the Taiwan Institute of Chemical Engineers*. 2012;43:15-23.
89. Lek S BA, Dimopoulos I, Lauga J, Moreau J. Improved estimation using neural networks of the food consumption of fish populations. *Marine Fresh Water Research*. 1995;46:1229-1236.
90. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw*. May 2006;19(4):408-415.
91. Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform*. May 2009;10(3):315-329.
92. Bishop C. Neural Networks for Pattern Recognition. *Oxford: Oxford University Press*. 1995.
93. Chen H, Zhang, J., Xu, Y., Chen, B., & Zhang, K. . Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules in CT scans. *Expert Systems with Applications*. 2012.
94. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*. Oct-Dec 2002;35(5-6):352-359.
95. Kim SM, Han H, Park JM, et al. A comparison of logistic regression analysis and an artificial neural network using the BI-RADS lexicon for ultrasonography in

- conjunction with introobserver variability. *J Digit Imaging*. Oct 2012;25(5):599-606.
96. Forsberg JA, Sjoberg D, Chen QR, Vickers A, Healey JH. Treating metastatic disease: Which survival model is best suited for the clinic? *Clin Orthop Relat Res*. Mar 2013;471(3):843-850.
 97. Shi HY, Lee KT, Wang JJ, Sun DP, Lee HH, Chiu CC. Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: a nationwide study. *J Gastrointest Surg*. Nov 2012;16(11):2126-2131.
 98. Shi HY, Hwang SL, Lee KT, Lin CL. In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. *J Neurosurg*. Apr 2013;118(4):746-752.
 99. Oermann EK, Kress MA, Collins BT, et al. Predicting survival in patients with brain metastases treated with radiosurgery using artificial neural networks. *Neurosurgery*. Jun 2013;72(6):944-951; discussion 952.
 100. Shi HY, Lee KT, Lee HH, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. *PLoS One*. 2012;7(4):e35781.
 101. Caocci G, Baccoli R, Vacca A, et al. Comparison between an artificial neural network and logistic regression in predicting acute graft-vs-host disease after unrelated donor hematopoietic stem cell transplantation in thalassemia patients. *Exp Hematol*. May 2010;38(5):426-433.
 102. Hiller B, Bradtke J, Balz H, Rieder H. CyDAS: a cytogenetic data analysis system. *Bioinformatics*. Apr 1 2005;21(7):1282-1283.
 103. Cutler P, ed *Problem Solving in Clinical Medicine*. Baltimore, Maryland: Lippincott, Williams & Wilkins; 1998.
 104. Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*. 2004;32:2616-2641.
 105. Pallant J, ed *A step by step guide to data analysis using IBM SPSS*. 5th ed. New York, NY: McGraw-Hill companies; 2013.
 106. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. Feb 20 2007;115(7):928-935.
 107. Sigaux F, Berger R, Bernheim A, Valensi F, Daniel MT, Flandrin G. Malignant lymphomas with band 8q24 chromosome abnormality: a morphologic continuum extending from Burkitt's to immunoblastic lymphoma. *Br J Haematol*. Jul 1984;57(3):393-405.
 108. Slavutsky I, Andreoli G, Gutierrez M, Narbaitz M, Lucero G, Eppinger M. Variant (8;22) translocation in lymphoblastic lymphoma. *Leuk Lymphoma*. Mar 1996;21(1-2):169-172.
 109. Sawyer JR, Waldron JA, Jagannath S, Barlogie B. Cytogenetic findings in 200 patients with multiple myeloma. *Cancer Genet Cytogenet*. Jul 1 1995;82(1):41-49.
 110. Yoshioka T, Miura I, Kume M, et al. Cytogenetic features of de novo CD5-positive diffuse large B-cell lymphoma: chromosome aberrations affecting 8p21 and 11q13 constitute major subgroups with different overall survival. *Genes Chromosomes Cancer*. Feb 2005;42(2):149-157.

