Unraveling expression and DNA methylation landscapes in cancer

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades
DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

Vorgelegt von
Diplom-Biomathematikerin Lydia Hopp
geboren am 15. Juli 1987 in Werdau

Die Annahme der Dissertation wurde empfohlen von:
1. Professor Dr. Peter F. Stadler, Universität Leipzig
2. Professor Dr. Rainer Spang, Universität Regensburg

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 20.09.2017 mit dem Gesamtprädikat magna cum laude.

Acknowledgments

At this point, I would like to thank everyone who has been involved in this work:

Foremost, I am very grateful to my supervisor, Hans Binder, for remarkable support, for scientific guidance, for the opportunity to work on this thesis at the IZBI, and for providing funding necessary for this work.

Furthermore, I would like to thank Jörg Galle, Peter Stadler, and Markus Löffler for their support.

Moreover, I thank all current and former colleagues (and friends) of the IZBI and of the Chair of Bioinformatics of the University of Leipzig, especially Anke, Arsen, David, Edith, Gero, Henry, Mario, Lilit, Milan, and Volkan. It was a pleasure!

I particularly thank Corinna, Petra and Jens, who always offered kind advice and provided assistance in problems beyond science: bureaucracy, organization and IT-problems.

Additionally, I must express my very profound gratitude to my dear family -Mum, Dad, Hanna, my grandparents and Benni- for contributing indirectly to my work by continuous encouragement and putting trust in me. Thank you!

This PhD thesis was conducted within the framework of the Leipzig Interdisciplinary Research Centre for Civilization Diseases (LIFE Center, University of Leipzig). In that regard I acknowledge the financial support from the European Social Fund (ESF).

Abstract

Cancer is a complex, heterogeneous disease and associated with a pluralism of distinct molecular events occurring on multiple layers of cell activity. It is a disease of genomic regulation driven by genetic and epigenetic mechanisms. Consideration of these regulatory levels is inevitable for understanding cancer genesis and progression. Improved high-throughput techniques developed in the last decades enable a highly resolved view on these mechanisms but at the same time the technologies produce an incredible amount of molecular data. Hence it needs advances in computational methods to master the data.

In this thesis we demonstrate how to cope with high-dimensional data to characterize molecular aspects of cancer. The main aim of this thesis is to develop and to apply bioinformatics methods to unravel molecular mechanisms, with special focus on gene expression and epigenetics, underlying cancer. Therefore, we selected two cancer entities, B-cell lymphoma and glioblastoma, for a more detailed, exemplary study.

Bioinformatics methods dealing with molecular cancer data have to tackle tasks like data integration, dimension reduction, data compression and proper visualization. One effective method that fulfills the mentioned tasks is self organizing map (SOM) machine learning, a technique to 'organize' complex, multivariate data. We present an analytic framework based on SOMs that aims at characterizing single-omics landscapes, here either regarding genome wide expression or methylation, to describe the heterogeneity of cancer on the molecular level. Molecular data of each sample is presented in terms of 'individual' maps, which enable their evaluation by visual inspection. The portrayal method also realizes comprehensive downstream analysis tasks such as marker selection and clustering of co-regulated features into modules, stratification of cases into subtypes, knowledge discovery, function mining and pathway analysis. Further, we describe how to detect and to correct outlier samples.

In a novel combining approach all these analytic tasks of the single-omics SOM are embedded in a workflow to integratively analyze gene expression and DNA methylation data of unmatched patient cohorts. We showed that this approach provides detailed insights into the transcriptome and methylome landscapes of cancer. Furthermore, we developed a new inter-omics method based on SOM machine learning for the combined analysis of gene expression and DNA methylation data obtained from the same patient cohort. The method allows the visual inspection of the data landscapes of each sample on a personalized and class-related level, where the relative contribution of each of both data entities can be tuned either to focus on expression or methylation landscapes or on a combination of both.

Using the single-omics SOM approach, we studied molecular subtypes of B-cell lymphoma based on gene expression data. The method disentangles tumor heterogeneity and provides suited markers for the cancer subtypes. We proposed a refined subtyping of

B-cell lymphoma into four subtypes, rather than a previously assumed three-group classification. In a second application of the single-omics SOM we studied a gene expression data set concerning glioblastoma for which we confirmed an established four-subtype classification. Our results suggested a similar gene activation pattern as observed in the lymphoma study characterized by an antagonistic switching between transcriptional modes related to immune response and cell division.

Our integrative study on a larger lymphoma cohort comprising additional subtypes confirmed previous results about the role of stemness genes during development and maturation of B-cells. Their dysfunctions in lymphoma are governed by widespread epigenetic effects altering the promoter methylation of the involved genes, their activity status as moderated by histone modifications, and also by chromatin remodeling. We identified subtype-specific signatures that associate with epigenetic effects such as remodeling from transcriptionally inactive into active chromatin states, differential promoter methylation, and the enrichment of targets of transcription factors such as *EZH2* and *SUZ12*.

While studying the transcription of epigenetic modifiers in lymphoma and healthy controls, we found that the expression levels of nearly all modifiers are strongly disturbed in lymphoma and concluded that the epigenetic machinery is highly deregulated. Our results suggested that Burkitt's lymphoma and diffuse large B-cell lymphoma differ by an imbalance of repressive and poised promoters, which is associated with an imbalance of the activity of histone- and DNA-modifying enzymes.

Our inter-omics method was applied to a high-grade glioblastomas. Their expression and methylation landscapes were segmented into modes of co-expressed and co-methylated genes, which reflect underlying regulatory modes of cell activity. We found antagonistic methylation and gene expression changes between the *IDH1* mutated and *IDH1* wild type subtypes, which affect predominantly poised and repressed chromatin states. Therefore we assume that these effects deregulate developmental processes either by their blockage or by aberrant activation.

Our methods presented in this thesis enable a holistic view on high-dimensional molecular data collected in large-scale cancer studies. The examples chosen illustrate the mutual dependence of regulatory effects on genetic, epigenetic and transcriptomic levels. Our finding revealed that epigenetic deregulation in cancer must go beyond simple schemes using only a few modes of regulation. By applying the tools and methods described above to large-scale cancer cohorts we could confirm and supplement previous findings about underlying cancer biology.

Table of Contents

1 Introduction

1.1 CHALLENGES IN MOLECULAR CANCER MEDICINE

Just a century ago infectious diseases like influenza, pneumonia and tuberculosis led the list of main causes of death. But now, according to the Federal Office of Statistics, cancer is the second most common cause of death after cardiovascular systems diseases with large impact for healthcare.

Understanding cancer biology is vital in order to develop new and more effective treatment approaches. A key characteristic for cancer is a rapid, uncontrolled growth of abnormal cells potentially located at any part of the body. For many reasons cancer is a disease with currently limited treatment possibilities and poor prognosis: It is not a uniform, but a complex disease affecting multiple layers of the cellular machinery:

Genetic layer: Since many years it is known that mutations and copy number variations represent a potential cause for various cancers suggesting that cancer is a genetic disease. For instance, gain of whole chromosome 7 and loss of chromosome 10 are indicators of early stage of glioblastoma (GBM) formation [1]. Also characteristic for GBM patients is that they carry genetic defects such as mutated *IDH1* gene, often being associated with

abnormal DNA methylation pattern [2]. The advent of genetic techniques showed that usually a battery of mutations accompanies most cancer types. This makes clear that this disease is related rather to the accumulation of a certain 'spectrum' of genetic defects than to single defects that solely cause the disease as in case of monogenetic diseases.

Layer of genomic regulation: The genome is regulated by epigenetic mechanisms and it has been found that genetic and epigenetic events are mutually dependent in tumorigenesis: Epigenetic deregulation may result in mutations while an altered epigenome in turn could be caused by mutations in genes transcribing epigenetic modifiers [3]. In case of lymphomagenesis, critical mutations were reported in genes such as *KMT6* (alias *EZH2,* coding for an enzyme, which methylates Lys-9 and Lys-27 of histone H3) leading to a gain of function preferentially in diffuse large B-cell lymphomas (DLBCL) [4–6].

Layer of clonal evolution: Researchers attribute evolution by clonal selection as one reason for the genetic complexity of cancer and for failure of therapeutic treatment. With the help of clonal selection tumor cells are able to adapt to their environment, and to develop mechanisms, for example, to escape from immune response. In this way they can select and accumulate mutations that lead to cell proliferation and typically result in formation of metastases [7].

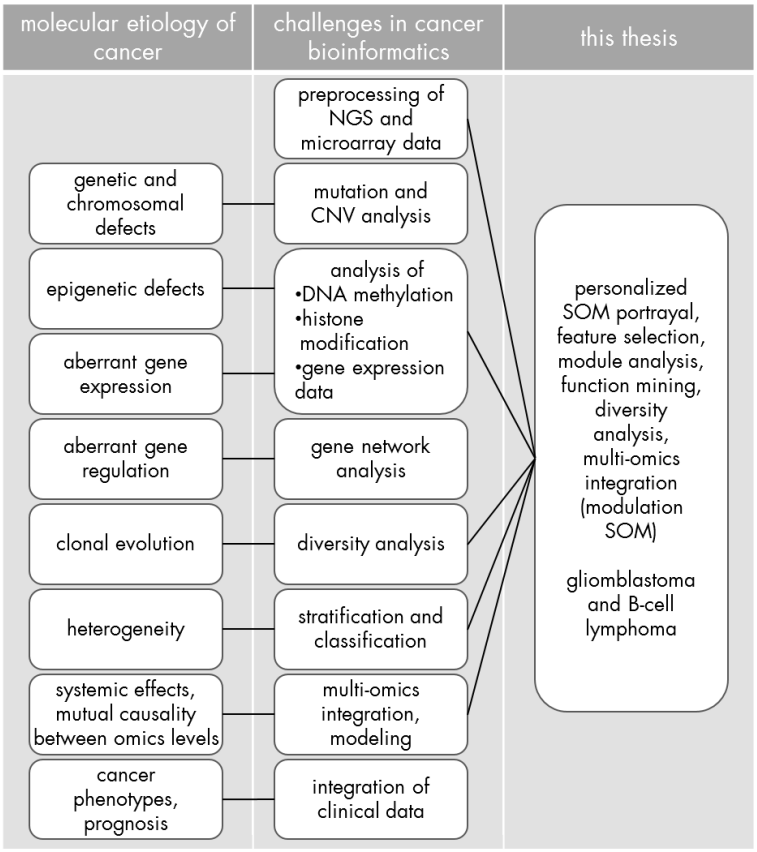


Figure 1: Challenges in molecular cancer medicine and the resulting demands on cancer bioinformatics. In the last column one can find the topics we dealt with in this thesis.

Pathway layer: Cells are constantly exposed to genotoxic stress, which causes for instance double-strand breaks of the DNA [8]. The damage caused should either be repaired or trigger cell apoptosis processes via a properly working DNA damage response (DDR) and repair proteins. In case of a malfunctioning repair machinery or a disrupted apoptotic pathway (due to age or environment) the risk for development of malignant tumor increases. Mutations of key DDR genes were for instance observed in DLBCL [9].

Taken together, research has shown that cancer is a heterogeneous disease meaning that one and the same type of cancer can be associated with a pluralism of distinct initial molecular events, leading to disturbed gene activity (see left panel of Figure 1). In consequence different biological functions and disease progression are characteristic and can lead to distinct molecular subtypes of a cancer entity. In other words, one cancer type splits on molecular level into three, four or even more subtypes, which constitute in principle disjunctive diseases in terms of genesis and progression and often with different prognosis and therapy options. Hence it is crucial to explore the molecular landscape of cancer and to stratify it into possible subtypes, and to characterize their specifics in terms of molecular markers, function and clinical relevance.

1.2 NEW GENOMIC TECHNOLOGIES AND CHALLENGES IN CANCER BIOINFORMATICS

Powerful high-throughput technologies, such as microarrays and next generation sequencing, have been developed in the last 20 years and enabled to study diseases on molecular level with high resolution, especially on genomics, transcriptomics and epigenomics levels. This revolution in the field of data acquisition gave rise to an overwhelming flood of molecular data. Depending on the biological material and the high-throughput technology used one obtains the abundance of ten thousands of mRNA transcripts per sample, millions of mutations, methylation levels of hundred thousands of DNA CpG sites and modification levels of histone side chains of millions of nucleosomes. In recent years, large-scale profiling of tumors were undertaken by means of projects such as The Cancer Genome Atlas (TCGA) [10], The Cancer Cell Line Encyclopedia [11] or the International Cancer Genome Consortium (ICGC) [12], which aim at characterizing cancer on the molecular and cellular level. These studies allowed to discover the heterogeneity of the underlying regulatory mechanisms and to assign them to molecular cancer subtypes. Although the data collection process of TCGA project ended in 2013, data analysis is still ongoing, revealing how challenging the incredible amount of high-dimensional heterogeneous cancer data is with great demands on bioinformatics methods [13–16]: Researchers need adequate tools to extract the information content of the data in an effective and intelligent way. This includes algorithmic tasks such as preprocessing, filtering of the data, feature

selection, linkage with the functional context in order to characterize and classify the different cancer subtypes, integration of clinical data and data obtained from multiple omics realms to achieve a systems-level understanding of the heterogeneity of cancer phenotypes, and finally it needs proper visualization of the data landscapes (also see middle column of Figure 1).

Especially, the latter task is very important because an intuitive visualization of massive data clearly promotes the quality control, the discovery of the intrinsic structure, functional data mining and finally the generation of hypotheses. We aim at adapting a holistic 'view' on the gene activation patterns rather than to consider single genes or single pathways. This view requires methods, which support an integrative and reductionist approach to disentangle the complex gene-phenotype interactions related to cancer genesis and progression.

1.3 SELF ORGANIZING MAPS AND PORTRAYAL OF MOLECULAR LANDSCAPES

One effective method that meets the requirements listed in the previous section is self organizing map (SOM) machine learning. SOMs are neural networks, which have been introduced by Kohonen [17]. First applications of SOMs on gene expression profiles were carried out by Törönen et al. [18] and Tamayo et al. [19].

A bioinformatics analysis pipeline based on SOMs has been developed by Wirth et al. [20], which enables a holistic view on high-dimensional molecular data collected in large-scale studies. Already being applied to numerous high-dimensional single-omics data sets regarding for instance human tissues [21] or time series experiments, the SOM-based pipeline has proven its capabilities as a reliable tool for clustering, dimension reduction and visualization: The portraying method transforms the multitude of different profiles inherent in a multidimensional data set into a two-dimensional map. The data map obtained can be simply 'read' by visual inspection revealing the number of relevant clusters of co-regulated genes in terms of disjunctive 'spots' and their mutual correlation structure. Furthermore, it provides a general framework for analytic tasks such as feature selection, integration of concepts of molecular function and systems tracking with individual resolution.

In order to meet the requirements of cancer bioinformatics the single-omics SOM pipeline presented in chapter 3 was adjusted and supplemented by a multitude of in-house methods and measures like additional spot-, entropy- and variance measures, novel adaptions for training of DNA methylome data, gene set lists with disease specific signatures sets, and a list of genes coding for epigenetic modifiers allowing to systematically study their deregulation in cancer.

1.4 OBJECTIVES

The aim of this thesis is to develop and apply SOM-based bioinformatics methods that enable the analysis of molecular high-throughput data collected in large cancer cohort studies with the special focus on epigenetic (dys-)regulation of transcription. We will unravel molecular mechanisms underlying cancer in the specific case of B-cell lymphoma and glioblastoma as proof-of principle applications.

Particular tasks addressed in this thesis are:

- portrayal of complex molecular data landscapes with individual resolution, in terms of gene expression and DNA methylation
- identification of suited markers for diagnosis of cancer subtypes that disentangle tumor heterogeneity and to discuss their relevance in terms of cancer biology
- re-evaluation and characterization of molecular subtypes described previously and their mutual comparison across the cancer entities
- joint analysis of gene expression and DNA methylation data to compare classification schemes originating from the different data types and analysis of the mutual associations between them
- the study of potential modes of epigenetic regulation in cancer subtypes under consideration of chromatin states, chromatin modifying enzymes, DNA methylation and gene expression

We will adapt SOM machine learning to these tasks and provide suited analysis tools. SOM represents a suitable method to achieve our goals and to face most of the challenges of cancer bioinformatics mentioned in section 1.1 (see also Figure 1). It already has proven to be efficient dealing with high-throughput data [22,23]. In this thesis the SOM pipeline is complemented and applied to several cancer data sets concerning gene expression and DNA methylation.

This thesis is divided into four main parts: First we give a short review about biological background and high-throughput technologies that are used to produce data concerning transcriptomics and DNA methylation. Secondly, chapter 3 is devoted to methodical questions: We describe details of SOM portrayal of high-throughput data and additional tools addressing different data analysis tasks. Those methods are exemplified by means of a prostate cancer study. Thirdly, each of the application chapters (4 and 5) is subdivided into three parts: At first, the cancer entities (B-cell lymphoma and glioblastoma multiforme, respectively) are analyzed based on gene expression data. Secondly, epigenetic mechanisms driving carcinogenesis are examined with regard to DNA methylation data and its impact on transcriptomics. Thirdly, integrative methods are presented to combine transcriptomics and epigenomics. This is either realized by studying the expression of genes coding for epigenetic modifiers or by a novel multi-omics approach, based on SOMs, exemplified by a glioblastoma cohort. In the final part a summary of the results is given and a conclusion is drawn.

2 Biological and experimental background

The central dogma of molecular biology, introduced by Francis Crick [24], combines three different transfer types of biological information, namely replication of DNA, transcription of DNA in RNA and translation of the obtained RNA into proteins. With regard to the basic mechanisms it still retains its validity but in the last two decades a lot of research has been done revealing different levels and mechanisms so that the dogma has to be revised. A cascade of regulatory mechanisms and interacting molecules must be complemented to the so far unidirectional central dogma from the perspective of systems biology [25].

2.1 BIOLOGICAL BACKGROUND

2.1.1 TRANSCRIPTOMICS

The DNA is divided into (protein-) coding and non-coding regions. In the coding segments of the DNA genes are stored. The term gene expression is used to describe the initial process leading to the synthesis of a gene product, usually a protein, from the information contained within a gene. The first step is called transcription, the process by which DNA is transcribed into RNA, another nucleic acid. There are several types of transcripts, named according to their functionality for instance mRNAs (messenger RNA), non-coding RNAs and small RNAs. The entirety of transcripts in a cell is called transcriptome and one of the main goal in transcriptomics is to quantify the expression level of the transcripts under certain conditions measured in terms of mRNA.

2.1.2 EUKARYOTIC CHROMATIN STRUCTURE

What makes the difference between for instance a hepatocyte and a skin cell if not the DNA sequence being the same in both cells within one organism? The answer to that question is expression of thousands of genes determining a cell's function. The gene expression state during differentiation is controlled by transcriptional regulation governed by chromatin state.

Understanding chromatin structure is crucial for transcriptional regulation. The structure of DNA is divided in four different types (see Figure 2). The primary structure is basically the nucleotide sequence of DNA itself, made up of the 4 bases adenine (A), thymine (T), cytosine (C) and guanine (G). The secondary structure, also known as α-helix is formed by complementary base pairing of two nucleotide strands by means of hydrogen bridges. The spatial, helical arrangement of the double strand is termed tertiary structure and the quaternary structure refers to the formation of complexes of DNA and proteins, also called nucleosomes. Repeating units of those nucleosomes form the chromatin to package the two complementary DNA strands, each consisting of approximately 3 billion nucleotides and distributed over 23 pairs of chromosomes, in the nucleus of every human cell. For a nucleosome a 147 bp long fraction of DNA is wrapped around a histone octamer consisting of two copies of the four core histones [26]. The linker histone represents the connection between 3′ and 5′ end of DNA wrapped around the histone octamer therefore stabilizing the nucleosome and is relevant to form higher-order chromatin. One can distinguish between two types of chromatin, namely euchromatin and heterochromatin determining the accessibility of the DNA. The latter is known to mediate transcriptional silencing due to its more condensed and closed structure while euchromatin also known as open chromatin contains genes being actively transcribed as a consequence to its loose packaging [26].

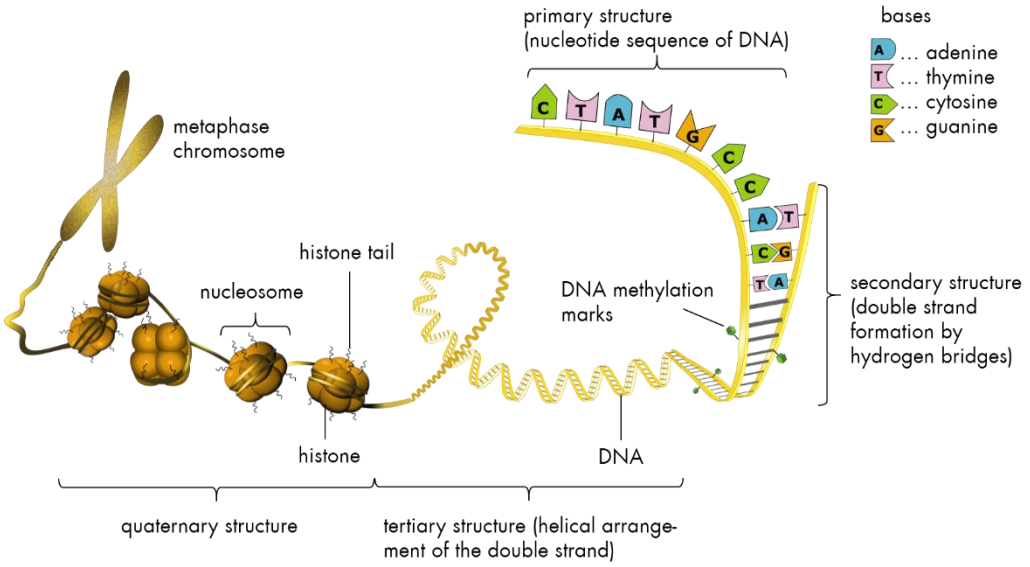


Figure 2: Chromatin structure. Adapted from [27].

2.1.3 EPIGENETICS

Epigenetics is sometimes described as a second layer that coats the DNA. The common definition according to today's understanding was given by Russo as 'the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence' [28]. It controls the chromatin structure and accessibility, determines the functionality and the phenotype of a cell by turning genes on or off [26]. Among the most prominent epigenetic mechanisms are posttranslational modifications of histone tails, methylation of DNA and chromatin remodeling [29], [30]. Roadmaps Epigenomics (http://www.roadmapepigenomics.org/) is one of the most well-known consortia concerning epigenetic data as it provides more than a hundred publicly available human reference epigenomes. One of the objectives of the consortium is to present a platform to study the role of epigenetics in the genesis and progression of human diseases.

Human cancer was firstly linked to epigenetics in 1983 when a global DNA hypomethylation was observed by Feinberg et al. [31]. Through the years epigenetics also became a hot topic in cancer therapy. Up to now several epigenetic-related anti-cancer drugs have been developed and approved that basically inhibit DNA methylation or histone modifications [32].

2.1.3.1 DNA METHYLATION

Methylation of the 5' carbon of cytosine at CpG (cytosine-guanine dinucleotide) sites is the most studied epigenetic mark. It is assumed that due to methylation of cytosines present in the promoter region of genes, either binding of transcription factors (TFs) may be hindered or mediators of chromatin remodeling complexes bind to those methylated cytosines,

which consequently promotes silencing of the downstream gene [33]. But meanwhile researchers have found that this mechanism doesn't apply to each gene concluding that quantitative relation of gene expression and promoter methylation is not yet fully understood [34].

Due to the asymmetric arrangement of CpG on the DNA (methylated cytosines lie diagonally to each other on both strands) methylation marks are stable and can be re-established after replication and inherited to daughter strands [35]. The methylation process is catalyzed by two kinds of DNA methyltransferase enzymes (DNMTs). The first one is called *de novo* DNMT, represented by *DNMT3A* and *DNMT3B*, which initially methylate the DNA during embryonic development. The maintenance methyltransferase *DNMT1* is responsible to inherit the methylation marks to the daughter strand after replication [36,37].

A variety of biological processes like aging, X-chromosome inactivation in females, development and genomic imprinting have been attributed to DNA methylation while altered methylation was brought into context with complex diseases like heart disease or cancer [38]. Hypermethylation of the promoters of tumor suppressor genes or genes involved in cell cycle or DNA repair pathways is often associated with tumorigenesis [3].

2.1.3.2 HISTONE MODIFICATION AND THE HISTONE CODE

Chromatin structure and transcription are to some extend regulated by chemical modifications like methylation, acetylation, phosphorylation, ubiquitylation, and sumoylation of the N-terminal tail of the core histones H2A, H2B, H3 and H4 [39]. Similar to DNA methylation enzymes catalyze the posttranslational modification reactions of arginine, lysine and serine residues at histone tails. Depending on the site and type, histone modification can have various effects. The most studied marks are acetylation and methylation with both chemical modifications having different effects on transcription: Acetylation of lysines on H3 or H4 generally promoting transcriptional activation due to its capability to decondense chromatin. The impact of methylated histone residues on expression is more diverse depending on the site, the residue (arginine or lysine) and the degree of methylation (lysines may be mono-, di- or trimethylated) [40]. High levels of H3K4 (histone H3 at lysine 4) methylation in promoter regions are correlated with high transcription rates while trimethylation of H3K9 and H3K27 have been associated with transcriptional silencing. Furthermore, H3K36me3 (trimethylation of H3K36) marks are found in the bodies of actively transcribed genes [41]. In 2001 Jenuwein and Allis introduced the histone code hypothesis implying that combinations of histone modifications may evoke changes in the chromatin state and lead to modified regulatory mechanisms of gene expression [42].

An aberrant activity state of histone-modifiers has also been implicated in cancer genesis. For instance upregulation of *EZH2* (methylates H3K9 and H3K27) has been associated with metastatic prostate cancer [43].

2.1.4 CARCINOGENESIS THROUGH (EPI-)GENOMIC DYSREGULATION

Figure 3a provides a schematic overview of central ingredients of carcinogenesis arising as a consequence of a modified chromatin state [44,45]: Chromatin states and also the more subtle activity states of gene promoters represent essential determinants of gene transcription shaping cell function and the production of chromatin modifying enzymes (Figure 3b). These enzymes model the chromatin states via writing and erasing of epigenetic marks attached. Such marks are then read by the chromatin (re-)modeling machinery, which potentially leads to changes of chromatin structure with possible consequences for gene expression.

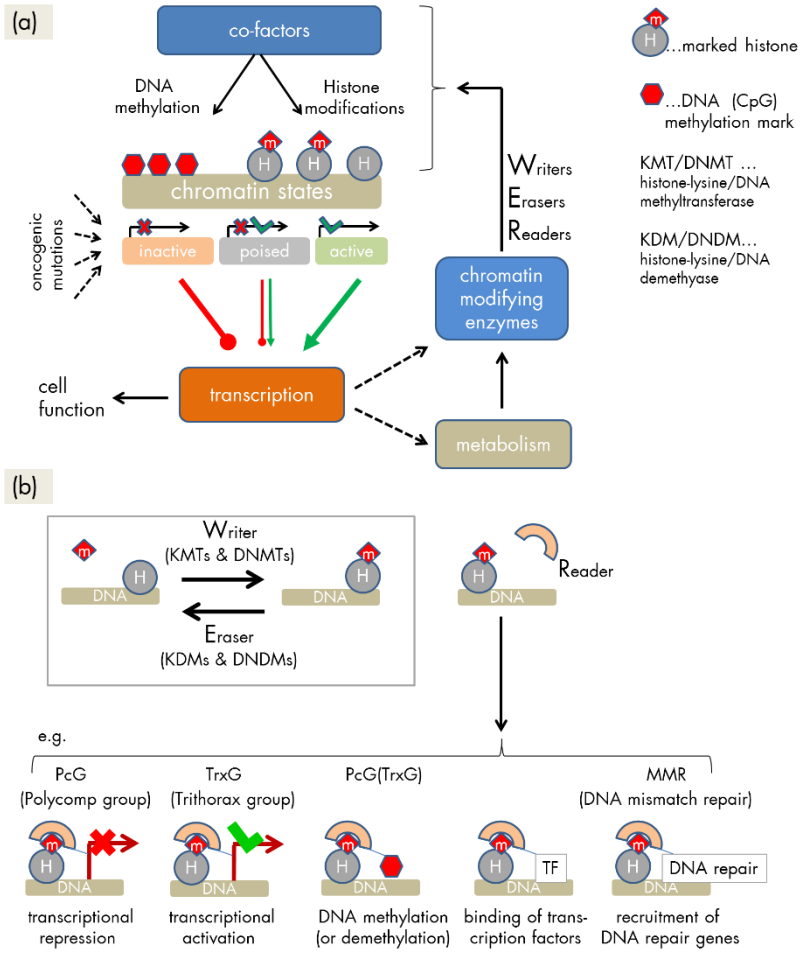


Figure 3: Carcinogenesis through (epi-)genomic dysregulation: **(a)** Circuit of epigenetic regulation: Different chromatin states are induced by histone modifications, which in concert with DNA methylation, modulate transcription of the affected genes, resulting in the production of chromatin modifying enzymes, which again regulate the formation of different chromatin states. This feedback loop is further modulated by metabolites, which serve as cofactors. Oncogenic mutations can disbalance this network giving rise to malignant cellular states. **(b)** Writers and erasers are chromatin modifying enzymes that add or remove epigenetic marks, respectively. Readers recognize such marks and induce specific molecular 'actions' (activation or repression of gene expression, writing or erasing of DNA or histone marks or recruitment of TFs or of DNA-repair genes).

The activity of chromatin modifying enzymes within this regulatory circuit represents one important determinant of epigenetic regulation. Mutations of genes coding epigenetic modifiers are initiating events in cancer that can induce an 'avalanche' of downstream epigenetic effects. They can start with the aberrant expression of chromatin modifying enzymes, which leads to aberrant epigenetic marks and then to aberrant chromatin states and finally to aberrant cellular activities. Mutations not directly targeting epigenetic modifiers can also induce analogous 'avalanches' of epigenetic deregulation, if, for example, they hit TFs, which downstream regulate the expression of epigenetic modifiers. De-regulation of the epigenetic machinery can also be mediated by the metabolome, e.g. if mutations of genes, encoding metabolic enzymes, modify metabolites acting as inhibitors or activators of epigenetic enzymes. For example, mutations of the gene, which codes for isocitrate dehydrogenase 1 (*IDH1*) disturb the DNA methylation machinery and induce special types of brain cancer by alterations of the activity of epigenetic enzymes [46]. Perturbations of chromatin-modifying mechanisms are among the central oncogenic pathways inducing human cancer [47].

2.2 HIGH-THROUGHPUT TECHNOLOGIES

2.2.1 GENE EXPRESSION QUANTIFICATION IN LARGE SCALE STUDIES

Gene expression profiling is used to simultaneously measure the activity of thousands of genes at a particular time in healthy and diseased states [48]. It has gained great significance in biology and biomedical research when it comes to identifying biological processes of large scale studies. The most frequently used methods for measuring gene expression are microarrays and RNA-sequencing (RNA-Seq) [49].

Starting in 1997, microarrays have revolutionized fields like molecular biology, medicine and pharmacy. One can distinguish between commercial platforms measuring the gene expression of the whole genome and custom arrays, which only spot probes with the sequence of genes of interest [50]. Depending on the manufacturer, the protocol or reagents vary, but the principle of microarrays remains the same and is called hybridization. Oligonucleotides with unknown sequence of length up to 70bp bind to probes of known sequences attached in ordered fashion on a solid surface like glass or silicon by complementary base pairing [51].

The workflow of RNA-Seq experiments is quite different: After isolation of RNA it is converted into cDNA, which is further fragmented into small pieces with all being of the same length [52]. So-called adapters of known sequence are ligated to the fragments. The fragments are then sequenced to a specific depth using a sequencing machine. At this point the laboratory work ends and data analysts get the sequenced reads and quality scores to further quantify RNA being present in a sample at a certain time.

Taken together, the main difference between microarrays and RNA-Seq is that for microarrays only the predefined set of probes and therefore only a limited set of genes' expression can be measured, while for RNA-Seq the expression of both mRNAs, small and non-coding RNAs is detected not limited to specific gene loci [48]. Furthermore, structural variants can be discovered. However, gene expression arrays are still accepted in bioscience and widely used not least because they are less expensive than RNA-Seq experiments, and depending on the purpose of the study researchers should decide whether to use microarrays or sequencing method.

2.2.2 MEASUREMENT OF DNA METHYLATION

Like measuring gene expression levels also for methylation there exist microarrays to detect DNA methylation rates and a sequencing alternative called bisulfite sequencing. Both methods work on the same principle of bisulfite conversion: The DNA is treated with sodium bisulfite, which doesn't have an impact on the sequence except for unmethylated cytosine being transformed to uracil [53]. Illumina's Infinium HumanMethylation450 BeadChip arrays is the most frequently used microarray platform for measuring DNA methylation [53]. The array spots 485,512 preselected CpGs distributed over the entire genome covering 99% of known genes. In contrast bisulfite sequencing offers greatest genomic coverage but clearly exceeds the costs in comparison with microarrays.

2.2.3 IDENTIFICATION OF HISTONE MODIFICATION SITES

Chromatin immunoprecipitation (ChIP) is generally used to detect DNA-protein interactions, for instance to identify binding sites of TFs to promoter regions of genes. This method is also applicable for the detection of chemical modifications of histone proteins using appropriate antibodies. The sequencing based technology is called ChIP-Seq, which has surpassed ChIP-chip being the DNA microarray based counterpart [54].

There are several large consortia that produce and provide genome-wide data regarding genomic regulation. For example the ENCODE (ENCyclopedia Of DNA Elements) project was launched in 2003 dedicated to the determination of all functional and regulatory elements of the human genome. The NIH Roadmap Epigenomics Consortium has gathered the so far largest collection of epigenomes, among others also in terms of ChIP-Seq data sets, derived from human tissues and primary cells [55].

Ernst et al. [56] developed a computational system based on a hidden Markov model called ChromHMM, which uses combinations of histone modifications to predict chromatin states. The Roadmap Epigenomics project provides ChromHMM based segmentation across 127 epigenomes of different tissue types. Later on we make use of the whole-genome segmentation into chromatin states in terms of healthy lymphoblastoid and neuronal progenitor cells in order to get an insight into the possible mechanisms of chromatin remodeling in various cancer entities.

3 SOM portrayal of high-throughput data

Self-organizing maps (SOM) portray molecular phenotypes with individual resolution. We developed an analysis pipeline based on SOM machine learning, which allows the comprehensive study of large scale clinical data. For all SOM analyses throughout this thesis we used the R-package 'oposSOM', which is publically available from the Bioconductor repository [20]. Further details concerning the SOM pipeline are provided in [57].

The potency of the method is demonstrated in a selected application studying the diversity of gene expression in prostate cancer progression (PCP), which has been published in:

Hopp, L., Wirth, H., Fasold, M., & Binder, H. (2013). Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. *Systems Biomedicine*, 1(2).

3.1 DATA

In Table 1 we compose a list of all data sets used within this thesis. Additional information can be found in supplement section 7.1 and in the associated sections.

Table 1: Data sets used throughout this thesis.

cancer entity	gene expression		DNA methylation	
	subtypes (sample size)	platform, GEO accession number, references	subtypes (sample size)	platform, GEO accession number, references
prostate cancer progression (PCP)	chapter 3			
	PIN (13), PCA_low (12), PCA_high (20), MET (17) , controls: BHP (22)	non-commercial spotted Human 20K Hs6 arrays, GSE6099, Tomlins et al. [58]		
B-cell lymphoma	section 4.1			
	mBL (44), non-mBL (129), intermediate (48)	Affymetrix HT HG-U133A arrays, GSE4475, Hummel et al. [59]		
	section 4.2			
	mBL (85), non-mBL (287), IntL (307), FL (121), BCL (64), controls: B-cells (17), GCB-cells (13), lymphoma cell line (32), tonsils (10)	Affymetrix HT HG-U133A, Hummel et al. [59]	DLBCL (54), mBL (18), IntL (16), FL (14), MCL (10), MM (14), controls: B-cells (5), GCB-cells (2)	GoldenGate Methylation Cancer Panel I, Martin-Subero [60]
	section 4.3			
	mBL (62), DLBCL (204), IntL (255), FL (3), BCL (36), controls: lymphoma cell lines (32), B-cells (17), GCB-cells (13), tonsils (10)	Affymetrix HT HG-U133A, GSE4475, GSE10172, GSE22470, GSE48184, GSE43677, Hummel et al. [59]		
glioblastoma multiforme (GBM)	section 5.1			
	MES (50), PN (45), NL(26), CL (32), control: NOR (11)	Affymetrix HT HG-U133A, Verhaak et al. [61]		
	section 5.2			
	MES (5), RTKI (6), RTKII (3), IDH (7), G43 (4), K27 (5), control: fetal (3)	Affymetrix HT HG-U133A, Sturm et al. [62]	G34 (18), K27(18), MES (36), IDH (19), RTKI (23), RTKII (22), controls: adult (2), fetus (4)	Illumina Human Methylation450 BeadChip, GSE36278, Sturm et al. [62]
	MES (50), PN (45), CL (32), control: NOR (10)	Affymetrix HT HG-U133A, Hopp et al. [61]		
	MES (21), CL (23), PN-IDH-mut (12), control: PN-IDH-wt (14)	Affymetrix HT HG-U133_Plus_2, GSE53733, Reifenberger et al. [63]		
	section 5.3			
	MES (16), RTKI (4), RTKII (16), IDH (3)	Affymetrix HT HG-U133A, TCGA, Sturm et al. [62]	MES (16), RTKI (4), RTKII (16), IDH (3)	Illumina Human Methylation450 BeadChip, TCGA, Sturm et al. [62]

3.2 PREPROCESSION: PREPARING THE DATA

Gene expression

Raw probe intensity values of gene expression Affymetrix arrays were calibrated and summarized into one expression value per probe set using the hook method [64,65]. For arrays other than Affymetrix, we downloaded already preprocessed expression data. To ensure comparability, expression values of the sample arrays were further quantile-normalized [66], which transfers the expression states of all samples into one common distribution.

Then the expression data E was transformed in log-scale and centralized with respect to the mean value of each gene $n = 1,\ldots,N$ averaged over all samples $j = 1,\ldots,J$

$$\Delta e_{nj} = \log_{10} E_{nj} - \frac{1}{J}\sum_{j=1}^{J} \log_{10} E_{nj}. \qquad \textbf{Eq.(1)}$$

This definition of differential expression and differential methylation refers to the mean expression/methylation level of each gene n in the data set studied. Centralization using Eq.(5) emphasizes further analysis on differential values (logFC, fold change units) independent of the respective absolute expression levels. Hence, a Δe_{nj} of zero means that the gene is expressed according to its mean expression value. If not stated otherwise, we use the terms over- and underexpression throughout the thesis for $\Delta e_{nj} > 0$ and $\Delta e_{nj} < 0$, respectively. For expression data we define 'profiles' given as data vector for each gene with the sample-related values as elements;

$$\Delta e_{n\cdot} = (\Delta e_{n1}, \ldots, \Delta e_{nJ}) \qquad \textbf{Eq.(2)}$$

and 'states', given as data vector for each sample with the gene-related values as elements,

$$\Delta e_{\cdot j} = (\Delta e_{1j}, \ldots, \Delta e_{Nj}). \qquad \textbf{Eq.(3)}$$

DNA methylation

For methylation data we used the 'M'-scale (ratio of methylated to unmethylated) instead of β-scale. β values are defined as the relative methylation level, which can vary between values of zero (no methylation) and unity (full methylation). For SOM analyses β values of gene n were transformed into M values

$$M_{nj} = \frac{\beta_{nj}}{1 - \beta_{nj}}\ , \qquad \textbf{Eq.(4)}$$

which theoretically cover the range between minus infinity (no methylation) to plus infinity (full methylation). M values are statistically more valid because they avoid heteroscedasticity of differential methylation values for large ($\beta > 0.8$) and small ($\beta < 0.2$) β values [67]. For intermediate β range ($0.2 < \beta < 0.8$) β and M are nearly linearly correlated. Genes located on chromosomes (Chr) X and Y were excluded from further analyses to minimize gender specific effects [22].

In analogy to preprocessing of expression data, differential methylation was calculated by taking the logarithm of methylation data und centralizing them with respect to the mean value of each gene *n* averaged over all samples *j*

$$\Delta m_{nj} = \log_{10} M_{nj} - \frac{1}{J}\sum_{j=1}^{J} \log_{10} M_{nj}. \qquad \textbf{Eq.(5)}$$

A Δm_{nj} of zero means that the gene is methylated according to its mean methylation value. If not stated otherwise, we use the terms hyper- and hypomethylation for $\Delta m_{nj} > 0$ and $\Delta m_{nj} < 0$, respectively. Methylation 'profiles' are given as data vectors for each gene *n* with the sample-related values as elements $\Delta m_{n\cdot} = (\Delta m_{n1}, \ldots, \Delta m_{nJ})$ while methylation 'states' are given as data vectors for each sample *j* with the gene-related values as elements $\Delta m_{\cdot j} = (\Delta m_{1j}, \ldots, \Delta m_{Nj})$.

3.3 SOM TRAINING

The preprocessed data is used to train a self-organizing map (SOM). It translates the high-dimensional data given as *N* x *J* matrix (*N*: number of genes, *J*: number of samples) into a *K* x *J* matrix (*K*: number of so-called metagenes) of reduced dimensionality $K \ll N$ ($N \sim 10^4$ and $K \sim 10^3$) using an unsupervised learning algorithm. The metagene profiles are obtained via iterative machine learning, while clustering the gene profiles on a two-dimensional quadratic grid of $\sqrt{K}$ tiles per x- and y-dimension using Euclidean distance as similarity measure. The final SOM consists of regions of similar metagene profiles. As the number of input genes *N* exceeds the number of nodes *K* in the grid, each metagene serves as a representative prototype of a 'mini-cluster' of real genes with similar profiles. It reflects the differential expression or methylation of the prototypic metagene compared to its profile-averaged value, $\Delta e_{kj} = e_{kj} - e_{k\cdot}$ and $\Delta m_{kj} = m_{kj} - m_{k\cdot}$, where e_{kj} and m_{kj} are the logged expression and methylation values of metagene *k* in sample *j* and $e_{k\cdot}$ and $m_{k\cdot}$ are the respective profile means, respectively.

Please note, that for each application throughout this thesis the SOM size (*K*) was chosen according to match the criterion of robustness of the SOM.

3.4 SOM STAINING: PORTRAYAL

Method

Each samples meta-state is visualized by color coding the two-dimensional mosaic of metagenes according to their feature values in the respective sample. Please note that the assignment of the genes to metagene clusters and therefore also their position in the SOM is identical in all sample portraits. Hence, the coloring at a certain position in the map refers to the same genes in all individual portraits allowing the direct comparison of their

expression or methylation levels between the maps. We first normalize the metagene data in each state to the range [–1, +1] and then color code the obtained mosaics: The 'log*FC*'-scale linearly transforms the normalized logged fold change of metagene *k* in sample *j*, $\log FC = \Delta e_{kj}^{norm}$ and $\log FC = \Delta m_{kj}^{norm}$ into green to maroon for $\Delta e_{kj}^{norm} \geq 0$ and $\Delta e_{kj}^{norm} \geq 0$, and green to dark blue for $\Delta m_{kj}^{norm} \leq 0$ and $\Delta m_{kj}^{norm} \leq 0$. The color patterns emerge as smooth textures representing the fingerprint of transcriptional activity or methylation state of each sample, respectively.

Average subtype-specific portraits are calculated as the mean value of each relative metagene measure over all phenotype portraits of one subtype, $\Delta e_{kc} \equiv \langle \Delta e_{kj} \rangle_{j \in c}$ and $\Delta m_{kc} \equiv \langle \Delta m_{kj} \rangle_{j \in c}$ (*c* is the class index of each subtype) followed by normalization and coloring in log*FC*-scale. They reflect subtype-specific expression and methylation patterns while leveling out the heterogeneity of the individual feature states and outliers. Furthermore population of the metagenes and the variance of metagene profiles can be visualized using the same mosaic structure. Details are provided in supplement section 7.2.

Example

PCP expression data was preprocessed as given in section 3.2 and supplemental section 7.1.1. SOM machine learning transforms the whole genome expression pattern into one colored mosaic image per sample. In case of PCP about four thousand single genes are distributed over 1,600 (40x40) tiles, each tile representing one metagene. Figure 4 portrays the expression states of PCP. We sort these expression portraits into different groups according to previous classifications into progression stages ranging from benign prostatic hyperplasia (BHP) and prostatic intraepithelial neoplasia (PIN) to low-grade (PCA_low), high-grade (PCA_high), and metastatic (MET) prostate cancer [68]. Exemplary portraits of individual samples are shown in log*FC*-scale highlighting areas of strong over- and underexpression.

We calculated the mean SOM-portrait (large images in Figure 4) of each class by averaging the expression values of each metagene over all class members. This averaging cancels out individual, highly fluctuating features and this way it amplifies consistent class-specific features. Each small mosaic exhibits a characteristic texture of the respective cancer sample. The expression portraits in log*FC*-scale reveal a handful of over- and underexpression spots, which selectively characterize different cancer subtypes such as BHP, PIN, PCA_low, PCA_high and MET prostate cancer in Figure 4. One observes either relative stable and consistent spot-patterns (e.g. for MET) or relatively heterogeneous and volatile patterns (e.g. for the PCA_high samples).

Some spots were observed in more than one PCP-stage. As a rule of thumb the spots of subsequent stages, and also of the final MET- and initial BHP-stages, tend to overlap. In consequence, the stage-specific spot pattern 'rotates' along the border of the map in clockwise direction with progressing cancer.

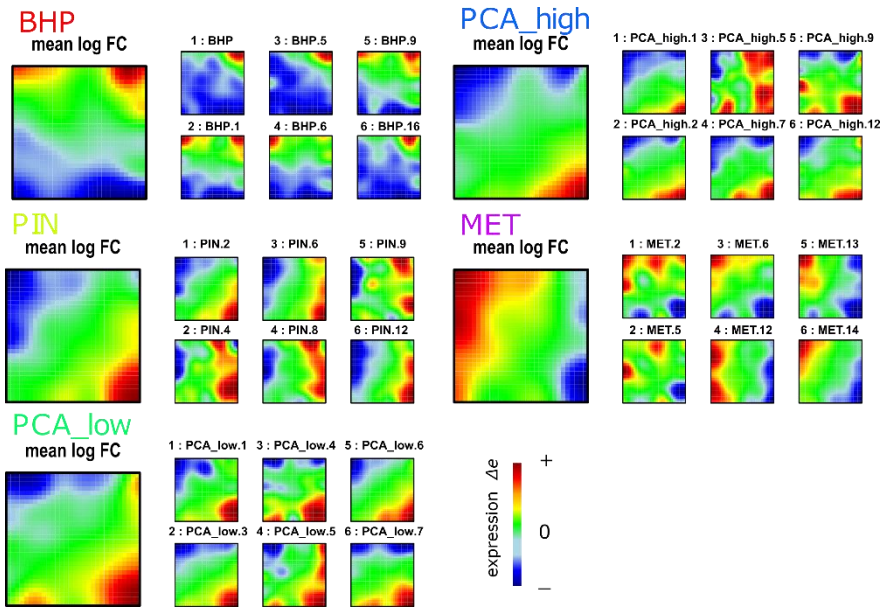


Figure 4: SOM gallery of PCP stages. The small mosaic images refer to selected individual tumor samples assigned to the five PCP stages while the large images represent the respective mean portraits of each stage. The images use a log*FC*-scale, where *FC* denotes the fold change of the expression of each metagene with respect to its mean expression in all samples. A complete gallery of all sample portraits is available in supplementary material of [69].

In summary, SOM-imaging portrays the individual expression landscapes of each sample in terms of characteristic color textures, which enable visual inspection of subtype-specific spot-like features representing clusters of differentially expressed and co-regulated genes. Simple averaging over groups of samples amplifies class-specific features.

3.5 SAMPLE SIMILARITY ANALYSIS: HETEROGENEITY

Method

Sample similarity analysis aims at establishing mutual relations between the phenotypes studied, e.g., to extract a hierarchy of similarities or to estimate mutual distances between the feature states. Similarity analysis compares the feature states as seen by the SOM portraits. It consequently uses the expression or methylation of metagenes instead of single genes as the basal data, which has the advantage of improving the representativeness and resolution of the results [23,70].

We applied second-level SOM analysis as proposed by Guo et al. [71] to visualize the similarity relations between the individual SOM-metagene patterns. We used the *K* metagene profiles of the *J* samples to perform sample-wise clustering.

Another method called independent component analysis (ICA) [72] was applied to the SOM-metagenes using the R-package 'fastICA'. It distributes the samples in the space spanned by the components of minimum mutual statistical dependence. These components

point along the directions of maximum information content in the data, which is estimated by their deviation from a (non-informative) normal distribution [73]. ICA is based on the covariance matrix calculated in terms of Pearson correlation coefficients *r* between all metagene values of pairwise combinations of samples. The correlation matrix was visualized using pairwise correlation maps (PCM), maximum spanning tree- (MST) and correlation cluster net (CN)-representations.

MST's are a well-established concept in graph theory. The algorithm interprets the distance matrix as a complete graph in which the edge weights correspond to the distances. The MST is the spanning tree that connects all vertices of that graph with the smallest sum of edge weights. It thus represents effectively the 'shortest' distance between two nodes in the graph resulting in a chain-like structure. MST's have been shown to be useful for clustering and classification of cancer subtypes using microarray data [74]. For the MST calculation we used the spantree function of the R-package 'igraph'.

A second correlation based representation is supplied by the CN. This unweighted graph is constructed by connecting the nodes (*i.e.* the samples), whose pairwise correlation coefficient *r* exceeds a given threshold (here $r_{threshold} = 0.5$). This graph supplements the sparse MST with a more detailed and network-like overview about the sample correlation structure. It implies more connections as the MST and thus considers also weaker mutual correlations.

Finally, we also applied the neighbor-joining algorithm (R-package 'ape') to represent similarity relations based on the Euclidean distances between the samples in terms of similarity trees [75]. The distances between pairs of samples in the tree are in scale. In contrast to MST-representation the phylogenetic tree allows to identify 'bush-like' clusters and to estimate the degree of mutual dissimilarity between them.

Example

The 2nd level SOM representation of PCP is given in Figure 5a with the mean regions occupied by the samples of each of the different PCP-stages being illustrated by the largely overlapping colored polygons. The first and final stages can be well distinguished, whereas the intermediate stages PIN, PCA_low and PCA_high are found essentially in the same region of the map. Note the correspondence between the spot patterns in the mean portraits of the subtypes and the symmetry of their arrangement in the 2nd level SOM: The rotating spot-pattern of the PCP-stages transforms into an U-shaped trajectory of subsequent stages in the 2nd level SOM reflecting the fact that a significant part of the genes are similarly expressed in the final MET-stage and in the initial BHP-stage, but differently expressed in the intermediate PIN- and PCA-stages (see the spot pattern in Figure 4).

As a complementary method, independent component analysis (ICA) was applied to the SOM portraits of all samples of each cancer progression stage (see Figure 5b). The samples are similarly distributed in ICA-space as in the 2nd level SOM.

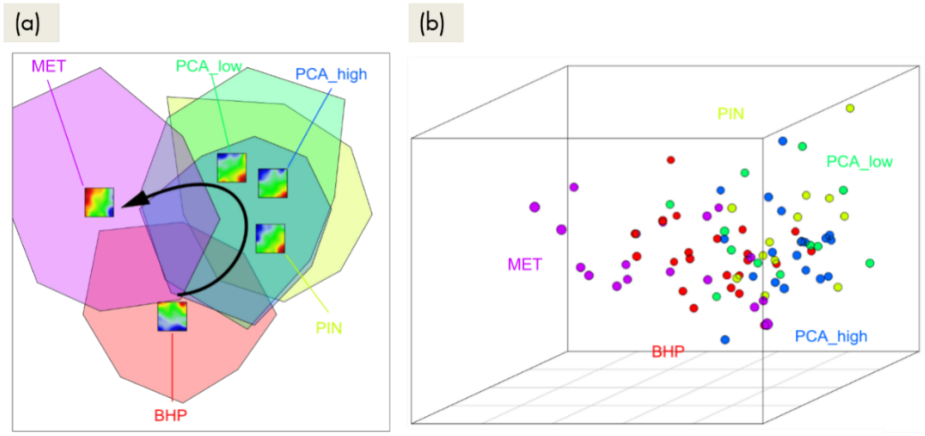


Figure 5: The 2nd level SOM and ICA similarity analysis of PCP stages. (**a**) The 2nd level SOM polygon representations of PCP stages. Note that the spot pattern in the mean expression maps of PCP virtually rotates with progressing cancer giving rise to a U-shaped trajectory in the map (see arrows); (**b**) Three dimensional ICA .

The PCM given in Figure 6a visualizes the correlation coefficients for all sample pairings, which are arranged according to their subtype assignments (see the color bars along the borders of the map). Maroon and red colored tiles assign strong correlations and thus pairwise combinations of similar portraits and blue colored tiles assign anti-correlated portraits where usually overexpressed regions have switched into underexpressed ones and *vice versa*. The samples of the same tumor stage were grouped together to visualize the intra- and inter-class similarity of the samples (see color bars along the edges of the map). For example, BHP samples are predominantly anti-correlated with PIN and PCA samples but partly correlated with MET samples (see also the respective anti-correlated or correlated spot pattern of the mean portraits per class). The covariance structure of the data is visualized using the MST and the CN representations shown in Figure 6b-c. Importantly, all similarities are based on the metagenes, which provide a better resolution than single gene-based similarity analysis [23,71].

The MST and the CN of the PCP-samples (Figure 6b and c respectively) show a backbone-like structure reflecting the temporal progression of the respective stages of prostate cancer. The mutual distance among them increases with progressing cancer as a rule of thumb. MET-samples are however again found near BHP-samples. This finding is consistent with the U-shaped arrangement of the PCP-stages in 2nd level SOM (see Figure 5a), which suggests larger similarity between the first and final stages than between the first and intermediate PCA_high- and PCA_low-stages. These similarity relations transform into star-like dendrograms, which are obtained using the neighbor-joining algorithm based on Euclidean distance metrics (see Figure 6d). One can see that the more localized MET-subtype tends to aggregate into separate branches whereas the intermediate PIN- and PCA-subtypes occupy diffuse branches.

The different similarity plots thus provide complementary information with emphasis on their distribution in two dimensions (2nd level SOM), the independence of the underlying

features (ICA) and the strongest correlations between the samples (MST), which is further disentangled in the CN- and dendrogram-plots. The MST, CN, and dendrograms partly reveal finer details such as the compactness or fuzziness of mutual relations. On the other hand, the 2nd level SOM provides the direct link to the original SOM images and the ICA projects the similarities between the samples in scale of the mutual similarities.

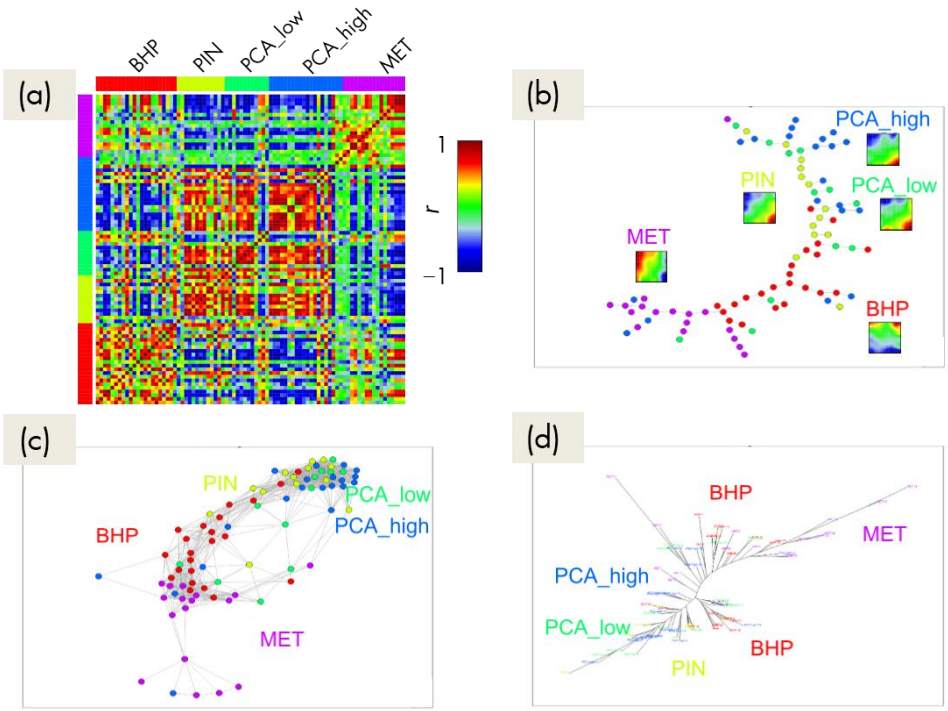


Figure 6: Similarity analysis of PCP. (**a**) The PCM visualizes the Pearson correlation coefficient of all pairwise combinations of sample portraits. Each subtype is characterized by a more or less pronounced brown-to-red square along the diagonal line, which reflects self-similarity of samples of the same type. Off-diagonal brown and blue regions refer to correlated and anti-correlated SOM-spot pattern, respectively. (**b**) The MST is shown together with the mean SOM portraits of each subtype. (**c**) The CN translates the PCM into a graph structure. The nodes are given by the samples and the edges connect positively correlated sample pairs ($r > 0.5$). (**d**) 'Phylogenetic' cluster tree.

3.6 DETECTING CO-REGULATED MODULES: 'SPOT' SELECTION

Method

The SOM algorithm arranges similar metagene profiles in neighbored tiles of the map whereas more different ones are located more distantly. In consequence, neighbored metagenes tend to be colored similarly owing to their similar values. Therefore, the obtained mosaic portraits show typically a smooth texture with red and blue spot-like regions referring to clusters of over- and underexpressed or hyper- and hypomethylated metagenes. These blurry images portray the feature landscape of each particular sample in terms of a

visual image. Metagenes from the same spot are co-regulated in the experimental series, whereas different, well-separated red spots in the same image refer to metagenes upregulated in the particular sample but differently regulated in other samples because of their different profiles. Each spot can consequently be interpreted as a module of a group of metagenes (and of associated single genes) showing concerted profiles.

We define over- (and under-) expression and hyper- (and hypo-) methylation spots by applying a simple 98th-percentile (and 2nd-percentile) criterion, which selects the respective fraction of the metagenes showing largest (or smallest) measures in each sample. Hence, the spots obtained are individual properties depending on the measure of the particular metagene in each sample. They can change their size from phenotype to phenotype and they can even disappear or transform from an over- into an underexpression spot and from hyper- into hypomethylation spot or *vice versa*.

The abundance of each spot is calculated as the relative frequency of appearance of each spot $s = A, B, \ldots$ in the samples of each cancer subtype,

$$x_{sc} = \frac{j_{sc}}{J_c} \qquad \textbf{Eq.(6)}$$

where J_c is the total number of samples per subtype c and m_{sc} is the number of portraits showing a particular spot s among those samples. The spot abundances are represented as stacked bar plot for each spot. The integral abundance, $X_s = \sum_c x_{sc}$, can be interpreted as the mean number of classes showing a particular spot. Its maximum value X^{max} equals the number of classes considered.

Example

In the next step, we analyzed the spot patterns of PCP SOM portraits to identify differences and common properties shared between the cancer stages. Unique or more common spots can provide information about the functional impact of gene activities specific to cancer subtype. Figure 7a shows the so-called overexpression summary map of PCP, which collects all spots with overexpression observed in the individual PCP portraits into one master map (see also [23]). Each distinct region of metagenes in the portraits exceeding the overexpression threshold defines a spot on the overexpression map, labeled by capital letters in Figure 7c. In total, we identified 15 such spots, 'A' to 'O', for PCP. Figure 7b visualizes the mean expression level across the metagenes of each spot for all samples. This heatmap thus provides an overview over the subtype-specific expression activity in each spot. For example, spot 'C' and partly also spot 'H' are selectively overexpressed in the BHP subtype, and spot 'N' in the PIN subtype, whereas spots 'A' and 'M' show sample specific activity, not specific to any subtype.

Our spot selection algorithm thus identifies both rare and frequent spot patterns. In the next step we assessed the relative frequency x_{sc} of each spot (see Eq.(6)). As shown in Figure 7d, only 6 out of 15 detected spots are relatively frequent ($x_{sc} > 0.2$) and the most

abundant spot 'N' is found in about 30% to 50% of all intermediate-staged samples, however, being relatively unspecific for tumor subtypes. Ten spots are observed in MET samples, reflecting that the expression patterns of the metastatic cancer samples are highly diverse with spots located in nearly all regions of the map. In contrast, the PIN and BHP samples show only 3 and 4 spots with overexpression, respectively. Analogous results from analyzing spots with underexpression are in line with these observations (data not shown here). Spots such as 'A', 'D', 'E', 'M', and 'O' are very rare with $x_{sc} < 0.1$.

In order to discover covariance between the metagene expression profiles in different spots, we calculated pairwise correlation maps and maximum spanning trees exploring relationships between spots (see supplementary material of [69]). As a rule of thumb, neighboring spots are strongly positively correlated and spots located in opposite corners of the map are often strongly anti-correlated. For example, spots 'E,' 'I', and 'L' are highly correlated (Figure 7c, blue dashed lines), whereas the spots 'N' and 'H' are anti-correlated (Figure 7c, red dashed lines).

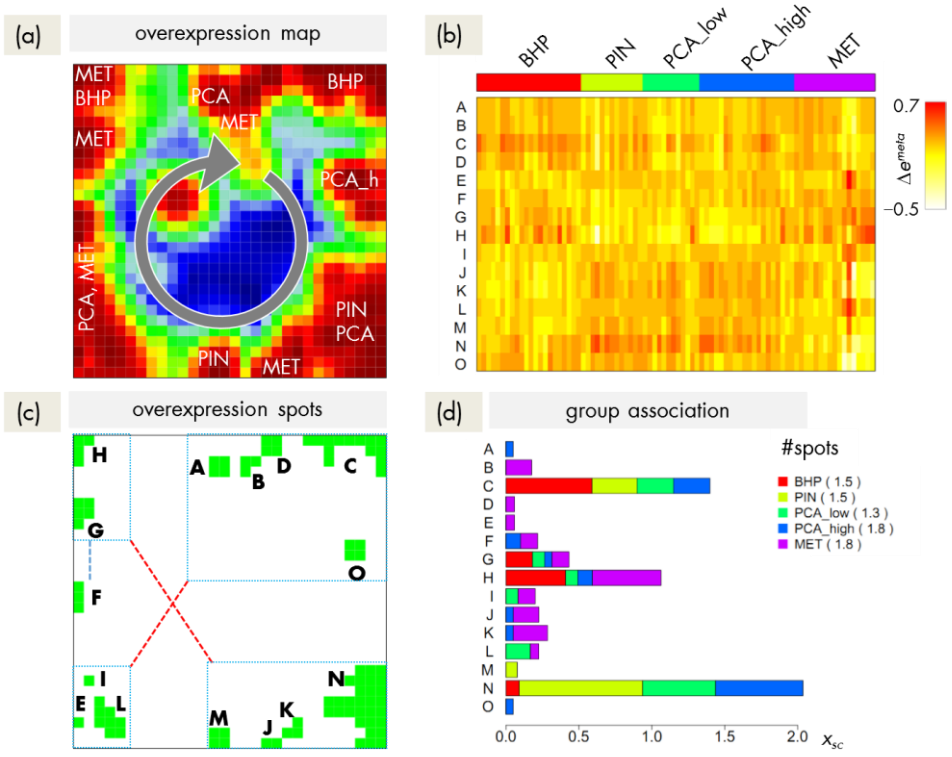


Figure 7: Overexpression spot characteristics of PCP. (**a**) In the overexpression summary map PCP stages associated with particular spots are indicated. The arrow represents the appearance of overexpression spots with cancer progression. (**b**) The heatmap shows the mean metagene expression for each spot 'A'…'O'. (**c**) Construction of the overexpression spot map defining the spots used for further analysis. Spots are labeled by capital letters. Correlated and anti-correlated spots are indicated by blue and red dashed lines, respectively. (**d**) The bar plot shows the fraction of samples of each subtype, which exhibits a given spot. The total bar length represents the overall frequency, while colors indicate the frequency by subtype. The average numbers of spots in the portraits of each subtype are given in parentheses in the top right legend.

3.7 FUNCTION MINING: GENE SET PROFILES AND POPULATION MAPS

Method

Co-regulated genes of each module can be assumed to be functionally related according to the 'guilt-by-association' principle [76]. Gene set analysis aims at identifying the functional context of these co-regulated modules. This method estimates the enrichment of groups of predefined genes (so-called gene sets) in gene lists, which are obtained independently, for example from SOM-spot analysis (see [77] for a critical review and references cited therein). A large and diverse collection of such sets can be derived from gene ontology (GO) gene annotation database [78] using the biomaRt interface [79]. Particularly, we included: (i) GO gene sets, composed of 'biological process' (BP), 'molecular function' (MF) and 'cellular component' (CC); (ii) canonical pathways, compiled from Biocarta, KEGG and Reactome; (iii) curated gene sets taken from the literature on chemical and genetic perturbations ('literature sets'); (iv) tissue specific gene sets determined previously [80]; and (v) 'special' gene sets taken from the literature on the cancer types addressed in the particular study.

The 'enrichment analysis' includes 'overrepresentation' analysis, 'overexpression' analysis, and their combination [80,81]. Overrepresentation estimates the probability to find members of a given set in a list, e.g., the genes included in a spot cluster, compared with their random appearance independent of their expression scores. We considered overrepresented sets with $p \leq 10^{-4}$, which ensures reasonable adjustment for false positives in the multiple testing problem. Contrarily, the term 'overexpression' defines the deviation between the mean expression value averaged over the set-members included in a spot-cluster and the mean expression value of genes independent of their overrepresentation. The gene set Z-score (GSZ) merges both gene set overrepresentation and overexpression approaches. For details see [80,82].

In addition to GSZ profiles we generated so-called gene set population maps to visualize the distribution of the genes of a selected set in the SOM portraits. This population map color codes the number of genes taken from the set in each of the tiles of the mosaic image. It ranges from white (no gene) to maroon (maximum number per tile observed for the particular gene set). Recall that each gene refers to one and the same metagene in all samples and thus it occupies a fixed position in all SOM portraits.

Finally we generated gene set overrepresentation heatmaps using an algorithm described previously [80]. We merged the top three gene sets per spot in a sample. Redundant gene sets were removed and represented by their minimum *p*-value. The resulted non-redundant global list of gene sets was converted into the GSZ enrichment heatmaps by applying hierarchical clustering.

Example

We applied gene set overrepresentation analysis to each spot-cluster using a collection of about 6000 predefined gene sets. Based on the functional context of the overrepresented gene sets obtained, we assign a short label to each detected spot (see Figure 8a).

