Statistical analysis of microarray based DNA methylation data

vorgelegt von
Diplom-Ingenieur
Fabian Model
aus Berlin

von der Fakultät IV für Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Herr Prof. Dr. H. Ehrig
Gutachter: Herr Prof. Dr. U. Kockelkorn
Herr Prof. Dr. M. Opper
Herr Prof. Dr. M. Ebert

Tag der wissenschaftlichen Aussprache: 12.07.2007

Berlin 2007
D 83

Abstract

Over the past few years interest in epigentic mechanisms, especially DNA methylation, has increased dramatically. The fundamental importance of epigenetic changes has been established, particularly in oncology. Aberrant DNA methylation occurs early in oncogenesis, is stable, and can be assayed in tissues and body fluids. Therefore genes with aberrant methylation can provide clues for understanding tumor pathways and are attractive candidates for detection of early neoplastic events. However, large-scale analysis of candidate genes has been hampered by the lack of high throughput assays for methylation detection. The introduction of the first microarray for DNA methylation analysis addressed this problem by allowing the measurement of several hundred selected CpG dinucleotides in parallel. DNA microarray technology has already revolutionized mRNA expression analysis. It also introduced a plethora of statistical problems such as control and maintenance of data quality and handling of high dimensional and usually under-determined marker selection or classification problems.

In this thesis novel statistical methods for the analysis of DNA methylation microarray data are developed. Starting from a simple generative model of the microarray measurement process algorithms for normalization, variance stabilization and DNA methylation rate estimation are derived. These pre-processing methods allow for an optimal estimation of DNA methylation patterns from the microarray hybridization intensities of a given biological specimen. A methodology for microarray quality and process control is introduced that estimates the quality of individual microarrays based solely on the distribution of the actual measurements without requiring repeated experiments. It can be used to reliably detect systematic experimental errors resulting in an improvement of overall data quality. Subsequently it is demonstrated how phenotypic classes can be predicted from microarray measurements by combining feature selection and discriminant analysis. By comparing several feature selection methods it is shown that the right dimension reduction strategy is of crucial importance for the classification performance. Methods for DNA microarray quality control, feature selection and class prediction are derived in a generic fashion that makes them equally applicable to DNA methylation and mRNA expression microarray data.

The developed methods are applied in a large microarray study to identify DNA methylation markers specific for colorectal neoplasia. In this study 43 candidate genes were probed with DNA from 89 colorectal adenocarcinomas, 55 colorectal polyps, 31 inflammatory bowel disease, 115 extracolonic

cancers, and 67 healthy tissues. The 20 most discriminating markers are highly methylated in colorectal neoplasia ($AUC > 0.8$; $P < 0.0001$). Normal epithelium and extracolonic cancers reveal significantly lower methylation. Results are validated on an independent sample set by real-time PCR. The discovered markers with high specificity for colorectal cancer have potential as blood-based screening markers whereas markers that are specific for multiple cancers could potentially be used as prognostic indicators or biomarkers for therapeutic response monitoring. The results clearly demonstrate that DNA methylation microarrays in combination with the developed analysis methods constitute a valuable tool for the discovery of novel epigenetic tumor markers and DNA methylation research in general.

Keywords:
DNA methylation, Microarray, Data analysis, Colorectal cancer

Zusammenfassung

Innerhalb der letzten Jahre hat das Interesse an epigenetischen Mechanismen, insbesondere der DNA-Methylierung, dramatisch zugenommen. Die fundamentale Bedeutung epigenetischer Veränderungen wurde insbesondere in der Onkologie etabliert. Aberrierende DNA-Methylierung entsteht in einem frühen Stadium der Onkogenese, ist stabil und kann in Geweben und Körperflüssigkeiten nachgewiesen werden. Daher können Gene mit aberrierender DNA-Methylierung Hinweise zum Verständnis von Signaltransduktionswegen in Tumoren liefern und sind attraktive Kandidaten für die Detektion früher neoplastischer Veränderungen. Allerdings wurde eine groß angelegte Analyse von Kandidatengenen durch einen Mangel an Hochdurchsatzmethoden zur Methylierungsmessung gehemmt. Die Einführung des ersten Microarrays zur Messung von DNA-Methylierung hat dieses Problem gelöst indem es die gleichzeitige Messung mehrerer hundert ausgewählter CpG-Dinukleotide erlaubt. DNA-Microarray-Technologie hat bereits die Analyse von mRNA Expression revolutioniert. Sie hat allerdings auch eine Unmenge statistischer Probleme wie die der Qualitätskontrolle, der Markerselektion und der Klassifikation in hochdimensionalen Datenräumen aufgeworfen.

In dieser Arbeit werden neuartige statistische Methoden zur Datenanalyse von DNA-Methylierungs-Microarrays entwickelt. Ausgehend von einem einfachen generativen Modell des Microarray-Messprozesses werden Algorithmen zur Normalisierung, Varianzstabilisierung und Bestimmung der DNA-Methylierungsrate hergeleitet. Diese Vorverarbeitungsmethoden erlauben eine optimale Schätzung der DNA-Methylierungsmuster einer gegebenen Probe aus den Microarray-Hybridisierungsintensitäten. Es wird eine Methodik zur Qualitäts- und Prozesskontrolle eingeführt, die es erlaubt die Qualität individueller Microarrays nur auf der Basis der eigentlichen Messwerte und ohne zusätzliche replizierte Experimente zu bestimmen. Dies erlaubt systematische experimentelle Fehler zuverlässig zu detektieren und damit die Datenqualität zu erhöhen. Weiterhin wird gezeigt wie phenotypische Klassen auf der Basis von Microarraymesswerten vorhergesagt werden können indem Verfahren der Merkmalsselektion und Diskriminanzanalyse verbunden werden. Durch den Vergleich verschiedener Merkmalsselektionsverfahren wird gezeigt, dass die richtige Strategie zur Dimensionsreduktion von entscheidender Bedeutung für eine gute Klassifikationsleistung ist. Die vorgestellten Methoden zur Qualitätskontrolle, Merkmalsselektion und Klassifikation sind so generisch, dass sie sowohl auf DNA-Methylierungs- als auch mRNA-Microarrays anwendbar sind.

Die entwickelten Methoden werden auf eine große Microarraystudie zur Identifizierung von DNA-Methylierungsmarkern für Kolorektalkarzinome angewandt. In dieser Studie wurden 43 Kandidatengene auf DNA von 89 kolorektalen Adenokarzinomen, 55 kolorektalen Polypen, 31 chronisch entzündlichen Darmerkrankungen, 115 nicht kolorektalen Karzinomen und 67 gesunden Gewebeproben gemessen. Die 20 meistdiskriminierenden Marker sind hochgradig methyliert in kolorektalen Neoplasien ($AUC > 0.8$; $P < 0.0001$). Normales Epithelium und nicht kolorektale Karzinome zeigen signifikant geringere Methylierung. Die Resultate wurden mittels real-time PCR auf einem Satz unabhängiger Gewebeproben validiert. Die entdeckten Markergene mit hoher Spezifitaet für kolorektale Karzinome sind potentielle Marker für einen blutbasierten Früherkennungstest. Markergene die spezifisch für mehrere Arten von Karzinomen sind könnten als prognostische Indikatoren oder Biomarker für die Therapieüberwachung benutzt werden. Die Resultate zeigen klar, dass DNA-Methylierungsmicroarrays in Kombination mit den entwickelten Analysemethoden ein äußerst wertvolles Werkzeug zur Entdeckung neuer Tumormarker und zur Erforschung von DNA-Methylierung im Allgemeinen darstellen.

Schlagwörter:
DNA-Methylierung, Microarray, Datenanalyse, Kolorektalkarzinom

Contents

1 Introduction **1**
- 1.1 DNA methylation . . . 2
  - 1.1.1 Biology . . . 2
  - 1.1.2 Cancer diagnostics . . . 5
  - 1.1.3 Measurement of DNA methylation . . . 7
- 1.2 Analysis of DNA microarray data . . . 11
  - 1.2.1 Microarray technology . . . 11
  - 1.2.2 Preprocessing . . . 11
  - 1.2.3 Quality control . . . 13
  - 1.2.4 Data interpretation . . . 14
- 1.3 Objectives and outlook . . . 18

2 Measuring DNA methylation **19**
- 2.1 Measurement process . . . 21
  - 2.1.1 Sample preparation . . . 21
  - 2.1.2 Microarray preparation . . . 28
  - 2.1.3 Hybridization and image analysis . . . 33
- 2.2 A statistical model of hybridization . . . 36
  - 2.2.1 Within chip noise . . . 36
  - 2.2.2 Between chip noise and normalization . . . 43
  - 2.2.3 Expected hybridization intensities . . . 48
- 2.3 Quantification of DNA methylation . . . 56
  - 2.3.1 Methylation scores . . . 56
  - 2.3.2 Ratios and differences of CG and TG oligos . . . 58
  - 2.3.3 A maximum likelihood estimator . . . 62
- 2.4 Results . . . 67

3 Controlling quality **74**
- 3.1 Microarray data and typical sources of error . . . 76
- 3.2 Detecting outlier chips with robust PCA . . . 77
  - 3.2.1 Methods . . . 77

3.2.2 Results . . . 81
3.3 Statistical process control . . . 81
3.3.1 Methods . . . 81
3.3.2 Results . . . 85

4 Class prediction and feature selection 89
4.1 Support Vector Machines . . . 90
4.2 Feature selection . . . 91
4.2.1 Principle Component Analysis . . . 92
4.2.2 Fisher criterion and t-test . . . 95
4.2.3 Backward elimination . . . 97
4.2.4 Exhaustive search . . . 97

5 Identification of CRC methylation markers 99
5.1 Materials and methods . . . 101
5.1.1 Patient samples . . . 101
5.1.2 DNA extraction . . . 102
5.1.3 Genome-wide identification of differentially methylated sequences . . . 102
5.1.4 Gene array . . . 103
5.1.5 MethyLight assays . . . 103
5.1.6 Statistical analysis . . . 104
5.2 Results . . . 104
5.2.1 Genome-wide discovery . . . 104
5.2.2 Gene array study . . . 105
5.2.3 Marker validation with MethyLight assays . . . 113

6 Discussion 116
6.1 Measuring DNA methylation . . . 116
6.2 Controlling quality and stability of microarray experiments . . . 117
6.3 Class prediction and feature selection . . . 118
6.4 Identification and validation of colorectal neoplasia-specific methylation markers . . . 118
6.5 Conclusions . . . 120

Bibliography 121

A Datasets 139
A.1 Methylation estimation . . . 139
A.2 Quality control . . . 139
A.3 Class prediction . . . 140

A.4 Marker selection 142

B List of symbols **143**

Chapter 1

Introduction

Tremendous progress has been made in molecular genetics since Watson and Crick published the double helix structure of DNA in 1953 [165]. Fifty years later the complete sequence of the human genome with its more than 3 billion bases is known and the annotation of known and predicted genes is stabilizing [84, 24].

Building on this knowledge scientific focus moves to understanding gene function and regulation. A thorough understanding of how the DNA code is interpreted and translated into RNA and proteins is especially essential for biomedical research since many human diseases are associated with alterations in gene sequence, gene expression, protein structure and protein modifications.

The focus of many genomic research studies is the investigation of messenger RNA (mRNA) or protein concentration in cells and tissues under varying conditions [67, 4, 105, 128]. A whole toolbox of new technologies has been developed to facilitate time and cost efficient experiments in this area. One of the technologies with the highest impact on modern research is the DNA microarray. It enables investigators to measure mRNA expression of several thousand transcripts in parallel. The rapid development of this technology has resulted in large, complex datasets but statistical methods to analyze them are not well established. A minimum consensus on how to solve the most basic problems in microarray data analysis has evolved over the last few years but most problems remain topics of active research [5].

The DNA sequence gives the blueprint for all possible states of a cell in terms of sequences that could be transcribed into mRNA and translated into proteins. RNA expression and protein analysis give a snapshot of this cell state at one point in time. In between the DNA sequence information, which is constant for an individual, and the amounts of generated mRNAs

and proteins which vary for every cell and over time, complex organisms have an additional *epigenetic* layer of information.

The term *epigenetics* defines all meiotically and mitotically heritable changes in gene expression that are not coded in the DNA sequence itself [46]. Epigenetics can, for instance, explain why the different cell types of an organism share identical DNA sequences but show broad morphological and functional diversity.

Methylation of DNA is the most extensively studied of epigenetic mechanisms, and is associated with a wide range of critical biological processes. In this thesis we will develop statistical methods that will allow us to measure and interpret DNA methylation patterns with the help of DNA microarrays.

1.1 DNA methylation

1.1.1 Biology

DNA methylation in vertebrates is a chemical modification of the cytosine nucleotide in which the 5-carbon position is enzymatically modified by the addition of a methyl group, such that cytosines can occur in a methylated or unmethylated state (see Fig. 1.1). Methylation of cytosines in higher eukaryotes occurs only in the sequence context of cytosine followed by guanine, a CpG dinucleotide, and is the only genetically programmed DNA modification in mammals.

The CpG dinucleotide is underrepresented in the human genome, likely because methylated cytosines are prone to deamination producing thymine, resulting in a G/T mismatch. This mutagenic property is postulated to have driven CpG depletion during evolution. Most of the CpG dinucleotides in the human genome are methylated (between 60-70%). However, CpG rich clusters of between three hundred and several thousand base pairs, so called CpG islands, are found close to the 5' regulatory regions of many genes and are generally not methylated. CpG islands that have a majority of their CpG dinucleotides unmethylated are referred to as *hypomethylated* whereas islands with a majority of methylated CpGs are called *hypermethylated*.

Hypermethylation of a CpG island is usually associated with transcriptional silencing of the neighboring gene (see Fig. 1.2). The symmetrical addition of the methyl group changes the appearance of the major groove of the double helix and directly influences transcription by altering the binding of sequence specific transcription factors, repressors and insulators [47]. An indirect reinforcement of the transcriptionally silent state is mediated by proteins that can bind to methylated CpGs. These proteins, which are

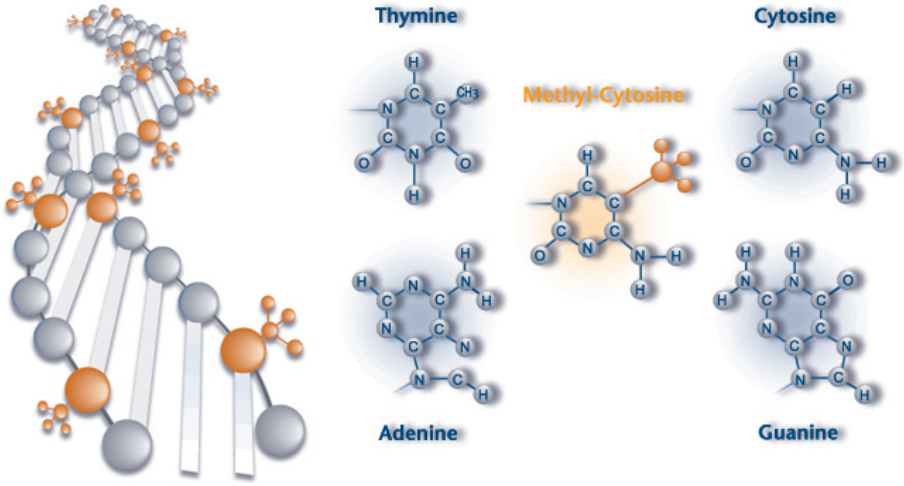

Figure 1.1: The four plus one bases. The DNA double helix is composed of 4 bases: adenine (A), thymine (T), cytosine (C), and guanine (G). Each base on one strand forms a bond with just one kind of base on the other strand, called a "complementary" base: A bonds with T, and C bonds with G. A special form of cytosine, the 5-methylcytosine, carries the methylation information. In higher eukaryotes it only occurs in the sequence context of guanine as CpG dinucleotide. The complementary CpG dinucleotides on the two strands have usually identical methylation status.

called methyl-CpG binding proteins, recruit histone deacetylases and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin termed heterochromatin [73, 162]. However, methylation does not cause transcriptional silencing in every case. When a negative regulatory element such as a silencer is hypermethylated expression of the associated gene can actually increase [91]. Furthermore, there is a group of genes that appear not to be regulated by DNA methylation, since their promoter regions are hypomethylated in all cell types independent of transcriptional activity [13].

DNA methylation has been shown to play a key role in the following genetic mechanisms:

- Tissue differentiation. Cell specific methylation plays a key role in the differentiation of cell types [6, 108].

- Silencing of repetitive elements and endogenous transposons [173].

- X chromosome inactivation. The silencing of one X chromosome in all human female cells is associated with DNA methylation. In this case hypermethylation of the complete X chromosome acts synergistically with a noncoding RNA from the Xist gene. Activity of the other X

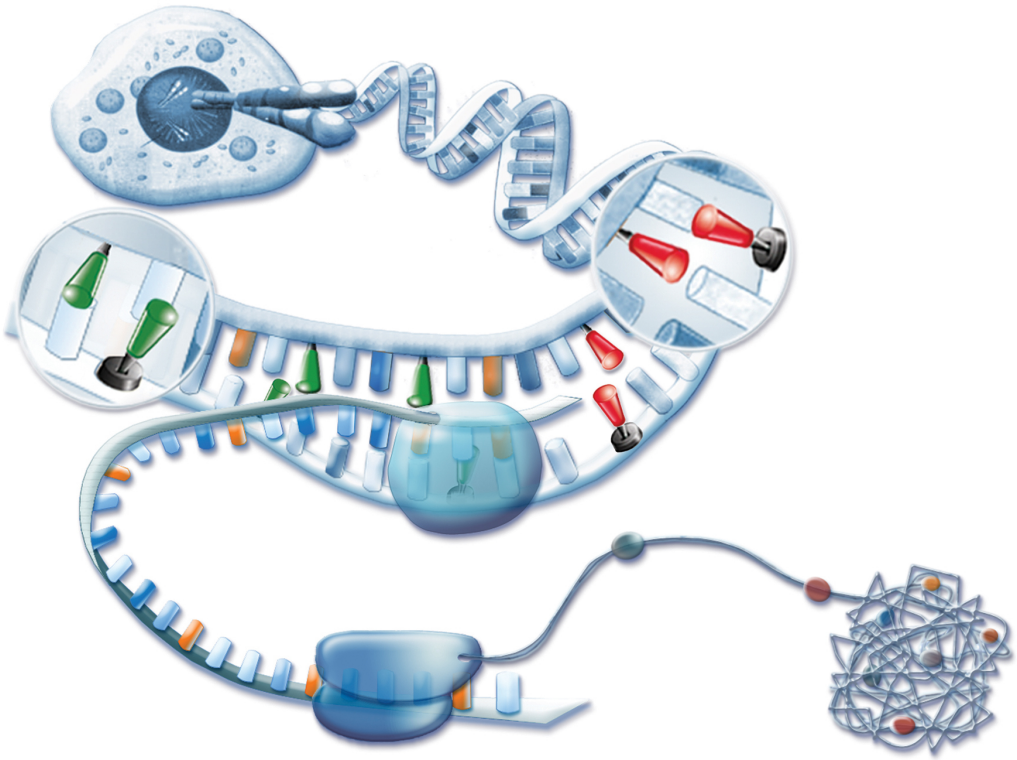

Figure 1.2: DNA methylation as an epigenetic switchboard for gene deactivation. The figure depicts the DNA double helix of one chromosome in the cell nucleus with the CpG methylation status symbolized by red (hypermethylated) and green (hypomethylated) switches. In hypomethylated areas transcription factors can bind and transcribe the respective gene into messenger RNA (mRNA) that in turn gets translated into a protein. For most genes the hypermethylation of their associated CpG island results in transcriptional silencing. In some cases hypermethylation of a negative regulating element such as a silencer can result in transcriptional activation. In both scenarios DNA methylation effectively turns on or off the transcription of a gene into mRNA and thus also controls the generation of the associated protein.

chromosome is ensured by transcriptional silencing of its Xist gene via hypermethylation [127].

- Imprinting. Hypermethylation of either the paternal or maternal allele causes asymmetric expression of some genes in a parent of origin specific manner [133].

- Gene - environment interaction. It has been shown that DNA methylation patterns are changed by environmental effects like exposure to xenobiotics during mammalian development [102], different diets [164], and stress [143].

Experimental evidence shows that DNA methylation is essential for embryogenesis and development in mammals [132]. It is maintained and propagated to new cell generations by DNA methyl transferases (DNMT) [12]. The exact mechanism of how methylation patterns are initially established during implantation of the zygote and later regulated is still unknown.

1.1.2 Cancer diagnostics

The medical significance of DNA methylation is illustrated in a number of human carcinomas, for which dramatic changes of DNA methylation patterns have been reported for tumors compared to normal tissues or cells. The most common alterations are a genome wide hypomethylation and gene specific hypermethylation. Genome wide hypomethylation mainly affects repetitive sequences in satellite DNA and centromeres causing a general loss of genome stability [89, 125].

Silencing of tumor suppressor genes by promoter hypermethylation usually affects genes involved in DNA repair, detoxification, cell cycle regulation or apoptosis [53, 74, 92, 54]. Knudson's two hit hypothesis postulates that for the development of a malignant cell both alleles of a tumor suppressor gene have to be inactivated [97]. Promoter hypermethylation leading to gene silencing can be one of those hits. Together with other events like mutation or loss of heterozygosity (LOH) promoter hypermethylation can completely deactivate a tumor supressor gene and cause malignancy of a cell [70]. Since in contrast to genetic mutations, epigenetic alterations of tumor DNA are potentially reversible they could be interesting targets for future therapeutics [94, 46].

An application of DNA methylation that is realizable in the near future is the development of biomarkers for diagnosis of cancer. In particular the hypermethylation of specific tumor suppressor genes has considerable advan-

tages compared to tumor markers based on single nucleotide polymorphisms (SNPs), mRNA or protein analysis:

- Promoter hypermethylation occurs early in tumoro-genesis and can be specific for certain tumor types.
- Hypermethylation of certain genes does not exist in normal cells. For these markers hypermethylation is a distinct qualitative and specific sign of malignancy and can be detected in a background of normal cells with high sensitivity.
- Compared to mRNA and protein measurements methylation patterns are very stable over time.
- Methylation is a chemically stable modification of DNA and is not affected by typical histopathological treatments such as paraffin embedding.
- Methylation can be absolutely quantified in relation to the total amount of DNA. This enables easy comparison between different measurements.
- The methylation signal is easily amplifiable via PCR.
- In contrast to single nucleotide polymorphisms (SNPs) DNA methylation signals occur at distinct and well defined genomic locations.

Therefore DNA methylation analysis can be used for a variety of applications in cancer diagnosis. One is the classification of tissue samples taken either from a biopsy of a suspicious lesion or from a surgically removed tumor [1, 115, 117, 111]. Typical diagnostic questions that have to be answered based on these tissue samples are:

- Malignancy
  Is the tumor benign or malignant?
- Prognosis
  How aggressive is the tumor? Will the patient have a relapse after surgery?
- Prediction of therapy response
  How will the tumor respond to a certain treatment? Is a particular chemo-therapy necessary? How much will it improve the patient's odds for not having a relapse?

Technically, fresh frozen or paraffin embedded tissue samples are the optimal source material for methylation analysis since they provide sufficient amounts of DNA that comes almost completely from the tumor tissue of interest. The disadvantage is that these samples usually require an invasive procedure that carries a certain risk, is unpleasant for the patient and of course that the tumor has to be actually diagnosed and located.

Another application of DNA methylation analysis is the detection of cancer in remote samples. Due to their uncontrolled growth and high rate of cell necrosis tumors can shed relatively high amounts of their DNA into body fluids such as blood or urine [100, 159]. By using sensitive detection methods that can identify methylated tumor DNA biomarkers in an excess of normal DNA, it is possible to diagnose cancer based on a simple blood or urine sample test. This kind of analysis does not require any invasive procedure, is very convenient for the patient, and therefore promises a high compliance in screening programs aimed at asymptomatic populations. Since many cancers are curable when detected early, population wide cancer screening promises a dramatic reduction in mortality and is the most promising way for fighting this disease.

A third application of DNA methylation in cancer diagnostics is the identification of patients that are at risk of developing a cancer over the course of their lives. This kind of predisposition can be caused by a loss of imprinting (LOI). An example is the gene *Insulin Growth Factor II* (IGF2) that is usually methylated on the maternal allele - resulting in expression of only the paternal allele. The loss of maternal imprinting is found in children with Wilms tumors [124] and it has been shown that loss of IGF2 imprinting increases the risk to develop colorectal cancer [36, 35]. Since LOI is a defect that arises during germline development, it is present in all patient cells and can be conveniently detected in blood.

In this thesis we will present data from different areas of cancer diagnostics. The final part of the thesis will focus on the identification of DNA methylation markers for the early detection of colorectal cancer.

1.1.3 Measurement of DNA methylation

For the analysis of DNA methylation, sensitive and quantitative methods are needed to detect even subtle changes in the degree of methylation, as biological samples often represent a heterogeneous mixture of different cells, e.g. tumor and non-tumor cells. A variety of techniques for the study of DNA methylation have been developed over the last years [61, 99, 149, 139]. All methods have different advantages and disadvantages with regard to quantitative accuracy, sensitivity, genome coverage and precise investigations of

individual CpG positions (see Fig. 1.3). Therefore the choice of method mainly depends on the desired application. DNA methylation measurement techniques can be roughly classified into methods analysing the total amount of methylcytosine in a sample, those based on methylation sensitive enzymatic digestion of genomic DNA and those relying on bisulphite conversion.

One of the most widely used techniques for the monitoring of global changes in the methylation level is HPLC following a quantitative hydrolysis of the DNA sample to single nucleotides [48]. Increased sensitivity with smaller amounts of DNA can be achieved by capillary electrophoresis or mass spectrometry [62, 7, 63]. In situ hybridisation methods with methylcytosine specific antibodies allow the detection of methylated sequences on a cell to cell basis [114]. However, since global methylation analysis is per definition completely unspecific it is not usable for most diagnostic purposes.

Traditionally methylation patterns have been analyzed by digestion of genomic DNA with methylation sensitive restriction endonucleases and subsequent detection by Southern blotting or PCR amplification [141]. Restriction landmark genomic scanning (RLGS) permits the genome wide quantitative assessment of epigenetic alterations between samples by digestion with a methylation sensitive enzyme and subsequent radio labeling of the created endonuclease sites [31, 137]. Differential methylation is analysed by comparing spot intensities on a two dimensional gel. Methods like methylated CpG island amplification [150], methylation sensitive arbitrarily primed PCR [68] and differential methylation hybridization [171] compare genome wide methylation patterns between two samples or two pools of samples. They are based on the restriction digest of DNA with methylation sensitive enzymes, followed by size fragmentation, PCR amplification and comparative analysis of hybridisation patterns to a microarray with DNA probes for CpG islands or gel spot patterns. Due to their dependence on restriction sites accessible to methylation sensitive restriction enzymes, only CpG sites found within these sequences can be analyzed and incomplete cleavage may give rise to false positive results. Nevertheless restriction based analysis methods are an excellent tool for the genome wide discovery of CpG sites that are differentially methylated for a given diagnostic question.

The introduction of sodium bisulphite conversion of genomic DNA has revolutionized the field of DNA methylation analysis [64]. Bisulphite treatment of genomic DNA samples results in the hydrolytic deamination of nonmethylated cytosines to uracils, while methylated cytosines are resistant to conversion [163]. After PCR amplification the methylation status at a given position is manifested in the ratio C (former methylated cytosine) to T (former nonmethylated cytosine) and can be analyzed as a virtual C/T polymorphism in the bisulphite treated DNA.

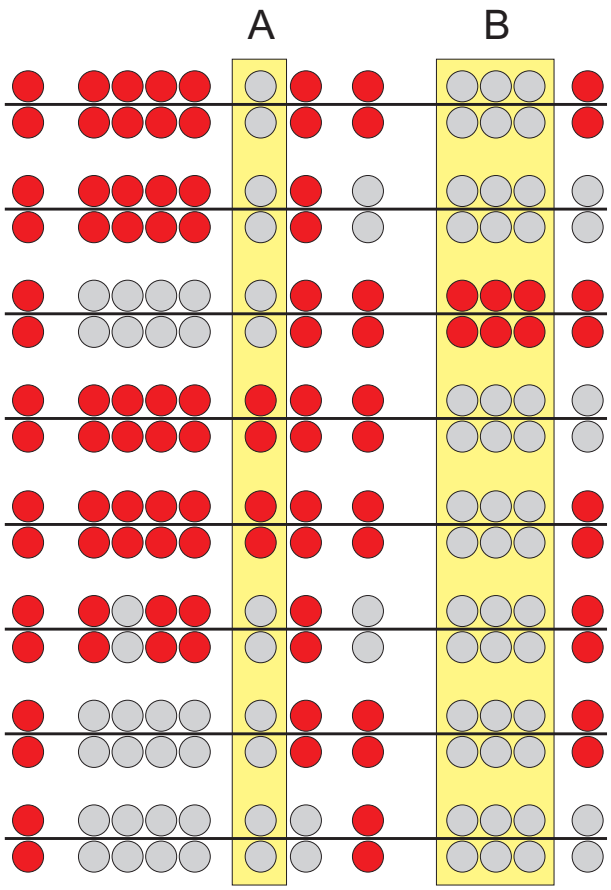

Figure 1.3: Principles of DNA methylation analysis. A genomic DNA sample usually consists of a heterogeneous mix of DNA molecules that are derived from many different cells. In this figure, each horizontal bar represents an entire double stranded haploid genome. Eight such haploid genomes are aligned above each other. Circles represent cytosine residues in context of CpG dinucleotides on the top or bottom strand of the DNA double helix. Methylated cytosines are represented by red circles, unmethylated cytosines by gray circles. DNA methylation analysis methods measure either each individual CpG methylation status (cloned bisulphite sequencing), the amount of methylated cytosines at one CpG position - e.g. the number of red dots in column A (methylation sensitive restriction methods, direct bisulphite sequencing, methylation microarrays), the amount of specific cytosine methylation patterns at one set of CpG positions - e.g. the number of completely red blocks in columns B (MSP, methylation microarrays) or the overall amount of cytosine methylation - i.e. the total number of red dots (global methylation analysis).

A commonly applied method for the assessment of the methylation status is either direct sequencing or sequencing of subclones of bisulphite treated DNA [64, 95]. It is so far the only method that allows a thorough analysis of multiple, closely neighboring CpG positions. Cloned bisulphite sequencing can be regarded as the gold standard of methylation analysis since it enables the measurement of the methylation status of every individual CpG dinucleotide in a sample (see Fig. 1.3). However, cloning is extremely labor intensive and costly and thus not suitable for large numbers of samples or genomic locations. Direct bisulphite sequencing is an efficient alternative but has relatively low accuracy and sensitivity [101].

Another popular method for the analysis of bisulphite converted DNA is methylation-specific PCR (MSP). It permits the amplification of small blocks of CpG sites with three pairs of primers for amplification, complementary to the methylation pattern of interest (either methylated, nonmethylated or a mixture) as well as a control for complete bisulphite conversion [75]. The main advantage of MSP is the high sensitivity that enables the detection of the target allele in the presence of a huge excess of other alleles and the detection of differentially methylated positions in body fluids [75, 55]. The biased amplification makes quantitation in a variable background difficult. Quantitation is improved by fluorescence based real-time PCR assays like MethyLight [45] or HeavyMethyl [32].

The detection method this thesis will focus on is the analysis of bisulphite converted DNA by hybridization onto microarrays[66, 1, 115]. In this technology selected genes are amplified by PCR from the bisulphite treated DNA using fluorescently labeled primers. Unmethylated CpG dinucleotides are converted to TG and originally methylated CpG sites are conserved. Pairs of PCR primers are multiplexed and designed complementary to DNA segments containing no CpG dinucleotides. This allows unbiased amplification of many alleles in one reaction. All PCR products from an individual sample are then mixed and hybridized to glass slides carrying a pair of immobilized oligonucleotides for each CpG position. Each of these detection oligonucleotides is designed to hybridize to the bisulphite converted sequence around a specific CpG site which was originally either unmethylated (TG) or methylated (CG). Hybridization conditions are selected to allow the detection of the single nucleotide differences between the TG and CG variants. Oligonucleotide hybridization intensities can then be used to derive the proportion of methylated CpG dinucleotides at the respective genomic locations.

1.2 Analysis of DNA microarray data

1.2.1 Microarray technology

A DNA microarray (also referred to as gene chip, DNA chip, or simply just chip) is a collection of multiple DNA segments attached to a solid surface, such as glass, plastic or silicon chip forming an array for the purpose of measuring DNA or mRNA concentrations. The affixed DNA segments are known as probes (also referred to as oligomeres or spots), thousands of which can be used in a single DNA microarray [146]. Each probe is designed to match a specific sequence of the target DNA or mRNA transcript. By observing the hybridization intensities of extracted and fluorescently labeled DNA or mRNA from a biological specimen binding to these probes it is possible to look at the sequence, the methylation status or the mRNA expression level of thousands of genes at once. Therefore DNA microarrays are one of the most popular technologies in molecular biology today [105]. Applications of microarray technology include marker identification, tissue classification, and discovery of new tissue subtypes [67, 4, 1].

There are two principle types of DNA microarrays. In two channel microarrays (sometimes for historic reasons also referred to as spotted microarrays, even though both types of microarrays can be spotted), the probes are synthesized oligonucleotides, cDNA (reverse transcribed DNA copies of mRNA) or small fragments of PCR products corresponding to mRNAs. This type of array is typically hybridized with cDNA from two samples to be compared (e.g. patient and control) that are labeled with two different fluorophores. The samples can be mixed and hybridized to one single microarray that is then scanned, allowing the visualization of up- and down-regulated genes in one experiment [140]. In single channel microarrays, the probes are usually oligonucleotides that are designed to match parts of the sequence of known or predicted mRNAs. These microarrays give estimations of the absolute value of gene expression and the comparison of two conditions requires the use of two separate microarrays [60].

Recently it has been shown that microarrays can also be used to detect DNA methylation and that results are comparable to mRNA expression analysis [66, 1, 115].

1.2.2 Preprocessing

Prior to the biological interpretation of the data a number of preprocessing transformations are usually applied to the raw measurement values from a microarray experiment. The major goals of these preprocessing steps are the

minimization of noise caused by technical variations or systematic errors and the transformation of the data into a format suitable for analysis.

Normalization

Typically, the first transformation applied to mRNA expression data is normalization. It adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparisons can be made. There are a number of reasons why data should be normalized, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, differences in hybridization or detection efficiencies between microarrays, and systematic biases in the measured expression levels. There are many approaches to normalizing expression levels. The simplest, total intensity normalization, assumes that the total amount of mRNA in all samples is constant. It removes between array bias by rescaling the hybridization intensities of all microarrays to have identical median or mean [146, 131, 5]. Quantile normalization makes the even stronger assumption that the distribution of mRNA concentrations is constant for every sample and homogenizes array intensity distributions accordingly [146, 131]. Since mRNA levels in cells are generally not constant, some methods use only a subset of genes (so called house-keeping genes) that are assumed to have constant expression levels. There are a number of alternative approaches such as linear regression, rank invariant methods [146], Chen's ratio statistics [26], centralization [174], or lowess normalization [42, 131]. Some methods not only remove between array bias but also address other systematic sources of errors coming from different dyes, spotting robots, spotting pins etc [42, 172, 131, 146].

Since in contrast to theoretically unlimited mRNA concentrations DNA methylation is a proportion of a fixed DNA amount all these methods are not directly applicable to DNA methylation chips and have to be adapted. In this thesis we will introduce a form of total intensity normalization that takes advantage of the methylation array specific complementary detection probes for methylated and unmethylated DNA.

Variance stabilization

Many traditional statistical methodologies, such as regression or the analysis of variance, are based on the assumptions that the data are normally distributed (or at least symmetrically distributed), with constant variance not depending on the mean of the data. If these assumptions are violated, the statistician may choose either to develop some new statistical techniques

that account for the specific ways in which the data fail to comply with the assumptions, or to transform the data. Where possible, data transformation is generally the easier of these two options [18, 135].

Microarray data fail rather dramatically to conform to the canonical assumptions required for analysis by standard techniques. They show a strong dependency between average hybridization intensity and variance [134]. It is therefore common practice to log transform microarray data prior to analysis [142, 5, 146]. However, the simple log transformation is unable to stabilize the variance for low intensity measurements due to the influence of additive noise components in this regime. Several authors have proposed the use of a generalized log transformation to address this problem [44, 80].

In this thesis we will derive and compare DNA methylation scores based on the simple log transformation as well as an adapted version of the generalized log transformation.

Calibration

Since nRNA and DNA concentrations are proportional to the observed hybridization intensities microarrays can be used to compare concentrations between different biological specimens. However, intercept and slope of the linear relation between concentrations and hybridization intensities are different for every oligonucleotide, vary with experimental conditions and are generally difficult to determine. An absolute quantification of DNA or mRNA concentrations is therefore not trivial and requires some form of calibration. Recent approaches calculate absolute mRNA expression levels based on parameter estimations from control measurements with spiked mRNA [50].

In this thesis we derive a method for the accurate estimation of DNA methylation proportions based on a small number of global calibration measurements of artificially methylated and unmethylated DNA.

1.2.3 Quality control

Despite the popularity of microarray technology, there remain serious problems regarding measurement accuracy and reproducibility. Considerable effort has been put into the understanding and correction of effects such as background noise, measurement signal noise on a slide and different dye efficiencies [22, 152, 42, 5, 142, 146]. When error sources are systematic and known one can try to reduce noise by normalization. However, it has not been clear until now how to handle variations between single slides and systematic alterations between slide batches if the error sources are not known

a priori since the above discussed normalization methods are not applicable in this case.

Between slide variations are so problematic because it is difficult to explicitly model the numerous different process factors that may distort the measurements. Some examples are concentration and amount of spotted probe during array fabrication, the amount of labeled target added to the slide and the general conditions during hybridization [152]. Other common but often neglected problems are handling errors such as accidental exchange of different probes during array fabrication [96]. These effects can arbitrarily affect single slides or whole slide batches. The latter is especially dangerous because it introduces a systematic error and can lead to false biological conclusions if confounded with phenotype annotations or treatment conditions.

There are several ways to reduce between slide variance and systematic errors. Removing obvious outlier chips based on visual inspection is an easy and effective way to increase experimental robustness. A more costly alternative is to do repeated chip experiments for every single biological sample and obtain a robust estimate for the average signal. With or without chip repetitions randomized block design can further increase certainty of biological findings.

In this thesis we will introduce methods to better control the stability of the microarray production process. Process stability control is well known in many areas of industrial production where multivariate statistical process control (MVSPC) is used routinely to detect significant deviations from normal working conditions. The major tool of MVSPC is the T^2 control chart, which is a multivariate generalization of the popular univariate Shewhart control procedure [112, 116]. We will show how this methodology can be adapted for the quality control of high dimensional microarray data.

1.2.4 Data interpretation

After the raw microarray data has been preprocessed and quality approved actual biological interpretation of the data can begin. Data interpretation methods fall into two general groups depending on the use of data annotations. Unsupervised methods ignore sample and gene annotations and just identify common patterns in the measurement data. Supervised methods on the other hand directly use data annotations like gene or sample classifications in combination with the measurement data itself to answer biological questions. All analysis methods can be applied to either gene or sample profiles.

Clustering

Clustering algorithms are unsupervised methods that group gene or sample profiles according to their similarity. Similarity is defined by a distance function, e.g. Euclidean distance, Manhattan distance or Pearson correlation. Hierarchical clustering methods group profiles in a tree diagram (dendrogram) so that similar gene or sample profiles are connected to each other [109, 49]. Partitional clustering methods like k-means clustering [41] or vector quantization identify a number of typical prototype profiles (also referred to as cluster centers or codebook vectors) each representing one cluster. All measurement profiles are then assigned to one of these clusters. Since clustering is not hypothesis driven it is a purely exploratory analysis method. It can be used to generate new hypotheses about gene or tissue subclasses [49, 67].

In this thesis we will use hierarchical clustering to explore the methylation patterns of colon tissue and colorectal neoplasia.

Dimension reduction

Dimension reduction algorithms project high dimensional gene or sample profiles into a lower dimensional subspace. They can be used for visualization purposes (usually a projection into 2 or 3 dimensions) or for preprocessing purposes (e.g. prior to classification algorithms). Most dimension reduction methods are unsupervised and do not take gene or sample information into account. Principle component analysis (PCA) projects data into an orthogonal subspace while retaining a maximum amount of variance [109]. Methods like independent component analysis (ICA) [9] or factor analysis [109] can also project into non-orthogonal subspaces. Correspondence analysis (CA) is able to generate a projection into a lower dimensional subspace while retaining associations between genes and samples [59]. Multi dimensional scaling (MDS) [109] and self organizing maps (SOM) [98, 67] are popular methods that embed high dimensional data into lower dimensions while retaining the original data topology.

There are also some methods like principle component regression, canonical correlation analysis (CCA) [109, 170] and partial least squares (PLS) [123, 122] that combine dimension reduction with supervised data interpretation tasks like classification or regression.

In this thesis we will use different variations of principle component analysis to derive algorithms for microarray quality control and feature selection.

Hypothesis testing and marker selection

The most common task in microarray data analysis is the identification of genes that are differentially expressed in two sample sets with different phenotypic attributes (e.g. normal and cancer tissues). A gene is differentially expressed if the mRNA expression levels of the samples from one set are on average higher than in the other set. Differentially expressed genes are referred to as mRNA markers. In analogy to expression we will call differentially methylated genes or CpG positions methylation markers.

The major problem of marker selection is to distinguish between real differential expression and differences in sample averages that are simply caused by chance. A variety of statistical tests exist to address this problem. They all compute the probability for the null hypothesis that an observed difference in sample sets has occurred randomly. If this probability is small the null hypothesis can be rejected and the respective gene has significant differential expression. The most commonly used test is Student's t-test. It has the disadvantage that it relies on the normal distribution of the two sample classes. A popular non-parametric alternative is the Wilcoxon or Mann-Whitney test [57]. For experimental settings where more than two classes have to be compared, analysis of variance (ANOVA) or the non-parametrical Kruskall-Wallice test are used [30].

A fundamental problem of microarray data analysis is the high data dimensionality - thousands of genes are measured in parallel. For the identification of differentially expressed genes that means several thousand hypotheses have to be tested in parallel resulting in a huge multiple testing problem. The significance level of each individual gene has to be corrected for the fact that there are several thousand chances of generating a false positive result. A variety of methods exist to address this problem [43, 157, 146]. A simple approach is the Bonferroni correction. However, it is very conservative and sacrifices a certain amount of statistical power. More practical and widely used approaches are the re-sampling method of Westfall and Young [167] and the false discovery rate (FDR) approach of Benjamini and Hochberg [11].

In this thesis we will use a multivariate generalization of the t-test to define control limits for microarray quality control. We will also use the t-test and the Wilcoxon test at several points for feature and marker selection.

Classification and Regression

One of the most important applications of microarray analysis is the prediction of phenotypical sample properties from mRNA expression or DNA methylation levels. Typical diagnostic applications are prediction of tissue

malignancy [166, 39], tumor type [67, 1] or treatment response [111, 144, 126]. Classification and regression are supervised learning problems. From a given set of annotated examples (training set, e.g. mRNA expression data with classification into normal and tumor samples) a prediction rule has to be learned. In the case of a classification problem the class of unknown samples or genes can then be predicted by the learned prediction rule. In the case of a regression problem it is a continuous value (e.g. patient survival time) that can be predicted.

A rich literature on classification and regression algorithms exists [14, 109, 158, 27, 41]. Popular classification algorithms for microarray data analysis are Fisher's linear discriminant, k nearest neighbor methods, neural networks, classification trees and support vector machines [10, 168, 23, 65].

In addition to classical statistical regression methods like linear regression, generalized linear regression, principle component regression and partial least squares (PLS) some classification algorithms like neural networks and support vector machines can be generalized to perform regression [14, 158, 41]. A special case of regression that is particularly important for diagnostic applications is a set of methods that can work with censored data. This problem typically arises when observing the survival time of patients after a certain treatment. For many patients monitoring until death is not possible. They will only be monitored for a certain time until they leave the study for various reasons. Cox regression is the most widely used method for this kind of incomplete data [33, 148, 126].

