METHOD DEVELOPMENT FOR QUANTITATIVE METHYLATION ANALYSIS BY DIRECT BISULFITE SEQUENCING, RAW DATA PROCESSING AND ANALYSIS OF THE HUMAN EPIGENOME PROJECT

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie-, Pharmazie-, Bio- und Naturwissenschaften
der Universität des Saarlandes

von
Jörn Lewin

Saarbrücken
12.08.2007

Tag des Kolloquiums:	23.01.2008
Dekan:	Prof. Dr. Uli Müller
Berichterstatter:	Prof. Dr. Jörn Walter Prof. Dr. Thomas Lengauer

Abstract

Epigenetic deals with flexible biochemical information layers that lie on top of the relatively stable DNA sequence and are involved in control of the structure and functionality of the DNA. One of the most easily analyzed epigenetic layers is DNA methylation.
The first part of this work describes the development and assessment of new algorithms enabling methylation quantification by interpretation of sequencing electropherogram data from direct bisulfite sequencing. It is shown that the use of the algorithms on data from PCR products is a suitable replacement for subcloning and sequencing of about 10 subclones - a prerequisite for efficient high throughput DNA methylation studies by direct sequencing, such as carried out in the Human Epigenome Project (HEP).
The second part demonstrates the possibility to compensate for artifacts and signal echos in raw data from direct bisulfite sequencing using a deconvolution algorithm.
In the third part of this thesis the data of the HEP is analyzed. The HEP is the first large-scale project providing high resolution methylation data in 12 healthy human tissue types on 3 chromosomes analyzed, with a view to answering biological questions. It is shown that differential methylation between healthy tissues is a common phenomenon - especially in conserved non coding sequences, how CpG density and proximity to functional genomic sites influence the methylation profile and in how far CpGs tend to be organized in co-methylated blocks.

Many parts of this work, which was originally planned as cumulative thesis, were previously published in articles (Lewin *et al.*, 2004; Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004) and will overlap with their content.

Zusammenfassung

Epigenetik beschäftigt sich mit dynamischen biochemischen Informationsebenen, welche die im Vergleich dazu relativ stabile DNS beeinflussen und eine Rolle bei der Kontrolle der Struktur und Funktion der DNS spielt. DNS Methylierung ist eine der am besten untersuchbaren epigenetischen Ebenen.
Diese Arbeit beschreibt im ersten Teil die Entwicklung eines neuen Algorithmus, der quantitative Methylierungsmessung auf der Grundlage von Elektropherogramm Daten aus der direkten Sequenzierung von PCR Produkten von Bisulfit behandelter DNS ermöglicht. Es wird gezeigt, daß die Verwendung des Algorithmus mit Daten von PCR Produkten eine brauchbare Alternative zu Subklonierung und Sequenzierung von ca. zehn Subklonen ist und damit effiziente DNS Methylierungsstudien durch Hochdurchsatzsequenzierung ermöglicht.
Der zweite Teil der Arbeit beschreibt die Möglichkeit mit Hilfe eines Dekonvolutions Algorithmus Artefakte und Signal Echos in Sequenzierungsrohdaten zu kompensieren.
Der dritte Teil behandelt die Analyse und die biologische Erkenntnisse aus den Daten des HEP, dem ersten hochauflösenden Methylierungsdatensatz für drei Chromosomen in zwölf Geweben. Es wird gezeigt, daß differentielle Methylierung zwischen gesunden Geweben weit verbreitet ist, welchen Einfluß CpG Dichte und Nachbarschaft zu funktionalen genomischen Bereichen auf das DNS Methylierungsprofil haben und in wieweit benachbarte CpGs dazu tendieren, sich in comethylierten Einheiten zu organisieren.

Viele Teile dieser Arbeit, die ursprünglich als kumulative Arbeit geplant war, haben einen hohen Überlapp mit bestehenden Veröffentlichungen (Lewin *et al.*, 2004; Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004).

Ausführliche Zusammenfassung

Epigenetik ist ein wichtiger Bereich der molekularen Genetik, der sich mit dynamischen biochemischen Informationsebenen beschäftigt, welche die im Vergleich dazu relativ stabile DNS beeinflussen. DNS Methylierung, Chromatinmethylierung, - acetylierung und -phosphorilierung spielen eine Rolle bei der Kontrolle des Status, der Struktur und Funktion der DNS. DNS Methylierung ist aufgrund ihrer Stabilität und Zugänglichkeit eine der am besten untersuchbaren epigenetischen Ebenen, die mit vielen regulativen Funktionen im Genom assoziiert wird.