'Inflammatory response' and 'cell division', two general hallmarks of cancer, are not among the leading gene sets in any of the spots. The respective GSZ profiles, however, show that 'inflammatory response' is selectively activated in the BHP and MET stages, whereas 'cell division' genes are overexpressed in the MET stage only (Figure 8c and d). The population maps of these gene sets indicate that the respective genes accumulate in the regions of more than one overexpression spot. For example, larger concentrations of genes related to 'cell division' are found in spots, for which the leading biological processes and cellular components are 'RNAPII activity' (spot 'G') and 'ribosome' (spot 'N'), respectively, whereas genes related to 'inflammation' accumulate in spots assigned to 'mitochondrion' (spots 'J' and 'K') and 'nucleosome' (spots 'A' and 'B') (Figure 8a).

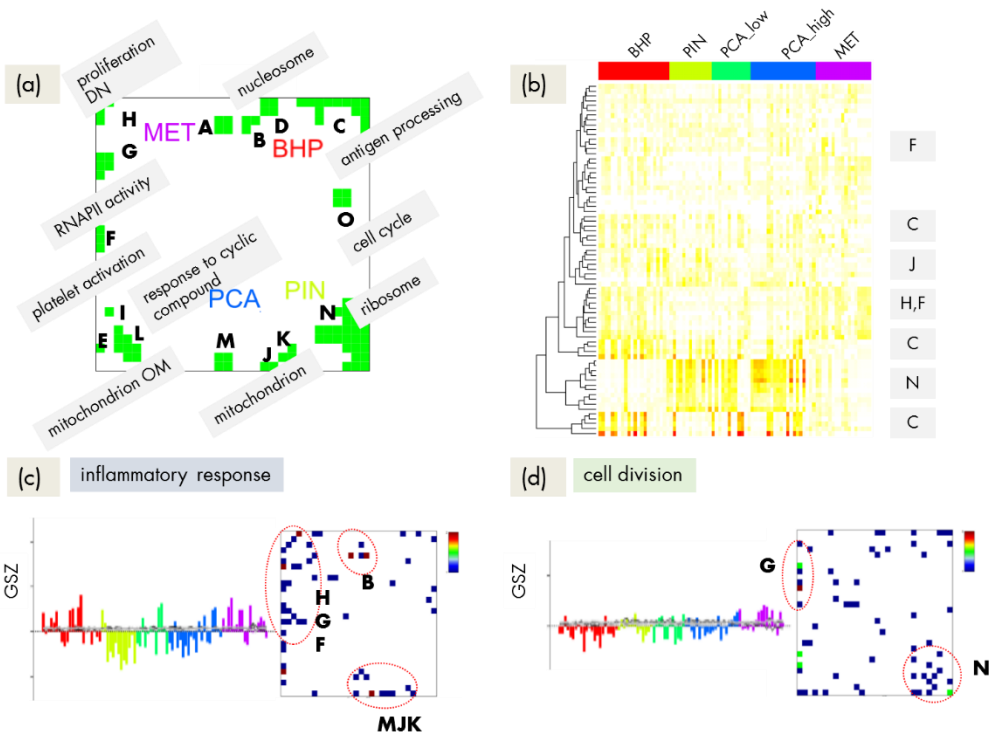


Figure 8: Gene set enrichment analysis of PCP. (**a**) The spot summary map shows the functional context of the most abundant spots (boxed labels) together with the associated stages (MET, BHP, PIN, PCA). (**b**) The overrepresentation heatmap of gene sets for the GO-term 'biological process' provides an overview. The letters on the right refer to the spots identified in (a). (**c**) and (**d**) Overexpression profile and map of the 'inflammatory response' and 'cell division' gene sets, respectively. The red dotted ellipses in the map indicate the spots of strongest enrichment. The full list of enriched gene sets, overrepresentation heatmaps of different gene set categories, and a gallery of the overexpression profiles and maps are given in supplementary material of [69].

For a more general overview of overrepresented gene sets, we generated gene set enrichment heatmaps to survey a larger collection of biological functions potentially contributing to the expression landscape. These heatmaps collect gene sets significantly overrepresented in the SOM portrait spots in a sample-specific fashion, and cluster them according to their degree of overrepresentation. Figure 8b shows the heatmap for gene sets associated with the GO-term 'biological process' and enriched in spots of the PCP SOM portraits. The one-way clustering separates the gene sets in agreement with their spot associations: For example spot 'N' mainly collects gene sets overexpressed in the PIN and also the PCA_low and PCA_high subtypes, whereas spot 'C' contains gene sets overexpressed in BHP. The heatmap also shows that gene sets from the spot 'H' tend to be overexpressed in MET. Detailed inspection of the heatmap reveals that the 'ribosome' spot 'N' contains additional gene sets such as 'translation' and 'gene expression'. These sets refer to different levels in the GO hierarchy, partly giving rise to overlapping groups of genes, which in consequence, trivially link similar expression patterns [83]. Here we neglect any interdependency due to such an overlap in gene sets, which may also arise across different GO categories and the curated gene sets from the literature. This redundancy might, however, highlight alternative aspects of annotated gene function.

3.8 MAPPING SUBTYPE-SPECIFIC SIGNATURE SETS

Method

To extract subtype-specific differential expression landscapes, we calculated difference maps, representing each metagene k in the mean SOM portrait of each subtype c according to:

$$diff_{kc} = \Delta e_{kc} - \mathrm{sign}(\Delta e_{kc}) \cdot \min(\max(|\Delta e_{kc'}|)_{c' \neq c}, |\Delta e_{kc}|) \qquad \textbf{Eq.(7)}$$

Eq.(7) selects specifically over- and underexpressed metagenes in a subtype. Particularly, $diff_{kc} > 0$ (or $diff_{kc} < 0$) means the expression of subtype c in metagene k exceeds (or falls below) the respective metagene expression in all other subtypes considered. $diff_{kc} = 0$ is obtained, if the relative expression of the metagene selected is unspecific for subtype c.

Example

PCP signature genes associated with different molecular concepts such as 'gluthathione metabolism' (specifically overexpressed in BHP), 'androgen signaling' (overexpressed in PIN and PCA_low), 'protein biosynthesis' (overexpressed in PIN and PCA), and 'cell cycle' (overexpressed in MET) were taken from publication [68]. We calculated the GSZ scores for those gene sets. The GSZ profiles in the left panels of Figure 9 confirm enrichment of expected biological pathways in the respective PCP stage and underexpression in the re-

maining stages. The location of the considered subtype-specific gene sets in the SOM landscape can be found in the population map in the next panel. The red dotted circles indicate the highest populated regions per gene set confirming that genes from these sets accumulate within the subtype-specific overexpression spots extracted from the SOM-portraits (see mean portraits).

To extract unique, subtype-specific spot patterns, we calculated difference maps (Figure 9, rightmost columns). Red, blue and white colored regions refer to positive, negative and indifferent $diff_{kc}$–values, respectively. The spots detected in the difference maps largely agree with features seen in the mean portraits of the respective subtypes. Non-specific features, however, (such as the spot 'N', found in several subtypes, Figure 7b and d) disappear by applying Eq.(7), as expected.

Hence, the simple tile-by-tile processing of metagene expression values identifies regions of class-specific over- and underexpression in the SOM-portraits. These regions are confirmed when compared with gene sets extracted from independent statistical analyses applied in the original paper. The mean and difference portraits thus provide a simple and intuitive aproach to localize class-specific spots in the maps.

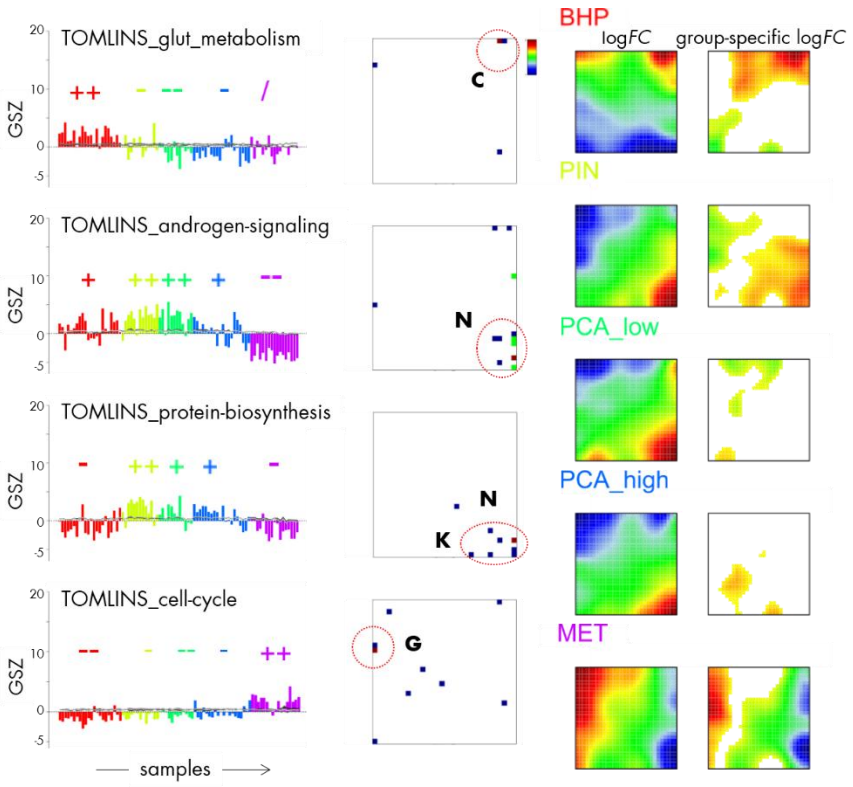


Figure 9: Stage-specific differential expression of PCP reported earlier in [58]. GSZ profiles; corresponding gene set population maps; subtype mean in log*FC*-scale; and difference portraits. In the GSZ profiles, each bar represents one sample, color coded according to subtype-membership. +, -, / signs above indicate over-, under- and indifferent expression, respectively. The red dashed ellipses in the gene set population maps indicate gene sets accumulating in distinct regions of the map, which to a good approximation agree with the subtype-specific spots in average and difference portraits.

4 B-cell lymphomas

B-cells are lymphocytes that are an essential component of the adaptive immune system. Immature 'naïve' B- (NB) cells are produced in the bone marrow, which then migrate to germinal centers (GC) where they differentiate into mature B-lymphocytes (Figure 10). These GC are central to the formation of B-cell-mediated immunity: B-cells undergo immunoglobulin somatic hypermutation and clonal expansion via intense proliferation in the dark zone. Subsequently they migrate to the light zone where they transform into long-lived memory B-cells and terminally differentiate to plasma cells that produce high-affinity antibodies. B-cell development is a multi-level process, which is driven by epigenetic regulation, incorporating DNA methylation and histone modifications, to induce the cell-specific gene expression pattern [84].

Dysfunction of epigenetic regulation represents a common and important feature of B-cell lymphomas. For example, GCB-cells are prone to instability in their cytosine DNA methylation patterns leading to aberrant methylation patterns in lymphoma, which display variable degrees of epigenetic heterogeneity [85–87]. Moreover, polycomb group (PcG) proteins, a subset of histone-modifying enzymes known to be crucial for B-cell maturation and differentiation, play a central role in malignant transformation of B-cells [88]. Genes *de novo* methylated in all lymphoma enrich in polycomb targets and share a similar stem cell-like epigenetic pattern [86]. Available evidence suggests that different diseases arise from oncogenic B-cell clones at a distinct stage of differentiation, ranging from NB-cells to plasma cells. These tumors of the lymphoid tissues represent one of the most heterogeneous malignancies owing to the wide spectrum of types of B-cells from which they can arise and also due to the heterogeneous microenvironment in the lymphatic organs providing a multitude of different niches for tumor progression. Many B-cell malignancies derive from germinal center B-cells, most likely because of the high proliferation rate of these cells and the high activity of mutagenic processes. This category includes diffuse large B-cell lymphomas (DLBCL), follicular lymphomas (FL), Burkitt's lymphomas (BL) and mantle cell lymphoma (MCL). Mature B-cell malignancies in addition include leukemias derived from B-cells that

have passed through the GC such as B-cell chronic lymphocytic leukemia (B-CLL), which is a stage of small lymphocytic lymphoma. Multiple myeloma (MM) is an incurable B-cell neoplasia arising from malignant plasma cells, which originates in illegitimate immunoglobulin heavy chain (*IGH*) switch recombinations.

Morphologic features of lymphomas resemble lymphocytes at distinct differentiation stages serving as basis for their histological classification. Alternatively, the rapidly emerging information obtained from molecular high-throughput gene expression studies creates a series of expression-based classification schemes [59,85,89–91], which distinguish, for example, molecular Burkitt's lymphoma (mBL), non-mBL, and intermediate lymphoma (IntL).

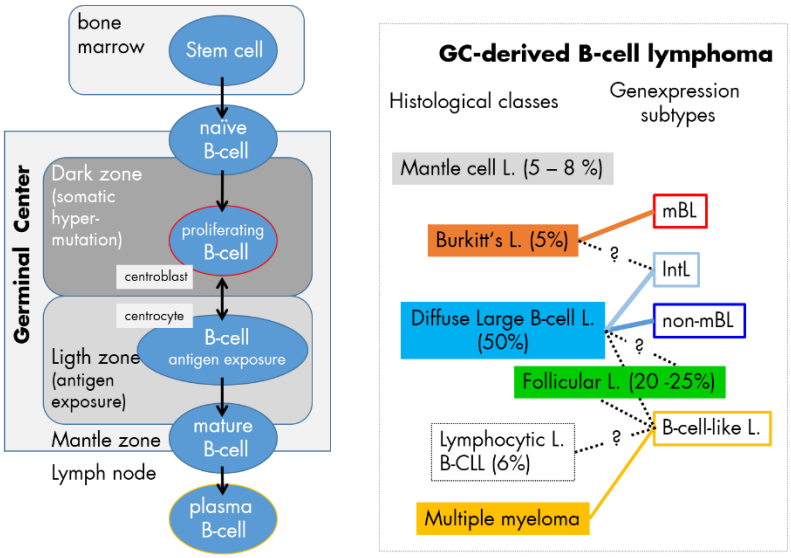


Figure 10: Developmental and maturation stages of B-cells provide a wide spectrum of cell-of-origin- and micro-environmental conditions for different histological classes of B-cell lymphoma. Their relation to the histological classes is partly unclear mainly due to the absence of clear-cut borderlines between the molecular and histological signatures and because of transformations between the classes. Incidence rates in percent of all B-cell lymphoma were taken from http://www.cancerresearchuk.org/cancer-info/cancerstats/types/nhl.

This chapter is based on the following 3 scientific publications:

Hopp, L., Lembcke, K., Binder, H., & Wirth, H. (2013). Portraying the Expression Landscapes of B-Cell Lymphoma- Intuitive Detection of Outlier Samples and of Molecular Subtypes. *Biology*, 2(4), 1411-1437.

Hopp, L., Nersisyan, L., Löffler-Wirth, H., Arakelyan, A., & Binder, H. (2015). Epigenetic Heterogeneity of B-Cell Lymphoma: Chromatin Modifiers. *Genes*, 6(4), 1076.

Hopp, L., Löffler-Wirth, H., & Binder, H. (2015). Epigenetic heterogeneity of B-cell lymphoma: DNA-methylation, gene expression and chromatin states. *Genes*, 6(3), 812-840.

4.1 GENE EXPRESSION LANDSCAPE OF LYMPHOMAS

Our application of SOM machine learning to lymphoma expression data aims at characterizing the heterogeneity of the genome wide expression landscapes and at describing the molecular cancer subtypes. Further, we describe how to detect and to correct outlier samples using their portraits. Finally, we propose a more detailed molecular subtype classification of the lymphoma samples. The classification of the lymphoma samples was used as given in Hummel et al. [59]: molecular Burkitt's lymphoma (mBL), non-mBL and intermediate. For details concerning the cohort and preprocessing of the data see section 3.1 and supplement section 7.1.2.

4.1.1 SOM PORTRAITS

We applied the SOM machine learning algorithm as described in section 3.2 that transforms the whole genome expression pattern of the 22,283 'single' genes into metagene expression data of dimension $K \times J = 2500 \times 221$.

Figure 11 shows the expression portraits of selected lymphoma samples arranged according to their previous classification into subtypes [59]. The individual portraits reveal a handful of clusters of co-expressed metagenes frequently observed. These over- and under-expression spots selectively characterize the different lymphoma subtypes: Samples of the mBL and non-mBL subtypes are mostly characterized by spots of overexpressed metagenes in top-right and bottom-left corners of the map, respectively. However, many additional spots can be observed in the portraits, indicating additional functional modules activated in the respective samples. Samples of the intermediate subtype show more volatile patterns with overexpressed metagenes frequently tending to occupy the top-left and bottom-right corners of the SOM. The full gallery of the 221 SOM portraits and supporting maps are given in the supplementary file of [91].

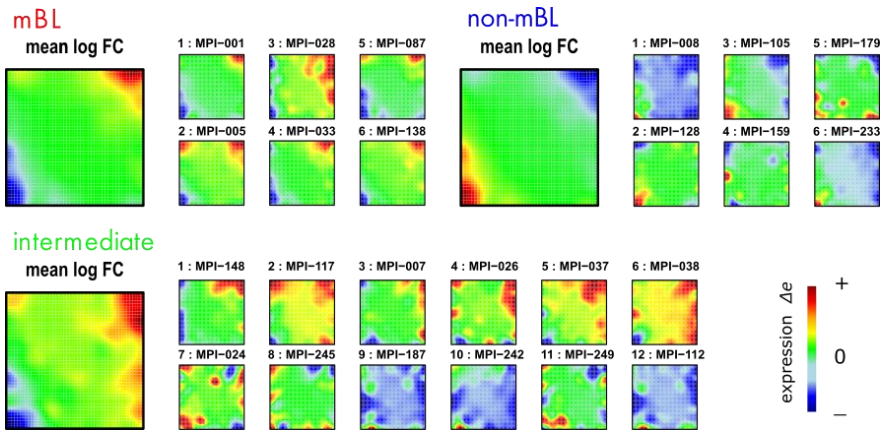


Figure 11: SOM gallery of lymphoma subtypes with a resolution of 50 × 50 metagenes: The small mosaic images refer to selected individual tumor samples assigned to the mBL, non-mBL and intermediate subtypes. The larger images represent the respective mean subtype portraits.

In support of the observations from the individual portraits we found that the mBL and non-mBL subtypes are characterized by two spots in opposite corners of the map: One spot in the top-right corner is overexpressed and the other one in the bottom-left corner is underexpressed in mBL samples and *vice versa* in non-mBL samples, revealing the antagonistic character of their expression patterns (see large mean subtype portraits). These subtype-specific spots collect highly populated, highly variable and well resolved metagenes (data not shown here).

4.1.2 SAMPLE DIVERSITY: THE THREE-SUBTYPE APPROACH

We generated a PCM (Figure 12a), which visualizes the correlation of all pairwise combinations of sample portraits (see section 3.5). The compact red square of mBL sample couples reflects the strong similarity between their expression landscapes whereas the blue off-diagonal area formed between the mBL and non-mBL samples indicates their anti-correlated expression states. Note that the pairings between non-mBL samples, although correlated, reveal a much fuzzier pattern due to the more heterogeneous expression states compared to the mBL subtype. The samples of the intermediate subtype either correlate with the mBL or non-mBL samples or with both in some cases.

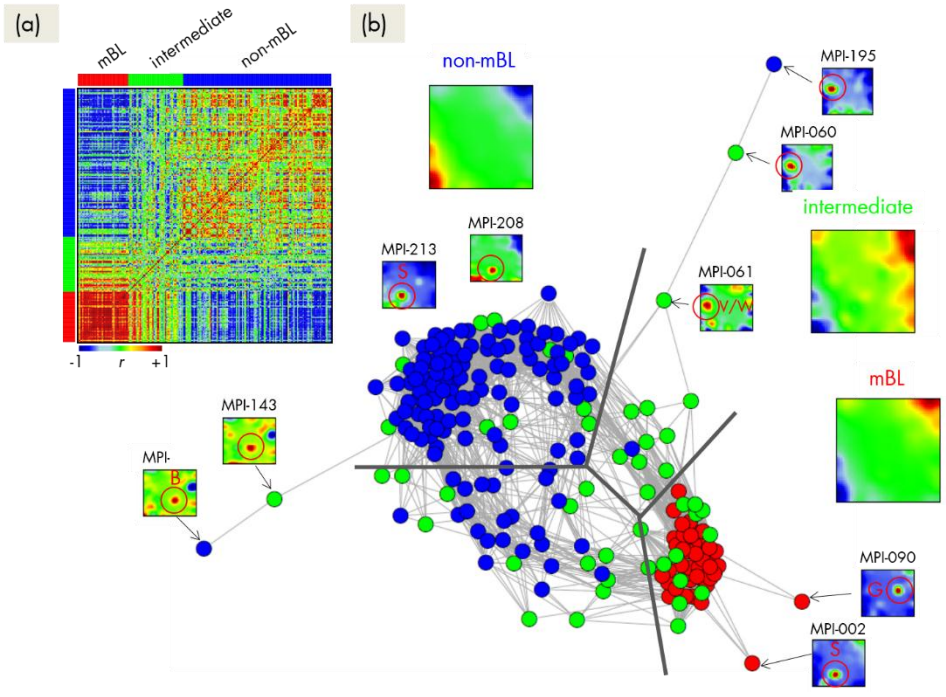


Figure 12: Pairwise correlation analysis of all lymphoma samples: (**a**) In the PCM red colors indicate positive and blue colors negative correlations between the samples. (**b**) The correlation network: Mean subtype portraits are given within the figure (large maps). Outlier nodes are highlighted by arrows. The SOM portraits of the respective samples are shown by small maps. The red circles and the spot letters (as assigned in section 4.1.3) indicate the outlier spots differing from the subtype-specific patterns (compare these individual sample portraits with the mean subtype portraits).

Visual inspection of the CN in Figure 12b (for details see section 3.5) shows that the mBL and non-mBL samples accumulate into well separated clusters whereas samples of the intermediate subtype heterogeneously spread over the region between these two clusters. Interestingly, these intermediate samples distribute along two disjunctive branches of the CN, which both link the mBL and non-mBL clusters. These two separate branches also include a fraction of the mBL and non-mBL samples (see the dark grey lines in Figure 12b roughly separating the clusters and branches). This distribution of the intermediate subtype samples reflects the heterogeneous spot characteristics of the subtypes as discussed above.

A few samples are located far away from their subtype-specific cluster and/or from the majority of the other samples in the CN. Those samples are usually characterized by rare or unique spots as indicated in Figure 12b. We will address this issue in more detail later.

In summary, the CN of lymphoma samples forms a 'donut-like' structure composed of alternating compact and fuzzier clusters. The former ones refer to the main subtypes and the latter ones to two distinct groups of samples mainly assigned to the intermediate subtype. The mutual correlation analysis as seen by the CN in combination with the SOM portraits thus provides additional information complementing other similarity analyses applied (not shown here).

4.1.3 CLUSTERS OF CO-EXPRESSED GENES CHARACTERIZE THE SUBTYPES AND OUTLIER FEATURES

We analyzed the spot patterns in order to identify specific properties of the lymphoma subtypes. Figure 13a shows the overexpression summary map (see [23] and section 3.6), collecting 23 overexpression spots labeled with capital letters 'A' – 'W' (Figure 13b).

Recall that our spot selection algorithm neglects the abundance of each spot in the individual portraits and identifies both rare (e.g., observed in only one sample) and frequent spot modules. The overexpression heatmap in Figure 13c visualizes the spot expression profiles. The color ranges from blue representing the lowest mean expression values, to red representing the highest values. The heatmap provides an overview of the degree of subtype-specific expression in each of the spot modules. For example, spot 'L' and partly also spot 'K' are selectively overexpressed in samples of the mBL subtype, while spot 'O' is characteristic for the non-mBL subtype. Contrary, more ubiquitous spots as 'N' as well as rare spots as 'A' or 'G' lack of subtype-specific overexpression. Note that frequent spots are usually located in the peripheral part of the map (*i.e.*, in the corners and along the edges) whereas rare spots tend to accumulate in the central part.

We use the spot information and the mean subtype portraits to assign subtype labels to the most prominent and specific spot modules (Figure 13a): Spots 'L' and 'K' are ascribed to mBL while spot 'O' is prominent in non-mBL samples. Those three spot modules contain marker genes overexpressed in the respective subtypes as validated below. Spots 'J' and 'Q', also frequently observed in the sample portraits, are assigned to the intermediate

subtype. Interestingly, they constitute two alternative intermediate states located in between the main subtypes mBL and non-mBL as indicated by the arrows in Figure 13a.

Please recall that the training algorithm distributes the metagenes in such a way that strongly correlated profiles are located at adjacent positions in the map whereas metagenes with anti-correlated profiles tend to occupy more distant regions, e.g., in the opposite corners of the map. The results of the spot correlation analysis are visualized in Figure 13b. One sees that, for example, the mBL marker spots 'K' and 'L' are highly correlated and usually appear together in the sample portraits whereas the anti-correlated overexpression spots 'K' and 'O' will not be observed together in the same expression portrait.

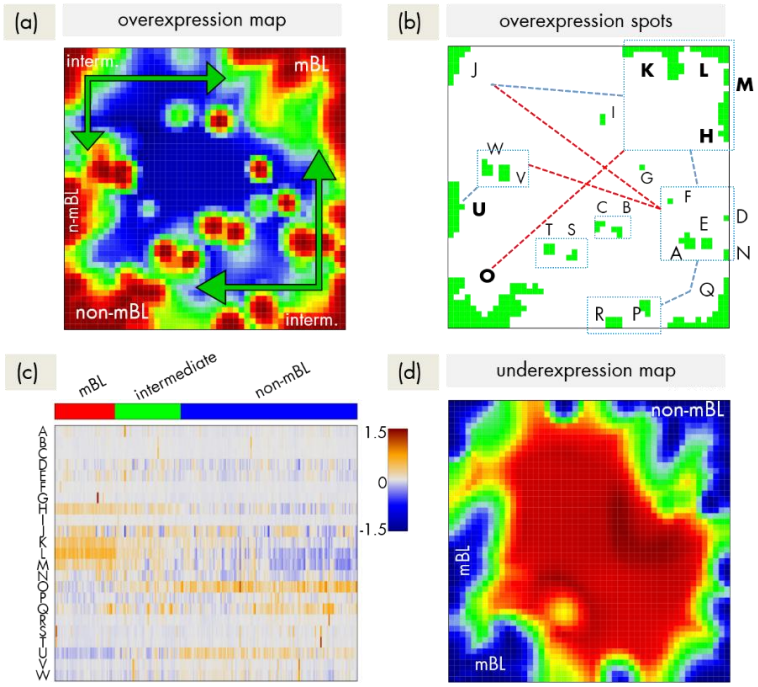


Figure 13: Spot module characteristics of lymphoma: (**a**) The overexpression summary map of lymphoma. Subtypes frequently showing the respective spots are indicated. (**b**) The overexpression spot map and (**c**) the overexpression heatmap. For details see caption of Figure 7. (**d**) The underexpression summary map collects all underexpressed spots observed in the individual portraits. Note the antagonistic nature of mBL and non-mBL expression: Spots overexpressed in mBL become underexpressed in non-mBL and *vice versa* (compare with panel a).

For this data set, we also detected 11 global underexpression spots emerging as blue regions in the SOM portraits (see Figure 13d). Position and size of most of the detected underexpression spots agree with those of the overexpression spots. Hence, overexpression of the respective metagenes in part of the samples changes into underexpression in other samples. For the analyses described in this study, we therefore use only the overexpression spots detected without loss of essential information. Interestingly, virtually no blue underexpression spot was detected in the central area of the map indicating that the rare overexpression spots do not show this dualism. Below we will show that these spots potentially constitute clusters of outlier genes, the expression of which is affected by bias effects.

In summary, the heterogeneous expression patterns observed in the individual portraits can be condensed to a few major expression modules represented by over- and underexpression spots. This way the relevant dimension of the data set is reduced by three orders of magnitude from about 20,000 single genes to approximately 12 frequent spot modules.

4.1.4 FUNCTION MINING: INFLAMMATION-VERSUS-PROLIFERATION

According to section 3.7 we applied gene set overrepresentation analysis to each spot-cluster taking into account a collection of more than 6,000 predefined gene sets. For each spot we obtained a list of gene sets ranked according to increasing *p*-value estimating the probability that genes of the set are found within the spot by chance.

We assigned a short notation to each of the spots (see Figure 14a) putting the genes accumulated in the respective spot into functional context. Some spots are obviously related to processes associated with general hallmarks of cancer such as 'inflammation' and 'cell division' (spots 'O' and 'K' respectively). Panel b of Figure 14 depicts the GSZ-expression profiles (left part) and the population maps (right part) of those two leading gene sets. The profiles clearly reflect the fact that the respective processes are selectively over- or underexpressed in a subtype-specific fashion. While 'inflammatory response' is activated in the non-mBL subtype, genes annotated to the gene set 'cell division' are active in the mBL subtype. The respective gene set population maps reveal that the associated genes accumulate in the regions of spots overexpressed in the respective subtype, as expected.

Neighboring spots of strongly correlated profiles can be assigned to related biological processes: The 'cell division' spot is surrounded by spots assigned to 'transcription factor binding', 'chromatin' and 'transcription' according to the most overrepresented gene sets in each of the spots. Note that, although related, these neighboring spots are usually characterized by subtle differences in their expression profiles and presumably also by fine differences in the functional context of the overrepresented gene sets. Population maps and overexpression spot maps therefore represent complementary tools for discovering the functional context of the expression landscapes. The results so far show that the lymphoma samples split into pairs of subtypes differing by the antagonistic activation of processes related to 'inflammation' and 'immune response' on one hand and to 'cell division' and the 'transcriptional and translational machinery' on the other hand (non-mBL-vs-mBL).

To validate the subtype-specific spot patterns identified above, we included the signature set that differentiates between mBL and non-mBL subtypes provided by Hummel et al. [59] (see Figure 14c). As expected, genes of this set clearly accumulate in the subtype-specific spots 'L' and 'O' assigned to mBL and non-mBL, respectively.

Another important question is about the possible origin of the rare spots in the central part of the map. In Figure 14d, we show the characteristics of two gene sets related to tissue specific gene expression in 'tonsils' [23,80] and to 'drug response' ('drug metabolism, cytochrome P450 (CYP)', see [92]), respectively. Their genes strongly accumulate in localized regions of the map agreeing with the positions of the rare spots 'S' and 'G',

respectively. Both gene sets are overexpressed in only few samples suggesting that the respective samples are outliers contaminated either with healthy tissue or affected by patient specific medication. As those effects are not related to the cancer studied they reflect systematic biases of the respective expression patterns.

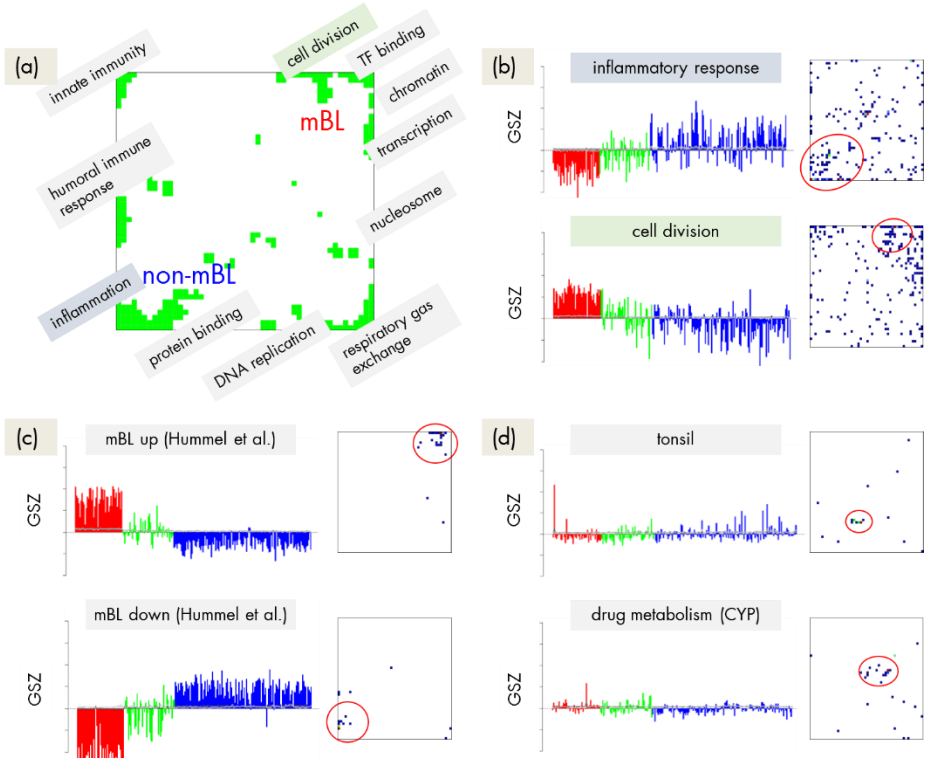


Figure 14: Functional analysis of lymphoma: (**a**) The functional context of the most abundant spots is assigned according to the topmost overexpressed gene sets in each of the spots. (**b**) – (**d**) GSZ profiles and population maps are shown for gene sets accumulating in the mBL and non-mBL specific overexpression spots as indicated by the red ellipses (panel b), for mBL-vs-non-mBL signature sets published previously [59] (panel c) and for sets accumulating in rare spots (panel d).

4.1.5 DETECTION AND CORRECTION OF OUTLIERS

Large tumor sample collections are prone to different effects not (or not directly) related to the expression profiles of the diseased tissue such as contaminations with healthy tissue (brain, blood etc.), different levels of RNA quality after extraction and wet lab preparation, technical biases due to day-to-day variations of hybridizations, and data recordings. Moreover, biological patient-to-patient variance is typically high and can be caused by other factors than the disease under study. The noisy character of the GSZ profiles and also the scatter of the global expression characteristics manifest this variability of the data. The development, selection, and appropriate application of suited methods of quality control aiming at identifying, understanding, and possibly also removing such effects represent a separate complex topic not addressed here in detail. However, our portraying approach offers a simple and direct option to check the whole-genome expression landscapes of the

individual samples by visual inspection of their molecular 'faces'. Particularly one searches for conspicuous spot patterns that clearly deviate from that of the majority of samples assigned to the same class.

Inspection of the CN in Figure 12b reveals a series of samples, which are located outside of the main network body. The portraits of these outlier samples exhibit overexpression spot patterns deviating from the subtype-specific patterns identified in terms of their mean SOM portraits. Particularly the spots ´G´, ´S´, and ´W´ are identified in the outlier sample portraits (red circles in Figure 12b; see Figure 13b for spot-letter assignments). Here, we exemplarily focus on spot ´S´, located in the bottom-left region of the SOM being strongly overexpressed in samples MPI-002, MPI-208, and MPI-213 (see Figure 12b). The topmost enriched gene set in this spot is the ´tonsil´-set. It was extracted as the tonsil-signature from a large expression data set of healthy human tissues previously analyzed with the SOM pipeline [23,80]. Enrichment of this set suggests that overexpression of spot ´S´ is caused by contamination of the tumor biopsy with adjacent healthy lymph node tissue.

Panel a in the left part of Figure 15 shows the GSZ profile and the population map of the ´tonsil´-set. The GSZ profile reveals strong overexpression of the set in a number of samples independent of their subtype assignment. The corresponding genes mainly accumulate in spot ´S´. Selected samples, which possess this particular spot in their portraits are shown in panel c. They can already be identified as potential outliers by simple visual inspection of the SOM portrait gallery (see supplementary material of [91]). We highlighted the samples in the GSZ profile (panel a) and in the CN (panel b) by arrows. Note however that not all of these samples protrude as clear outliers in the CN. Despite the strong overexpression of the contamination spot ´S´, the overall expression state of e.g. samples MPI-208 and MPI-213 obviously resemble those of the unbiased samples.

In a simple correction step we removed the genes included in the outlier spot from the whole data set (see red circle in the population maps in Figure 15). This procedure can be repeated for other contamination spots identified: For example, spot ´G´ was found to be related to 'drug metabolism' (´cytochrome p450´, see Figure 14d and sample MPI-090 in Figure 12b), presumably due to individual medication of the patient. Spots ´V´/´W´ show an intense increase in expression of the 'G-antigen-family' for unknown reasons (samples MPI-060, MPI-061, and MPI-195 in Figure 12b).

After removing strongly biased genes from the training data, we generated a new SOM. Note that, depending on the purpose, also re-evaluation of only parts of the analyses may be sufficient. The right part of Figure 15 shows the results after correction for tonsil-contamination accumulated in spot ´S´. The corresponding GSZ profile shows a more uniform expression of the gene set after correction. The respective sample portraits now show the characteristic spot signatures of the respective subtypes, *i.e.*, of mBL for MPI-002 and non-mBL for MPI-208 and MPI-213. Especially the outlier sample MPI-002 is now located within the mBL cluster in the CN, such that it attains a more compact shape.

In summary, the combination of individual portraits, enrichment analysis and the CN provides a framework for easy and intuitive detection of outlier spots and samples. After correction, more reliable expression landscapes of the samples are obtained.

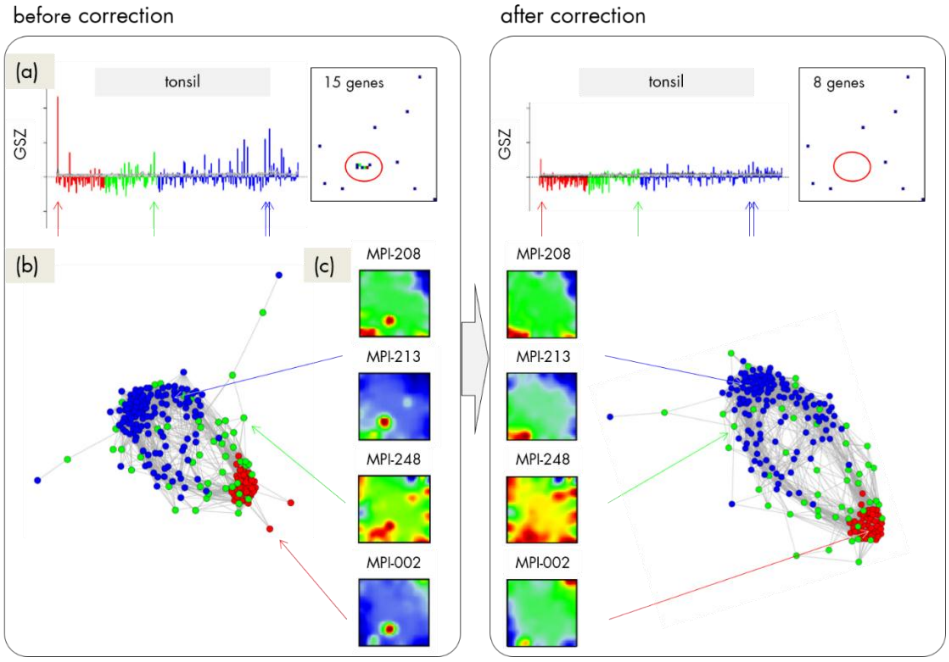


Figure 15: Correction of outlier samples contaminated with healthy lymph node tissue. The left and right parts of the figure refer to the uncorrected and corrected data, respectively. (**a**) GSZ profile and population map of the ´tonsil´ gene set: The signature is not characteristic for one of the subtypes and their genes accumulate in spot ´S´ of the map. (**b**) Correlation network of the lymphoma data set. (**c**) SOM portraits of selected outlier samples. The arrows point to the position of these samples in the CN and in the GSZ profile.

4.1.6 ALTERNATIVE SUBTYPING OF B-CELL LYMPHOMAS INTO FOUR SUBTYPES

Our analysis so far suggests that the samples assigned to the intermediate subtype split up into two separate branches. These two branches are characterized by overexpression spots in the bottom-right (spot 'Q') and top-left (spot 'J') part of the expression portraits, respectively (compare the first and the second row of the intermediate sample portraits in Figure 11). CN analysis clearly shows two distinct sample groups forming two continuous transition ranges linking the compact mBL and non-mBL clusters. These transition ranges include samples of the intermediate and also of the mBL and non-mBL types (Figure 12b). These results suggest the existence of four subtypes partly differing from the classification into three subtypes discussed so far.

In order to further verify this hypothesis, we applied a modified 'prototype-guided' k-Means clustering of the metadata to segregate the samples into these four subtypes. k-Means is an algorithm, which iteratively assigns the samples to so-called cluster prototypes showing the minimal mutual Euclidean distance and subsequently computes new

prototypes as the centroids of the members of each cluster [93]. k-Means requires predefinition of a desired cluster number, while the initial prototypes are usually chosen randomly or initialized from the data [94]. The SOM portraits now constitute another option to initialize the prototypes: They can be established using selected expression patterns observed in the portraits such as the most prominent overexpression spots. Particularly, we define initial prototypic expression portraits showing a selected spot pattern for each subclass with values $\max(\Delta e_{kj})$ for metagenes within the spot and 0 for metagenes outside. These prototypic spot patterns are then used to assign the samples to the respective clusters in the standard k-Means algorithm.

In this study spot 'K' initializes the new mBL-like subtype mBL*, spot 'O' the non-mBL-like subtype non-mBL* and spots 'J' and 'Q' the two new intermediate subtypes intermediate A and intermediate B, respectively. Figure 16a shows the obtained four cluster centroids after convergence of the k-Means algorithm. They represent the mean portraits of the four new subtypes mBL*, intermediate A, intermediate B and non-mBL*. Note that the mean portraits of the mBL* and non-mBL* subtypes closely resemble that of the initial mBL and non-mBL classes, respectively (compare with Figure 11). In contrast, the mean portraits of the new intermediate A and intermediate B subtypes clearly differ from that of the initial intermediate subtype and from that of the mBL* and non-mBL* patterns.

We re-colored the CN plot according to the new subtype classification (Figure 16b). The mBL* and non-mBL* clusters are more compact compared to the initial mBL and non-mBL clusters (compare with Figure 12b). The expression landscapes of the new groups obtained are obviously more homogeneous. The samples of the two intermediate subtypes accurately accumulate along the two separated branches linking the mBL* and non-mBL* clusters except a certain region of overlap in the center of the CN.

In the next step, we compare the robustness of the old and new subtype cluster assignments by applying the bootstrap clustering approach. Therefore, k-Means clustering is repeatedly applied to a subset of samples chosen randomly from the complete set of samples. The mean metagene expression states of the subtypes are used as initial cluster prototypes. The fraction of proper assignments of samples in agreement with their actual class assignment then defines a robustness score of each sample: A bootstrap stability score of 1 means that the respective sample is always found in the correct subtype, while a score of 0.5 means that the sample is assigned properly in only 50% of the resampling repetitions. For the previous classification into three subtypes, the stability scores of the intermediate and non-mBL subtype samples show a broad distribution with scores of 0.5 and below. The new four subtype classification is clearly more robust ($p < 10^{-4}$, Wilcoxon signed-rank test), reflecting a more consistent and stable clustering of the samples (Figure 16c). Only a small number of relatively uncertainly assigned samples are found even in the transition ranges between the different clusters.

To further validate the results of our k-Means approach by an independent method we applied consensus clustering [95] (see supplement section 7.3), which supported the four-class approach.

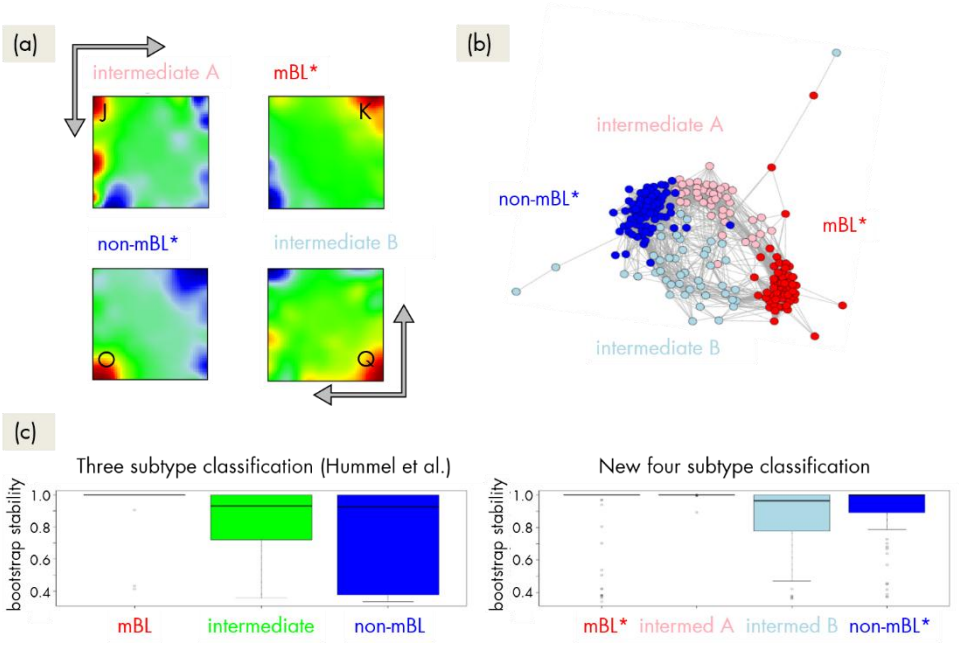


Figure 16: k-Means clustering into four lymphoma subtypes: (**a**) Mean expression portraits of the four new subtypes. The grey arrows indicate the spot pattern transitions from mBL* to non-mBL* via intermediate A or B. (**b**) CN colored according to the new subtypes obtained. (**c**) Bootstrap stability score of three (left part) and four (right part) subtype classification.

4.1.7 CHARACTERIZATION OF THE NEW SUBTYPES

The four new subtypes are defined by their distinct expression patterns and their particular functional contexts, *i.e.*, they represent molecular subtypes. The question arises if these molecular subtypes associate with selected genetic, clinical, or alternative molecular phenotypes collected independently [59]. Previously published patient phenotypic data was used to characterize the newly defined subtypes in the cohort studied [96]. These included data from immunohistochemical staining against *CD10*, *BCL2*, *BCL6*, and *MUM1*, data from interphase fluorescence *in situ* hybridization (FISH) for *IGH*, *MYC*, *BCL6* and *BCL2* loci, overall survival, age, and gender. We calculated the frequency distribution of patients for each of the characteristics over the four subtypes. Table S 3 reveals associations between these characteristics and the subtypes in terms of enriched or depleted patient numbers (*p*-values are obtained from Fisher's exact test).

For mBL* and non-mBL* one finds analogous frequency distributions of a series of characteristics as described in previous studies, e.g., the age dependency [59], the effect of the *MYC*-gene translocation [59], different immune-phenotypes [97] and the GCB-ABC-signature [98]. Nearly 90% of the lymphoma samples assigned to the non-mBL* and to intermediate A & B subtypes are classified as diffuse large B-cell lymphoma (DLBCL) suggesting a close similarity between these three subtypes. A series of characteristics such as the IG-MYC status and immune-phenotypes *CD10*, *BCL6*, and *BCL2* support this result.

However, the new intermediate A and B subtypes also show specific properties. Interestingly, the tumors with the activated B-cell (ABC) signature are clearly overrepresented in the intermediate A subtype, whereas the alternative germinal center B-cell (GCB) signature clearly depletes in this subtype. They also show differential characteristics with respect to the appearance of genetic aberrations (*MYC* translocation and immunoglobulin heavy chain (*IGH*) break) and to the *BCL2* immune-phenotype: Firstly, the IG-MYC translocation is more frequently found in the intermediate B subtype compared with the intermediate A and the non-mBL* lymphoma. Secondly, intermediate A lymphomas less frequently show the *IGH* break and the *BCL2*+ immuno-phenotype than the other subtypes. Thirdly, intermediate B and non-mBL* lymphomas possess slightly enriched populations of t(14;18)(q32;q21) translocations, which juxtapose the *BCL2* oncogene to the *IGH* locus.

It turned out that each of the subtypes is characterized by different hallmarks of cancer, e.g., proliferation and high transcriptional and translational activity in mBL*; activated immune response and inflammation in non-mBL*, innate immunity in the intermediate A subtype and up-regulated expression of common cancer gene signatures [99] in the intermediate B subtype. Generic, *MYC*-related poor prognosis gene signatures [100] are associated with the mBL* and, to a lesser extent, intermediate A subtypes. Moreover, we found that intermediate A subtype lymphomas show expression signatures of activated B-cells and strong dissimilarity with expression landscapes of GCB-cells and healthy lymph node tissue suggesting different cell-of-origins. On the level of gene regulation, the decomposition of lymphoma into four subtypes obviously further diversifies into different modes, which in turn reflect driving effects on the genetic and epigenetic levels. The understanding of these molecular mechanisms thus requires the combined analysis of genetic, epigenetic and transcriptional data.

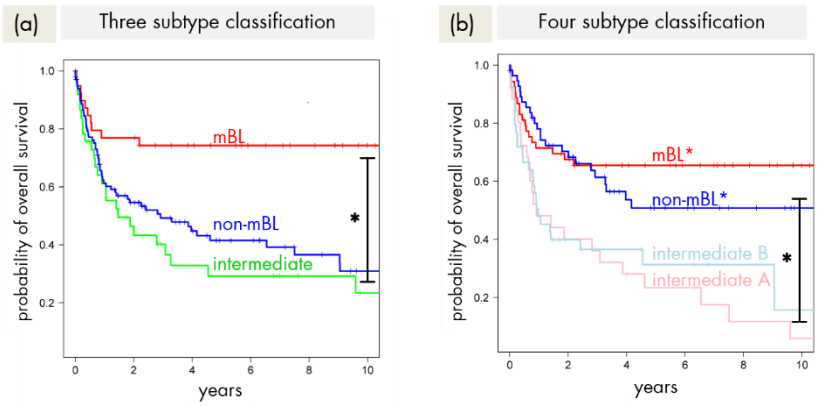


Figure 17: Kaplan-Meier survival curves of lymphoma of (**a**) the original three subtypes and (**b**) the new four subtype classifications. Tick marks indicate patients alive at the time of last follow-up. Subtype-specific survival curves are compared using log-rank test and the respective *p*-values are indicated within the figures.

Finally, we generated Kaplan-Meier diagrams to estimate the probability of subtype-specific overall patient survival as a function of time [101]. Figure 17a and b show the curves for the three and four subtype classifications, respectively. Based on the original definition by Hummel et al., patients with mBL lymphomas show significantly better survival rates as intermediate and non-mBL patients ($p < 10^{-3}$ in log-rank test, see also [59]). In contrast, our new classification now reveals that both mBL* and non-mBL* patients show better survival rates than patients of the intermediate A & B subtypes. Assignment of lymphoma to either of the two intermediate subtypes roughly halves the survival rate. The diversification of lymphoma subtypes thus clearly impacts prognosis.

A recent study also proposed new classes of B-cell lymphoma based on a correlation gene set analysis and using a larger patient collective [102]. This study excluded mBL samples from the patient cohort and divided the remaining DLBCL cases into three classes. Their expression signatures and phenotypic characteristics show certain similarities with our non-mBL*, intermediate A and B subtypes; however, they also differ in other properties, for example in the assignment of cell-of-origin properties and of energy metabolism signatures.

4.1.8 CONCLUSION

We applied single omics SOM approach to patient expression data of mature aggressive B-cell lymphomas to characterize the specifics of the genome-wide expression landscapes in different molecular subtypes of lymphoma. We presented a straightforward strategy to identify outlier samples and modules, e.g., due to contaminations of tumor samples with healthy tissue, and to correct them. Furthermore, we found indications for a finer subtype classification of aggressive B-cell lymphoma into four subtypes. Samples were classified using a spot-guided and metagene-based k-Means clustering method. The robustness and consensus-cluster stability of the new four subtypes exceeds that of previous three class approaches. The functional and clinical impact of the new subtypes was discussed. The two intermediate subtypes of heterogeneous molecular signatures are associated with poor survival prognosis compared with the more homogeneous mBL* and non-mBL* subtypes.

Our case study shows that analyzing gene expression landscapes with the tools presented here facilitates information mining in such huge data sets and eventually promotes our understanding of cancer biology.

4.2 DNA METHYLATION LANDSCAPE OF LYMPHOMAS AND ITS IMPACT ON TRANSCRIPTION

In this section we focus on germinal center derived B-cell lymphoma and multiple myeloma. The molecular mechanisms underlying genesis, progression and also mutual transformations between the subtypes is not clear in many details. Changing gene expression signatures are strongly linked to perturbations of epigenetic mechanisms. Understanding molecular mechanisms of lymphoma thus requires a combined view including gene expression, epigenetics and also genetic factors affecting B-cell biology. The lymphoma methylation samples were classified according to Martin-Subero [60] as diffuse large B-cell lymphoma (DLBCL), molecular Burkitt's lymphoma (mBL), intermediate lymphoma (IntL), follicular lymphoma (FL) and mantle cell lymphoma (MCL). Further the data set contained multiple myeloma (MM), healthy B-cells and germinal center B-cells (GCB) as reference. As we attempted an integrative analysis of DNA methylation and gene expression data in this study and the methylation data sets comprised several subtypes exceeding those analyzed in 4.1, a larger cohort of gene expression samples was needed to be considered here. Gene expression cases taken from MMML (Molecular mechanisms of malignant lymphoma, described in [59]) cohort were assigned to the same classes as mentioned for methylation data. For details concerning the cohort and preprocessing of the data see sections 3.1 and 7.1.3.

4.2.1 HIGH-DIMENSIONAL DATA PORTRAYING

Preprocessed gene-centric expression and methylation data were clustered using SOM machine learning. Three different SOMs were trained using (i) methylation β values (MetSOM), (ii) centralized β values with respect to the mean β of a gene averaged over all samples (DmetSOM), and (iii) centralized log-expression data (DexSOM). For each run we used a quadratic grid of size 50×50 to distribute the expression or methylation profiles. Note that genes are arranged differently in each of the SOM trainings. For comparison we mapped groups of selected genes in each of the SOM maps as described below.

4.2.1.1 ABSOLUTE DNA METHYLATION PORTRAYING IDENTIFIES HYPO- AND HYPERMETHYLATED GENES

Figure 18a shows the gallery of the mean DNA methylation portraits for all classes studied. Red and blue regions in the images refer to genes with high and low methylation levels of the probed CpG regions with β values near one and zero, respectively. The map can be segmented into regions containing genes hyper- and hypomethylated in lymphomas and a region with almost invariantly methylated genes in between (Figure 18b and c). Groups of signature genes with characteristic methylation profiles can be extracted from spots assigned using Arabic numbers '1' – '6' (Figure 18c). The methylation maps thus

provide genes hyper- and hypomethylated in lymphomas compared with B-cells and also genes with almost invariant methylation levels. For example, genes in region '5' are clearly on high methylation level in B-cells and on lower level and thus hypomethylated in lymphomas.

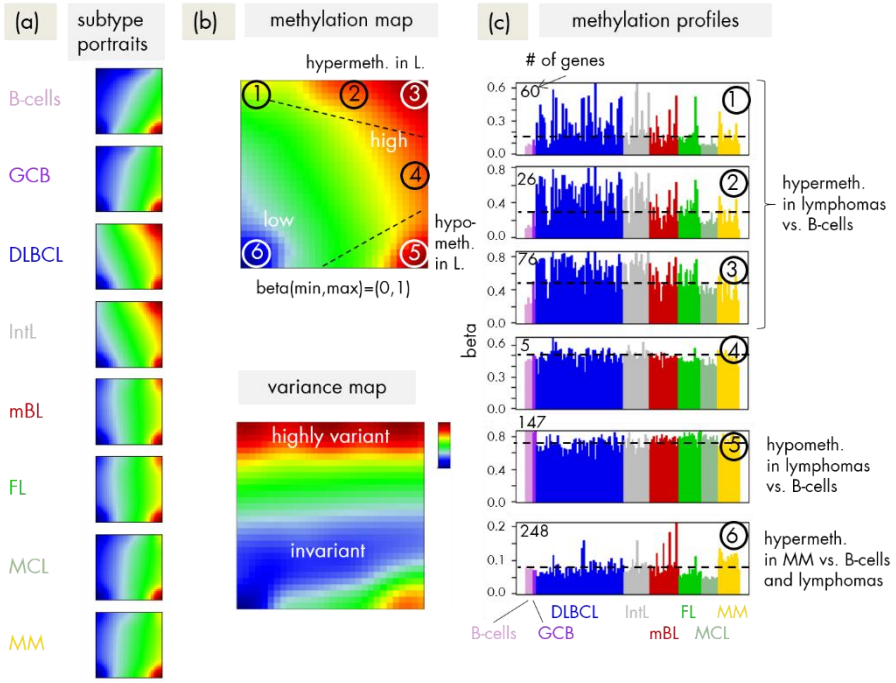


Figure 18: SOM portraying of DNA methylation landscapes of lymphomas (MetSOM). (**a**) SOM portraits of histological lymphoma classes and of controls. Red and blue colors assign regions containing genes of high and low methylation levels, respectively. (**b**) The methylation overview map summarizes regions hypermethylated in any classes compared with the others in red. The methylation variance map identifies regions of highly variable (red) and almost invariant (blue) β values. (**c**) The methylation profiles show the mean methylation level among the samples of genes taken from the 'spot' regions '1' – '6' assigned in the methylation overview map. Horizontal dashed lines serve as guide for the eye showing the mean β level of the respective spot averaged over all samples. Assignments as 'hyper-' or 'hypomethylated' refer to relative methylations compared with B-cells.

4.2.1.2 DIFFERENTIAL METHYLATION PORTRAYING BETTER RESOLVES DIFFERENCES BETWEEN THE LYMPHOMA CLASSES

In a previous work it was shown that the analysis of centralized values better resolves subtle differences between the samples [103]. We therefore calculated a second SOM using centralized methylation values (DmetSOM), where the mean β value of each gene averaged over all samples was subtracted from its actual β value. Centralization rather focuses the view on methylation changes between the samples independent of the absolute methylation level of the genes. In the obtained DmetSOM portraits we identified five spot-clusters numbered 'i' – 'v', which provide differential methylation profiles reflecting specific

hyper- and hypomethylation of selected lymphoma classes compared with B-cells (Figure 19a - c).

Invariantly methylated genes accumulate in the center of the map, whereas the variable genes occupy different regions near the border in a profile-specific manner. Mapping of the methylation clusters '1' – '6' obtained from the MetSOM (previous subsection) into the DmetSOM reveals mostly a one-to-one relationship (Figure 19d). For example, spot 'v' referring to genes specifically hypomethylated in B-cell, mBL and MCL compared to DLBCL distributes over spot '1' and, to a less degree spot '2'. This result simply means that most of the genes undergoing hypo- or hypermethylation between the different sample classes show predominantly an initially high or low methylation level, respectively. We therefore restrict our further analysis to the clusters 'i' – 'v' in the DmetSOM.

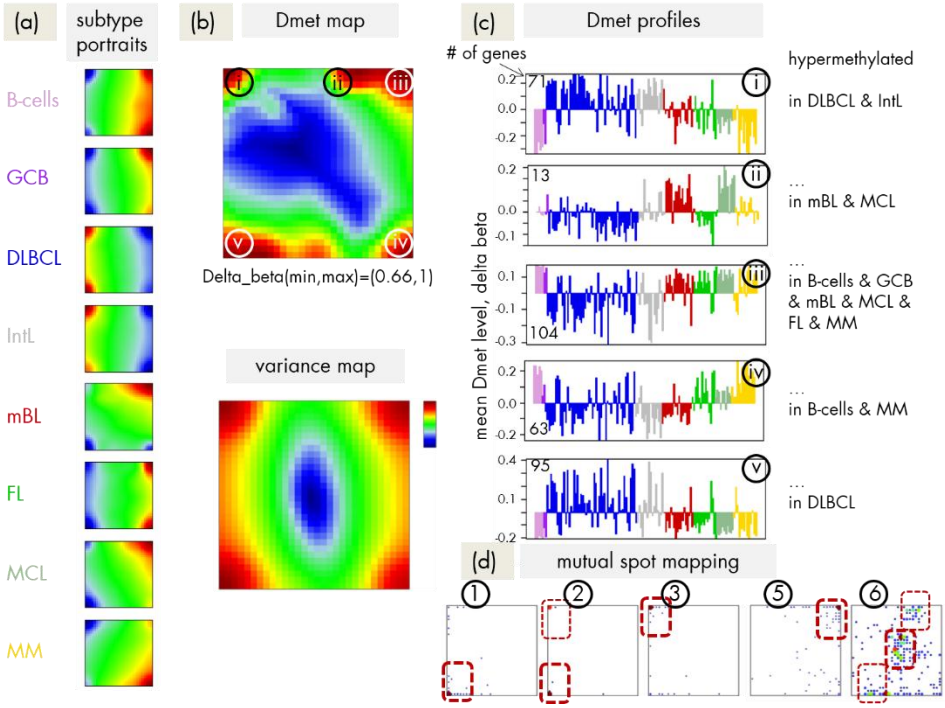


Figure 19: Differential methylation portraying of lymphoma and controls (DmetSOM). See legend of Figure 18 for a detailed description of the panels (**a**) – (**c**). The DmetSOM better resolves differential methylation between the lymphoma classes (compare with the MetSOM in Figure 18). (**d**) Genes from methylation spots '1' – '6' of the MetSOM were mapped into the DmetSOM (each dot marks a metagene occupied by at least one gene from each of the spots '1' – '6', respectively). One finds almost a 1:1 relationship between the spots except for spot '6', which 'hides' spot 'ii'.

Gene set enrichment analysis provides first ideas about the functional context of the genes in the spot modules (Table 2). Spots 'i' and 'v' hypermethylated in DLBCL and IntL enrich genes related to the 'formation of the polycomb repressive complex' (PRC2), which controls cellular development and differentiation [104]. Interestingly, genes from these spots are hypermethylated also in other cancers such as colorectal cancer (CRC) and high and low grade glioma. *Vice versa*, hypomethylated genes in DLBCL and IntL (spot 'iii') are also consistently hypomethylated in CRC and glioma suggesting parallels in epigenetic

regulation between different cancer types. Genes hypermethylated in B-cells and MM (spot 'iv') are associated with 'immune processes', whereas genes hypermethylated in mBL and MCL (spot 'ii') enrich processes related to 'cell proliferation' and 'cell cycle activity'.

Table 2: Functional context of the differentially methylated gene clusters.

Spots			Regulated classes		Functional context:
Dmet-spot	Met-spot	mean Met-level	Dmet up[1]	Dmet down[1]	enriched gene sets[2]
i	2,3	intermediate	DLBCL, IntL	B-cell, GCB, FL, MM	CIMP high-vs-low hypermethylated; hypermethyated in primary glioblastoma [105]; hypermethylated_in-cancer-and-ageing [106]; hypermethylated-in-CRC[107]; *NANOG*- *SUZ12*-, and *EED*- targets [Wang];
ii	6	low	BL, MCL	DLBCL, FL	*MYC*-targets [108]; GO_BP: G1/S-transition in mitotic cell cycle; GO_BP: cell cycle
iii	5	high	B-cell, GCB, mBL, FL, MCL, MM	DLBCL, IntL	Hypomethylated in CRC; CIMP high-vs-low hypomethylated [107]; Hypermethylated in adult brain [62]; Hypomethylated in secondary glioblastoma [105]
iv			B-cell, MM	DLBCL, IntL, mBL	GO_BP: immune response; hypomethylated in glioma [105]; GO_CC: nuclear chromatin; NKF-beta down in mBL [85]; *IL21*-targets down[109]
v	1,2	low, intermediate	DLBCL, IntL	B-cell, GCB, mBL, FL, MCL, MM	*SUZ12*-targets[110]; hypermethylated in grade 3 astrocytoma and grade 2 oligodendroglioma [105]; hypermethylated in low grade glioma[111]; hypermethylated in CRC[107]; low expression TF [112]

[1] Sample classes showing high (Dmet up) or low (Dmet down) methylation levels, respectively.
[2] Enrichment of predefined gene sets in the spot-lists of genes (Dmet-spot and/or Met-spots) was calculated as described in [80]. Gene sets were taken from literature or from gene ontology (GO) categories biological process (BP) or cellular component (CC).

Next we investigated the diversity landscape of the methylation portraits of lymphoma and reference samples. The calculated similarity network reveals two main clusters, which can be assigned to samples methylated either similarly to B-cells or to DLBCL (Figure 20). The essentially two main spot patterns of the mean DmetSOM portraits shown in Figure 19a directly reflect the separation between two main sample clusters seen in Figure 20: The samples with DLBCL-like methylation patterns preferentially show red hypermethylation spots in the left part of the portraits (spots 'i' and 'v', see also Table 2), whereas the B-cell-like methylation patterns is characterized by red hypermethylation spots in the right part of the map (spots 'ii' to 'iv'). These patterns are strongly anti-correlated, *i.e.*, hypermethylation

is opposed by hypomethylation for many genes when compared with mean methylation level averaged over all samples.

The DLBCL-like methylation cluster contains most of the DLBCL (69%) and IntL (81%) samples but also a certain number of FL (14%), mBL (28%) and MM (14%). On the other hand, also the second cluster of B-cell-like methylation contains 25% of the DLBCL and 19% of the IntL samples. Hence, methylation of the lymphoma classes is characterized by a certain degree of fuzziness. The gallery of individual DmetSOM portraits shown in supplementary material of [113] indicates that, e.g. two of the FL samples show clearly a DLBCL-like methylation characteristics, whereas the majority of the FL are compatible with B-cell-like methylation patterns. Note also that the B-cell-like methylation cluster reveals a fine structure, which separates MM and B-cells on one hand and mBL, GCB-cells and MCL on the other hand. This fine structure is related to subtle methylation differences between hypermethylation spots 'ii' – 'iv' (Figure 19 and Table 2). Finally note that the similarity analysis is based on a relatively small selection of less than 800 genes only, which might distort similarity relations if relevant groups of genes are under- or overrepresented.

Lists of genes from the regions 'i' – 'v' are given in supplementary material of [113].

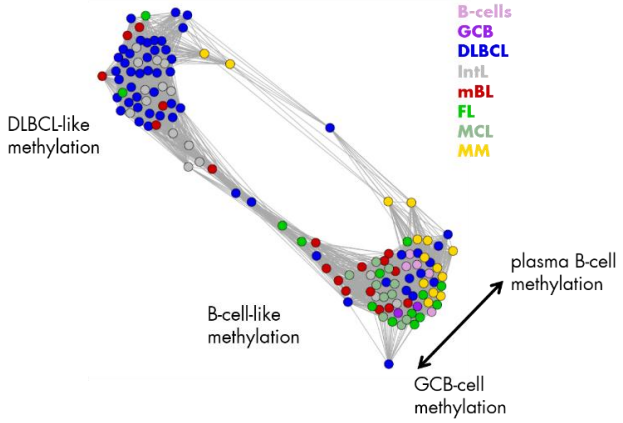


Figure 20: Similarity network of the methylation landscapes of the lymphoma samples studied. Each circle refers to one sample colored according to its class assignment. Two main cluster can be distinguished, which include samples of B-cell-like and DLBCL-like methylation. See section 3.5 for details.