The major problem of all classification and regression algorithms for DNA methylation and mRNA expression data analysis alike is the high dimension of the input space with hundreds or thousands of genes as compared to the usually small number of available samples. Even the support vector machine algorithm that is designed to overcome this problem still suffers from these extreme conditions. Therefore selection of a minimal set of genes or *features* with optimal predictive power is of crucial importance for good performance. A wide variety of feature selection approaches exist in the statistical and machine learning literature [14, 16, 168, 10]. One common approach is to construct or rank features independent of the learning machine that does the actual classification. These algorithms are called filter methods [16]. Another approach is to use the learning machine itself for feature selection. These techniques are called wrapper methods and try to identify the features that are important for the generalization capability of the machine [16].

There is a close relationship between the problems of marker and feature selection. A differentially expressed or methylated marker can always be used as a feature for a classifier and will have some predictive power. However, there might be other markers that have as good or even better classification

performance. On the other hand not every valuable feature has to be a marker. One can construct theoretical examples of features that combined give perfect classification but individually show no difference between classes (e.g. an exclusive *OR* combination of two genes). Generally marker selection aims at finding a set of genes that individually give the best univariate class separation whereas feature selection aims at finding the set of genes that as a multivariate classifier give optimal classification performance.

In this thesis we will show that simple differential methylation filters like the Fisher criterion [14] constitute very powerful feature selection methods and give excellent classification performance on DNA methylation data when combined with a support vector machine.

1.3 Objectives and outlook

The objective of this thesis is to establish statistical methods that enable researchers to use DNA microarrays to measure DNA methylation and draw meaningful biological conclusions.

In chapter 2 we introduce a generative model for the microarray measurement process and derive optimal preprocessing methods for the quantification of DNA methylation from observed hybridization intensities. In chapter 3 we focus on quality control and develop statistical methods to identify and avoid experimental errors in large-scale microarray studies. Chapter 4 shows how DNA methylation microarrays can be used to reliably classify tumor tissue samples by combining feature selection methods and support vector machine classifiers. Chapter 5 applies the developed methods and demonstrates how DNA methylation microarrays can be used to identify markers for the early detection of colorectal cancer. Finally chapter 6 discusses the presented results.

Chapter 2

Measuring DNA methylation with oligonucleotide microarrays

In this chapter algorithms will be derived that take the raw hybridization intensity signals from a methylation specific oligonucleotide microarray and use them to quantify the proportion of methylated DNA strands in a given biological sample for a specific CpG position.

The process of measuring DNA methylation consists of several biotechnical steps in the laboratory. On the one hand the biological specimen has to be prepared to actually make DNA methylation visible and to amplify the signal. On the other hand the DNA microarray as the measurement device has to be prepared to facilitate the CpG dinucleotide specific quantitation of methylation.

After biological sample and DNA microarray are independently prepared they are brought together in the hybridization step. During this step the different CpG dinucleotide specific probes on the microarray react with the amplified sample resulting in methylation specific signals on the microarray. These methylation signals can then be read by an optical scanner providing the input for the analysis algorithms. Fig. 2.1 gives an overview of the whole measurement process.

The following sections will give an overview of the biotechnical process steps, quantify the measurement process with a generative statistical model and finally derive algorithms to estimate methylation.

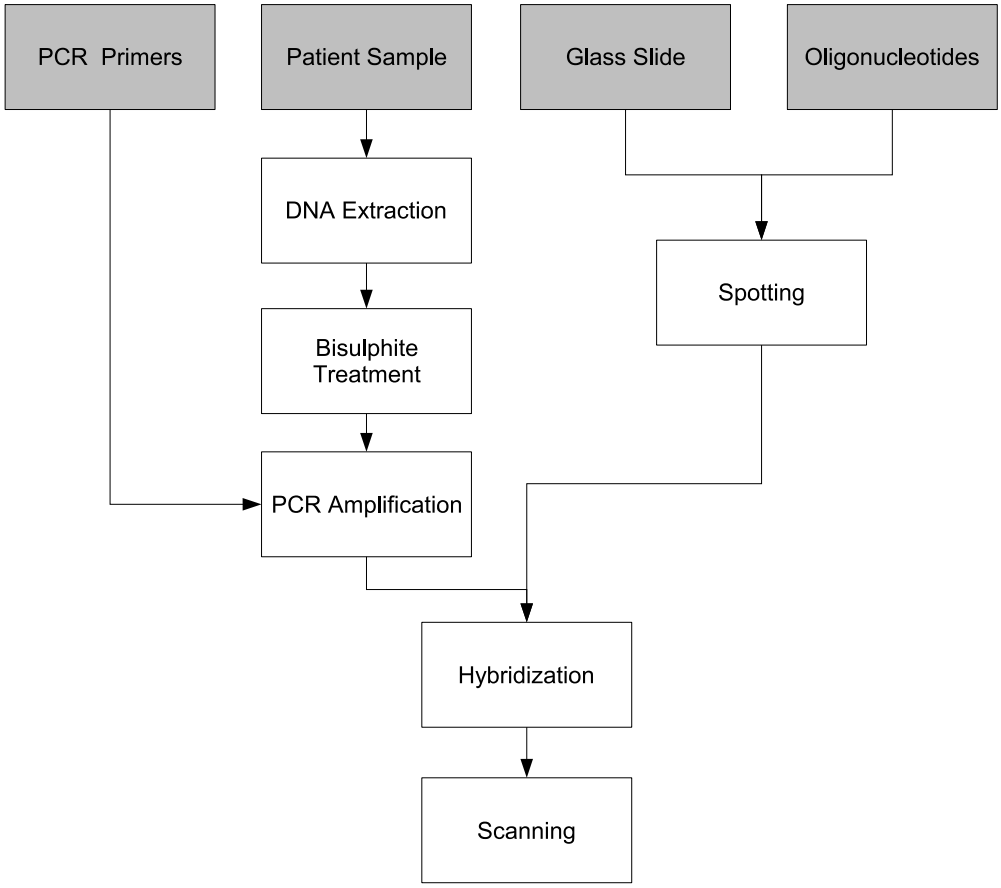

Figure 2.1: Overview of microarray based DNA methylation measurement process. The top row of grey boxes represents the main physical components of the microarray measurement process: the original patient sample to be measured, the raw glass slide, PCR primers and detection oligonucleotides. These components are processed and combined in several steps shown as white boxes. In the last step the finished microarray containing the final signals is scanned and can then be analyzed.

2.1 Measurement process

2.1.1 Sample preparation

Most of the time the biological specimen that have to be analyzed are tissue samples. Typical examples are biopsies of a surgically removed tumor, parts of normal tissue adjacent to a removed tumor or blood.

DNA extraction

The first process step for any kind of tissue sample is the extraction of the DNA. In order to do this the cell boundaries have to be crushed and the DNA has to be separated from all the other cell components. The concrete protocols for doing this vary and depend on the type of tissue sample. However, the two most important factors influencing the measurement quality are always amount and degradation of the extracted DNA.

DNA degradation describes the problem that DNA strands can break during sample preparation impeding the following amplification step. For the typically used fresh tissue samples DNA degradation is not a problem. However, it becomes a critical factor when using paraffin embedded tissue [145].

The amount of extracted DNA N^{DNA} simply describes the number of DNA strands available for analysis. For a given CpG position p a certain number N_p^{DNA+} of strands will be upmethylated and a certain number N_p^{DNA-} will be downmethylated. However, independent of the CpG position p their sum is always the total amount of DNA

$$N^{DNA} = N_p^{DNA+} + N_p^{DNA-}.$$

Note that the two complimentary DNA strands from the same allele have identical methylation. Depending on which DNA strands our detection technology measures we count only 3' or 5' strands or both. In the later case N^{DNA}, N_p^{DNA+} and N_p^{DNA-} would all be even numbers.

What we want to estimate from our DNA sample is m_p, the expected proportion of methylated DNA in a certain tissue of interest at CpG position p. Given our DNA sample the obvious way to estimate m_p is to simply compute the proportion of methylated DNA

$$\hat{m}_p = \frac{N_p^{DNA+}}{N^{DNA}} = 1 - \frac{N_p^{DNA-}}{N^{DNA}} = \frac{N_p^{DNA+}}{N_p^{DNA+} + N_p^{DNA-}}. \tag{2.1}$$

In practice there will always be a difference between the observed methylation rate $\hat{m}_p$ and the expected methylation rate m_p of the pure tissue of interest. This difference is caused by the following two sampling processes.

On the lowest level there is the process of DNA sampling. Given a homogeneous tissue consisting of an infinite number of cells, N^{DNA} strands get selected as specimen. In a pure tissue sample each of these alleles is with probability m_p methylated at CpG position p. The probability of observing a certain methylation rate $\hat{m}_p$ is given by the following binomial distribution

$$P(\hat{m}_p) = \binom{N^{DNA}}{\hat{m}_p N^{DNA}} m_p^{\hat{m}_p N^{DNA}} (1 - m_p)^{(1-\hat{m}_p)N^{DNA}}. \tag{2.2}$$

The estimator $\hat{m}_p$ is unbiased and has a standard deviation of $\sqrt{\frac{m_p(1-m_p)}{N^{DNA}}}$ [21].

In the current microarray process implemented at Epigenomics the required minimum amount of DNA is 10 ng per sample and PCR reaction (see section on PCR below). Assuming a weight of 0.004 ng for one allele of human DNA this corresponds to 2500 different 5' or 3' strands. That means the standard deviation of the methylation estimate is bounded by

$$SDV[\hat{m}_p] = \sqrt{\frac{m_p(1-m_p)}{2500}} \leq \sqrt{\frac{0.5^2}{2500}} = 0.01,$$

which is neglectable for most practical purposes. However, it should be noted that for applications that measure paraffin embedded tissues or body fluids the amount of target material can be considerably lower and DNA sampling becomes a critical issue.

Unfortunately in practice the fewest tissues are homogeneous. A real tumor sample for instance can consist of 90% fat or adjacent normal tissue and only as few as 10% tumor cells. The tumor cells themselfes can also be highly inhomogeneous and represent different pathological subtypes. This inhomogeneity of real tissues results in the problem of tissue sampling. Bias and variance of $\hat{m}_p$ introduced by tissue sampling can be very severe and depend on the concrete tissue type, method and quality of surgery and quality of pathological analysis and dissection. It is hard to estimate and remains as a major noise component in the data. The only way to avoid bias and variance caused by tissue sampling is by doing micro dissection. With this method single tissue cells of interest are selected under the microscope. Although it can improve data quality dramatically, it is very labor intensive, needs special equipment and produces only small amounts of DNA.

Bisulphite treatment

Because methylation is a relatively minor modification of DNA it is practically invisible for all classical methods of DNA analysis, particularly PCR

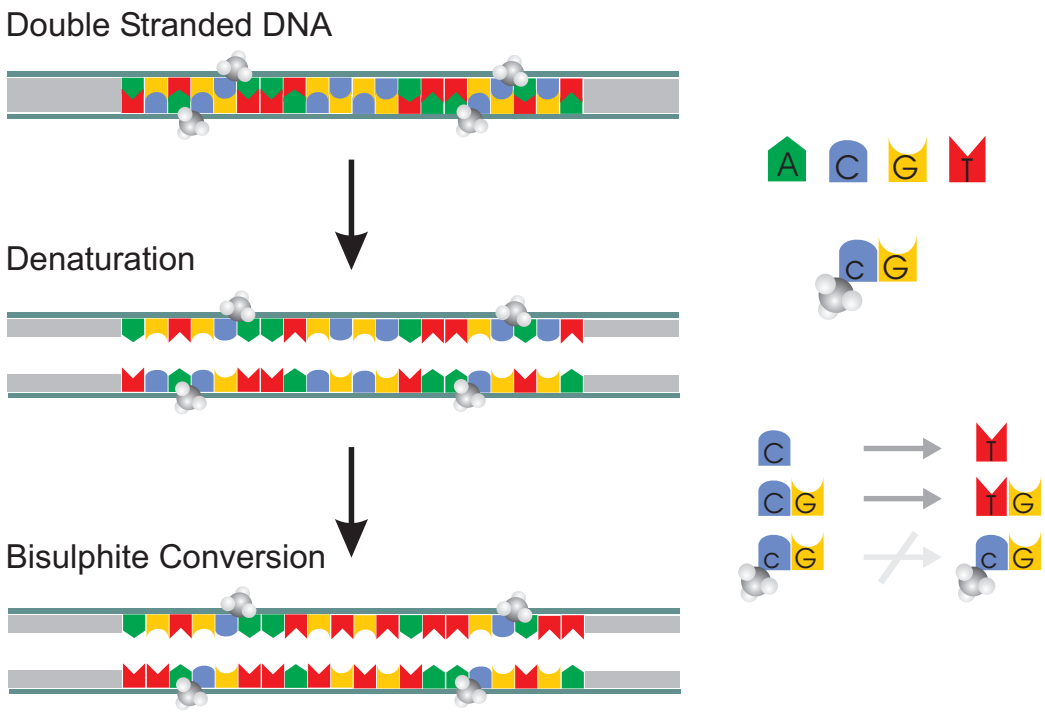

Figure 2.2: Cytosine conversion by bisulphite treatment. The double-stranded genomic DNA from the original patient sample is first separated into individual strands (denaturation). Then all Cytosines (blue symbols) that are not protected by a methyl group (grey molecules) are converted to Uracil which looks like Thymidine (red symbols) to the TAQ-Polymerase. Note that Cytosines can only be methylated in CpG context (i.e. the C is followed by a G). After bisulphite treatment all unmethylated Cytosines in the two DNA strands are converted to Thymidines. The two DNA strands are therefore not complimentary anymore.

and hybridization. After the DNA is extracted its methylation signature has therefore to be converted into sequence information that is visible to the following steps of the measurement process. This is achieved by treating the extracted genomic DNA with bisulphite which converts all unmethylated Cytosines into Urazil. The Urazil molecule in turn is read by the following PCR amplification as Thymidine. This results in an effective translation of all unmethylated Cytosines into Thymidine. Fig. 2.2 shows an example of the complete conversion process. Note that the two DNA strands are not complementary anymore after bisulphite treatment.

Control experiments performed at Epigenomics indicate that the bisulphite conversion is not always perfect. Depending on the sequence context and resulting secondary structures of the DNA the bisulphite reaction can be inhibited. This can result in bisulphite DNA strands with unmethylated

Cytosines which would create a bias toward upmethylation in the following measurement procedure. However, in almost all cases bisulphite conversion rates are above 95%. Therefore we will ignore imperfect conversion here.

PCR

After the methylation signal of the target DNA is made visible as sequence alteration by bisulphite treatment it has to be amplified with the polymerase chain reaction (PCR). PCR generates an exponential amount of copies of small stretches of DNA specified by a forward and a reverse primer. In the same reaction it also attaches a flouorecent label to the generated copies. See Fig. 2.3 for a detailed explanation of PCR. Note that methylation information is not copied by PCR since the polymerase does not discriminate between methylated and unmethylated Cytosines. Only the bisulphite induced sequence changes remain after amplification.

The primers for PCR are designed such that they do not contain any CpG sites. In this way it is assured that there is no bias in amplifying methylated or unmethylated DNA.

The small excerpts of DNA copied by PCR are called fragments. The specific set of fragments generated by one PCR on one DNA sample is called amplificate. It represents the amplification of one specific genomic region defined by the respective primer pair. The term amplificate is therefore often used in a more abstract meaning as a synonym for the genomic region itself.

For the standard microarray process at Epigenomics 64 of these amplificates with a length between 100 and 800 base pairs are used. The design of these amplificates is an essential step in the experimental design and has to focus on genomic regions (usually promotor areas of genes or adjacent CpG islands) with known importance to the biological question of interest. Note that PCR primers can be designed for both bisulphite DNA strands resulting in 2 different possible amplificates. See Fig. 2.6 for an example. The a priori information about interesting genomic regions can either come from specific whole genome discovery experiments [31, 68] or from the literature.

Performing all 64 PCR reactions per sample individually would be far too labor intensive. Instead one can perform several PCR amplifications in one reaction by pooling the primer pairs. The choice of the primer pairs to pool for this multiplex PCR (mPCR) is non-trivial and has to be optimized in order to minimize cross-reactivity between the different primers and amplification products [136].

In the microarray process currently implemented at Epigenomics primers are pooled to 8-plexes. To generate the full set of 64 amplificates 8 of these

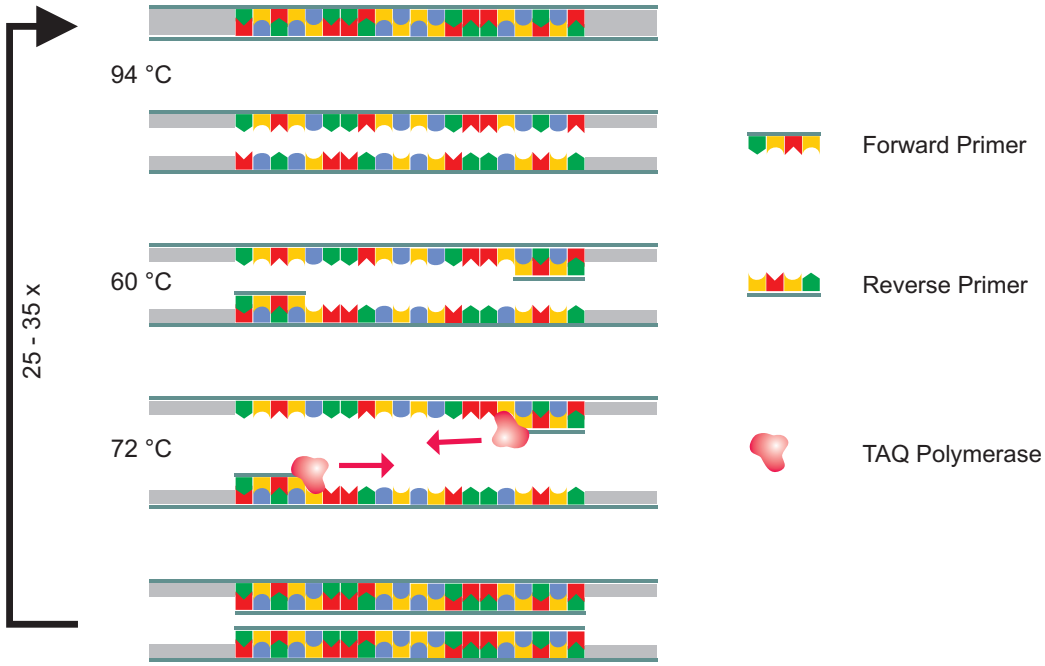

Figure 2.3: Polymerase chain reaction (PCR). The PCR reaction consists of a series of 25 to 35 cycles. Each cycle consists of three steps. (1) The double-stranded genomic or partially double-stranded bisulphite converted DNA has to be heated to 94-96C in order to separate the strands. This step is called denaturing; it breaks apart the hydrogen bonds that connect the two DNA strands. Prior to the first cycle, the DNA is often denatured for an extended time to ensure that both the template DNA and the primers have completely separated and are now single-strand only. Time: 1-2 minutes. (2) After separating the DNA strands, the temperature is lowered so the primers can attach themselves to the single DNA strands. This step is called annealing. The temperature of this stage depends on the primers and is usually 5C below their melting temperature (45-60C). A wrong temperature during the annealing step can result in primers not binding to the template DNA at all, or binding at random. Time: 1-2 minutes. (3) Finally, the TAQ-Polymerase has to fill in the missing strands. It starts at the annealed primer and works its way along the DNA strand. This step is called elongation. The elongation temperature depends on the TAQ-Polymerase. The time for this step depends both on the TAQ-Polymerase itself and on the length of the DNA fragment to be amplified. As a rule-of-thumb, 1 minute per 1000 base pairs. There is one primer for each of the two complimentary strands. Every cycle each primer initiates the generation of a new complimentary strand starting from its own binding site towards the 3' end. Both primers together result in an exponential amplification of the DNA fragment between them.

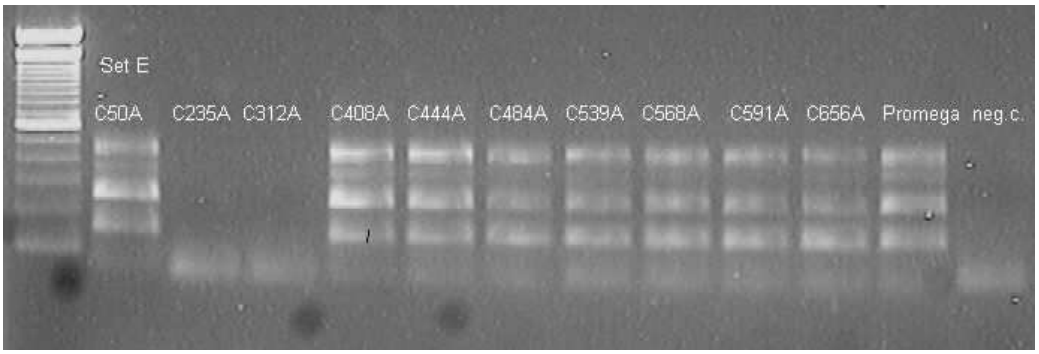

Figure 2.4: Gel electrophoresis for mPCR control. Lane 1: 100 bp marker; Lane 2-11: multiplex PCR performance of one primer set on 10 test samples. Lane 12: positive control (Promega DNA), Lane 13: H2O control. Each of the 10 test samples is checked for presence of all amplification products. This is done by comparing the bands between the test samples and the positive control (Promega DNA). In this example samples C235A and C312A show insufficient amplification.

8-plex mPCRs are performed for each sample. Because the mPCR reactions are so complex and depend heavily on DNA quality each individual amplification result is controlled for presence of desired amplificates and absence of undesired byproducts by gel electrophoresis. See Fig. 2.4 for an example. However, gel electrophoresis is only a very crude control. Because of its limited resolution with regard to different amplification products there remains still a high likelihood of undetected additional byproducts or missing amplificates. Although additional byproducts may later cause undesired cross-hybridization signals (see Section 2.2.3) they are not as critical as missing amplificates. An only weakly amplified fragment will cause a more noisy methylation estimate later on. A completely missing fragment will cause an undefined methylation measurement in the later process steps and result in an outlier.

A practical approach to at least detect amplificates with a high likelihood of failure is to perform capillar electrophoresis measurements on a set of randomly selected samples. See Fig. 2.5 for an example. Amplificates with a high likelihood of failure can then be excluded from the later analysis steps. However, with the current technology it is not possible to avoid a considerable amount of noise and outliers caused by different amplification efficiencies and completely failing amplifications.

In the following we will assume a perfectly unbiased and efficient PCR resulting in identical proportions of bisulphite converted methylated DNA

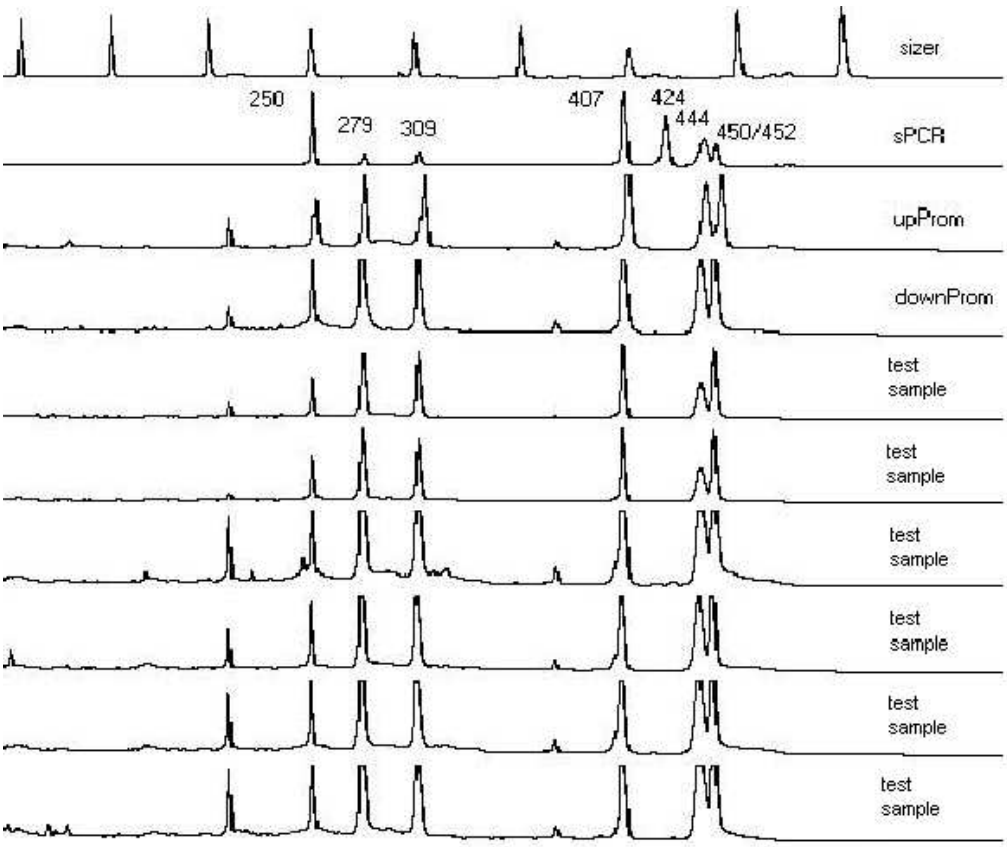

Figure 2.5: Capillar electrophoresis for mPCR control. Each row represents the mPCR product of a individual test sample or control (sPCR - mix of single PCR products, upProm - artificially upmethylated Promega DNA, downProm - artificially downmethylated Promega DNA). The sPCR row shows the amplificate peaks that should be present. In the regular mPCR samples additional peaks corresponding to unwanted byproducts can be observed. However, the main quality control criteria is the presence of all 8 single PCR peaks in the multiplex PCR products of the test samples.

before and after amplification. With 100% PCR efficiency the number of total fragments N^{PCR} after n_C PCR cycles is given as

$$N^{PCR} = 2^{n_C} N^{DNA}.$$

Since the methylation pattern was translated into simple sequence information and the PCR primers themselfes do not cover any CpG sites we can assume unbiased amplification of all methylation patterns. Therefore the numbers of originally methylated and unmethylated fragments for a given CpG position p are $N_p^{PCR+} = 2^{n_C} N_p^{DNA+}$ and $N_p^{PCR-} = 2^{n_C} N_p^{DNA-}$ respectively. According to Eq. 2.1 the proportion of methylated DNA can be estimated as

$$\hat{m}_p = \frac{N_p^{DNA+}}{N_p^{DNA+} + N_p^{DNA-}} \overset{\text{PCR}}{=} \frac{2^{n_C}}{2^{n_C}} \frac{N_p^{DNA+}}{N_p^{DNA+} + N_p^{DNA-}} = \frac{N_p^{PCR+}}{N_p^{PCR+} + N_p^{PCR-}}. \tag{2.3}$$

Ignoring outliers caused by not working amplifications $\hat{m}_p$ is still distributed according to Eq. 2.2.

2.1.2 Microarray preparation

Bisulphite treatment converts the methylation signal to a change in DNA sequence. Unmethylated CpG positions are converted to TpG, whereas methylated CpG positions remain unchanged. PCR amplifies this methylation dependent sequence alteration for DNA fragments of interest and attaches a fluorescent label. DNA methylation can now be measured by designing oligo nucleotide probes complementary to the methylation induced sequence alterations and by placing these specific oligo probes at different locations on a microarray.

Oligo nucleotide design

Oligo nucleotides (also called oligomeres or short oligos) are synthesized short stretches of "artificial DNA" that can bind (hybridize) to the fragments produced by PCR. The length of the detection oligo nucleotides used at Epigenomics is usually around 20 base pairs. In order to measure methylation oligo nucleotide probes have to be designed either to detect unconverted (originally methylated) CpG positions or to detect CpG positions converted to TpG (originally unmethylated CpG positions). Here we will call the class of oligo nucleotides designed to detect methylated CpG positions CG-oligos. The class of oligo nucleotides designed to detect unmethylated CpG positions will be called TG-oligos.

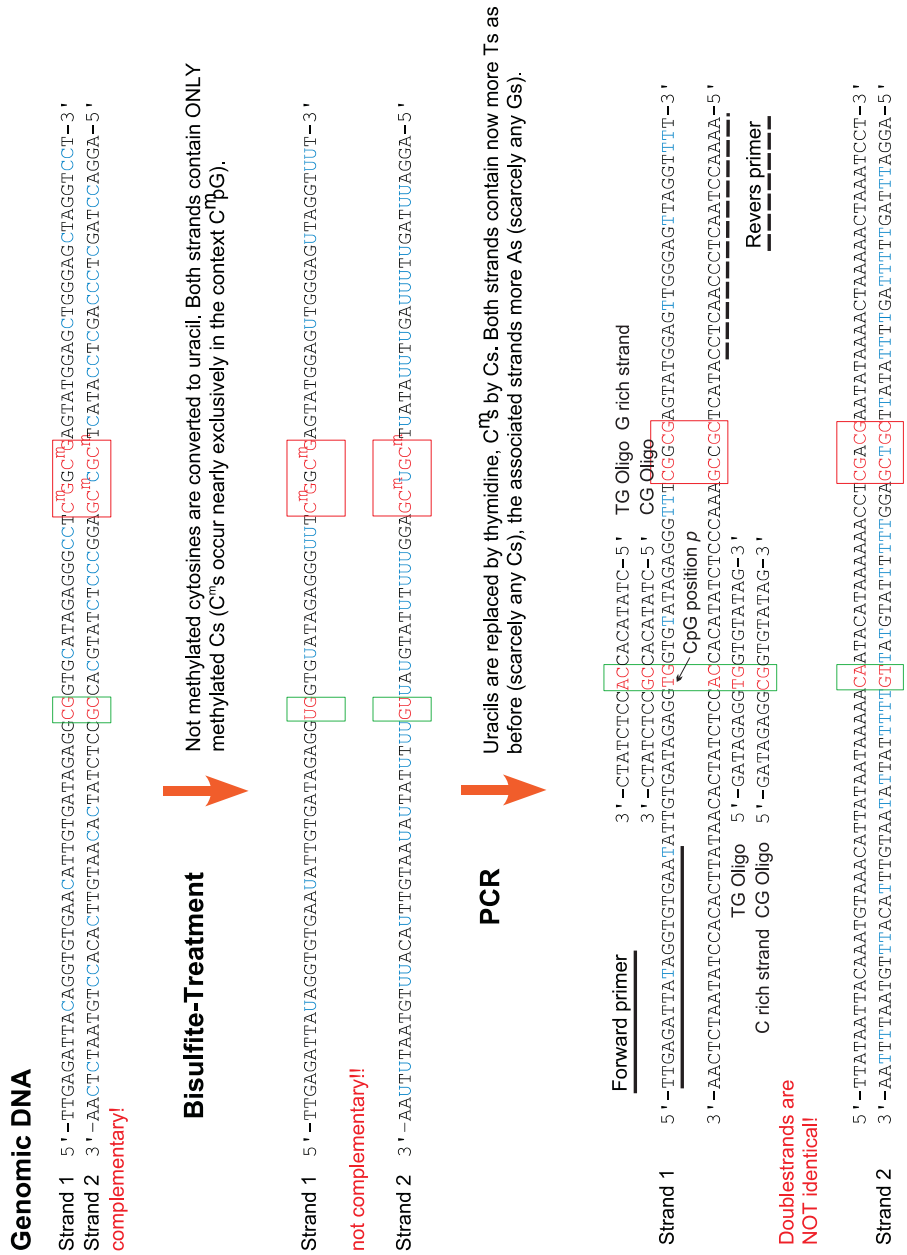

Figure 2.6: Oligo and primer design. The figure shows how a double-stranded genomic DNA is bisulphite converted and amplified. This results in 2 different amplificates with 4 different kinds of fragments that can be detected via complementary detection oligomeres. Typical primer, CG and TG oligo examples are shown for the first bisulphite strand and its resulting amplificates. Note that detection oligos are usually designed for the C rich strand because of higher achievable melting temperature differences. However, a design for the G rich strand is possible as well and example detection oligos are shown in the figure. The non-methylation specific detection oligo is in this case not really a TG oligo but has an AC dinucleotide in its center.

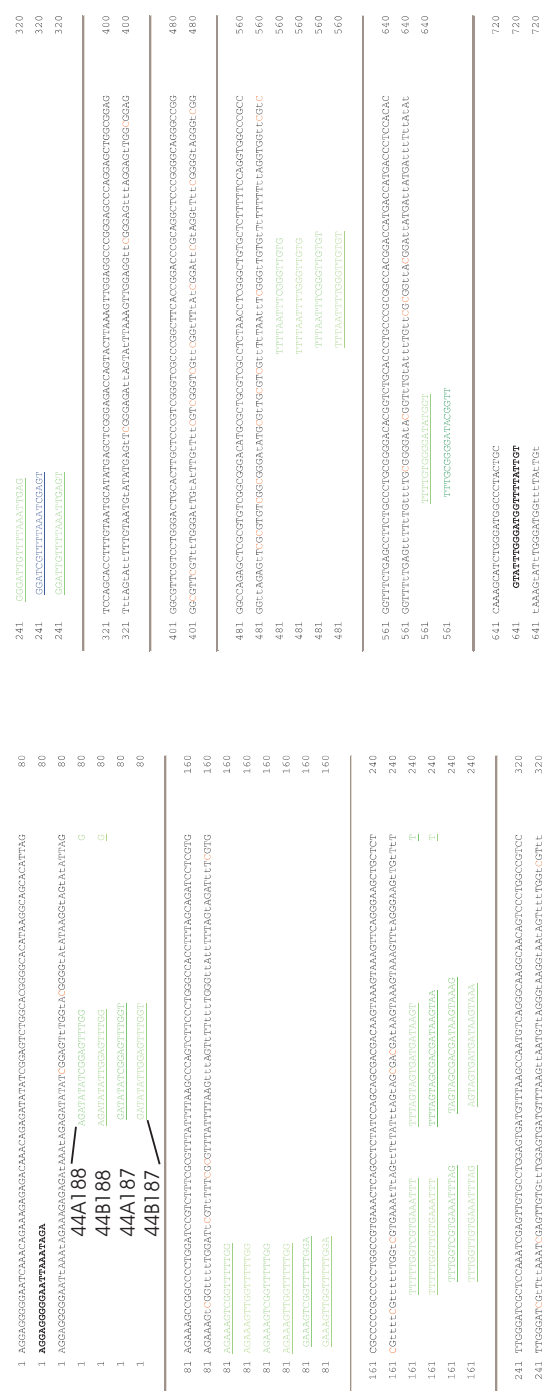

Figure 2.7: Example of designed primers and detection oligo families for the *Homo sapiens estrogen receptor 1* (ESR1) gene. Figure shows genomic sequence, bisulphite converted sequence (t = converted Cytosines not in CpG context, red C = Cytosine in CpG context), designed primer sequences (printed bold face) and designed detection oligos (different colors representing different melting temperatures). The oligos covering the CpG at position 44 are marked and will be used in later examples.

CG- and TG-oligos are always designed to cover one or more CpG positions and several adjacent base pairs invariant to methylation. Fig. 2.6 shows a detailed example for an oligo design. The non-CpG base pairs of an oligo make it specific for a certain CpG position in the DNA. The CpG base pairs determine the oligo class (CG- or TG-oligo) and make the oligo sensitive to changes in methylation. The right proportion between sequence specific and methylation sensitive base pairs is crucial for a good working oligo. Typical oligos cover between 1 and 3 CpG dinucleotides and have a total length of 16 to 22 base pairs.

We call a set of neighboring CpG dinucleotides a CpG cluster. In the following we will assume that neighboring CpG sites are co-methylated (either all CpG sites in the CpG cluster are methylated or all CpG sites are not methylated). Mathematically we will treat a CpG cluster equivalent to a single CpG position. The cluster will simply be represented by methylation state and position of its first CpG dinucleotide.

Usually it is possible to design different oligos of at least theoretically identical quality by shifting, shrinking or expanding the sequence. We call a set of oligos querying the same set of CpG positions an oligo family. Fig. 2.7 shows an example of oligo families for the *Homo sapiens estrogen receptor 1* (ESR1) gene. What is important is that every detection oligo is designed to query the methylation information of exactly one CpG position or CpG cluster. That means we can define a unique mapping from oligos to CpG positions:

$$p(q) := \text{ CpG position } p \in \mathcal{P} \text{ that oligo } q \text{ is binding to,} \tag{2.4}$$

where p is the genomic location of a CpG dinucleotide or CpG cluster and q is an index specifying an oligo (i.e. a specific oligo sequence). Since there can be several oligos querying the same CpG site this mapping is generally not one-to-one and therefore not invertible.

Optimal oligo design is a complex task and has to take into account sequence context around the CpG position of interest, sequences of all other amplificates measured with the same microarray and chemical conditions of the hybridization reaction. For this thesis optimality criteria for oligo design will be ignored. However, we will have to model the statistical behavior of real oligos. This includes methylation sensitivity and at least partially sequence specificity.

We refer to oligos that are designed to measure DNA methylation at a specific CpG position as detection oligos. They constitute the majority of all oligos on a microarray. In order to control the conditions of the hybridization reaction it is usefull to include some control oligos beside the detection oligos.

These control oligos have sequences that do not match any part of any PCR fragment.

The simplest control oligo is a negative control. Since its sequence does not match with any of the amplificates it should show no signal. In practice it can be used to measure the degree of sequence unspecific hybridization as well as unspecific background signal and background noise.

The second type of control oligo used in the Epigenomics array process is the positive control. This oligo type is designed to hybridize with another control oligo or control PCR fragment of known concentration that is added into the original PCR mix. The addition of an artificial control oligo or amplificate with fixed concentration is called spiking. Positive control oligos and their hybridization reaction with the spiked control oligos or fragments can be used to verify that hybridization conditions are specific enough.

Spotting

After the designed oligo nucleotide probes are synthesized they are spotted onto glass slides. A standard glass slide is 75 mm high, 25 mm wide and 1 mm thick. It can hold 2048 spots for detection oligos, each with a diameter of about $580\mu m$. The standard Epigenomics 64 gene chip contains spots for:

- 64 amplificates corresponding to up to 64 different genes
- 4 CG oligos per amplificate
- 4 TG oligos per amplificate
- 4 repetitions for each individual oligo,

resulting in a total of $64 * (4 + 4) * 4 = 2048$ spots per chip.

Control oligos are spotted in between 8x8 blocks of detection oligos and do not reduce the number of 2048 detection oligo spots. There are typically several different negative control oligos spotted in 4-fold redundancy. The number of different positive control oligos is usually between 1 and 3 but spotted in 32-fold redundancy.

The concrete process of transferring and fixating the synthesized oligos onto the glass slides is rather complex. It involves several pipeting and spotting steps performed by specialized robots as well as several additional compounds for slide activation and oligo immobilization. In principle the concrete spotting process has no influence on the final interpretation of the methylation data. However, if production conditions are changed (e.g. oligos get resynthesized with a different concentration or an immobilization buffer is exchanged) or if errors occur (e.g. two spotting plates with different oligo

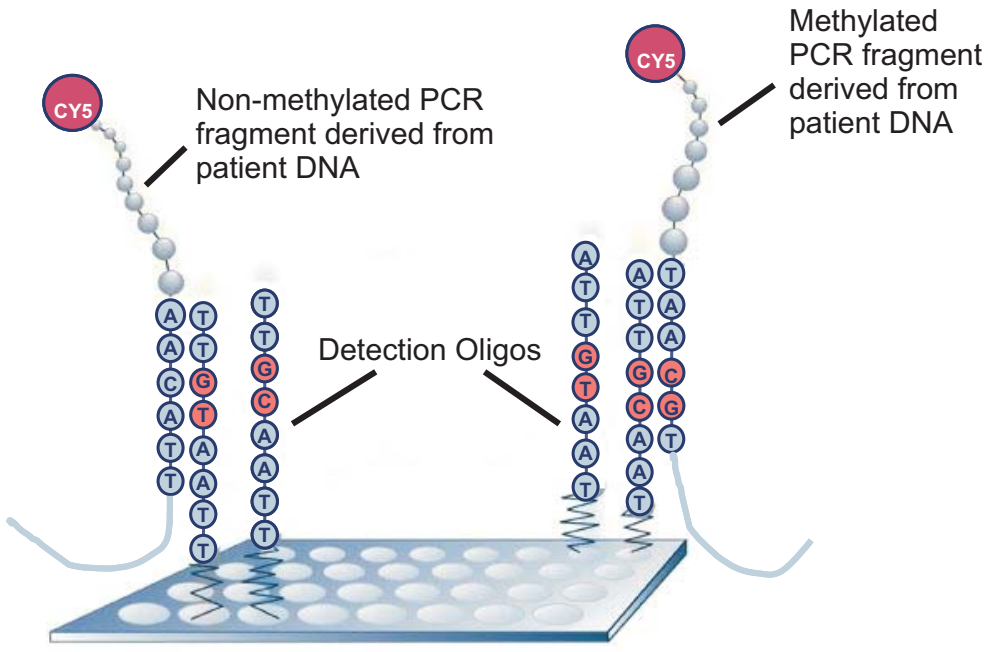

Figure 2.8: Methylation specific hybridization on microarray. The figure shows how methylated and unmethylated fragments bind to their respective CG and TG detection oligos. On the left hand side a bisulphite converted unmethylated amplificate (CG converted to TG, corresponding to CA on the complimentary strand) binds to its matching TG oligo. On the right hand side a unconverted methylated amplificate (CG stays CG, corresponding to CG on the complimentary strand) binds to its matching CG oligo.

sets get confused) then this has an immediate and severe effect on the final data. Usually very typical outlier patterns are generated. Chapter 3 will give several examples and show how to systematically control for problems in the microarray production process itself.

2.1.3 Hybridization and image analysis

In the final step of the microarray process the bisulphite treated and PCR amplified sample DNA is dissolved in a hybridization buffer and washed over the spotted glass slide.

During this process the dissolved amplificates will bind to the spotted oligos. This binding process is called hybridization. Fig. 2.8 visualizes the binding of amplificates to spotted oligos. The temperature of the hybridization reaction is kept constant at a level that is below the melting temperature of matching oligo-amplificate pairs but above the melting temperature

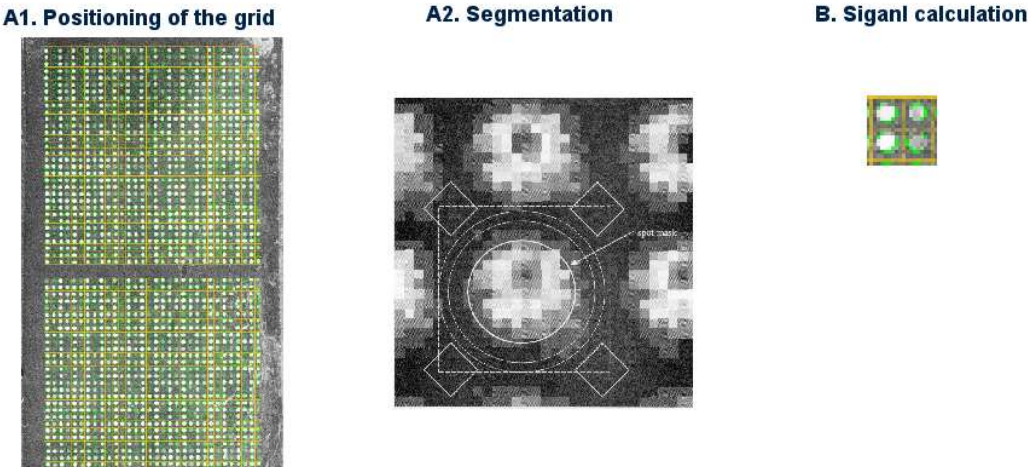

Figure 2.9: Grid finding and spot intensity estimation. In order to identify the specific oligonucleotide spots on a microarray and estimate their respective intensities the following steps are performed by the chip evaluation software: A1) Rotation and translation of the known grid are estimated. Each cell of the alligned grid should contain exactly one spot. A2) Within each grid cell the pixels belonging to the respective spot are identified. B) Median spot intensity, spot diameter and median background intensity are computed for every spot.

of miss-matching oligo-amplificate pairs. Typical hybridization temperatures are around 42C.

