Bioinformatical analysis of B-cell lymphomas

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Steffen Blenk
aus Bamberg
Würzburg 2007

Für meine Frau Isabella.

Eingereicht am: .....

Mitglieder der Promotionskommission:
Vorsitzender: Prof. Dr. Martin J. Müller
Gutachter: Prof. Dr. Thomas Dandekar
Gutachter: Prof. Dr. Erich Buchner

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

Contents

Chapter 1

Summary

Zusammenfassung

Hintergrund: Die Häufigkeit von Non-Hodgkin-Lymphomen (NHL), den am meisten beobachteten Krebserkrankungen, steigt weiter an. Von den aggressiven Non-Hodgkin-Lymphomen (NHL) macht das "großzellige, diffuse B-Zell-Lymphom" (DLBCL) den größten Anteil aus. Durch Genexpressionsmuster wurden zwei Subtypen definiert: ACB ("Activated B-like DLBCL") und GCB ("Germinal Center B-like DLBCL"). Die Patienten der Gruppe ABC sterben ohne Therapie oft innerhalb weniger Monate, weil der ABC Typ einen aggressiveren Krankheitsverlauf aufweist. Ein weiteres, von einer malignen Entartung der B-Lymphozyten ausgehendes Lymphom, ist das "Mantelzell Lymphom" (MCL). Es tritt selten auf und ist ebenfalls mit einer schlechten Prognose verbunden. Eine vollständige Heilung nach der Therapie ist sehr selten.

Methoden: In diesem Projekt wurden diese B-zell Lymphome mit bioinformatischen Methoden untersucht, um auf molekularer Ebene neue Eigenschaften oder bisher unentdeckte Zusammenhänge zu finden. Das würde das Verständnis und damit auch die Therapie voranbringen. Dafür standen uns Überlebens-, Genexpressions- und chromosomale Aberrationsdaten zur Verfügung. Sie sind die bevorzugte Wahl der Mittel, um genetische Veränderungen in Tumorzellen zu bestimmen. Hierbei fallen oft große Datenmengen an, aus welchen man mit bioinformatischen Methoden vorher unerkannte Trends und Hinweise identifizieren kann.

Ergebnisse (MCL): Explorative Analysen sowohl der Genexpressions- (zweite Hauptachse der Korrespondenz Analyse) als auch der chromosomalen Aberrationsdaten des Mantelzell-Lymphom zeigten uns hierbei, daß es trotz der linearen Korrelation zwischen der veröffentlichten Proliferationssignatur und der Überlebenszeit sinnvoll ist, in den Patienten (n=71) zwei Ausprägungen zu betrachten: Patienten mit schlechter und mit guter Prognose. Statistische Tests (moderate t-test, Wilcoxon rank-sum test) dieser beiden Typen zeigten Unterschiede im Zellzyklus und ein Netzwerk von Kinasen auf, welche für den Unterschied zwischen guter und schlechter Prognose verantwortlich sind. Sieben Gene (CENPE, CDC20, HPRT1, CDC2, BIRC5, ASPM, IGF2BP3) konnten gefunden werden, die eine ähnliche gute Prognose für Überlebenszeiten ermöglichen, wie eine früher veröffentlichte Proliferationssignatur mit 20 Genen. Außerdem konnten chromosomale Banden durch eine explorative Analyse mit der Prognose assoziiert werden (Chromosom 9: 9p24, 9p23, 9p22, 9p21, 9q33 and 9q34).

Ergebnisse (DLBCL): Durch geeignete Normalisierung der Genexpressionsdaten von 248 DLBCL-Patienten trennte der Signatur basierte Predictor die Risikogruppen nun besser auf. Eine ähnlich gute Auftrennung konnte von uns sogar mit sechs Genen erreicht werden. Die explorative Analyse der Genexpressionsdaten (S. Blenk, J. Engelmann, M. Weniger, J. Schultz, M. Dittrich, A. Rosenwald, H. K. Müller-Hermelink, T. Müller and T. Dandekar; Cancer Informatics, *in press*) konnte die Subtypen ABC und GCB als valide Gruppen bestätigen. In den Genen, die ABC und GCB unterscheiden, ergab sich eine Häufung in späten und frühen Zellzyklusstadien. Klassische Lymphommarker, neu aufgefundene spezielle Gene und Zellzyklusgene bilden ein Netzwerk, das die ABC und GCB Gruppen klassifizieren und Unterschiede in deren Regulation erklären kann (ASB13, BCL2, BCL6, BCL7A, CCND2, COL3A1, CTGF, FN1, FOXP1, IGHM, IRF4, LMO2, LRMP, MAPK10, MME, MYBL1, NEIL1 and SH3BP5; unterstrichene Gene sind überexprimiert in ABC). Dies ist auch für die Diagnose, Prognose und Therapie (Zytostatika) interessant.

Summary

Background: The frequency of the most observed cancer, Non Hodgkin Lymphoma (NHL), is further rising. Diffuse large B-cell lymphoma (DLBCL) is the most common of the NHLs. There are two subgroups of DLBCL with different gene expression patterns: ABC ("Activated B-like DLBCL") and GCB ("Germinal Center B-like DLBCL"). Without therapy the patients often die within a few months, the ABC type exhibits the more aggressive

behaviour. A further B-cell lymphoma is the Mantle cell lymphoma (MCL). It is rare and shows very poor prognosis. There is no cure yet.

Methods: In this project these B-cell lymphomas were examined with methods from bioinformatics, to find new characteristics or undiscovered events on the molecular level. This would improve understanding and therapy of lymphomas. For this purpose we used survival, gene expression and comparative genomic hybridization (CGH) data. In some clinical studies, you get large data sets, from which one can reveal yet unknown trends.

Results (MCL): The published proliferation signature correlates directly with survival. Exploratory analyses of gene expression and CGH data of MCL samples (n=71) revealed a valid grouping according to the median of the proliferation signature values. The second axis of correspondence analysis distinguishes between good and bad prognosis. Statistical testing (moderate t-test, Wilcoxon rank-sum test) showed differences in the cell cycle and delivered a network of kinases, which are responsible for the difference between good and bad prognosis. A set of seven genes (CENPE, CDC20, HPRT1, CDC2, BIRC5, ASPM, IGF2BP3) predicted, similarly well, survival patterns as proliferation signature with 20 genes. Furthermore, some bands could be associated with prognosis in the explorative analysis (chromosome 9: 9p24, 9p23, 9p22, 9p21, 9q33 and 9q34).

Results (DLBCL): New normalization of gene expression data of DLBCL patients revealed better separation of risk groups by the 2002 published signature based predictor. We could achieve, similarly well, a separation

with six genes. Exploratory analysis of gene expression data could confirm the subgroups ABC and GCB. We recognized a clear difference in early and late cell cycle stages of cell cycle genes, which can separate ABC and GCB. Classical lymphoma and best separating genes form a network, which can classify and explain the ABC and GCB groups. Together with gene sets which identify ABC and GCB we get a network, which can classify and explain the ABC and GCB groups (ASB13, BCL2, BCL6, BCL7A, CCND2, COL3A1, CTGF, FN1, FOXP1, IGHM, IRF4, LMO2, LRMP, MAPK10, MME, MYBL1, NEIL1 and SH3BP5; underlined genes are more highly expressed in ABC; S. Blenk, J. Engelmann, M. Weniger, J. Schultz, M. Dittrich, A. Rosenwald, H. K. Müller-Hermelink, T. Müller and T. Dandekar; Cancer Informatics, *in press*).

Altogether these findings are useful for diagnosis, prognosis and therapy (cytostatic drugs).

Chapter 2

Introduction

Lymphomas are cancers originating in the lymphatic system. They arise if a lymphocyte starts to proliferate in an uncontrolled way, crowding out healthy cells and creating tumors. These lymphocytes may spread from one site to other parts of the body. Lymphomas are divided into two major groups, the Hodgkin Lymphoma (HL) and all other lymphomas fall into the category of Non Hodgkin Lymphomas (NHL). The incidence of Non Hodgkin lymphoma cases has almost doubled over the last 20 years (Fisher and Fisher, 2004). The reasons are uncertain, although there are some known correlations between NHL and infection with HIV or Epstein-Barr virus. In this thesis the "Diffuse Large B-cell Lymphoma", most common, and the "Mantle Cell Lymphoma" one of the rarest of the NHLs, are investigated. Their causes are uncertain too, but there is some knowledge about their biology. Large scale array data can shed light on typical disease markers and genes. We investigate gene expression arrays from over 240 diffuse large B-cell lymphoma patients, find key genes for prognosis and distinguishing between different subtypes of this

disease. Furthermore, from a set of gene expression and CGH data from 71 cyclin D1-positive mantle cell lymphoma patients we identify key genes and indication for long and short surviving patients.

2.1 Immune System

As the thesis mainly analyzes the gene expression in B-cell cancers some facts about the genetics of the immune system need to be introduced. Together with the T-lymphocytes the B-lymphocytes form the two main types of lymphocytes, subtypes of leucocytes (white blood cells). They are both essential components of the adaptive immune system, giving cellular and humoral immune responses.

B-lymphocytes or B-cells are produced in the bone marrow and are very important for the humoral immune response. Their main function is to produce antibodies against soluble antigens. In their development each stage represents a change in the genome content at the antibody loci. The antibodies are composed of two light (L) and two heavy (H) chains. Many variations of these chains exist due to somatic recombination and mutation and also through gene recombination of and within the H and the L chain loci. The main antibody factories are the plasma B-cells. Memory B-cells develop from activated B-cells and are specific to the antigen encountered during the primary immune response.

T-lymphocytes are responsible for cell mediated immunity. They are developed in the thymus, which is the reason for the abbreviation "T", but they arise in the bone marrow. Several different subtypes of T cells exist,

each with a distinct function, for example the cytotoxic T cells and helper T cells. The cytotoxic T cells (Tc cells) are responsible for cell mediated immune response, by destroying infected cells. Their T-cell receptor (TCR) allows them to monitor all cells of the body with the help of the Major Histocompatibility Complex (MHC) I proteins, which are expressed and presented on the surface by nearly all human cells. MHC II proteins are anchored only on antigen presenting cells (APCs), like B-cells, and are recognised by helper T cells (Th cells). The MHC T cell interactions are mediated by the glycoprotein co-receptors CD8 for MHCI and CD4 for MHCII molecules, whereas CD8 ("cluster of differentiation 8") is expressed on the surface of Tc cells and CD4 ("cluster of differentiation 4") on Th helper cells. The most known MHC genes are HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA and HLA-DRB1 (HLA: Human Leukocyte Antigen). The first three genes express MHC I and the last six the MHC II proteins, both encoded by chromosome 6p.

Another important part of humoral immune response are the so called germinal centers (GC) in the lymphoid tissue. After B-cells are activated by antigens they migrate to the GCs, where the memory B-cells are born by undergoing isotype switching and somatic mutation resulting in a more accurate antigen binding. The B-cells there proliferate rapidly.

2.2 Cell cycle

Cancer cells are uncontrolled proliferating cells, the cell cycle is continuously kept active, therefore the cell cycle is subject in this chapter. The cell cycle

takes place between cell divisions and triggers the following division. There are four distinct phases in the common cell cycle in eukaryotes: G1, S, G2, and M. The three first are called interphase, as the cell prepares itself for the division.

In the G1 phase, or "GAP 1" phase, the cell performs its usual metabolic activities. Fully differentiated cells enter from G1 the G0 state, where they can remain indefinitely. If the cell is going to divide, it increases the amount of cytoplasm. After a distinct "point of no return", the restriction point, the cell has to go on into the S, the "synthesis" phase, where its DNA is duplicated. After the S phase, the "GAP 2" (G2) phase is used by the cell for preparing the mitosis by growth and producing new proteins. In the M phase, the cell segregates the duplicated chromosomes, so both daughter cells are diploid again with one chromosome from the mother and one from the father. In detail the M phase consists of the mitosis, separating the chromosomes between the two daughter cells and the cytokinesis, which divides the cytoplasm. The mitosis passes through the stages prophase, metaphase, anaphase, and telophase, and each of these is defined by several processes and morphological properties. After the telophase a fiber ring composed of actin around the center of the cell contracts and pinches the cell into two daughter cells, the so called cytokinesis. Then the cell cycle is complete and the daughter cells are in G1 again. The duration of the G2 and the M phase is relatively short.

The most important known molecule classes for the regulation of the cell cycle are the cyclins and the cyclin-dependent kinases (CDKs). CDKs perform phosphorylation that activates or inactivates target proteins to co-

ordinate the entry into the next cell cycle phase. Cyclins themselves are characterized by periodicity in protein abundance throughout the cell cycle. Bound to CDKs, they function as regulatory subunits, as CDKs are inactive in their absence. Different cyclins exhibit distinct expression and degradation patterns which contribute to the temporal coordination of the cell cycle. Once activated the cyclin CDK complexes prepare the cell for entry in the next phase by promoting the gene expression of transcription factors or the degradation of inhibitors.

2.3 Classifications of lymphomas

For effective and successful treatments a more detailed classification of human cancer has been shown to be clinically useful. The most common classification for lymphomas is the distinction between Hodgkin lymphomas, with about 15%, and Non-Hodgkin Lymphomas (NHL), with around 85% of human lymphomas. Dr. Thomas Hodgkin defined the Hodgkin lymphoma 1832. These two subgroups are heterogeneous. Further classifications are defined on the basis of morphologic and molecular parameters, described in the "**R**evised **E**uropean-**A**merican **L**ymphoma" (REAL) classification by the "**W**orld **H**ealth **O**rganization" (WHO). Unfortunately it is very likely, that some morphologic subtypes consist of more than one disease, as reflected by WHO's classification, which includes several morphologic and immunophenotypic variants. Such is the case for "Diffuse Large B-cell Lymphoma" (DLBCL), the most common type of Non Hodgkin lymphoma accounting for about 40%. "Mantle Cell Lymphoma" (MCL) is also a member of the NHL

and accounts for about 5%-6% of lymphomas.

2.3.1 Diffuse Large B Cell Lymphoma

Diffuse Large B-cell Lymphoma (DLBCL) is an aggressive cancer of the mature B-lymphocytes and the most frequent B-cell NHL. It usually occurs between adolescence and old age, but the frequency is clearly higher in people over 60 years of age, and it is more common in men than women. Around 40% of these lymphoma patients are cured with therapy, but the causes for MCL are unknown yet.

Diagnosis relies on morphological, immunophenotypic and laboratory parameters. In clinical situations, the International Prognostic Index (IPI) (1993) is often used to predict outcome. The IPI estimates the threat facing patients with risk factors: age; tumor stage; serum lactate dehydrogenase concentration; performance status and the number of extranodal disease sites. The detailed calculation is not relevant here, but each risk factor increases the value with its level. For example, the age counts as "high level" if the patient is over 60 years and an elevated concentration of serum lactate dehydrogenase (LDH) too. LDH serves here as an indicator for cells with a high rate of turnover which is characteristic for cancer cells. The combination of factors correlates with the risk groups, whereas patients in the high risk group have clearly less than a 50% chance of 5-year survival. On the molecular level, gene expression signatures have been defined that predict outcome in DLBCL independent of the IPI (Rosenwald et al., 2002), hoping it will be more useful than the IPI.

The most frequent treatment is chemotherapy, which applies anti-cancer (cytotoxic) drugs to destroy the cancer cells. This cures a large number of patients, and even if a cure is not possible, this treatment can control the disease for a number of years. Other therapies are inclusive of: high-dose chemotherapy in combination with bone marrow or stem cell infusions, radiotherapy, steroid therapy and monoclonal antibody therapy.

An advance in classification for the DLBCL was reported in 2000 as two new subtypes were proposed on the basis of gene expression profiling: the "Activated B-like DLBCL" (ABC) and the "Germinal Center B-like DLBCL" (GCB) subtype. The first one is more aggressive and overall the survival rate is much lower than the other. As the names indicate, the ABC DLBCL cells show a gene expression pattern similar to in vitro activated B-cells and the GCB DLBCL cells similar to the germinal center B-cells.

ABC and GCB gene expression patterns differ in thousands of genes. One of them is the "B-cell CLL/lymphoma 6" (BCL6), which plays an important role in development of B-cells and of DLBCL cancer. BCL6 is a zinc finger transcription factor, which acts as a sequence-specific repressor of transcription. It is essential for differentiation of mature B-cells into germinal center B-cells during an immune response (Ye et al., 1997; Dent et al., 1997). GCB patients show a higher expression of BCL6, than ABC patients (Alizadeh et al., 2000; Rosenwald et al., 2002; Wright et al., 2003). The target genes, regulated by BCL6, were identified (Shaffer et al., 2000) and they are not surprisingly lower expressed in GCB than in ABC. Some of them are known to be induced when B-cells are activated through the antigen receptor. Known as "cyclin D2" (CCND2), "CD69 molecule" (CD69), "CD44

molecule" (CD44), and "chemokine (C-C motif) ligand 3" (CCL3, also referred as MIP-1-alpha or SCYA3). "Cyclin-dependent kinase inhibitor 1B" (p27, Kip1 or CDKN1B) is also a target of BCL6 and a negative regulator of cell cycle progression. As BCL6 blocks that inhibitor, it could explain the proliferation rate of germinal center B-cells.

It should be mentioned here that BCL6 is deregulated by chromosomal translocations in about 20% of DLBCLs (Pasqualucci et al., 2003). Nevertheless, these translocations cannot explain all BCL6 deregulations. Although GCB cases show a high gene expression of BCL6, they do not correlate to these translocations. A further property of GCB gene expression which differs from ABC cases is the somatic hypermutation of immunoglobulin genes, as they derive from germinal center B-cells (Lossos et al., 2000).

An oncogenic event in DLBCL is the translocation t(14;18), which leads to a higher expression of "B-cell CLL/lymphoma 2" (BCL2), because then BCL2 lies closely to the enhancers of the immunoglobulin heavy chain locus. This translocation occurs in 45% of the GCB cases, and was not found in ABC cases. However, most of the ABC patients have a high gene expression value of BCL2 mRNA (Rosenwald et al., 2002; Wright et al., 2003).

The amplification of the c-rel (REL) locus on chromosome 2, was also not found in ABC patients (Rosenwald et al., 2002; Bea et al., 2005) occurring in about 16% of the GCB cases. REL, or "v-rel reticuloendotheliosis viral oncogene homolog (avian)" encodes a member of the NF-κB family. NF-κBs are anti apoptotic transcription factors. The NF-κB pathway is highly expressed in ABC but not in GCB patients. Complementary to this ABC patients show a high expression of NF-κB targets (Davis et al., 2001). So the

NF-κB pathway could be a promising therapeutic target for ABC patients.

As the known subgroups of DLBCL differ in their biology and their gene expression, a molecular diagnosis method was developed, which estimates the probability of being ABC or GCB. Applied to gene expression values of 58 DLBCL patients measured with Affymetrix chips, it classified the subgroups successfully, with some remaining unidentified (Wright et al., 2003).

2.3.2 Mantle Cell lymphoma

Mantle cell lymphoma (Swerdlow and Williams, 2002) usually infiltrates the mantle zone of the lymph nodes, the filtering components of the lymphatic system. Although patients receive a variety of chemotherapies, there is no cure nor survival prolonging treatment as yet. MCL is associated with a poor prognosis, and it remains incurable with the current chemotherapeutic approaches. It usually occurs between the late 30s to old age, but the highest frequency of MCL is in the people over 50 years of age and it is three times more common in men than women. Despite response rates of 50-70% with many regimens, the disease typically progresses after chemotherapy. The median survival time is approximately 3 years (range, 2-5 y); the 10-year survival rate is only 5-10%. Although it is also a cancer of the B-lymphocytes, its biology is completely different from DLBCL. It arises from malignant B lymphocytes in the mantle zone, a part of the lymph nodes, which surrounds the follicle centers of the lymph nodes. These cancer lymphocytes start growing eliminating mantle zone and changing the size of the lymph node. Furthermore, they penetrate rapidly into other lymph nodes and organs.

The most known and important biological event is the translocation between chromosome 11 and 14 (t(11;14)(q13;q32)) (Bogner et al., 2006) in about 50% of all MCL cases. As a result the "cyclin D1" (CCND1) gene is brought under the control of the IgH enhancer leading to an overexpression. As shown (Rosenwald et al., 2003), some cases show a higher CCND1 expression, with an associated shorter survival. Different observed expressed forms of CCND1 in MCL account for that observation (Rimokh et al., 1994; Lebwohl et al., 1994). Although the translocation can be observed in about half of the MCL cases, the CCND1 overexpression is a more constant observation. Staining cells for increased levels of cyclin D1 provides an excellent marker for specific diagnosis. CCND1 forms heterodimers with the kinases CDK4 or CDK6 and functions as their regulatory subunit. These CDKs trigger the G1/S phase transition of the cell cycle (Sherr and McCormick, 2002). So this translocation disturbs the regulation of the cell cycle.

Another important event is the deletion of the INK4a/ARF locus in about every fifth patient (Pinyol et al., 1997, 1998, 2000; Rosenwald et al., 2003). It encodes the tumor suppressor proteins, p16INK4a and p14ARF (Sherr and McCormick, 2002), which both can arrest the cell cycle. The former one by inhibiting the interaction of CDK4 and CDK6 with D cyclins, as CCND1, and the latter one by blocking the degradation of "tumor protein p53" (p53), a well known protein, playing important roles in apoptosis and cell cycle, especially the transition from G0 to G1.

2003 Rosenwald et al defined a proliferation signature based predictor, in which the expression values of 20 genes estimate the length of survival. This signature was created by fitting gene expression values to survival data in

a supervised manner. By identifying patient subsets that differed by more than five years in median survival, the authors showed the advantage of this predictor compared with 2.7 years of other methods (Velders et al., 1996; Argatoff et al., 1997; Bosch et al., 1998; Räty et al., 2002).

Furthermore, they showed that the differences in CCND1 abundance synergized with INK4a/ARF locus deletions to affect proliferation rate and survival. Another advantage of this signature is its property to measure the proliferation rate indepently of events as INK4a/ARF deletion or CCND1 overexpression, which both are independently occuring, and integrating those effects, which affect the cell cycle process. These results led to a proposed model of abnormal cell cycle regulation in MCL and point to a cell cycle inhibiting therapy.

2.4 Gene expression microarrays

A microarray is a spatial array of oligonucleotide probes, which are arranged on a small solid support surface. These probes, representing nucleotide sequences in known genes, are located in such a way that the position and the nucleotide sequence of each probe are known. The length of the probes varies between 15 and 25 nucleotides, but up to 60 are possible. The DNA from the sample being studied is isolated, fragmented and tagged with a fluorescent dye. After that the DNA fragments are incubated with the chip. DNA or RNA from the samples, which is complementary to the DNA in the microarray, binds to the probes and the unbound DNA/RNA is washed away. The surface of the microarray is then scanned with a laser beam and

the data obtained is used to produce a visible image. Colour intensity indicates the extent of hybridization of the different probes. In this way RNA can be measured that may be translated into active proteins. The term for this is gene expression analysis. With tens of thousands of probes on each slide, the microarrays accelerated the science in general and mainly in the topic for genetic tests. Even whole genome chips are available. There are several different technologies for producing chips, for example printing with fine-pointed pins on glass chips, photolithographies, even ink jet printing. In data collection and statistical analysis different technologies have also evolved.

Two main types of arrays are mentioned here: single channel and two channel microarrays. The first one is incubated with samples, marked by one dye. The advantage of this chip type is the estimation of the absolute value of gene expression, but the comparison of two groups needs two chips. The most well known are the commercially available Affymetrix chips. The two channel microarrays are usually hybridized with a cDNA mix of two samples, which are marked with two different dyes, for example Cy5 a red and Cy3 a green fluorescence. You can measure the up and down regulated genes of two groups with one chip, however, the disadvantage is in the lack of absolute gene expression levels.

Special arrays for different aims can be created, mostly by using the two channel technology. A prominent example is the "Lymphochip", a microarray especially for lymphomas (Alizadeh et al., 1999).

Of course the limitations of microarrays should be kept in mind. So the measurement of transcription of genes includes not the protein expression.

In every gene expression experiment the transcription is only an indication of the real protein expression. Especially boutique arrays are further limited by lacking the genes of the complete genome. On the other hand boutique arrays enable more samples to be measured, because of the much cheaper source, and deliver a more stable and low-noise result. In regard to these conditions one may ask, if microarrays are an adequate tool at all. But diseases with clonal beginning and gene associated changes are best investigated by microarrays. For example, it has been shown that microarrays can subclassificate leukemia (Kohlmann et al., 2003). Only microarrays could find those clinically relevant acute leukemia subgroups.

2.5 CGH - Comparative genomic hybridization

Another broadly used method for investigating oncogenetic events in cancers is the comparative genomic hybridization (CGH). This method measures copy number changes of chromosome regions, as gains, losses, amplification or even no change. Fluorescently labeled samples of interest and normal DNA hybridize to either normal human metaphase preparations or array-CGH, a slide containing defined DNA probes. With the colour ratio the DNA gain or loss in the sample is measured. As mentioned it measures only copy number changes, so balanced chromosomal changes and the location of the rearranged sequences remain undetected.

2.6 Thesis Project

The thesis project (part of project B-36) is funded, by the "**I**nterdisziplinäres **Z**entrum für **K**linische **F**orschung" (IZKF) Würzburg. The overall aim of the B-36 project is a better understanding of tumor progression and pathogenesis in B-cell lymphoma. Besides gaining knowledge about the biology of the disease itself, a better understanding supports the treatment success. Therefore gene expression- and CGH-data of the collaborating IZKF groups of Prof. Dr. Müller-Hermelink, Dr. Rosenwald and Dr. Kneitz are analysed. Through bioinformatical analyses of this data and pathway modelling, key factors can be identified. The validation is done by experimental work. In this way diagnostic factors and potential pharmacological targets become available. Another task mentioned here is the build up of a database, which stores data, results and delivers the bioinformatical methods for finding key genes and events by a web interface. This database is now available at the URL http://gepat.bioapps.biozentrum.uni-wuerzburg.de/GEPAT/ (Weniger et al., 2007).

In detail my part of the project used statistical analysis for identifying key factors as well as pathway modelling. Therefor the DLBCL and MCL data, generated with the Lymphochip and delivered by Prof. Müller-Hermelink and Dr. Rosenwald, were analysed bioinformatically. The first step, the normalization of gene expression data for samples comparison and analysis, was performed by my colleague Julia Engelmann. The database, which accelerated my work enormously, was established by my colleague Markus Weniger.

The causes for DLBCL and MCL are unknown. Medicine focuses on investigating clinically relevant subgroups for more accurate treatment. We investigate here the differences between the ABC and GCB subgroups and search for potential entities in MCL cases, hoping to gain new knowledge and improvements in their treatment.

For the interactions between pathogen/tumor and host the "Active Analyzer of Interaction Networks" is proposed in chapter 4. It contributes database search, network analysis and enlargement as a small Java based tool. It was exemplary applied to actin polymerization.

The gene expression and CGH-data of the MCL samples are analysed in chapter 5. The published proliferation signature (Rosenwald et al., 2003) of 20 genes enabled patients to be categorised into four risk groups. Here, the aim was to find key events and entities within the MCL cases. Exploratory analysis of gene expression and CGH-data revealed a patients clustering matching the separation by the median of their proliferation signatures. Following this, a classification according to the proliferation signature was done, which was supported and confirmed by exploratory analysis of gene expression and CGH-data. Moreover the regulatory differences and cascades implicated in the group differences are further promoted by a novel application of the non-parametric Wilcoxon rank-sum test on CGH data e.g. specific changes on chromosome 9. As a clinical application a new seven gene predictor is derived from these gene markers distinguishing efficiently long and short living patients.

In chapter 6 the DLBCL analysis is described. The gene expression measurements of the Lymphochip revealed the ABC and GCB subgroups and

their distinguishing genes as well as a signature based survival predictor. We tried to find a new handy survival predictor in these gene expression data and the supported survival data. Simultaneously, the analysis focuses on the ABC and GCB classification by gene expression measurements. Therefor three gene sets were investigated in order to establish how well they are able to distinguish between the subgroups. The associated knowledge and experimental data from the protein interaction database STRING delivered a network for these gene sets.

The discussion is in chapter 7. The general aim of this project was to find key events as genes and pathways to gain new knowledge about the two B-cell lymphomas DLBCL and MCL. The results will help to understand the biology of DLBCL and MCL and to investigate new treatment targets.

Chapter 3

Materials and Methods

We used as hardware a Personal Computer with a 2.00 GHz Intel Pentium Prozessor and 512 MByte RAM. For more time consuming calculations or high-throughput experiments a Transtec Linux Cluster (10 dual Xeon processors each with 1GB RAM) was used.

3.1 Software development

Eclipse The program ACTIN was developed with the open source development platform "Eclipse". It is, besides others, an "Integrated Developement Environment" (IDE) which allows working with different programming languages. There are also plugins for markup languages. We chose the programming language Java, (Java 2 SDK, SE Version 1.4.2_03) as it is available for all common operating systems and computers.

Data formats The "**Ex**tensible **M**arkup **L**anguage" (XML) (W3C, a) enables defining data formats and exchanging a wide variety of data. Data for-

mats are created by defining tags. The "Document Type Definition" (DTD) specifies the syntax of an XML. One example of an XML based data format is the following protein interaction format. A work group of the "Human Proteome Organization" (HUPO), the "Proteomics Standards Initiative" (PSI), defined the protein interaction data format PSI **M**olecular **I**nteraction (PSI MI) (Kaiser, 2002; Hermjakob et al., 2004a). It describes protein interactions and was developed to facilitate data comparison, exchange and allows data integration across experiments. This data format is supported by databases such as "Biomolecular Interaction Network Database" (BIND) (Bader et al., 2003; Gilbert, 2005), Cellzome, "Database of Interacting Proteins" (DIP) (Salwinski et al., 2004), "Human Protein Reference Database" (HPRD) (Peri et al., 2003), "European Bioinformatics Institute's" (EMBL-EBI) (Stoesser et al., 2003), IntAct (Hermjakob et al., 2004b), "Molecular Interactions" (MINT) (Zanzoni et al., 2002; Chatr-aryamontri et al., 2007), and STRING. These databases are not synchronised with each other and their inhouse data formats are not compatible. PSI MI enables adapting and investigating the protein interaction infomation in thes databases. The PSI MI data format consists of some parts, from which ACTIN extracts the "interactorList" und die "interactionList". The former one lists annotations of each protein interactor and the latter one describes the protein interactions. The "e**X**tensible **S**tylesheet **L**anguage" (XSL) (W3C, b), a transformation language, was chosen to convert XML files into HTML (W3C, c) files.