4.2.1.3 GENE EXPRESSION PORTRAYING USING EXTENDED MMML-COHORT DATA

We characterized the heterogeneity of gene expression landscapes of lymphoma in detail (see [91,114]) in an analogous approach as used above for differential methylation data. Figure 21 summarizes the main results of the DexSOM analysis showing the mean SOM expression portraits of lymphoma subtypes and controls in panel a, the spot summary and variance maps (panel b) and the respective spot profiles (panel c). Please recall that for the lymphoma expression data set analyzed in section 4.1 only the subtypes mBL, non-mBL and intermediate were considered. In the present DexSOM additional B-cell lymphoma subtypes and control samples were trained to ensure a joint analysis with methylation data, which comprises comparable subtypes.

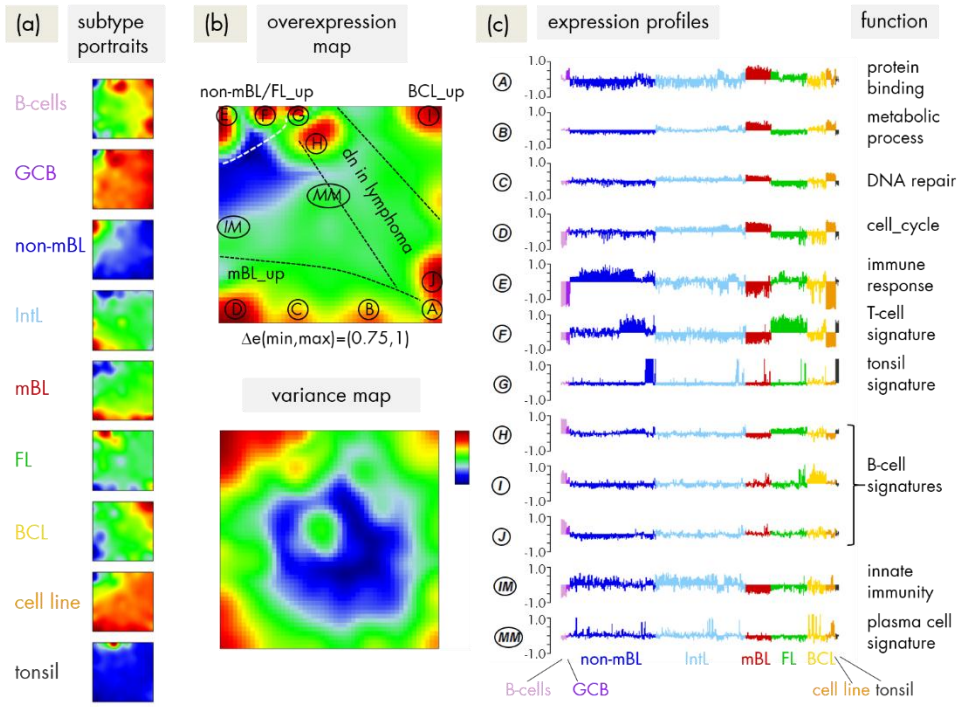


Figure 21: SOM portraying of the expression landscape of lymphoma (DexSOM). See caption of Figure 18 for details. Expression classes were color-coded such that their color agrees with the respective histological class in the methylation data set. Most of the spot modules detected can be clearly assigned to distinct lymphoma classes providing lists of signature genes, which are up-regulated in the respective sample classes and which are associated with distinct biological functions. For example, mBL (spots 'A' – 'D') and DLBCL (spot 'E') are related first of all to genes promoting proliferation and immune response, respectively.

4.2.2 MUTUAL MAPPING OF EXPRESSION AND METHYLATION MODULES REVEALS POSITIVE AND NEGATIVE CORRELATIONS

After separate SOM analysis of DNA methylation and gene expression data we linked both types of analyses in the next step to detect mutual relations between promoter methylation and gene expression. In a first attempt we mapped the approximately 800 genes considered on the methylation arrays into the gene expression landscape of lymphoma (DexSOM) and color-coded their methylation level (see supplementary material of [113]). No densely populated areas of uniquely methylated genes were found indicating a fuzzy relationship between co-methylated and co-expressed genes. Possibly this mutual mapping on gene level provides a suited approach if the methylation assay probes all genes, which are also considered in the expression assay.

In the next step we considered groups of co-methylated genes separately: Genes of the Dmet-spots 'i' – 'v' (see Figure 19 and red circles in Figure 22a) were mapped into the MetSOM and DexSOM where they clearly accumulate in distinct regions as indicated by

the dotted red rectangles in Figure 22a. This result reflects the fact that groups of co-methylated genes are also co-expressed in a class-specific fashion as confirmed also by the respective methylation and expression profiles shown in the right part of Figure 22a.

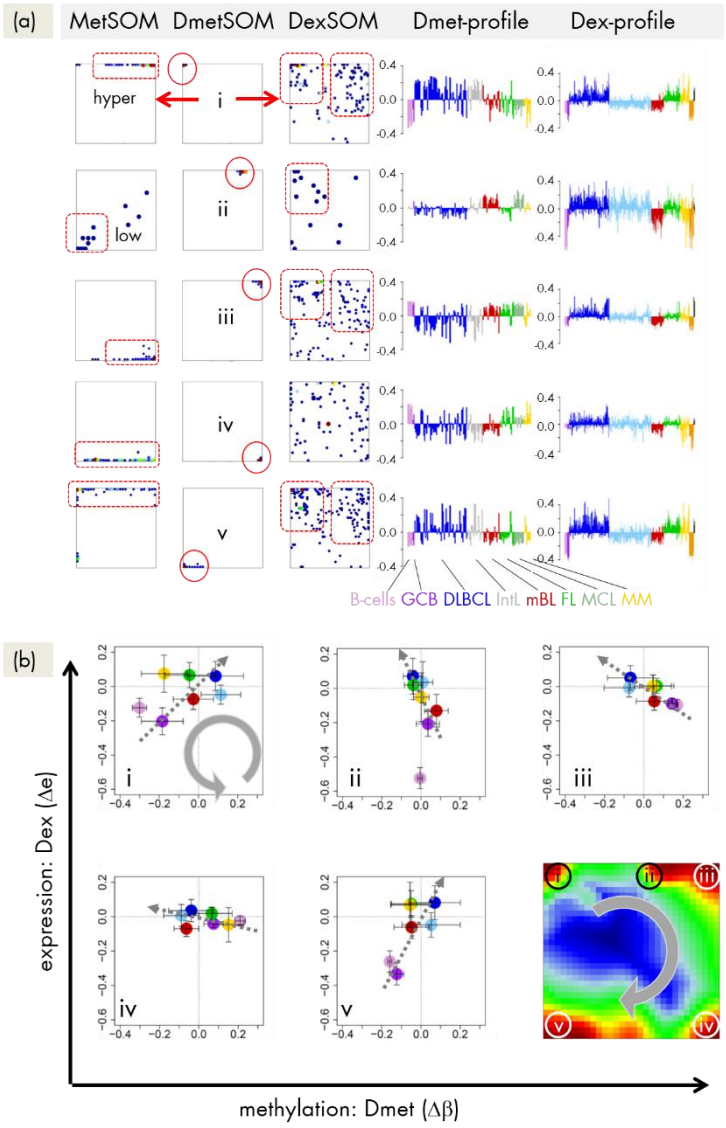


Figure 22: Mapping of differentially methylated genes into the lymphoma expression SOM: (**a**) The DmetSOM-gene clusters 'i' – 'v' (red circles) were mapped into the Met- and DexSOM where they accumulate in specific areas (red rectangles). The red arrows illustrate the mapping direction. The Dmet- and Dex-profiles reveal class-specific correlations between methylation and expression data, which are plotted in panel (**b**) for genes taken from each of the Dmet-spots 'i' – 'v'. The class-specific mean methylation and expression levels of the gene groups were plotted in x and y direction, respectively, each class represented by a colored dot. The error bars indicate the standard deviation of the sample data of each class. The dotted arrows point from the GCB-cell to the DLBCL dots thus serving as indicator for the slope of the mutual association between the methylation and expression data. Note that the clockwise arrangement of the spots 'i' – 'v' in the DmetSOM transforms into counter-clockwise arrangement of the data cloud in the correlation plots.

To better resolve the mutual relations we correlate class-averaged mean methylation and expression levels of the gene groups taken from each of the methylation spots 'i' – 'v' in panel b of Figure 22. Spots 'i' and 'v' are characterized by a positive correlation: *i.e.* hypermethylation in DLBCL with respect to B- and GCB-cells is accompanied by overexpression in DLBCL with respect to the healthy cell controls. MM and partly FL show concerted coexpression with DLBCL but still similar methylation compared with B and GCB-cells. The other lymphoma classes behave similarly however with smaller effects. Genes from spot 'iii' show a negative correlation, where differential methylation changes sign compared with spots 'i' and 'v' but differential expression does not. In other words, overexpression in DLBCL is associated with hyper- (spots 'i' and 'v') and hypo- (spot 'iii') methylation as well. Recall that all three spots 'i', 'iii', and 'v' are also functionally related: They enrich genes differentially methylated in other cancer types and related to 'PRC2 formation'. Spot 'ii' (related to 'proliferation', see Table 2) collects genes weakly responding to methylation but strongly to differential expression for most of the lymphoma classes. Note that in mBL hypermethylation of spot 'ii' genes associates with underexpression of the respective genes. In contrast, spot 'iv' (related to 'immune response') weakly responds to expression changes but strongly to differential methylation.

We also mapped the spot-clusters of co-expressed genes extracted from the expression SOM into the methylation SOM to assess mutual correlations (see supplementary material of [113]). Most of the effects observed are weaker than for the co-methylated gene clusters 'i' – 'v' presumably due to causal relations between promoter methylation and gene expression leading to the dilution of correlations in the opposite direction. On the other hand, the data clearly reveals expression changes between lymphomas and the reference B- and GCB-cells, which are accompanied by marked differential methylation effects in both positive and negative directions as well.

Hence, we observe positive and negative correlations between expression and methylation changes by mapping clusters of co-methylated genes into expression space and *vice versa*. Most pronounced effects are observed between B/GCB-cells and DLBCL in correspondence with the sample diversity analysis (Figure 20) but also the other lymphoma subtypes show gradual and specific effects roughly in the same order as illustrated in Figure S 1c.

4.2.3 MAPPING OF FUNCTIONAL GENE SETS: INFLAMMATION AND DEVELOPMENTAL GENES ARE PRONE TO ABBERANT METHYLATION

Next, we analyzed a series of functional gene sets in an analogous fashion as the spot modules in the previous subsection (Figure 23). The obtained characteristics can be grouped into different patterns. '*MYC*-targets' [115] and 'transcription factors (TF) associated with high gene expression levels' [112] give rise to large expression differences between the lymphoma classes but almost negligible methylation effects. 'TFs associated with low expression levels' and 'G-protein receptors' show a similar relation between expression

and methylation changes where the expression levels of the lymphoma classes however swap their order in the correlation plot. The latter effect can be directly extracted from the areas of highest population densities of the genes in the DexSOM: The 'high expression' genes enrich in the lower part of the DexSOM whereas the 'low expression genes' preferentially occupy areas near the left and right upper corners of the map (see the dotted red rectangles in Figure 23). The third group of 'PRC2-related' genes gives rise to marked class-specific expression and methylation changes. Interestingly, the expression characteristics of the 'PRC2-group' and of the 'low expression' group are almost identical whereas their methylation characteristics differ largely in amplitude. It seems that the 'PRC2-related' gene sets specifically select genes, which change expression and methylation in a lymphoma-specific fashion, whereas the 'low expression' gene sets contain genes, which show main effects in the expression domain only. This difference can be rationalized by the fact that a large fraction of these genes is affected 'indirectly' by downstream co-regulation of gene expression without alterations of promoter methylation.

The next group of 'age related genes' can be interpreted as a subgroup of the 'PRC2-related' and 'low expression' genes, which occupies essentially only the right upper region of the DexSOM. In the correlation plot one sees that this restriction strongly reduces the variance of the expression values between the lymphoma classes whereas the alterations of methylation are similar to the 'PRC2-related' gene sets. This result implies that 'PRC2-related' genes are governed by more diverse regulation mechanisms of gene expression than the 'age-related' genes. Note, however, that the gene set 'developmental regulators' being part of the 'age-related' group also collects genes referring to the 'formation of the polycomb complex' [116]. These genes were obtained from gene expression measurements whereas the 'ageing-associated hypermethylated genes' [117] are extracted from DNA methylation studies, which explains the larger response of the latter ones in the methylation dimension.

The last 'CIMP'-group genes accumulate in the top left region of the DexSOM. They consequently share similarities with the groups of 'PCR2-related' and 'low expression genes' whose genes also accumulate in this region of the map. The methylation effect of the gene sets 'inflammatory response' is small but more pronounced for the 'GCIMP'-gene set extracted from glioma data [46]. Other 'CIMP- and GCIMP-related' gene sets obtained in colorectal and in brain cancer studies, respectively, also respond in the methylation dimension (data not shown, see also Table 2).

In summary, we found two main combined methylation/expression patterns exemplified by the 'high-expression' and 'PCR2-related' groups, where only the latter is characterized by both expression and methylation changes. The latter group can be further split into 'CIMP-like' and 'age-related' genes.

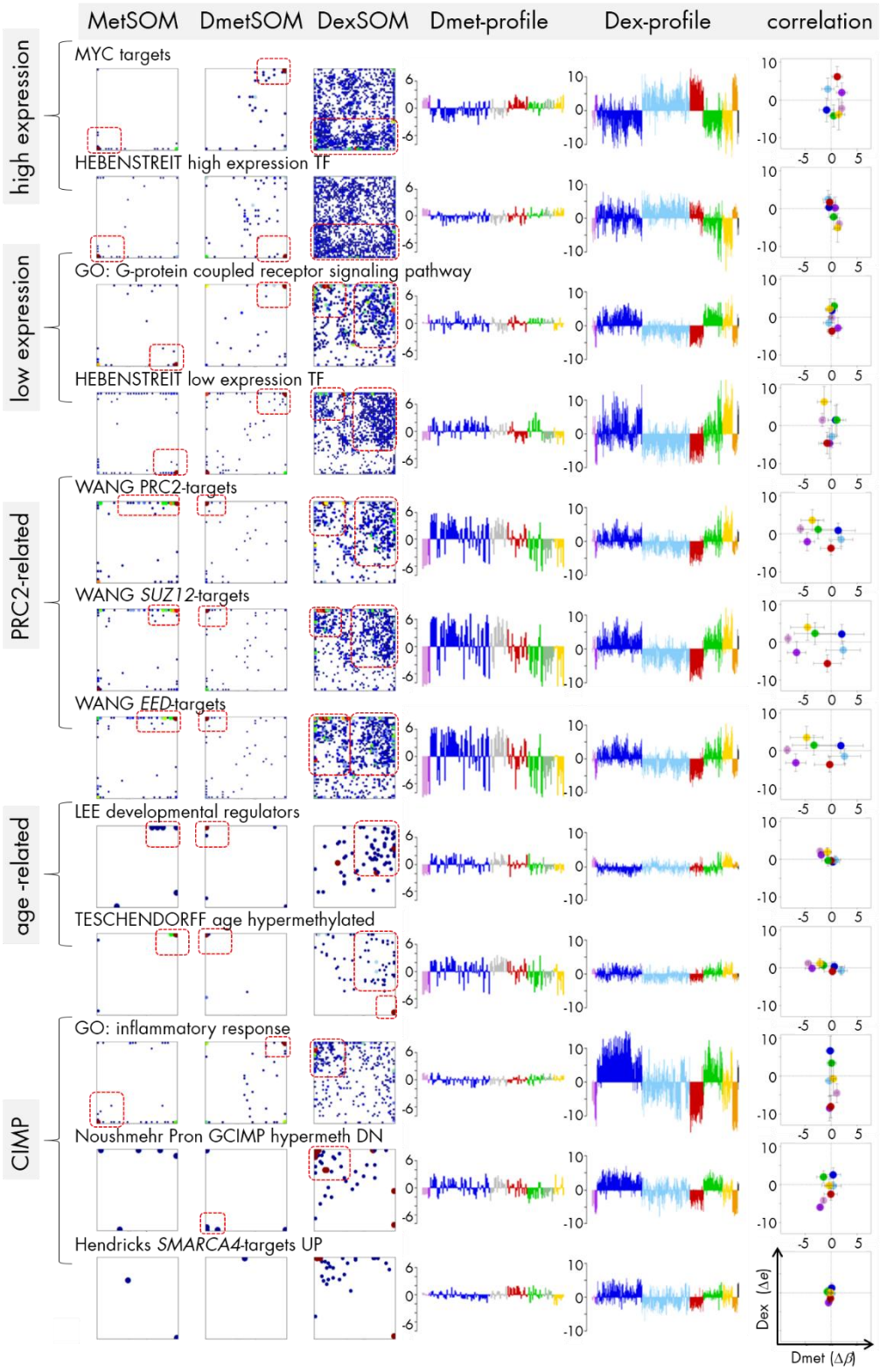


Figure 23: Mapping of selected functional gene sets into the lymphoma methylation and expression SOM (see legend of Figure 22 for assignments). Gene sets were taken from [46,110,112,115–119]. The combined methylation-expression data groups into five different patterns as indicated by the brackets and the designations given at the left part of the figure.

4.2.4 *EZH2*-TARGETS STRONGLY DEREGULATE IN LYMPHOMAS

EZH2 is the catalytic subunit of PRC2 and mediates transcriptional repression through its histone methyltransferase activity that trimethylates H3K27 [120,121]. *EZH2* is upregulated in normal GCB-cells and it is implicated in lymphomagenesis. It binds partially to the same targets in GCB-cells as in embryonic stem cells (ESC), which are preferentially H3K27me3-marked and thus transcriptional inactive [120]. *EZH2* in normal GCB-cells represses tumor suppressor genes, thus driving cellular proliferation. Similar to the regulatory state in stem cells, it prevents premature differentiation and maintains transcriptional silencing already present in NB-cells [120].

We analyzed *EZH2*-targets and H3K27me3-marked genes determined in ESC, (naïve) B- and GCB (centroblasts) cells obtained by means of ChIP-chip experiments [120]. The respective sets of genes were mapped into our different SOMs (see supplementary material of [113]). The methylation and expression characteristics of both *EZH2*-target and H3K27me3-marked genes in ESC, B- and GCB-cells among the systems studied here are very similar and closely resemble those of the 'PRC2-related' genes. As a result we found that H3K27me3-marked *EZH2*-targets are transcriptional repressed in healthy B- and GCB-cells, and that repression of these targets is mostly maintained in mBL but at least partly turns into activation in DLBCL, IntL, FL and MM. These expression changes are paralleled by hypomethylation in B- and GCB-cells and in MM on one hand and hypermethylation in DLBCL and IntL on the other hand. Stratification of genes with respect to anti-correlation between expression of *EZH2* and that of its targets [120] specifically selects genes from DexSOM spots 'E' and 'F' supporting this view because *EZH2*-mediated trimethylation of H3K27 is expected to inactivate the expression of the *EZH2*-target genes. Interestingly, *de novo EZH2*-targets in centroblasts compared with NB-cells reverse expression levels and to a less degree also methylation levels in MM, B- and GCB-cells. It was suggested that *EZH2* upregulation during the transition from NB-cells to centroblasts reactivates a stem cell-like repression program, which is not present in NB-cells and possibly featuring increased self-renewal and proliferative potential [120].

4.2.5 CHROMATIN STATES AND THEIR POSSIBLE REMODELING

Higher-order chromatin structure is emerging as an important regulator of gene expression. Alterations of gene expression programs can be induced by the remodeling of chromatin states, which for example facilitate transcription in open regions of euchromatin, but prevent gene expression in densely packed regions of heterochromatin. These different states of chromatin conformation are governed by the arrangement of nucleosomes being the central structural elements of DNA packing in the nucleus. In turn, the arrangement of nucleosomes is modulated by chemical modifications of distinct amino acids in the side chains of the histone units forming the nucleosomes. A whole battery of such modifications and their combinatorial patterns are able to tune the transcriptional activity of the affected genes by influencing the functional state of gene's structural elements such as enhancers

and promoters, and also stages of the transcriptional process such as transcriptional elongation, transition, activation, and repression [122].

To get an insight into the possible mechanism of chromatin remodeling in lymphoma we make use of the chromatin states identified in GM12878 lymphoblastoid cells (LBC), which imitate immature lymphocytes [119]. The chromatin states were calculated from ChIP-Seq data of a series of histone modifications using a hidden Markov model [123] (see section 2.2.3 for details). We mapped the respective chromatin regions of each state on the human genome and collected the genes included in each of the eleven chromatin states into one gene set, and then mapped them into the lymphoma methylation and expression SOMs to assess their methylation and expression characteristics as described above (Figure S 7).

We found close correspondence between the methylation/expression properties of groups of chromatin states and the groups of functional gene sets identified above: Genes with chromatin states strongly promoting transcription ('active Txn' states), namely the states 'active promoters', 'transcriptional elongation' and 'transcriptional transition', 'weak transcription' and also the state 'weak promoters' closely resemble the characteristics of the 'high expression' gene sets shown in Figure 23. Contrarily, transcriptionally inactive states ('poised promoters' and 'repressed promoters') share close similarity with the 'PCR2-related' gene sets. The state 'heterochromatin' resembles the 'low expression' gene sets. Note that the 'Txn-inactive' states and 'heterochromatin' show a nearly mirror symmetrical profile of gene expression compared with the 'Txn-active' states with low expression levels in mBL and IntL and high levels in BCL (MM), non-mBL and FL. The state 'strong enhancers' forms a separate group, which differs from the functional gene sets considered. It's methylation and expression profiles virtually agrees with that of the 'active Txn' states except for the expression level in mBL, which turns from high activity in 'active Txn' states into low activity in the 'strong enhancer' state and *vice versa* for FL.

Interestingly, the transcriptional inactive chromatin states (and also 'PRC2-related' genes) show the largest variability of DNA methylation between the classes with lowest levels in healthy GCB and B-cells, intermediate levels in MM, FL and mBL, and high levels in DLBCL and IntL. Thus they resemble the order of overall methylation variability shown in Figure S 1. These methylation changes were paralleled by positively correlated alterations of gene expression. Genes located in 'heterochromatin' show virtually the same class-dependence of gene expression but almost no variation in methylation. Hence, genes becoming activated in 'heterochromatin' are obviously affected by other mechanisms not associated with methylation changes of their promoters.

Note that the assignment of chromatin states refers to the lymphoblastoid cell line but not to the lymphoma classes studied here. Generally one expects that 'Txn-active' states show higher gene expression levels than 'Txn-inactive' states and 'heterochromatin'. This trivial relation implies to use the mean transcriptional activity of the chromatin states in lymphoma as a measure to estimate the correspondence between the nominal chromatin state referring to lymphoblastoid cells and the real one in lymphoma. For an overview we

stratified the expression levels of the chromatin states in the different lymphoma classes into high, moderate and low levels based on the GSZ profiles shown in Figure S 7 and visualized them in Figure 24a: The expression level observed in GCB-cells, mBL and IntL is indeed high in 'Txn-active' chromatin states and low in 'Txn-inactive' chromatin states. Thus the real expression levels agree with the nominal ones suggesting global correspondence between the chromatin states in the reference cells and that in mBL and IntL. Contrarily, the expression levels in BCL, FL, and non-mBL disagree with the expression levels expected for the nominal chromatin states. This switching of gene activity between these two groups of samples suggests remodeling of chromatin in BCL, FL, and also non-mBL compared with lymphoblastoid cells and thus also with mBL, IntL, and GCB-cells. Note also that the activity patterns of B-cells, tonsils and also of BCL differs from that of GCB-cells suggesting remodeling of chromatin between healthy (pre- and post-GC) B-cells and GCB-cells. Moreover the similar expression patterns of B-cells and of BCL supports the plasma cell characteristics of BCL differing from the characteristics of the GC-derived lymphoma subtypes.

To assess the relation between DNA hypermethylation in lymphomas and the chromatin states we calculated the percentage of overlap-genes from the different chromatin states also found in the set 'hypermethylated in DLBCL' taken from [60]. The overlap of hypermethylated genes is only about 10% for transcriptional active states but much higher (50% - 90%) for transcriptional inactive states. Hence, activation of the latter states in DLBCL/non-mBL and FL seems to be accompanied by hypermethylation of a large fraction of genes being inactive in lymphoblastoid cells, mBL, and IntL.

Finally, we transferred the expression levels of selected gene sets discussed above into the tabular form for direct comparison with that of the chromatin states (Figure 24b). '*MYC*-target' genes are expressed in parallel with 'Txn-active' states among the systems studied. This agreement suggests that the '*MYC*-targets' are found predominantly in chromatin regions active in mBL, IntL, GCB-cells, the cancer cell line and lymphoblastoid cells (95% overlap between '*MYC*-targets' and 'active promoters'). In contrast, gene sets related to 'inflammation' and 'G-protein receptor activity', both hypermethylated in DLBCL, accumulate in chromatin states inactive in the reference system but activated in DLBCL, IntL, FL and BCL.

In summary, gene sets referring to distinct chromatin states in the reference cells show well distinguished expression and DNA methylation characteristics either agreeing or disagreeing with the expression level expected in the nominal chromatin states. Disagreement indicates chromatin remodeling in IntL and non-mBL and especially in B-cells and BCL compared with mBL and GCB-cells. Hence, one can distinguish three groups of samples showing characteristic expression patterns of genes assigned to different chromatin states. They comprise (i) mBL, GCB-cells and cancer cell lines; (ii) BCL, MM, and (pre- and post-GC) B-cells and (iii) IntL, DLBCL, and FL.

(a) mean gene expression in selected chromatin states

chromatin state [1]	gene expression [2]: +...high, x...moderate, -...low									% hyper-meth. in DLBCL ~
reference	lymphomas [3]					controls				
lymphoblastoid cells	mBL	IntL	non-mBL	FL	BCL	B-cells	GCB-cells	tonsils	cell line	
active states	+	+	x	x	-	-	+	-	+	10
weak promoter	+	+	+	x	-	-	x	-	+	30
strong enhancer	x	+	+	+	-	-	+	-	x	30
weak enhancer	x	+	+	+	-	-	x	-	x	60
inactive states & heterochr.	-	-	+	+	+	x	-	+	-	50-90
repetetive CNV	x	x	x	+	x	x	x	x	x	30

(b) mean gene expression of selected functional gene sets

function	gene expression [2]: +...high, x...moderate, -...low								
	lymphomas [3]					controls			
gene set [4]	mBL	IntL	non-mBL	FL	BCL	B-cells	GCB-cells	tonsils	cell line
MYC-targets	+	+	-	-	-	-	+	-	+
PRC2-related	x	x	x	x	+	+	x	x	x
hypermethylated upon ageing&cancer	-	-	x	x	+	x	x	x	-
hypermethylated in DLBCL	-	-	+	+	+	-	-	+	-
inflammation & stroma	-	-	+	+	x	x	x	+	-
G-protein coupled receptors	-	-	+	+	+	x	-	x	-

Figure 24: Mean gene expression level of selected gene sets in lymphoma and reference systems: **(a)** selected chromatin states and **(b)** selected gene sets. The gene expression level was stratified into high (red), moderate (green) and low (blue) levels using the respective GSZ profiles.

[1] Chromatin states were defined in [123] with respect to the associated histone marks (see, e.g. Fig. 1b in [123]). The most characteristic marks are in active states (e.g. active promoters): H3K4me3/me2, H3K27ac, H3K9ac; weak promoter: H3K4me3/me2; strong enhancer: H3K4me1/me2, H3K27ac, H3K9ac; weak enhancer: H3K4me1/me2; inactive states (e.g. inactive and poised promoters): H3K27me3, H3K4me2; heterochromatin: no mark; repetitive CNV: all marks.

[2] gene expression levels of the chromatin states (Figure S 7) and functional gene sets (Figure 23).

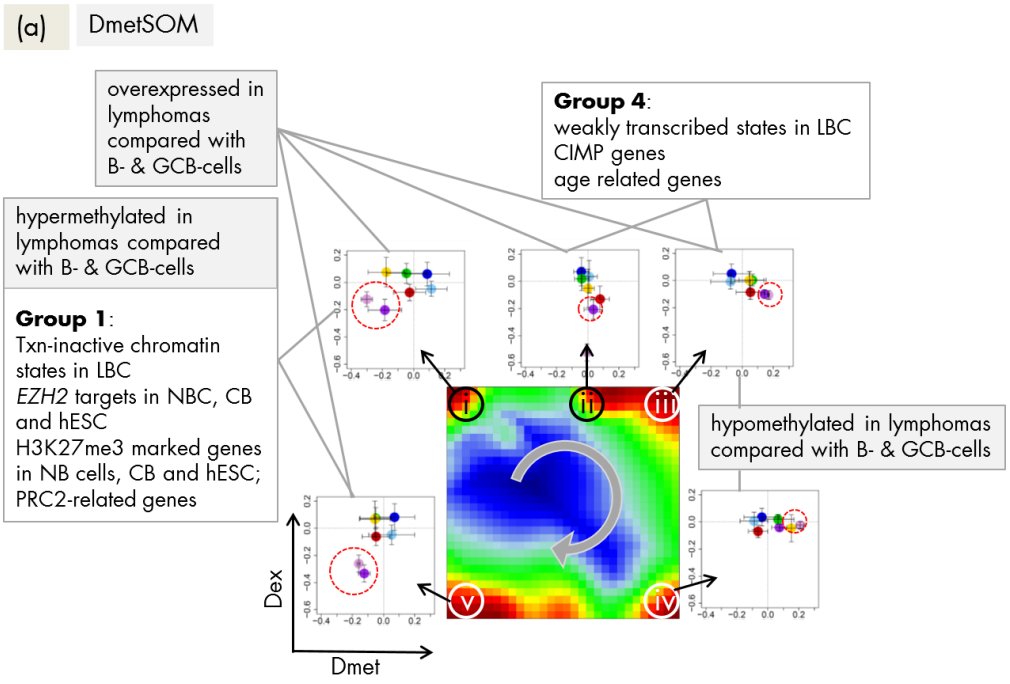
[3] assignment of lymphoma classes refers to the expression classes introduced in Figure 21.

[4] Gene sets were taken from [115] ('*MYC*'), [116] ('PRC2 developmental regulators'), [117] ('hypermethylated upon ageing and cancer'), [60] ('hypermethylated in DLBCL'), [89,124] ('inflammation and stroma') and GO ('G-protein coupled receptor activity and signaling pathway'); see also Figure 23.

4.2.6 FUNCTIONAL CONTEXT OF DIFFERENTIAL METHYLATED AND EXPRESSED GENES

Mutual correlation plots between the mean expression and methylation levels of the genes of each of the spot-modules revealed different patterns with impact for underlying epigenetic mechanisms of genomic regulation (Figure 25). We identified groups of genes mostly affected by methylation with only tiny expression changes (e.g. DmetSOM-spots 'iii' and 'iv' and DexSOM-spot 'G'), *vice versa*, groups of genes with almost invariant methylation levels but strongly varying expression (e.g. DmetSOM-spots 'ii' and DexSOM-spots 'D' and 'E'), and groups with strongly positive (spots 'i', 'v' and 'H' and 'J') and negative (e.g. spots 'iii', 'A' and 'I') correlations between expression and methylation levels in the different sample classes. Moreover, the Dmet- and DexSOM disentangle genes systematically hyper- and hypomethylated and/or over- and underexpressed in lymphoma compared with healthy B- and GCB-cells (see Figure 25). Hence, SOM portraying served as an effective sorting machine to extract different modes of co-regulation between expression and methylation mechanisms specifically characterizing lymphoma and differentiating also between the lymphoma subtypes.

To assign the functional meaning to the spot modules, especially in the context of underlying epigenetic mechanisms, we applied enrichment analysis using a multitude of predefined gene sets related to categories such as biological function (e.g. 'inflammation', 'cell development and ageing'), targets of different TFs (e.g. '*MYC*', 'high and low expression TFs') and epigenetic modulators (e.g. *EZH2*, *SUZ12*, PRC2), different chromatin states in reference lymphoblastoid cells and also genes differently expressed and methylated in other cancers (e.g. CIMP and GCIMP genes in colorectal cancer and glioma, respectively).



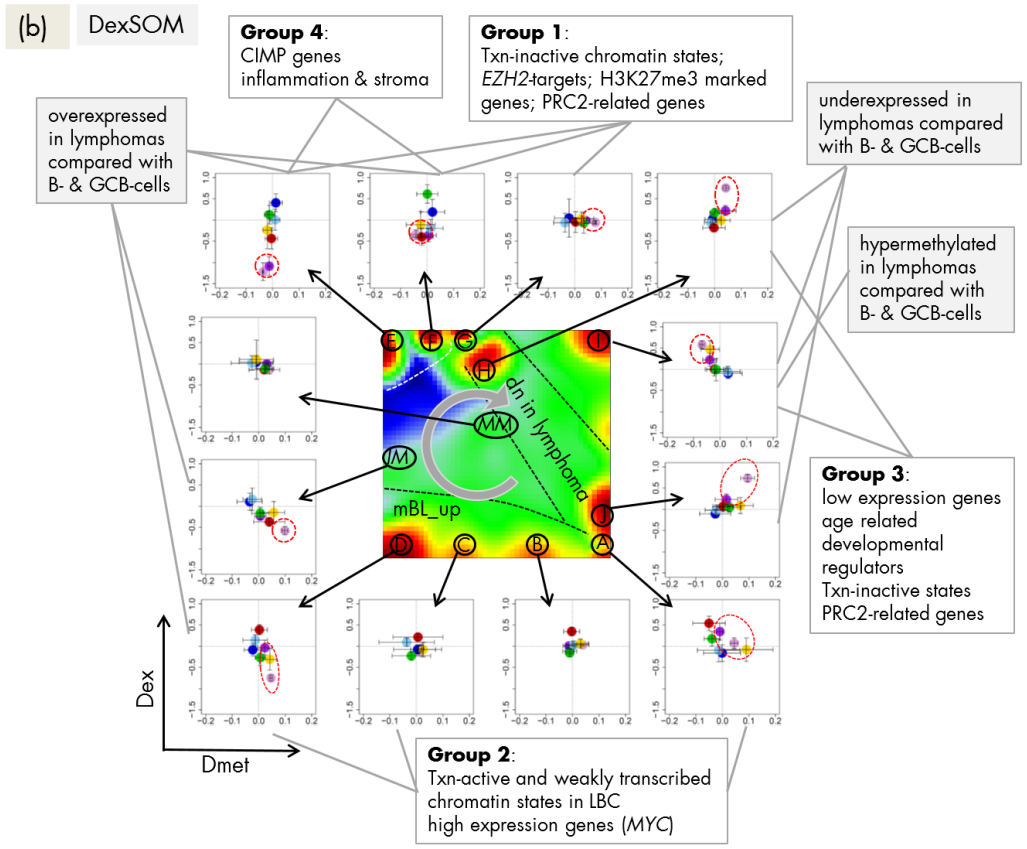


Figure 25: Integrative view on differential methylation and gene expression in lymphomas and on the related functional context. Spot modules of co-methylated genes were extracted from (**a**) the DmetSOM and (**b**) the DexSOM. The class-specific correlation plots for each spot reveal systematic methylation and expression changes in both maps many of them being associated with functional gene sets. Especially, differential methylation and expression with respect to healthy controls (B- and GCB-cells, see red dotted circles) as well as systematic differences between lymphoma subtypes (e.g., mBL, DLBCL and MM) were sorted in a systematic fashion in both SOM maps.

Interestingly, we found pronounced similarities of the expression and methylation signatures of gene sets from different categories in the lymphoma data, which indicate mutual relations between them. Particularly, the spot-modules can be sorted roughly into four main groups (see Figure 25):

- Group 1 is enriched in PRC2- and *EZH2*-targets, related to transcriptionally inactive states in LBC and shows strong variation in expression and methylation levels being hypermethylated and overexpressed in lymphomas compared with the controls
- Group 2 comprises transcriptionally active chromatin states, TFs related to highly expressed genes and *MYC*-targets. It promotes cell proliferation and shows strong expression changes especially between mBL on high and the controls on low levels, but virtually no differential methylation
- Group 3 accumulates mostly in the top right part of DexSOM and contains ageing and developmental genes, and low expression TF genes. It overlaps with group 1 with re-

spect to the enriched chromatin states and part of the PRC2- and *EZH2*-targets. Expression of these genes is down regulated in lymphomas compared with the controls but the methylation can differ in both directions.

- Group 4 accumulates in the top left part of the DexSOM and contains CIMP/GCIMP genes, genes related to 'inflammation and stroma', *SMARCA4*-targets and another part of the PRC2- and *EZH2*-targets. These genes are strongly upregulated in DLBCL, IntL and partly FL, and downregulated in the controls and BL. They show moderate methylation changes being slightly hypermethylated in lymphomas.

4.2.7 EPIGENETIC REGULATION IN LYMPHOMAS AS SEEN BY GENE EXPRESSION AND DNA METHYALTION

Figure 26 schematically illustrates and summarizes our results in the light of B-cell and lymphoma biology. Healthy B-cells pass essentially three relevant compartments, the dark and light zone of the GC and 'outside-of-the-GC', which subsumes plasma, lymph node and also bone marrow (see also Figure 10). The associated types of B-cells can transform into the different lymphoma classes as illustrated by the red arrows in Figure 26a. The triangular shape of the scheme is motivated by the three different types of lymphoma classes, which point to similarities with GC dark zone (DZ) B-cells in terms of proliferative activity, GC light zone (LZ) B-cells in terms of inflammatory signatures, and pre- and post-GCB-cells in terms of (healthy) B-cell signatures (see also [91,114]).

The colored 'ramps' code for alterations in gene expression and/or methylation between the lymphoma classes, which associate with the groups of genes defined in the previous subsection and which were specified with respect to changing chromatin states (Figure 26b). Group 1 genes give rise to increasing differential expression and methylation between lymphomas and healthy B-cells with largest effect in DLBCL. We suggest that the strong alterations in gene expression manifest chromatin remodeling from PRC-repressed and poised chromatin states into active ones associated with hypermethylation in lymphomas. Hence, group 1 genes are obviously of central importance for a mechanism of lymphomagenesis transforming healthy GCB-cells into malignant ones. Recall that the largest differential effect of these genes in gene expression and methylation is observed for DLBCL. Along the axis linking BL and DLBCL the expression changes are counterbalanced by group 2 genes, which strongly upregulate in BL compared with DLBCL almost without methylation changes. Presumably this trend is mainly caused by the activation of *MYC* in mBL (and also selected *MYC*-positive IntL cases), which in turn amplifies the expression of already transcribed genes giving rise to a sort of hyperactivation of the transcriptional state without strong DNA methylation effects and chromatin remodeling. Group 3 and 4 genes mainly differentiate between DLBCL and MM however in opposite directions. Both groups show alterations in gene expression and methylation as well, and thus partly resembling group 1 genes in their molecular determinants. Particularly, group 1, 3, and 4 genes contain PRC2- and *EZH2*-targets showing that repressed and poised promoter states play a

pivotal role in cell fate decisions of GCB-cells and in their transformation into cancerogenic states.

B-cells employ epigenetic mechanisms to generate effective memory responses resembling epigenetic reprogramming of stem cells upon cell fate decisions. Particularly, the transition from NB-cells permits GCB-cells to generate the differential response to antigenic challenges and to differentiate toward plasma cell fates. Deregulation of the underlying epigenetic determinants such as DNA methylation [125] and/or chromatin activity states can be assumed to potentially disturb or even to prevent normal differentiation of B-cells leading to malignant lymphomas.

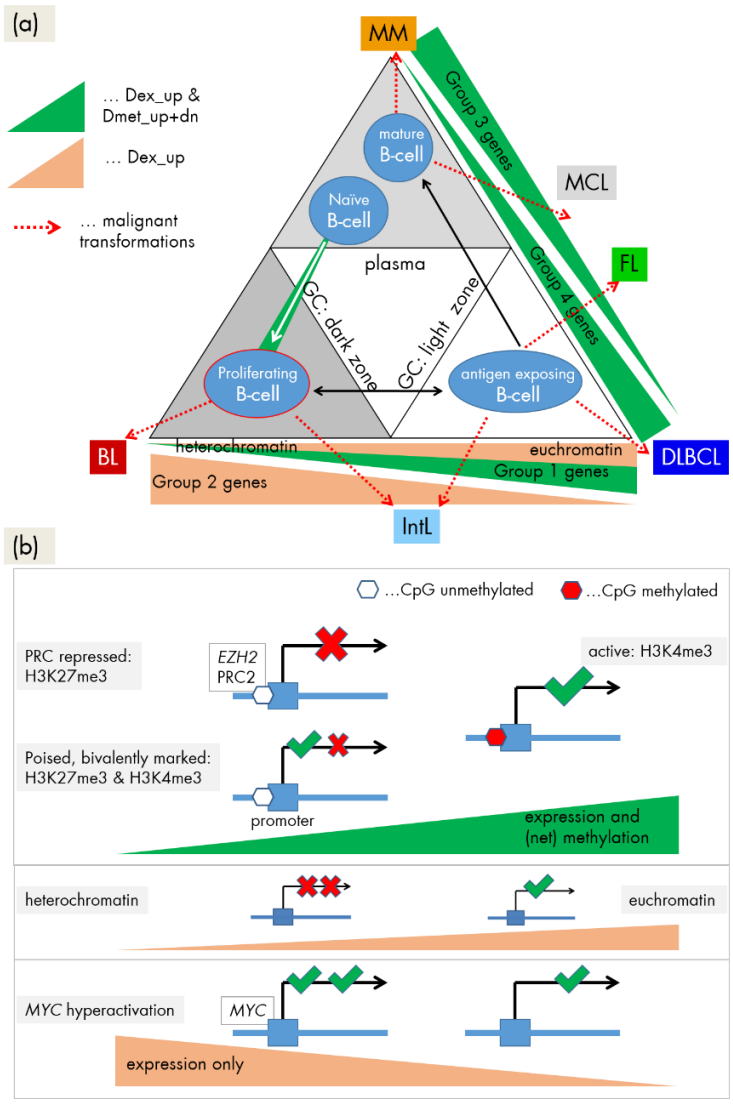


Figure 26: Epigenetic regulation scheme of lymphoma: (**a**) Scheme illustrating lymphoma heterogeneity with respect to their cell of origin and groups of affected genes. Different lymphoma subtypes can originate from GCB-cells located in the DZ of the GC (centroblasts), from its LZ (centrocytes) or from maturated plasma B-cells as indicated by the dotted red arrows. (**b**) Associated chromatin states and their remodeling due to altering histone modifications affecting transcription. The green ramp codes increasing expression and methylation associated with chromatin remodeling from inactive

and poised to active states. The orange ramp codes increasing expression without methylation changes either due to chromatin remodeling from hetero- to euchromatin or due to *MYC* hyperactivation. Gene groups are specified in Figure 25.

Many promoters in ESCs are in poised chromatin states defined by both H3K27me3 and H3K4me3 histone marks. Those bivalent states allow the cell to either activate or maintaining repressed the affected genes when needed for development hence ensuring robust differentiation [126]. In view of this basal mechanism it appears not surprisingly that bivalent chromatin states in the reference lymphoblastoid cells are strongly affected by expression and methylation changes observed in group 1, 3 and 4 genes. These bivalent promoters possibly ensure the plasticity of the genome to switch between the functional requirements in the different compartments of the GC.

Recent studies suggest that *EZH2* upregulation during the transition of NB-cell to proliferating GCB-cell (centroblast) reactivates a stem cell-like repression program not present in NB-cells and possibly featuring increased self-renewal and proliferative potential. This program accomplishes a proliferative function in GCB-cells, which makes them prone for malignant transformation into lymphoma [120]. PCR2-mediated repression seems to be almost independent of DNA methylation in normal B-cells (including proliferating centroblasts). However in lymphoma DNA methylation of these genes clearly changes, where many hypermethylated genes are targeted by PCR2 also found in stem cells [87] and centroblasts [120]. Methylation in the promoters of PCR2 genes can also associate with the opposite effect by destabilizing inactive chromatin states and thus promoting their remodeling into active ones, e.g. in group 3 genes in DLBCL. Our analysis suggests also the parallel remodeling of heterochromatin into transcriptionally active euchromatin without clear alterations of the methylation of the promoters of the involved genes.

4.2.8 CONCLUSION

From a methodical viewpoint our study shows, that integrative SOM portraying of expression and methylation data together with function mining using a battery of gene sets provides detailed insights into the regulatory landscape affecting the transcriptome and methylome and delivers a hypothesis for epigenetic mechanisms of lymphomagenesis. Our analysis is based on unmatched data sets with respect to the cancer cases used. We expect considerably improvement of the method for matched data sets.

Our study confirms previous results about the role of stemness genes during development and maturation of B-cells and the dysfunction of these regulatory programs in lymphomas presumably locking them in more proliferative or more immune-reactive states referring to GCB-cell functionalities in the dark and light zone of the GC. These dysfunctions are governed by epigenetic effects altering the promoter methylation of the involved genes, their activity status as moderated by histone modifications and also by higher-order chromatin structures, which emerge as an important regulator of gene expression.

4.3 TRANSCRIPTIONAL ACTIVITY OF CHROMATIN MODIFIERS IN LYMPHOMAS

Mutations affecting epigenetic and transcriptional modifiers are also frequently found in B-cell lymphomas [44,127]. Large-scale disruptions of DNA methylation and histone modification patterns are emerging hallmarks of these diseases. B-cell lymphomas represent a very heterogeneous cancer entity due to their complex cell of origin background. It is characterized by heterogeneous DNA methylation and gene expression patterns, which strongly vary between different lymphoma subtypes (see [67], [118] and above). These patterns also indicate profound chromatin remodeling between the cancer subtypes and also between different stages of B-cell differentiation (see section 4.2.5). In this section, we study the transcriptional activity of more than 50 epigenetic modifiers in different lymphoma subtypes and healthy controls. We ask how the expression landscape of this disease is modulated by these enzymes. Furthermore different modes of epigenetic regulation are discussed and a review of existing knowledge about selected modifiers in the context of B-cell and lymphoma biology is given.

We demonstrate how the cartography of epigenetic modifiers using SOMs helps to interpret their behavior in terms of factors that mediate writing, erasing and/or reading of epigenetic marks and how they contribute to cancer genesis and progression.

A subgroup of samples of the gene expression cohort considered in section 4.2 was analyzed here thus being divided into the same subtypes. For details concerning the cohort and preprocessing of the data see sections 3.1 and 7.1.4.

4.3.1 TRANSCRIPTION AND DNA METHYLATION UNDER CONTROL OF EPIGENETIC MODIFIERS

Mutual coupling of writer/eraser activities

Figure 27 illustrates a part of the epigenetic mechanisms regulating gene activity in terms of a simple scheme. They comprise histone modifications, DNA methylation, regulatory interactions and feedback loops between them. We take into account here only three histone modifications, namely trimethylations (me3) of H3K4, H3K9, and H3K27 and DNA methylation of CpGs in the promoter regions of affected genes. Each modification is described as a balance between writing and erasing reactions catalyzed by methyltransferases (for writing methylation marks to histone-lysines and DNA-CpGs) and demethylases (for erasing methylation marks from histone-lysines and DNA-CpGs), respectively. We will use the abbreviations KDM and KMT for histone lysine demethylases and methyltransferases, respectively, and DNDM and DNMT for DNA-CpG demethylases and methyltransferases, respectively. The scheme provides an idea how histone modifications couple each with another, with DNA methylation, and with gene activity.

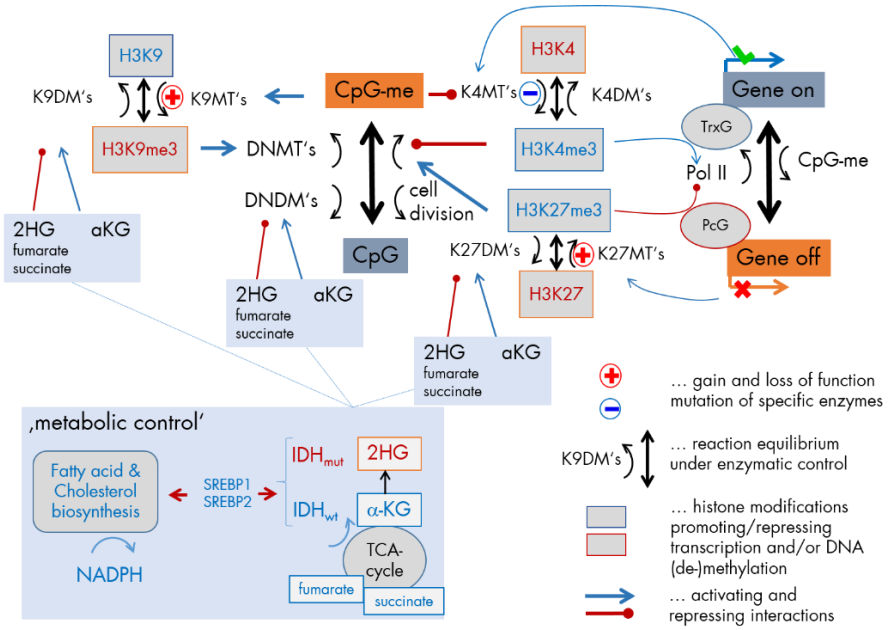


Figure 27: Transcription and DNA (promoter-) methylation under enzymatic control: The scheme summarizes selected regulatory paths affecting histone- and CpG-methylation and gene expression via different histone and DNA methylating and demethylating enzymes.

Gene expression and DNA methylation is regulated by the battery of enzymes, which either activate or inhibit transcription. For example, trimethylation of H3K4 at the promoter is assumed to activate transcription of the respective genes mediated by trithorax group proteins (TrxG). DNA methylation impacts transcription indirectly by reducing the reading capability of TrxG via reduced binding capability of H3K4me3-methyltransferases [128] and via recruitment of H3K9-KTMs [129]. Complexes containing DNA *de novo* methyltransferases (*DNMT3A/B* and *L*) are assumed to be recruited by H3K9me3 (for *DNMT3A/B*) [130] and repelled by H3K4me3 (for *DNMT3L*) [131]. This mechanism defines a positive feedback loop of *de novo* DNA methylation via H3K9me3 and *DNMT3A/B* recruitment and a negative one via H3K4me3 and *DNMT3L* inhibition. Another positive feedback loop promoting DNA methylation is formed via H3K27me3 and DNMT recruitment [132].

H3K4me3 and H3K27me3 at the promoter act antagonistically, leading to transcriptional activation and repression of the affected genes, respectively. They also have an impact on the regulation of developmental genes in fate decisions [133]. Both processes require reader-writer complexes, namely TrxG and polycomb group proteins (PcG) [134,135], respectively. The latter ones form polycomb repressive complexes (PRC), either PRC1 or PRC2, which act in sequential manner to stably maintain gene repression (see also Figure 3b). PRC2 writes H3K27me3, which is subsequently read by PRC1 creating a silent chromatin state. DNA methylation is also affected by the maintenance methyltransferase *DNMT1* to recover methylation marks at the newly synthesized DNA strands after cell division. High methylation and presumably also proliferation rates of the cells require

high *DNMT1* activities for methylation maintenance [136]. Bivalently (with H3K4me3 and H3K27me3) marked histones give rise to so-called poised promoters, which are 'easy switchable' between active or inactive transcriptional programs by erasing either the H3K27me3 or H3K4me3 marks, respectively.

Mutations of EZH2 and MLL2 potentially induce hypermethylation

Genome screening in patients with lymphoma have detected a series of mutations in genes involved in the epigenetic regulation of transcription [4,137–141] (see Table 3). Mutations of critical role in lymphomagenesis occur in genes, such as *KMT6* (alias *EZH2*, being a K27MT) leading to a gain of function preferentially in DLBCL [4–6], and *KMT2B/2D* (alias *MLL2*, being a K4MT) giving rise to its loss of function in FL, as well as in DLBCL [142] (see red plus and blue minus signs in Figure 27 and also in the simplified scheme shown in Figure 28a). Our scheme suggests, in this particular case, an increase in DNA methylation and a decrease in gene activity based on the altered activities of these enzymes (see Figure 28b for illustration). Resulting hypermethylation will affect PcG- and TrxG-related genes as well. Indeed, net-hypermethylation of PRC2-target genes was reported for DLBCL, compared with healthy B- and GCB-cells (see section 4.2.1.2). This methylation change is induced by hyper-trimethylation of H3K27 as found recently in enzyme activity experiments [143].

Epigenetics under metabolic control

IDH1 and *IDH2* (for short *IDH1/2*) catalyze the interconversion of isocitrate and α-ketoglutarate (α-KG alias 2-oxoglutarate). α-KG is a tricarboxylic acid (TCA) cycle intermediate and an essential cofactor for many enzymes, including Jumonji C (JmjC) domain containing KDMs such as *KDM2A*, *4B/C*, *5C* and TET-family DNDMs [144]. Cancer-associated *IDH1/2* mutations alter the enzymes such that they reduce α-KG to the structurally similar metabolite (R)-2-hydroxyglutarate (2-HG). α-KG generates nicotinamide adenine dinucleotide phosphate (NADPH), whereas mutant *IDH1/2* converts α-KG into 2-HG and consumes the reducing agent NADPH. 2-HG has been shown to inhibit JmjC-KDMs and TET-DNDMs leading to aberrant epigenetic modifications in tumor cells [145–147]. The inhibitory effect of 2-HG is expected to have a similar effect on our regulatory network as the mutations of *EZH2* and *MLL2* (Figure 28c). Mutations of *IDH1/2* are frequent events in tumors such as gliomas [46,63,148] and leukemia [149]. *IDH1/2* mutations are, however, rather scarce in lymphomas [150] and cannot account for such parallel effects.

One can, however, hypothesize that intermediate products of the TCA-cycle, such as succinate and fumarate, have a similar effect on epigenetics like 2-HG thus proving a possible explanation of the observed effects [144] (Figure 28c). Recall that widespread metabolic alterations allow tumor cells to remain and proliferate in certain tumor microenvironments [151]. Among lymphomas, especially BL but partly also IntL are characterized

by high proliferative activity, strongly activated energy metabolism and mitochondrial function, which were often paralleled by activated *c-MYC* expression (see section 4.1.7 and [167]). Such massive metabolic changes suggest interference with epigenetic regulation via modifying enzymes responding to metabolites. Activating interaction, e.g., due to high abundance of α-KG is expected to demethylate histone lysines and DNA CpGs and to activate expression, *i.e.*, alterations not corresponding to the observed ones (Figure 28d). However, other intermediate products of the TCA-cycle, namely succinate and fumarate, are shown to counteract α-KG. Their enhanced production in metabolically activated lymphoma subtypes possibly explains the observed trend [144] (Figure 28c).

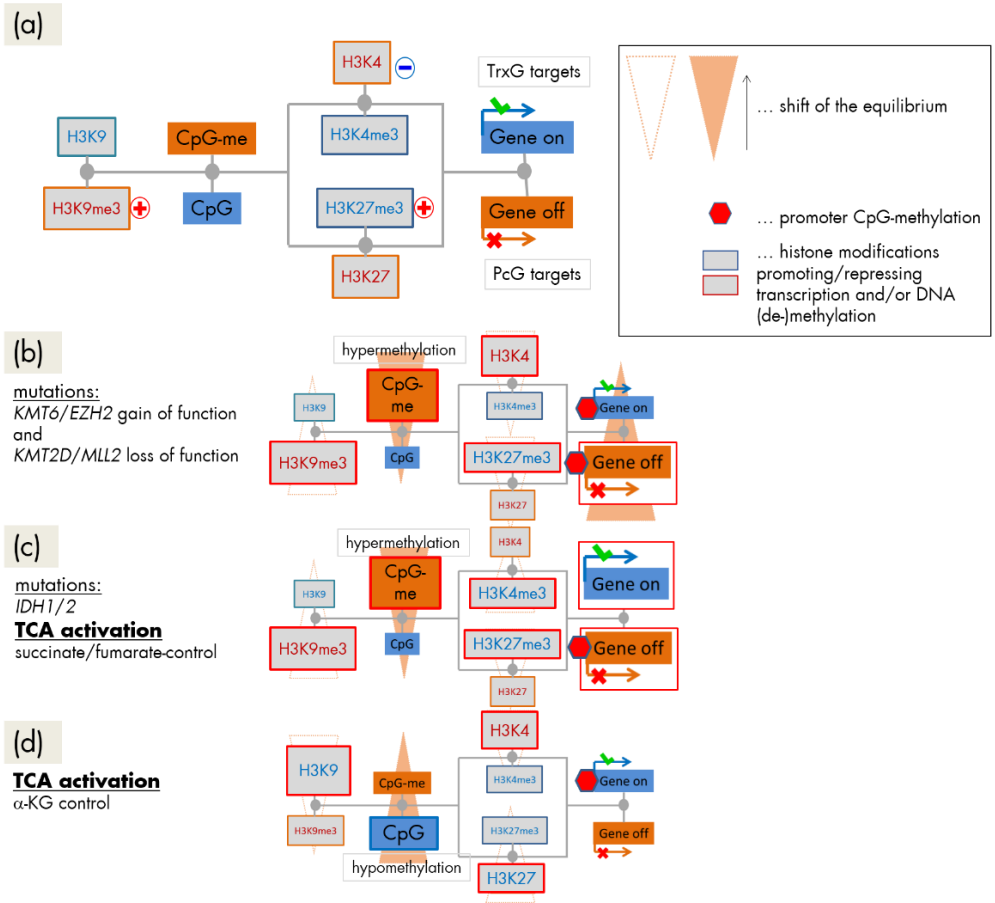


Figure 28: Effect of selected mutations of epigenetic modifiers and of TCA metabolism on CpG methylation and gene expression: (**a**) Simplified sketch of the scheme shown in Figure 27. Plus and minus signs indicate gain and loss of function mutations, respectively; (**b**) The scheme suggests that mutations of *EZH2* and *MLL2* result in DNA-hypermethylation and repression of transcription. Alterations are shown by enlarged boxes and red frames; (**c**) Mutated *IDH1* produces 2-HG that inhibits a series of KDMs. It results in hypermethylation and altered gene expression. The same trend is expected for increased TCA-activities with increased levels of fumarate and succinate, both inhibiting the same modifiers as 2-HG; (**d**) TCA activation with increased amounts of α-KG will demethylate DNA and repress expression. Note that global changes of methylation are paralleled by local ones due to the hypothesized alterations of gene expressions: Red marks indicate promoter methylations in (b) – (d) recruited by PcG-repressed and/or TrxG-deactivated genes.

Table 3: Chromatin modifiers with possible relevance for lymphoma: An overview.

Target	Type[1]	Enzyme	Alias	Mark	txn[2]	mut[3]	Dex[4]	Spot[5]	Comment[6]
	Writer **E**raser **R**eader			**Me**	act rep		**LvsBc**		
H3K4	W	***KMT2A***	*MLL*		act		−	IJ	TrxG:MLL complex
		KMT2B	*MLL2, KMT2D*		act	x	−	J	loss of function in DLBCL/FL, TrxG:MLL complex
		KMT2F	*SETD1A*		act		−	(H)	
		KMT2G	*SETD1B*	Me3	act		−	H	
		KMT3C*	*SMYD2*	Me2/Me3	act		x	(MM)	
		KMT3E	*SMYD3*	Me2/Me3	act		+	(D)	
		PRDM9	*MSBP3, PFM6*	Me3	act		−	(I)	
		SETMAR*	*METNASE*		act		x	B	
	E	***KDM1A****	*LSD1, AOF2*	Me1/Me2	rep		+	B	'gene body cleaner'
		KDM5A	*JARID1A, RBBP2*	Me2/Me3	rep		x	(A)	JmjC, 'gene body cleaner'
		KDM5B	*JARID1B, PLU1*		rep		−	(F)	JmjC
		KDM5C	*JARID1C, SMCX*		rep		+	(I)	JmjC
H3K9	W	***KMT1C****	*EHMT2, G9A*	Me1/Me2	rep		x	(B)	
		KMT1D*	*EHMT1, GLP*	Me1/Me2	rep		+	(D)	
		KMT1E	*SETDB1*	Me3	rep		x	(I)	
		KMT6(A)*	*EZH2*		rep	+	+	D	gain of function in cancer/DLBCL/FL, PRC2 complex
		KMT8	*PRDM2, RIZ*		rep		−	(D)	missense mutation in DLBCL
	E	***KDM1A****	*LSD1, AOF2*		act		+	B	
		KDM3A	*JMJD1, TSGA*	Me1/Me2	act		−	(IM)	JmjC
		KDM3B	*JHDM2B*		act		x	(C)	JmjC
		KDM4A*	*JMJD2*	Me3	act		−	(J)	JmjC
		KDM4B	*JHDM3B*	Me3	act		−	J	JmjC
		KDM4C*	*JHDM3C*	Me3	act		−	I	JmjC
		KDM4D	*JMJD2D*	Me2/Me3	act		x	(B)	JmjC
		KDM7A*	*JHDM1D*	Me2	act		−	H	JmjC
		MINA	*MDIG, ROX*	Me3	act		+	(B)	
H3K27	W	***KMT1C****	*EHMT2, G9A*		rep		x	(B)	
		KMT1D*	*EHMT1, GLP*		rep		+	(D)	
		KMT6(A)*	*EZH2*		rep	+	+	D	gain of function in cancer/DLBCL/FL, PRC2 complex

		KMT6B	*EZH1*		rep		−	J	PRC2 complex
		WHSC1	*NSD2, MMSET*		rep		+	D	mutated in BL and MCL, opens chromatin
	E	***KDM6A***	*UTX*	Me2/Me3	act		−	(IM)	
		KDM6B	*JMJD3*	Me2/Me3	act		−	J	involved in inflammatory response, JmjC
		KDM7A*	*JHDM1D*	Me2	act	x	−	H	JmjC
H3K36	E	***KDM2A***	*FBXL11, JHDM1A*	Me2			−	(H)	JmjC
		KDM4A*	*JMJD2*	Me3	rep		−	(J)	JmjC
		KDM4C*	*JHDM3C*	Me3	rep		−	I	JmjC
		KDM8	*JMJD5*	Me2	rep		x	(B)	JmjC
	W	***KMT2H***	*ASH1L*		act		−	(I)	
		KMT3A	*SETD2, SET2*	Me3	act		−	J	recruits MMR
		KMT3B	*NSD1, STO*				−	(J)	
		KMT3C*	*SMYD2*	Me2	act		x	(MM)	
		SETMAR*	*METNASE*	Me2	act		x	B	
H3K79	W	***KMT4***	*DOT1L*		act	x	x	(MM)	loss of function in lymphomas
DNA	W	***DNMT1***			rep		+	D	maintenance
		DNMT3A			rep		x	(D)	*de novo* methylation
		DNMT3B			rep		+	B	*de novo* methylation
		DNMT3L			rep		+	(D)	induces *de novo* DNA methylation by recruitment or activation of DNMT3
	E	***TET3***			act		−	I	
	R, E	***MBD2***			act/ rep		−	(I)	mediates CpG-methylation signal

[1] Here we consider only KMTs and DNMTs as epigenetic writers and KDMs and DNDMs as erasers. Epigenetic readers possess effector domains and recognize and bind to modified residues. Many 'classical' TFs (that 'read' special DNA binding motifs) are also epigenetic readers because their binding to DNA is also governed by epigenetic marks (see also [45,153] and Figure 3b).

[2] Expected net effect on the transcriptional activity of the affected genes. In general there is no one-to-one relation between a certain epigenetic modifier and the change of gene expression. Combinations of modifiers and their marks give rise to a large variety of options (also called chromatin code). Here we assign the proposed effects of chromatin marks on gene expression according to GeneCards (www.genecards.org).

[3] Activating/gain of function (+) or deactivating/loss of function (x) mutation observed in lymphoma.

[4] Differential expression with respect to B-cells: +…up; − …down; x…indifferent.

[5] Spot cluster: e.g., 'A'… gene belongs to spot 'A'; (A)…gene is found near spot 'A' in the map; spot characteristics: 'B', 'C', 'D': up in BL and down in DLBCL/FL; 'F': up in FL; 'I': up in BCL and MM and down in BL and partly DLBCL; 'J': up in B- and GCB-cells and down in lymphomas; 'H': up in B-cells, tonsils and FL, down in BL; see also section 4.2.1.3.

[6] Mutation data and assignments to lymphoma classes were taken from [4,137–140].

* Enzymes marked with asterisks perform multiple roles by catalyzing more than one lysine side chain.

In summary, our simple scheme predicts the activation of repressed PcG-related genes paralleled by DNA-hypermethylation after mutations of the genes *EZH2* and/or *MLL2*, which both code for KMTs. Mutations of *IDH1/2* or an increased TCA-activity are suggested to have a similar effect on histone and DNA methylation. However *IDH1/2* mutations are scarce in lymphoma suggesting alternative mechanisms that couple metabolism with epigenetics.

4.3.2 THE EXPRESSION SOM COORDINATE SYSTEM

For results of the lymphoma gene expression SOM training see section 4.2.1.3. The spot map (Figure 29a, compare with Figure 21) selects defined spot areas ('A' – 'J', 'IM', and 'MM') representing clusters of co-expressed genes being overexpressed in a certain sample class. Accordingly, the map can be segmented into areas of characteristic differential expression between the lymphoma classes and healthy controls (e.g., BL_vs_DLBCL means that the area contains genes overexpressed in BL compared with DLBCL), see Figure 29c. The dashed borderlines between these areas serve as guide for the eye inspection only. In reality, the areas are fuzzy without clear-cut borderlines. For the sake of a simple and clear description, we divide the map into four quadrants (Q1–Q4) and a central region (Z) (see Figure 29b). Q1 can be assigned to genes upregulated in 'MM' (MM_up), Q2 contains diverse deregulation patterns, Q3 can be assigned to genes specifically upregulated in BL (BL_up), Q4 is assigned to DLBCL_up and FL_up, and, finally, Z includes genes of almost invariant expression.