Over time the hybridization reaction converges to an equilibrium where the rate at which new oligo-amplificate duplexes are formed is equal to the rate at which the already formed duplexes dissolve. This equilibrium is usually reached after several hours. A typical hybridization time in the Epigenomics chip process is 16 hours. After the hybridization time is over the glass slide is taken out of the hybridization buffer and washed so that only amplificates that are bound to an oligo are left on the surface.

After hybridization and washing are finished the microarray is scanned by a laser scanner. The dye labels attached to the amplificates which in turn are bound to their matching oligos on the microarray are excited by the laser and emit light of a certain wave length (e.g. Cy5 labels with peak emission at 670nm). Typically all amplificates are labeled with the same dye. By scanning the whole microarray an image is created whose intensities are proportional to the number of bound amplificates at the respective array position.

The first data analysis step is to identify the oligo spots on the microarray image and estimate their intensities. Spots are identified by alligning the

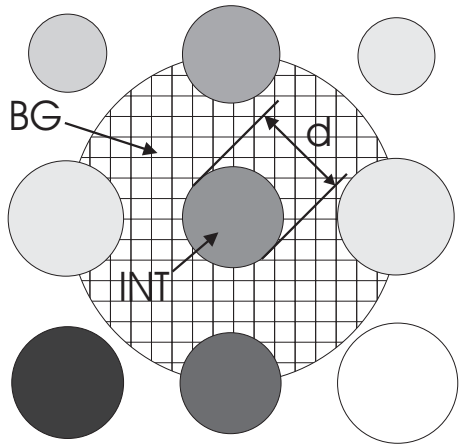

Figure 2.10: Measures for hybridization intensity. A single spot is characterized by its median intensity INT, its diameter d (measured in nm) and the local background intensity BG. From these characteristics the following three alternative hybridization intensity measures can be computed. Median Intensity: $MI = INT$. Effective Median Intensity: $EMI = INT - BG$. Effective Median Volume: $EMV = 0.4\pi d^2(INT - BG)$.

known spotting grid to the observed image. Then a circular region that covers the bright pixels within the respective grid cell is identified as the spot itself. The rest of the grid cell is assumed to be background. Fig. 2.9 shows a typical scan of an Epigenomics microarray with the alligned spotting grid.

For each identified spot we can measure the spot intensity INT by computing the median intensity of all pixels within the circular spot region. We can also measure the background intensity BG and the spot diameter d. With these three parameters we can define the following alternative measures for hybridization intensity and the amount of bound amplificates:

$$\begin{aligned} MI &= INT \\ EMI &= INT - BG \\ EMV &= 0.4\pi d^2(INT - BG), \end{aligned} \tag{2.5}$$

where MI stands for median intensity, EMI for effective median intensity and EMV for effective median volume. See Fig. 2.10 for a schematic visualization. In the following sections these three measures will be compared with regard to their usefullness for estimating the proportion of methylated DNA in a sample.

2.2 A statistical model of hybridization

In order to estimate methylation from hybridization intensities we have to understand the systematic and stochastic sources of error of the microarray process. In the following sections we will derive simple generative models for hybridization signals. First we will quantify stochastic variations for hybridization intensities on a single microarray (within chip noise) and on different microarrays (between chip noise). Then we will derive a simple model for systematic deviations of hybridization intensities.

2.2.1 Within chip noise

A generative model

What we observe in a microarray experiment are oligo intensities O_q, with index q specifying the oligo and the value of O_q itself given by one of the three measures defined in Eq. 2.5. These hybridization intensities can be used to measure DNA methylation because of the following dependencies:

- DNA methylation is proportional to the amount of originally methylated amplificates in a sample: $m_p \propto N_p^{PCR+}$ (see Eq. 2.3).
- For a given volume of the hybridization reaction the concentration of a originally methylated amplificate is proportional to its amount.
- The absolute number and concentration of stable oligo-amplificate duplexes at equilibrium is proportional to the concentration of the respective target amplificate [169].
- Assuming that the microarray image scanner is linear the theoretically expected hybridization intensity I_q is proportional to the number of bound labeled amplificates at oligo q.
- The observed hybridization intensities are on average proportional to the theoretically expected hybridization intensity: $E[O_q] \propto I_q$ (see noise model below).

Since every oligo can be mapped to a specific CpG site (Eq. 2.4) and we assume that only the intended target CpG site on the target amplificate binds to the oligo we get the following correlations between oligo intensities and CpG methylation:

$$\begin{aligned} E[O_q] \propto I_q \propto N_{p(q)}^{PCR+} \propto m_{p(q)} \quad & \text{for CG oligos} \\ E[O_q] \propto I_q \propto N_{p(q)}^{PCR-} \propto 1 - m_{p(q)} \quad & \text{for TG oligos} \end{aligned} \tag{2.6}$$

As for every measurement device the dependence between the expected hybridization intensity I_q and an actual observed hybridization intensity O_q is distorted by systematic and stochastic errors. In order to optimally estimate I_q and ultimately the methylation frequency m_p from a set of repeated hybridization observations of the same oligo q $\mathcal{O}_q = \{O_{q,i}, i = 1 \dots n_r\}$ we have to understand and model these sources of error.

When we look at a set of repeated measurements from several spots of the same oligonucleotide on the same chip we observe that the standard deviation linearly increases with the average hybridization intensity of the oligo. This dependence between variance and intensity can be observed on most DNA microarray platforms and seems to be independent of the particular oligo sequence [134, 106]. For a standard Epigenomics microarray this relationship is shown in Fig. 2.11.

The ratio between standard deviation and average intensity for a given oligonucleotide is refered to as the coefficient of variance

$$CV_q = \frac{\sqrt{Var[O_q]}}{E[O_q]}. \tag{2.7}$$

The easiest generative model with a linear dependence between standard deviation and intensity is a log normal distribution:

$$O_q = I_q e^{\eta_q}, \tag{2.8}$$

where $\eta_q = N(0, \sigma_{\eta_q})$ is a normal distribution with mean 0 and standard deviation σ_{η_q}. Mean and variance of this log normal intensity distribution are given as [56]:

$$\begin{aligned} E[O_q] &= E[I_q e^{\eta_q}] = I_q \sqrt{e^{\sigma^2_{\eta_q}}} \\ Var[O_q] &= Var[I_q e^{\eta_q}] = I_q^2 e^{\sigma^2_{\eta_q}} (e^{\sigma^2_{\eta_q}} - 1). \end{aligned} \tag{2.9}$$

It follows that the coefficient of variance for this model is constant:

$$CV_q^{LN} = \frac{\sqrt{Var[O_q]}}{E[O_q]} = \frac{\sqrt{I_q^2 e^{\sigma^2_{\eta_q}} (e^{\sigma^2_{\eta_q}} - 1)}}{I_q \sqrt{e^{\sigma^2_{\eta_q}}}} = \sqrt{e^{\sigma^2_{\eta_q}} - 1}. \tag{2.10}$$

We can estimate I_q and σ_{η_q} from a set $\mathcal{O}_q$ of repeated measurements of an oligo q as

$$\begin{aligned} \hat{I}_q &= \exp\left(\frac{1}{|\mathcal{O}_q|} \sum_{O \in \mathcal{O}_q} \log O\right) \\ \hat{\sigma}^2_{\eta_q} &= \frac{1}{|\mathcal{O}_q| - 1} \sum_{O \in \mathcal{O}_q} \left(\log O - \log \hat{I}_q\right)^2. \end{aligned} \tag{2.11}$$

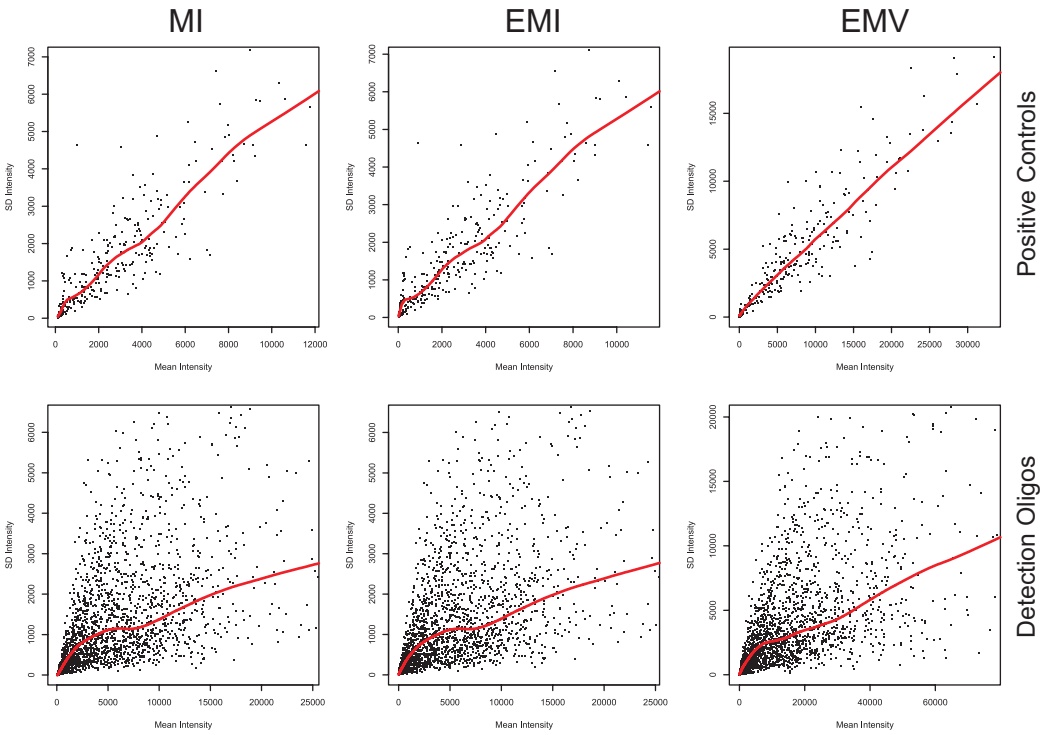

Figure 2.11: Mean-SD dependence of raw hybridization intensities O_q. The upper row shows 7 positive control oligos on 48 microarrays. The lower row shows 50 detection oligos on the same 48 microarrays. The three columns correspond to the three different intensity measures MI, EMI and EMV. Each point shows mean hybridization intensity vs. standard deviation for one oligo on a single microarray computed from the respective spot repetitions on the respective microarray. The red line shows the average standard deviation for a given intensity computed by a lowess fit. Each microarray contained positive controls in 32-fold redundancy and detection oligos in 4-fold redundancy. Note that SD estimates for the detection oligos are very imprecise due to their estimation from only 4 data points.

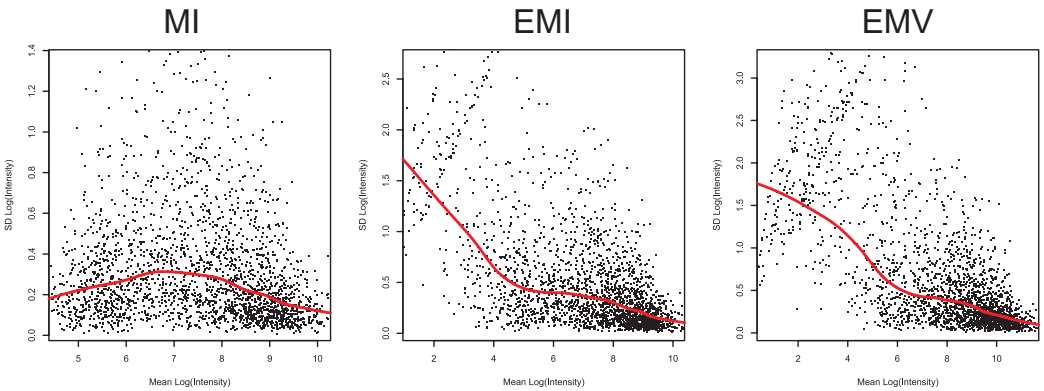

Figure 2.12: Mean-SD dependence of log transformed hybridization intensities. The three plots correspond to the three different intensity measures MI, EMI and EMV. Each point shows mean vs. standard deviation of the log transformed hybridization intensities for one detection oligo on a single microarray computed from the respective spot repetitions on the respective microarray. Plots were generated from a total of 48 microarrays and 50 detection oligos. Hybridization intensities below 1 were cut off and set to 1 prior to the log transformation. The red line shows the average standard deviation for a given intensity computed by a lowess fit. Each microarray contained detection oligos in 4-fold redundancy. Note that SD estimates are very imprecise due to their estimation from only 4 data points.

Usually we have several different oligos with repeated measurements. Typically these oligos will have different expected intensities I_q. But as shown in Fig. 2.11 their CVs are approximately constant for a concrete experimetal series of identical arrays. This is only an approximation because in practice the CV may vary with the location of the respective oligo on the chip due to locally different hybridization conditions.

Assuming that oligo sequence and location dependent noise have no essential impact we can estimate the average standard deviation σ_η by pooling the single variance estimates of all oligos q from oligo set $\mathcal{D} \subseteq \mathcal{Q}$ as

$$\hat{\sigma}_\eta^2 = \frac{1}{|\mathcal{D}|} \sum_{q \in \mathcal{D}} \sigma_{\eta_q}^2, \tag{2.12}$$

given that the number of repeated measurements of all oligos in $\mathcal{D}$ is identical.

According to Eq. 2.8 the logarithm of the observed hybridization intensities O_q should be normally distributed and have constant variance σ_η^2. The dependence between standard deviation and intensity on logarithmic data is shown in Fig. 2.12. Obviously the variance of the logarithmic data is only approximately stable for the MI measurement values. For the EMI and EMV values Eq. 2.8 holds only for higher intensities. For very small intensities

the variance is not decreasing with the intensity anymore. The explanation for this behaviour is a small oligo independent base hybridization with a gaussian noise characteristic. Because the MI values will never fall below the relatively high background intensity the oligo base hybridization is neglectable when taking the logarithm. However, since the background corrected EMI and EMV values can approach zero or even be smaller than zero the base hybridization noise becomes important. We can easily incorporate this additive base hybridization into Eq. 2.8 and get the following generative model [134]:

$$O_q = I^{BG} + I_q e^{\eta} + \epsilon, \tag{2.13}$$

where I^{BG} is the average base hybridization intensity assumed to be constant for all oligos and $\epsilon = N(0, \sigma_\epsilon)$ is a normal distribution with mean 0 and standard deviation σ_ϵ.

Using Eq. 2.9 we can easily derive mean and variance of this intensity distribution:

$$\begin{aligned} E[O_q] &= E[I^{BG} + I_q e^{\eta} + \epsilon] = I^{BG} + I_q \sqrt{e^{\sigma_\eta^2}} \\ Var[O_q] &= Var[I^{BG} + I_q e^{\eta} + \epsilon] = {I_q}^2 S_\eta^2 + \sigma_\epsilon^2, \end{aligned} \tag{2.14}$$

where $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$.

To determine the parameters of the model we can use oligos with known hybridization properties. $\sigma_{\eta_q}^2$ can be estimated from positive control oligos or from very intense detection oligos by using Eq. 2.12. The additive noise term can be neglected for high I_q. Negative control oligos can be used to estimate I^{BG} and σ_ϵ. From a set $\mathcal{N} \subseteq \mathcal{O}$ of negative control oligos spotted in n_r-fold replication the parameters can be estimated as

$$\begin{aligned} \hat{I}^{BG} &= \frac{1}{n_r |\mathcal{N}|} \sum_{q \in \mathcal{N}} \sum_{i=1}^{N} O_{q,i} \\ \hat{\sigma}_\epsilon^2 &= \frac{1}{n_r |\mathcal{N}| - 1} \sum_{q \in \mathcal{N}} \sum_{i=1}^{N} \left(O_{q,i} - \hat{I}^{BG} \right)^2. \end{aligned} \tag{2.15}$$

Because we simply pooled measurements from different negative control oligonucleotides we assumed that unspecific background hybridization intensity and variance are independent of the actual negative control oligo sequence. In practice the set of negative controls will contain some oligonucleotides that show a significant amount of unspecific hybridization resulting in overestimation of $\hat{I^{BG}}$ and $\hat{\sigma_\epsilon^2}$. The estimates can be considerably improved by using

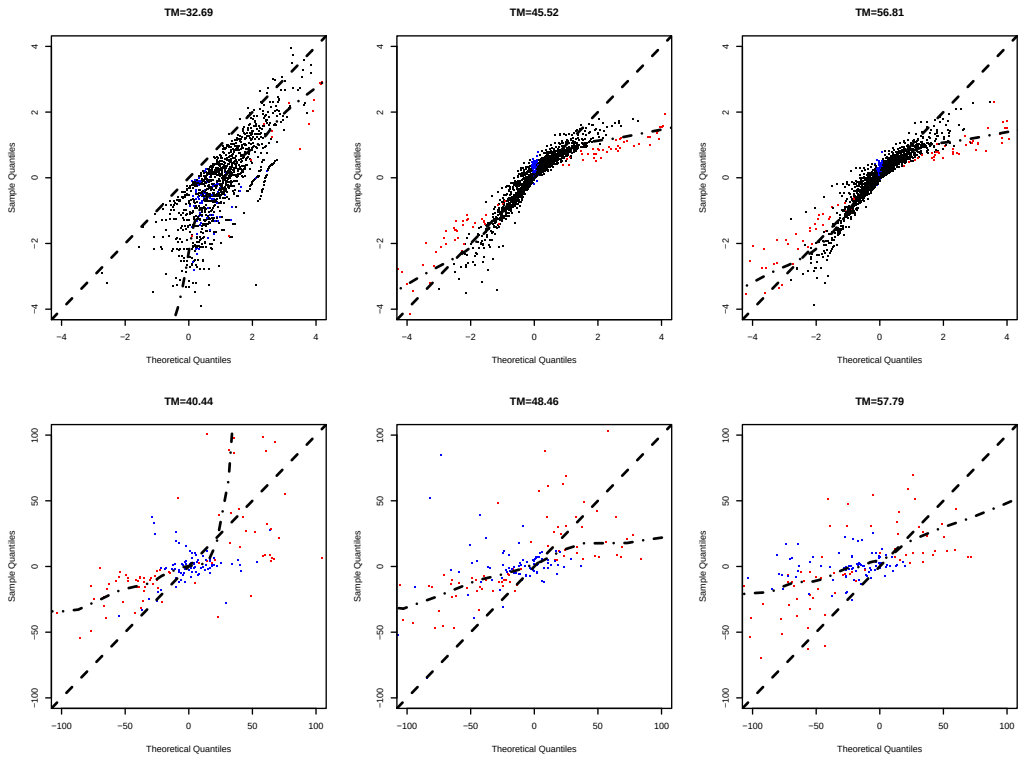

Figure 2.13: QQ-plots of single chip EMI data vs. single chip noise model. The upper row shows QQ-plots in log-log transformed intensity space for the positive controls with the lowest, median and highest melting temperature (TM). The lower row shows QQ-plots in raw intensity space for the negative controls with the lowest, median and highest TM. A single point represents model and observed distribution quantile for a single oligo and chip. Each plot shows quantiles from 48 different chips. According to their number of repeated spots positive control oligos have 32 different quantiles, negative control oligos only 4. Red points show $\frac{1}{32}, \frac{32}{32}$ quantiles for positive controls and $\frac{1}{4}, \frac{4}{4}$ quantiles for negative controls. Blue points show $\frac{16}{32}, \frac{17}{32}$ quantiles for positive controls and $\frac{2}{4}, \frac{3}{4}$ quantiles for negative controls. Dashed lines show the diagonal, dotdashed lines the average QQ-plot computed by a lowess fit. QQ-plots for MI and EMV data look similar.

robust estimates for location and scale instead of mean and variance. For instance median and median absolute deviation:

$$\begin{aligned} \hat{I}^{BG} &= med(O_{q,i}|q \in \mathcal{N}, i = 1...n_r) \\ \hat{\sigma}_\epsilon &= mad(O_{q,i}|q \in \mathcal{N}, i = 1...n_r). \end{aligned} \tag{2.16}$$

Fig. 2.13 shows quantile-quantile plots between the model from Eq. 2.13 and some positive and negative control oligos from several chips. It demonstrates that even for the extreme cases of positive and negative controls the hybridization model of Eq. 2.13 is a reasonably good approximation. However, even for the positive and negative control oligos with median melting temperatures, which should be representative for the detection oligos, there are clear differences at the tails of the positive control oligo distributions and there seems to be a systematic overestimation of the variance of the negative controls.

Variance stabilization

Many standard statistical methods used for the interpretation of microarray data assume that the data are normally distributed or have at least constant variance independent of the mean. The classical approach to stabilize the variance of microarray data is to use a simple log transformation of the raw intensity values [106, 115, 1]. However, as seen in Fig. 2.12 this transformation fails to stabilize the variance for oligos with low intensities.

We would like to find a smooth function $T(O_q)$ that stabilizes the variance for the additive hybridization noise model of Eq. 2.13.

The asymptotic variance $AV[T(O_q)]$ of the transformed intensities can be computed using the delta method as [44, 80]

$$AV[T(O_q)] = \left(\frac{\partial T}{\partial O}(I_q + I^{BG})\right)^2 Var[O_q]. \tag{2.17}$$

This delta method variance estimate is based on a first order Taylor expansion of T around the median intensity $med(O_q) = I_q + I^{BG}$ and gives a reasonably good approximation since most values of O_q will be close to its median.

We seek a $T(O_q)$ so that $AV[T(O_q)]$ is constant. We set

$$AV[T(O_q)] = \left(\frac{\partial T}{\partial O}(I_q + I^{BG})\right)^2 Var[O_q] = c, \tag{2.18}$$

where c is some constant. Inserting Eq. 2.14 and solving for T gives

$$\begin{aligned}\left(\frac{\partial T}{\partial O}(I_q + I^{BG})\right)^2 &= \frac{c}{Var[O_q]} \\ &= \frac{c}{{I_q}^2 S_\eta^2 + \sigma_\epsilon^2} \\ \Leftrightarrow \frac{\partial T}{\partial O}(I_q + I^{BG}) &= \frac{c}{\sqrt{{I_q}^2 S_\eta^2 + \sigma_\epsilon^2}}. \end{aligned} \tag{2.19}$$

And with $O := I_q + I^{BG}$ we get

$$\begin{aligned}\Leftrightarrow \frac{\partial T}{\partial O}(O) &= \frac{c}{\sqrt{(O - I^{BG})^2 S_\eta^2 + \sigma_\epsilon^2}} \\ \Leftrightarrow \int \frac{\partial T}{\partial O}(O) dO &= \int \frac{c}{\sqrt{(O - I^{BG})^2 S_\eta^2 + \sigma_\epsilon^2}} dO. \end{aligned} \tag{2.20}$$

For $c = S_\eta^2$ a solution for this equation is [21]

$$T(O_q) = \ln\left(O_q - I^{BG} + \sqrt{(O_q - I^{BG})^2 + \frac{\sigma_\epsilon^2}{S_\eta^2}}\right). \tag{2.21}$$

Since we had to set $c = S_\eta^2$ we also fixed the asymptotic variance of the transformed data to S_η^2.

The transformation parameters I^{BG}, σ_ϵ and $S_\eta^2 = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$ can be estimated using Eq. 2.12 and Eq. 2.16. Fig. 2.14 shows that if O_q is large the constants I^{BG} and $\frac{\sigma_\epsilon^2}{S_\eta^2}$ become neglectable and $T[\cdot]$ is a simple log transformation. However, when O_q is small or even negative $T[\cdot]$ still gives consistent results with constant variance. Because of these properties $T[\cdot]$ is also referred to as generalized log transformation $GLog[\cdot]$ [135]. Fig. 2.15 shows the stabilized dependence between mean and variance after applying the generalized log transformation.

2.2.2 Between chip noise and normalization

The last sections modeled the intensity distributions of repeatedly spotted oligos on the same microarray. When we look at identical oligonucleotides on different chips that were hybridized with the same sample we observe additional deviations not explained by within chip noise (see Fig. 2.16a). This additional noise component is called between chip noise. It is caused by

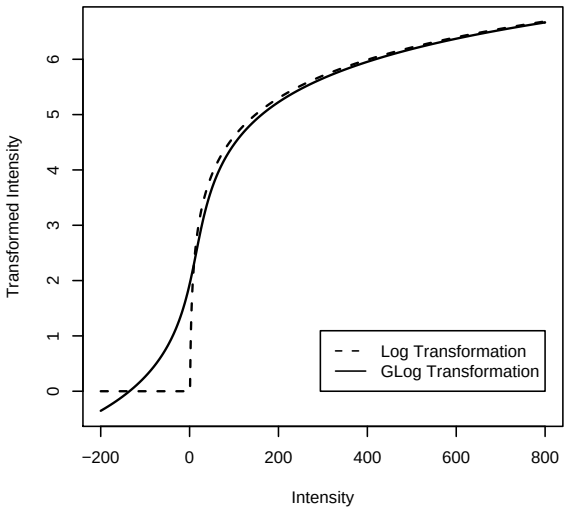

Figure 2.14: Relation between Log and GLog transformations. $Log[I]$ and $GLog[I] - log(2)$ are plotted over a typical range of intensities. Realistic parameters for the GLog transformation were estimated from the *Calibration* dataset (see Appendix A, EMI data, $S_\eta = 0.7$, $\sigma_\epsilon = 25$, $I^{BG} = 15$).

slightly different hybridization conditions and different scanner settings. In the Epigenomics microarray process the intensity amplification of the scanner is optimized manually by an operator to compensate for different overall hybridization intensities. However, considerable scaling differences between chips remain. Here we will simply model this between chip noise as an arbitrary scaling of all oligo intensities on a chip c by a constant factor f_c [146].

For a set of chip repetitions $\mathcal{C} = \{c|c \in \{1, \ldots, n_c\}$ that were all hybridized with PCR product from the same sample the scaling factors $f_c; c \in \mathcal{C}$ can be easily estimated from the 50%-quantile of the overall intensity distribution as

$$f_c = \frac{med_{q \in \mathcal{Q}}(O_q^c)}{med_{c \in \mathcal{C}}(med_{q \in \mathcal{Q}}(O_q^c))}. \tag{2.22}$$

However, formally this equation does not hold when the different chips were hybridized with different samples. Fig. 2.16a shows that samples with different overall degrees of methylation have very different intensity distributions. The median oligo intensity on chips hybridized with an unmethylated sample (left most red line in Fig. 2.16a) is significantly lower than on chips hybridized with a completely methylated sample (right most red line in Fig. 2.16a). This difference can be seen even though individual array repetitions of the same sample show considerable differences. This shows that a change of the average hybridization intensity of an array can be caused by either a technical variation like hybridization conditions or by a higher concentration of target

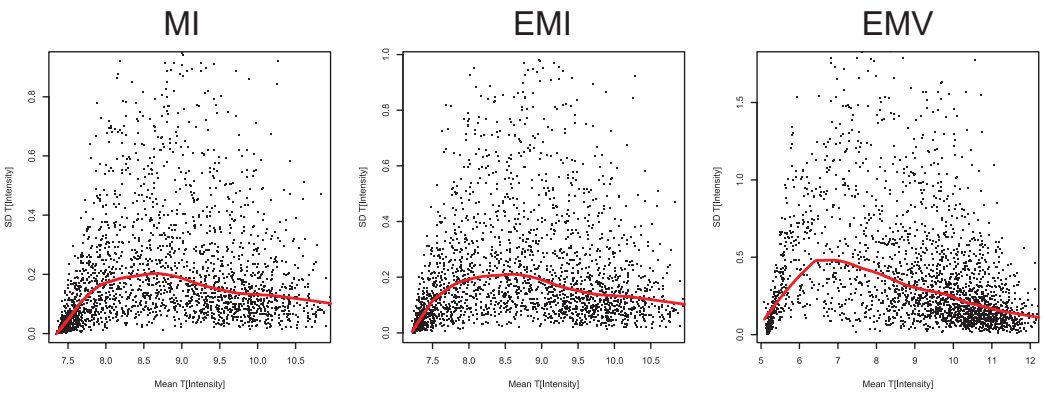

Figure 2.15: Mean-SD dependence of per microarray varaince stabilized hybridization intensities. The three plots correspond to the three different intensity measures MI, EMI and EMV. The generalized log transformation with paramteres estimated from each individual array was used for variance stabilization. Each point shows mean vs. standard deviation of the GLog transformed hybridization intensities for one detection oligo on a single microarray computed from the respective spot repetitions on the respective microarray. Plots were generated from a total of 48 microarrays and 50 detection oligos. The red line shows the average standard deviation for a given intensity computed by a lowess fit. Each microarray contained detection oligos in 4-fold redundancy. Note that SD estimates are very imprecise due to their estimation from only 4 data points.

molecules in the original sample. The first variation is a systematic experimental bias and we want to eliminate it by normalization. The second variation is what we actually want to measure. The separation between these two effects is one of the major difficlulties in the normalization of mRNA microarrays [146, 174].

In the case of our methylation microarrays we have a considerable advantage. The standard Epigenomics array design always contains identical numbers of CG and TG oligos for each CpG position. These oligos show an inverse hybridization behaviour for different amounts of methylated DNA in the target sample. The CG oligo intensity increases with higher degrees of methylation; TG oligo intensity decreases. As a result the sum of CG and TG oligo intensities of a single oligo family (the set of oligos binding to the same CpG position or CpG cluster) is approximately constant and independent from the degree of methylation at the respective CpG position. Fig. 2.16b shows the distribution of average oligo family intensities defined as

$$mean_{q\in\mathcal{F}_p} O_q^c := \frac{1}{|\mathcal{F}_p|} \sum_{q\in\mathcal{F}_p} O_q^c, \tag{2.23}$$

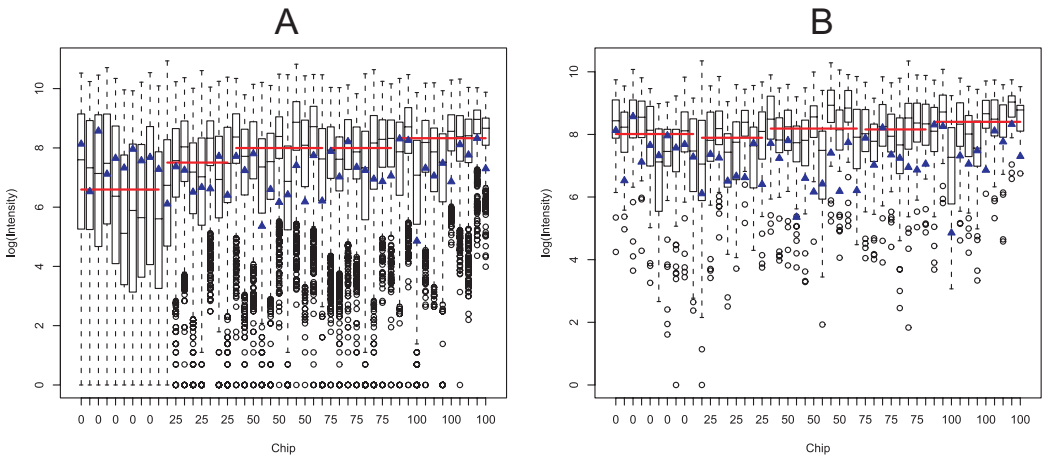

Figure 2.16: Distribution of detection oligo hybridization intensities. A) Single oligo intensities B) Average oligo family intensities. For individual microarrays hybridized with mixtures of artificially methylated DNA (x-axis, 0%, 25%, 50%, 75% and 100% methylation) the distribution of individual oligo intensities or average oligo family intensities are shown as boxplots. Red lines are median intensities for chips of identical methylation rate. Blue triangls are median positive control oligo intensities.

with $\mathcal{F}_p = \{q|p = p(q)\}$ as the set of all oligos querrying CpG position p. No systematic difference between the different samples can be observed.

The estimation of the overall scaling factor f_c from Eq. 2.22 can be easily modified to work on average oligo family intensities:

$$f_c = \frac{med_{p\in\mathcal{P}}(mean_{q\in\mathcal{F}_p}(O_q^c))}{med_{c\in\mathcal{C}}(med_{p\in\mathcal{P}}(mean_{q\in\mathcal{F}_p}(O_q^c)))}, \tag{2.24}$$

where $\mathcal{P}$ is the set of all CpG positions covered by oligos on the chip.

Normalizing the chips with this factor results in a moderate reduction of between chip variance (see Fig. 2.17). A considerable amount of variance remains due to higher order differences in intensity distribution. However, as can be seen by comparing Fig. 2.16 and Fig. 2.18 the median within sample intensities do not change while the within sample noise is reduced (individual chip intensity distributions are more similar). This means the normalization completely retains biologically relevant between sample variation while reducing the between array noise.

After normalization the between chip variability is minimized and repeated measurements of the same sample on different chips should be approximately distributed according to Eq. 2.13. This means the derived variance stabilizing transformation from Eq. 2.21 should be applicable. Fig. 2.19 shows that the resulting data have indeed approximately constant variance.

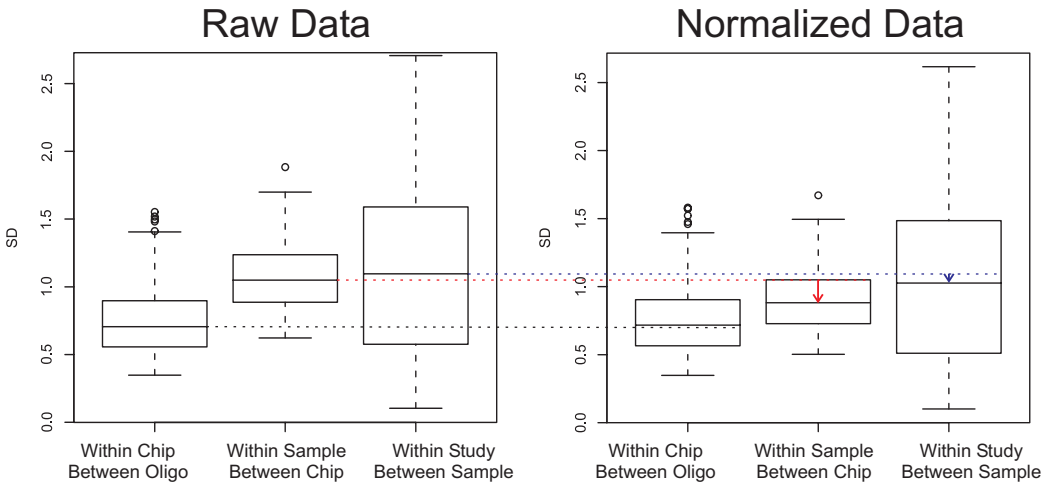

Figure 2.17: Distributions of within chip (between 4 oligo replications), within sample (between 10 chip replications) and within study (between 0%, 25%, 50%, 75% and 100% samples) standard deviations over 50% brightest oligos in log intensity space. Raw data is plotted on the left, normalized data on the right. Dashed horizontal lines are median standard deviations of raw data. Red and blue arrows in the right plot indicate the between chip and between sample variance reductions after normalization.

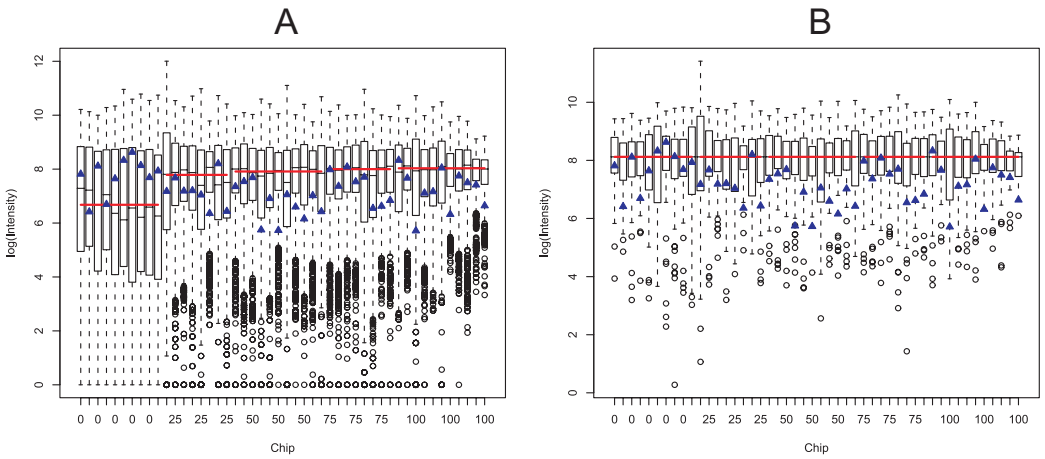

Figure 2.18: Distribution of detection oligo hybridization intensities after normalization. A) Single oligo intensities B) Average oligo family intensities. For individual microarrays hybridized with mixtures of artificially methylated DNA (x-axis, 0%, 25%, 50%, 75% and 100% methylation) the distribution of normalized individual oligo intensities or normalized average oligo family intensities are shown as boxplots. Red lines are median intensities for chips of identical methylation rate. Blue triangls are median positive control oligo intensities.

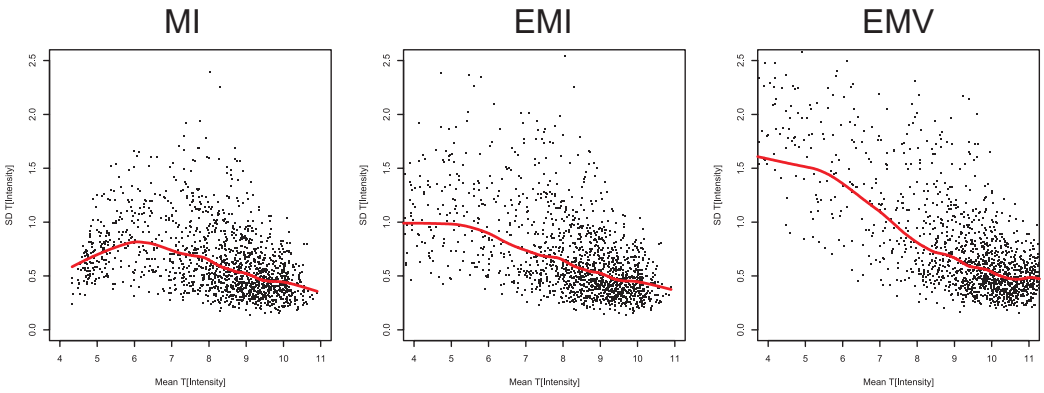

Figure 2.19: Mean-SD dependence of globally varaince stabilized hybridization intensities. The three plots correspond to the three different intensity measures MI, EMI and EMV. The generalized log transformation with paramteres estimated from all arrays after normalization was used for variance stabilization. Each point shows mean vs. standard deviation of the normalized and GLog transformed hybridization intensities for one detection oligo and one methylation level computed from the respective spot and chip repetitions from the respective methylation level. Plots were generated from a total of 5 methylation levels and 50 detection oligos. The red line shows the average standard deviation for a given intensity computed by a lowess fit. Each microarray contained detection oligos in 4-fold redundancy and at least 8 chips were hybridizaed per methylation level.

However, it seems the variance stabilization works better for MI and EMI data as compared to the EMV data. Fig. 2.20 shows that the distribution of the EMI data is even close to normal.

2.2.3 Expected hybridization intensities

In the last two sections we have derived noise models for stochastic variations of hybridization intensities within and between single microarrays. We have shown that these stochastic variations depend on the expected hybridization intensities of the oligos in a specific hybridization reaction but are independent of the respective oligo sequence characteristics. On the other hand the expected hybridization intensity of an oligomere is a direct function of its sequence and the amount of matching amplificates in the hybridization reaction. In this section we will derive a simplified model for the very complex kinetics of the hybridization reaction that will enable us later on to construct practical algorithms for the computation of methylation proportions from observed hybridization intensities.

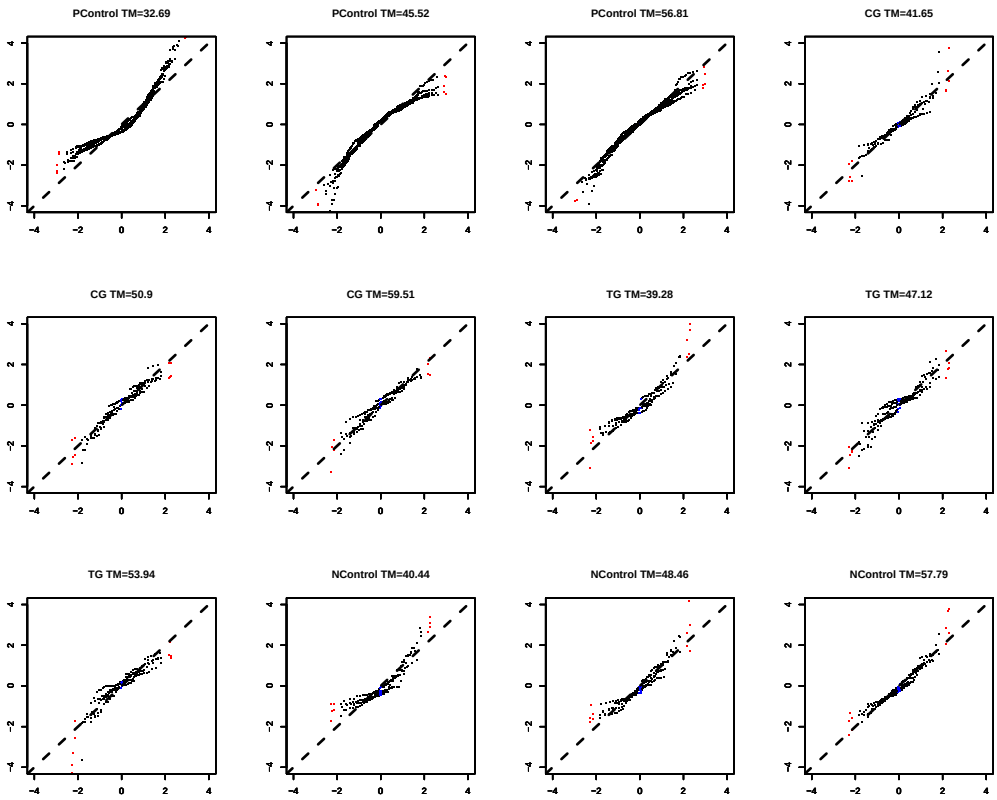

Figure 2.20: QQ-plots of normalized and variance stabilized EMI data vs. normal distribution. The plots show QQ-plots in GLog transformed, mean centered and variance normalized intensity space for positive controls, CG detection oligos, TG detection oligos and negative control oligos with the respective lowest TM, median TM and highest TM. A single point represents normal and observed distribution quantile for a single oligo and all chips and spot repetitions from the same methylation level. Each plot shows quantiles from 5 different methylation levels. According to their number of repeated spots (32 for positive control and 4 for detection and negative control oligos) and the number of chip repetitions per methylation level (minimum 8) the different plots show a varying number of quantiles. Red points show the respective lowest and highest quantile, blue points the two middle quantiles. Dashed lines show the diagonal, dotdashed lines the average QQ-plot computed by a lowess fit. QQ-plots for MI and EMV data look similar.

Hybridization kinetics

The simple hybridization reaction between two complementary nucleic acid molecules in solution can be described by the following dissociation reaction [169, 129, 37]

$$R + L \underset{k^r}{\overset{k^f}{\rightleftharpoons}} C, \tag{2.25}$$

where R is the concentration of oligomeres available for hybridization, L is the concentration of target amplificates available for hybridization and C is the concentration of bound duplexes. k_f and k_r are the respective reaction rates for formation and deformation of duplexes.

Assuming that there is an excess of available oligomeres compared to the number of target molecules the foreward hybridization reaction follows a pseudo first order kinetics [169, 129]. The concentrations at equilibrium are connected by the quotient of the reaction rates:

$$C = \frac{k^f}{k^r} L. \tag{2.26}$$

This model assumes that oligo and amplificate are in solution and is only an approximation of the complex hybridization kinetics on a solid surface microarray [51, 15]. But it shows that the number of amplificates hybridized to an oligo is proportional to the concentration of the respective amplificate. This is true for all oligomere-amplificate pairs. The designed pairs with matching sequences as well as oligo-amplificate pairs with missmatching sequences. The hybridization of amplificates is specific to matching oligomeres because the reverse rate k^r is much higher for missmatch duplexes than for match duplexes [37]. The difference between matching k^r and missmatching k^r is very sequence dependent and determines the amount of observed unspecific (all amplificates bind to an oligo) and cross hybridization (a specific not matching amplificate binds in addition to the matching amplificate).