3.2 Data sets

Both data sets contain besides the gene expression measurements censored survival data for the patients.

DLBCL data The DLBCL study includes and re-analyzes raw data from a well documented study (Rosenwald et al., 2002). We have access to all data, in which the subgroups were defined by hierarchical clustering. Our study analyzes a modified data set as follows: more patients (a total of 248 patients, each patient array included 12196 gene spots corresponding to 3717 genes), different classifications and number of cases (12.3

MCL data As the DLBCL data set, the MCL gene expression data (n=71) were obtained from cDNA arrays containing genes preferentially expressed in lymphoid cells or genes known or presumed to be part of cancer development or immune function ("Lymphochip" microarrays (Alizadeh et al., 1999)). Additionally the dataset is completed by comparative genomic hybridization (CGH) data for each patient (n=71). The CGH values are associated with following meanings: 1 = "loss"; 2 = "gain"; 3 = "amplification". The "amplification" is understood here as a multiplier of 4, compared to 0, which denotes "no change". The samples were collected from cyclin D1-positive patients of several hospitals in the "Lymphoma and Leukemia Molecular Profiling Project" (LLMPP) (Rosenwald et al., 2003).

3.3 Explorative data analysis

For the explorative analysis principal component, and correspondence analysis were used.

The Principal Components Analysis (PCA) method used here is implemented in R (R Development Core Team, 2006). The aim is to reduce multidimensional datasets to lower dimensions containing the most important properties of the data for explorative analysis. By linear transformation the dimensions are combined in principal axes, and so the data are positioned into a new coordinate system in such way, that the greatest variance lies on the first principal component, the second greatest variance on the second one, etc. The advantage is in keeping the subspace with the largest variance. The top few linear combinations typically contain the most information in the data and serve as good overview.

The R package "**M**odern **A**pplied **S**tatistics with **S**" (MASS) (Venables and Ripley, 2002) was used here for the correspondence analysis (CA), another explorative analysis. Its concept is similar to PCA, but the data are additionally scaled and the rows and columns are treated equivalently. So it can be that some unimodal distributed measurements can be recognized by CA but not by PCA.

For both methods, PCA and CA, different calculation algorithms exist to get the results.

The R library vegan (Oksanen et al., 2007) was used for the constrained or Canonical Correspondence Analysis (CCA) (Ter Braak, 1986). In this package Legendre & Legendre's algorithm is used (Legendre and Legendre,

1998). A weighted linear regression on constraining variables is applied to the Chi-square transformed data. Here we used the chromosomes as constraining variables. Then a correspondence analysis is performed on the fitted values.

Unbiased class discovery, used the ISIS ("Identifying Splits with clear Separation") method (von Heydebreck et al., 2001). This method searches for binary class distinctions in the gene expression levels in an unsupervised fashion. The "Diagonal Linear Discriminant (DLD) score" quantifies for every found bipartition how strongly the two classes are separated. It is a type of "diagonal linear discriminant analysis" (DLDA). Scoring each vertex in a graph in which they stand for bipartitions and are defined as neighbours if they differ only in one sample, the search focuses on local maxima. As the whole graph has got 2^{n-1} nodes it is not applicable to score all possible bipartitions with the DLD score. So an efficient heuristic finds candidate partitions.

3.4 Statistical analysis

All statistical analyses applied to the DLBCL data set were performed using the statistical software package R (Ihaka and Gentleman, 1996; R Development Core Team, 2006) with its specific libraries. For data normalization, the Bioconductor package, an open source software project for the analysis and comprehension of genomic data (Gentleman et al., 2004), including methods such as vsn, loess and scaling methods, was used. Based on diagnostic plots we chose gene expression normalization using within-array and between-array normalization methods. The within-array normalization "loess" (W.S. Cleve-

land and Shyu, 1992; Yang et al., 2001, 2002) adjusts expression log-ratios in such a way that they average to zero within each array to make genes on one array comparable with each other. For between-array normalization, the "scale" method proposed by Yang et al. (2001, 2002) and further explained by Smyth and Speed (2003) was applied. This method scales log-ratios to have the same median-absolute-deviation (MAD) across arrays. By this, log-ratios are normalized to show similar variance across a batch of arrays. To detect differentially expressed genes the Bioconductor (Gentleman et al., 2004) package "limma" (Smyth, 2005) was used. It is a state of the art gene expression analysis framework. Its special strength is the improved robust statistics based on linear models and a moderated t-test statistics corrected for multiple measurements to detect differentially expressed genes (Smyth, 2005, 2004). It fits linear models on the gene expression values of each gene with respect to the groups which are compared. After that the method applies empirical Bayes shrinkage of the standard errors. Due to its robustness the method can be applied to experiments with a small number of samples. Each spot is now analyzed individually and not just pooled as was done in the previous analysis(Rosenwald et al., 2002). Furthermore, we investigated how robust the data is regarding advanced normalization.

The R package "survival" was used to calculate the Cox regression hazard model (Andersen and Gill, 1982; Therneau et al., 1990), which investigates the influence of gene expression values on survival time. The survival data contain as usual censored information. A sample contains censored data if the event of interest (death) did not happen within the observation time. This lack of information can be handled by the Cox regression hazard model.

It estimates the effect of variables on the time an event takes to happen. The variables and events stand here for gene expression values and death. The model defines the log-hazard as

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

whereas $\alpha(t)$ is the unspecified baseline hazard and x_i the ith covariate. In contrast to the variables the baseline hazard function $\alpha(t)$, remains unspecified. As the covariates, here the gene expression values, enter the model linearly the only condition is the constancy of variable's effects on survival. The package used here applies the Andersen and Gill (Andersen and Gill, 1982) counting process formulation of the Cox proportional hazards regression model. Using the coefficients with their according gene expression values, the outcome predictor score for each patient can be calculated. The coefficients relate directly to hazard so that a positive value indicates a bad outcome prognosis and a negative one a positive effect of the variable. We used here the Wald test to determine the significance of the association between the model and the outcome. Furthermore, Kaplan Meier estimates were also derived. They can show the survival difference of conditions in the course of time by plotting the according curves on the abscissa against the probability of survival ordinate. The plot shows decreasing horizontal steps, the decreasing survival probability over time, and the vertical marks indicate censored information. In our case, the single lines represent different risk groups.

"Prediction Analysis of Microarrays" (PAM) was used for supervised

analyses. PAM (Tibshirani et al., 2002) performs a nearest shrunken centroid method to identify a subset of genes that best characterize samples as ABC or GCB DLBCL. It computes a standardized centroid for each class and shrinks the prototypes for a given classification error threshold. In the resulting list the obtained optimal (for the given error) shrunken centroid identifier is followed by the number of genes it contains. The chosen classifier is validated by ten-fold cross-validation. Smaller gene sets typically show larger error rates. However, often many almost equally good performing classifiers exist, showing very similar error rates. In this case, following the parsimony idea, we opted for the one containing the smallest number of genes. The proposed best gene set used for our analysis (31 spots) is labeled in the plot by an "x" character.

To identify all protein-protein network interactions and their analysis we used the "Search Tool for the Retrieval of Interacting Genes/Proteins" (STRING)-version 6.3 (http://string.embl.de/) (von Mering et al., 2005). STRING is a database of known and predicted protein-protein interactions. The interaction information arises from genomic context, experiments, other databases, coexpression and textmining. Here we used it with a Bayesian confidence level of 0.400 (medium confidence) and a custom limit of 0 (this means, that only direct interactions of the considered genes and their proteins are shown).

Gene sequences were delivered by the Ensembl database (Hubbard et al., 2007) (http://www.ensembl.org/). It provides, besides other organisms, the genome for homo sapiens and very useful search features, creating a user friendly database search.

Most of the statistical analyses on the MCL data set were performed using the "**G**enome **E**xpression **P**athway **A**nalysis **T**ool" (GEPAT), a web-based platform for annotation, analysis and visualization of microarray gene expression data (Weniger et al., 2007). The analysis and visualization methods in the context of genomic, proteomic and metabolic scope are integrated in an easy to use, interactive graphical user interface. The database performs the analyses with Bioconductor (Gentleman et al., 2004) and some of its packages.

For identification of differentially expressed genes, it uses also the "limma" package (Smyth, 2005) which offers moderate t-statistics. For multiple testing correction it offers three methods from which we chose the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). For identifying all protein-protein network interactions GEPAT uses also the STRING database (von Mering et al., 2005) version 6.3. The STRING database comprises known and predicted protein-protein interactions.

3.5 CGH analysis

The Wilcoxon rank-sum test (Wilcoxon, 1945; Bauer, 1972; Hollander and Wolfe, 1999), a non-parametric statistical significance test, was applied to the CGH data. It tests here each of the chosen bands against the null hypothesis H_0 in which there is no statistically significant difference between our proposed two MCL subgroups in specific chromosomes. The compared groups, therefore, consist of the values -1, 0, 1 and 2. The method tests if the ranked values are equally distributed – the null hypothesis – or more

group like distributed – the two sided alternative H_1. An exact p-value gives a rejection probability of the null hypothesis.

The R package "survival" was used to calculate the Cox regression hazard models. It examines the correlation between the given measurements and the survival data. For the exploratory analysis of the CGH-data as well as for the new predictor of MCL overall survival, we used the Wald test to determine the significance of the association between the model and the outcome.

Results

Chapter 4

Interlude:

A network analyzer - ACTIN

Protein network interactions are a focus of tumor-host-interaction. A tool developed for this is the "Active Analyzer of Interaction Networks" (ACTIN). The aim was network extension which was exemplary applied on the actin polymerization. In order to model the actin polymerization in silicio, an appropriate program was developed, which simulates any cascade depending on user input. It provides depth-first simulation for the investigated network topology, which can be given in by an ASCII file or by hand. The user interface allows to modify individual nodes and simulates the resulting network with every single step. It provides furthermore network extension by searching protein interaction partners in PSI MI (See chapter 3) supporting databases. So the existing network can be extended and simulated again. ACTIN is a pure Java application and delivers a clear model also in complex situations.

The program was presented at the "German Conference on Bioinformatics 2004" in Bielefeld. On the following two pages the poster abstract is shown.

Simulating actin polymerization cascades with ACTIN

Steffen Blenk and Thomas Dandekar

Dept. of bioinformatics, biocenter, D-97074 university of Würzburg

Introduction

All multi-cellular organisms use signalling cascades for transducing information. Because many pathogenes use specific signalling cascades it is important to understand the processing of pathogen transducing signals, the actual time course and key steps of processing events (activation, inhibition). In our study we considered the actin polymeryzation signalling cascade which is modified upon invasion of pathogens such as Lysteria monocytogenes. *L. monocytogenes* is quite hardy and resists the deleterious effects of freezing, drying, and heat remarkably well. Most *L. monocytogenes* strains are pathogenic to some degree. The aim at searching for new starting points for medical therapy by modulation of pathogen input signals or inhibition of an activation signal leads to questions for fundamental research. Before implementing a new anti-pathogen strategy it is necessary to know how the different activatory and inhibitory stimuli interact, in particular given a certain set of stimuli, (i) what will be the outcome? (ii) What happens if a certain node or input is modified?

Many different approaches have been proposed to study regulatory networks, e.g. Petri nets [Heiner et al., 2004][Chen et al., 2003], custom written simulations with differential equations [Lee et al., 2003] or even virtual cells [Tomita, 2001].

The present study shows a very fast and direct modelling approach using HashMaps. The advantages are (i) fast simulation speed and (ii) Virtual unlimited size of simulated network.

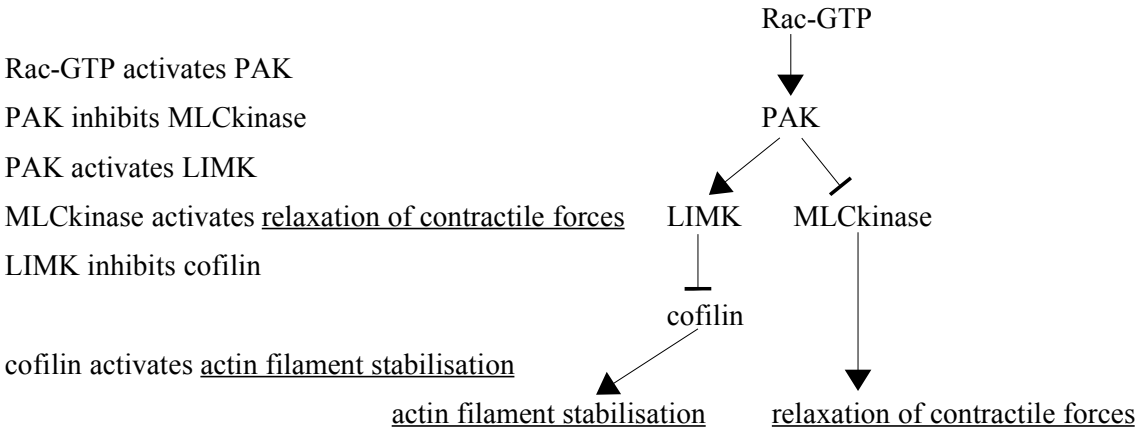
Results and Discussion

The program was implemented in Java 2 SDK, SE Version 1.4.2_03 using an 2.00 GHz Intel Pentium Prozessor with 512 MByte RAM.

User surface and program run: To run a simulation, an ASCII formatted Input File represents the network. The additional protein interactions are loaded from all databases, which provide their information in HUPO`s PSI MI format (DIP, MINT, EBI, …).

The defined topology can be changed and allows the user to introduce and consider new network interactions. A user surface allows further to change the presence of individual components of the network. The output consists of the end-elements, the end effectors of the cascade, which the program identifies by searching a way through the cascade, considering inhibitory and activatory stimuli and processing through the cascade.

Example:



In the example (subgraph of the complete actin polymerization network) we have to find a condition for actin filament stabilisation. In this simple netmap, if the user sets the node LIMK to “not present” then the actin filament stabilisation can be reached provided that the set of the conditions of the input nodes is set to activatory (in the simple example this requires only ras-GTP as active). This example is part of the

regulation of actin-based motility we used for our study. The model for this function is a middle-sized network of 23 elements and 6 end reactions, which are different types of actin polymerization (biological data after [Ahmadian et al., 2002]).

2. Performance:

Run time increases with additional entries only slowly arithmetically, because the Algorithm uses simple indexing by directly hashing. As [Schmidt et al., 2001] showed in their work, in which they used hashes for an artificial intelligence system, the direct indexing provides the advantage of working very fast.

Further advantages of our implementation are that results can be stored in different formats: HTML, XML and MI-XML.

The MI-XML formatted results can be adapted to other databanks.

3. Application to the actin nucleation network:

Database searches provide further interaction partners and indications for new not known interaction partners. Using the algorithm a number of different input and output conditions can be analyzed, in the middle sized network we study 22 nodes which already provide 2^{22} possibilities how input can be processed before the action (actin polymerization) comes about. Our large network of over 108 nodes provides 2^{108} possibilities for processing to the algorithm. The simulation allows us to rapidly test the network behaviour on this wide range of conditions and to see which set of conditions is necessary for a distinct pathway allowing actin nucleation under certain biological pre-conditions.

Another option the network simulator allows is to test network robustness. Some networks need to react to weak stimuli, whereas they must be able to maintain their state when exposed even to very strong stimuli [Bar-Yam et al., 2004]. The robustness is affected by the topology [Ingolia 2004]. In the current model we want to identify pivotal nodes (where removal is not robust for output behaviour) for further therapeutic modification of response.

Even large networks can be simulated with the present implementation with rapid processing and output. The program directly sums up stimuli at each node conveniently using the hash map and can process large quantities of data. However, further effort will include additional important information such as modification and/or different processing times of individual nodes and conditions to better separate and identify the biological important paths leading to the various effector reactions when the cascade is properly switched by the correct biological stimuli.

Availability: The executable is available on request from the authors together with a protocol for use.

References:

[Heiner et al., 2004] M. Heiner and I. Koch and J. Will. *Model validation of biological pathways using Petri nets-demonstrated for apoptosis.* Biosystems. 2004

[Chen et al., 2003] M. Chen and R. Hofestadt. *Quantitative Petri net model of gene regulated metabolic networks in the cell.* In Silico Biol. 2003;

[Tomita, 2001] M. Tomita. *Whole-cell simulation: a grand challenge of the 21st century.* Trends Biotechnol. 2001

[Schmidt et al., 2001] R. Schmidt and L. Gierl. *Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes.* Artif Intell Med. 2001

[Bar-Yam et al., 2004] Y. Bar-Yam and I. R. Epstein. *Response of complex networks to stimuli.* Proc Natl Acad Sci U S A. 2004

[Ingolia 2004]N. T. Ingolia. *Topology and robustness in the Drosophila segment polarity network.* PLoS Biol. 2004

[Ahmadian et al., 2002] M. R. Ahmadian and A. Wittinghofer and G. Schmidt. *The actin filament architecture: tightly regulated by the cells, manipulated by pathogens.* EMBO Rep. 2002

[Lee et al., 2003] E. Lee and A. Salic and R. Kruger and R. Heinrich and M. W. Kirschner. *The Roles of APC and Axin Derived from Experimental and Theoretical Analysis of the Wnt Pathway.* PLoS Biol. 2003

The tool was furthermore applied to test and model different kinase network topologies of the human actin polymerization network used by pathogens.

Chapter 5

Analysis of gene expression: Long and short surviving MCL patients

This second study is an effort to improve molecular insights and markers of the disease for better diagnosis and potential therapeutic strategies. The study thus looked both at CGH and gene expression data to improve diagnosis in this respect as well as new molecular markers in addition to the well known ones such as the characteristic antigens (shared with blood cells from which the tumor may develop) CD5, CD 20 and FMC7.

In our study we used gene expression data from 71 cyclin D1-positive patients and coupled these to data on their corresponding chromosomal aberrations (n=71) and pathway modelling. Different bioinformatical techniques applied involve the new databank-system GEPAT (Weniger et al., 2007) as well as pathway alignment and gene context methods (von Mering et al.,

2005).

Some morphological subtypes of MCL are known (Argatoff et al., 1997; Bosch et al., 1998). The relations between the proliferation of the tumor and the shorter survival of patients were recognized resulting in prognostic markers fitting markers or clinical parameters to the survival time (Argatoff et al., 1997; Bosch et al., 1998; Räty et al., 2002; Velders et al., 1996; Rosenwald et al., 2003). After exploratory analysis of gene expression and CGH-data, we classified patients according to the proliferation signature of Rosenwald et al. (Rosenwald et al., 2003), a gene expression based predictor of survival. Therefor we separated the patients according to the median of the (precalculated) proliferation signature values. This leads to the two subgroups "small" and "big", which are from now on referred to as "s" and "b" and, which are supported and confirmed by. We identify aurora kinases, further disregulated cell cycle genes linked to CDC2 and regulatory differences and cascades implicated in the group differences, which represent an interaction network. A separation according this molecular marker was not done yet and helps to reveal further differences between long time and short time survival in MCL. The analysis of the corresponding CGH-data from chromosomes VII and IX supports the classification and tests single bands. Additionally a seven gene predictor is derived distinguishing long and short living patients.

Moreover, we investigated in an extended diffuse large B-cell lymphoma (DLBCL) data set (chapter 6) the gene expression differences between MCL and DLBCL in the MAPK cascade. Clinical implications from the analysis are discussed.

5.1 Exploratory identification of the Mantle Cell Lymphoma subgroups "survival" (s) and "bad prognosis" (b)

Some parameters were found which are useful for defining subgroups of mantle cell lymphoma and allow a separation according to the survival time. The survival time is the most obvious and biological meaningful parameter in which subgroups should show a significant difference in determining individual clinical treatment.

Gene expression data

After extensive reannotation using GEPAT (Weniger et al., 2007), we selected 3000 genes with the highest variance and applied correspondence analysis (Figure 5.1). We know, that the proliferation signature values represent a continuum. Our analyses show that we can define two clinical useful types. We already found that the second axis is able to, almost perfectly, separate patients according to the median of the proliferation signature values in a multidimensional data space, confirming the stability and reliability of the two subgroups in the data. In other words, the second axis distinguishes between the poorer or better prognosis.

So the proliferation signature (Rosenwald et al., 2003) was re-examined by exploratory data analysis not only by the genes of the proliferation signature but also by a huge amount of genes. We ranked the 71 MCL patients according to their proliferation signature values and separated them accord-

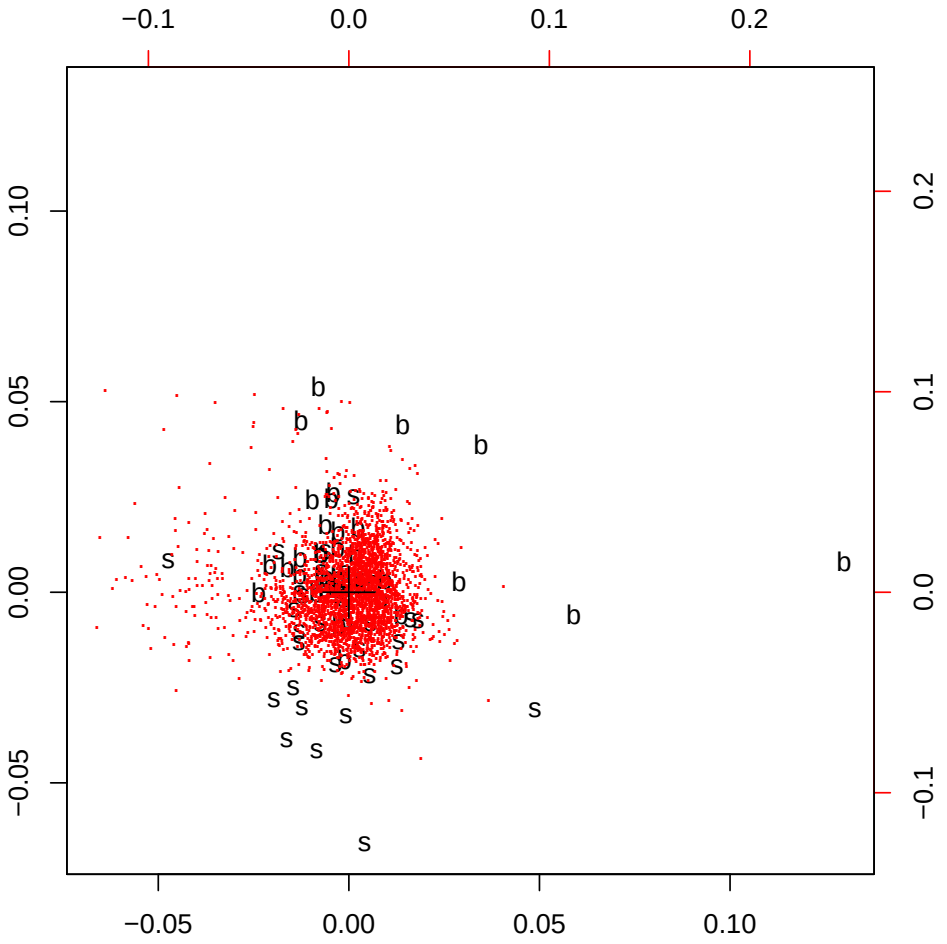


Figure 5.1: ***Correspondence analysis identifies the two Mantle cell lymphoma subgroups.*** **The gene expression data are projected on the first two principal axes. The patients can be clearly separated by this exploratory analysis considering the 3.000 genes (red dots) of the highest variance. In the correspondence plot this is indicated by the horizontal separation line. The patients are labelled with "s" and "b" which represent the separation by the median of the proliferation signature into two different entities. Patients with a proliferation signature value smaller than the median are marked with "s" (survival) and the other patients with "b" (bad prognosis).**

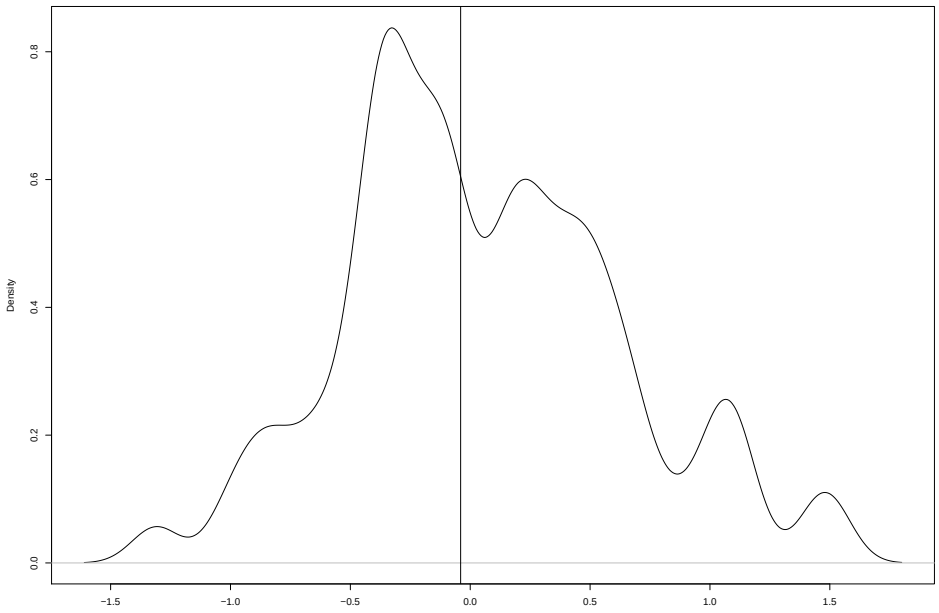


Figure 5.2: ***Density plot of proliferation signature values in MCL patients.*** **The 71 proliferation signature values are ranked on the x-axis. The vertical line marks the median. The values do not clearly build two clusters. The outcome correlates directly to the value. Exploratory analyses reveal clusters wich represent a separation by the median.**

ing to the median (Figure 5.2). We defined two groups - "s" for survival and "b" for bad prognosis. Patients with a high proliferation signature value tend to have a poorer prognosis than patients with a low proliferation signature value.

CGH data

Now to each single chromosome of the CGH data, the exploratory data analyses correspondence analysis (Figure 5.3) and principal component analysis (Figure 5.4) were applied. Both methods are useful for exploring information and structures in data in order to get a first impression. Principal

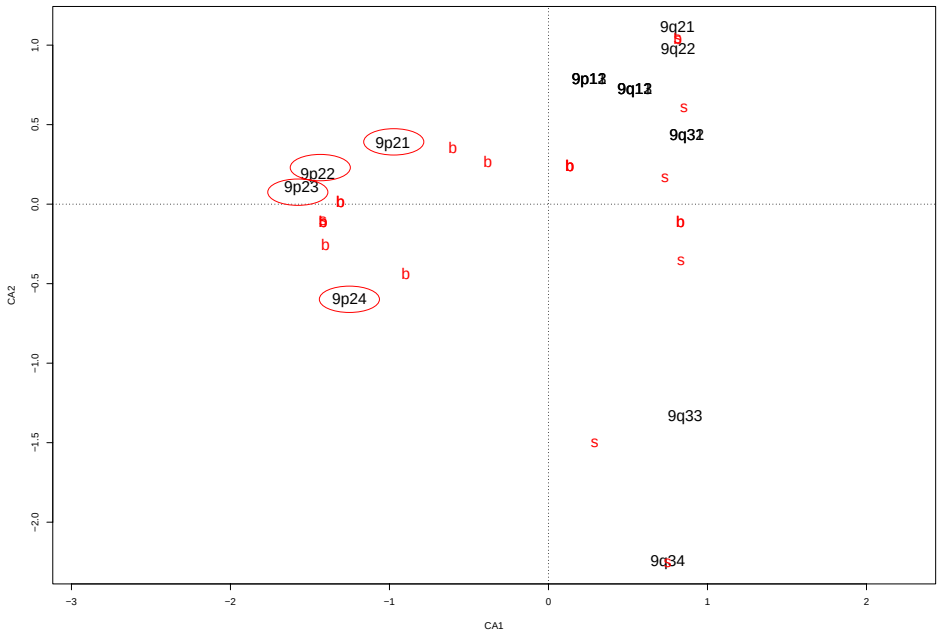


Figure 5.3: ***Correspondence analysis of chromosome 9 over the "s" and "b" group.*** **The first order factor axis separates almost completely these two groups. It is also obvious that the first four bands 9p24, 9p23, 9p22, 9p21 attract most of all b-patients. This leads to the assumption, that these four bands are responsible for the difference of the longer living "s" and the shorter living "b" patients. The second order factor axis separates at first glance strongly the last two bands 9q33 9q34 from the rest.**

components analysis reduces multidimensional data sets to lower dimensions for analysis. Correspondence analysis works similarly, but scales the data, which results in rows and columns being equivalent. Regarding CGH-data without prior knowledge and through these well known and smart methods provides an unprejudiced picture of the chromosomal aberrations.