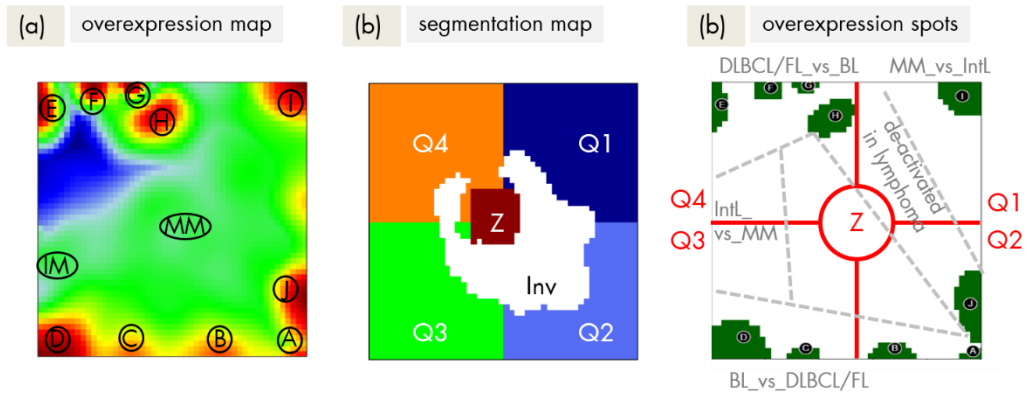


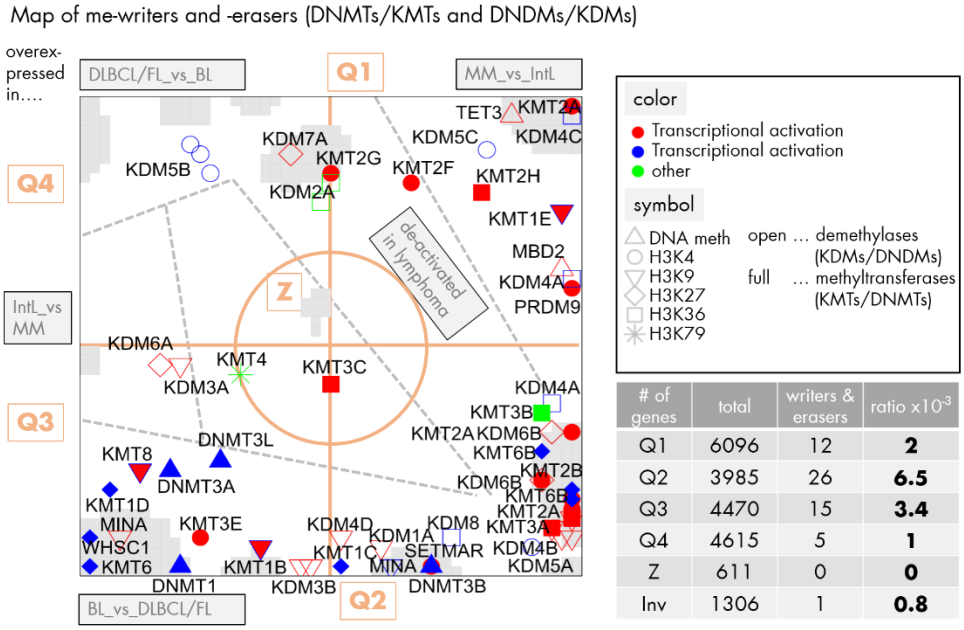
Figure 29: Expression SOM characteristics of lymphoma: (**a**) The overexpression spot summary map shows all overexpression spots in red, which were detected in the lymphoma cohort studied [114]; (**b**) and (**c**) We segmented the map into areas of characteristic differential expression between the lymphoma classes and healthy controls: Four quadrants Q1–Q4 and a central area Z. In addition, we separately considered genes with invariant expression profiles, which populate the blue area in the variance map (Inv).

4.3.3 EXPRESSION CARTOGRAPHY OF EPIGENETIC MODIFIERS

In this subsection, we aim to verify the predictions made above. Using lymphoma expression data we systematically monitor the expression levels of about 50 methylating and demethylating enzymes in different lymphoma subtypes and healthy controls to document their heterogeneity in regulating gene activity (see Table 3 for an overview). Note that the enzymatic activity is modulated by a series of post-transcriptional and -translational factors (such as posttranslational modifications, local accessibilities and concentrations of cofactors), which are beyond our data.

SOM expression map of epigenetic modifiers

For a holistic view, we make use of SOM-portrayal method, which locates the genes coding the modifying enzymes into a quadratic map. The map allows to deduce the expression characteristics of a gene from its location in the map (see section 3.4).



# of genes	total	writers & erasers	ratio x10^{-3}
Q1	6096	12	**2**
Q2	3985	26	**6.5**
Q3	4470	15	**3.4**
Q4	4615	5	**1**
Z	611	0	**0**
Inv	1306	1	**0.8**

Figure 30: Mapping of writers and erasers of epigenetic methylation marks into the gene expression landscape (DexSOM) of lymphoma. The map is segmented into regions of specific differential expression between the lymphoma classes by dashed lines. The mode of differential expression is indicated in the grey boxes. Spot-clusters of differentially expressed genes are grey-colored. The map is further divided into four quadrants Q1–Q4 and a central region Z (see Figure 29b and c for segmentation of the SOM). Epigenetic modifiers are labeled as shown in the right part of the figure. A few genes are redundant because more than one probe set interrogate them (e.g., *KDM4A, KDM6B, and KMT6B*). The redundant probe sets are located in close proximity reflecting strongly correlated expression profiles (see Table 3 for assignment of the enzymes). We counted the number of genes and of writers and erasers of methylation marks in each of the areas. These modifiers were strongly enriched in Q2, on intermediate levels in Q1 and Q3 and on low levels in Q4, Inv and Z (see table).

The overview map shown in Figure 30 summarizes the location of genes encoding writers and erasers of histone-lysine and of DNA-CpG methylation marks listed in Table 3. Interestingly, the genes encoding epigenetic modifiers strongly accumulate in Q2, to a less degree in Q1 and Q3, but they are almost absent in Q4 (Figure 30). This asymmetric distribution reflects the fact that gene expression of the majority of methylating and demethylating enzymes is either up- or down-regulated in lymphomas compared with B-cells and/or activated in BL compared with DLBCL and FL. Enzymes up-regulated in DLBCL and/or FL compared with BL are however rather scarce.

According to the histone code hypothesis numerous of histone modifications are assumed to regulate gene expression of the associated genes. In Figure 30, we color coded the assumed effect on transcription by symbols being red for activation, blue for silencing and green for unknown effect. Enzymes promoting gene expression are slightly enriched in Q1, which contains genes down-regulated in lymphomas. Thus, activating marks in Q1 correspond to the expression in B-cells. On the other hand, Q2 and Q3 are more puzzling, as those reflect no preference for activating and de-activating marks.

DNA methylating enzymes: DNA-MTs and -DMs

DNMTs and DNDMs accumulate in opposite corners of the map shown in Figure 31a (Q3 *versus* Q1) being up- and down-regulated, respectively, in lymphomas compared with B-cells. The *de novo* methyltransferases *DNMT3A,B,L* show only moderate effect also between GCB- and B-cells in agreement with previous results [125]. All of them, but especially *DNMT3A*, have a high gene activity in BL and a relatively low one in the other lymphoma subtypes. The maintenance methyltransferase *DNMT1*, on the other hand is strongly up-regulated in lymphomas and GCB-cells showing also maximum activity in BL. In the context of the GCB phenotype, Shaknovich et al. [125] attributes several functions to *DNMT1* like chromatin condensation/de-condensation, maintenance of genomic DNA methylation and also repair of double strand DNA break [125].

Note that proliferative activity is extraordinarily high in BL, also requiring high activities of CpG methylation maintenance and DNA repair processes. Demethylases, on the other hand, are on highest expression levels in MM revealing a strong antagonism of DNA methylation and demethylation between MM and DLBCL (and BL), between BL and DLBCL and partly between B- and GCB-cells. Note that DNA methylation is not a simple sum of DNMT activities. A comparison of the epigenomes between normal and cancerous stem cells, and between pluripotent and differentiated states shows that the presence of at least two DNMTs is required for differential DNA methylation effects [154]. Moreover DNA (de-) methylating enzymes operate often in concert with histone modifications.

DNA demethylases of the TET-family play an important role in fine regulation of DNA methylation. Their inactivation leads to the establishment of DNA hypermethylation phenotype [155]. *TET3* locates near spot 'I' meaning low expression levels in DLBCL and IntL but relatively strong ones in MM and B-cells. This inactivation in DLBCL and IntL indeed accom-

panies with aberrant methylation patterns partly resembling methylator phenotypes observed in other cancer types such as colorectal cancer and glioma (see section 4.2.3). The possible role of TET-family proteins in coupling mechanisms with the TCA-metabolism will be discussed below. *MBD2* is a methyl-CpG-reader that has been reported to be both a transcriptional repressor and a DNDM [156]. In lymphoma it shows a similar expression profile as *TET3*.

Histone K27MTs and DMs

Figure 31b–f disentangles the KMTs and KDMs according to the position of the methylated/de-methylated lysines in the H3 subunit. Firstly, we see strong activation of *EZH2/KMT6*(*A*) in all lymphoma subtypes compared with B-cells (Figure 31b). A high fraction of DLBCL and of grade 3 FL harbors *EZH2* mutations, suggesting that these mutations are early events of lymphomagenesis [117]. *EZH2* gain of function mutation presumably favors the emergence of malignant disease by suppressing anti-proliferative and differentiation processes [6,88,121,157]. In GCB-cells *EZH2* bivalent chromatin domains are built at key regulatory regions to temporarily repress GCB-cell differentiation. These physiological effects are amplified by somatic mutations through enhanced silencing of *EZH2*-targets leading to malignant transformation into highly proliferative lymphoma types [121]. Our data shows that *EZH2* expression in all lymphoma subtypes except for BL is decreased compared with GCB-cells, but it is increased compared with B-cells.

EZH1/KMT6B, another H3K27-methyltransferase (and a homolog of *EZH2*) and the KDMs *UTX/KDM6A* and *JMJD3/KDM6B* show roughly antagonistic profiles compared with *EZH2* as they are strongly deactivated in lymphomas. *KDM6A* and B play an important role in the differentiation of tissues from embryonic stem cells (ESC), where their deactivation impairs differentiation [158,159]. Moreover, somatic mutations of *KDM6A* have been found in a number of cancer types indicating the importance of this enzyme in tumorigenesis. The antagonistic changes of *EZH2* and *KDM6A* in lymphomas suggest that the methyltransferase and demethylase act in a concerted fashion and shift the methylation equilibrium towards trimethylated H3K27, which promotes repressive transcriptional states.

EZH1 safeguards ESC identities and maintains repression in resting cells [160]. It is more abundant in non-proliferative adult organs and acts transcriptionally as antagonist of *EZH2*, which *de novo* establishes H3K27me3 in dividing cells [161–163]. Our data thus supports the view that activation of *EZH2* in lymphoma 'over-represses' suppressors of proliferative programs, whereas de-activation of *EZH1* 'under-represses' maintenance suppressor of resting cells presumably thus destabilizing their state.

WHSC1, another K27MT, changes in lymphomas in a similar way as *EZH2*. It is frequently mutated in MCL [164] and partly also in BL [165]. It has been suggested that *WHSC1* mutations in these lymphoma types are associated with open chromatin in their cell(s) of origin [165]. Note that *WHSC1* and other genes located in Q3 (spot 'D' in the DexSOM) are up-regulated in BL and partly GCB-cells thus supporting this view (see also section 4.2.5, where we assign these genes to euchromatin states in BL). *WHSC1* is also

activated in MM where it is thought to open chromatin structure [166]. Solely the enzyme *KDM7A* is located in Q1 (spot 'H') due to the fact that it is specifically activated in FL and MM. This enzyme de-methylates H3K27me2.

In summary, the expression profiles of genes coding for H3K27 (de)-methylating enzymes reflect concerted deregulation of repressed (including also poised) genes in lymphomas. These genes seem to become 'over-repressed', which presumably results in a loss of plasticity of cellular programs. In consequence, the cells become unable to return into an active state as required for healthy GCB-cell function.

Histone K4MTs and DMs

Genes encoding methyltransferases for Lys-4 accumulate in Q2 (mainly in spot 'J') thus resembling the profiles of a series of K27MTs and especially K27DMs (see Figure 31b and c for comparison). This region also contains *KMT2B* (*MLL2*) frequently carrying a loss of function mutation in about 30% of DLBCL and 90% of FL patients [142,167]. Down-regulation of this gene is indeed observed in lymphomas (Figure 31c). *KMT2B* is assumed to act as a central tumor suppressor [142] and to de-activate TrxG-related genes. Another gene of the *MLL*-group, *KMT2A* (*MLL*) shows a virtually identical profile suggesting similar function. *KMT2A* is targeted by chromosomal translocations deactivating this gene on Chr 11q23 in lymphomas [168].

K4 modifying enzymes are depleted in Q3 and partly enriched in Q1 in sharp contrast to K27 modifiers, reflecting the partly antagonistic role of both types of modifiers in either promoting or repressing transcription. Only *KMT3E* is found in Q3 (near spot 'D'). *KMT3E* (*SMYD3*) knockdown causes cell cycle arrest and induction of apoptosis [169]. Hence, it's up-regulation in lymphomas and especially in BL associates with the opposite effects leading to increased proliferation and anti-apoptotic 'cancer hallmark' activities.

Interestingly, the H3K4 demethylase *KDM5B* (alias *JARID1B*) is among the very few genes found in Q4, which contains genes specifically up-regulated in DLBCL and FL. *KDM5B* acts as 'gene body cleaner' of near promoters and enhancers of bivalent (*i.e.*, weakly transcribed) genes by demethylating their gene bodies during ESC self-renewal and differentiation [170]. This mechanism ensures correct expression of the affected genes. The H3K4 demethylase *KDM1A* (alias *LSD1*) also demethylates H3K4me3 in gene bodies, however, in inactive genes. Interestingly, *KDM1A* is located in Q2 (spot 'B'), which roughly antagonistically switches with respect to Q4 (spot 'F') co-expressed with *KDM5B*. These results suggest that stemness genes with bivalently marked promoters co-regulate with *KDM5B* activity because their functionality is maintained by this enzyme, while genes repressed in ESC differentiation co-regulate with *KDM1A* activity. Proper function of these enzymes maintains developmental genes in their bivalent-active or repressive state.

In summary, writers and erasers of H3K4me3 tend to show an antagonistic behavior compared with the respective H3K27-modifiers, which corresponds to their mostly antagonistic effect on transcription. Up-regulation of K4DMs and deactivation of K4MTs seems to lead to under-activation of tumor suppressors controlling, e.g., apoptosis and proliferation.

Histone K9MTs and DMs

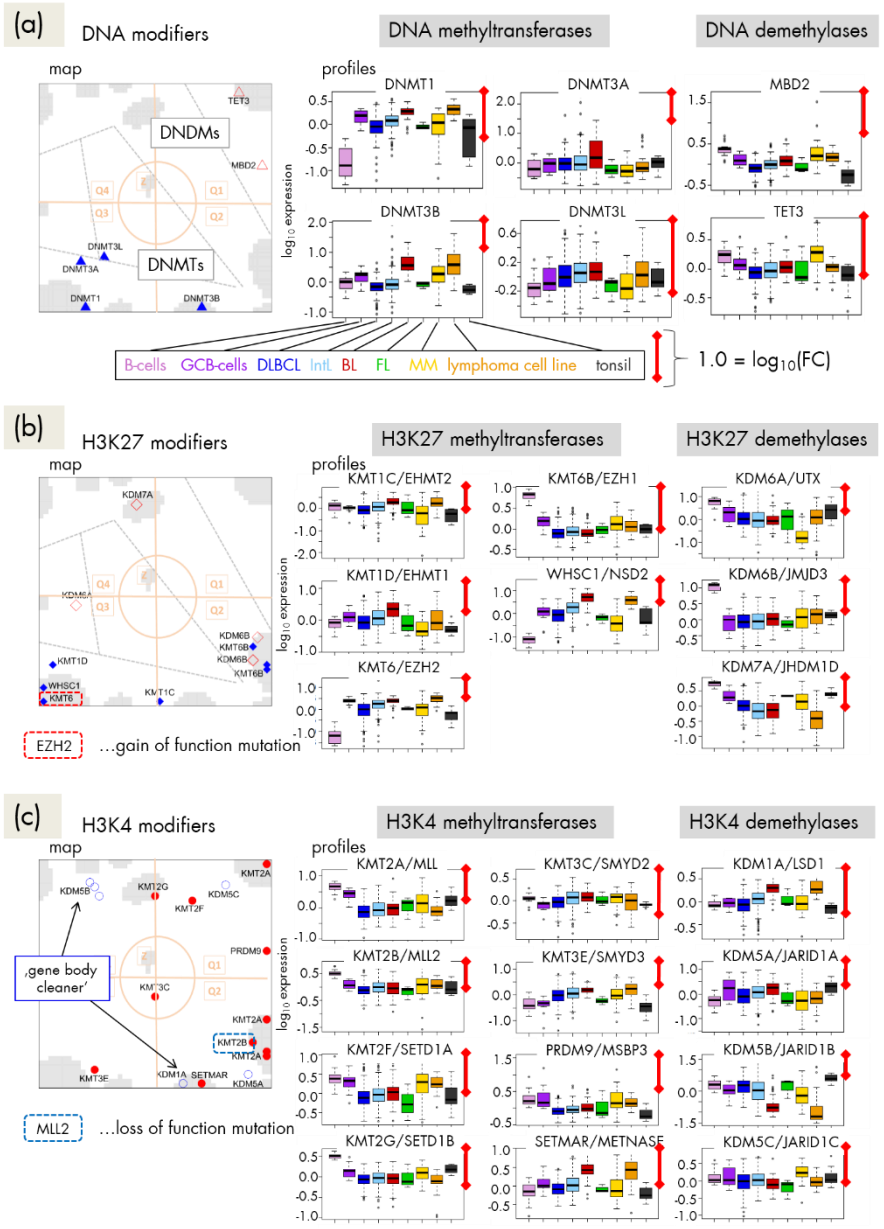
H3K9me3 promotes CpG-methylation and gene deactivation (Figure 27). The distribution of H3K9 (de-) methylating enzymes in the DexSOM shares similarities with that of H3K27me3 representing the second deactivating mark considered here (compare Figure 31b and d). Particularly, K9MTs accumulate in Q3 (spot 'D') being up-regulated in lymphomas and especially in BL. Note that part of the modifiers (e.g., *EZH2/KMT6*) affect both H3K9 and H3K27. Interestingly, K9DMs tend to occupy a wide area in the map ranging from Q3 (spots 'C') over Q2 (especially spots 'B' and 'J') to Q1 (spot 'I') thus showing activation (Q3) and deactivation (Q1 and Q2) in lymphomas compared with B-cells (see also the profiles in Figure 31d). For example, *KDM4D* (Q3) demethylating H3K9me3 [171] specifically up-regulates in BL. Other members of the KDM4-family—*KDM4B* (spot 'J'), *KDM4A* and *KDM4C* (spot 'I')—are deactivated in lymphoma thus presumably promoting H3K9me3 and DNA CpG methylation. Contrarily, these enzymes are overexpressed in other cancers such as breast, colorectal, lung and prostate cancer because they are required for efficient cancer cell growth [172,173]. *KDM4C* shows the most pronounced effect among them, being active in B-cells and MM, on intermediate level in GCB-cells and on lowest level in DLBCL, thus suggesting a certain role in DNA-hypermethylation observed in lymphomas. The activity of this enzyme clearly anti-correlates with the energy metabolism possibly due to TF regulation via *SREBP1* and/or intermediates of the TCA cycle inhibiting its activity (see below).

Histone K36MTs and DMs

Part of the KDM4-family also demethylates H3K36me3 thus activating expression according to the histone code because of its role in transcriptional elongation. This dual function is thought to repress aberrant transcription [174]. H3K36 marks distribute over the gene body and perform fine tuning of expression by interacting with RNA Polymerase II. They also play a role in nucleosome positioning, alternative splicing and exon activation [175]. Interestingly, H3K36me3 is required as reading mark for DNA repair proteins by acting as chromatin switch, which makes DNA accessible for double strand repair [176]. For example, high levels of *KMT3A* (*SETD2*) ensure accurate homologous DNA repair in human cells [177]. *KMT3A* expression down-regulates in lymphomas and especially in DLBCL (Figure 31e, Q2, spot 'J') suggesting reduced potential for DNA repair. On the other hand, the associated demethylase *KDM4A* counteracting *KMT3A* shows a similar expression profile indicating a more complex effect. Other members of the KMT3 family (*KMT3B* and *KMT3C*) co-regulate with *KMT3A* thus promoting H3K36me3 demethylation in lymphomas. In summary, K36MTs and K36DMs both accumulate in Q1 and Q2 with homogeneous expression profiles reflecting their down regulation in lymphomas compared with B-cells.

Histone K79MT

KMT4-mediated H3K79 di- and tri-methylation is essential for embryogenesis and hematopoiesis. The sole enzyme responsible for H3K79 methylation considered here is *KMT4* (*DOT1L*) promoting transcription by stimulating its elongation phase [178]. *KMT4* is located in the central zone of the map (Z) near spot 'MM' (Figure 31f) up-regulated only in MM suggesting a specific role of this enzyme in this lymphoma class. *DOT1L* has attracted the interest concerning the emergence of *MLL*-rearranged leukemia, where mistargeting of *DOT1L* leads to aberrant H3K79 methylation [179].



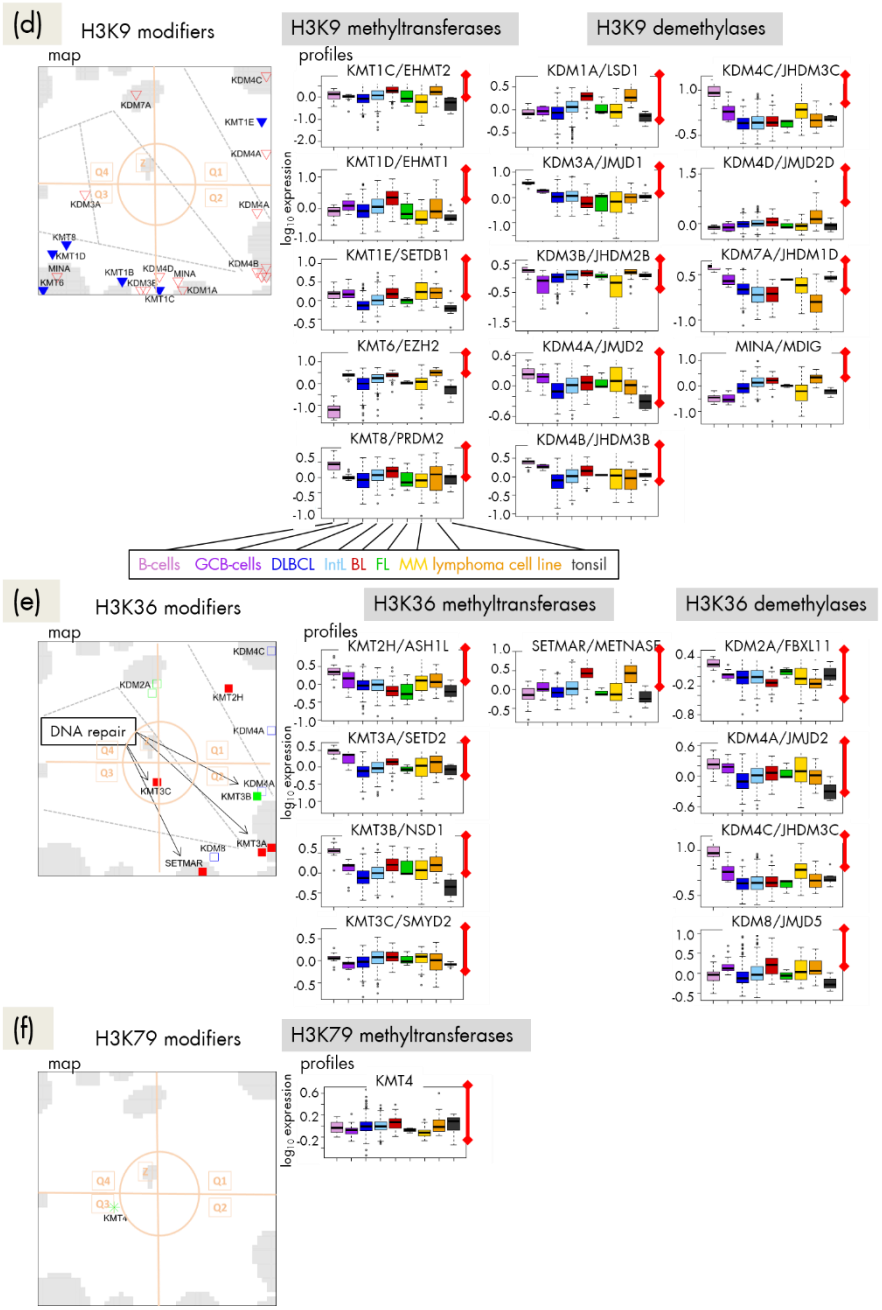
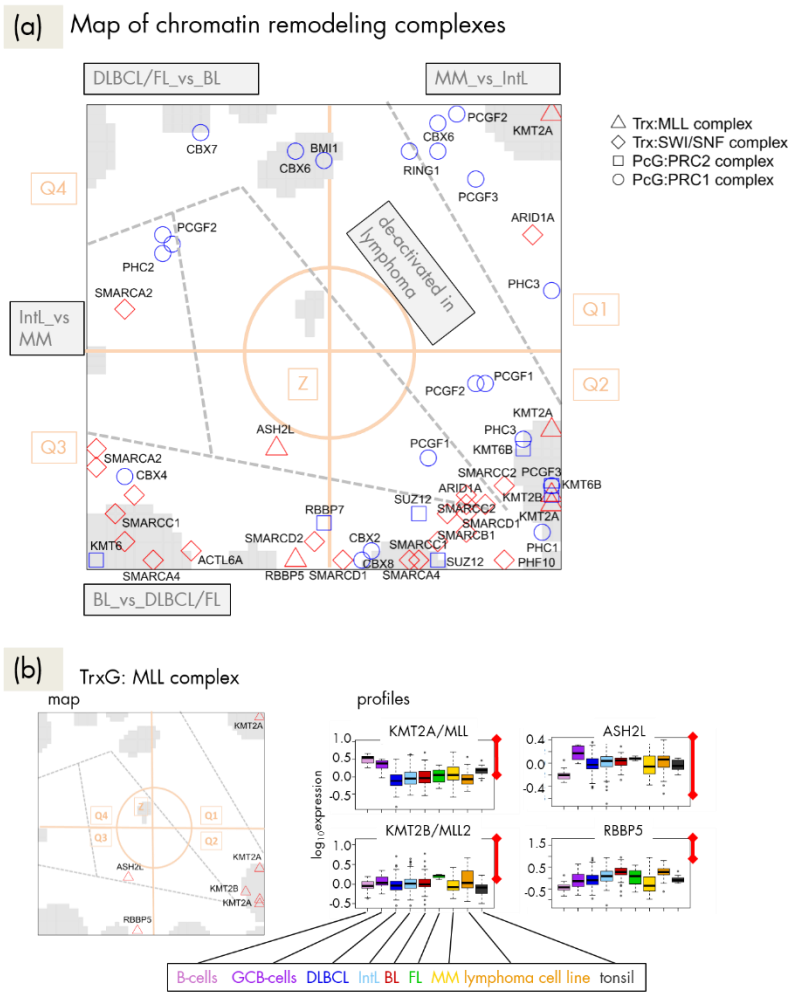


Figure 31: Groupwise mapping of writers and erasers of methylation marks at CpGs and lysine side chains of histone subunit H3 and their expression profiles: (**a**) DNA, (**b**) H3K4, (**c**) H3K27, (**d**) H3K9, (**e**) H3K36, and (**f**) H3K79. The symbols and their coloring are assigned in Figure 30. Modifiers (de-)marking more than one lysine residue are shown several times. Note the different scales of ordinate-axes. The length of the red scale-bar refers to a fold change (FC) between two expression values of one order of magnitude.

4.3.4 EXPRESSION CARTOGRAPHY OF CHROMATIN REMODELING COMPLEXES

A large fraction of modifiers discussed in the previous subsection acts in concert each with another and forms different kinds of functional complexes together with writers. In this subsection we regrouped the enzymes and complemented them with relevant writer-proteins in a complex-related order to discover their expression profiles in the cohort studied. The overview map shown in Figure 32a reveals accumulation of the complex-related genes (except for PRC1-related genes) in Q2 and Q3 reflecting a further narrowing of the regulatory space compared with the full set of modifying enzymes (compare with Figure 30a). Recall that Q2 (especially spot 'J') collects enzymes down-regulated in lymphomas, whereas Q3 contains genes up-regulated in lymphomas with high expression levels in BL and relatively low levels in DLBCL and FL. Hence, the map reflects a dual antagonism of activation/de-activation patterns, namely (i) up in lymphomas and down in B-cells and *vice versa* and (ii) up in BL and down in DLBCL and FL where, however, the antagonistic mode is almost lacking.



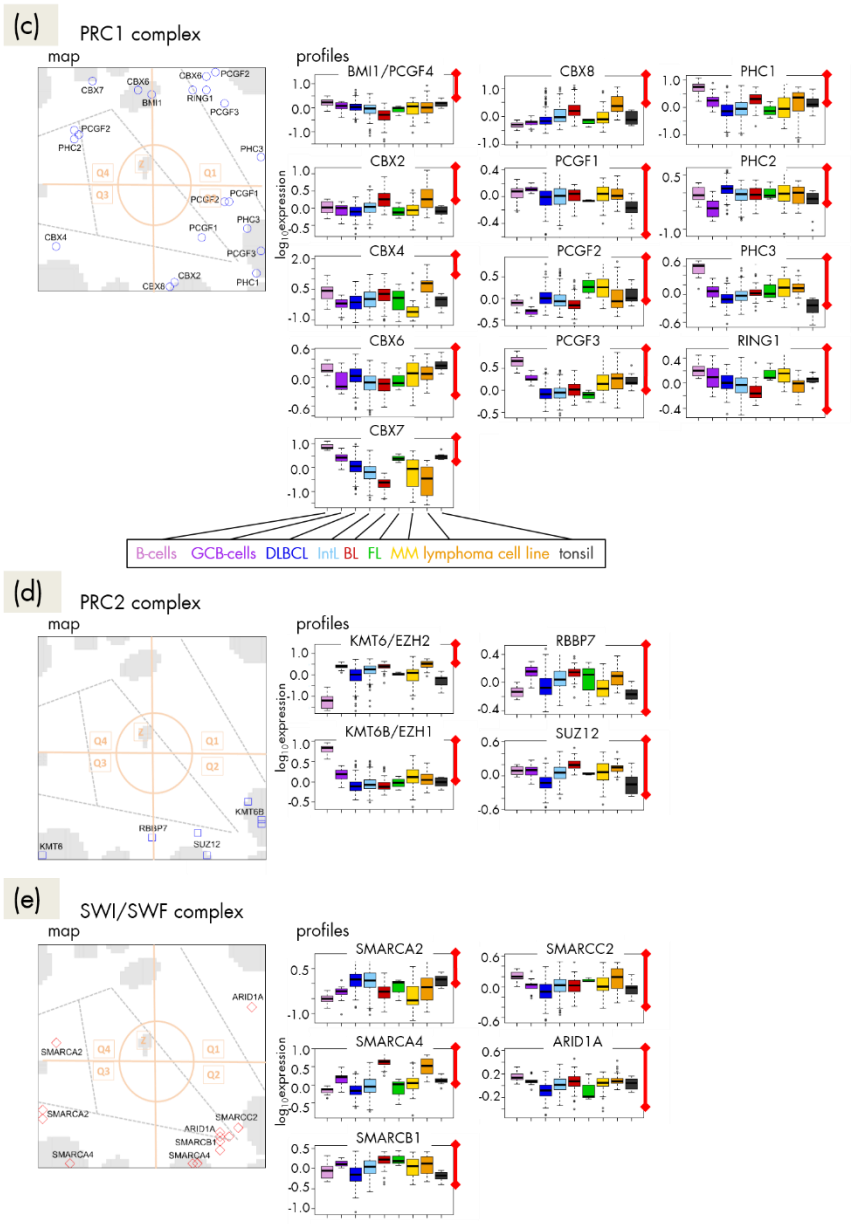


Figure 32: Map of ingredient-genes of chromatin modifying complexes: (**a**) overview map, (**b**) TrxG/MLL-, (**c**) PRC1-, (**d**) PRC2-, and (**e**) ATP-dependent chromatin remodeling SWI/SWF complexes, respectively.

TrxG/MLL complex

TrxG/MLL is a reader-writer complex that leads to H3K4 trimethylation and activates expression [162,180]. The retinoblastoma binding protein 5 (*RBBP5*) and *ASH2L* are conserved subunits of the MLL complex, which form a heterodimer with intrinsic methyltransferase activity required for methylation of H3K4 [181,182]. Expression of both compounds is high in proliferative GCB-cells and BL, and low in B-cells, MM, and partly DLBCL

(Figure 32b). Their profiles partly diverge from that of the KMTs showing partly antagonistic changes.

PRC1 complex

The PRC1 complex stabilizes gene repression established by PRC2 beforehand (see below). Main compounds are chromobox homolog (CBX)-family proteins (alias heterochromatin protein, HP1), which are essential for heterochromatin formation and stabilization. Part of them (*CBX2*, *4* and *8*) are located in Q3 and partly Q2, whereas the others (*CBX6* and *7*) are found in the opposite half of the map (Q4 and Q1) revealing either up-regulation in BL and down-regulation in DLBCL and FL or *vice versa*, respectively (Figure 32c). Large amounts of CBX proteins seem to serve as reservoir for heterochromatin formation to bind to the nucleosomes upon request and stabilize repressive chromatin states during cell differentiation [183]. The antagonism of CBX expression between BL and DLBCL suggests chromatin remodeling between euchromatin in BL and heterochromatin in DLBCL as hypothesized in section 4.2.7. *CBX2* binds to genes deactivated by H3K27me3 [184]. Such genes accumulate in spot 'B' together with *CBX2* indicating coregulation and particularly deactivation in DLBCL (see next subsection). Elevated expression of *CBX7* and of *BMI1* (alias *PCGF4*) in DLBCL and FL (Q4) are related to aberrant regulatory programs inducing high-grade tumor transformation and chemotherapy resistance [88]. Other compounds of PRC1 are polycomb group RING finger (*PCGF*) and polyhomeotic homolog (*PHC*) proteins accumulating in Q1 and Q2 thus indicating down-regulation in GCB-derived lymphomas. Importantly, different ingredients of PRC1 fulfill different roles in chromatin condensation and they bind also to different genomic loci thus defining different subgroups of PRC1 [184], which possibly explains the different profiles of CBX and of *PCGF/PHC* proteins.

PRC2 complex

PRC2 catalyzes methylation of H3K27me3 through its 'enhancer of zeste' (*EZH*) constituents (see above). Other compounds are *SUZ12* and *RBBP7* both required for the establishment of specific expression programs needed for differentiation of ESCs [185]. All PRC2 compounds studied are found in Q2 and Q3 (Figure 32d), which well correspond to the distribution of CBX-proteins discussed above: PRC2 genes in Q2 and Q3 are *de novo* and temporarily repressed in DLBCL by H3K27me3. Afterwards they transform into permanently repressed heterochromatin by CBX-PRC1 binding. Note also that *SUZ12* and PRC2-targets strongly enrich in Q1 and Q4 (see section 4.2.6), containing genes, which antagonistically switch compared with Q2 and Q3 genes. This suggests that PRC2- and *SUZ12*-targets up-regulate in DLBCL owing to 'over-suppressing' their regulators. Taken together altered, subtype-specific expression of PRC1 and PRC2 genes is a leitmotif in lymphomas suggesting significant regulatory roles of PRC1 and PRC2 in development of both normal B-lymphocyte and lymphomas [88].

SWI/SWF complex

The SWI/SWF complex belongs to the ATP-dependent chromatin remodeling complexes. They remodel chromatin (and particularly the packing of the nucleosomes) to make DNA accessible during transcription, replication and DNA repair [186]. Notably, the ingredient genes of this complex accumulate also in Q2 and Q3 (Figure 32e) together with PRC2 and CBX-PRC1 genes. Hence, SWI/SWF genes regulate in concert with the main regulatory modes differentiating BL and DLBCL and partly also lymphomas and healthy B-cells. Most of the SWI/SWF compounds are highly expressed in BL and weakly expressed in DLBCL, which also supports our view of extended remodeling from open euchromatin to closed heterochromatin between both lymphoma subtypes (see section 4.2.5). Possibly euchromatin is maintained in BL by high activity of SWI/SWF-compounds and low activity of CBX-PRC1-compounds (see above). One of the genes coding SWI/SWF-compounds, *SMARCA4*, is frequently mutated in BL [4,137]. Our data reveals a strong overexpression of *SMARCA4* in this subtype (Figure 32e). Targets inhibited by *SMARCA4* [119] are found in Q4 to be down-regulated in BL (section 4.2.6).

4.3.5 DEREGULATION OF EPIGENETIC MODIFIERS GOVERNS HETEROGENEITY OF LYMPHOMAS

Dysregulation of epigenetic writer-eraser equilibria diminish plasticity of B-cells during maturation

In Figure 28 we discussed different scenarios of oncogenic perturbations in terms of a simplified scheme of epigenetic regulation. The systematic analysis of transcriptional activities of epigenetic modifiers presented in the previous subsections now enables us to compare the expected with the observed changes (Figure 33a). Compared with B-cells, the equilibria of histone methylation reactions shifts in direction of methylated H3K9 and H3K27 and demethylated H3K4 if one uses the expression data as a proxy for enzyme activities. These shifts suggest the increase of repressed and the decrease of active promoters in lymphomas accompanied by DNA hypermethylation, *i.e.*, similar alterations as expected for *EZH2* and *MLL2* mutations (compare with the scenario in Figure 28b). Hence, the latter mutations and the expression changes of the enzymes suggest similar effects on DNA methylation and gene activities.

The diversification of lymphoma data into different subtypes and healthy controls enables a refined view, for example, on the changes of enzyme expression between different stages of B-cell development in the GC. For some of the enzymes (e.g., *KMT6/EZH2* and *KDM6B/JMJD3*) one finds similar expression levels in GCB-cells and in part of the lymphoma subtypes: For example, *EZH2* is silenced in resting B-cells but massively up-regulated in GCB-cells, which undergo rapid proliferation and immunoglobulin affinity matura-

tion. *JMJD3* shows nearly the opposite trend being highly active in B-cells but nearly inactive in GCB-cells and lymphomas. A similar, although less pronounced trend is found for *KMD6A/UTX*, another relevant *K27DM* [159].

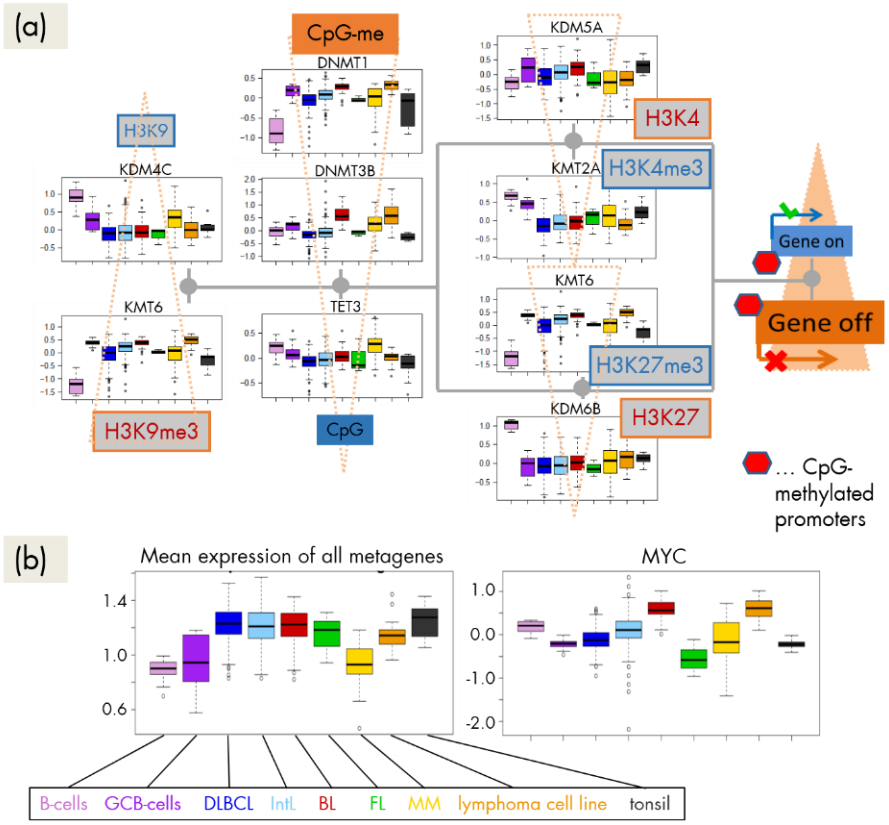


Figure 33: Dysregulation of epigenetic writer-eraser equilibria: (**a**) The expression profiles of the methyltransferases and demethylases of H3K9, H3K4, H3K27 and of DNA-CpGs suggest a shift of expression of the affected genes towards repressed and CpG-methylated promoter states. The scheme is redrawn from Figure 28 and supplemented by selected expression profiles of the respective enzymes determined from the lymphoma cohort studied. The triangles indicate the shift of the methylation-demethylation reactions in lymphomas compared with B-cells deduced from the expression profiles from the respective enzymes. (**b**) Expression profiles of *MYC* and of the mean total expression averaged over all SOM-metagenes of all samples. Total expression is consistently activated in lymphomas except for MM compared with B- and GCB-cells, whereas *MYC* is on high level in BL and IntL carrying genetic activating *MYC* defects.

These results reflect alterations of cellular programs during lymphocyte development, which are accompanied or even governed by epigenetic mechanisms. H3K27 trimethylation of PcG-targets in healthy B-cells leads to cell reprogramming [187], e.g., to transform NB-cells into highly proliferative GCB-cells. This reprogramming potentially also includes H3K4 methylation, which in concert with H3K27 methylation forms bivalent promoter domains. These combined histone marks are required to poise genes for activation or deactivation in response to developmental and differentiation indications [188]. The resolution of the bivalent domains is mediated by K4DMs and K27DMs. Mutations of *MLL2* and *EZH2*

genes may both perturb this equilibrium. In consequence associated regulations go awry and lymphomas can emerge. Perturbations in the fine balance of GCB-cell proliferation, differentiation and antigen exposure are assumed to lock GCB-cells in an immature and proliferative state, which in collaboration with other lesions induce lymphoma.

Comparison of the transcriptional activities of the enzymes between the lymphoma subtypes reveals subtle differences, which suggest different types and degrees of disturbed equilibria. In all example profiles shown in Figure 33a one sees a monotonous increase of the mean enzyme expression from DLBCL over IntL to BL suggesting a continuous shift of the histone methylation equilibria. Previously (see section 4.2.5) we presented indications for pronounced chromatin remodeling between BL and DLBCL affecting first of all transformations between repressed, poised and active promoter states. These changes of promoter states potentially ensure alternative activation of proliferative (in BL and partly IntL), inflammatory, and developmental (in DLBCL and partly FL and IntL) expression programs and they are accompanied by aberrant DNA methylation in the promoter regions of the affected genes. Note that H3K4 and especially H3K27 methylation can tune not only developmental and 'stemness' genes but also inflammatory processes needed to respond to external stimuli [189]. These different types of genes have in common that their function requires a high degree of plasticity for cell fate decisions. These decisions should induce different kinds of functional differentiation including maturation stages of the cells, their proliferative and metabolic activity and also the ability for adequate immune response.

Activation of gene expression and of TCA metabolism in lymphomas associates with epigenetics

Bivalent and repressed promoters are prerequisites for the plasticity of the B-cells required during their maturation in the GC. These genes can serve as hubs in TF networks that switch whole cascades of downstream genes either as suppressors, activators and/or enhancers of their transcriptional activity. In consequence, suppression of anti-proliferative programs and/or activation of inflammatory processes is assumed to govern molecular mechanisms in lymphomas with respect to these functionalities. Increased proliferation requires up-regulation of the molecular machineries required for transcription and translation. In addition it needs activation of the metabolism delivering the energy needed for these processes as indeed observed in BL and IntL [91].

To judge this overall balance we calculated the mean total expression level of each sample using the metagene expression data obtained in our SOM analysis. The mean sample expression clearly reveals a bimodal distribution with high expression levels in GC-derived lymphomas on one hand and with low expression levels in healthy B- and GCB-cells and in MM sharing similar expression signatures with B-cells (Figure 33b). This result clearly supports the view that malignant transformations from B- and/or GCB-cells into GC-derived lymphomas are paralleled by the massive upregulation of the transcriptional activity in the cells. Interestingly, the profile of total expression anti-correlates with the expression profiles of genes located in spot 'I' and particularly with that of *KDM4C* shown in

Figure 34a. This result suggests that total gene expression in lymphomas and B-cells is related to the TCA-energy metabolic activity, which in turn couples with the expression of epigenetic modifiers and particularly with *KDM4C* demethylating H3K9me3. The low level of *KDM4C* in lymphomas (except for MM) promotes trimethylation of H3K9 and recruitment of DNMTs, which are on high level in lymphomas (see *DNMT1* in Figure 31b). In final consequence, one expects increased CpG-methylation in agreement with the scheme in Figure 33.

On the other hand, our data indicates subtle differences of the expression of a series of genes between B- and GCB-cells. Particularly, GCB-cells show higher total expression (Figure 33b) and higher activity of KEGG-TCA- and NADPH-related genes compared with B-cells (Figure 34c). This difference is possibly governed by a shift of the H3K9-methylation equilibrium, which suggests also changes in DNA promoter methylation (Figure 33a) of genes affecting the energy metabolism. We indeed identified differential methylation patterns between B- and GCB-cells, where increased methylation is found for PRC2-targets and repressed bivalent chromatin states in GCB-cells (see section 4.2.5). This result suggests that chromatin remodeling in the GC switches the state of metabolic activity between GCB and B-cells.

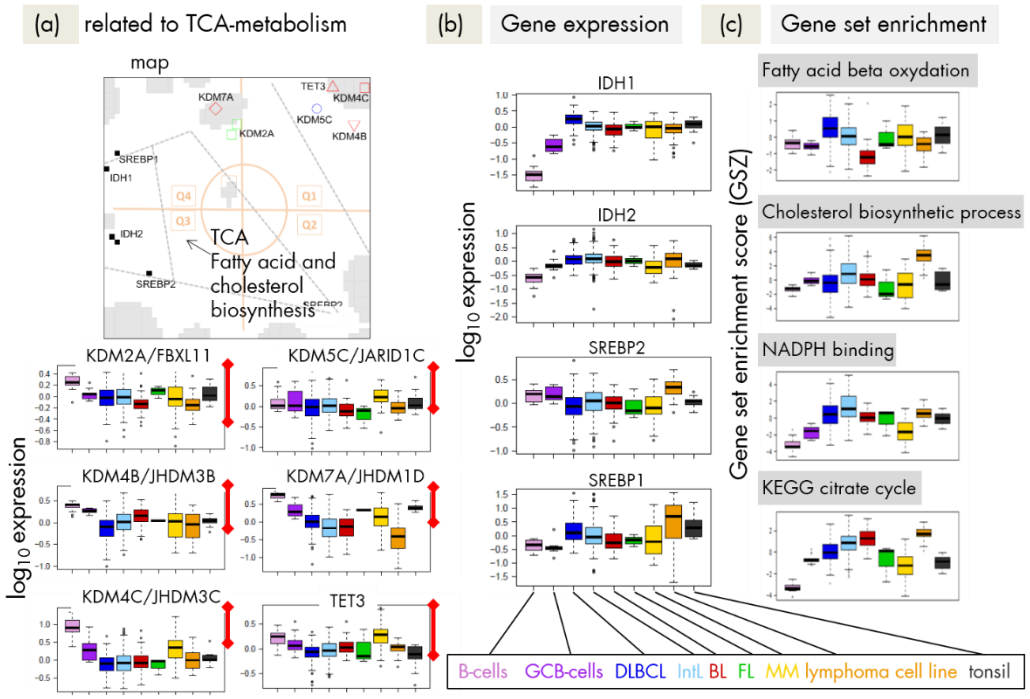


Figure 34: TCA-cycle-related epigenetic compounds: (**a**) Overview map and profiles of histone JmjC- and DNA TET-demethylases; (**b**) Expression profiles of genes coding for TCA-related enzymes and (**c**) Gene set enrichment profiles of GO-gene sets related to TCA.

We considered also another possible mechanism of global activation of transcription. Particularly, *MYC* can act as a universal amplifier of gene expression by hyper-activating genes via transcriptional pause release [190,191]. In consequence genes once activated by other mechanisms can be expected to become hyper-activated due to aberrant

MYC-overexpression. The expression profile of *MYC* (Figure 31b) however considerably differs from that of global expression. *MYC* is on high level in BL and to a less degree also in IntL compared with DLBCL and FL, mainly due to genetic defects amplifying the *MYC* gene in BL and part of IntL. TCA metabolic activity better associates with the global transcriptional level in GC-derived lymphomas, suggesting mutual relations and possible consequences for epigenetics as discussed above.

Asymmetric activation of methyl-writers and -erasers

Our study clearly shows that the expression of nearly all enzymes considered alters markedly between the lymphoma subtypes. For a holistic view we make use of the fact that SOM cartography maps the genes in an organized way. The structure of the map provides information about the underlying regulatory net because the arrangement of spots reflects their mutual co-variance structure (Figure 29c). We assigned the location of the epigenetic modifiers in the map to the respective spots (see Table 3), and with a more coarse resolution to the quadrants Q1 to Q4. Interestingly, we found strong depletion of epigenetic modifiers in Q4 opposed by their enrichment in Q2 and particularly also in Q1 and Q3 (Figure 30). In the next step we rearranged the network of expression modules to better resolve its covariance structure (Figure 35). It clearly reveals a 'backbone' of mutually correlated modules, which sequentially connects spots from Q1 to Q3. A second backbone is formed by correlated spots mostly located in Q4 and partly in Z (spot 'MM') and Q3 ('IM'). It forms an almost separated entity connected via anti-correlated edges (in red) from the first, main backbone. The spots and thus also the respective quadrants contain co-regulated genes specifically up-regulated in different subtypes as indicated in Figure 35. Importantly, almost each of the regulatory modes also affects a group of epigenetic modifiers. In other words, (de)regulation of epigenetics covers the whole transcriptional landscape of lymphoma. Moreover, the epigenetic modifiers enrich within the spot clusters when compared with the total number of genes in the spots (Fishers exact test: $p = 3.7 \cdot 10^{-6}$). Hence, epigenetic modifiers are affected by (de-)regulatory effects with higher probability than expected by chance.

The network can be decomposed into a subnet, which mainly refers to genes that antagonistically switch between BL on one hand and DLBCL/FL on the other hand. Interestingly, this subnet accumulates methyltransferases in Q3 that tend to repress gene expression of their target genes leading to antagonistic expression profiles in Q4 (the detailed assignment of enzymes to each of the spots is given in supplementary material of [192]). The imbalance between the gene expression of methyltransferases and demethylases between Q3 and Q4 can be rationalized partly by the requirement of maintenance methylation of DNA-CpG and histone methylation marks after cell division and DNA replication, which requires high activities of methyltransferases. The question whether upregulation of KMT and DNDM expression in the highly proliferative subtypes BL and partly IntL ensures maintenance of DNA and histone methylation patterns or whether it leads to progressive loss of methylation requires further studies.

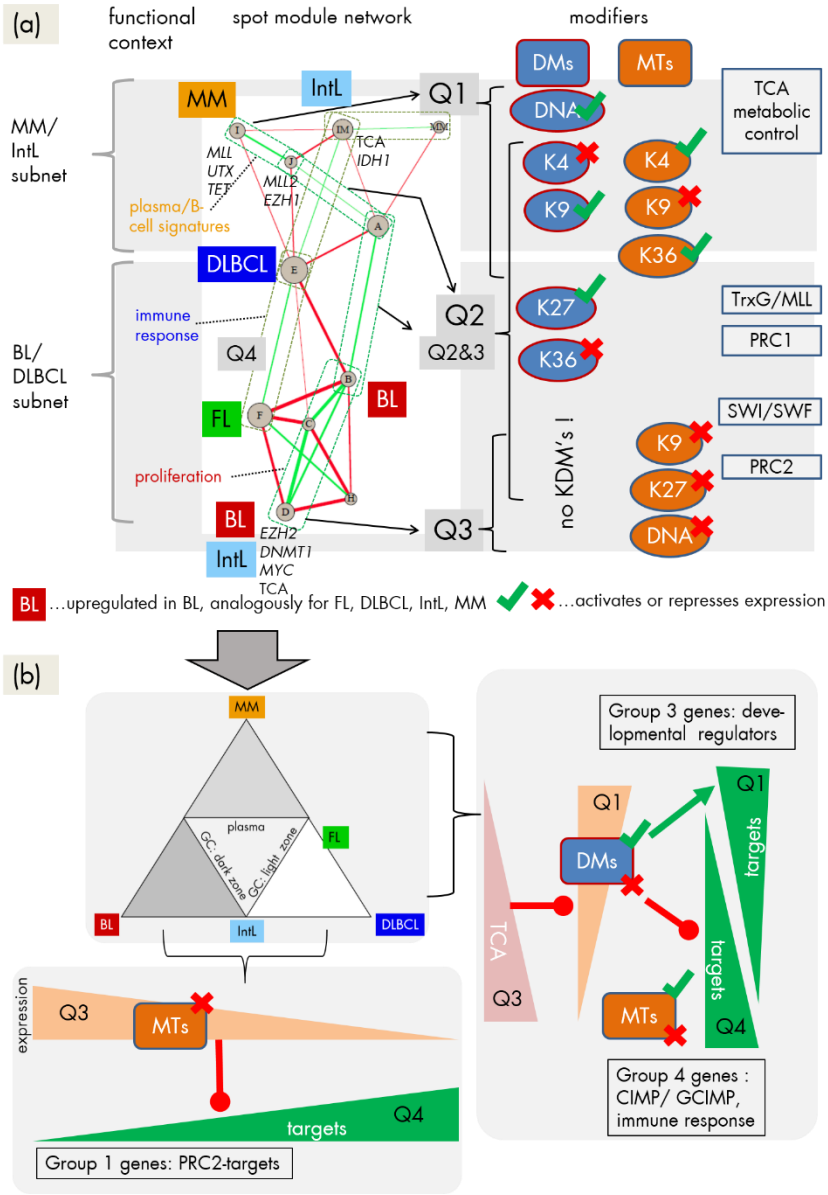


Figure 35: Asymmetric activation of methyl-writers and -erasers: (**a**) Network of expression modules governing lymphoma heterogeneity. The nodes refer to the spot clusters extracted from the SOM analysis (for functional assignments see section 4.2.1.3). Green and red edges indicate positive and negative correlations with $|w| > 0.3$ using the weighted topological overlap correlation measure (see [69] for details concerning weighted topological overlap). A 'backbone' of correlated modules can be assigned to the Q1–Q3 quadrants of the SOM used to map enzyme activities. Accumulation of different types of modifiers and complexes in Q1–Q3 is shown in the right part. Q4 is almost depleted from modifiers. A detailed list of the modifiers found in each spot is given in supplementary material of [192]. (**b**) The DLBCL/BL and MM/IntL subnets explain the expression changes of three different groups of genes identified in section 4.2.6 (see text).

Figure 35b illustrates this antagonism between BL and DLBCL using a triangular scheme of lymphocyte development and lymphoma heterogeneity. Particularly the K27MTs in Q3 are expected to inhibit PRC2-targets, which indeed accumulate in the anti-correlated region

Q4. These genes were subsumed as group 1 genes in section 4.2.6, being hypermethylated and overexpressed in lymphomas compared with the controls. In addition, compounds of the PRC1 and SWI/SWF complexes co-regulate with these methyltransferases and their targets. This parallel suggests that the stabilization of repressed promoters and the opening/closing of chromatin are mechanisms that change the expression patterns between BL and DLBCL.

Another subnet in Figure 35a contains genes that switch between MM and IntL. It accumulates methyltransferases and demethylases in Q1 that repress or activate expression. Most of the demethylases are JmjC-family enzymes, which are repressed by TCA products such as fumarate and succinate as illustrated in the right panel of Figure 35b. These demethylases together with the KMTs in Q1 then either repress or activate expression of their targets giving rise to group 3 and group 4 genes as genes that antagonistically change their expression between Q1 and Q4 (see section 4.2.6). These gene groups are enriched in 'developmental regulators', genes related to 'immune response', 'PRC2-targets' and also 'CIMP/GCIMP' genes hypermethylated in colon and brain cancer, respectively.

Both subnets overlap in Q2 collecting most of the epigenetic modifiers including activating and repressing ones without clear preference (Figure 35a). This overlap region contains modules that switch expression between (GC)B-cells and lymphomas. We hypothesize that the underlying modes regulate transcriptional programs differentiating between healthy B- and GCB-cells. Other enzymes localize near spots 'H' and 'MM' referring to early and late stages of B-cell maturation, respectively [114].

In summary, network analysis of the epigenetic modifiers identifies two subnets related to differential expression between BL and DLBCL, and between MM and IntL subtypes, respectively. The former subnet is governed by methyltransferases upregulated in BL and repressing transcription of their target genes. The latter one contains demethylases, which are presumably under metabolic control and which can activate and/or repress their targets.

4.3.6 CONCLUSION

Lymphomas show a very diverse pattern of transcriptional activity of histone and DNA methylating and demethylating enzymes and of associated reader complexes. Basic epigenetic functions in healthy B-cells seem to ensure a high level of plasticity for cell fate decisions between biological functions, such as proliferation, immune response and differentiation that sequentially switch on and off during B-cell maturation in the GC. Repressed and poised promoter states of key regulatory genes seem to play a pivotal role in this process. The fine balance between histone modifications activating or repressing transcription is governed by methylation equilibria of lysine histone side chains and of DNA CpGs.

In lymphomas this balance becomes disturbed in a subtype-specific fashion leading to deregulations of functional programs, which, in final consequence, induce lymphomagenesis. Driver mutations directly affecting epigenetic modifiers, such as *EZH2*

and *MLL2*, represent one type of initial events causing malignant transformation in lymphocytes [142]. Another option can be seen in the massive upregulation of the energy metabolism in the cell and metabolic coupling with epigenetics, where metabolites act as cofactors of JmjC-type demethylases. Finally, also indirect effects, e.g., if disturbed gene regulations affect epigenetic modifiers with downstream consequences for the epigenome are possible.

The main result of our systematic study is the finding that the expression levels of nearly all 50 enzymes studied markedly change between the sample-classes considered. Lymphoma biology apparently associates with deregulation of large parts of the epigenetic machinery of the cell. Preliminary results on enzymes affecting histone marks other than methyl groups, such as acetyl groups, support this view. Hence, understanding of epigenetic deregulation in lymphoma must go beyond simple schemes using only a few modes of regulation. We showed that the systematic 'cartography' of epigenetic modifiers onto the expression landscape of a disease using SOM machine learning as the basic technique enables a holistic view on the heterogeneity of (de-) regulation by epigenetic modifiers. A comprehensive, data driven network analysis provided indications that (de-) regulation of epigenetic enzymes is associated with virtually all modes of transcriptional regulation identified in lymphomas.

On the other hand, our network analysis showed that BL and DLBCL differ by the imbalance of repressive and poised promoters, which is governed first of all by methyltransferases and to a less degree by demethylases. The underrepresentation of demethylases in this regulation has the interesting consequence that in DLBCL only a small amount of modifying enzymes becomes upregulated, whereas BL is characterized by massive activation of modifiers.

5 Glioblastomas

Glioblastoma multiforme (GBM), the most frequent incurable brain tumor with an average survival time of approximately one year is a heterogeneous disease due to its resistance to therapeutic approaches [193]. Moreover, the molecular foundations of lower-grade gliomas (LGGs, WHO grade II and III) remain less well characterized than those of their fully malignant grade IV GBM counterpart.

Based on gene expression data derived from 200 GBM patients four subclasses termed Proneural, Mesenchymal, Neural and Classical were defined [61]. The Proneural class has been associated with mutations of *PDGFRA* (platelet derived growth factor receptor α) or *IDH1* while *EGFR* (epidermal growth factor receptor) was found mutated in Classical and *NF1* in Mesenchymal subtype [194]. Later on the Proneural cases have been further divided into GCIMP (glioma-CpG island methylator phenotype)-positive and -negative depending on their *IDH1* mutation status, with the former ones showing better outcomes.

Also for the genesis of GBM corrupt epigenetic regulation plays a crucial role. For instance promoters of *CDKN2A*, *RB1*, *PTEN*, *TP53* and *MGMT* have been reported to show hypermethylation in GBM cases. Regarding DNA methylation data Sturm et al. [62] defined 5 GBM classes, that showed a strong association with age of the patients: In young patients mutations of *H3F3A* (coding for histone H3.3) was reported, more precisely those mutations, which lead to amino acid substitutions at K27 (median age 10.5 years) or G34 (18 years) were termed K27 or G34 subclass, respectively. They are age-wise followed by RTKI 'PDGFRA' (36 years), IDH (40 years) and RTKII 'Classic' (58 years) subtypes. The names for RTKI 'PDGFRA' and RTKII 'Classic' classes were chosen according to amplification of *PDGFRA* and *EGFR* (compare expression subgroups), respectively.

This section is based on the following 3 scientific publications:

Hopp, L., Wirth, H., Fasold, M., & Binder, H. (2013). Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. *Systems Biomedicine*, 1(2).

Hopp, L., Willscher, E., Löffler-Wirth, H., & Binder, H. (2015). Function Shapes Content: DNA-Methylation Marker Genes and their Impact for Molecular Mechanisms of Glioma. *Journal of Cancer Research Updates*, 4(4), 127-148.

Hopp, L., Löffler-Wirth, H. , Galle, J. , & Binder, H. (2017). Combined portrayal of gene expression and DNA methylation landscapes disentangles modes of epigenetic regulation in glioma. In Preparation.

5.1 GENE EXPRESSION LANDSCAPE OF GLIOBLASTOMAS

We apply the SOM pipeline to gene expression profiles of a glioblastoma multiforme cohort in order to characterize the specifics of the genome wide expression landscapes in different subtypes and functionally interpret those using SOM portraying, similarity analysis and enrichment techniques. We are furthermore interested to detect possible outliers. As given in Verhaak et al. [61], the 164 samples were assigned to Mesenchymal (MES), Proneural (PN), Neural (NL), Classical (CL) GBM-subtypes and to normal healthy brain (NOR) for comparison. For details concerning the cohort and preprocessing of the data see section 3.1 and supplement section 7.1.5.

5.1.1 SOM PORTRAITS

In Figure 36 a gallery of exemplary expression portraits of glioblastoma multiforme tumors is shown. The expression portraits in log*FC*-scale reveal a handful of over- and underexpression spots, which selectively characterize different cancer subtypes such as MES, PN, NL and CL subtypes of GBM. One observes either relative stable and consistent spot-patterns (e.g. for MES- and PN-samples) or relatively heterogeneous and volatile patterns (e.g. for the CL- and NL-samples).

Mean SOM-portraits of each class were calculated, amplifying consistent class-specific features. For example, the MES-GBM subtype and normal brain tissue are characterized by two spots in opposite corners of the map, one of which being overexpressed and the other one being underexpressed in MES- and *vice versa* in NOR-samples. These class-specific spots collect highly populated, variable and resolved metagenes (see supplementary material of [69]). The mean portraits of the other three GBM-subtypes are more diffuse: The PN-, CL-, and NL-subtypes are characterized by two or three specific spots per subtype.

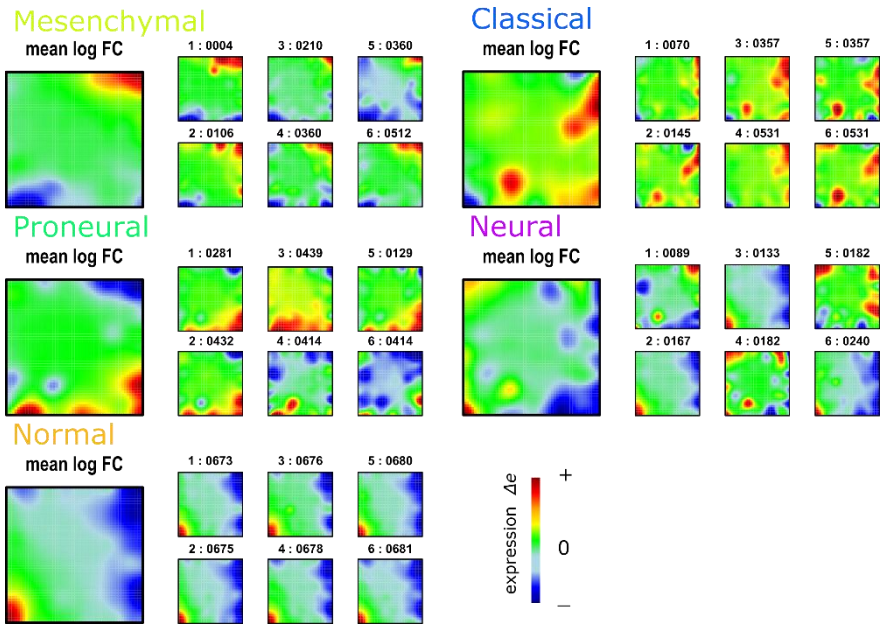


Figure 36: SOM gallery of glioblastoma multiforme subtypes: The small mosaic images refer to selected individual brain samples assigned to four GBM-subtypes and healthy brain tissue (normal). The large images represent the subtype mean SOM portraits. A complete gallery of all sample portraits is available in supplementary material of [69].

5.1.2 SAMPLE DIVERSITY ANALYSIS SUPPORTS FOUR-SUBTYPE CLASSIFICATION

In the 2nd level SOM the samples of the four GBM subtypes accumulate in different, well separated regions of the map (Figure 37a and b). This result supports the specification of subtypes taken from ref. [61]. The ten normal brain tissue samples occupy a very narrow area in the top right corner of the map. Their portraits most closely resemble that of the NL subtype: Both mean portraits show a common overexpression spot in the bottom left corner, which is not present in the mean portraits of the other GBM-subtypes.

As a complementary method, ICA was applied to the SOM portraits of all samples of each cancer subtype. The three dimensional and two dimensional ICA-plots of the GBM study are shown in Figure 37c and d, respectively. The samples are similarly distributed in ICA-space as in the 2nd level SOM. Additional information can be extracted from the distribution of the cancer subtypes along the independent component axes IC1, IC2, and IC3. The GBM-subtypes mainly arrange in the IC1/IC2-plane whereas the NOR-reference samples are separated away from most of the cancer samples in direction of IC3 axis. The MES- and PN-subtypes systematically differ in their IC2-coordinate whereas distinction of NL- and CL-subtypes in their IC1-coordinate can be observed. Hence, the former and the latter two subtypes are obviously characterized by two sets of genes, which change independently between the two pairwise combinations of subtypes. These subtype-specific features, in turn, are mostly independent of the features, which differentiate between cancer and normal samples along IC3. Also part of the NL-samples varies in direction of this component reflecting the partly similar expression pattern of NL- and NOR-samples.

