Freie Universität Berlin

Fachbereich für
Mathematik und Informatik

Transcriptional Profiling of Aggressive Lymphoma

Stefan Bentink

June 2009

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Rainer Spang
Prof. Dr. Harald Stein

Gutacher:
Prof. Dr. Martin Vingron
Prof. Dr. Rainer Spang
Prof. Dr. Harald Stein

Tag der mündlichen Prüfung: 9. Dezember 2009

For my Mother

Preface

Acknowledgements This work was carried out in the *Computational Diagnostics* group of the Department of Computational Molecular Biology at the Max Planck Institute for Molecular Genetics in Berlin and after moving in the *Computational Diagnostics* group of the Institute of Functional Genomics at the University of Regensburg. I thank all past and present colleagues for the pleasant working atmosphere and fruitful scientific and non-scientific discussions. Furthermore, the thesis research has been carried out in the framework of the research network "Molecular Mechanisms in Malignant Lymphoma" (MMML), supported by the Deutsche Krebshilfe. Members of the MMML are listed in the appendix.

I am grateful to my supervisors *Rainer Spang* and *Martin Vingron*, especially Rainer for giving me the opportunity to carry out my thesis research in the lymphoma research project. This gave me the opportunity to collaborate with experts from many different disciplines. The thesis research focuses on the computational and statistical aspects of clinical lymphoma research, which is impossible to carry out without collaboration. I thank all my collaborators, especially *Michael Hummel*, *Harald Stein*, *Reiner Siebert*, *Markus Löffler*, *Lorenz Trümper*, *Dido Lenze*, *Maciej Rosolowski*, *Hilmar Berger*, *Dirk Hasenclever*, *Wolfram Klapper*, *Andreas Rosenwald*, *German Ott*, *Burkhard Hirsch*, *Anke Ehlers*, *Judith Dierlamm* and *Eva M. Murga Penas* for their contributions.

While working on this thesis I met many people who made contributions either direct or indirect via fruitful discussions. In particular I greatfully thank *Marion Rother*, *André Mäurer*, *Dennis Kostka*, *Stefanie Scheid*, *Florian Markowetz*, *Jochen Jäger*, *Juby Jacob*, *Inka Appel*, *Hannelore Kaspar*, *Corinna Unger*, *Karin Eberhart*, and *Claudio Lottaz*.

I am especially grateful to Claudia my girlfriend, my mother, and my whole family for encouraging me throughout the time it took to finish the present work.

Publications Most contents of this thesis have been published in peer-reviewed journals. The molecular definition of Burkitt lymphoma described in chapter 3 has been published in the *New England Journal of Medicine* [42]. The subsequent application of the Burkitt classifier to childhood lymphoma has been published in *Blood* [46]. The publication of pathway activation patterns from chapter 6 can be found in *Leukemia* [11]. The contents of chapter 5 can be found in a publication in *Haematologica* [23]. Furthermore, I presented the pathway activation patterns in plenary session at the 10th International Conference on Malignant Lymphoma, Lugano, Switzerland, 4-7 June, 2008 [1], and the Burkitt classifier from chapter 3 together with the

GCB/ABC classifier from chapter 5 in a session on molecular classifiers at the International Workshop on Aggressive Lymphoma in Göttingen, Germany, 12-15 September, 2007. A list of selected publication related to lymphoma, where I contributed during thesis research can be found in the appendix.

Figures Figure 1.2 has been taken from Wikipedia.org, where it is published under GNU Free Documentation License by Wikipedia user Mikael Häggström. Figures 1.1 and 1.4 have been downloaded from web page of the National Human Genome Research Institute http://www.genome.gov. They are published there with the remark: "All of the illustrations in the Talking Glossary of Genetics are freely available and may be used without special permission."

Stefan Bentink Berlin, June 2009

Thesis Contributions

Lymphoma is the fifth most frequent cancer in North America and Western Europe. This thesis is concerned with transcriptional profiling of diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL) using supervised and semi-supervised machine learning methodology. It investigates two aspects of lymphoma classification in detail.

Diagnosis of Burkitt lymphoma. The distinction of BL and DLBCL based on traditional diagnostic criteria is often imprecise. Expert pathologist disagree frequently. Nevertheless, an accurate diagnostic distinction is mandatory for treatment decision.

Functional Stratification. Traditional molecular biological inference is based on hypothesis-driven intervention (e.g. via mutagenesis or over-expression of genes) in cellular systems to gain insight into molecular mechanisms. However, human cancer cells in their natural environment are not accessible to interventional assays. Thus, clinical microarray studies predominantly provide purely observational data.

The contributions of the present work are:

- The introduction of the semi-supervised learning problem of core group extension. Starting from a small set of unambiguously diagnosed tumors, the problem is to find additional cases similar to the core group from an unlabeled pool of tumors without diagnosis.
- The development of an Expectation-Maximization (EM) based Algorithm to core group extension.
- The generation of a linear signature allowing a quantitative and reproducible diagnostic distinction of BL and DLBCL implementing the core group extension strategy.
- The development of a semi-supervised learning method allowing stratification of tumors from clinical microarray studies based on data from hypothesis-driven interventional cell line assays.
- The generation of a novel functional stratification of DLBCL.

Contents

Part I

Theory

Chapter 1

Introduction

1.1 Gene expression and differentiation

Transcription and translation. Life is organized in small units called cells. The average size of cells is in the dimension of several micrometers. The number of cells constituting a living organism ranges from one like in yeast up to several billion like in human. Biological processes in living organism are controlled at the level of individual cells, each containing a full copy of the organism's overall blueprint, the genome. A cell is organized on three levels;

1. an archive storing the genome,
2. a machinery that makes temporary copies of sections of the genome, and
3. the production units finally implementing the genome.

The archive and the production units are spatially separated. The temporary copies enable and control the flow of information from the archive to the production units. The genome is consists of deoxyribonucleic acid (DNA). The information flow from the DNA to the production sites operates via messenger ribonucleic acid (mRNA) molecules, which are copies of small stretches of the information encoded in the DNA. The final production units are the ribosomes. They synthesize proteins based on DNA-instructions delivered via mRNA molecules. Proteins are the realization of the information encoded in the DNA. In human more than 18.000 different proteins[1] exist and thus the DNA holds different sections, each encoding the instructions to produce another protein. We call a section of an organism's DNA encoding the instructions to synthesize a single protein a gene. The process implementing the information flow from DNA to protein is called gene expression, where the step from DNA to mRNA is called transcription, and the step from mRNA to protein is called translation. The information flow has also been described as the central dogma of molecular biology [19].

DNA. DNA is a polymeric molecule consisting of a sequence of the 4 nucleosides adenosine, guanosine, cytosine and thymidine connected via phosphate ester bounds.

[1] http://www.ensembl.org/Homo_sapiens/Info/Index

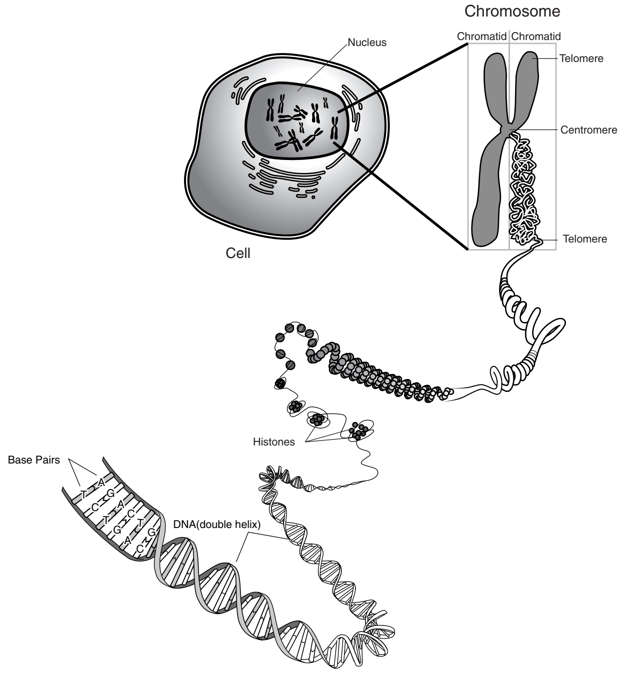

Figure 1.1: Chromosome. DNA is located in the nucleus of cells. It is packed in chromosomes. We denote the end of a chromosome as telomere and the central region as centromere. The centromere splits chromosomes asymmetrically in long arms (q) and short arms (p)

Each nucleoside consists of a ribose sugar and one of the 4 bases adenine (A), guanine (G) cytosine (C) and thymine (T). Given two independent DNA strands, base T of the first strand can form hydrogen bonds with base A of the second strand, while base G forms hydrogen bonds with base C. If two strands can pair with each other with respect to the A/T and G/C base-pairing rules they are complementary to each other. For example ACGT is complementary to TGCA. A genome consists of two complementary DNA strands that pair with each other via hydrogen bonds and arrange as the DNA double helix. The blueprint of life is encoded in the sequence of the four letters A,C,G and T. The human genome comprises approximately 4 billion base pairs splitting into 23 DNA-molecules organizing in 23 chromosomes (figure 1.1). Each cell except the germ cells contains two copies of the double strand.

Chromosome. A chromosome is a structure of proteins (histones) and a single or a pair of two homologous coiled DNA-molecules the chromatids. In its most condensed form during the so-called metaphase of the cell cycle a chromosome looks like illustrated in figure 1.1. Two chromatids pair with each other at the centromere. The centromere splits the chromatids into a longer and a shorter arm. We denote the long arm with q and the short arm with p. Stained microscopic preparations of metaphase chromosomes show characteristics lighter and darker banding patterns, which are numbered along the chromosome. The notation 18q21 denotes band 21 on the long arm of chromosome 18.

Differential gene expression and gene regulation. A human consists of billions of cells each containing the same genome. Thus each individual cell has the potential to run each biological process encoded in the genome. The cells are organized in organs and organ systems — accumulations of cells of specialized function. For example, a liver cell is committed to metabolism and detoxification, while a muscle cell contributes to the motility of an organism. Both, the liver and the muscle cell contain the same genetic information. Nevertheless, each of them only expresses those genes, which are necessary to perform the cell type-specific biological processes (either metabolism and detoxification or motion). We refer to this phenomenon as differential gene expression. Differential gene expression comprise not only qualitative aspect wether a gene is expressed or not. Gene expression can furthermore differ quantitatively, either in the amount of mRNA that is transcribed from DNA, or in the amount of protein that is translated from the mRNA. The process controlling when, where, and how much of a gene is expressed is called gene regulation.

Proliferation and differentiation. A human develops over a long period of time. Starting from a fertilized egg all of us have been an embryo, have been a baby, a child, and so on. We increase in size, and we develop more and more the abilities, which make us adult humans. From the perspective of cells this development requires two important biological processes: proliferation and differentiation. While proliferation increases the number of cells via duplication, differentiation specializes cells to certain biological tasks. Stem cells are cells that bear the ability to differentiate into many different kinds of more specialized cells. If a stem cell can develop into any kind of cell, it is called pluripotent. If the developmental potential of a stem cell is restricted to a certain branch, it is called multipotent. For example, the bone-marrow contains stem cells, which have the potential to differentiate into any kind of blood cell. They are called multipotent hematopoietic stem cells.

Cancer. During development cells arrange in organs, where they become specialized to certain biological functions. The high level of organization of the human body requires that each individual cell is exactly instructed when to proliferate, in which kind of cell to differentiate and when to stop proliferation. Cancer arises, if genetic damages decouple a cell from the regulatory circuits of cell proliferation and differentiation and instead give rise to uncontrolled proliferation. A genetic damage (hit) can render (transform) all kinds of cells into cancer cells. It is a common concept in the taxonomy of cancers to classify tumors with respect to the non-cancerous cell of origin of the cancer cell. Thus, we denote cancers arising from developing blood cells as hematopoietic cancers.

Hematopoietic cancer. Hematopoiesis is the process of blood cell formation. All human blood cells develop from the same type of cell, the multipotent hematopoietic stem cell (HSC). HSCs are located in the bone-marrow and give rise to distinct lineages of blood cell development. This development ends in differentiated blood cells like B-lymphocytes, plasma cells, T-lymphocyte or macrophages (figure 1.2). We can roughly classify three types of hematopoietic cancers: leukemias, lymphomas and myelomas. Leukemias arise from early myeloid or lymphoid precursors of the blood

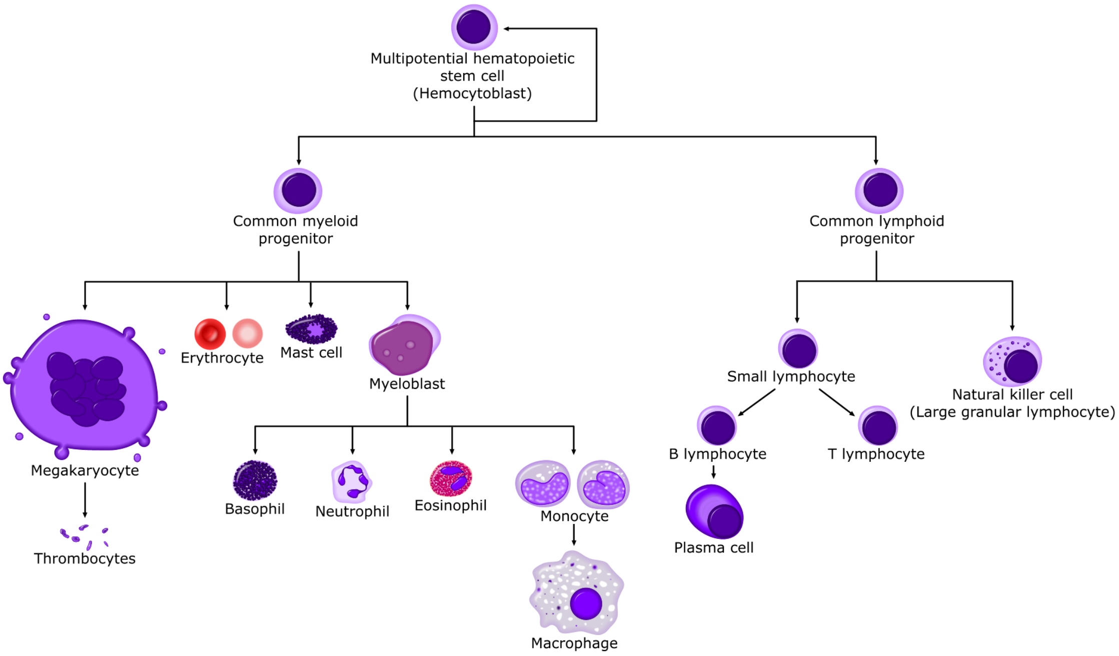
Multipotential hematopoietic stem cell (Hemocytoblast)
Common myeloid progenitor
Common lymphoid progenitor
Erythrocyte
Mast cell
Myeloblast
Megakaryocyte
Thrombocytes
Basophil
Neutrophil
Eosinophil
Monocyte
Macrophage
Small lymphocyte
Natural killer cell (Large granular lymphocyte)
B lymphocyte
T lymphocyte
Plasma cell

Figure 1.2: Hematopoiesis. Many different kinds of blood cells exist. Their differentiation process however starts from a common progenitor cell in the bone marrow, the multipotent hematopoietic stem cell (HSC).

cells. Myelomas arise from plasma cells representing a very late developmental stage of B-lymphocytes. In this thesis I will discuss the classification of lymphomas. Lymphomas comprise a heterogeneous pool of cancers arising from B- or T-lymphocytes (figure 1.2). The developmental stage of the cell of origin of lymphomas is between that of leukemias and myelomas.[64]

Lymphoma. Lymphoma is the fifth most frequent cancer in both men and women with about 19 per 100, 000 new cases in the United States per year [30], which is representative for ”Westernized” countries in North America and Europe [30]. The cell of origin is either a B- or T-lymphocyte harboring one or more characteristic rearrangements of the genome of the kind shown in figure 1.4 [50]. Lymphomas are classified into distinct disease entities. We classify them with respect to the cell of origin into T- and B-cell lymphomas. B-cell lymphomas comprise about 95% of all lymphomas. We further delineate Hodgkin lymphomas (HL) from all other lymphomas summarized under non-Hodgkin lymphomas (NHL). HL represents about 12% of all lymphomas. Thus, the majority of lymphomas are NHLs splitting up into various subtypes of different clinical and biological presentation. The diagnostic distinctions are based on a comprehensive set of genetic and histopathologic criteria collected within the ”World Health Organization (WHO) Classification of Tumors of Haematopoietic and Lymphoid Tissues” (current version: [87]). We distinguish indolent (slow-growing) from aggressive forms of lymphoma. This thesis focuses on two aggressive forms of non-Hodgkin lymphoma; diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma

(BL).

The majority of all lymphomas are of B-cell origin. The physiological role of native B-cells is to provide a receptor (B-cell receptor) specifically detecting pathogens like bacteria, viruses or parasites. Different B-cells provide different B-cell receptors, each specific to a different pathogen. Due to the great diversity of pathogenes, there is also a great diversity of B-cell receptors, which is generated during early B-cell development in the bone marrow (see figure 1.2). If a B-cell expresses the final version of its B-cell receptor and meets a pathogen that specifically binds to this receptor, the B-cell becomes activated and enters into the germinal center reaction. The germinal center is a formation of different kinds of hematopoietic cells establishing in lymphoid tissues upon the infection with a pathogen. We all know the process as swollen lymph nodes while suffering a common cold. During the germinal center reaction the B-cell receptor is further optimized, such that it detects the pathogen better. A series of physiological changes of the B-cells accompanies the process of B-cell activation and the germinal center reaction. It is widely accepted that many lymphomas arise from the B-cells at the different developmental stages of the germinal center reaction [36].

1.2 Diffuse large B-cell and Burkitt lymphoma

Diffuse large B-cell lymphoma. DLBCL is the most frequent lymphoma, accounting for 30-40% of all lymphoid neoplasms [89]. The diversity of DLBCL with respect to clinical presentation and outcome, and its pathological and biological heterogeneity suggest that DLBCL comprise several disease entities that may require different therapeutic approaches [32, 17, 88].

Burkitt lymphoma. BL comprises three different subtypes of aggressive B-cell lymphoma, the endemic BL, the sporadic BL, and the immunodeficiency BL [13]. Denis Burkitt first described the disease in 1958 as tumor involving the jaws in African children [15]. This variant is called endemic BL. Endemic BL is commonly observed in equatorial Africa, in children aged between 4 to 7 [13]. Furthermore it shows a strong association to an infection with the Epstein-Bar-Virus (EBV) [100]. The histopathological characteristics (microscopic appearance of tumor sections) of endemic BL are not limited to Africa. We also find tumors expressing the same characteristics but no EBV infection in many parts of the world. Tumors outside Africa showing the histopathological characteristics of endemic BL are called sporadic BL. Sporadic BL accounts for 1-2% of all adult lymphomas in western Europe and the United States [103]. In addition to sporadic BL one delineates a third variant of BL, which is associated to infections with the human immunodeficiency virus (HIV). We refer to it as immunodeficiency BL. Sporadic and immunodeficiency BL comprise only a very small fraction of adult NHL in non-endemic areas, however in childhood NHL sporadic BL comprises more than 30% of all cases. Thus BL is the most prominent

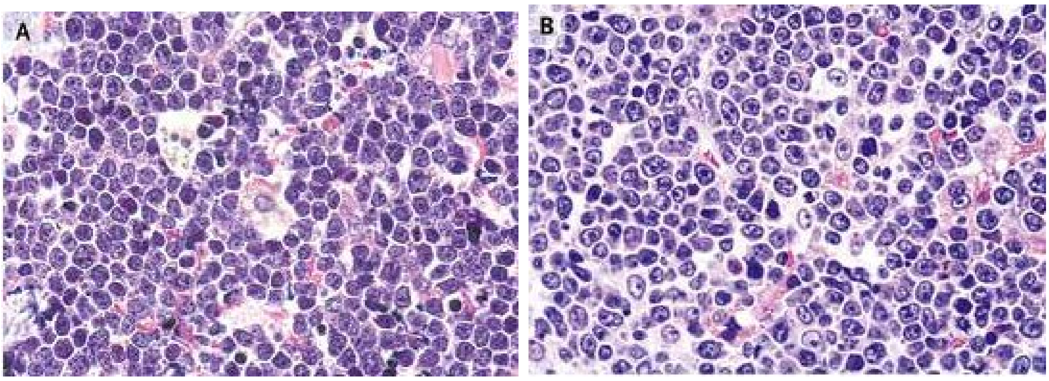

Figure 1.3: Histomorphologic appearance of DLBCL and BL. Shown are microscopic tissue sections stained with hematoxylin and eosin. Panel A shows the classic morphologic appearance of Burkitt lymphoma, which is called the ”starry sky” picture. It is caused by the white areas representing macrophages interspersing with the tumor cells stained in deep blue. Panel B shows in contrast the morphologic appearance of diffuse large-B-cell lymphoma. (Figure reproduced from [42])

childhood NHL in regions where BL is not endemic including western Europe and the United States [76].

Treatment of DLBCL and BL. DLBCL and BL are treated differently [35]. With the use of chemotherapy regimens that involve methotrexate and cytarabine, cure rates for sporadic Burkitt lymphoma approach 90 percent in children and 70 percent in adults [24]. Diffuse large-B-cell lymphoma, by contrast, is not only biologically but also clinically heterogeneous [32]. Treatment with a combination of chemotherapy based on cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) and the monoclonal antibody rituximab can induce lengthy remissions in many patients [17]. Approximately 30 percent of patients with diffuse large-B-cell lymphoma, however, have disease that is resistant to this treatment or relapse soon after receiving it [24, 68].

Diagnostic distinction of DLBCL and BL. The accurate differential diagnosis of BL and DLBCL is mandatory for treatment decision. The diagnostic distinction of the two entities is based on the combination of histopathology, immunohistochemistry and fluorescent in situ hybridization (FISH):

Histopathology. Histopathology is the microscopic inspection of tissue sections obtained by surgery or biopsy. For lymphoma diagnostics the microscope samples are prepared from the tumor material. The sections are stained according to standard protocols (e.g. with hematoxylin and eosin). The different lymphoma subtypes show specific microscopic characteristics. Diagnosis must be confirmed by experienced expert pathologists especially trained on this task. Figure 1.3 shows tissue sections representing the typical histomorphologic representation of BL (figure 1.3 A) and DLBCL (figure 1.3 B).

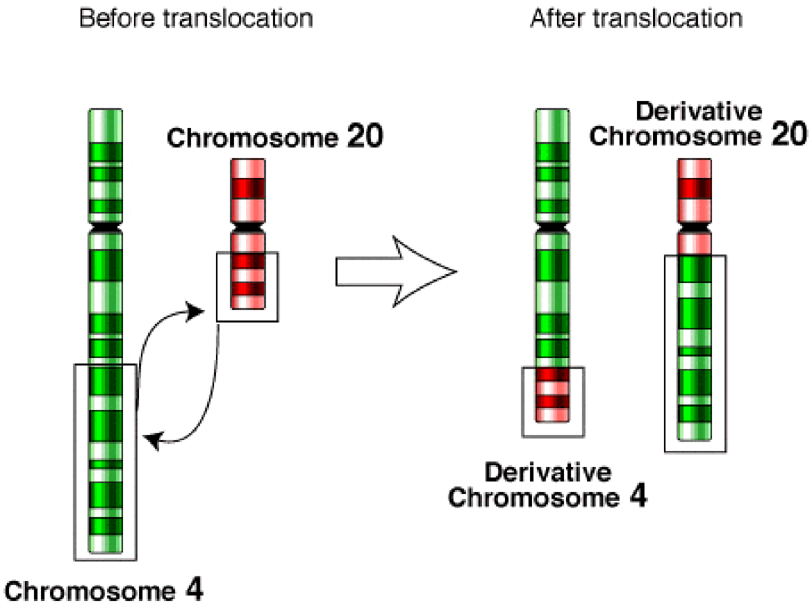

Figure 1.4: Translocation of the 4th and the 20th chromosome. The human genome is organized in 23 pairs of homologous chromosomes. We often observe chromosomal translocations in cancer cells. In the present example chromosome 4 and chromosome 20 have exchanged small fragments with each other. This mechanism can alter the expression of genes located on the rearranged chromosomal regions. Furthermore it can cause the fusion of genes located directly on the breakpoint region. The latter generates fusion genes with a novel functional spectrum.

Immunohistochemistry. Immunohistochemistry is a method detecting the expression of individual proteins in histological tissue sections. If we aim on detecting the expression of protein A in a tissue section, we need an antibody (anti-A), which is a small protein that specifically binds to protein A. The antibody anti-A is labeled with a dye, and added to the tissue section. Anti-A binds to protein A, and the dye accumulates in the regions of the tissue section were protein A is expressed. The accumulation of dye and thus the expression of protein A is assessed by microscopic inspection. As for the morphologic diagnosis, also the interpretation of immunohistochemical stainings requires expert pathologists.

Fluorescent in situ hybridization (FISH). Often the presence of a particular chromosomal rearrangement (figure 1.4) is characteristic for a diagnostic entity. Molecular probes (short stretches of DNA) can be designed, which are complementary to the chromosomal DNA sequences next to the breakpoint regions of the rearranged chromosomes. Fluorescent in-situ hybridization (FISH) is a diagnostic technique where the presence of specific chromosomal rearrangements is detected via fluorescent molecular probes and fluorescent microscopy.

According to the WHO [87] we can delineate BL from DLBCL by joining histopathology, immunohistochemistry and FISH analysis. We consider a tumor sample as Burkitt lymphoma, if

- it shows a characteristic BL histomorphology,
- the expression of the proteins CD20, BCL6 and CD10 can be detected by immunohistochemistry,
- the absence of the expression of the proteins BCL2 and CD5 can be confirmed by immunohistochemistry,
- the fraction of proliferating cells is $\geq 95\%$, as assessed by immunohistochemistry of the proliferation marker protein Ki-67,
- and the oncogene MYC on chromosome 8 is translocated to the chromosomal locus of the immunoglobulin heavy chain (IGH) on chromosome 14 or to either of the loci of the immunoglobulin light chains (IGK or IGL) on chromosomes 2 or 22.

Lymphomas carrying the MYC to IGH, IGK or IGL translocation we refer to as ”IG-MYC”. We detect them by fluorescence in situ hybridization. The most frequent translocation in IG-MYC lymphomas is that between chromosome 8 and chromosome 14. It is denoted translocation t(8;14). The oncogene MYC can furthermore be translocated to regions not involving any immunoglobulin locus. We refer to that kind of lymphoma as ”non-IG-MYC”. [42]

Even with the use of current diagnostic criteria, the distinction of BL and DLBCL is not precise; agreement among expert pathologists on the histopathological diagnosis of classic Burkitt lymphoma, an intermediate variant (atypical Burkitt lymphoma), and diffuse large B-cell lymphoma is only 53% [89, 56]. High throughput transcriptional profiling has become a valuable technology for molecular phenotyping of cancer that might help to improve on the diagnostic definition of Burkitt lymphoma.

1.3 Lymphoma transcriptional and genomic profiling

Transcriptional profiling. Cells are specialized to different biological functions depending on the tissue or organ they constitute. They express different genes according to their specialization. Today, we can record the information on expression levels of thousands of genes in a single assay. The success of genome sequencing projects has led to the development of the DNA microarray. This measurement platform has been developed in the mid 1990 (see e.g. [77]) and measures gene expression at the level of mRNA. We thus refer to its application as transcriptional profiling. Microarrays yield global gene expression finger prints of different organs, tissues, cells, and tumors. The diagnostic potential of transcriptional profiling has been first shown in 1999 [34] by using transcriptional profiles to discriminate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The technology entered lymphoma classification in 2000, when Alizadeh et al. [3] have proposed a novel concept to subdivide diffuse large B-cell lymphomas into an activated B-cell like (ABC) and germinal center B-cell like (GCB) transcriptional phenotype.

GCB-like and ABC-like DLBCL. Alizadeh et al. [3] stratify DLBCL into ABC-like (activated B-cell like) and GCB-like (germinal center B-cell like) lymphomas. The taxonomy was chosen, because the ABC-like lymphoma transcriptional profiles are similar to those of non-cancerous activated B-cells, and the GCB-like lymphoma transcriptional profiles are similar to non-cancerous germinal center B-cells. GCB-lymphomas represent B-cells during the germinal center reaction, while ABC may represent a post germinal center B-cell developmental stage [53]. The proposal is based on two assumptions:

1. The transcriptional profile of each lymphoma resembles the transcriptional profile of the non-cancerous cell it arose from, the cell of origin.
2. Native B-cells change their transcriptional profile while they pass though the developmental stages of B-cell activation and germinal center reaction. Different developmental stages express different transcriptional profiles.

Notably, patients with a lymphoma expressing the GCB-like transcriptional profile have a favorable prognosis as compared to the ABC-like lymphomas. The findings of Alizadeh et al. could be verified in subsequent lymphoma transcriptional profiling studies [62, 74].

Genomic profiling. A transcriptional profile is a snap shot of the physiological state of a tissue or cell. It can change over time for example in response to changing environmental conditions. In contrast, the construction plan encoded in the genomic DNA sequence remains constant and is the same for each kind of body cell. The situation changes, if mutations take place, which alter the genomic DNA sequence in individual cells. Mutations can occur at the level of individual base-pairs via insertions, deletions or substitutions of single nucleotides. Furthermore, parts of the chromosomes containing whole stretches of DNA can be lost, gained or translocated between different chromosomes. The latter we refer to as chromosomal mutations or aberrations. They are frequent in lymphoma. Translocations are balanced mutations as illustrated in figure 1.4. No change in the net amount of genomic DNA is associated with translocations. In contrast, gains and losses of parts of chromosomes are unbalanced mutations, since they represent a net gain or loss of genomic DNA.

Unbalanced mutations are accessible to genome-wide analysis technology. In comparative genomic hybridization (CGH) tumor DNA and normal DNA each labeled with a different fluorescent dye are added to normal chromosomes. The labeled tumor and normal DNA bind to (hybridize with) the chromosomal DNA with respect to the complementary base-pairing rules. Imbalances of the DNA amount between the normal and the tumor DNA are assessed by microscopic image processing of the label intensities of tumor and normal DNA hybrids with the chromosomes. Array CGH is a microarray-based version of CGH with an increased resolution to assess copy-number variations of chromosomes or parts thereof.

The presence of balanced translocations is not visible to genome-wide CGH methodology. Instead, translocation is tested separately by fluorescent in situ hybridization (FISH).

The GCB- and ABC-like subtypes of DLBCL tend to show distinct patterns of chromosomal aberrations. A genomic translocation involving the immunoglobulin heavy chain gene on chromosome 14 and the BCL2 gene on chromosome 18 – t(14;18)) – is unique to GCB lymphomas [45]. ABC lymphomas frequently show gains of regions on chromosome 3q and 18q with ”q” denoting the long arms of chromosomes 3 and 18, and losses on 6q. GCB lymphomas have frequently gained regions located on chromosome 12q [9]. Nevertheless, these are statistical associations. A definition of ABC and GCB via specific chromosomal abberations is not possible.

1.4 The MMML lymphoma data set

A comprehensive clinical study of a series of lymphoma samples involving state-of-the-art diagnostics (histopathology, immunohistochemistry and FISH) in conjunction with genome-wide transcriptional and genomic profiling requires joint efforts of experts from different disciplines. The research project ”Molecular Mechanisms in Malignant Lymphoma” (MMML) is a joint initiative of more than 15 contributing German groups supported by the ”Deutsche Krebshilfe”. The goal is a comprehensive characterization of aggressive B-cell lymphoma using genome-wide diagnostic approaches. Furthermore, the underlying molecular mechanisms in malignant lymphoma are investigated. This thesis focuses on the analysis of the transcriptional profiles from 220 mature aggressive B-cell lymphomas collected within the MMML project in its initial phase, and 36 additional cases included in a subsequent study of pediatric and childhood mature aggressive B-cell lymphoma. The study includes DLBCL, BL and cases that cannot be further subclassified to either of those categories. The following data has been collected:

Clinical and patient characteristics. For the majority of the cases information on sex, age and survival times is available. Furthermore, Ann Arbor stage at time of diagnosis is available, which is an established staging parameter in Hodgkin and Non-Hodgkin lymphoma discriminating between four levels of tumor spread (stage I-IV).

Consensus histopathological diagnosis. Each case has been evaluated by a panel of at least 6 expert hematopathologists to assign a consensus histopathological diagnosis of BL or DLBCL. Cases were labeled B-NHL high-grade, if the experts did not agree on a consensus diagnosis.

Immunohistochemistry of important diagnostic marker proteins. The presence of protein expression of the following diagnostic lymphoma markers has been assessed by immunohistochemistry: CD20, BCL6, CD10, BCL2 and CD5. Furthermore, the proliferation rate of the tumors has been assessed by scoring the fraction of cells expressing the proliferation marker Ki-67.

FISH analysis screening for the most relevant genomic translocations. The presence of chromosomal breakpoints and translocations has been assessed by fluorescent in situ hybridization. Cases were screened for the presence of IG-MYC, non-IG-MYC and IGH-BCL2 translocations. Furthermore, the presence of breakpoints at the locus of the BCL6-gene on chromosome 3 has been analyzed.[59, 94, 60]

Array CGH data. Array-based comparative genomic hybridization has been performed for 185 out of the 220 cases, applying an array platform that tiles the genome with 2799 DNA fragments (clones) [78, 29]. The data has been used to compute a genomic complexity score defining the number of chromosomal aberrations per case.

Transcriptional profiling. The mRNA of each tumor was extracted and subjected to transcriptional profiling with the U133A GeneChip platform provided by Affymetrix. A single U133A GeneChip provides quantitative measurements from 22283 individual features denoted as probe-sets. In the majority of cases, the transcriptional level of a human gene is measured by a single U133A probe-set. However, more complex human genes are measured by multiple alternative probe-sets.