The results indicate a strong correlation to the first four bands of chromosome 9, 9p24, 9p23, 9p22 and 9p21 and the proposed two subgroups "s" and "b". In the correspondence plot Figure 5.3, the four bands mentioned before attract most patients of the subgroup "b" and the first factor axis

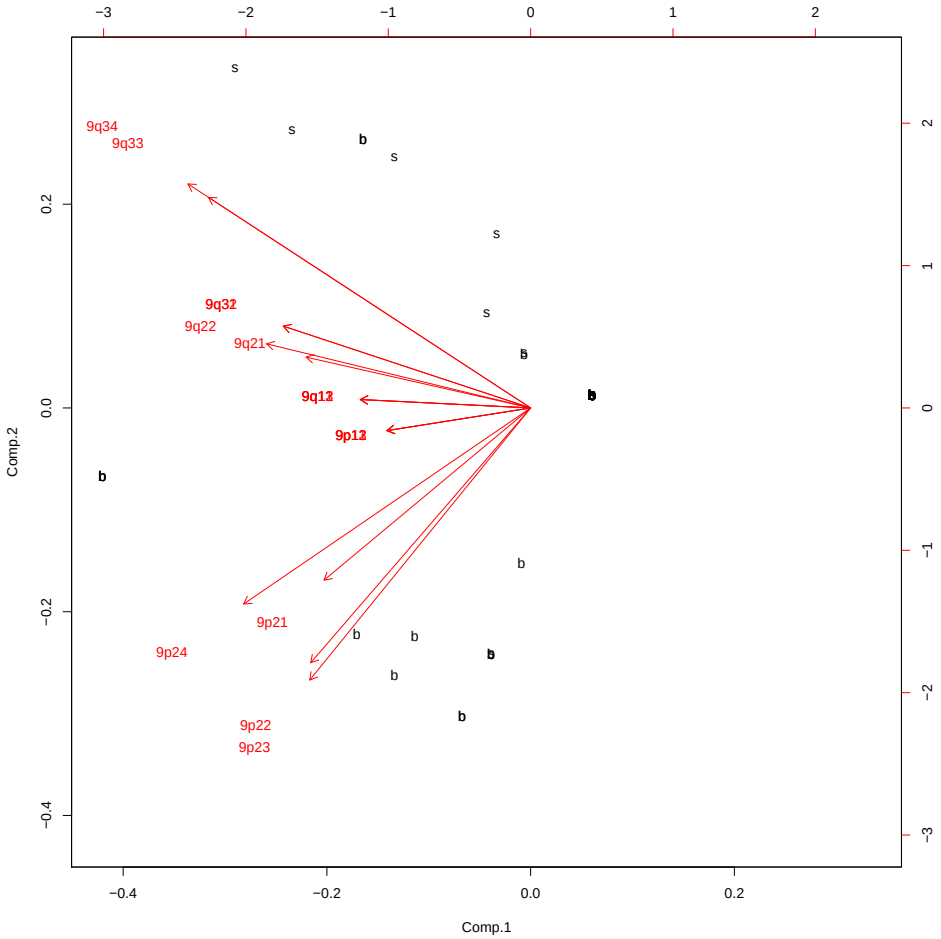


Figure 5.4: ***Principal Components Analysis of chromosome 9 bands separating the "s" and "b" group.*** **The second principal component separates almost all patients of the subgroup “b” from the remain. They are grouped together close to the first four vectors, corresponding to the first four bands 9p24, 9p23, 9p22, 9p21, which go into the same direction and are of similar length. Remarkable are the vectors of the bands 9q33 and 9q34. They also are of similar length and go exactly into the same direction. Along their length, they congregate almost all patients of the type “s”. This leads to the assumption, that the first four and the last two bands of chromosome 9 play a crucial role for “s” and “b” classification.**

separates almost completely the two groups. Therefore, these four bands seem to play a role towards a better prognosis. The bands 9q33 and 9q34 are located relatively far away from the remaining ones. In Figure 5.4, the second principal component groups almost all the "b" patients near the four bands 9p24, 9p23, 9p22, 9p21; vectors of these show similar length and similar direction. As in the correspondence analysis the bands 9q33 and 9q34 are grouped together. Here, their vectors show very similar length and the same direction. Along their length almost all "s" samples congregate. These results indicate that the these 6 bands, the first four and the last two bands of chromosome 9, are connected with the subgroups "s" and "b". So these exploratory analyses support the "s" and "b" classification.

Cox regression hazard model applied to CGH data

Further exploratory data analysis was performed to merge the survival time and the CGH-data by the Cox regression hazard model. To avoid problems with the regression we changed the data in such a way as no data points occur with a value of 0, which is associated with "no change". The values and their meanings are now: "loss": -1; "no change": 1; "gain": 2; "amplification": 3. So only the "loss" data values are transformed into -1. A univariate Cox regression hazard model was performed on all available bands of the CGH-data of all 71 patients. The above mentioned four bands of chromosome 9 delivered, amongst others, the most significant results. The resulting bands are "9p24", "9p23", "9p22", "9p21", "9q31" and "9q32". Note that the previous analyses revealed four bands, which intersect with these.

These two different entities were further examined. As proliferation marks the cancer progress, on average "s"-patients have a better prognosis than "b"-patients.

5.2 A gene expression based survival predictor

To improve survival predictions we searched with univariate Cox regression hazard analysis for highly significant genes, which correlate strongly with the overall survival time of the first 50 MCL samples, our training set. A four gene predictor with the genes CDC2, ASPM, tubulin-α and CENP-F (1) could not be tested, as after reannotation by GEPAT, mapping of CENP-F seems not sure anymore. Predictors with 4, 5 or 6 genes delivered not the desired predictive power (data not shown). So we identified for this task a set of seven genes, which again includes the well known "cell division cycle 20 homolog" (CDC20) and the "cell division cycle 2" (CDC2). The former one is known to be required for two microtubule-dependent processes, nuclear movement prior to anaphase and chromosome separation (Sethi et al., 1991). As a member of the Ser/Thr protein kinase family the latter is a catalytic subunit of the protein kinase complex "M-phase promoting factor" (MPF), which is necessary for G1/S and G2/M phase transitions of eukaryotic cell cycle. It is regulated by cyclins (Norbury and Nurse, 1991, 1992). HPRT1, the "hypoxanthine phosphoribosyltransferase 1" is located on chromosome X. It is known to be involved in colon cancer and Lesch-Nyhan syndrome

SpotID	Gene	GeneID	Official Full Name
6558	CENPE	ENSG00000138778	Centromeric protein E
7495	CDC20	ENSG00000117399	Cell division cycle protein 20 homolog
7892	HPRT1	ENSG00000165704	Hypoxanthine-guanine phosphoribosyltransferase
7019	CDC2	ENSG00000170312	Cell division control protein 2 homolog
7376	BIRC5	ENSG00000089685	Baculoviral IAP repeat-containing protein 5
6422	ASPM	ENSG00000066279	Abnormal spindle-like microcephaly-associated protein
5923	IGF2BP3	ENSG00000136231	IGF-II mRNA-binding protein 3

Table 5.1: *The genes of the survival predictor.* Univariate Cox regression hazard analysis revealed these seven genes best correlating with the survival time (see Material and Methods). The first column indicates the gene accession number in the data set, the second the gene name, followed by the official full name. The genes are ordered by their significance in decreasing order. CENPE ist the most significant gene.

(Jinnah et al., 2000). With CDC20, HPRT1 and CDC2 three of the strongest predictor genes match with three genes from the 20 genes proliferation signature of Rosenwald et al.. However, a good prediction power was obtained with additional four genes and a new compact survival predictor could be derived from this (Table 5.1).

As with the gene CENPF in the proliferation signature, there is one member of the centromere proteins in our predictor. The kinesin-like motor protein "centromere protein E" (CENPE) accumulates in the G2 phase of the cell cycle. It is supposed to be responsible for mammalian chromosome movement or spindle elongation or even both (Yen et al., 1992). The

"baculoviral IAP repeat-containing 5" (BIRC5) gene prevents apoptotic cell death and is a member of the "inhibitor of apoptosis" (IAP) gene family. It has been established that it is expressed in most tumours and in lymphoma (Ambrosini et al., 1997). It takes part in controlling cell proliferation and in regulation of cell lifespan. Additionally, it is supposed to participate in the spindle checkpoint and associates with AURKB (Beardmore et al., 2004). ASPM, the "asp (abnormal spindle) homolog" is essential for normal mitotic spindle function in embryonic neuroblasts (Bond et al., 2002). The protein encoded by "insulin-like growth factor 2 mRNA binding protein 3" (IGF2BP3), overexpressed in some human tumours, is found in the nucleolus. It binds there to the 5' UTR of the insulin-like growth factor II leader 3 mRNA and may repress translation of insulin-like growth factor II during late development (Müeller-Pillasch et al., 1997; Monk et al., 2002; Nielsen et al., 1999).

The seven genes were used to calculate a multivariate Cox regression hazard model and with its coefficients, a gene expression based survival estimator separated all 71 patients into two subgroups (Figure 8.1). Compared to proliferation signature's ability to distinguish two risk groups, the seven gene predictor does it similarly well (Figure 5.5). The correlation between this classification and the "s" and "b" groups of the proliferation signature (Figure 8.2) is overall about 0.62 and in the validation set (patients 51 - 71) it is 0.81.

A correspondence analysis of the 3000 genes with the highest variance showed clear clustering of these two entities (Figure 5.6).

Compared with the sample grouping of the proliferation signature (Figure

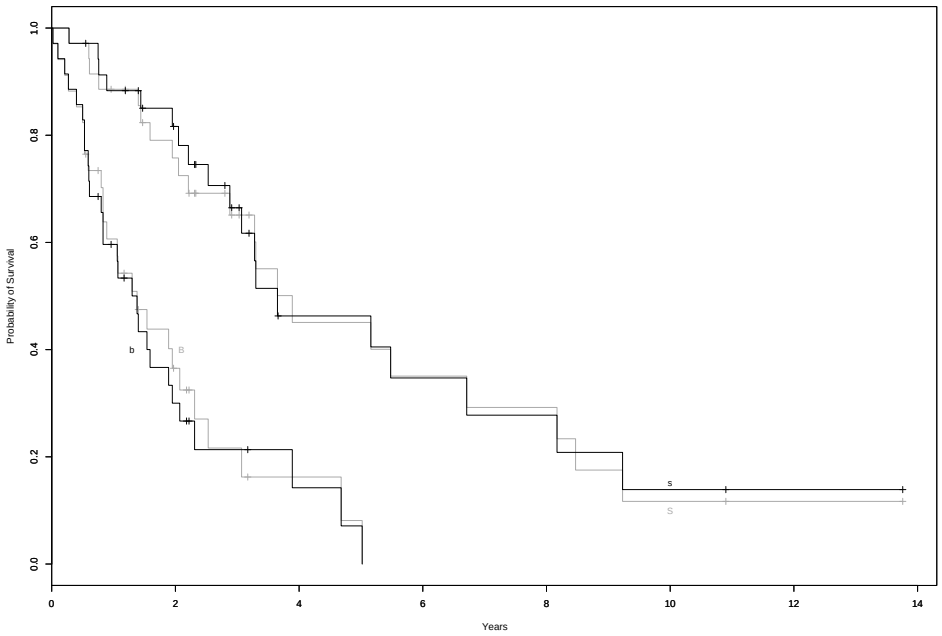


Figure 5.5: ***Kaplan Meier plot of survival data in MCL subgroups.*** **The x-axis denotes the course of time in years and the y-axis marks the probability of survival. Both, the proposed proliferation signature (black) and the seven genes predictor (grey) separate clearly two risk groups in the survival data. The overlap between the patients of the two classifications is relatively high.**

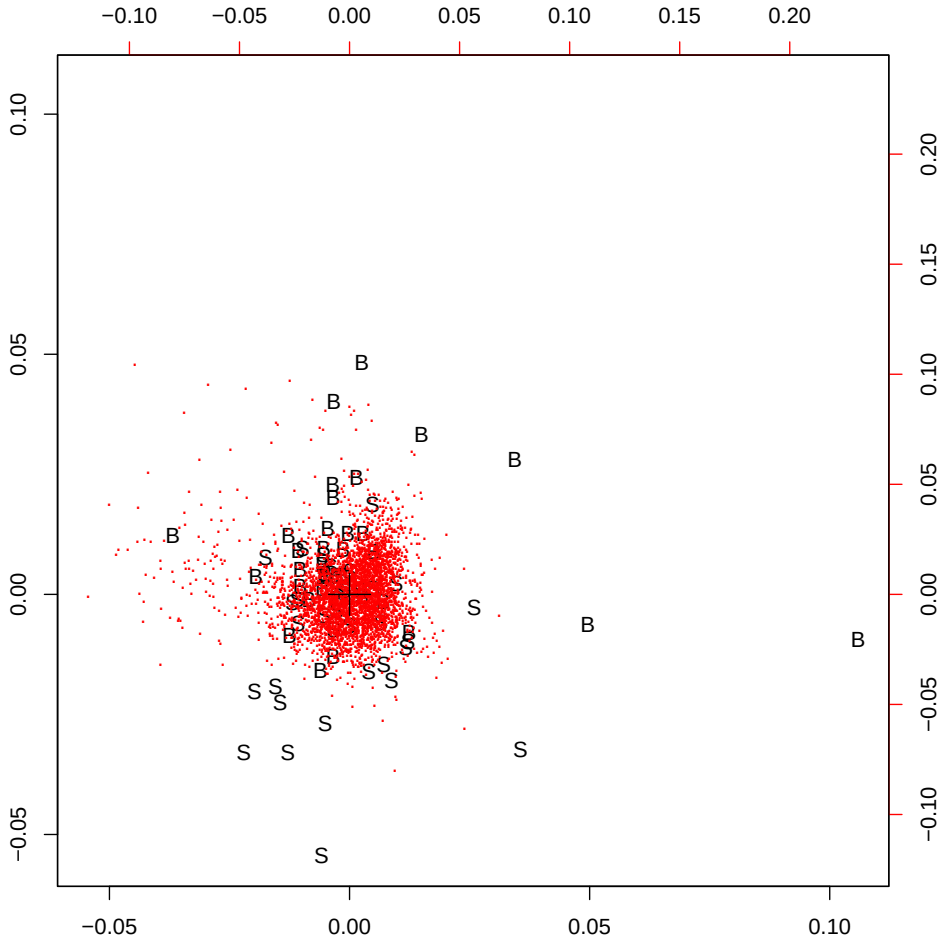


Figure 5.6: ***Correspondence analysis separates two MCL subgroups derived by the 7 genes survival predictor.*** **The 3.000 genes with highest variance (red dots) separate between the two subgroups, which were delivered by the seven gene predictor and are drawn as "S" and "B". They were separated by the median of the predictor values. In contrast to the proliferation signature based predictor (Figure 5.1), the patients here show a little more overlap, but cluster clearly.**

5.1), the samples show a little overlap, but are once again clearly separated. These results, taken together, give a clear indication, that the seven gene predictor is able to distinguish the patients almost like the proliferation signature, but with less effort.

5.3 Protein networks and interactions differentiating between two lymphoma subgroups

Following this we searched for differentially expressed key genes in the two subgroups "s" and "b". After applying a moderate t-test the well known cell division cycle 2(CDC2) gene, also known as CDK1, shows the most significance difference between "s" and "b" patients (Table 5.2).

It is important for the transition from G1 to S and G2 to M (Aleem et al., 2005; Malumbres and Barbacid, 2007), and its interaction partners show a significant up/downregulation as can be seen in the protein-protein interaction plot created with the hand curated "Human Protein Reference Database" (HPRD) database (Peri et al., 2003) (Figure 5.7). This indicates the genes being relevant for the pathogenicity of MCL.

Furthermore, our data show differences for genes encoding the direct interactions between CDC2, the Serine/Threonine kinase WEE1 and the tyrosine phosphatase "cell division cycle 25C" (CDC25). WEE1 catalyzes the inhibitory tyrosine phosphorylation of CDC2/cyclin B kinase, and appears to coordinate the transition between DNA replication and mitosis by protecting the nucleus from cytoplasmically activated CDC2 kinase. CDC25 directs de-

SpotID	Gene	Fold change	p-value	EnsemblID
7019	CDC2	1.3737029	1.8651454E-13	ENSG00000170312
6632	NP-057427.3	0.94384	3.4574367E-13	ENSG00000117724
3399	UHRF1	1.1446086	1.5513529E-12	ENSG00000034063
5112	NP-060880.2	1.0916529	1.5513529E-12	ENSG00000123485
6994	AURKB	1.4594886	1.5513529E-12	ENSG00000178999
6388	MKI67	1.5062114	1.7304206E-12	ENSG00000148773
6721	Q9Y645	1.2185314	3.2408542E-12	ENSG00000140451
7024	BUB1	1.2488679	3.2408542E-12	ENSG00000169679
6392	NP-057427.3	1.3208085	3.2902188E-12	ENSG00000117724
5726	MKI67	1.4871315	3.6012686E-12	ENSG00000148773
6029	NP-057427.3	1.2980943	5.249176E-12	ENSG00000117724
7423	BIRC5	1.3726515	6.49239E-12	ENSG00000089685
4985	ASPM	1.3310171	7.281489E-12	ENSG00000066279
5754	KIF23	1.2461857	1.6424877E-11	ENSG00000137807
5271	ASPM	1.3205649	2.2259293E-11	ENSG00000066279
6104	KIF23	1.1683029	2.4981522E-11	ENSG00000137807

Table 5.2: ***More significant genes separating good (s) and bad (b) prognosis.*** **The most significant genes after a moderate t-test between the groups "s" and "b" with Benjamini and Hochberg multiple testing correction. The gene "cell division cycle 2" (CDC2), which is important for the transition G1 to S and G2 to M shows the biggest difference in gene expression between the two groups. This indicates that these cell cycle transitions are part of the difference between the two groups.**

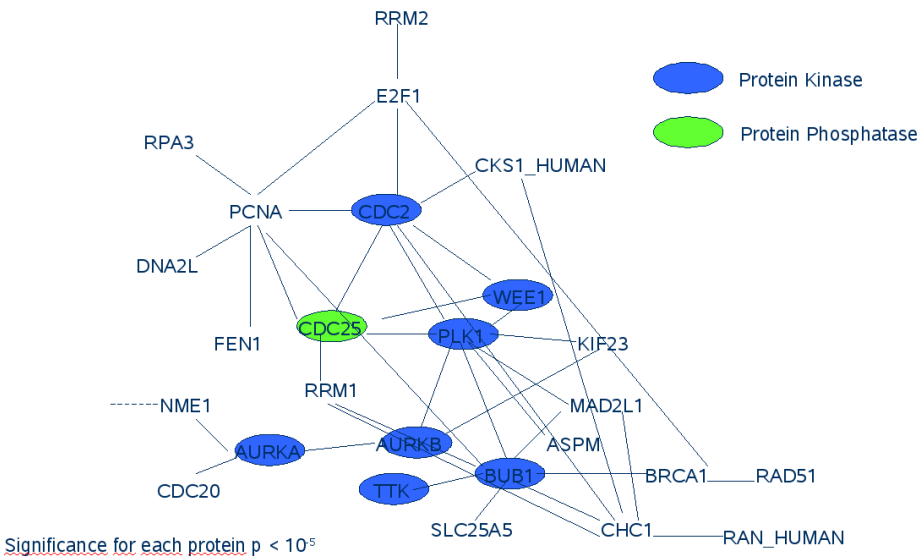


Figure 5.7: ***Protein interaction network (HPRD) of significantly different expressed genes.*** **The genes encoding these proteins show a significant expression difference between the "s" and "b" group (moderate t-test). Remarkably CDC2 is involved in a small interaction network of protein kinases and almost all of these interaction partners(CDC25, WEE1, AURKB, AURKA, BUB1) are associated with the cell cycle.**

phosphorylation of the cyclin B-bound CDC2 and triggers entry into mitosis (Schafer, 1998). However, a new finding are the expression differences in the Aurora kinases for "s" and "b". These associate with microtubules during chromosome movement and segregation during mitosis, whereas the kinase "budding uninhibited by benzimidazoles 1 homolog" (BUB1) is involved in spindle checkpoint function. "Aurora kinase B" (AURKB) localizes to microtubules near kinetochores, "Aurora kinase A" (AURKA) localizes to centrosomes (Lampson et al., 2004). BUB1 partly functions by phosphorylating CDC20, a member of the mitotic checkpoint complex, and activating the spindle checkpoint (Tang et al., 2004). Aside from other proteins, the checkpoint machinery consists of the kinases Bub1, Mps1, BubR1 and Aurora B (AURKB). It is possible that they phosphorylate diffusible key substrates and provide a way to amplify and strengthen the "wait-anaphase" signal in group "b".

It is remarkable and unknown for MCL, that a relatively tight network of cell cycle regulating phosphatases and kinases (CDC25, WEE1, AURKB, AURKA, BUB1) results in an up- or down regulation if the "b" and "s" group are compared (Figure 5.7).

The network of interaction partners of CDC2 from the STRING server is shown in Figure 5.8. STRING is a database with documented and predicted protein-protein interactions to which we mapped the genes with expression differences. We have found that CDC2, E2F1, PCNA, CDC25C, WEE1 and NCL, show high expression values in group "b".

Four proteins differently expressed in "s" and "b" join the interaction partners mentioned above. The "proliferating cell nuclear antigen" (PCNA),

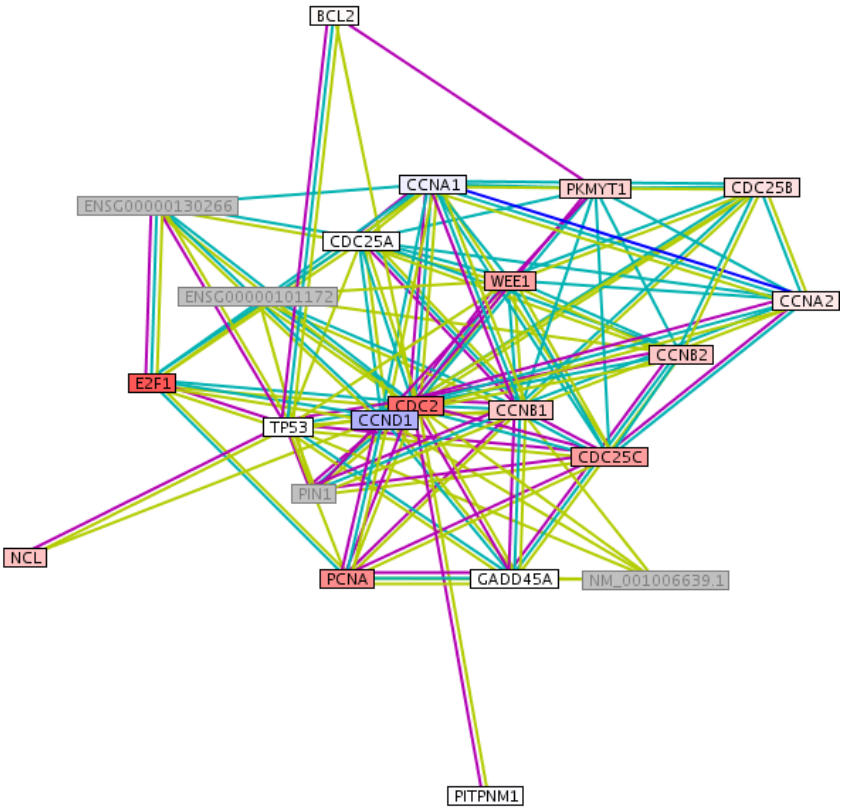


Figure 5.8: ***Differences in gene expression of interaction partners of CDC2 in MCL subgroups.*** **In this network figure, red indicates high expression and blue low expression in the subgroup "b" of the proliferation signature. White indicates no gene expression diïiňĂerence and grey the unavailability of the gene in our data set. "Cell division cycle 2" (CDC2) gene interacts in different manners with "cyclin D1" (CCND1), "cell division cycle 25C"(CDC25C), "proliferating cell nuclear antigen" (PCNA), "E2F transcription factor 1" (E2F1) and WEE1. CDC2 and CCND1 are both required for the G1/S transition (Aleem et al., 2005; Malumbres and Barbacid, 2007; Schafer, 1998). The genes WEE1 and CDC25C phosphorylate and dephosphorylate the complexes bound with CDC2 in a cell cycle regulating manner. The "proliferating cell nuclear antigen" (PCNA) is involved in DNA replication whereas "E2F transcription factor 1" (E2F1) controls cell cycle and mediates cell proliferation and apoptosis. A cell cycle regulated transcription activator "Nucleolin" (NCL) shows little difference.**

a cofactor of DNA polymerase delta, helps to increase the processivity of leading strand synthesis during DNA replication in group "b". Because of its ability to interact with multiple partners, it is involved in Okazaki fragment processing, DNA repair, translesion DNA synthesis, DNA methylation, chromatin remodeling and cell cycle regulation (Maga and Hubscher, 2003). The "E2F transcription factor 1" (E2F1) is a member of the E2F family of transcription factors and plays a crucial role in the control of the cell cycle. This protein can mediate both cell proliferation and p53-dependent/independent apoptosis (Crosby and Almasan, 2004) and is less expressed in "s". "Nucleolin" (NCL), an abundant multifunctional phosphoprotein of proliferating and cancerous cells (Lapeyre et al., 1987; Derenzini et al., 1995; Srivastava and Pollard, 1999), was identified as cell cycle regulated transcription activator (Grinstein et al., 2006) and is highly expressed in "b". CDC2 also interacts here with CCND1. CDC2 and CCND1 are both required for the G1/S transition. Note that with the exception of CDC2, CCND1 and E2F1 these genes were not known to play a role for survival in MCL.

Moreover, the interaction partners of cyclin D1 are significantly and differently expressed genes (Figure 5.9; the colour coding is the same as in Figure 5.8). Whereas CCND1 and CDK4 are assumed to be involved in cell cycle progression of MCL, MYC is suspected of increasing MCL's proliferation rate. FOS, JUN and MYBL2 are known to play a role in cancer, but not explicitly in MCL. Whereas FOS ("v-fos FBJ murine osteosarcoma viral oncogene homolog") and JUN ("jun oncogene") are weakly down regulated in "b" the remaining ones like MYC ("v-myc myelocytomatosis viral oncogene homolog (avian)"), MYBL2 ("v-myb myeloblastosis viral oncogene

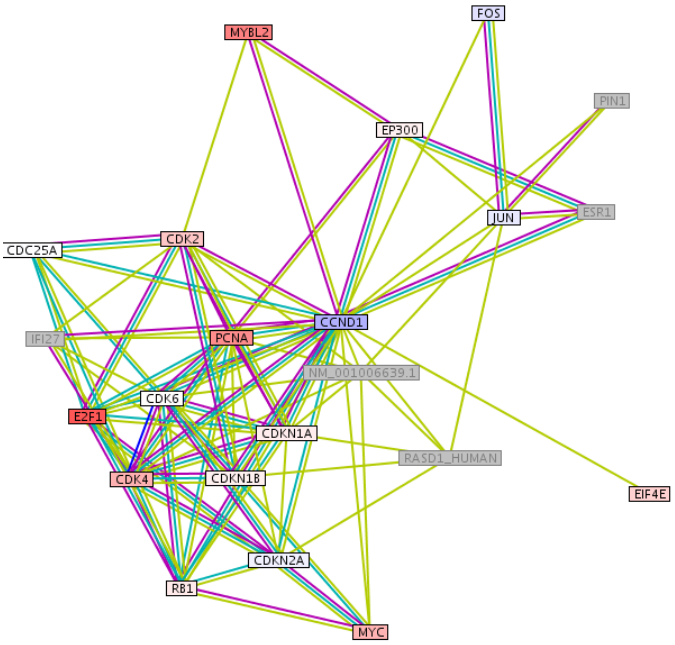


Figure 5.9: ***Protein interaction partners of CCND1: Different gene expression in MCL subgroups.*** **The colors red, blue and grey mean "over expressed", "down regulated" (in "b") and "not available in the data set". FOS encodes for a leucine zipper protein and plays a role in regulation of cell proliferation, differentiation, transformation and tumourigenesis (Milde-Langosch, 2005). The JUN protein interacts directly with specific target DNA sequences to regulate gene expression (Hartl et al., 2003) and is involved in tumorigenesis by cooperating with oncogenic alleles of Ras, an activator of the mitogen activated protein kinases (Weiss and Bohmann, 2004). MYC and MYBL2 play a role in cell cycle progression and act as transcription factors. MYC is also associated with apoptosis, cellular transformation, cell growth, proliferation, differentiation, and a variety of hematopoietic tumors, leukemias and lymphomas (Eisenman, 2001; Marcu et al., 1992; Pelengaris et al., 2002), and was part of the original proliferation signature (Rosenwald et al., 2003). MYBL2 has been shown to play a role in the G1/S transition (Golay et al., 1992) and proliferation (Sala and Watson, 1999) and is known to be regulated by CCND1 (Horstmann et al., 2000; Cesi et al., 2002). CDK4 and CDK6 are important regulators of cell cycle transition from G1 to S, phosphorylate, and thus regulate the activity of tumor suppressor protein Rb (Schafer, 1998).**

homolog (avian)-like 2"), CDK4 ("Cyclin-dependent kinase 4") and CDK6 show higher gene expression values. The high expression of transcription factors and cell cycle regulating factors in the "b" group emphasizes the poor prognosis for this group.

5.4 CGH data from chromosomes VII, IX support the classification and add new genes

To test the indication exploratory analysis of CGH-data revealed, we applied the Wilcoxon rank-sum test on the CGH data and compared the two groups "s" and "b". The null hypothesis corresponds to no differences between the two entities. The resulting p-values for every band of chromosome 9 are compared in Figure 5.10. They strongly indicate the significance of the first four bands 9p24, 9p23, 9p22 and 9p21. These bands have MCL related genes such as "cyclin-dependent kinase inhibitor 2B" (CDKN2B), "cyclin-dependent kinase inhibitor 2A" (CDKN2A) and "tumor protein p53" (TP53). TP53 mutations are associated with the blastoid variant of MCL and with a poorer prognosis. The bands 9q33 and 9q34 have less significance. To visualize this result more clearly Figure 8.3 plots the densities of the p-values. A peak in the density indicates significant bands of the Wilcoxon test.