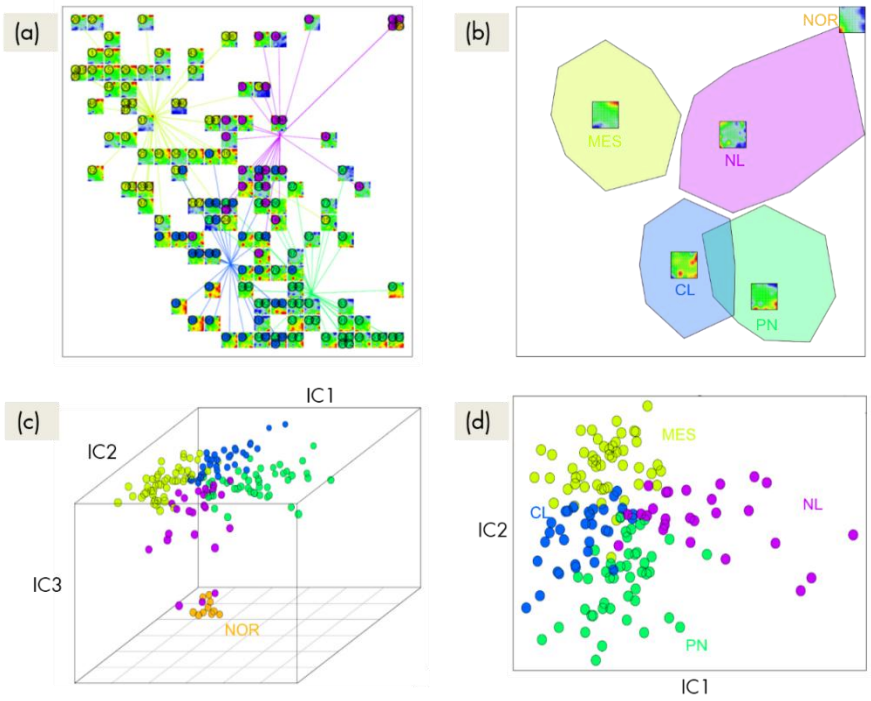


Figure 37: 2nd level SOM and ICA similarity analysis of GBM cancer subtypes: (**a**) The position of each GBM-sample is marked by the respective 1st level SOM image. Samples of the same GBM-subtype are connected by lines drawn to the centroid of the respective class. (**b**) shows essentially the same 2nd level SOM as in panel a. The mean regions occupied by the samples of each of the four subtypes are illustrated by the colored polygons. The mean SOM portraits of each GBM-subtype are located in the center of the respective polygon. The four GBM-subtypes occupy roughly the four quadrants of the map whereas the 10 normal tissue samples aggregate into one tile in the top-right corner of the map. (**c**) The three-dimensional distribution of samples is shown in the space spanned by the three leading independent components IC1 – IC3. (**d**) The projection of the GBM-subtypes into the IC1/IC2-plane.

The MST- and especially the CN-plots of the GBM-subtypes (Figure 38a and b) reveal similarities between the subtypes, which are less evident in the 2nd level SOM: For example, the NL and PN subtypes share more similarities with the NOR-reference samples than the CL- and MES-subtypes, which on the other hand, are relatively similar. Note also that the PN- and MES-samples accumulate within compact clusters whereas the CL- and NL-clusters are fuzzier. These subtypes form a continuum between the MES- and PN-forms, which distribute along two separate branches. The CN forms a 'donut-like' structure composed of alternating compact and fuzzy clusters. The latter ones refer to intermediate NL- and CL-subtypes 'linking' the MES- and PN-subtypes in the compact clusters.

The more localized MES- and PN-subtypes in the phylogenetic cluster tree visualization (see Figure 38c) tend to aggregate into separate branches whereas the intermediate NL- and CL-subtypes again occupy diffuse branches in between. The dendrogram of GBM reveals that the NL-samples group along a separate branch together with the NOR-samples.

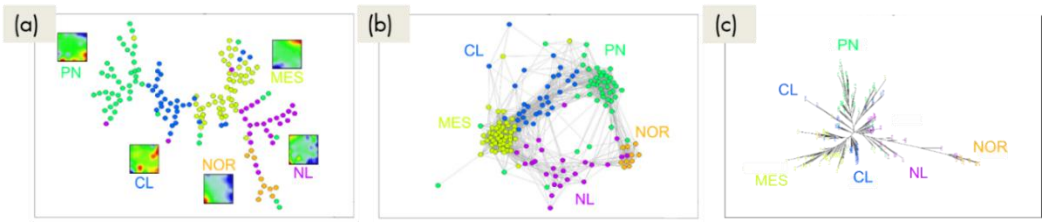


Figure 38: Similarity analysis of GBM: (**a**) The MST is shown in the left part together with mean SOM portraits of each subtype. (**b**) The middle and (**c**) right part show the CN and phylogenetic cluster tree, respectively.

5.1.3 CLUSTERS OF CO-EXPRESSED GENES

Recall, that the most prominent features are the over- and underexpression spots formed by neighbored metagenes of similar profiles, which in turn represent clusters of correlated and thus potentially co-regulated genes strongly over- and/or underexpressed in a subset of samples. Alternative options of module selections are discussed in supplementary material of [69].

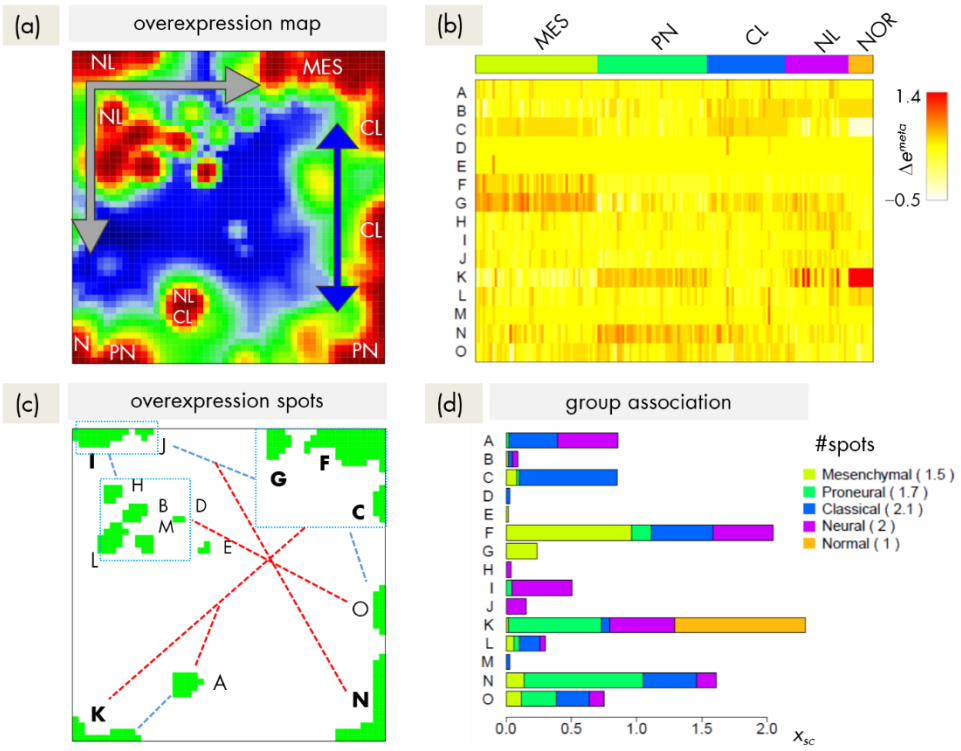


Figure 39: Overexpression spot characteristics of GBM: (**a**) The overexpression summary map collects all overexpression spots observed in the individual profiles into one map (see also section 3.6 for details). Classes showing the respective spots are indicated in the map. (**b**) The overexpression spot heatmap. (**c**) The overexpression spot map. (**d**) The spot-abundance bar plot shows the fraction of samples of each subtype, which exhibit a given spot. For details see caption of Figure 7.

In the next step we analyze these spot patterns to identify differences and common properties shared between the cancer subtypes. Such more unique or more ubiquitous spots are potential candidates for extracting the functional impact of the specifics of gene activity in each of the cancer subtypes. Figure 39a shows the overexpression summary map of GBM, which collects all overexpression spots observed in the individual GBM portraits into one master map (see also [23]). In total, we identified 15 overexpression spots 'A' to 'O' in GBM, being labeled using capital letters (Figure 39c). The spot-expression heatmap in Figure 39b provides an overview of the subtype-specific expression in each of the spot clusters. For example, spots 'G' and partly also spot 'F' are selectively overexpressed in samples of the MES-subtype and spot 'I' in the NL-subtype whereas spots 'M' and 'O' are more ubiquitous lacking subtype-specific overexpression (see also spot-abundance bar plot in Figure 39d).

5.1.4 FUNCTION MINING: AGAIN INFLAMMATION-VERSUS-PROLIFERATION

We applied gene set overrepresentation analysis (see methods section 3.7) and assigned a short notation to each of the GBM-spots (see Figure 40a). Selected spots of GBM are obviously related to processes generally associated with cancer physiology such as 'inflammation' (spot 'F') and 'cell division' (spot 'N'). Panels c and d of Figure 40 depict the GSZ-expression profiles and the population maps of the gene sets 'inflammatory response' and 'cell division' for GBM data. The profiles clearly reflect the fact that the respective processes are selectively activated and de-activated in a subtype-specific fashion: 'Inflammatory response' in the MES and 'cell division' in the PN subtype. The respective gene set population maps reveal that the associated genes accumulate in the regions of spots overexpressed in the maps of the different subtypes (compare with Figure 41).

Figure 40b shows the heatmap for gene sets referring to the GO-term 'biological process' enriched in spots of the GBM-samples. The one-way clustering separates the gene sets in agreement with their spot associations: For example, spot 'F' mainly collects gene sets overexpressed in the MES- and also the NL-subtype whereas the adjacent spot 'G' contains gene sets overexpressed in the MES- and CL- subtype. The heatmap also shows that gene sets from spot 'K' tend to be overexpressed in healthy brain samples and the NL- and PN-subtypes as well. It further assigns gene sets overexpressed in the PN-subtype to spot 'N'.

In addition to one-way clustering heatmaps, we also performed two-way clustering of gene sets and samples to detect inconsistencies in the class labeling of the samples. The resulting heatmaps for the literature gene sets (GSEA2) reveal that the cancer-related gene sets essentially form two clusters with strong enrichment in spots highly overexpressed in the MES subtype (spot 'F' and 'G') and the PN subtype (spot 'N'), respectively. This seems to reflect common gene activation patterns present in different tumors associated with either 'inflammation' (for MES) or 'cell division' (for PN). See supplementary material of [69] for supporting results and complementary analyses.

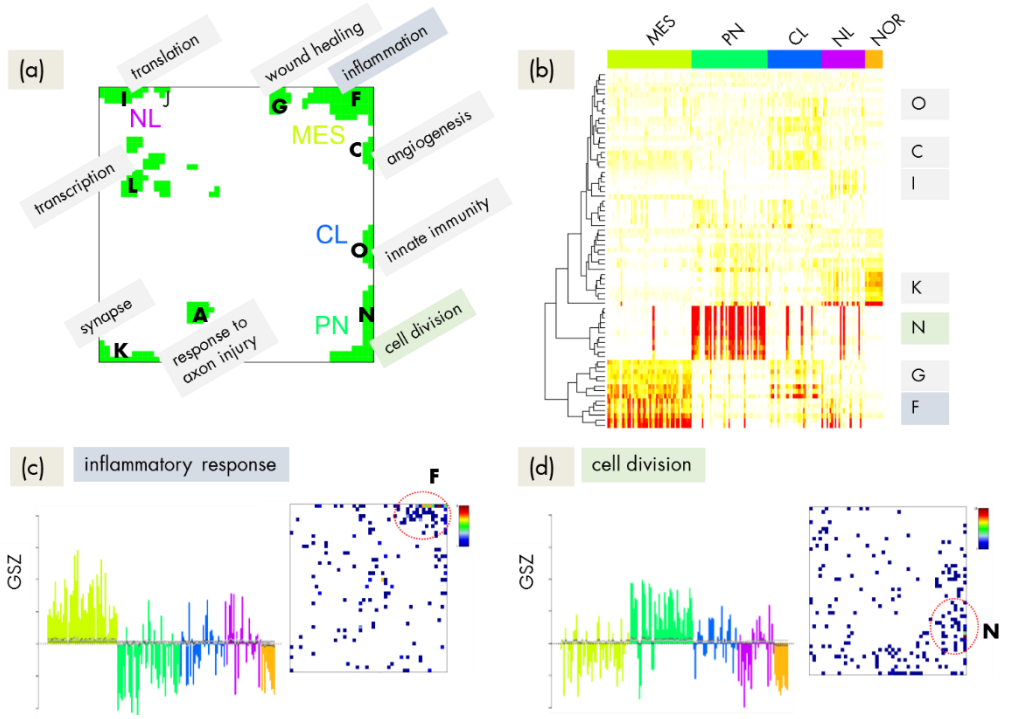


Figure 40: Gene set enrichment analysis of GBM: **(a)** The spot summary map. **(b)** The overrepresentation heatmap of gene sets referring to the GO-term biological process. **(c)** and **(d)** overexpression profile and map of the gene sets 'inflammatory response' and 'cell division', respectively. See legend of Figure 8 for details.

5.1.5 SUBTYPE-SPECIFIC DIFFERENTIAL EXPRESSION DISENTANGLES HETEROGENEITY OF TRANSCRIPTIONAL PROGRAMS

Most of the spots observed in the studied cancer portraits appear in multiple cancer subtypes with non-negligible abundance (see Figure 39d). Such non-specific spots are mostly filtered out by calculating the mean SOM-portrait of each cancer subtype. The remaining overexpression spots in the mean portraits are candidates for sets of genes, which are specifically overexpressed in the respective subtype.

Genes, which are specifically and significantly overexpressed in each of the four GBM subtypes have been determined independently using SAM (significance analysis of microarrays, [195]) where each subtype was compared to the other three subtypes. We treated the obtained signature genes as 'gene sets' and calculated their GSZ profiles in the samples studied (leftmost panel in Figure 41). The profiles in confirm the fact that, indeed, each signature set is specifically overexpressed in the respective subtype and underexpressed in the remaining three subtypes of GBM. The NL-specific signature shows overexpression also in the healthy brain tissue, which was not taken into account extracting specific signature genes in [61].

Figure 41 also shows the overrepresentation population map for each of the signature sets. They clearly reveal that genes of each of the sets accumulate in the spots of subtype-

specific overexpression, which were identified using the mean maps, and with higher specificity, the difference maps (for details see section 3.8).

Hence, the mean and difference potraits allow to extract specific signature spots for each cancer subtype. Spot-specific significance analysis then allows extracting lists of signature genes from each of the spots [195]. Interestingly, the signature genes of the PN- and CL-subtypes distribute over more than one, mostly well separated overexpression spot. They obviously belong to different functional modules of co-expressed genes.

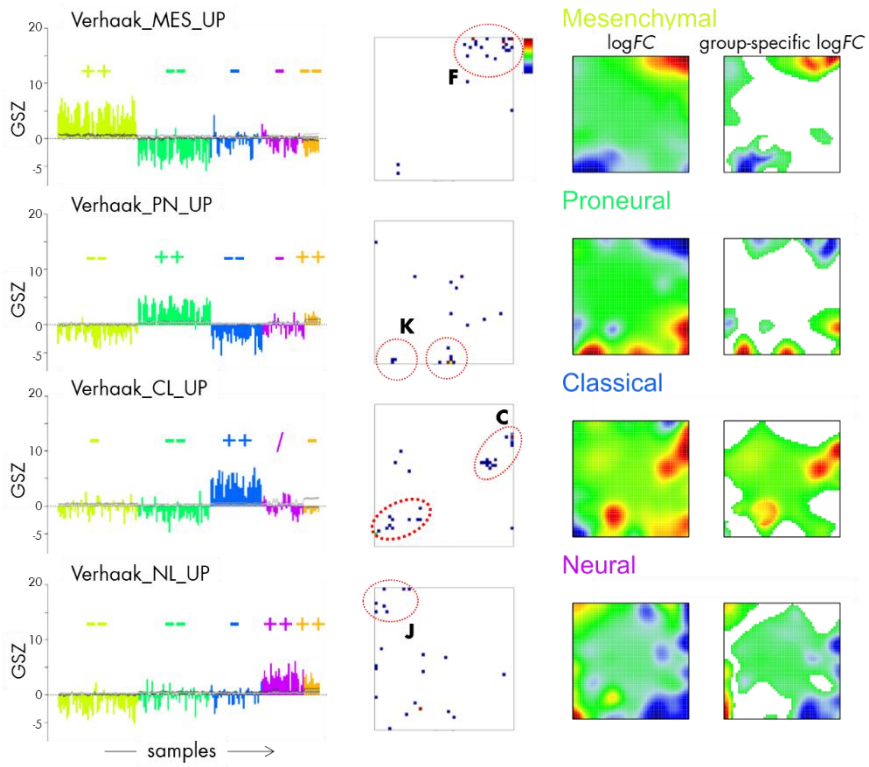


Figure 41: Subtype-specific genes of GBM. From the left to the right: (i) GSZ-expression profiles of the subtype-specific gene sets taken from [61]; (ii) the respective gene set population maps; (iii) subtype-related mean and (iv) difference portraits. For details see legend of Figure 9.

5.1.6 MAPPING GLOBAL TWO-GROUP DIFFERENTIAL EXPRESSION MASKS DETAILS OF GLIOMA TRANSCRIPTION

A recent study identified 1,236 significantly differentially expressed genes in a GBM-versus-normal brain difference analysis of TCGA samples without special emphasis on GBM subtypes [196]. We extracted three sets from the ranked list of these genes (425 strongly and 426 less strongly downregulated genes and 376 upregulated ones). They accumulate essentially in spots 'K' (downregulated genes) and 'C' and 'N' (upregulated genes) of our SOM (see the gene set population maps in Figure 42a). This spot-pattern closely agrees with the strongest over- and underexpression spots observed in the images of normal brain tissue and in the mean portraits averaged over all GBM samples studied

(compare the gene set population maps and the mean portraits of normal brain and of all GBM samples shown in Figure 42a and b, respectively). Note however that the expression levels of the mean NOR- and GBM-maps are strongly antagonistic, *i.e.* strongly upregulated spots in the NOR-portrait become strongly downregulated in the GBM-portrait and *vice versa*. This symmetry of the expression landscapes simply reflects the fact that the expression amplitudes of the NOR-samples largely exceed that of the GBM-samples. The gene sets extracted thus cover only a small part of the expression modules detected in our SOM analysis (Figure 42c). Especially spots of weak differential expression but of potentially high impact for the different GBM subtypes remain undetected. The GSZ profiles reveal that the selected genes only weakly differentiate between the MES-, PN-, and CL-subtypes (Figure 42). The profiles also show that the NL-subtype partly follows the expression pattern of normal brain.

These results illustrate the benefits of our portraying approach: It provides a detailed view on the compartmentalization of the expression landscape into different modules, which allows their separate analysis in terms of biological function. One might easily miss such details when using the simple differential disease-versus-normal approach.

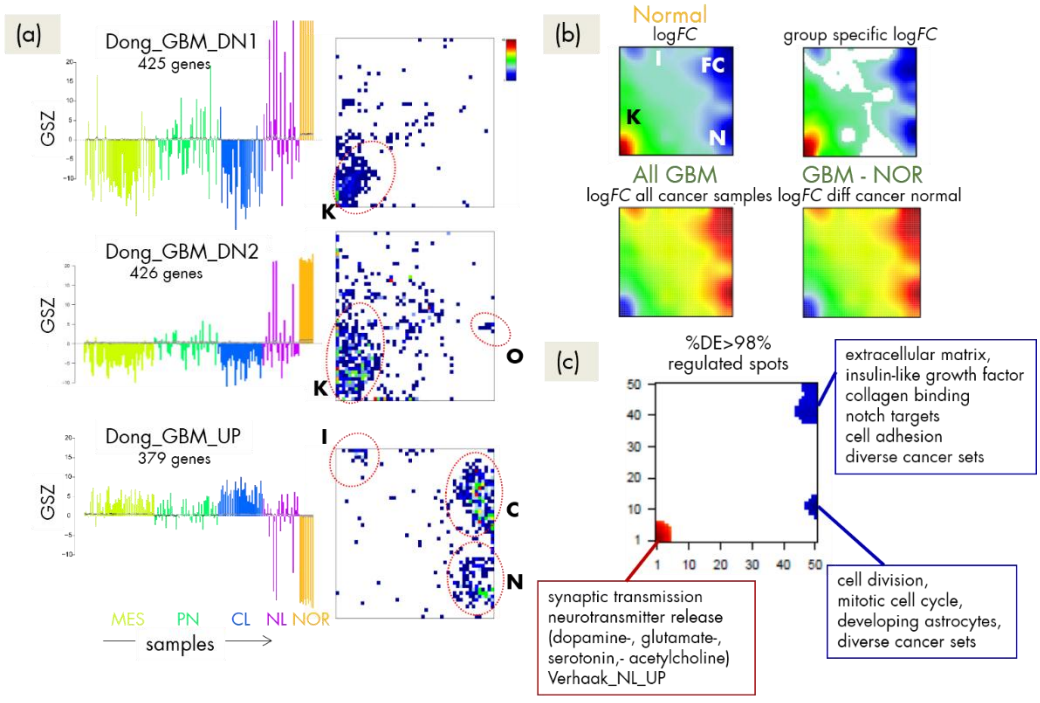


Figure 42: Differentially down- and up-regulated genes in a GBM versus normal study: **(a)** GSZ profiles and the population maps of the genes taken from Dong et al. [196] and split into three sets. **(b)** The mean log*FC* and difference maps of normal brain samples (first row), and of all GBM-samples studied (second row). **(c)** Selected spots and their functional context.

5.1.7 CATEGORIZING THE GENE SETS: GO-TERMS, CANCER-, AND CELL TYPE-RELATED GENES

Neighboring spots of strongly correlated metagene expression profiles can be assigned to related BPs: As shown in Figure 40, the 'inflammation' spot 'F' in GBM is close to spots assigned to 'wound healing' and 'angiogenesis'; the 'cell division' spot 'N' is close to spot 'O' labeled 'innate immunity', where 'stress activated signaling' was the most strongly overrepresented gene set. Note that, although related, these neighboring spots are usually characterized by subtle differences in their expression profiles and presumably also by fine differences in the functional context of the overrepresented gene sets. In Figure 43a we provide GSZ profiles and population maps of a series of gene sets selected from the GO-terms 'biological process' (BP), 'cellular component' (CC) and 'molecular function' (MF), which change in concert with 'inflammation' and 'cell division'. The population maps clearly reveal these subtle differences: For example, GSZ profiles of both 'immune response' and 'wound healing' change together with 'inflammation' and accumulate in adjacent but different regions of the maps. The population maps of 'angiogenesis' and, to a lesser degree, of 'wound healing', give rise to the overexpression of the respective GSZ profiles in CL, while underexpression was observed for other 'inflammation'-like gene sets.

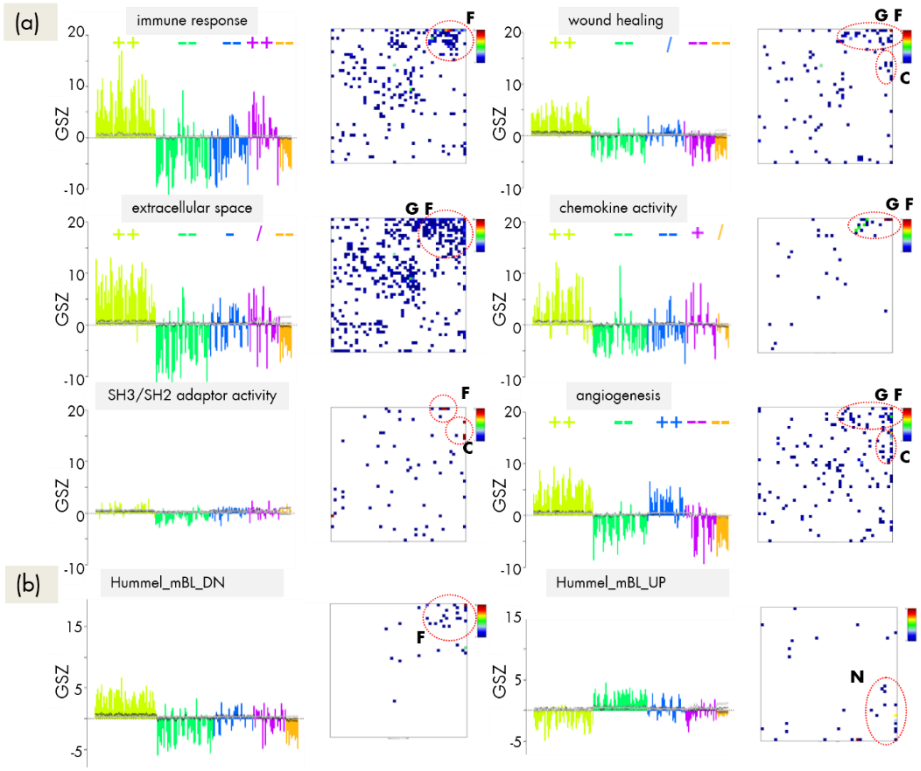


Figure 43: GSZ profiles and population maps of a series of gene sets in GBM. **(a)** Selected profiles and population maps of gene sets acting in concert with 'inflammatory response' in GBM. Regions of overrepresentation in the maps are indicated by red dotted ellipses. The letters refer to the respective overexpression spots. **(b)** Enrichment of cancer sets in GBM: Sets up- and down-regulated in Burkitt's lymphoma [59].

The results so far show that GBM splits into subtypes differing by the antagonistic activation of BP related to 'inflammation' and 'immune response' vs. processes related to 'cell division' and 'transcriptional and translational machinery' (namely MES vs. PN). We observed a similar separation of subtypes related to 'inflammation' and 'cell division' in B-cell lymphoma (BL, see section 4.1.4). In order to evaluate the degrees of similarity between both GBM and BL cancer entities in this respect, we studied the enrichment of signature gene sets derived from BL in GBM (Figure 43b): It turned out that the two signature gene sets up- and downregulated in the BL subtypes accumulated in spots 'F' and 'N' (Figure 40), which are overexpressed in the MES and the PN subtypes, respectively. This result suggests a more generic nature of the underlying processes related to 'inflammation' and 'cell division' in cancers.

Gene sets related to 'innate immunity' are found overrepresented in spot 'O' of the GBM map, which is overexpressed in the CL- and PN- and partly also in MES-subtypes. The intermediate CL- and NL-subtypes of GBM are characterized by spots 'A' and 'O', 'C' (CL-subtype) and 'A', 'I', 'J', and partly 'K' (NL-subtype, Figure 39). The latter spot 'K' is also characteristic for healthy brain tissue. It is therefore not surprising that it contains overrepresented populations of gene sets related to 'nervous processes' such as 'synaptic transmission' and 'neurotransmitter secretion' (data not shown here). The NL-subtype however differs from the healthy brain tissue mainly by the appearance of spots 'I' and 'J' (see Figure 41), which contain overrepresented gene sets related to 'translation', such as 'ribosome' and 'mitochondrion'. The CL subtype-specific spot 'C' is assigned to 'angiogenesis', and thus it reflects a common cancer process.

We further analyzed the relationship between gene sets and the cell type or tissue specificity in order to understand the biological meaning of the GBM subtypes. We collected the gene set enrichment level from the brain transcriptome database [197] as proposed before [61]. Mature cell types such as neurons, oligodendrocytes, astrocytes, and cultured astroglial cells may be of interest for their primary associations with tumor subtypes and as inherent signatures retained from progenitor cells. In agreement with earlier studies [61], we found subtype-specific enrichment of signatures: 'Oligodendrocytic' (in PN and NL), 'astrocytic' (in CL and NL), 'neuronal' (in NL), and 'cultured astroglia' (in MES and partly CL); see supplementary material of [69]. We also tested signatures for 'developing astrocytes' (enriched in PN and partly NL) and 'nervous tissue' (enriched in NL and NOR).

Our SOM mapping and profiling of the different signature sets, however, provides a finer assignment to the different GBM subtypes: The 'oligodendrocytic', 'neuronal', and 'nervous system' genes accumulate preferentially in spot 'K', which is overexpressed in normal brain tissue. Its key property in GBM is the antagonistic upregulation in NL and downregulation in CL subtypes. In contrast, the 'astrocytic' signature genes accumulate in spot 'A', with the corresponding upregulation in NL and CL subtypes and downregulation in MES and PN subtypes. Hence, the co-located spots 'A' and 'K' can be associated with different regulation patterns especially in NL and CL subtypes, while these can be associated with different cell types. Interestingly, the 'astrocyte' signature strongly resembles that

of the 'aging brain_DN' set of genes, which reduce their activity in the aging cortex [198] and the GO-term 'negative regulation of cell death'. The signatures of 'nervous tissue' and 'developing astrocytes' are enriched in spots 'K' and 'O', respectively. They similarly respond in an antagonistic up-vs-down fashion in PN and MES subtypes. Note that spot 'O' was associated also with BP related to 'cell division' such as 'mitosis', 'DNA repair', and undifferentiated cancer. Finally, while the signature genes of 'cultured astroglia' also accumulate in spot 'O', they are found primarily in spots 'G' and 'F', showing upregulation in the MES subtype and antagonistic downregulation in normal brain tissue.

5.1.8 CONTAMINATIONS, OUTLIERS AND MISCLASSIFIED SAMPLES IN GBM

Possible reasons for the occurrence of contaminations are given in section 4.1.5. Inspection of the individual SOM portraits combined with similarity and gene set enrichment analyses provide a framework of hand-in-hand options to detect and to correct strongly biased samples.

In Figure 44 we re-plotted the CN similarity plot of GBM together with selected individual portraits of samples, which are located either outside of the main clusters and/or within an apparently 'false cluster'. For example, samples no. 326 and 156 originally assigned to the MES- and PN-subtypes are found within the 'wrong' area of the net near the green PN- and yellow MES-cluster, respectively. Comparison of the portrait of sample 156 (and partly 321) with the mean portraits of the MES-and PN-subtypes reveals that its expression landscape obviously represents a combination of both expression signatures where the MES-signature more heavily contributes to the mixture than the PN-signature in contradiction to the original class assignment taken from [61]. Another heterogeneous group of samples (e.g. no. 290, 152, 358) form a set of outliers near the blue CL-cluster. Inspection of the respective portraits reveals that a few overexpression spots (e.g., 'L', 'B', and 'D') are obviously responsible for this behavior: They are not observed in the majority of the remaining CL-samples. A similar argumentation applies to outlier samples no. 326, 84 and 87 showing strong expression of spot 'n1'. Note also that these groups of outliers are mostly heterogeneous, *i.e.* they contain samples assigned to different subtypes.

These 'outlier'-spots are mostly relatively rare and unspecific for one of the GBM-subtypes (see e.g. the abundance bar plot for spots 'L' and 'D' in Figure 39d). This result suggests that these features are presumably caused by contaminations of non-tumor cells or by treatment effects and thus they are not or not directly related to GBM. Gene set analysis shows, that spot 'B', for example, contains an enriched number of genes related to 'xenobiotics' and 'drug metabolism'.

Hence, misclassifications of samples can be caused by the mixing of different subtypes and also by outlier features, which are presumably not related to cancer, but which make samples of different subtypes similar. These examples demonstrate that our portraying approach not only detects potential outliers and misclassified samples but in addition helps

researchers to generate hypotheses about the origin of these effects and also to extract more detailed information from the data, for example, by applying spot-related functional analysis.

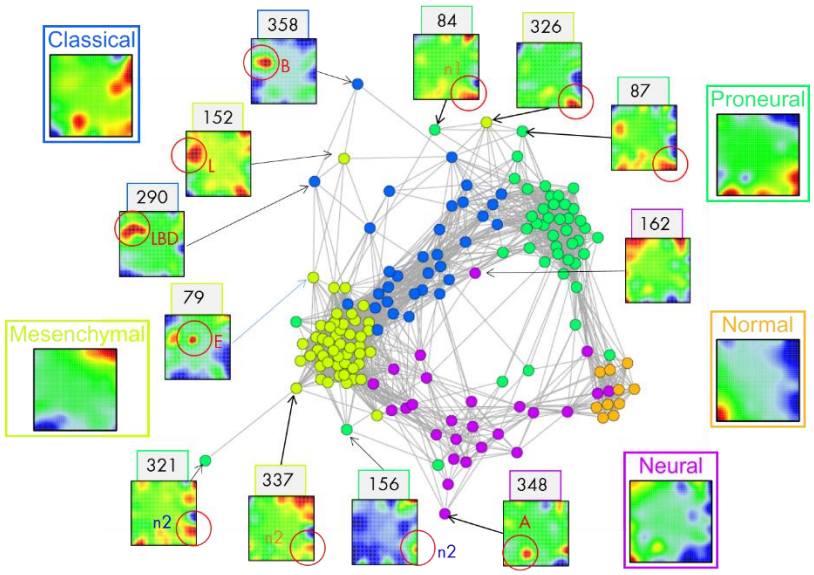


Figure 44: Outliers and misclassified samples in GBM are indicated in the CN-similarity plot by arrows together with the respective sample portraits. The subtype-averaged mean portraits are shown for comparison at the left and right margins of the figure. The red circles and the letters assign the spots causing the partly atypical properties of the samples.

5.1.9 CONCLUSION

We applied the single-omics SOM method to expression profiles of glioblastoma multiforme to characterize the specifics of the genome wide expression landscapes in different subtypes of cancer. Our method simultaneously detects features, which are differentially expressed and correlated in their profiles in the set of samples studied. Functionally related genes often merge into larger aggregates, which can then be interpreted as functional modules. Characteristic differences between subtypes can be clearly identified and further analyzed using metagene profiles representing the intrinsic correlation groups.

Summarizing, the GBM subtypes studied here can be divided into two 'localized' and two 'intermediate' ones. The localized subtypes are characterized by the antagonistic activation of processes related to immune response and cell division, commonly observed also in other cancers. In contrast, each of the intermediate subtypes forms a heterogeneous continuum of expression states linking the localized subtypes. Both 'intermediate' subtypes were characterized by distinct expression patterns related to translational activity (upregulated in NL) and innate immunity (upregulated in CL).

Our case study demonstrated that analyzing gene expression landscapes in the context of a compendium of molecular concepts is useful in understanding cancer biology.

5.2 DNA METHYLATION LANDSCAPE OF GLIOBLASTOMAS

We re-analyzed microarray DNA methylation data published in a previous study on pediatric and adult brain tumors and non-neoplastic controls [62] to get a detailed insight into the methylation landscapes of gliomas and the functional impact of sets of DNA methylation marker genes for molecular mechanisms of cancer diversity, genesis and progression.

For DNA methylation the GBM adult samples were classified into five molecular subtypes as described in Sturm et al. [62]: MES, IDH, RTKI and RTKII. The pediatric GBM split into two subtypes namely G34 and K27. Additionally fetal and adult controls were considered. Three expression data sets as studied in Sturm et al. [62], Hopp et al.[61] and Reifenberger et al. [63] were used to establish associations with methylation data. For details concerning the cohort and preprocessing see section 3.1 and supplement section 7.1.6.

5.2.1 HIGH-DIMENSIONAL DATA PORTRAYING

For SOM analysis we used either *M* values (MetSOM, Eq.(4)) or centralized *M* values (DmetSOM, Eq.(5)). DmetSOM attenuates methylation changes independent of the methylation level whereas MetSOM directly considers absolute methylation levels and thus enables to distinguish highly methylated from weakly methylated genes. MetSOM has the advantage to resolve modules of co-methylated genes in more detail with higher granularity [113]. For details see section 4.2.1.2.

5.2.1.1 ABSOLUTE METHYLATION LEVELS IN GBM AND HEALTHY BRAIN (METSOM)

In the next step, SOM data portrayal was applied to the gene-centric methylation data including all glioma samples and the non-neoplastic brain samples serving as reference. The method 'projects' the methylation data onto a two-dimensional grid of 40x40 pixels. Appropriate color-coding then visualized the methylation landscapes of each sample in terms of its individual methylation portrait (not shown). Figure 45a shows the gallery of mean portraits for all classes studied. Red and blue regions in the images refer to genes with high and low methylation levels of the probed CpG regions, respectively. Hence, the map can be segmented into regions containing genes of high and low methylation levels of their promoters and in regions containing genes with strongly variant and almost invariant methylation levels (Figure 45b). The regions of variant and of invariant genes thus include regions of high and low mean methylation levels as well.

Groups of genes with characteristic methylation profiles can be extracted from the map using a correlation metrics (Figure 45c). Accordingly, the methylation landscape divides into regions of hyper- and hypomethylated genes in almost all samples and in regions showing differential methylation effects between them as indicated in the figure. We calculated the mean methylation level and its variance separately for each subtype using the

individual methylation portraits (Figure 45d). One sees that IDH, RTKII and, to a less degree mesenchymal tumors are globally hypermethylated with respect to the controls whereas G34, K27, and RTKI are globally hypomethylated. The variance of the methylation level reflects the coarseness of the methylation landscapes of the subtypes. The decreased variance in gliomas compared with the controls reflects smoother landscapes in the tumors with more balanced methylation levels between the genes on the average.

In summary, methylation changes in gliomas comprise both, hyper- and hypomethylation in a subtype-specific fashion. SOM mapping identifies genes with different methylation levels and specific alterations of the methylation levels between the subtypes.

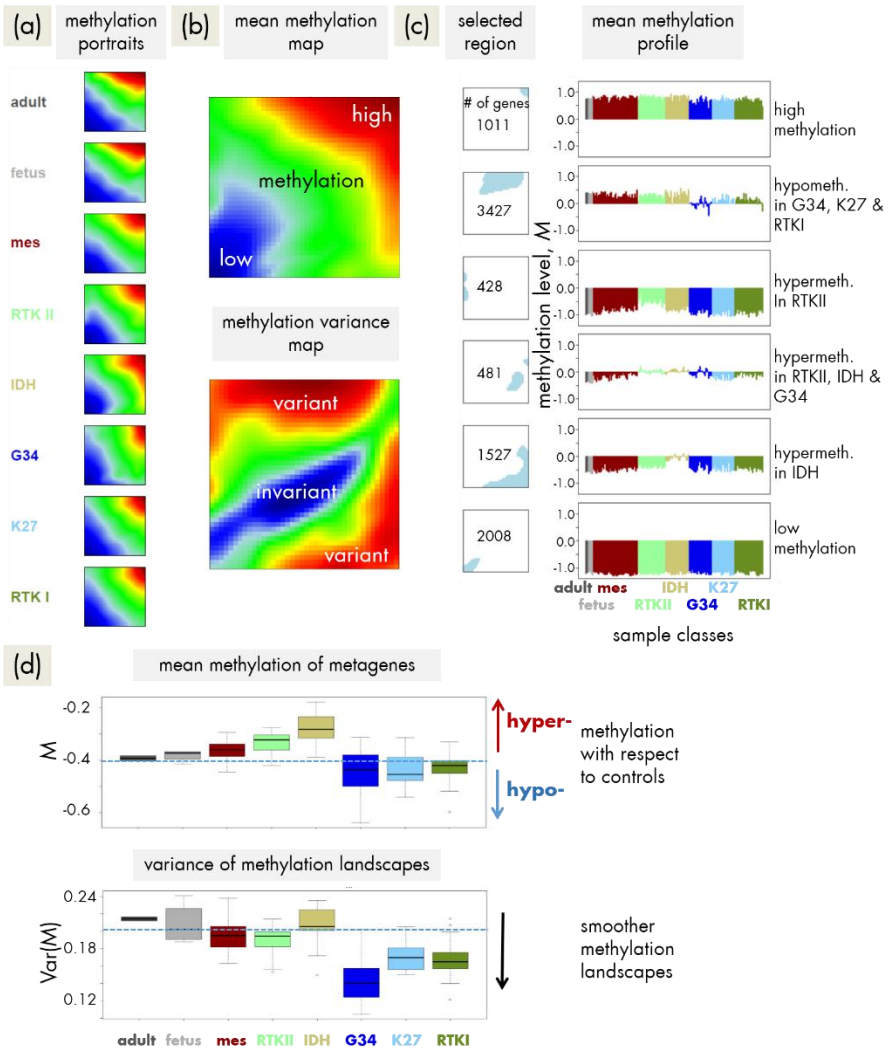


Figure 45: MetSOM portrayal of the methylation landscapes of GBM subtypes: (**a**) SOM portraits of glioma subtypes and of healthy controls. (**b**) The methylation overview map visualizes regions of high (red) and low (blue) methylation levels. The methylation variance map identifies regions of genes showing highly variable (red) and almost invariant (blue) methylation. (**c**) Selected regions of the map show different methylation profiles among the samples. (**d**) Mean methylation level and variance of the classes studied.

5.2.1.2 RELATIVE (-CENTRALIZED) METHYLATION (DMETSOM)

In the next step we trained a second SOM using centralized methylation values (DmetSOM) where the mean methylation level of each gene averaged over all samples was subtracted from its actual methylation *M* value. Centralization focuses the view on methylation changes between the samples independent of the absolute methylation level of the genes and it improves resolution with respect to differential markers that distinguish the different classes [113]. The class-averaged mean DmetSOM portraits shown in Figure 46a are clearly more diverse than the respective MetSOM portraits shown in Figure 45a. One clearly identifies similar textures of the maps of non-neoplastic brain (adult and fetal) and mesenchymal GBM and of K27 and RTKI GBM, respectively. The similarity net in Figure 46a more clearly visualizes the mutual similarities of individual methylation landscapes of the samples based on the mutual (Pearson) correlation coefficients between them, which were color-coded in the heatmap in Figure 46b. The classes can be roughly grouped into three superclusters, which we assign as 'brain-like' because of the only small and moderate methylation changes in GBM; as (hyper-) glioma CpG methylator phenotype (GCIMP) and as hypomethylator phenotype (CHOP) based on the global methylation drifts in GBM as suggested before in [62]. The brain-like and CHOP (and partly also GCIMP) groups show mainly anti-correlated methylation landscapes meaning that large groups of genes concertedly 'switch' their methylation levels between these groups (see the blue off-diagonal areas in Figure 46b).

Note that each class forms its own cloud of samples in the similarity net, which still reflects its own specifics within each of the supercluster (see below). On the other hand, one observes a certain degree of fuzziness between the subtypes. For example, the K27 and RTKI sample clouds partly overlap. In the supplementary material of [199] we provide the individual sample portraits sorted for each GBM subtype using hierarchical clustering trees. Part of the samples shows methylation landscapes, which can be interpreted as mixtures of different subtypes (e.g. of K27, RTKI, and G34) or as mixtures with healthy brain methylation characteristics (part of the mesenchymal and RTKII samples). The 'personalized' portrayal of the samples enables the detailed assignment of these mixed characteristics.

The silhouette plot in Figure 46c evaluates the robustness of class assignment for all samples. It reveals that the IDH, G34, RTKII, partly K27 and the controls form relatively robust classes whereas mesenchymal and especially RTKI are rather unambiguously assigned mainly due to overlapping characteristics with non-neoplastic fetal brain and G34 GBM, respectively (see the color bar in Figure 46c, which annotates the 'best class membership'). Note that our robustness analysis is based on gene-centric whole-genome methylation landscapes and thus it does not contradict the classification proposed in [62], which is based on the 8,000 most variant CpG probes. Our robustness analysis however illustrates the degree of fuzziness of class assignment, which reflects the mutual overlap between them and possibly also common biological factors that drive tumorigenesis.

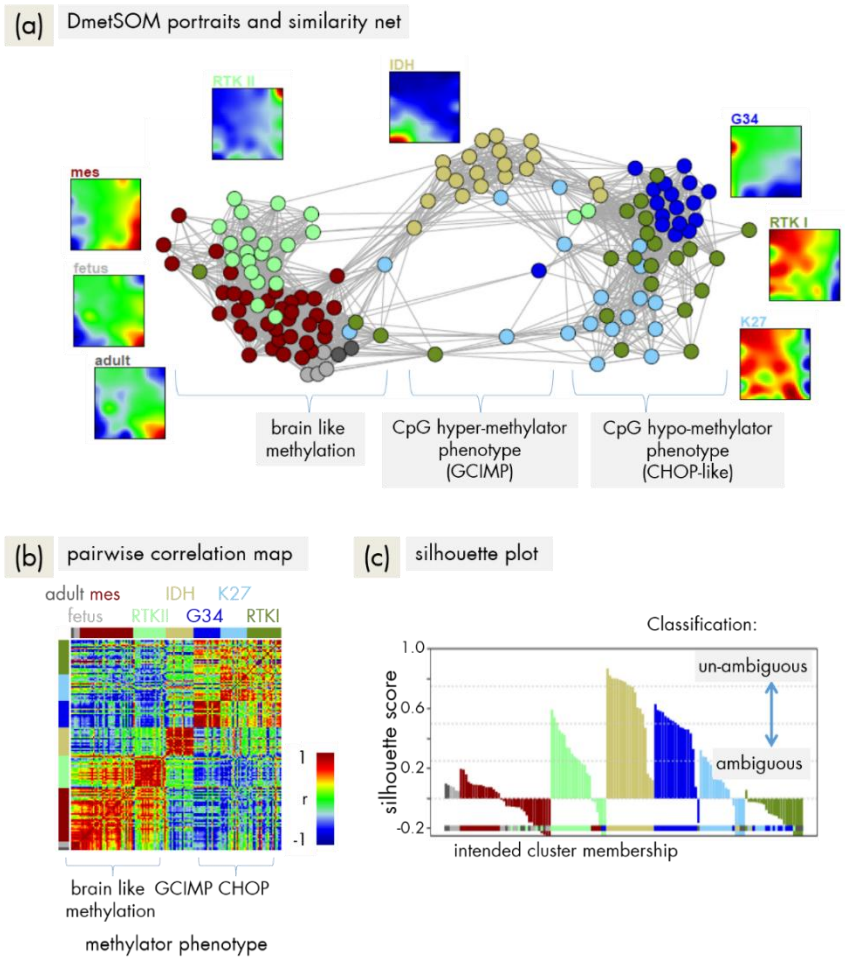


Figure 46: SOM portrayal of centralized GBM methylation data (DmetSOM): (**a**) SOM mean portraits of the GBM subtypes and of the controls and similarity net of the samples studied. Samples with strong mutual correlation coefficients are connected by lines. The sample classes can be divided into three main groups as indicated. (**b**) The pairwise correlation heatmap visualizes the mutual correlation coefficient for all pairwise combinations of samples. (**c**) The silhouette plot estimates the quality of classification of samples into methylation subtypes. Negative values indicate preference for other subtypes, which are assigned as color bar below.

5.2.2 FUNCTION MINING

The summary map in Figure 47a colors regions hypermethylated in any of the subtypes in red. After appropriate segmentation (methylation spot were detected analogously to expression modules, see methods section 3.6) we identified 12 spot-clusters each containing between nearly two-thousand and sixty single genes. Six of these spot regions labelled 'A' – 'F' show profiles with subtype-specific differential methylation whereas six additional 'satellite' spots ('A1' – 'E1') reveal more complex profiles (Figure 47b). For example, the methylation profiles of GBM samples in spots 'D' and 'D1' resemble each other whereas methylation of the controls completely changes sign. The methylation profiles of the genes in most of the spots are highly correlated providing significance levels beyond $p < 10^{-64}$

using a q^2-test statistics [103,200]. Importantly, two of the satellite spots of less variant genes refer to high (spot 'B1') or low ('E1') methylation levels in all systems studied. In the following we will focus on the main spots and the latter two satellite spots. Lists of genes in each spot are provided in supplementary material of [199].

DmetSOM analysis is based on centralized *M* values to increase sensitivity to methylation changes relative to the mean *M* value of each gene. In general one however asks for cancer specific methylation changes relative to the healthy controls. We therefore analyzed difference SOM with respect to the mean methylation map of non-neoplastic brain tissue of adults. The differential methylation landscapes support the superclusters of brain-like, GCIMP and CHOP-like methylation patterns (Figure S 3). Moreover, one sees that spot 'A1' is hypomethylated and spot 'D1' hypermethylated in all GBM compared with the healthy brain.

Extended spot statistics reveals that spots 'C', 'E', and 'F' are highly sensitive (nearly each sample of the respective subtype shows this spot) and specific (virtually no other subtypes show this spot) as hypermethylation markers for the IDH, G34, and RTKII subtypes, respectively (Figure 47c). The respective areas of the map thus can be interpreted as fingerprint regions as indicated in Figure 47a. The spot number distributions for each of the subtypes presents that most of the samples of all classes show only one or two spots (Figure 47d). However, part of the GBM samples and especially that of the MES- and RTKI- subtypes show up to five hypermethylation spots in parallel this way reflecting the high degree of fuzziness of these classes on feature level.

Gene set enrichment analysis provides first ideas about the functional context of the genes in the spot modules (Table S 4). Spots 'D' and 'E' are associated with BPs already found in gene expression analysis of GBM such as 'immune response' and 'meiosis', respectively (see section 5.1.7). For example, hypomethylation of genes from spot 'D' in MES is related to 'immune response'. It associates with high expression levels of 'immune response' genes in MES suggesting anti-correlation between DNA methylation and expression. Spots 'C' and 'E' are hypermethylated in IDH and G34, respectively. They enrich genes supporting the formation of the PRC2 and also functionally related genes such as *EED*- and *SUZ12*-targets, which control cellular development and differentiation [104]. These processes correlate with repressive and poised chromatin states defined by H3K27me3 and/or H3K4me3 histone marks in brain tissue and stem cells [201,202]. Sets of affected genes consequently enrich in these spots 'C' and 'E', as expected. We also find marker gene sets studied in previous DNA methylation and gene expression studies of GBM: For example, methylation markers for GBM of the GCIMP type [46] strongly enrich in the IDH hypermethylation spot 'C'. Interestingly, genes from this spot are hypermethylated also in other cancers such as colorectal cancer (CIMP-type CRC) and B-cell lymphoma.

In summary, spot-segmentation of the SOM of centralized methylation data provides sets of marker genes, which are specifically regulated in different glioma subtypes and which are well characterized in terms of previous knowledge. In the following subsections we will address the latter result more in detail.

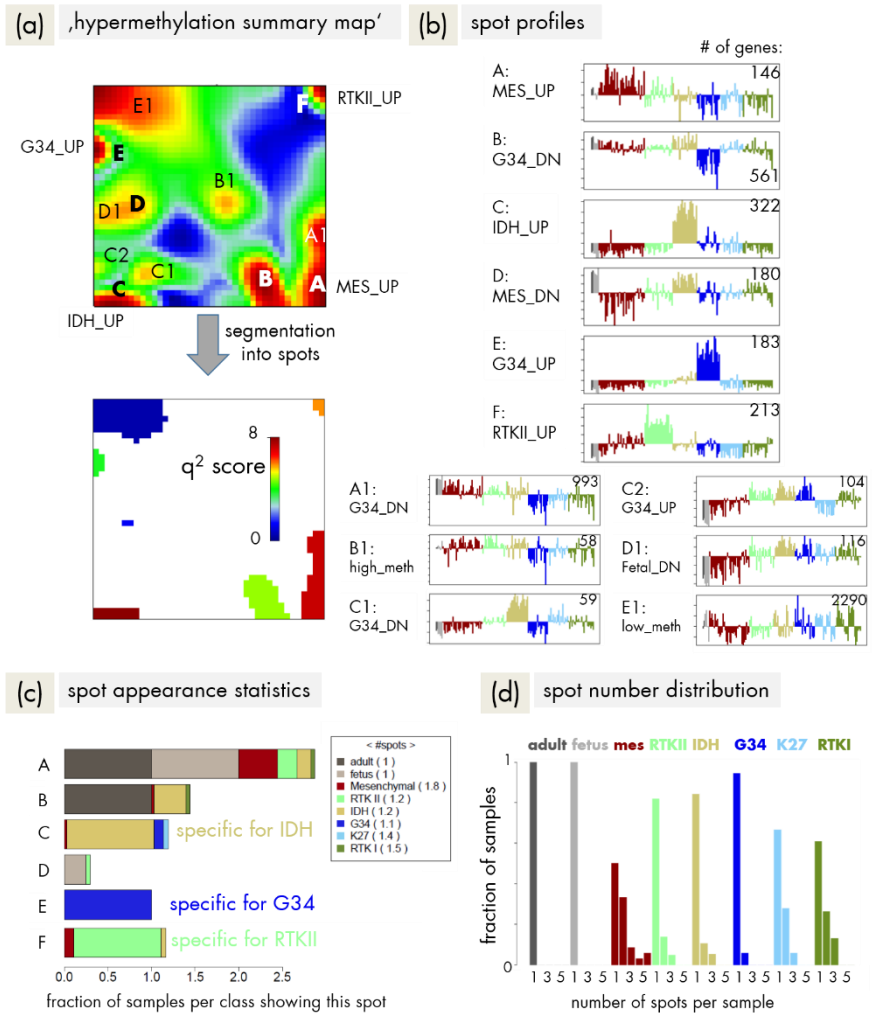


Figure 47: Segmentation of the DmetSOM into spot modules of co-methylated genes in GBM: (**a**) The hypermethylation summary map indicates regions hypermethylated in any of the classes compared with any other one in red. Each of the 'spot' regions is labeled as indicated. Segmentation of the map provides defined spot regions. Their color codes the q^2 significance score, which is minimal for cluster 'E1'. (**b**) The methylation spot profiles reveal unique hyper- or hypomethylation of selected classes for the six main spots labeled by capital letters. Six satellite spots show more subtle profiles compared with the respective main spots. (**c**) The spot statistics assigns the fraction of samples of each class that shows one of the main spots. A bar length of unity for one subtype means that all samples show this spot. (**d**) The spot number distributions show that the controls express exclusively one spot. Also most of the GBM samples in each subtype show only one spot. However, also GBM samples with three and even five spots (MES subtype) exist, reflecting the increased heterogeneity of their methylation landscapes.

5.2.3 PREVIOUS KNOWLEDGE: GBM-SPECIFIC SIGNATURE SETS

Previous DNA methylation studies on gliomas have published sets of marker genes for different molecular and histological subtypes [46,105,111,203,204]. We mapped them into the DmetSOM for analysis in terms of gene set maps and profiles (Figure 48).

The genes extracted in Noushmehr et al. [46] as 'hypermethylated and deactivated in GCIMP' clearly show hypermethylation in the GCIMP IDH subtype also in our data. However one also finds increased methylation of the G34 subtype suggesting a mixture of mainly IDH but also of G34 signature genes. Mapping of this genes set into the DmetSOM indeed reveals two regions of high local densities near the signature spot 'C' (for IDH subtype) and 'E' (for G34 subtype).

Christensen et al. [105] published a series of signature genes determined as hypermethylated in different groups of low grade gliomas (LGG) relatively to healthy controls including different WHO gradings (II or III) and histological diagnoses (astrocytoma, oligodendroglioma, oligoastrocytoma). All our maps and profiles in Figure 48 except for one show mainly the IDH signature thus indicating a common methylation patterns in LGG independent of WHO grade and histological assignment. The only exception is the methylation signature of primary GBM, which can be interpreted as a mixture of IDH and RTKII cases in the respective data. Other authors found the RTKII–signature for GBM-hypermethylation (see the data of Martinez et al. [203] in Figure 48 and also [204]). The resulting 'hypermethylation signature' obviously strongly depends on the composition of the cohort used for extracting marker gene sets. This result agrees with the fact that the incidence of each of the three subtypes RTKII (classical), MES (mesenchymal), and IDH (proneural) in random adult GBM cohorts is roughly comparable [61,63]. Without stratification into these subtypes one gets consequently a mixture of the respective signatures as observed. Note in this context that the signature of the MES subtype is consistently observed as 'hypomethylated' in GBM in a series of gene sets taken from [105,111]. Contrarily, the IDH (proneural) cases dominate with usually about 80% of all cases in LGGs [148]. The resulting signatures of different LGG strata are consequently close to that of the IDH subtype as observed. We will further discuss this point below in the context of expression signature genes.

To estimate the similarity of different gene sets one usually counts the number of overlapping genes and represents them in terms of Venn diagrams. Note, however, that for example the gene set of Noushmehr et al. 'hypermethylated in GBM' overlaps with each of the 'hypermethylated in LGG' sets of Christensen et al. by only a few genes. The percentage of overlap refers to less than 10% of the total number of genes in the Noushmehr et al. set. On first sight this result suggests the lack of similarity between these sets. Our analysis using gene set mapping however provided the opposite result. We clearly found similar enrichment profiles and enrichment maps of the different sets. It is an important benefit of our method to detect similarities between different marker sets even in the case of a small overlap between them. Such a small overlap between different but similar sets can be simply rationalized by the application of conservative significance thresholds in the selection algorithms for marker genes. High significance levels for differential expression in the original data however can neglect 'still affected' and thus functionally related genes that can become significant in one but not in alternative studies.

In summary, DNA methylation signature genes from alternative GBM studies well agree with our spot signatures. The IDH (proneural) methylation signature dominates in LGG

largely independent of WHO grade and histological diagnosis. Especially in GBM the sets reflect mixtures of the subtypes, which are present in the cohorts used for extraction of gene sets (typically IDH, classical, and mesenchymal). SOM mapping of gene sets robustly identifies similarities between different gene sets even under conditions of noisy compositions. Our approach outperforms overlap-measures as often used in terms of Venn diagrams.

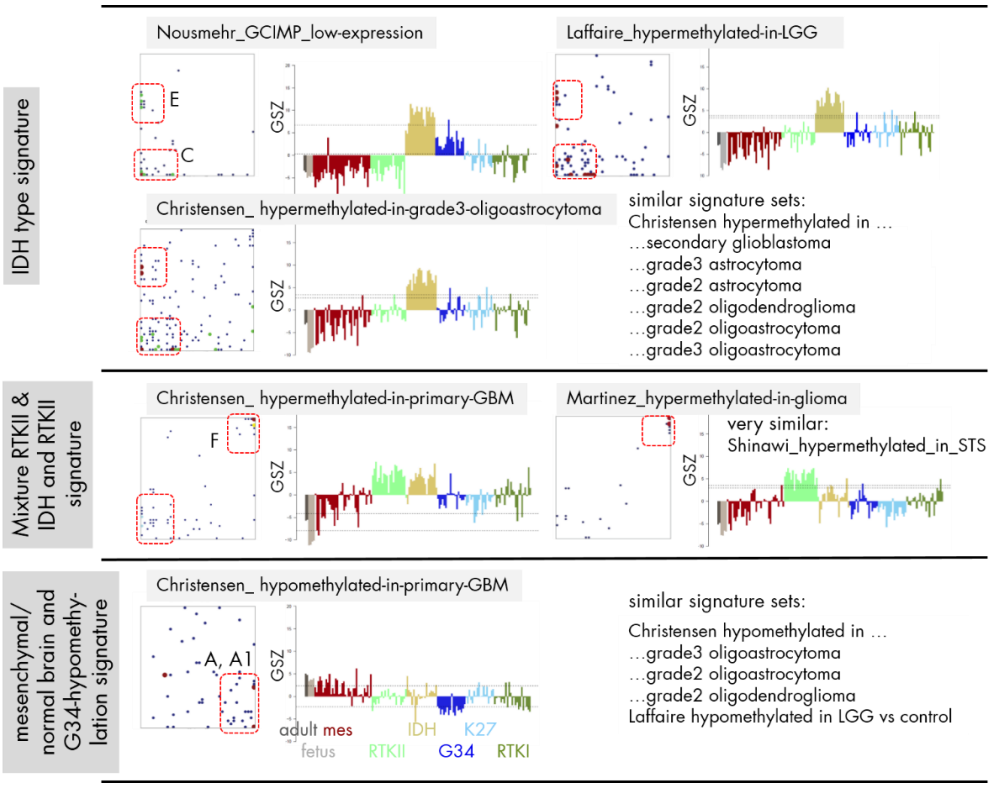


Figure 48: Mapping of methylation marker gene sets for gliomas taken from refs. [46,105,111,203,204]: The gene set maps show the distribution of marker genes in the DmetSOM. The genes accumulate in different spot areas as indicated by the red dashed frames. The GSZ profiles reveal subtype-specific methylation effects. Nearly all sets collecting hypermethylation marker genes show an IDH_UP-signature, which partly mixes with the RTKII_UP signature. Sets with very similar signatures are listed without showing the data.

5.2.4 PREVIOUS KNOWLEDGE: MARKER SETS OF OTHER CANCER ENTITIES

We previously found that GCIMP marker genes from glioma studies also differentiate between subtypes of B-cell lymphoma representing a completely different cancer entity [113]. *Vice versa*, DNA methylation gene sets from previous studies for B-cell lymphoma [113] and for colon cancer [205] also enrich in selected spots of the DmetSOM of gliomas studied here (Table S 4). This result motivated us to analyze these sets more in detail using gene set maps and profiles as described in the previous subsection. Genes, hypermethylated in the CIMP-high subtype in CRC and also genes hypermethylated in DLBCL accumulate in spots 'F' and 'C' thus revealing mixed characteristics of the RTKII and IDH subtypes in gliomas (see supplement section 7.5). This agreement between different

cancers also extends to spots 'A' and 'B', which accumulate genes hypomethylated in G34 gliomas, CRC and also DLBCL compared with BL, another subtype of B-cell lymphoma. Hence, the IDH and RTKII subtypes of GBM share similarities with the hypermethylator phenotypes in CRC and lymphomas. On the other hand the G34 subtype resembles the respective hypomethylator subtypes in lymphomas and CRC. These striking agreements suggest general mechanisms of aberrant DNA methylation in different cancer entities.

5.2.5 ASSOCIATIONS BETWEEN GENE EXPRESSION AND PROMOTER METHYLATION

In the next step we analyzed the association between gene expression and DNA methylation of the spot genes using matched samples taken from [62] (also see section 5.1) and also independent expression data [63,69] for which we matched the classes with the methylation data studied here (see Figure 49a and Table S 2). The hypermethylation spots of the MES (spot 'A'), IDH ('C'), and RTKII ('F') subtypes consistently reveal strong anti-correlation between promoter methylation and gene expression in all three analyses. The same result was obtained for the G34 subtype in the matched sample data. The independent GBM expression data does not show this effect because it doesn't contain pediatric cases. For other spots one observes the absence of systematic expression changes despite marked methylation effects (spot 'C2'), positive correlations ('B') and also neither marked expression nor methylation effects for the hyper- and hypomethylation spots 'B1' and 'E1', respectively (data not shown).

In Figure 49b we explicitly show the expression profiles of the genes from selected methylation spot sets in the Verhaak-reference data set as analyzed in section 5.1. One clearly sees that hypermethylation of the promoters of the selected genes in a selected subtype accompanies strong downregulation of gene expression of these genes. Importantly, the expression profiles respond in a subtype-specific fashion. This result reflects the important fact that the methylation classes show also class-specific expression effects and thus a close mutual relation between gene expression and DNA methylation.

To further proof this relation we mapped gene expression marker sets for LGGs (WHO grade II and III) and GBM (grade IV) into the DmetSOM to estimate their DNA methylation status (see supplementary material of [199]). In general, we found strong subtype-specific effects thus confirming the close relation between expression and methylation. For example, LGGs with a co-deletion on Chr 1 and 19 as a hallmark of oligodendroglioma show the RTKII hypermethylation signature (see supplementary material of [199]). Grade II and III LGGs differ in the methylation level of RTKII and IDH signature genes on one hand and of G34 signature genes on the other hand. Hence, we again found a mixing between different methylation classes in the subcohorts selected. The expression classes proposed by Gorovets et al. [206] for LGGs can be assigned to a brain-like_UP methylation signature (neuroblastic LGG), a mixed RTKII and IDH signature (early progenitor LGGs) and an IDH_UP signature (pre-glioblastoma, PG; see supplementary material of [199]). Note that

genes hypermethylated in IDH tumors (IDH_UP) are on low expression level in *IDH1*-mut tumors but on high level in *IDH1*-wt tumors such as PG. Hence, hypermethylation signatures of *IDH1*-mut tumors correspond to overexpression signatures of *IDH1*-wt tumors and *vice versa* due to the anti-correlation between expression and methylation effects.

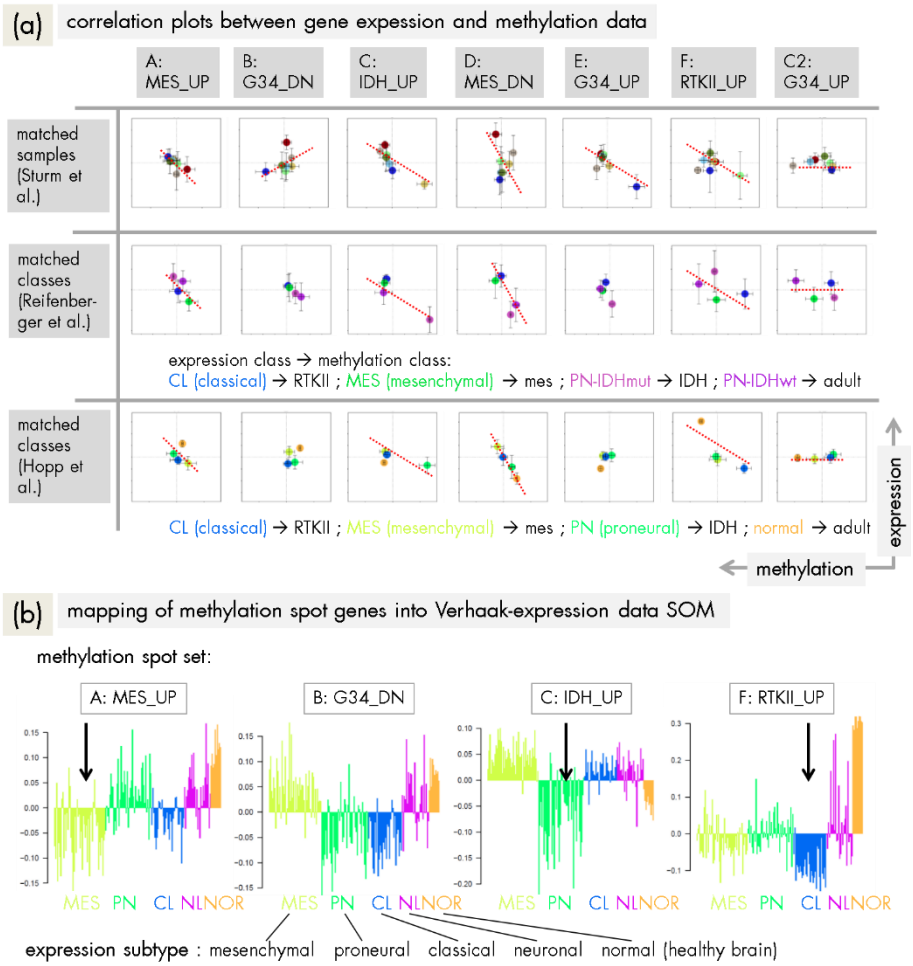


Figure 49: Correlation between GBM DNA methylation and gene expression: (**a**) Correlation plots between matched DNA methylation and gene expression data of the spot genes reveal preferentially anti-correlated changes as indicated by the red dotted lines, which serve as a guide for the eye. The matching rules for the classes are given within the figure. For details concerning the correlation plots see caption of Figure 22. (**b**) Gene expression profiles of the methylation spot sets in the GBM expression data analyzed in section 5.1: Hypermethylation sets (MES_UP, IDH_UP, RTKII_UP) associate with underexpression in the respective subtype as indicated by the arrows. Note that the color code for the GBM subtypes was chosen from the original papers [63,69].

This anti-concerted assignment of methylation and expression signatures is evident also in the expression signatures of GBM (see supplementary material of [199]): Genes, overexpressed in *IDH1*-wt tumors of the mesenchymal and/or classical subtypes are mostly hypermethylated in *IDH1*-mut, -proneural tumors. A sketchy use of terms like 'IDH_UP-sig-

nature' can imply incorrect associations because 'over'-methylation in the IDH subtype associates with 'over'-expression of another one, namely in *IDH1*-wt mesenchymal and/or classical subtypes.

Please note also, that deactivation of gene expression by DNA methylation of gene promoters represents only one possible mechanism how DNA methylation affects transcription. Alternative mechanisms are discussed, which for example explain also correlated changes between gene expression and DNA methylation. For example, a methylated DNA sequence motif can take on a new function by creating a novel DNA binding site for transcriptional activators that could not be predicted from sequence information alone. Such mechanisms expand the functional role of DNA methylation in gene regulation, being capable to regulate active and repressive gene states in a site-specific manner [207].