After the hybridization experiment is performed and all amplificates not bound in stable duplexes are removed by the washing step the measured fluorescent oligo intensities on the microarray are proportional to the concentration of stable duplexes C.

A model of hybridization

We model the outcome of a single hybridization experiment as follows. A tissue sample is represented by its set of amplificates $\mathcal{R}$ with concentrations $\boldsymbol{a}$ (a_r concentration of amplificate r, with $r \in \{1, \ldots, |\mathcal{R}|\}$). Note that if PCR amplification would have perfect efficiency for all amplificates all a_r would

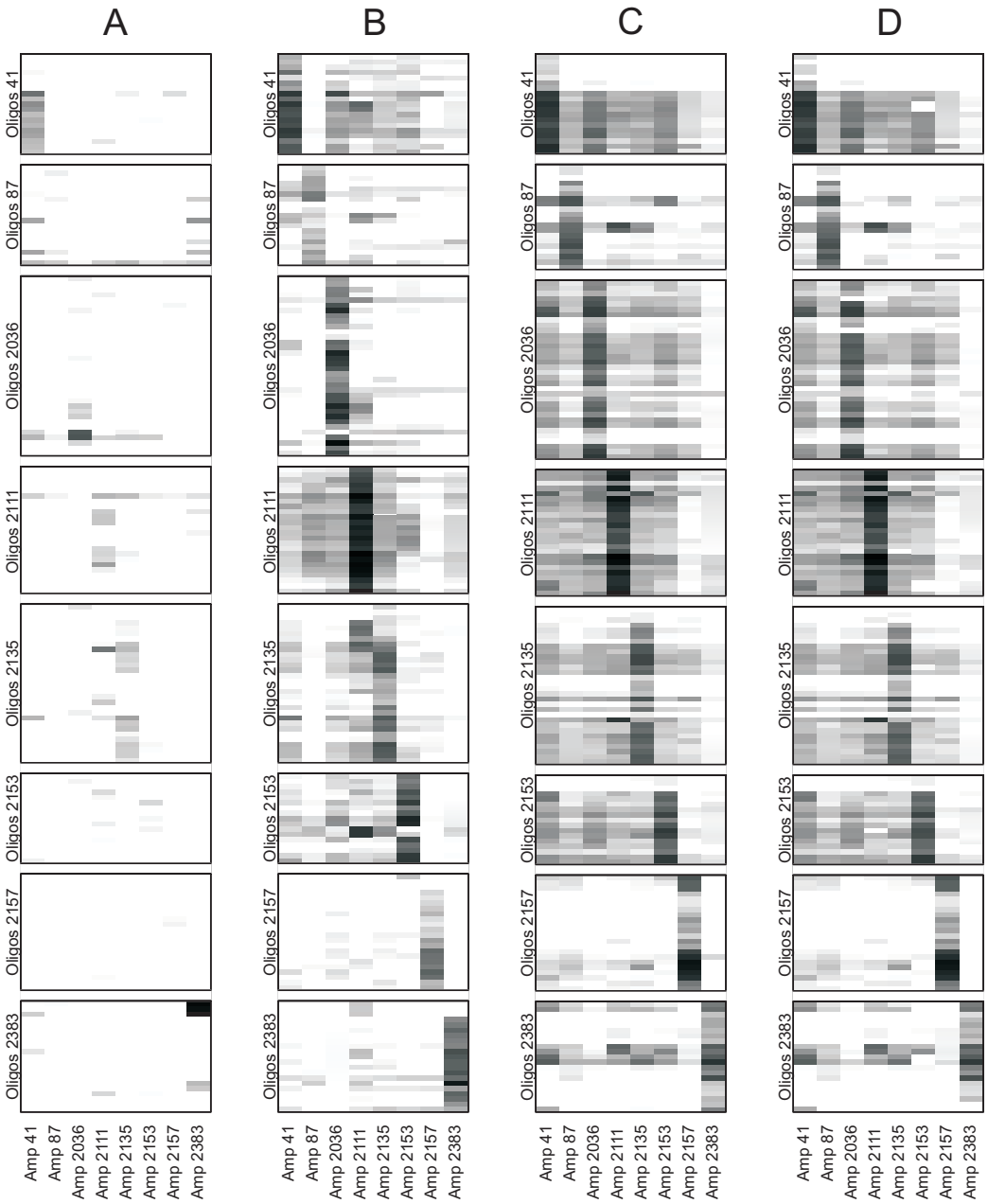

Figure 2.21: Equilibrium constant matrices. The figures show experimental estimates of the equilibrium constant matrices for a microarray design with 8 amplificates. Rows are single oligos, columns are amplificates. Grey values code Log transformed equilibrium constant estimates, where white corresponds to very small or negative constants and black to very high constants. CG and TG oligo constants are shown separately. For both oligo types the respective missmatch matrix and the difference matrix between match and missmatch are shown: A) K^- for CG oligos B) $K^+ - K^-$ for CG oligos C) K^+ for TG oligos D) $K^- - K^+$ for TG oligos

be identical with $a_r = N^{PCR}/V$ (with V being the volume of the solution). Each amplificate can occur with different methylation patterns depending on the number of covered CpG positions and the methylation of the original DNA sample. To simplify the following derivations we assume here that CpG sites within an amplificate are comethylated (have all identical methylation status). To obtain results for individual CpG clusters each cluster can be treated as a "virtual" amplicon.

Using the comethylation assumption the methylation unspecific amplificate concentration vector $\boldsymbol{a}$ can be expressed as the sum of methylated and unmethylated amplificate concentrations as

$$\boldsymbol{a} = \boldsymbol{a}^- + \boldsymbol{a}^+. \tag{2.27}$$

Again these concentrations are simply the volume normalized fragment numbers $a_r^- = N_p^{PCR-}/V$ and $a_r^+ = N_p^{PCR+}/V$ with amplificate r covering CpG position p.

At equilibrium the amount of oligo-amplificate duplexes at each oligo is then given by the following equation:

$$\boldsymbol{o} = K^- \boldsymbol{a}^- + K^+ \boldsymbol{a}^+, \tag{2.28}$$

where $[\boldsymbol{o}]_q, q \in \{1, \ldots, |\mathcal{Q}|\}$ is the concentration of stable oligo-amplificate duplexes at oligo q and $[K^-]_{qr} = \frac{k_{qr}^f}{k_{qr}^{r,-}}$ and $[K^+]_{qr} = \frac{k_{qr}^f}{k_{qr}^{r,+}}$ are the matrices of equilibrium constants between oligo q and unmethylated or methylated amplificate r. Note that only the reverse rate coefficient is different for unmethylated and methylated amplificates because duplex formation is primarily dependent on the interaction frequency of oligos and amplificates, whereas duplex deformation depends on the sequence match and the resulting melting temperature of the duplex [37, 15].

To determine the equilibrium constant matrices experimentally we simplify the experimental conditions so that either all amplificates are completely methylated or completely unmethylated and the concentartion of all amplificates is equal. Furthermore we label exactly one amplificate $r, r \in \{1, \ldots, |\mathcal{R}|\}$ of the amplificate set $\mathcal{R}$ with the fluorecent dye CY3 and all remaining amplificates with the fluorecent dye CY5. When we measure the fluorecence intensities of the respective microarray on the CY3 channel the observed oligo intensities are proportional to the equilibrium constants $[K]_{.,r}$ for amplificate r. Repeating this experiment with unmethylated and methylated DNA for all amplificates from set $\mathcal{R}$ gives us estimates for the equilibrium constant matrices K^- and K^+. Fig. 2.21 shows examples for such estimates.

In analogy to Eq. 2.4 we define the following mapping between oligos and amplificates:

$$r(q) := \{r \in \mathcal{R} \mid \text{Amplificate } r \text{ that oligo } q \text{ is binding to}\}. \tag{2.29}$$

Additionally we define the following indicator function to identify oligos q that were designed to detect an amplificate r:

$$\delta^{OA}(q,r) := \begin{cases} 1, & r(q) = r \\ 0, & r(q) \neq r. \end{cases} \tag{2.30}$$

Ideally the equilibrium constant matrices should be positive for all matching oligo-amplificate pairs $\{(q,r)|\delta^{OA}(q,r) = 1\}$ and 0 for all other pairs. This corrsponds to a situation where stable duplexes are only formed between matching oligos and amplificates. Additionally the K^+ matrix should only have positive elements for CG oligomeres and the K^- matrix for TG oligomeres. However, Fig. 2.21 shows that in reality:

- Many oligos have a unspecific base hybridization independent of the amplificate sequence.
- Some oligos show cross hybridization with specific missmatch amplificates.
- Many TG oligos show a high affinity for the matching methylated amplificate, i.e. these TG oligos are not methylation specific.

Although we showed that it is possible to model the complete hybridization behaviour of a specific microarray design and estimate its equilibrium constants experimentally this procedure is not practical. The experimental effort for real world arrays with 64 or more amplificates is prohibitive.

A simplified model of hybridization

In order to use a hybridization model in practice for the estimation of DNA methylation proportions from observed hybridization intensities we have to be able to determine all model parameters in a cost effective way with minimal experimental effort. In the following we will therefore simplify Eq. 2.28 until the number of parameters is sufficiently reduced to allow for an easy estimation procedure. Some of the necessary simplifications and assumptions will be quite severe and have to be kept in mind when interpreting model predictions and results.

The majority of the parameters in the hybridization model of Eq. 2.28 are the oligo-amplificate missmatch elements corresponding to the equilibrium constants for unspecific and cross hybridization. In order to simplify the hybridization model and facilitate an estimation of all paramters in practice we will assume that cross hybridization can be neglected. The equilibrium constant matrices from Eq. 2.28 then simplify to

$$\begin{aligned} \left[K^-\right]_{qr} &= k_q^{0,-} + k_{qr}^- \delta^{OA}(q,r) \\ \left[K^+\right]_{qr} &= k_q^{0,+} + k_{qr}^+ \delta^{OA}(q,r), \end{aligned} \tag{2.31}$$

where $k_q^{0,-}$ and $k_q^{0,+}$ model the oligo specific but amplificate unspecific background hybridization and k_{qr}^- and k_{qr}^+ model the oligo-amplificate specific hybridization. Since we only retain the matching oligo-amplificate equilibrium constants a pure oligo based indexing is sufficient:

$$\begin{aligned} k_q^- &:= k_{qr}^- \text{ with } \delta^{OA}(q,r) = 1 \\ k_q^+ &:= k_{qr}^+ \text{ with } \delta^{OA}(q,r) = 1. \end{aligned} \tag{2.32}$$

Using Eq. 2.27 we can then rewrite Eq. 2.28 as

$$\begin{aligned} \boldsymbol{o} &= K^+ \boldsymbol{a}^+ + K^- \left(\boldsymbol{a} - \boldsymbol{a}^+\right) \\ &= \left(K^+ - K^-\right) \boldsymbol{a}^+ + K^- \boldsymbol{a} \\ o_q &= \left(k_q^{0,+} - k_q^{0,-}\right) \|\boldsymbol{a}\|_1 + \left(k_q^+ - k_q^-\right) a^+_{r(q)} + \\ &\quad k_q^{0,-} \|\boldsymbol{a}\|_1 + k_q^- a_{r(q)}. \end{aligned} \tag{2.33}$$

Now we make the assumption that the unspecific hybridization is independent of the concrete amplificate methylation patterns:

$$k_q^0 := k_q^{0,+} = k_q^{0,-}. \tag{2.34}$$

Under the additional assumption that the total amplificate concentration vector $\boldsymbol{a}$ is identical (or made identical by normalization) for each hybridization experiment we can define the following constants:

$$\begin{aligned} k_q &:= k_q^+ - k_q^- \\ b_q &:= k_q^0 \|\boldsymbol{a}\|_1 + k_q^- a_{r(q)}. \end{aligned} \tag{2.35}$$

This simplifies Eq. 2.33 to

$$o_q = k_q a^+_{r(q)} + b_q. \tag{2.36}$$

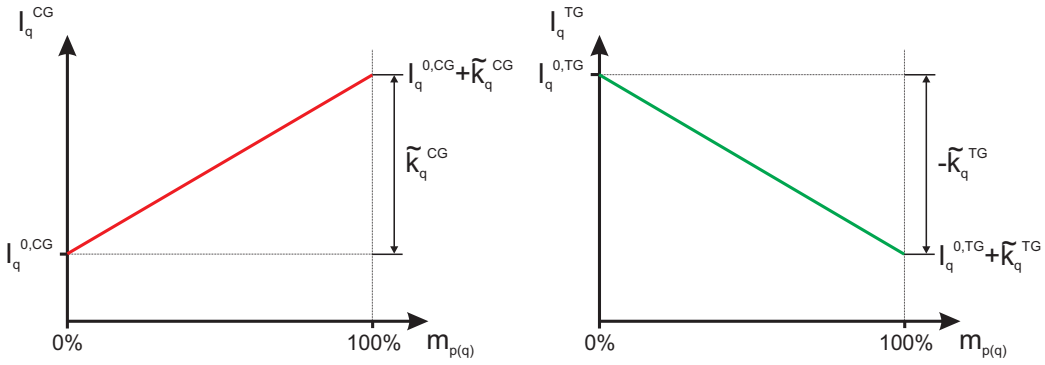

Figure 2.22: Simple hybridization intensity model. The plots show the linear dependence between hybridization intensities and DNA methylation rate for CG (left) and TG (right) oligomeres.

Normalizing this equation to the total amount of amplificate $a_{r(q)}$ and using the proportionality between amplificate concentration and hybridization intensity gives

$$I_q = \tilde{k}_q m_{p(q)} + I_q^0, \tag{2.37}$$

where I_q is the expected intensity, $\tilde{k}_q$ is the product of the original equilibrium constant difference k_q and the concentration-intensity conversion factor, $m_{p(q)}$ is the methylation proportion of the amplificate queried by oligo q and I_q^0 is the unspecific background intensity of oligo q. Fig. 2.22 visualizes this hybridization model for CG and TG oligos.

Although we derived Eq. 2.37 only for comethylated amplificates we can generalize it to single CpG dinucleotides that are queried by our detection oligos. In this case the methylation proportion $m_{p(q)}$ is the number of individual amplificates methylated at the CpG position queried by oligo q devided by the total number of copies of this amplificate.

The following list summarizes the major assumptions we made when deriving the simplified hybridization model of Eq 2.37:

- Excess of oligomeres compared to the number of target molecules available for hybridization
- No cross hybridization
- Amplificate unspecific background hybridization is independent of amplificate methylation status
- Identical amplificate concentrations for all samples or normalization of signal intensities

These assumptions imply that all columns of the missmatch matrix K^- are identical and that the match-missmatch difference matrix $K^+ - K^-$ is diagonal. A comparison with the experimental results of Fig. 2.21 show that these assumptions are only approximations. That has to be kept in mind when interpreting model predictions and results.

2.3 Quantification of DNA methylation

In the previous section we derived a quantitative model that explains how the intensity values reported by our microarrays are generated from DNA mixtures with arbitrary methylation patterns. In this section we will derive algorithms that infere the methylation state of the original DNA from the observed microarray intensities.

2.3.1 Methylation scores

A standard Epigenomics microarray contains exactly one CG and one TG oligomere for each CpG cluster that is queried. These CG and TG oligos are spotted adjacent to each other in 4 pairs distributed over the array surface. As mentioned in Section 2.1.2 it is possible to design several overlapping oligos to query the same CpG position. However, to keep things simple we will start with using the standard convention that there is exactly one CG/TG oligo pair per CpG position. These matching pairs will be referenced by the same oligo index q which has a one to one mapping to the respective CpG position $p = p(q)$. The observed and theoretically expected hybridization intensities corresponding to these oligo pairs will be referenced as $O_{q,i}^{CG}$, $O_{q,i}^{TG}$, $I_{q,i}^{CG}$ and $I_{q,i}^{TG}$, with i as the replicate index.

Usually a sample is hybridized onto more than one microarray to generate more repeated measurements and facilitate a more accurate estimate of methylation. The hybridization intensities of these replicates have to be aggregated into a single score for each CG/TG oligo pair. This methylation score should reflect the amount of methylation in the respective sample at the CpG position queried by the respective CG/TG oligo pair.

The methylation score $S_{p(q)}\left(\{O_{q,i}^{CG}, O_{q,i}^{TG}\}_{i=1\ldots n_r}\right)$ is a function of the n_r observed intensity pairs $\{O_{q,i}^{CG}, O_{q,i}^{TG}\}_{i=1\ldots n_r}$ from the CG/TG oligo pair q designed to measure CpG position p. We assume that the data has been normalized to remove between chip variations and do not distinguish between oligo repetitions on the same chip or on different chips. What we are interested in is the relation between the observed methylation score S_p and the real proportion of methylated DNA at CpG position p. This relation is given

by the conditional probability distribution $P(S_p|m_p)$, which ideally has the following properties:

1. **Monotonicity** $\frac{d}{dm_p}E[S_p|m_p] > 0$
   The expected methylation score S_p should be monotone increasing with the true methylation m_p.

2. **Discrimination** $\frac{\frac{d}{dm_p}E[S_p|m_p]}{\sqrt{Var[S_p|m_p]}} > c_{min}$
   Small changes in methylation should be detectable. That means the change of the score compared to its standard deviation should be high enough. An alternative measure of discrimination for a methylation score is the probability of measuring a higher score at a higher methylation level: $P(S'_p > S_p|m'_p, m_p)$ with $m'_p > m_p$. In practice this probabiliy can be approximated by the area under the ROC curve (AUC) between score measurements at methylation levels m'_p and m_p [71].

3. **Linearity** $E[S_p|m_p] \propto m_p$
   A certain change in the score always corresponds to the same change in methylation.

4. **Identity / Accuracy** $E[(S_p(m_p) - m_p)^2|m_p] < c_{max}$
   The methylation score actually reports the methylation proportion with a certain accuracy. Overall accuracy can be decomposed in the following two components:

   (a) **Bias** $E[S_p|m_p] - m_p$
   This is the systematic error component.

   (b) **Precision/Variance** $Var[S_p|m_p]$
   This is the stochastic error component.

5. **Variance Stability** $Var[S_p|m_p] = c$
   The variance of the score is independent from the measured methylation.

Monotonicity and good discrimination are necessary properties of a good methylation score. They enable us to make essential biological observations like "sample A is hypermethylated compared to sample B". Linearity and identity enable us to make more detailed statements about biology like "sample A is 2-fold hypermethylated compare to sample B" or "25% of the DNA in sample A is methylated". A score with constant variance simplifies the following data analysis steps because most statistical standard methods assume additive or even white noise. Note that the constant variance property

partially contradicts the identity property which requires the methylation score to be in the intervall $[0, 1]$.

With the noise model from Eq. 2.13 and the hybridization model of Eq. 2.37 we have a generative model of the observed hybridization intensities O_q given the methylation of the respective CpG $m_{p(q)}$:

$$\begin{aligned} O_q^{CG} &= I^{BG} + \left(\tilde{k}_q^{CG} m_{p(q)} + I_q^{0,CG}\right) e^{\eta} + \epsilon. \\ O_q^{TG} &= I^{BG} + \left(\tilde{k}_q^{TG} m_{p(q)} + I_q^{0,TG}\right) e^{\eta} + \epsilon. \end{aligned} \tag{2.38}$$

Based on this generative model we are able to predict the properties of a methylation score given the global noise parameters σ_η, σ_ϵ and I^{BG} (estimated as described in Section 2.2.1) and given the hybridization parameters of the involved CG and TG oligomeres $\tilde{k}_q$ and I_q^0.

The additive structure of the hybridization noise makes it impossible to derive the methylation score distributions $P(S_p|m_p)$ in closed form. But we can use the model to numerically determine the methylation score distributions for typical parameter values.

In the following sections we will derive four methylation scores with different properties. We will start with the simple methylation proportion and log ratio scores previously described in the DNA methylation microarray and sequencing literature [115, 1, 101]. Then we will introduce a generalized log ratio score based on the variance stabilizing transformation derived in Section 2.2.1. Finally we will derive a maximum likelihood score that takes full advantage of the generative hybridization model derived in the previous sections.

2.3.2 Ratios and differences of CG and TG oligos

A straight foreward way to combine CG and TG oligo intensities is to use ratios and differences. A high CG oligo intensity indicates a high methylation and should therefore be the minuend in a difference and the dividend in a ratio. A high TG oligo intensity indicates a low methylation and should be the subtrahend in a difference and the divisor in a ratio.

An advantage of using ratios or log-differences of CG and TG oligo intensities is that they are invariant to a global and even local intensity rescaling. Since CG and TG oligos are right next to each other on a typical microarray both intensities are scaled by the same factor which cancels out by taking the ratio or the log difference. A normalization as described in Section 2.2.2 is therefore unnecessary.

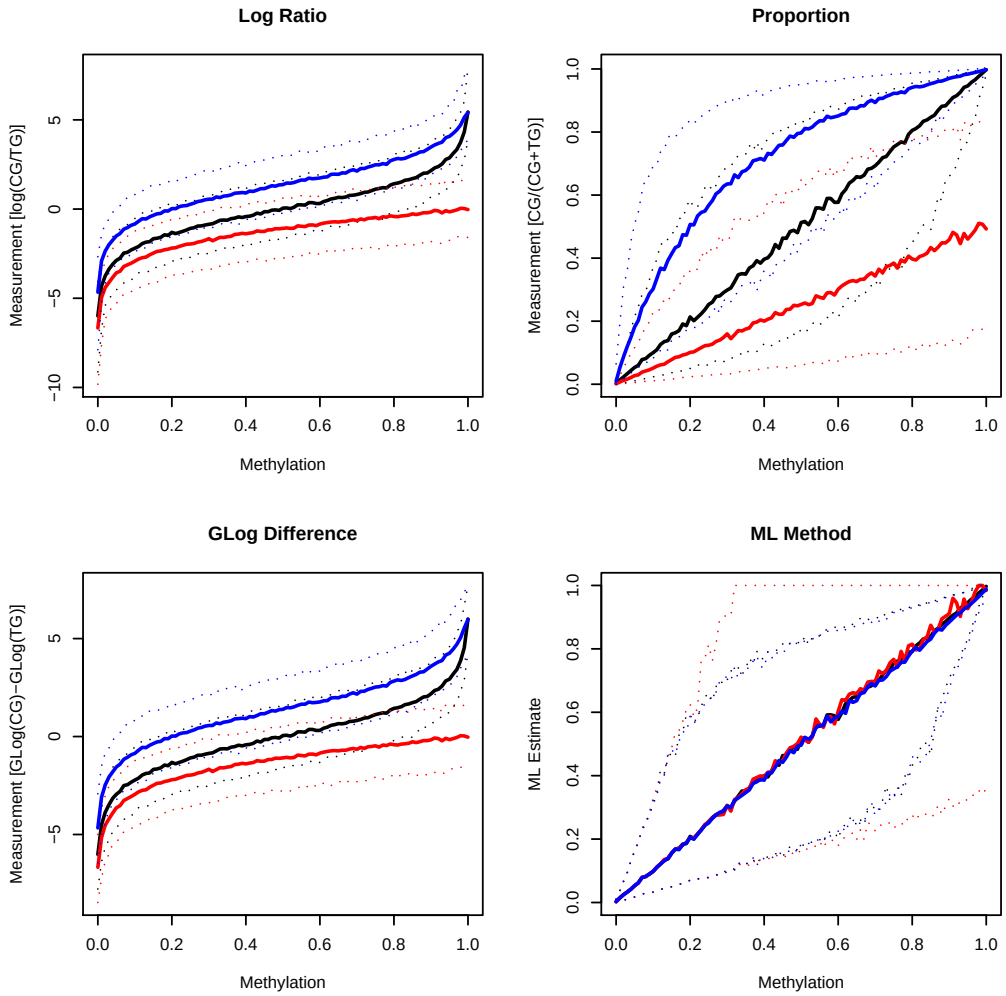

Figure 2.23: Methylation score distributions. The four plots show the distributions of the log ratio score, the generalized log difference score, the proportion score and the maximum likelihood score. For every methylation level the respective median score (solid line) and the 5% and 95% quantiles (dotted lines) are shown. Score distributions were numerically simulated by generating 1000 random samples according to the generative model of Eq. 2.38. Realistic model parameters were estimated from the *Calibration* dataset (see Appendix A, EMI data, $S_\eta = 0.7$, $\sigma_\epsilon = 25$, $I^{BG} = 15$). The black curve shows the scores for an ideal oligo pair: $I_q^{0,CG} = 0$ and $I_q^{0,TG} = \tilde{k}_q^{CG} = -\tilde{k}_q^{TG} = 5000$. The red curve shows the scores for an TG oligo with unspecific background hybridization: $I_q^{0,CG} = 0, I_q^{0,TG} = -2\tilde{k}_q^{TG}$ and $\tilde{k}_q^{CG} = -\tilde{k}_q^{TG} = 5000$. The blue curve shows the scores for an TG oligo with reduced dynamic range: $I_q^{0,CG} = 0$ and $\tilde{k}_q^{CG} = 5000$, $I_q^{0,TG} = -\tilde{k}_q^{TG} = 2500$.

Methylation proportion

When we further simplify Eq. 2.37 and assume that the equilibrium constant differences and background intensities of all matching CG-TG oligo pairs are equal:

$$\begin{aligned} \tilde{k}_q^{CG} &= -\tilde{k}_q^{TG} \\ I_q^{0,CG} &= 0 \\ I_q^{0,TG} &= -\tilde{k}_q^{TG}, \end{aligned} \tag{2.39}$$

then $I_q^{CG} + I_q^{TG}$ is constant and the proportion of methylated DNA at CpG position $p(q)$ is equal to $I_q^{CG}/(I_q^{CG} + I_q^{TG})$:

$$\begin{aligned} \frac{I_q^{CG}}{I_q^{CG} + I_q^{TG}} &= \frac{\tilde{k}_q^{CG} m_{p(q)}}{\tilde{k}_q^{CG} m_{p(q)} + \tilde{k}_q^{TG} m_{p(q)} + I_q^{0,TG}} \\ &= \frac{\tilde{k}_q^{CG} m_{p(q)}}{\tilde{k}_q^{CG} m_{p(q)} - \tilde{k}_q^{CG} m_{p(q)} + \tilde{k}_q^{CG}} \\ &= \frac{m_{p(q)}}{m_{p(q)} - m_{p(q)} + 1} \\ &= m_{p(q)}. \end{aligned} \tag{2.40}$$

In analogy to this proportion of expected hybridization intensities we can define the methylation proportion score on the observed intensities as

$$S_{p(q)}\left(\{O_{q,i}^{CG}, O_{q,i}^{TG}\}_{i=1\ldots n_r}\right) = med_{i=1\ldots n_r}\left(\frac{max(O_{q,i}^{CG}, c)}{max(O_{q,i}^{CG}, c) + max(O_{q,i}^{TG}, c)}\right), \tag{2.41}$$

where c is a small positive intensity. This score basically ignores the hybridization noise. It only makes sure that no negative, undefined or infinite scores can occure by cutting off negative or zero intensities.

As the simulation results in Fig. 2.23 show, the proportion score gives very good estimates of the methylation proportion when the assumptions of Eq. 2.39 are met. However, it gets arbitrarily rescaled by a unspecific hybridization background and even becomes non-linear when CG and TG oligo have different equilibrium constants. Its variance is not stable, especially at the extreme methylation levels of 0% and 100% where the variance rapidly converges to 0.

Variance stabilized differences

If our main goal is to derive a score with stable variance we can simply start with the variance stabilized CG and TG intensities. Assuming the simple

log-normal distribution model of Eq. 2.8 we can define the simple log ratio score $\log\left(\frac{I_q^{CG}}{I_q^{TG}}\right) = \log(I_q^{CG}) - \log(I_q^{TG})$ as

$$S_{p(q)}\left(\{O_{q,i}^{CG}, O_{q,i}^{TG}\}_{i=1\ldots n_r}\right) = med_{i=1\ldots n_r}\left(\log\left(\frac{max(O_{q,i}^{CG}, c)}{max(O_{q,i}^{TG}, c)}\right)\right), \quad (2.42)$$

where c is a small positive intensity. Because this score is the difference of two normaly distributed random variables it is itself normal and has constant variance. However, as we have shown in Section 2.2.1 the log-normal property is only approximately correct for high intensities. In the low intensity regime the simple log ratio score makes only sure that no undefined or infinite scores can occure by cutting off negative or zero intensities.

In order to better handle low intensity values we can use the generalized log transformation from Eq. 2.21 to define the generalized log difference score as

$$S_{p(q)}\left(\{O_{q,i}^{CG}, O_{q,i}^{TG}\}_{i=1\ldots n_r}\right) = med_{i=1\ldots n_r}\left(T[O_{q,i}^{CG}] - T[O_{q,i}^{TG}]\right). \quad (2.43)$$

This score handles all intensity values in a consistent way by taking the additive background noise into account. Since the transformed CG and TG intensities are approximately normaly distributed we can expect the GLog difference score to be approximately normal with constant variance.

Fig. 2.23 shows that both scores have indeed constant variance over the full range of methylation proportions. However, for very high or very low methylation rates the scores become highly non-linear. Both scores have no easy to interprete relation to the absolute methylation level. It is only clear that they are monotone increasing with the methylation proportion.

The numerical simulation shows also that the log ratio and the GLog difference score give almost identical results. This is not surprising since for a working CG/TG oligo pair at least one of two oligos will always have a relatively high intensity. As shown in Fig. 2.14 Log and GLog transformation are approximately identical for high intensities. Since Log ratio and GLog difference scores are dominated by the oligo with the higher intensity both transformations give almost identical results. Note that this is not necessarily true for not properly working CG/TG oligo pairs with weak hybridization signals on both oligos.

2.3.3 A maximum likelihood estimator

With the noise model from Eq. 2.13 and the hybridization model of Eq. 2.37 we have a generative model of the observed hybridization intensities O_q given the methylation of the respective CpG position $m_{p(q)}$:

$$O_q = I^{BG} + \left(\tilde{k}_q m_{p(q)} + I_q^0\right) e^{\eta} + \epsilon. \tag{2.44}$$

This equation is identical for CG and TG oligomeres and from now on we will drop the requirement of matching CG/TG oligo pairs. For the remainder of this section the oligo index q will refer to a single CG or TG oligomere and a CpG position p can be covered by an arbitrary set of CG and/or TG oligomeres $\mathcal{Q}_p$.

Using the maximum likelihood (ML) framework [14] we can derive optimal estimates for the methylation m_p from the observed hybridization intensities. Compared to the simple methylation scores from the previous section the ML score will have the advantage that it takes background hybridization and equilibrium constants into account. An additional advantage is that the ML score can naturally combine measurements from oligo families $\mathcal{Q}_p$ (i.e. signals from arbitrary numbers of CG or TG detection oligos with different sequences all querying the same CpG position).

Methylation likelihood

In Section 2.2.1 we described how the noise parameters of Eq. 2.44 can be estimated from repeated measurements of detection and negative control oligos. The background intensities and equilibrium constants have to be estimated from specific calibration experiments. The easiest calibration experiment is to measure 100% methylated DNA (e.g. SSS1 treated blood DNA) and 0% methylated DNA (e.g. Phi29 amplified DNA). From these calibration measurements we get direct estimates for the expected hybridization intensities of 0% methylation ($I_q^{0,CG}$, $I_q^{0,TG}$) and 100% methylation ($I_q^{1,CG}$, $I_q^{1,TG}$). The equilibrium constants can be expressed with these intensity estimates as

$$\begin{aligned} \tilde{k}_q^{CG} &= I_q^{1,CG} - I_q^{0,CG} \\ \tilde{k}_q^{TG} &= I_q^{1,TG} - I_q^{0,TG}. \end{aligned} \tag{2.45}$$

Note that since CG oligos are designed to bind to methylated CpG dinucleotides we can expect $I_q^{1,CG} > I_q^{0,CG}$, resulting in $\tilde{k}_q^{CG} > 0$. For TG oligos that are designed to bind to unmethylated CpG dinucleotides we expect $I_q^{1,TG} < I_q^{0,TG}$, resulting in $\tilde{k}_q^{TG} < 0$. These expectations can be used as criteria to exclude not properly working oligos.

Using the fact that the computation of equilibrium constants in Eq. 2.45 is identical for CG and TG oligos we can rewrite the expected hybridization intensity of Eq. 2.37 as

$$I_q = I_q^0 + m_{p(q)}(I_q^1 - I_q^0), \tag{2.46}$$

where I_q^0 and I_q^1 are the expected hybridization intensities of oligo q on unmethylated and methylated DNA respectively.

The standard procedure for measuring I_q^0 and I_q^1 currently implemented at Epigenomics is to hybridize 5 microarrays with artificially downmethylated DNA and 5 microarrays with artificially upmethylated DNA within each experimental study. With the usual 4fold oligo redundancy per chip this means that after between array normalization I_p^0 and I_p^1 can be estimated from 20 independent observations (e.g. by simply taking the median intensity and substracting the I^{BG} estimate).

Using the generalized log transformation from Eq. 2.21 the variance stabilized intensity distribution of oligo q can then be approximated with a normal distribution as

$$\begin{aligned} T[O_q] &= N\left(T[I_q + I^{BG}], S_\eta\right) \\ &= N\left(\ln\left(I_q + \sqrt{I_q^2 + c}\right), S_\eta\right) \\ &= N\big(\ln\big(I_q^0 + m_{p(q)}(I_q^1 - I_q^0) \\ &\quad + \sqrt{(I_q^0 + m_{p(q)}(I_q^1 - I_q^0))^2 + c}\big), S_\eta\big). \end{aligned} \tag{2.47}$$

Note that constant variance does not necessarily imply normally distributed. However, as we have shown in Fig. 2.20 the transformed microarray data is indeed close to normally distributed. This approximation simplifies the following computations considerably.

Accordingly the error between variance stabelized model prediction at oligo q and an observed hybridization intensity $O_{q,i}$ is given as

$$\begin{aligned} E_q^i(m_{p(q)}) &= T[O_{q,i}] - T[I_q + I^{BG}] \\ &= T[O_{q,i}] - \ln\left(I_q + \sqrt{I_q^2 + c}\right) \\ &= T[O_{q,i}] - \ln\big(I_q^0 + m_{p(q)}(I_q^1 - I_q^0) \\ &\quad + \sqrt{(I_q^0 + m_{p(q)}(I_q^1 - I_q^0))^2 + c}\big). \end{aligned} \tag{2.48}$$

The likelihood of observing a set of N hybridization intensities for oligo q is

$$L_q(m_{p(q)}) = \prod_{i=1}^{n_r} \frac{1}{\sqrt{2\pi}S_\eta} e^{-\frac{E_q^i(m_{p(q)})^2}{2S_\eta^2}}. \tag{2.49}$$

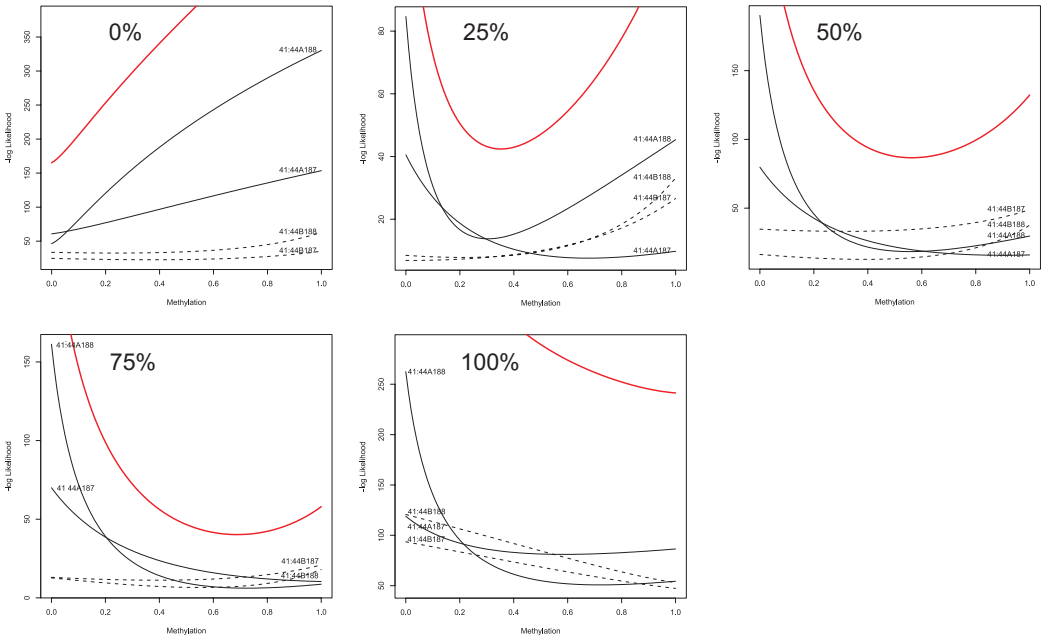

Figure 2.24: Likelihood functions for an oligo family of the ESR1 gene from the *Calibration* dataset (see Appendix A). Individual CG (black solid line) and TG (black dotted line) oligo likelihood functions and the resulting oligo family likelihood function (red solid line) are shown for hybridization experiments with 5 different methylation rates of the ESR1 gene.

For an oligo family $\mathcal{Q}_p$ covering the same CpG position p the total family likelihood $L_{\mathcal{Q}}$ is the product of the individual oligo likelihoods:

$$L_{\mathcal{Q}}(m_p) = \prod_{q \in \mathcal{Q}_p} L_q(m_{p(q)}). \tag{2.50}$$

Note that the oligo family $\mathcal{Q}_p$ can contain arbitrary numbers of CG and TG oligos. This natural aggregation of information from different oligos is one of the major advantages of the maximum likelihood approach.

Fig. 2.24 shows an example of individual CG and TG oligo and resulting total family likelihood functions on one CpG position of the ESR1 gene at 5 different methylation states. One can see that CG oligos contribute the most information at low methylation rates, whereas TG oligos contribute the most information at high methylation rates. Generally the information contribution of TG oligos is small compared to CG oligos due to their lower melting temperature gap between match and mismatch duplexes and the resulting higher affinity for unspecific hybridizaton (see also Fig. 2.21).

Minimization

Eq. 2.49 gives the likelihood of observing a set of hybridization intensities $\{O_{q,i}\}_{i=1\ldots n_r}$ given a certain methylation rate $m_{p(q)}$ at CpG position p. Following the maximum likelihood framework [14] we can estimate the most likely methylation rate m_p given the observed intensity data $\{O_{q,i}\}_{i=1\ldots n_r}$ by minimizing the negative log likelihood

$$\begin{aligned} l_q(m_{p(q)}) &= -\ln L_q(m_{p(q)}) \\ &= \sum_{i=1}^{n_r} \Big(-\ln \frac{1}{\sqrt{2\pi} S_\eta} + \frac{E_q^i(m_{p(q)})^2}{2S_\eta^2} \Big) \\ &\propto \sum_{i=1}^{n_r} E_q^i(m_{p(q)})^2. \end{aligned} \tag{2.51}$$

The corresponding negative log likelihood of a complete oligo family $\mathcal{Q}_p$ is

$$\begin{aligned} l_{\mathcal{Q}}(m_p) &= -\ln L_{\mathcal{Q}}(m_p) \\ &= \sum_{q \in \mathcal{Q}_p} l_q(m_p) \\ &\propto \sum_{q \in \mathcal{Q}_p} \sum_{i=1}^{n_r} E_q^i(m_p)^2, \end{aligned} \tag{2.52}$$

with all oligos $q \in \mathcal{Q}_p$ querying the same CpG position p. The numerical value of m_p that minimizes Eq. 2.52 is the most likely methylation responsible for our observations. Therefore the maximum likelihood score is defined as

$$\begin{aligned} S_p(\{O_{q,i}\}_{q\in\mathcal{Q}_p,i=1\ldots n_r}) &= min_{m_p} l_{\mathcal{Q}}(m_p) \\ &= min_{m_p} \sum_{q\in\mathcal{Q}_p} \sum_{i=1}^{n_r} E_q^i(m_p)^2. \end{aligned} \tag{2.53}$$

Note that m_p is per definition constrained to be within $[0, 1]$.

There exists a variaty of numerical methods to perform the actual minimization [130, 14]. Results in this thesis were all generated by using the R optimize function (a combination of golden section search and successive parabolic interpolation) [20].

Confidence intervals

In addition to the maximum likelihood estimate for the methylation rate m_p we can also derive approximate confidence intervalls for this estimate from the observed Fisher information [38]. The observed Fisher information J is given by the second derivative of the negative log likelihood from Eq. 2.52 as

$$\begin{aligned} J(m_p) &= \frac{\partial^2}{\partial m_p^2} l_{\mathcal{Q}}(m_p) \\ &= \frac{1}{2S_\eta^2} \sum_{q\in\mathcal{Q}_p} \sum_{i=1}^{n_r} \frac{\partial^2}{\partial m_p^2} E_q^i(m_p)^2 \\ &= \frac{1}{S_\eta^2} \sum_{q\in\mathcal{Q}_p} \frac{(I_q^1 - I_q^0)^2}{(I_q^2 + c)^{3/2}} \left(n_r \sqrt{I_q^2 + c} + I_q \sum_{i=1}^{n_r} \left(T[O_{q,i}] - \ln(I_q + \sqrt{I_q^2 + c}) \right) \right), \end{aligned} \tag{2.54}$$

with $I_q = I_q^0 + m_p(I_q^1 - I_q^0)$. The confidence intervals for the maximum likelihood estimate of S_p of the methylation rate m_p can then be given as

$$S_p \pm \frac{k}{\sqrt{J(S_p)}}, \tag{2.55}$$

where k is the appropriate z critical value of the normal distribution (e.g. 1.96 for 95% confidence). Note that despite sometimes larger numerical estimates from the Fisher information approximation the confidence intervalls are constrained to $[0, 1]$.

2.4 Results

Here we will compare the four derived methylation scores (log difference, glog difference, proportion and maximum likelihood score) with each other. For this comparison we use the data from the *Calibration* dataset (see Appendix A) which provides an extensive amount of replicated measurements for several DNAs with known methylation state (0%, 25%, 50%, 75%, 100%).

In the following we will investigate a variety of performance criteria for our methylation scores. These criteria follow the methylation score properties defined in Section 2.3 and are based on the calibration curve - the relation between true methylation values and reported methylation score values. Of course every CpG position and its corresponding oligo family has its own calibration curve. An example for a CpG position in the ESR1 gene is shown in Fig. 2.25.

In order to make a general statement about the performance of the different methylation scores we will now look at the distribution of the major criteria over all oligo families of the *Calibration* chip (see Appendix A) and over all of our three raw intensity measures (MI, EMI, EMV). Distributions of calibration curves are shown in Fig. 2.26. As expected only the log difference and glog difference scores have approximately constant variance over all methylation levels. The methylation proportion and ML score have a much smaller variance at 0% methylation than at higher methylation levels. On the other hand while all four scores are monotonically increasing only the methylation proportion and ML scores report values close to the true methylation rates. The two difference scores are not designed to report direct methylation rates and an analysis of accuracy or bias is therefore pointless.

For practical applications as disease classification or marker identification the most important property of a methylation score is discrimination - its ability to discriminate different methylation levels from each other. Fig. 2.27 shows the distribution of AUC (area under the ROC curve) values for all adjacent methylation levels of all scores and raw intensity measures. For all four methylation scores the EMI and EMV raw intensity measures give better discrimination between 0% and 25% then the simple MI measure. The discrimination ability of the four different scores is similar with exception of the ML score that seems to have a slight disadvantage at the 25%-50% discrimination.