The Wilcoxon rank sum test revealed similar results for chromosome 7. Here, the bands 7p21, 7p15, 7p14 are potentially related to the classification of "s" and "b" patients. Now the log p-values and their densities are plotted against the bands in Figure 5.11 and in Figure 8.4. The explorative analyses

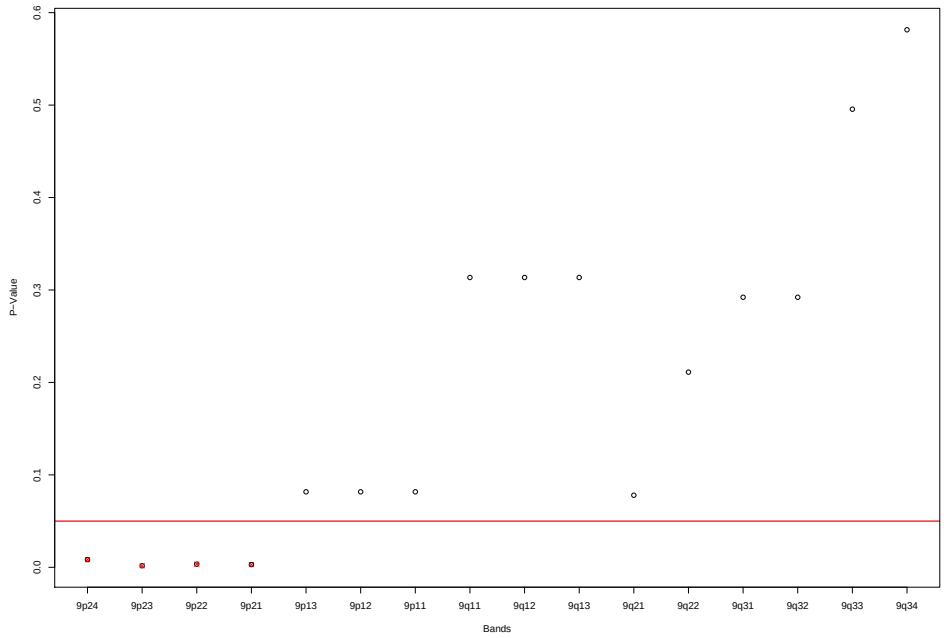


Figure 5.10: ***P values of the Wilcoxon test for the bands of chromosome 9.*** **This figure plots the bands of Chromosome 9 on the x-axis against the p-values of the Wilcoxon test (y-axis), which tested each band between the two groups "s" and "b". The p-values of the first four bands 9p24, 9p23, 9p22, 9p21 are very small, compared to the remaining ones. This affirms the proposed subgroups "s" and "b" and indicates that the first four bands have a relation to this classification.**

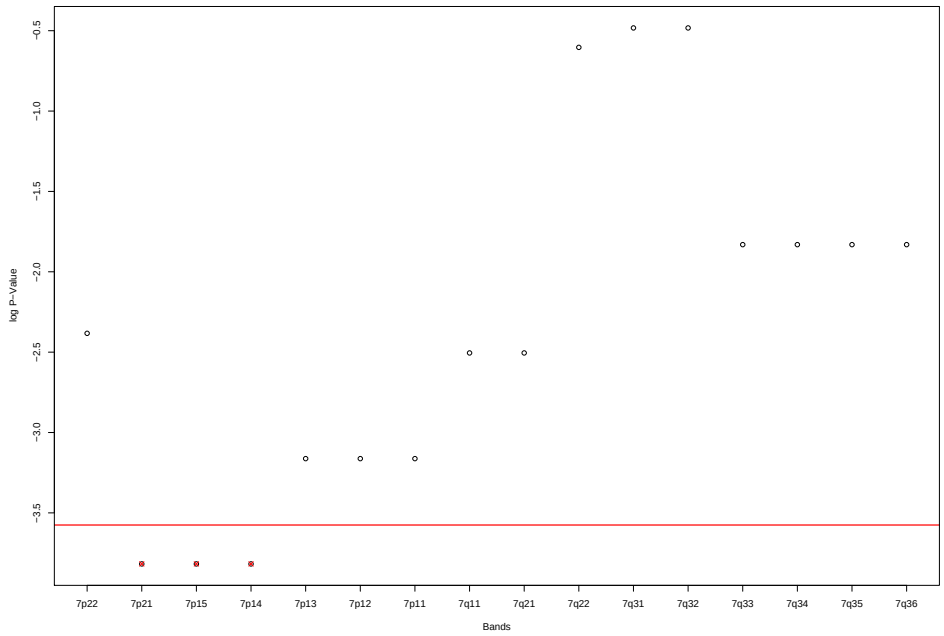


Figure 5.11: ***P-values of the Wilcoxon test for the bands of chromosome 7.*** **The Wilcoxon test was applied to all bands of chromosome 7 over the two groups "s" and "b". The bands of chromosome 7 (x-axis) are plotted against the log p-values (y-axis). Three bands show a very low p-value: 7p21, 7p15, 7p14. As the four bands of chromosome 9, they could have a relation to the "s" - "b" classification.**

of chromosome 7 did not reveal such a clear relation as in chromosome 9.

The proposed subgroups are defined by the gene expression based proliferation signature, which acts as a survival predictor, and are supported very well by the CGH data of chromosome 9. We checked the location of the signature genes as we wondered if they were on chromosome 9 or 7, however, this was not the case. Also the genes of the gene network in Figure 5.8 are located elsewhere. We investigated the gene expression data of these bands as none of the previously mentioned results could explain the relationship between the subgroups and the subgroup-separating CGH-data of chromosome

SpotID	Gene	Start bp	End bp	p-value	Official Full Name
7865	HSPA5	1.270e+07	1.270e+07	0.0304	Heat shock 70kDa protein 5
7073	PPP6C	1.269e+07	1.269e+07	0.0338	Protein phosphatase 6, catalytic subunit
7430	PBX3	1.275e+07	1.277e+07	0.0338	Pre-B-cell leukemia homeobox 3
1694	PTGS1	1.241e+07	1.241e+07	0.0393	Prostaglandin-endoperoxide synthase 1
7687	QSCN6L1	1.382e+07	1.382e+07	0.0393	Quiescin Q6 sulfhydryl oxidase 2

Table 5.3: ***The best "s" and "b" separating genes of chromosome 9 bands 9p24, 9p21, 9q33, and 9q34.*** **A moderate t-test revealed the following ones as the genes with the highest significance. Althogh the significance is weak, it is quite remarkable that these genes here show a distinct clustering on the basis of genomic positions, which can be observed in Figure 8.5**

9. Again a moderate t-test was applied to rank genes differentially expressed between "s" and "b". The top five are listed in Table 5.3, e.g. the "Heat Shock 70kDa protein 5" and a catalytic subunit of "Protein Phosphatase 6". Several of their functions implicate that they are critical in cancer development. Their genomic positions revealed a quite remarkable clustering of these genes, shown in Figure 8.5. Three of the genes seem to be located very closely to each other.

The "heat shock 70kDa protein 5" (HSPA5), also referred to as 'immunoglobulin heavy chain-binding protein' (BiP) targets misfolded proteins for degradation, and has an anti-apoptotic property. It is present in a wide variety of cancer cells and cancer biopsy tissues and contributes to tumor

growth and drug resistance of cancer cells (Li and Lee, 2006). The PPP6C gene encodes for a catalytic subunit of the Ser/Thr phosphatases, the "protein phosphatase 6 catalytic subunit" (Stefansson and Brautigan, 2006). The pre-B-cell leukemia transcription factor 3 (PBX3) reveals extensive homology to PBX1, a human homeobox gene involved in t(1;19) translocation in acute pre-B-cell leukemias. However, in contrast to PBX1, the expression of PBX3 is not restricted to particular states of differentiation or development (Monica et al., 1991). It is also known that if HoxB8, a homeobox gene identified as a cause of leukaemia, binds to the Pbx cofactors it blocks differentiation in certain cell types (Knoepfler et al., 2001). "Prostaglandin-endoperoxide synthase 1" (PTGS1) is the key enzyme in prostaglandin biosynthesis, and is known to play a role in the human colon cancer (Garavito and Mulichak, 2003; Wiese et al., 2003). The expression of the alternative splice variants is differentially regulated by cytokines and growth factors (DeWitt, 1991; Hla et al., 1993; Herschman, 1994). Very little is known about "quiescin Q6-like 1" (QSCN6L1), except its major role in regulating the sensitization of neuroblastoma cells for IFN-gamma-induced apoptosis (Wittke et al., 2003). Similar obvious clustering on chromosome 7 could not be observed.

Chapter 6

Analysis of gene expression: Survival and subgroups in DLBCL

In this chapter the raw data from a well documented study (Rosenwald et al., 2002) are analyzed. However, now more patients are involved in an enlarged data set with a total of 248 patients. Each patient array included 12196 gene spots corresponding to 3717 genes, generated with the "Lymphochip" (Alizadeh et al., 2000). With this specialized microarray Alizadeh et al. (Alizadeh et al., 2000) investigated the gene expression patterns of "diffuse large B-cell lymphoma" (DLBCL), "follicular lymphoma" (FL) and "chronic lymphocytic leukemia" (CLL). They identified two until then unknown distinct types of the DLBCL by gene expression profiling. As described in detail in chapter one the "activated B-cell-like DLBCL" (ABC) group has a worse overall survival than the "germinal center B-cell-like DLBCL" (GCB) group.

von Heydebreck et al. (2001) applied their new class discovery method ISIS on a subset of 62 samples and 4026 clones of the data by Alizadeh and colleagues (Alizadeh et al., 2000) and were able to show evidence for ABC and GCB next to CLL and FL in an unsupervised manner. Here ISIS was also applied and it confirmed the classical subgroups ABC DLBCL and GCB DLBCL independent from hierarchical clustering. Furthermore, it supports those subgroups being homogeneous entities in the data..

The survival analysis of Rosenwald et al. (2002), revealed gene expression signatures (collating several genes) and based on this an outcome predictor of survival. The constituents are the "Germinal-center B-cell signature", "MHC class II signature", "Lymph-node signature", "Proliferation signature" and the gene "BMP6". The predictor has a greater prognostic power in classifying patients into risk groups than the IPI (see above).

Starting with 36 well known DLBCL prognosis genes from the literature, Lossos et al. (Lossos et al., 2004) found a six gene based outcome predictor and applied it to the data sets of Alizadeh et al. (Alizadeh et al., 2000) and of Rosenwald et al. (Rosenwald et al., 2002). The latter one is an ongoing study and thus an extension and revision of the old data from Rosenwald et al. (2002) was possible (see Material and Methods).

To find better prognosis predictors, our analysis includes the expression values for these 36 well known DLBCL prognosis genes (Lossos et al., 2004). However, we use a data set enlarged to previous studies and we apply more adequate tools from the Bioconductor library (Gentleman et al., 2004) to derive better predictors than e.g. the six-spot predictor derived previously by (Lossos et al., 2004). We examine the data set to validate the marker gene

classification into exactly two pathological entities by a unbiased statistical classification analysis and show further that these confirms the classical subgroups ABC DLBCL and GCB DLBCL independent from gene expression signatures. After that, we advance the analysis of the raw data and obtain a simplified predictor with good quality for clinical prognosis (6 instead of 17 genes). More importantly, we identify and demonstrate that expression of early and late cell cycle genes distinguishes well the pathological entities ABC and GCB DLBCL. Furthermore, we show that the most significant gene expression differences found including the cell cycle genes as well as classical markers and best separating genes can be integrated into a compact key regulatory network showing clear expression differences between both diffuse large B-cell-lymphoma subgroups. This finding is verified comparing it to the average distribution of genes on the Lymphochip and the connection distances between them as well as confirming key gene expression differences found in our main data set by new analysis of further gene expression data (Shipp et al., 2002). The introduced methods can also be applied to other studies of gene expression analysis in cancer. Now a picture emerges where a central regulatory circuit tunes immune signatures, apoptotic and proliferation pathways in different ways between ABC and GCB DLBCL.

6.1 Survival analysis

The raw data include microarray data and survival data from 240 patients with diffuse large B-cell lymphoma as well as their pathological classification. For normalization the methods “loess” (W.S. Cleveland and Shyu, 1992; Yang

et al., 2001, 2002) and "scale" (Yang et al., 2001, 2002; Smyth and Speed, 2003) were used, as we are aware, for the first time.

Survival Prognosis of immune signatures applied to the data set The International Prognostic Index (IPI) score is often used in the clinical setting to differentiate lymphomas into low, medium and high risk cases. The immune signatures by Rosenwald et al. (2002) were shown to be independent from the clinical IPI score and useful predictors within the low, medium and high IPI risk groups on their data set.

We tested the performance of advanced normalization methods, namely the methods "loess" and "scale" on our enlarged data set. The IPI score is considered here only as an independent and established clinical prognosis marker. On normalized data of 240 patients and considering all individual spots we utilised Kaplan Meier plots (Figure 8.6) and reveal the efficient performance of the gene expression profiles (Rosenwald et al., 2002) also for this enlarged set of data using the improved normalization procedure. The low risk IPI group in the renormalized and enlarged data is not as well separated between the best and worst quartile as in Rosenwald et al. (Rosenwald et al., 2002). The separation of the high risk group is virtually unchanged. However, in the medium risk group a better separation was achieved by the renormalization and single spot analysis of all patient data. For the medium risk patients a better separation into high and low risk is particularly important for prognosis. The aim of the previous study mentioned above (Rosenwald et al., 2002) and our new study is of course to derive even better and gene expression based predictors of survival than the IPI score, which is consid-

ered here only as an independent and established clinical prognosis marker. However, these results support this notion with enlarged data and further developed normalization methods. This methodi, including the advanced normalization, can also be applied to any other microarray data set.

Simplified prognosis predictors for cox hazard regression survival analysis The immune signature requires the measurement of gene expression for a battery of genes. Next, we investigated whether a combination of fewer array spots and gene expression measurements is able to achieve a similarly good classification.

Multivariate analysis is computational prohibitive for more than 4 spots (results in Table 8.1, Table 8.2). However, by univariate analysis we could systematically test the capability of the gene expression values from individual spots to separate patients with good and bad prognosis in Kaplan-Meier plots. Within each of the three IPI classes the gene expression measurement should recognize and separate well the best patient quartile (with good prognosis) from the worst patient quartile (with poor prognosis). In contrast, a sub-optimal combination of spots confuses these patients and achieves no good separation in the plots. The univariate analysis was done with all genes and the 160 patients from the training-set and we identified the spots for the best describing outcome. Taken together in a multivariate model they form a predictor separating best and worse quartiles for all three IPI categories including the 80 patients from the validation-set. Results show, that five spots (details in Supplemental Material) are about equal to the six gene predictor of Lossos et al. (Lossos et al., 2004). Note that the 5 spot predictor also con-

Gene name	Gene Description
HLA-DPα	Major histocompatibility complex, class II, DP alpha 1
HLA-DQα	Major histocompatibility complex, class II, DQ alpha 1
HLA-DRb5	Major histocompatibility complex, class II, DR beta 1
SEPT1	Serologically defined breast cancer antigen NY-BR-24=Similar to DIFF6
EIF2S2	Eukaryotic translation initiation factor 2 subunit 2
IDH3A	Isocitrate dehydrogenase 3 (NAD+) alpha

Table 6.1: ***Optimal molecular survival predictor applying six genes.*** **The gene symbol (left side) is followed by the gene description. Three of these genes are HLA major histocompatibility complex genes (HLA).**

siders different splicing forms in HLA-DRB5. However six spots (Table 6.1) – corresponding to 6 genes – even are an improvement for this classification task. The five-spot-predictor includes the following spots: HLA-DPa, Brca, HLA-DQa, and two clones of HLA-DRB5. The six-spot-predictor includes the HLA-DPa, HLA-DQa, HLA-DRB5, Brca, ETIF2 and ID3A genes and shows an improved performance (Figure 6.1). The separation of the best and worst quartiles in the three IPI classes is comparable to the prediction success of the complete signature according to Rosenwald et al. ((Rosenwald et al., 2002), Figure 3) and classifies different patient quartiles better than the set proposed by Lossos et al. (LMO2, BCL6, FN1, CCND2, SCYA3 and BCL2 for overall survival in DLBCL). Our predictor is delivered by bioinformatical analysis of gene expression measurements, whereas Lossos et al. (2004) used RT-PCR. Our method, however can also be applied to RT-PCR. Moreover, we tested the influence of the high correlation between the genes HLA-DPa, HLA-DQa and HLA-DRB5 on the quality of the predictor. Therefore we estimated the survival probability with predictors of non correlated genes from

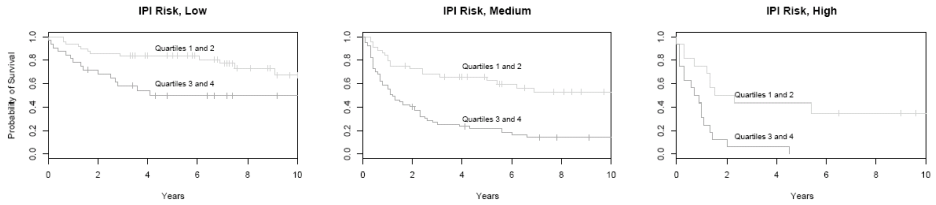


Figure 6.1: ***Prognosis prediction applying the molecular predictor of 6 genes after improved normalization.*** **Kaplan-Meier plots are estimated by a Cox-Regression Hazard model of the genes listed in Table 6.1. Normalization was improved applying the "loess" method. The x-axis corresponds to the time, measured in years and the y-axis denotes the probability of survival, predicted for the risk group. The predictor was applied over all patients in each single risk group, namely "low", "medium" and "high". The plots show large differences in the survival rate for all risk groups.**

the univariate analysis (data not shown). However, the survival probability yields no improvement in the results (data not shown).

6.2 An older DLBCL classification with ABC, GCB and Type 3

The patient classifications ABC, GCB and "Type 3" found by Rosenwald et al. (2002) are outdated. Type 3, supposed to include more than one type of DLBCL because of its heterogeneity, was rejected. So a more recent and accurate classification is available and used in the whole thesis. But in this chapter we introduce a remarkable result obtained with the old one. Within this rejected classification from 240 patients correspond 73 to ABC, 115 to GCB, and 52 are annotated to "Type 3". Until there is no more evidence for "Type 3", the following results have to be considered as interesting theory.

The genetic differences between the three DLBCL entities, ABC, GCB,

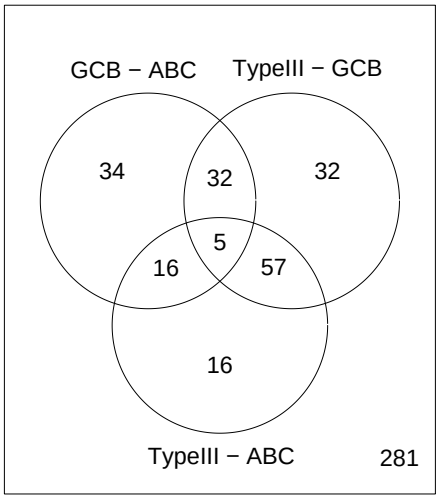
and Type 3 were now investigated through consideration of periodically expressed human genes associated with the cell cycle, a trigger of cell differentiation and proliferation. Within the cell cycle lies the crucial decision of abnormal proliferation.

The 240 samples were separated according to the three subtypes. We compared the gene expression measurements of cell cycle genes between the subgroups by moderate t-test. Significant differently expressed cell cyle genes were found in all comparisons for each single group. 32 were characteristic for GCB, 16 for ABC and 57 for Type 3. The intersection of all tests contains 5. 281 measurements showed no significant different expression for the subgroups (Figure 6.2(a)). We tested not only the cell cycle genes but all available genes (Figure 6.2(b)).

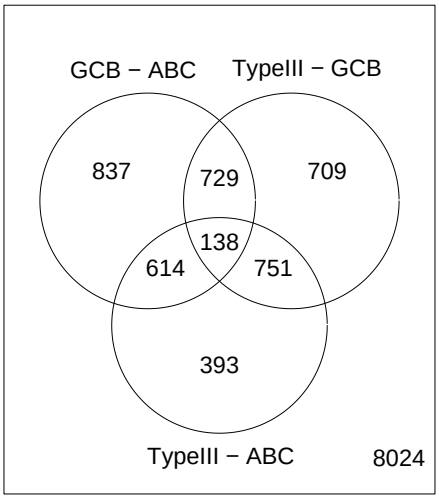
With this random noise we obtained almost the same result: almost all cell cycle genes were in the significant gene groups. This indicates that the result is independent of other gene expression values.

We then tested if the cell cycle result especially the peak for Type 3, reflects only a random result. As 473 is the amount of mapped cell cycle measurements, we randomly chose 1000 times 473 values randomly from all Lymphochip spots and each time we performed the moderate t-test. The result (Figure 6.3) very clearly shows that now there is no significant bigger amount of genes in the Type 3 describing intersection. Furthermore, the study revealed that the cell cycle result does not reflect a random result. On the contrary the comparison of random results and cell cycle results indicates a Type 3 specific subset of cell cycle genes.

As a result the cell cycle genes, which specifically assign B-cell lymphomas



(a) Differently expressed genes of the cell cycle set



(b) Differently expressed genes of the data set

Figure 6.2: ***Venn Diagrams of differently expressed genes.*** **Each circle represents a comparison between two subgroups. The overlap of circles gives the intersection of differently expressed genes. 6.2(a) shows cell cycle gene expression differences between the three B-cell lymphoma groups. 6.2(b) represents the same test with all "Lymphochip" genes, which include the cell cycle genes. Almost all cell cycle genes of 6.2(a) occur in the outlined groups in 6.2(b). The number on the bottom corner in on the right represents genes with "no significant difference" in the comparisons.**

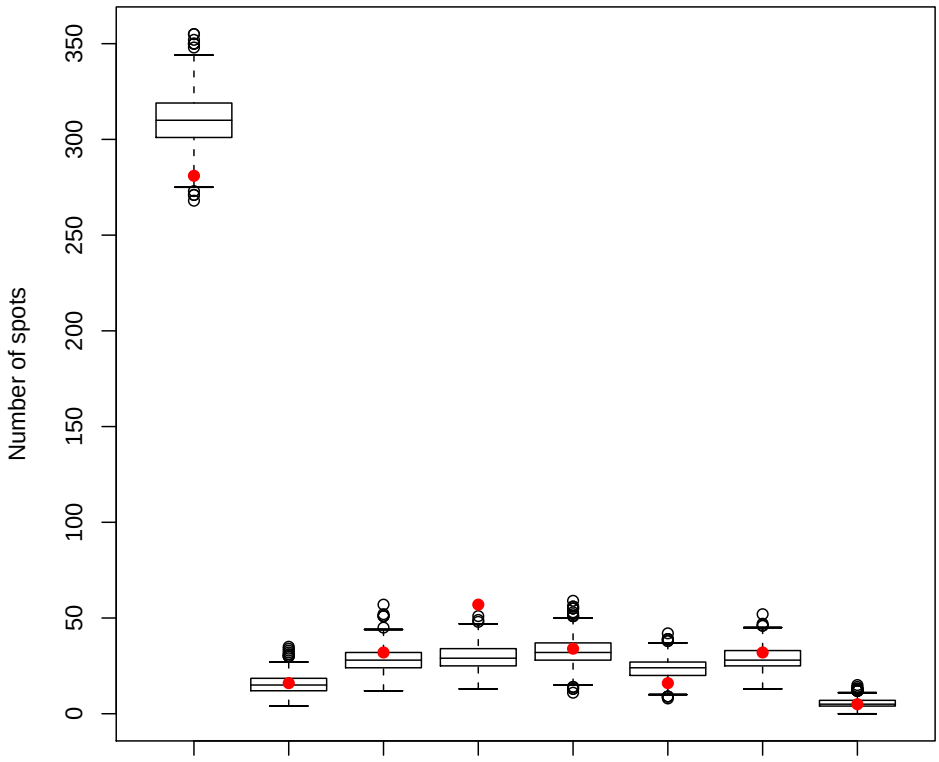


Figure 6.3: ***Differently expressed genes of randomly chosen genes.*** **To test if the peak of cell cycle genes in Type 3 was a random result, we randomly randomly chose the same amount of spots as in moderate t-test for cell cycle genes (473) and again we applied a moderate ttest with the same group comparisons. This procedure was performed 1000 times. The x-axis represents the group comparisons and the y-axis delivers the amount of differently expressed spots. The x-indices are from left to right: "Not significant", "Type3-ABC", "Type3-GCB", "(Type3-GCB) + (Type3-ABC)", "GCB-ABC", "(GCB-ABC) + (Type3- ABC)", "(GCB-ABC) + (Type3-GCB)", "overall intersection". The box plots indicate the distribution of the resulting gene amounts after random selection. The red dots signify the resulting gene amounts of moderate t-test, shown in Figure 6.2(a). Obviously the differently expressed genes in the Type 3 describing group "Type 3-GCB ∩ Type 3-ABC" (red dot in the fourth index on x-axis) show clearly a higher value than test results from randomly chosen genes (box plot).**

to Type 3, permit investigaton of the cell cycle states and allow us make an assumption in regard to which state Type 3 is predominantly occuring. As Type 3 patients are a heterogeneous group, a reclassification was performed and some of the Type 3 patients are now part of ABC or GCB type DLBCL-patients.

6.3 Statistical validation of ABC DLBCL and GCB DLBCL

The ABC and GCB DLBCL subgroups were originally introduced on the basis of gene expression profiling. There has been some suggestion that certain DLBCL form a third group (Hans et al., 2004). Furthermore, it is interesting to see whether this classification is also valid for an enlarged data set by an unsupervised classification method. Statistical analysis on the data (50 best separating genes from 3000 and a total of 150 patients) allows us to decide, independently of any pre-clustering of specific marker genes, whether there are 2, 3 or more lymphoma subgroups and whether they overlap with groups according to other group definitions (e.g. pathology).

The ISIS method (von Heydebreck et al., 2001) classifies data into two groups without prior (unsupervised) knowledge of the grouping. It investigates systematically all likely bipartitions on the gene expression data from our data set (see chapter 3) and gives a DLD score (diagonal linear discriminant score) for each partition. As a gene-expression profile based control, the samples were revised and 82 patients have been originally classified as

ABC subtype, 112 as GCB subgroup, for 48 patients no previous classification was available. A maximum sample size of 150 patients each for ISIS run considered 3000 measurements and delivered 50 best separating genes. Figure 6.4 shows the bipartitions ranked according to their DLD score. The bipartitions with the three highest DLD scores support and, in fact, identify each the two pathological entities ABC and GCB, though for this bipartition search only bipartitions were considered and no marker gene or signature pre-classification into ABC or GCB was applied. We further searched for subgroups within these splits but found no support for this. An appropriate bipartition could not be observed using previously putatively classified Type 3 patients and the ABC or GCB samples (data not shown). A further subgroup within the ABC or GCB entities is not validated by ISIS. We conclude that the precise separation into these two subgroups is thus well supported even by an unbiased statistical method independent of predefined expression signatures.

6.4 Genes best distinguishing DLBCL subgroups ABC from GCB

In all following chapters the most recent classification is used. Now from 248 patients 82 correspond to the ABC subtype whereas 112 are annotated to the GCB subgroup, for 48 patients thus a gene expression-based classification into either group was not available. The resting ones are associated with PMBL and MCL.

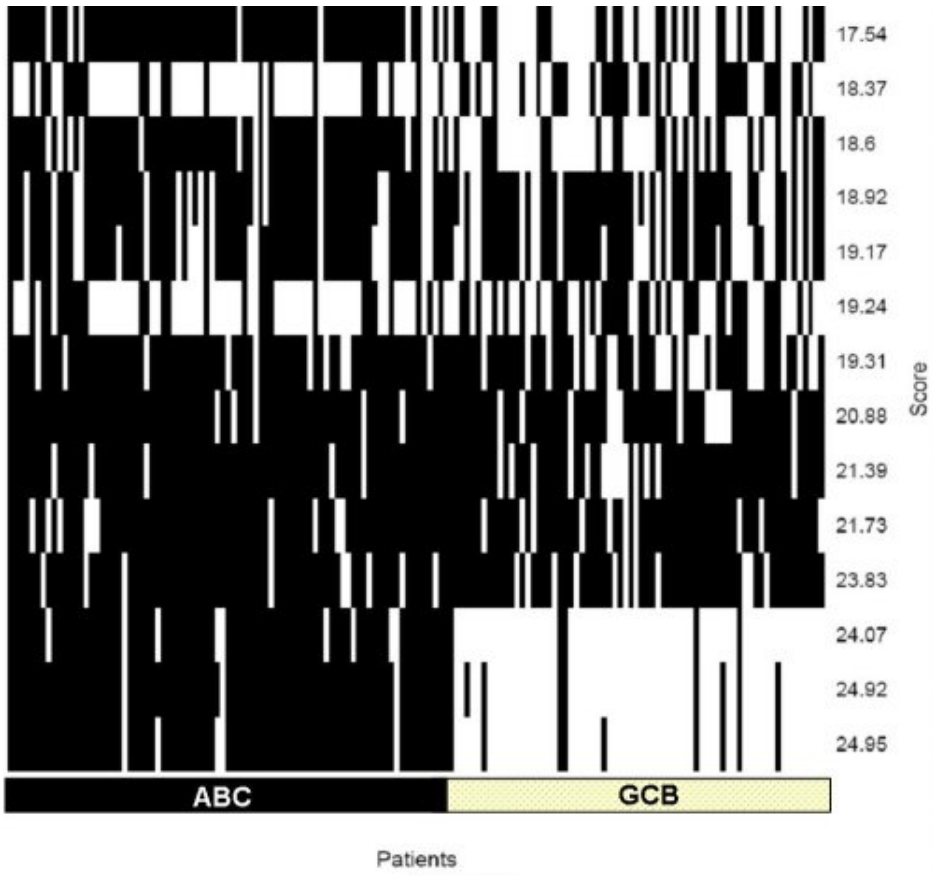


Figure 6.4: ***ABC and GCB DLBCL are clearly separated by unsupervised statistical analysis.*** **Optimal bipartitions of patients are calculated by ISIS based on optimal bipartition subsets of genes (50). Every column of the x-axis represents a patient. On the bottom, the DLBCL-type of the patient is labeled. On the y-axis every row shows the bipartitions ranked in increasing score of separation quality. The three best bipartitions show a very consistent and clear signal separating the ABC from the GCB patients. The unsupervised method ISIS reveals the ABC-GCB classification independent of proliferation signatures. No evidence for a previously suggested third group "Type 3"âĂİ was found. Only a few patients are falsely assigned if compared to the DLBCL gene signature assignment, some of the patients are consistently wrongly assigned over many optimal bipartitions. The unsupervised method ISIS reveals the ABC-GCB classification based on 50 genes, reflecting that the difference in gene expression between both tumour subgroups is the major signal in the data.**

The nearest shrunken centroid analysis using the R-package PAM ("Prediction Analysis of Microarrays") allows us to identify the best separating genes between the two subgroups ABC and GCB DLBCL with the smallest cross-validation error. Figure 8.7 in the supplement plots gene numbers of classifiers versus the resulting error rates. We identified a subset of 18 genes (31 spots), at which the associated optimal classifier indicates a good prediction power even with a small number of genes: Larger gene sets yield classifiers showing similar error rates (see Materials and Methods). However, smaller gene sets (less than 22 genes) result in inferior classification (Figure 8.7 upper plot). In detail, the error rate for the single subgroups differs between ABC and GCB DLBCL. Whereas GCB DLBCL is correctly predicted with few genes, the error for ABC DLBCL increases strongly (Figure 8.7 lower plot). However, for clinical application both entities have to be separated well and should not be confused. The optimal set of 18 genes is listed in the Table 8.8. Based on this gene subset a separation of ABC and GCB DLBCL with an overall cross validation error of 6.2% was achieved, see top of Figure 8.7. 5 out of 82 ABC DLBCL samples were falsely predicted as GCB, which corresponds to an error rate of 6.10%. The false prediction of 7 GCB samples out of 112 corresponds to an error rate of 6.25% (Figure 8.7 lower plot; error rates in Table 8.3).