5.2.6 METHYLATION OF GBM SUBTYPES ASSOCIATES WITH CELLULAR PROGRAMS AND THEIR (DE-)ACTIVATION BY CHROMATIN REMODELING

Functional analysis of the spot lists of genes revealed specific functional modes and states of gene activity, which associate with the different sets of markers and thus also with the methylation subtype (Table S 4). To study the biological context more in detail we generated one-way clustered heatmaps of gene sets referring to the GO-category BP (Figure 50a), to chromatin states of brain tissue (Figure 50b), to regulators in poorly differentiated cells [208], to repressive, poised, and active histone methylation states [201] and also special gene sets with notably profiles (see supplementary material of [199]).

Firstly, one finds two 'limiting' profiles characterized by (i) high methylation of the brain-like classes and low methylation of CHOP-like classes and (ii) by the respective antagonistic CHOP-like_UP/brain-like_DN profile. The former profile comprises functions like 'neurological systems process', 'immune response' (Figure 50a), 'TFs associated with low expression levels' in mammalian cells in general [112], 'fatty acid metabolism' and partly 'transcriptional active chromatin states' (Figure 50b). These profiles are characterized by strong hypomethylation of G34, K27, and RTKI compared with the other GBM subtypes and also healthy brain. The second types of profiles (ii) are associated with high methylation levels of 'cell cycle' (Figure 50a), 'ribosomal', 'mitochondrial' genes, 'high transcription TFs', 'hypoxia', 'DNA repair' and 'ageing', partly *EZH2*-targets, *MYC*-, *NOTCH*- and *SOX2*-targtes (see supplementary material of [199]) and 'heterochromatin states' (Figure 50b) in brain-like classes and low methylation in CHOP. These two groups (i) and (ii) of antagonistic DNA methylation are mainly responsible for the two superclusters established in the similarity plots (Figure 46a and b).

Note that group (ii) associates with highly methylated genes that enrich in and near spot 'E1'. Recall that high methylation levels correlate mostly with low gene activities. Hence, 'high transcription TFs' are repressed by DNA methylation in group (ii) and packed

into closed chromatin states whereas lower methylation levels associate with active chromatin states. The situation reverses in group (i), where 'low transcription TFs' and 'active chromatin states' become repressed by high methylation levels.

In between these two 'limiting' states one finds a third type of profiles (iii) with uniquely high methylation in RTKII, IDH or G34 and also mixtures of them. These states enrich inactive chromatin states with repressed and/or poised promoters, developmental and tissue-differentiation genes and PRC2-targets and related genes: Targets of *EED*, *SUZ12* and *EZH2*, the catalytic subunit of a H3K27 methyltransferase. These results are supported by histone modification data, which shows that type (iii) profiles associate with repressive H3K27me3 and bivalent H3K27me3 and H3K4me3 marks (see supplementary material of [199]). This data also shows that so-called high CpG promoters are mainly involved in repression of these genes whereas repressed low CpG promoters associate partly with type (i) brain-like_UP methylation profiles.

Interestingly, G34 tumors associate with strong hypermethylation of genes related to promoter opening and telomere end packing. Pediatric GBM and especially G34 tumors show alternative lengthening of telomeres (ALT) mediated by homologous recombination and supported by mutations of the *ATRX* gene, which mediates histone assembly in subtelomeric regions [194]. We found strong hypermethylation of genes encoding histones H1 and partly also H 2 and H3 thus suggesting aberrant expression and in final consequence aberrant nucleosome assembly and aberrant telomere maintenance function.

Hence, IDH, G34, and also RTKII are characterized by DNA methylation and thus transcriptional repression of genes, which obviously suppress tumorigenesis in healthy brain. Hypermethylation of PRC2 repressed targets and of poised promotors is a molecular hallmark of many cancer types [208] including B-cell lymphomas [60,113] and CRC. This ubiquitous property partly explains the similar signatures of high CpG methylator phenotypes in gliomas, colon cancer and lymphomas. This agreement is further supported by overlapping chromatin states in the healthy tissues: Especially genes with poised promoter states (TssP) agree to about 50% (of about 3000 genes) in brain tissue and colon and brain and lymphoblastoid cells as well.

These results show that methylation effects associate with different chromatin states, which in turn enable different modes of gene activity in terms of transcriptional programs. Global hypermethylation of the brain-like and IDH subtypes and global hypomethylation of the CHOP-like subtypes associates with open chromatin states, which are either transcriptional active in the RTKII and also mesenchymal subtypes or inactive in IDH and partly RTKII and G34 subtypes. Methylation of closed chromatin counteracts the global net methylation tendencies, *i.e.* it is associated with reduced methylation in the brain-like and IDH subtypes and increased methylation in the CHOP-like subtypes. Note that the assignment of chromatin states is based on healthy brain data (mid frontal lobe, mf lobe), which presumably only partly can be applied to the diseased brain. Hence, methylation effects associate with changes of the chromatin states, for example if highly methylated nominal active promoter states transform into inactive ones or even into heterochromatin.

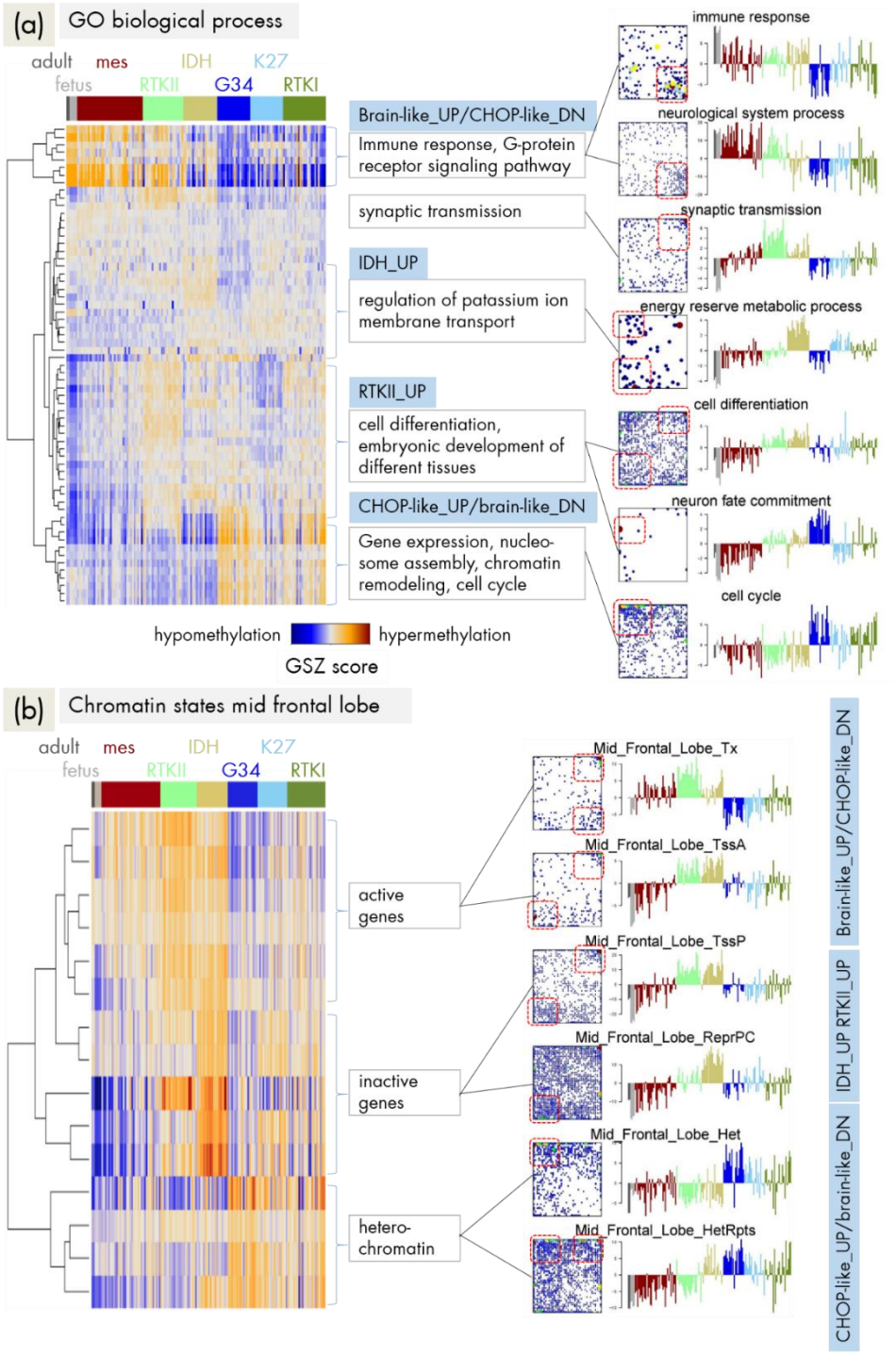


Figure 50: Methylation heatmap of genes referring to (**a**) the GO-term BP and (**b**) genes assigned to different chromatin states in healthy brain (mf lobe). Colors maroon to blue indicate high to small methylation levels, respectively. Chromatin states were grouped into active ones (e.g. Tx, Txn, TssA), inactive (ReprPC, Quies, TssP) and closed/heterochromatin (Het, HetRpts, ZNF) roughly agreeing with the clustering of methylation patterns shown in the right part of the figure.

5.2.7 DISCUSSION

SOM portrayal of marker sets resolves heterogeneity of DNA methylation across glioma subtypes, cancer entities and different cohorts

Our study focused on DNA methylation data stratified with respect to molecular subtypes of adult and pediatric GBM and healthy brain controls. Using centralized methylation data we identified clusters of co-methylated genes among the samples studied. The DmetSOM disentangles genes systematically hyper- and hypomethylated in gliomas compared with healthy brain and it extracts systematic methylation differences between the glioma subtypes. To assign the functional meaning to the spot modules we applied enrichment analysis using a multitude of pre-defined gene sets related to categories such as BP (e.g. inflammation, cell development and ageing), targets of different transcription factors (e.g. *MYC*, *NANOG*-, high and low expression TFs) and epigenetic modulators (e.g. *EED*, *SUZ12*; PRC2, *EZH2*), different chromatin states in reference mf lobe tissue, genes differently methylated in other cancers (e.g. CIMP in CRC and methylation subtypes of B-cell lymphoma) and also of marker gene sets for differential methylation and expression between glioma subtypes obtained in independent studies.

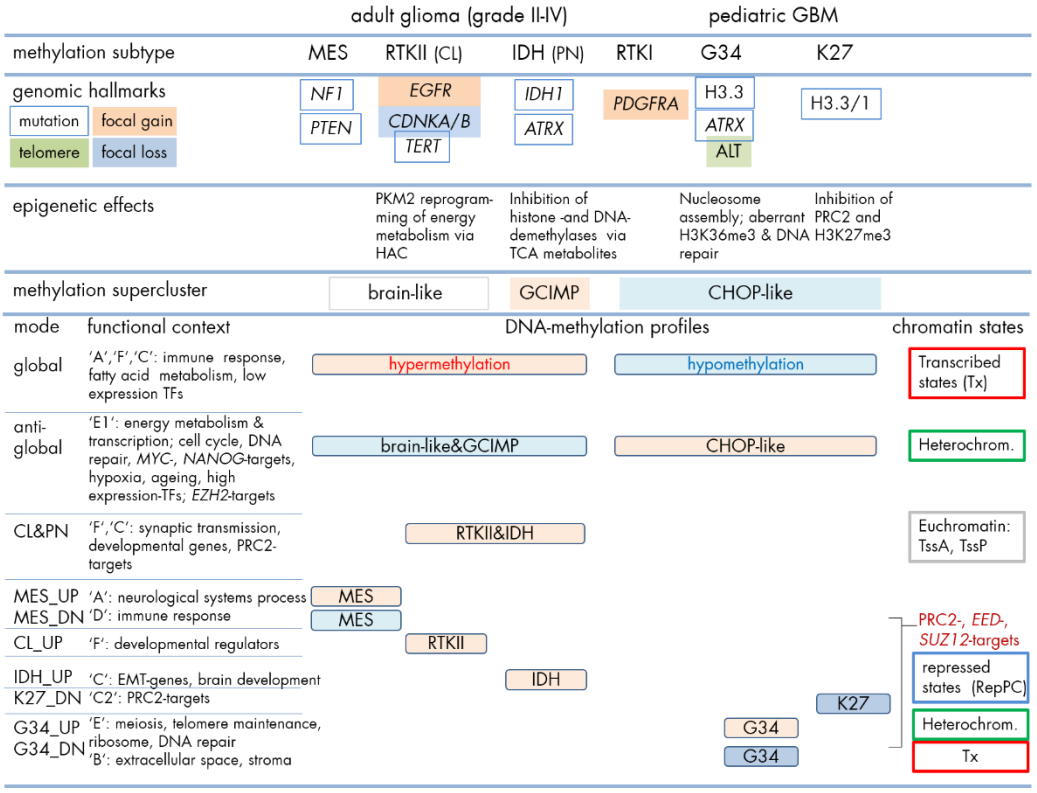


Figure 51: Overview scheme summarizing genomic hallmarks of adult and pediatric GBM subtypes, epigenetic mechanisms, and regulatory modes of promotor methylation and gene activity extracted from our analysis. The functional context associates with the spot clusters of genes obtained from DmetSOM analysis. The chromatin states refer to healthy mf lobe tissue. Their assignment to the

regulatory modes suggests specific targets for DNA methylation: For example, transcribed states in healthy brain are prone to global hypermethylation in brain-like and GCIMP tumors and prone to global hypomethylation in CHOP-like states. The antagonistic mode of methylation affects mainly heterochromatin in healthy brain. Note that promoter methylation mostly anti-correlates with gene activity: E.g., energy metabolism becomes upregulated in brain-like and GCIMP tumors compared with CHOP-like ones. PRC2-, *EED*-, and *SUZ12*-targets are hypermethylated and thus transcriptionally repressed in RTKII, IDH, and G34 but activated in MES and K27 tumors by hypomethylation.

Interestingly, we found pronounced subtype-specific methylation signatures of gene sets from different glioma studies. The signatures indicate a common scheme of aberrant gene regulation in LGGs and adult and pediatric GBM. The GCIMP signature is found across most of the glioma studies as a basal hallmark of *IDH1*-mut tumors. However, our analysis finds also mixed methylation signatures in many cases especially for histological classes, which often represent mixtures of different molecular subtypes. Hence, methylation signatures enable the further 'de-mixing' of histological classes according to molecular variants. This result supports recent studies showing that DNA-based molecular profiling of GBMs distinguishes biologically distinct tumor groups and provides prognostically relevant information beyond histological classification [148]. On the other hand, molecular profiling is hardly suited for reliable distinction of tumor grades due to grade-independent mechanisms.

We also found pronounced correlation between gene expression and methylation signatures of gliomas. It reflects coupled mechanisms of methylation and gene activity. Whether DNA methylation profiling provides a more robust and clinically useful platform for GBM subgrouping remains to be tested. The enrichment of DNA methylation signatures of other cancer entities in gliomas suggests general oncogenic methylation mechanisms.

Methylation marker sets reveal molecular mechanisms of gliomas

DNA methylation acts as an epigenetic modification in vertebrate DNA. It has become clear that the DNA and histone lysine methylation systems are highly interrelated and rely mechanistically on each other for normal chromatin function [35]. Controlling the timing and placement of DNA methylation in the genome is essential for normal cellular function and its dysfunction de-regulates cell activities. Figure 51 summarizes the main results of our study by relating methylation profiles, glioma subtypes, biological functions and chromatin states each to another.

The global methylation profile, namely hypermethylation in brain-like and GCIMP tumors and hypomethylation in pediatric GBM and RTKI, associates with the biological processes immune response and fatty acid metabolism. This mode is counterbalanced by antagonistic methylation changes, which can be assigned to cell cycle activity and energy metabolism. In healthy brain these modes accumulate genes from different chromatin states, namely transcribed states and silent heterochromatin, respectively. This result suggests that transcribed states in healthy brain become suppressed by DNA hypermethylation in brain-like and GCIMP subtypes whereas silent heterochromatin becomes possibly activated due

to hypomethylation of the affected genes in these tumors. In contrast, methylation levels in the CHOP-like pediatric GBM and RTKI correspond to the chromatin states assigned in the healthy brain. These results suggest chromatin remodeling between brain-like and GCIMP on one hand and CHOP-like tumors on the other hand. In other words, global methylation effects seem to associate with a different chromatin organization in the methylation superclusters.

These global changes were further modulated by a series of methylation effects, which refer to only a few or even single subtypes and thus define their specificity. We found hypermethylation of genes normally activated in stem cells, combined with preferential repression of polycomb-regulated genes (PRC2-, *EED*-, and *SUZ12*-targets) in RTKII, IDH, and also G34 tumors. These genes are enriched in chromatin states assigned to repressed and bivalent promoters with H3K27me3 or H3K27me3 and H3K4me3 marked histones, respectively. This methylation signature is generally found in poorly differentiated tumors [208] and, for example, also in B-cell lymphoma [60,113] indicating 'suppression of tumor suppressors' associated with tissue-specific cell differentiation [201,209,210]. Interestingly, targets of TFs involved in development and differentiation (*OCT4*, *NANOG*, *SOX2*) and also *MYC* are antagonistically methylated compared with the PRC2-targets thus suggesting different regulatory modes for more repressed and more active genes, respectively. A similar dualism was previously suggested in terms of high and low transcription TFs in metazoan which associate with high and low gene expression levels of their targets, respectively [112]. The split between both types of TFs was recently established also in lymphomas [113]. The high transcription TFs show generally a low DNA methylation level in the brain-like and GCIMP tumors. Contrarily, low transcription TFs are associated with high methylation levels reflecting the expected anti-correlated activation pattern between methylation and gene expression. In CHOP-like tumors this relation however reverses showing hypermethylation of high expression TFs and hypomethylation of low expression TFs and thus apparently improper expression levels in these tumors, which possibly reflects chromatin remodeling as discussed above.

Also K27 tumors show enrichment of PRC2-target genes becoming however hypomethylated in terms of DNA methylation and low levels of repressive H3K27me3 histone marks as well. We conclude that these genes are transcriptionally more active in K27 gliomas compared with the other subtypes in agreement with [211,212]. Hence, the Lys27 mutation of H3.3 associates with the global reduction of repressive histone marks of the H3K27me3 type, activation of gene expression and DNA de-methylation.

Aberrant hypermethylation in IDH tumors of the GCIMP type is induced mostly by the IDH1 mutation leading to inhibition of histone-lysine- and DNA de-methylases carrying the Jumonji-domain via intermediate metabolites of the citrate cycle which act as their coenzymes [144]. In tumors of the RTK-types epigenetic dysregulation associates also with metabolic reprogramming, namely with aberrant activation of the pyruvate kinase M2 (*PKM2*) isoform, a glycolytic enzyme involved in ATP generation and pyruvate production, which plays an essential role in tumor metabolism and growth. It also functions as a protein kinase

that phosphorylates and/or acetylates histones during transcription and chromatin remodeling with consequences for CpG methylation [213]. RTKI tumors together with K27 and G34 show hypermethylation of genes related to 'pyruvate metabolism', 'ATP binding', 'mitochondrion', and 'ribosome cellular components' suggesting transcriptional down regulation of the energy metabolism and protein synthesis. Interestingly, 'targets of *EZH2*', a compound of PRC2 catalyzing the formation of H3K27me3, show similar methylation profiles, which also resemble those of genes up-regulated upon ageing and under hypoxia. Subtle differences between the methylation profiles 'pyruvate metabolism' and 'ATP binding'/ 'mitochondrion' in IDH gliomas on one hand and G34 and RTKI on the other hand however suggest different mechanisms of metabolic control in these subtypes.

G34 tumors show specific hypermethylation of genes associated with 'telomere length maintenance' (Reactome sets 'packaging of telomere lengths' and 'pol I promoter opening'), 'histone assembly', and 'DNA repair' suggesting increased genomic instability of this subtype. G34 tumors display an 'ALT' (alternative lengthening of telomeres) phenotype presumably mediated by homologous recombination and caused by the mutation of the *ATRX* gene and possibly also by the G34 mutation of H3.3 itself [62,214]. Aberrant DNA repair functionality in G34 is possibly associated with DNA hypermethylation of the respective genes and aberrant methylation markings of H3K36me3 required for proper recruitment of the DNA-repair machinery [176,177,194]. Interestingly we find also strong hypermethylation of genes referring to ribosome and mitochondrial functions in G34 suggesting a deactivation of transcriptional and energy-metabolic processes in this subtype.

Taken together, these findings illustrate a widespread functional role of DNA methylation in gene regulation in gliomas essentially contributing to the heterogeneity of glioma subtypes and strongly affecting the underlying molecular mechanisms of cell function.

5.2.8 CONCLUSION

Sets of differential methylation genes in gliomas represent surrogate markers of molecular mechanisms governing (epi-)genomic dysregulation. DNA methylation phenomena are complex ensuring complex tuning of gene function. Consideration of this regulatory level is inevitable for understanding cancer genesis and progression. It provides suited markers for diagnosis of glioma subtypes and disentangles tumor heterogeneity.

5.3 COMBINED PORTRAYAL OF GENE EXPRESSION AND DNA METHYLATION IN GLIOBLASTOMAS

In this section we aim at extending the SOM portrayal method of single-omics data to the combined portrayal of two data types where we selected gene expression and DNA methylation data because of their impact for tumor biology.

We selected TCGA data of high grade glioma as the tumor entity for combined portrayal because of the extended pre-work characterizing gene expression and DNA methylation data and because of established classification schemes (see, e.g., [46,61–63] and references cited therein). On the other hand, particular modes of gene regulation in glioma subtypes governed by mutations of the *IDH1* gene or of receptor-tyrosine kinases (RTKs) need specification in terms of involved genes, affected chromatin states and co-regulated chromatin modifying enzymes and will be studied here. Therefore we used data obtained from Sturm et al. [62] with the samples being assigned to Mesenchymal (MES), RTKI 'PDGFRA' (RTKI), RTKII 'Classic' (RTKII), and IDH molecular GBM subtypes. This data set is just a small subset of the samples analyzed in 5.2 as in terms of this multi-omics integration presented here we needed both gene expression and methylation samples collected for the same patients studied. For details concerning the cohort and preprocessing of the data see section 3.1 and supplement section 7.1.7.

The first part of this section is devoted to methodical issues such as the modulation SOM algorithm, the characterization of the data landscapes obtained and the identification of regulatory modes of co-expressed and co-methylated genes. In the second part we focus on the functional interpretation of these modes and their relations to glioma key genes, to chromatin states in healthy brain and the transcription of selected chromatin modifying enzymes. The second part thus aims at providing a comprehensive view on epigenetic factors affecting gene activity in glioma that can be extracted from our combined analysis. It extends our previous integrative SOM studies on epigenetic regulation in B-cell lymphomas (sections 4.2 and 4.3) and DNA methylation in gliomas (section 5.2).

5.3.1 PREPROCESSING: CENTRALIZATION AND HARMONIZATION

As input we used gene-centered expression (E_{nj}) and methylation (M_{nj}) data obtained from microarray experiments. Both expression and methylation data were preprocessed as given in section 3.2. For combined analysis expression and methylation data were transformed into a unique, 'harmonized' scale by normalizing them with respect to the mean absolute value averaged over all data $\Delta e^{*}_{nj} = \Delta e_{nj}/\langle|\Delta e|\rangle_{all}$ and $\Delta m^{*}_{nj} = \Delta m_{nj}/\langle|\Delta m|\rangle_{all}$, where $\langle|\ldots|\rangle$ denotes averaging of absolute values. This harmonization makes the scales of expression and methylation data mutually comparable.

5.3.2 MODULATION SOM: PORTRAYAL OF COMBINED EXPRESSION AND METHYLATION STATES

Method

In the next step after centralization and harmonization we merged expression and methylation profiles into combined profiles

$$\Delta d^*_{n\bullet} = (w \cdot \Delta e^*_{n\bullet}, (1-w) \cdot \Delta m^*_{n\bullet}), \qquad \textbf{Eq.(8)}$$

where both data were combined with different mutual weights, w, chosen from the data intervals $w = [0,1]$ and $(1-w) = [1,0]$, respectively (see Figure 52 for a schematic overview of the data processing pipeline). The combined profile vectors are then clustered into prototypic profiles by applying SOM machine learning using our implementation 'oposSOM' [20], which was previously described in detail [21,80,91]. The metagene profiles are given by

$$\Delta d^*_{k\bullet} = (w \cdot \Delta e^*_{k\bullet}, (1-w) \cdot \Delta m^*_{k\bullet}), \qquad \textbf{Eq.(9)}$$

with $k = 1,\ldots,K$ (K is the number of metagenes).

After SOM training meta- and single gene expression and methylation data are back transformed into their original scales for visualization and further downstream analysis, *i.e.* $\Delta d^*_{kj} \rightarrow \Delta e_{kj}, \Delta m_{kj}$ and $\Delta d^*_{nj} \rightarrow \Delta e_{nj}, \Delta m_{nj}$, by keeping the cluster associations between metagenes and single genes. Accordingly, each sample studied is characterized by its state of metagene expression and methylation, $\Delta e_{\bullet j} = (\Delta e_{1j},\ldots,\Delta e_{Kj})$ and $\Delta m_{\bullet j} = (\Delta m_{1j},\ldots,\Delta m_{Kj})$, respectively. In analogy also each metagene is characterized by its profile of expression and methylation values in all the samples studied, $\Delta e_{k\bullet} = (\Delta e_{k1},\ldots,\Delta e_{kj})$ and $\Delta m_{k\bullet} = (\Delta m_{k1},\ldots,\Delta m_{kj})$. The expression and methylation states are visualized by color coding the metagene values in the quadratic mosaic grid used for SOM training. This way one obtains two images per sample, which separately 'portray' its expression and methylation landscapes (Figure 52). Importantly, each metagene is located at the same position in both maps. It is associated with the same cluster of single genes because of the joint training of expression and methylation data. Therefore, both maps can be directly compared each with another, e.g. to identify regions of specific combinations of expression and methylation data. For this aim it is desirable to merge these separate expression and methylation portraits into one joint image, which enables to directly identify combinatorial properties such as genes whose down-regulated expression is associated with hypermethylation or *vice versa*. With this aim we calculate the signed square root co-variance (ScoV) for each metagene and each sample,

$$c_{kj}(\Delta e, \Delta m) = \text{sign}(\Delta e_{kj} \Delta m_{kj}) \cdot \sqrt{\Delta e_{kj}, \Delta m_{kj}}, \qquad \textbf{Eq.(10)}$$

which provides the state vector $c_{\bullet j} = (c_{1j},\ldots,c_{Kj})$. It defines the covariance landscape of the expression and methylation data of each sample using the SOM-mosaic arrangement and

a suited color code. As standard we used a sample-related scale, which colors the minimum and maximum values of each state in dark blue and red, respectively. SOM portraits of expression, methylation and of combined covariance data were obtained for each sample. Mean expression, methylation and covariance landscapes for each subtype were obtained by averaging the respective metagene values over the samples of each class.

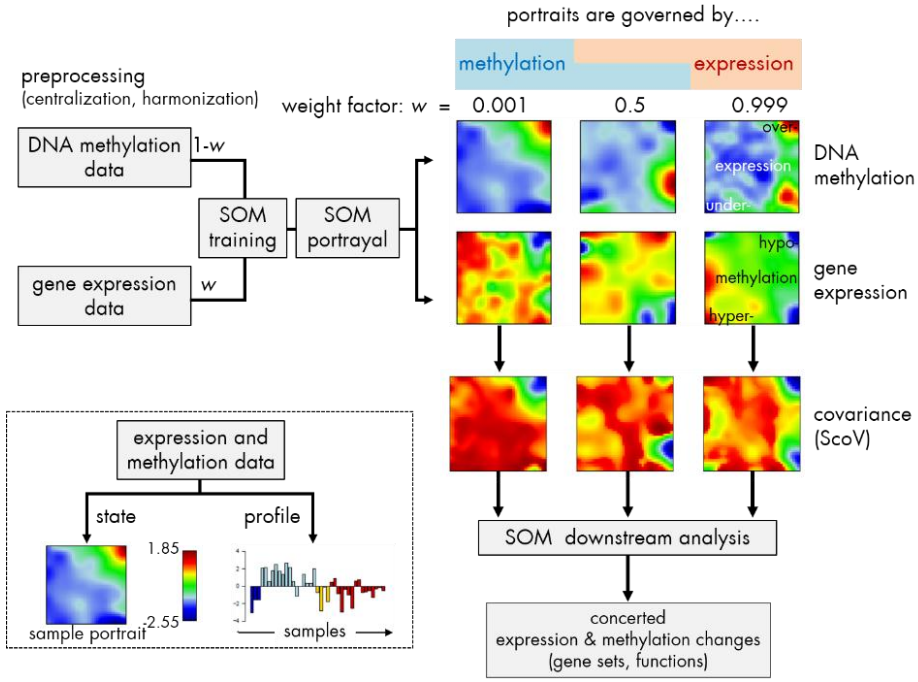


Figure 52: Schematic overview of the modulation SOM method (see text).

Application

SOM machine learning transforms the whole gene expression and promoter methylation pattern of about twenty thousand single genes into three mosaic images per sample, one for methylation and expression data each and one for the covariance (ScoV) between them (Figure 53), each of size 40x40. The weight-factor *w* tunes the SOM structure from 'governed predominantly by co-methylation of the genes among the samples' ($w = 0.001$) to 'governed predominantly by co-expression' ($w = 0.999$) via 'governed by both' ($w = 0.5$) in between. In other words, $w = 0.001$ essentially provides the methylation landscape, which is modulated by the expression data whereas $w = 0.999$ provides the expression landscape modulated by methylation data. An equally weighted landcape is obtained with $w = 0.5$. The red and blue spots in the images assign clusters of overexpressed/-methylated and underexpressed/-methylated genes, respectively. The covariance (ScoV) map combines this data to identify clusters of concerted (red) and anti-concerted (blue) changes of expression and methylation. Importantly, for each *w* the genes are located at the same position in each of the three maps and thus they can be directly compared. E.g., sample 1 (upper part of Figure 53) shows one overexpression and one hypomethylation spot virtually located at the same position of the expression and

methylation maps at $w = 0.001$, which combine into a blue spot in the ScoV-map indicating anti-correlation between DNA promoter methylation and expression of the respective genes. With increasing weigth the structure of the SOM changes and becomes progressively governed by the expression data. The second sample was chosen from another GBM-subtype (IDH-subtype). The SOM portraits show an almost different spot structure compared with sample 1 (MES-subtype), which indicates different methylation and expression landscapes. The full gallery of SOM images of all samples studied is given in the supplementary material of [215].

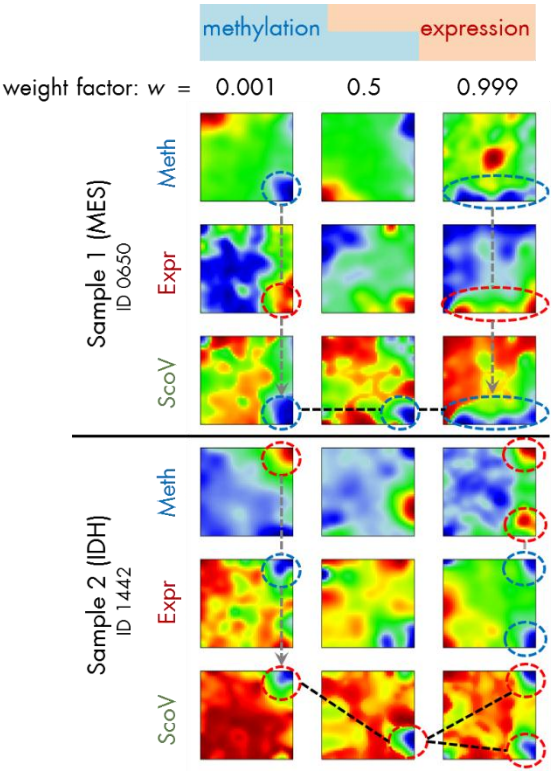


Figure 53: Expression, methylation and ScoV portraits of two selected GBM samples obtained after SOM training using three different weight factors, which provide SOM portraits that are governed by the methylation data ($w = 0.001$), expression data ($w = 0.999$) or both ($w = 0.5$). The circles indicate clusters of genes with anti-correlated methylation and expression changes.

5.3.3 SAMPLE DIVERSITY IN METHYLATION AND EXPRESSION PORTRAITS

To assess how SOM-transformed gene expression and methylation data reflect the diversity of the samples we calculated similarity networks based on the correlation coefficients between the metagene values of all pairwise combinations of samples (Figure 54). Methylation data well separates the samples into the four different GBM subtypes, considered. This result is expected because the subtypes were defined according to the methylation characteristics of the samples, however based on CpG-island level data and not integrated promoter metagene data. Our results thus confirm this classification for promoter methylation data. The SOM clustering into metagene-data depends on the chosen weight

coefficient *w* giving rise to slightly varying similarity plots, which however all identify the subtypes properly (Figure 54). Interestingly, also the expression data provides well separated clusters of the IDH, MES, and RTKII subtypes thus indicating associations between expression and methylation changes observed.

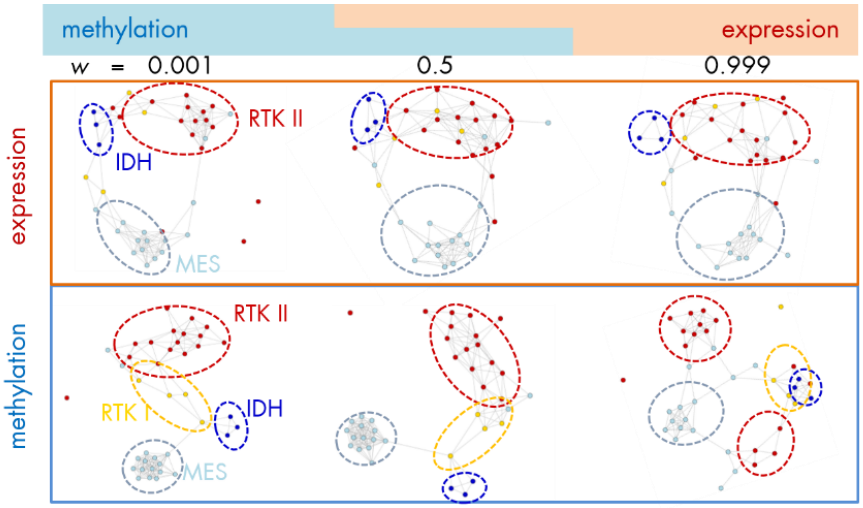


Figure 54: Diversity analysis of the GBM samples was performed in terms of CNs of the expression and methylation SOM images for different weight factors. The dashed circles indicate clusters formed by samples of the different subtypes.

5.3.4 EXPRESSION, METHYLATION AND COMBINED PORTRAITS OF GLIOMA SUBTYPES

For a better description of the subtypes we calculated the mean expression, methylation and ScoV-portraits of each class by averaging the respective values of each metagene over all class members (Figure 55a). The mean ScoV-portrait of each subtype shows characteristic blue spot patterns representing clusters of genes with anti-correlated methylation and expression values where either hypermethylation combines with underexpression or hypomethylation with overexpression (see the red and blue frames in Figure 55a, respectively). The most indicative spots were assigned by 'M1'–'M3' in the ScoV-methylation driven landscape ($w = 0.001$), 'E1'–'E5' in the ScoV-expression driven landscape ($w = 0.999$) and 'C1'–'C3' in the combined landscape ($w = 0.5$).

For example, for $w = 0.001$ the IDH subtype is characterized by an underexpression spot in the top-right corner, which at the same time is hypermethylated thus resulting in a blue spot in the ScoV map assigned as 'M1'. It indicates anti-correlated changes of methylation and expression of the included genes. The MES-subtype shows in both expression and methylation landscapes an almost mirror symmetrical pattern with regard to the RTKII subtype where blue regions convert into red ones and *vice versa*. These patterns show that expression and methylation levels of many genes change in an antagonistic way between these two subtypes. Moreover, the MES and RTKII are characterized by the spots 'M2' and 'M3' in the ScoV-map, respectively, whereas RTKI shows a superposition of all three main

spots 'M1'–'M3'. For $w = 0.5$ and $w = 0.999$ the spot patterns vary where, e.g., 'M1' first transforms into 'C1' and then into 'E1' and 'E2' for the IDH subtype. The question arises whether these spot modules are formed by the same genes or different ones and why a singular spot splits into two in the different landscapes.

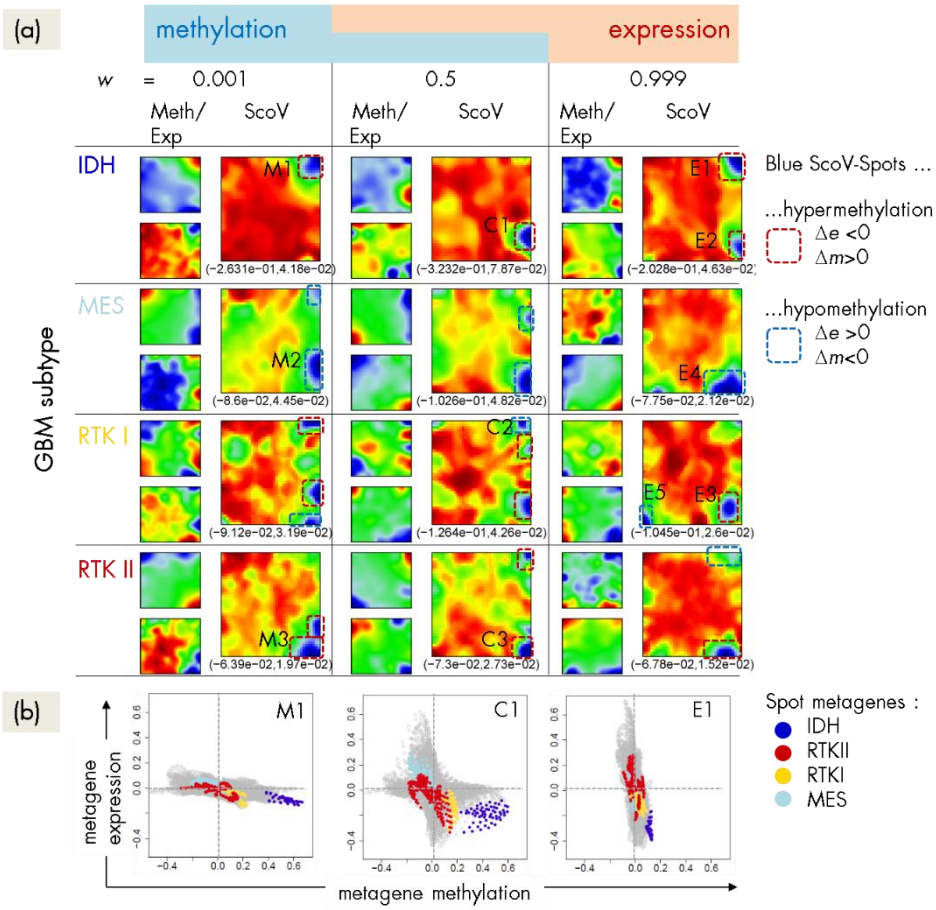


Figure 55: Gallery of mean expression, methylation and ScoV portraits of glioma subtypes. (**a**) The small mosaic images visualize the mean expression (upper) and methylation (lower) landscapes. The large images are the mean ScoV-portraits. (**b**) Correlation plots of SOM-metagene expression and methylation values for different *w*. The variance of the data is dominated either by expression ($w = 0.001$) or by methylation ($w = 0.999$) values whereas at $w = 0.5$ the data shows a combination of both modes. The colored dots are metagenes of the spots 'M1', 'C1', and 'E1', which are characteristic for blue ScoV-spots of the IDH subtype. The grey dots show the remaining metagenes not included in the respective spots.

5.3.5 BASIC AND MODULATED STRUCTURE OF THE SOM

The landscapes and consequently also the location of spots and their number change after tuning the weight factor between $w = 0.001$ and 0.999. The SOM structure is governed by both methylation and expression data, where either one is considered with major weight and the other one with minor weight. The major weight (e.g. $w = 0.999$ for expression data) determines the basic structure of the landscape whereas the minor weight

(e.g. (1– *w*) = 0.001 for methylation data in this map) modulates this basic structure. The plot of expression-vs-methylation metagene data shows that the major weight component determines whether methylation or expression data shows largest variance (Figure 55b, *w* = 0.001 and 0.999, respectively). For *w* = 0.5 the metagene data splits into two main components virtually aligning along the methylation and expression axes, respectively (Figure 55b, *w* = 0.5). The blue colored dots mark metagenes of selected ScoV-spots, comprising genes of anti-concerted methylation and expression values in the IDH subtype. The correlation values tend to align along a diagonal line from top left to down right in the quadrants, which are defined by hypermethylation and underexpression values however with a marked systematic deviation towards the major component for *w* = 0.001 and 0.999, respectively. This analysis shows that the different SOMs reflect different self-organization modes of the data with the consequence that genes with strongly variant methylation, expression or both are collected together in the spot clusters.

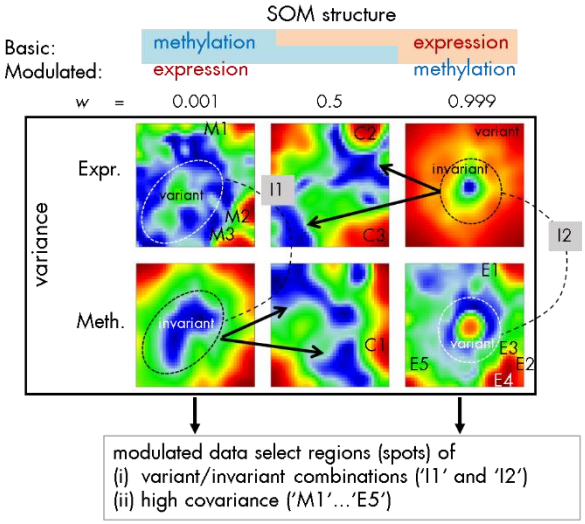


Figure 56: Variance maps of the expression (row above) and methylation (row below) metagene profiles in the SOM for different weight factors *w*. The color sequence blue-green-red indicates increasing variance from low to high. The spot structure is governed by the modulated data. The spots of variant/invariant combination split into different modes at *w* = 0.5 (see arrows).

For a more systematic view we make use of variance maps [21] shown Figure 56, which color code the variance of the metagene profiles in each pixel between blue (invariant expression/methylation) and red (highly variant expression/methylation). The maps reveal two main topological characteristics: Firstly, one detects regions of genes with virtually invariant expression and variant methylation in the center of the maps for *w* = 0.999 (see dashed circles around the spot 'I2' in the right part of Figure 56) and genes with virtually invariant methylation and variant expression in the center of the maps for *w* = 0.001 (region 'I1' in the left part of Figure 56). In other words, SOM training when governed by expression data (*w* = 0.999) clusters genes with almost invariant expression profiles together, which however show still substantial variability of their methylation data. *Vice versa*, SOM training when governed by methylation data (*w* = 0.001) clusters genes

with almost invariant methylation profiles together, which show moderate expression changes. Hence, the maps for $w = 0.001$ and 0.999 identify genes in these two orthogonal situations. SOM training at $w = 0.5$ provides a map structure reflecting the transition between them where the variant/invariant combinatorial clusters split into two or more regions (see arrows in Figure 56).

Secondly, the SOM governed by expression data ($w = 0.999$) locates genes with highly variant expression profiles along the borders of the map (right part of Figure 56) whereas the SOM governed by methylation data ($w = 0.001$) locates genes with highly variant methylation profiles along the borders of the map (left part of Figure 56). These red regions of high variance get modulated by the methylation and expression data in the first and second case, respectively. The ScoV-spots were formed by regions of high variance of the modulated data.

5.3.6 SPOT GENES OVERLAP AND SPOT 'MELTING'

Next we asked for the mutual similarity between the blue negative-ScoV-spots determined for different w in terms of the fraction of overlapping genes, *i.e.* of genes found together in all pairwise combinations of spots. The pairwise spot overlap matrix was visualized as a heatmap (Figure 57b) and as a spot overlap network (Figure 57a) where the edges connect spots of substantial overlap in the ScoV-summary maps. One sees that, for example, spot 'M1' overlaps with 'C2', which in turn overlaps with 'E1' and 'E2' however to a weaker degree. In other words, one finds mutual footprints of the genes included in the spot clusters between the different maps.

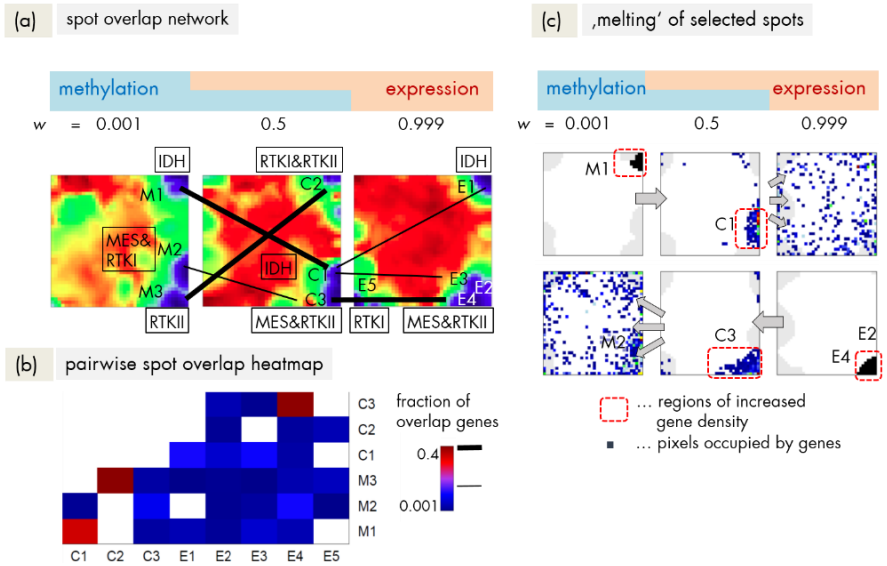


Figure 57: Spot overlap analysis provides fractions of overlapping genes in all pairwise combinations of spots. (**a**) and (**b**) spot combinations of relative large overlaps were connected to form the overlap network, which provides a footprint of genes between the spots in the three types of maps. (**c**) Genes aggregated in selected spots progressively 'melt' and 'dissolve' over the whole map upon tuning w. Metagenes occupied by at minimum one gene are indicated by dots in the maps.

For an alternative visualization we 'track' the genes of a spot selected in one of the map in the other maps (Figure 57c). This representation shows that the spot-structure progressively 'dissolves' with changing w thus reflecting the alteration of self-organizing properties of the map. The spot clusters observed at the intermediate $w = 0.5$ form sort of a consensus clusters connecting the spots observed in the two other maps at $w = 0.001$ and $w = 0.999$. Hence, tuning the basic components (methylation or expression) redistributes the genes and selects different genes in the spot clusters.

5.3.7 FUNCTION MINING OF SPOTS MODULES

The functional context of the genes included in each of the spot-clusters considered was studied using gene set analysis as described in section 3.7. We extracted a series of consensus functional modes across the maps, which reflect the footprints of overlapping genes discussed above. Modes I and II are governed by genes hypermethylated and underexpressed in IDH and, in consequence, overexpressed in RTKII (mode I) and MES (II) subtypes (see the expression-vs-methylation plots in Figure 58). These modes enrich marker genes for GBM subtypes with non-mutated *IDH1* [61,63,69]. In addition, especially mode I is enriched with genes located on 'Chr 7', which shows pronounced copy number gains as a hallmark of *IDH1*-wt gliomas [63], associated with overexpression of the respective genes. Here the methylation subtype RTKII [62] largely agrees with the classical subtype (CL) [61].

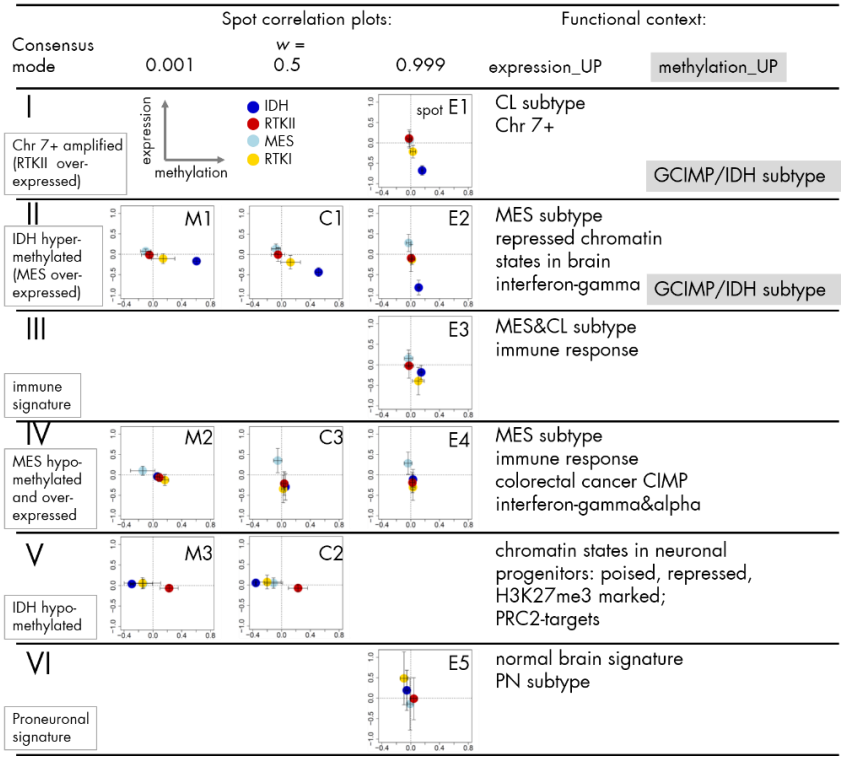


Figure 58: Functional analysis of ScoV-spot modules provides distinct consensus modes, which can be assigned to previous knowledge.

Mode III reflects an 'immune signature', which is observed in tumors of the MES and CL subtypes whereas mode IV, although showing a similar functional context, is more specific for MES in terms of overexpressed and hypomethylated genes. Mode V (and partly also mode II) is enriched with genes located in 'repressed' and 'poised' chromatin states in neuronal progenitor (NP) cells and with PRC2-targets, which are found to be affected by aberrant DNA methylation in glioma and other cancer entities [106,113,199]. Finally, mode VI shows proneural characteristics enriching also genes transcriptionally active in 'healthy brain'. Additional important modes can be assigned to spots 'I1' and 'I2' (Figure 56) showing either almost invariant methylation or expression, respectively. The former one is characterized by genes in 'active and transcribed' chromatin states and also 'MYC-targets' related to high 'cell cycle activity' whereas the latter mode enriches TFs associated with 'low expression' levels [112] and genes encoding 'G-protein coupled receptors' with function in 'olfactory transduction' (see supplementary material of [215]). Interestingly, the latter mode also enriches genes found hypermethylated in the 'CIMP phenotype of colorectal cancer' [22], which suggests susceptibility of the same genes for aberrant methylation patterns in different cancer entities in agreement with our previous analyses [199].

5.3.8 GENE SET MAPS

Gene set analysis as applied in the previous section estimates enrichment of genes of a set within a spot cluster of genes. In this section we apply the complementary approach of gene mapping, which enables us to visually inspect the distribution of sets of genes in the expression landscape and thus to study their local accumulation in distinct regions of the map. First, we mapped sets of genes hypermethylated or overexpressed in selected glioma subtypes, which were taken from previous analysis (see section 5.2.2) and studies [46,61,199](Figure 59a, b). The former sets accumulate in distinct spot clusters of our methylation map ($w = 0.001$). They can be assigned to our main consensus modes I, II, IV and V as defined in Figure 58 and show strong anti-correlation between promoter methylation and expression (see red dashed lines in the correlation plots). Gene sets overexpressed in the CL and MES subtype accumulate in our spots 'E1', and 'E2' and 'E4', respectively, whereas genes upregulated in the PN subtype form distinct clusters near the left border of the expression map ($w = 0.999$). Theses clusters were not explicitly considered as spot clusters here because of their moderate negative values in the ScoV-map not meeting the threshold criterion for spot selection. However, the respective regions reveal areas of hypermethylation and underexpression in the methylation and expression maps of the IDH subtype (Figure 55a). Hence, gene set mapping in general confirms the results of expression and methylation analyses from independent studies in the data studied here. Moreover, maps of gene sets extracted from glioma subtypes not included in our data set such as pediatric and low grade glioma nevertheless form gene clusters in our data landscape of adult, high-grade glioma (see supplementary material of [215]) thus indicating functional overlap between these tumor groups without clear borderlines. Finally, genes

highly expressed in 'healthy brain' accumulate in spot 'E5' not evident in glioma but, interestingly, show partly similarities concerning their methylation patterns with RTKII (spot 'M3').

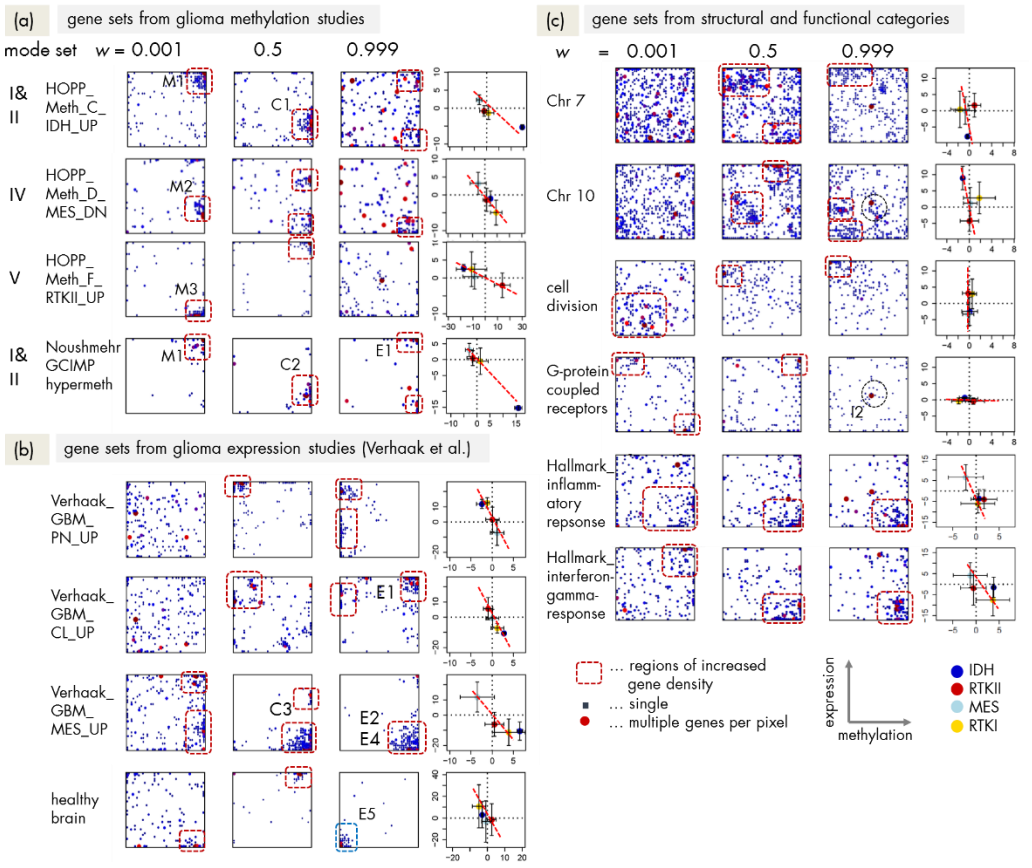


Figure 59: Maps and expression-methylation correlation plots of gene sets of selected categories. (**a**) Gene sets taken from methylation studies [46,62,199] provide genes hypermethylated in glioma subtypes whereas (**b**) gene sets taken from expression studies [21,61] provide genes overexpressed in the respective glioma subtypes and healthy brain. (**c**) Gene sets collect genes located at Chr 7 and 10, of the GO terms, and of hallmarks of cancer [216]. The correlation plots show the mean expression GSZ score of each subtype as a function of the mean methylation GSZ score of each subtype.

As a third group we mapped gene sets assigned to selected structural and functional categories (Figure 59c). Copy number gains on 'Chr 7' and losses on 'Chr 10' are genetic hallmarks of a large fraction of *IDH1*-wt glioma [63,217](see also Figure 60a below). Genes located on these chromosomes aggregate in distinct areas of the map, which collect genes over- and underexpressed in the *IDH1*-wt (*i.e.* all subtypes except IDH) gliomas thus reflecting a dose-response relation between copy numbers and gene expression. Interestingly, the area of increased local density of 'Chr 7' genes partly agrees with the local enrichment of genes related to 'cell cycle activity'. Hence, gains on 'Chr 7' in the non-IDH subtypes associate with increased proliferation of the cancer cells where the changes in gene expression largely dominate compared with DNA methylation changes. Interestingly,

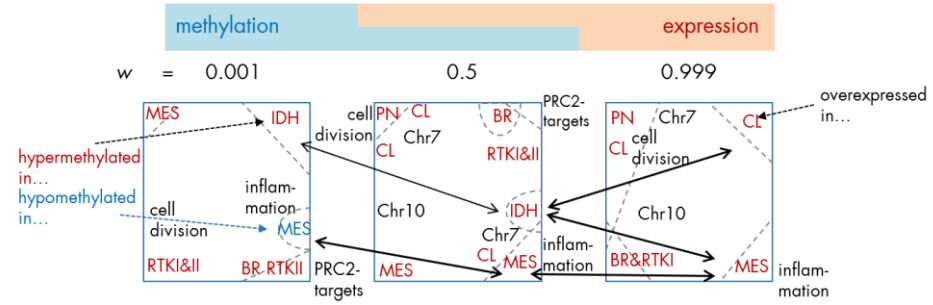
these areas of proliferative function agree with areas of genes carrying low DNA methylation levels at their promoters, which confirms their highly activated state (see supplementary material of [215]). Genes related to 'inflammation', 'interferon gamma response' and also 'stroma' accumulated in areas upregulated in MES accompanied by moderate methylation changes. In contrast, genes encoding 'G-protein coupled receptors' show strong methylation changes, which however are not paralleled by expression changes as indicated by their accumulation in spot 'I2'. 'PRC2-targets' and genes with 'poised promotors' show a similar behavior (see supplementary material of [215]).

5.3.9 MOLECULAR LANDSCAPES AND KEY GENES

We summarized the information about differentially methylated and expressed genes in the different subtypes and their functional context together into SOM-maps. In this way molecular landscapes governed by promoter methylation, gene expression or both are defined as discussed above (Figure 60a). The maps reveal rough rules of thumb: (i) Methylation and expression subtypes can be assigned to distinct areas of the map where groups of genes show characteristic differential methylation and/or expression; (ii) these areas can be associated with selected functional categories such as 'cell division' and 'inflammation' specifically activated in MES together with 'stromal signatures'; (iii) also chromosomal defects associate with part of these areas, e.g. gains on Chr 7 associate with overexpression of genes in CL. The fact that subtypes can be characterized by more than one group of genes in different regions of the map (e.g. MES) indicates an intrinsic heterogeneity in the subtypes not resolved by the classification of glioma used here. The consensus modes I - III connect areas attributed to genes (hypermethylated) in the IDH subtype in the methylation map (spot 'M1') with two areas ('E1' & 'E2') attributed to genes overexpressed in CL and MES subtypes. This 'switching' of subtypes trivially reflects the anti-correlation between promoter methylation and gene expression. A similar situation applies to the region enriched with Chr 10 genes: Due to the copy number loss in many *IDH1*-wt glioma the area is attributed to the PN-subtype that lacks defects at Chr 10 without reduced expression.

Part b of Figure 60 documents the correspondence between glioma subtypes defined in DNA methylation and gene expression studies together with key (epi-)genetic defects, *i.e.* genes frequently mutated and/or hypermethylated in glioma [61,62]. Mapping of these genes into the data landscapes reveals: (i) Part of them (*CDKN2A*, *PTEN*, and *CDKN2B*) show distinct methylation differences between the subtypes suggesting that in addition to the mutational effect also methylation changes modify the activity of these genes, especially in IDH. Cell cycle dysregulation via aberrant functions of cyclin dependent kinase inhibitors is a widespread mechanism in tumorigenesis leading to uncontrolled cell divisions [218] with impact for gliomas [219]. (ii) Other genes, especially the RTK's *EGFR* and *PDGFRA*, are dominated by expression differences especially between RTKI and RTKII samples due to chromosomal defects and mutations, as expected [62]. *MGMT* shows expression and methylation differences between IDH and the other subtypes [220].

(a) data landscape



(b) genomic characteristics of glioma subtypes

methylation subtype	genomic defects	expression subtype	genomic defects	
IDH (IDH_mut)	*IDH1*-mut	Proneural (PN)	*PDGFRA* CN+, *IDH1*-mut and wt, *TP53*-mut	
RTKI	*PDGFRA* (Chr 4+)			
RTKII	*EGFR* (Chr 7+), *CDKN2A/2B* (Chr 9-)	Classical (CL)	*EGFR* CN+, *CDKN2A/2B* CN-	Chr 7+: *EGFR*, *MET* Chr 10-: *PTEN*
MES		MES	*NF1* (Chr 17- & mut)	
normal brain (BR)		NOR		

(c) mapping of glioma key genes

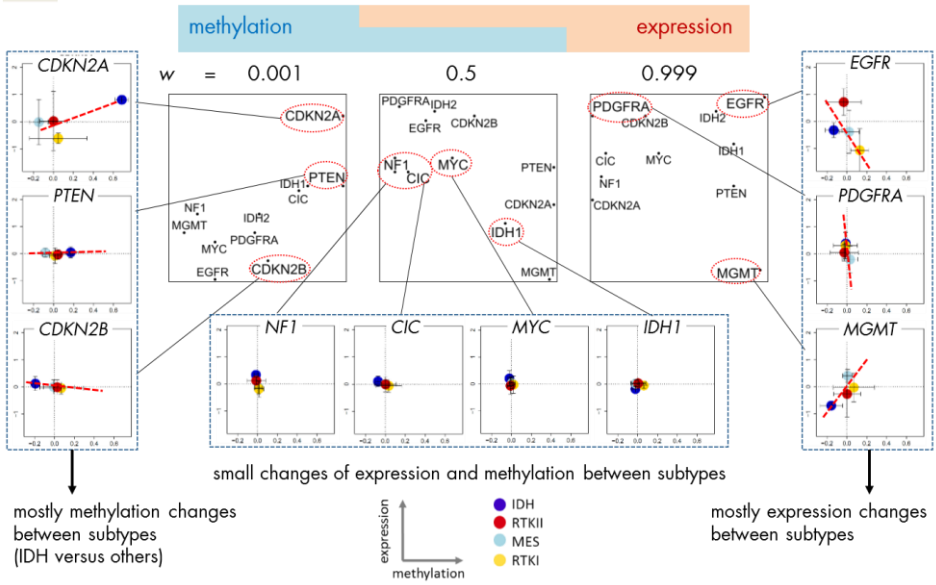


Figure 60: Molecular landscapes and key GBM genes. (**a**) Molecular landscapes, (**b**) genetic characteristics of the subtypes and (**c**) maps of key GBM genes.

(iii) A third group of key genes manifests only tiny expression and methylation changes between the subtypes, which however doesn't exclude their impact for gene regulation. For example, there is no doubt that mutation of *IDH1*, although only weakly affected in terms of gene expression and promoter methylation, has strong effect on genome wide methylation pattern in glioma. The *IDH1*-mutation in first instance doesn't change the activity of the gene but instead modifies the outcome of its enzymatic action, namely it leads to

the production of the metabolite 2-HG instead of α-KG where 2-HG inhibits the activity of chromatin modifying enzymes with consequences for global DNA methylation patterns (see below) [149,221]. Hence, *IDH1* mutation serves as the main driver for the formation of the IDH subtype without evident methylation and/or expression changes of this gene [222].

5.3.10 CHROMATIN STATES

Aberrant DNA methylation and expression levels in cancer are often related to alterations of the chromatin organization compared with healthy, non-neoplastic tissue [223]. We analyzed sets of genes assigned to distinct chromatin states in healthy brain (mf lobe) and NP cells taken from [55] (see section 2.2.3) to discover systematic changes of their expression and methylation levels between the glioma subtypes as possible indications for alterations of the chromatin states in glioma. We found one cluster enriched with poised and repressed chromatin states in the reference tissues showing strong anti-correlation between promoter methylation and gene expression in the glioma subtypes as indicated by the plus and minus signs in the heatmap in Figure 61. Note that the promoter methylation level in MES is close to that in healthy brain see section 5.2.1.2. IDH and RTKII subtypes show pronounced hypermethylation of genes in these chromatin states paralleled by their transcriptional repression. Hypermethylation of repressed (e.g. RepPC) and poised (TssP) promotors is a molecular hallmark of many cancer types [208] including B-cell lymphomas [60,113,224], CRC [22] and melanomas [225]. This deactivation can be assumed to 'suppress processes that suppress' tumor development and this way facilitate tumorigenesis. Compared with non-targets, genes repressed by PRC2-targets are more likely to show a promoter DNA hypermethylation pattern specific for cancer. This process supports a stem cell origin of cancer in which gene expression is long-term repressed resulting in a continuous self-renewal state of the cell, possibly causing malignant transformation [226].

This cluster further splits into two subclusters, c1 and c2, which differ in the mutual degree of methylation in IDH (larger in c1) and RTKII (larger in c2) GBMs and in the opposite trends of methylation in the RTKI subtype (hypermethylated in c1 and hypomethylated in c2) suggesting different modes of epigenetic deregulation of gene activities. The *IDH1*-mutation in IDH leads to inhibition of KDMs and DNDMs via metabolites of the TCA-cycle (see below) whereas in RTKII one suggests reprogramming of the energy metabolism via histone acetyltransferases [32]. Both subclusters also differ in the types of promoters (repressed in c1 and poised in c2) and in the reference tissue types (NPs in c1 and mf lobe in c2). Hence, molecular mechanisms of DNA methylation obviously specifically affect genes in poised and repressed states with impact for brain development. Here, developmental genes and poised promoters are more prone to hypermethylation in RTKII whereas repressed promoters and heterochromatin states of healthy brain are more strongly affected in IDH.

Gene SOM-maps of these chromatin states indicate a wide distribution of these genes despite their accumulation in distinct areas (see right part of Figure 61). Hence, the mean effect splits into modes that were identified as consensus modes above (Figure 58). For example, mode V enriches poised, repressed and PRC2 chromatin states in NP cells, which associate with hypermethylation in RTKII.

Other clusters in the heatmap in Figure 61 can be attributed to active chromatin states or heterochromatin with either dominating effects on expression or methylation in the different subtypes however without pronounced anti-correlation between both expression and methylation levels as in c1 and c2. These anti-concerted changes between both data entities in these clusters reflect mutual interactions between gene expression and the methylation level in their promoters and they obviously associate primarily with genes in poised and repressed chromatin.