1.5 Thesis organization

This theses is concerned with the statistical analysis of transcriptional profiles of lymphomas. It is organized as follows:

Introduction to molecular signatures. A DNA microarray is a powerful diagnostic device. Given a transcriptional profile as input, a molecular signature is a mathematical function with a diagnosis as output. This diagnosis can be a disease category, the presence of a genetic aberration or the activity of a particular oncogene. Chapter 2 will give an introduction to the statistical learning methodology and terminology behind molecular signatures.

Defining Burkitt lymphoma by a novel core-group extension approach. Even with the use of current diagnostic criteria, the distinction of Burkitt lymphoma and diffuse large B-cell lymphoma is not precise. Part II proposes a novel semi-supervised statistical learning strategy to derive a robust, quantitative and reproducible molecular signature of Burkitt lymphoma. In chapters 3 and 4 we will discuss two different implementations of this strategy.

GCB, ABC and their relation to gains of chromosome 18q. The classification of GCB and ABC is purely based on transcriptional profiling. In chapter 5 we test for the presence of these DLBCL subtypes in the MMML data set. Furthermore, we will analyze the relationship of the ABC transcriptional phenotype and a gain of chromosome 18q, which we frequently observe in ABC-like DLBCL.

Pathway activation patterns in mature aggressive B-cell lymphoma. Mature aggressive B-cell lymphomas are heterogenous with respect to their clinical presentation, their genetic make-up and their histopathological appearance. They possibly comprise several distinct disease entities arising from different mechanisms of disease. A controlled assay in a cell line allows us to design experiments in line with a certain hypothesis of disease mechanism. In chapter 6 we will use gene expression signatures generated from oncogene over-expression experiments in cell lines to predict oncogene activity on transcriptional profiles of mature aggressive B-cell lymphomas, yielding a novel pathway-based stratification of that disease.

Chapter 2

Molecular signatures and supervised learning

2.1 Molecular signatures as diagnostic devices

Establishing transcriptional profiles as diagnostic devices is a well established research direction [93, 102, 97]. The common situation is that we are given a set of gene expression profiles from tumors of n patients. For each individual patient we have a class label that assigns a (sub-)type to his/her tumor. The objective is to predict these labels from the observed transcriptional profiles. We are in the well defined framework of supervised machine/statistical learning [55]. Let us consider two different scenarios:

1. We are given a partition of a cancer into several distinct entities based on some gold standard diagnostic criteria. However, the diagnostic tests that define the gold-standard classification are expensive and elaborate.
2. A particular cancer can follow two completely different clinical courses. While one fraction of the patients responds well to the treatment, another fraction does not. So far, no test exists that predicts the therapy response prior to treatment.

In the first example, a single microarray can replace a whole panel of elaborate and expensive tests. In the second example, if biology - detectable by microarrays - causes the different response to treatment, patients would benefit from a microarray-driven stratification prior to treatment into aggressive and less aggressive therapy regimens. Both scenarios can be formalized as supervised classification problems. The purpose of the first one is diagnosis, and the purpose of the second one is prediction of treatment response.

Notation. We store n gene expression profiles each containing transcriptional levels of $i = 1, 2, ...p$ genes in a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Each row of $\mathbf{X}$ contains the expression profile of a single patient's tumor. The j-th patient's diagnosis is $y_i \in \mathcal{K} = \{k\}_{k=1}^{K}$ were K is the number of distinct disease entities. The set of diagnoses of all patients constitute the label vector $\mathbf{y} \in \mathbb{R}^n$. A clinical microarray study produces a data set $\mathcal{D}$ consisting of the label profile pairs $(\mathbf{X}, \mathbf{y}) = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_j, y_j)$.

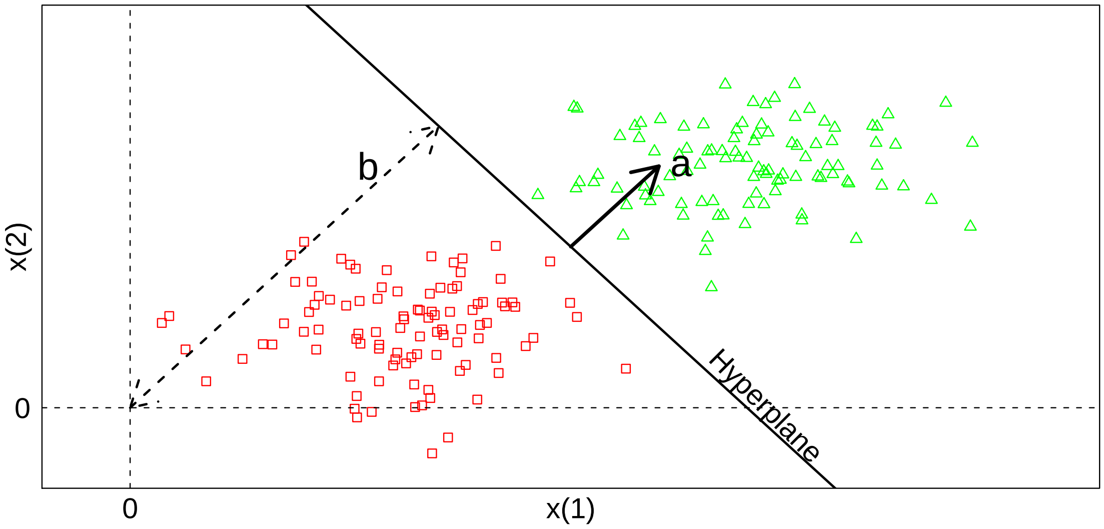

Figure 2.1: Linear separation of two dimensional data. Each red square and green triangle represents a sample in the two dimensional feature space. They can be separated by a hyperplane, which is a line in two dimensions. We can sufficiently describe it by the normal vector **a** and the offset b.

The goal is to learn from the example data set $\mathcal{D}$ a rule that correctly predicts a novel patient's tumor sub-type y from its expression profile $\mathbf{x}$.

Linear Signatures. Consider the p-dimensional space spanned by the expression values of p genes. We can classify a sample $\mathbf{x}$ represented as vector $\mathbf{x} = (x_1, x_2, ...x_p)$ of p gene expression values by the linear function

$$f(\mathbf{x}) = \sum_{i=1}^{p} a_i x_i + b. \tag{2.1}$$

If we are given a set of $j = 1, 2, ..., n$ samples from two different classes, where each sample is represented as vector $\mathbf{x}_j$ of gene expression values, equation 2.1 defines a classification rule based on the normal vector $\mathbf{a}$ and an offset b. The vector $\mathbf{a} = (a_1, a_2, ..., a_p)$ is perpendicular to a hyperplane separating the two classes. Varying b moves the hyperplane along $\mathbf{a}$ (figure 2.1). The sum in equation 2.1 projects p-dimensional data onto the line spanned by $\mathbf{a}$. If $f(\mathbf{x}) < 0$, a sample $\mathbf{x}$ is assigned to the first class and if not, it is assigned to the second class. The linear function 2.1 with its parameters $\mathbf{a}$ and b defines a linear signature. Supervised learning is the procedure to derive a parameter constellation for $\mathbf{a}$ and b from an example data set $\mathcal{D}$. Several distinct supervised learning algorithms have been proposed to derive linear signatures from microarray data, e.g. support vector machines, logistic regression and linear discriminant analysis (reviewed in [55]). We will now have a closer look at diagonal linear discriminant analysis (DLDA), which is a special case of discriminant analysis. Furthermore, we will restrict to the two-class case as shown in figure 2.1 throughout this thesis.

2.2 Diagonal linear discriminant analysis

The parameters $\mathbf{a}$ and b in equation 2.1 define a linear decision boundary between two classes. Considering figure 2.1, many boundaries exist and there is no unique constellation of the parameters $\mathbf{a}$ and b. Thus, we need an additional objective criterion to find a unique parameter constellation that finally defines a molecular signature. In linear discriminant analysis (LDA) [25] we therefore impose an explicit model on the shape of the data, i.e. we consider the data as mixture of multivariate Gaussian distributions with one component for each class k and a common covariance matrix. The parameters can be estimated directly from the data and have a unique solution, and in case of $K = 2$ yielding a single separating hyperplane between the two classes. The classification function $\mathcal{C}(\mathbf{x})$ of an LDA model assigns a novel sample $\mathbf{x}$ to the most likely class k

$$\mathcal{C}(\mathbf{x}) = argmin_k(\mathbf{x} - \mu_k)\hat{\Sigma}^{-1}(\mathbf{x} - \mu_k)', \tag{2.2}$$

where μ_k denotes the centroids of the k mixture components and $\hat{\Sigma}^{-1}$ denotes the inverse of the joint covariance matrix of the gene expression values. In transcriptional profiling we have more genes than samples and the inversion of the covariance matrix is thus not possible. In diagonal linear discriminant analysis (DLDA) we cope with that problem by imposing a common diagonal covariance matrix to all classes, which has zero entries outside the main diagonal and the variances of the individual genes on the diagonal. This matrix is always invertible, if the diagonal contains no zero entries. We define the DLDA classifier as

$$\mathcal{C}(\mathbf{x}) = argmin_k \sum_{i=1}^{p} \frac{(x_i - \mu_{i,k})^2}{\sigma_i^2}, \tag{2.3}$$

where x_i represent the expression of the i'th gene in a sample $\mathbf{x}$, $\mu_{i,k}$ represents the mean expression of the i'th gene in the k'th class, and σ_i^2 represents the within-class variance of gene i, which is the same for all classes. We estimate the p individual gene expression means μ_i and variances σ_i^2 as sample means and pooled variances on an example data set $\mathcal{D}$:

$$\hat{\mu}_{i,k} = \bar{x}_{i,k}, \tag{2.4}$$

$$\hat{\sigma}_i^2 = \frac{1}{(n-K)} \sum_{k=1}^{K} \sum_{j \in \mathcal{C}_k} (x_{i,j} - \bar{x}_{i,k})^2, \tag{2.5}$$

where $\mathcal{C}_k$ denotes the set of positive non-zero integers indexing the samples of the k'th class.

2.3 Performance

Misclassification error. A molecular signature predicts labels as output from transcriptional profiles as input. The misclassification error rate quantifies the performance of a molecular signature with respect to its ability to predict the labels correctly. We distinguish the misclassification error rate on training data from the misclassification error rate on test data. The training data set represents the samples that we use as example data to fit and select a model.

The intended use of a molecular signature is to predict labels of novel samples, i.e. to provide a diagnosis of new patients. The samples in the test data set serve as example for these new patients. It comprises samples that have been excluded from the model fitting step. The misclassification error rate on the test data is a realistic guess of the performance of the signature with respect to its diagnostic use. In contrast the training error rate is too optimistic.

Cross validation. We are given an example data set $\mathcal{D}$. In order to learn a signature from the example data, and to estimate its performance, we need to split $\mathcal{D}$ into a training set $\mathcal{D}_{train}$ and a test set $\mathcal{D}_{test}$. The training set serves as data for parameter estimation, and the test data allows performance evaluation. Each sample included in $\mathcal{D}_{test}$, decreases the number of samples in $\mathcal{D}_{train}$, thus decreasing the number of samples available for parameter estimation, making this step less precise. On the other hand, if we choose the number of samples in $\mathcal{D}_{test}$ small, the estimate of the misclassification error becomes less precise. An alternative procedure is cross validation, using each sample in $\mathcal{D}$ twice, once as training example and once as test example: In n-fold cross validation we randomly split $\mathcal{D}$ in n equally sized disjoint subsets: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_n\}$. Then we loop over $1, 2, ..., n$, each time defining $\mathcal{D} \backslash \mathcal{D}_n$ as training set and $\mathcal{D}_n$ as test set. In each loop we learn a molecular signature on $\mathcal{D} \backslash \mathcal{D}_n$ and predict the samples in $\mathcal{D}_n$ yielding an unbiased prediction of each sample in $\mathcal{D}$.

2.4 Gene selection

The regulation of transcription is complex. Each tissue runs biological processes specific to its function. The expression of genes that are not required in a tissue is shut down, while the same genes are activated in another tissue. The concerted action of pathway and tissue specific gene activity is known as differential gene expression. A molecular signature should discriminate between different disease-entities. Due to the complex mechanisms of gene regulation, we need to take into account that only a subset of genes is involved in the biological differences between the disease-entities we want to distinguish. Thus we expect only a subset of genes to be informative for classification. Indeed, the performance of a signature to correctly predict labels of novel samples can be improved by training the signature only on subsets of informative genes [25].

A common approach of gene selection is univariate thresholding, where we rank genes based on a univariate classification score. The Wilcoxon rank sum statistic and the

t-statistic are prominent choices in this context. The latter is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{2.6}$$

where $\bar{x}_1$ and $\bar{x}_2$ denote the means of the two classes we wish to compare and $\hat{\sigma}$ denote the pooled within standard deviation (see equation 2.5). Only genes with a univariate classification score above a certain threshold are included into the signature. Thus the threshold represents an additional model parameter controlling the number of genes of a molecular signature. The method of choice to select this threshold parameter is cross validation, where we select the threshold yielding a molecular signature with a minimal cross validated misclassification error.

2.5 Nearest shrunken centroids (NSC)

The *nearest shrunken centroids* method (NSC) [90] extends the framework of DLDA with a shrinkage approach to gene selection. In the context of transcriptional profiling the performance of NSC is in the same range as those of more sophisticated classification methods like logistic regression and support vector machines [52, 96]. The core of the NSC method is DLDA, which classifies a new sample $\mathbf{x}$ according to the closest of k class centroids, where the distances are standardized by the within class standard deviation. NSC modifies these distances to:

$$\delta_k(\mathbf{x}) = \sum_{i=1}^{p} \frac{(x_i - \bar{x}'_{i,k})^2}{(\hat{\sigma}_i + s_0)^2} - 2\ln \pi_k, \tag{2.7}$$

and classifies by

$$\mathcal{C}(x) = argmin_k \delta_k(x). \tag{2.8}$$

The distance in equation 2.7 is very similar to the DLDA classification function in equation 2.3; x_i denotes the expression of the i'th gene in a sample $\mathbf{x}$ and $\hat{\sigma}_i$ denotes the square root of the pooled variance of gene i (see equation 2.5). Nevertheless, there are three important differences:

- NSC calculates distances to shrunken centroids denoted by $\bar{x}'_{i,k}$ instead of original centroids. Shrinkage is the topic of the next paragraph.
- NSC takes class prior probabilities π_k into account. When estimating it by the relative size of the classes in the training data it gives higher classification weight to classes more frequent in the overall population.
- In NSC variances are regularized by a fudge factor s_0, which is a positive constant guarding against extremely small denominators as described in [92].

Shrinkage. Gene selection is implemented in a shrinkage approach: The class centroids are shrinked towards the overall centroid. This causes genes whose within class mean is close to the overall mean to be discarded from the analysis. Let

$$d_{i,k} = \frac{\bar{x}_{i,k} - \bar{x}_i}{m_k \cdot (\hat{\sigma}_i + s_0)}, \tag{2.9}$$

where $m_k = \sqrt{1/n_k + 1/n}$ makes the denominator in equation 2.9 equal to the estimated standard error of the numerator. Thus $d_{i,k}$ is a t-statistic for gene i comparing class k to the average class. The class centroids are shrinked toward the overall centroids, and the shrinkage is tuned by a parameter Δ:

$$d'_{i,k} = sign(d_{i,k})(|d_{i,k}| - \Delta)_+. \tag{2.10}$$

The expression $(|d_{i,k}| - \Delta)_+$ denotes the positive part of $|d_{i,k}| - \Delta$, which is zero if $|d_{i,k}| - \Delta \leq 0$ and $|d_{i,k}| - \Delta$ otherwise. We discard genes where $d'_{i,k}$ becomes zero. The resulting shrinked class centroid is

$$\bar{x}'_{i,k} = \bar{x}_i + m_k \cdot (\hat{\sigma}_i + s_0) d'_{ik}. \tag{2.11}$$

Class posterior probabilities. If we plug in $\bar{x}'_{i,k}$ in equation 2.7, we can compute class posterior probabilities in line with the suggestion of the authors of NSC [90]:

$$\hat{p}_k(\mathbf{x}) = \frac{e^{-\frac{1}{2}\delta_k(\mathbf{x})}}{\sum_{l=1}^{K} e^{-\frac{1}{2}\delta_l(\mathbf{x})}}. \tag{2.12}$$

Adaptive model selection. The previous paragraphs describe nearest shrunken centroids classifiers. While fitting an NSC classifier, the means, variances and class proportions can be directly computed from the training data. Let $\hat{\Sigma} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_p^2)$ denote the diagonal covariance matrix, where the element $\hat{\sigma}_i^2$ is the squared, pooled within class standard deviation of gene i (see equation 2.5) in the training data, and $\hat{\mu}_k = (\bar{x}_{1,k}, \bar{x}_{2,k}, ..., \bar{x}_{p,k})$ denote the k class centroids, where the element $\bar{x}_{i,k}$ is the training set mean of gene i in class k. The parameter $\hat{\pi}_k$ denotes the relative class sizes in the training data.

The shrinkage parameter Δ is selected from a set of multiple candidate values. The objective is to obtain a model of optimal performance on test data. Thus, Δ is selected via cross-validation as described on page 18. Given the random partition of $\mathcal{D}$ into the equally sized subsets $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n$, each $\mathcal{D} \backslash \mathcal{D}_n$ defines a training set to estimate $\Theta_{\mathcal{D} \backslash \mathcal{D}_n} = \{\hat{\Sigma}, \hat{\mu}_k, \hat{\pi}_k\}$. We select Δ according to:

$$\Delta_{opt} = argmin_{\Delta \in \mathbb{R}^+} \sum_n \sum_{j \in \mathcal{D}_n} l(\mathbf{x}_j, \mathcal{C}_\Delta(\mathbf{x}_j | \Theta_{\mathcal{D} \backslash \mathcal{D}_n}), y_j), \tag{2.13}$$

where $\mathbf{x}_j$ denotes a gene expression profile from the test set $\mathcal{D}_n$, $\mathcal{C}_\Delta(\mathbf{x}_j | \Theta_{\mathcal{D} \backslash \mathcal{D}_n})$ is its classification based on the training set classifier, and y_j is the true class label of $\mathbf{x}_j$. The function $l(\cdot)$ defines the 0/1 loss function, which is 1, if $\mathcal{C}_\Delta(\mathbf{x}_j | \Theta_{\mathcal{D} \backslash \mathcal{D}_n})$ does not correctly predict y_i and to 0 otherwise.

Validation The number of misclassifications is a common performance measure of a molecular classifier. However, we must not use the same data that has been used to derive the signature also for assessing its performance. This would yield and over-optimistic estimate as described in [81, 4]. To derive a molecular signature using NSC and assess its performance we can proceed in four steps:

- Randomly split data set into a training set and a test set.
- Put training set into a cross validation loop to select the optimal shrinkage parameter Δ.
- Use whole training set to derive a single NSC classifier using Δ as selected above.
- Apply this signature to the independent test data to assess its predictive performance.

If we split the data into a training set and a test set, the test set must not be used during the training phase. Thus we have fewer samples available to estimate the model parameters. This compromises the estimated model. Cross validation allows using each sample twice, once as test sample and once as training sample. Furthermore in each loop of a 10-fold cross validation we use 90% of the data to estimate the model parameter and only 10% to validate them. In contrast, if we split the data set half-half into a training and a test set, we can only use 50% of the data to estimate the model parameters. We expect a weaker model as compared the model estimated on 90% of the data. Thus cross-validation seems to be superior to the training set/test set approach of model validation. However, in each cross validation loop we train the signature anew. Thus the cross-validated misclassification error does not provide a performance measure of a single, fixed signature. It rather provides the performance of the learning algorithm. We still need an independent validation set to obtain a misclassification error of a given signature.

2.6 The compound covariate predictor

The compound covariate is a proposal of John Tukey [91]. He criticizes using a full regression model involving the estimation of covariances in a setting where we wish to relate several dozens of covariates to survival in a few hundred patients. A compound covariate is a linear combination of the basic covariates being studied, with each covariate having its own coefficient or weight in a linear combination [73, 82]. In transcriptional profiling the situation is more extreme than in the setting Tukey had in mind. Nonetheless, Tukey's proposal performs well in framework of transcriptional profiling [39, 101].

In line with [73] the proposal consists of a parameter estimation step and a gene selection step. It applies to two-class classification tasks. The value of the compound covariate for a sample $\mathbf{x}$ is:

$$c(\mathbf{x}) = \sum_{i \in \mathcal{G}} t_i x_i, \tag{2.14}$$

where t_i denotes the t-score of gene i comparing the two classes we wish to discriminate. G denotes a set of positive integers indexing the genes selected prior to classification. Radmacher et al. [73] suggest that $\mathcal{G}$ should contain the genes with the largest absolute t-scores comparing the transcriptional levels in the two classes. They suggest to either predefine the number of genes prior to classification, or to predefine a certain significance level α the genes must meet with respect to a t-test for differential expression between the classes.

A novel gene expression profile $\mathbf{x}$ is assigned to class $k = 1$ if

$$|\bar{c}^{(1)} - c(\mathbf{x})| < |\bar{c}^{(2)} - c(\mathbf{x})| \tag{2.15}$$

and to class $k = 2$ otherwise; $\bar{c}^{(1)}$ and $\bar{c}^{(1)}$ denote the mean values of $c(\mathbf{x})$ within the two classes. The decision rule defined by equation 2.15 classifies samples with respect to the nearest center $\bar{c}^{(k)}$.

Relationship between the compound covariate and DLDA. Equation 2.3 defines a diagonal linear discriminant classifier as:

$$\mathcal{C}(\mathbf{x}) = argmin_k \sum_{i=1}^{p} \frac{(x_i - \bar{x}_{i,k})^2}{\hat{\sigma}_i^2}.$$

In the two-class classification problem one assigns a sample $\mathbf{x}$ to class $k = 1$, if

$$\sum_{i=1}^{p} \frac{(x_i - \bar{x}_{i,1})^2}{\hat{\sigma}_i^2} < \sum_{i=1}^{p} \frac{(x_i - \bar{x}_{i,2})^2}{\hat{\sigma}_i^2}, \tag{2.16}$$

and thus if

$$\sum_{i=1}^{p} \frac{\bar{x}_{i,1} - \bar{x}_{i,2}}{\hat{\sigma}_i^2}(x_i - \frac{\bar{x}_{i,1} + \bar{x}_{i,2}}{2}) > 0. \tag{2.17}$$

The quotient $(\bar{x}_{i,1} - \bar{x}_{i,2})/\hat{\sigma}_i^2$ defines a normal vector (gene weights), which is multiplied with gene expression values x_i shifted by an offset $(\bar{x}_{i,1} + \bar{x}_{i,2})/2$.

The weights of the compound covariate predictor are defined by $t_i = \sqrt{n}(\bar{x}_{i,1} - \bar{x}_{i,2})/\hat{\sigma}_i$. Thus we can write equation 2.14 as

$$c(\mathbf{x}) = \sqrt{n} \sum_{i=1}^{p} \frac{\bar{x}_{i,1} - \bar{x}_{i,2}}{\hat{\sigma}_i} x_i. \tag{2.18}$$

Both DLDA and compound covariate predictors are based on the computation of a linear combination of gene expression values $s = a_1x_1 + a_2x_2 + ... + a_px_p$. However, DLDA uses variances, while the compound covariate uses the standard deviation in the denominator. Note, DLDA and compound covariate predictors estimate the weights a_i separately gene by gene and not jointly involving the estimation of covariances. Thus both methods are in line the compound idea of John Tukey [91]. Throughout this thesis we will follow these ideas. We will start with a molecular signature for Burkitt lymphoma based on a nearest shrunken centroids classifier.

Part II

Burkitt lymphoma

Chapter 3

A molecular definition of Burkitt lymphoma

The distinction between Burkitt lymphoma (BL) and diffuse large-B-cell lymphoma is not reliably reproducible with the use of the current criteria of morphology, immunophenotype, and genetic abnormalities [56, 89]. The Burkitt translocation or its variants, which juxtapose the locus of the MYC oncogene and one of the three immunoglobulin (IG) loci, are present in almost all Burkitt lymphomas [38, 14]. Nevertheless, MYC translocations are not specific for Burkitt's lymphoma since they also occur in other lymphomas, including diffuse large-B-cell lymphoma (DLBCL). In the latter, chromosomal breakpoints at the MYC locus are recurrently associated with non-IG partner loci and complex chromosomal alterations [5, 47, 48, 6, 26, 54, 33, 65].

The imprecise distinction between BL and DLBCL on diagnosis may lead to the inadequate treatment of some patients with a mature aggressive B-cell lymphoma. Studies involving gene-expression profiling indicate that DLBCL comprise two or more biologic subgroups with different clinical behaviors [3, 74, 62, 9]. Prior to our publication [42], and the one of Dave et al. [20], no signature of gene expression that distinguishes BL from DLBCL has been established.

In this chapter we derive a molecular signature for the differential diagnosis of Burkitt lymphoma and diffuse large B cell lymphoma. The analysis is based on 220 specimens of mature aggressive B-cell lymphomas including DLBCL, BL and unclassifiable cases. They have been collected and subjected to transcriptional profiling within the research project "Molecular Mechanisms in Malignant Lymphoma" (MMML, see section 1.4). The cases have been comprehensively characterized with respect to the WHO diagnostic criteria for Burkitt lymphoma [87]. These are

- the presence of a characteristic histomorphology,
- the presence of protein expression of CD20, BCL6 and CD10, and the absence of BCL2 and CD5 respectively,
- a fraction of proliferating (Ki-67 protein expressing) cells $\geq 95\%$,
- and the presence of an IG-MYC genetic translocation.

Table 3.1: Diagnostic criteria for Burkitt lymphoma. The differential diagnosis between Burkitt lymphoma and diffuse large B-cell lymphoma is based on a panel of histomorphological, immunohistochemical and genetic criteria. This table shows the frequencies of the diagnostic criteria in the MMML data set.

Criterion	Present	Absent	no classification
Burkitt or Burkitt-like histomorphology	36	165	19
IG-MYC chromosomal translocation	59	155	6
CD10 protein expression	94	115	11
BCL6 protein expression	165	33	22
BCL2 protein expression	153	61	6
CD20 protein expression	212	5	3
CD5 protein expression	22	184	16
Fraction of Ki-67 protein expressing cells$\geq$95%	51	162	7

Table 3.1 shows the frequencies of the individual WHO-BL criteria in the 220 MMML lymphoma samples. Only 8 of the 220 lymphomas satisfy all criteria at the same time. Thus, a stringent application of the WHO classification scheme labels only 8 of the tumor samples as BL. Nevertheless, expert pathologists expect more BL cases hidden in the pool of 212 remaining cases that have not been classified as BL by the WHO criteria. The problem is, which of the remaining cases are Burkitt lymphomas. Note, even among expert pathologists the agreement on the histopathological diagnosis of classic Burkitt lymphoma, atypical Burkitt lymphoma, and diffuse large B-cell lymphoma is only 53% [89, 56]. In order to achieve a quantitative and more reliable diagnostic distinction between Burkitt lymphoma and diffuse large B-cell lymphoma, we now derive a molecular signature.

3.1 Core group extension

3.1.1 Problem

In supervised learning as described in chapter 2, there is no doubt on the class labels, and they serve as certain indication for training, model selection and evaluation of a molecular signature. The present scenario is different. We have a data set $\mathcal{D}$ containing 220 lymphomas. 8 of them are labeled as BL according to the WHO, and 212 remain unlabeled, due to inconclusive diagnostic criteria. However the unlabeled cases might contain further BLs. Let us denote the set of 8 certain BL cases with C for core group, the 212 remaining cases with R, and the Burkitt cases hidden in R with E for extension. The set $(C \cup E)$ denotes the extended core group and the set

Table 3.2: Subsets in data set $\mathcal{D}$.

Name	Description	Number
$\mathcal{D} = C \cup R$	Overall data set	220
C	Core group of Burkitt lymphomas satisfying all WHO criteria	8
R	Remaining cases that do not meet the WHO criteria for Burkitt lymphoma ($C \cap R = \{\varnothing\}$)	212
$E \subset R$	Hidden Burkitt lymphomas in R that do not meet the WHO criteria	unkown
$(C \cup E)$	Set of all Burkitt lymphomas	unknown $(8 + \#E)$
$(R \backslash E)$	Set of non-Burkitt lymphomas	unknown $(212 - \#E)$

$(R\backslash E)$ denotes the remaining cases without E. We know the cases in C and in R, but not the cases in E. Our goal is to identify E. Furthermore we want to learn a molecular signature that classifies novel cases as members of the sets $(C \cup E)$ or $(R\backslash E)$. In terms of the Burkitt classification problem $(C \cup E)$ denotes "true", molecular Burkitt lymphomas and $(R\backslash E)$ denotes non-molecular Burkitt lymphomas. Table 3.2 summarizes the notation. Our problem is related (but not identical) to the well studied problem of supervised classification. We consider it as core group extension (COGE) problem.

3.1.2 Naive approach

In order to implement an algorithm for core group extension we use the supervised Nearest Shrunken Centroids (NSC) approach for the training step without any modification. However, we will deviate from supervised analysis to select and evaluate the final molecular signature.

Training. NSC starts with defining a number (default 30) of possible threshold parameters Δ yielding signatures with different numbers of genes. The candidate values for Δ are chosen to be equally spaced between 0 (no shrinkage, all genes remain in the signature) and an upper limit, which removes all genes from the classifier. The upper limit of Δ is chosen such that the shrinked distance to the overall centroid is exactly zero for the top ranking gene i in absolute value of equation 2.9.

In core group extension we train a two class NSC classifier for each candidate value of Δ. We consider the set C as class one and the set R as class two. In line with equation 2.7

$$\delta_C^{\Delta}(\mathbf{x}) = \sum_{i=1}^{p} \frac{(x_i - \bar{x}'_{i,C})^2}{(\hat{\sigma}_i + s_0)^2} - 2 \ln \pi_C, \tag{3.1}$$

and

$$\delta_R^{\Delta}(\mathbf{x}) = \sum_{i=1}^{p} \frac{(x_i - \bar{x}'_{i,R})^2}{(\hat{\sigma}_i + s_0)^2} - 2\ln \pi_R, \tag{3.2}$$

The parameters π_C and π_R denote the expected proportions of the classes in the population. We don't know the proportions in the population, and the present data set does not represent the proportions in the populations. Thus we set $\pi_C = \pi_R = \frac{1}{2}$. The pooled within-class standard deviation of gene i in C and R is

$$\hat{\sigma}_i^2 = \frac{1}{(n-2)}[\sum_{j\in C}(x_{ij} - \bar{x}_{i,C})^2 + \sum_{j\in R}(x_{ij} - \bar{x}_{i,R})^2]. \tag{3.3}$$

In line with equation 2.12 we compute a core group posterior probability

$$\hat{p}_C^{\Delta}(\mathbf{x}) = \frac{e^{-\frac{1}{2}\delta_C^{\Delta}(\mathbf{x})}}{e^{-\frac{1}{2}\delta_C^{\Delta}(\mathbf{x})} + e^{-\frac{1}{2}\delta_R^{\Delta}(\mathbf{x})}}. \tag{3.4}$$

We refer to $\hat{p}_C^{\Delta}(\mathbf{x})$ as core group index or molecular Burkitt index. In the present problem C is a small but homogeneous group of "true" Burkitt lymphomas, and R is a large and heterogenous pool of all other lymphomas in $\mathcal{D}$. We expect R to contain additional Burkitt lymphomas denoted as hidden set E, but with $\#E \ll \#R$. We expect C to be representative for all Burkitt lymphomas $C \cup E$, and that R has almost the population characteristics of $R\backslash E$. In conclusion, we naively consider the NSC signature trained on the classification task C versus R as signature discriminating between $C \cup E$ and $R\backslash E$. Thus

$$\hat{p}_{C\cup E}^{\Delta}(\mathbf{x}) \approx \hat{p}_C^{\Delta}(\mathbf{x}). \tag{3.5}$$

Classification. Standard NSC uses the classification function

$$\mathcal{C}^{(C,R)}(\mathbf{x}) = argmin_k \delta_k(\mathbf{x}), \tag{3.6}$$

with $k \in \{C, R\}$ and $\delta_k(\mathbf{x})$ as defined in equations 3.1 and 3.2. Our goal is to derive a molecular signature discriminating between $C \cup E$ and $R\backslash E$. We naively compute $\mathcal{C}^{(C,R)}(\mathbf{x})$ instead. However, in order to express our belief that there is uncertainty in the naive classification, we make use of the molecular Burkitt index $\hat{p}_C^{\Delta}(\mathbf{x})$ and introduce an intermediate gray zone of unclassified samples. We classify the lymphomas in $\mathcal{D}$ as molecular Burkitt lymphomas (mBL) and non-molecular Burkitt lymphomas (non-mBL) by

$$\mathcal{C}_{Burkitt}^{\Delta}(\mathbf{x}) = \begin{cases} \text{mBL} & \text{if } \hat{p}_C^{\Delta}(\mathbf{x}) > 0.95 \\ \text{non-mBL} & \text{if } \hat{p}_C^{\Delta}(\mathbf{x}) < 0.05 \\ \text{intermediate} & \text{if } 0.05 \geq \hat{p}_C^{\Delta}(\mathbf{x}) \geq 0.95. \end{cases} \tag{3.7}$$

Model selection. We run the standard NSC approach with m (default: $m = 30$) different values for the threshold parameter Δ yielding m different classification models. We subject each of them to cross validation as described in section 2.5. In standard NSC we choose Δ_{opt} out of $\{\Delta_m\}_{m=1}^{30}$ by minimizing the cross validated misclassification error rate. In the core group extension scenario we do not have the misclassification as performance measure, because most class labels are unknown. At this point we deviate from standard NSC and select Δ_{opt} from $\{\Delta_m\}_{m=1}^{30}$ while maximizing sparseness and core group sensitivity of the resulting model, i.e. we choose the signature, which

- contains a minimal number of genes,
- while classifying all samples in C as molecular Burkitt lymphoma with $\hat{p}_C^{\Delta}(\mathbf{x}) > 0.95$.