6.5 Functional relationship of classical lymphoma marker genes and the genes from the best separating set

How well distinguish well known classical markers for lymphoma between these two subtypes of DLBCL? For this we collected classical lymphoma genes from literature (Monti et al., 2005; Lee et al., 2003; Willis et al., 1999; Polo et al., 2004; Rosenwald et al., 2002) and identified 35 genes in our data set which represent classical markers involved in lymphoma pathogenesis. Furthermore, we added the three metabolic enzymes LDH (IPI score prognosis marker), IDH and PDH (Table 8.4). Altogether these 38 genes correspond to 180 spots. PAM analysis identified from these genes a set of 9 well distinguishing genes (21 spots) (Table 8.5, Table 8.6), with an overall error rate of 14% (10% for the training set; 15% for the validation group). Thus these classical genes require more spots and their separation is less than the optimal prediction set above (Figure 8.7). To investigate the relationship between the best separating gene set identified above (see Figure 8.7) and the classical lymphoma marker genes, we merged them and performed the analysis again. We found, however, that here the best separating genes (Figure 8.7) achieve top ranks in this task (Table 8.7). It is only the "mitogen-activated protein kinase 10" (MAPK10), the best classical lymphoma marker, that reaches top ranks with 3 spots. BCL6 as the next best classical marker only reaches rank 31. Perhaps in contrast to expectation, the classical lymphoma genes do not add much information to the set of best separating diagnostic

genes. Below we show that classical marker genes are close to the central regulatory network best separating GCB and ABC DCBCL. However, the classical marker genes achieve not such a high rank in the separation task and hence are not that clearly differentially regulated as are the genes from the best-separating set.

6.6 Differences in the cell cycle

As cell cycle is critical for cancer cell division we next investigated by PAM analysis (see Material and Methods), whether the functional group of cell cycle genes alone could separate the two B-cell lymphoma groups.

We identified 473 spots in the data set, which correspond and are homologous to the cell cycle genes found by de Lichtenberg et al. (2005). These genes are annotated according to expression in the cell cycle state (100 steps between 0 and 99 for a full cell cycle). The separation between the lymphoma subgroups improves as more genes are used. We show here the result for 77 genes (Table 8.8, Table 8.9; error rate of 15.4%; classification optimum, see Materials and Methods). This set of cell cycle genes yields low error rates combined with a medium sized gene set. The genes are listed in Table 8.9. They mainly reflect the late cell cycle states. We asked how their cell cycle stages behave compared to the total of cell cycle genes. Figure 6.5 compares the complete cell cycle genes in our data set with the subset of 77 genes in a density plot.

The line over the coloured area indicates all cell cycle states of the whole chip and the bold dashed line the subset of 77 genes. The densities of this

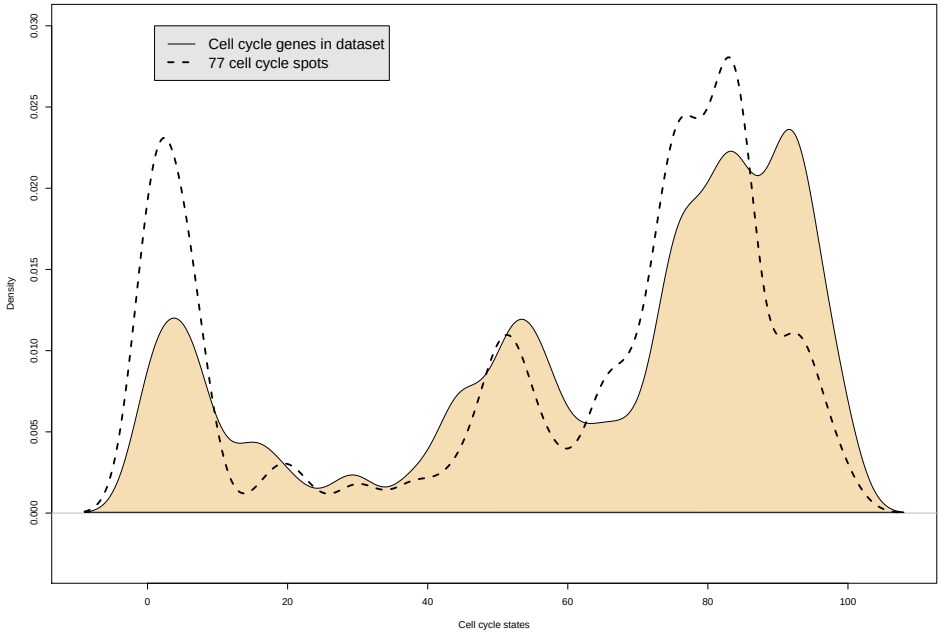


Figure 6.5: ***Early and late cell cycle genes are overrepresented in the best separating cell cycle gene set.*** **The density plot compares the distribution of different cell cycle gene sets. The x-axis denotes the cell cycle states from 0 to 99 (complete cell cycle), the y-axis represents the relative frequencies. The black line over the coloured area shows the density of all mapped cell cycle genes of de Lichtenberg et al. (2005) in the enlarged data set. The area is coloured for easier comparison. The dashed line represents the density of the optimal separating subset of cell cycle genes (77 spots). It is obvious that the subset, compared with the mapping of all cell cycle genes, has got two big peaks, one in the early and one in the later cell cycle states. These peaks indicate cell cycle gene expression differences between the subgroups ABC and GCB in these states.**

gene sets clearly differ in the early (0% - 18%) and in the late (75% - 85%) states ($p = 6.65 \cdot 10^{-10}$; Wilcoxon one sided test).

Additionally the cell cycle spots, which show the biggest difference in gene expression values between the two groups, are in the late states 72, 80, 84 and 85. Supplemental Figure 8.8 (MA-plot: middle intensity of the genes against difference in expression of both lymphoma subgroups) plots the whole data set and shows their expression values according to the ABC and GCB subgroups, cell cycle spots are drawn according to their state between 0 and 99 with colours from red to yellow. Spots with very big differences between the subgroups are labeled with their gene name and cell cycle state. They concern mainly the late cell cycle states. It is quite remarkable, that in the gene expression based two dimensions of the MA-plot these cell cycle states stay together and form a cluster. These results indicate a clear subgroup difference in the cell cycle states. Important differences concern the genes: butyrophilin-like 9 (BTNL9), early B-cell factor (EBF), cyclin G2 (CCNG2), TSC22 domain family, member 1 (TSC22D1), interleukin 6 (IL6), immediate early response 5 (IER5), TIMP metallopeptidase inhibitor 1(TIMP1) and v-maf musculoaponeurotic fibrosarcoma oncogene homolog (MAF). The function of these genes indicates specific differences in the behaviour of the two tumour subgroups.

6.7 Network analysis of differentially expressed key genes

To follow up these findings, we next investigated how the differentially expressed genes in the two subgroups are specially connected, and how their respective gene products interact with each other. To analyze this systematically, different large scale protein interaction databases were investigated. We first tried to apply the HPRD database (Peri et al., 2003) (hand curated protein interactions) but the data in this database are too sparse to cover all lymphoma markers analyzed. In contrast, the much larger meta- and protein-protein interactions database STRING (von Mering et al., 2005) allowed us to establish an interaction network. (Figure 6.6, details in Figure 8.9) Note that this analysis focuses on the clearly differently expressed genes in ABC and GCB (Table 8.7). The resulting interactions and functional classification of the genes identified are shown as a network in Figure 8.9. Classical lymphoma gene markers (dark grey boxes) as listed in Table 8.5 combine and interact with the compact cluster of the most powerful differentiating genes (white boxes) for the whole data set (Table 8.10) as delivered by PAM. The connections are mainly found by textmining and only the two interactions between BCL6 - IRF4 and between SH3BP5 - MAPK10 are available from the HPRD data set (experimental/biochemical Data) as a direct physical interaction (blue). The different article sources re-examine the interaction predictions for different tumor entities: “DLBCL”, “no cancer disease”âĂİand “other cancer”. Note that these categories support the interactions from three different view points (Figure 6.6).

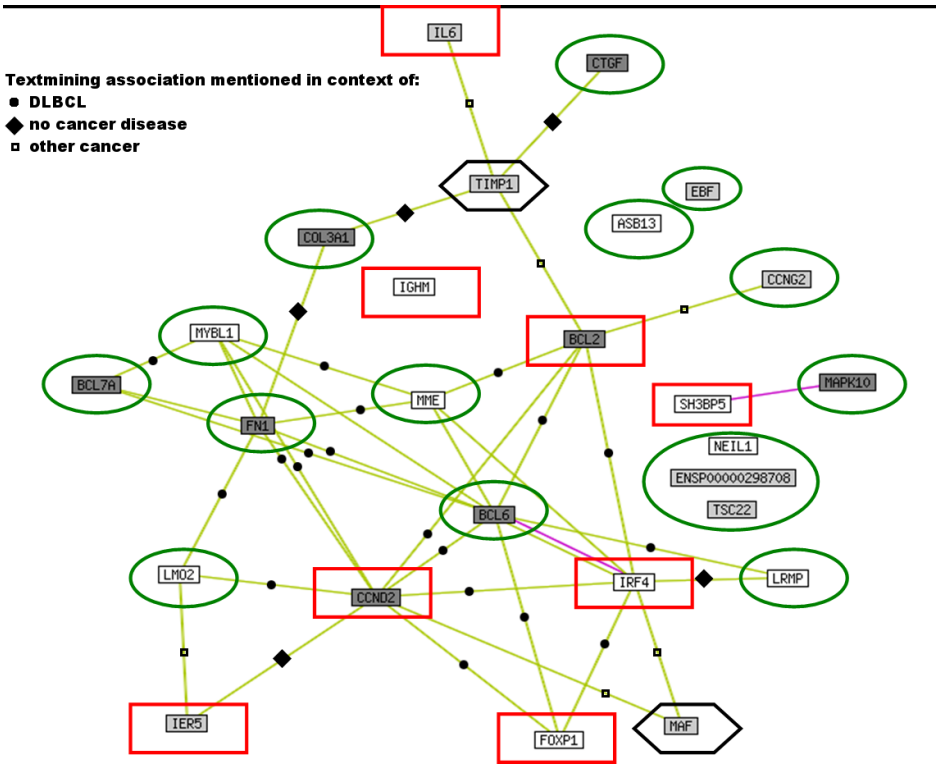


Figure 6.6: *Regulatory network differently regulated in ABC and GCB B-cell lymphomas.* Functional protein association network using interactions predicted by the STRING database: the most powerful predictive genes in the PAM analysis (white boxes; see Figure 8.9), classical textbook lymphoma genes (dark grey boxes), additional the cell cycle genes (light grey boxes; see Figure 8.8: 5 of these 8 cell cycle genes are connected directly with the network. TIMP1 even connects the so far uninvolved classical lymphoma gene CTGF with the network. This indicates how well the cell cycle genes fit to the existing graph). The new connections are confirmed by text mining of PubMed abstracts(circles: "DLBCL", diamonds: "no cancer disease", empty square: "other cancer"); these different data complement each other. The genes with a significantly higher expression in the ABC group are marked by a red rectangle. Green ellipses mark higher expression in GCB. Black hexagons mark genes which have a very high average gene expression value in both entities and are an important part for the network.

Validation of the central network including further tests. As a control for this finding of a compact regulatory network separating both entities regarding gene expression, we establish that all Lymphochip genes are equally distributed in regard to the human interactome and not pre-clustered in this respect (Figure 8.10). Moreover, the characteristic path length for randomly picked protein genes from the Lymphochip is 3.985 (Figure 8.11) and clearly longer than the direct interactions (path to lengths 1 or 2) found for differentially regulated network (Figure 8.9).

STRING found 11 of the 18 best separating genes and 8 of the 9 separating classical lymphoma genes as members of this dense interaction network. The remaining 8 genes, 7 from the first mentioned set and 1 from the latter one, are not part of the database. As "cyclin D2" (CCND2) occurs in these two subsets we obtain a protein-protein association network of 18 nodes. Regarding network regulation the underlined genes are more highly expressed in ABC, all others are more highly expressed expressed in GCB subtype: ASB13, BCL2, BCL6, BCL7A, CCND2, COL3A1, CTGF, FN1, FOXP1, IGHM, IRF4, LMO2, LRMP, MAPK10, MME, MYBL1, NEIL1 and SH3BP5 (Table 8.11). The network of different regulated genes, which was found, is predicted to closely interact with and influence each other. This has been determined by evidence from the STRING database, HPRD database and various interaction evidence types specifically collated by STRING). The different characteristics of the network can partly be rationalized. Protein functions involved in the network include stimulation of proliferation, block of proliferation, apoptosis, differentiation and immune cell specific functions (Table 6.2).

Functional categories	**Gene**	**Description**
Proliferation		
	CCND2	cyclin D2, regulates G1 to S transition of CDK4/CDK6; CTGF, fibroblast growth factor
	MAPK10	map kinase 10
	MYBL1	transcriptional activator in the proliferation of neurons, spermatogenic and B-lymphoid cells (recognition sequence: 5'YAAC(GT)G-3')
	ASB13	ankyrin repeat and sox box-containing protein 13, mediates protein-protein interactions, sox box couples suppressors of cytokine signalling and binding partners with elongin B and C complex to target them for degradation
	SH3BP5	SH3 domain binding protein, targets protein-protein interaction
Block of proliferation		
	MME	synonyms CALLA, common acute lymphocytic leukemia antigen, the synonym CD10 stresses its properties as a tumor suppressor gene

	BCL7A	putative tumor suppressor gene in T-cell lymphoma
Apoptosis		
	BCL2	integral outer mitochondrial protein to block apoptosis
	BCL6	transcriptional repressor, necessary for germinal center formation in lymph nodes
Differentiation		
	CTGF	fibroblast differentiation
	FOXP1	forkhead box P1
	LMO2	LIM domain only 2 transcription factor for hematopoetic development
	LAMP	expressed in lymphoid cells during development
	COL3A1	collagen type III
	FN1	fibronectin 1, cell adhesion
	NEIL1	base excision repair
Immune cell specific		
	IGHM	immunoglobulin heavy chain gene
	IRF4	interferon regulatory factor 4

Table 6.2: ***Regulatory network of genes best distinguishing ABC and GCB.*** **The genes of the network in Figure 8.9 are associated to the functional categories "Proliferation", "Block of proliferation", "Apoptosis", "Differentiation" and "Immune cell specific", by their annotation. Most of them are part of the antagonists "Proliferation" and "Block of proliferation". This indicates the complex regulation and importance of proliferation in the determination of ABC and GCB lymphomas. Classical lymphoma genes (see Table 8.4) known previously are given in italics.**

Interestingly, both subgroups reveal clear differences in these specific pathways and sub-networks with their regulation. Furthermore, the large collection of protein-protein interactions from the STRING database shows that all these different proteins separating the two subgroups are connected by first order interactions. Moreover, 5 of the 8 cell cycle genes identified in Figure 8.8 above, to be regulated differently are directly interacting with this regulatory network (Figure 6.6). The genes with a significantly higher expression in the ABC group are marked by a red rectangle, whereas green ellipses mark higher expression in GCB. These differences are an interesting pointer for a more specific anti-cancer treatment. In order to validate the gene expression differences that were found, we show that several of the key

gene expression differences identified are found again after analyzing further data from Shipp et al. ((Shipp et al., 2002; Wright et al., 2003); Table 8.12).

Gene functions for separating genes. The shorter survival of patients with ABC DLBCL is connected to pathways regulated differently from GCB DLBCL. Thus the well known BCL2 as a central apoptosis blocker is more highly expressed in order to allow tumor cell survival in ABC DLBCL. BCL6, a transcriptional repressor important for B-cell differentiation, is downregulated in ABC DLBCL. Altogether, apoptosis is less highly expressed in the ABC DLBCL subtype. Furthermore, the low gene expression values of the genes MME, a proliferation blocker, the proliferation promoting CCND2 and BCL7A, and the high expression values of SH3BP5 in the ABC DLBCL patients stimulate proliferation. Both the immune cell specific genes IGHM and IRF4 are more highly expressed in ABC DLBCL, however, all genes which are associated with differentiation are downregulated. In conclusion, this network indicates a downregulation of apoptosis and differentiation for the ABC DLBCL patients whereas the proliferation and immune cell stimulating genes are upregulated.

From the cell cycle genes which are connected to the network, IL6 and IER5 higher values are shown in the ABC group whereas BTNL9 and CCNG2 show an upregulation in the GCB group. For the latter it is known that CCNG2 and IL6 block the proliferation.

Do the clear gene expression differences between both subgroups reflect only differences in B-cell specific regulation? In order to gain a first impression, regarding T-cell regulatory pathways from our data, we tested whether

notch target genes, important in T cell differentiation (Reizis and Leder, 2002), are regulated differently in the two groups. Notch proteins are transmembrane receptors, which trigger the signalling pathway. They alter gene expression and are often repressed in many cancers. We conducted a search for differential regulation of notch target genes in diffuse large B-cell lymphomas (DLBCL). Notch target genes in Drosophila are regulated by GY-box-, Brd-box-, and K-box-class microRNAs (Lai et al., 2005). The boxes are found in the 3'-UTRs. We mapped all genes of the "Lymphochip" to the transcripts annotated in Ensemble database. We screened these and found candidate notch target genes, whose transcripts bear the target sequences that have been mentioned. All three boxes were found in the genes given in supplementary Table 8.13. From these transcripts the "Deoxycytidine kinase" gene (ENSG00000156136, DCK) and the "Translocation associated membrane protein 2" (ENSG00000065308, TRAM2) show clear gene expression differences between the ABC and GCB subgroups.

Chapter 7

Discussion

Limitations A certain limitation is that we support only acyclic networks. This means no cycles are allowed in the network and the cascade itself cannot be a cycle. Nevertheless giving the same protein or reaction as a start node and an end node allows to trace a cycle semi-automatically. The user should appreciate the function of the algorithm, especially the handling of inhibiting relations so that he is able to enter a correctly designated pathway. However this can be advantageously exploited for teaching.

7.1 Comparing Lymphoma subgroups

We identified key factors, genes and networks in order to improve understanding of tumor progression in lymphoma. We investigated survival data and gene expression differences between two entities of diffuse large B-cell lymphoma. We have described a compact prognostic predictor and a key network which separates gene regulation in ABC (Activated B-like DLBCL)

from GCB (Germinal Center B-like DLBCL) including cell cycle regulating differences.

The second project explored the severe mantle cell lymphoma. The most accurate and well known survival predictor was shown to reveal new relationships which were accessible by the effective bioinformatical analysis of gene expression combined with CGH data. CGH data were used to support the proliferation signature based patient classification. The results led to a biological functional network. Additionally a seven gene predictor was proposed, which distinguishes the two risk groups well.

In both projects microarray data were used to get results. As mentioned in chapter 2.4 microarrays are an appropriate measurement tool for diseases with gene associated changes. Further, using microarrays to obtain so many gene measurements makes it possible to investigate relationships between genes or gene sets. This is essential for understanding such diseases and subsequently their treatment. Finally, the advantage of such measurements is the possibility to combine them with data and additional information from different sources, as was done here with the cell cycle stages. As a result this data enabled effective production of relevant results.

Many marker genes for diagnosis and prognosis have been found previously by other authors and are discussed in the following text. Compared with these, our results were delivered by analysis of existing data and additional knowledge such as new patient classification or gene sets of interest. Generally, the data were enlarged and investigated bioinformatically. As a result our markers and found relationships include more than one biological aspect of these two lymphomas and result in an overall picture. The results

are discussed now in detail.

7.2 MCL

The MCL study consolidates gene expression and CGH-data regarding MCL subgroups with good or bad prognosis to an overall picture. These subgroups are indicated and confirmed by exploratory analyses. This picture highlights as yet unknown relationships and differences between patients from these groups.

Correspondence analysis of gene expression values indicated a statistical valid classification into the longer living "s" and the shorter living "b" subgroups. These were defined by the median of the outcome predictor score derived by proliferation signature (Rosenwald et al., 2003) as a discriminator.

A new prognostic indicator, similar to the proliferation signature, was developed with gene expression values of only seven genes, which are, undoubtedly, a much smaller gene set than the 20 genes of the proliferation signature. With the key genes CDC20, HPRT1 and CDC2 the seven-gene-predictor matches with three genes from the 20 genes proliferation signature. Moreover, the four genes CENPE, BIRC5, ASPM and IGF2BP3 add to its predictive power and are associated with chromosome movement, inhibition of apoptosis and tumors. It was shown that a four gene predictor (CDC2, ASPM, tubulin-alpha, CENP-F) (Rosenwald et al., 2003) is also able to predict length of survival with high statistical significance. Besides the fact, that the proliferation signature is more efficient and powerful than the four gene model, our model meets extensive re-annotation of the genes through

the clone IDs.

Also the CGH data supported the classification of “b” and “s”. The association of alterations in chromosomal regions and outcome of MCL patients was shown (Beà et al., 1999; Allen et al., 2002; Kohlhammer et al., 2004).

Gene expression analysis comparing the “s” and “b” groups, delivered mostly cell cycle related genes and their protein interactions, which determine prognosis. We systematically both confirmed and identified additional genes which were also found to be differentially expressed. Differently expressed interaction partners of the most significant gene CDC2 and the well known marker CCND1 revealed a network picture, which ensures the crucial role of the cell cycle in MCL. Thereby we confirmed the MCL relevant genes CDC2, CCND1, CDK4, MYC and E2F1 and found such genes as CDC25, WEE1, AURKB, AURKA, BUB1, PCNA, FOS, JUN and MYBL2 interesting specifically for MCL.

The Wilcoxon rank sum test reveal relations between the bands 9p24, 9p23, 9p22 and 9p21 and the difference in prognosis of the subgroups. Investigation of those bands and their most significant differentially expressed genes revealed a cluster of genes with properties such as “differentiation blocking”, “anti apoptotic” and “apoptosis inducing”. Supporting our finding, the band 9p21 was suggested, by microarray analysis, to be useful in MCL (Rubio-Moscardo et al., 2005). Less convincingly, some bands of chromsome 7 also confirmed the classification. Also here the the annotations and properties of embedded genes are known, but further data are required to better explain the relation between gene functions and survival. CGH-data may improve the power of gene expression based predictors in MCL and influence the gene

expression (Salaverria et al., 2007). Besides others, the band 9p21 was associated with a poor clinical outcome, which affirms our finding. But this study extends these results in two ways.

1. exploratory analysis shows here for the first time, that in fact CGH-data alone point and support two different MCL groups with clearly different prognosis.

2. CGH-data point here directly to several genes regulated differently in these two subgroups.

The analysis found after careful reannotation of involved genes two entities of MCL patients which could be supported by exploratory analysis, gene expression values, CGH-data, network analysis and literature mining. We obtain an improved classification of MCL subgroups in which differentially expressed genes led to a small protein interaction network of kinases. A seven gene predictor appears as an easy to measure prognosis indicator for clinical use. The Wilcoxon rank sum test was for the first time applied successfully to a CGH data set in this study as well as the PCA, which nicely focus on chromosome 9. Following the indicated bands, we found differentially expressed, cancer related genes. We conclude that the combination of gene expression and CGH-data reveals new impressions of MCL.

7.3 DLBCL

This study strives to improve marker gene detection for prognosis and subtype diagnosis of Diffuse Large B-cell Lymphomas (DLBCL) through the

application of a range of methods that are also useful for other gene expression measurements in cancer. This was achieved on different levels:

An adequate normalization of the gene expression intensities, applying the loess method (W.S. Cleveland and Shyu, 1992; Yang et al., 2001, 2002) allowed a better separation for good and poor outcome quartiles of survival, in particular for patients with medium IPI score where a better separation is important in order to give an accurate prognosis.

We investigated simplified predictors: multivariate analysis showed that a four-spot predictor does not perform well. Univariate analysis found that a six spot predictor is able to reflect prognosis better than quartiles in previous six-spot predictors (Lossos et al., 2004) or a tested five spot predictor, in particular for high risk patients.

Simplified prognosis predictors are of great importance in cancer treatment. In regard to diffuse large B-cell lymphomas, the IPI score (199, 1993) is a clinical standard. Our analysis identified a simplified predictor that is very useful and reliable for clinical prognosis (6 instead of 17 gene spots). Such a simplified prognosis predictor can be more easily applied in clinical settings than a measurement of a complex gene array signature. This finding will be also useful for clinical monitoring e.g. applying RT-PCR (Lossos et al., 2003).

The classification of diffuse large B-cell lymphoma into two pathological entities has been established by marker genes and their expression for some time (Alizadeh et al., 2000). A third entity has been discussed (Hans et al., 2004) but was disputed again in the light of recent data. The present statistical analysis where we applied the ISIS method provides an independent

and unsupervised method to support and identify these two subgroups. In addition to the work of von Heydebreck et al., ISIS analysis here clearly indicates, for a large data set, the separation of patients into the two subgroups ABC and GCB through an unbiased and independent statistical method.

Integrated picture of all gene regulation differences. Following this the statistical analysis identified all genes which clearly distinguish the ABC and GCB DLBCL subgroups, including differences in early and late cell cycle, which could be exploited for a differential cytostatic therapy in the two subgroups. To get an integrated picture of these differences we considered all the identified gene expression differences in order to obtain a detailed description of the differences between both DLBCL subgroups regarding regulation of the cellular network. We show that immune signatures, apoptotic and proliferation pathways are tuned in different ways between ABC and GCB DLBCL. A central circuit of genes is formed by genes that distinguish both lymphoma subgroups and are regulated differently. We also verified this for other data after completion of the first analysis. For the data by Shipp et al. (Shipp et al., 2002; Wright et al., 2003) once again key genes from the central network shown in Figure 8.9 are confirmed as having a significant different regulation in this totally different data and patient set (Table 8.12). Classical lymphoma genes are either directly or indirectly interacting with it.

Besides this central network other pathways are also implicated, we showed that two Notch pathway targets are specifically upregulated.

The different predictors shown in this study were the best predictors

according to PAM curves and statistical analysis, and also it gave clear improvements in determining prognosis compared to previous studies (Alizadeh et al., 1999; Lossos et al., 2004). Furthermore, our hypothesis is not pure speculation but directly includes enlarged gene expression data on 248 patients and individual analysis of 12196 array spots compared to pooled data whereas fewer patients were used in a number of older studies (Alizadeh et al., 1999; Lossos et al., 2004). Significant marker genes were found in this study by different statistical methods (PAM, ISIS, LIMMA). Clearly, using other methods (e.g. support vector machine) different gene sets can be obtained. However, the influential gene selection of PAM has previously been shown (Tibshirani et al., 2002). The different gene sets were validated against each other including classical marker genes and furthermore by analysis of new data (Shipp et al., 2002; Wright et al., 2003). The genes found are shown to form a compact interaction network obtained by another independent analysis method. Furthermore, the delineated regulatory network adds biological data and data from large-scale interaction databases to show that the identified marker genes are in fact members of a close interacting regulatory network, with molecular functions that mirror the differences in pathology of the two subgroups GCB and ABC DLBCL. The statistical analysis of the cell cycle genes and their associated cell cycle states indicates a possible target for therapy. Differences between the ABC and GCB DLBCL subgroups are at the beginning and the end of the M-phase and the early part of the G1 phase. Inhibiting early cell cycle genes, overexpressed in ABC, adding known cytostatic drugs such as mitosis inhibitors and early G1 blocker, may be particularly useful for ABC DLBCL patients. A more

detailed therapy profile would take the further differences in regulation into account.

Conclusion. The present analysis reveals through the use of an array of methods a detailed picture for molecular markers differentiating GCB and ABC DLBCL for prognosis and diagnosis. We apply it to GCB and ABC DLBCL for clinical use in determining prognosis and diagnosis, this included efficient six spot predictors. The entities ABC and GCB DLBCL have been confirmed by statistical analysis independent of gene expression signatures, a third entity could not be supported. The resulting genes with altered expressions were found to form a tightly connected regulatory network including cell cycle genes, apoptosis and immune differentiation implicated in the clinical severity of ABC DLBCL compared to the GCB DLBCL subtype.

7.4 Interlude: ACTIN

ACTIN is a Java based tool which delivers a simulation with the Depth-First-Search. The visualization of the simulation is carried out by showing the found traces by writing the labels of the visited nodes. The program directly sums up stimuli at each node conveniently using the hash map and can process large quantities of data. In a network of 1000 nodes there are up to $2^{1000} = 10^{300}$ possibilities how input can be processed before the action comes about. Future effort will include additional important information such as modification and/or different processing times of individual nodes and conditions to better separate and identify the biological important paths

leading to the various effector reactions when the cascade is properly switched by the correct biological stimuli.

7.5 General Challenges

The combination of knowledge and data from different sources is – amongst others – a characteristic property of bioinformatics and critical for biological insight into cell differentiation and cancer. The fascination in this study was enhanced by the discovery of new knowledge and the subsequent conclusions. Subsequent to this caution needs to be exercised to ensure an accurate observation and of the results. Also here, in two lymphoma analyses different sources like experimental data, databases and knowledge from literature were combined. The results of the two single lymphoma analyses are now discussed together, how they benefit and contribute to biological knowledge and the development of treatment.

Both tasks conclude that the cell cycle plays an important role in survival and tumor progression. A general problem that occurs in every gene expression analysis, is that results are based on previously used normalization methods. Notable here is that alternative normalization methods can have different effects on results (Irizarry et al., 2003; Quackenbush, 2002; Smyth and Speed, 2003). As there is no "best" normalization method we cannot reflect on this topic here, and have to assume that we have chosen the most appropriate method available. We mention here, that recently a modified "lowess" normalization method for special boutique two-colour microarrays as the Lymphochip was proposed (Oshlack et al., 2007).