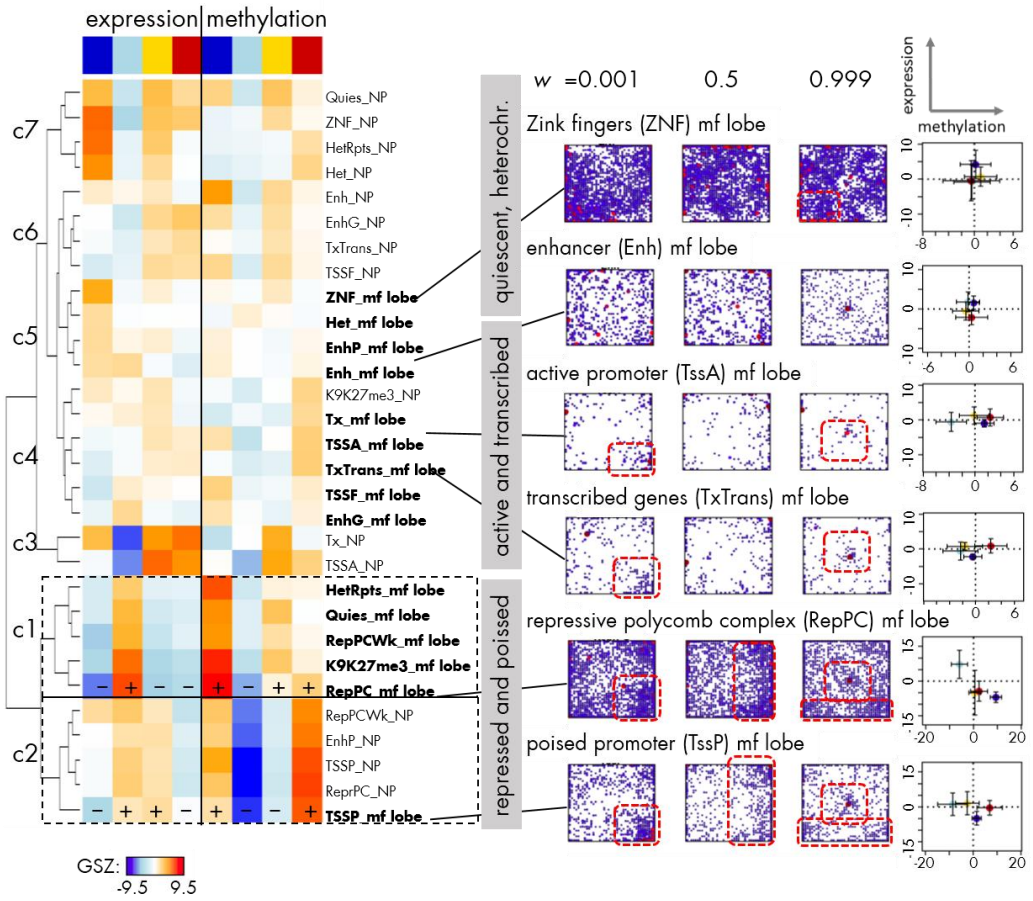


Figure 61: Mean expression and DNA promoter methylation of sets of genes referring to different chromatin states in neuronal progenitors (NP) and mid frontal lobe (mf lobe) taken from [55]. Chromatin states were defined in [122,123]. The right part of the figure shows gene set maps of selected chromatin states and the respective correlations between mean expression and methylation in units of the GSZ-score. The clusters c1 and c2 reveal anti-correlated expression and methylation levels in the subtypes as indicated by the '+' and '-' signs. Other clusters are dominated either by variation of the mean methylation (e.g. c4) or mean expression (e.g. c5 and c7) levels.

5.3.11 CHROMATIN MODIFIERS

Chromatin states and DNA methylation are regulated by a large battery of chromatin modifying enzymes whose transcriptional activity is expected to vary between the glioma subtypes. We analyzed DNA methylation and gene expression of more than 30 chromatin modifying enzymes either directly catalyzing DNA methylation or indirectly mediating DNA methylation and gene expression via methylation (KMT) or demethylation (KDM) of selected lysine side chains of histone H3. The heatmap in Figure 62a reveals two main clusters of enzymes with either up- or downregulated expression in the IDH subtype mostly showing anti-correlated promoter methylation and predominantly low expression in MES. Expression of many enzymes thus seems to be modulated by DNA methylation in their promoter region.

Mapping of the enzymes considered into the methylation and expression landscapes reveals that only a few of them locate in or near the characteristic ScoV-clusters accumulating genes with strongly anti-correlated expression and methylation profiles as discussed above (Figure 62b). Among them are, for example, the methyltransferases *DNMT3A* and *KMT6* (alias *EZH2*), ensuring de-novo DNA methylation and trimethylation of H3K27, respectively. Both mechanisms potentially support DNA hypermethylation and transcriptional deactivation where *KMT6* also contributes to the formation of PRC2-repressed chromatin states. Co-expression of *DNMT3A* and *KMT6* was also found in B-cell lymphoma (see section 4.3.3), which presumably reflects their interaction as enzymatic components of the PRC2 [227]. Note that activity of these enzymes is reduced in IDH-glioma because of suggested suppression of PRC2 function in IDH. We will discuss this phenomenon below. An antagonistic regulation mode (methylation down and expression up in IDH) includes KDMs that demethylate H3K9me2/me3 and H3K27me3. Both processes co-regulate with genes included in the spot-cluster 'M2'. It contains the oncogene *KDM4A*, overexpression of which associates with poor prognosis in many cancers [172].

The majority of enzymes however accumulates in an extended region of the map showing enhanced cell division activity paralleled by a weak total DNA methylation level in all subtypes (see also Figure 55 and Figure 59). In contrast, another region, which associates with immune response functions is largely depleted from enzymes meaning that neither expression nor methylation of modifying enzymes strongly co-regulates with genes associated with immunity-related function. An analogous asymmetric distribution of chromatin modifying enzymes was reported for gene activation patterns in lymphoma (see section 4.3.3). It was rationalized partly by the requirement of maintenance methylation of DNA and histone methylation marks after cell division and DNA replication. These processes require enhanced activities of these enzymes to re-establish the methylation state of DNA and histone side chains after cell division along the newly synthesized DNA. Moreover, it demonstrates that high transcriptional activity of these enzymes accompanies by a low DNA methylation level of their promoters. In summary, the enzyme machinery affecting DNA methylation shows aberrant expression and methylation changes between the glioma

subtypes, which in turn can be assumed to modify gene regulation via a multitude of feedback mechanisms.

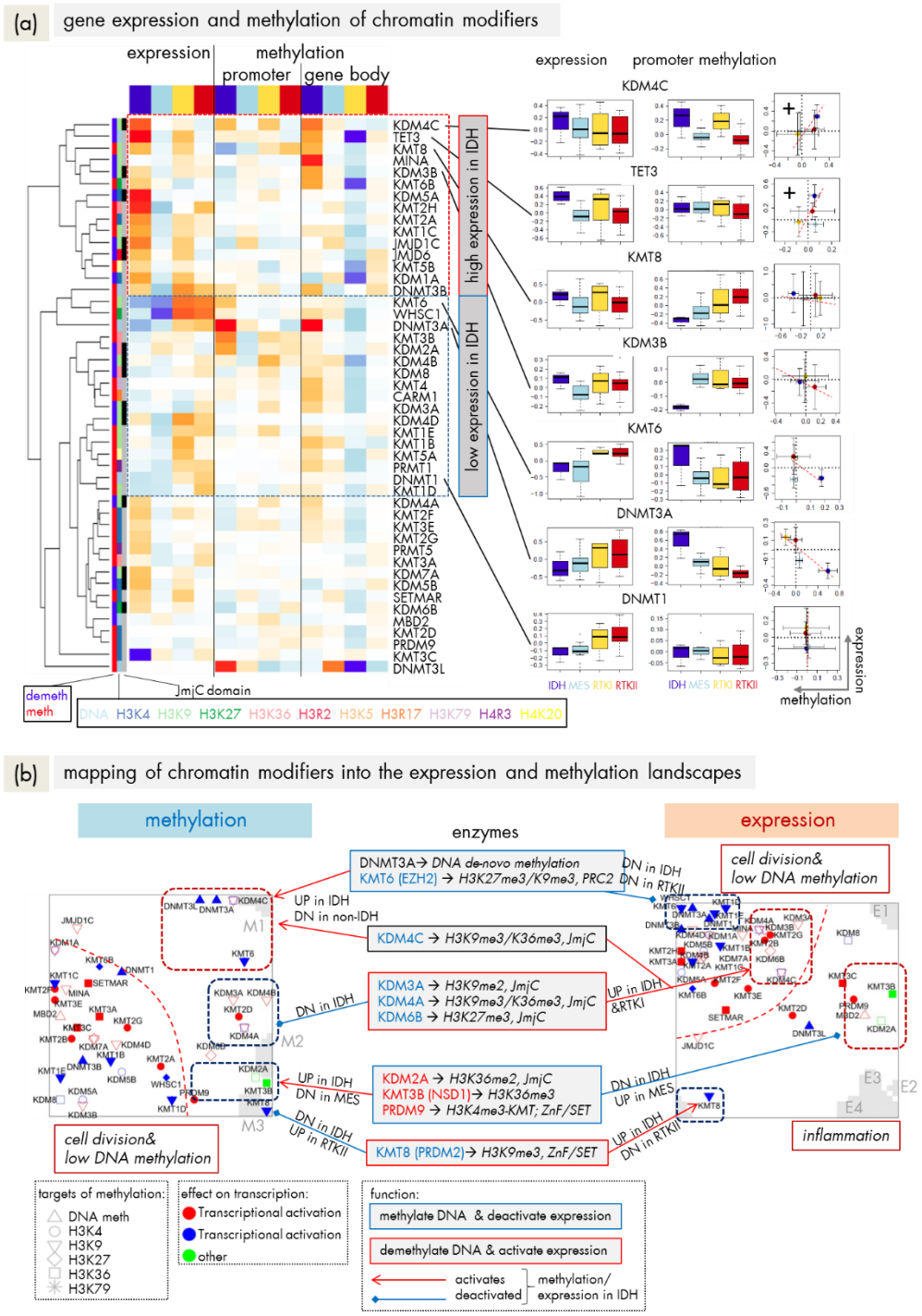


Figure 62: Expression and DNA-(promoter and gene body) methylation of genes encoding chromatin modifying enzymes: **(a)** The heatmap divides into clusters of enzymes over- and under-expressed in IDH showing also almost anti-correlated promoter methylation levels. For the enzymes *TET3* and *KDM4C* one finds positively correlated variations of expression and methylation. **(b)** SOM maps of the chromatin modifying enzymes: Only a few of them accumulate in or near the clusters of

genes with strong anti-correlated expression and methylation changes (rectangular frames). Most genes encoding chromatin modifying enzymes accumulate in the region with moderate changes of expression and methylation between the subtypes and they mostly associate with an increased cell cycle activity. The region that associated with inflammation is almost depleted with enzymes.

5.3.12 REGULATION OF DNA METHYLATION, CHROMATIN STATES AND GENE EXPRESSION

In the previous subsections we studied promoter DNA methylation and gene expression in the context of chromatin states in healthy neuronal tissues and of chromatin modifying enzymes, which together form essential ingredients of the molecular machinery ensuring epigenetic regulation of cellular programs. The scheme in Figure 63a illustrates the main interactions between these ingredients in a simplified fashion: Promoter methylation in glioma predominantly represses gene expression of the affected genes. DNA methylation marks are written or erased by DNMT and DNDM, respectively, whose activity thus adjusts the methylation level along the DNA. KMTs and KDMs adjust their methylation status and in consequence the activity of the affected genes according to the chromatin code [228] where transcriptional repression via H3K27me3 is linked with PRC2. In turn, the modification status of the histones also affects the DNA methylation often via interactions with DNMTs where activating histone modifications tend to suppress promoter methylation and *vice versa*. The methylation status of the histones largely determines the chromatin state and its transcriptional activity.

The dominant mechanism by which *IDH1* mutations are oncogenic in IDH subtype is the 2-HG mediated inhibition of JmjC KDMs and TET DNDMs [145] resulting in a shift of the methylation-demethylation equilibria towards hypermethylation of DNA and repressive methylation marks H3K9me3 and H3K27me3 [222]. The widespread disturbance of expression of the enzyme-machinery reported above indeed supports this but it suggests also additional options leading, e.g. to hypomethylation effects in IDH by repression of demethylases of activating histone marks. The mutation landscape of glioma revealed a large number of mutations in chromatin modifying enzymes, which however, except those in *IDH1*, *ATRX,* and partly *KMT2C* (alias *MLL3*) and *KMT2D* (alias *MLL2*), were found in only less than 1% of the cases and thus rarely can be seen as drivers of tumorigenesis [217]. Alternative mechanisms, e.g. via tumor-induced changes of cell activity and their coupling with the epigenetic machinery are candidates to promote tumorigenesis. For example, increased cell division activity in cancer requires adjustment of the maintenance machinery for methylation marks via activation of KMTs and DNMTs. Even small disturbances in this regulation possibly accumulate remarkable shifts in cellular programs during clonal evolution of tumor cells potentially leading to different cancer subtypes. Mechanisms of metabolic coupling via different intermediate products of the TCA-cycle inhibiting KDMs potentially enabling such shifts have been suggested previously [144,192]. In RTK tumors epi-

genetic dysregulation seems to associate with metabolic reprogramming, namely with aberrant activation of the enzyme *PKM2* [194], which plays an essential role in tumor metabolism and is involved in glycolysis to generate ATP and pyruvate [213].

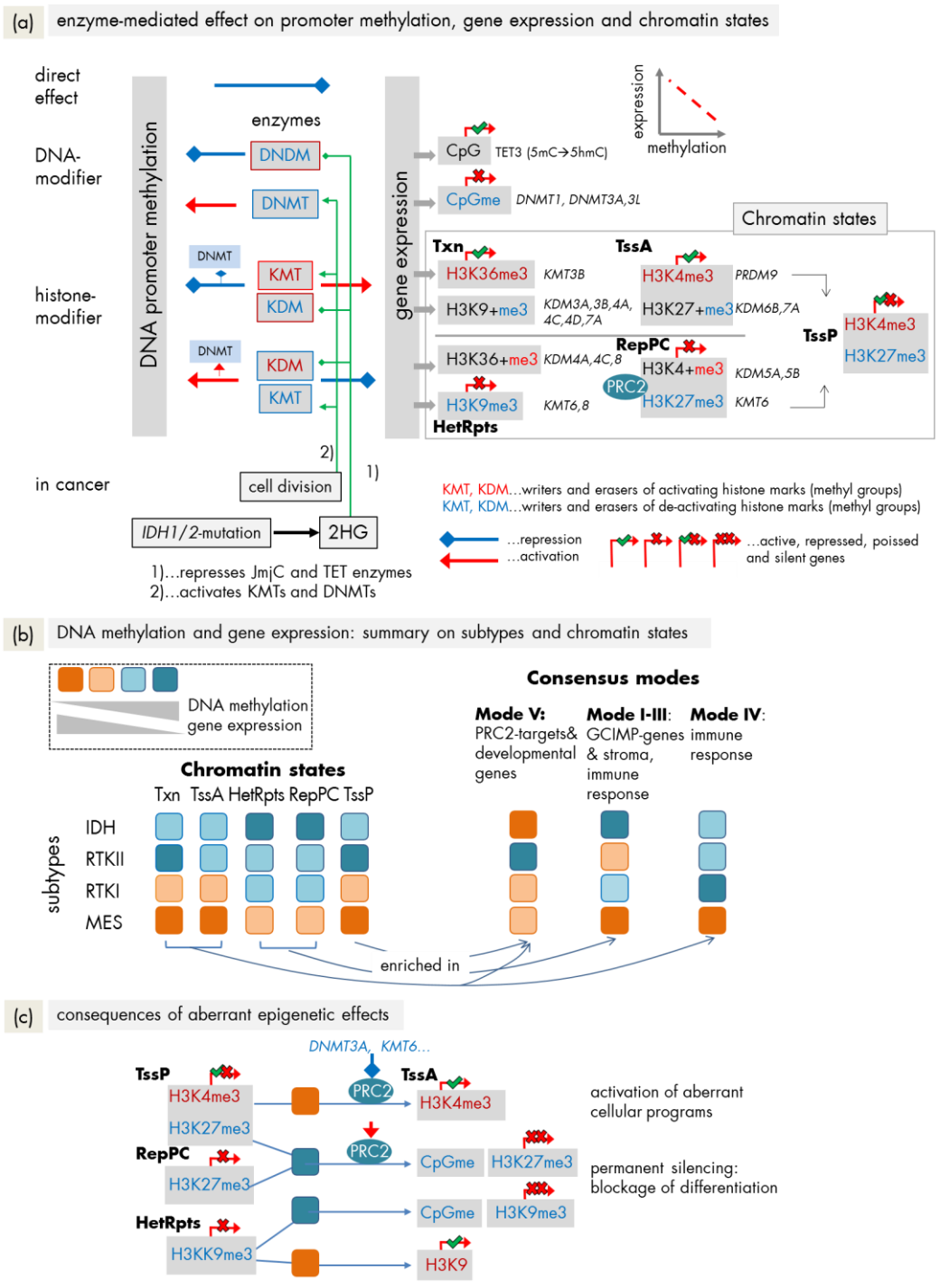


Figure 63: Co-regulation of DNA methylation, gene expression and chromatin states: **(a)** Scheme linking DNA-promoter methylation and gene expression with the activity of epigenetic enzymes, the methylation state of histone lysine side chains and associated chromatin states. Mutation of *IDH1/2* leads to inhibition of demethylation reactions via metabolic coupling and in consequence to enduring

silence of developmental genes. **(b)** Summary of DNA methylation and gene expression changes of genes from selected chromatin states in the GBM subtypes. Net methylation and expression levels of chromatin states in the subtypes are indicated by the color of the boxes where IDH are consistently hypermethylated in all chromatin states selected. The consensus modes enrich these states but partly show different patterns, e.g. RTKII are hypermethylated in mode V. **(c)** Epigenetic regulation leading to aberrant activation or silencing of genes.

Our SOM analysis revealed highest net DNA methylation levels throughout all chromatin states in IDH and lowest methylation in MES with intermediate levels in RTKI and RTKII (Figure 63b). Strongest hypermethylation is observed in repressed and poised states of IDH. The combination of repressive histone marks in these states with reduced levels of KDMs for these marks and also of DNDMs is expected to promote the shift of the reaction equilibrium towards DNA methylation. In contrast, active states are obviously less prone to hypermethylation because the activating histone marks inhibit DNA methylation.

We disentangled these overall patterns into selected consensus modes, which are characterized by concerted alterations of methylation and expression levels among the subtypes (Figure 58). Modes I-IV roughly agree with these overall trends (Figure 63b). In contrast, mode V shows inverse methylation and expression changes, namely strong hypermethylation in RTKII and hypomethylation in IDH. This mode is enriched in a subset of poised chromatin states including PRC2-targets and developmental genes, which suggests their transcriptional deactivation in RTKII and activation in IDH. Interestingly, we found increased promoter methylation and reduced expression of genes encoding the enzymes *DNMT3A* and *KMT6* in IDH, which suggests suppression of repression by PRC2 and thus activation of PRC2-targets in IDH and the opposite trend in RTKII as indeed observed. In other words, deactivation of PRC2 promoting enzymes by DNA hypermethylation of their promoters is expected to activate developmental genes targeted by PRC2, a mechanism that potentially shapes aberrant cellular programs in the emerging tumor cells (Figure 63c). Contrarily, high activity of PRC2 promoting enzymes in the other subtypes and especially in RTKII represses associated genes and, in combination with the hypermethylation of their promoters, will enhance their repression and lead to enduring silencing and the blockage of cellular-differentiation. In general, PRC2-targets are more likely to show a promoter DNA hypermethylation pattern specific for cancer compared with non-targets. This process supports a stem cell origin of cancer in which gene expression is long-term repressed resulting in a continuous self-renewal state of the cell, possibly causing malignant transformation [226].

5.3.13 GENE BODY DNA METHYLATION

We restricted DNA methylation analysis to the promoter region of the genes. Part of the enzymes such as *KDM4C* and *KDM3A* (alias *SETD2*)*/B* de-methylate H3K36me3, a histone modification required for transcriptional elongation, DNA-repair, and efficient nucleotide synthesis and DNA-replication [177,229]. It acts rather along the body of the genes and not in their promoter region. To account for this we calculated the integral

methylation level of all probed CpGs along the body of each gene. For chromatin modifying enzymes we found that the methylation of the gene body mostly differs from the methylation pattern of the respective promoter region (Figure 62a): Particularly, gene bodies in RTKI are consistently hypomethylated and in IDH almost consistently hypermethylated indicating that methylation patterns of the gene body are less diverse than that of the promoter region. Moreover, differential methylation of the gene body shows no pronounced anti-correlation with gene expression as found for promoter methylation in glioma. We mapped gene body methylation into the SOM calculated using promoter methylation values (see supplementary material of [215]). We found that gene body data only weakly maps onto the promoter methylation landscape thus revealing almost disjunctive methylation patterns. The SOM modulation method offers one option to study relations between them.

5.3.14 CONCLUSION

DNA methylation of CpGs in gene promoters and gene expression are mutually-dependent effects that both regulate activity of cellular programs. We presented a new method based on SOM machine learning that enables an integrative view on gene expression and DNA methylation data. The method 'portraits' the expression and methylation landscapes for each sample and cancer subtype and thus allows their visual inspection on a personalized and class-related basis. The relative contribution of each of both data entities can be tuned either to focus on expression or methylation landscapes or on a combination of both. We applied the method to gene expression and promoter methylation data of gliomas, a tumor entity, which classifies into a series of molecular subtypes differing in DNA methylation and gene expression as well. Expression and methylation landscapes were segmented into modules of co-expressed and co-methylated genes, which reflects underlying regulatory modes of cell activity. Expression and methylation modules are typically anti-correlated suggesting a common functional background. We also found modes of co-expressed genes without co-methylation effect and *vice versa*. We identified different modes of combined gene expression and DNA methylation changes between the subtypes, assigned their functional context in terms of activated cellular programs, and related them to chromatin states in healthy brain and to the expression of selected chromatin modifying enzymes with consequences for DNA methylation and gene expression. Interestingly, we found antagonistic methylation and gene expression changes between the IDH (*IDH1*-mut proneural) and RTKII (classical) subtypes, which affect predominantly poised and repressed chromatin states in healthy brain tissue. These effects deregulate developmental processes either by their blockage or by aberrant activation leading to inappropriate cellular functions. The examples chosen illustrated that integral analysis of gene expression, DNA methylation and, in final consequence, also genetic defects is required to disentangle molecular factors of complex diseases.

6 Summary and Conclusion

The fundamental subject of this thesis was to develop and to apply bioinformatics methods based on SOM machine learning in order to unravel molecular mechanisms underlying cancer in the specific case of lymphoma and glioblastoma with special focus on gene expression and epigenetic mechanisms affecting gene activities. Methodical challenges were hereby (i) the high diversity of cancer on the molecular level that requires appropriate computational methods for stratification and visual evaluation of the data landscapes for each individual case, for example to identify their individual specifics; (ii) the extraction of suited features that characterize the subtypes and the respective functional context requiring strategies beyond a case-control two group comparison and the evaluation and consideration of weak effect sizes; (iii) the integration of different data types such as transcriptome and (DNA-) methylome data and their joint analysis, to extract mutual associations as candidates for possible causal effects important in the context of genomic regulation and particularly of its dysfunction promoting tumorigenesis.

We demonstrated that our SOM portrayal method well meets these requirements and enables the detailed molecular characterization of cancer. Particularly, the method provides a holistic view on high-dimensional data collected in large-scale studies. The portraying method transforms the multitude of different modes inherent in a multidimensional data set into a two-dimensional map for each sample. This map can be simply 'read' by visual inspection revealing relevant clusters of co-regulated genes and their functional context by applying knowledge mining. Importantly, features were selected using the concept of 'spots' based on the assumption of co-regulation, which is more sensitive than the concept of maximum differential effect size. Furthermore those spots enable finer identification of functional modules. We demonstrated that SOM portrayal provides a general framework for analytic tasks such as feature selection, integration of concepts of molecular function and systems tracking with individual resolution. Furthermore, the method has proven its

value to detect contaminations, which we observed for gene expression studies. We presented a workflow that suggests a way to cope with contaminated samples caused by either inaccurate biopsies or variations in the sample preparation process. This method enables one to correct the data without need for exclusion of affected samples, which would mean a loss of valuable information.

We illustrated the potency of this method by analyzing the transcriptome and DNA-methylome landscapes of B-cell lymphoma and glioma. Particular objectives in these studies were: (i) the re-evaluation and characterization of molecular subtypes described previously and their mutual comparison across the cancer entities; (ii) the joint analysis of gene expression and DNA methylation data to compare classification schemes originating from the different data types and to analyze mutual associations between them; (iii) the study of potential modes of epigenetic regulation in the cancer subtypes under consideration of chromatin states, chromatin-modifying enzymes, DNA methylation and gene expression.

Regarding a big cohort study of mature aggressive B-cell lymphoma patients we reanalyzed previously published gene expression microarray data and proposed a more detailed molecular subtype classification of the samples. It turned out that each of the newly defined subtypes is characterized by different hallmarks of cancer, e.g., proliferation and high transcriptional and translational activity in mBL*, activated immune response and inflammation in non-mBL*, innate immunity in the intermediate A subtype and up-regulated expression of common cancer gene signatures in the intermediate B subtype. Furthermore we found that the survival prognosis for the two intermediate subtypes is even worse compared with the more homogeneous mBL* and non-mBL* subtypes.

The gene expression study of glioblastoma revealed that the GBM subtypes can be divided into two 'localized' and two 'intermediate' ones. The localized subtypes MES and PN were characterized by the antagonistic activation of processes related to immune response and cell division, respectively. In contrast, each of the 'intermediate' subtypes formed a heterogeneous continuum of expression states linking the 'localized' subtypes. In general, we observed a similar separation of subtypes related to inflammation and cell division in both B-cell lymphoma and glioma, which suggests a more generic nature of the underlying processes related to molecular hallmarks of cancer such as inflammation and cell division.

It is generally known that epigenetics regulates gene expression and that changes in the epigenome may lead to carcinogenesis. We examined epigenetic mechanisms driving tumorigenesis and particularly the possible role of chromatin remodeling in the transformations from healthy into malignant B-cells and from healthy brain tissue into malignant tissue. With the help of our integrative method based on correlation of both gene expression and DNA methylation data measured in lymphoma cohorts we found groups of genes showing characteristic expression and methylation signatures among the subtypes studied. These signatures are associated with epigenetic effects such as remodeling from transcriptionally inactive into active chromatin states, differential promoter methylation and the enrichment of targets of transcription factors such as *EZH2* and *SUZ12*.

We also applied our methods to DNA methylation data of brain cancer in order to extract sets of differentially methylated genes, to demonstrate their functional impact and to discuss their relevance in terms of glioma biology. We showed that the intrinsic structure of this methylation data is compatible with a multitude of signature sets extracted from independent cohorts including DNA methylation and gene expression data thus reflecting their common biological background. We showed that the specifics of biological functions of different glioma subtypes shape the content of these marker sets. In turn, including not only standard functional information according, e.g. to different gene ontology terms but also chromatin states of the healthy brain enabled us to study epigenetic mechanisms of glioma progression and the associated interplay between gene activity and methylation. The enrichment of DNA methylation signatures of other cancer entities in gliomas suggests general oncogenic mechanisms of aberrant DNA methylation.

Furthermore, we systematically studied the expression of more than 50 genes that code for histone and DNA (de)methylating enzymes in lymphomas and healthy controls. As a main result, we found that the expression levels of nearly all enzyme encoding genes become markedly disturbed in lymphomas, suggesting deregulation of large parts of the epigenetic machinery. We discussed the effect of DNA promoter methylation and of transcriptional activity in the context of mutated epigenetic modifiers such as *EZH2* and *MLL2*. Also for glioma the chromatin modifiers showed similar deregulation as observed for lymphomas. Therefore we concluded that the expression of many enzymes seems to be modulated by DNA methylation in their promoter region. We found that the enzyme machinery affecting DNA methylation shows aberrant expression and methylation changes between the glioma subtypes, which in turn can be assumed to modify gene regulation via a multitude of feed-back mechanisms.

Last but not least we presented an inter-omics approach demonstrated on matched glioblastoma cases of gene expression and promoter DNA methylation data. Both omics-profiles were trained together with various weighting factors. This method allows for direct comparison of matched individual single-omics portraits and for extraction of co-regulated gene modules showing concerted, anti-concerted or single-omics driven changes of expression and methylation.

A couple of open questions remain not addressed in this thesis. For example for both B-cell lymphoma and glioma we studied the expression of genes coding for epigenetic enzymes but the role of the enzymes should rather be estimated based on their chemical activities than only on their expression levels. Also the list of enzymes considered has to be extended beyond methylation to include also other histone modifications, such as acetylation, ubiquitylation, and others, for which one can expect also massive deregulation effects in cancer. In addition meta-analyses including different cancer entities are required to identify more ubiquitous and more specific modes of epigenetic regulation. Another point is that the understanding of molecular mechanisms of cancer requires integrative analysis of omics data including not only gene expression and DNA methylation but also, e.g., mutations, chromatin states and proteomic data. Although in this dissertation the method of joint

analysis was applied to transcriptomics and epigenomics data, in particular gene expression and DNA methylation, it is not restricted only to those but can be adapted to other omics. Nonetheless, this thesis contributed to the research of cancer entities like B-cell lymphoma and glioblastoma and previous findings about their underlying biology have not only been confirmed but also supplemented by our analyses.

7 Supplement

7.1 MATERIAL AND PREPROCESSING DETAILS

7.1.1 PCP GENE EXPRESSION DATA

Prostate cancer progression (PCP) microarray data are available under GEO accession number GSE6099 (104 non-commercial spotted Human 20K Hs6 arrays). The original study [58] addressed the molecular mechanisms associated with gene expression changes in the course of prostate cancer progression using laser capture microdissection by means of 84 samples from 44 individuals. The samples used were assigned to five stages of cancer progression ranging from benign prostatic hyperplasia (BHP, 22 samples) and prostatic intraepithelial neoplasia (PIN, 13) to low-grade (PCA_low, Gleason score 3, 12 samples), high-grade (PCA_high, Gleason score 4–5, 20 samples), and metastatic (MET, 17) prostate cancer.

7.1.2 LYMPHOMA GENE EXPRESSION DATA

Microarray data of lymphoma are available under GEO accession number GSE4475 (data from 221 Affymetrix HT HG-U133A arrays). This study used biopsy specimens of mature aggressive B-cell lymphoma with a tumor content of at least 70 percent. The classification of lymphoma samples into different subtypes is used as provided by Hummel et al. [59]: Of all 221 lymphomas, 44 were assigned to the mBL (molecular Burkitt's lymphoma) signature and 129 to non-mBL signature. 48 cases form an intermediate group, representing the transition zone between the mBL and non-mBL groups. The expression data was preprocessed as given in section 3.2.

7.1.3 LYMPHOMA DNA METHYLATION AND GENE EXPRESSION

DNA methylation data

Microarray-derived DNA methylation rates of 1,410 CpGs (GoldenGate Methylation Cancer Panel I; Illumina, San Diego, CA) of in total 133 samples obtained from hematological neoplasms and reference systems were taken from [60]. The CpGs were located in the range from -1500 bp to +500 bp around the TSS of 768 genes thus serving as markers for their promoter methylation. Methylation data was given in units of beta values estimating the level of methylation between values of zero (no methylation) and unity (full methylation) for each promoter. Differential methylation defines the difference between beta values of two states, e.g. between lymphoma and healthy B-cells, where hyper- and hypomethylation assigns positive and negative differences (delta beta values), respectively. Integral differential methylation was calculated as mean differential methylation separately averaged over all positive and negative delta beta values. Please take into account that for SOM analysis of differential methylation (DmetSOM) we used centralized methylation data, which are calculated as the difference between the beta value of a given promoter in a given sample and its mean value averaged over all samples studied.

For B-cells we find a bimodal shape of the frequency distribution of β values among the genes studied with maxima near zero (completely de-methylated CpG sites) and unity (completely methylated, see Figure S 1a). The respective distributions of β values in lymphoma are characterized by a wide loss of this bimodality where especially the fraction of highly methylated genes with β values near unity markedly decreases. Accordingly, the distributions of β value alterations of the genes in the different systems compared with their methylation in B-cells are tailed to both, positive and negative values reflecting hypo- and hypermethylation of the respective genes (Figure S 1b).

The integral hyper- and hypomethylation of all genes considered reveals the progressively increasing disturbance of DNA methylation in lymphoma being largest in DLBCL and IntL, but being relatively small in MM, FL, MCL and also mBL (Figure S 1c). This trend agrees with the results of previous studies reporting the gain of epigenetic heterogeneity (in terms of differential methylation with respect to the reference state of healthy B-cells) with progressive aggressiveness of lymphoma being largest in DLBCL [230,231]. Except for MCL, we find a global hypermethylation of the genes in lymphoma compared with B-cells (Figure S 1d). On the other hand, the variance of β values in each of the samples strongly decreases in lymphoma mainly due to the decrease or even loss of bimodality reported above (Figure S 1e). In summary, methylation changes in lymphoma comprise both, hyper- and hypomethylation effects leading to a loss of bimodality of promoter methylation with maxima at low and high β values and to more balanced methylation landscapes, where promoter regions tend to become methylated on intermediate β-levels.

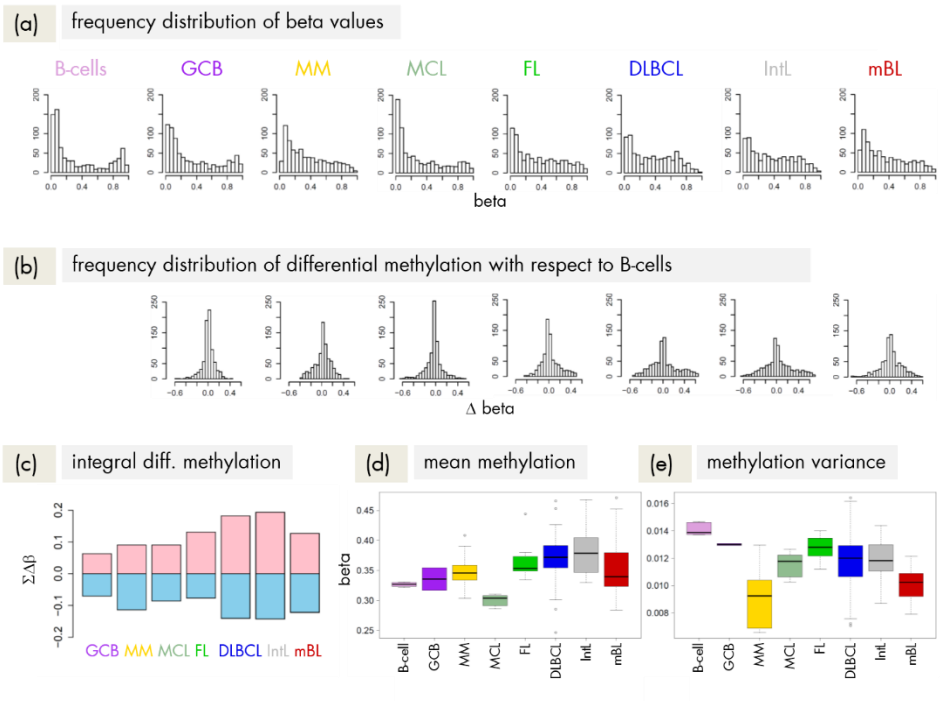


Figure S 1: DNA methylation summary characteristics of lymphoma and of healthy B- and GCB-cells. (**a**) The frequency distribution of the promoter methylation β values of B-cells shows two maxima referring to almost not- and completely methylated promoters, respectively. (**b**) and (**c**) The distributions of β values lose this bimodality to a large degree in lymphoma, where weakly and intermediately methylated genes become hypermethylated and highly methylated genes become hypomethylated compared with healthy B-cells. (**d**) The total methylation level increases and (**e**) the variability of methylation among the genes in each of the samples decreases.

Gene expression data

Expression data were taken from the MMML (Molecular mechanisms of malignant lymphoma) cohort described in [59] comprising 936 samples. Lymphoma samples were classified into five molecular subtypes as described above and [122]: mBL (85 samples), non-mBL (287), IntL (307), FL (121), and B-cell-like lymphoma (BCL, 64). According to pathological diagnosis, the molecular subtypes refer predominantly to BL (mBL), DLBCL (non-mBL) and MM (BCL). Further the cohort contains B-cells (17), GCB-cells (13), a lymphoma cell line (32) and tonsils (10) as reference. The microarray expression data (Affymetrix HT HG-U133A) were processed as described previously (see section 3.2). The B-cells subsume naïve pre- and mature post-GCB-cells, which show virtually indistinguishable gene expression patterns. The GCB-cells are centroblasts with strongly activated proliferative cellular programs.

7.1.4 LYMPHOMA GENE EXPRESSION DATA FOR PORTRAYAL OF CHROMATIN MODIFIERS

A subgroup (632 samples) of the publically available gene expression data (GEO accession numbers GSE4475, GSE10172, GSE22470, GSE48184, GSE43677) considered in section 4.2 were taken [59]. The cohort contains 62 mBL (BL), 204 DLBCL (non-mBL), 255 IntL, 3 FL, 36 BCL samples and 17 healthy B-cells, 13 GCB-cells, 32 lymphoma cancer cell lines and 10 tonsil samples as control.

7.1.5 GBM GENE EXPRESSION DATA

A public available cancer data set of a patient cohort study regarding glioblastoma multiforme (GBM) was analyzed. Microarray data are available on 'The Cancer Genome Atlas' (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/). We downloaded level 1 data of 153 GBM and 10 normal brain tissue specimen hybridized on Affymetrix HT HG-U133A arrays comprising raw intensities of 22.777 single genes. We used the classification of tumor subtypes given in [61]: The samples were assigned to Mesenchymal (MES, 50 samples), Proneural (PN, 45), Neural (NL, 26), Classical (CL, 32) GBM-subtypes and to normal healthy brain (11) for comparison. The latter specimens were taken from adjacent brain tissue of GBM patients. The expression data was preprocessed as given in section 3.2.

7.1.6 GBM DNA METHYLATION DATA

DNA Methylation data

DNA methylation data of 136 GBM and 6 control samples were taken from ref. [60,62] (available under GEO Series accession number GSE36278). The data refer to pediatric and adult GBM and to non-neoplastic cerebellum specimen as controls (Table S 1). GBM samples were classified according to the methylation clusters identified in [62]. Accordingly, the pediatric GBM split into two subtypes carrying mutations of the *H3F3A* gene, which affect two different amino acids of histone H3.3, namely G34 or K27, respectively. The adult GBM were classified into four subtypes labeled according to correlations with genetic defects. These genetic hallmarks constitute mutations of the *IDH1* gene ('IDH' subtype) and focal copy number (CN) amplifications of the *PDGFRA* ('RTKI' subtype) or *EGFR* ('RTKII' subtype) gene both coding receptor tyrosine kinases (RTK). The RTKII cases are called 'classical' because they enrich combined gain of CNs at Chr 7 and loss of CNs at Chr 10 both representing a hallmark of *IDH1* wild type GBM [62]. The 'mesenchymal' subtype shows a lower incidence of GBM typical CN alterations.

Microarray-derived DNA methylation data (Illumina HumanMethylation450 BeadChip) of the 136 GBM and 6 control samples were taken in terms of β values of 485,512 CpGs. Methylation levels were estimated in a gene centric way by averaging the CpG-related β

values over genomic regions of the promoters of each gene ranging from 1500 bp upstream the transcription start site (TSS) to the TSS (Figure S 2). The methylation data was further preprocessed as given in section 3.2.

Table S 1: DNA methylation data set (Sturm et al. [62]).

subtype	n		genetic hallmark[1]	expression subtype[2]
adult	2	control		
fetus	4	control		
MES(enchymal)	36	adult GBM		mesenchymal
RTKII (classical)	22	adult GBM	*CDKN2A* (CN loss), *EGFR* (CN amplification)	classical
RTKI (PDGFRA)	23	adult GBM	*PDGFRA* (CN amplification)	
IDH	19	adult GBM	*IDH1* (mut)	proneural
G34	18	pediatric GBM	*H3F3A*/ G34 (mut)	
K27	18	pediatric GBM	*H3F3A*/ K27 (mut)	

[1] See, e.g., [194] for an overview.
[2] According to [61].

On average, CpG-related β value reveal a smoothly decaying methylation level upstream of the TSS of the genes and relatively noisy methylation in their first exon (Figure S 2a). The frequency distribution of gene centric β values shows a typical bimodal shape with maxima near zero (completely de-methylated CpG sites) and unity (completely methylated CpG sites, see Figure S 2b). The distribution of the IDH-subtype clearly reveals a trend towards global hypermethylation: The fraction of weakly methylated genes decreases while the fraction of highly methylated genes increases compared with the distributions in the healthy controls. On the other hand the distribution of the G34-subtype shows the opposite effect and thus a trend towards global hypomethylation (see the arrows in Figure S 2b).

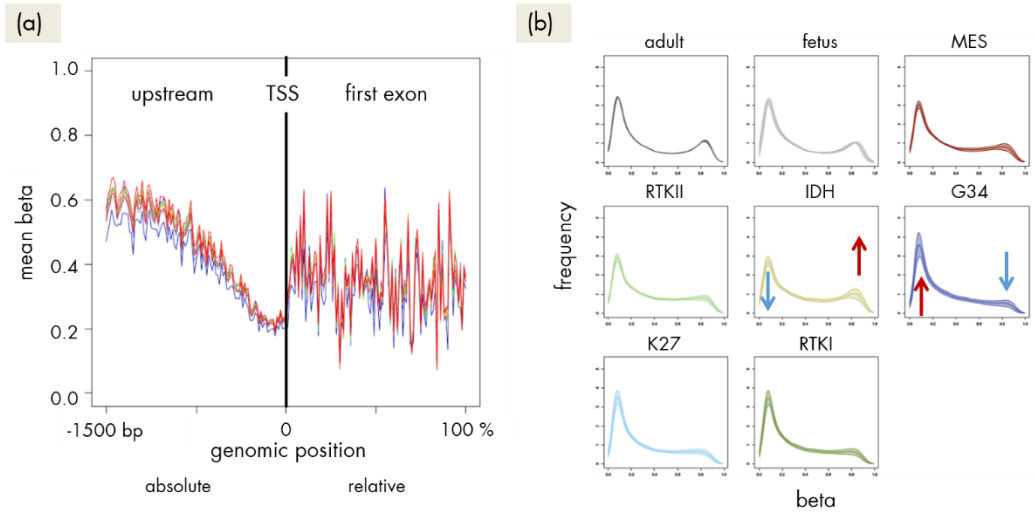


Figure S 2: Global β methylation characteristics: (**a**) Mean methylation level as a function of the genomic position relative to the TSS. CpG-β values were averaged over all genes for each subtype

(the colors were assigned in b); (**b**) frequency distribution of β values for the GBM subtypes and controls. The arrows serve as a guide for the eye to indicate methylation changes leading to global hyper- or hypomethylation in IDH- and G34-type GBM compared with healthy controls.

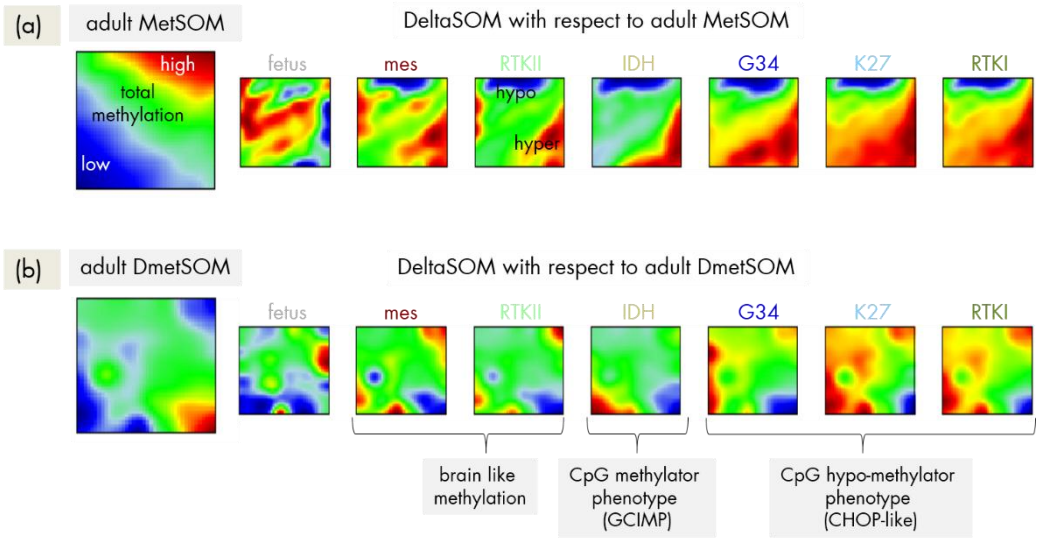


Figure S 3: Differential methylation analysis with respect to adult healthy brain. (**a**) Difference MetSOM portraits (metagene-methylation data are subtracted pixelwise) show subtype-specific hypermethylation in GBM. (**b**) Difference DmetSOM portraits reveal global hyper- and hypomethylation in spots 'F' and 'A1', respectively.

Gene expression data

Three expression data sets were used to establish associations with methylation data (see Table S 2). Microarray expression data of 30 matched samples and 3 unmatched fetal controls were taken from [62]. They comprise the same subtypes as the methylation data. A second set of expression data was taken from [61] and processed and analyzed previously (see section 5.1 and [69]). This data comprises healthy brain, mesenchymal, classical, proneural and neural GBM, which were matched with the classes of the methylation data. The third data set was taken from [63]. It consists of GBM with mesenchymal, classical, proneural with IDH1/2-mutational and proneural with IDH1/2-wild type characteristics.

Table S 2: Gene expression data of GBM.

methylation classes	Sturm et al. [62] matched samples	Hopp et al.[61] matched classes	Reifenberger et al. [63] matched classes
adult		healthy (n=10)	proneural *IDH1*-wt (n=14)
fetal	fetal (n=3)		
MES	mesenchymal (5)	mesenchymal (50)	mesenchymal (21)
RTKII	RTKII (3)	classical (32)	classical (23)
RTKI	RTKI (6)		
IDH	IDH (7)	proneural (45)	proneural *IDH1*-mut (12)
G34	G43 (4)		
K27	K27 (5)		

7.1.7 GBM MATCHED GENE EXPRESSION AND DNA METHYLATION DATA

Gene expression data

Microarray based gene expression (Affymetrix HT HG-U133A arrays) and DNA methylation (Illumina 450K arrays) data are available on 'The Cancer Genome Atlas' (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/). We downloaded data of glioblastoma multiforme (GBM) batch 111 containing specimen of 39 patients. We used the classification of tumor subtypes given in [62]: According to distinct DNA methylation clusters the samples were assigned to Mesenchymal (MES, 16 samples), RTKI 'PDGFRA' (RTKI, 4), RTKII 'Classic' (RTKII, 16), and IDH (3) molecular GBM subtypes. We used level 1 (raw data) gene expression data.

DNA Methylation data

We used level 3 CpG-related DNA methylation data (β values) matched to the same patients as considered for gene expression data. CpG DNA methylation data were mapped to the promoter region of each gene ranging from 2kb upstream up to 200bp downstream of the transcription start site (TSS) of each gene using RefSeq mRNA annotation and averaged to get one methylation β value for each gene promoter available.

7.2 SUPPORTING MAPS AND METAGENE VARIABILITY

Method

Additional information, such as the population (number of single genes per metagene mini-cluster) and the variance of metagene expression profiles can be visualized using the same mosaic structure as in the expression and methylation portraits. The additional information is then color coded using proper scales. For example the variance map visualizes the variance of the metagenes in each of the tiles,

$$var_k = \frac{1}{M-1}\sum_j (\Delta e_{kj} - \Delta e_{k.})^2 = \frac{1}{M-1}\sum_j \Delta e_{kj}{}^2 \qquad \textbf{Eq.(11)}$$

with $\Delta e_{k.} = 0$. We also calculate the orthogonal variability of the metagene expression landscape of each SOM image,

$$var_j = \sum_k \frac{1}{K-1}\sum_k (\Delta e_{kj} - \Delta e_{.j})^2 \qquad \textbf{Eq.(12)}$$

where $\Delta e_{.j} = 0$ is the mean differential expression averaged over all metagenes of sample *j*.

Example

SOM-machine learning scales the difference between the expression profiles of adjacent metagenes inversely to their population, *i.e.*, adjacent metagene profiles become more similar for highly populated metagenes. This way the method tends to distribute the single genes over as much as possible tiles.

The population map of PCP shown in Figure S 4a reveals that the single genes distribute inhomogeneously among the tiles of the mosaic. Highly populated metagenes (see yellow and red tiles) predominantly group along the edges of the map whereas only a few genes were distributed to the central area. Tiles in the central area refer to genes with virtually invariant expression in all samples studied. These invariant genes give rise to the dark blue spot in the central area of the variance map (Figure S 4b). Both, invariant and empty metagenes carry essentially no specific information as classification markers in transcriptional profiling. Hence, the tiles occupied by empty and invariant genes form regions not suited for differential expression analysis between the cancer progression stages studied.

The more variant and higher populated metagenes reveal an underlying spot-like pattern preferentially along the boundaries of the map (red areas), which agrees with the over- and underexpression spots detected in the SOM mosaics of individual samples.

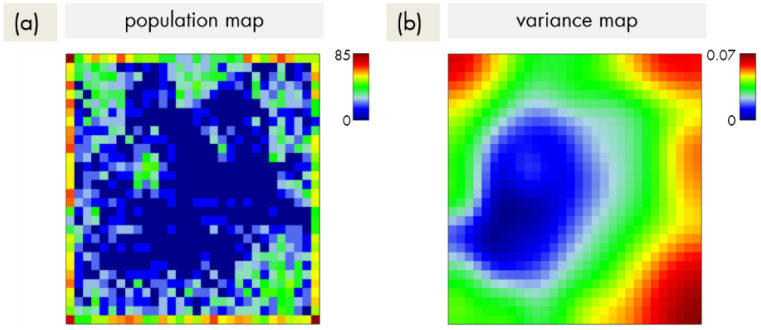


Figure S 4: Supporting maps characterizing the SOM trained for PCP: **(a)** The population map visualizes the number of single genes per metagene cluster. Highly populated metagenes accumulate along the edges (red tiles). **(b)** The metagene variance map color codes the variance of the metagene profiles. Virtually invariant metagene profiles form the central blue spot whereas highly variant ones are found in the peripheral regions of the map.

7.3 CONSENSUS CLUSTERING OF B-CELL LYMPHOMA

Consensus clustering aims at reaching a consensus on the number of classes in the data and at judging reliability of the class assignment of the samples. We applied the R-package 'ConsensusClusterPlus' [232] for portioning the samples into *c* classes using hierarchical clustering with *c* ranging from two to six. For each *c*, one obtains a consensus matrix, reflecting the fraction of common class memberships for all pairwise combinations of samples estimated in a series of resampling runs (details are given in [95]).

Figure S 5a – c shows the heatmaps of the consensus matrix for two to four classes, respectively. Pairs of samples, robustly assigned to the same cluster, accumulate within one of the blue squares along the diagonal of the heatmap. The two-class approach basically divides the samples into an mBL-like and a non-mBL-like cluster (Figure S 5a). The three-class approach essentially splits the samples into the mBL/intermediate/non-mBL subtype structure as proposed in [59] (Figure S 5b). The four-class consensus clustering resembles our new subtype classification with the two intermediate subtypes (Figure S 5c). The five- and six-cluster approaches virtually do not change this result: The additional fifth and sixth clusters collect only one and three outlier samples, respectively (data not shown).

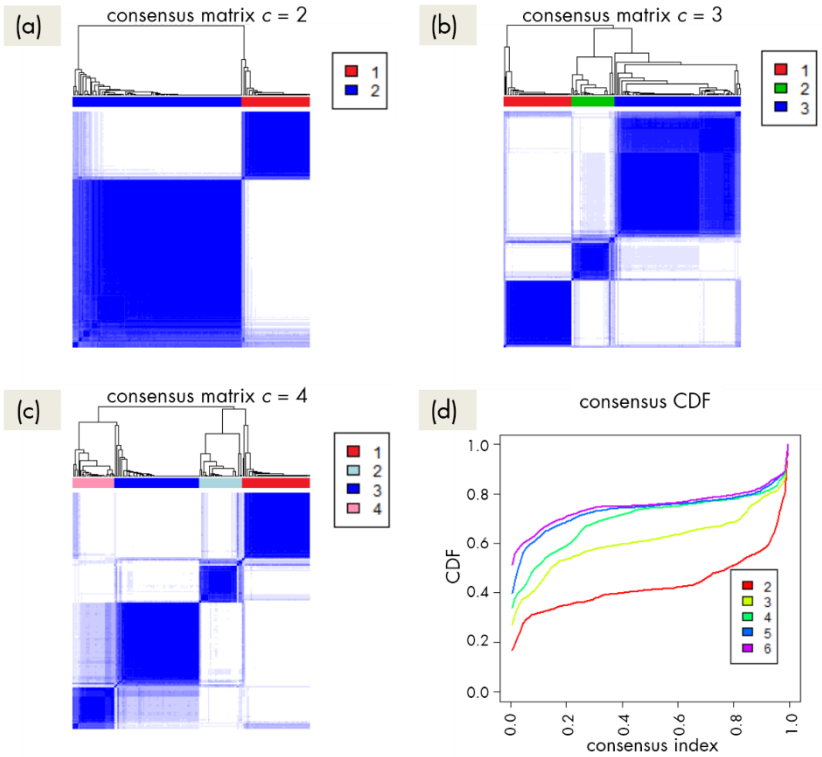


Figure S 5: Consensus clustering: (**a**) – (**c**) Cluster-heatmaps of the consensus matrices for class numbers ranging from two to four, respectively. Pairs of samples frequently found in one joint class accumulate in the blue regions along the diagonal of the map. (**d**) Cumulative distribution function (CDF) for class numbers ranging from two to six.

The cumulative distribution function (CDF) aggregates the consensus values up to a certain fractional co-occurrence of sample pairs. The CDF thus reflects the 'degree of heterogeneity' of a consensus matrix using one curve such that clusterings with different c can be directly compared with the purpose to identify the optimal class number [95]. The incremental change between CDF curves with increasing c serves as a measure to judge whether increasing the class number leads to a marked increase of cluster's stability or not. The obtained CDFs in Figure S 5d support the four-class approach: The CDF converges for $c > 3$ showing only small incremental changes with further increasing c. Note that the increment between $c = 4$ and 5 is caused by a single-sample cluster.

7.4 PHENOTYPIC CHARACTERIZATION OF NEW LYMPHOMA SUBTYPES

Table S 3: Phenotypic and molecular characterization of the four new subtypes (data taken from [96]). Percentages refer to the total number of samples. Parameters are not available for all samples. *P*-values are calculated using Fisher's exact test.

Characteristic			Lymphoma subtype				*p*-value
			mBL*	intermediate A	intermediate B	non-mBL*	
Total	number of patients	221	62 (28%)	42 (19%)	44 (20%)	73 (33%)	
Age	<20 y	32 (14%)	26 (42%)	0 (0%)	1 (2%)	5 (7%)	<0.001
	21–65 y	92 (42%)	27 (44%)	14 (33%)	22 (50%)	29 (40%)	
	>66 y	95 (43%)	9 (15%)	27 (64%)	20 (45%)	39 (53%)	
Gender	male	127 (57%)	40 (65%)	26 (62%)	23 (52%)	38 (52%)	0.44
	female	91 (41%)	22 (35%)	15 (36%)	20 (45%)	34 (47%)	
Diagnosis	BL	15 (7%)	15 (24%)	0 (0%)	0 (0%)	0 (0%)	<0.001
	Atypical BL	20 (9%)	16 (26%)	3 (7%)	0 (0%)	1 (1%)	
	DLBCL	164 (74%)	24 (39%)	37 (88%)	38 (86%)	65 (89%)	
	Mature aggressive BL, unclassifiable	18 (8%)	5 (8%)	2 (5%)	5 (11%)	6 (8%)	
Ann Arbor stage	I or II	72 (33%)	25 (40%)	9 (21%)	15 (34%)	23 (32%)	0.37
	III or IV	82 (37%)	19 (31%)	15 (36%)	22 (50%)	26 (36%)	
Response to treatment	Complete remission	68 (31%)	27 (44%)	8 (19%)	10 (23%)	23 (32%)	0.40
	Complete remission, unconfirmed	18 (8%)	4 (6%)	2 (5%)	6 (14%)	6 (8%)	
	No change	2 (1%)	0 (0%)	0 (0%)	1 (2%)	1 (1%)	
	Partial response	16 (7%)	1 (2%)	3 (7%)	5 (11%)	7 (10%)	
	Progress	24 (11%)	7 (11%)	4 (10%)	7 (16%)	6 (8%)	
Molecular classification Hummel [59]	mBL	44 (20%)	44 (71%)	0 (0%)	0 (0%)	0 (0%)	<0.001
	intermediate	48 (22%)	18 (29%)	11 (26%)	10 (23%)	9 (12%)	
	non-mBL	129 (58%)	0 (0%)	31 (74%)	34 (77%)	64 (88%)	
GCB-ABC classification Wright [98]	ABC	58 (26%)	2 (3%)	26 (62%)	15 (34%)	15 (21%)	<0.001
	GCB	120 (54%)	53 (85%)	10 (24%)	18 (41%)	39 (53%)	
	unclassified	43 (19%)	7 (11%)	6 (14%)	11 (25%)	19 (26%)	
Translocations *MYC* translocation	IG-MYC	60 (27%)	49 (79%)	1 (2%)	6 (14%)	4 (5%)	<0.001
	non-IG-MYC	15 (7%)	6 (10%)	5 (12%)	2 (5%)	2 (3%)	
	neg	144 (65%)	7 (11%)	36 (86%)	35 (80%)	66 (90%)	
BCL6 Break	pos	37 (17%)	2 (3%)	9 (21%)	11 (25%)	15 (21%)	0.002
	neg	179 (81%)	59 (95%)	32 (76%)	31 (70%)	57 (78%)	
IGH Break	pos	115 (52%)	53 (85%)	11 (26%)	23 (52%)	28 (38%)	<0.001
	neg	103 (47%)	9 (15%)	30 (71%)	20 (45%)	44 (60%)	
t(14;18) translocation	pos	25 (11%)	5 (8%)	2 (5%)	6 (14%)	12 (16%)	0.19
	neg	193 (87%)	57 (92%)	40 (95%)	37 (84%)	59 (81%)	
Immunohistochemistry	*CD10* low	114 (52%)	3 (5%)	33 (79%)	26 (59%)	52 (71%)	<0.001
	high	96 (43%)	56 (90%)	6 (14%)	14 (32%)	20 (27%)	
	BCL2 low	62 (28%)	38 (61%)	2 (5%)	7 (16%)	15 (21%)	<0.001
	high	153 (69%)	22 (35%)	39 (93%)	35 (80%)	57 (78%)	
	BCL6 low	34 (15%)	5 (8%)	9 (21%)	7 (16%)	13 (18%)	0.21
	high	168 (76%)	52 (84%)	29 (69%)	32 (73%)	55 (75%)	
	MUM1 low	66 (30%)	29 (47%)	7 (17%)	8 (18%)	22 (30%)	0.001
	high	139 (63%)	27 (44%)	33 (79%)	32 (73%)	47 (64%)	
	KI67 low	125 (57%)	17 (27%)	26 (62%)	26 (59%)	56 (77%)	<0.001
	high	89 (40%)	44 (71%)	15 (36%)	14 (32%)	16 (22%)	

7.5 MARKER SETS OF B-CELL LYMPHOMAS AND COLORECTAL CANCER DIFFERENTIATE ALSO BETWEEN GLIOMA CLASSES

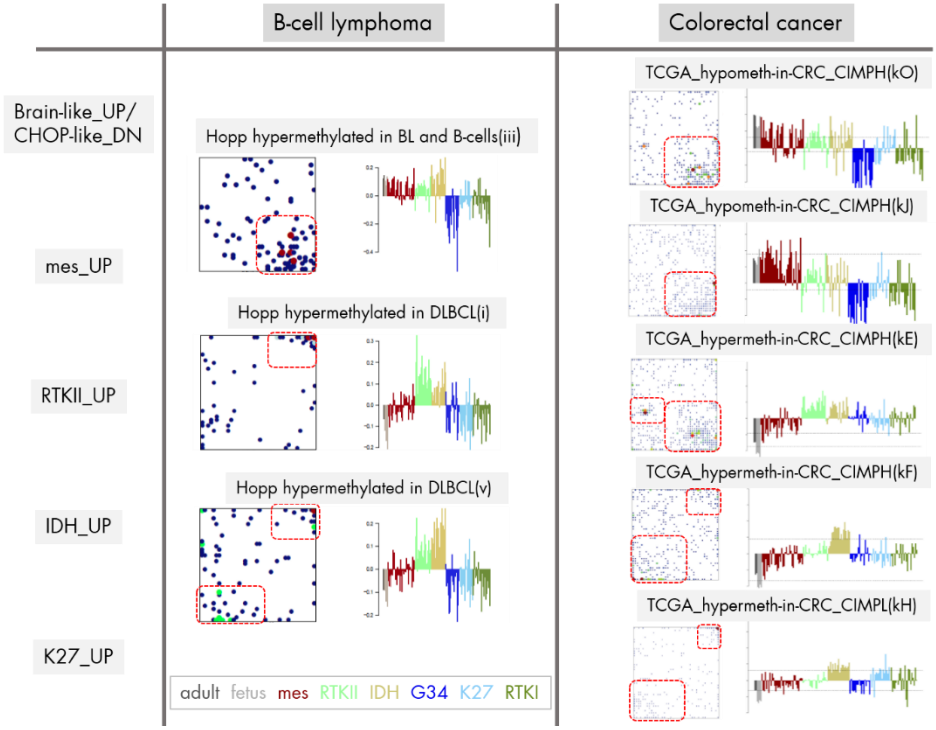


Figure S 6: Mapping of methylation-signature gene sets of B-cell lymphoma and of colorectal cancer into the DmetSOM of glioma. The gene sets were determined using SOM spot analysis in recent studies on DNA methylation data in [113] and [205], respectively. The red frames indicate regions of increased local densities of genes. The profiles indicate subtype-specific hyper- (and hypomethylation) in glioma.

7.6 CHROMATIN STATES IN LYMPHOMAS

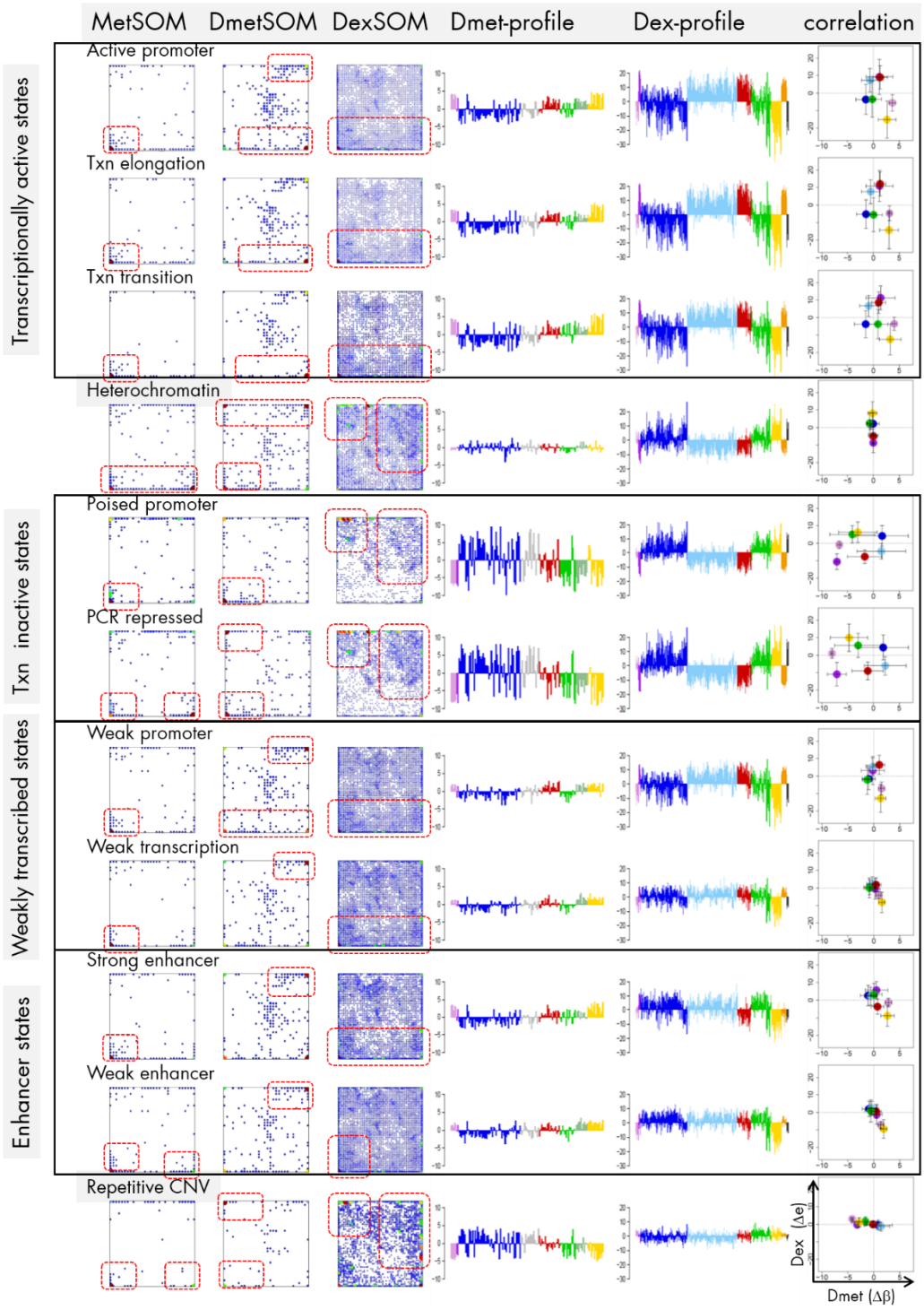


Figure S 7: Mapping of genes referring to different chromatin states as determined in lymphoblastoid cells using ChIP-Seq and a Hidden Markov model [122,123]. According to the methylation and expression characteristics in the lymphoma data set these gene sets can be grouped into five types.

7.7 FUNCTION MINING OF GBM METHYLATION SPOT MODULES

Table S 4: Sets of methylation marker genes and their functional context.