Solution. We plug in Δ_{opt} in $\mathcal{C}_{Burkitt}^{\Delta}(\mathbf{x})$ (equation 3.7) and assign the lymphomas to the sets E and $R \backslash E$:

- E: We assign all samples in R classified as mBL to the set E denoting the core extension.
- $R \backslash E$: We assign all samples in R classified as non-mBL to the set $R \backslash E$ denoting a heterogenous pool of other lymphomas purified from the molecular Burkitt lymphomas.
- Intermediate: We do not assign the intermediate cases to any of the sets mentioned above.

Trivial solution. It is important to note that there is always at least one solution meeting the model selection criteria, which we denote as trivial solution. If we shrink such that no gene remains in the signature, $\delta_k^{\Delta_{opt}}(\mathbf{x})$ in equations 3.1 and 3.2 purely depend on π_C and π_R. If we additionally define $\pi_C = 1$ and $\pi_R = 0$, all samples become a molecular Burkitt lymphoma with $\hat{p}_C^{\Delta}(\mathbf{x}) > 0.95$. We do not consider this solution as useful, however we keep in mind that it exists, and that it meets our model selection criteria.

Evaluation The parameter Δ_{opt} defines an NSC classification model. A common approach in supervised learning is to evaluate such model by its ability to predict gold-standard labels correctly (performance). Here we don't know the class labels except those of the core group, and we cannot evaluate our signature by performance. Instead we evaluate the selected model based on different considerations: The final solution of the core group extension algorithm depends on the initial choice of the cases that enter the core group. Due to possible disagreement between expert pathologists [89, 56], this choice can vary. Thus, we expect a valid core group extension solution at least to be robust against uncertainty in the composition of the core group. In order to assess this robustness, we select Δ_{opt}, then we choose with replacement a certain number B (here $B = 1000$) of bootstrap samples of the core group C_{boot} and combine them with the remaining cases R. This results in B bootstrapped (training) data sets $\mathcal{D}_{boot} = C_{boot} \cup R$. We recompute the signature defined by Δ_{opt} from each bootstrapped data set resulting in B posteriors $\{\hat{p}_{C_{boot},l}^{\Delta_{opt}}(\mathbf{x}_j)\}_l^B$ for each sample $j \in \mathcal{D}$.

The variation of the $\{\hat{p}^{\Delta_{opt}}_{C_{boot},l}(\mathbf{x}_j)\}^B_l$ for a sample j indicates the variability of the signature with respect to uncertainty of the core group.

3.2 Results

Data preparation

In order to derive a molecular signature for Burkitt lymphoma we study gene expression profiles of the 220 lymphomas in $\mathcal{D}$. One case has been measured twice so that we have 221 gene expression profiles. The Affymetrix HGU133A platform used to generate the transcriptional profiles holds 22283 probe-sets. Eight lymphomas satisfy the WHO Burkitt criteria. We assign them to C, and split the remaining cases randomly into training ($n = 105$ cases) and test set ($n = 107$ cases) balanced for the histological diagnosis (see table 3.1). We assign the cases in the training set to R, run core group extension on $C \cup R$, and use the test set to evaluate the signature on independent data.

Core group extension

Training We train the signature using the default parameters of the NSC as implemented in the statistical programming language R [72] (*pamr*, Prediction Analysis for Microarrays [37]). The label vector contains two class labels (C = core BL and R = non-core BL). We define a uniform prior over the two classes $\pi_C = \pi_R = \frac{1}{2}$ and compute a set of 30 different threshold values for the parameter Δ as described in section 3.1.2. We compute a model for each Δ. Table 3.3 shows the number of genes included in each of the 30 models.

Model selection In order to select Δ_{opt} we cross validation. We randomly split the training set into 8 equally sized buckets each containing one of the 8 core BL cases in C, and loop over the 8 buckets. In each loop, we set aside one of the buckets and train an NSC model for each level of Δ on the remaining 7 buckets. We use these models to predict the core group posteriors of the cases in the "left-out-bucket". This yields prediction for each training case. Based on this prediction we select the optimal core group extension model. Figure 3.1 shows the cross validated Burkitt index for different values of Δ. The dashed horizontal lines define borderlines between molecular Burkitt lymphomas (mBL, index > 0.95), non-molecular Burkitt lymphomas (non-mBL, index < 0.05), and intermediate cases ($0.05 \leq$ index ≤ 0.95). We select the 74 probe-set model, since it is the most sparse model with respect to the number of genes that classifies all cases in C as molecular Burkitt lymphoma.

Evaluation Shrinkage with $\Delta_{opt} \approx 3$ yields 74 probe sets remaining in the signature. Figure 3.2 visualizes the distribution of 1000 mBL-indices $\{\hat{p}^{\Delta_{opt}}_{C_{boot},l}(\mathbf{x}_j)\}^{1000}_l$ for each sample j resulting form B=1000 bootstrap classifiers obtained from C_{boot} versus

Table 3.3: Varying Δ. The table lists the number of genes selected for 30 candidate Burkitt classifiers that differ with respect to the level of the threshold parameter Δ.

Δ	n genes	Δ	n genes	Δ	n genes
0.0000000	22283	1.6687858	1620	3.3375717	30
0.1668786	18445	1.8356644	1162	3.5044502	18
0.3337572	14931	2.0025430	803	3.6713288	10
0.5006357	11926	2.1694216	547	3.8382074	6
0.6675143	9297	2.3363002	367	4.0050860	5
0.8343929	7226	2.5031787	237	4.1719646	4
1.0012715	5498	2.6700573	163	4.3388431	3
1.1681501	4141	2.8369359	103	4.5057217	2
1.3350287	3022	3.0038145	74	4.6726003	1
1.5019072	2251	3.1706931	47	4.8394789	0

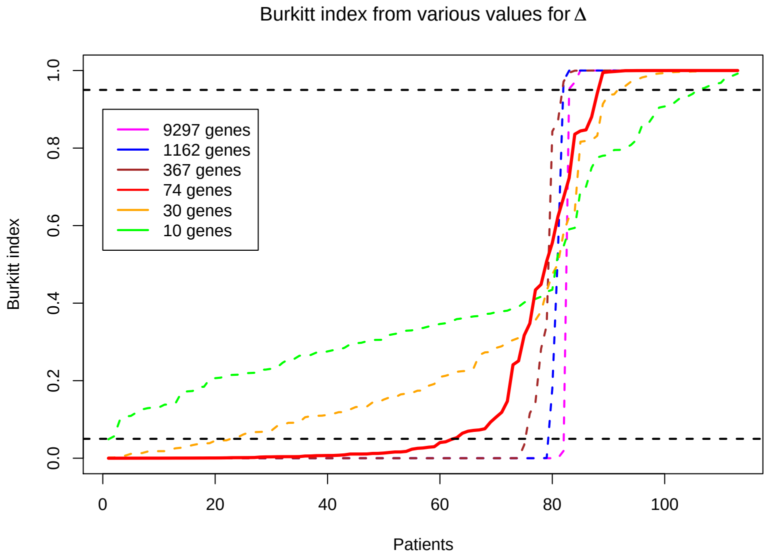

Figure 3.1: Varying Δ. The lines show the cross validated Burkitt index for different values of Δ. The dashed horizontal lines define cut off values deciding a case is a molecular Burkitt lymphoma (mBL, *index* > 0.95), a non-molecular Burkitt lymphoma (non-mBL, *index* < 0.05) or a case assigned an intermediate, unclassified state ($0.05 <$ *index* < 0.95). We select the 74 gene (probe-set) model as the molecular Burkitt classifier (red line). This model keeps the number genes small, while classifying core Burkitt lymphomas correctly.

R classification models. Box 1 contains a detailed explanation of the bootstrap results. Furthermore we assess robustness by counting how often an mBL patient (set E) changes into an non-mBL patient (set $R \backslash E$) during the bootstrap, and vice versa. The rate of such label changes is less than $3^0/_{00}$. This indicates that, even though robustness not explicitly entered in the model, the resulting model turns out to be remarkably robust. Relearning a classifier with Δ_{opt} on the complete training data yields the final mBL signature.

Figure 3.2 furthermore visualizes the results obtained from the application of the bootstrap and the final signature to the test set of 105 lymphomas. Even though the test set is not part of the learning and model selection step, the observations on the training data are in line with the test data. The proportions of mBL, intermediate and non-mBL cases are comparable, and the composition of the groups with respect to morphology and genetics are consistent (figure 3.2). Thus, the signature applies to independent test data. For the following discussion of the biological and clinical findings we merge the training and the test set, yielding 8 core Burkitt lymphomas C, 36 additional molecular Burkitt lymphomas E, 128 non-molecular Burkitt lymphomas $R \backslash E$, and 48 intermediate cases.

Box 1: Details on figure 3.2 Each bar of the plot at the top of the figure describes the average number of chromosomal abberations of the tumors measured by array comparative genomic hybridization. We refer to that quantity as genomic complexity. The dotted horizontal line represents the mean complexity of each group. The second plot shows the stability of the core group extension with respect to random perturbations of the core group data (bootstrap analysis). The frequency of the perturbed mBL-signature index scores (from 0 to 1, bottom of plot to top) obtained from 1000 runs of the algorithm is indicated by color (very low frequency, orange; low, yellow; medium, green; high, blue; very high, red). The vertical lines delineate the three groups of lymphomas (mBL, intermediate, and non-mBL) - as well as the core group of cases - and the dashed horizontal lines indicate the index-score cutoffs defining the mBL group (0.95) and the non-mBL group (0.05). Among the mBL cases, the index score is close to 1 for all bootstrap perturbations, whereas in the non-mBL group it is near 0, demonstrating the stability of the signature. The mBL-signature index scores resulting from the non-bootstrapped signatures are represented as a dashed curve. Below, the heat map shows the gene-expression levels of the 74 mBL-signature probe-sets, with 1 probe-set shown per row. Bright blue indicates a low level of expression (3 SD below the average of all cases), bright yellow indicates a high level of expression (3 SD above the average), and black the average level of expression across all samples. The cases are ordered from left to right on the basis of decreasing mBL-signature index score, given below the heat map. Green represents a high index score (mBL), and red a low index score (non-mBL). The color gradient in the intermediate group highlights the continuous transition of the index score between the mBL and non-mBL cases. The MYC translocation partners are shown according to type: IG-MYC fusion (bright green), non-IG-MYC fusion (dark

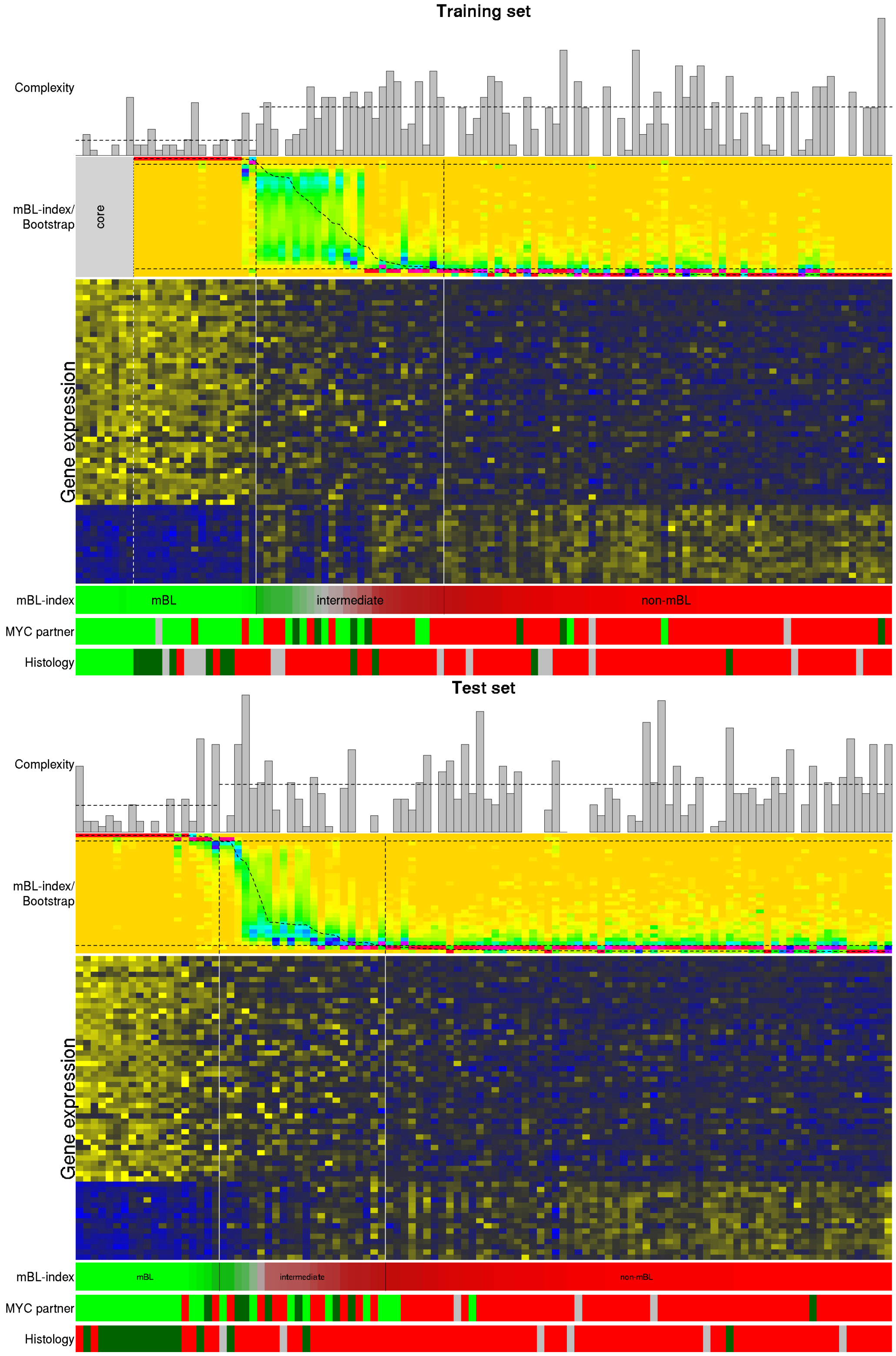

Figure 3.2: Identification by core group extension of cases with an mBL signature. The figure compiles the result of the core group extension algorithm and the most important biological findings on the training and test set. Box 1 describes the figure in detail. (Figure reproduced from [42])

green), MYC-breakpoint absent (red), and no data available (gray). Finally, the histologic diagnosis is shown at the bottom. Bright green indicates Burkitt lymphoma in the core group; dark green, atypical Burkitt lymphoma; red, diffuse large-B-cell lymphoma; and gray, unclassifiable mature aggressive B-cell lymphoma.

Biological and clinical findings

Lymphomas expressing the mBL Signature The 74-gene model identifies 36 lymphomas with an mBL-signature index score of greater than 0.95 (set E) in addition to the 8 core Burkitt lymphomas (C), for a total of 44 mBL cases ($C \cup E$). The 8 core cases are similar to the additional 36 cases with regard to age of the patients and genetic features, as well as clinical course. Of the additional 36 mBL cases, 21 are categorized as atypical Burkitt lymphomas by expert pathologists because of their Burkitt-like morphology or their immunophenotype. It is important to note that 11 of the 36 cases in E have the distinctive morphologic appearance of diffuse large-B-cell lymphoma. The remaining four mBL cases have the morphologic appearance of mature aggressive B-cell lymphoma but can not be further classified histologically. With regard to the immunophenotype, the BL markers CD10 and BCL6 are consistently expressed in 42 and 39 mBL cases. BCL2 typically absent in BL is detectable at a low level in seven of the mBL cases and at a high level in two if them.

B-Cell Lymphomas not expressing the mBL Signature Of all 220 lymphomas, 176 have an mBL-signature index score of less than 0.95. Of these 176 cases, 128 have an mBL-signature index score of less than 0.05 and are assigned to the non-mBL group ($R \backslash E$). The remaining 48 cases have an mBL-signature index score between 0.05 and 0.95 and we do not assign them to the mBL or non-mBL group. We assign them to the intermediate group, representing the transition zone between the mBL and non-mBL groups.

With few exceptions (12 cases), the histologic diagnosis in the non-mBL cases is diffuse large-B-cell lymphoma. The histologic diagnosis of 39 (81 percent) of the intermediate cases is also diffuse large-B-cell lymphoma. Non-mBL and intermediate cases show strong concordance regarding age distribution, immunophenotype, growth fraction (Ki-67 score), and chromosomal complexity.

Genetic Aberrations and Gene Expression 43 mBL cases were tested for the presence of MYC translocations by using fluorescence in situ hybridization. All but five cases (88 percent) carry an IG-MYC fusion and one of these five have both non-IG-MYC and IGH-BCL2 fusions. In the 38 mBL cases with IG-MYC fusion, IGH-BCL2 fusion and BCL6 breakpoints are absent. The average number of chromosomal aberrations as revealed by array comparative genetic hybridization (chromosomal complexity score) is low in the 38 mBL cases with IG-MYC fusion but is high in the 5 mBL cases without IG-MYC fusion.

Table 3.4: Morphologic, immunohistochemical and Genetic characteristics of lymphomas.

Characteristic	**Lymphoma**							
	All		mBL		Non-mBL		Intermediate	
	number (percent[1])							
Total	220		44		128		48	
Age at diagnosis								
$<$ 60 years	100	(46)	40	(91)	39	(31)	21	(45)
$\geq$ 60 years	118	(54)	4	(9)	88	(69)	26	(55)
Sex								
Female	91	(43)	13	(30)	56	(45)	22	(47)
Male	123	(57)	30	(70)	68	(55)	25	(53)
Morphologic diagnosis								
Burkitt lymphoma (core group)	8	(4)	8	(18)	0	(0)	0	(0)
Atypical Burkitt lymphomas[2]	28	(13)	21	(48)	3	(2)	4	(8)
Diffuse large B-cell lymphoma	165	(75)	11	(25)	115	(90)	39	(81)
Mature aggressive B-cell lymphoma, unclassifiable	18	(8)	4	(9)	9	(7)	5	(10)
Burkitt leukemia–lymphoma[3]	1	(< 1)	0	(0)	0	(0)	0	(0)
CD10 expression[4]								
Absence	115	(55)	0	(0)	95	(79)	20	(43)
Presence	94	(45)	42	(100)	26	(21)	26	57)
BCL6 expression[4]								
Absence	33	(17)	0	(0)	26	(23)	7	(15)
Presence	165	(83)	39	(100)	87	(77)	39	(85)
BCL2 expression[4]								
Absence	61	(29)	33	(79)	20	(16)	8	(17)
Presence	153	(71)	9	(21)	104	(84)	40	(83)
Ki-67-score[4]								
$<$ 95%	162	(76)	15	(34)	107	(88)	40	(85)
$\geq$ 95%	51	(24)	29	(66)	15	(12)	7	(15)
MYC translocation[5]								
IG-MYC	59	(28)	38	(88)	5	(4)	16	(33)
Non-IG-MYC	15	(7)	1	(2)	4	(3)	10	(21)
MYC-negative	140	(65)	4	(9)	114	(93)	22	(46)
IGH-BCL2 fusion[5]								
Absent	192	(88)	43	(98)	111	(89)	38	(79)
Present	25	(12)	1	(2)	14	(11)	10	(21)
BCL6 breakpoint[5]								
Absent	177	(83)	43	(100)	94	(76)	40	(85)
Present	36	(17)	0	(0)	29	(24)	7	(15)
Chromosomal complexity score[6]								
Low $<$ 6%	74	(40)	31	(79)	30	(29)	13	(31)
High $\geq$ 6%	111	(60)	8	(21)	74	(71)	29	(69)

Genetic group								
MYC-simple	35	(17)	29	(76)	0	(0)	6	(13)
MYC-complex	33	(16)	5	(13)	9	(7)	19	(40)
MYC-negative	140	(67)	4	(11)	114	(93)	22	(47)

[1] Percentages refer to the number of cases that can be evaluated; data are not available for all cases. Percentages may not total 100 because of rounding. mBL denotes molecular Burkitt lymphoma.

[2] These lymphomas show a Burkitt-like morphology or a Burkitt morphology with an atypical immunohistochemical marker expression.

[3] A Burkitt leukemia is a leukemia, where the tumor cells show the characteristics of Burkitt lymphoma cells.

[4] The expression of CD10, BCL6, BCL2, and Ki-67 has been determined by immunohistochemical analysis.

[5] The MYC and BCL6 breakpoints and the IGH-BCL2 and MYC fusions have been analyzed by fluorescent in situ hybridization (FISH).

[6] Chromosomal complexity was determined with array-based comparative genomic hybridization.

The frequency of MYC breakpoints (regardless of translocation partner) is lower in the intermediate and non-mBL groups than in the mBL group. MYC breakpoints are common in the intermediate group, whereas they are less frequent in the non-mBL group. Non-IG partners are frequently involved in MYC translocation in both the intermediate group and the non-mBL group. Among the 35 MYC-positive intermediate and non-mBL cases, 16 have a concurrent IGH-BCL2 fusion, BCL6 breakpoint, or both. The chromosomal complexity score is significantly higher in the intermediate and non-mBL groups than in the mBL group regardless of the presence of MYC breakpoints or absence of MYC breakpoints.

On the basis of these data, we can distinguish three main cytogenetic groups within the mature aggressive B-cell lymphomas. The first we call ”MYC-simple”: lymphomas with IG-MYC fusions and a low chromosomal complexity score (<6) that do not have IGH-BCL2 fusions and BCL6 breakpoints. The second we call ”MYC-complex”: all lymphomas with non-IG-MYC fusions or all lymphomas with IG-MYC fusions that have a high chromosomal complexity score (>5), an IGH-BCL2 fusion, or BCL6 breakpoint, or any combination of these. The third we call ”MYC-negative”, comprising MYC-negative lymphomas. The mBL group predominantly consists of MYC-simple lymphomas; the non-mBL group predominantly consisted of MYC-negative lymphomas. In contrast, the intermediate group contained most of the MYC-complex cases but also occasional MYC-simple and several MYC-negative cases.

Molecular and Clinical Characteristics Clinical information on the patients is only available for a subset of the cases. However, no significant differences can

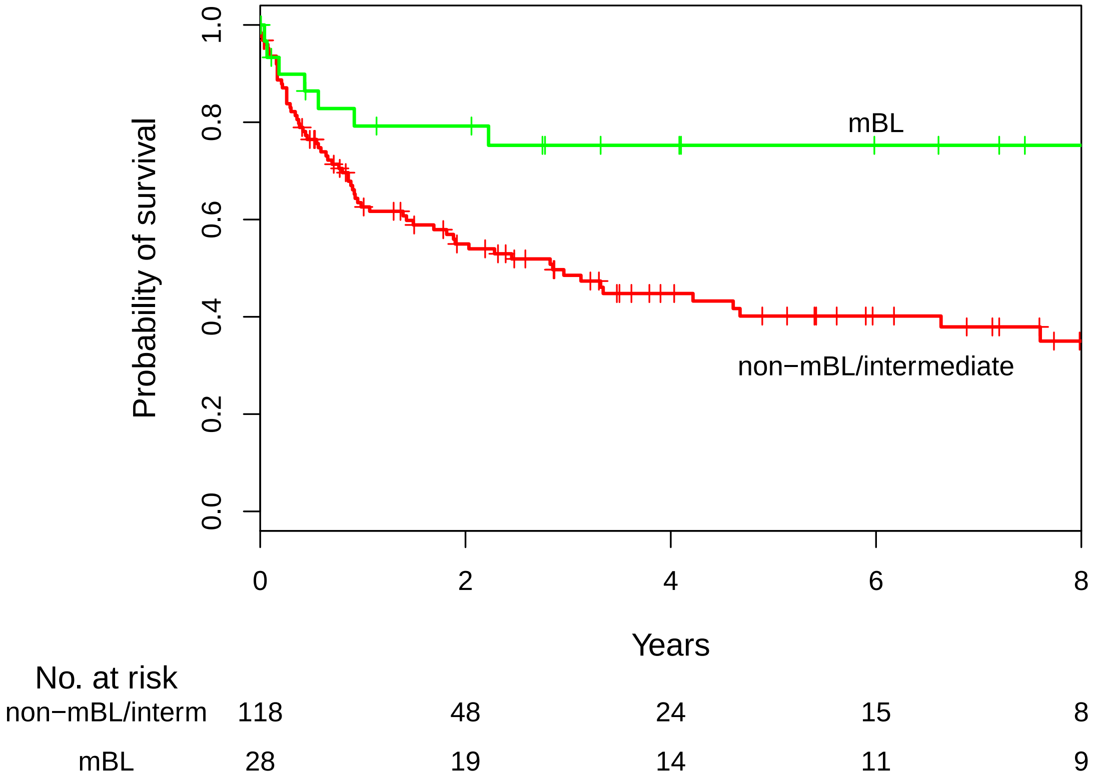

Figure 3.3: Kaplan-Meier estimates of survival according to the mBL signature. Overall survival among patients with positive for an mBL-signature is significantly greater than that among the patients with non-mBL or intermediate lymphoma (P=0.003 by the log-rank test [58, 67]). Tick marks denote patients alive at the time of last follow-up.

be observed regarding the morphologic characteristics, immunophenotype, or gene-expression pattern between the 146 patients with survival data available and the 74 patients without or with incomplete clinical information. Thus, patients with clinical information available do not differ from patients without clinical information available, and the analysis of the survival data from a subset of the patients is representative for the whole data set. We find, patients with lymphomas classified as mBL have better five-year survival rate than patients with non-mBL or intermediate lymphomas (75 percent vs. 39 percent, P=0.003 by log-rank test for different survival [58, 67], see figure 3.3). However, we know, patients with mBL are younger on average, which possibly explains the favorable diagnosis of mBL. Thus we model the survival of the patients in the different molecular subtypes by multivariate linear Cox regression [18]. This analysis takes further confounding factors into account. Here the confounding factors whose negative influence on survival is known are the age of patients (dichotomized in patients < 60 years and those ≥ 60 years) and the stage of the tumor (Ann Arbor stage dichotomized in I and II versus III and IV). The results of multivariate Cox-regression analysis show that the favorable outcome among pa-

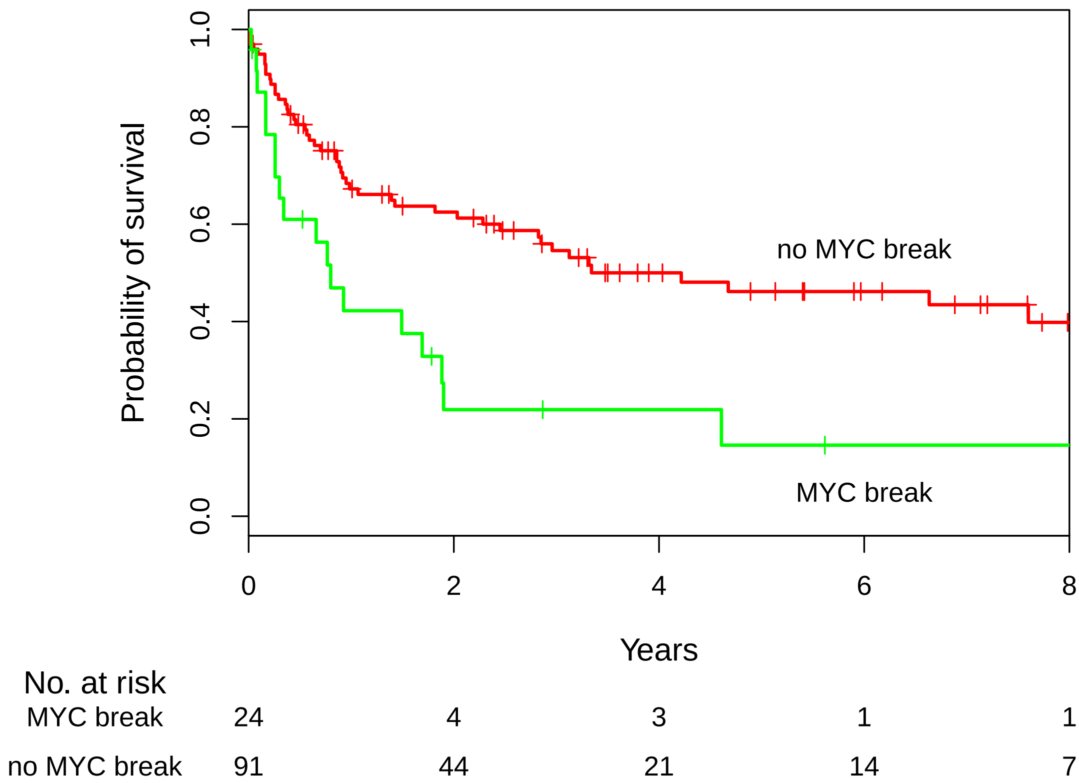

Figure 3.4: Kaplan-Meier estimates of survival according to the MYC break. Overall survival among non-mBL/intermediate patients with MYC breaks (IG-MYC and non-IG-MYC) is significantly worse than that among non-mBL/intermediate patients without a MYC break (P=0.005 by the log-rank test [58, 67]). Tick marks denote patients alive at the time of last follow-up.

tients with mBL can largely be explained by the onset of the disease at a young age and the limited stage of the disease (hazard ratio for death of non-mBL/intermediate lymphomas, 1.47; 95% CI: 0.57-3.81; P=0.43). However, we can show adverse effects of the MYC-breakpoint independent of age and stage among the 82 patients with non-mBL or intermediate lymphomas for whom clinical information is available. The presence of a MYC breakpoint, as compared to its absence, is associated with a poor five-year survival rate (15 percent vs. 44 percent, P=0.005 by log-rank test for different survival [58, 67], see figure 3.4). In Cox-regression restricted to the non-mBL and intermediate group, the presence of a MYC breakpoint - which occurs mainly in MYC-complex lymphomas - is associated with a significantly worse survival, independently of Ann Arbor stage, age, and the GCB/ABC signature, which we discuss in chapter 5 (hazard ratio for death of MYC-break positive lymphomas, 2.85; 95% CI: 1.43-5.68; P=0.003).

3.3 Validation on, and application to independent data sets

3.3.1 Pediatric mature aggressive B-cell lymphoma

Lymphomas are the third most common group of cancers in children and adolescent. Non-Hodgkin lymphoma (NHL), which account for approximately 60% of all lymphomas, represent 6% of all malignancies in children up to 14 years of age (German Childhood Cancer Registry, GCCR[1]). The spectrum of NHL occurring in children and adolescent differs strikingly from adults. Whereas indolent lymphomas are frequent in adults, the vast majority of lymphomas in children and adolescent are aggressive lymphomas, mainly mature aggressive B-cell lymphomas including particularly Burkitt lymphoma and diffuse large B-cell lymphoma.

With currently available combination chemotherapy for both BL and DLBCL an overall survival rate of 90% and more can be reached in children [99, 66]. In most pediatric study groups BL and DLBCL, although recognized by the WHO classification as distinct lymphoma entities, are currently treated according to the same treatment protocols in children [99, 66, 16]. The stratification of treatment intensity is based on clinical risk factors like stage, but not the histopathological diagnosis. In adults, only BL is treated with protocols initially used in children [24, 83]. For DLBCL CHOP-like regiments are the standard for adult patients [99, 66, 69, 42]. To gain further insights into the molecular characteristics of pediatric mature aggressive B-cell lymphoma, the MMML performed comprehensive molecular profiling, including gene expression, array-CGH, fluorescent in situ hybridization and immunohistochemistry on 54 patients $\leq$ 14 years. Of them, 49 were treated within prospective clinical trials of the German NHL-BFM (non-Hodgkin Berlin-Frankfurt-Münster) study group [46]. Furthermore, 18 of the cases discussed here are already part of the initial MMML data set of 220 lymphomas, yielding a total 36 newly analyzed cases.

Figure 3.5 shows the results of the pediatric lymphoma profiling study in a similar way as figure 3.2. Among the 54 patients of age 14 years or younger at presentation, the morphologic diagnoses is BL (n = 16, 30%), atypical BL (BL-like, n = 10, 19%), DLBCL (n = 16, 30%), follicular lymphoma (FL grade 3, n = 2, 4%) and high-grade B-NHL not further classified (n = 10, 19%). Gene expression profiling reveals 34 mBL (63%), 11 intermediates (20%) and 9 (17%) non-mBL lymphomas in the group of pediatric patients. Morphologic BL/atypical BL as well as mBL defined by gene expression are more frequent in children than adults (49% vs 11% and 63% vs 11%, respectively). Vice versa, DLBCL and non-mBL are less frequent in children than in adults (30% vs 82% and 17% vs 67%, respectively). The percentages of intermediates are similar in the 2 age groups (20% in children vs 22% in adults).