With high throughput analysis and univariate Cox regression hazard models a gene expression based survival predictor was created. As was shown in another high throughput study (Rosenwald et al., 2002) and a MHC class II gene expression analysis (Rimsza et al., 2004) this gene family is correlated with poor patient survival. The proposed predictor reflects this, consisting of 50% MHC class II genes. The unsatisfactory treatment success of the IPI classification led to gene expression based predictors which show obvious improved prognostic accuracy as most probably they are associated with the cause, progress and biology of cancer disease. Interestingly, it has been recently shown that including the IPI with dimension reduction methods can improve the accuracy of prognosis compared to only gene expression based predictors (Li, 2006), but including the IPI as a simple covariate in the Cox regression hazard model did not lead to an improved predictor in this study (data not shown).

Several other predictors can be found with different methods. Here a publication should be mentioned, which re-evaluates and extends this statistical methodology applied to the same data set used in this thesis (Segal, 2006) and another one, which found three genes of our own predictor (SEPT1, HLA-DPα, EIF2S2) to be associated with survival time using Bayesian variable selection (Sha et al., 2006). HLA-DPα was also identified by (Rosenwald et al., 2002) and (Gui and Li, 2005) to be correlated with survival.

Regarding the differences in cell cycle states between the ABC and GCB DLBCL we conclude that a combination of gene expression data and cell cycle time series experiments can produce reliable results even if the gene expression data were not measured in a time series experiment. We are

aware, that this kind of combination was done for the first time for this data set. As a result new information was found from the well known gene expression data set, resulting in new potential targets for treatment of ABC and GCB DLBCL, at the end of M-phase and the early part of G1 phase.

The analysis of the rejected Type 3 confronts us with a result that raises a question. Why there is a signal and what does it reflect?

MCL analysis delivered differences between patients with poor and good prognosis in gene expression and CGH data, and the classification was stimulated by explorative analysis and literature (Rosenwald et al., 2003). The differences between the two types of MCL patients help in the understanding of a lymphoma, that to date remains incurable. Using the Cox regression hazard models, a useful gene set was defined which distinguishes patients with good and poor prognosis. An alternative to gene expression measurements is the immunohistochemical measurement of prognostic markers. With Repp86 a proliferation marker for MCL patients expressed in cell cycle phases S, G2 and M delivers similar survival curves as Rosenwald et al. (2003) (Schrader et al., 2005).

The undoubted crucial role of CCND1 in MCL becomes more complex as point mutations and genomic deletions in CCND1 influence the proliferation rate and hence the survival (Wiestner et al., 2007). Unfortunately this information cannot be taken into account in the data set used here.

The screening for potential drug targets using bioinformatics and genomic information proceeds (Ricke et al., 2006; Krasky et al., 2007) and the investigation of two well known data sets in this thesis led to potential drug and treatment targets of two Lymphomas.

Microarrays and cancer in general Genetic diseases are usually attributed to a set of genes, proteins and other factors. As protein expression is very difficult to measure, gene expression analysis usually investigates transcription of genes. Microarrays make investigation of gene expression and their connection to function an easy to use approach (Brown and Botstein, 1999) and presumably they will become a standard laboratory tool (Saito, 2006). Their use was proved (Spellman et al., 1998) and the understanding of cell properties continues to grow. Although limitations and problems are known (Oshlack et al., 2007; Chaudhuri, 2005) there is a lot of optimism surrounding this technology (Strauss, 2006; Jayapal and Melendez, 2006). Also science in general and especially treatment of cancer benefit as microarrays are helpful in drug discovery and validation of therapeutics. They are particularly, helpful in forming hypothesis (Gerhold et al., 2002), identifying biological pathways (Quackenbush, 2001; Eisen et al., 1998) and have enormous potential (Cole et al., 1999).

Cancer is a set of abnormal cells dividing consistently. The cause of most types of human cancers is unknown, but as uncontrolled growth and proliferation is their common property, scientists suspect key events associated with cell cycle and proliferation. The comparative analyses of gene expression in this study pointed to cell cycle relevant genes. As the stages of human cell cycle genes are known and were successfully applied here, we propose that there will be further promising experiments in which the expression of human cell cycle genes in lymphoma or general in cancer samples are measured. As in this task not all human cell cycle genes could be mapped on the chip, microarrays for the human genome or boutique arrays with these genes

should be used. Another approach could be a meta-analysis of existing gene expression data.

Bibliography

Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H. Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, Elaine M Hurt, Hong Zhao, Lauren Averett, Liming Yang, Wyndham H Wilson, Elaine S Jaffe, Richard Simon, Richard D Klausner, John Powell, Patricia L Duffey, Dan L Longo, Timothy C Greiner, Dennis D Weisenburger, Warren G Sanger, Bhavana J Dave, James C Lynch, Julie Vose, James O Armitage, Emilio Montserrat, Armando López-Guillermo, Thomas M Grogan, Thomas P Miller, Michel LeBlanc, German Ott, Stein Kvaloy, Jan Delabie, Harald Holte, Peter Krajci, Trond Stokke, Louis M Staudt, and Lymphoma/Leukemia Molecular Profiling 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.

Susan G Fisher and Richard I Fisher. The epidemiology of non-hodgkin's lymphoma. *Oncogene*, 23(38):6524–6534, Aug 2004. doi: 10.1038/sj.onc.1207843. URL `http://dx.doi.org/10.1038/sj.onc.1207843`.

A predictive model for aggressive non-hodgkin's lymphoma. the international non-hodgkin's lymphoma prognostic factors project. *N Engl J Med*, 329 (14):987–994, Sep 1993.

B. H. Ye, G. Cattoretti, Q. Shen, J. Zhang, N. Hawe, R. de Waard, C. Leung, M. Nouri-Shirazi, A. Orazi, R. S. Chaganti, P. Rothman, A. M. Stall, P. P. Pandolfi, and R. Dalla-Favera. The bcl-6 proto-oncogene controls germinal-centre formation and th2-type inflammation. *Nat Genet*, 16(2):161–170, Jun 1997. doi: 10.1038/ng0697-161. URL `http://dx.doi.org/10.1038/ng0697-161`.

A. L. Dent, A. L. Shaffer, X. Yu, D. Allman, and L. M. Staudt. Control of inflammation, cytokine expression, and germinal center formation by bcl-6. *Science*, 276(5312):589–592, Apr 1997.

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. doi: 10.1038/35000501. URL `http://dx.doi.org/10.1038/35000501`.

George Wright, Bruce Tan, Andreas Rosenwald, Elaine H Hurt, Adrian Wiestner, and Louis 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):9991–9996, Aug 2003. doi: 10.1073/pnas.1732008100. URL `http://dx.doi.org/10.1073/pnas.1732008100`.

A. L. Shaffer, X. Yu, Y. He, J. Boldrick, E. P. Chan, and L. M. Staudt. Bcl-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity*, 13(2):199–212, Aug 2000.

Laura Pasqualucci, Oxana Bereschenko, Huifeng Niu, Ulf Klein, Katia Basso, Roberta Guglielmino, Giorgio Cattoretti, and Riccardo Dalla-Favera. Molecular pathogenesis of non-hodgkin's lymphoma: the role of bcl-6. *Leuk Lymphoma*, 44 Suppl 3:S5–12, 2003.

I. S. Lossos, A. A. Alizadeh, M. B. Eisen, W. C. Chan, P. O. Brown, D. Botstein, L. M. Staudt, and R. Levy. Ongoing immunoglobulin somatic mutation in germinal center b cell-like but not in activated b cell-like diffuse large cell lymphomas. *Proc Natl Acad Sci U S A*, 97 (18):10209–10213, Aug 2000. doi: 10.1073/pnas.180316097. URL `http://dx.doi.org/10.1073/pnas.180316097`.

Silvia Bea, Andreas Zettl, George Wright, Itziar Salaverria, Philipp Jehn, Victor Moreno, Christof Burek, German Ott, Xavier Puig, Liming Yang, Armando Lopez-Guillermo, Wing C Chan, Timothy C Greiner, Dennis D Weisenburger, James O Armitage, Randy D Gascoyne, Joseph M Connors, Thomas M Grogan, Rita Braziel, Richard I Fisher, Erlend B Smeland, Stein Kvaloy, Harald Holte, Jan Delabie, Richard Simon, John Powell, Wyndham H Wilson, Elaine S Jaffe, Emili Montserrat, Hans-Konrad Muller-Hermelink, Louis M Staudt, Elias Campo, Andreas Rosenwald, and Lymphoma/Leukemia Molecular Profiling 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.

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.

Steven H Swerdlow and Michael E Williams. From centrocytic to mantle cell lymphoma: a clinicopathologic and molecular review of 3 decades. *Hum Pathol*, 33(1):7–20, Jan 2002.

Christian Bogner, Christian Peschel, and Thomas Decker. Targeting the proteasome in mantle cell lymphoma: a promising therapeutic approach. *Leuk Lymphoma*, 47(2):195–205, Feb 2006. doi: 10.1080/10428190500144490. URL `http://dx.doi.org/10.1080/10428190500144490`.

Andreas Rosenwald, George Wright, Adrian Wiestner, Wing C Chan, Joseph M Connors, Elias Campo, Randy D Gascoyne, Thomas M Grogan, H. Konrad Muller-Hermelink, Erlend B Smeland, Michael Chiorazzi, Jena M Giltnane, Elaine M Hurt, Hong Zhao, Lauren Averett, Sarah Henrickson, Liming Yang, John Powell, Wyndham H Wilson, Elaine S Jaffe, Richard Simon, Richard D Klausner, Emilio Montserrat, Francesc Bosch, Timothy C Greiner, Dennis D Weisenburger, Warren G Sanger, Bhavana J Dave, James C Lynch, Julie Vose, James O Armitage, Richard I Fisher, Thomas P Miller, Michael LeBlanc, German Ott, Stein Kvaloy, Harald Holte, Jan Delabie, and Louis M Staudt. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2):185–197, Feb 2003.

R. Rimokh, F. Berger, C. Bastard, B. Klein, M. French, E. Archimbaud, J. P. Rouault, B. Santa Lucia, L. Duret, and M. Vuillaume. Rearrangement of ccnd1 (bcl1/prad1) 3' untranslated region in mantle-cell lymphomas and t(11q13)-associated leukemias. *Blood*, 83(12):3689–3696, Jun 1994.

D. E. Lebwohl, R. Muise-Helmericks, L. Sepp-Lorenzino, S. Serve, M. Timaul, R. Bol, P. Borgen, and N. Rosen. A truncated cyclin d1 gene encodes a stable mrna in a human breast cancer cell line. *Oncogene*, 9(7):1925–1929, Jul 1994.

Charles J Sherr and Frank McCormick. The rb and p53 pathways in cancer. *Cancer Cell*, 2(2):103–112, Aug 2002.

M. Pinyol, L. Hernandez, M. Cazorla, M. Balbín, P. Jares, P. L. Fernandez, E. Montserrat, A. Cardesa, C. Lopez-Otín, and E. Campo. Deletions and loss of expression of p16ink4a and p21waf1 genes are associated with

aggressive variants of mantle cell lymphomas. *Blood*, 89(1):272–280, Jan 1997.

M. Pinyol, F. Cobo, S. Bea, P. Jares, I. Nayach, P. L. Fernandez, E. Montserrat, A. Cardesa, and E. Campo. p16(ink4a) gene inactivation by deletions, mutations, and hypermethylation is associated with transformed and aggressive variants of non-hodgkin's lymphomas. *Blood*, 91(8):2977–2984, Apr 1998.

M. Pinyol, L. Hernández, A. Martínez, F. Cobo, S. Hernández, S. Beà, A. López-Guillermo, I. Nayach, A. Palacín, A. Nadal, P. L. Fernández, E. Montserrat, A. Cardesa, and E. Campo. Ink4a/arf locus alterations in human non-hodgkin's lymphomas mainly occur in tumors with wild-type p53 gene. *Am J Pathol*, 156(6):1987–1996, Jun 2000.

G. A. Velders, J. C. Kluin-Nelemans, C. J. De Boer, J. Hermans, E. M. Noordijk, E. Schuuring, M. H. Kramer, W. A. Van Deijk, J. B. Rahder, P. M. Kluin, and J. H. Van Krieken. Mantle-cell lymphoma: a population-based clinical study. *J Clin Oncol*, 14(4):1269–1274, Apr 1996.

L. H. Argatoff, J. M. Connors, R. J. Klasa, D. E. Horsman, and R. D. Gascoyne. Mantle cell lymphoma: a clinicopathologic study of 80 cases. *Blood*, 89(6):2067–2078, Mar 1997.

F. Bosch, A. López-Guillermo, E. Campo, J. M. Ribera, E. Conde, M. A. Piris, T. Vallespí, S. Woessner, and E. Montserrat. Mantle cell lymphoma: presenting features, response to therapy, and prognostic factors. *Cancer*, 82(3):567–575, Feb 1998.

Riikka Räty, Kaarle Franssila, Heikki Joensuu, Lasse Teerenhovi, and Erkki Elonen. Ki-67 expression level, histological subtype, and the international prognostic index as outcome predictors in mantle cell lymphoma. *Eur J Haematol*, 69(1):11–20, Jul 2002.

A. Alizadeh, M. Eisen, R. E. Davis, C. Ma, H. Sabet, T. Tran, J. I. Powell, L. Yang, G. E. Marti, D. T. Moore, J. R. Hudson, W. C. Chan, T. Greiner, D. Weisenburger, J. O. Armitage, I. Lossos, R. Levy, D. Botstein, P. O. Brown, and L. M. 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*, 64:71–78, 1999.

Alexander Kohlmann, Claudia Schoch, Susanne Schnittger, Martin Dugas, Wolfgang Hiddemann, Wolfgang Kern, and Torsten Haferlach. Molecular

characterization of acute leukemias by use of microarray technology. *Genes Chromosomes Cancer*, 37(4):396–405, Aug 2003. doi: 10.1002/gcc.10225. URL http://dx.doi.org/10.1002/gcc.10225.

Markus Weniger, Julia C Engelmann, and JÃűrg Schultz. Genome expression pathway analysis tool–analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics*, 8:179, 2007. doi: 10.1186/1471-2105-8-179. URL http://dx.doi.org/10.1186/1471-2105-8-179.

W3C. The extensible markup language., a. URL http://www.w3.org/XML/.

Jocelyn Kaiser. Proteomics. public-private group maps out initiatives. *Science*, 296(5569):827, May 2002. doi: 10.1126/science.296.5569.827. URL http://dx.doi.org/10.1126/science.296.5569.827.

Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jérôme Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K. Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The hupo psi's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, Feb 2004a. doi: t926. URL http://dx.doi.org/t926.

Gary D Bader, Doron Betel, and Christopher W V Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, Jan 2003.

Don Gilbert. Biomolecular interaction network database. *Brief Bioinform*, 6(2):194–198, Jun 2005.

Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004. doi: 10.1093/nar/gkh086. URL http://dx.doi.org/10.1093/nar/gkh086.

Suraj Peri, J. Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan,

Babylakshmi Muthusamy, T. K B Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Zhixing Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V. Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, Oct 2003. doi: 10.1101/gr.1680803. URL `http://dx.doi.org/10.1101/gr.1680803`.

Guenter Stoesser, Wendy Baker, Alexandra van den Broek, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Francesco Nardone, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara, and Robert Vaughan. The embl nucleotide sequence database: major new developments. *Nucleic Acids Res*, 31(1):17–22, Jan 2003.

Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. Intact: an open source molecular interaction database. *Nucleic Acids Res*, 32 (Database issue):D452–D455, Jan 2004b. doi: 10.1093/nar/gkh052. URL `http://dx.doi.org/10.1093/nar/gkh052`.

Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: a molecular interaction database. *FEBS Lett*, 513(1):135–140, Feb 2002.

Andrew Chatr-aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. Mint: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–D574, Jan 2007. doi: 10.1093/nar/gkl950. URL `http://dx.doi.org/10.1093/nar/gkl950`.

W3C. The extensible stylesheet language family., b. URL `http://www.w3.org/Style/XSL/`.

W3C. The publishing language of the world wide web., c. URL http://www.w3.org/html/.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL http://www.R-project.org. ISBN 3-900051-07-0.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

J. Oksanen, R. Kindt, P. Legendre, and R. B. O'Hara. *vegan: Community Ecology Package*, 2007. URL http://cran.r-project.org/. R package version 1.8-4.

C. J. F. Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179, 1986.

P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, 2nd edition, 1998.

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–S114, 2001.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. doi: 10.1186/gb-2004-5-10-r80. URL http://dx.doi.org/10.1186/gb-2004-5-10-r80.

E. Grosse W.S. Cleveland and W.M. Shyu. *Local regression models*, chapter 8. Wadsworth & Brooks/Cole, 1992.

Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. *Normalization for cDNA microarray data*, volume 4266, pages 141–152. Proceedings of SPIE, 2001.

Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, Feb 2002.

Gordon K Smyth and Terry Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–273, Dec 2003.

Gordon K Smyth. *Limma: linear models for microarray data*, pages 397–420. Springer, New York, 2005.

Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):Article3, 2004. doi: 10.2202/1544-6115.1027. URL `http://dx.doi.org/10.2202/1544-6115.1027`.

P. Andersen and R. Gill. Cox's regression model for counting processes, a large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

T Therneau, P Grambsch, and T Fleming. Martingale based residuals for survival models. *Biometrika*, March 1990.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002. doi: 10.1073/pnas.082099299. URL `http://dx.doi.org/10.1073/pnas.082099299`.

Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue): D433–D437, Jan 2005. doi: 10.1093/nar/gki005. URL `http://dx.doi.org/10.1093/nar/gki005`.

T. J P Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox,

V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617, Jan 2007. doi: 10.1093/nar/gkl996. URL `http://dx.doi.org/10.1093/nar/gkl996`.

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 (Methodological)*, 57:125–133, 1995.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:8083, 1945.

David F. Bauer. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67:687–690, 1972.

Myles Hollander and Douglas A. Wolfe. *Nonparametric statistical inference*. John Wiley & Sons, New York, second edition, 1999.

N. Sethi, M. C. Monteagudo, D. Koshland, E. Hogan, and D. J. Burke. The cdc20 gene product of saccharomyces cerevisiae, a beta-transducin homolog, is required for a subset of microtubule-dependent cellular processes. *Mol Cell Biol*, 11(11):5592–5602, Nov 1991.

C. Norbury and P. Nurse. Cyclins and cell cycle control. *Curr Biol*, 1(1): 23–24, Feb 1991.

C. Norbury and P. Nurse. Animal cell cycles and their control. *Annu Rev Biochem*, 61:441–470, 1992. doi: 10.1146/annurev.bi.61.070192.002301. URL `http://dx.doi.org/10.1146/annurev.bi.61.070192.002301`.

H. A. Jinnah, L. De Gregorio, J. C. Harris, W. L. Nyhan, and J. P. O'Neill. The spectrum of inherited mutations causing hprt deficiency: 75 new cases and a review of 196 previously reported cases. *Mutat Res*, 463(3):309–326, Oct 2000.

T. J. Yen, G. Li, B. T. Schaar, I. Szilak, and D. W. Cleveland. Cenp-e is a putative kinetochore motor that accumulates just before mitosis. *Nature*, 359(6395):536–539, Oct 1992. doi: 10.1038/359536a0. URL `http://dx.doi.org/10.1038/359536a0`.

G. Ambrosini, C. Adida, and D. C. Altieri. A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma. *Nat Med*, 3(8):917–921, Aug 1997.

Victoria A Beardmore, Leena J Ahonen, Gary J Gorbsky, and Marko J Kallio. Survivin dynamics increases at centromeres during g2/m phase transition and is regulated by microtubule-attachment and aurora b kinase activity. *J Cell Sci*, 117(Pt 18):4033–4042, Aug 2004. doi: 10.1242/jcs.01242. URL http://dx.doi.org/10.1242/jcs.01242.

Jacquelyn Bond, Emma Roberts, Ganesh H Mochida, Daniel J Hampshire, Sheila Scott, Jonathan M Askham, Kelly Springell, Meera Mahadevan, Yanick J Crow, Alexander F Markham, Christopher A Walsh, and C. Geoffrey Woods. Aspm is a major determinant of cerebral cortical size. *Nat Genet*, 32(2):316–320, Oct 2002. doi: 10.1038/ng995. URL http://dx.doi.org/10.1038/ng995.

F. Müeller-Pillasch, U. Lacher, C. Wallrapp, A. Micha, F. Zimmerhackl, H. Hameister, G. Varga, H. Friess, M. Büchler, H. G. Beger, M. R. Vila, G. Adler, and T. M. Gress. Cloning of a gene highly overexpressed in cancer coding for a novel kh-domain containing protein. *Oncogene*, 14 (22):2729–2733, Jun 1997. doi: 10.1038/sj.onc.1201110. URL http://dx.doi.org/10.1038/sj.onc.1201110.

D. Monk, L. Bentley, C. Beechey, M. Hitchins, J. Peters, M. A. Preece, P. Stanier, and G. E. Moore. Characterisation of the growth regulating gene imp3, a candidate for silver-russell syndrome. *J Med Genet*, 39(8): 575–581, Aug 2002.

J. Nielsen, J. Christiansen, J. Lykke-Andersen, A. H. Johnsen, U. M. Wewer, and F. C. Nielsen. A family of insulin-like growth factor ii mrna-binding proteins represses translation in late development. *Mol Cell Biol*, 19(2): 1262–1270, Feb 1999.

Eiman Aleem, Hiroaki Kiyokawa, and Philipp Kaldis. Cdc2-cyclin e complexes regulate the g1/s phase transition. *Nat Cell Biol*, 7(8):831–836, Aug 2005. doi: 10.1038/ncb1284. URL http://dx.doi.org/10.1038/ncb1284.

Marcos Malumbres and Mariano Barbacid. Cell cycle kinases in cancer. *Curr Opin Genet Dev*, 17(1):60–65, Feb 2007. doi: 10.1016/j.gde.2006.12.008. URL http://dx.doi.org/10.1016/j.gde.2006.12.008.

K. A. Schafer. The cell cycle: a review. *Vet Pathol*, 35(6):461–478, Nov 1998.

Michael A Lampson, Kishore Renduchitala, Alexey Khodjakov, and Tarun M Kapoor. Correcting improper chromosome-spindle attachments during cell

division. *Nat Cell Biol*, 6(3):232–237, Mar 2004. doi: 10.1038/ncb1102. URL http://dx.doi.org/10.1038/ncb1102.

Zhanyun Tang, Hongjun Shu, Dilhan Oncel, She Chen, and Hongtao Yu. Phosphorylation of cdc20 by bub1 provides a catalytic mechanism for apc/c inhibition by the spindle checkpoint. *Mol Cell*, 16(3):387–397, Nov 2004. doi: 10.1016/j.molcel.2004.09.031. URL http://dx.doi.org/10.1016/j.molcel.2004.09.031.

Giovanni Maga and Ulrich Hubscher. Proliferating cell nuclear antigen (pcna): a dancer with many partners. *J Cell Sci*, 116(Pt 15):3051–3060, Aug 2003. doi: 10.1242/jcs.00653. URL http://dx.doi.org/10.1242/jcs.00653.

Meredith E Crosby and Alexandru Almasan. Opposing roles of e2fs in cell proliferation and death. *Cancer Biol Ther*, 3(12):1208–1211, Dec 2004.

B. Lapeyre, H. Bourbon, and F. Amalric. Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence. *Proc Natl Acad Sci U S A*, 84(6):1472–1476, Mar 1987.

M. Derenzini, V. Sirri, D. Trerè, and R. L. Ochs. The quantity of nucleolar proteins nucleolin and protein b23 is related to cell doubling time in human cancer cells. *Lab Invest*, 73(4):497–502, Oct 1995.

M. Srivastava and H. B. Pollard. Molecular dissection of nucleolin's role in growth and cell proliferation: new insights. *FASEB J*, 13(14):1911–1922, Nov 1999.

Edgar Grinstein, Ying Shan, Leonid Karawajew, Peter J F Snijders, Chris J L M Meijer, Hans-Dieter Royer, and Peter Wernet. Cell cycle-controlled interaction of nucleolin with the retinoblastoma protein and cancerous cell transformation. *J Biol Chem*, 281(31):22223–22235, Aug 2006. doi: 10.1074/jbc.M513335200. URL http://dx.doi.org/10.1074/jbc.M513335200.

Karin Milde-Langosch. The fos family of transcription factors and their role in tumourigenesis. *Eur J Cancer*, 41(16):2449–2461, Nov 2005. doi: 10.1016/j.ejca.2005.08.008. URL http://dx.doi.org/10.1016/j.ejca.2005.08.008.

M. Hartl, A. G. Bader, and K. Bister. Molecular targets of the oncogenic transcription factor jun. *Curr Cancer Drug Targets*, 3(1):41–55, Feb 2003.

Carsten Weiss and Dirk Bohmann. Deregulated repression of c-jun provides a potential link to its role in tumorigenesis. *Cell Cycle*, 3(2):111–113, Feb 2004.

R. N. Eisenman. Deconstructing myc. *Genes Dev*, 15(16):2023–2030, Aug 2001. doi: 10.1101/gad928101. URL http://dx.doi.org/10.1101/gad928101.

K. B. Marcu, S. A. Bossone, and A. J. Patel. myc function and regulation. *Annu Rev Biochem*, 61:809–860, 1992. doi: 10.1146/annurev.bi.61.070192.004113. URL http://dx.doi.org/10.1146/annurev.bi.61.070192.004113.

Stella Pelengaris, Mike Khan, and Gerard Evan. c-myc: more than just a matter of life and death. *Nat Rev Cancer*, 2(10):764–776, Oct 2002. doi: 10.1038/nrc904. URL http://dx.doi.org/10.1038/nrc904.

J. Golay, G. Cusmano, and M. Introna. Independent regulation of c-myc, b-myb, and c-myb gene expression by inducers and inhibitors of proliferation in human b lymphocytes. *J Immunol*, 149(1):300–308, Jul 1992.

A. Sala and R. Watson. B-myb protein in cellular proliferation, transcription control, and cancer: latest developments. *J Cell Physiol*, 179(3):245–250, Jun 1999. doi: 3.0.CO;2-H. URL http://dx.doi.org/3.0.CO;2-H.

S. Horstmann, S. Ferrari, and K. H. Klempnauer. Regulation of b-myb activity by cyclin d1. *Oncogene*, 19(2):298–306, Jan 2000. doi: 10.1038/sj.onc.1203302. URL http://dx.doi.org/10.1038/sj.onc.1203302.

V. Cesi, B. Tanno, R. Vitali, C. Mancini, M. L. Giuffrida, B. Calabretta, and G. Raschellà. Cyclin d1-dependent regulation of b-myb activity in early stages of neuroblastoma differentiation. *Cell Death Differ*, 9(11):1232–1239, Nov 2002. doi: 10.1038/sj.cdd.4401103. URL http://dx.doi.org/10.1038/sj.cdd.4401103.

Jianze Li and Amy S Lee. Stress induction of grp78/bip and its role in cancer. *Curr Mol Med*, 6(1):45–54, Feb 2006.

Bjarki Stefansson and David L Brautigan. Protein phosphatase 6 subunit with conserved sit4-associated protein domain targets ikappabepsilon. *J Biol Chem*, 281(32):22624–22634, Aug 2006. doi: 10.1074/jbc.M601772200. URL http://dx.doi.org/10.1074/jbc.M601772200.

K. Monica, N. Galili, J. Nourse, D. Saltman, and M. L. Cleary. Pbx2 and pbx3, new homeobox genes with extensive homology to the human proto-oncogene pbx1. *Mol Cell Biol*, 11(12):6149–6157, Dec 1991.

P. S. Knoepfler, D. B. Sykes, M. Pasillas, and M. P. Kamps. Hoxb8 requires its pbx-interaction motif to block differentiation of primary myeloid progenitors and of most cell line models of myeloid differentiation. *Oncogene*, 20(39):5440–5448, Sep 2001. doi: 10.1038/sj.onc.1204710. URL `http://dx.doi.org/10.1038/sj.onc.1204710`.

R. Michael Garavito and Anne M Mulichak. The structure of mammalian cyclooxygenases. *Annu Rev Biophys Biomol Struct*, 32:183–206, 2003. doi: 10.1146/annurev.biophys.32.110601.141906. URL `http://dx.doi.org/10.1146/annurev.biophys.32.110601.141906`.

Frederick W Wiese, Patricia A Thompson, James Warneke, Janine Einspahr, David S Alberts, and Fred F Kadlubar. Variation in cyclooxygenase expression levels within the colorectum. *Mol Carcinog*, 37(1):25–31, May 2003. doi: 10.1002/mc.10115. URL `http://dx.doi.org/10.1002/mc.10115`.

D. L. DeWitt. Prostaglandin endoperoxide synthase: regulation of enzyme expression. *Biochim Biophys Acta*, 1083(2):121–134, May 1991.

T. Hla, A. Ristimäki, S. Appleby, and J. G. Barriocanal. Cyclooxygenase gene expression in inflammation and angiogenesis. *Ann N Y Acad Sci*, 696:197–204, Nov 1993.

H. R. Herschman. Regulation of prostaglandin synthase-1 and prostaglandin synthase-2. *Cancer Metastasis Rev*, 13(3-4):241–256, Dec 1994.

Isabel Wittke, Ruprecht Wiedemeyer, Andrea Pillmann, Larissa Savelyeva, Frank Westermann, and Manfred Schwab. Neuroblastoma-derived sulfhydryl oxidase, a new member of the sulfhydryl oxidase/quiescin6 family, regulates sensitization to interferon gamma-induced cell death in human neuroblastoma cells. *Cancer Res*, 63(22):7742–7752, Nov 2003.

Izidore S Lossos, Debra K Czerwinski, Ash A Alizadeh, Mark A Wechser, Rob Tibshirani, David Botstein, and Ronald Levy. Prediction of survival in diffuse large-b-cell lymphoma based on the expression of six genes. *N Engl J Med*, 350(18):1828–1837, Apr 2004. doi: 10.1056/NEJMoa032520. URL `http://dx.doi.org/10.1056/NEJMoa032520`.

Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo C T Aguiar, Michelle Gaasenbeek, Michael Angelo,

Michael Reich, Geraldine S Pinkus, Tane S Ray, Margaret A Koval, Kim W Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S Neuberg, Eric S Lander, Jon C Aster, and Todd R Golub. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, 8(1):68–74, Jan 2002. doi: 10.1038/nm0102-68. URL `http://dx.doi.org/10.1038/nm0102-68`.