Spot	UP	DN	Functional context: Enriched gene sets[1]	top 10 genes[2]
A	MES		olfactory receptor activity (MF); G-protein coupled receptor signaling pathway (BP), neurological systems process (BP), colon cancer: CIMP_methylation_DN, CIMP_expression_UP [205]	*ANGPTL1, BCAN, LAMA4, APOC1, TUT1, FADS1, OR4C46, OR11H6, CDH19, GDF5OS,*
B		G34	extracellular region (CC); keratin filament (CC); colon cancer: CIMP_methylation_DN [205]	*PRR33, VIP, FGF17, EMB, USP44, CCR7, HOXB1, LHX5, PRKCD, C1orf64*
C	IDH		hallmark epithelial mesenchymal transition (cancer), GCIMP_signature genes: silenced_by_methylation [46]; colon cancer: CIMP_methylation_UP [205]; Christensen_ methylated_in_LGG [105]; Benporath_H3K27me3_ in_ES [208]; brain development (BP), Meissner_brain_HCP_with-H3K4me3_and_ H3K27me3 [233], Verhaak_classical_expression_UP [69]	*MT3, SPATA6L, OSBPL1A, TCEA2, MEOX2, ZNF3, L3MBTL4, KIAA0101, TMEM106A, PLLP*
D	controls	MES	immune response (BP), cytokine mediated signaling pathway (BP)	*TLR4, RTN4, NR2F2, VIM, TMEM140, NMI, PAXIP1−AS2, DHRS4, CISD2, TM4SF18*
E	G34		*EED*-targets, *SUZ12*-targets, PRC2-targets, H3K27me3 [208]; RNA-Poll_opening (Reactome); meiosis and telomere maintenance (Reactome)	*INHBB, MORN3, NAB2, PCDH10, FGGY, LMCD1, DPYSL3, RASD1, MANF, IGFBP7*
F	RTKII		*EED*-targets, *SUZ12*-targets, PRC2-targets, H3K27me3 [208]; H3K27me3 in HCP [201]; Brain HCP with H3K27me3, with H3K4me3 and H3K27me3 [233], developmental regulators [116]	*PCDHAC1, ZSCAN1, GALNT9, ROBO2, CEP126, POPDC3, EXO5, GRIN3A, HSPA1L, KCNB2*
A1	controls, MES	G34	Olfactory receptor activity (MF), neurological system process (BP), keratinization (BP)	*RPRD2, DPP10, FBLIM1, OR51B4, OR8J3, STX3, ACSM1, OR6Y1, SPTA1, CYB5R2*
B1		high methylation	Hallmark bile acid metabolism, Sensory perception of taste (BP), cell-cell junction (CC)	*ANKRD7, COX7A2, RGS21, LINC01588, KRTAP21−3, NUPR1L, RNASEH2C, HRH4, C5 SLC13A4*
C1	IDH	G34	*SUZ12*-targets, PRC2-targets [208]	*PTGER4, PAX7, IRX4, ACVR1C, OTX1, TTI2, TMEM61, SPIN1, MOXD1, SLC6A5*

C2	G34	control, MES, K27	Cell adhesion (BP), calcium ion binding (MF), *EED*-targets, PRC2-targets, *SUZ12*-targets [208], ES_WITH_H3K27ME3 [233]	*HOXC9, FMN1, ATP8B1, ST6GAL1, EVX2, SFTA3, TBX5, GJA3, GAD2, PAX5*
D1	IDH	control, MES	Nervous system development (BP), hemophilic cell adhesion (BP), LINDVALL_IMMORTALIZED_BY_TERT_UP	*PAX6, FOXB2, VSX1, MKX, COBL, MTA3, PDGFA, ST8SIA4, SH3BP4, C9orf135*
E1		low methylation	KIM_*MYC*-targets [108]	*EPM2AIP1, ZNF300, KCNH6, SLC35G1, ZNF580, AUNIP, DLL3, TSHZ3, ZNF311, MIB1*

[1] Enrichment of predefined gene sets in the spot-lists of genes was calculated as described in [80]. Gene sets were taken from literature or from gene ontology (GO) categories biological process (BP) or cellular component (CC). Only gene sets with GSZ-enrichment $p < 10^{-5}$ were taken into account.
[2] Genes are ranked with decreasing correlation coefficient with the spot profile. Full gene lists together with significance measures (*p*-values of correlation and differential t-tests and false discovery rates) are given in supplementary material of [199]. The lists contain also genes not included in the functional gene sets.

Index of Abbreviations

BCL	B-cell-like lymphoma
BHP	Benign prostatic hyperplasia
BP	Biological process
CC	Cellular component
ChIP	Chromatin immunoprecipitation
CHOP	CpG hypomethylator phenotype
Chr	Chromosome
CL	Classical
CN	Correlation net
CpG	Cytosine-guanine dinucleotide
CRC	Colorectal cancer
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
DNDM	DNA demethylase
DNMT	DNA methyltransferase
DZ	Dark zone
ENCODE	ENCyclopedia Of DNA Elements
Enh	Enhancer
ESC	Embryonic stem cells
FL	Follicular lymphoma
GBM	Glioblastoma multiforme
GC	Germinal center
GCB	Germinal center B-cells
(G)CIMP	(Glioma) CpG methylator phenotype
GO	Gene ontology
GSZ	Gene set Z-score
ICA	Independent component analysis
ICGC	International Cancer Genome Consortium
IntL	Intermediate lymphoma
JmjC	Jumonji C
KDM	Lysine demethylase
KMT	Lysine methyltransferase
LGG	Low-grade gliomas
LZ	Light zone
mBL	Molecular Burkitt's lymphoma
MCL	Mantle cell lymphoma
MES	Mesenchymal
MET	Metastatic prostate cancer

MF	Molecular function
mf lobe	Mid frontal lobe
MM	Multiple myeloma
MMML	Molecular mechanisms of malignant lymphoma
MST	Maximum spanning tree
NB	Naïve B-cell
NL	Neural
NOR	Normal
NP	Neuronal progenitors
PCA_high	High-grade prostate cancer
PCA_low	Low-grade prostate cancer
PcG	Polycomb group
PCM	Pairwise correlation maps
PCP	Prostate cancer progression
PIN	Prostatic intraepithelial neoplasia
PN	Proneural
PRC	Polycomb repressive complex
RepPC	Repressive polycomb complex
RNA	Ribonucleic acid
RTK	Receptor tyrosine kinases
ScoV	Signed square root co-variance
SOM	Self organizing maps
TCA	Tricarboxylic acid
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TrxG	Trithorax group
TSS	Transcription start site
TssA	Active promoter
TssP	poised promoter
TxTrans	Transcribed genes
ZNF	Zink fingers

List of Tables

List of Figures

Bibliography

[1] T. Ozawa, M. Riester, Y.-K. Cheng, J.T. Huse, M. Squatrito, K. Helmy, N. Charles, F. Michor, E.C. Holland, Most Human Non-GCIMP Glioblastoma Subtypes Evolve from a Common Proneural-like Precursor Glioma, Cancer Cell, 26 (2014) 288–300.

[2] A.L. Cohen, S.L. Holmen, H. Colman, IDH1 and IDH2 mutations in gliomas, Curr. Neurol. Neurosci. Rep., 13 (2013) 345.

[3] J.S. You, P.A. Jones, Cancer Genetics and Epigenetics: Two Sides of the Same Coin?, Cancer Cell, 22 (2012) 9–20.

[4] R. Schmitz, R.M. Young, M. Ceribelli, S. Jhavar, W. Xiao, M. Zhang, G. Wright, A.L. Shaffer, D.J. Hodson, E. Buras, X. Liu, J. Powell, Y. Yang, W. Xu, H. Zhao, H. Kohlhammer, A. Rosenwald, P. Kluin, H.K. Müller-Hermelink, G. Ott, et al., Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics, Nature, 490 (2012) 116–20.

[5] R.D. Morin, N.A. Johnson, T.M. Severson, A.J. Mungall, J. An, R. Goya, J.E. Paul, M. Boyle, B.W. Woolcock, F. Kuchenbauer, D. Yap, R.K. Humphries, O.L. Griffith, S. Shah, H. Zhu, M. Kimbara, P. Shashkin, J.F. Charlot, M. Tcherpakov, R. Corbett, et al., Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin, Nat. Genet., 42 (2010) 181–5.

[6] K. Lund, P.D. Adams, M. Copland, EZH2 in normal and malignant hematopoiesis, Leukemia, 28 (2014) 44–9.

[7] I.A. Rodriguez-Brenes, D. Wodarz, Preventing clonal evolutionary processes in cancer: Insights from mathematical models, Proc. Natl. Acad. Sci. U. S. A., 112 (2015) 8843–50.

[8] J.D. Choi, J.-S. Lee, Interplay between Epigenetics and Genetics in Cancer, Genomics Inf., 11 (2013) 164–173.

[9] N.F.C.C. de Miranda, R. Peng, K. Georgiou, C. Wu, E. Falk Sörqvist, M. Berglund, L. Chen, Z. Gao, K. Lagerstedt, S. Lisboa, F. Roos, T. van Wezel, M.R. Teixeira, R. Rosenquist, C. Sundström, G. Enblad, M. Nilsson, Y. Zeng, D. Kipling, Q. Pan-Hammarström, DNA repair genes are selectively mutated in diffuse large B cell lymphomas, J. Exp. Med., 210 (2013) 1729–42.

[10] T. Cancer, G. Atlas, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature, 455 (2008) 1061–8.

[11] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C.J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M.F. Berger, J.E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F.A. Mapa, et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature, 483 (2012) 603–307.

[12] T.J. Hudson (Chairperson), W. Anderson, A. Aretz, A.D. Barker, C. Bell, R.R. Bernabé, M.K. Bhan, F. Calvo, I. Eerola, D.S. Gerhard, A. Guttmacher, M. Guyer, F.M. Hemsley, J.L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusuda, D.P. Lane, F. Laplace, et al., International network of cancer genome projects, Nature, 464 (2010) 993–998.

[13] G.H. Fernald, E. Capriotti, R. Daneshjou, K.J. Karczewski, R.B. Altman, Bioinformatics challenges for personalized medicine, Bioinformatics, 27 (2011)

1741–1748.
[14] M. Pop, S.L. Salzberg, Bioinformatics challenges of new sequencing technology, Trends Genet., 24 (2008) 142–149.
[15] A. Sboner, X.J. Mu, D. Greenbaum, R.K. Auerbach, M.B. Gerstein, The real cost of sequencing: higher than you think!, Genome Biol., 12 (2011) 125.
[16] E.R. Mardis, The $1,000 genome, the $100,000 analysis?, Genome Med., 2 (2010) 84.
[17] T. Kohonen, Self Organizing Maps, Springer, Berlin, Heidelberg, New York, (1995).
[18] P. Törönen, M. Kolehmainen, G. Wong, E. Castrén, Analysis of gene expression data using self-organizing maps, FEBS Lett., 451 (1999) 142–146.
[19] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. U. S. A., 96 (1999) 2907–12.
[20] H. Löffler-Wirth, M. Kalcher, H. Binder, oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor, Bioinformatics, 31 (2015) 3225–7.
[21] H. Wirth, M. Loffler, M. Bergen, H. Binder, Expression cartography of human tissues using self organizing maps, BMC Bioinformatics, 12 (2011) 306.
[22] H. Binder, L. Hopp, K. Lembcke, H. Wirth, Personalized disease phenotypes from massive OMICs data, (n.d.).
[23] H. Wirth, M. Löffler, M. Von Bergen, H. Binder, Expression cartography of human tissues using self organizing maps, BMC Bioinformatics, 12 (2011) 306.
[24] F. Crick, Central dogma of molecular biology, Nature, 227 (1970) 561–3.
[25] S. Franklin, T.M. Vondriska, Genomes, proteomes, and the central dogma, Circ. Cardiovasc. Genet., 4 (2011) 576.
[26] J. Minarovits, F. Banati, K. Szenthe, H.H. Niller, Epigenetic Regulation, Adv. Exp. Med. Biol., 879 (2016) 1–25.
[27] M. Maleszewska, B. Kaminska, Is Glioblastoma an Epigenetic Malignancy?, Cancers (Basel)., 5 (2013) 1120–1139.
[28] V.E.A. Russo, A.D. Riggs, R.A. Martienssen, Epigenetic mechanisms of gene regulation, New York, 1996.
[29] X. Zhang, J.A. Kuivenhoven, A.K. Groen, Forward Individualized Medicine from Personal Genomes to Interactomes, Front. Physiol., (2015).
[30] K. Muegge, Lsh, a guardian of heterochromatin at repeat elements, Biochem. Cell Biol., 83 (2005) 548–54.
[31] A.P. Feinberg, B. Vogelstein, Hypomethylation distinguishes genes of some human cancers from their normal counterparts, Nature, 301 (1983) 89–92.
[32] V.A. DeWoskin, R.P. Million, The epigenetics pipeline, Nat. Rev. Drug Discov., 12 (2013) 661–2.
[33] D. Bhattacharjee, S. Shenoy, K.L. Bairy, DNA Methylation and Chromatin Remodeling: The Blueprint of Cancer Epigenetics, Scientifica (Cairo)., 2016 (2016) 6072357.
[34] S. Lou, H.-M. Lee, H. Qin, J.-W. Li, Z. Gao, X. Liu, L.L. Chan, V. KL Lam, W.-Y. So, Y. Wang, S. Lok, J. Wang, R.C. Ma, S.K.-W. Tsui, J.C. Chan, T.-F. Chan, K.Y. Yip, A. Bird, S. Cokus, S. Feng, et al., Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation, Genome Biol., 15 (2014) 408.

[35] N.R. Rose, R.J. Klose, Understanding the relationship between DNA methylation and histone lysine methylation, Biochim. Biophys. Acta, 1839 (2014) 1362–72.
[36] T. Chen, Y. Ueda, J.E. Dodge, Z. Wang, E. Li, Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b, Mol. Cell. Biol., 23 (2003) 5594–605.
[37] T. Phillips, The Role of Methylation in Gene Expression, Nat. Educ., 1 (2008) 116.
[38] B. Kinde, H.W. Gabel, C.S. Gilbert, E.C. Griffith, M.E. Greenberg, Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2, Proc. Natl. Acad. Sci. U. S. A., 112 (2015) 6800–6.
[39] M. Rasool, A. Malik, M.I. Naseer, A. Manan, S. Ansari, I. Begum, M.H. Qazi, P. Pushparaj, A.M. Abuzenadah, M.H. Al-Qahtani, M.A. Kamal, S. Gan, The role of epigenetics in personalized medicine: challenges and opportunities, BMC Med. Genomics, 8 Suppl 1 (2015) S5.
[40] C. Dupont, D.R. Armant, C.A. Brenner, Epigenetics: definition, mechanisms and clinical perspective, Semin. Reprod. Med., 27 (2009) 351–7.
[41] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-Resolution Profiling of Histone Methylations in the Human Genome, (n.d.).
[42] N. Hattori, T. Ushijima, Compendium of aberrant DNA methylation and histone modifications in cancer, Biochem. Biophys. Res. Commun., (2014).
[43] S. Varambally, S.M. Dhanasekaran, M. Zhou, T.R. Barrette, C. Kumar-Sinha, M.G. Sanda, D. Ghosh, K.J. Pienta, R.G.A.B. Sewalt, A.P. Otte, M.A. Rubin, A.M. Chinnaiyan, The polycomb group protein EZH2 is involved in progression of prostate cancer, Nature, 419 (2002) 624–629.
[44] S. Sharma, T.K. Kelly, P.A. Jones, Epigenetics in cancer, Carcinogenesis, 31 (2010) 27–36.
[45] R.A. Varier, H.T.M. Timmers, Histone lysine methylation and demethylation pathways in cancer, BBA - Rev. Cancer, 1815 (2011) 75–89.
[46] H. Noushmehr, D.J. Weisenberger, K. Diefes, H.S. Phillips, K. Pujara, B.P. Berman, F. Pan, C.E. Pelloski, E.P. Sulman, K.P. Bhat, R.G.W. Verhaak, K.A. Hoadley, D.N. Hayes, C.M. Perou, H.K. Schmidt, L. Ding, R.K. Wilson, D. Van Den Berg, H. Shen, H. Bengtsson, et al., Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma, Cancer Cell, 17 (2010) 510–22.
[47] P. Chi, C.D. Allis, G.G. Wang, Covalent histone modifications–miswritten, misinterpreted and mis-erased in human cancers, Nat. Rev. Cancer, 10 (2010) 457–69.
[48] Y. Guo, Q. Sheng, J. Li, F. Ye, D.C. Samuels, Y. Shyr, Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data, PLoS One, 8 (2013) e71462.
[49] K.J. Mantione, R.M. Kream, H. Kuzelova, R. Ptacek, J. Raboch, J.M. Samuel, G.B. Stefano, Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq, Med. Sci. Monit. Basic Res., 20 (2014) 138–42.
[50] P.C. Roberts, Gene expression microarray data analysis demystified, Biotechnol. Annu. Rev., 14 (2008) 29–61.
[51] D. Gershon, Microarray technologyAn array of opportunities, Nature, 416 (2002) 885–891.
[52] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet., 10 (2009) 57–63.

[53] M.L. Wright, M.G. Dozmorov, A.R. Wolen, C. Jackson-Cook, A.R. Starkweather, D.E. Lyon, T.P. York, Establishing an analytic pipeline for genome-wide DNA methylation, Clin. Epigenetics, 8 (2016) 45.

[54] H. O'Geen, L. Echipare, P.J. Farnham, Using ChIP-seq technology to generate high-resolution profiles of histone modifications, Methods Mol. Biol., 791 (2011) 265–86.

[55] R.E. Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, V. Amin, J.W. Whitaker, M.D. Schultz, L.D. Ward, A. Sarkar, G. Quon, R.S. Sandstrom, M.L. Eaton, Y.-C. Wu, et al., Integrative analysis of 111 reference human epigenomes, Nature, 518 (2015) 317–330.

[56] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, Nat. Methods, 9 (2012) 215–216.

[57] H. Wirth, M. Loffler, M. von Bergen, H. Binder, Expression cartography of human tissues using self organizing maps, BMC Bioinformatics, 12 (2011) 306.

[58] S. a Tomlins, R. Mehra, D.R. Rhodes, X. Cao, L. Wang, S.M. Dhanasekaran, S. Kalyana-Sundaram, J.T. Wei, M. a Rubin, K.J. Pienta, R.B. Shah, A.M. Chinnaiyan, Integrative molecular concept modeling of prostate cancer progression, Nat. Genet., (2007).

[59] M. Hummel, S. Bentink, H. Berger, W. Klapper, S. Wessendorf, T.F.E. Barth, H.-W. Bernd, S.B. Cogliatti, J. Dierlamm, A.C. Feller, M.-L. Hansmann, E. Haralambieva, L. Harder, D. Hasenclever, M. Kühn, D. Lenze, P. Lichter, J.I. Martin-Subero, P. Möller, H.-K. Müller-Hermelink, et al., A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling, N. Engl. J. Med., 354 (2006) 2419–30.

[60] J.I. Martin-Subero, O. Ammerpohl, M. Bibikova, E. Wickham-Garcia, X. Agirre, S. Alvarez, M. Brüggemann, S. Bug, M.J. Calasanz, M. Deckert, M. Dreyling, M.Q. Du, J. Dürig, M.J.S. Dyer, J.-B. Fan, S. Gesk, M.-L. Hansmann, L. Harder, S. Hartmann, W. Klapper, et al., A comprehensive microarray-based DNA methylation study of 367 hematological neoplasms, PLoS One, 4 (2009) e6986.

[61] R.G.W. Verhaak, K.A. Hoadley, E. Purdom, V. Wang, Y. Qi, M.D. Wilkerson, C.R. Miller, L. Ding, T. Golub, J.P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B.A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H.S. Feiler, et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1, Cancer Cell, 17 (2010) 98–110.

[62] D. Sturm, H. Witt, V. Hovestadt, D.-A. Khuong-Quang, D.T.W. Jones, C. Konermann, E. Pfaff, M. Tönjes, M. Sill, S. Bender, M. Kool, M. Zapatka, N. Becker, M. Zucknick, T. Hielscher, X.-Y. Liu, A.M. Fontebasso, M. Ryzhova, S. Albrecht, K. Jacob, et al., Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma, Cancer Cell, 22 (2012) 425–37.

[63] G. Reifenberger, R.G. Weber, V. Riehmer, K. Kaulich, E. Willscher, H. Wirth, J. Gietzelt, B. Hentschel, M. Westphal, M. Simon, G. Schackert, J. Schramm, J. Matschke, M.C. Sabel, D. Gramatzki, J. Felsberg, C. Hartmann, J.P. Steinbach, U. Schlegel, W. Wick, et al., Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling, Int. J. Cancer, 135 (2014) 1822–31.

[64] H. Binder, S. Preibisch, "Hook"-calibration of GeneChip-microarrays: theory and

algorithm, Algorithms Mol. Biol., 3 (2008) 12.
[65] H. Binder, K. Krohn, S. Preibisch, "Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures, Algorithms Mol. Biol., 3 (2008) 11.
[66] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics, 19 (2003) 185–93.
[67] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W.A. Kibbe, L. Hou, S.M. Lin, Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC Bioinformatics, 11 (2010) 587.
[68] S.A. Tomlins, R. Mehra, D.R. Rhodes, X. Cao, L. Wang, S.M. Dhanasekaran, S. Kalyana-Sundaram, J.T. Wei, M.A. Rubin, K.J. Pienta, R.B. Shah, A.M. Chinnaiyan, Integrative molecular concept modeling of prostate cancer progression, Nat. Genet., 39 (2007) 41–51.
[69] L. Hopp, H. Wirth, M. Fasold, H. Binder, Portraying the expression landscapes of cancer subtypes A case study of glioblastoma multiforme and prostate cancer, Syst. Biomed., 1 (2013) 1–23.
[70] H. Wirth, M. von Bergen, J. Murugaiyan, U. Rösler, T. Stokowy, H. Binder, MALDI-typing of infectious algae of the genus Prototheca using SOM portraits, J. Microbiol. Methods, 88 (2012) 83–97.
[71] Y. Guo, G.S. Eichler, Y. Feng, D.E. Ingber, S. Huang, Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers, J. Biomed. Biotechnol., 2006 (2006) 69141.
[72] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Netw., 13 411–30.
[73] W. Liebermeister, Linear modes of gene expression determined by independent component analysis, Bioinformatics, 18 (2002) 51–60.
[74] M. Riester, C. Stephan-Otto Attolini, R.J. Downey, S. Singer, F. Michor, A differentiation-based phylogeny of cancer subtypes, PLoS Comput. Biol., 6 (2010) e1000777.
[75] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol., 4 (1987) 406–25.
[76] J. Quackenbush, Genomics Microarrays–guilt by association, Science, 302 (2003) 240–1.
[77] J.J. Goeman, P. Bühlmann, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics, 23 (2007) 980–7.
[78] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology The Gene Ontology Consortium, Nat. Genet., 25 (2000) 25–9.
[79] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, A. Kasprzyk, BioMart Central Portal–unified access to biological data, Nucleic Acids Res., 37 (2009) W23–7.
[80] H. Wirth, M. von Bergen, H. Binder, Mining SOM expression portraits: feature selection and integrating concepts of molecular function, BioData Min., 5 (2012) 18.
[81] M. Ackermann, K. Strimmer, A general modular framework for gene set enrichment analysis, BMC Bioinformatics, 10 (2009) 47.
[82] P. Törönen, P.J. Ojala, P. Marttinen, L. Holm, Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function, BMC

Bioinformatics, 10 (2009) 307.
[83] A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, Bioinformatics, 22 (2006) 1600–7.
[84] F. Alberghini, V. Petrocelli, M. Rahmat, S. Casola, An epigenetic view of B-cell disorders, Immunol. Cell Biol., 93 (2015) 253–260.
[85] S.S. Dave, K. Fu, G.W. Wright, L.T. Lam, P. Kluin, E.-J. Boerma, T.C. Greiner, D.D. Weisenburger, A. Rosenwald, G. Ott, H.-K. Müller-Hermelink, R.D. Gascoyne, J. Delabie, L.M. Rimsza, R.M. Braziel, T.M. Grogan, E. Campo, E.S. Jaffe, B.J. Dave, W. Sanger, et al., Molecular diagnosis of Burkitt's lymphoma, N. Engl. J. Med., 354 (2006) 2431–42.
[86] Y. Jiang, A. Melnick, The epigenetic basis of diffuse large B-cell lymphoma, Semin. Hematol., 52 (2015) 86–96.
[87] J.I. Martín-Subero, M. Kreuz, M. Bibikova, S. Bentink, O. Ammerpohl, E. Wickham-Garcia, M. Rosolowski, J. Richter, L. Lopez-Serra, E. Ballestar, H. Berger, X. Agirre, H.-W. Bernd, V. Calvanese, S.B. Cogliatti, H.G. Drexler, J.-B. Fan, M.F. Fraga, M.L. Hansmann, M. Hummel, et al., New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling, Blood, 113 (2009) 2488–97.
[88] G.G. Wang, K.D. Konze, J. Tao, Polycomb genes, miRNA, and their deregulation in B-cell malignancies, Blood, 125 (2015) 1217–25.
[89] G. Lenz, G.W. Wright, N.C.T. Emre, H. Kohlhammer, S.S. Dave, R.E. Davis, S. Carty, L.T. Lam, A.L. Shaffer, W. Xiao, J. Powell, A. Rosenwald, G. Ott, H.K. Muller-Hermelink, R.D. Gascoyne, J.M. Connors, E. Campo, E.S. Jaffe, J. Delabie, E.B. Smeland, et al., Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways, Proc. Natl. Acad. Sci. U. S. A., 105 (2008) 13520–5.
[90] 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, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403 (2000) 503–11.
[91] L. Hopp, K. Lembcke, H. Binder, H. Wirth, Portraying the Expression Landscapes of B-CellLymphoma-Intuitive Detection of Outlier Samples and of Molecular Subtypes, Biology (Basel)., 2 (2013) 1411–37.
[92] Cytochrome p450 and chemical toxicology, Chem. Res. Toxicol., 21 (2008) 70–83.
[93] S. Lloyd, Least squares quantization in PCM, Inf. Theory, IEEE Trans., 28 (1982) 129–137.
[94] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, (2007) 1027–1035.
[95] S. Monti, P. Tamayo, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Mach. Learn., (2003).
[96] W. Klapper, M. Kreuz, C.W. Kohler, B. Burkhardt, M. Szczepanowski, I. Salaverria, M. Hummel, M. Loeffler, S. Pellissery, W. Woessmann, C. Schwänen, L. Trümper, S. Wessendorf, R. Spang, D. Hasenclever, R. Siebert, Patient age at diagnosis is associated with the molecular characteristics of diffuse large B-cell lymphoma, Blood, 119 (2012) 1882–7.
[97] H. Stein, M. Hummel, [Burkitt's and Burkitt-like lymphoma Molecular definition and

value of the World Health Organisation's diagnostic criteria], Pathologe, 28 (2007) 41–5.
[98] G. Wright, B. Tan, A. Rosenwald, E.H. Hurt, A. Wiestner, L.M. Staudt, A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma, Proc. Natl. Acad. Sci. U. S. A., 100 (2003) 9991–6.
[99] Y. Lu, Y. Yi, P. Liu, W. Wen, M. James, D. Wang, M. You, Common human cancer genes discovered by integrated gene-expression analysis, PLoS One, 2 (2007) e1149.
[100] A. Wolfer, B.S. Wittner, D. Irimia, R.J. Flavin, M. Lupien, R.N. Gunawardane, C.A. Meyer, E.S. Lightcap, P. Tamayo, J.P. Mesirov, X.S. Liu, T. Shioda, M. Toner, M. Loda, M. Brown, J.S. Brugge, S. Ramaswamy, MYC regulation of a "poor-prognosis" metastatic cancer cell state, Proc. Natl. Acad. Sci. U. S. A., 107 (2010) 3698–703.
[101] E.L. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations on JSTOR, J. Am. Stat. Assoc., 53 (1958) 457–481.
[102] M. Rosolowski, J. Läuter, D. Abramov, H.G. Drexler, M. Hummel, W. Klapper, R.A.F. Macleod, S. Pellissery, F. Horn, R. Siebert, M. Loeffler, Massive transcriptional perturbation in subgroups of diffuse large B-cell lymphomas, PLoS One, 8 (2013) e76287.
[103] H. Binder, H. Wirth, A. Arakelyan, K. Lembcke, E.S. Tiys, V.A. Ivanisenko, N.A. Kolchanov, A. Kononikhin, I. Popov, E.N. Nikolaev, L. Pastushkova, I.M. Larina, Time-course human urine proteomics in space-flight simulation experiments, BMC Genomics, 15 Suppl 1 (2014) S2.
[104] E. Walker, J.L. Manias, W.Y. Chang, W.L. Stanford, PCL2 modulates gene regulatory networks controlling self-renewal and commitment in embryonic stem cells, Cell Cycle, 10 (2011) 45–51.
[105] B.C. Christensen, A.A. Smith, S. Zheng, D.C. Koestler, E.A. Houseman, C.J. Marsit, J.L. Wiemels, H.H. Nelson, M.R. Karagas, M.R. Wrensch, K.T. Kelsey, J.K. Wiencke, DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma, J. Natl. Cancer Inst., 103 (2011) 143–53.
[106] A.E. Teschendorff, X. Liu, H. Caren, S.M. Pollard, S. Beck, M. Widschwendter, L. Chen, The dynamics of DNA methylation covariation patterns in carcinogenesis, PLoS Comput. Biol., 10 (2014) e1003709.
[107] P.W. Ang, M. Loh, N. Liem, P.L. Lim, F. Grieu, A. Vaithilingam, C. Platell, W.P. Yong, B. Iacopetta, R. Soong, Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features, BMC Cancer, 10 (2010) 227.
[108] Y.H. Kim, L. Girard, C.P. Giacomini, P. Wang, T. Hernandez-Boussard, R. Tibshirani, J.D. Minna, J.R. Pollack, Combined microarray analysis of small cell lung cancer reveals altered apoptotic balance and distinct expression signatures of MYC family gene amplification, Oncogene, 25 (2006) 130–8.
[109] S. Bentink, S. Wessendorf, C. Schwaenen, M. Rosolowski, W. Klapper, A. Rosenwald, G. Ott, A.H. Banham, H. Berger, A.C. Feller, M.-L. Hansmann, D. Hasenclever, M. Hummel, D. Lenze, P. Möller, B. Stuerzenhofecker, M. Loeffler, L. Truemper, H. Stein, R. Siebert, et al., Pathway activation patterns in diffuse large B-cell lymphomas, Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K, 22 (2008) 1746–54.
[110] G.G. Wang, J. Song, Z. Wang, H.L. Dormann, F. Casadio, H. Li, J.-L. Luo, D.J. Patel, C.D. Allis, Haematopoietic malignancies caused by dysregulation of a

chromatin-binding PHD finger, Nature, 459 (2009) 847–51.
[111] J. Laffaire, S. Everhard, A. Idbaih, E. Crinière, Y. Marie, A. de Reyniès, R. Schiappa, K. Mokhtari, K. Hoang-Xuan, M. Sanson, J.-Y. Delattre, J. Thillet, F. Ducray, Methylation profiling identifies 2 groups of gliomas according to their tumorigenesis, Neuro. Oncol., 13 (2011) 84–98.
[112] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, S.A. Teichmann, RNA sequencing reveals two major classes of gene expression levels in metazoan cells, Mol. Syst. Biol., 7 (2011) 497.
[113] L. Hopp, H. Löffler-Wirth, H. Binder, Epigenetic Heterogeneity of B-Cell Lymphoma: DNA Methylation, Gene Expression and Chromatin States, Genes (Basel)., 6 (2015) 812–40.
[114] H. Loeffler-Wirth, L. Hopp, M. Kreuz, R. Siebert, M. Loeffler, H. Binder, A holistic view on the expression landscape of Germinal center-derived B-cell lymphomas, Submitted, (2016).
[115] B.E. Bernstein, E. Birney, I. Dunham, E.D. Green, C. Gunter, M. Snyder, An integrated encyclopedia of DNA elements in the human genome, Nature, 489 (2012) 57–74.
[116] T.I. Lee, R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar, B. Chevalier, S.E. Johnstone, M.F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H.L. Murray, J.P. Zucker, B. Yuan, G.W. Bell, E. Herbolsheimer, N.M. Hannett, K. Sun, et al., Control of developmental regulators by Polycomb in human embryonic stem cells, Cell, 125 (2006) 301–13.
[117] A.E. Teschendorff, J. West, S. Beck, Age-associated epigenetic drift: implications, and a case of epigenetic thrift?, Hum. Mol. Genet., 22 (2013) R7–R15.
[118] Y. Kondo, L. Shen, A.S. Cheng, S. Ahmed, Y. Boumber, C. Charo, T. Yamochi, T. Urano, K. Furukawa, B. Kwabi-Addo, D.L. Gold, Y. Sekido, T.H.-M. Huang, J.-P.J. Issa, Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation, Nat. Genet., 40 (2008) 741–50.
[119] K.B. Hendricks, F. Shanahan, E. Lees, Role for BRG1 in cell cycle control and tumor suppression, Mol. Cell. Biol., 24 (2004) 362–76.
[120] I. Velichutina, R. Shaknovich, H. Geng, N.A. Johnson, R.D. Gascoyne, A.M. Melnick, O. Elemento, EZH2-mediated epigenetic silencing in germinal center B cells contributes to proliferation and lymphomagenesis, Blood, 116 (2010) 5247–55.
[121] W. Béguelin, R. Popovic, M. Teater, Y. Jiang, K.L. Bunting, M. Rosen, H. Shen, S.N. Yang, L. Wang, T. Ezponda, E. Martinez-Garcia, H. Zhang, Y. Zheng, S.K. Verma, M.T. McCabe, H.M. Ott, G.S. Van Aller, R.G. Kruger, Y. Liu, C.F. McHugh, et al., EZH2 Is Required for Germinal Center Formation and Somatic EZH2 Mutations Promote Lymphoid Transformation, Cancer Cell, 23 (2013) 677–692.
[122] J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome, Nat. Biotechnol., 28 (2010) 817–25.
[123] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B.E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types, Nature, 473 (2011) 43–9.
[124] S.S. Dave, G. Wright, B. Tan, A. Rosenwald, R.D. Gascoyne, W.C. Chan, R.I. Fisher, R.M. Braziel, L.M. Rimsza, T.M. Grogan, T.P. Miller, M. LeBlanc, T.C. Greiner, D.D. Weisenburger, J.C. Lynch, J. Vose, J.O. Armitage, E.B. Smeland, S. Kvaloy, H. Holte, et al., Prediction of survival in follicular lymphoma based on

molecular features of tumor-infiltrating immune cells, N. Engl. J. Med., 351 (2004) 2159–69.
[125] R. Shaknovich, L. Cerchietti, L. Tsikitas, M. Kormaksson, S. De, M.E. Figueroa, G. Ballon, S.N. Yang, N. Weinhold, M. Reimers, T. Clozel, K. Luttrop, T.J. Ekstrom, J. Frank, A. Vasanthakumar, L.A. Godley, F. Michor, O. Elemento, A. Melnick, DNA methyltransferase 1 and DNA methylation patterning contribute to germinal center B-cell differentiation, Blood, 118 (2011) 3559–69.
[126] P. Voigt, W.-W. Tee, D. Reinberg, A double take on bivalent promoters, Genes Dev., 27 (2013) 1318–38.
[127] R. Shaknovich, A. Melnick, Epigenetics and B-cell lymphoma, Curr. Opin. Hematol., 18 (2011) 293–9.
[128] J.P. Thomson, P.J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A.R.W. Kerr, A. Deaton, R. Andrews, K.D. James, D.J. Turner, R. Illingworth, A. Bird, CpG islands influence chromatin structure via the CpG-binding protein Cfp1, Nature, 464 (2010) 1082–6.
[129] N. Fujita, S. Watanabe, T. Ichimura, S. Tsuruzoe, Y. Shinkai, M. Tachibana, T. Chiba, M. Nakao, Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression, J. Biol. Chem., 278 (2003) 24132–8.
[130] N. Feldman, A. Gerson, J. Fang, E. Li, Y. Zhang, Y. Shinkai, H. Cedar, Y. Bergman, G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis, Nat. Cell Biol., 8 (2006) 188–94.
[131] S.K.T. Ooi, C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S.-P. Lin, C.D. Allis, X. Cheng, T.H. Bestor, DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA, Nature, 448 (2007) 714–7.
[132] J.A. Hagarman, M.P. Motley, K. Kristjansdottir, P.D. Soloway, Coordinate Regulation of DNA Methylation and H3K27me3 in Mouse Embryonic Stem Cells, PLoS One, 8 (2013) e53880.
[133] H. Liu, Y. Chen, J. Lv, H. Liu, R. Zhu, J. Su, X. Liu, Y. Zhang, Q. Wu, Quantitative epigenetic co-variation in CpG islands and co-regulation of developmental genes, Sci. Rep., 3 (2013) 2576.
[134] C. Köhler, C.B.R. Villar, Programming of gene expression by Polycomb group proteins, Trends Cell Biol., 18 (2008) 236–43.
[135] R.J. Klose, S. Cooper, A.M. Farcas, N.P. Blackledge, N. Brockdorff, Chromatin sampling–an emerging perspective on targeting polycomb repressor proteins, PLoS Genet., 9 (2013) e1003717.
[136] A. Jeltsch, R.Z. Jurkowska, New concepts in DNA methylation, Trends Biochem. Sci., 39 (2014) 310–8.
[137] C. Love, Z. Sun, D. Jima, G. Li, J. Zhang, R. Miles, K.L. Richards, C.H. Dunphy, W.W.L. Choi, G. Srivastava, P.L. Lugar, D.A. Rizzieri, A.S. Lagoo, L. Bernal-Mizrachi, K.P. Mann, C.R. Flowers, K.N. Naresh, A.M. Evens, A. Chadburn, L.I. Gordon, et al., The genetic landscape of mutations in Burkitt lymphoma, Nat. Genet., 44 (2012) 1321–5.
[138] M. Sarris, K. Nikolaou, I. Talianidis, Context-specific regulation of cancer epigenomes by histone and transcription factor methylation, Oncogene, 33 (2014) 1207–17.
[139] H. Shen, P.W. Laird, Interplay between the cancer genome and epigenome, Cell, 153 (2013) 38–55.

[140] J.P. Vaqué, N. Martínez, A. Batlle-López, C. Pérez, S. Montes-Moreno, M. Sánchez-Beato, M.A. Piris, B-cell lymphoma mutations: improving diagnostics and enabling targeted therapies, Haematologica, 99 (2014) 222–31.
[141] Y.R. Chung, E. Schatoff, O. Abdel-Wahab, Epigenetic alterations in hematopoietic malignancies, Int. J. Hematol., 96 (2012) 413–27.
[142] R.D. Morin, M. Mendez-Lago, A.J. Mungall, R. Goya, K.L. Mungall, R.D. Corbett, N.A. Johnson, T.M. Severson, R. Chiu, M. Field, S. Jackman, M. Krzywinski, D.W. Scott, D.L. Trinh, J. Tamura-Wells, S. Li, M.R. Firme, S. Rogic, M. Griffith, S. Chan, et al., Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma, Nature, 476 (2011) 298–303.
[143] C.J. Sneeringer, M.P. Scott, K.W. Kuntz, S.K. Knutson, R.M. Pollock, V.M. Richon, R.A. Copeland, Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas, Proc. Natl. Acad. Sci. U. S. A., 107 (2010) 20980–5.
[144] M. Xiao, H. Yang, W. Xu, S. Ma, H. Lin, H. Zhu, L. Liu, Y. Liu, C. Yang, Y. Xu, S. Zhao, D. Ye, Y. Xiong, K.-L. Guan, Inhibition of α-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors, Genes Dev., 26 (2012) 1326–38.
[145] W. Xu, H. Yang, Y. Liu, Y. Yang, P. Wang, S.-H. Kim, S. Ito, C. Yang, P. Wang, M.-T. Xiao, L. Liu, W. Jiang, J. Liu, J. Zhang, B. Wang, S. Frye, Y. Zhang, Y. Xu, Q. Lei, K.-L. Guan, et al., Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases, Cancer Cell, 19 (2011) 17–30.
[146] E.-H. Shim, C.B. Livi, D. Rakheja, J. Tan, D. Benson, V. Parekh, E.-Y. Kho, A.P. Ghosh, R. Kirkman, S. Velu, S. Dutta, B. Chenna, S.L. Rea, R.J. Mishur, Q. Li, T.L. Johnson-Pais, L. Guo, S. Bae, S. Wei, K. Block, et al., L-2-Hydroxyglutarate: an epigenetic modifier and putative oncometabolite in renal cancer, Cancer Discov., 4 (2014) 1290–8.
[147] C. Lu, P.S. Ward, G.S. Kapoor, D. Rohle, S. Turcan, O. Abdel-Wahab, C.R. Edwards, R. Khanin, M.E. Figueroa, A. Melnick, K.E. Wellen, D.M. O'Rourke, S.L. Berger, T.A. Chan, R.L. Levine, I.K. Mellinghoff, C.B. Thompson, IDH mutation impairs histone demethylation and results in a block to cell differentiation, Nature, 483 (2012) 474–8.
[148] M. Weller, R.G. Weber, E. Willscher, V. Riehmer, B. Hentschel, M. Kreuz, J. Felsberg, U. Beyer, H. Löffler-Wirth, K. Kaulich, J.P. Steinbach, C. Hartmann, D. Gramatzki, J. Schramm, M. Westphal, G. Schackert, M. Simon, T. Martens, J. Boström, C. Hagel, et al., Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups, Acta Neuropathol., 129 (2015) 679–93.
[149] J.-A. Losman, W.G. Kaelin, What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer, Genes Dev., 27 (2013) 836–852.
[150] J. Zhang, V. Grubor, C.L. Love, A. Banerjee, K.L. Richards, P.A. Mieczkowski, C. Dunphy, W. Choi, W.Y. Au, G. Srivastava, P.L. Lugar, D.A. Rizzieri, A.S. Lagoo, L. Bernal-Mizrachi, K.P. Mann, C. Flowers, K. Naresh, A. Evens, L.I. Gordon, M. Czader, et al., Genetic heterogeneity of diffuse large B-cell lymphoma, Proc. Natl. Acad. Sci. U. S. A., 110 (2013) 1398–403.
[151] J. Yun, J.L. Johnson, C.L. Hanigan, J.W. Locasale, L. Galluzzi, A. Arcaro, R. Schafer, Interactions between epigenetics and metabolism in cancers, (2012).
[152] P. Caro, A.U. Kishan, E. Norberg, I.A. Stanley, B. Chapuy, S.B. Ficarro, K. Polak,

D. Tondera, J. Gounarides, H. Yin, F. Zhou, M.R. Green, L. Chen, S. Monti, J.A. Marto, M.A. Shipp, N.N. Danial, Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma, Cancer Cell, 22 (2012) 547–60.

[153] M. Jakovcevski, S. Akbarian, Epigenetic mechanisms in neurological disease, Nat. Med., 18 (2012) 1194–204.

[154] B. Jin, J. Ernst, R.L. Tiedemann, H. Xu, S. Sureshchandra, M. Kellis, S. Dalton, C. Liu, J.-H. Choi, K.D. Robertson, Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells, Cell Rep., 2 (2012) 1411–24.

[155] K. Williams, J. Christensen, K. Helin, DNA methylation: TET proteins-guardians of CpG islands?, EMBO Rep., 13 (2012) 28–35.

[156] N. Detich, J. Theberge, M. Szyf, Promoter-specific activation and demethylation by MBD2/demethylase, J. Biol. Chem., 277 (2002) 35791–4.

[157] M. Caganova, C. Carrisi, G. Varano, F. Mainoldi, F. Zanardi, P.-L. Germain, L. George, F. Alberghini, L. Ferrarini, A.K. Talukder, M. Ponzoni, G. Testa, T. Nojima, C. Doglioni, D. Kitamura, K.-M. Toellner, I. Su, S. Casola, Germinal center dysregulation by histone methyltransferase EZH2 promotes lymphomagenesis, J. Clin. Invest., 123 (2013) 5009–22.

[158] W. Jiang, J. Wang, Y. Zhang, Histone H3K27me3 demethylases KDM6A and KDM6B modulate definitive endoderm differentiation from human ESCs by regulating WNT signaling pathway, Cell Res., 23 (2013) 122–30.

[159] T. Swigut, J. Wysocka, H3K27 demethylases, at long last, Cell, 131 (2007) 29–32.

[160] J. Son, S.S. Shen, R. Margueron, D. Reinberg, Nucleosome-binding activities within JARID2 and EZH1 regulate the function of PRC2 on chromatin, Genes Dev., 27 (2013) 2663–77.

[161] X. Shen, Y. Liu, Y.-J. Hsu, Y. Fujiwara, J. Kim, X. Mao, G.-C. Yuan, S.H. Orkin, EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency, Mol. Cell, 32 (2008) 491–502.

[162] B. Henriquez, F.J. Bustos, R. Aguilar, A. Becerra, F. Simon, M. Montecino, B. van Zundert, Ezh1 and Ezh2 differentially regulate PSD-95 gene transcription in developing hippocampal neurons, Mol. Cell. Neurosci., 57 (2013) 130–43.

[163] R. Margueron, G. Li, K. Sarma, A. Blais, J. Zavadil, C.L. Woodcock, B.D. Dynlacht, D. Reinberg, Ezh1 and Ezh2 Maintain Repressive Chromatin through Different Mechanisms, Mol. Cell, 32 (2008) 503–518.

[164] S. Beà, R. Valdés-Mas, A. Navarro, I. Salaverria, D. Martín-Garcia, P. Jares, E. Giné, M. Pinyol, C. Royo, F. Nadeu, L. Conde, M. Juan, G. Clot, P. Vizán, L. Di Croce, D.A. Puente, M. López-Guerra, A. Moros, G. Roue, M. Aymerich, et al., Landscape of somatic mutations and clonal evolution in mantle cell lymphoma, Proc. Natl. Acad. Sci. U. S. A., 110 (2013) 18250–5.

[165] J. Zhang, D. Jima, A.B. Moffitt, Q. Liu, M. Czader, E.D. Hsi, Y. Fedoriw, C.H. Dunphy, K.L. Richards, J.I. Gill, Z. Sun, C. Love, P. Scotland, E. Lock, S. Levy, D.S. Hsu, D. Dunson, S.S. Dave, The genomic landscape of mantle cell lymphoma is related to the epigenetically determined chromatin state of normal B cells, Blood, 123 (2014) 2988–96.

[166] E. Martinez-Garcia, R. Popovic, D.-J. Min, S.M.M. Sweet, P.M. Thomas, L. Zamdborg, A. Heffner, C. Will, L. Lamy, L.M. Staudt, D.L. Levens, N.L. Kelleher,

J.D. Licht, The MMSET histone methyl transferase switches global histone methylation and alters gene expression in t(4;14) multiple myeloma cells, Blood, 117 (2011) 211–20.
[167] L. Pasqualucci, V. Trifonov, G. Fabbri, J. Ma, D. Rossi, A. Chiarenza, V.A. Wells, A. Grunn, M. Messina, O. Elliot, J. Chan, G. Bhagat, A. Chadburn, G. Gaidano, C.G. Mullighan, R. Rabadan, R. Dalla-Favera, Analysis of the coding genome of diffuse large B-cell lymphoma, Nat. Genet., 43 (2011) 830–7.
[168] T. Gindin, V. Murty, B. Alobeid, G. Bhagat, MLL/KMT2A translocations in diffuse large B-cell lymphomas, Hematol. Oncol., 33 (2015) 239–46.
[169] L.-B. Chen, J.-Y. Xu, Z. Yang, G.-B. Wang, Silencing SMYD3 in hepatoma demethylates RIZI promoter induces apoptosis and inhibits cell proliferation and migration, World J. Gastroenterol., 13 (2007) 5718–24.
[170] B.L. Kidder, G. Hu, K. Zhao, KDM5B focuses H3K4 methylation near promoters and enhancers during embryonic stem cell self-renewal and differentiation, Genome Biol., 15 (2014) R32.
[171] M. Zoabi, P.T. Nadar-Ponniah, H. Khoury-Haddad, M. Usaj, I. Budowski-Tal, T. Haran, A. Henn, Y. Mandel-Gutfreund, N. Ayoub, RNA-dependent chromatin localization of KDM4D lysine demethylase promotes H3K9me3 demethylation, Nucleic Acids Res., 42 (2014) 13026–38.
[172] W.L. Berry, R. Janknecht, KDM4/JMJD2 histone demethylases: epigenetic regulators in cancer cells, Cancer Res., 73 (2013) 2936–42.
[173] B.L. Gregory, V.G. Cheung, Natural variation in the histone demethylase, KDM4C, influences expression levels of specific genes including those that affect cell growth, Genome Res., 24 (2014) 52–63.
[174] T. Bartke, M. Vermeulen, B. Xhemalce, S.C. Robson, M. Mann, T. Kouzarides, Nucleosome-interacting proteins regulated by DNA and histone methylation, Cell, 143 (2010) 470–84.
[175] E.J. Wagner, P.B. Carpenter, Understanding the language of Lys36 methylation at histone H3, Nat. Rev. Mol. Cell Biol., 13 (2012) 115–26.
[176] C.-C. Pai, R.S. Deegan, L. Subramanian, C. Gal, S. Sarkar, E.J. Blaikley, C. Walker, L. Hulme, E. Bernhard, S. Codlin, J. Bähler, R. Allshire, S. Whitehall, T.C. Humphrey, A histone H3K36 chromatin switch coordinates DNA double-strand break repair pathway choice, Nat. Commun., 5 (2014) 4091.
[177] S.X. Pfister, S. Ahrabi, L.-P. Zalmas, S. Sarkar, F. Aymard, C.Z. Bachrati, T. Helleday, G. Legube, N.B. La Thangue, A.C.G. Porter, T.C. Humphrey, SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability, Cell Rep., 7 (2014) 2006–18.
[178] R.K. Slany, The molecular biology of mixed lineage leukemia, Haematologica, 94 (2009) 984–93.
[179] C.M. McLean, I.D. Karemaker, F. van Leeuwen, The emerging roles of DOT1L in leukemia and normal development, Leukemia, 28 (2014) 2131–2138.
[180] A. Patel, V. Dharmarajan, V.E. Vought, M.S. Cosgrove, On the mechanism of multiple lysine methylation by the human mixed lineage leukemia protein-1 (MLL1) core complex, J. Biol. Chem., 284 (2009) 24242–56.
[181] M.M. Steward, J.-S. Lee, A. O'Donovan, M. Wyatt, B.E. Bernstein, A. Shilatifard, Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes, Nat. Struct. Mol. Biol., 13 (2006) 852–4.
[182] F. Cao, Y. Chen, T. Cierpicki, Y. Liu, V. Basrur, M. Lei, Y. Dou, An Ash2L/RbBP5 heterodimer stimulates the MLL1 methyltransferase activity through coordinated

substrate interactions with the MLL1 SET domain, PLoS One, 5 (2010) e14102.
[183] C. Vincenz, T.K. Kerppola, Different polycomb group CBX family proteins associate with distinct regions of chromatin using nonhomologous protein sequences, Proc. Natl. Acad. Sci. U. S. A., 105 (2008) 16572–7.
[184] Z. Gao, J. Zhang, R. Bonasio, F. Strino, A. Sawai, F. Parisi, Y. Kluger, D. Reinberg, PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes, Mol. Cell, 45 (2012) 344–56.
[185] D. Pasini, A.P. Bracken, J.B. Hansen, M. Capillo, K. Helin, The polycomb group protein Suz12 is required for embryonic stem cell differentiation, Mol. Cell. Biol., 27 (2007) 3769–79.
[186] L. Tang, E. Nogales, C. Ciferri, Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription, Prog. Biophys. Mol. Biol., 102 (2010) 122–128.
[187] G. Fragola, P.-L. Germain, P. Laise, A. Cuomo, A. Blasimme, F. Gross, E. Signaroldi, G. Bucci, C. Sommer, G. Pruneri, G. Mazzarol, T. Bonaldi, G. Mostoslavsky, S. Casola, G. Testa, Cell reprogramming requires silencing of a core subset of polycomb targets, PLoS Genet., 9 (2013) e1003292.
[188] S. Chen, Y. Shi, A new horizon for epigenetic medicine?, Cell Res., 23 (2013) 326–8.
[189] F. De Santa, M.G. Totaro, E. Prosperini, S. Notarbartolo, G. Testa, G. Natoli, The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing, Cell, 130 (2007) 1083–94.
[190] Z. Nie, G. Hu, G. Wei, K. Cui, A. Yamane, W. Resch, R. Wang, D.R. Green, L. Tessarollo, R. Casellas, K. Zhao, D. Levens, c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells, Cell, 151 (2012) 68–79.
[191] P.B. Rahl, C.Y. Lin, A.C. Seila, R.A. Flynn, S. McCuine, C.B. Burge, P.A. Sharp, R.A. Young, c-Myc regulates transcriptional pause release, Cell, 141 (2010) 432–45.
[192] L. Hopp, L. Nersisyan, H. Löffler-Wirth, A. Arakelyan, H. Binder, Epigenetic Heterogeneity of B-Cell Lymphoma: Chromatin Modifiers, Genes (Basel)., 6 (2015) 1076–112.
[193] F.B. Furnari, T. Fenton, R.M. Bachoo, A. Mukasa, J.M. Stommel, A. Stegh, W.C. Hahn, K.L. Ligon, D.N. Louis, C. Brennan, L. Chin, R.A. DePinho, W.K. Cavenee, Malignant astrocytic glioma: genetics, biology, and paths to treatment, Genes Dev., 21 (2007) 2683–710.
[194] D. Sturm, S. Bender, D.T.W. Jones, P. Lichter, J. Grill, O. Becher, C. Hawkins, J. Majewski, C. Jones, J.F. Costello, A. Iavarone, K. Aldape, C.W. Brennan, N. Jabado, S.M. Pfister, Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge, Nat. Rev. Cancer, 14 (2014) 92–107.
[195] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A., 98 (2001) 5116–21.
[196] H. Dong, H. Siu, L. Luo, X. Fang, L. Jin, M. Xiong, Investigation gene and microRNA expression in glioblastoma, BMC Genomics, 11 Suppl 3 (2010) S16.
[197] J.D. Cahoy, B. Emery, A. Kaushal, L.C. Foo, J.L. Zamanian, K.S. Christopherson, Y. Xing, J.L. Lubischer, P.A. Krieg, S.A. Krupenko, W.J. Thompson, B.A. Barres, A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function, J. Neurosci., 28

(2008) 264–78.
[198] T. Lu, Y. Pan, S.-Y. Kao, C. Li, I. Kohane, J. Chan, B.A. Yankner, Gene regulation and DNA damage in the ageing human brain, Nature, 429 (2004) 883–91.
[199] L. Hopp, E. Willscher, H. Löffler-Wirth, H. Binder, Function Shapes Content: DNA-Methylation Marker Genes and their Impact for Molecular Mechanisms of Glioma, J. Can. Res. Updates, 4 (2015) 127–148.
[200] J. Lauter, E. Glimm, M. Eszlinger, Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate, Stat. Neerl., 59 (2005) 298–312.
[201] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature, 448 (2007) 553–60.
[202] A. Meissner, T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein, C. Nusbaum, D.B. Jaffe, A. Gnirke, R. Jaenisch, E.S. Lander, Genome-scale DNA methylation maps of pluripotent and differentiated cells, Nature, 454 (2008) 766–70.
[203] R. Martinez, J.I. Martin-Subero, V. Rohde, M. Kirsch, M. Alaminos, A.F. Fernandez, S. Ropero, G. Schackert, M. Esteller, A microarray-based DNA methylation study of glioblastoma multiforme, Epigenetics, 4 (2009) 255–64.
[204] T. Shinawi, V.K. Hill, D. Krex, G. Schackert, D. Gentle, M.R. Morris, W. Wei, G. Cruickshank, E.R. Maher, F. Latif, DNA methylation profiles of long- and short-term glioblastoma survivors, Epigenetics, 8 (2013) 149–56.
[205] H. Binder, L. Hopp, K. Lembcke, H. Wirth, Personalized Disease Phenotypes from Massive OMICs Data, in: IGI Global, 1AD.
[206] D. Gorovets, K. Kannan, R. Shen, E.R. Kastenhuber, N. Islamdoust, C. Campos, E. Pentsova, A. Heguy, S.C. Jhanwar, I.K. Mellinghoff, T.A. Chan, J.T. Huse, IDH mutation and neuroglial developmental features define clinically distinct subclasses of lower grade diffuse astrocytic glioma, Clin. Cancer Res., 18 (2012) 2490–501.
[207] S.B. Rothbart, B.D. Strahl, Interpreting the language of histone and DNA modifications, Biochim. Biophys. Acta, 1839 (2014) 627–43.
[208] I. Ben-Porath, M.W. Thomson, V.J. Carey, R. Ge, G.W. Bell, A. Regev, R.A. Weinberg, An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors, Nat. Genet., 40 (2008) 499–507.
[209] C.T. Watson, G. Disanto, G.K. Sandve, F. Breden, G. Giovannoni, S. V Ramagopalan, Age-associated hyper-methylated regions in the human brain overlap with bivalent chromatin domains, PLoS One, 7 (2012) e43840.
[210] G. Li, C. Warden, Z. Zou, J. Neman, J.S. Krueger, A. Jain, R. Jandial, M. Chen, Altered expression of polycomb group genes in glioblastoma multiforme, PLoS One, 8 (2013) e80970.
[211] P. Voigt, D. Reinberg, Putting a halt on PRC2 in pediatric glioblastoma, Nat. Genet., 45 (2013) 587–9.
[212] Epigenetic Dysregulation Promotes Gene Activation in Pediatric Glioma, Cancer Discov., 3 (2013) OF15–OF15.
[213] L. Chen, Y. Shi, S. Liu, Y. Cao, X. Wang, Y. Tao, PKM2: the thread linking energy metabolism reprogramming with epigenetics in cancer, Int. J. Mol. Sci., 15 (2014) 11435–45.
[214] K. Kannan, A. Inagaki, J. Silber, D. Gorovets, J. Zhang, E.R. Kastenhuber, A.

Heguy, J.H. Petrini, T.A. Chan, J.T. Huse, Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma, Oncotarget, 3 (2012) 1194–203.

[215] L. Hopp, H. Löffler-Wirth, J. Galle, H. Binder, Combined portrayal of gene expression and DNA methylation landscapes disentangles modes of epigenetic regulation in glioma, Prep., (2017).

[216] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J.P. Mesirov, P. Tamayo, The Molecular Signatures Database Hallmark Gene Set Collection, Cell Syst., 1 (2015) 417–425.

[217] C.W. Brennan, R.G.W. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S.R. Salama, S. Zheng, D. Chakravarty, J.Z. Sanborn, S.H. Berman, R. Beroukhim, B. Bernard, C.-J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S.A. Shukla, G. Ciriello, et al., The somatic genomic landscape of glioblastoma, Cell, 155 (2013) 462–77.

[218] J.O. Funk, Cell Cycle Checkpoint Genes and Cancer, in: Encycl. Life Sci., John Wiley & Sons, Ltd, Chichester, UK, 2006.

[219] W. Liu, G. Lv, Y. Li, L. Li, B. Wang, Downregulation of CDKN2A and Suppression of Cyclin D1 Gene Expressions in Malignant Gliomas, J. Exp. Clin. Cancer Res., 30 (2011) 76.

[220] M. Weller, R. Stupp, G. Reifenberger, A.A. Brandes, M.J. van den Bent, W. Wick, M.E. Hegi, MGMT promoter methylation in malignant gliomas: ready for personalized medicine?, Nat. Rev. Neurol., 6 (2010) 39–51.

[221] P.S. Ward, J.R. Cross, C. Lu, O. Weigert, O. Abel-Wahab, R.L. Levine, D.M. Weinstock, K.A. Sharp, C.B. Thompson, Identification of additional IDH mutations associated with oncometabolite R(-)-2-hydroxyglutarate production, Oncogene, 31 (2012) 2491–8.

[222] S. Turcan, D. Rohle, A. Goenka, L.A. Walsh, F. Fang, E. Yilmaz, C. Campos, A.W.M. Fabius, C. Lu, P.S. Ward, C.B. Thompson, A. Kaufman, O. Guryanova, R. Levine, A. Heguy, A. Viale, L.G.T. Morris, J.T. Huse, I.K. Mellinghoff, T.A. Chan, IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype, Nature, 483 (2012) 479–483.

[223] M.L. Suva, N. Riggi, B.E. Bernstein, Epigenetic Reprogramming in Cancer, Science (80-.)., 339 (2013) 1567–1570.

[224] H. Kretzmer, S.H. Bernhart, W. Wang, A. Haake, M.A. Weniger, A.K. Bergmann, M.J. Betts, E. Carrillo-de-Santa-Pau, G. Doose, J. Gutwein, J. Richter, V. Hovestadt, B. Huang, D. Rico, F. Jühling, J. Kolarova, Q. Lu, C. Otto, R. Wagener, J. Arnolds, et al., DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control, Nat. Genet., 47 (2015) 1316–25.

[225] T. Gerber, E. Willscher, H. Loeffler-Wirth, L. Hopp, D. Schadendorf, M. Schartl, U. Anderegg, G. Camp, B. Treutlein, H. Binder, M. Kunz, T. Gerber, E. Willscher, H. Loeffler-Wirth, L. Hopp, D. Schadendorf, M. Schartl, U. Anderegg, G. Camp, B. Treutlein, et al., Mapping heterogeneity in patient-derived melanoma cultures by single-cell RNA-seq, Oncotarget, 5 (2016).

[226] M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D.J. Weisenberger, M. Campan, J. Young, I. Jacobs, P.W. Laird, Epigenetic stem cell signature in cancer, Nat. Genet., 39 (2007) 157–158.

[227] D.M. Roy, L.A. Walsh, T.A. Chan, Driver mutations of cancer epigenomes, Protein Cell, 5 (2014) 265–296.

[228] T. Jenuwein, C.D. Allis, Translating the histone code, Science, 293 (2001) 1074–80.
[229] S.X. Pfister, E. Markkanen, Y. Jiang, S. Sarkar, M. Woodcock, G. Orlando, I. Mavrommati, C.-C. Pai, L.-P. Zalmas, N. Drobnitzky, G.L. Dianov, C. Verrill, V.M. Macaulay, S. Ying, N.B. La Thangue, V. D'Angiolella, A.J. Ryan, T.C. Humphrey, Inhibiting WEE1 Selectively Kills Histone H3K36me3-Deficient Cancers by dNTP Starvation, Cancer Cell, 28 (2015) 557–568.
[230] N. Chambwe, M. Kormaksson, H. Geng, S. De, F. Michor, N.A. Johnson, R.D. Morin, D.W. Scott, L.A. Godley, R.D. Gascoyne, A. Melnick, F. Campagne, R. Shaknovich, Variability in DNA methylation defines novel epigenetic subgroups of DLBCL associated with different clinical outcomes, Blood, 123 (2014) 1699–708.
[231] S. De, R. Shaknovich, M. Riester, O. Elemento, H. Geng, M. Kormaksson, Y. Jiang, B. Woolcock, N. Johnson, J.M. Polo, L. Cerchietti, R.D. Gascoyne, A. Melnick, F. Michor, Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity, PLoS Genet., 9 (2013) e1003137.
[232] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics, 26 (2010) 1572–3.
[233] A. Meissner, Epigenetic modifications in pluripotent and differentiated cells, Nat. Biotechnol., 28 (2010) 1079–88.

Curriculum vitae

Name:	Lydia Hopp
Date of birth:	15.07.1987
Place of birth:	Werdau, Germany
Nationality:	German

EDUCATION

Since 10/2014	PhD student Interdisciplinary Centre for Bioinformatics - IZBI, University of Leipzig
10/2011 - 10/2014	PhD student Research group 'Systems medicine', LIFE, University of Leipzig, Interdisciplinary Centre for Bioinformatics - IZBI, University of Leipzig
12/2010 - 10/2011	Research fellow Interdisciplinary Centre for Bioinformatics - IZBI, University of Leipzig
07/2008 - 11/2008	Study abroad Massey University Palmerston North, New Zealand
10/2005 - 10/2010	Diploma in Biomathematics Ernst- Moritz- Arndt- Universität Greifswald Main courses: Analysis/ Mathematical optimization, Molecular biology

List of publications

*: Contributed equally

Publications included in this thesis:

Hopp, L.*, Wirth, H.*, Fasold, M. & Binder, H.: Portraying the expression landscapes of cancer subtypes: A glioblastoma multiforme and prostate cancer case study. *Systems Biomedicine* 2013

Hopp, L., Lembcke, K., Binder, H. & Wirth, H.: Portraying the Expression Landscapes of B-Cell Lymphoma- Intuitive Detection of Outlier Samples and of Molecular Subtypes. *Biology* 2013

Hopp, L., Willscher, E., Löffler-Wirth, H. & Binder, H.: Function Shapes Content: DNA-Methylation Marker Genes and their Impact for Molecular Mechanisms of Glioma. *Journal of Cancer Research Updates* 2015

Hopp, L., Nersisyan, L., Löffler-Wirth, H., Arakelyan, A. & Binder, H.: Epigenetic Heterogeneity of B-Cell Lymphoma: Chromatin Modifiers. *Genes* 2015

Hopp, L., Löffler-Wirth, H. & Binder, H.: Epigenetic heterogeneity of B-cell lymphoma: DNA-methylation, gene expression and chromatin states. *Genes* 2015

Hopp, L., Löffler-Wirth, H. , Galle, J. & Binder, H.: Combined portrayal of gene expression and DNA methylation landscapes disentangles modes of epigenetic regulation in glioma. In Preparation 2017

Publications not included in this thesis:

Binder, H., Fasold, M., Hopp, L., Cakir, V., Bergen, M.v. & Wirth, H.: Molecular phenotypic portraits - exploring the 'OMEs' with individual resolution. *HIBIT conference 2011 proceedings*

Steiner, L.*, Hopp, L.*, Wirth, H., Galle, J., Binder, H., Prohaska, S. & Rohlf, T.: A global genome segmentation method for exploration of epigenetic patterns. *PLoS ONE* 2012

Wirth, H.*, Cakir, V.*, Hopp, L. & Binder, H.: Analysis of miRNA expression using machine learning. *Methods in Molecular Biology* 2014

Cakir, V.*, Wirth, H.*, Hopp, L. & Binder, H.: miRNA expression landscapes in stem cells, tissues and cancer. *Methods in Molecular Biology* 2014

Binder, H., Hopp, L., Lembcke, K., Wirth, H.: Personalized Disease Phenotypes from Massive OMICs Data. In B. Wang, R. Li, & W. Perrizo (Eds.) Big Data Analytics in Bioinformatics and Healthcare (pp. 359-378). Hershey, PA: *Medical Information Science Reference*. 2015

Keller, M., Hopp, L., Liu, X., Wohland, T., Rohde, K., Cancello, R., et al.: Genome-wide methylome and transcriptome analysis in human adipose tissue unravels novel candidate genes for obesity. *Molecular Metabolism* 2016

Hamidouche, Z., Rother, K., Przybilla, J., Krinner, A., Clay, D., Hopp, L., et al.: Bistable epigenetic states explain age-dependent decline in mesenchymal stem cell heterogeneity. *Stem Cells* 2016

Gerber, T.*, Willscher, E.*, Loeffler-Wirth, H., Hopp, L., Schadendorf, D., Schartl, M., et al.: Mapping heterogeneity in patient-derived melanoma cultures by single-cell RNA-seq. *Oncotarget* 2016

Löffler-Wirth, H., Kreuz, M., Hopp, L., …, Siebert, R., Löffler, M., et al.: A holistic view on the expression landscape of mature aggressive B-cell lymphoma. In Preparation 2017

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 27.01.2017

Lydia Hopp