Only the proportion and the ML score are designed to report direct methylation proportions. Fig. 2.28 and Fig. 2.29 show bias and precision of these scores. It is obvious that the simple proportion score has a very strong tendency to underestimate methylation values due to the unspecific binding characteristic of many TG oligos. The ML score on the other hand

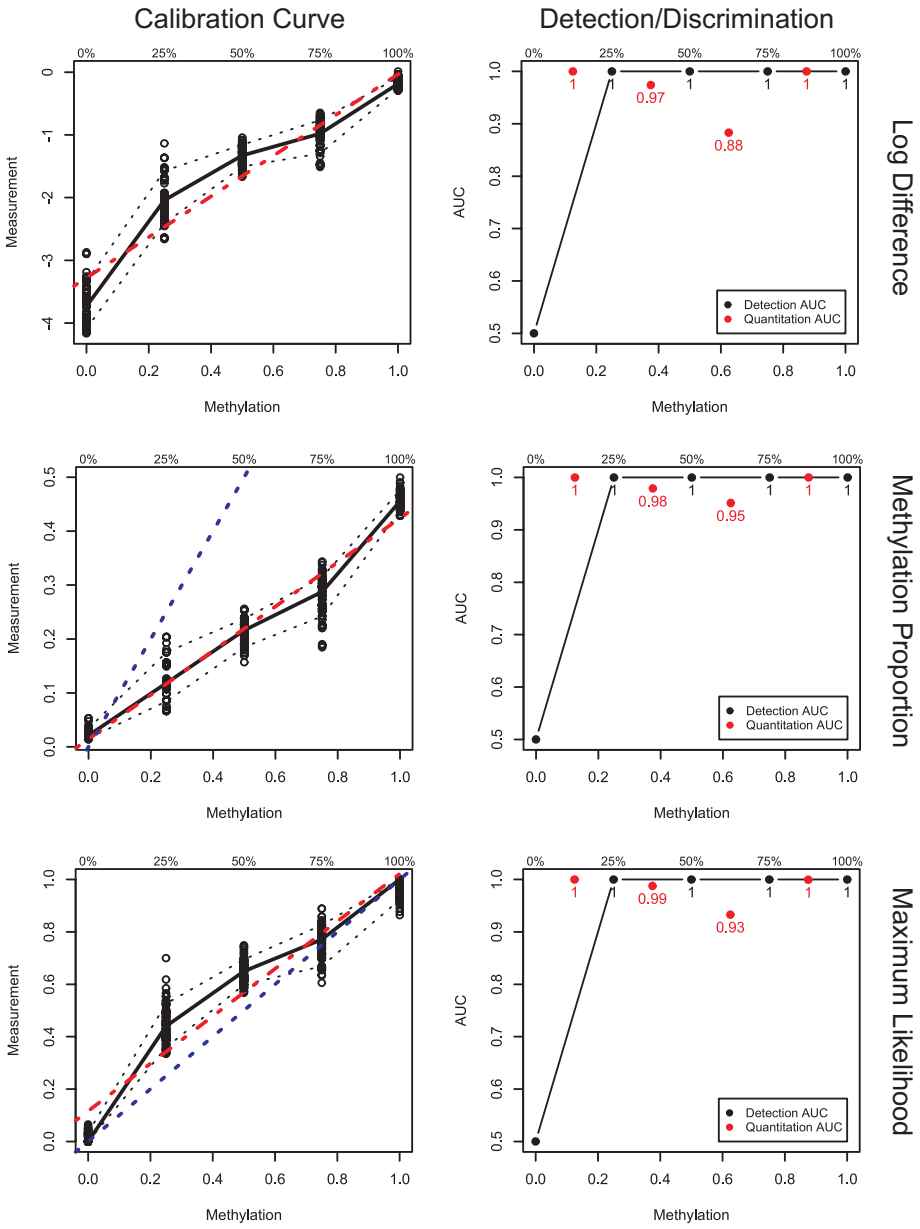

Figure 2.25: Calibration and detection curves for an oligo family of the ESR1 gene. Every row shows results for the log difference, the proportion and the maximum likelihood methylation score. Results for the glog difference score were identical to the log difference score. The plots to the left show the calibration curves. The x-axis is the true methylation proportion of the ESR1 gene. The y-axis is the reported value of the respective methylation score. Individual points correspond to measurement values from individual microarrays, solid black line is the median methylation score, dotted black lines are the 10% and 90% quantiles of the reported methylation scores. The dashed red line is the linear fit of all reported scores. The dashed blue line is the diagonal, i.e. the ideal calibration curve of a unbiased methylation score. The plots to the right show area under the ROC curve (AUC) values for the discrimination between 0% methylation and all other methylation levels (Detection AUC, black curve) and adjacent methylation levels (Quantitation AUC, red points).

	Log Difference	GLog Difference	Methylation Proportion	Maximum Likelihood
Monotonicity	++	++	++	++
Linearity	-	-	+	++
Discrimination	++	++	++	+
Bias	- -	- -	-	++
Precision			+	+
Variance Stability	+	++	- -	- -

Table 2.1: Comparison of methylation scores. Scores are rated as good (++), sufficient (+), insufficient (-) or bad (- -) in every category.

takes the oligo specific properties into account and is on average almost unbiased. On the precision side the ML score shows smaller variations at the extreme methylation levels of 0% and 100% but the proportion score seems to be less variable at te methylation range between 25% and 75%. However, it has to be noted that the dynamic range of the proportion score considerable smaller at higher methylation levels due to its strong bias. When this is taken into account both scores have about equal precision at the 75% methylation level.

Table 2.4 gives an overview how well our four methylation scores behave with regard to the different performance criteria. With regard to the different raw intensity measures all scores show better discrimination characteristics with background correction. The EMI measure seems to give the best detection and discrimination rates with all scores.

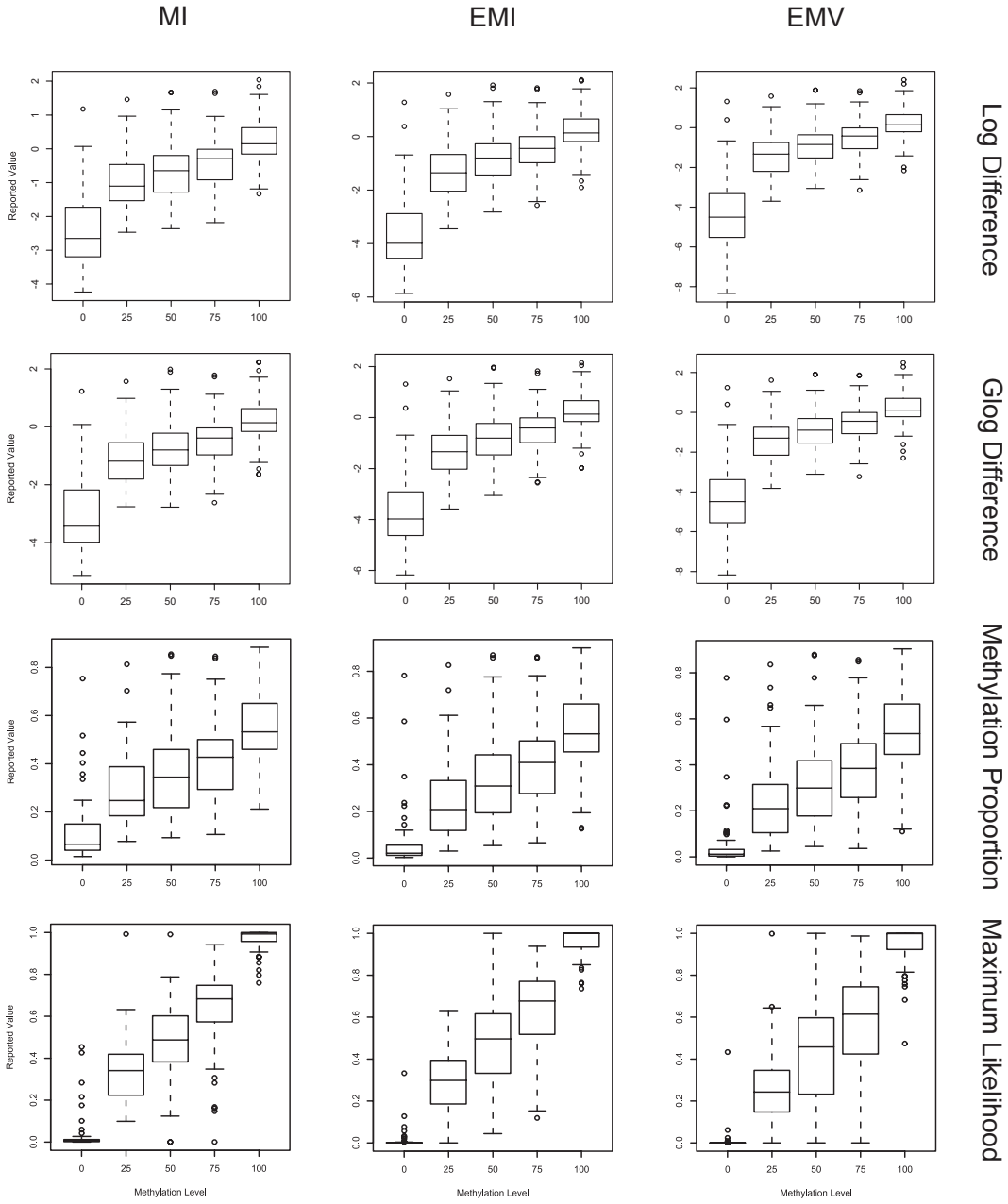

Figure 2.26: Calibration curve distributions. Distributions are shown for every combination of methylation scores (rows: log difference score, glog difference score, methylation proportion score, maximum likelihood score) and raw intensity measures (columns: MI, EMI, and EMV). Each plot shows the distribution of methylation scores from all oligo families of the *Calibration* chip (see Appendix A). The x-axis corresponds to the methylation level of the measured control DNA. The y-axis shows the reported values of the respective methylation score.

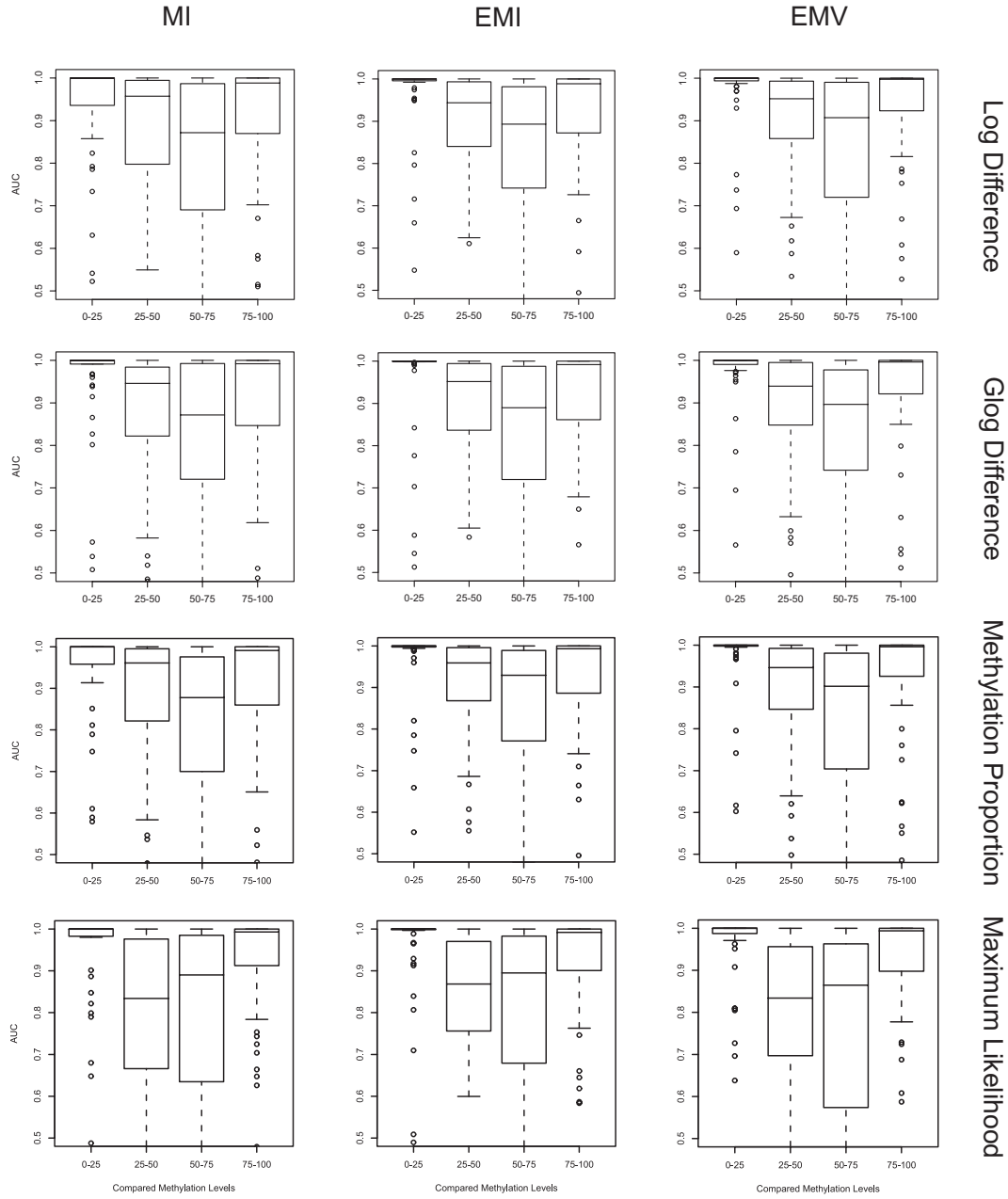

Figure 2.27: Discrimination distributions. Distributions are shown for every combination of methylation scores (rows: log difference score, glog difference score, methylation proportion score, maximum likelihood score) and raw intensity measures (columns: MI, EMI, and EMV). Each plot shows the distribution of AUC values between two adjacent methylation levels from all oligo families of the *Calibration* chip (see Appendix A). The x-axis shows the different pairs of methylation levels of the measured control DNA that are compared. The y-axis shows the distribution of AUC values for the respective comparison.

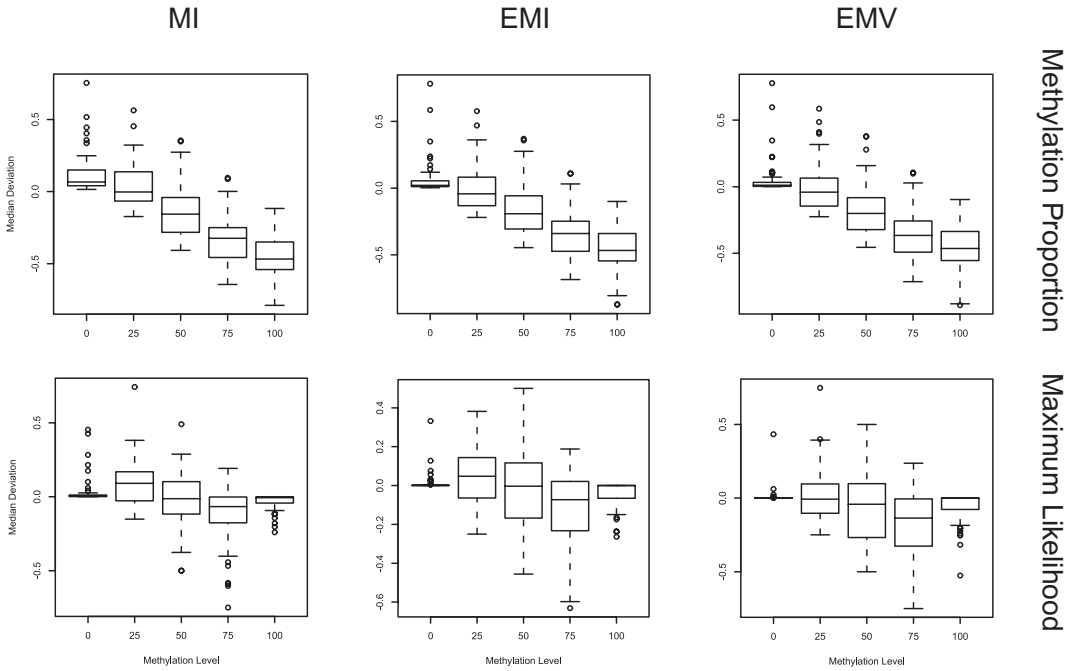

Figure 2.28: Bias distributions. Distributions are shown for every combination of quantitative methylation scores (rows: methylation proportion score, maximum likelihood score) and raw intensity measures (columns: MI, EMI, and EMV). Each plot shows the distribution of bias (median deviation of reported score values from true methylation values) of the repective methylation scores from all oligo families of the *Calibration* chip (see Appendix A). The x-axis corresponds to the methylation level of the measured control DNA. The y-axis shows the bias of the reported values of the respective methylation score.

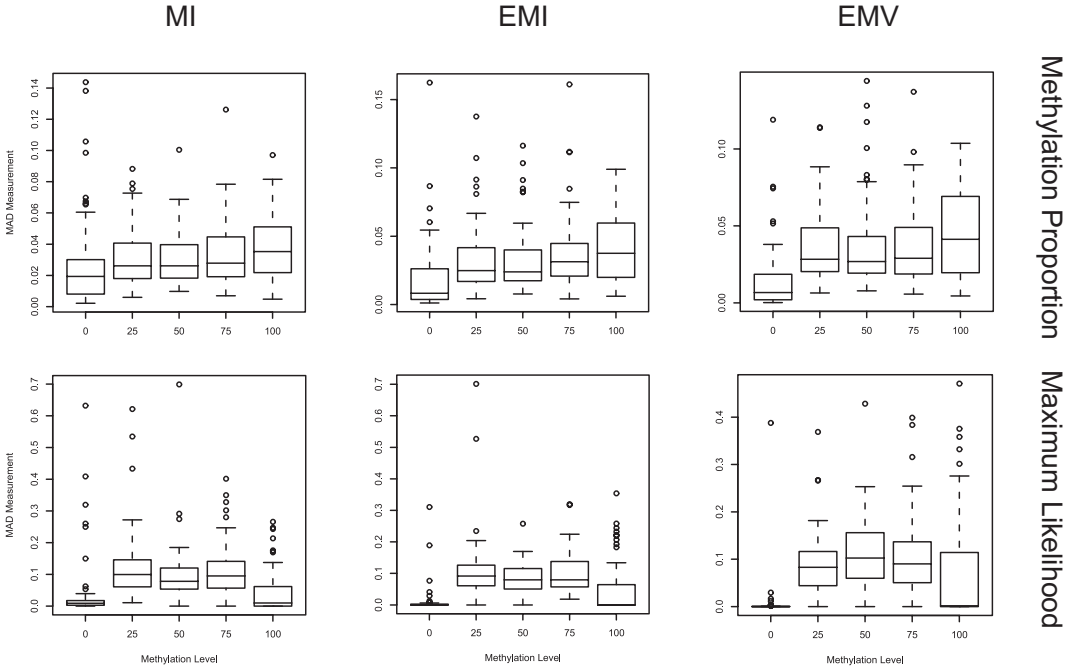

Figure 2.29: Precision distributions. Distributions are shown for every combination of quantitative methylation scores (rows: methylation proportion score, maximum likelihood score) and raw intensity measures (columns: MI, EMI, and EMV). Each plot shows the distribution of presicion (median absolute deviation from median of reported score values) of the repective methylation scores from all oligo families of the *Calibration* chip (see Appendix A). The x-axis corresponds to the methylation level of the measured control DNA. The y-axis shows the precision of the reported values of the respective methylation score.

Chapter 3

Controlling quality and stability of microarray experiments

Microarray production is rapidly evolving towards a high throughput industry. Therefore it seems natural to apply the principles of multivariate statistical process control (MVSPC) to statistical quality control of microarray experiments. However, most of the relevant process parameters of a microarray experiment cannot be measured routinely in a high throughput environment. As an alternative, we propose to use the measurement values of the microarrays themselves to control the stability of the production process. However, these measurements are extremely high dimensional and contain outliers, prohibiting the application of standard MVSPC methods. We show that it is nevertheless possible to apply MVSPC techniques, when using robust PCA [81] to remove outliers and reduce data dimensionality.

Furthermore, we introduce novel methods that provide additional information about the nature of a process error (e.g. probe permutation vs. change in probe concentration). We demonstrate on three large DNA methylation microarray datasets that this technique is a powerful tool to detect process errors in microarray experiments.

The rest of the chapter is structured as follows. In the first section we give a short introduction to the process that generated our microarray data and point out typical sources of artefacts. In the second section we demonstrate how robust PCA can be used to detect abnormal hybridizations. This is an essential prerequisite for the application of statistical process control to microarray data. Finally MVSPC is introduced in the third section and we develop a method to check whether all essential conditions stay constant over the course of an experimental series.

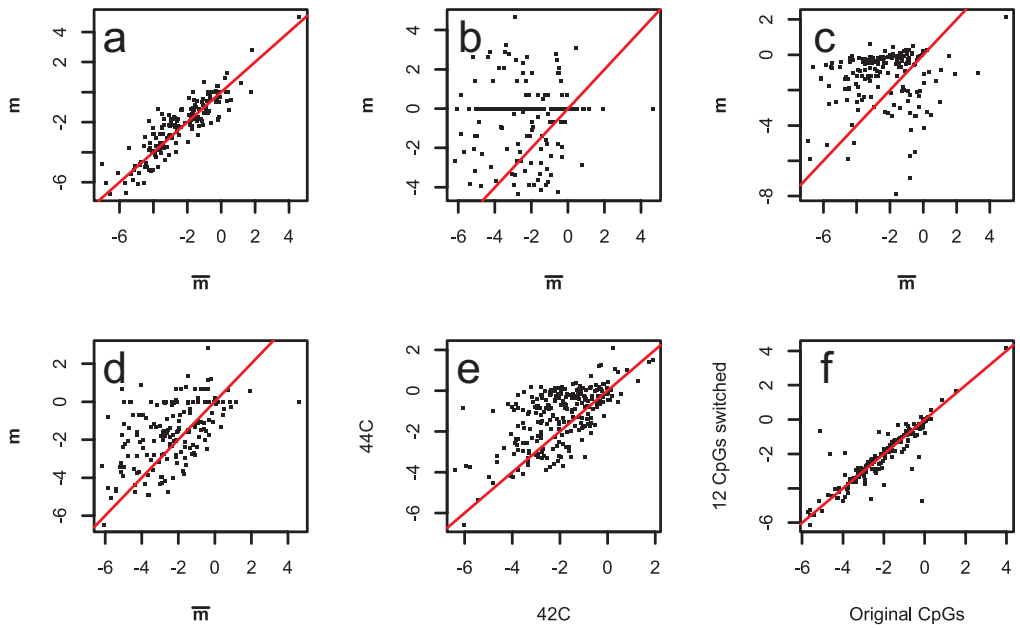

Figure 3.1: Typical artefacts in microarray based methylation analysis. The plots show the correlation between single or averaged methylation profiles. Every point corresponds to a single CpG position, the axis-values are log ratios. **a)** A normal chip, showing good correlation to the sample average. **b)** A chip classified as "unacceptable" by visual inspection. Many spots showed no signal, resulting in a log ratio of 0 after thresholding the signals to $\epsilon > 0$. **c)** A chip classified as "good". Hybridization conditions were not stringent enough, resulting in saturation. In many cases pairs of CG and TG oligos showed nearly identical high signals, giving a log ratio around 0. **d)** A chip classified as "acceptable". Hybridization signals were weak compared to the background intensity, resulting in a high amount of noise. **e)** Comparison of group averages over all 64 ALL/AML chips hybridized at 42°C and all 48 ALL/AML chips hybridized at 44°C. **f)** Comparison of group averages over 447 regular chips from the lymphoma dataset and the 200 chips with a simulated accidental probe exchange during slide production, affecting 12 CpG positions.

3.1 Microarray data and typical sources of error

In the following, n_q is the number of CG-TG oligo pairs per slide, n_s is the number of biological samples in the study and n_c is the number of hybridized chips in the study. For a specific oligo pair $q \in \{1, \ldots, n_q\}$, the frequency of methylated alleles in sample $s \in \{1, \ldots, n_s\}$, hybridized onto chip $c \in \{1, \ldots, n_c\}$ can then be quantified by one of the methylation scores from Section 2.3 as $d_{cq} = S_{p(q)}\left(\{O^{CG}_{q,i,c}, O^{TG}_{q,i,c}\}_{i=1\ldots n_r}\right)$. Since we will often assume normality of our data in the following sections the methylation score should have stable variance. Througout this chapter we will use the log difference score. We refer to a single hybridization experiment c as experiment or chip. The resulting set of measurement values is the methylation profile $\boldsymbol{d}_c = (d_{c1}, ..., d_{cn_q})'$.

In order to illustrate typical error sources we use the Lymphoma dataset (see Appendix A) with its more than 9 repeated hybridization experiments c for every single biological sample s. With this high number of replications for each biological sample the corresponding average methylation profile $\bar{\boldsymbol{d}}_s$ can be reliably estimated. Here we use the L_1-median

$$\bar{\boldsymbol{d}}_s = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{c \in \mathcal{C}_s} \|\boldsymbol{d}_c - \boldsymbol{x}\|_2 \tag{3.1}$$

to compute a robust estimate for the methylation profile of biological sample s from its set of repetitions $\mathcal{C}_s$. Outlier chips can then be relatively easily detected by their strong deviation from the sample methylation profile $\bar{\boldsymbol{d}}_s$.

Fig. 3.1a shows a typical chip classified as "good" by visual inspection. The small random deviations from the sample median are due to the approximately normally distributed experimental noise. A typical chip classified as "unacceptable" by visual inspection is shown in Fig. 3.1b and can be easily identified by the fact that many of the oligo pairs gave no signal which results in a log ratio of zero. The opposite case is shown in Fig. 3.1c. This chip has very strong hybridization signals and was classified as "good" by visual inspection. However, obviously the hybridization conditions have been too unspecific and most of the oligos were saturated. Fig. 3.1d shows a chip classified as "acceptable". Many of these chips give good measurements, however some of them have such weak correlation with the true methylation profile that they should be regarded as outliers.

Other potential error sources such as changing concentrations or handling errors during slide production will influence whole chip batches. Variations in hybridization buffer or salt concentration will systematically affect the

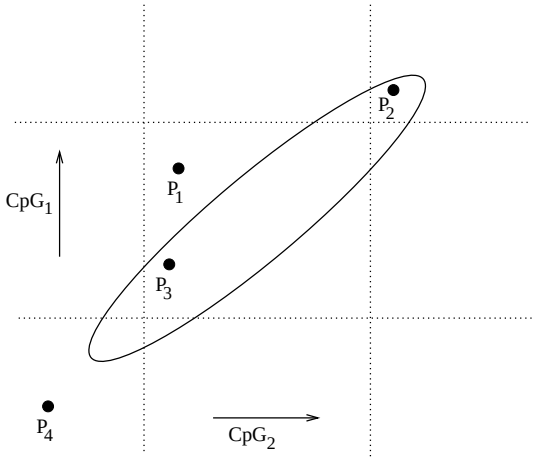

Figure 3.2: Comparison between univariate (central rectangle) and multivariate (ellipse) upper confidence intervals. P_1 is not detected as outlier by univariate t_k-distance, but by multivariate T^2-statistic. P_2 is erroneously detected as outlier by the univariate t_k-distance, but not by the multivariate T^2-statistic. For P_3 (non-outlier) and P_4 (outlier) both methods give the same decisions.

melting temperature of the spotted oligos. Fig. 3.1e shows this systematic effect by comparing hybridizations at two different temperatures. Finally, Fig. 3.1f shows the simulation of an accidental probe exchange during slide production, affecting 12 CpG positions.

After identifying possible error sources the question remains how to reliably detect them, if they cannot be avoided with absolute certainty. Our objective is to exclude single outlier chips from the analysis and to detect systematic changes in experimental conditions as early as possible in order to facilitate a fast recalibration of the production process.

In the following we will introduce a method to detect systematic errors which does not rely on repeated hybridization experiments and makes no explicit assumptions about error sources. This will be achieved in three major steps. First outliers are removed by robust PCA. Then classical PCA is used for dimension reduction. Finally methods from MVSPC are applied to detect changes in experimental conditions.

3.2 Detecting outlier chips with robust PCA

3.2.1 Methods

As a first step we aim to detect single outlier chips. In contrast to statistical approaches based on image features of single slides [22] we will use the overall distribution of the whole experimental series. This is motivated by the fact that although image analysis algorithms will successfully detect bad hybridization signals, they will usually fail in cases of unspecific hybridization. The idea is to identify the region in measurement space where most of the chips $\boldsymbol{d}_c, c = 1 \ldots n_c$, are located. The region will be defined by its center

and an upper limit for the distance between a single chip and the region center. Chips with deviations higher than the upper limit will be regarded as outliers.

A simple approach would be to separately define for every dimension (in our case oligo pair) q the deviation of a chip c from the centerμ_q as

$$t_q(c) = \frac{|d_{cq} - \mu_q|}{s_q}, \tag{3.2}$$

where $\mu_q = (1/n_c)\sum_c d_{cq}$ is the mean and $s_q^2 = 1/(n_c - 1)\sum_c (d_{cq} - \mu_q)^2$ is the sample variance overall chips. Assuming that the d_{cq} are normally distributed, t_q multiplied by a constant follows a t-distribution with $n_c - 1$ degrees of freedom. This can be used to define the upper limit of the admissible region for a given significance level α [109].

However, a separate treatment of the different dimensions is only optimal when they are statistically independent. As Fig. 3.2 demonstrates it is important to take into account the correlation between different dimensions. It is possible that a point which is not detected as an outlier by a component wise test is in reality an outlier (e.g. P_1 in Fig. 3.2). On the other hand, there are points that will be erroneously detected as outliers by a component wise test (e.g. P_2 in Fig. 3.2). Because microarray data have usually a very high correlation, it is better to use a multivariate distance concept instead of the simple univariate t_q-distance. A natural generalization of the t_q-distance is given by Hotelling's T^2 statistic, defined as

$$T^2(c) = (\boldsymbol{d}_c - \boldsymbol{\mu})' S^{-1} (\boldsymbol{d}_c - \boldsymbol{\mu}) \tag{3.3}$$

with mean $\boldsymbol{\mu} = (1/n_c)\sum_{c=1}^{n_c} \boldsymbol{d}_c$ and sample covariance matrix $S = 1/(n_c - 1)\sum_{c=1}^{n_c} (\boldsymbol{d}_c - \boldsymbol{\mu})(\boldsymbol{d}_c - \boldsymbol{\mu})'$. Assuming that the $\boldsymbol{d}_c$ are multivariate normally distributed, T^2 multiplied by a constant follows a F-distribution with n_q degrees of freedom for the numerator and $n_c - n_q$ for the denominator. This can be used to define the upper limit of the admissible region for a given significance level α [109].

Two problems arise when we want to use the T^2-distance for microarray data:

1. For less chips n_c than dimensions n_q, the sample covariance matrix S is singular and not invertible.

2. The estimates for $\boldsymbol{\mu}$ and S are not robust against outliers [107].

The first problem can be addressed by using principle component analysis (PCA) to reduce the dimensionality of the measurement space [109]. This is

done by projecting all measurement profiles $\boldsymbol{d}_c$ onto the first k eigenvectors with the highest variance. As a result we get the k-dimensional centered vectors $\tilde{\boldsymbol{d}}_c = P_{PCA}(\boldsymbol{d}_c - \boldsymbol{\mu})$ in eigenvector space. After the projection, the covariance matrix $\tilde{S} = diag(\tilde{s}_1, \ldots, \tilde{s}_d)$ of the reduced space is a diagonal matrix and the T^2-distance of Eq. 3.3 is approximated by the T^2-distance in the reduced space

$$\tilde{T}^2(c) = \sum_{r=1}^{k} \frac{\tilde{d}_{cr}^2}{\tilde{s}_r^2}. \tag{3.4}$$

Under the assumption that the true variances are equal to the observed variances $\tilde{s}_r$, $\tilde{T}^2$ follows a χ^2 distribution with k degrees of freedom. This can be used to define the upper limit of the admissible region for a given significance level α. However, the problem remains that the estimated eigenvectors and variances $\tilde{s}_r$ are not robust against outliers.

We propose to solve the problem of outlier sensitivity by using robust principle component analysis (rPCA) [81]. rPCA finds the first k directions with the largest scale in data space, robustly approximating the first k eigenvectors. The algorithm starts with centering the data with a robust location estimator. Here we use the L_1 median

$$\boldsymbol{\mu}_{L1} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{c=1}^{n_c} \|\boldsymbol{d}_c - \boldsymbol{x}\|_2. \tag{3.5}$$

In contrast to the simple component wise median this gives a robust estimate of the distribution center that is invariant to orthogonal linear transformations such as PCA [107].

Then all centered observations are projected onto a finite subset of all possible directions in measurement space. The direction with maximum robust scale is chosen as an approximation of the largest eigenvector (e.g. by using the Q_n estimator [34]). After projecting the data into the orthogonal subspace of the selected "eigenvector" the procedure searches for an approximation of the next eigenvector. Following Hubert et al. we have simply chosen the finite set of possible directions as the set of centered observations themselves. Note that in our experience the concrete choice of robust estimators for location and scale has no crucial impact on the results.

After obtaining a robust projection of the data into a k-dimensional subspace we can compute the outlier insensitive $\tilde{T}^2$-distance and its respective upper limit of the admissible region $\tilde{T}^2_{UCL}$, also referred to as the upper control limit (UCL). For a given significance level α it is computed as

$$\tilde{T}^2_{UCL} = \chi^2_{k,1-\alpha}. \tag{3.6}$$

Every observation $\boldsymbol{d}_c$ with $\tilde{T}^2(c) > \tilde{T}^2_{UCL}$ is regarded as an outlier.

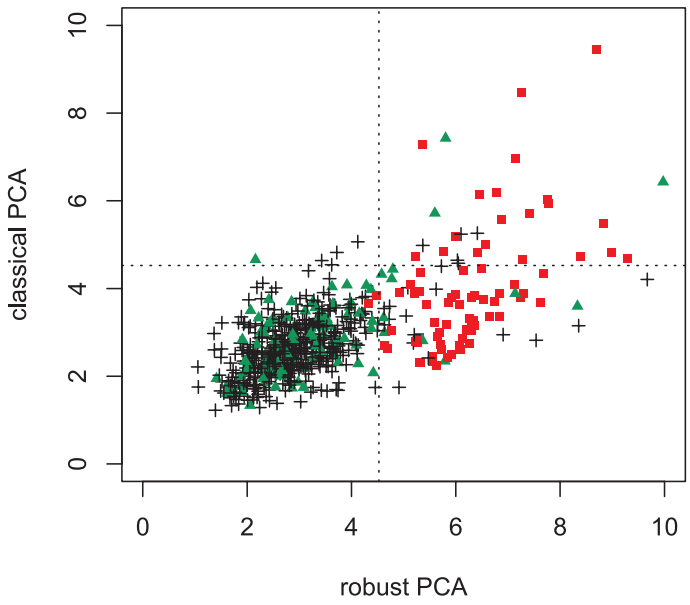

Figure 3.3: $\tilde{T}^2$-Distances of robust PCA versus classical PCA for the Lymphoma dataset. The $\tilde{T}^2_{UCL}$-values are shown as two dotted lines. Chips, right of the vertical line are detected as outlier by robust PCA. Chips above the horizontal line are detected as outlier by classical PCA. Chips classified as "unacceptable" by visual inspection are shown as squares, "acceptable" chips as triangles and "good" chips as crosses. Note that the "good" chips detected as outliers by rPCA have all been confirmed to show saturated hybridization signals. The $\tilde{T}^2_{UCL}$-values are calculated with $k = 10$ and significance level $\alpha = 0.025$.

3.2.2 Results

We tested the rPCA algorithm by comparing its performance to classical PCA on the Lymphoma dataset. The results are shown in Fig. 3.3.

The rPCA algorithm detected 97% of the chips with "unacceptable" quality, whereas classical PCA only detected 29%. 10% of the "acceptable" chips were detected as outliers by rPCA, whereas PCA detected 3%. rPCA detected 21 chips as outliers which were classified as "good". These "good" chips have all been confirmed to show saturated hybridization signals, not identified by visual inspection. This means that rPCA is able to detect nearly all cases of outlier chips identified by visual inspection. Additionally rPCA detects microarrays which have unconspicous image quality but show an unusual hybridization pattern.

An obvious concern with this use of rPCA for outlier detection is that it relies on the assumption of normal distribution of the data. If the distribution of the biological data is highly multi-modal, biological subclasses may be wrongly classified as outliers. To quantify this effect we simulated a very strong cluster structure in the Lymphoma data by shifting one of the smaller subclasses by a multiple of the standard deviation. Only when the measurements of all 174 CpG of the subclass where shifted by more than 2 standard deviations a considerable part of the biological samples were wrongly classified as outliers.

This situation can only be reliably detected when there are repeated hybridization experiments for every sample. In this case the fraction of outlier chips per sample can be computed. A high fraction would indicate a biological cause. We used a threshold of 50% outlier chips per sample to detect outliers resulting from biological effects. However, we never encountered such a situation in our datasets.

3.3 Statistical process control

3.3.1 Methods

In the last section we have seen how outliers can be detected solely on the basis of the overall data distribution. Statistical process control expands this approach by introducing the concept of time. The idea is to observe the variables of a process for some time under perfect working conditions. The data collected during this period form the so called historical dataset (HDS). Under the assumption that all variables are normally distributed, the mean $\boldsymbol{\mu}_{HDS}$ and the sample covariance matrix S_{HDS} of the historical dataset fully describe the statistical behavior of the process.

Given the historical dataset it becomes possible to check at any time point i how far the current state of the process has deviated from the perfect state by computing the T^2-distance between the ideal process mean $\boldsymbol{\mu}_{HDS}$ and the current observation $\boldsymbol{d}_i$. This corresponds to Eq. 3.3 with the overall sample estimates $\boldsymbol{\mu}$ and S replaced by their reference counterparts $\boldsymbol{\mu}_{HDS}$ and S_{HDS}. Any change in the process will cause observations with greater T^2-distances. To decide whether an observation shows a significant deviation from the HDS we compute the upper control limit as

$$T^2_{UCL} = \frac{p(n+1)(n-1)}{n(n-p)} F_{p,n-p,1-\alpha}, \tag{3.7}$$

where p is the number of observed variables, n is the number of observations in the HDS, α is the significance level and F is the F-distribution with p degrees of freedom for the numerator and $n-p$ for the denominator. Whenever $T^2 > T^2_{UCL}$ is observed the process has to be regarded as significantly out of control [112].

In our case the process to control is a microarry experiment and the only process variables we have are the log ratios of the actual hybridization intensities. A single observation is then a chip $\boldsymbol{d}_i$ and the HDS of size N_{HDS} is defined as $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_{N_{HDS}}\}$. We have to be aware of a few important issues in this interpretation of statistical process control. Firstly, our data has a multi-modal distribution which results from a mixture of different biological samples and classes. Therefore the assumption of normality is only a rough approximation and T^2_{UCL} from Eq. 3.7 should be regarded with caution. Secondly, as we have seen in the last sections, microarray experiments produce outliers, resulting in transgression of the UCL. This means that sporadic violations of the UCL are normal and do not indicate that the process is out of control. The third issue is that we have to use the assumption that a microarray study will not systematically change its data generating distribution over time. Therefore the experimental design has to be randomized or block randomized, otherwise a systematic change in the true biological data would be interpreted as an out of control situation (e.g. when all patients with the same disease subtype are measured in one block). Finally, the question remains what time means in the context of a microarray experiment. Beside the biological variation in the data, there are a multitude of different parameters which can systematically alter the final hybridization intensities. The experimental series should stay constant with regard to all of them. In our experience the best initial choice is to order the chips by their date of hybridization, which shows a very high correlation to most process parameters of interest.

1. Order chips according to the parameter of interest, e.g. date of hybridization.
2. Take the set of ordered chips $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_{n_c}\}$, remove outliers with rPCA for computing the first k eigenvectors with classical PCA
3. Project the set of all ordered chips $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_{n_c}\}$ into the k-dimensional subspace spanned by the computed eigenvectors
4. Select the first N_{HDS} chips $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_{N_{HDS}}\}$ as historical dataset, remove outliers with rPCA for computing $\boldsymbol{\mu}_{HDS}$ and S_{HDS}.
5. For every time index $i \in \{1, ..., n_c\}$
   (a) Compute T^2-distance between $\boldsymbol{d}_i$ and $\boldsymbol{\mu}_{HDS}$.
   (b) If $\frac{N_{CDS}}{2} < i < n_c - \frac{N_{CDS}}{2}$
      i. Select $\{\boldsymbol{d}_{i-N_{CDS}/2}, ..., \boldsymbol{d}_i, ..., \boldsymbol{d}_{i+N_{CDS}/2}\}$ as current dataset, remove outliers with rPCA for computing $\boldsymbol{\mu}_{CDS}$ and S_{CDS}.
      ii. Compute T_w^2-distance between $\boldsymbol{\mu}_{HDS}$ and $\boldsymbol{\mu}_{CDS}$.
      iii. Compute L-distance between S_{HDS} and S_{CDS}.
6. Generate T^2 control chart by plotting T^2, T_w^2 and L.

Figure 3.4: Algorithm for generating a T^2 control chart. Major parameters of the algorithm are the subspace dimensions and the window sizes N_{HDS} and N_{CDS}. Here we have always used the same number of principle components k for the robust PCA and the embedding and set the window sizes to five times the number of free parameters in the covariance estimate $N_{HDS} = N_{CDS} = 5\frac{k(k+1)}{2}$.

Although it is certainly interesting to look how single hybridization experiments $\boldsymbol{d}_i$ compare to the HDS, we are more interested in how the general behavior of the chip process changes over time. Therefore we define the current dataset (CDS) as $\{\boldsymbol{d}_{i-N_{CDS}/2}, \ldots, \boldsymbol{d}_i, \ldots, \boldsymbol{d}_{i+N_{CDS}/2}\}$, where i is the time of interest. This allows us to look at the data distribution in a time interval of size N_{CDS} around i. In analogy to the classical setting in statistical process control we can define the T^2-distance between the HDS and the CDS as

$$T_w^2(i) = (\boldsymbol{\mu}_{HDS} - \boldsymbol{\mu}_{CDS})^T \bar{S}^{-1} (\boldsymbol{\mu}_{HDS} - \boldsymbol{\mu}_{CDS}), \tag{3.8}$$

where $\bar{S}$ is calculated from the sample covariance matrices S_{HDS} and S_{CDS} as

$$\bar{S} = \frac{(N_{HDS} - 1)S_{HDS} + (N_{CDS} - 1)S_{CDS}}{N_{HDS} + N_{CDS} - 2}. \tag{3.9}$$

Although it is possible to use the T_w^2-distance between the historical and current dataset to test for $\boldsymbol{\mu}_{HDS} = \boldsymbol{\mu}_{CDS}$, this information is relatively meaningless. The hypothesis that the means of HDS and CDS are equal would almost always be rejected, due to the high power of the test. What is of more interest is T_w itself, which is the amount by which the two sample means differ in relation to the standard deviation of the data.

In order to see whether an observed change of the T_w^2-distance comes from a simple translation, it is also interesting to compare the two sample covariances S_{HDS} and S_{CDS}. A translation in $\log(CG/TG)$ space means that the hybridization intensities of HDS and CDS differ only by a constant factor (e.g. a change in probe concentration). This situation can be detected by looking at

$$\begin{aligned} L(i) = 2\Big[\ln|\bar{S}| \quad &- \quad \frac{N_{HDS} - 1}{N_{HDS} + N_{CDS} - 2} \ln|S_{HDS}| \\ &- \quad \frac{N_{CDS} - 1}{N_{HDS} + N_{CDS} - 2} \ln|S_{CDS}| \Big], \end{aligned}$$

which is the test statistics of the likelihood ratio test for different covariance matrices [72]. It gives a distance measure between the two covariance matrices (i.e. $L = 0$ means equal covariances).

Before we can apply the described methods to a real microarray dataset we have again to solve the problem that we need a non-singular and outlier resistant estimate of S_{HDS} and S_{CDS}. In contrast to the last section, the simple approximation of S_{HDS} by its first principle components will not work here. The reason is that changes in the experimental conditions outside the HDS will not necessarily be represented in the first principle components of S_{HDS}.