111. Aamot HV, Torlakovic EE, Eide MB, Holte H, Heim S. Non-Hodgkin lymphoma with t(14;18): clonal evolution patterns and cytogenetic-pathologic-clinical correlations. *J Cancer Res Clin Oncol*. Jul 2007;133(7):455-470.
112. Adam P, Steinlein C, Schmid M, et al. Characterization of chromosomal aberrations in diffuse large B-cell lymphoma (DLBL) by G-banding and spectral karyotyping (SKY). *Cytogenet Genome Res*. 2006;114(3-4):274-278.
113. Sander S, Calado DP, Srinivasan L, et al. Synergy between PI3K signaling and MYC in Burkitt lymphomagenesis. *Cancer Cell*. Aug 14 2012;22(2):167-179.
114. Rodig SJ, Kutok JL, Paterson JC, et al. The pre-B-cell receptor associated protein VpreB3 is a useful diagnostic marker for identifying c-MYC translocated lymphomas. *Haematologica*. Dec 2010;95(12):2056-2062.
115. Havelange V, Ameye G, Theate I, et al. Patterns of genomic aberrations suggest that Burkitt lymphomas with complex karyotype are distinct from other aggressive B-cell lymphomas with MYC rearrangement. *Genes Chromosomes Cancer*. Jan 2013;52(1):81-92.
116. Johnson NA, Al-Tourah A, Brown CJ, Connors JM, Gascoyne RD, Horsman DE. Prognostic significance of secondary cytogenetic alterations in follicular lymphomas. *Genes Chromosomes Cancer*. Dec 2008;47(12):1038-1048.
117. Watanuki J, Hatakeyama K, Sonoki T, et al. Bone marrow large B cell lymphoma bearing cyclin D3 expression: clinical, morphologic, immunophenotypic, and genotypic analyses of seven patients. *Int J Hematol*. Sep 2009;90(2):217-225.
118. Yeh YM, Chang KC, Chen YP, et al. Large B cell lymphoma presenting initially in bone marrow, liver and spleen: an aggressive entity associated frequently with haemophagocytic syndrome. *Histopathology*. Dec 2010;57(6):785-795.
119. Pienkowska-Grela B, Witkowska A, Grygalewicz B, et al. Frequent aberrations of chromosome 8 in aggressive B-cell non-Hodgkin lymphoma. *Cancer Genet Cytogenet*. Jan 15 2005;156(2):114-121.
120. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. Aug 2010;19(3):227-229.
121. Capello D, Scandurra M, Poretti G, et al. Genome wide DNA-profiling of HIV-related B-cell lymphomas. *Br J Haematol*. Jan 2010;148(2):245-255.
122. Stoffel A, Filippa D, Rao PH. The p73 locus is commonly deleted in non-Hodgkin's lymphomas. *Leuk Res*. Dec 2004;28(12):1341-1345.
123. Berglund M, Enblad G, Flordal E, et al. Chromosomal imbalances in diffuse large B-cell lymphoma detected by comparative genomic hybridization. *Mod Pathol*. Aug 2002;15(8):807-816.
124. Chen W, Houldsworth J, Olshen AB, et al. Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas. *Blood*. Mar 15 2006;107(6):2477-2485.
125. Takeuchi I, Tagawa H, Tsujikawa A, et al. The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of B-cell lymphomas and genetic regions involved in lymphomagenesis. *Haematologica*. Jan 2009;94(1):61-69.
126. Kasugai Y, Tagawa H, Kameoka Y, Morishima Y, Nakamura S, Seto M. Identification of CCND3 and BYSL as candidate targets for the 6p21

- amplification in diffuse large B-cell lymphoma. *Clin Cancer Res.* Dec 1 2005;11(23):8265-8272.
127. Maria Murga Penas E, Schilling G, Behrmann P, et al. Comprehensive cytogenetic and molecular cytogenetic analysis of 44 Burkitt lymphoma cell lines: secondary chromosomal changes characterization, karyotypic evolution, and comparison with primary samples. *Genes Chromosomes Cancer.* Jun 2014;53(6):497-515.
 128. Choi WW, Weisenburger DD, Greiner TC, et al. A new immunostain algorithm classifies diffuse large B-cell lymphoma into molecular subtypes with high accuracy. *Clin Cancer Res.* Sep 1 2009;15(17):5494-5502.
 129. Carey CD, Gusenleitner D, Chapuy B, et al. Molecular Classification of MYC-Driven B-Cell Lymphomas by Targeted Gene Expression Profiling of Fixed Biopsy Specimens. *J Mol Diagn.* Jan 2015;17(1):19-30.
 130. Burgesser MV, Gualco G, Diller A, Natkunam Y, Bacchi CE. Clinicopathological features of aggressive B-cell lymphomas including B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell and Burkitt lymphomas: a study of 44 patients from Argentina. *Ann Diagn Pathol.* Jun 2013;17(3):250-255.
 131. Soldini D, Georgis A, Montagna C, et al. The combined expression of VPREB3 and ID3 represents a new helpful tool for the routine diagnosis of mature aggressive B-cell lymphomas. *Hematol Oncol.* Sep 2014;32(3):120-125.
 132. Moncini S, Salvi A, Zuccotti P, et al. The role of miR-103 and miR-107 in regulation of CDK5R1 expression and in cellular migration. *PLoS One.* 2011;6(5):e20038.
 133. Wrobel G, Roerig P, Kokocinski F, et al. Microarray-based gene expression profiling of benign, atypical and anaplastic meningiomas identifies novel genes associated with meningioma progression. *Int J Cancer.* Mar 20 2005;114(2):249-256.
 134. Onnis A, De Falco G, Antonicelli G, et al. Alteration of microRNAs regulated by c-Myc in Burkitt lymphoma. *PLoS One.* 2010;5(9).