[1] http://info.imsd.uni-mainz.de/K_Krebsregister/english/

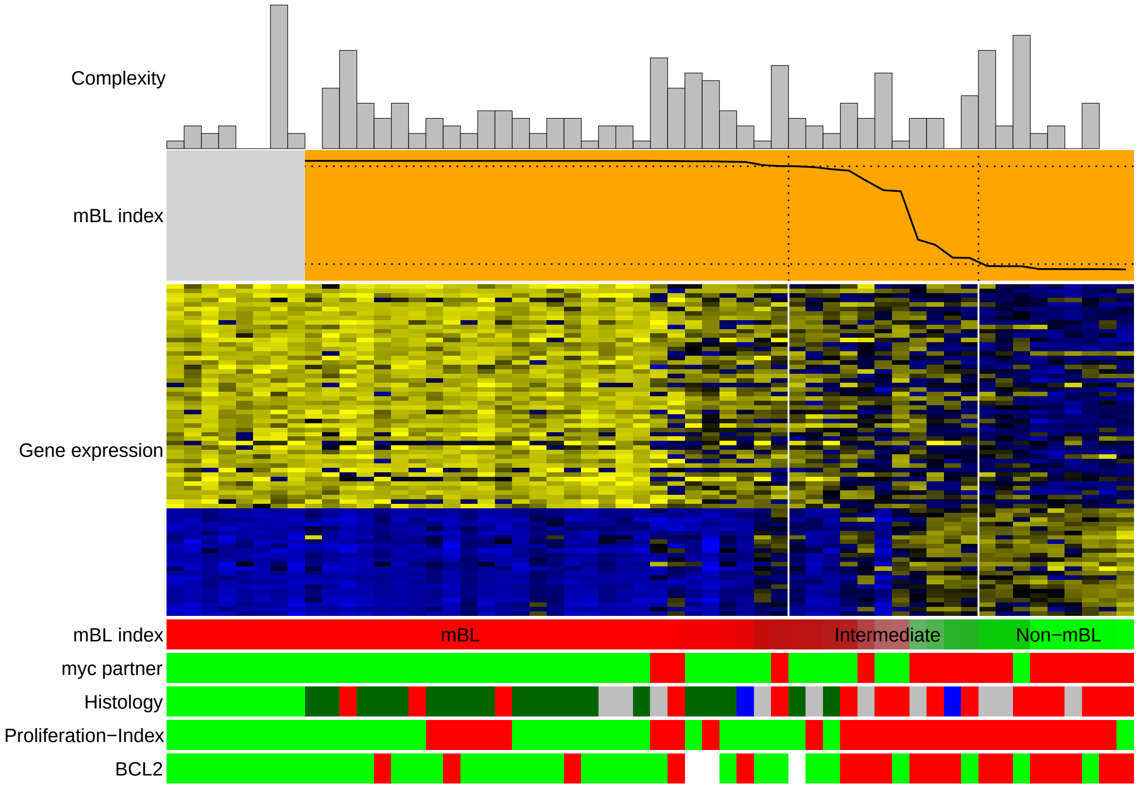

Figure 3.5: Core group extension applied to 54 pediatric lymphomas [46]. The panel compiles the same information on the 54 pediatric lymphomas as figure 3.2 except that bootstrap predictions are not available. The gray box hiding the mBL index above the leftmost cases indicates the 8 core Burkitt lymphomas of the initial MMML data set we added to the plot. The bar plot on top of the panel encodes the frequency of genomic gains and losses as revealed by array-CGH, followed by a plot of the mBL-index and a heatmap encoding the expression of the mBL signature genes. The first bar below shows again the mBL-index, where green encodes a high mBL-index and red encodes a low mBL-index. The next bar encodes in green the presence of a MYC translocation, and in red its absence. The histologic diagnosis is encoded in the bar below, with bright green indicating core-Burkitt lymphomas; dark green Burkitt lymphomas not fulfilling all WHO-criteria; red, diffuse large B-cell lymphomas; blue, follicular lymphoma; and gray, unclassifiable mature aggressive B-cell lymphoma. The level of the proliferation index indicating the number of proliferating tumor cells as revealed by Ki-67 protein expression is encoded below with green for Ki-67>95%, and red for Ki-67<95%. Finally BCL2 protein expression is indicated in green for negativity and red for positivity. White bars indicate "not assessable". (Figure reproduced from [46])

Remarkably, of the 16 morphologically defined DLBCL in children, 5 (31%) become reclassified as mBL by the gene expression signature, with 3 of them being IG-MYC positive suggesting that they are indeed biologic BL. In contrast to pediatric lymphomas, only 11 of 165 morphologically defined DLBCL in the initial series of 220 lymphomas become reclassified as mBL (table 3.4). Thus, gene expression profiling leads to a reclassification of morphologically diagnosed DLBCL more frequent in children (31%) than in the overall population (6.7%). On the other hand, the morphologic diagnosis of BL/atypical BL correlates well with the molecular diagnosis mBL in children. Only 2/26 morphologically defined BL (1 BL and 1 atypical BL, together 7.7%) do not become classified as mBL by gene expression profiling. They were classified as intermediate lymphomas.

In conclusion, the mBL-signature applies to pediatric lymphomas and provides evidence, that the frequency of Burkitt lymphomas in children is higher than expected. It is already known that morphologic DLBCL are less frequent in children than in adults [63]. Nevertheless, our data indicate that non-mBL (the molecular counterpart of morphologic/histopathologic DLBCL) exists in pediatric patients. However, gene expression profiling reveals that non-mBL is even rarer in the pediatric age group than anticipated by morphology and immunohistochemistry alone. Within the group of pediatric patients one third of the morphologically diagnosed DLBCL become classified as non-mBL by gene expression profiling and the others as unclassifiable (intermediates) or mBL. These data might explain the high frequency of MYC breaks reported for pediatric patients [70], because the group of morphologically defined DLBCL in children seems to contain a higher rate of ”contamination” with lymphomas with a mBL expression profile (set E, see table 3.2) than their adult counterpart. Although adequate studies are lacking, initial reports suggest that morphologically defined DLBCL with an mBL gene expression signature might benefit from BL therapy protocols [20, 42, 40]. Because the therapeutic strategies in BL and DLBCL are the same in children, the problem of assigning a patient with mature aggressive B-cell lymphoma to an insufficient therapy based on the morphologic/histopathologic diagnosis is clinically less relevant in children than in adults at the current stage. However, future targeted therapeutic strategies might differ between mBL and non-mBL and thus might need a precise distinction of mBL and non-mBL.

3.3.2 The LLMPP consortium data set

The Leukemia/Lymphoma molecular profiling project (LLMPP) is a consortium of research groups predominantly but not exclusively from North America. From the LLMPP Dave et al. [20] published a molecular profiling study on Burkitt lymphoma and diffuse large B-cell lymphoma similar to the MMML study. The data we will refer to as LLMPP data comprises 303 lymphoma gene expression profiles from tumors that have been initially classified as Burkitt lymphoma (n=71), diffuse large B-cell lymphoma (n=223), and cases that could be not further sub-classified (n=9). Even though the statistical method of Dave et al. differs from ours (they use a compound covariate predictor, see chapter 2), the authors arrive at very similar conclusions: Cases exist with a BL-signature together with an expert diagnosis of DLBCL and vice-versa. Furthermore, the authors show, that cases with a Burkitt signature that have been treated with the DLBCL protocol have a significantly worse survival than BL-signature positive patients treated with the more aggressive BL protocol. These findings demonstrates the clinical importance of an accurate diagnosis of Burkitt lymphoma. Unfortunately, we cannot apply the signature of Dave et al. to our data set directly, since it was produced on another, custom Affymetrix oligonucleotide microarray (LymphDx 2.7k), yielding data on another scale than those generated with Affymetrix HGU133A microarrays. However, Dave et al. subjected a subset of 99 samples in parallel to Affymetrix HGU133plus2 microarrays. We can combine those with the MMML data set generated on Affymetrix HGU133A microarrays[2] allowing us to apply our core group extension based signature to 99 cases of Dave et al. The comparison of the MMML mBL-signature and the LLMPP BL-signature on the 99 cases yields:

	BL (Dave et al.)	non-BL (Dave et al.)
non-molecular Burkitt lymphomas (MMML)	0	34
intermediate (MMML)	2	30
molecular Burkitt lymphomas (MMML)	31	2

Figure 3.6 demonstrates the application of the MMML mBL-signature to the 99 LLMPP lymphomas. While we are able to predict most of the LLMPP Burkitt lymphomas correctly, we assign almost half of the non-BL lymphomas to the intermediate group. However, a closer look at figure 3.6 reveals that most of the non-BL LLMPP cases classified as intermediate by our signature have an mBL-index close to the lower cut-off at 0.05. A slight adjustment of the mBL-signature to the LLMPP data is enough to consistently classify most of the intermediate cases as non-mBL. Note, that the LLMPP samples have been collected within a different consortium, they have been processed in a different lab, and the lab-equipment to process HGU133plus2 microarrays differs from that used for HUG133A microarrays. Despite these systematic differences, the mBL-signature directly applies to the LLMPP data set.

Conclusion. Although the mBL-classifier was trained to discriminate between C and R it performs well in discriminating between $C \cup E$ and $R \backslash E$ (see table 3.2 and section 3.1.2). Furthermore, we could validate the mBL-signature on an independent data set from the LLMPP. Nevertheless, the mBL-signature was trained in a naive way. In the next chapter we will follow an iterative procedure together with a normal mixture model to learn a classifier directly on $C \cup E$ and $R \backslash E$ instead of C and R.

[2] The Affymetrix HGU133plus2 GeneChip represents an enhanced microarray design containing all probe-sets at once, originally distributed on two separate GeneChips (HGU133A and HGU133B).

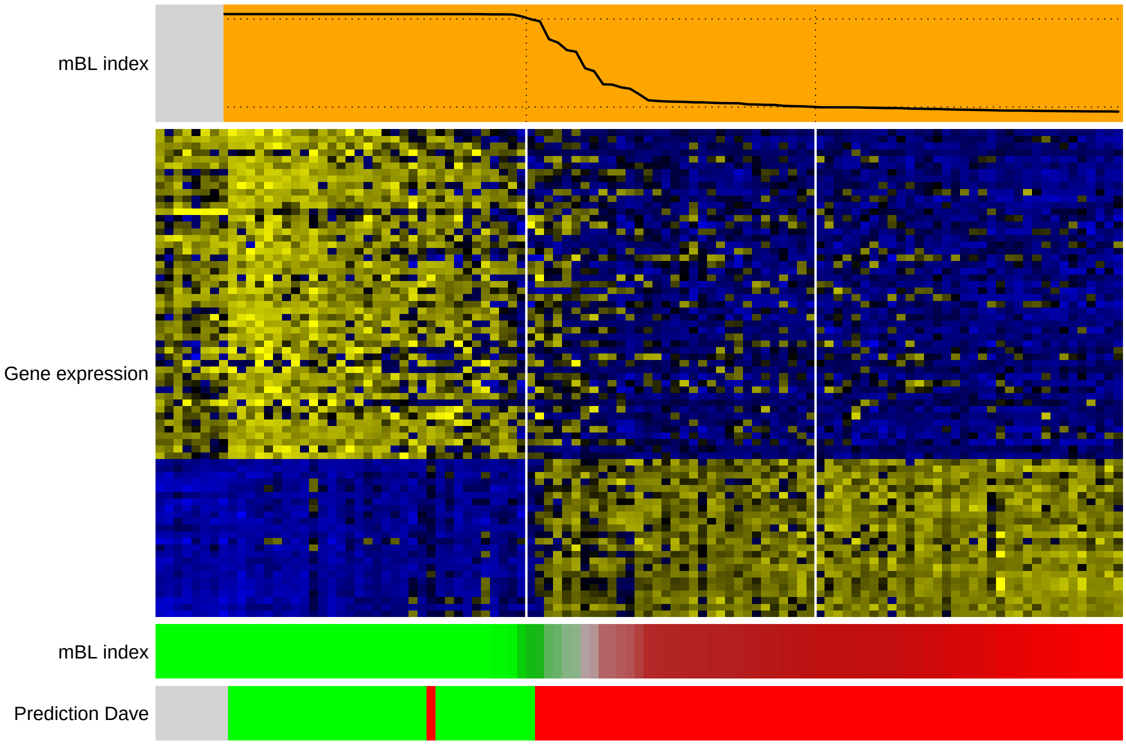

Figure 3.6: Core group extension applied 99 LLMPP lymphomas [20]. The panel compiles the same information on 99 LLMPP lymphomas as figure 3.2 except that bootstrap predictions are not available. The gray box hiding the mBL index above the leftmost cases indicates the 8 core Burkitt lymphomas of the initial MMML data set that we added to the plot. The bar on at the bottom of the panel shows encodes the prediction of the LLMPP BL-signature, with green encoding BL and red encoding non-BL (DLBCL, respectively).

Chapter 4

A model based clustering approach to core group extension

In section 3 we have approached the core group extension problem from the supervised learning perspective using a variant of linear discriminant analysis. We consider the resulting model as naive, since it has been derived from incomplete class labels. In this chapter we will approach the core group extension problem from the unsupervised perspective. We will apply a Gaussian mixture model based clustering method to partition $\mathcal{D}$ into k disjoint sets. However, we will deviate from the fully unsupervised framework in two points:

- We will restrict the search space of possible cluster solutions to those yielding two clusters, where one cluster contains all cases in C as defined in table 3.2 and the second cluster contains none of the cases in C.
- We will use the labels of C to perform gene selection prior to clustering.

The first part of this chapter describes the core group extension method based on a univariate Gaussian mixture. The second part proposes a novel gene selection strategy, and in the third part we combine the novel core group extension algorithm and the gene selection strategy to the Burkitt lymphoma problem.

4.1 Core group extension revisited

The Burkitt lymphoma classification is a core group extension problem as described in chapter 3. We have a data set $\mathcal{D}$ of 220 lymphomas. 8 of them are labeled as BL according to the WHO, and 212 remain unlabeled, due to inconclusive diagnostic criteria. However the unlabeled cases might contain further BLs. We denote the set of 8 confirmed BL cases with C for core group, the 212 remaining cases with R, and the Burkitt cases hidden in R with E for extension. The set $(C \cup E)$ denotes the extended core group and the set $(R \backslash E)$ denotes the remaining cases without E. We know the cases in C and in R, but not the cases in E. Our goal is to identify E. Furthermore we want to learn a molecular signature that classifies novel cases as members of the sets $(C \cup E)$ or $(R \backslash E)$. In terms of the Burkitt classification problem $(C \cup E)$ denotes "true", molecular Burkitt lymphomas and $(R \backslash E)$ denotes non-molecular Burkitt lymphomas. In this chapter we adapt univariate Gaussian mixture modeling to the core group extension problem.

4.2 Expectation-maximization for Gaussian mixture models

The underlying model of Gaussian mixture model based clustering is a data generating distribution with density $f(\mathbf{x}) = \prod_{j=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{x}_j | \mu_k, \Sigma_k)$, where $\phi_k(\mathbf{x}_j | \mu_k, \Sigma_k)$ are K normal densities for K components in $\mathcal{D}$. The membership of the samples in the K components is unknown. The goal is to find a constellation of the parameters μ_k,Σ_k, and π_k that maximizes the likelihood $\mathcal{L}(\mathbf{X} | \pi_k, \mu_k, \Sigma_k)$ of the observed data $\mathbf{X}$.

$$\mathcal{L}(\mathbf{X} | \pi_k, \mu_k, \Sigma_k) = \prod_{j=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{x}_j | \mu_k, \Sigma_k), \tag{4.1}$$

where μ_k and Σ_k denote the location and covariance of the K normal components, and π_k denotes the proportions of the components. An efficient way to estimate the parameters is the expectation-maximization (EM) algorithm [22, 61] an iterative procedure of alternating E-steps (E=expection) and M-steps (M=maximization). During the **E-step** we estimate the conditional class probabilities p_k given a certain parameter constellation for μ_k,Σ_k, and π_k:

$$\hat{p}_k(\mathbf{x}_j) = \frac{\pi_k \phi_k(\mathbf{x}_j | \mu_k, \Sigma_k)}{\sum_{l=1}^{K} \pi_l \phi_l(\mathbf{x}_j | \mu_l, \Sigma_l)}. \tag{4.2}$$

During the **M-step** we estimate the parameters μ_k,Σ_k, and π_k given $\hat{p}_k(\mathbf{x}_j)$ from the preceding E-step. The estimates of μ_k and π_k have closed-form expressions [31]:

$$\begin{aligned} \hat{\mu}_k =& \frac{1}{n_k} \sum_{j=1}^{n} \hat{p}_k(\mathbf{x}_j) \mathbf{x}_j, \\ \hat{\pi}_k =& \frac{n_k}{n} \\ n_k =& \sum_{j=1}^{n} \hat{p}_k(\mathbf{x}_j). \end{aligned} \tag{4.3}$$

The estimation of $\hat{\Sigma}_k$ is less straight forward (see [31] for details). One possible choice is

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{j=1}^{n} \hat{p}_k(\mathbf{x}_j)(\mathbf{x}_j - \hat{\mu}_k)(\mathbf{x}_j - \hat{\mu}_k)^T. \tag{4.4}$$

However, in analysis we have more genes than samples and the covariance matrix given in equation 4.4 is not invertible. Nevertheless, we need to invert $\hat{\Sigma}_k$ to compute

$\hat{p}_k(\mathbf{x}_j)$. Alternatively, like in supervised discriminant analysis described in chapter 2 we can assume independence of genes and choose the covariance matrix to be diagonal. In the next section, we will follow yet another approach, where we do not have to estimate covariances. We will project the data first on a univariate line, and fit the mixture model to the univariate projection.

EM converges to a local maximum of the likelihood after several iterations of alternating E- and M- steps [22]. However, the final solution depends on the starting point of the iterations. Prominent methods to choose the starting point are hierarchical clustering in the multivariate case and quantile-based partitioning in the univariate case. Both are part of a comprehensive model based EM-clustering strategy proposed in [31]. In the next section we adapt univariate Gaussian mixtures [31] to core group extension of Burkitt lymphomas.

4.3 Core group extension expectation maximization (COREEM)

4.3.1 Implementation

Assume we are given a normal vector $\mathbf{a} \in \mathbb{R}^p$, then the projection $s_j = \mathbf{a}^T\mathbf{x}_j$ with $j = 1, 2, ..., n$ indexing n samples defines a univariate partially labeled data set $\mathbf{s} \in \mathbb{R}^n$. We refer to s_j as gene expression index aggregating the expression level of p genes in a univariate score. It represents the core group if $j \in C$. We now model the data in $\mathbf{s}$ as univariate mixture of two normal distributions we refer to as components. The first component should represent Burkitt lymphomas comprising the set $C \cup E$ (index $k = 1$) and the second component non-Burkitt lymphomas comprising the set $R \backslash E$ (index $k = 2$). We use the partial label information and modify the E-step of the standard EM-algorithm defined in equation 4.2. This yields the extension step of our novel core group extension model. We refer to it as core group extension expectation maximization (COREEM):

E-step

$$\hat{p}_{C\cup E}(s_j) = \begin{cases} \dfrac{\hat{\pi}\phi(s_j|\hat{\mu}_1, \hat{\sigma}_1^2)}{\hat{\pi}\phi(s_j|\hat{\mu}_1, \hat{\sigma}_1^2) + (1-\hat{\pi})\phi(s_j|\hat{\mu}_2, \hat{\sigma}_2^2)}, & \text{if} \quad j \notin C \\ 1, & \text{if} \quad j \in C \end{cases}, \tag{4.5}$$

$$\hat{p}_{R\backslash E}(s_j) = 1 - \hat{p}_{C\cup E}(s_j).$$

where $\phi(s_j|\hat{\mu}_k, \hat{\sigma}_k^2)$ denotes the density of the univariate normal distribution:

$$\phi(s_j|\hat{\mu}_k, \hat{\sigma}_k^2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_k} \exp -\frac{1}{2}(\frac{s_j - \hat{\mu}_k}{\hat{\sigma}_k})^2 \tag{4.6}$$

The M-step is like in the standard EM algorithm for univariate two-component Gaussian mixtures:

M-step

$$\begin{aligned} n_1 = \sum_{j=1}^{n} \hat{p}_{C\cup E}(s_j), \quad n_2 = \sum_{j=1}^{n} \hat{p}_{R\setminus E}(s_j), \\ \hat{\mu}_1 = \sum_{j=1}^{n} \hat{p}_{C\cup E}(s_j) s_j / n_1, \quad \hat{\mu}_2 = \sum_{j=1}^{n} \hat{p}_{R\setminus E}(s_j) s_j / n_2 \\ \hat{\pi} = n_1/n, \end{aligned} \tag{4.7}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{j=1}^{n} \hat{p}_{C\cup E}(s_j)(s_j - \hat{\mu}_1)^2}{n_1}, \quad \hat{\sigma}_2^2 = \frac{\sum_{j=1}^{n} \hat{p}_{R\setminus E}(s_j)(s_j - \hat{\mu}_2)^2}{n_2}$$

Initialization. The solution of EM depends on the starting point [31]. A prominent choice is to split the data at the quantiles into k equally sized initial classes, where k depends on the number of presumed components (see [31] for details). We set $\hat{p}_k = 1$, if a sample belongs to the initial class k and $\hat{p}_k = 0$ otherwise. Given this initial choice of $\hat{p}_k$ we start the algorithm with an M-step. Here, we are in the core group extension scenario. Thus we initialize the EM algorithm with two initial classes and set $\hat{p}_{C\cup E}(s_j) = 1$ if $j \in C$ and $\hat{p}_{C\cup E}(s_j) = 0$ otherwise.

Classification. We run core group extension expectation maximization until convergence and obtain estimates for $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\pi}_1 = \hat{\pi}$, and $\hat{\pi}_2 = 1 - \hat{\pi}$ from the final M-step. We set up a classification rule and assign a sample j with expression index s_j to the core group if

$$\hat{\pi}_1 \phi(s_j | \hat{\mu}_1, \hat{\sigma}_1^2) \geq \hat{\pi}_2 \phi(s_j | \hat{\mu}_1, \hat{\sigma}_1^2). \tag{4.8}$$

This can be written as

$$\frac{(s_j - \hat{\mu}_1)^2}{\hat{\sigma}_1^2} + \ln \hat{\sigma}_1^2 - 2 \ln \hat{\pi}_1 \geq \frac{(s_j - \hat{\mu}_2)^2}{\hat{\sigma}_2^2} + \ln \hat{\sigma}_2^2 - 2 \ln \hat{\pi}_2. \tag{4.9}$$

The critical value s_{crit} defining the classification boundary is the solution of

$$\begin{aligned} \frac{(s_{crit} - \hat{\mu}_1)^2}{\hat{\sigma}_1^2} + \ln \hat{\sigma}_1^2 - 2 \ln \hat{\pi}_1 = \frac{(s_{crit} - \hat{\mu}_2)^2}{\hat{\sigma}_2^2} + \ln \hat{\sigma}_2^2 - 2 \ln \hat{\pi}_2 \\ , s_{crit}^2(\hat{\sigma}_2^2 - \hat{\sigma}_1^2) + 2 s_{crit}(\hat{\mu}_2 \hat{\sigma}_1^2 - \hat{\mu}_1 \hat{\sigma}_2^2) + \hat{\mu}_1^2 \hat{\sigma}_2^2 - \hat{\mu}_2^2 \hat{\sigma}_1^2 - (\ln \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \frac{\hat{\pi}_1^2}{\hat{\pi}_2^2}) \hat{\sigma}_1^2 \hat{\sigma}_2^2 = 0. \end{aligned} \tag{4.10}$$

We solve a quadratic equation of the form

$$a x^2 + b x + c = 0 \tag{4.11}$$

with

$$\begin{aligned} a &= (\hat{\sigma}_2^2 - \hat{\sigma}_1^2) \\ b &= 2(\hat{\mu}_2\hat{\sigma}_1^2 - \hat{\mu}_1\hat{\sigma}_2^2) \\ c &= \hat{\mu}_1^2\hat{\sigma}_2^2 - \hat{\mu}_2^2\hat{\sigma}_1^2 - (\ln\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\frac{\hat{\pi}_1^2}{\hat{\pi}_2^2})\hat{\sigma}_1^2\hat{\sigma}_2^2. \end{aligned} \tag{4.12}$$

Note that equation 4.11 yields two solutions for s_{crit}. Figure 4.1 demonstrates the reason why there are two decision boundaries: The likelihood of the red class is larger than that of the green class on both sides (tails) of the green mixture component. The region where we classify a samples as red is not connected. Here, we consider only the value between the two component means s_{crit_1} and ignore the second "outer" solution of equation 4.11. We will discuss this later in section 4.5. The expression index s_j is a linear combination $s_j = \mathbf{a}^T\mathbf{x}_j$ of the gene expression vector $\mathbf{x}_j$ and weights $\mathbf{a}$. Hence we have

$$f(\mathbf{x}) = \sum_{i=1}^{p} a_i x_i + s_{crit}, \tag{4.13}$$

where $\mathbf{x}$ is in the extended core group if $f(\mathbf{x}) > 0$.

So far, we have described the final step of core group extension expectation maximization using a predefined set of p signature genes and predefined weights $\mathbf{a}$. In the following section we proceed with describing, how we select a set of signature genes and the weights $\mathbf{a}$.

4.4 Gene selection and expression index

4.4.1 Filtering and weighting genes

In supervised classification we have completely labeled data. A common approach of gene selection is univariate thresholding, where we rank genes based on a univariate classification score. The t-statistic is a prominent choice in this context. It is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{4.14}$$

where $\bar{x}_1$ and $\bar{x}_2$ denote the means of the two classes we wish to compare and $\hat{\sigma}$ denote the pooled within standard deviation (see equation 2.5). Only genes with a univariate classification score above a certain threshold are included into a molecular signature. We assume: We want to discriminate two classes. All cases in the same class are drawn from the same distribution while the distributions for the two classes are different. The t-score selects genes based on the gene-per-gene differences of univariate distributions. In core group extension the core group cases (C) are assumed to be drawn from the same distribution, while the remaining cases in R are drawn

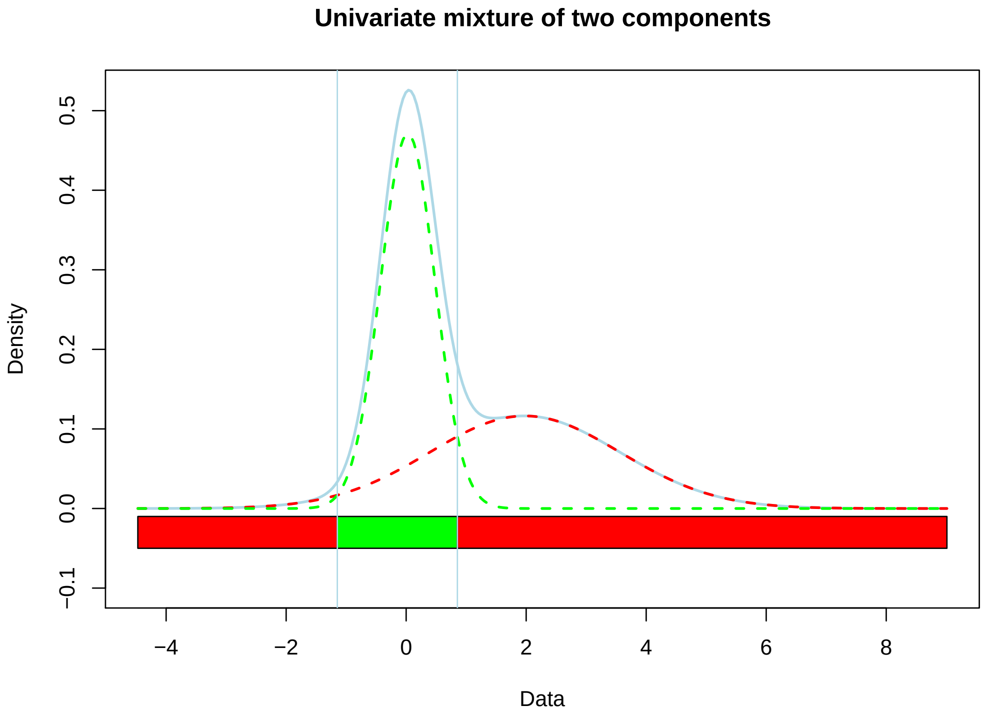

Figure 4.1: A univariate mixture of two Gaussian distributions has two decision boundaries. Shown are the densities of two mixing univariate Gaussian distributions (dashed green and red lines). The green density has a low variance as compared to the red. The tails of the red density are wider than those of the green density. This causes that the decision region (bar below the density plot) for the red class is not connected, but stretches along both sides of the very narrow green class. In case of equal variances we have only a single decision boundary.

from a mixture. Part of them are drawn from E following the core group distribution while the others are drawn from $R\backslash E$ following a different distribution. The mean of the a priori group R is not representing a population mean but is contaminated by hidden cases from E. Consequently the variances estimated across the R cases are to large due to the contamination of this group. We take this into account and deviate from the standard t-score for gene selection. We ignore the mean and the standard deviation of the cases in R. We select a gene i, if the distance of its core group mean differs from the overall mean

$$t_{C,i} = \frac{\bar{x}_{i,C} - \bar{x}_i}{\hat{\sigma}_{C,i} + s_0}, \tag{4.15}$$

where $\hat{\sigma}_{C,i}$ denotes the standard deviation of gene i in the core group C. We ignore $\hat{\sigma}_R$. The constant $s_0 = \text{median}(\{\hat{\sigma}_{C,i}\}_i^p)$ is a fudge factor, which prevents high scoring genes due to extremely low $\hat{\sigma}_{C,i}$ as described in [92]. We rank genes according to absolute values of $t_{C,i}$, yielding

$$\mathbf{x}^{(p^*)} = x_{(1)}, x_{(2)}, ..., x_{(p^*)}, \tag{4.16}$$

which contains the gene expression values of the p^* top scoring genes with respect to $|t_{C,i}|$. Furthermore, we follow ideas of the compound covariate predictor (see section 2.6 on page 21) and use $a_i = t_{C,i}$ as weights to compute $s_j = \mathbf{a}^T \mathbf{x}_j$. Hence

$$s_j^{(p^*)} = \sum_{i=1}^{p^*} t_{C,(i)} \cdot x_{(i),j} \tag{4.17}$$

denotes the gene expression index of the p^* top scoring genes in sample j. Next we choose the optimal number of genes p^*.

4.4.2 Choosing the number of genes in the compound score

The score $t_{C,i}$ in equation 4.15 ranks genes, and we use it as weights $\mathbf{a}$ in $s_j = \mathbf{a}^T \mathbf{x}_j$. What remains to be decided is the choice of p^*, the number of genes included in the model. In supervised classification, cross validation can be used to determine the optimal number of signature genes p^* minimizing the number of misclassifications. The labels decide whether a classification is correct or not. In core group extension part of the cases are unlabeled. For these cases we can not reliable distinguish between a correct and an incorrect classification.Instead define another objective for model selection:

Let $\mathbf{s}^{(p^*)} = s_1^{(p^*)}, s_2^{(p^*)}, ..., s_n^{(p^*)}$ be the index aggregating the expression values of the p^* top ranking genes in n samples. A subset of the samples is labeled as core group C. Running core group extension expectation maximization as described in section 4.3 yields a parameter constellation from the last M-step (equation 4.7) after the likelihood has converged to a local maximum. The M-step yields $\hat{\mu}_1^{(p^*)}$, $\hat{\mu}_2^{(p^*)}$, $\hat{\sigma}_1^{(p^*)}$ and $\hat{\sigma}_2^{(p^*)}$. We define

$$d(p^*) = \frac{|\hat{\mu}_1^{(p^*)} - \hat{\mu}_2^{(p^*)}|}{\frac{1}{2}(\hat{\sigma}_1^{(p^*)} + \hat{\sigma}_2^{(p^*)})} \tag{4.18}$$

to score the discrimination between the two normal mixture components resulting from core group extension expectation maximization. We run the algorithm for different p^* and choose p^* maximizing $d(p^*)$.

4.5 Results

We will now apply core group extension expectation maximization to the same data set $\mathcal{D}$ as described in section 3.2. It contains 221 gene expression profiles from 220 aggressive B-cell lymphomas, and we start by splitting $\mathcal{D}$ into a core group C of 8

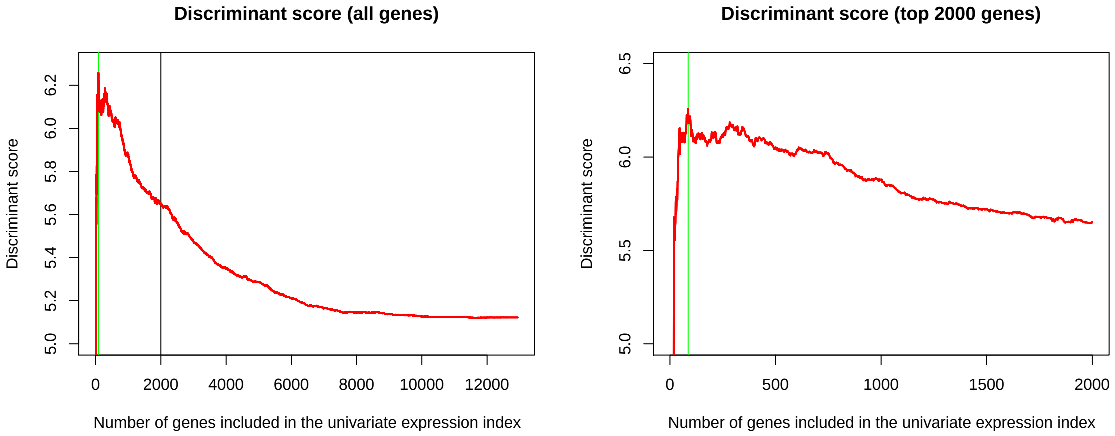

Figure 4.2: Discriminant scores of two-component Gaussian data models. Both plots show different discriminant scores on the y-axis. They have been computed in line with equation 4.18 for different values of p^* (x-axis). Each p^* yields different gene expression indices and subsequently another core group extension expectation maximization solution. The plot on the left shows the discriminant scores across the whole data set. The plot on the right zooms into the left plot (indicated by the vertical black line) and shows only the result for the top 2000 scoring genes with respect to $|t_{C,i}|$. The discriminant score has its maximum at $p^* = 86$ genes (vertical green line).