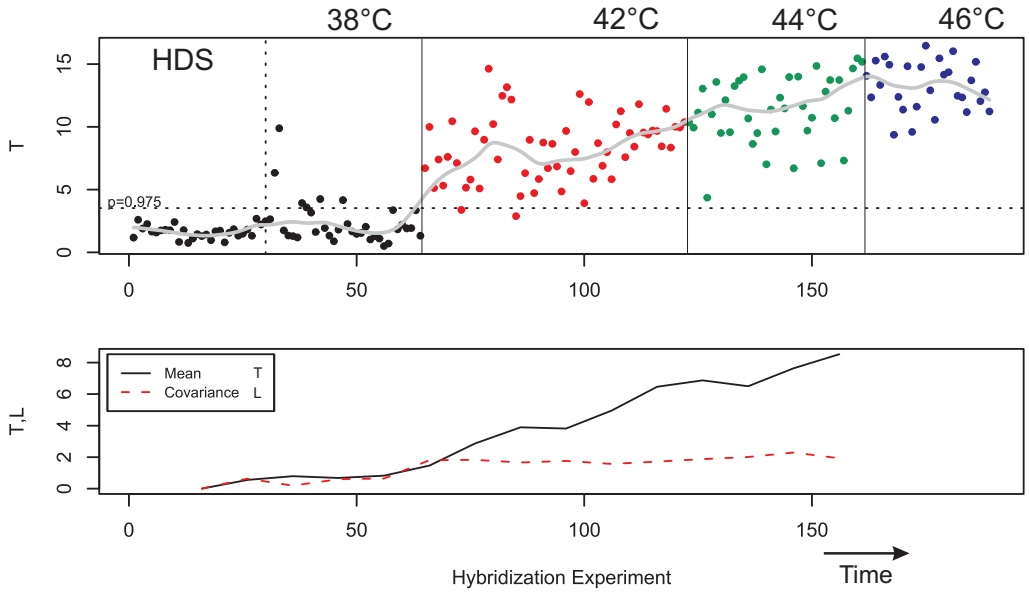

Figure 3.5: T^2 control chart of Temperature experiment. The same ALL/AML samples were hybridized at 4 different temperatures. The upper plot shows the T-distance of all 207 hybridizations to the HDS, where the grey curve shows the running average as computed by a lowess fit [160]. The lower plot shows the T_w- and L-distance between HDS and CDS with a window size of $N_{HDS} = N_{CDS} = 30$.

The solution is to first embed all the experimental data into a lower dimensional space by PCA. This works, because any significant change in the experimental conditions will be captured by one of the first principle components. S_{HDS} and S_{CDS} can then be reliably computed in the lower dimensional embedding. The problem of robustness is simply solved by first using robust PCA to remove outliers before performing the actual embedding and before computing the sample covariances. A summary of our algorithm is given in Fig. 3.4.

With the computed values for T^2, T_w^2 and L we can now generate a plot that visualizes the quality development of the chip process over time, a so called T^2 control chart.

3.3.2 Results

The first example is shown in Fig. 3.5, which demonstrates how our algorithm detects a change in hybridization temperature (see Appendix A for dataset description). As can be expected, the T^2-value grows with an increase in hybridization temperature. The systematic increase of the L-distance indicates that this is not only caused by a simple translation in methylation space.

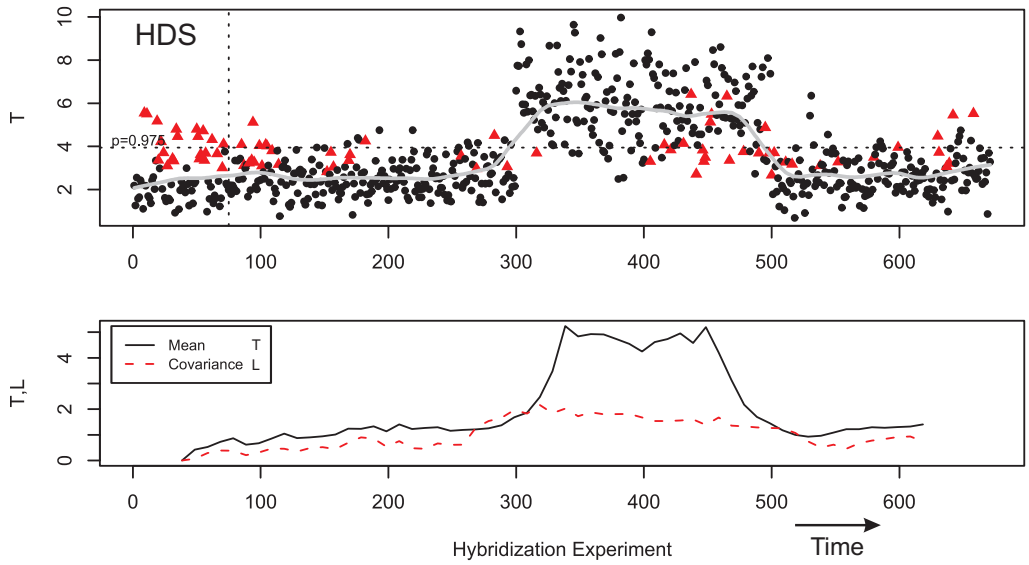

Figure 3.6: T^2 control chart of simulated probe exchange in the Lymphoma dataset. Between chips 300 and 500 an accidental oligo probe exchange during slide production was simulated by rotating 12 randomly selected CpG positions. The upper plot shows the T-distance of all 647 hybridizations, where the grey curve shows the running average as computed by a lowess fit [160]. Triangular points are chips classified as "unacceptable" by visual inspection. The lower plot shows the T_w- and L-distance between HDS and CDS with a window size of $N_{HDS} = N_{CDS} = 75$.

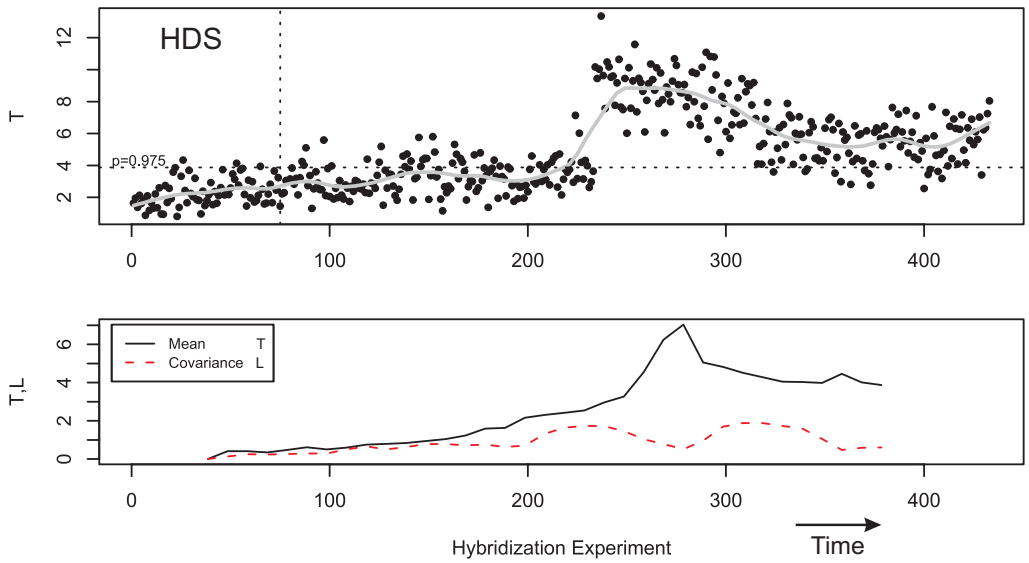

Figure 3.7: T^2 control chart of ALL/AML study. Over the course of the experiment a total of 46 oligomeres for 35 different CpG positions had to be re-synthesized. Oligos were replaced at time indices 234 and 315. The upper plot shows the T-distance of all 433 hybridizations, where the grey curve shows the running average as computed by a lowess fit [160]. The lower plot shows the T_w- and L-distance between HDS and CDS with a window size of $N_{HDS} = N_{CDS} = 75$.

The process has to be regarded as clearly out of control, because almost all chips are above the UCL after the temperature change and the process center has drifted more than $T_w = 4$ standard deviations away from its original location.

Fig. 3.6 shows how our method detects the simulated handling error in the Lymphoma dataset (see Appendix A). The affected chips can be clearly identified by the significant increase in the T^2-distances as well as by their change in the covariance structure.

Finally, Fig. 3.7 shows the T^2 control chart of the ALL/AML study (see Appendix A). It clearly indicates that the experimental conditions significantly changed two times over the course of the study. A look at the L-distance reveals that the covariance within the two detected artefact blocks is identical to the HDS. A change in covariance can be detected only when the CDS window passes the two borders. This clearly indicates that the observed effect is a simple translation of the process mean.

The major practical problem is now to identify the reasons for the changes. In this regard the most valuable information from the T^2 control chart is the time point of process change. It can be cross-checked with the laboratory protocol and the process parameters which have changed at the same time

can be identified. In our case the two process shifts corresponded to the time of replacement of re-synthesized probe oligos for slide production, which were obviously delivered at a wrong concentration. After exclusion of the affected CpG positions from the analysis the T^2 chart showed normal behavior and the overall noise level of the dataset was significantly reduced.

Chapter 4

Class prediction and feature selection

The probably most important application of microarray technology from a scientific as well as from a clinical point of view is the classification of tissue types, especially the prediction of tumor malignancy, aggressiveness and response to treatment [67, 1, 111, 4, 126]. In order to perform a methylation based tissue class prediction we will use the well known support vector machine algorithm [158, 27]. This algorithm has shown outstanding performance in several areas of application and has already been successfully used to classify mRNA expression data [10, 168, 23, 65]. The major problem of all classification algorithms for methylation and expression data analysis alike is the high dimension of input space compared to the small number of available samples. Although the support vector machine is designed to overcome this problem it still suffers from these extreme conditions. Therefore feature selection is of crucial importance for good performance [16, 168, 10] and we give special consideration to it by comparing several methods on our methylation data.

The dataset we use as an example (see Appendix A) consists of cell lines and primary tissue obtained from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). A total of 17 ALL and 8 AML samples were included. The methylation status of these samples was evaluated at 81 CpG dinucleotide positions.

The rest of this chapter is organised as follows. In the first section, we give a short introduction to the support vector machine and describe our experimental setting. In the second section, we address the problem of feature selection by introducing and comparing several methods. Results on the leukemia dataset are given for all methods.

4.1 Support Vector Machines

In our case, the task of cancer classification consists of constructing a machine that can predict the leukemia subtype (ALL or AML) from a patients methylation pattern. For every patient sample this pattern is given as the average[1] methylation scores $d_{iq} = mean_{c \in \mathcal{C}_s} S_{p(q)} \left(\{O^{CG}_{q,k,c}, O^{TG}_{q,k,c}\}_{k=1...n_r} \right)$, where i is the respective patient sample index and q a specific oligo pair. The complete patient methylation profile is given by the vector $\boldsymbol{d}_i = (d_{i1}, ..., d_{in_q})'$. Througout this chapter we will use the log difference score.

Based on a given set of training examples $D = \{\mathbf{d}_i : \mathbf{d}_i \in R^{n_q}\}$ with known diagnosis $Y = \{y_i : y_i \in \{ALL, AML\}\}$ a discriminant function $f : R^{n_q} \rightarrow \{ALL, AML\}$, where n_q is the number of oligo pairs (and for this dataset also CpG positions), has to be learned. The number of misclassifications of f on the training set $\{D, Y\}$ is called training error and is usually minimized by the learning machine during the training phase. However, what is of practical interest is the capability to predict the class of previously unseen samples, the so called generalization performance of the learning machine. This performance is usually estimated by the test error, which is the number of misclassifications on an independent test set $\{D', Y'\}$.

The major problem of training a learning machine with good generalization performance is to find a discriminant function f which on the one hand is complex enough to capture the essential properties of the data distribution, but which on the other hand avoids over-fitting the data. The Support Vector Machine (SVM) tries to solve this problem by constructing a linear discriminant that separates the training data and maximises the distance to the nearest points of the training set. This maximum margin separating hyperplane minimizes the ratio between the radius of the minimum enclosing sphere of the training set and the margin between hyperplane and training points. This corresponds to minimising the so called radius margin bound on the expected probability of a test error and promises good generalization performance [158].

Of course there are more complex classification problems, where the dependence between class labels y_i and features $\mathbf{d}_i$ is not linear and the training set can not be separated by a hyperplane. In order to allow for non-linear discriminant functions the input space can be non-linearly mapped into a potentially higher dimensional feature space by a mapping function $\Phi : \mathbf{d}_i \mapsto \Phi(\mathbf{d}_i)$. Because the SVM algorithm in its dual formulation uses only the inner product between elements of the input space, the knowledge of the kernel function $k(\mathbf{d}_i, \mathbf{d}_j) = \langle \Phi(\mathbf{d}_i) \cdot \Phi(\mathbf{d}_j) \rangle$ is sufficient to train the SVM.

[1] Every hybridisation experiment was at least 3 times repeated and the results averaged.

	Training Error 2 Features	Test Error 2 Features	Training Error 5 Features	Test Error 5 Features
Linear Kernel				
Fisher Criterion	0.01	0.05	0.00	0.03
Golub's Method	0.01	0.05	0.00	0.04
t-Test	0.05	0.13	0.00	0.08
Backward Elimination	0.02	0.17	0.00	0.05
PCA	0.13	0.21	0.05	0.28
No Feature Selection†	0.00	0.16		
Quadratic Kernel				
Fisher Criterion	0.00	0.06	0.00	0.03
Golub's Method	0.00	0.06	0.00	0.05
t-Test	0.04	0.14	0.00	0.07
Backward Elimination	0.00	0.12	0.00	0.05
PCA	0.10	0.30	0.00	0.31
Exhaustive Search	0.00	0.06	-	-
No Feature Selection†	0.00	0.15		

Table 4.1: Performance of different feature selection methods. † The SVM was trained on all 81 features.

It is not necessary to explicitly know the mapping Φ and a non-linear SVM can be trained efficiently by computing only the kernel function. Here we will only use the linear kernel $k(\mathbf{d}_i, \mathbf{d}_j) = \langle \mathbf{d}_i \cdot \mathbf{d}_j \rangle$ and the quadratic kernel $k(\mathbf{d}_i, \mathbf{d}_j) = (\langle \mathbf{d}_i \cdot \mathbf{d}_j \rangle + 1)^2$.

In the next section we will compare SVMs trained on different feature sets. In order to evaluate the prediction performance of these SVMs we used a cross-validation method [14]. For each classification task, the samples were partitioned into 8 groups of approximately equal size. Then the SVM predicted the class for the test samples in one group after it had been trained using the 7 other groups. The number of misclassifications was counted over 8 runs of the SVM algorithm for all possible choices of the test group. To obtain a reliable estimate for the test error the number of misclassifications were averaged over 50 different partitionings of the samples into 8 groups.

4.2 Feature selection

The simplest way for applying a SVM to our methylation data is to use every CpG position as a separate dimension, not making any assumption about the interdependence of CpG sites from the same gene. On the leukemia subclassification task the SVM with linear kernel trained on this 81 dimensional input space had an average test error of 16%. Using a quadratic kernel did not significantly improve the results (see Table 4.1). An obvious explanation

for this relatively poor performance is that we have only 25 data points (even less in the training set) in a 81 dimensional space. Finding a separating hyperplane under these conditions is a heavily under-determined problem. And as it turns out, the SVM technique of maximising the margin is not sufficient to find the solution with optimal generalization properties. It is necessary to reduce the dimensionality of the input space while retaining the relevant information for classification. This should be possible because it can be expected that only a minority of CpG positions has any connection with the two subtypes of leukemia.

4.2.1 Principle Component Analysis

The probably most popular method for dimension reduction is principle component analysis (PCA) [14]. For a given training set D, PCA constructs a set of orthogonal vectors (principle components) which correspond to the directions of maximum variance. The projection of D onto the first k principle components gives the 2-norm optimal representation of D in a k-dimensional orthogonal subspace. Because this projection does not explicitly use the class information Y, PCA is an unsupervised learning technique.

In order to reduce the dimension of the input space for the SVM we performed a PCA on the combined training and test set $\{D, D'\}$ and projected both sets on the first k principle components. This gives considerably better results than performing PCA only on the training set D and is justified by the fact that no label information is used. However, the generalization results for $k = 2$ and $k = 5$, as shown in Table 4.1, were even worse than for the SVM without feature selection. The reason for this is that PCA does not necessarily extract features that are important for the discrimination between ALL and AML. It first picks the features with the highest variance, which are in this case discriminating between cell lines and primary patient tissue (see Fig. 4.1a), i.e. subgroups that are not relevant to the classification task. As is shown in Fig. 4.2, features carrying information about the leukemia subclasses appear only from the 9th principle component on. The generalization performance including the 9th component is significantly better than for a SVM without feature selection. However, it seems clear that a supervised feature selection method, which takes the class labels of the training set into account, should be more reliable and give better generalization.

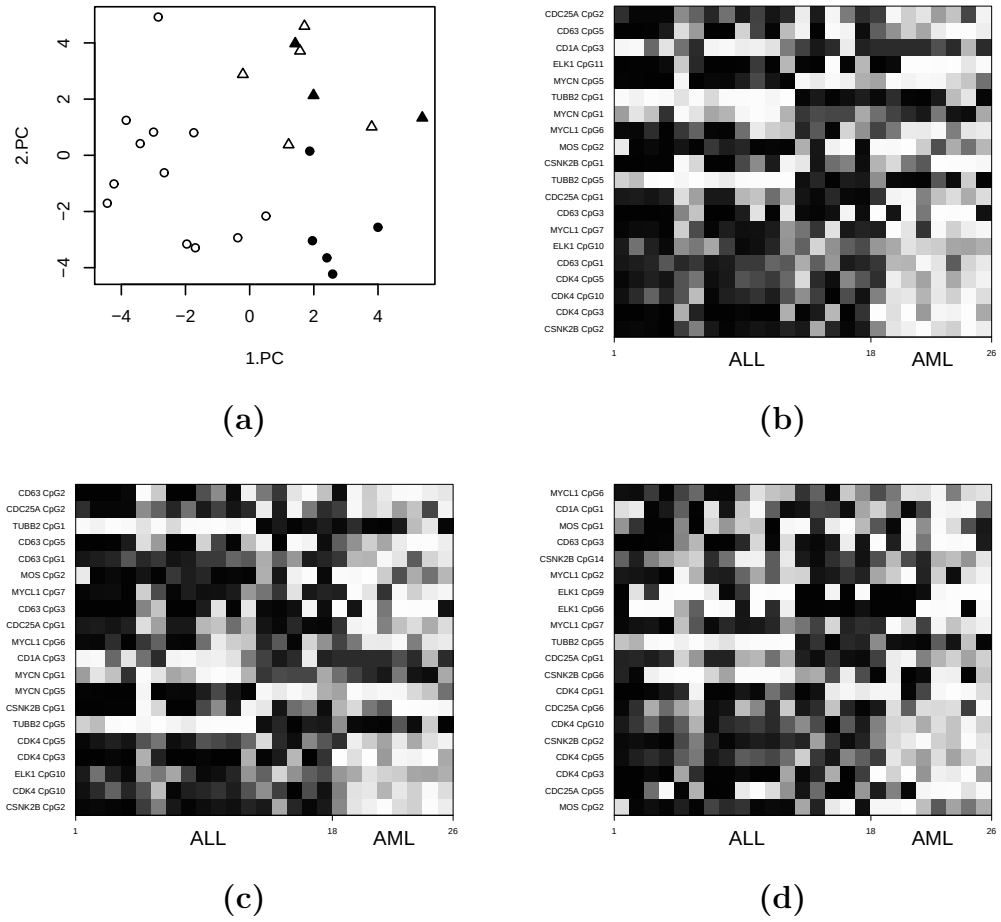

Figure 4.1: Feature selection methods. **a)** Principle component analysis. The whole dataset was projected onto its first 2 principle components. Circles represent cell lines, triangles primary patient tissue. Filled circles or triangles are AML, empty ones ALL samples. **b)** Fisher criterion. The 20 highest ranking CpG sites according to the Fisher criterion are shown. The highest ranking features are on the bottom of the plot. High probability of methylation corresponds to black, uncertainty to grey and low probability to white. **c)** Two sample t-test. **d)** Backward elimination.

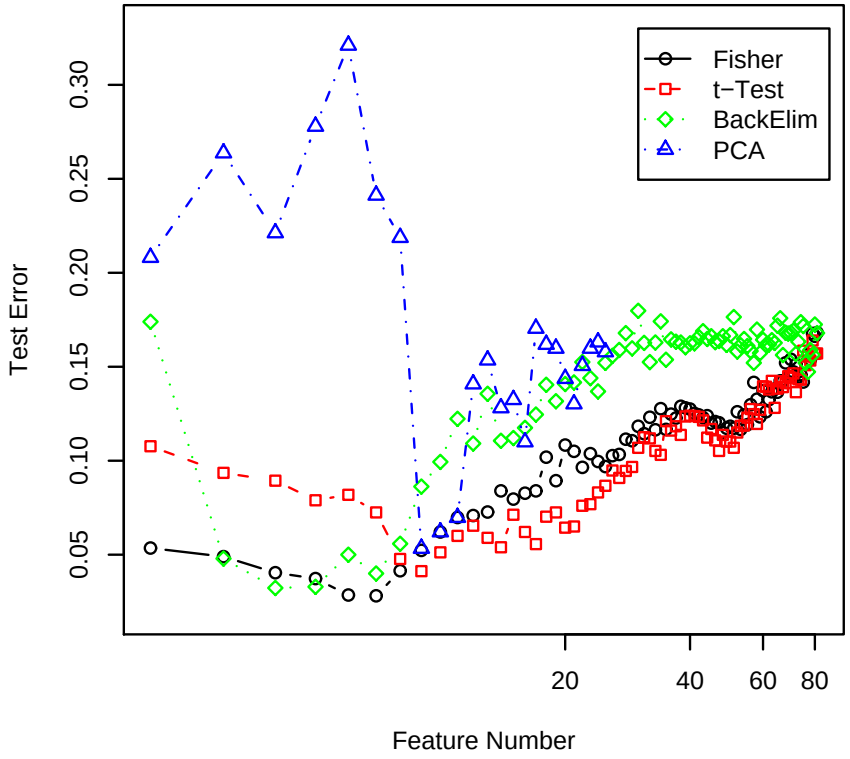

Figure 4.2: Dimension dependence of feature selection performance. The plot shows the generalization performance of a linear SVM with four different feature selection methods against the number of selected features. The x-axis is scaled logarithmically and gives the number of input features for the SVM, starting with two. The y-axis gives the achieved generalization performance. Note that the maximum number of principle components corresponds to the number of available samples. The performance of Golub's method was very similar to the Fisher criterion and is not shown.

4.2.2 Fisher criterion and t-test

A classical measure to asses the degree of separation between two classes is given by the Fisher criterion [14]. In our case it gives the discriminative power of the kth CpG as

$$J(k) = \frac{(\mu_k^{ALL} - \mu_k^{AML})^2}{\sigma_k^{ALL^2} + \sigma_k^{AML^2}},$$

where $\mu_k^{ALL/AML}$ is the mean and $\sigma_k^{ALL/AML}$ is the standard deviation of all d_{ik} with $y_i = ALL/AML$. The Fisher criterion gives a high ranking for CpGs where the two classes are far apart compared to the within class variances. Fig. 4.1b shows the methylation profiles of the best 20 CpGs according to the Fisher criterion. The very similar criterion

$$G(k) = \frac{|\mu_k^{ALL} - \mu_k^{AML}|}{\sigma_k^{ALL} + \sigma_k^{AML}}$$

was used by Golub and coworkers for their ALL/AML classification based on mRNA expression data [67]. Its relation to the Fisher criterion is given by

$$G^2(k) = J(k)\left(1 + \frac{2\sigma_k^{ALL}\sigma_k^{AML}}{\sigma_k^{ALL^2} + \sigma_k^{AML^2}}\right)^{-1},$$

which shows the preference of Golub's ranking for features with different within class variances compared to the Fisher criterion.

Another approach to rank CpGs by their discriminative power is to use a test statistic for computing the significance of class differences. Here we assumed a normal distribution of the methylation levels of a CpG position within a class and used a two sample t-test to rank the CpGs according to the significance of the difference between the class means [113]. Fig. 4.1c shows the ranking, which is very similar to the Fisher criterion because a large mean difference and a small within class variance are the important factors for both methods.

In order to improve classification performance we trained SVMs on the k highest ranking CpGs according to the Fisher criterion, Golub's method or t-test. Fig. 4.3 shows a trained SVM on the best two CpGs from the Fisher criterion. The test errors for $k = 2$ and $k = 5$ are given in Table 4.1. The results show a dramatic improvement of generalization performance. Using the Fisher criterion for feature selection and $k = 5$ CpGs the test error was decreased to 3% compared to 16% for the SVM without feature selection. Fig. 4.2 shows the dependence of generalization performance from the selected dimension k and indicates that especially the Fisher criterion gives

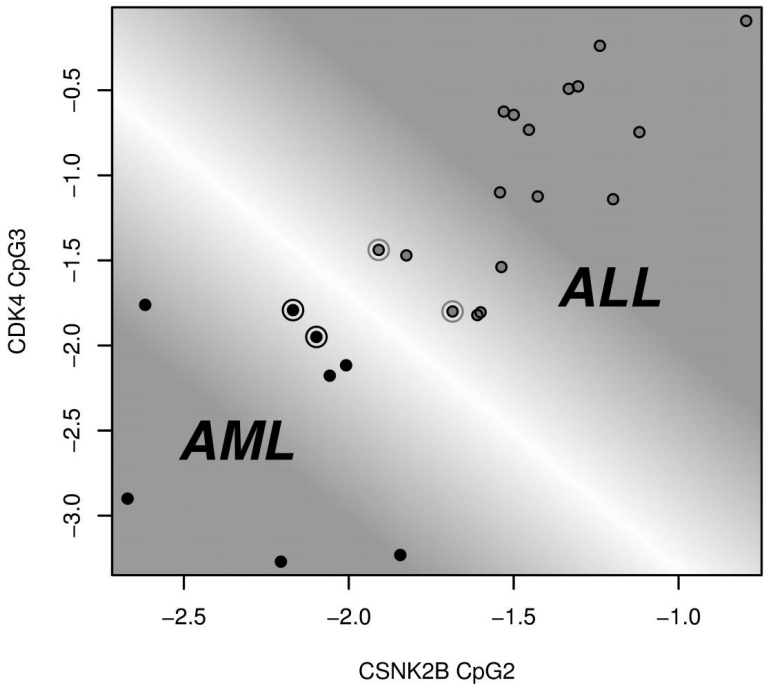

Figure 4.3: Support Vector Machine on two best features of the Fisher criterion. The plot shows a SVM trained on the two highest ranking CpG sites according to the Fisher criterion with all ALL and AML samples used as training data. The black points are AML, the grey ones ALL samples. Circled points are the support vectors defining the white borderline between the areas of AML and ALL prediction. The grey value of the background corresponds to the prediction strength.

dimension independent good generalization for reasonable small k. The performance of Golub's ranking method was equal or slightly inferior to the Fisher criterion on our dataset, whereas the t-test performance was considerably worse for small feature numbers.

Although the described CpG ranking methods give very good generalization, they have some potential drawbacks. One problem is that they can only detect linear dependencies between features and class labels. A simple XOR or even OR combination of two CpGs would be completely missed. Another drawback is that redundant features are not removed. In our case there are usually several CpGs from the same gene which have a high likelihood of comethylation. This can result in a large set of high ranking features which carry essentially the same information. Although the good results seem to indicate that the described problems do not appear in our dataset, they should be considered.

4.2.3 Backward elimination

PCA, Fisher criterion and t-test construct or rank features independent of the learning machine that does the actual classification and are therefore called filter methods [16]. Another approach is to use the learning machine itself for feature selection. These techniques are called wrapper methods and try to identify the features that are important for the generalization capability of the machine. Here we propose to use the features that are important for achieving a low training error as a simple approximation. In the case of a SVM with linear kernel these features are easily identified by looking at the normal vector $\mathbf{w}$ of the separating hyperplane. The smaller the angle between a feature basis vector and the normal vector the more important is the feature for the separation. Features orthogonal to the normal vector have obviously no influence on the discrimination at all. This means the feature ranking is simply given by the components of the normal vector as w_k^2. Of course this ranking is not very realistic because the SVM solution on the full feature set is far from optimal as we demonstrated in the last subsections. A simple heuristic is to assume that the feature with the smallest w_k^2 is really unimportant for the solution and can be safely removed from the feature set. Then the SVM can be retrained on the reduced feature set and the procedure is repeated until the feature set is empty. Such a successive feature removal is called backward elimination [16]. The resulting CpG ranking on our dataset is shown in Fig. 4.2d and differs considerably from the Fisher and t-test rankings. It seems backward elimination is able to remove redundant features. However, as shown in Table 4.1 and Fig. 4.2 the generalization results are not better than for the Fisher criterion. Furthermore, backward elimination seems to be more dimension dependent and it is computationally more expensive. It follows that at least for this dataset the simple Fisher criterion is the preferable feature selection technique.

4.2.4 Exhaustive search

A canonical way to construct a wrapper method for feature selection is to evaluate the generalization performance of the learning machine on every possible feature subset. Cross-validation on the training set can be used to estimate the generalization of the machine on a given feature set. What makes this exhaustive search of the feature space practically useless is the enormous number of $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ different feature combinations and there are numerous heuristics to search the feature space more efficiently (e.g. backward elimination) [16].

Here we only want to demonstrate that there are no higher order correlations between features and class labels in our dataset. In order to do this we exhaustively searched the space of all two feature combinations. For every of the $\binom{81}{2} = 3240$ two CpG combinations we computed the leave-one-out cross-validation error of a SVM with quadratic kernel on the training set. From all CpG pairs with minimum leave-one-out error we selected the one with the smallest radius margin ratio. This pair was considered to be the optimal feature combination and was used to evaluate the generalization performance of the SVM on the test set.

The average test error of the exhaustive search method was with 6% the same as the one of the Fisher criterion in the case of two features and a quadratic kernel. For five features the exhaustive computation is already infeasible. In the absolute majority of cross-validation runs the CpGs selected by exhaustive search and Fisher criterion were identical. In some cases suboptimal CpGs were chosen by the exhaustive search method. These results clearly demonstrate that there are no second order combinations of two features in our dataset that are important for an ALL/AML discrimination. We expect that higher than second order combinations of more than two features can not be detected reliably with such a limited sample size. Therefore the Fisher criterion should be able to extract all classification relevant information from our dataset.

Chapter 5

Identification and validation of colorectal neoplasia-specific methylation markers

Although colorectal cancer is the second most common cause of malignant death in industrialized countries, the mechanisms and pathways of the neoplastic events associated with this complex disease are not well understood. Genetic alterations in colorectal neoplasia have been studied extensively as candidate markers for detection and analysis of the disease [87, 161, 58, 29, 88], but much less is known about epigenetic changes, including aberrant methylation of genes. Several genes have been shown to be preferentially hypermethylated in both colorectal cancer and premalignant adenomas with dramatic effects on the expression of their resultant proteins [77, 76, 92, 93, 90] indicating that silencing of tumor suppressor genes or other genes in tumor pathways can occur both from mutation events and/or aberrant methylation.

Application of expression-based microarray profiling has proven effective in distinguishing RNA profile differences between tumor types and classes providing information for understanding of tumor pathways [67, 4]. Recently, this technology has been adapted to methylation-based microarray profiling which can distinguish the epigenetic methylation profile of samples from large groups of patients [1, 66]. This type of analysis detects methylation ratios at CpG positions that have been amplified by polymerase chain reaction (PCR) from bisulfite-modified genomic DNA. By evaluating modified DNA from different patient populations, these arrays have been used to identify methylation markers that distinguish among types of tumors, differentiate tumors from normal tissue and predict clinical outcome [1]. Since the methylation microarray requires larger amounts of DNA it is not appli-

cable to clinical situations where only low levels of DNA are available from samples such as biopsies or body fluids. To achieve sensitive detection of DNA methylation from such sources, real-time PCR methods, for example MethyLight, can be used to distinguish patient profiles [45].

From a clinical perspective more accurate detection markers are needed to improve the effectiveness and efficiency of both the screening and surveillance of colorectal neoplasia. Aberrantly methylated genes represent attractive candidate markers for this purpose, as cancer-specific methylation changes occur early in tumorigenesis [78], appear to be stable, yield a positive amplifiable signal, and can be assayed with high analytical sensitivity. Unfortunately, many or the more commonly described methylation markers in the literature such as ER, MGMT, MLH1 and CDKN2a have not been adequately tested for specificity to a target cancer by simultaneously analyzing methylation status across multiple tumor types and normal tissue. As a result many of these most widely investigated markers are not suitable for specific detection of a particular disease. For example, methylation of the gene CDKN2a (p16) has been reported to be found in blood from patients with numerous types of cancer including oral cancer, gastric cancer, melanoma, non-small cell lung cancer, hepatocellular cancer, and bladder cancer in a number of independent studies [119, 154, 83, 110, 28, 156]. Clearly, methylation of this gene is important in neoplastic progression, but its utility as a specific marker for a single cancer in a screening application is questionable. Furthermore, CDKN2a has been shown to be methylated in blood from individuals with non-cancerous diseases, albeit at a lower rate [28, 79].

Because of the genetic heterogeneity of colorectal neoplasia, multiple genetic markers may be required for acceptable tumor detection rates [2, 40]. Since methylation occurs early and in distinct genomic areas, it might be possible to achieve high clinical sensitivity with a smaller number of methylated DNA markers [77]. Feasibility studies have shown that aberrantly methylated DNA markers can be assayed from serum or plasma [78, 175, 120, 52, 19, 155, 69], and from stool [3, 118, 25] to detect colorectal cancer. However, robustly conducted genome-wide searches are needed to identify methylated DNA sequences that optimally discriminate colorectal neoplasia from other tissues and normal blood components.

In this study, we report the use of a genome-wide PCR-based discovery process to identify sequences that are differentially methylated between colorectal neoplasia, normal colon tissue and peripheral blood lymphocytes (PBLs) from healthy age-matched individuals. We provide validation of these differential methylation markers via use of both methylation microarrays and real-time PCR for discrimination of colorectal neoplasia compared to healthy mucosa and age-matched healthy PBLs and also to other disease states,

including actively inflamed epithelia and malignant tissues. The markers identified are consistent with the concept that hypermethylation is an important proponent of tumorogenesis since several of the candidates found in our genome-wide screening have recently been implicated as being involved in the neoplastic process and several candidates from our literature-based search previously reported to be involved in cancer were verified in this study. The high accuracy of these markers suggests that the sensitive, methylation specific real-time PCR assays described in this study may be useful for detection of disease at early stages in blood and for interrogation of neoplastic pathways. Based on our comprehensive analysis of these candidate markers in diverse tissue types we suggest potential applications for the markers.

5.1 Materials and methods

5.1.1 Patient samples

Institutional review boards at all participating sites approved this study.

Genome-wide discovery

Differentially methylated sequences were identified using pathologically verified colonic tissue samples obtained from the National Disease Research Interchange (NDRI, Philadelphia, PA/USA), the Cooperative Human Tissue Network (CHTN, Nashville, TN/USA) and ILSbio (Chestertown, MD/USA). These included 25 adenocarcinomas, 6 adenomas, and 42 tumor-free control tissues. Normal blood for peripheral blood lymphocyte isolation was obtained from Puget Sound Blood Center (Seattle, WA).

Gene array

Pathologically verified tissues were obtained from surgical procedures or endoscopic biopsies performed at the Mayo Clinic (Rochester, MN), Semmelweis University Clinic (Budapest, Hungary) or University Hospital Carl Gustav Carus (Dresden, Germany). All normal tissues were obtained from patients endoscopically verified as absent of lesions and without a history of neoplasia. The total sample set included 358 patient DNAs and two control DNAs. The patient DNAs were extracted from 29 normal colon samples, 31 inflammatory bowel disease (IBD), 55 colon polyps (45 polyps $<$1cm, 10 polyps $\geq$1cm), 89 colorectal cancers (30 Dukes A/B, 56 Dukes C/D, 1 unknown, 2 high grade polyps$\geq$1cm), 116 non-colonic cancer samples from liver (9), bile duct (10), pancreas (10), lung (squamous and adenocarcinoma)

(38), breast (28), prostate (5), esophagus (6), stomach (10), PBL (14) and normal tissue from sites other than colon: esophagus mucosa (7), gastric mucosa (7), liver (10). Additionally one control sample of unmethylated human DNA (Molecular Staging), and one control sample of enzymatically methylated DNA (SssI, NEB) was included. All colon and lung tissues were matched by age/sex as well as location in the colon and the lung (central and peripheral).

MethyLight assays

Pathologically verified tissues were obtained from surgical procedures or endoscopic biopsies performed at the Mayo Clinic (Rochester, MN), Semmelweis University Clinic (Budapest, Hungary) or University Hospital Carl Gustav Carus (Dresden, Germany) or by commercial sample collections performed by Asterand (Detroit, MI), Integrated Lab Services (Research Triangle Park, NC) and Clinomics (Pittsfield, MA) in accordance with a provided specimen collection protocol. The total sample set included 149 patient DNAs from normal colon tissue (18), pathologically normal colon tissue adjacent to tumor (28), normal PBLs (25), IBD (9), colon polyps (11), colorectal cancers (28), breast cancer (15) and liver cancer (15). Not all assays were run on all samples because of limited DNA amounts.

5.1.2 DNA extraction

DNA extraction of snap-frozen surgical samples for discovery was performed using Genomic Tip-500 columns (Qiagen, Valencia, CA). Extraction for the microarray and real-time PCR assays was optimized by sample type including tissue sections from snap-frozen tissue, frozen surgical specimens and snap-frozen small biopsies. Surgical specimens from University Hospital Carl Gustav Carus were extracted using Genomic Tip-100 columns. Frozen tissue sections from Mayo Clinic were extracted using a MagNa Pure device (Roche Applied Science; Indianapolis, IN). DNA from biopsies performed at Semmelweis University Clinic was prepared using Qiagen buffers and the High Pure PCR Template Preparation Kit (Roche Applied Science).

5.1.3 Genome-wide identification of differentially methylated sequences

To identify markers with high specificity for colon cancer we used pooled genomic DNA from colonic normal, adenomas and adenocarcinoma tissue and

analyzed them using the previously described methods, methylation specific-arbitrarily primed PCR (MS-APPCR) [104] and methylated CpG island amplification (MCA) [151].

Patient samples used in these experiments were divided into three age groups: >65 years of age, 50 to 65 years, and <50 years. Samples were also divided into 4 types depending on the extent of disease. 1) normal adjacent tissue (NAT) (>6cm from tumor) or no disease, 2) adenomas, 3) cancer with no nodal involvement or metastasis (N0M0), and 4) advanced disease with nodal involvement (N1-2,M0) and/or metastasis (N1-2,M1). For each of these age and disease groups 3-5 patient samples were combined into one pool. In addition methylation patterns of all cancerous and pre-cancerous conditions from all age groups were compared to age-matched normal peripheral blood lymphocytes.

5.1.4 Gene array

The microarray was performed as described in Chapter 2 with oligonucleotides covering regions of 43 discovery and literature-derived genes and 2 control genes. For the discovery derived genes primer pairs and oligonucleotides were designed around the identified differentially methylated sequence whenever possible. Multiple primer pairs and oligonucleotides were designed for some genes for a total of 54 amplicons and a total of 248 oligonucleotide pairs. Each oligonucleotide contained 2-3 CpG sites. Hybridization conditions allowed the detection of single nucleotide differences. Additionally 8 negative control oligonucleotides with random sequences were included to facilitate estimation of unspecific background hybridization. The methylation proportion of each oligonucleotide was estimated from 4 spot repetitions per microarray and on average 4 hybridization repetitions per sample using the maximum likelihood score (see Section 2.3). Unmethylated human DNA (Molecular Staging, New Haven, CT) and enzymatically methylated control DNAs (SssI; New England Biomedical) were used to calibrate the data. Amplicons for all discovery genes, candidate genes and control genes used in the combined array are shown in Appendix Table A.1.

5.1.5 MethyLight assays

The MethyLight assays were performed on the ABI Prism 7900 (Applied Biosystems) using standard TaqMan chemistry as previously reported by Eads et al [45]. Standard curves for each assay were established using CpGenome Universal Methylated DNA (Serologicals/Chemicon) at concentrations between 31.6 pg/ul and 31.6 ng/ul DNA. Sample DNA was diluted

to 2 ng/ul and aliquoted into strip tubes for 3 assays. 10 ng DNA/rxn was tested in duplicate for each assay. A methylation unspecific assay for β actin was used to determine total bisulfite-treated DNA concentration for each sample.

5.1.6 Statistical analysis

Analysis of the gene array data was performed on log10-transformed methylation proportions averaged over all CpG positions from the same gene by computing the mean. Hierarchical clustering of the gene array data was performed by using the simple 2-norm as distance metric between samples and between genes. Samples and genes were clustered using Ward's minimum variance method [160]. Fisher's exact test was used to test for association between clustering results and phenotypes. AUC values were estimated using the trapezoidal rule. P-values were computed with a Wilcoxon test. A simple cut-off classifier was used for classification. Sensitivity and specificity were estimated by 200 bootstrapping runs that randomly divided the dataset into training (about 2/3 of the samples) and test set (about 1/3 of the samples). For every bootstrap run the cut-off was set to 95% specificity on the respective training set. Sensitivity and specificity were then computed from the respective test set. We report median sensitivity and specificity values from the 200 bootstrap runs as well as 90% confidence intervals (5% and 95% quantiles of the bootstrap estimates). For the two-marker panel analysis the reported panel value for each sample was computed by taking the maximum of the two individual marker measurement values. MethyLight analysis was performed on the ratio of methylated DNA (measured by the respective marker assay) to total bisulfite DNA (measured by the β actin assay). Ideally this ratio results in a number in the range $[0, 1]$ and represents the proportion of methylated DNA in the respective sample. DNA amounts were estimated from the respective standard curves by linear regression. Replicate marker measurements were averaged.

5.2 Results

5.2.1 Genome-wide discovery

The discovery process resulted in over 500 unique sequences that were potential candidates for colorectal cancer biomarkers. The differentially methylated sequences identified using MS-APPCR [104] and MCA [86] were scored and prioritized using the following scoring variables:

- appearance using multiple discovery methods
- appearance in multiple pools of like samples
- located within a CpG island
- located within the promoter region of a gene
- located near or within predicted or known genes
- known to be associated with disease
- class of gene (transcription factor, growth factor, tumor suppressor, oncogene)
- repetitive element.

Under this scoring schema, a sequence received a point for each of the above criteria, and received a score of -8 for having repetitive sequence content greater than 50%. Therefore, the highest score possible was 7; the lowest was -8. Scores were automatically calculated for each sequence using genomic annotations from the Ensembl database (http://www.ensembl.org).

Using the scoring criteria above along with manual review of the sequences, 30 sequences were selected for microarray analysis (Table A.1). Sequences with significant (>50%) repetitive element content were eliminated from consideration. Our comprehensive database of sequences derived from internal genome-wide discovery experiments allowed us to also eliminate sequences found using other previously tested tumor types. Selected sequences scored 1 or greater with the majority scoring 3 or more.

5.2.2 Gene array study

For additional confirmation of the methylation state of the potential markers we constructed a methylation specific gene array containing oligonucleotides representing the 30 selected genome-wide discovery sequences and also 13 potential methylation biomarkers from the literature (Table A.1). Additional genes were chosen from a previous microarrray study of literature-derived sequences and selected based on involvement in neoplasia and performance discrimination for colorectal cancer vs. pathologically normal colon tissue (data not shown). In our discovery experiments, the exon 1 region of the TMEFF2 gene was identified as being hypermethylated. Since the promoter region of this gene had been described as differentially methylated in the literature [103] and was also shown to discriminate between CRC and healthy

colon in the previous microarray study (data not shown), this region was included as a candidate sequence. TMEFF2 methylation measurements from promoter and exon 1 region are highly correlated (between amplificate correlation R=0.76) and were therefore aggregated and treated as one locus for further analysis.