Christine P Hans, Dennis D Weisenburger, Timothy C Greiner, Randy D Gascoyne, Jan Delabie, German Ott, H. Konrad Müller-Hermelink, Elias Campo, Rita M Braziel, Elaine S Jaffe, Zenggang Pan, Pedro Farinha, Lynette M Smith, Brunangelo Falini, Alison H Banham, Andreas Rosenwald, Louis M Staudt, Joseph M Connors, James O Armitage, and Wing C Chan. Confirmation of the molecular classification of diffuse large b-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1):275–282, Jan 2004. doi: 10.1182/blood-2003-05-1545. URL `http://dx.doi.org/10.1182/blood-2003-05-1545`.

Stefano Monti, Kerry J Savage, Jeffery L Kutok, Friedrich Feuerhake, Paul Kurtin, Martin Mihm, Bingyan Wu, Laura Pasqualucci, Donna Neuberg, Ricardo C T Aguiar, Paola Dal Cin, Christine Ladd, Geraldine S Pinkus, Gilles Salles, Nancy Lee Harris, Riccardo Dalla-Favera, Thomas M Habermann, Jon C Aster, Todd R Golub, and Margaret 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. doi: 10.1182/blood-2004-07-2947. URL `http://dx.doi.org/10.1182/blood-2004-07-2947`.

J. W. Lee, N. J. Yoo, Y. H. Soung, H. S. Kim, W. S. Park, S. Y. Kim, J. H. Lee, J. Y. Park, Y. G. Cho, C. J. Kim, Y. H. Ko, S. H. Kim, S. W. Nam, J. Y. Lee, and S. H. Lee. Braf mutations in non-hodgkin's lymphoma. *Br J Cancer*, 89(10):1958–1960, Nov 2003. doi: 10.1038/sj.bjc.6601371. URL `http://dx.doi.org/10.1038/sj.bjc.6601371`.

T. G. Willis, D. M. Jadayel, M. Q. Du, H. Peng, A. R. Perry, M. Abdul-Rauf, H. Price, L. Karran, O. Majekodunmi, I. Wlodarska, L. Pan, T. Crook, R. Hamoudi, P. G. Isaacson, and M. J. Dyer. Bcl10 is involved in t(1;14)(p22;q32) of malt b cell lymphoma and mutated in multiple tumor types. *Cell*, 96(1):35–45, Jan 1999.

Jose M Polo, Tania Dell'Oso, Stella Maris Ranuncolo, Leandro Cerchietti, David Beck, Gustavo F Da Silva, Gilbert G Prive, Jonathan D Licht, and Ari Melnick. Specific peptide interference reveals bcl6 transcriptional and

oncogenic mechanisms in b-cell lymphoma cells. *Nat Med*, 10(12):1329–1335, Dec 2004. doi: 10.1038/nm1134. URL http://dx.doi.org/10.1038/nm1134.

Ulrik de Lichtenberg, Lars Juhl Jensen, Søren Brunak, and Peer Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307 (5710):724–727, Feb 2005. doi: 10.1126/science.1105103. URL http://dx.doi.org/10.1126/science.1105103.

Boris Reizis and Philip Leder. Direct induction of t lymphocyte-specific gene expression by the mammalian notch signaling pathway. *Genes Dev*, 16(3):295–300, Feb 2002. doi: 10.1101/gad.960702. URL http://dx.doi.org/10.1101/gad.960702.

Eric C Lai, Bergin Tam, and Gerald M Rubin. Pervasive regulation of drosophila notch target genes by gy-box-, brd-box-, and k-box-class micrornas. *Genes Dev*, 19(9):1067–1080, May 2005. doi: 10.1101/gad.1291905. URL http://dx.doi.org/10.1101/gad.1291905.

S. Beà, M. Ribas, J. M. Hernández, F. Bosch, M. Pinyol, L. Hernández, J. L. García, T. Flores, M. González, A. López-Guillermo, M. A. Piris, A. Cardesa, E. Montserrat, R. Miró, and E. Campo. Increased number of chromosomal imbalances and high-level dna amplifications in mantle cell lymphoma are associated with blastoid variants. *Blood*, 93(12):4365–4374, Jun 1999.

Jeannette E Allen, Rachael E Hough, John R Goepel, Sarah Bottomley, Gill A Wilson, Helen E Alcock, Margaret Baird, Paul C Lorigan, Elisabeth A Vandenberghe, Barry W Hancock, and David W Hammond. Identification of novel regions of amplification and deletion within mantle cell lymphoma dna by comparative genomic hybridization. *Br J Haematol*, 116(2):291–298, Feb 2002.

Holger Kohlhammer, Carsten Schwaenen, Swen Wessendorf, Karlheinz Holzmann, Hans A Kestler, Dirk Kienle, Thomas F E Barth, Peter MÃűller, German Ott, JÃűrg Kalla, Bernhard Radlwimmer, Armin Pscherer, Stephan Stilgenbauer, Hartmut DÃűhner, Peter Lichter, and Martin Bentz. Genomic dna-chip hybridization in t(11;14)-positive mantle cell lymphomas shows a high frequency of aberrations and allows a refined characterization of consensus regions. *Blood*, 104(3):795–801, Aug 2004. doi: 10.1182/blood-2003-12-4175. URL http://dx.doi.org/10.1182/blood-2003-12-4175.

Fanny Rubio-Moscardo, Joan Climent, Reiner Siebert, Miguel A Piris, Jose I Martín-Subero, Inga NielÃđnder, Javier Garcia-Conde, Martin J S Dyer, Maria Jose Terol, Daniel Pinkel, and Jose A Martinez-Climent. Mantle-cell lymphoma genotypes identified with cgh to bac microarrays define a leukemic subgroup of disease and predict patient outcome. *Blood*, 105 (11):4445–4454, Jun 2005. doi: 10.1182/blood-2004-10-3907. URL `http://dx.doi.org/10.1182/blood-2004-10-3907`.

Itziar Salaverria, Andreas Zettl, Sílvia Beà, Victor Moreno, Joan Valls, Elena Hartmann, German Ott, George Wright, Armando Lopez-Guillermo, Wing C Chan, Dennis D Weisenburger, Randy D Gascoyne, Thomas M Grogan, Jan Delabie, Elaine S Jaffe, Emili Montserrat, Hans-Konrad Muller-Hermelink, Louis M Staudt, Andreas Rosenwald, and Elias Campo. Specific secondary genetic alterations in mantle cell lymphoma provide prognostic information independent of the gene expression-based proliferation signature. *J Clin Oncol*, 25(10):1216–1222, Apr 2007. doi: 10.1200/JCO.2006.08.4251. URL `http://dx.doi.org/10.1200/JCO.2006.08.4251`.

I. S. Lossos, D. K. Czerwinski, M. A. Wechser, and R. Levy. Optimization of quantitative real-time rt-pcr parameters for the study of lymphoid malignancies. *Leukemia*, 17(4):789–795, Apr 2003. doi: 10.1038/sj.leu.2402880. URL `http://dx.doi.org/10.1038/sj.leu.2402880`.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003. doi: 10.1093/biostatistics/4.2.249. URL `http://dx.doi.org/10.1093/biostatistics/4.2.249`.

John Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, Dec 2002. doi: 10.1038/ng1032. URL `http://dx.doi.org/10.1038/ng1032`.

Alicia Oshlack, Dianne Emslie, Lynn Corcoran, and Gordon Smyth. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biol*, 8(1):R2, Jan 2007. doi: 10.1186/gb-2007-8-1-r2. URL `http://dx.doi.org/10.1186/gb-2007-8-1-r2`.

Lisa M Rimsza, Robin A Roberts, Thomas P Miller, Joseph M Unger, Michael LeBlanc, Rita M Braziel, Dennis D Weisenberger, Wing C Chan, H. Konrad Muller-Hermelink, Elaine S Jaffe, Randy D Gascoyne, Elias

Campo, Deborah A Fuchs, Catherine M Spier, Richard I Fisher, Jan Delabie, Andreas Rosenwald, Louis M Staudt, and Thomas M Grogan. Loss of mhc class ii gene and protein expression in diffuse large b-cell lymphoma is related to decreased tumor immunosurveillance and poor patient survival regardless of other prognostic factors: a follow-up study from the leukemia and lymphoma molecular profiling project. *Blood*, 103(11):4251–4258, Jun 2004. doi: 10.1182/blood-2003-07-2365. URL `http://dx.doi.org/10.1182/blood-2003-07-2365`.

Lexin Li. Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22(4):466–471, Feb 2006. doi: 10.1093/bioinformatics/bti824. URL `http://dx.doi.org/10.1093/bioinformatics/bti824`.

Mark R Segal. Microarray gene expression data with linked survival phenotypes: diffuse large-b-cell lymphoma revisited. *Biostatistics*, 7(2):268–285, Apr 2006. doi: 10.1093/biostatistics/kxj006. URL `http://dx.doi.org/10.1093/biostatistics/kxj006`.

Naijun Sha, Mahlet G Tadesse, and Marina Vannucci. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22(18):2262–2268, Sep 2006. doi: 10.1093/bioinformatics/btl362. URL `http://dx.doi.org/10.1093/bioinformatics/btl362`.

Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, Jul 2005. doi: 10.1093/bioinformatics/bti422. URL `http://dx.doi.org/10.1093/bioinformatics/bti422`.

Carsten Schrader, Dirk Janssen, Peter Meusers, Günter Brittinger, Jens U Siebmann, Reza Parwaresch, and Markus Tiemann. Repp86: a new prognostic marker in mantle cell lymphoma. *Eur J Haematol*, 75(6):498–504, Dec 2005. doi: 10.1111/j.1600-0609.2005.00540.x. URL `http://dx.doi.org/10.1111/j.1600-0609.2005.00540.x`.

Adrian Wiestner, Mahsa Tehrani, Michael Chiorazzi, George Wright, Federica Gibellini, Kazutaka Nakayama, Hui Liu, Andreas Rosenwald, H. Konrad Muller-Hermelink, German Ott, Wing C Chan, Timothy C Greiner, Dennis D Weisenburger, Julie M Vose, James O Armitage, Randy D Gascoyne, Joseph M Connors, Elias Campo, Emilio Montserrat, Francesc Bosch, Erlend B Smeland, Stein Kvaloy, Harald Holte, Jan Delabie, Richard I Fisher, Thomas M Grogan, Thomas P Miller, Wyndham H

Wilson, Elaine S Jaffe, and Louis M Staudt. Point mutations and genomic deletions in cyclin d1 create stable truncated mrnas that are associated with increased proliferation rate and shorter survival in mantle cell lymphoma. *Blood*, Feb 2007. doi: 10.1182/blood-2006-08-039859. URL http://dx.doi.org/10.1182/blood-2006-08-039859.

Darrell O Ricke, Shaowen Wang, Richard Cai, and Dalia Cohen. Genomic approaches to drug discovery. *Curr Opin Chem Biol*, 10(4):303–308, Aug 2006. doi: 10.1016/j.cbpa.2006.06.024. URL http://dx.doi.org/10.1016/j.cbpa.2006.06.024.

A. Krasky, A. Rohwer, J. Schroeder, and P. M. Selzer. A combined bioinformatics and chemoinformatics approach for the development of new antiparasitic drugs. *Genomics*, 89(1):36–43, Jan 2007. doi: 10.1016/j.ygeno.2006.09.008. URL http://dx.doi.org/10.1016/j.ygeno.2006.09.008.

P. O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nat Genet*, 21(1 Suppl):33–37, Jan 1999. doi: 10.1038/4462. URL http://dx.doi.org/10.1038/4462.

Hirohisa Saito. [microarray as a standard laboratory technique and as an unprecedented tool for understanding systems biology]. *Rinsho Byori*, 54 (7):732–737, Jul 2006.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

Joydeep D Chaudhuri. Genes arrayed out for you: the amazing world of microarrays. *Med Sci Monit*, 11(2):RA52–RA62, Feb 2005.

Evelyn Strauss. Arrays of hope. *Cell*, 127(4):657–659, Nov 2006. doi: 10.1016/j.cell.2006.11.005. URL http://dx.doi.org/10.1016/j.cell.2006.11.005.

Manikandan Jayapal and Alirio J Melendez. Dna microarray technology for target identification and validation. *Clin Exp Pharmacol Physiol*, 33 (5-6):496–503, 2006. doi: 10.1111/j.1440-1681.2006.04398.x. URL http://dx.doi.org/10.1111/j.1440-1681.2006.04398.x.

David L Gerhold, Roderick V Jensen, and Steven R Gullans. Better therapeutics through microarrays. *Nat Genet*, 32 Suppl:547–551, Dec 2002. doi: 10.1038/ng1042. URL http://dx.doi.org/10.1038/ng1042.

J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–427, Jun 2001. doi: 10.1038/35076576. URL `http://dx.doi.org/10.1038/35076576`.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868, Dec 1998.

K. A. Cole, D. B. Krizman, and M. R. Emmert-Buck. The genetics of cancer–a 3d model. *Nat Genet*, 21(1 Suppl):38–41, Jan 1999. doi: 10.1038/4466. URL `http://dx.doi.org/10.1038/4466`.

Chapter 8

Supplementary Material

8.1 MCL

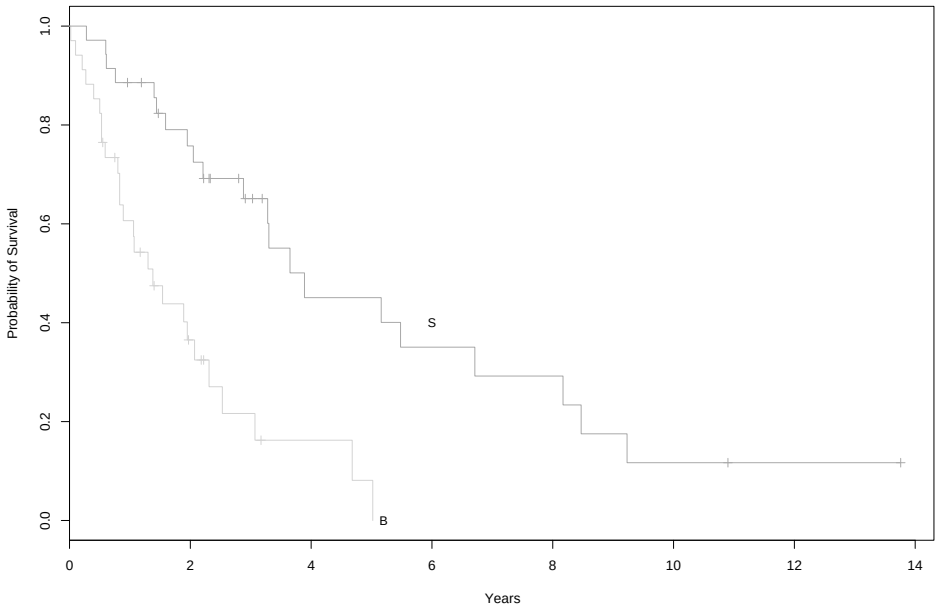


Figure 8.1: ***Kaplan Meier plot of prognosis prediction applying a set of seven genes as a molecular predictor.*** **The seven genes with the strongest influence on survival time were chosen by univariate Cox-Regression Hazard models. Applied to all patients it deliveres two distinct clearly separated risk groups. The x-axis denotes the course of time in years and the y-axis marks the probability of survival.**

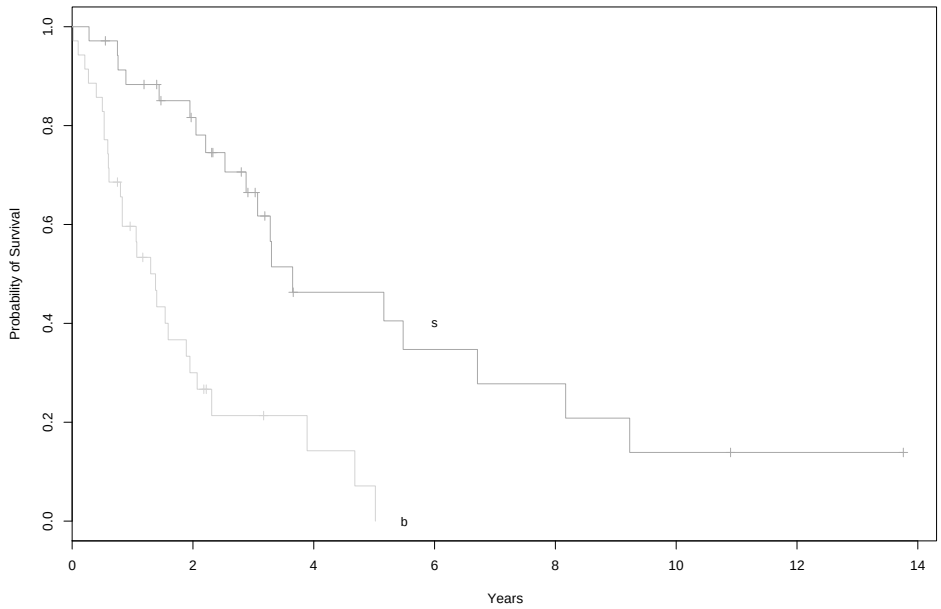


Figure 8.2: ***Kaplan Meier plot of prognosis prediction applying the genes of proliferation signature as a molecular predictor.*** **A Cox-Regression Hazard model of proliferation signature genes delivered the two distinct risk groups drawn in the Kaplan-Meier plots. The x-axis represents the time in years and the y-axis denotes the probability of survival. The predictor was applied over all patients. There is a clear difference of the survival rate between the two risk groups.**

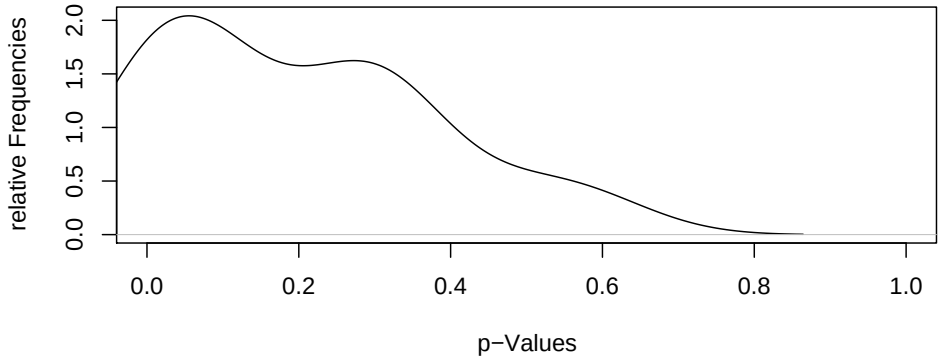


Figure 8.3: ***Density plot of p-values of the Wilcoxon test for the bands of chromosome 9.*** **The p-values of Wilcoxon test for the bands (x-axis) of chromosome 9 over the subgroups "s" and "b" are represented in their relative frequencies (y-axis). The peak of the first bands indicates that signal of the test ranges from p-value 0 to 0.1. The p-values of the first four bands 9p24, 9p23, 9p22, 9p21 vary between these limits. This affirms the proposed subgroups "s" and "b" and indicates that the first four bands have a relation to this classification.**

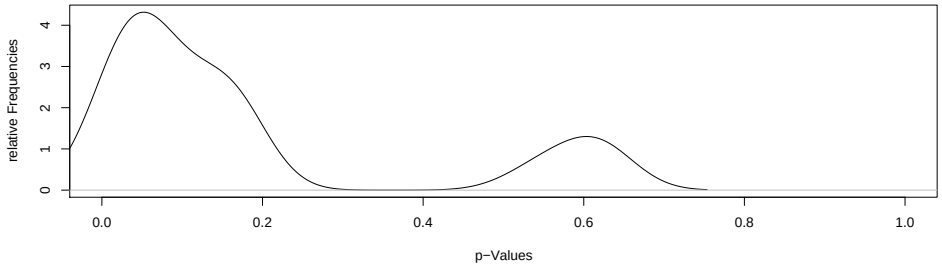


Figure 8.4: ***Density plot of p-values of the Wilcoxon test for the bands of chromosome 7.*** **The p-values from the Wilcoxon test applied on the bands of chromosome 7 are plotted against their relative frequencies. A peak occurs between the limits of 0 and 0.1. The p-values of some bands vary between these limits. These bands are the significant signal of the performed test, affirm the proposed subgroups "s" and "b" and could have a relation to this classification.**

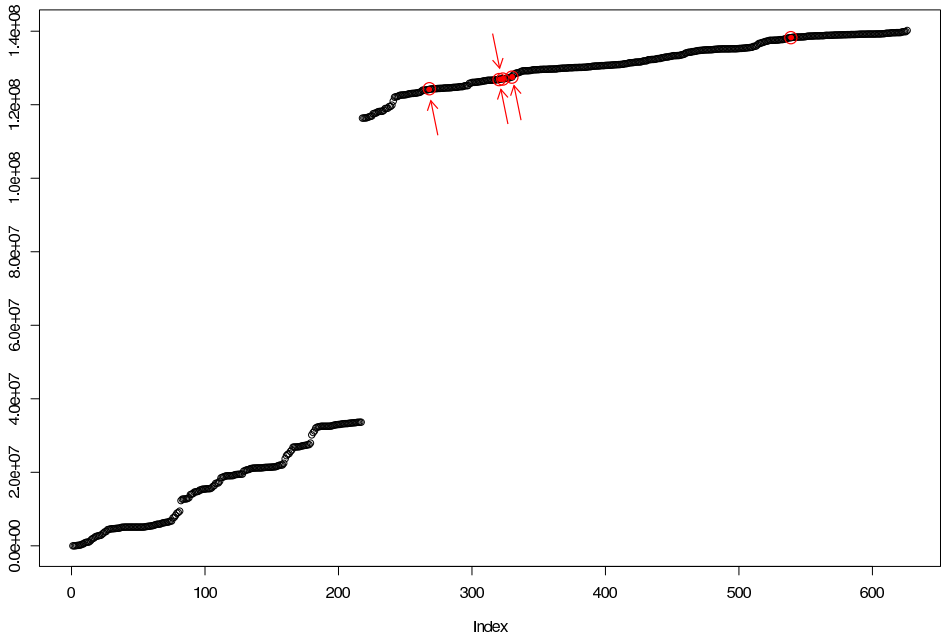


Figure 8.5: ***Plotted base pair positions of genes on Chromosome 9.*** **Here all genes, which are located on the bands 9p24, 9p21, 9q33 and 9q34 of chromosome 9 are sorted and plotted according to their starting genomic position. The positions are plotted on the y axis. The x-axis represents the genes. A moderate t-test revealed the best "s" and "b" separating genes in our dataset in these bands. Their starting points are drawn in red. Remarkably three are close to each other.**

8.2 DLBCL

A multivariate analysis is desirable in order to systematically identify spots which describe the outcome and cooperate well with each other in the Cox regression hazard model. However, this requires a huge search space of combinations to be tested. To reduce this, we considered only four spot combinations of (i) the gene spots suggested by Rosenwald et al. (2002), (ii) the 36 important genes for diffuse large B-cell lymphoma chosen by Lossos et al. (2004) and (iii) the LDH-, IDH-, and PDH gene spots (the latter to better reflect IPI-scores). Cox Regression Hazard analysis was performed on all possible four tuples of these 153 indicator spots testing 160 patients (several days of calculation time on a LINUX cluster with 20 nodes of Pentium IV CPUs). Table 8.1 shows the gene content of the ten best multivariate four-spot-predictors (the next best combinations after removing these spots is found in Table 8.2). The best multivariate four-spot combination is compact and small, but neither as good as the five spot predictor in results nor as good as the signatures from Rosenwald et al. (2002). The analysis further shows that there is a correlation with survival prediction for the clinical parameter LDH (Table 8.2), but the prediction based on this well known parameter (part of the IPI score) is even worse than the results shown in Table 8.2. In contrast (see below), the new five-spot and six-spot predictors identified by univariate analysis will be useful heuristics for diagnosis and clinic, e.g. to identify risk quartiles and subgroups (Figure 8.6).

Nr.	**Gene**
1	BCL6
2	BRAF
3	ARAF1
4	RAF1
5	RAS
6	MEK
7	MAP
8	HLA-DPα
9	HLA-DQα
10	HLA-DRα
11	HLA-DRα
12	α-Actinin
13	COL3A1
14	Connective-tissue growth factor
15	FN1
16	KIAA0233

17	PLAUR
18	E2IG3
19	NPM3
20	BMP6
21	CASP10
22	POU2AF1
23	CDKN2A
24	MYC
25	BCL2
26	FCGR2B
27	CyclinD1
28	NFKB2
29	PAX5
30	BCL10
31	CDK6
32	DDX6
33	BCL7A
34	CyclinD2
35	IL-10
36	LDH
37	IDH
38	PDH

Table 8.4: *Classical lymphoma genes.* Lymphoma associated genes were collected from literature and were also found in the data set. Furthermore, we added the metabolic enzymes "lactate dehydrogenase" (LDH), "isocitrate dehydrogenase" (IDH) and "pyruvate dehydrogenase" (PDH). The latter are represented in the data by the genes PDHB, PDHA1, IDH3A, IDH3G, IDH3B, IDH1, IDH3B, IDH3A, LDHB and LDHA.

SpotID	**Gene Name**
24376	*Centerin
17496	MYBL1
28014	MYBL1
19326	IGHM
19254	MME
33991	FOXP1

19384	MAPK10
19375	FOXP1
16049	IGHM
26454	SH3BP5
22118	KIAA0864
24787	CCND2
24787	CCND2
28979	LMO2
15914	MAPK10
19346	SH3BP5
15864	MME
19238	LMO2
30263	ASB13
19291	MYBL1
19312	NEIL1
25036	FLJ12363
26385	MME
19227	LOC96597
22122	IRF4
16886	LRMP
24480	KIAA1039
27378	LRMP
27379	LRMP
24729	IRF4
27673	LRMP
19348	*Similar to
24429	BCL6
28472	MAPK10
26516	*Similar tlone=417048
19268	BCL6
32529	@Homo sapiH08 (LOC152137) Sur_clone=232 2321
17646	BCL2

Table 8.7: ***Combined classifier for lymphoma subtypes.*** **The resulting gene list that distinguishes ABC and GCB if the PAM analysis is performed only on the 31 best spots merged with the well known lymphoma genes. Marked in grey are the 31 best spots from all twelve thousand spots compared. Remarkably, the two classical lymphoma marker genes MAPK10 and CCND2 reach a similar quality in distinguishing ABC and GCB as the best separating ones.**

SpotID	EnsembleID	Cell cycle state	Gene
24927	ENSG00000165810	85	BTNL9
33929	ENSG00000165810	85	BTNL9
26913	ENSG00000138764	72	CCNG2
24750	ENSG00000136244	80	IL6
32430	ENSG00000162783	56	IER5
24491	ENSG00000165810	85	BTNL9
30172	ENSG00000138764	72	CCNG2
24930	ENSG00000187837	69	HIST1H1C
24725	ENSG00000011007	59	TCEB3
24908	ENSG00000118515	83	SGK
30355	ENSG00000164330	84	EBF
32096	ENSG00000164330	84	EBF
31931	ENSG00000164543	18	STK17A
26081	ENSG00000180447	80	GAS1
19374	ENSG00000124762	21	CDKN1A
24969	ENSG00000164330	84	EBF
24647	ENSG00000164330	84	EBF
34708	ENSG00000118515	83	SGK
27774	ENSG00000134058	92	CDK7
26401	ENSG00000118515	83	SGK
26725	ENSG00000164330	84	EBF
28881	ENSG00000163918	52	RFC4
17786	ENSG00000102804	1	TSC22D1
24613	ENSG00000102804	1	TSC22D1
33901	ENSG00000100644	2	HIF1A
27538	ENSG00000171656	96	ETV5
27952	ENSG00000179583	76	CIITA

34557	ENSG00000052841	2	TTC17
30021	ENSG00000099953	95	MMP11
27704	ENSG00000164330	84	EBF
26992	ENSG00000102804	1	TSC22D1
26344	ENSG00000138764	72	CCNG2
24832	ENSG00000163918	52	RFC4
26080	ENSG00000163739	76	CXCL1
33329	ENSG00000179583	76	CIITA
17290	ENSG00000134058	92	CDK7
30922	ENSG00000185658	5	BRWD1
26162	ENSG00000135541	91	AHI1
34288	ENSG00000134884	48	NA
33646	ENSG00000185658	5	BRWD1
26951	ENSG00000102804	1	TSC22D1
24977	ENSG00000153936	92	HS2ST1
16661	ENSG00000123080	75	CDKN2C
25942	ENSG00000145050	49	ARMET
22163	ENSG00000169926	6	KLF13
17405	ENSG00000178573	30	MAF
27275	ENSG00000100644	2	HIF1A
30415	ENSG00000164330	84	EBF
34484	ENSG00000151150	50	ANK3
33221	ENSG00000065809	2	FAM107B
32218	ENSG00000179583	76	CIITA
29637	ENSG00000145632	99	PLK2PLK2
27939	ENSG00000179583	76	CIITA
27328	ENSG00000108984	44	MAP2K6
28792	ENSG00000099326	53	ZNF42
30725	ENSG00000175455	65	CCDC14
16736	ENSG00000136244	80	IL6
30874	ENSG00000081320	77	STK17B
28707	ENSG00000123080	75	CDKN2C
33336	ENSG00000175455	65	CCDC14
15871	ENSG00000168310	7	IRF2
28640	ENSG00000100526	0	CDKN3
28748	ENSG00000136244	80	IL6
28430	ENSG00000168310	7	IRF2
26084	ENSG00000128590	38	DNAJB9
30859	ENSG00000117650	93	NEK2
28674	ENSG00000138061	66	CYP1B1
16127	ENSG00000138061	66	CYP1B1

24868	ENSG00000012963	52	C14orf130
30508	ENSG00000081320	77	STK17B
34108	ENSG00000169926	6	KLF13
16053	ENSG00000173757	83	STAT5B
16091	ENSG00000100526	0	CDKN3
33594	ENSG00000179583	76	CIITA
32924	ENSG00000185658	5	BRWD1
32766	ENSG00000135164	74	DMTF1
16597	ENSG00000109971	0	HSPA8

Table 8.8: ***Cell cycle gene set that best distinguishes ABC and GCB subgroup.*** **The genes are annotated by their spot ID, ensemble gene-ID and their gene name. Additionally the cell cycle states are given. The latter parameter shows a strong signal in the early and late cell cycle states compared with all available cell cycle states in the data set.**