Burkitt lymphomas, a training set R of 105 mature aggressive B-cell lymphomas, and a test set $\mathcal{D}_{test}$ of 107 mature aggressive B-cell lymphomas. Furthermore, we apply an additional processing step to the data: Affymetrix microarrays measure transcription with probe-sets. In most cases, a single probe-set measures a single gene. However, in case of complex genes, Affymetrix designed multiple probe-sets for the same gene. In order to remove redundant information in our final core group extension model, we average expression values obtained from different probe-sets that measure the same gene. From the 22283 different probe-set expression values, we obtain $p = 12938$ different gene expression values.

In line with equation 4.15 we compute $t_{C,i}$ for each gene $i = 1, 2, 3, ..., 12938$ comparing the expression of each gene in the core group C with its expression in all samples $C \cup R$. We rank genes according to decreasing absolute values of $t_{C,i}$ and include the top p^* genes into the computation of the gene expression indices $\mathbf{s}^{(p^*)}$ in line with equation 4.17. This yields a univariate data set $\mathcal{D}_{uni}^{(p^*)}$ with $n = \#C + \#R = 113$ samples. We derive a data set $\mathcal{D}_{uni}^{(p^*)}$ for each $p^* = 1, 2, 3, ..., 12938$, and subject them one after the other to core group extension expectation maximization. Figure 4.2 shows the discriminant scores of the resulting models as defined in equation 4.18.

The gene expression index $\mathbf{s}^{(p^*)}$ yields a core group extension expectation maximization solution with maximal discriminant score $d(p^*)$, if $p^* = 86$. Figure 4.3 shows the density of the resulting model. The model has two decision boundaries $s_{crit_1} = 3.09$

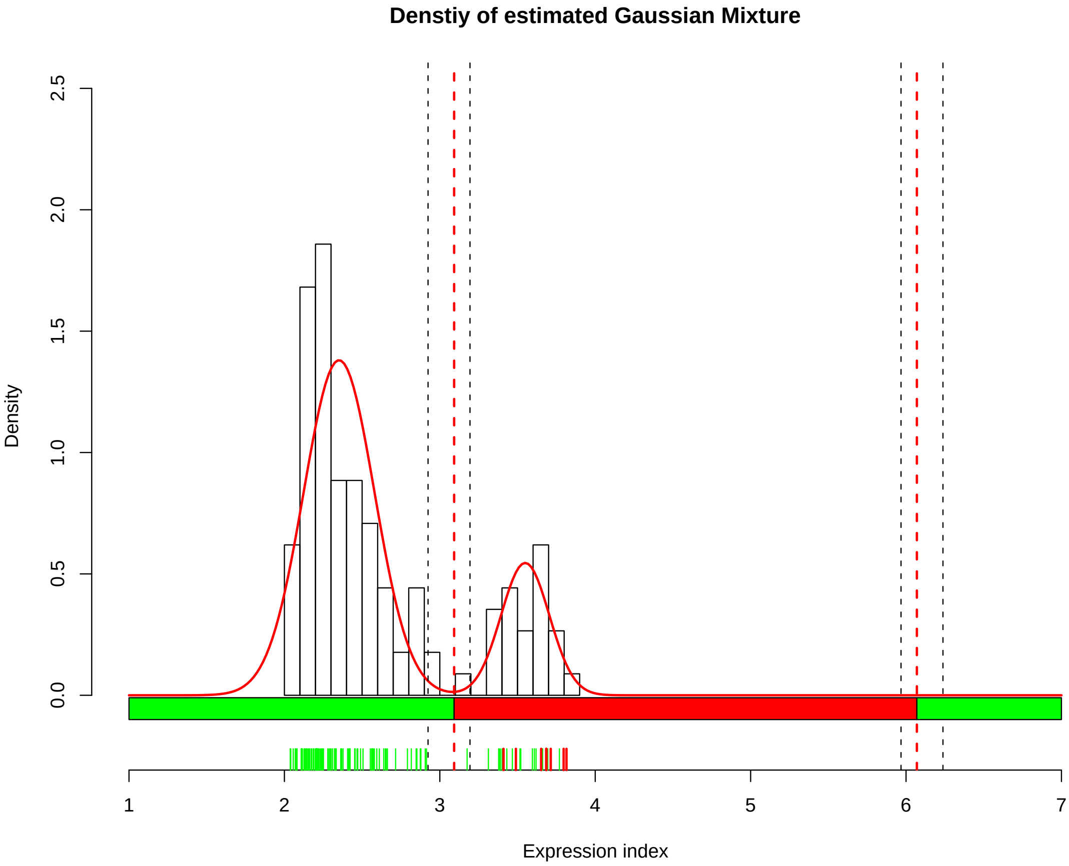

Figure 4.3: Density of the optimal core group extension expectation maximization model. The plot shows a histogram of the optimal core group extension model of a 86-gene expression index in 113 lymphomas. The red line is the 2-component Gaussian mixture density estimated by core group extension expectation maximization. The level of the expression index of the core Burkitt lymphomas C is highlighted by red ticks below the histogram. The green ticks highlight the expression index of the remaining lymphomas R. The bar along the x-axis below the histogram indicates the classification of the samples with respect of the class probabilities obtained from equation 4.5. If an individual sample has an expression index in the range indicated by the red region the posterior probability of the core group component is higher. If it has an expression index in the green region, the posterior probability of the non core group component is higher. The classification depends on the parameter π_k, taking into account different proportions of the two classes. The dashed vertical red lines highlight the decision boundaries obtained from the core group extension expectation maximization. The dashed vertical black lines highlight the range of the decision boundaries, if we vary the class proportions between $\pi_{1,2} = (0.01, 0.99)$ and $\pi_{1,2} = (0.99, 0.01)$.

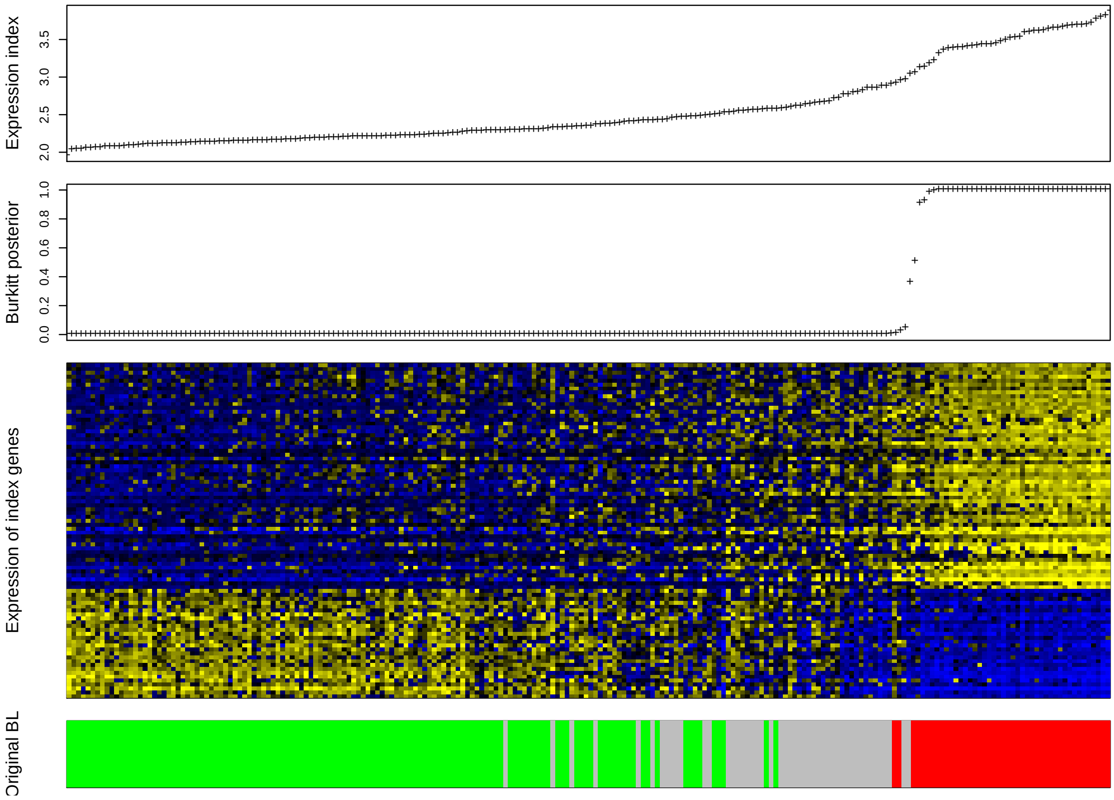

Figure 4.4: Core group extension expectation maximization model of 220 lymphomas. The panels from top to down show: 1.) The Burkitt expression index $s_j^{(p^*)} = a^T \mathbf{x}_j^{(p^*)}$ of the individual samples j sorted along the x-axis in increasing order, 2.) the posterior class probability for Burkitt lymphoma sorted with respect to the Burkitt expression index and 3.) the expression of the $p^* = 86$ genes constituting the Burkitt index in the 220 lymphomas. The bar below highlights in red molecular Burkitt lymphomas as defined in chapter 3, in green non-molecular Burkitt lymphomas and in gray intermediate cases.

and $s_{crit_2} = 6.07$. The latter is far outside the range of the data, and we do not consider it for classification. We define

$$f(\mathbf{x}) = \sum_{i=1}^{86} t_{C,(i)} x_{(i)} + 3.09, \tag{4.19}$$

where (i) indexes the genes ordered with respect to decreasing absolute values of $t_{C,i}$. Cases are considered as Burkitt lymphoma, if $f(\mathbf{x}) > 0$ and as non-Burkitt lymphoma otherwise. If we classify the 113 training cases accordingly, we obtain 24 Burkitt lymphomas and 89 non-Burkitt lymphomas. The 24 Burkitt lymphomas include the 8 core cases. Modeling the 86-gene expression index yields the model parameters given in table 4.1.

The estimates for the class proportions $\hat{\pi}_1$ and $\hat{\pi}_2$ are 0.7882 and 0.2118. Thus, the estimated proportion of Burkitt lymphomas is about 20% of the population. This

Table 4.1: Core group extension expectation maximization model parameters estimated on 113 lymphoma gene expression profiles.

	Component 1	**Component 2**
Assigned subtype	non-Burkitt lymphoma	Burkitt lymphoma
Size (training set)	89	24
$\hat{\mu}_k$ (equation 4.7)	2.352	3.549
$\hat{\sigma}_k^2$ (equation 4.7)	0.0519	0.02401
$\hat{\pi}_k$ (equation 4.7)	0.7882	0.2118

information is included in the classification function defined in equation 4.19. The model depends on the composition of the data set. This is important, because the composition of the present training data does not represent the natural proportions of the subtypes among the population of mature aggressive B-cell lymphoma. Instead the proportion of Burkitt lymphomas is artificially high. There was a priority on including Burkitt lymphomas when the data was collected. In order to figure out how much the model prediction depends on π_k, we plug in the following extreme prior parameter constellations into the final Gaussian mixture model:

- $\pi_1 = 0.01$ and $\pi_2 = 1 - \pi_1 = 0.99 \Rightarrow s_{crit_{1,2}} = (2.9242, 6.2376)$,
- $\pi_1 = 0.99$ and $\pi_2 = 1 - \pi_1 = 0.01 \Rightarrow s_{crit_{1,2}} = (3.1941, 5.9677)$.

Even if we vary the π_k in an extreme range from 1 to 99%, only one sample changes its label based on the resulting classification. The decision boundaries vary along two regions of very low density of the data, if we vary π_k (figure 4.3). Since the true proportions of Burkitt lymphomas and non-Burkitt lymphomas are unknown and the classification of the samples is very stable with respect to varying π_k, we modify the classifier by exchanging the estimated class proportions with $\pi_1 = \pi_2 = 0.5$. The classification of the training data remains unaffected. The critical value s_{crit} separating the two mixture components changes to $s_{crit} = 3.0531$. Thus, we classify a sample $\mathbf{x}$ as Burkitt lymphoma, if

$$f(\mathbf{x}) = \sum_{i=1}^{86} t_{C,(i)} x_{(i)} + 3.05 > 0. \tag{4.20}$$

In order to test the classifier on an independent data set we applied it to our test data set of 107 lymphomas. We obtain a stratification of the test data into 18 Burkitt lymphomas and 89 non-Burkitt lymphomas. If we merge training and test set and compare the classification of equation 4.20 with the stratification described in chapter 3 we obtain:

	non-mBL	intermediate	mBL
non-Burkitt lymphomas (EM-based approach)	128	48	2
Burkitt lymphomas (EM-based approach)	0	0	42

The classification differs only in 2 cases, which have been classified as molecular Burkitt lymphoma by the nearest shrunken centroids procedure described in chapter 3, but not by core group extension expectation maximization. The two discrepant cases do not have a genomic translocation of the oncogene MYC, which is a gold-standard diagnostic criterion for Burkitt lymphoma. Figure 4.4 shows the final result of core group extension expectation maximization on the joint training and test set. The Burkitt posterior class probabilities shown in this figure have been computed by plugging in the estimated model parameters (table 4.1, but $\pi_1 = \pi_2 = 0.5$) into equation 4.2 (page 46). The Burkitt class probability obtained is more extreme than that of the naive nearest shrunken centroids approach in chapter 3. While the naive approach yields 48 intermediate cases with a Burkitt probability larger than 0.05 and smaller than 0.95, the core group extension expectation maximization based class probability is between 0.05 and 0.95 only in 4 cases. In conclusion, core group extension expectation maximization yields a model providing a clearer cut between Burkitt lymphomas and the remaining mature aggressive B-cell lymphomas. It leaves less cases unclassified, which improves its potential diagnostic use. At the same time, the major conclusions from the naive approach described in chapter 3 remain unchanged considering the novel classification. Notably, 42 of the 44 molecular Burkitt lymphomas defined in chapter 3 remain Burkitt lymphomas according to core group extension expectation maximization.

4.6 Validation on further data sets.

In order to further evaluate the core group extension expectation maximization derived Burkitt-signature we apply it to the LLMPP and the pediatric lymphoma data set introduced in chapter 3. The upper panel of figure 4.5 shows the gene expression index in 54 lymphomas from patients younger than 14 years analyzed within the pediatric molecular profiling project. The lower panel shows the same for 99 LLMPP lymphoma profiles. The signature applies to both data sets.

Pediatric lymphomas. If we compare the initial molecular Burkitt lymphoma classification of the pediatric data set as defined in chapter 3, and the novel gene expression index based classification we obtain:

	non-mBL	intermediate	mBL
non-Burkitt lymphomas (EM-based approach)	9	12	2
Burkitt lymphomas (EM-based approach)	0	0	31

As before the core group extension expectation maximization based signature classifies intermediate cases as non-Burkitt lymphomas. Furthermore, two of the cases, which have been classified earlier as molecular Burkitt lymphomas become reclassified as non-Burkitt lymphomas. The majority of the cases are classified as Burkitt lymphomas by both, the core group extension expectation maximization based and the nearest shrunken centroids based signature. Only 4 of the 54 pediatric cases have an intermediate Burkitt probability between 0.05 and 0.95 according to the novel model. Also in this data set, the cut between Burkitt and non-Burkitt cases is clearer.

LLMPP. The 99 LLMPP lymphomas shown in the lower panel of figure 4.5 have been initially classified by the Burkitt signature published by Dave et al. [20]. If we reclassify the same cases using our core group extension expectation maximization based signature we obtain

	BL (Dave et al.)	non-BL (Dave et al.)
non-Burkitt lymphomas (EM-based)	0	65
Burkitt lymphomas (EM-based)	33	1

We have only a single discrepant case. Notably, the samples profiled in Dave et al. [20] have been collected within a different consortium (the LLMPP), have been processed in a different lab, and have been subjected to a different Affymetrix microarray platform (HUG133plus2). Nevertheless, the signature trained on the MMML data set directly applies to the LLMPP data set.

In conclusion, the Burkitt lymphoma signature is robust and reproducible in different data sets. Furthermore, different statistical methodology applied to different data sets yield the same groups. Transcriptional profiling thus provides strong evidence that Burkitt lymphoma is biologically different from diffuse large B-cell lymphoma, and that a molecular signature provides a quantitative basis for the diagnostic distinction of Burkitt lymphoma from diffuse large B-cell lymphoma. We know finish the discussion of Burkitt lymphoma at this point and continue with strategies to the molecular stratification of diffuse large B-cell lymphoma.

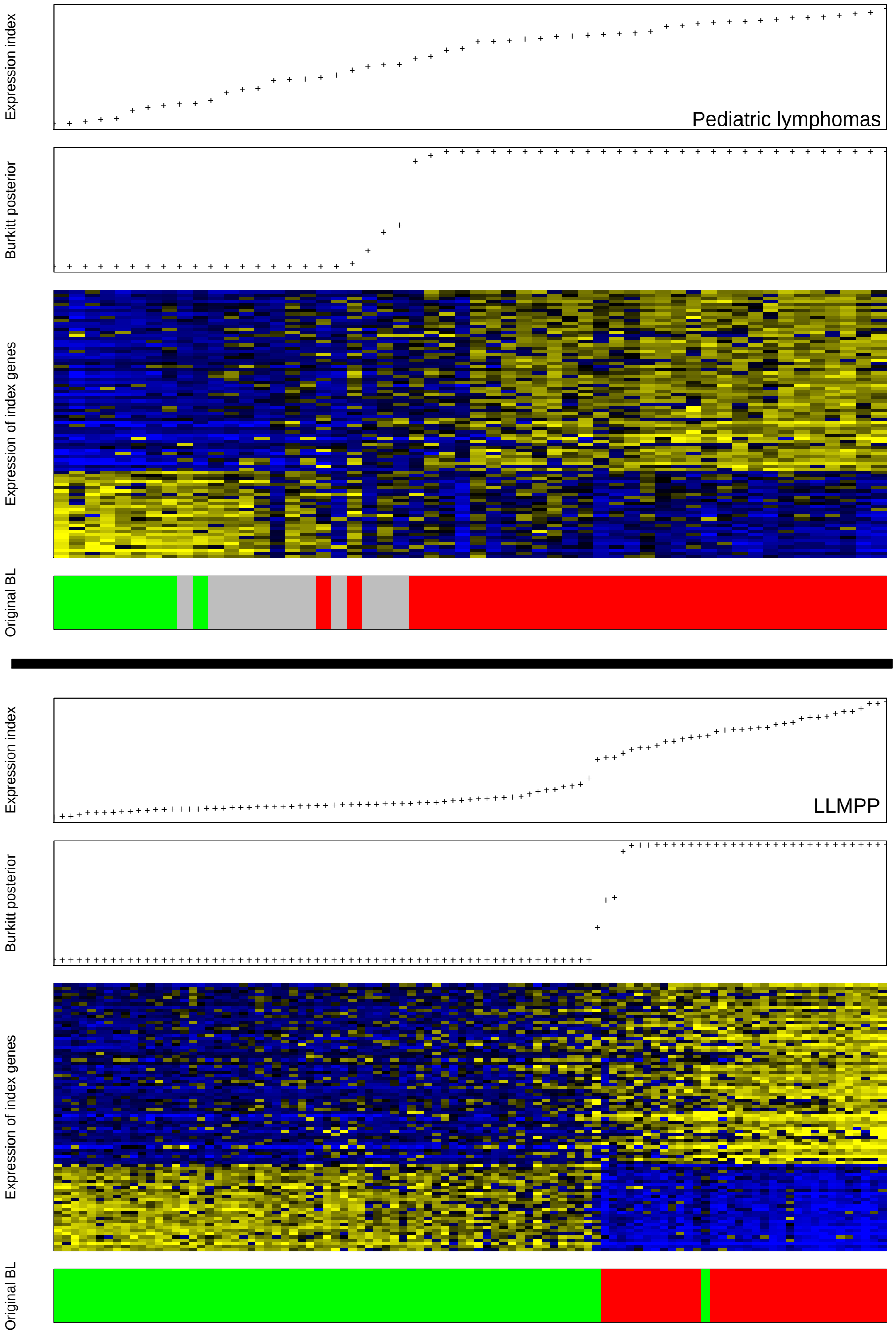

Figure 4.5: Core group extension expectation maximization model applied to 54 pediatric lymphomas (top, [46]) and 99 LLMPP lymphomas (bottom, [20]). See legend of figure 4.4 for details. The only deviation from the legend of figure 4.4 is the bar below the lower panel showing the LLMPP data. This bar does not encode the mBL-signature as described in chapter 3 but indicates in red the BL-classification as defined in the original publication of the data [20].

Part III

Molecular stratification of diffuse large B-cell lymphoma

Chapter 5

GCB, ABC and their relation to gains of chromosome 18q

In this chapter we explore germinal center B-cell-like (GCB-like) and activated B-cell-like (ABC-like) lymphomas in the MMML data set. Furthermore, we will have a closer look at the MALT1 gene, which is a close neighbor to BCL2 an important oncogene in lymphomagenesis on chromosomal band 18q21 and has an important impact on NF-κB activation [75]. Genomic gains of regions on chromosome 18q are frequent events in ABC-like lymphomas [9], and in most cases jointly increase the chromosomal copy number of BCL2 and MALT1 together. We are interested in the question, if there is an ABC independent relationship between transcription and the presence of a genomic gain of 18q/MALT1.

Gain of MALT1 on chromosome 18q21 has been evaluated by a combination of array comparative hybridization (array CGH) and fluorescence in situ hybridization (FISH) as described in Dierlamm et al. [23]. The GCB and ABC labels will be predicted from transcriptional profiles by using a compound covariate predictor following Wright et al. [101].

5.1 Molecular features of GCB and ABC

The germinal center B-cell-like/ activated B-cell-like (GCB/ABC) signature stratifies diffuse large B-cell lymphoma (DLBCL) based on transcriptional profiling [3, 74]. Patients with a GCB-lymphoma on average have a better overall survival as compared to ABC-lymphoma patients [3, 74]. Additional to the prognostic potential of the GCB/ABC stratification, several genetic aberrations have been identified that occur at different frequencies in GCB and ABC lymphomas [41, 9, 44]. However, their presence does not allow a classification to GCB or ABC. Some of them involve the genomic locus of BCL2 on chromosome band 18q21, and thus altering the expression of this important oncogene in lymphoma pathogenesis. GCB-like DLBCL frequently show a t(14;18) genomic translocation [41]. BCL2 is target of this translocation and thus becomes overexpressed in GCB lymphomas. In contrast, in ABC lymphomas we observe BCL2 protein overexpression in the absence of the genomic translocation t(14;18) [44]. The molecular mechanisms underlying BCL2 overexpression in absence of t(14;18) are unknown. However, ABC-like DLBCL has constitutively activated the transcription factor nuclear factor (NF)-κB [21, 28, 51], and BCL2 is a target gene for transcriptional regulation through NF-κB. Thus BCL2 upregulation in ABC may be due to NF-κB activity. Another possible mechanism increasing BCL2 protein expression is a chromosomal gain of the BCL2 locus, as demonstrated by Iqbal et al. [44]. BCL2 is located on chromosome 18q and chromosome 18q is frequently gained in ABC-like lymphomas [9].

5.2 The compound covariate predictor for GCB and ABC

Wright et al. [101] propose a molecular signature comprising the transcriptional levels of 27 genes to distinguish GCB from ABC lymphomas. They have derived this signature from a custom microarray platform called Lymphochip [2], and suggest a compound covariate predictor (see section 2.6) to obtain GCB and ABC labels from Lymphochip transcriptional profiles. In line with equation 2.14 they compute

$$c(\mathbf{x}) = \sum_{i \in G} t_i x_i, \tag{5.1}$$

where t_i denotes t-scores for differential expression between GCB and ABC of the $i = 1, 2, ..., 27$ signature genes observed on Lymphochip data. They compute $c(\mathbf{x}_j)$ for each sample j and model it as a mixture of two univariate normals. Given the densities of two univariate normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2

$$\begin{aligned} p(x|k=1) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp -\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2 \\ p(x|k=2) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp -\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2, \end{aligned} \tag{5.2}$$

they obtain a maximum likelihood classification of a new sample x as member of class $k = 1$ if

$$p(x|k=1) > p(x|k=2) \tag{5.3}$$

Replacing x with the linear predictor score $c(\mathbf{x})$, μ_1 with $\bar{c}^{(1)}$, μ_2 with $\bar{c}^{(2)}$ and the two standard deviations with empirical estimates $\hat{\sigma}_1$ and $\hat{\sigma}_2$ yields

$$\frac{(c(\mathbf{x}) - \bar{c}^{(1)})^2}{\hat{\sigma}_1^2} + \ln \hat{\sigma}_1^2 < \frac{(c(\mathbf{x}) - \bar{c}^{(2)})^2}{\hat{\sigma}_2^2} + \ln \hat{\sigma}_2^2. \tag{5.4}$$

The two variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ of $c(\mathbf{x})$ within the two classes C_1 and C_2 are estimated as:

$$\begin{aligned}\hat{\sigma}_1^2 &= \frac{1}{n_1 - 1}\sum_{j \in C_1}(c(\mathbf{x}_j) - \bar{c}^{(1)})^2 \\ \hat{\sigma}_2^2 &= \frac{1}{n_2 - 1}\sum_{j \in C_2}(c(\mathbf{x}_j) - \bar{c}^{(2)})^2.\end{aligned} \tag{5.5}$$

Finally, Wright et al. [101] compute class posterior probabilities that a sample with an expression profile $\mathbf{x}$ is member of class one by

$$\hat{p}(k = 1|\mathbf{x}) = \frac{p(c(\mathbf{x})|k = 1)}{p(c(\mathbf{x})|k = 1) + p(c(\mathbf{x})|k = 2)}. \tag{5.6}$$

Given that class one represents GCB and class two represents ABC, Wright et al. [101] classify lymphomas

$$\mathcal{C}_{GCB/ABC}(\mathbf{x}) = \begin{cases} \text{GCB} & \text{if} \quad \hat{p}(k = 1|\mathbf{x}) > 0.9 \\ \text{ABC} & \text{if} \quad \hat{p}(k = 1|\mathbf{x}) < 0.1 \\ \text{unclassified} & \text{if} \quad 0.1 \geq \hat{p}(k = 1|\mathbf{x}) \geq 0.9. \end{cases} \tag{5.7}$$

A cross platform signature approach The original GCB/ABC signature is based on data generated with the Lymphochip [2], which is a substantially different microarray platform than that provided by Affymetrix. A direct application of signatures derived from the Lymphochip to data generated with Affymetrix microarrays is not possible, since the measurements are on different scales and the genes present on both platforms differ. Wright et al. [101] propose a cross platform translation of their Lymphochip signature to Affymetrix microarrays. They describe a special case, where the goal is to apply the Lymphochip signature to a lymphoma data set generated with Affymetrix HU6800 microarrays [80]. Of the 27 Lymphochip signature genes 14 are present on HU6800 microarrays. Some of these 14 genes are measured by more than one probe-set of the Affymetrix microarray. In this case, Wright et al.[101] suggest to average multiple measures of the same gene within each sample. Due to the technological differences between the Lymphochip and the Affymetrix platform, Wright et al.[101] shift and scale the Affymetrix gene intensities to the same mean and variance across the samples than in the corresponding Lymphochip data.

We follow the strategy proposed by Wright et al. [101] to add the GCB/ABC labels to our series of 220 lymphoma expression profiles, which is based on Affymetrix HGU133A microarrays. We are able to identify HGU133A probe-sets for 15 out of the 27 Lymphochip signature genes listed in table 5.1. Like in Wright et al. [101], we derive a "truncated" signature using only the subset of genes present of both the Lymphochip and the Affymetrix platform and start with testing, if this "truncated" model still predicts at least the original Lymphochip data correctly. The Lymphochip data set and the gene weights accompany the publication of Wright et al. online [101]. Table 5.2 compares the GCB/ABC-predictions of the original Lymphochip

Table 5.1: Genes selected for GCB/ABC compound covariate predictor. The table lists the Lymphochip features of the GCB/ABC predictor that have counterparts on Affymetrix HGU133A GeneChips. Note, several lymphochip features match multiple Affymetrix probe-sets.

Lymphochip	Affymetrix	Symbol	Description
19346	201810_s_at, 201811_x_at	SH3BP5	SH3-domain binding protein 5 (BTK-associated)
16049	209374_s_at, 211632_at, 211634_x_at, 211635_x_at, 211637_x_at, 211638_at, 211650_x_at, 211908_x_at, 212827_at, 214916_x_at, 215621_s_at, 216363_at, 216371_at, 216372_at, 216491_x_at, 216510_x_at, 217035_at, 217198_x_at, 217217_at	IGHM	immunoglobulin heavy constant mu
24729	204562_at	IRF4	interferon regulatory factor 4
24899	209193_at	PIM1	pim-1 oncogene
27565	207691_x_at, 209473_at	ENTPD1	ectonucleoside triphosphate diphosphohydrolase 1
17227	209827_s_at	IL16	interleukin 16 (lymphocyte chemoattractant factor)
29385	207655_s_at	BLNK	B-cell linker
16858	200951_s_at, 200953_s_at	CCND2	cyclin D2
26385	203434_s_at, 203435_s_at	MME	membrane metallo-endopeptidase (neutral endopeptidase, enkephalinase, CALLA, CD10)
24904	212975_at	KIAA0870	
24429	203140_at	BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)
27673	204674_at, 35974_at	LRMP	lymphoid-restricted membrane protein
17496	213906_at	MYBL1	v-myb myeloblastosis viral oncogene homolog (avian)-like 1
17218	204249_s_at	LMO2	LIM domain only 2 (rhombotin-like 1)
28338	203723_at	ITPKB	inositol 1,4,5-trisphosphate 3-kinase B

Table 5.2: The GCB/ABC model is originally based on a Lymphochip data set. The table shows the classification of the original Lymphochip samples classified based on the full 27 gene model (rows) and the "truncated" 15-gene model (columns).

	ABC	GCB	unclassified
ABC	67	0	4
GCB	0	105	6
unclassified	7	7	44

data using the 27-gene signature to the predictions obtained from the "truncated" 15-gene signature. We do not observe any label swapped between GCB and ABC. In conclusion, the 15-gene signature performs well in recovering the original 27-gene prediction.

In order to add the GCB/ABC labels to Affymetrix data, Wright et al. [101] shift and scale Affymetrix gene intensities to the same mean and variance across the samples than in the corresponding Lymphochip data. It is important to note that this procedure explicitly assumes that the composition of the two study populations is the same, the population underlying the Lymphochip data, and the population underlying the Affymetrix data. Our data set comprises Burkitt lymphomas and diffuse large B-cell lymphomas, while the Lymphochip data set underlying the GCB/ABC signature only comprises diffuse large B-cell lymphomas. In order to make the two data sets comparable we remove the molecular Burkitt lymphomas from our series and restrict the GCB/ABC analysis to non-mBL and intermediate cases mainly comprising cases of histomorphological diagnosis diffuse large B-cell lymphoma. We obtain 79 of our non-mBL/intermediate lymphomas classified as GCB, and 58 as ABC. 39 remain unclassified with respect to equation 5.7. Figure 5.1 shows the expression of the signature genes and the class probabilities of GCB and ABC in 176 non-mBL and intermediate MMML lymphomas. Since GCB/ABC represents a purely transcriptional profiling based classification of lymphomas, we cannot compare the results from the compound covariate prediction to some external classification that allows validation. However, we can confirm the plausibility of our results by a several important observations:

- In line with previous reports [3, 74, 101] patients with GCB lymphomas have a significantly better overall survival rate.

- It has been reported that t(14;18) genomic translocations exclusively occur in GCB lymphomas [41]. Our data set contains 25 cases with t(14;18) translocations; 1 of them occurs in an mBL tumor, 1 in an ABC lymphoma, and 23 in GCB lymphomas. Thus, we can confirm the over-representation of t(14;18) translocations in GCB lymphomas.

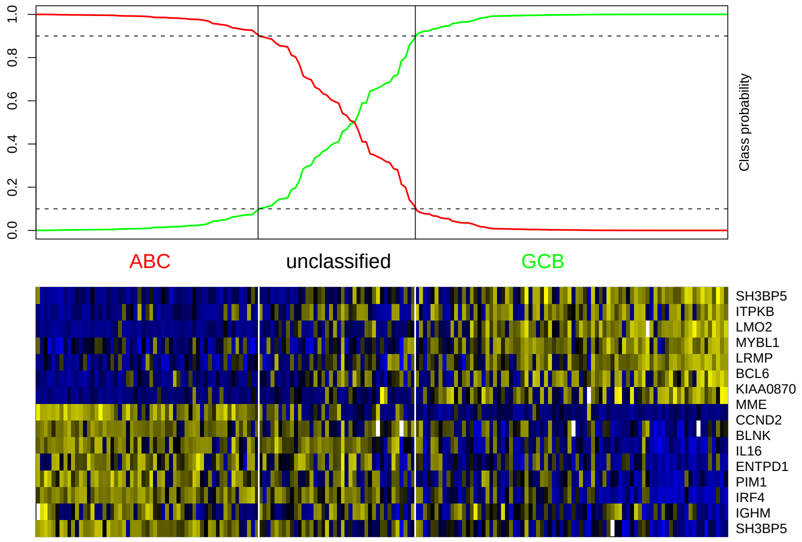

Figure 5.1: Expression of GCB/ABC signature genes in MMML data. The top panel shows GCB (green) and ABC (red) class probabilities of 176 non-mBL and intermediate MMML data set lymphomas (x-axis). Probabilities have been computed according to equation 5.6, and samples have been classified according to equation 5.7. The lower panel shows the expression of the 15 signature genes (table 5.1), where bright yellow encodes expression 3 SD above the mean over all cases and light blue encodes expression 3 SD below the mean.

- We know that gains of chromosome 18q are more frequent in ABC lymphomas than in GCB lymphomas [9], and we can confirm this observation on the MMML data set [23].

The last item regarding the association of ABC lymphomas and gains of regions on chromosome 18q will be the topic of the next section.