We determined the ability of the 43 differentially methylated gene regions to discriminate between colorectal cancer and other tissues using a large, highly diverse sample set containing colorectal cancer tissue and tissue samples from other types of cancers, colon inflammatory conditions, colon polyps and numerous histopathologically determined normal tissues.

Hierarchical clustering

To identify systematic similarities in the overall methylation patterns of samples and genes we performed a hierarchical clustering on the entire gene set and the set of 204 colon-derived tissue DNA samples (Fig. 5.1). The majority of normal and inflammatory colon samples fall into a cluster that shows no methylation on most genes (Cluster N: 25 normal, 29 inflammatory, 12 colon polyp and 16 CRC samples). The other cluster (Cluster C) consists predominantly of neoplastic samples and is clearly separated into two sub-clusters (C1 and C2), which show different degrees of hypermethylation. The sub-cluster with the strongest methylation is composed only of neoplastic tissue (Cluster C1: 28 colon polyp and 38 CRC samples). The other sub-cluster shows an intermediate degree of methylation and includes some histologically normal and inflammatory samples (Cluster C2: 4 normal, 2 inflammatory, 15 colon polyp and 35 CRC samples). There is no significant association between the two neoplastic sub-clusters and tumor stage or grade. However, there are a significantly higher number of adenomas larger than 1 cm in the sub-cluster C1 than in the sub-clusters N and C2 (C1: 14 colon polyps$\geq$1cm, 15 colon polyps$<$1cm; C2: 2 colon polyps$\geq$1cm, 13 colon polyps$<$1cm; N: 2 colon polyps$\geq$1cm, 11 colon polyps$<$1cm; $P < 0.01$).

As can be expected from the directed selection of candidate sequences for the microarray study the overall clustering results show a clear separation between normal and inflammatory samples in cluster N on the one side and polyp and CRC samples in cluster C on the other side. Cluster N contains 90% of the non-neoplastic samples. Cluster C contains 81% of the neoplastic and pre-neoplastic samples. The majority of discriminatory markers are hypermethylated in polyp and CRC samples from cluster C and show a typical CpG island methylator phenotype (CIMP) [85]. Polyp and CRC samples in Cluster N on the other hand are samples not methylated for the majority of discriminatory markers tested and appear to be a CIMP negative population.

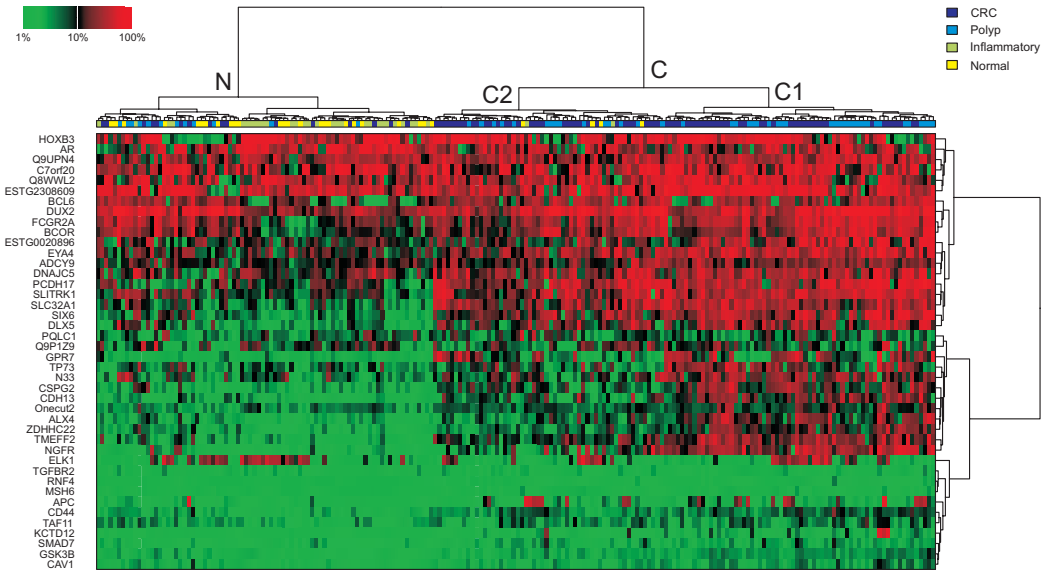

Figure 5.1: Hierarchical Clustering of all 204 colon-derived tissue samples and all 42 loci from the gene array. Columns are patient samples and rows are genomic loci. Row annotations give the gene name. The class information was unknown to the clustering algorithm. The average degree of methylation of each genomic locus in each sample is represented by the decadic logarithm of the methylation proportion ranging from below 1% methylated alleles (green) to methylation of all alleles (red). There are three main tissue clusters labeled as N, C1 and C2. Cluster N composition: 25 (30%) normal colon, 29 (35%) inflamed colon, 12 (15%) colon polyp and 16 (20%) colon cancer samples. Cluster C1 composition: 28 (42%) colon polyp and 38 (58%) colon cancer samples. Cluster C2 composition: 4 (7%) normal colon, 2 (4%) inflamed colon, 15 (27%) colon polyp and 35 (63%) colon cancer samples. Between cluster comparison: C1 has the highest degree of methylation and contains 46% of the neoplastic and pre-neoplastic samples. C2 contains 10% of the non-tumor and 35% of the neoplastic and pre-neoplastic samples. N contains 90% of the non-tumor samples, 18% of the CRC samples and 22% of the colon polyp samples.

The observed strong similarity between the CRC and colon polyp samples is supported by previous studies that show early alterations in methylation in pre-cancerous conditions of the colon [138, 8]. Based on the clustering results, all subsequent analyses of the data combine the CRC and colon polyp samples for comparison to normal tissue, other cancers and IBD.

Individual marker performance

To quantify the influence of non-colon derived tissues on the classification performance of individual markers we analyzed the dataset in two different ways. First we looked at the complete sample set. Here the negative class consisted of 214 samples from normal colon (29), inflammatory colon (31), PBL (14) and other normal (10 liver, 7 stomach, and 7 esophagus) and non-colorectal cancer tissues (28 breast cancer, 38 lung cancer, 9 liver cancer, 10 pancreatic cancer, 10 bile duct cancer, 10 stomach cancer, 5 prostate cancer, and 6 esophageal cancer). The positive class was composed of 144 CRC and polyp samples. 30 markers were highly significant with $P < 0.0001$. 10 markers showed a very strong class separation with an area under the ROC curve (AUC) of $\geq$0.8 (Fig. 5.2). The sensitivity of these strong markers ranged between 35% and 52% at a specificity level of 95%.

In a second analysis we looked only at colon-derived tissues. In this case the negative class consisted of 60 samples from normal and inflammatory colon. The positive class was again composed of 144 colon cancer and polyp samples. Despite the lower sample number as compared to the full dataset, 29 markers were highly significant with $P < 0.0001$. 19 markers showed a very strong class separation with an AUC of $\geq$0.8 (Fig. 5.2). The sensitivity of these strong markers ranged between 44% and 81% at a specificity level of 95%. The omission of non-colon derived tissues resulted in a strong increase of ΔAUC$\geq$0.05 for 17 markers and a strong decrease of ΔAUC$\leq$-0.05 for 5 markers. Classification results of all individual markers are summarized in Fig. 5.2.

Marker panel performance

Using a panel of markers does not significantly improve performance over the best single marker, TMEFF2. The best two-marker panel is TMEFF2 plus NGFR. This panel has a sensitivity of 55% (CI 44%-68%) at 95% specificity in the classification of all samples (+5% compared to TMEFF2 alone). The sensitivity in classifying only colon derived tissue samples is 85% (CI 75%-93%) at 95% specificity (+4% compared to TMEFF2 alone).

Gene Name	Source	Gene Array						MethyLight			
		All Tissues			Colon Derived Tissues			All Tissues			
		AUC	*P-Value*	Sensitivity @ 95% Specificity	AUC	*P-Value*	Sensitivity @ 95% Specificity	AUC	*P-Value*	Sensitivity	Specificity
TMEFF2 (HPP1)	Candidate APPCR	**0.90**	<0.0001	0.50 [0.34,0.75]	**0.94**	<0.0001	0.81 [0.71,0.89]	0.88	<0.0001	0.83 [0.57,1.00]	0.95 [0.84,1.00]
ZDHHC22	APPCR	**0.88**	<0.0001	0.51 [0.36,0.65]	**0.87**	<0.0001	0.52 [0.35,0.73]	0.85	<0.0001	0.68 [0.50,0.86]	0.95 [0.85,1.00]
SLITRK1	MCA	**0.86**	<0.0001	0.50 [0.36,0.65]	**0.85**	<0.0001	0.58 [0.35,0.72]	0.72	0.0086	0.38 [0.00,0.72]	0.95 [0.84,1.00]
SLC32A	APPCR	**0.85**	<0.0001	0.51 [0.36,0.61]	**0.89**	<0.0001	0.67 [0.55,0.80]	0.77	0.0012	0.60 [0.25,1.00]	0.95 [0.84,1.00]
DLX5	APPCR	**0.84**	<0.0001	0.42 [0.23,0.56]	**0.83**	<0.0001	0.55 [0.41,0.69]				
GSK3B	Candidate	**0.84**	<0.0001	0.37 [0.17,0.50]	**0.89**	<0.0001	0.58 [0.40,0.75]				
NGFR	APPCR	**0.83**	<0.0001	0.52 [0.39,0.66]	**0.87**	<0.0001	0.71 [0.61,0.82]	0.82	<0.0001	0.73 [0.57,0.91]	0.95 [0.84,1.00]
PCDH17	MCA	**0.83**	<0.0001	0.49 [0.38,0.59]	**0.82**	<0.0001	0.52 [0.37,0.74]				
TUSC3 (N33)	Candidate	**0.82**	<0.0001	0.39 [0.23,0.55]	0.74	<0.0001	0.30 [0.14,0.49]	0.84	<0.0001	0.60 [0.33,1.00]	0.95 [0.83,1.00]
BCOR	APPCR	**0.80**	<0.0001	0.35 [0.20,0.47]	**0.82**	<0.0001	0.41 [0.27,0.56]				
RNF4*	APPCR	0.79	<0.0001	0.38 [0.19,0.52]	**0.84**	<0.0001	0.56 [0.27,0.71]				
SIX6	APPCR	0.77	<0.0001	0.28 [0.17,0.45]	**0.85**	<0.0001	0.55 [0.34,0.75]				
FCGR2A	APPCR	0.77	<0.0001	0.28 [0.14,0.48]	**0.81**	<0.0001	0.43 [0.26,0.60]				
CD44	Candidate	0.77	<0.0001	0.13 [0.06,0.28]	**0.86**	<0.0001	0.50 [0.29,0.65]				
CSPG2	Candidate	0.76	<0.0001	0.24 [0.12,0.38]	**0.83**	<0.0001	0.62 [0.34,0.73]				
DUX2	MCA	0.75	<0.0001	0.25 [0.09,0.43]	0.70	<0.0001	0.16 [0.00,0.43]				
CDH13	Candidate	0.74	<0.0001	0.35 [0.24,0.44]	**0.81**	<0.0001	0.62 [0.49,0.75]				
BCL6	APPCR	0.73	<0.0001	0.31 [0.18,0.47]	0.78	<0.0001	0.35 [0.16,0.56]	0.59	0.0001	0.17 [0.00,0.43]	1.00 [0.97,1.00]
GPR7	APPCR	0.73	<0.0001	0.23 [0.13,0.33]	**0.82**	<0.0001	0.58 [0.43,0.71]				
Onecut2	APPCR	0.72	<0.0001	0.14 [0.06,0.25]	0.77	<0.0001	0.27 [0.11,0.46]				
ADCY9	APPCR	0.72	<0.0001	0.32 [0.22,0.41]	0.72	<0.0001	0.33 [0.20,0.48]				
TAF11	APPCR	0.71	<0.0001	0.25 [0.12,0.38]	0.66	0.0004	0.27 [0.07,0.47]				
CAV1	Candidate	0.70	<0.0001	0.21 [0.13,0.32]	0.75	<0.0001	0.33 [0.20,0.51]				
SMAD7	APPCR	0.68	<0.0001	0.18 [0.09,0.29]	0.71	<0.0001	0.18 [0.08,0.45]	0.61	0.037	0.14 [0.00,0.38]	0.95 [0.86,1.00]
KCTD12	APPCR	0.67	<0.0001	0.16 [0.08,0.28]	0.77	<0.0001	0.44 [0.29,0.61]				
DNAJC5	APPCR	0.67	<0.0001	0.21 [0.13,0.29]	0.73	<0.0001	0.28 [0.12,0.47]				
EYA4	Candidate	0.66	<0.0001	0.11 [0.06,0.18]	**0.81**	<0.0001	0.54 [0.39,0.68]	0.88	<0.0001	0.55 [0.20,0.88]	0.95 [0.83,1.00]
ALX4	APPCR	0.65	<0.0001	0.12 [0.05,0.21]	**0.83**	<0.0001	0.55 [0.43,0.69]	0.75	0.0006	0.54 [0.20,0.83]	0.95 [0.82,1.00]
ESTG0020896	APPCR	0.64	<0.0001	0.27 [0.16,0.38]	0.63	0.0042	0.33 [0.25,0.44]				
ESTG2308609	APPCR	0.62	0.0002	0.18 [0.10,0.27]	0.56	0.15	0.17 [0.06,0.36]				
Q8WWL2	APPCR	0.59	0.0027	0.09 [0.04,0.15]	0.53	0.48	0.10 [0.04,0.20]				
TP73	Candidate	0.59	0.0024	0.06 [0.02,0.11]	0.68	<0.0001	0.28 [0.18,0.44]				
Q9UPN4	MCA	0.59	0.0039	0.13 [0.06,0.22]	0.63	0.0038	0.23 [0.09,0.35]				
APC	Candidate	0.59	0.0057	0.04 [0.00,0.10]	**0.81**	<0.0001	0.46 [0.17,0.58]				
HOXB3	APPCR	0.59	0.0044	0.15 [0.07,0.23]	0.55	0.27	0.09 [0.04,0.16]				
MSH6	Candidate	0.58	0.013	0.06 [0.02,0.12]	0.64	0.0010	0.15 [0.02,0.32]				
PQLC1	MCA	0.54	0.17	0.04 [0.00,0.10]	0.51	0.86	0.02 [0.00,0.09]				
Q9P1Z9	APPCR	0.51	0.70	0.07 [0.02,0.13]	0.54	0.35	0.15 [0.07,0.25]				
C7orf20	MCA	0.50	0.92	0.09 [0.04,0.15]	0.53	0.56	0.10 [0.04,0.18]				
ELK1	Control	0.46	0.16	0.07 [0.00,0.13]	0.42	0.082	0.04 [0.00,0.10]				
TGFBR2	Candidate	0.46	0.17	0.04 [0.00,0.08]	0.50	0.92	0.06 [0.00,0.14]				
AR	Candidate	0.37	<0.0001	0.02 [0.00,0.06]	0.33	0.0001	0.02 [0.00,0.08]				

Figure 5.2: Single marker classification performance for microarray (left) and MethyLight assays (right). Shown is the area under the ROC curve (AUC), P-value (Wilcoxon) and sensitivity at 95% specificity (median value plus 90% confidence interval estimated by bootstrapping) for each marker. For the microarray data results are given for the classification of all samples (CRC and adenomas (N=144) vs. normal colon, inflammatory colon and other normal and cancerous tissues (N=221)) and for the classification of only colon derived tissues (CRC and adenomas (N=144) vs. inflammatory and normal colon (N=60)). Markers with a very strong class separation (AUC$\geq$0.8) in the overall OR colon only classification are printed bold. For the MethyLight data results are only given for the classification of all samples (CRC and adenomas (N=39) vs. normal adjacent colon, inflammatory colon and other cancerous tissues (N=110), for some assays sample numbers were lower due to insufficient DNA amounts). Specificity confidence intervals of the microarray data at 95% were very similar between different markers (for all tissues all intervals covered by: [0.87,1.00]; for colon derived tissues all intervals covered by: [0.77,1.00]). For the MethyLight data specificity is given explicitly for every marker.

Paneling does not significantly increase sensitivity of the markers for colorectal cancer over TMEFF2 alone because all of our markers detect the same subset of CRC and polyp samples. The CIMP positive cancer cluster shown in Fig. 5.1 (Cluster C) includes 81% of the colorectal carcinoma and polyps in the study. TMEFF2 alone is heavily hypermethylated on 67% of these CIMP positive samples (78/116 CIMP samples with TMEFF2 methylation >10%). Only 11% of the remaining CIMP negative CRC and polyp samples show TMEFF2 hypermethylation (3/28 CIMP negative samples with TMEFF2 methylation >10%). Since no additional marker shows significant hypermethylation on the CIMP negative samples or significantly higher methylation levels than TMEFF2 on the CIMP positive samples, overall marker complementarity is minimal.

Distribution of methylation frequencies

To further understand the behavior of the strongest markers on different tissue types we looked at the distribution of methylation frequencies on all 358 samples grouped into 5 major tissue classes. For this analysis CRC and polyp as well as normal and inflammatory colon tissue samples were combined since their respective methylation rate distributions were similar. Fig. 5.3 shows box plots of the major tissue classes for all markers from the gene array with an AUC of $\geq$0.8. Median methylation levels of all tissue subclasses are shown in Fig. 5.4 and detailed box plots for all tissue subclasses can be found in Fig. 5.5. The overall low degree of methylation of markers GSK3B, RNF4 and CD44 is a result of the poor correlation between different CpG positions within the same amplicon (median between CpG correlations: GSK3B R=0.27, RNF4 R=0.28, CD44 R=0.41, all other 17 most discriminating markers R$\geq$0.47) and indicates a lack of co-methylation within the CpG island.

Generally all markers with the exception of APC show hypermethylation of the colorectal cancer class compared to the healthy colon and the PBL class. However, the methylation patterns of our markers differ considerably with regard to the non-colonic healthy tissues and non-colonic cancer classes. Markers TMEFF2, ZDHHC22, SLITRK1, SLC32A1, DLX5, GSK3B, NGFR, PCDH17, N33 and BCOR differentiate colon neoplasia very well from the majority of other tissues (AUC$\geq$0.8). All show strong hypermethylation of colorectal cancer compared to other tissues with varying differences between the other tissue classes. Other markers such as RNF4, SIX6, CD44, CSPG2, CDH13, GPR7, EYA4, ALX4 and APC show only small or no differences between colorectal cancer and the non-colonic cancer class. N33 shows significant hypermethylation of colorectal cancer compared to normal colon but

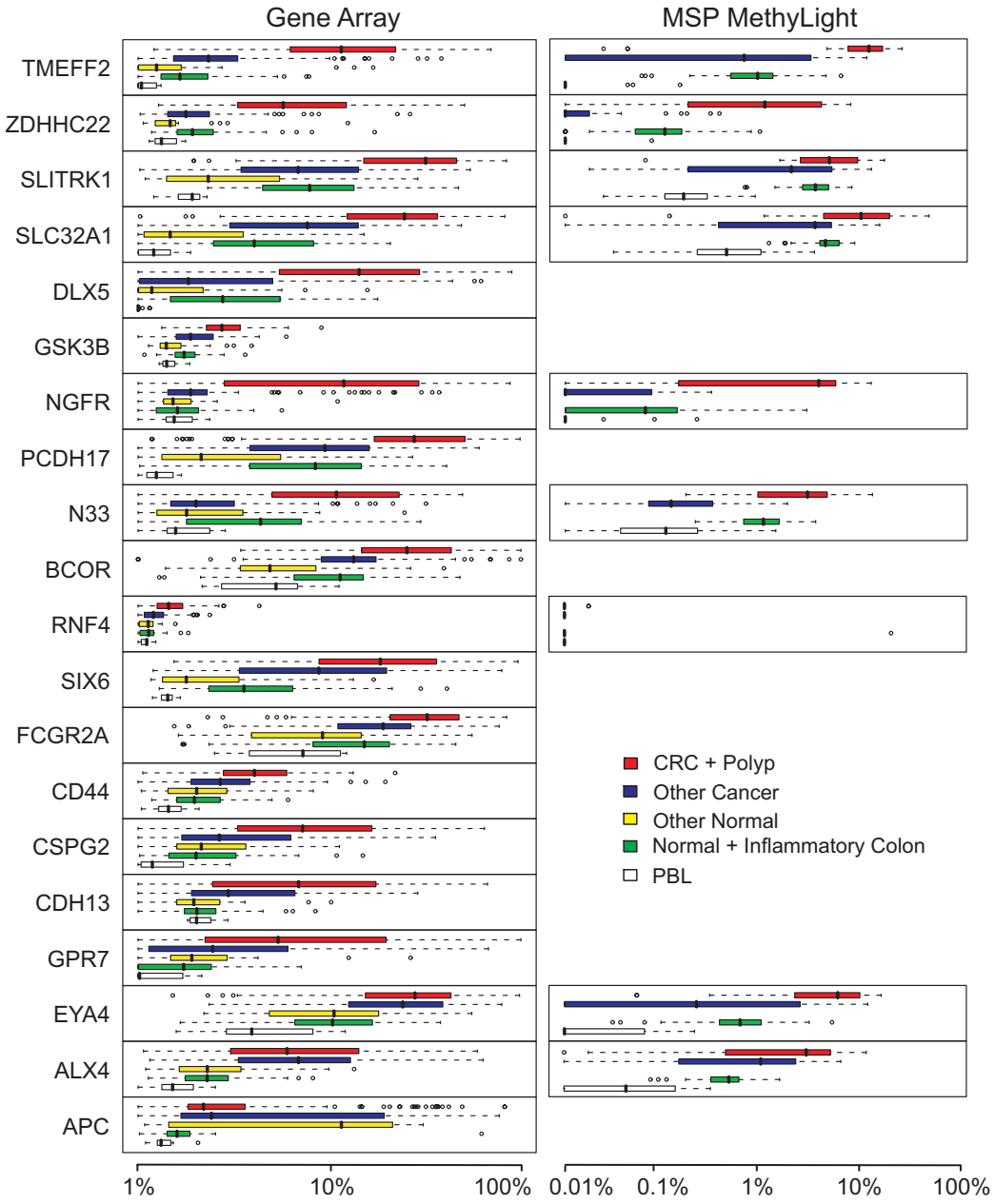

Figure 5.3: Methylation levels of different tissue classes. For each marker gene (rows) the distribution of methylation levels in the major tissue classes is visualized by a box plot. The left column shows methylation levels from microarray analysis. Horizontal axis is percent methylation with 1-100% methylation scale. The right column shows methylation levels from real time MethyLight analysis. Horizontal axis is 0.01-100% methylation. Individual box plots show the middle 50% of the data, the middle line is the median, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Methylation measurement values outside the whisker range are plotted as individual points.

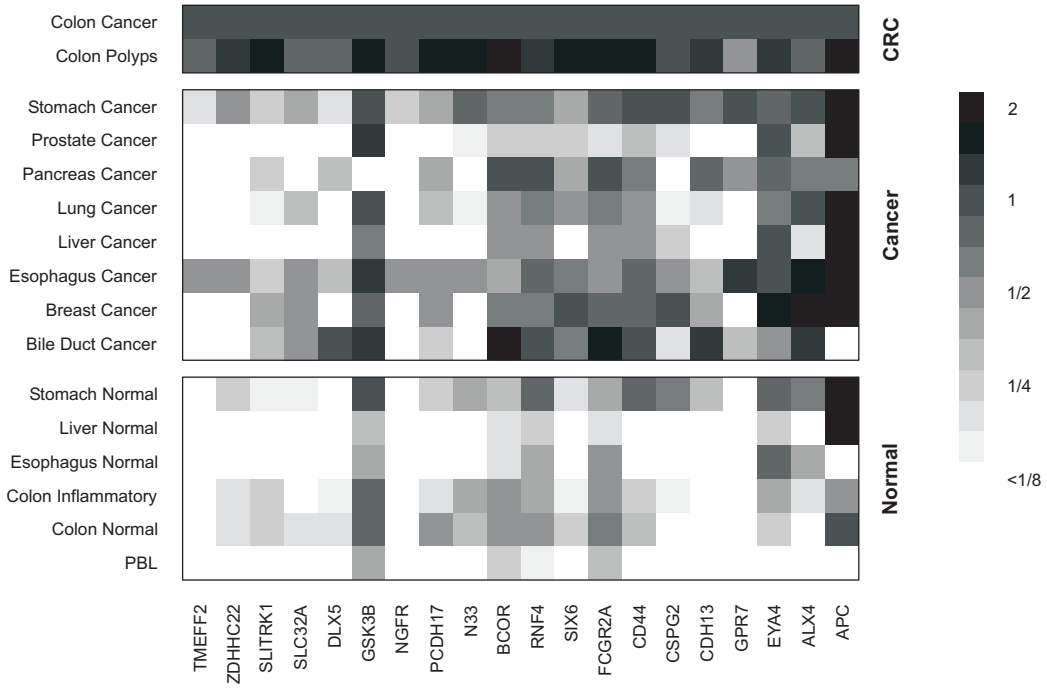

Figure 5.4: Relative methylation levels of normal and non-colorectal cancer tissue classes in comparison to CRC and polyps. For each tissue class (rows) and each marker gene (columns) the median methylation level is plotted as fold change over the median CRC methylation level. Fold changes are restricted to a range of 2-fold hypermethylation and 8-fold hypomethylation over the median CRC methylation level. See Fig. 5.5 for box plots of all subclasses.

also gives a very strong discrimination between colon tissue (normal colon and colorectal cancer) and most other tissues. All of our markers show some degree of hypermethylation in stomach and esophageal cancer tissue and to some lower extent in normal stomach tissue.

5.2.3 Marker validation with MethyLight assays

We developed 11 real-time MethyLight assays for markers that were designated as having strong to poor performance on the gene array. 9 of the markers had very high performance (AUC$\geq$0.8) in the colon tissue only classification (TMEFF2, ZDHHC21, SLITRK1, SSLC32A, NGFR, N33, RNF4, EYA4, and ALX4). 2 markers with poorer performance (BCL6, SMAD7) were tested because although the array results were not strong, original discovery scoring of these sequences was high (6 and 4, respectively) and this information would also allow us to further correlate array performance results with real-time assay results. For TMEFF2 the real-time assay was designed in the promoter region of the gene. Classification performance of the MethyLight assays was estimated on an independent sample set with a negative class of normal and normal adjacent colon (46 samples), inflammatory colon (9 samples), and other cancerous tissues (15 breast cancer and 15 liver cancer). The positive class was composed of 39 colon cancer and adenoma samples. Of the 9 MethyLight assays for the strongest gene array markers, five (TMEFF2, ZDHHC21, NGFR, N33, EYA4) were highly significant with $P < 0.0001$ and showed a substantial class separation with an AUC of $\geq$0.8 (Fig. 5.2). Three assays (SLITRK1, SSLC32A and ALX4) showed a significant but weaker class separation ($P < 0.009$; AUC$\geq$0.72). RNF4, a strong candidates from the gene array could not be reproduced using real-time PCR analysis since almost all amplifications yielded no product. This is likely due to a lack of significant co-methylation of CpG sites within the assay region but was not further investigated. The 2 poorer performing gene array markers (BCL6, SMAD7) showed poor results with their corresponding real-time MethyLight assays, confirming results obtained using the gene array. The sensitivity of the five strongest markers ranged between 55% and 83% at a specificity level of 95%. Fig. 5.3 shows the methylation frequency distributions of the 9 MethyLight assays for the most discriminative gene array markers. The scale of methylation level for these real-time assays is extended to 0.01% methylation as compared to 1% methylation used for the gene array data because of the increased sensitivity of real-time PCR. At this level of analytical sensitivity, TMEFF2, ZDHHC22 and NGFR are completely negative on PBL and show high specificity with regard to other tissues indicating these markers may be excellent candidates for blood-based early detection

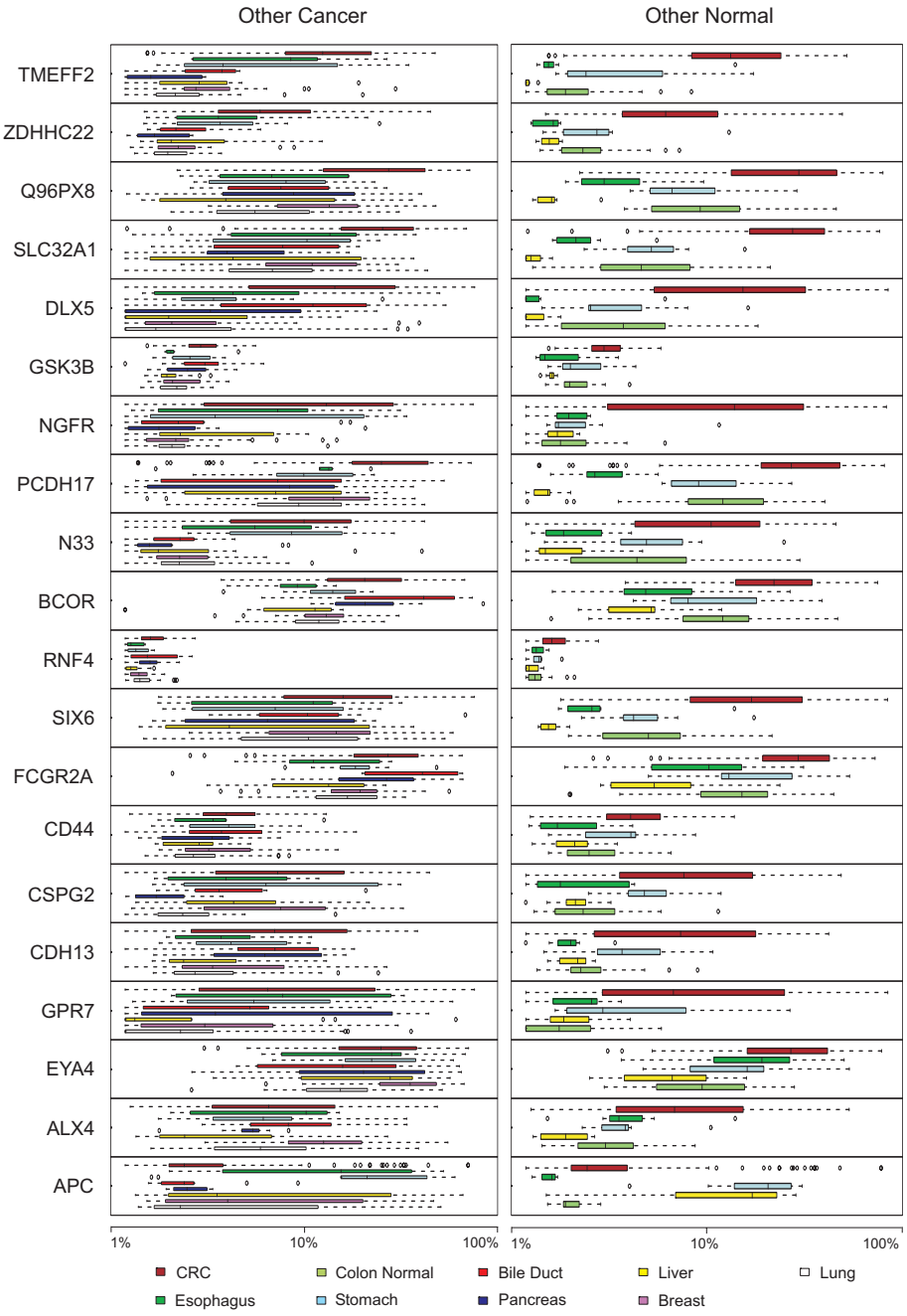

Figure 5.5: Detailed methylation distribution of non-colonic tissue classes from microarray analysis. The left column shows methylation levels of non-colorectal cancers. Colorectal cancer (CRC) is given as a reference. The right column shows methylation levels of non-colonic normal tissues. Normal Colon and CRC are given as references. Horizontal axis is percent methylation with 1-100% methylation scale. Gene names are shown on vertical axis. Individual box plots show the middle 50% of the data, the middle line is the median, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Methylation measurement values outside the whisker range are plotted as individual points.

applications. TMEFF2, ZDHHC22 and EYA4 all have minimal overlap of methylation levels between CRC and normal and inflammatory colon tissue making them potential candidates for stool based assays or molecular classification tests, however quantitation would be necessary for these analyses.

Chapter 6

Discussion

6.1 Measuring DNA methylation

We have derived a simple generative model that quantitatively explains the hybridization intensities of DNA methylation microarrays. It takes into account systematic biases from unspecific background hybridization as well as stochastic deviations from the microarray hybridization reaction. Based on this model we derived algorithms for variance stabilization, microarray normalization, and finally quantification of DNA methylation.

The derived methylation scores have different properties and the choice of which score to use depends on the application requirements. The log ratio score is very simple and can detect very small changes in methylation while providing almost constant variance of measurement noise over the whole methylation range. However, it does not provide a clear relation to the actual proportion of methylated DNA except for the simple monotonicity property that higher score values correspond to higher methylation. For normally working oligonucleotide pairs the log ratio score results are virtually identical to the more complicated generalized log ratio score. Therefore the log ratio is the score of choice for quality control and pure classification or marker selection applications where a direct estimation of the proportion of methylated DNA is not necessary.

The maximum likelihood score can provide unbiased estimates of the relative amount of methylated DNA in a given sample. It does this by taking information from dedicated calibration experiments into account and therefore has a clear advantage compared to the overly simple proportion score. An obvious disadvantage of the maximum likelihood score is that it cannot provide a constant variance of the measurement noise. Depending on the dataset, taking the logarithm of the methylation estimates can help to

generate more symmetrical distributions of measurement values. The maximum likelihood score is the score of choice when direct quantification of DNA methylation proportions and the comparison of measurement values with other DNA methylation measurement technologies are important.

The presented results show that DNA methylation microarrays together with the proper pre-processing algorithms can accurately quantify relative amounts of methylated DNA in tissue samples and constitute a promising high throughput tool for DNA methylation research.

6.2 Controlling quality and stability of microarray experiments

We have shown that robust principle component analysis and techniques of statistical process control can be used to detect flaws in microarray experiments. Robust PCA has proven to be able to automatically detect nearly all cases of outlier chips identified by visual inspection, as well as microarrays with inconspicous image quality but saturated hybridization signals. With the T^2 control chart we introduced a tool that facilitates the detection and assessment of even minor systematic changes in large-scale microarray studies.

A major advantage of both methods is that neither rely on an explicit modeling of the microarray process since they are solely based on the distribution of the actual measurements. Having successfully applied our methods to the example of DNA methylation data, we assume that the same results can be achieved with other types of microarray platforms. The sensitivity of the methods improves with increasing study sizes, due to their multivariate nature. This makes them particularly suitable for medium to large-scale experiments in a high throughput environment.

The retrospective analysis of a study with our methods can greatly improve results and avoid misleading biological interpretations. When the T^2 control chart is monitored in real time a given quality level can be maintained in a very cost effective way. On the one hand, this allows for an immediate correction of process parameters. On the other hand, this makes it possible to specifically repeat only those slides affected by a process artefact. This guarantees high quality while minimizing the number of repetitions.

A general shortcoming of T^2 control charts is that they only indicate that something went wrong, but not exactly what was the source. Therefore we have used the time point at which a significant change happened in order to identify the responsible process parameter. We have shown that

changes in covariance structure provide additional information and permit discrimination between different problems like changes in probe concentration and accidental handling errors. However, further work will be necessary to facilitate an efficient detection of error sources.

6.3 Class prediction and feature selection

We have demonstrated that in order to achieve reliable predictions on the basis of small training set sizes the selection of relevant features is necessary, even for advanced learning algorithms such as the support vector machine. For classification tasks where the class information is directly correlated to single CpG dinucleotide markers, the simple Fisher criterion is a powerful and efficient feature selection strategy. For more complex problems it will be necessary to derive feature selection algorithms that can remove or combine redundant features and handle higher order feature dependencies.

Our results clearly demonstrate that microarray based methylation analysis combined with supervised learning techniques can reliably predict known tumor classes. Classification results were comparable to mRNA expression data and our results suggest, that methylation analysis can be applied to other kinds of tissue. Well documented tissue samples with patient history can typically be obtained only as archived specimens. This strongly limits the amount and number of tissues available for expression analysis [17]. The methylation approach has the potential to overcome this fundamental limitation: through the mere fact that since DNA is the object of study, extraction of usable material is possible form archived samples. This enables the examination of methylation patterns in large numbers of archived specimen with comprehensive clinical records and removes one of the major limitations for the discovery of complex biological processes by statistical means.

6.4 Identification and validation of colorectal neoplasia-specific methylation markers

Using the combined approach of genome-wide methylation discovery with candidate marker identification, followed by microarray analysis and real-time PCR verification, we identified a set of highly methylated sequences that are present in colorectal neoplasia. We identified markers such as TMEFF2, NGFR and ZDHHC22 that have high specificity in the diverse sample set and may be useful in clinical applications such as non-invasive screening for detection of colorectal neoplasia in either blood or stool based tests. We also

found markers that discriminate between normal colon and tissue with early neoplastic changes which have potential for use in molecular classification of colon tissue to more accurately determine early neoplastic changes, tumor aggressiveness or treatment response. For example TMEFF2, ZDHHC22, and EYA4 could be useful for molecular classification of early stages of disease in applications such as inflammatory bowel disease surveillance.

Many genes identified during our discovery and validation process have not been reported to be methylated in the setting of cancer biology and may provide insight into gene regulation. BCOR has been shown to be associated with genes involved in cancer or regulation of cell growth. Recent studies indicate BCOR is a transcriptional repressor of BCL6, a proto-oncogene and is an important transcriptional regulator of embryogenesis [82, 121]. Inactivation of this gene by methylation in the promoter region could provide a selective advantage for malignant cell growth. NGFR, also known as p75 (NTR), was recently identified as a tumor suppressor gene that induces apoptosis in malignant cells [147]. No association with methylation in the promoter region of this gene or inhibition of this gene by methylation has been previously described. Other identified genes such as SLITRK1 and SLC32A1 have neither been associated with cancer nor reported as having aberrant methylation in their promoter regions. Interestingly another solute carrier family member, SLC5A8 has been implicated as a tumor suppressor and also shown to be methylated in both gastric cancer cell lines and primary gastric cancers [153]. Clearly these genes warrant further investigation into their roles in malignant transformation.

Since greater than 90% of the marker candidates identified in the methylation array study could be validated by real-time PCR (MethyLight) analysis, these data support the use of our process to identify and confirm methylation biomarkers. By using a broad genome-wide method to identify initial candidates along with a systematic selection system to differentiate those candidates with characteristics most likely to be biologically important followed by verification on methylation microarrays and finally validation using real-time PCR we have clearly shown that valuable biomarkers for oncological diagnostic applications, such as TMEFF2, ZDHHC22, and NGFR can be found.

It is also evident that the markers identified in this study do not identify all colorectal cancer tissues. The lack of increase in sensitivity with panelling of the markers and the inability to identify all colorectal tumors with these panels is thought to be due to the manner in which our markers were identified and also potentially due to biology. Since at all stages in our process we identified and tested our markers in relation to healthy samples and other cancers, we have eliminated many markers that are methylated

to any degree in these tissues. For example, GSK3B, EYA4 and APC were not identified in our discovery process and although very highly methylated in colorectal cancer and adenomas they are also methylated in other cancers and healthy tissues. Due to the use of pooling in our initial genome wide discovery experiments we also introduced a bias towards markers that show hypermethylation on a majority of CRC samples. The signal of a hypothetical marker having hypermethylation only on a small subclass of CRC samples would have been effectively diluted out by the pooling procedure. However, biologically one can question whether methylation changes occur in all colorectal tumors. Indeed we observed that many of the tumors with increased methylation in one marker, exhibit increased methylation in multiple regions as also reported by Issa [86]. Are the remaining samples a CIMP negative population? Follow on marker identification studies will therefore be focused on studying the tissues that are methylation negative for the current marker set in order to answer this question.

In addition, further analysis of these candidate marker genes with close attention to their association with clinical variables such as age, sex, colonic location, smoking history, family history, and others that have been shown to be key predictors of cancer phenotype and clinical outcomes could provide additional insight into their potential as biomarkers. Further prospective studies of these markers, based on real-time PCR assays, either in a remote sample amenable to population screening or in biopsy samples from longitudinal studies are indicated.

6.5 Conclusions

Taken together, we have developed a package of algorithms that addresses all major aspects of data analysis for DNA methylation microarrays. We are able to accurately measure the proportion of methylated DNA in a given tissue sample including a strict control for single array quality as well as subtle changes of study conditions over time. These methylation measurements can then be used to build optimal predictors for tissue classification, discover new tissue subclasses or to select marker genes for further development of diagnostic tests. The exciting new opportunities this technology provides are demonstrated by our identification, detailed description and validation of several promising new DNA methylation markers for the early detection of colorectal cancer.

Bibliography

[1] P Adorján, J Distler, E Lipscher, F Model, J Müller, C Pelet, A Braun, A R Florl, D Gütig, G Grabs, A Howe, M Kursar, R Lesche, E Leu, A Lewin, S Maier, V Müller, T Otto, C Scholz, W A Schulz, H H Seifert, I Schwope, H Ziebarth, K Berlin, C Piepenbrock, and A Olek. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.*, 30(5):e21, 2002.

[2] D A Ahlquist, J E Skoletsky, K A Boynton, J J Harrington, D W Mahoney, W E Pierceall, S N Thibodeau, and A P Shuber. Colorectal cancer screening by detection of altered human DNA in stool: feasibility of a multitarget assay panel. *Gastroenterology*, 119(5):1219–1227, Nov 2000.

[3] D S A Ahlquist, K K Klatt, J J Harrington, and J M Cunningham. Novel use of hypermethylated DNA markers in stool for detection of colorectal cancer: a feasibility study. *Gastroenterology*, 122(Suppl. A40), 2002.

[4] 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 Jr Hudson, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, R Levy, W Wilson, M R Grever, J C Byrd, D Botstein, P O Brown, and L M Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.

[5] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, 2006.

[6] J T Attwood, R L Yung, and B C Richardson. DNA methylation and the regulation of gene transcription. *Cell Mol Life Sci*, 59(2):241–57, 2002.

[7] P Babinger, I Kobl, W Mages, and R Schmitt. A link between DNA methylation and epigenetic silencing in transgenic Volvox carteri. *Nucleic Acids Res*, 29(6):1261–71, 2001.

[8] Alfa H C Bai, Joanna H M Tong, Ka-Fai To, Michael W Y Chan, Ellen P S Man, Kwok-Wai Lo, Janet F Y Lee, Joseph J Y Sung, and Wai K Leung. Promoter hypermethylation of tumor-related genes in the progression of colorectal neoplasia. *Int J Cancer*, 112(5):846–853, Dec 2004.

[9] A J Bell and T J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput*, 7(6): 1129–1159, Nov 1995.

[10] A Ben-Dor, L Bruhn, N Friedman, I Nachman, M Schummer, and Z Yakhini. Tissue classification with gene expression profiles. In *Proceedings of the fifth annual international conference on computational molecular biology*, 2001. in press.

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

[12] T H Bestor. The DNA methyltransferases of mammals. *Hum Mol Genet*, 9(16):2395–2402, Oct 2000.

[13] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, Jan 2002.

[14] C M Bishop. *Neural networks for pattern recognition.* Oxford University Press, New York, 1995.

[15] J Bishop, S Blair, and A M Chagovetz. A competitive kinetic model of nucleic Acid surface hybridization in the presence of point mutants. *Biophys J*, 90(3):831–840, Feb 2006.

[16] A Blum and P Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[17] D D L Bowtell. Options available - from start to finish - for obtaining expression data by microarray. *Nature genetics suppl.*, 21:25–32, 1999.

[18] G.E.P. Box and D.R. Cox. An analysis of transformations. *J. Roy. Stat. Soc. Series B*, 26:211–252, 1964.

[19] J Brabender, H Usadel, K D Danenberg, R Metzger, P M Schneider, R V Lord, K Wickramasinghe, C E Lum, J Park, D Salonga, J Singer, D Sidransky, A H Holscher, S J Meltzer, and P V Danenberg. Adenomatous polyposis coli gene promoter hypermethylation in non-small cell lung cancer is associated with survival. *Oncogene*, 20(27):3528–3532, Jun 2001.