Ensemble gene ID	**Cell cycle state**	**Gene symbol**
ENSG00000011007	59	TCEB3
ENSG00000012963	52	C14orf130
ENSG00000052841	2	TTC17
ENSG00000065809	2	FAM107B
ENSG00000081320	77	STK17B
ENSG00000099326	53	ZNF42
ENSG00000099953	95	MMP11
ENSG00000100526	0	CDKN3
ENSG00000100644	2	HIF1A
ENSG00000102804	1	TSC22D1
ENSG00000108984	44	MAP2K6
ENSG00000109971	0	HSPA8
ENSG00000117650	93	NEK2
ENSG00000118515	83	SGK
ENSG00000123080	75	CDKN2C
ENSG00000124762	21	CDKN1A
ENSG00000128590	38	DNAJB9
ENSG00000134058	92	CDK7
ENSG00000134884	48	NA
ENSG00000135164	74	DMTF1

ENSG00000135541	91	AHI1
ENSG00000136244	80	IL6
ENSG00000138061	66	CYP1B1
ENSG00000138764	72	CCNG2
ENSG00000145050	49	ARMET
ENSG00000145632	99	PLK2PLK2
ENSG00000151150	50	ANK3
ENSG00000153936	92	HS2ST1
ENSG00000162783	56	IER5
ENSG00000163739	76	CXCL1
ENSG00000163918	52	RFC4
ENSG00000164330	84	EBF
ENSG00000164543	18	STK17A
ENSG00000165810	85	BTNL9
ENSG00000168310	7	IRF2
ENSG00000169926	6	KLF13
ENSG00000171656	96	ETV5
ENSG00000173757	83	STAT5B
ENSG00000175455	65	CCDC14
ENSG00000178573	30	MAF
ENSG00000179583	76	CIITA
ENSG00000180447	80	GAS1
ENSG00000185658	5	BRWD1
ENSG00000187837	69	HIST1H1C

Table 8.9: ***The cell cycle genes, which were chosen to distinguish the ABC and the GCB group.*** **The cell cycle genes annotated by their ensemble gene-ID and their gene name. Additionally the cell cycle states are annotated. The latter parameter shows a strong signal in the early and late cell cycle states compared with all available cell cycle states in the data set.**

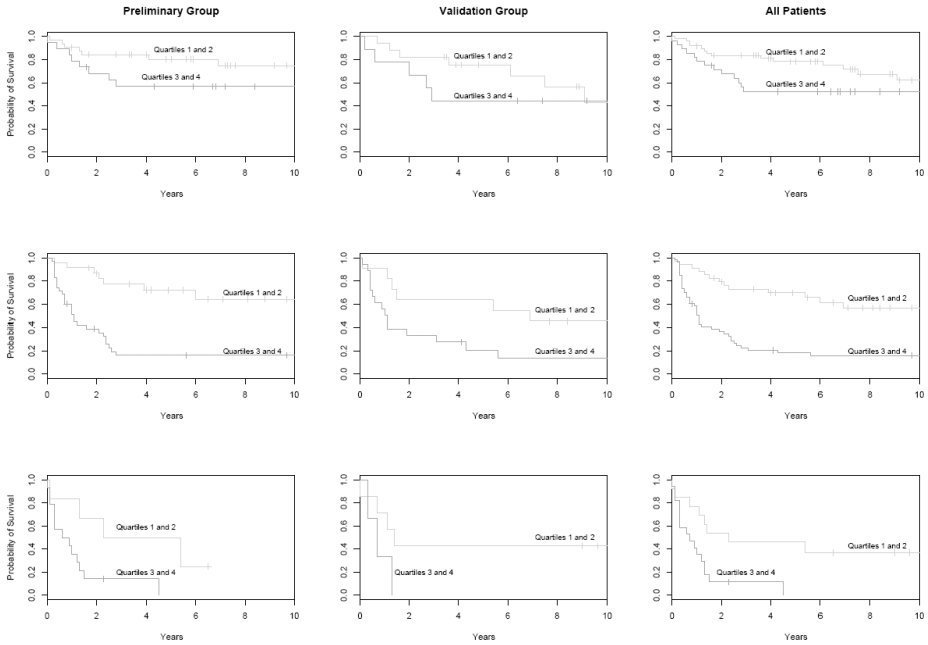


Figure 8.6: ***Kaplan Meier plots of the IPI groups.*** **The Kaplan Meier plots estimated by the molecular predictor of Rosenwald et al. applied to the new normalized gene expression data of the 240 diffuse large B-cell lymphoma patients. The plots show different groups according to their IPI risk and the training set as Training, Validation and all patients. The left column represents the training-group, the middle one the validation group and the right one all patients. The rows show the IPI risk groups. The first row shows low risk, the second one the medium risk and the last one the high risk patients. The x-axis is the time in years and the y-axis the probability of survival.**

Nr.	Multivariate	Cox	regression	hazard
1	HGAL	Germ-S	ACTa1	HLA-DRA
2	HGAL	CD54(2)	ACTa1	HLA-DRA
3	HGAL	CD54(2)	HLA-DRA(2)	ACTa1
4	HGAL	CD54(2)	HLA-DRA(3)	ACTa1
5	HGAL	ACTa1	HLA-DRA	CD54
6	HGAL	MHCIIDQa1	CD54(2)	ACTa1
7	HGAL	CD54(2)	MHCIIDRb	ACTa1
8	HGAL	Germ-S	MHCIIDRb	ACTa1
9	HGAL	Germ-S	HLA-DRA(2)	ACTa1
10	HGAL	Germ-S	HLA-DRA(3)	ACTa1

Table 8.1: ***Multivariate Cox regression hazard models.*** **A heuristic search of multivariate Cox regression hazard models revealed this 10 best fitting models. All possible multivariate Cox regression hazard models of four 4 genes from 36 important genes for diffuse large B-cell lymphoma and the metabolic genes LDH, IDH and PDH were calculated and these ten gene sets fit best. Genes are abbreviated according to GenBank nomenclature.**

Nr.	Multivariate	Cox	regression	hazard
1	CD10	IRF4	HLA-DRb5	LDH(2)
2	IRF4(2)	BCL7A	HLA-DRb5	LDH(2)
3	MYC	IRF4(2)	HLA-DRb5	LDH
4	MYC	IRF4(2)	HLA-DQa1	LDH
5	PLAU	IRF4	BCL7A	HLA-DRb5
6	IRF4	BCL7A	HLA-DRb5	LDH(2)
7	PLAU	IRF4(2)	BCL7A	HLA-DRb5
8	IRF4	BCL6	BCL7A	HLA-DRb5
9	CD10	IRF4(2)	HLA-DRb5	LDH(2)
10	MYC	IRF4(2)	HLA-DRb5	LDH(2)

Table 8.2: ***Next best multivariate Cox regression hazard models.*** **If the genes appearing in Table 8.1 are removed, and the heuristic search of multivariate Cox regression hazard models is redone, these ten models are the next best fitting. The genes are represented by their GenBank abbreviation. The metabolic marker LDH from the IPI score occurs in the four best fitting models as well as in the the majority of the models.**

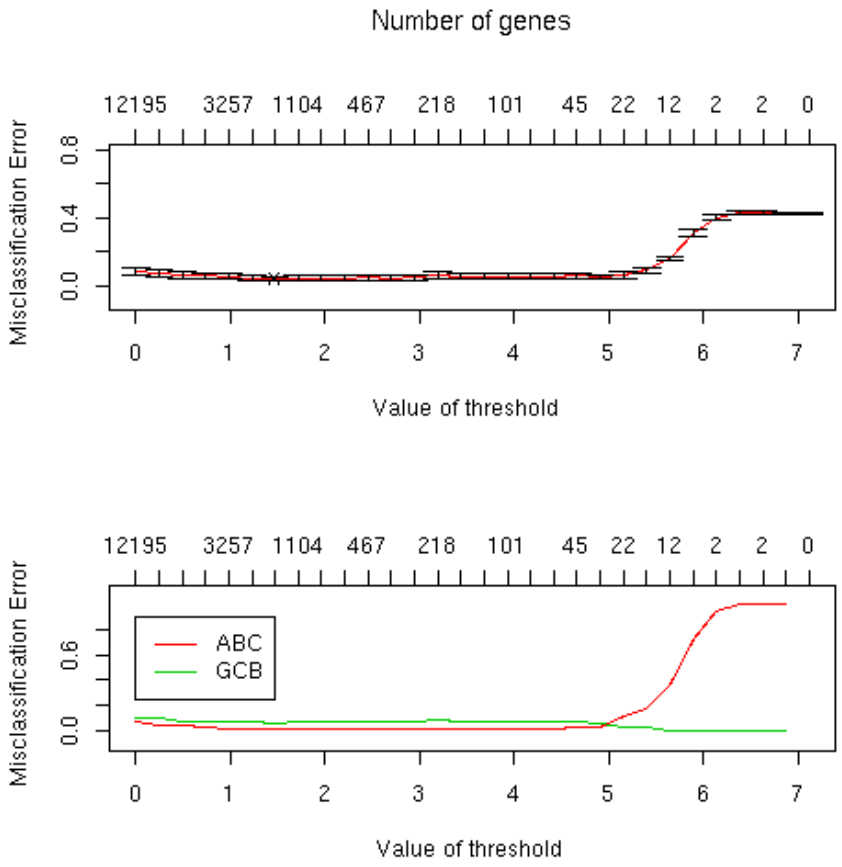


Figure 8.7: ***PAM misclassification error of the ABC and GCB subgroups over all genes of the enlarged dataset.*** **The upper plot shows the overall error while the lower one shows the subgroup specific errors. In both, the various thresholds on the lower x-axis correspond to different numbers of genes, labeled on the upper x-axis. The y-axis represents the error and ranges from 0 to 1. The good overall performance of PAM requires only few genes to decrease the error dramatically. The error rate decreases strongly between the thresholds of 6 and 5, which represent the amount of shrinkage. Hence we chose a threshold below 5 with the corresponding set of best separating genes (an optimal choice with few errors and a low number of genes). The performance for the single subgroups shows a big difference between ABC and GCB. Whereas GCB shows a good performance even with few genes, the prediction quality of ABC decreases dramatically. This indicates a complex pattern of gene expression in ABC patients which is defined in more than 15 genes.**

Type	ABC	GCB	Class Error rate
ABC	77	5	0.06097561
GCB	7	105	0.06250000
Overall	error	rate	= 0.062

Table 8.3: ***Confusion matrix of misclassification.*** **Confusion matrix of the cross validated PAM analysis with the threshold 4.906, applied to all available spots. From 80 ABC patients 5 were predicted as GCB and from 112 GCB patients 7 were predicted as ABC. The overall error rate is the mean of the two single group error rates.**

Nr.	Gene
1	FN1
2	BCL6
3	CTGF
4	BCL2
5	MAPK10
6	CCND2
7	COL3A1
8	KIAA0233
9	BCL7A

Table 8.5: ***Classical marker genes of lymphoma disease distinguish between ABC and GCB lymphoma subtype.*** **(PAM analysis; error rates for this gene set: TR:10% VAL:15.38%; F:CV:14%)**

SpotID	Gene Name
19384	MAPK10
24787	CCND2
15914	MAPK10
24429	BCL6
28472	MAPK10
19268	BCL6
16858	CCND2
17646	BCL2
16789	BCL2
19361	COL3A1
26535	BCL6
28859	BCL2
24367	BCL2
17791	FN1
16016	FN1
16732	FN1
31398	FN1
19379	FN1
27499	KIAA0233
24415	BCL7A
29222	CTGF

Table 8.6: ***Lymphochip spots of known lymphoma genes.*** **180 spots, which are known to deal with lymphoma were tested to distinguish between ABC and GCB subtype by PAM analysis. Successful genes are given in descending order (gene set error rate:TR:10% VAL:15.38%; F:CV:14%).**

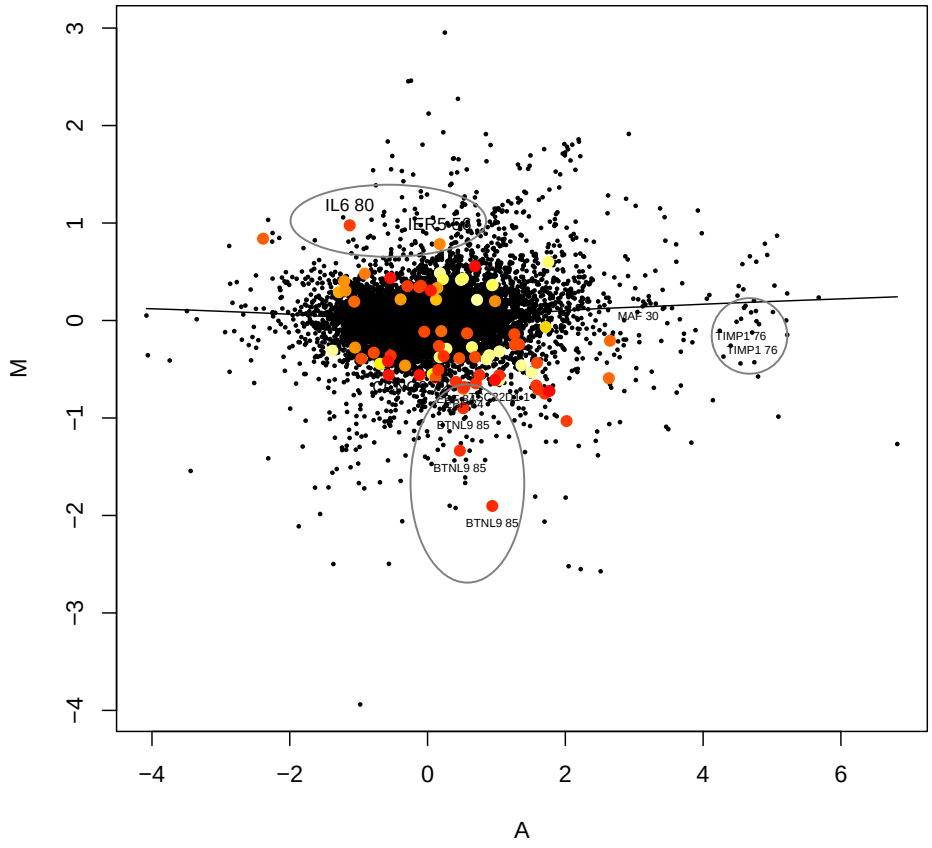


Figure 8.8: ***Cell cycle genes with extreme expression differences shown by a MA-plot of normalized gene expression values.*** **The M values on the y-axis correspond to the gene expression difference between the ABC and GCB patient medians and the A values on the x-axis correspond to the average expression of all genes in both groups. The coloured points represent the 77 cell cycle spots chosen by PAM analysis. The colour scale ranges from yellow to red, whereas yellow is annotated to cellcycle state 0 and red to state 99. Additionally some cell cycle genes show more extreme A values (circle). They are labeled with their names and their cell cycle state. Remarkably, some genes associated with a late cell cycle state cluster together regarding their gene expression values in both dimensions (ellipse). Again, late cell cycle states indicate a high difference in the M value (difference in gene expression) between the two subgroups. A locally weighted regression smoothing line (lowess) shows that systematic and random variations are well controlled by the normalization procedure: Its shape fits almost perfectly the horizontal line.**

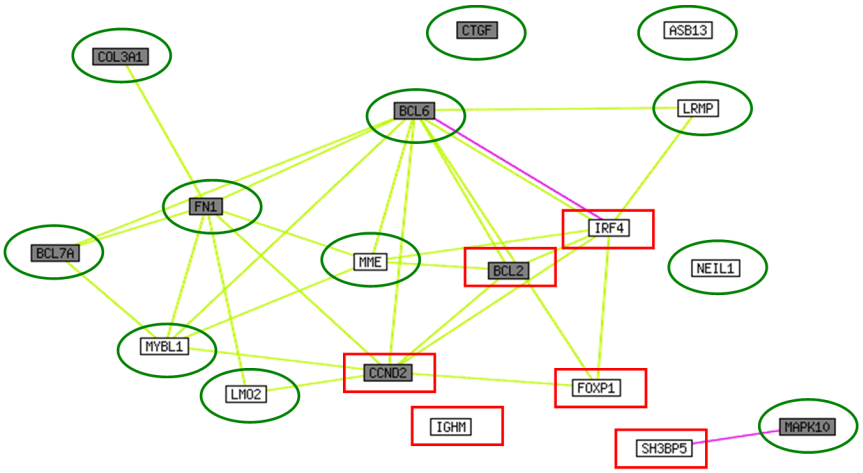


Figure 8.9: ***Regulatory network differently regulated in ABC and GCB B-cell lymphomas.*** **This figure shows the resulting network and interaction pattern with each other for the best separating genes applying data from the STRING meta-database of protein interactions. Classical lymphoma genes and best separating gene set form a tight network with the best separating genes in the center. Shown are the strongly connected network members. They consist of (i) classical lymphoma marker genes (grey boxes), and (ii) the most powerful predictive genes in the PAM analysis (white boxes). Genes which show a significant higher expression in the ABC subgroup are marked by a red rectangle. They are associated to proliferation, block of proliferation, apoptosis, differentiation and specific for immune cells, as most of the remaining ones. Green ellipses mark higher expression in GCB. The almost fully connected gene network demonstrates that both classes of genes are well participating in the interaction network according to the string meta-database. Furthermore, the string analysis shows that almost all connections between both classes – the yellow coloured edges – are based from biochemical literature (mainly Medline reports). Only the interaction of "interferon regulatory factor 4" (IRF4) and "B-cell CLL/lymphoma 6" (BCL6) is confirmed by large-scale interaction screen experiments.**

Nr.	Gene
1	MYBL1
2	*Centerin
3	FOXP1
4	LOC96597
5	SH3BP5
6	KIAA0864
7	IRF4
8	ASB13
9	*Similar to human endogenous retrovirus-4 Clone=417048
10	NEIL1
11	MME
12	IGHM
13	LMO2
14	LOC152137
15	KIAA1039
16	LRMP
17	FLJ123633
18	CCND2

Table 8.10: ***Genes which distinguish best between ABC and GCB according PAM analysis.*** **From all twelve thousand spots from the lymphoma chip, the listed genes distinguish best between ABC and GCB according to PAM analysis. The best separating genes are written on the top.**

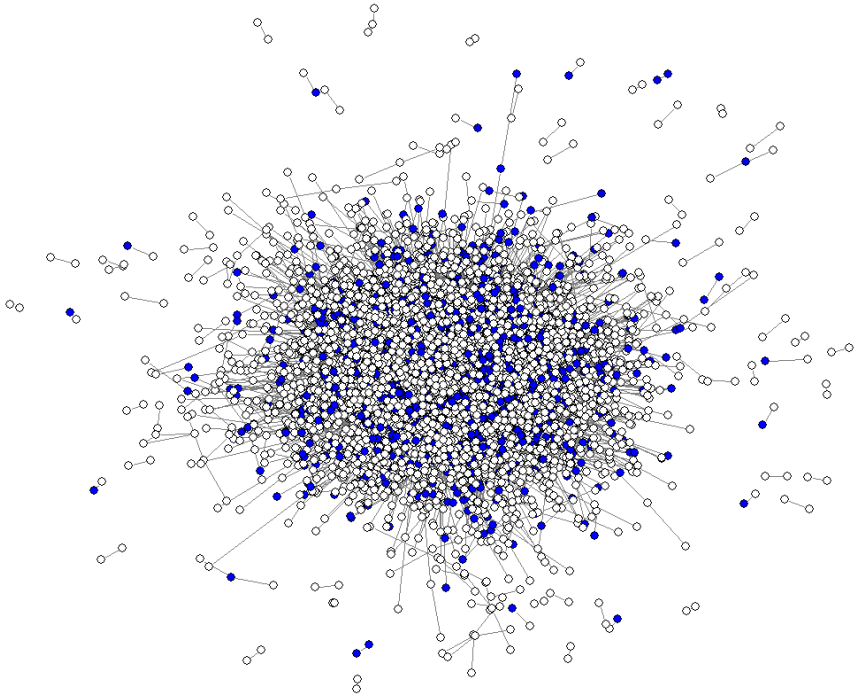


Figure 8.10: ***The Lymphochip genes in the human interactome.*** **This plot shows the human interactome as a protein interaction network. The proteins (circles) of the lymphochip are filled out. Interactions are drawn as a line. Characteristic path length and the longest one are 4.642 and 15, respectively.**

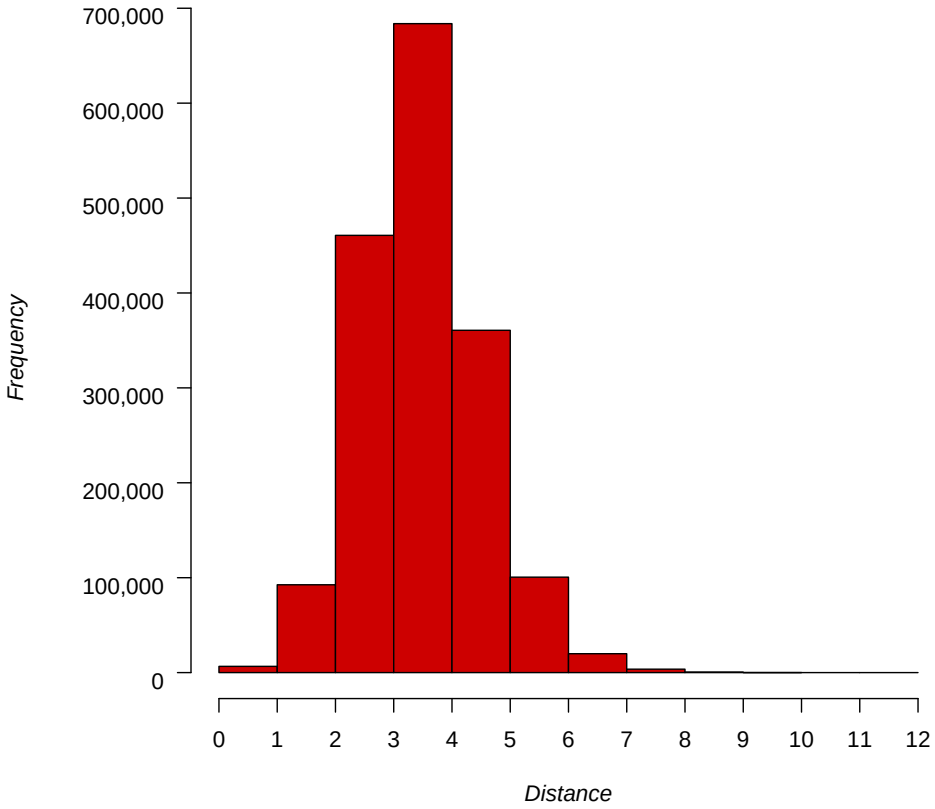


Figure 8.11: ***Histogram of the protein interaction distances.*** **The genes of the Lymphochip were mapped to the protein interaction graph in the human interactom. The histogram shows the occurring distances of these genes in the interactome. The longest distance is 11 whereas the characteristic path length is 3.985.**

Gene	ABC	GCB
ASB13	-	+
MYBL1	-	+
MME	-	+
MAPK10	-	+
LRMP	-	+
LMO2	-	+
FN1	-	+
CTGF	-	+
COL3A1	-	+
BCL6	-	+
BCL7A	-	+
NEIL1	-	+
SH3BP5	+	-
BCL2	+	-
CCND2	+	-
IRF4	+	-
IGHM	+	-
FOXP1	+	-

Table 8.11: ***Gene expression values of the main regulatory network distinguishing ABC and GCB.*** **Genes from Figure 8.9 and their gene expression values in the subgroups ABC and GCB are shown. The symbol "-" indicates a lower gene expression than "+". In this network, more genes of the more aggressive ABC type have a lower gene expression than the GCB type.**

Genes	p-value	T-value
CCND2	6.260705e-06	5.56939706
BCL6	2.490035e-02	-2.34449786
BCL2	1.843571e-03	3.43618678
IRF4	2.082072e-07	6.49044833
LMO2	3.820841e-07	-6.66162303
MAPK10	3.888633e-02	-2.15403094

Table 8.12: ***T-test result of network genes in another data set.*** **The genes from the proposed STRING-network in Figure 8.9 were used to apply a T-test between the ABC and the GCB group in the gene expression data of Shipp et al.. The authors Wright et al. found some evidence for these DLBCL groups in there. The most obvious rejection of the null hypothesis is delivered by IRF4, LMO2, CCND2, BCL2, BCL6 and MAPK10, which are also part of the predictor of Wright et al..**

GeneID	TranscriptID	Description
ENSG00000156136	ENST00000286648	Deoxycytidine kinase
ENSG00000148158	ENST00000277244	Sorting nexin family member 30
ENSG00000179388	ENST00000317216	Early growth response protein 3
ENSG00000198833	ENST00000361212	Ubiquitin-conjugating enzyme E2 J1
ENSG00000198833	ENST00000361333	Ubiquitin-conjugating enzyme E2 J1
ENSG00000065308	ENST00000182527	Translocation associated membrane protein 2
ENSG00000170584	ENST00000302764	NudC domain containing protein 2
ENSG00000074706	ENST00000265198	phosphoinositide-binding protein PIP3-E
ENSG00000134108	ENST00000256496	ADP-ribosylation factor-like 10C)

Table 8.13: ***List of potential Notch target transcripts.*** **For all genes of the Lymphochip, all available transcripts annotated in ensemble were screened for the GY, Brd and K boxes. Only these transcripts bear all three boxes, GY, Brd and K in the 3'-UTRs. They are possible candidates to be regulated by the Notch signalling pathway. Moreover, the Deoxycytidine kinase (ENSG00000156136) and the Translocation associated membrane protein 2 (ENSG00000065308) show different gene expression values between the ABC and GCB subgroups.**

Chapter 9

Curriculum vitae

Steffen Blenk
Email: steffen.blenk@biozentrum.uni-wuerzburg.de
Telephone: +49-9549-2606199

Education

1999-2004	Master of Science program in biology Major fields of study: Biochemistry, Genetics, Biotechnology Fields of specialization: Bioinformatics
1998-1999	Basic military service
1989-1998	Secondary school completed with Abitur (approximately equivalent to A-levels), Neusprachliches Franz-Ludwig-Gymnasium in Bamberg
1985-1989	Primary school, Grundschule Oberaurach in Trossenfurt
1989-1998	Neusprachliches Franz-Ludwig-Gymnasium in Bamberg Intensive Courses: Biology, Physics

Work experience

01/2007-06/2007	Scientific coworker, SFB/TR34: "Pathophysiology of Staphylococci", University of Würzburg
01/2007	Scientific coworker, "Molecular and cellular basis of behavioural plasticity using Drosophila olfactory learning", University of Würzburg

2004-2006	Scientific coworker, IZKF B-36/-(1): "Analysis of gene-expression and pathways involved in molecular pathogenesis", University of Würzburg
2002-2003	Research Assistant, SFB 554: "Mechanism and Evolution of Arthropod Behavior", University of Würzburg
2000	BOSCH Laboratory (Quality Control)
1999	BOSCH Technical Internship

Publications

DLBCL	**S. Blenk**, J. Engelmann, M. Weniger, J. Schultz, M. Dittrich, A. Rosenwald, H. K. Müller-Hermelink, T. Müller and T. Dandekar. "Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): Analysis of molecular predictors, signatures, cell cycle state and patient survival.", *Cancer Informatics*, 2007 (*in press*).
MCL	**S. Blenk**, J. Engelmann, S.Pinkert, M. Weniger, J. Schultz, A. Rosenwald, H. K. Müller-Hermelink, T. Müller and T. Dandekar. "Explorative data analysis in MCL reveals gene expression networks implicated in survival and prognosis and is supported by explorative CGH analysis.", *Cancer Informatics*, 2007 (*submitted*).

Conferences contributions

2004	German Conference of Bioinformatics 2004 - Bielefeld, October 4-6, 2004
2006	International Bioinformatics Symposium - Würzburg, July 27, 2006

Languages

German	native speaker
English	fluent
French	basic user

Würzburg, August 16, 2007

Chapter 10

Danksagung

An dieser Stelle möchte ich allen Personen, die mir während meiner Doktorarbeit geholfen haben, meinen Dank aussprechen.

Zuerst und vornehmlich möchte ich mich bei meinem Doktorvater Prof. Dr. Thomas Dandekar für die Bereitstellung des interessanten Themas bedanken. Ohne seine kompetente Betreuung, seine stets offene Tür und die vielen hilfreichen Anregungen wäre es mir nicht möglich gewesen, diese Arbeit anzufertigen. Ebenfalls möchte ich Dr. Tobias Müller für seine konstruktiven Fragen und seine Ratschläge danken. Zusammen mit Prof. Dr. Jörg Schulz betreuten sie das gesamte Projekt. Außerdem danke ich Prof. Dr. Erich Buchner für die Durchsicht der Arbeit.

Von meinen Arbeitskollen möchte ich in erster Linie meinen Kollegen, Julia Engelmann, Stefan Pinkert und Markus Weniger für die freundliche Zusammenarbeit Dank sagen. Neben diesen namentlich genannten möchte ich dem gesamtem Lehrstuhl meinen Dank aussprechen. Desweiteren gilt mein Dank dem IZKF Würzburg für die Projektförderung, stellvertretend seinem Vorstandsmitglied und Sprecher Professor Dr. med. Müller-Hermelink. Ihm und Herrn Dr. Rosenwald vielen Dank für die konstruktive Kritik und das Datenmaterial.

Außerhalb des Lehrstuhls haben mir meine Familie und meine Freunde allgemeinen und moralischen Beistand gegeben und mich unterstützt, wo es nur möglich war. Ich danke hiermit meinen lieben Eltern und Geschwistern, meiner Tante Martha, unserer Nachbarin Manuela Eberhart, meinem Vorbild Jan Hyzak, dem ich unbeschreiblich viel verdanke und meinen Freunden Simon Herbert, Jens Hisham Naim und Marcus Stüben. Ihnen allen schulde ich speziellen Dank und wünsche ihnen alles Gute für ihr weiteres Leben.

Auch denjenigen, die jetzt nicht namentlich erwähnt wurden, und mich in irgendeiner Form unterstützt haben, möchte ich recht herzlich dafür danken.

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Dissertation selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die Dissertation hat in gleicher oder ähnlicher Form in noch keinem anderen Prüfungsverfahren vorgelegen. Außer einem Diplom in Biologie habe ich bisher keine anderen akademische Grade erworben oder versucht zu erwerben.

Würzburg, August 16, 2007