5.3 Dissecting the dependency between ABC and 18q/MALT1 gains

Array CGH and FISH analysis has been carried out on 116 DLBCL out of the 220 MMML lymphoma samples (see [23]). This joint genetic analysis provides a binary label indicating, if an 18q/MALT1 gain is present (44 cases) or absent (72 cases) in each individual lymphoma. The compound covariate predictor provides the cell of origin labels GCB (49 cases), ABC (41 cases) and unclassified (26 cases) for the same 116 samples. We exclude the unclassified cases expressing an ambiguous cell of origin signature and explore the dependency of GCB, ABC and the 18q/MALT1 gain in 90 lymphoma samples. The starting point is a 2×2 contingency table representing absolute frequencies of an 18q/MALT1 gain in GCB and ABC lymphomas:

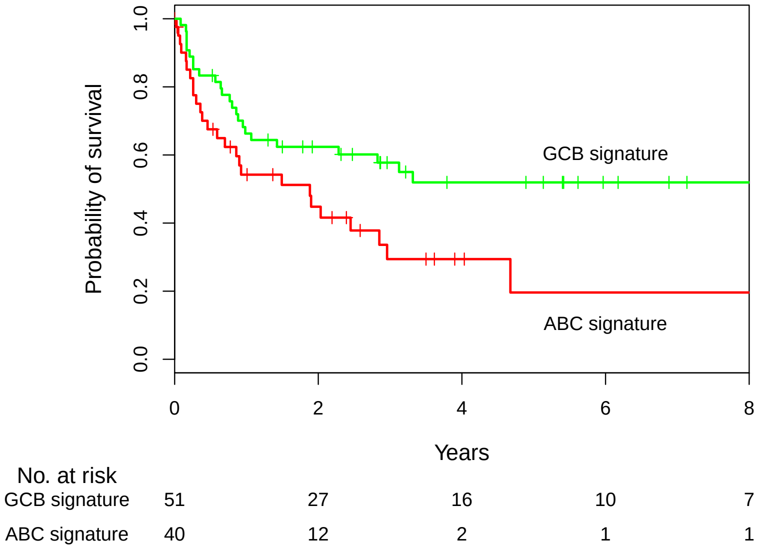

Figure 5.2: Kaplan-Meier estimates of survival according to the GCB/ABC signature. Overall survival among non-mBL/intermediate patients with an ABC-like transcriptional phenotype is significantly worse than that among non-mBL/intermediate patients with a GCB-like transcriptional phenotype (P=0.003 by the log-rank test).

	ABC	GCB
18q/MALT1 gain	29	10
no 18q/MALT1 gain	12	39

In order to test, if our data represents a significant deviation from the null hypothesis of an equal number of 18q/MALT1 gains in both, ABC and GCB lymphomas, we perform a Fisher's exact test (p-value: 1.9×10^{-6}). Thus, we reject the null hypothesis of equal proportions and conclude: 18q/MALT1 gains are more frequent in ABC lymphomas. Nevertheless, the presence of the genomic imbalance is not restricted to ABC lymphomas. We frequently observe GCB cases harboring an 18q/MALT1 gain. Thus, it is not possible to designate the presence of an 18q/MALT1 gain as potential pathogenicity mechanism specific for ABC lymphoma. Instead, we assume that the 18q/MALT1 gain is independent of the cell of origin GCB/ABC signature. The two factors 18q/MALT1 and GCB/ABC rather overlap with each other without any causative relationship. In order to identify the effect of 18q/MALT1 on gene expression taking into account GCB/ABC as confounder, we choose a linear modeling approach.

5.3.1 Methods

We use a regression approach to assess the problem (see e.g. Simon et al. [82]). Given two factors describing two different phenotypic characteristics and using the original notation of Simon et al. [82], we can model gene expression as a dependent variable y depending on the two predictor variables z and w:

$$y = \alpha + \beta_1 z + \beta_2 w + e \tag{5.8}$$

In the present situation we can think of z as a binary indicator variable with 1 indicating the presence of an 18q/MALT1 gain and 0 indicating the absence. Furthermore, w denotes a binary indictor with 1 encoding the ABC and 0 the GCB transcriptional phenotype. The coefficient α denotes an offset and each coefficient β represents the association between the level of the corresponding factor and the variance of the gene expression levels across the samples. We are interested in wether or not there is a significant association between gene expression and the factor z while taking into account the confounder w. Thus, we perform a test on the null hypothesis $H_0 : \beta_1 = 0$ and infer a p-value accordingly. In the MMML data set we measure transcriptional levels with 22283 different probe-sets in parallel. We use two different approaches to test the resulting data for an association of gene expression and the presence of an 18q/MALT1 gain, while taking into account GCB/ABC as confounder. We use a probe-set by probe-set approach and a global testing approach.

An individual probe-set test for differential gene expression

In this approach we fit the linear model from equation 5.8 separately to each of the 22283 probe-sets. We model the observed measurements of each individual probe-set as response variable depending on the two predictor variables 18q/MALT1 and GCB/ABC. We aim on identifying individual probe-sets, where we can reject the null hypothesis of no association between the 18q/MALT1 gain and gene expression. Thus, we perform 22283 independent tests in parallel. The "linear models for microarrays" (limma) of Gordon Smyth [84] provides the framework to implement the analysis. In line with the Benjamini and Hochberg [10] we control the false discovery rate (FDR) to deal with the problem of multiple hypothesis testing.

A global test for differential gene expression

Global testing is an alternative to testing each individual probe-set separately. In the global approach we test a single joint null hypothesis, instead of multiple individual null hypotheses. The joint null hypothesis is: The observed measures of multiple probe-sets together are random fluctuations only. Such a test allows for asking the question of wether or not a certain factor causes transcriptional changes at all. The GlobalANCOVA (global analysis of covariance) approach by Mansmann and Meister [57] and Hummel et al. [43] provides methodology to fit joint linear models to observed data from multiple probe-sets. It allows to infer a p-value on the global joint null hypothesis that a certain factor of a linear model does not change the joint transcription of a given set of probe-sets. At the same time we take into account further covariates as confounding factors. We use GlobalANCOVA to test for the global impact of 18q/MALT1 gains on lymphoma transcriptional profiles. Furthermore, we perform global tests on sets of probe-sets defined by chromosomal location.

5.3.2 Results

In order to determine differential gene expression between lymphoma cases with and without a gain of 18q/MALT1, we precede in several steps. First, we apply a global test for differential expression between 18q/MALT1-positive and 18q/MALT1-negative cases. We ignore GCB/ABC as possible confounding factor and model the global gene expression y using GlobalANCOVA with a single covariate:

$$y = \alpha + \beta z + e, \tag{5.9}$$

where z denotes a binary predictor variable indicating the presence (44 cases) or absence (72 cases) of an 18q/MALT1 gain. GlobalANCOVA returns an empirical permutation p-value of 0.004, if we test for the null hypothesis $H_0 : \beta = 0$. We implement an individual probe-set analysis using limma with the same linear model. We control the FDR at a level of 0.05 and obtain a list of 135 probe-sets showing differential expression between 18q/MALT1-positive and 18q/MALT1-negative cases. Notably, the list includes the probe-sets specific for MALT1 and BCL2. We conclude: There is differential gene expression between 18q/MALT1-positive and 18q/MALT1-negative cases. In the next step, we take into account that the ABC transcriptional phenotype strongly overlaps with the presence of an 18q/MALT1 gain and include this information in the linear model:

$$y = \alpha + \beta_1 z + \beta_2 w + \beta_3 zw + e_j, \tag{5.10}$$

were z denotes a binary variable encoding the presence and absence of an 18q/MALT1 gain and w denotes a binary variable encoding the ABC and GCB phenotype. The third coefficient $\beta_3 zw$ is an interaction term, i.e. we take into account that the effect of z depends on the state of w and vice versa. An example for an interaction, would be the observation of differences of the effect of an 18q/MALT1 gain depending on wether it occurs in the ABC or GCB phenotype. We do not test for the presence of an interaction, but we take it into account in our model.

Note, the analysis yielding the results from the linear model with 18q/MALT1 positivity as single binary predictor (equation 5.9) is based on the whole series of 116 lymphomas, where the information on an 18/MALT1 gain is available (44 18q/MALT1-positive, 72 18q/MALT1-negative). In order to analyze the data in line with the more complex model (equation 5.10) with GCB/ABC as additional predictor variable we exclude 26 samples unclassified with respect to the GCB/ABC signature.

GlobalANCOVA returns a permutation p-value of 0.41, if we test for the influence of an 18q/MALT1 gain on gene expression, while taking GCB/ABC into account. Thus, there is no statistical evidence for global ABC/GCB independent effects of an 18q/MALT1 gain.

Next we include prior knowledge on the chromosomal location of genes. We apply GlobalANCOVA to test for the influence of 18q/MALT1 to each chromosome (excluding X and Y chromosomes, 22 tests) separately. We adjust the resulting $n_{test} = 22$ p-values for multiple testing using the Bonferroni correction ($p_{adjust} = min(p \times n_{test}; 1)$). This test identifies chromosomes with global changes in gene expression between 18q/MALT1-positive and 18q/MALT1-negative lymphomas. Only chromosome 18 displays global changes in gene expression, which are independent of the ABC/GCB status ($p_{adjust} < 0.03$, figure 5.3). To further elucidate this finding, we analyze the probe-sets located on chromosome 18 in a probe-set by probe-set test using limma with the model specified in equation 5.10. Of the 332 probe-sets located on chromosome 18 we are able to generate a list of 47 affected probe-sets, if we control the false discovery rate at a level of 0.05.

In summary, 18q/MALT1 gains occur at a significantly higher frequency in ABC than in GCB lymphomas. Differential gene expression can be detected between 18q/MALT1-positive and -negative cases, at a global and an individual probe-set level. However, the significance of this observation is lost, if we add GCB/ABC as additional explanatory variable to the model. Nevertheless, probe-sets located in the amplified region of chromosome 18 show a consistent level of differential expression despite the confounding GCB/ABC stratification. This observation allows us to speculate that there is a gene-dosage effect of the genes located within the amplified region of chromosome 18, i.e. the transcriptional level of 18q genes correlates with the copy number of that chromosomal region. Furthermore, we can show this effect to be independent of GCB and ABC, and thus gains of 18q are possibly not the pathological event leading to ABC-like lymphomas.

Introduced in 2000 [3] the GCB/ABC stratification of diffuse large B-cell lymphoma has become an integral part of the classification scheme of lymphoma. In addition several genetic aberrations have been identified that show associations to GCB and ABC [41, 9, 44] allowing for speculation on possible disease mechanisms. However, it is important to note that GCB, ABC and their associations to certain genetic abberations are observations on data from human tumor samples. Since, traditional modes of biological inquiry often require the manipulation of a biological system [85], it is difficult or impossible to achieve insight into the actual disease mechanisms just by observing tumor samples. In the next chapter we will discuss an analysis strategy combining interventional experiments conducted in line with the traditions of molecular biology with observational clinical microarray data revealing highly recurrent patterns of pathway activation in mature aggressive B-cell lymphoma.

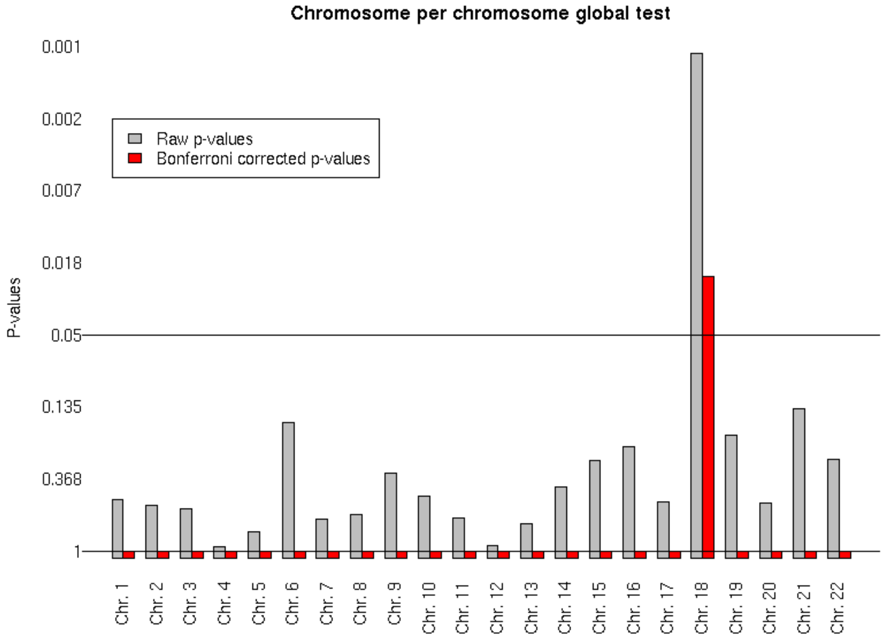

Figure 5.3: Global tests on influence of 18q/MALT1 gains on the expression of genes located on different chromosomes. The length of the bars encode the level of p-values. The scale of the y-axes is inverted and the ticks are on a logarithmic scale. Thus, long bars represent low p-values. Each pair of bars represent the result of another chromosome-specific global test for differential expression induced by an 18q/MALT1 gain. The gray bars denote the original GlobalANCOVA p-values and the red bars denote the Bonferroni corrected p-values. Only chromosome 18 shows a significant level of differential gene expression.

Chapter 6

Pathway activation patterns in malignant B-cell lymphoma

6.1 Introduction

Deregulation of cell signaling pathways controlling cell growth and cell survival is a common feature of all cancers. Modern molecularly targeted cancer drugs intervene in cell signaling compensating for pathway deregulation. Hence characterizing tumors with respect to pathway activation will become crucial for treatment decisions.

Bild et al. [12] have shown for carcinomas that oncogenic pathway signatures artificially induced in non-malignant breast epithelial cells can predict outcome and treatment efficiency. The authors transfected quiescent primary human mammary epithelial cells (HMECs) singly with each of the five human oncogenes: MYC, activated RAS, SRC, E2F3, and activated β catenin. They trained five discriminatory classifiers of oncogenic pathway activation on expression profiles from transfected HMECs and controls using a supervised Batesian classification model [97], and applied them to expression profiles of epithelial neoplasms including ovarian, breast, and lung cancer to predict the activity of each of the pathways in tumors. In this chapter we will extend the same approach to diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL) using the original HMEC cell line data from Bild et al.[12].

A naive approach to pathway signatures. We aim at using the original data from the oncogene over-expression experiments of Bild et al. [12] to predict oncogene expression in DLBCL and BL. The experimental design of Bild et al. [12] is the following: In a series of replicated biological experiments they grew primary mammary epithelial cell cultures (HMEC). They transfected them singly with each of the five human oncogenes MYC, activated RAS, SRC, E2F3, and activated β catenin, or with a control transgene (GFP=green fluorescent protein), subjected the samples to microarrays, and trained binary classifiers of pathway activity, each based on the comparison of the respective oncogene and the control microarray samples. As described earlier, molecular classifiers are based on a set of genes that show differential

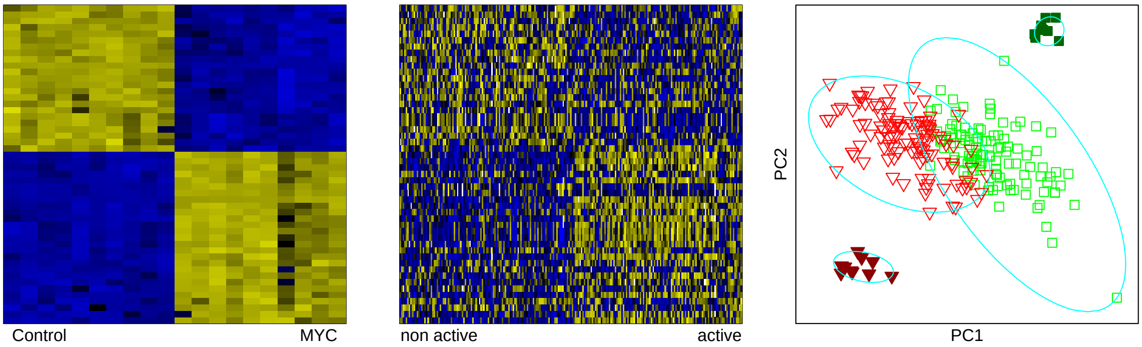

Figure 6.1: Expression of genes induced upon MYC-transfection of epithelial cell lines. The heat maps display expression levels of 50 genes most differentially expressed genes between control and MYC-transfected HMECs. Yellow encodes expression above the overall mean of a gene, blue expression below that mean and black average expression levels. The heat map on the left shows the expression in the HMEC samples. The heat map in the middle shows the expression of the same genes in lymphomas, where only little structure becomes apparent. The plot on the right displays all samples in the space spanned by the first two principal components generated by this gene set. Dominant is the difference between HMECs (filled symbols) and lymphomas (empty symbols). Green squares indicate MYC-transfected HMECs as well as corresponding lymphomas with gene expression more in the direction of the MYC transfected cell lines, red triangles indicate the control transfected status, respectively. (Figure reproduced from [11])

expression between conditions/classes. Bild et al. [12] followed these ideas and selected genes ”most highly correlated” with the oncogene vs. control comparison. We re-analyzed the original data of Bild et al. [12], selected the 50 genes with largest absolute t-scores (see equation 4.14) in the MYC transfection vs. control comparison, and explored the expression of them in our series of 220 mature aggressive B-cell lymphomas. The stratification of the lymphomas obtained from the genes selected on the HMEC data is not convincing (figure 6.1). Only very little structure becomes apparent in the lymphomas. Most likely, this failure is due to global expression differences between epithelial cell lines and lymphoid tumors caused by the different biology of the two tissues. We will now introduce a heuristic semi-supervised gene selection algorithm that selects genes based on two criteria: The genes must show differential expression in the HMEC oncogene vs. control comparison and the genes must yield a consistent – pathway based – stratification of the lymphomas. The method is a modification of the class-finding algorithm ISIS (”ISIS” for identifying splits of clear separation) [95].

6.2 Identification of splits of clear separation (ISIS)

Semi-, un-, or supervised? The objective of supervised classification is: Given a training set of label/data pairs $(y, \mathbf{x})_j$, find a function that correctly predicts the labels y_j given the data $\mathbf{x}_j$. In un-supervised classification one is interested in finding an optimal partition of the samples, given the data and an objective function (e.g. the likelihood of a Gaussian mixture model defined by equation 4.1). We can refer to un-supervised classification as clustering. In semi-supervised classification we combine ideas from clustering and supervised classification and device the objective: Find a partition of the samples that is optimally supported by the data (e.g. by maximizing the data likelihood) and assigns samples with known distinct labels to disjoint clusters and samples with known identical class labels to the same cluster. Core-group extension is a semi-supervised classification problem (see chapters 3 and 4). In chapter 4 we have approached it by modeling a gene expression index as univariate Gaussian mixture. The gene expression index has been computed from genes that we have selected prior to modeling by their correlation with the incomplete core group class labels. In contrast to semi-supervised classification, we don't have class labels in the un-supervised case, and we thus cannot perform gene selection. Nevertheless, we measure the expression of several thousand genes in transcriptional profiling. Different genes are associated with different biological pathways, and different pathways are associated with different diseases. Thus, different sets of genes support different stratifications of samples. In contrast clustering, which finds an optimal partition of samples given a set of genes, we device the objective of class discovery: Given a large set of possible genes, find subsets of them each supporting a tight clustering of the samples. The class-finding algorithm ISIS is a heuristic method implementing this objective: It searches bi-partitions of the samples supported by small subsets of genes.

The ingredients of the ISIS algorithm. A full description of ISIS can be found in [95]. The algorithm comprises two components: 1.) A scoring function based on diagonal linear discriminant analysis (DLD score) of two classes. 2.) A heuristic search for local maxima of the DLD score in the search space of 2^{n-1} possible bi-partitions of samples. The heuristic maximization part consists of a global candidate search and a local maximization of these candidate splits.

DLD score. The DLD score quantifies the discrimination of a given bi-partition of samples into n_1 members of class one and n_2 members of class two. It is defined on p genes with highest absolute two-sample t-score (see equation 4.14). In line with diagonal linear discriminant analysis each sample $\mathbf{x}_j$ is projected from the p-dimensional space onto the univariate line defined by

$$s_j = \sum_{i=1}^{p} \frac{\bar{x}_{i,1} - \bar{x}_{i,2}}{\hat{\sigma}_i^2}(x_{ij}), \tag{6.1}$$

were $\hat{\sigma}_i$ denotes the gene-per-gene pooled within class standard deviations (see equation 2.5). The projection s_j is a univariate expression index aggregating the expression of the p top t-scoring genes into a single number. The DLD score S_b is the two sample t-score resulting from the comparison of the values of s_j in the two candidate classes:

$$S_b = \frac{\bar{s}_1 - \bar{s}_2}{\sqrt{\hat{\sigma}_s^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{6.2}$$

were $\hat{\sigma}_s^2$ denotes the pooled standard deviation of s in the to candidate classes (see equation 2.5).

Search for candidate bi-partitions. The ISIS heuristic starts from a large set of candidate partitions generated from i cluster average gene expression profiles $\mathbf{x}_i^* = x_{i1}^*, x_{i2}^*, ..., x_{in}^*$. The average profiles are the product of hierarchical clustering of genes. For every profile i and every sample $j_{cut} = 1, ..., n$ the value $x_{ij_{cut}}^*$ defines a bi-partition given by the subsets $\mathbf{M}^- = \{j | x_{ij}^* \leq x_{ij_{cut}}^*\}$ and $\mathbf{M}^+ = \{j | x_{ij}^* > x_{ij_{cut}}^*\}$. A bi-partition is considered a candidate, if the absolute value of the t-score $t_{ij_{cut}} = t_i(\{\mathbf{M}^-, \mathbf{M}^+\})$ comparing $\mathbf{x}_i^*$ in the two sets $\mathbf{M}^-$ and $\mathbf{M}^+$ is larger than a certain quantile $1 - \alpha$ of the null distribution of $t_{ij_{cut}}$. Since the null distribution of $t_{ij_{cut}}$ in unknown, von Heydebreck et al. [95] simulated the expected quantiles of $t_{ij_{cut}}$ by a Monte Carlo approach.

Local optimization of candidate bi-partitions. The procedure described in the previous paragraph yields a set of candidate bi-partitions. Starting from each of them separately, von Heydebreck et al. [95] choose a greedy local optimization of the DLD score in the neighborhood of the candidate bi-partitions. Two bi-partition are considered neighbors if they differ in a single sample.

6.3 Biological assay instructed ISIS (BASIS)

ISIS is an un-supervised class-finding algorithm. In order to search for bi-partitions of tumor gene expression profiles, which are supported by the results of controlled cell line assays, we modified ISIS into a semi-supervised gene-selection procedure: Given the joint data set of tumors (in our special case lymphomas) and control or oncogene transfected cell lines, the goal of the modified algorithm is to identify genes that 1.) classify the cell lines into control and oncogene transfected samples and 2.) yield a well separated stratification of the tumors into two groups. One of these groups should show expression levels similar to the control cell line experiment and the second group should show expression levels similar to the cell line oncogene transfection experiment. We need to modify ISIS only with respect to candidate selection.

Modification of the candidate bi-partition search. In standard ISIS $\mathbf{x}_i^* = x_{i1}^*, x_{i2}^*, ..., x_{in}^*$ defines a cluster average gene expression profile, and $x_{ij_{cut}}^*$ defines a bi-partition given by the subsets $\mathbf{M}^- = \{j | x_{ij}^* \leq x_{ij_{cut}}^*\}$ and $\mathbf{M}^+ = \{j | x_{ij}^* > x_{ij_{cut}}^*\}$. It is considered as candidate bi-partition, if the t-statistic $t_{ij_{cut}} = t_i(\{\mathbf{M}^-, \mathbf{M}^+\})$ is large. In semi-supervised ISIS we merge the cell line and tumor data set. The

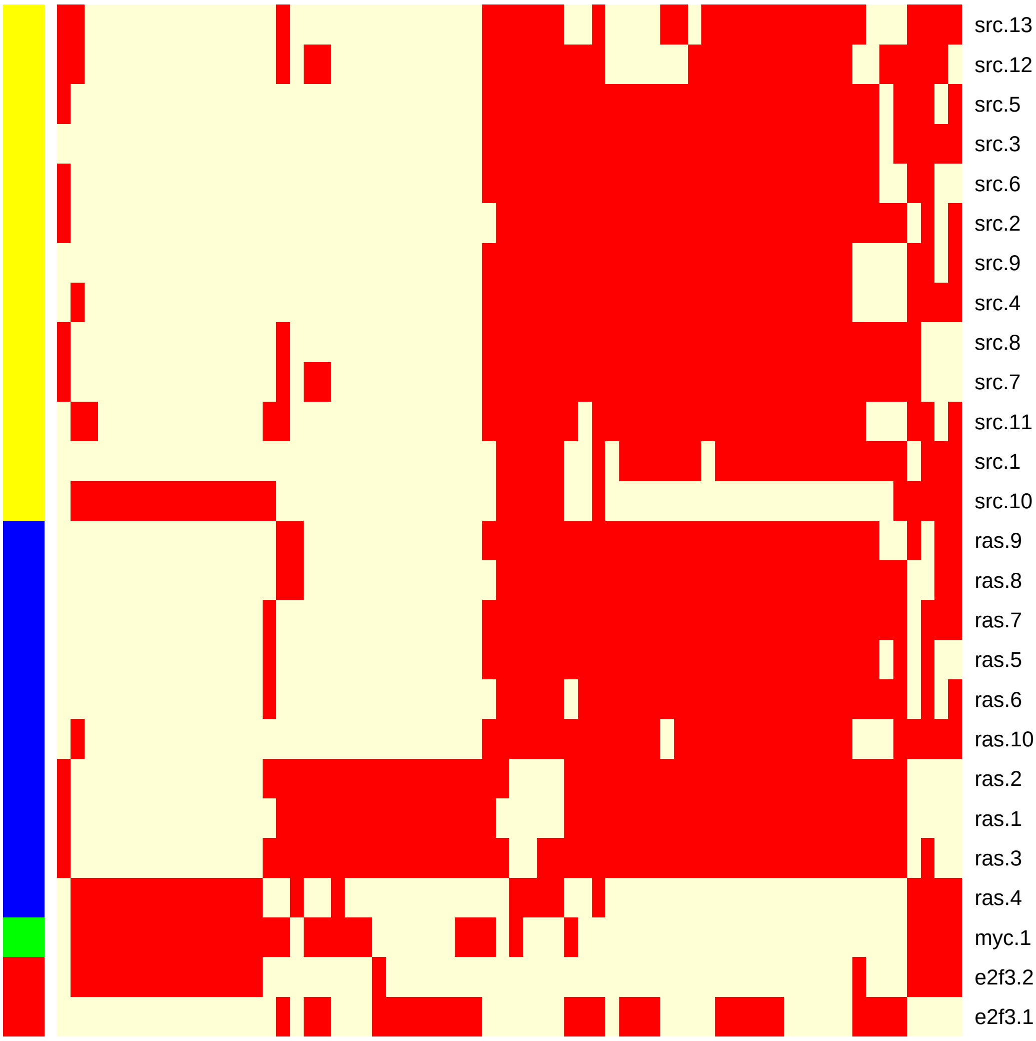

Figure 6.2: Lymphoma stratifications obtained from 26 stable BASIS solutions. Each column of the heatmap corresponds to an individual lymphoma and each row to one of 26 stable BASIS solutions. Red and beige encodes the expression levels of the genes, which constitute a bi-partition obtained as BASIS solution. Red indicates an expression level, which is similar to the oncogene transfected cell lines. Beige indicates that the same genes are expressed at a level similar to the control cell line. The color bar to the left highlights in yellow 13 bi-partitions obtained from BASIS on the combined data-set of lymphomas and SRC-transfected cell lines and in blue 10 bi-partitions resulting from combining lymphomas and RAS-transfected cell lines. Yellow indicates the unique solution for the MYC oncogene and red the two bi-partitions obtained from BASIS on lymphomas and E2F3-transfected cell lines. We do not obtain a stable BASIS solution if combining lymphomas and the β-catenin transfected cell lines. The bi-partitions are highly redundant. By visual inspection we choose MYC.1, E2F.1, E2F.2, SRC.2, SRC.10, RAS.1, RAS.4, and RAS.6 for further exploration.

samples are now partially labeled. They split into the tumors T (unlabeled), the control cell lines C (labeled as class one) and the oncogene transfected cell lines O (labeled as class two). A bi-partition defined by x^*_{ijcut} splits $\mathcal{D} = C \cup O \cup T$ into $\mathbf{M}^- = \{j|x^*_{ij} \leq x^*_{ijcut}\}$ and $\mathbf{M}^+ = \{j|x^*_{ij} > x^*_{ijcut}\}$. In semi-supervised ISIS a bi-partition must fulfill the condition below given by equations 6.3 in addition to large values of $t_{ijcut} = t_i(\{\mathbf{M}^-, \mathbf{M}^+\})$ to become a candidate bi-partition.

$$(C \subset \mathbf{M}^-) \wedge (O \subset \mathbf{M}^+) \vee (O \subset \mathbf{M}^-) \wedge (C \subset \mathbf{M}^+) \tag{6.3}$$

Local optimization of semi-supervised candidate bi-partitions. The procedure described in the last paragraph yields candidate bi-partitions of the tumors T. We proceed with the greedy local optimization as described in the standard ISIS algorithm [95].

Stability selection. In parts, the BASIS procedure can be considered supervised. Labeled cell line data enter the algorithm and we aim on finding a model that predicts the labels correctly. However, at the same time we search for a-priori unknown bi-partitions of the tumor samples. In that point we are in an un-supervised scenario and the validation becomes difficult, due to the lack of performance criteria like the error rate. Instead, we check the robustness of labels with respect to bootstrapping the training data:

1. Split data set into a training set $\mathcal{D}_{train}$ and a stability selection set $\mathcal{D}_{select}$.
2. Run BASIS to obtain bi-partitions of $\mathcal{D}_{train}$.
3. Choose bootstrap samples $\mathcal{D}_{train,boot}$ of the profiles in $\mathcal{D}_{train}$.
4. Train on $\mathcal{D}_{train,boot}$ a supervised DLDA model[1] of each bi-partition obtained in step 2.
5. Use each bootstrapped DLDA-model to predict labels on $\mathcal{D}_{select}$.
6. Repeat B-times from step 3.

For each bi-partition identified in step 2 we obtain B label predictions of each sample in $\mathcal{D}_{select}$. We select those bi-partitions from step 2 that yield the most consistent predictions on the samples in $\mathcal{D}_{select}$ during the bootstrap.

Consistency: We are given $B = 1000$ bootstrap predictions of the samples in $\mathcal{D}_{select}$. If an individual sample is predicted class 1 for 900 bootstrap samples and class 2 for 100, then the individual sample consistency c_j of the bootstrap predictions is $c_j = 90\%$ (see equation 6.4 below). The overall consistency is the average of the individual consistencies of all samples in the stability selection set. We define individual consistencies as

$$c_j = \max(\#\{n|y_{j,pred} = 1\}^B_{n=1}, \#\{n|y_{j,pred} = 2\}^B_{n=1})/B, \tag{6.4}$$

[1] The training phase of each DLDA model includes gene selection. The models are based on the p genes with highest absolute t-score

where $y_{j,pred}$ denotes a prediction of a sample j in the selection set by a DLDA model derived from bootstrapped training data. We obtain B such predictions for each sample in $\mathcal{D}_{select}$.

Like ISIS the proposed procedure identifies bi-partitions of clinical tumor gene expression profiles. However, their identification is not purely un-supervised but in-line with the results of controlled interventional cell line experiments. We thus refer to the procedure as *biological assay instructed identification of splits of clear separation* (BASIS).