[20] R Brent. *Algorithms for minimization without derivatives*. Prentice-Hall, 1973.

[21] I N Bronnstein and K A Semendjajew. *Taschenbuch der Mathematik*. Teubner, Leipzig, 1991.

[22] C S Brown, P C Goodwin, and P K Sorger. Image metrics in the statististical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. USA*, 98(16):8944–8949, July 2001.

[23] M P Brown, W N Grundy, D Lin, N Cristianini, C W Sugnet, T S Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA*, 97:262–267, 2000.

[24] P Carninci, T Kasukawa, S Katayama, J Gough, and M C Frith et al. The transcriptional landscape of the mammalian genome. *Science*, 309 (5740):1559–63, 2005.

[25] Wei-Dong Chen, Z James Han, Joel Skoletsky, Jeff Olson, Jerome Sah, Lois Myeroff, Petra Platzer, Shilong Lu, Dawn Dawson, Joseph Willis, Theresa P Pretlow, James Lutterbaugh, Lakshmi Kasturi, James K V Willson, J Sunil Rao, Anthony Shuber, and Sanford D Markowitz. Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. *J Natl Cancer Inst*, 97(15):1124–1132, Aug 2005.

[26] Y Chen, E R Dougherty, and M Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, 2:364–374, 1997.

[27] N Christianini and J Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.

[28] Hyung Jun Chu, Jeong Heo, Soo Boon Seo, Gwang Ha Kim, Dae Hwan Kang, Geun Am Song, Mong Cho, and Ung Suk Yang. Detection of aberrant p16INK4A methylation in sera of patients with liver cirrhosis

and hepatocellular carcinoma. *J Korean Med Sci*, 19(1):83–86, Feb 2004.

[29] D C Chung. The genetic basis of colorectal cancer: insights into critical pathways of tumorigenesis. *Gastroenterology*, 119(3):854–865, Sep 2000.

[30] W J Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, New York, 1999.

[31] J F Costello, M C Fruhwald, D J Smiraglia, L J Rush, G P Robertson, X Gao, F A Wright, J D Feramisco, P Peltomaki, J C Lang, D E Schuller, L Yu, C D Bloomfield, M A Caligiuri, A Yates, R Nishikawa, H Su Huang, N J Petrelli, X Zhang, M S O'Dorisio, W A Held, W K Cavenee, and C Plass. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat Genet*, 24(2):132–138, Feb 2000.

[32] Susan E Cottrell, Jurgen Distler, Nancy S Goodman, Suzanne H Mooney, Antje Kluth, Alexander Olek, Ina Schwope, Reimo Tetzner, Heike Ziebarth, and Kurt Berlin. A real-time PCR assay for DNA-methylation using methylation-specific blockers. *Nucleic Acids Res*, 32 (1):e10, 2004.

[33] D R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[34] C Croux and P J Rousseeuw. Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistic*, 1:411–428, 1992.

[35] Marcia Cruz-Correa, Hengmi Cui, Francis M Giardiello, Neil R Powe, Linda Hylind, Angela Robinson, David F Hutcheon, David R Kafonek, Sheri Brandenburg, Yiqian Wu, Xiaobing He, and Andrew P Feinberg. Loss of imprinting of insulin growth factor II gene: a potential heritable biomarker for colon neoplasia predisposition. *Gastroenterology*, 126(4): 964–70, 2004.

[36] Hengmi Cui, Marcia Cruz-Correa, Francis M Giardiello, David F Hutcheon, David R Kafonek, Sheri Brandenburg, Yiqian Wu, Xiaobing He, Neil R Powe, and Andrew P Feinberg. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science*, 299(5613):1753–5, 2003.

[37] Hongyue Dai, Michael Meyer, Sergey Stepaniants, Michael Ziman, and Roland Stoughton. Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays. *Nucleic Acids Res*, 30(16):e86, Aug 2002.

[38] Degroot and Schervish. *Probability and Statistics*. Addison-Wesley Publishing, 2002.

[39] S M Dhanasekaran, T R Barrette, D Ghosh, R Shah, S Varambally, K Kurachi, K J Pienta, M A Rubin, and A M Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–6, 2001.

[40] S M Dong, G Traverso, C Johnson, L Geng, R Favis, K Boynton, K Hibi, S N Goodman, M D'Allessio, P Paty, S R Hamilton, D Sidransky, F Barany, B Levin, A Shuber, K W Kinzler, B Vogelstein, and J Jen. Detecting colorectal cancer in stool with the use of multiple genetic targets. *J Natl Cancer Inst*, 93(11):858–865, Jun 2001.

[41] R O Duda, P E Hart, and D G Stork. *Pattern Classification*. John Wiley & Sons, New York, 2000.

[42] S Dudoit, Y H Yang, M J Callow, and T P Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report, University of Berkeley, August 2000.

[43] Sandrine Dudoit, Mark J van der Laan, and Katherine S Pollard. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol*, 3(1):Article13, 2004.

[44] B P Durbin, J S Hardin, D M Hawkins, and D M Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:105–110, 2002.

[45] C A Eads, K D Danenberg, K Kawakami, L B Saltz, C Blake, D Shibata, P V Danenberg, and P W Laird. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res*, 28(8):E32, 2000.

[46] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–63, 2004.

[47] Melanie Ehrlich. Expression of various genes is controlled by DNA methylation during mammalian development. *J Cell Biochem*, 88(5): 899–910, 2003.

[48] D Eick, H J Fritz, and W Doerfler. Quantitative determination of 5-methylcytosine in DNA by reverse-phase high-performance liquid chromatography. *Anal Biochem*, 135(1):165–71, 1983.

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

[50] Kristof Engelen, Bart Naudts, Bart De Moor, and Kathleen Marchal. A calibration method for estimating absolute expression levels from microarray data. *Bioinformatics*, 22(10):1251–1258, May 2006.

[51] David Erickson, Dongqing Li, and Ulrich J Krull. Modeling of DNA hybridization kinetics for spatially resolved biochips. *Anal Biochem*, 317(2):186–200, Jun 2003.

[52] M Esteller, M Sanchez-Cespedes, R Rosell, D Sidransky, S B Baylin, and J G Herman. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. *Cancer Res*, 59(1):67–70, Jan 1999.

[53] M Esteller, M Toyota, M Sanchez-Cespedes, G Capella, M A Peinado, D N Watkins, J P Issa, D Sidransky, S B Baylin, and J G Herman. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. *Cancer Res*, 60(9): 2368–71, 2000.

[54] Manel Esteller. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21(35):5427–40, 2002.

[55] Manel Esteller. Relevance of DNA methylation in the management of cancer. *Lancet Oncol*, 4(6):351–8, 2003.

[56] M Evans, N Hastings, and B Peacock. *Statistical Distributions*. John Wiley & Sons, New York, 2000.

[57] W J Ewens and G R Grant. *Statistical Methods in Bioinformatics*. Springer, New York, 2002.

[58] E R Fearon and B Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, Jun 1990.

[59] K Fellenberg, N C Hauser, B Brors, A Neutzner, J D Hoheisel, and M Vingron. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A*, 98(19):10781–10786, Sep 2001.

[60] S P Fodor, R P Rava, X C Huang, A C Pease, C P Holmes, and C L Adams. Multiplexed biochemical assays with biological chips. *Nature*, 364(6437):555–556, Aug 1993.

[61] Mario F Fraga and Manel Esteller. DNA methylation: a profile of methods and applications. *Biotechniques*, 33(3):632, 634, 636–49, 2002.

[62] Mario F Fraga, Esther Uriol, L Borja Diego, Maria Berdasco, Manel Esteller, Maria Jesus Canal, and Roberto Rodriguez. High-performance capillary electrophoretic method for the quantification of 5-methyl 2'-deoxycytidine in genomic DNA: application to plant, animal and human cancer tissues. *Electrophoresis*, 23(11):1677–81, 2002.

[63] Simonetta Friso, Sang-Woon Choi, Gregory G Dolnikowski, and Jacob Selhub. A method to assess genomic DNA methylation using high-performance liquid chromatography/electrospray ionization mass spectrometry. *Anal Chem*, 74(17):4526–31, 2002.

[64] M Frommer, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proc Natl Acad Sci USA*, 89:1827–1831, 1992.

[65] T Gaasterland and S Bekiranov. Making the most of microarry data. *Nature Genetics*, 24:204–206, 2000.

[66] R S Gitan, H Shi, C M Chen, P S Yan, and T H Huang. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res.*, 12(1):158–164, January 2002.

[67] T Golub, D Slonim, P Tamayo, C Huard, M Gaasenbeek, J Mesirov, H Coller, M Loh, J Downing, M Caligiuri, C Bloomfield, and E Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[68] M L Gonzalgo, G Liang, C H 3rd Spruck, J M Zingg, W M 3rd Rideout, and P A Jones. Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed PCR. *Cancer Res*, 57(4):594–599, Feb 1997.

[69] W M Grady, A Rajput, J D Lutterbaugh, and S D Markowitz. Detection of aberrantly methylated hMLH1 promoter DNA in the serum of patients with microsatellite unstable colon cancer. *Cancer Res*, 61(3): 900–902, Feb 2001.

[70] W M Grady, J Willis, P J Guilford, A K Dunbier, T T Toro, H Lynch, G Wiesner, K Ferguson, C Eng, J G Park, S J Kim, and S Markowitz. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat Genet*, 26(1):16–7, 2000.

[71] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29–36, Apr 1982.

[72] J Hartung and B Epelt. *Multivariate Statistik*. R Oldenbourg Verlag München Wien, 1995.

[73] Brian Hendrich and Susan Tweedie. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet*, 19(5):269–77, 2003.

[74] J G Herman, C I Civin, J P Issa, M I Collector, S J Sharkis, and S B Baylin. Distinct patterns of inactivation of p15INK4B and p16INK4A characterize the major types of hematological malignancies. *Cancer Res*, 57(5):837–41, 1997.

[75] J G Herman, J R Graff, S Myohanen, B D Nelkin, and S B Baylin. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A*, 93(18):9821–6, 1996.

[76] J G Herman, A Umar, K Polyak, J R Graff, N Ahuja, J P Issa, S Markowitz, J K Willson, S R Hamilton, K W Kinzler, M F Kane, R D Kolodner, B Vogelstein, T A Kunkel, and S B Baylin. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A*, 95(12):6870–6875, Jun 1998.

[77] James G Herman. Hypermethylation pathways to colorectal cancer. Implications for prevention and detection. *Gastroenterol Clin North Am*, 31(4):945–958, Dec 2002.

[78] James G Herman and Stephen B Baylin. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med*, 349(21): 2042–2054, Nov 2003.

[79] Charles R Holst, Gerard J Nuovo, Manel Esteller, Karen Chew, Stephen B Baylin, James G Herman, and Thea D Tlsty. Methylation of p16(INK4a) promoters occurs in vivo in histologically normal human mammary epithelia. *Cancer Res*, 63(7):1596–1601, Apr 2003.

[80] Wolfgang Huber, Anja von Heydebreck, Holger Sultmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104, 2002.

[81] M Hubert, P J Rousseeuw, and S Verboven. A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60:101–111, 2002.

[82] K D Huynh, W Fischle, E Verdin, and V J Bardwell. BCoR, a novel corepressor involved in BCL-6 repression. *Genes Dev*, 14(14):1810–1823, Jul 2000.

[83] Daisuke Ichikawa, Hiroshi Koike, Hisashi Ikoma, Daito Ikoma, Nobuyuki Tani, Eigo Otsuji, Kazuya Kitamura, and Hisakazu Yamagishi. Detection of aberrant methylation as a tumor marker in serum of patients with gastric cancer. *Anticancer Res*, 24(4):2477–2481, Jul 2004.

[84] IHGSC. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[85] J P Issa. CpG-island methylation in aging and cancer. *Curr Top Microbiol Immunol*, 249:101–118, 2000.

[86] J P Issa. The epigenetics of colorectal cancer. *Ann N Y Acad Sci*, 910: 140–153, Jun 2000.

[87] S H Itzkowitz. Colonic polyps and polyposis syndromes. In L F M Felman and M Sleisenger, editors, *Gastrointestinal and Liver Disease: Pathophysiology/Diagnosis/Management*, volume II, pages 2175–2214. Saunders, Philadelphia, 2002.

[88] Jeremy R Jass, Vicki L J Whitehall, Joanne Young, and Barbara A Leggett. Emerging concepts in colorectal neoplasia. *Gastroenterology*, 123(3):862–876, Sep 2002.

[89] W Ji, R Hernandez, X Y Zhang, G Z Qu, A Frady, M Varela, and M Ehrlich. DNA demethylation and pericentromeric rearrangements of chromosome 1. *Mutat Res*, 379(1):33–41, 1997.

[90] P A Jones. DNA methylation and cancer. *Cancer Res*, 46(2):461–466, Feb 1986.

[91] P A Jones and D Takai. The role of DNA methylation in mammalian epigenetics. *Science*, 293(5532):1068–70, 2001.

[92] Peter A Jones and Stephen B Baylin. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 3(6):415–28, 2002.

[93] M F Kane, M Loda, G M Gaida, J Lipman, R Mishra, H Goldman, J M Jessup, and R Kolodner. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Res*, 57 (5):808–811, Mar 1997.

[94] Adam R Karpf and David A Jones. Reactivating the expression of methylation silenced genes in human cancer. *Oncogene*, 21(35):5496–503, 2002.

[95] A Kerjean, A Vieillefond, N Thiounn, M Sibony, M Jeanpierre, and P Jouannet. Bisulfite genomic sequencing of microdissected cells. *Nucleic Acids Res*, 29(21):E106–6, 2001.

[96] J Knight. When the chips are down. *Nature*, 410:860–861, April 2001.

[97] A G Jr Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–3, 1971.

[98] T Kohonen. *Self-Organizing Maps*. Springer, New York, 1995.

[99] Peter W Laird. The power and the promise of DNA methylation markers. *Nat Rev Cancer*, 3(4):253–66, 2003.

[100] S A Leon, B Shapiro, D M Sklaroff, and M J Yaros. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res*, 37(3): 646–50, 1977.

[101] Jorn Lewin, Armin O Schmitt, Peter Adorjan, Thomas Hildmann, and Christian Piepenbrock. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics*, 20(17):3005–12, 2004.

[102] Shuanfang Li, Stephen D Hursting, Barbara J Davis, John A McLachlan, and J Carl Barrett. Environmental exposure, DNA methylation, and gene regulation: lessons from diethylstilbesterol-induced cancers. *Ann N Y Acad Sci*, 983:161–169, Mar 2003.

[103] G Liang, K D Robertson, C Talmadge, J Sumegi, and P A Jones. The gene for a novel transmembrane protein containing epidermal growth factor and follistatin domains is frequently hypermethylated in human tumor cells. *Cancer Res*, 60(17):4907–4912, Sep 2000.

[104] G Liang, C E Salem, M C Yu, H D Nguyen, F A Gonzales, T T Nguyen, P W Nichols, and P A Jones. DNA methylation differences associated with tumor tissues identified by genome scanning analysis. *Genomics*, 53(3):260–268, Nov 1998.

[105] D J Lockhart and E A Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827–836, 2000.

[106] I Lonnstedt and T P Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Stat. Sinica*, 12:111–139, 2002.

[107] H P Lopuhaä and P J Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.

[108] Victoria V Lunyak, Gratien G Prefontaine, and Michael G Rosenfeld. REST and peace for the neuronal-specific transcriptional program. *Ann N Y Acad Sci*, 1014(NIL):110–20, 2004.

[109] K V Mardia, J T Kent, and J M Bibby. *Multivariate Analysis*. Academic Press Harcourt Brace and Company, 1979.

[110] Alessandra Marini, Alireza Mirmohammadsadegh, Sandeep Nambiar, Annett Gustrau, Thomas Ruzicka, and Ulrich R Hengge. Epigenetic inactivation of tumor suppressor genes in serum of patients with cutaneous melanoma. *J Invest Dermatol*, 126(2):422–431, Feb 2006.

[111] John W M Martens, Inko Nimmrich, Thomas Koenig, Maxime P Look, Nadia Harbeck, Fabian Model, Antje Kluth, Joan Bolt-de Vries, Anieta M Sieuwerts, Henk Portengen, Marion E Meijer-Van Gelder, Christian Piepenbrock, Alexander Olek, Heinz Hofler, Marion Kiechle, Jan G M Klijn, Manfred Schmitt, Sabine Maier, and John A Foekens. Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer Res*, 65(10):4101–4117, May 2005.

[112] R L Mason and J C Young. Interpretive features of a T^2 chart in multivariate SPC. *Quality Progress*, 33(4):84–89, April 2000.

[113] W Mendenhall and T Sincich. *Statistics for engineering and the sciences.* Prentice-Hall, New Jersey, 1995.

[114] O J Miller, W Schnedl, J Allen, and B F Erlanger. 5-Methylcytosine localised in mammalian constitutive heterochromatin. *Nature*, 251(5476): 636–7, 1974.

[115] Fabian Model, Peter Adorján, Alexander Olek, and Christian Piepenbrok. Feature selection for DNA methylation based cancer classification. *Bioinformatics*, 17(1):S157–S164, 2001.

[116] Fabian Model, Thomas Konig, Christian Piepenbrock, and Peter Adorjan. Statistical process control for large scale microarray experiments. *Bioinformatics*, 18 Suppl 1:155–163, 2002.

[117] Fabian Model, Neal Osborn, David Ahlquist, Robert Gruetzmann, Bela Molnar, Ferenc Sipos, Orsolya Galamb, Christian Pilarsky, Hans-Detlev Saeger, Zsolt Tulassay, Kari Hale, Suzanne Mooney, Joseph Lograsso, Peter Adorjan, Ralf Lesche, Andreas Dessauer, Joerg Kleiber, Baerbel Porstmann, Andrew Sledziewski, and Catherine Lofton-Day. Identification and validation of colorectal neoplasia-specific methylation markers for accurate classification of disease. *Mol Cancer Res*, 5 (2):153–63, 2007.

[118] Hannes M Muller, Michael Oberwalder, Heidi Fiegl, Maria Morandell, Georg Goebel, Matthias Zitt, Markus Muhlthaler, Dietmar Ofner, Raimund Margreiter, and Martin Widschwendter. Methylation changes in faecal DNA: a marker for colorectal cancer screening? *Lancet*, 363 (9417):1283–1285, Apr 2004.

[119] Y Nakahara, S Shintani, M Mihara, S Hino, and H Hamakawa. Detection of p16 promoter methylation in the serum of oral cancer patients. *Int J Oral Maxillofac Surg*, 35(4):362–365, Apr 2006.

[120] Hiroshi Nakayama, Kenji Hibi, Masumi Taguchi, Tsunenobu Takase, Taiji Yamazaki, Yasushi Kasai, Katsuki Ito, Seiji Akiyama, and Akimasa Nakao. Molecular detection of p16 promoter methylation in the serum of colorectal cancer patients. *Cancer Lett*, 188(1-2):115–119, Dec 2002.

[121] David Ng, Nalin Thakker, Connie M Corcoran, Dian Donnai, Rahat Perveen, Adele Schneider, Donald W Hadley, Cynthia Tifft, Liqun Zhang, Andrew O M Wilkie, Jasper J van der Smagt, Robert J Gorlin, Shawn M Burgess, Vivian J Bardwell, Graeme C M Black, and Leslie G Biesecker. Oculofaciocardiodental and Lenz microphthalmia syndromes result from distinct classes of mutations in BCOR. *Nat Genet*, 36(4):411–416, Apr 2004.

[122] Danh V Nguyen. Partial least squares dimension reduction for microarray gene expression data with a censored response. *Math Biosci*, 193 (1):119–137, Jan 2005.

[123] Danh V Nguyen and David M Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, Jan 2002.

[124] O Ogawa, D M Becroft, I M Morison, M R Eccles, J E Skeen, D C Mauger, and A E Reeve. Constitutional relaxation of insulin-like growth factor II gene imprinting associated with Wilms' tumour and gigantism. *Nat Genet*, 5(4):408–12, 1993.

[125] M Okano, D W Bell, D A Haber, and E Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–57, 1999.

[126] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, Michael G Walker, Drew Watson, Taesung Park, William Hiller, Edwin R Fisher, D Lawrence Wickerham, John Bryant, and Norman Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351(27):2817–2826, Dec 2004. Clinical Trial.

[127] B Panning and R Jaenisch. DNA hypomethylation can activate Xist expression and silence X-linked genes. *Genes Dev*, 10(16):1991–2002, 1996.

[128] Eric Phizicky, Philippe I H Bastiaens, Heng Zhu, Michael Snyder, and Stanley Fields. Protein analysis on a proteomic scale. *Nature*, 422 (6928):208–215, Mar 2003.

[129] M J Pilling and P W Seakins. *Reaction Kinetics*. Oxford University Press, New York, 1995.

[130] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes in C*. Cambridge University Press, 1992.

[131] John Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, Dec 2002.

[132] W Reik, W Dean, and J Walter. Epigenetic reprogramming in mammalian development. *Science*, 293(5532):1089–1093, Aug 2001.

[133] W Reik and J Walter. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, 2(1):21–32, 2001.

[134] D M Rocke and B Durbin. A model for measurement error for gene expression arrays. *J Comput Biol*, 8(6):557–569, 2001.

[135] David M Rocke and Blythe Durbin. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, 19(8):966–972, May 2003.

[136] Tamas Rujan, Reinhold Wasserkort, and Armin Schmitt. Integrated primer design strategy for PCR amplification of bisulphite treated DNA. In *Annual International Conference on Intellitent Systems for Molecular Biology*, volume 10, 2002.

[137] Laura J Rush and Christoph Plass. Restriction landmark genomic scanning for DNA methylation in cancer: past, present, and future applications. *Anal Biochem*, 307(2):191–201, 2002.

[138] Silvia Sabbioni, Elena Miotto, Angelo Veronese, Elisa Sattin, Laura Gramantieri, Luigi Bolondi, George A Calin, Roberta Gafa, Giovanni Lanza, Giuliano Carli, Eros Ferrazzi, Carlo Feo, Alberto Liboni, Sergio

Gullini, and Massimo Negrini. Multigene methylation analysis of gastrointestinal tumors: TPEF emerges as a frequent tumor-specific aberrantly methylated marker that can be detected in peripheral blood. *Mol Diagn*, 7(3-4):201–207, 2003.

[139] Philipp Schatz. *Entwicklung eines Verfahrens zur Hochdurchsatzanalyse von DNA-Methylierung*. PhD thesis, Universität des Saarlandes, 2005.

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

[141] J Singer-Sam, J M LeBon, R L Tanguay, and A D Riggs. A quantitative HpaII-PCR assay to measure methylation of DNA from a small number of cells. *Nucleic Acids Res*, 18(3):687, 1990.

[142] Gordon K Smyth, Yee Hwa Yang, and Terry Speed. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol*, 224:111–136, 2003.

[143] Vincent Sollars, Xiangyi Lu, Li Xiao, Xiaoyan Wang, Mark D Garfinkel, and Douglas M Ruden. Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nat Genet*, 33(1):70–74, Jan 2003.

[144] Christos Sotiriou, Soek-Ying Neo, Lisa M McShane, Edward L Korn, Philip M Long, Amir Jazaeri, Philippe Martiat, Steve B Fox, Adrian L Harris, and Edison T Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–8, 2003.

[145] K Specht, T Richter, U Muller, A Walch, M Werner, and H Hofler. Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. *Am J Pathol*, 158 (2):419–429, Feb 2001.

[146] D Stekel. *Microarray Bioinformatics*. Cambridge University Press, Cambridge, 2003.

[147] Arshia Tabassum, Fatima Khwaja, and Daniel Djakiew. The p75(NTR) tumor suppressor induces caspase-mediated apoptosis in bladder tumor cells. *Int J Cancer*, 105(1):47–52, May 2003.

[148] T M Therneau. *Modeling survival data*. Springer, New York, 2000.

[149] Jörg Tost. *Tools for the elucidation of gene function and regulation - molecular haplotyping and epigenotyping*. PhD thesis, Universität des Saarlandes, 2003.

[150] M Toyota, C Ho, N Ahuja, K W Jair, Q Li, M Ohe-Toyota, S B Baylin, and J P Issa. Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res*, 59(10):2307–12, 1999.

[151] M Toyota, C Ho, N Ahuja, K W Jair, Q Li, M Ohe-Toyota, S B Baylin, and J P Issa. Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res*, 59(10):2307–2312, May 1999.

[152] G C Tseng, M K Oh, L Rohlin, J C Liao, and W H Wong. Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29(12):2549–2557, 2001.

[153] Masako Ueno, Minoru Toyota, Kimishige Akino, Hiromu Suzuki, Masanobu Kusano, Ayumi Satoh, Hiroaki Mita, Yasushi Sasaki, Masanori Nojima, Kazuyoshi Yanagihara, Yuji Hinoda, Takashi Tokino, and Kohzoh Imai. Aberrant methylation and histone deacetylation associated with silencing of SLC5A8 in gastric cancer. *Tumour Biol*, 25(3):134–140, May 2004.

[154] Paola Ulivi, Wainer Zoli, Daniele Calistri, Francesco Fabbri, Anna Tesei, Marco Rosetti, Marta Mengozzi, and Dino Amadori. p16INK4A and CDH13 hypermethylation in tumor and serum of non-small cell lung cancer patients. *J Cell Physiol*, 206(3):611–615, Mar 2006.

[155] Henning Usadel, Jan Brabender, Kathy D Danenberg, Carmen Jeronimo, Susan Harden, James Engles, Peter V Danenberg, Stephen Yang, and David Sidransky. Quantitative adenomatous polyposis coli promoter methylation analysis in tumor tissue, serum, and plasma DNA of patients with lung cancer. *Cancer Res*, 62(2):371–375, Jan 2002.

[156] M T Valenzuela, R Galisteo, A Zuluaga, M Villalobos, M I Nunez, F J Oliver, and J M Ruiz de Almodovar. Assessing the use of p16(INK4a) promoter gene methylation in serum for detection of bladder cancer. *Eur Urol*, 42(6):622–628, Dec 2002.

[157] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol*, 3(1):Article14, 2004.

[158] V Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.

[159] V Vasioukhin, P Anker, P Maurice, J Lyautey, C Lederrey, and M Stroun. Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br J Haematol*, 86(4):774–9, 1994.

[160] W N Venables and B D Ripley. *Modern Applied Statistics with S-PLUS.* Springer-Verlag New York, 1999.

[161] B Vogelstein, E R Fearon, S R Hamilton, S E Kern, A C Preisinger, M Leppert, Y Nakamura, R White, A M Smits, and J L Bos. Genetic alterations during colorectal-tumor development. *N Engl J Med*, 319 (9):525–532, Sep 1988.

[162] P A Wade. Methyl CpG-binding proteins and transcriptional repression. *Bioessays*, 23(12):1131–7, 2001.

[163] R Y Wang, C W Gehrke, and M Ehrlich. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res*, 8(20):4777–90, 1980.

[164] Robert A Waterland and Randy L Jirtle. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol*, 23(15):5293–5300, Aug 2003.

[165] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[166] J B Welsh, P P Zarrinkar, L M Sapinoso, S G Kern, C A Behling, B J Monk, D J Lockhart, R A Burger, and G M Hampton. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci U S A*, 98(3):1176–81, 2001.

[167] P H Westfall and S S Young. *Resampling-based multiple testing: examples and methods for p-value adjustment.* John Wiley & Sons, New York, 1993.

[168] J Weston, S Mukherjee, O Chapelle, M Pontil, T Poggio, and V Vapnik. Feature selection for svms. In *Advances in neural information processing systems*, volume 13, Cambridge, MA, 2001. MIT Press. in press.

[169] J G Wetmur. Hybridization and renaturation kinetics of nucleic acids. *Annu Rev Biophys Bioeng*, 5:337–361, 1976.

[170] Y Yamanishi, J-P Vert, A Nakaya, and M Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19 Suppl 1:323–330, 2003.

[171] P S Yan, M R Perry, D E Laux, A L Asare, C W Caldwell, and T H Huang. CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin Cancer Res*, 6(4):1432–8, 2000.

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

[173] J A Yoder, C P Walsh, and T H Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, 13(8):335–40, 1997.

[174] A Zien, T Aigner, R Zimmer, and T Lengauer. Centralization: A new method for the normalization of gene expression data. *Bioinformatics*, 17:S323–S331, 2001.

[175] Hongzhi Zou, Baoming Yu, Ren Zhao, Zhiwei Wang, Hui Cang, Donghua Li, Guoguang Feng, and Jing Yi. Detection of aberrant p16 methylation in the serum of colorectal cancer patients. *Zhonghua Yu Fang Yi Xue Za Zhi*, 36(7):499–501, Dec 2002.

Appendix A

Datasets

A.1 Methylation estimation

Calibration The first dataset we have used in Chapter 2 is a calibration experiment with a total of 48 microarrays that were hybridized with various mixtures of artificially methylated and unmethylated DNA. The microarrays contained 476 CG and TG detection oligos from 54 oligo families covering 8 different genes. The following proportions of methylated DNA in a background of unmethylated DNA were tested: 0% (10 chips), 25% (8 chips), 50% (11 chips), 75% (8 chips) and 100% (11 chips).

Cross hybridization The second dataset is from a microarray experiment conducted to quantify the extent of cross hybridization. The same calibration microarray as described above with its 476 oligos covering 8 different genes was used. Each microarray was hybridized with either fully methylated or fully unmethylated fragments. Exactly one of the eight amplificates was labeled with the fluorecent dye CY3 and all remaining seven amplificates with the fluorecent dye CY5. For each of these 16 combinations (2 methylation states times 8 different labeling states) about 6 microarray replicates were hybridized. We have used this dataset to estimate the equilibrium constant matrices of this particular microarray design.

A.2 Quality control

In Chapter 3 we use data from three microarray studies. In each study the methylation status of about 200 different CpG dinucleotide positions from promoters, intronic and coding sequences of 64 genes was measured.

Temperature Control Our first set of 207 chips comes from a control experiment where PCR amplificates of DNA from peripheral blood of 15 patients diagnosed with ALL or AML was hybridized at 4 different temperatures (38°C,42°C,44°C,46°C). We have used this dataset to prove that our method can reliably detect shifts in experimental conditions.

Lymphoma The second dataset with an overall number of 647 chips comes from a study where the methylation status of different subtypes of non-Hodgkin lymphomas from 68 patients was analyzed. All chips underwent a visual quality control, resulting in quality classification as "good" (proper spots and low background), "acceptable" (no obvious defects but uneven spots, high background or weak hybridization signals) and "unacceptable" (obvious defects). We have used this dataset to identify different types of outliers and showed how our methods detect them.

In addition we have simulated an accidental exchange of oligo probes during slide fabrication in order to demonstrate that such an effect can be detected by our method. The exchange was simulated in silico by permuting 12 randomly selected CpG positions on 200 of the chips (corresponding to an accidental rotation of a 24 well oligo supply plate during preparation for spotting).

ALL/AML Finally we have showed data from a second study on ALL and AML, containing 433 chips from 74 different patients. During the course of this study 46 oligomeres ran out of stock and had to be re-synthesized. As it turned out, some of them showed a significant change in hybridization behavior, due to synthesis quality problems. We have demonstrate how our algorithm successfully detected this systematic change in experimental conditions.

A.3 Class prediction

The dataset [1] consists of cell lines and primary tissue obtained from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). A total of 17 ALL and 8 AML samples were included. The methylation status of these samples was evaluated at 81 CpG dinucleotide positions located in CpG rich regions of the promoters, intronic and coding sequences of 11 genes. These were randomly selected from a panel of genes representing different pathways associated with tumour genesis. Two of the 11 selected genes are located on the X-chromosome.

Gene Name	Gene Description	Chromosome Location	Discovery Method	Score	Amplicon Location
DNAJC5	DnaJ (Hsp40) homolog, subfamily C, member 5γ γ cysteine string protein; γ-CSP	2p23.3	AP-PCR	2	intron 1
ALX4	Homeobox Protein Aristaless-Like 4	11p11.2	AP-PCR	5	intron 1
Q8WWL2	SPIR-2 Protein	16q24.3	AP-PCR	2	exon 4
DUX2	Double Homeobox 2	10q26.3	MCA	4	exon 1
KCTD12	Potassium channel tetramerisation domain containing 12; chromosome 13 open reading frame 2	13q22.3	AP-PCR	4	promoter
HOXB3	Homeobox Protein HOX-B3 (HOX-2G) (HOX-2.7)	17q21.32	AP-PCR	2	promoter
ZDHHC22	Zinc finger, DHHC domain containing 22 chromosome 14 open reading frame 59	14q24.3	AP-PCR	4	promoter
PQLC1	PQ loop repeat containing 1	18q23	MCA	1	???
FCGR2A	Low-affinity immunoglobulin gamma FC-region receptor II-A precursor	1q23.3	AP-PCR	1	upstream
ENSESTG020896	EST only	16p13.2	AP-PCR	3	promoter
TAF11	Transcription initiation factor TFIID 28 KDA subunit	6p21.31	AP-PCR	3	promoter
TMEFF2 (HPP1)	Transmembrane protein with EGF-like and two follistatin-like domains 2	2q32.3	AP-PCR	3	exon 1
Onecut2	One cut domain family member 2 (onecut-2 transcription factor) (OC-2)	18q21.31	AP-PCR	4	intron 1
SLITRK1	Slit and trk like 1 protein; slit and trk like gene 1	13q31.2	MCA	1	exon 1
NGFR	Tumor necrosis factor receptor superfamily member 16 precursor	17q21.33	AP-PCR	3	intron 1
GENSCAN037834	Prediction only	11q24.3	AP-PCR	1	exon 1
ADCY9	Homo sapiens adenylate cyclase 9	16p13.3	AP-PCR	3	promoter
Q9UPN4	5-azacytidine-induced protein 1	17q25.3	MCA	2	exon 9
SLC32A1	Solute carrier family 32	20q11.23	AP-PCR	3	exon 1
C7orf20	Chromosome 7 open reading frame 20	7p22.3	MCA	2	exon 1
PCDH17	Protocadherin 17	13q21.1	MCA	4	promoter
NPBWR1	Neuropeptides B/W receptor type 1 (G protein-coupled receptor 7)	8q11.23	AP-PCR	3	intron 1
RNF4	Ring Finger Protein 4	4p16.3	AP-PCR	4	promoter
DLX5	Homeobox protein DLX-5	7q21.3	AP-PCR	3	promoter
BCOR	BCL6 co-repressor	Xp11.4	AP-PCR	3	intron 1
SIX6	Sine oculis homeobox homolog 6	14q23.1	AP-PCR	4	intron 1
BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	3q27.3	AP-PCR	6	intron 1
Q9P1Z9	Homo sapiens mRNA for KIAA1529 protein	9q22.33	AP-PCR	2	intron 3
SMAD7	Mothers against decapentaplegic homolog 7	18q21.1	AP-PCR	4	intron 1
EYA4	Eyes Absent Homolog 4	6q23	Literature	N/A	promoter
MSH6	MutS homolog (MSH6)	2p16	Literature	N/A	promoter
APC	Adenomatosis polyposis coli (APC)	5q21-q22	Literature	N/A	promoter
CD44	CD44 antigen	11p13	Literature	N/A	promoter
CSPG2	Chondroitin sulfate proteoglycan 2 (versican)	5q14.3	Literature	N/A	promoter
CDH13	H-cadherin	16q24.2-q24.3	Literature	N/A	promoter
GSK3B	GSK3 Beta	3q13.3	Literature	N/A	promoter
TMEFF2 (HPP1)	Transmembrane protein with EGF-like and two follistatin-like domains 2	2q32.3	Literature	N/A	promoter
AR	Androgen receptor	Xq11.2-q12	Literature	N/A	promoter
TUSC3 (N33)	Candidate tumor suppressor 8p22	8p22	Literature	N/A	promoter
TGFBR2	Transforming growth factor beta receptor II	3p22	Literature	N/A	promoter
TP73	Tumor protein p73	1p36.3	Literature	N/A	promoter
CAV1	Caveolin-1	7q31.1	Literature	N/A	promoter

Table A.1: Sequences selected from genome-wide discovery or from literature for validation on oligonucleotide microarray.

A.4 Marker selection

The total sample set included 358 patient DNAs and two control DNAs. The patient DNAs were extracted from 29 normal colon samples, 31 inflammatory bowel disease (IBD), 55 colon polyps (45 polyps <1cm, 10 polyps ≥1cm), 89 colorectal cancers (30 Dukes A/B, 56 Dukes C/D, 1 unknown, 2 high grade polyps≥1cm), 116 non-colonic cancer samples from liver (9), bile duct (10), pancreas (10), lung (squamous and adenocarcinoma) (38), breast (28), prostate (5), esophagus (6), stomach (10), PBL (14) and normal tissue from sites other than colon: esophagus mucosa (7), gastric mucosa (7), liver (10). Additionally one control sample of unmethylated human DNA (Molecular Staging), and one control sample of enzymatically methylated DNA (SssI, NEB) was included.

The microarray was performed as described in Chapter 2 with oligonucleotides covering regions of 43 discovery and literature-derived genes and 2 control genes. For the discovery derived genes primer pairs and oligonucleotides were designed around the identified differentially methylated sequence whenever possible. Multiple primer pairs and oligonucleotides were designed for some genes for a total of 54 amplicons and a total of 248 oligonucleotide pairs. Each oligonucleotide contained 2-3 CpG sites. Additionally 8 negative control oligonucleotides with random sequences were included to facilitate estimation of unspecific background hybridization. Amplicons for all discovery genes, candidate genes and control genes used in the combined array are shown in Table A.1.

Appendix B

List of symbols

N^{DNA}	Number of DNA strands in extracted sample
N_p^{DNA+}, N_p^{DNA-}	Number of DNA strands in extracted sample that are methylated or unmethylated at CpG position p
N_p^{PCR+}, N_p^{PCR-}	Number of PCR fragments that originate from a DNA strand methylated or unmethylated at CpG position p
$\mathcal{P}$	Set of CpG positions
$\mathcal{Q}$	Set of oligomeres
$\mathcal{Q}_p$	Set of oligomeres querying the same CpG position p
$\mathcal{N}$	Set of negative control oligomeres
$\mathcal{R}$	Set of amplificates
$\mathcal{C}$	Set of microarrays
p	Specific CpG position $p \in \mathcal{P}$ or position index $p \in \{1, ..., \lvert\mathcal{P}\rvert\}$
q	Specific CG and/or TG oligomere $q \in \mathcal{Q}$ or oligomere index $q \in \{1, ..., \lvert\mathcal{Q}\rvert\}$
r	Specific amplificate $r \in \mathcal{R}$ or amplificate index $r \in \{1, ..., \lvert\mathcal{R}\rvert\}$
n_r	Number of spot replications per oligo
n_p	Number of CpG positions queried by oligos of a particular microarray layout
n_q	Number of oligos or oligo pairs on a perticular microarray layout
n_c	Number of microarrays in a dataset
n_s	Number of biological samples in a dataset

$O_{q,i}$	Observed hybridization intensity of oligo q, repetition i ($i \in \{1, ..., n_r\}$)
I_q	Theoretical / expected hybridization intensity of oligo q, $I_q \propto E[O_q]$
I^{BG}	Oligo independent base hybridization intensity
$\mathcal{O}_q$	Set of observed hybridization intensities from oligo q, $\mathcal{O}_q = \{O_{q,1}, ..., O_{q,n_r}\}$
f_c	Intensity scaling factor of chip c
σ_η	Standard deviation of multiplicative hybridization noise
σ_ϵ	Standard deviation of additive background hybridization noise
S_p	Methylation score at CpG position p
m_p	Expected proportion of DNA strands that are methylated in a pure tissue at CpG position p
d_{cq}	Methylation score measured on microarray c and CG-TG oligo pair q
d_{sp}	Methylation score measured on sample s and CpG position p
$\boldsymbol{d}_i$	Methylation profile of chip or sample i, $\boldsymbol{d}_i = (d_{i1}, ..., d_{in_q})'$
$\boldsymbol{a}$	Vector of total amplificate concentrations
$\boldsymbol{a}^+$, $\boldsymbol{a}^-$	Vector of methylated or unmethylated amplificate concentrations
k_{qr}^f	Duplex formation rate between oligo q and amplificate r
$k_{qr}^{r,+}$,$k_{qr}^{r,-}$	Duplex deformation rate between oligo q and amplificate r for methylated or unmethylated amplificates
K^+, K^-	Equilibrium constant matrix for methylated or unmethylated amplificates
$max(\cdot, \cdot)$	Maximum
$min(\cdot, \cdot)$	Minimum
$med(\cdot)$	Median
$mad(\cdot)$	Median absolute deviation
$mean(\cdot)$	Arithmetic mean
$Var[\cdot]$	Variance
$E[\cdot]$	Expectation
$\lvert \cdot \rvert$	Cardinality of a set
$T[\cdot]$	Generalized Log transformation
c	An arbitrary constant

Previously published work

The content of chapter 3 was published in:

F. Model, T. König, C. Piepenbrock and P. Adorjan, "Statistical process control for large scale microarray experiments", *Bioinformatics*, 18 Suppl 1, S155-63, 2002

The content of chapter 4 was published in:

F. Model, P. Adorjan, A. Olek and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification", *Bioinformatics*, 17 Suppl 1, S157-64, 2001

The content of chapter 5 was published in:

F. Model, N. Osborn, D. Ahlquist, R. Gruetzmann, B. Molnar, F. Sipos, O. Galamb, C. Pilarsky, H. Saeger, Z. Tulassay, K. Hale, S. Mooney, J. Lograsso, P. Adorjan, R. Lesche, A. Dessauer, J. Kleiber, B. Porstmann, A. Sledziewski and C. Lofton-Day, "Identification and validation of colorectal neoplasia-specific methylation markers for accurate classification of disease", *Molecular Cancer Research*, 5(2), 153-63, 2007

Other publications on DNA methylation microarrays:

P. Adorjan, J. Distler, E. Lipscher, F. Model, J. Müller, C. Pelet, A. Braun, A. Florl, D. Gütig, G. Grabs, A. Howe, M. Kursar, R. Lesche, E. Leu, A. Lewin, S. Maier, V. Müller, T. Otto, C. Scholz, W. Schulz, H. Seifert, I. Schwope, H. Ziebarth, K. Berlin, C. Piepenbrock and A. Olek, "Tumour class prediction and discovery by microarray-based DNA methylation analysis", *Nucleic Acids Res.*, 30(5), e21, 2002

W. Enard W, A. Fassbender, F. Model, P. Adorjan, S. Paabo and A. Olek, "Differences in DNA methylation patterns between humans and chimpanzees", *Curr Biol.*, 14(4), R148-9, 2004

J.W. Martens, I. Nimmrich, T. Koenig, M.P. Look, N. Harbeck, F. Model, A. Kluth, J. Bolt-de Vries, A.M. Sieuwerts, H. Portengen, M.E. Meijer-Van Gelder, C. Piepenbrock, A. Olek, H. Hofler, M. Kiechle, J.G. Klijn, M. Schmitt, S. Maier and J.A. Foekens, "Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer", *Cancer Res.*, 65(10), 4101-17, 2005

Acknowledgements

First I want to thank all the great people at Epigenomics Berlin, especially Péter Adorján, Tamas Rujan, Jürgen Distler, Cécile Pelet and Evelyne Becker, for introducing me to the exciting world of molecular biology and DNA methylation. Special thanks go to my fellow PhD students Thomas König, Joern Lewin, Philipp Schatz, Anne Fassbender, Claudia Ivascu and Reimo Tetzner for their scientific and moral support over the last years.

I thank Professor Ulrich Kockelkorn for critically reading this manuscript and supervising this thesis. I would also like to thank the other members of the STAT group at TU Berlin Juergen Schweiger, Nicole Krämer, Malte Kuss and Joerg Betzin for the many enlightening discussions.

Finally I have to thank all my colleagues and friends at Epigenomics Seattle, especially Cathy Lofton-Day, Theo deVos, Volker Liebenberg, Robert Day and Andrew Sledziewski, who taught me the basics of biology and medicine and always made me feel like at home.