Conserved oncogene inducible module. The core of the ISIS and BASIS algorithm is the identification of small sets of genes that induce bi-partitions of samples with respect to an optimal DLD score. BASIS searches for bi-partitions of samples instructed by labeled biological assay data. Here, we aim on combining epithelial cell lines as labeled biological assay data and lymphomas as unlabeled data that we intend to stratify. The biology of the underlying tissues (epithelial and lymphoid) is different. We can thus not expect to find the same expression patterns on epithelial cell cultures and on lymphoma tumor samples. The BASIS algorithm however explicitly searches for common features of the labeled biological assay data and the unlabeled data. It identifies patterns of gene expression, which are conserved in the different tissues. A set of genes that constitutes a BASIS bi-partition is a conserved gene expression module and since we consider oncogene activity we refer to it as *conserved oncogene inducible module.*

6.4 Results

BASIS run. Our goal of the present project was to use the original data from the oncogene over-expression experiments of Bild et al. [12] to predict oncogene expression in our series of 220 DLBCL or BL samples. In contrast to the naive approach described in section 6.1 for the MYC-transfection, we will now present the results from the BASIS approach. In order to derive conserved oncogene inducible modules, we split the lymphoma samples randomly into a training and a test set, balanced with respect to the genetic classification of the samples based on the MYC-break and the gene expression based classification described in chapter 3. The training set contains 25 IG-MYC-fusion positive mBL-samples, 10 IG-MYC-fusion positive intermediate samples, 10 MYC-negative intermediate samples and 55 MYC-negative non-mBL samples. The samples within each group have been chosen randomly from the data-set. In order to choose $\mathcal{D}_{train}$ and $\mathcal{D}_{select}$, we further split the training set in three equally sized batches ($n_1 = 33, n_2 = 33, n_3 = 34$) balanced for appearance of an IG-MYC-fusion and the mBL diagnosis. Two batches have been combined to $\mathcal{D}_{train}$ and the remaining batch has been used as stability selection set $\mathcal{D}_{select}$. $\mathcal{D}_{train}$ has been combined with the original data from [12] containing control HMEC samples and HMEC samples transfected with one of five human oncogenes. Each HMEC-microarray sample has been added twice to the merged data set, thus increasing

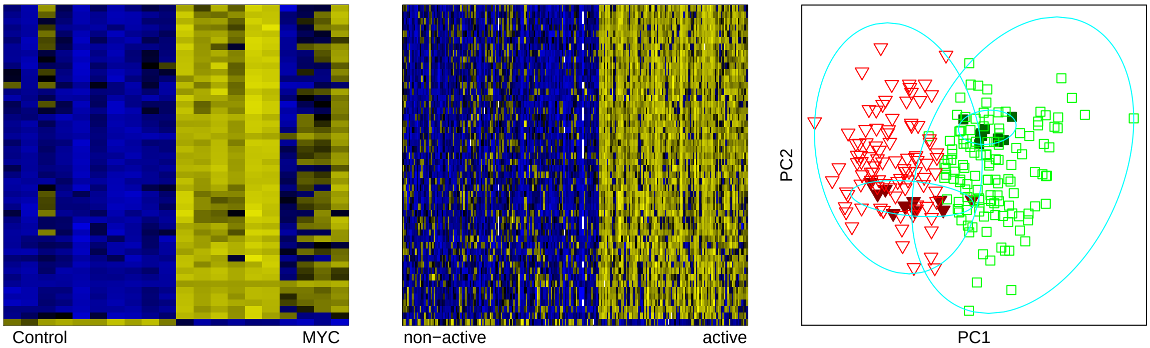

Figure 6.3: Expression of genes constituting the MYC.1-conserved oncogene expression module. The heat maps display expression levels of the 50 genes, which are members of the MYC.1 conserved oncogene expression module identified by BASIS. The heat map on the left shows their expression in the HMEC cell lines. The heat map in the middle column shows the expression of the same genes in lymphomas. The plot on the right displays all samples in the space spanned by the first two principal components generated by this gene set. The difference between HMECs (filled symbols) and lymphomas (empty symbols) is no longer dominant as in figure 6.1. Green squares indicate MYC-transfected HMECs as well as lymphomas with gene expression similar the MYC transfected cell lines, red triangles indicate the control transfected status, respectively. (Figure reproduced from [11])

the weight of the cell line assays in the BASIS-runs. In summary, we subjected five data-sets to semi-supervised BASIS analysis, each containing 66 lymphoma samples in $\mathcal{D}_{train}$, $2 \times 10 = 20$ control HMEC samples and $2 \times n_c$ HMEC samples transfected with either of the oncogenes (with $n_c = 10, 10, 7, 9$ or 9 for MYC, activated RAS, SRC, E2F3, or activated β catenin). For each of the five data-sets, we shifted the gene expression in HMEC samples by a gene-specific offset so that the gene-wise mean across samples becomes the same in lymphomas and HMECs. Each BASIS run returned a set of possible bi-partitions of the lymphoma samples, which classifies them into a group with gene expression similar to the control HMECs and another group with gene expression similar to the respective oncogene transfected cells. From stability filtering we obtained 26 stable bi-partitions (average consistency$>$ 95%). Figure 6.2 shows the respective 26 different stratifications of the lymphomas obtained. Some of them are highly redundant. Thus, we decided by visual inspection of figure 6.2 to further investigate only a subset of 8 of the 26 bi-partitions (MYC.1, E2F.1, E2F.2, SRC.2, SRC.10, RAS.1, RAS.4, and RAS.6). Note, we did not obtain any stable BASIS solution if combining lymphomas and the β-catenin transfected cell lines.

Conserved oncogene expression modules. Section 6.1 presents the results from a naive approach combining the MYC-transfection experiments of epithelial cell cultures with our series of 220 lymphomas. Figure 6.3 shows the set of genes (conserved oncogene expression module) of the MYC.1-BASIS solution. This gene set induces an apparent stratification of the lymphomas into two groups. One of the two groups show expression levels similar to the MYC-transfected cells and the other shows expression levels similar to the control cells. The BASIS algorithm identified a clear bi-partitions of unlabeled samples in-line with labeled biological assay data.

Pathway activation patterns. MYC, RAS, SRC and E2F3 oncogenes, but not β-catenin yield eight conserved oncogene inducible modules in a training set of $n = 100$ mature aggressive B-cell lymphomas (MYC.1, E2F3.1, E2F3.2, SRC.2, SRC.10, RAS.1, RAS.4, and RAS.6). Each of these oncogene inducible modules is either active or non-active in a lymphoma (activation state). Active means that a lymphoma expresses an oncogene inducible module at the level, which is similar to the expression level in the oncogene transfected cell lines. We combined the eight activation states to binary patterns, which we call a PAP for Pathway Activation Pattern. PAPs define non-overlapping groups of lymphomas from the perspective of pathway activation. Figure 6.4 shows the patterns together with the underlying expression data. Expression characteristics from the training set reemerge well in the test samples ($n = 120$). Importantly, 158 of the 220 (72%) mature aggressive B-cell lymphomas show only five of the possible $2^8 = 256$ PAPs. The remaining 62 samples display distinct, rarely or non recurrent patterns (recurrence in $< 5\%$ of the lymphomas). We subsume them in a heterogeneous pool we refer to as mind-L for molecularly individual lymphomas.

Burkitt lymphoma is characterized by a distinct pathway activity pattern. According to chapter 3 and references [42, 20] molecular Burkitt lymphoma (mBL) is a homogeneous lymphoma entity with respect to molecular, genetic and clinical features. Also here, we found only one recurrent pathway activation pattern (BL-PAP) among the mBL cases. This pattern is expressed in 39 of the 44 (89%) mBL cases. Vice versa, 39 of the 41 (95%) lymphomas displaying the pattern show the mBL signature. Consistent with activity of the MYC.1 module in the BL-PAP, in 38 of 40 of these lymphomas, where FISH data are available, a MYC-breakpoint has been detected.

Conserved module activity patterns and lymphoma stratification. In addition to the BL-PAP, we find four recurrent patterns in DLBCL, which we term PAP-1 to PAP-4 (figure 6.4). A total of 42 DLBCLs show the most frequent pattern PAP-1, which is exactly the inverse of the BL-PAP. Modules, which are active in PAP-1 are non-active in BL-PAP and vice-versa The second pattern, PAP-2, is present in 32 DLBCLs. PAP-1 and -2 lymphomas frequently expressed the BCL6 protein (28/37, or 76% and 25/31, or 81%, respectively) but rarely CD10 (7/33, or 18% and 4/27, or 13%, respectively). PAP-3 is the only pattern, except for BL-PAP, that display activation of the MYC.1 module, although this does not commonly arise through MYC translocation, as a break is present in only 3 of 27 cases suggesting alternative means of pathway activation. PAP-4, which we find in 16 lymphomas, is the only activation pattern more prevalent in females (11/16). Unlike BL, neither activated B-cell like DLBCLs (ABC-DLBCL, PAP-1: 8 cases, PAP-2: 20 cases, PAP-3: 14 cases, PAP-4: 5 cases; mind-L: 11 cases) nor germinal center like DLBCLs (GCB-DLBCL,

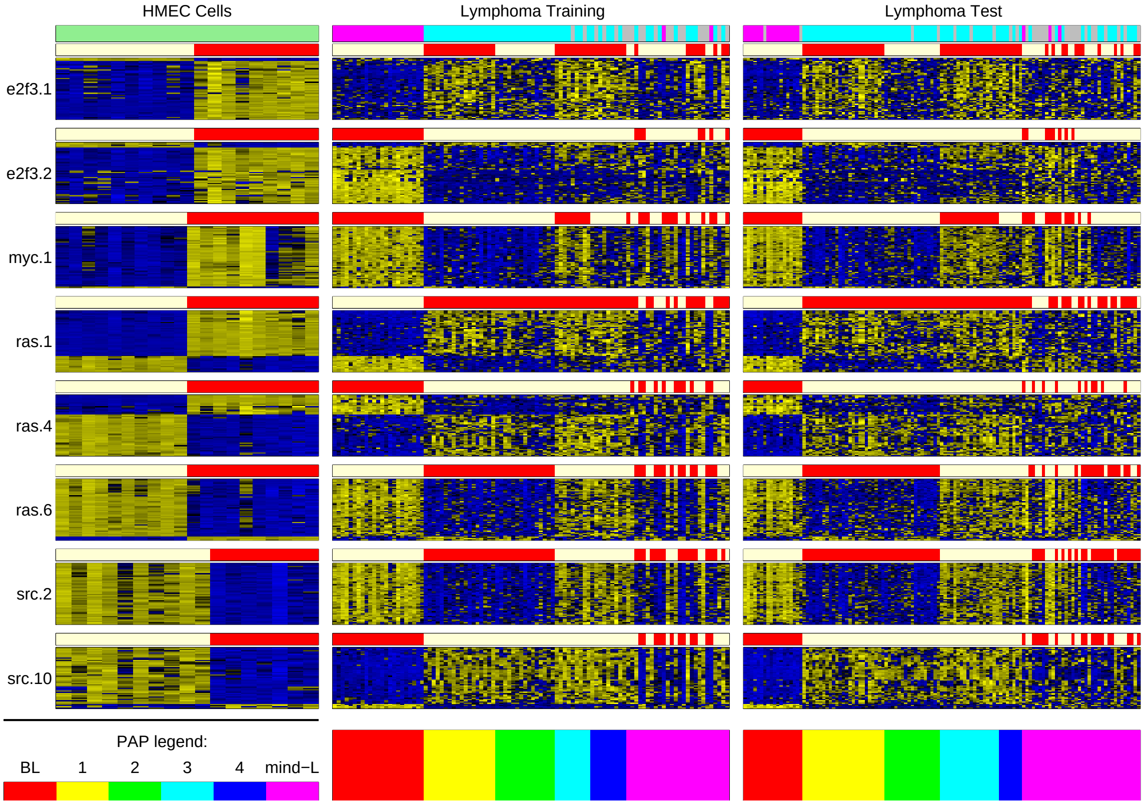

Figure 6.4: Stratification of lymphomas based on pathway activation patterns. The eight rows correspond to the conserved oncogene inducible modules. In each row the left heat map shows expression of the module genes in HMECs, the middle heat map shows the expression of the same genes in the training samples of the lymphoma data, while the right one refers to the test samples of the lymphoma data. The samples are sorted by module activation patterns starting with BL-PAP on the left and ending with mind-L on the right (see the color coding in the bar below the heat maps). Above each row of heat maps is a bar indicating module activation in red. The pattern of module activation is constant in each of the groups BL-PAP, PAP-1, PAP-2, PAP3 and PAP-4 but heterogeneous in the pool mind-L. The horizontal bar on top of all plots encodes the type of samples (lightgreen: HMECs, cyan: non-mBL, magenta: mBL, gray: intermediate). (Figure reproduced from [11])

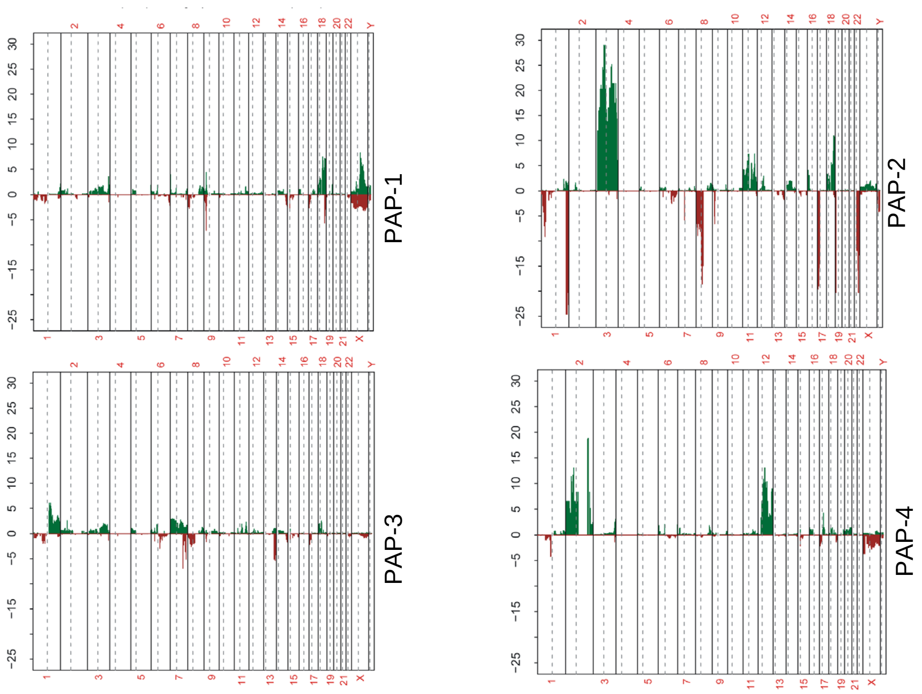

Figure 6.5: Specificity of chromosomal imbalances detected by Array-CGH for PAP 1-4.. The x-axis shows the genomic position of a clone. Chromosomes are separated by solid and chromosome arms by dashed vertical lines. The y-axis shows the chi-square score measuring over-representation of gains and losses in the respective PAP-groups. Scores corresponding to losses are shown with a negative sign for clarity. Over-representation tests are computed for the contrasts PAP-1 (top left), PAP-2 (top right), PAP-3 (bottom left) and PAP-4 (bottom right) vs. the respective remaining cases without the BL-PAP group. (Figure reproduced from [11])

PAP-1: 19 cases, PAP-2: 5 cases, PAP-3: 7 cases, PAP-4: 9 cases; mind-L: 37 cases) display a unique pathway activation pattern. Finally, mind-L includes 69% (33/48) of intermediate cases defined in chapter 3 with an mBL-index between mBL and non-mBL. In line with this, a high number of lymphomas with MYC-breakpoints are included in this pool (28/61, or 46% of mind-L cases), which turns out to be particularly enriched for lymphomas with non-IG-MYC fusion and MYC complex status (11/15, or 73% and 22/33, or 66% of all non-IG-MYC and MYC complex cases, respectively).

The PAP-groups show distinct patterns of chromosomal changes and FOXP1 expression.[2] [3] Data on genomic imbalances from 183 cases of our data set determined by array-based comparative genomic hybridization (CGH) have been described previously [42]. Every clone on an array-CGH was either classified as showing genomic gain, loss, normal copy number, or it was called missing if it could not be classified [49]. We computed a χ^2-test for each clone in order to test for over-representation of gains and losses in PAP groups. Missing values were removed before the computation of each statistic. Multiple testing adjustment was performed with the step-down minP method implemented in the R package multtest [71].

Despite PAP-2 differing transcriptionally from PAP-1 only with respect to the first E2F3 inducible module E2F3.1, it shows a profile of chromosomal changes markedly distinct, not only from PAP-1, but also from PAPs -3 and -4 (figure 6.5). The most frequent chromosomal changes in PAP-2 detected by array-CGH are gains at 18q21 and 3q27 (n=14/29 PAP-2 cases, or 48%). Gains at 18q21 containing the BCL2 and MALT1 oncogenes and 3q27 containing the BCL6 oncogene have been shown to be associated with ABC-DLBCL [9]. Notably, in PAP-2 also GCB-DLBCLs show MALT1/18q21 gains (2/5) and lack IGH-BCL2 fusions (n=5) like ABC-DLBCLs, indicating genetic similarity of GCB- and ABC-like lymphomas in PAP-2. The most specific chromosomal change in PAP-2 is a gain at 3p13. Indeed, of nine lymphomas displaying a 3p13 gain, eight belonged to PAP-2 and only one to PAP-3. The specifically gained region in PAP-2 at 3p13 contains FOXP1, which has been proposed as oncogene involved in lymphomagenesis as well as a prognostic marker in diffuse large B-cell lymphoma [27, 86, 98, 7, 8]. Immunohistochemical staining of the FOXP1 protein in 54 cases (PAP-1: $n_1 = 27$, PAP-2: $n_2 = 27$) reveals a significantly increased expression of the transcription factor in PAP-2 as compared to PAP-1 ($P = 0.006$ by a one-sided Wilcoxon's test for a shift of PAP-2 intensities toward higher expression). Furthermore, the 3p13-gain positive cases express the FOXP1 protein at high levels indicating an association between FOXP1 expression, 3p13-gains and the PAP-stratification.

The PAP groups are also present in an independent data-set. We derived the oncogenic pathway modules on a series of expression profiles generated with Affymetrix HGU133 gene chips. The LLMPP transcriptional profiling study already described in section 3.3 and published by Dave et al.[20] contains a similar data set with 303 lymphoma profiles. However the profiles have been generated with a custom Affymetrix oligonucleotide microarray (LymphDx 2.7k) with 2524 unique genes that are expressed differentially among the various forms of non-Hodgkins lymphoma. This array holds only a small fraction of the genes of the HGU133-arrays. Dave et al.[20] furthermore hybridized 99 tumors in parallel to both, the HGU133 platform and

[2] The analysis of array-CGH data in the context of pathway activation patterns has been gratefully carried out by Maciej Rosolowski from the Institute for Medical Informatics, Statistics and Epidemiology at the University of Leipzig.

[3] Protein expression of the FOXP1 protein has been gratefully evaluated by expert pathologists Wolfram Klapper (University of Kiel), Andreas Rosenwald (University of Wuerzburg), and German Ott (Robert-Bosch-Krankenhaus, Stuttgart)

LymphDx platform. To identify the pathway activation states on the 303 LymphDx profiles we need to translate the quantitative DLDA-based pathway signatures from the HGU133 gene chip scale to the scale of the LymphDx platform. We used the 99 tumors hybridized in parallel to both platforms the HGU133 and the LymphDx chip. We started by adding the pathway activation labels to the HGU133 profiles of the 99 tumors. Then, we used these added labels and the LymphDx profiles of the same 99 tumors to learn LymphDx-chip-based pathway classifiers. Each of the cross platform classifiers includes 25 genes, which are present on both platforms. We where thus able to confirm our pathway activation patterns on an independent data set of 303 mature aggressive B-cell lymphomas. The four recurring DLBCL-PAPs and the BL-PAP identified in our data-set are also the most recurrent PAPs in the 303 mature aggressive B-cell lymphomas of Dave et al.[20]. This indicates that the five PAPs widely cover the spectrum of mature aggressive B-cell lymphomas from the perspective of pathway activation constellations. Furthermore, 49 of the 55 lymphomas exhibiting the BL-PAP in the dataset of Dave et al.[20] also show the BL-signature defined in this paper confirming that the BL-PAP pattern is characteristic for BL.

Survival analysis. Clinical data including information on therapy and the two parameters age and Ann Arbor stage both used in the international prognostic index (IPI [79]) have been available for 134 cases of our own study and for 220 cases of Dave et al. [20]. We analyzed the association of survival with eight oncogenic modules and their combination to four recurrent PAPs (PAPs 14) by fitting multivariate Cox proportional hazard models [18]. The analysis was carried out for each study separately and for the pooled data taking both studies together. For the pooled data we used stratified Cox models assuming separate baseline hazard functions for both studies. The ABC signature is an established prognostic indicator and age and Ann Arbor stage are part of the IPI. In our analysis we have included the presence of an ABC signature, age (age>59 years) and stage (stage=III or IV) as covariates in the multivariate models, such that the estimated hazards are independent of them. Furthermore, the analysis was restricted to patients having received a treatment with a combination of chemotherapy based on cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) or similar and not belonging to the BL-PAP, which is strongly associated with mBL. Altogether, the survival analysis includes 81 of our own cases and 186 cases from Dave et al. [20].

Figure 6.6 summarizes the results of survival analysis of BL-PAP negative cases from both data sets, our own one and the one published by Dave et al.[20]. Shown are Cox proportional hazard ratios [18] estimated from multivariate analysis adjusting for ABC status, age, and Ann Arbor stage together with their 95% confidence intervals. In our own collection of lymphomas the PAP-1 group has a significantly better prognosis then non-PAP-1 DLBCLs (hazard ratio for death, 0.25; 95% CI: 0.1-0.65; P=0.004), while PAP-2 is a group with a significantly worse prognosis (hazard ratio for death, 2.45; 95% CI: 1.16-5.17; P=0.019). The same trends can be observed in the study of Dave et al.[20], although statistical significance is not reached. Addressing the question, as to which molecular features are responsible for the prognostic differences, we extended the survival analysis to the level of individual oncogene inducible modules. In the pooled analysis of both data-sets, patients with lymphomas showing activity of the first E2F3 inducible module, E2F3.1, have a significantly better prognosis (hazard ratio for death, 0.47; 95% CI: 0.33-0.67; P=0.00003). This prognostic effect, which is independent of the cell of origin signature and the established clinical risk factors age and Ann Arbor stage, is also visible in both individual studies. In contrast, patients with RAS.4 active lymphomas, which are restricted to and comprise almost one third of the mind-L lymphomas, display consistently worse outcome across both studies.

6.5 Discussion

By introducing conserved oncogenic transcriptional modules to the molecular pathology of lymphomas we have structured this disease from the perspective of oncogenic pathway activation. The PAPs identified four novel biologically homogenous subgroups among the DLBCLs. In contrast, our approach describes mBL as a lymphoma with a uniform oncogenic pathway pattern, in line with our recent molecular definition of this lymphoma type. Thus, mBL truly represents a single lymphoma entity.

More importantly, the PAPs identified four novel biological subgroups among the DLBCLs with homogenous pathway activation constellations. Most remarkable are the differences between the two largest groups PAP-1 and PAP-2. First, we observed strong prognostic effects. Second, PAP-2 is characterized by accumulated genetic aberrations on several chromosomes, which are found only on baseline frequencies in PAP-1 lymphomas. Moreover, protein expression of FOXP1 is significantly higher in PAP-2, in line with frequently observed gains on chromosome 3p13 around the locus of this gene. Notably, FOXP1 constitutes a target for IGH-translocations in DLBCL and MALT-type lymphomas [27, 86, 98] and expression of the FOXP1 protein, a member of the forkhead box (FOX) transcription factor family, has also been reported to be associated with poor prognosis in DLBCL [7, 8]. These differences are even more striking given that PAP-1 and PAP-2 only differ with respect to the activity of the E2F3.1 module.

We have introduced two concepts: oncogene-inducible modules and PAPs. Both have merits of their own. The first E2F3-inducible module, E2F3.1, and the RAS-inducible module, RAS.4, appear to be the strongest prognostic markers. However, modules do not group patients, as they overlap. A patient is not either E2F3.1-positive or RAS.4-positive, but can also have both features or none at all. Moreover, no single module on its own characterizes BLs. In contrast, PAPs define nonoverlapping lymphoma groups; a feature that is important in view of treatment decisions or molecularly stratified clinical studies.

All oncogene-inducible pathways are active in some DLBCL and inactive in others, supporting the hypothesis that DLBCL as a whole is a biologically heterogeneous lymphoma entity. However, the accumulation of 72% of BL and DLBCL cases in only 5 of the 256 possible PAPs shows that the biological processes underlying the conserved modules are not independently regulated in lymphomas. In contrast, their regulatory interaction characterizes mBL and the four biologically homogenous groups of DLBCL.

Among DLBCLs (PAPs 1-4), our survival analysis suggests that oncogene module activation patterns have clinical significance and are associated with overall survival. Similar effects were observed in an independent data set. Although statistical significance and reproducibility in a second data set were achieved, it is important to note that sample sizes are small, studies are retrospective and patients were not treated with today's state of the art treatment combining rituximab with CHOP (R-CHOP). Thus definitive conclusions concerning the prognostic value of both modules and PAPs require further studies, although certain trends are visible.

The oncogenes analyzed here have been chosen in the original study by Bild et al.[12] due to their prominent role in breast, ovarian and lung cancer. Although they do not give a complete picture of oncogenic pathway activity in lymphomas (for example BCL6, BCL2, MUM1 and BLIMP1, which have not been analyzed), the strong conservation of downstream transcriptional modules is remarkable and underlines the general importance of these pathways in tumor genesis.

The PAPs identified four novel biologically homogenous subgroups among the DLBCLs, which could guide the design of stratified prospective randomized studies on the efficiency of treatment modalities. In the future, PAPs could direct the development of inhibitors specific for oncogene-driven pathways characteristically activated in our pathway-defined lymphoma subgroups. Based on the conservation of the oncogenic modules across various solid and hematological cancers, targeted molecular-based therapies might well be effective in different kinds of tumors irrespective of localization or tissue derivation. Finally, conservation of oncogenic modules across cancers may also help to explain why some widely used anticancer drugs are potent in different cancers.

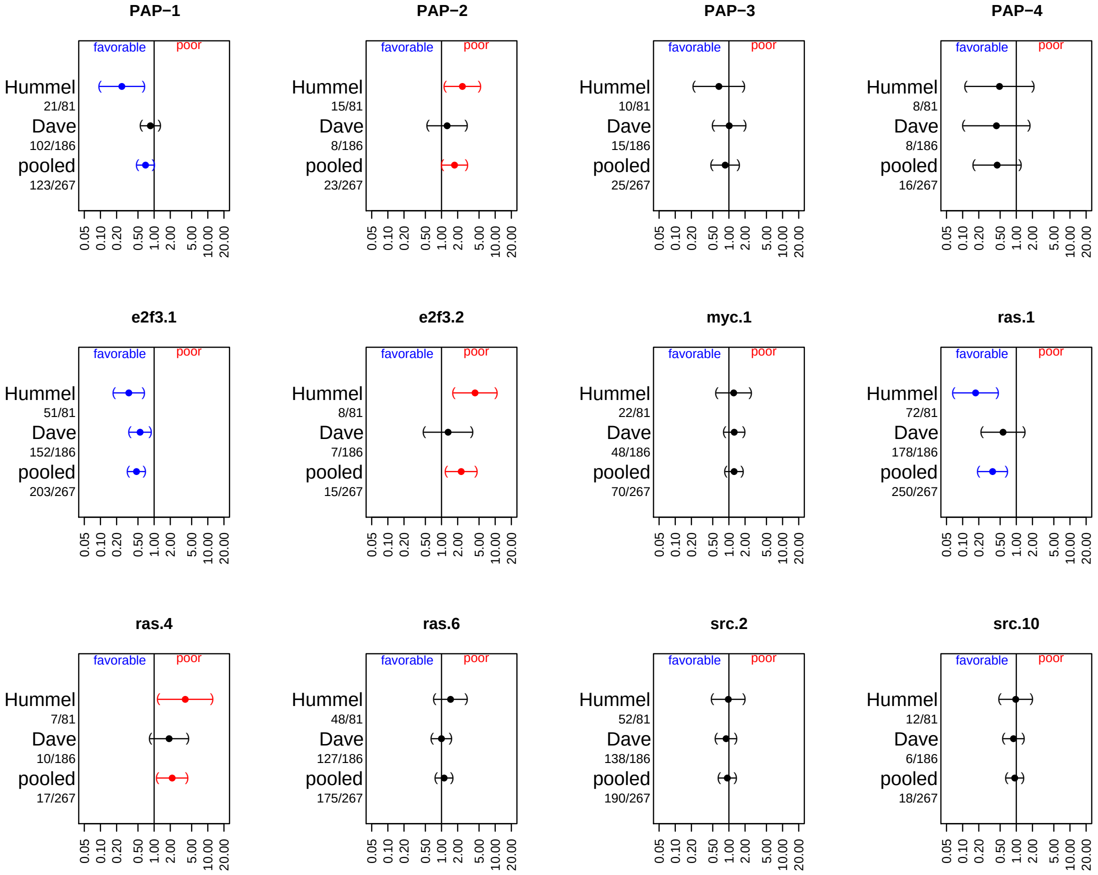

Figure 6.6: Survival analysis of BL-PAP negative lymphomas adjusted for ABC status, age and Ann Arbor stage. Shown are estimated hazard ratios (x-axes) and their 95% confidence intervals associated with PAPs 1-4 (top row) and the underlying eight conserved oncogenic modules (bottom rows). Within each plot rows refer to the two individual studies — our own one published by Hummel et al.[42] and the one published by Dave et al.[20]. The last row shows the result from the pooled analysis. Hazard ratios result from multivariate Cox models including ABC status, age and Ann Arbor stage and, thus, give estimations of risks independent of these known risk factors. For the pooled data we use stratified Cox models assuming separate baseline hazard functions for each study. Significantly favorable hazard ratios are highlighted in blue, significantly poor in red. We consider a hazard as significant, if the 95% confidence interval does not include the 1. The group sizes as well as the total number of samples are given on the left of each plot (n-group/n-total). (Figure reproduced from [11])

Part IV

Appendix

Summary and Conclusions

Today, we can record the information on expression levels of thousands of genes in a single assay. The success of genome sequencing projects has led to the development of the DNA microarray, yielding global gene expression finger prints of different organs, tissues, cells, and tumors. This thesis is concerned with the diagnostic use of microarrays, and their potential contribution to the systematics and taxonomy of lymphomas.

The diagnostic application of microarray data is often discussed in the framework of supervised learning, which is a well defined problem. We are given a set of data points and labels assigning the data points to discrete classes. The goal of supervised learning is to derive a mathematical function, which takes the data points as input and predicts the labels as output. In terms of a molecular diagnosis, a data point is a gene expression profile, and a label is a diagnosis. The theory of supervised learning is based on the assumption that the classification of the data points given by the labels is fixed. In contrast, the classification and taxonomy of lymphoma is not fixed. It is rather the product of ongoing empirical efforts of pathologists to stratify patients into meaningful subtypes with respect to treatment options. These efforts are based on available diagnostic and clinical parameters. The stratification is updated regularly due to the development of novel diagnostic techniques, and increasing knowledge from clinical trials. The diagnostic techniques are mainly based on genetics and molecular biology. In this thesis, we do not consider the microarray as diagnostic device predicting existing disease categories (supervised classification). We consider it as molecular readout that helps us to refine the existing taxonomy of lymphomas. The problem is rather semi-supervised than supervised, which reflects in the different research questions addressed in this thesis.

Defining Burkitt lymphoma. The differential diagnosis between Burkitt lymphoma (BL) and diffuse large-B-cell lymphoma (DLBCL) is not reliably reproducible with the use of the criteria defined by the World Health Organization (WHO). The imprecise distinction between BL and DLBCL on diagnosis may lead to inadequate treatment of lymphoma patients. Stringently applied, the WHO criteria yield a positive diagnosis of Burkitt lymphoma for 8 out of 220 lymphomas investigated in this thesis. Nevertheless, expert pathologists expect more Burkitt lymphomas among the 220 tumor samples, and the goal was to identify them. We denoted the problem as core group extension problem: Given microarray data and starting from a small core group of cases, the objective of core group extension is to find a set of signature genes, which distinguishes the core group from the majority of other cases. At the same time, additional cases should be identified that have expression levels coherent with the core group (across the a priori unknown set of signature genes). In chapters 3 and 4 we implement core group extension in two different semi-supervised approaches refining the definition of Burkitt lymphoma based on microarray data. In chapter 3 we approach core group extension from the supervised perspective, modifying the supervised nearest shrunken centroids method into a semi-supervised version. In chapter 4 we approach core group extension from the unsupervised perspective and modify a model based clustering approach into a semi-supervised version. Both methods yield very similar results on the same data. They identified 36 (chapter 3) or 34 (chapter 4) cases of BL in addition to the core group.

The germinal center and activated B-cell like signature. In chapter 5 we explore the presence of the germinal center-like (GCB-like) and the activated B-cell-like (ABC-like) transcriptional phenotypes of lymphoma in the 220 tumor samples investigated in this thesis. The GCB/ABC stratification is the result of one of the first published diagnostic applications of microarrays [3] and has been obtained by hierarchical clustering. We could confirm the presence of the subtypes in our data set, as well as the better overall survival of patients with GCB-like lymphomas as compared to ABC-like lymphomas.

Pathway based stratification of diffuse large B-cell lymphoma. Traditional molecular biological inference is based on experimental perturbations of biological systems. The problem of studying human cancer is that we cannot perturb this system in its natural environment. A clinical microarray study provides purely observational data and does not provide insight into disease mechanisms. However, since modern targeted drugs directly intervene in molecular processes causing disease, we need to study and predict disease mechanisms in individual tumors. In chapter 6 we discuss a joint analysis of traditional perturbation experiments and observational clinical microarray data, to stratify patients with respect to potential disease mechanisms. Also in this chapter we follow ideas of semi-supervised learning. Combining microarray data from over-expression of five human oncogenes in cell lines with our lymphoma data, yields a novel stratification of lymphomas into groups with distinct biological characteristics, genetic aberrations and prognosis.

Bibliography

[1] Proceedings and abstracts of the 10th international conference on malignant lymphoma, 4-7 june 2008, lugano, switzerland. *Ann Oncol*, 19 Suppl 4:iv31–i296, Jun 2008.

[2] A. Alizadeh, M. Eisen, R. Davis, C. Ma, H. Sabet, T. Tran, J. Powell, L. Yang, G. Marti, D. Moore, J. Hudson, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, I. Lossos, R. Levy, D. Botstein, P. Brown, and L. Staudt. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb Symp Quant Biol*, 64:71–8, 1999.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.

[4] C. Ambroise and G. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10):6562–6, May 2002.

[5] W. Y. Au, D. E. Horsman, R. D. Gascoyne, D. S. Viswanatha, R. J. Klasa, and J. M. Connors. The spectrum of lymphoma with 8q24 aberrations: a clinical, pathological and cytogenetic study of 87 consecutive cases. *Leuk Lymphoma*, 45(3):519–528, Mar 2004.

[6] W. Y. Au, D. E. Horsman, D. S. Viswanatha, J. M. Connors, R. J. Klasa, and R. D. Gascoyne. 8q24 translocations in blastic transformation of mantle cell lymphoma. *Haematologica*, 85(11):1225–1227, Nov 2000.

[7] A. H. Banham, J. M. Connors, P. J. Brown, J. L. Cordell, G. Ott, G. Sreenivasan, P. Farinha, D. E. Horsman, and R. D. Gascoyne. Expression of the foxp1 transcription factor is strongly associated with inferior survival in patients with diffuse large b-cell lymphoma. *Clin Cancer Res*, 11(3):1065–1072, Feb 2005.

[8] S. L. Barrans, J. A. L. Fenton, A. Banham, R. G. Owen, and A. S. Jack. Strong expression of foxp1 identifies a distinct subset of diffuse large b-cell lymphoma (dlbcl) patients with poor outcome. *Blood*, 104(9):2933–2935, Nov 2004.

[9] S. Bea, A. Zettl, G. Wright, I. Salaverria, P. Jehn, V. Moreno, C. Burek, G. Ott, X. Puig, L. Yang, A. Lopez-Guillermo, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. D. Gascoyne, J. M. Connors, T. M. Grogan, R. Braziel, R. I. Fisher, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, R. Simon, J. Powell, W. H. Wilson, E. S. Jaffe, E. Montserrat, H.-K. Muller-Hermelink, L. M. Staudt, E. Campo, A. Rosenwald, and L. M. P. Project. Diffuse large b-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood*, 106(9):3183–3190, Nov 2005.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[11] S. Bentink, S. Wessendorf, C. Schwaenen, M. Rosolowski, W. Klapper, A. Rosenwald, G. Ott, A. H. Banham, H. Berger, A. C. Feller, M.-L. Hansmann, D. Hasenclever, M. Hummel, D. Lenze, P. Mller, B. Stuerzenhofecker, M. Loeffler, L. Truemper, H. Stein, R. Siebert, R. Spang, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. Pathway activation patterns in diffuse large b-cell lymphomas. *Leukemia*, 22(9):1746–1754, Sep 2008.

[12] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, Jan 2006.

[13] K. A. Blum, G. Lozanski, and J. C. Byrd. Adult burkitt leukemia and lymphoma. *Blood*, 104(10):3009–3020, Nov 2004.

[14] L. M. Boxer and C. V. Dang. Translocations involving c-myc and c-myc function. *Oncogene*, 20(40):5595–5610, Sep 2001.

[15] D. BURKITT. A sarcoma involving the jaws in african children. *Br J Surg*, 46(197):218–223, Nov 1958.

[16] M. S. Cairo, M. Gerrard, R. Sposto, A. Auperin, C. R. Pinkerton, J. Michon, C. Weston, S. L. Perkins, M. Raphael, K. McCarthy, C. Patte, and F. A. B. L. I. S. Committee. Results of a randomized international study of high-risk central nervous system b non-hodgkin lymphoma and b acute lymphoblastic leukemia in children and adolescents. *Blood*, 109(7):2736–2743, Apr 2007.

[17] B. Coiffier. State-of-the-art therapeutics: diffuse large b-cell lymphoma. *J Clin Oncol*, 23(26):6387–6393, Sep 2005.

[18] D. R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah

Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

[19] F. Crick. On protein synthesis. In *The biological replication of macromolecules. Symposia of the Society for Experimental Biology*, number 12, pages 138–163, 1958.

[20] S. Dave, K. Fu, G. Wright, L. Lam, P. Kluin, E. Boerma, T. Greiner, D. Weisenburger, A. Rosenwald, G. Ott, H. Müller-Hermelink, R. Gascoyne, J. Delabie, L. Rimsza, R. Braziel, T. Grogan, E. Campo, E. Jaffe, B. Dave, W. Sanger, M. Bast, J. Vose, J. Armitage, J. Connors, E. Smeland, S. Kvaloy, H. Holte, R. Fisher, T. Miller, E. Montserrat, W. Wilson, M. Bahl, H. Zhao, L. Yang, J. Powell, R. Simon, W. Chan, L. Staudt, and . Molecular diagnosis of Burkitt's lymphoma. *N Engl J Med*, 354(23):2431–42, Jun 2006.

[21] R. E. Davis, K. D. Brown, U. Siebenlist, and L. M. Staudt. Constitutive nuclear factor kappab activity is required for survival of activated b cell-like diffuse large b cell lymphoma cells. *J Exp Med*, 194(12):1861–1874, Dec 2001.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[23] J. Dierlamm, E. M. M. Penas, S. Bentink, S. Wessendorf, H. Berger, M. Hummel, W. Klapper, D. Lenze, A. Rosenwald, E. Haralambieva, G. Ott, S. B. Cogliatti, P. Mller, C. Schwaenen, H. Stein, M. Lffler, R. Spang, L. Trümper, R. Siebert, and D. K. N. P. M. M. in Malignant Lymphomas". Gain of chromosome region 18q21 including the malt1 gene is associated with the activated b-cell-like gene expression subtype and increased bcl2 gene dosage and protein expression in diffuse large b-cell lymphoma. *Haematologica*, 93(5):688–696, May 2008.

[24] M. Divin, P. Casassus, S. Koscielny, J. Bosq, C. Sebban, C. L. Maignan, A. Stamattoulas, B. Dupriez, M. Raphal, J.-L. Pico, V. Ribrag, G. E. L. A., and G. O. E. L. A. M. S. Burkitt lymphoma in adults: a prospective study of 72 patients treated with an adapted pediatric lmb protocol. *Ann Oncol*, 16(12):1928–1935, Dec 2005.

[25] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Amer Stat Assoc*, 97:77–87, 2002.

[26] C. H. Dunphy, H. W. van Deventer, K. J. Carder, K. W. Rao, and G. A. Dent. Mature b-cell acute lymphoblastic leukemia with associated translocations (14;18)(q32;q21) and (8;9)(q24;p13). a burkitt variant? *Arch Pathol Lab Med*, 127(5):610–613, May 2003.

[27] J. A. L. Fenton, E. Schuuring, S. L. Barrans, A. H. Banham, S. J. Rollinson, G. J. Morgan, A. S. Jack, J. H. J. M. van Krieken, and P. M. Kluin. t(3;14)(p14;q32) results in aberrant expression of foxp1 in a case of diffuse large b-cell lymphoma. *Genes Chromosomes Cancer*, 45(2):164–168, Feb 2006.

[28] F. Feuerhake, J. L. Kutok, S. Monti, W. Chen, A. S. LaCasce, G. Cattoretti, P. Kurtin, G. S. Pinkus, L. de Leval, N. L. Harris, K. J. Savage, D. Neuberg, T. M. Habermann, R. Dalla-Favera, T. R. Golub, J. C. Aster, and M. A. Shipp. Nfkappab activity, function, and target-gene signatures in primary mediastinal large b-cell lymphoma and diffuse large b-cell lymphoma subtypes. *Blood*, 106(4):1392–1399, Aug 2005.

[29] H. Fiegler, P. Carr, E. Douglas, D. Burford, S. Hunt, C. Scott, J. Smith, D. Vetrie, P. Gorman, I. Tomlinson, and N. Carter. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, 36(4):361–74, Apr 2003.

[30] S. G. Fisher and R. I. Fisher. The epidemiology of non-hodgkin's lymphoma. *Oncogene*, 23(38):6524–6534, Aug 2004.

[31] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[32] K. Gatter and R. Warnke. *Pathology and genetics of tumours of haematopoietic and lymphoid tissues*, volume 3 of *World Health Organization classification of tumours*, chapter Diffuse large B-cell lymphoma, pages 171–174. Lyon, France: IARC Press, 2001.

[33] C. E. Gauwerky, K. Huebner, M. Isobe, P. C. Nowell, and C. M. Croce. Activation of myc in a masked t(8;17) translocation results in an aggressive b-cell leukemia. *Proc Natl Acad Sci U S A*, 86(22):8867–8871, Nov 1989.

[34] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 1999.

[35] N. L. Harris and S. J. Horning. Burkitt's lymphoma–the message from microarrays. *N Engl J Med*, 354(23):2495–2498, Jun 2006.

[36] N. L. Harris, H. Stein, S. E. Coupland, M. Hummel, R. D. Favera, L. Pasqualucci, and W. C. Chan. New approaches to lymphoma diagnosis. *Hematology Am Soc Hematol Educ Program*, pages 194–220, 2001.

[37] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. pamr: Pam: prediction analysis for microarrays. R package version 1.40.0.

[38] J. L. Hecht and J. C. Aster. Molecular biology of burkitt's lymphoma. *J Clin Oncol*, 18(21):3707–3721, Nov 2000.

[39] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene-expression profiles in hereditary breast cancer. *N Engl J Med*, 344(8):539–48, Feb 2001.

[40] C. Hoffmann, E. Wolf, C. Wyen, G. Fätkenheuer, J. V. Lunzen, H.-J. Stellbrink, A. Stoehr, A. Plettenberg, H. Jaeger, R. Noppeney, M. Hentrich, N. Goekbuget, D. Hoelzer, and H.-A. Horst. Aids-associated burkitt or burkitt-like lymphoma: short intensive polychemotherapy is feasible and effective. *Leuk Lymphoma*, 47(9):1872–1880, Sep 2006.

[41] J. Z. Huang, W. G. Sanger, T. C. Greiner, L. M. Staudt, D. D. Weisenburger, D. L. Pickering, J. C. Lynch, J. O. Armitage, R. A. Warnke, A. A. Alizadeh, I. S. Lossos, R. Levy, and W. C. Chan. The t(14;18) defines a unique subset of diffuse large b-cell lymphoma with a germinal center b-cell gene expression profile. *Blood*, 99(7):2285–2290, Apr 2002.

[42] M. Hummel, S. Bentink, H. Berger, W. Klapper, S. Wessendorf, T. F. E. Barth, H.-W. Bernd, S. B. Cogliatti, J. Dierlamm, A. C. Feller, M.-L. Hansmann, E. Haralambieva, L. Harder, D. Hasenclever, M. Kühn, D. Lenze, P. Lichter, J. I. Martin-Subero, P. Möller, H.-K. Müller-Hermelink, G. Ott, R. M. Parwaresch, C. Pott, A. Rosenwald, M. Rosolowski, C. Schwaenen, B. Stürzenhofecker, M. Szczepanowski, H. Trautmann, H.-H. Wacker, R. Spang, M. Loeffler, L. Trümper, H. Stein, R. Siebert, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. A biologic definition of burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med*, 354(23):2419–2430, Jun 2006.

[43] M. Hummel, R. Meister, and U. Mansmann. Globalancova: exploration and assessment of gene group effects. *Bioinformatics*, 24(1):78–85, Jan 2008.

[44] J. Iqbal, V. T. Neppalli, G. Wright, B. J. Dave, D. E. Horsman, A. Rosenwald, J. Lynch, C. P. Hans, D. D. Weisenburger, T. C. Greiner, R. D. Gascoyne, E. Campo, G. Ott, H. K. Müller-Hermelink, J. Delabie, E. S. Jaffe, T. M. Grogan, J. M. Connors, J. M. Vose, J. O. Armitage, L. M. Staudt, and W. C. Chan. Bcl2 expression is a prognostic marker for the activated b-cell-like type of diffuse large b-cell lymphoma. *J Clin Oncol*, 24(6):961–968, Feb 2006.

[45] J. Iqbal, W. Sanger, D. Horsman, A. Rosenwald, D. Pickering, B. Dave, S. Dave, L. Xiao, K. Cao, Q. Zhu, S. Sherman, C. Hans, D. Weisenburger, T. Greiner, R. Gascoyne, G. Ott, H. Müller-Hermelink, J. Delabie, R. Braziel, E. Jaffe, E. Campo, J. Lynch, J. Connors, J. Vose, J. Armitage, T. Grogan, L. Staudt, and W. Chan. BCL2 translocation defines a unique tumor subset within the germinal center B-cell-like diffuse large B-cell lymphoma. *Am J Pathol*, 165(1):159–66, Jul 2004.

[46] W. Klapper, M. Szczepanowski, B. Burkhardt, H. Berger, M. Rosolowski, S. Bentink, C. Schwaenen, S. Wessendorf, R. Spang, P. Möller, M. L. Hansmann, H.-W. Bernd, G. Ott, M. Hummel, H. Stein, M. Loeffler, L. Trümper, M. Zimmermann, A. Reiter, R. Siebert, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. Molecular profiling of pediatric mature b-cell lymphoma treated in population-based prospective clinical trials. *Blood*, 112(4):1374–1381, Aug 2008.

[47] S. Knezevich, O. Ludkovski, C. Salski, V. Lestou, M. Chhanabhai, W. Lam, R. Klasa, J. M. Connors, M. J. S. Dyer, R. D. Gascoyne, and D. E. Horsman. Concurrent translocation of bcl2 and myc with a single immunoglobulin locus in high-grade b-cell lymphomas. *Leukemia*, 19(4):659–663, Apr 2005.

[48] M. H. Kramer, J. Hermans, E. Wijburg, K. Philippo, E. Geelen, J. H. van Krieken, D. de Jong, E. Maartense, E. Schuuring, and P. M. Kluin. Clinical relevance of bcl2, bcl6, and myc rearrangements in diffuse large b-cell lymphoma. *Blood*, 92(9):3152–3162, Nov 1998.

[49] M. Kreuz, M. Rosolowski, H. Berger, C. Schwaenen, S. Wessendorf, M. Loeffler, and D. Hasenclever. Development and implementation of an analysis tool for array-based comparative genomic hybridization. *Methods Inf Med*, 46(5):608–613, 2007.

[50] R. Küppers. Mechanisms of b-cell lymphoma pathogenesis. *Nat Rev Cancer*, 5(4):251–262, Apr 2005.

[51] L. T. Lam, R. E. Davis, J. Pierce, M. Hepperle, Y. Xu, M. Hottelet, Y. Nong, D. Wen, J. Adams, L. Dang, and L. M. Staudt. Small molecule inhibitors of ikappab kinase are selectively toxic for subgroups of diffuse large b-cell lymphoma defined by gene expression profiling. *Clin Cancer Res*, 11(1):28–40, Jan 2005.

[52] J. Lee, J. Lee, M. Park, and S. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48:869–85, 2005.

[53] G. Lenz, G. W. Wright, N. C. T. Emre, H. Kohlhammer, S. S. Dave, R. E. Davis, S. Carty, L. T. Lam, A. L. Shaffer, W. Xiao, J. Powell, A. Rosenwald, G. Ott, H. K. Muller-Hermelink, R. D. Gascoyne, J. M. Connors, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, L. M. Rimsza, R. I. Fisher, D. D. Weisenburger, W. C. Chan, and L. M. Staudt. Molecular subtypes of diffuse large b-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci U S A*, 105(36):13520–13525, Sep 2008.

[54] E. G. Levine, D. C. Arthur, J. Machnicki, G. Frizzera, D. Hurd, B. Peterson, K. J. Gajl-Peczalska, and C. D. Bloomfield. Four new recurring translocations in non-hodgkin lymphoma. *Blood*, 74(5):1796–1800, Oct 1989.

[55] C. Lottaz, D. Kostka, and R. Spang. *Bioinformatics - From Genomes to Therapies*, volume 2, chapter Chapter 26, Classification of Patients, pages 957–991. Wiley-VCH, Weinheim, 2007.

[56] I. Magrath, E. Jaffe, and K. Bhatia. *Neoplastic hematopathology*, chapter Burkitt's lymphoma, pages 953–986. Philadelphia: Lippincott Williams & Wilkins, 2001.

[57] U. Mansmann and R. Meister. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med*, 44(3):449–53, 2005.

[58] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, Mar 1966.

[59] J. Martin-Subero, S. Gesk, L. Harder, W. Grote, and R. Siebert. Interphase cytogenetics of hematological neoplasms under the perspective of the novel WHO classification. *Anticancer Res*, 23(2A):1139–48, 2003.

[60] J. Martin-Subero, L. Harder, S. Gesk, B. Schlegelberger, W. Grote, J. Martinez-Climent, M. Dyer, F. Novo, M. Calasanz, and R. Siebert. Interphase FISH assays for the detection of translocations with breakpoints in immunoglobulin light chain loci. *Int J Cancer*, 98(3):470–4, Mar 2002.

[61] G. J. Mclachlan and T. Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, October 2007.

[62] S. Monti, K. J. Savage, J. L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, B. Wu, L. Pasqualucci, D. Neuberg, R. C. T. Aguiar, P. D. Cin, C. Ladd, G. S. Pinkus, G. Salles, N. L. Harris, R. Dalla-Favera, T. M. Habermann, J. C. Aster, T. R. Golub, and M. A. Shipp. Molecular profiling of diffuse large b-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5):1851–1861, Mar 2005.

[63] L. M. Morton, S. S. Wang, S. S. Devesa, P. Hartge, D. D. Weisenburger, and M. S. Linet. Lymphoma incidence patterns by who subtype in the united states, 1992-2001. *Blood*, 107(1):265–276, Jan 2006.

[64] T. I. Mughal, J. M. Goldman, and S. T. Mughal. *Understanding Leukemias, Lymphomas and Myelomas.* Taylor & Francis, an imprint of the Taylor & Francis Group, 2006.

[65] E. Nacheva, M. J. Dyer, P. Fischer, G. Stranks, J. M. Heward, R. E. Marcus, C. Grace, and A. Karpas. C-myc translocations in de novo b-cell lineage acute leukemias with t(14;18)(cell lines karpas 231 and 353). *Blood*, 82(1):231–240, Jul 1993.

[66] C. Patte, A. Auperin, M. Gerrard, J. Michon, R. Pinkerton, R. Sposto, C. Weston, M. Raphael, S. L. Perkins, K. McCarthy, M. S. Cairo, and F. A. B. M. B. I. S. Committee. Results of the randomized international fab/lmb96 trial for intermediate risk b-cell non-hodgkin lymphoma in children and adolescents: it is possible to reduce treatment for the early responding patients. *Blood*, 109(7):2773–2780, Apr 2007.

[67] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, 135(2):185–207, 1972.

[68] M. Pfreundschuh, L. Truemper, D. Gill, A. Osterborg, R. Pettengell, M. Trneny, K. Imrie, J. Walewski, P.-L. Zinzani, and M. Loeffler. First analysis of the completed mabthera international (mint) trial in young patients with low-risk diffuse large b-cell lymphoma (dlbcl): Addition of rituximab to a chop-like regimen significantly improves outcome of all patients with the identification of a very favorable subgroup with ipi=o and no bulky disease. In *Abstracts of the 46th Annual Meeting of the American Society of Hematology, New Orleans, December 1-7, 2004*, page 157.

[69] M. Pfreundschuh, L. Trümper, A. Osterborg, R. Pettengell, M. Trneny, K. Imrie, D. Ma, D. Gill, J. Walewski, P.-L. Zinzani, R. Stahel, S. Kvaloy, O. Shpilberg, U. Jaeger, M. Hansen, T. Lehtinen, A. Lpez-Guillermo, C. Corrado, A. Scheliga, N. Milpied, M. Mendila, M. Rashford, E. Kuhnt, M. Loeffler, and M. I. T. Group. Chop-like chemotherapy plus rituximab versus chop-like chemotherapy alone in young patients with good-prognosis diffuse large-b-cell lymphoma: a randomised controlled trial by the mabthera international trial (mint) group. *Lancet Oncol*, 7(5):379–391, May 2006.

[70] H. A. Poirel, M. S. Cairo, N. A. Heerema, J. Swansbury, A. Auprin, E. Launay, W. G. Sanger, P. Talley, S. L. Perkins, M. Raphal, K. McCarthy, R. Sposto, M. Gerrard, A. Bernheim, C. Patte, and F. A. B. M. B. . I. S. Committee. Specific cytogenetic abnormalities are associated with a significantly inferior outcome in children and adolescents with mature b-cell non-hodgkin's lymphoma: results of the fab/lmb 96 international study. *Leukemia*, 23(2):323–331, Feb 2009.

[71] K. S. Pollard, Y. Ge, S. Taylor, and S. Dudoit. *multtest: Resampling-based multiple hypothesis testing.* R package version 1.23.3.

[72] R Development Core Team. R: A language and environment for statistical computing. 2008. ISBN 3-900051-07-0.

[73] M. D. Radmacher, L. M. McShane, and R. Simon. A paradigm for class prediction using gene expression profiles. *J Comput Biol*, 9(3):505–511, 2002.

[74] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M.

Hurt, H. Zhao, L. Averett, L. Yang, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. Lpez-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, L. M. Staudt, and L. M. P. Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N Engl J Med*, 346(25):1937–1947, Jun 2002.

[75] J. Ruland, G. S. Duncan, A. Wakeham, and T. W. Mak. Differential requirement for malt1 in t and b cell antigen receptor signaling. *Immunity*, 19(5):749–758, Nov 2003.

[76] J. T. Sandlund, J. R. Downing, and W. M. Crist. Non-hodgkin's lymphoma in childhood. *N Engl J Med*, 334(19):1238–1248, May 1996.

[77] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

[78] C. Schwaenen, M. Nessling, S. Wessendorf, T. Salvi, G. Wrobel, B. Radlwimmer, H. Kestler, C. Haslinger, S. Stilgenbauer, H. Döhner, M. Bentz, and P. Lichter. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proc Natl Acad Sci U S A*, 101(4):1039–44, Jan 2004.

[79] M. A. Shipp. Prognostic factors in aggressive non-hodgkin's lymphoma: who has "high-risk" disease? *Blood*, 83(5):1165–1173, Mar 1994.

[80] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[81] R. Simon. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J Natl Cancer Inst*, 97(12):866–7, Jun 2005.

[82] R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Springer, 2004.

[83] S. Smeland, A. K. Blystad, S. O. Kvaly, I. M. Ikonomou, J. Delabie, G. Kvalheim, J. Hammerstrm, G. F. Lauritzsen, and H. Holte. Treatment of burkitt's/burkitt-like lymphoma in adolescents and adults: a 20-year experience from the norwegian radium hospital with the use of three successive regimens. *Ann Oncol*, 15(7):1072–1078, Jul 2004.

[84] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3, 2004.

[85] L. M. Staudt and S. Dave. The biology of human lymphoid malignancies revealed by gene expression profiling. *Adv Immunol*, 87:163–208, 2005.

[86] B. Streubel, U. Vinatzer, A. Lamprecht, M. Raderer, and A. Chott. T(3;14)(p14.1;q32) involving igh and foxp1 is a novel recurrent chromosomal aberration in malt lymphoma. *Leukemia*, 19(4):652–658, Apr 2005.

[87] S. Swerdlow, E. Campo, N. Harris, S. Pileri, H. Stein, J. Thiele, and J. Vardiman, editors. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues.* IARC Press, 2008.

[88] The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-hodgkin's lymphoma. *N Engl J Med*, 329(14):987–994, Sep 1993.

[89] The Non-Hodgkin's Lymphoma Classification Project. A clinical evaluation of the international lymphoma study group classification of non-hodgkin's lymphoma. *Blood*, 89(11):3909–3918, Jun 1997.

[90] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, May 2002.

[91] J. W. Tukey. Tightening the clinical trial. *Control Clin Trials*, 14(4):266–285, Aug 1993.

[92] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, Apr 2001.

[93] L. van 't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, Jan 2002.

[94] R. Ventura, J. Martin-Subero, M. Jones, J. McParland, S. Gesk, D. Mason, and R. Siebert. FISH Analysis for the Detection of Lymphoma-Associated Chromosomal Abnormalities in Routine Paraffin-Embedded Tissue. *J Mol Diagn*, 8(2):141–51, May 2006.

[95] A. von Heydebreck, W. Huber, A. Poustka, and M. Vingron. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17(Suppl 1):S107–14, 2001.

[96] L. Wessels, M. Reinders, A. Hart, C. Veenman, H. Dai, Y. He, and L. Veer. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–62, Oct 2005.

[97] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98(20):11462–7, Sep 2001.

[98] I. Wlodarska, E. Veyt, P. D. Paepe, P. Vandenberghe, P. Nooijen, I. Theate, L. Michaux, X. Sagaert, P. Marynen, A. Hagemeijer, and C. D. Wolf-Peeters. Foxp1, a gene highly expressed in a subset of diffuse large b-cell lymphoma, is recurrently targeted by genomic aberrations. *Leukemia*, 19(8):1299–1305, Aug 2005.

[99] W. Woessmann, K. Seidemann, G. Mann, M. Zimmermann, B. Burkhardt, I. Oschlies, W.-D. Ludwig, T. Klingebiel, N. Graf, B. Gruhn, H. Juergens, F. Niggli, R. Parwaresch, H. Gadner, H. Riehm, M. Schrappe, A. Reiter, and B. F. M. Group. The impact of the methotrexate administration schedule and dose in the treatment of children and adolescents with b-cell neoplasms: a report of the bfm group study nhl-bfm95. *Blood*, 105(3):948–958, Feb 2005.

[100] D. H. Wright. What is burkitt's lymphoma and when is it endemic? *Blood*, 93(2):758, Jan 1999.

[101] G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proc Natl Acad Sci U S A*, 100(17):9991–6, Aug 2003.

[102] E. Yeoh, M. Ross, S. Shurtleff, W. Williams, D. Patel, R. Mahfouz, F. Behm, S. Raimondi, M. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. Pui, W. Evans, C. Naeve, L. Wong, and J. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, Mar 2002.

[103] J. T. Yustein and C. V. Dang. Biology and treatment of burkitt's lymphoma. *Curr Opin Hematol*, 14(4):375–381, Jul 2007.

Zusammenfassung

Lymphome sind die fünfthäufigste Krebserkrankung in westlichen Staaten (Europa und Nordamerika). In dieser Arbeit geht es um die molekulare Charakterisierung des diffus großzelligen B-Zell Lymphoms (DLBCL) und des Burkitt Lymphoms (BL) mit Hilfe von Transkriptionsprofilen und überwachten und halbüberwachten maschinellen Lernverfahren. Zwei wesentliche Probleme der Lymphomklassifikation werden mit Hilfe von Transkriptionsprofilen untersucht.

Diagnostik des Burkitt Lymphoms. Die diagnostische Unterscheidung von BL und DLBCL ist oft nicht präzise. Das heißt, verschiedene Pathologen kommen hier oft zu verschiedenen Ergebnissen. Eine zuverlässige Unterscheidung der beiden Lymphomtypen ist unerlässlich für die Auswahl der Therapie.

Funktionale Stratifikation. Traditionelle molekularbiologische Untersuchungen beruhen darauf, dass man experimentell gezielt in biologische Prozesse eingreift (z.B. durch Mutagenese oder Überexperession), um diese besser verstehen zu können. Das Problem bei der Untersuchung von Krebs im Menschen ist, dass man den individuellen Tumor in seiner natürlichen Umgebung nicht experimentell untersuchen kann. Eine klinische Microarraystudie liefert lediglich Beobachtungsdaten.

Beiträge dieser Arbeit sind:

- Die Einführung des halbüberwachten Lernproblems der Kerngruppenerweiterung. Dabei werden ausgehend von einer sicher diagnostizierten Kerngruppe von Tumoren weitere Fälle gesucht, die die gleichen Eigenschaften haben, von denen man aber die Diagnose nicht kennt.
- Die Entwicklung eines *Expectation-Maximization* (EM) basierten Algorithmus zur zur Kerngruppenerweiterung.
- Die Generierung einer linearen Signatur zur quantitativen und reproduzierbaren diagnostischen Unterscheidung von BL und DLBCL mit Hilfe der Kerngruppenerweiterung.
- Die Entwicklung einer halbüberwachten Lernmethode, die es erlaubt Tumore in klinischen Genexpressionsstudien aufgrund der Daten aus hypothesengetriebenen Interventionsexperimenten in Zelllinien zu stratifizieren.
- Die Generierung einer neuen funktionalen Stratifikation von DLBCL.

Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

Publications

- **S. Bentink**, S. Wessendorf, C. Schwaenen, M. Rosolowski, W. Klapper, A. Rosenwald, G. Ott, A. H. Banham, H. Berger, A. C. Feller, M.-L. Hansmann, D. Hasenclever, M. Hummel, D. Lenze, P. Möller, B. Stuerzenhofecker, M. Loeffler, L. Truemper, H. Stein, R. Siebert, R. Spang, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. **Pathway activation patterns in diffuse large B-cell lymphomas.** *Leukemia*, 22(9):1746–1754, Sep 2008.

- M. Hummel, **S. Bentink**, H. Berger, W. Klapper, S. Wessendorf, T. F. E. Barth, H.-W. Bernd, S. B. Cogliatti, J. Dierlamm, A. C. Feller, M.-L. Hansmann, E. Haralambieva, L. Harder, D. Hasenclever, M. Kühn, D. Lenze, P. Lichter, J. I. Martin-Subero, P. Möller, H.-K. Müller-Hermelink, G. Ott, R. M. Parwaresch, C. Pott, A. Rosenwald, M. Rosolowski, C. Schwaenen, B. Stürzenhofecker, M. Szczepanowski, H. Trautmann, H.-H. Wacker, R. Spang, M. Loeffler, L. Trümper, H. Stein, R. Siebert, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. **A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling.** *N Engl J Med*, 354(23):2419–2430, Jun 2006.

- J. Dierlamm, E. M. M. Penas, **S. Bentink**, S. Wessendorf, H. Berger, M. Hummel, W. Klapper, D. Lenze, A. Rosenwald, E. Haralambieva, G. Ott, S. B. Cogliatti, P. Möller, C. Schwaenen, H. Stein, M. Löffler, R. Spang, L. Trümper, R. Siebert, and D. K. N. P. M. M. in Malignant Lymphomas”. **Gain of chromosome region 18q21 including the MALT1 gene is associated with the activated B-cell-like gene expression subtype and increased BCL2 gene dosage and protein expression in diffuse large B-cell lymphoma.** *Haematologica*, 93(5):688–696, May 2008.

- A. Ehlers, E. Oker, **S. Bentink**, D. Lenze, H. Stein, and M. Hummel. **Histone acetylation and DNA demethylation of B cells result in a hodgkin-like phenotype.** *Leukemia*, 22(4):835–841, Apr 2008.

- B. Hirsch, M. Hummel, **S. Bentink**, F. Fouladi, R. Spang, R. Zollinger, H. Stein, and H. Dürkop. C**D30-induced signaling is absent in Hodgkin's cells but present in anaplastic large cell lymphoma cells.** *Am J Pathol*, 172(2):510–520, Feb 2008.

- J. I. Martn-Subero, M. Kreuz, M. Bibikova, **S. Bentink**, O. Ammerpohl, E. Wickham-Garcia, M. Rosolowski, J. Richter, L. Lopez-Serra, E. Ballestar, H. Berger, X. Agirre, H.-W. Bernd, V. Calvanese, S. B. Cogliatti, H. G. Drexler, J.-B. Fan, M. F. Fraga, M. L. Hansmann, M. Hummel, W. Klapper, B. Korn, R. Küppers, R. A. F. Macleod, P. Möller, G. Ott, C. Pott, F. Prosper, A. Rosenwald, C. Schwaenen, D. Schübeler, M. Seifert, B. Stürzenhofecker, M. Weber, S. Wessendorf, M. Loeffler, L. Trümper, H. Stein, R. Spang, M. Esteller, D. Barker, D. Hasenclever, R. Siebert, and M. M.

in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. **New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling.** *Blood*, 113(11):2488–2497, Mar 2009.

- W. Klapper, M. Szczepanowski, B. Burkhardt, H. Berger, M. Rosolowski, **S. Bentink**, C. Schwaenen, S. Wessendorf, R. Spang, P. Möller, M. L. Hansmann, H.-W. Bernd, G. Ott, M. Hummel, H. Stein, M. Loeffler, L. Trümper, M. Zimmermann, A. Reiter, R. Siebert, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. **Molecular profiling of pediatric mature B-cell lymphoma treated in population-based prospective clinical trials.** *Blood*, 112(4):1374–1381, Aug 2008.

- M. Montesinos-Rongen, A. Brunn, **S. Bentink**, K. Basso, W. K. Lim, W. Klapper, C. Schaller, G. Reifenberger, J. Rubenstein, O. D. Wiestler, R. Spang, R. Dalla-Favera, R. Siebert, and M. Deckert. **Gene expression profiling suggests primary central nervous system lymphomas to be derived from a late germinal center B cell.** *Leukemia*, 22(2):400–405, Feb 2008.

- C. Schwaenen, A. Viardot, H. Berger, T. F. E. Barth, **S. Bentink**, H. Döhner, M. Enz, A. C. Feller, M.-L. Hansmann, M. Hummel, H. A. Kestler, W. Klapper, M. Kreuz, D. Lenze, M. Loeffler, P. Möller, H.-K. Müller-Hermelink, G. Ott, M. Rosolowski, A. Rosenwald, S. Ruf, R. Siebert, R. Spang, H. Stein, L. Truemper, P. Lichter, M. Bentz, S. Wessendorf, and M. M. in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. **Microarray-based genomic profiling reveals novel genomic aberrations in follicular lymphoma which associate with patient survival and gene expression status.** *Genes Chromosomes Cancer*, 48(1):39–54, Jan 2009.

Members of the MMML project

Recent and present members of the research network ”Molecular Mechanisms in Malignant Lymphoma” supported by the Deutsche Krebshilfe are:

- *Pathology group*: Wolfram Klapper, Monika Szcepanowski, Thomas Barth, Wolfram Bernd, Alfred Feller, Martin-Leo Hansmann, Peter Möller, German Ott, Hans-Konrad Müller-Hermelink, Andreas Rosenwald, Hans-Heinrich Wacker, Sergio Cogliatti, Michael Hummel, Harald Stein.
- *Genetics group*: Carsten Schwaenen, Swen Wessendorf, Heiko Trautmann, Jose-Ignacio Martin-Subero, Eugenia Haralambieva, Judith Dierlamm, German Ott, Andreas Rosenwald, Thomas Barth, Christiane Pott, Ralf Küppers, Reiner Siebert.
- *Bioinformatics group*: Maciej Rosolowski, Rainer Spang, Hilmar Berger, Stefan Bentink, Dirk Hasenclever, Markus Löffler.
- *Project coordination*: Benjamin Stürzenhofecker, Hilmar Berger, Michael Hummel, Lorenz Trümper.
- *Steering Committee*: Reiner Siebert, Harald Stein, Markus Löffler, Lorenz Trümper.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, June 2009

Stefan Bentink