Der erste Teil dieser Arbeit behandelt einen neuen Algorithmus, der quantitative Methylierungsmessung in DNS auf der Grundlage von Elektropherogramm Daten aus der direkten Sequenzierung von PCR Produkten von mit Bisulfit behandelter DNS ermöglicht. Die unter Verwendung des Algorithmus erzielten quantitativen Ergebnisse werden mit verschiedenen bekannten Testsytemen bewertet: Daten aus DNS Mixturen mit bekannter Methylierung, Daten von Gemischen bekannter Subklone von PCR Fragmenten und Methylierungsmessungen von anderen Platformen. Es wird gezeigt, daß die Verwendung des Algorithmus mit Daten von PCR Produkten eine brauchbare Alternative zu Subklonierung und Sequenzierung von ca. zehn Subklonen ist und damit effiziente DNS Methylierungsstudien durch Hochdurchsatzsequenzierung wie zum Beispiel das Humane Genom Projekt (HEP) ermöglicht.

Der zweite Teil der Arbeit beschreibt und evaluiert die Möglichkeit mit Hilfe von Dekonvolution Artefakt und Signal Echos in Sequenzierungsrohdaten zu kompensieren, die aufgrund der heterogenen Zusammensetzung von PCR Produkten aus Bisulfit behandelter DNS auftreten können. Es wird an realen Beispielen und generierten Modelldaten gezeigt, daß durch die Verwendung eines solchen Algorithmus eine Verbesserung von Daten mit Artefakten erzielt werden kann.

Der dritte Teil behandelt die Datenanalyse und die biologische Erkenntnisse des HEP, dem ersten hochauflösenden Methylierungsdatensatz für drei Chromosomen in 12 Geweben. Die Methylierung wird - basierend auf Annotationen der Gewebeproben und biologischem Kontext der genomischen Koordinaten der Messung - auf verschiedene Fragestellungen hin untersucht. Es wird unter anderem gezeigt, daß differentielle Methylierung zwischen gesunden Geweben weit verbreitet ist, vor allem in nicht kodierenden evolutionär kon-

servierten Bereichen. Es wird untersucht welchen Einfluß CpG Dichte und Nachbarschaft zu funktionalen genomischen Bereichen auf das DNS Methylierungsprofil haben, und gezeigt, daß die durchschnittliche Methylierung um Transkriptionsstartstellen (TSS) herum ein klares fast symmetrisches Profil mit einem Minimum an der Stelle der TSS zeigt. Der Zusammenhang des Einflusses von Faktoren wie dem Alter auf lokale und globale Methylierung wird untersucht, mit dem Ergebnis, daß im zugrundelegenden Datensatz keine systematischen Tendenzen gefunden werden können. Es wird charakterisiert, in wieweit benachbarte CpGs dazu tendieren, sich in comethylierten Einheiten zu organisieren.

Viele Teile dieser Arbeit, die ursprünglich als kumulative Arbeit geplant war, haben einen hohen Überlapp mit bestehenden Veröffentlichungen (Lewin *et al.*, 2004; Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004).

Contents

Chapter 1

Introduction

This chapter provides an overview of DNA methylation and methods to detect and measure it with a focus on DNA sequencing methods, and relevant background about four color trace files as potential information source for quantitative DNA methylation measurements. The biology and basic techniques are explained to allow non biologists access to the themes of this thesis. I will explain why the research of the epigenome is of high importance and why DNA methylation is the epigenetic layer of choice to be assessed. I will show that in order to understand genetics, proteomics, cell differentiation, individual cell behavior and even cancerogenesis it is useful to understand the epigenome.

1.1 From Genome to Epigenome - understanding the cells individuality

Currently the complete DNA sequences of the human and many other genomes are sequenced (Consortium, 2004; Lander *et al.*, 2001) and available in databases (Curwen *et al.*, 2004). This source of information is the result of many large international sequencing studies. Based on this and additional experimental data many genetic functions are known about the genomes whereby this information is mainly based on the four letter code of the DNA: protein coding regions, variations that are connected to heritable phenotypes, diseases or health risks, recognition patterns for protein binding sites and more (Lewin, 2003).

Though the four letter code sequence provides a lot of information about a species or an individual organism, it does not explain how single cells or differentiated tissues run and control their individual cellular programs using specific subsets of the genomic information. In order to understand cellular programs, genetic information must be investigated within its context using a biological characteristic about which even less is known, the **epigenome**. The epigenome carries further information based on **chemical modifications of DNA and chromatin** that are associated with the **three dimensional structure** of the chromatin, its functionality, the activity of the **transcriptome** and therefore the composition of the **proteome** (see Fig. 1.1). The fact that the genomic DNA information layer is static in the process of a dividing cell developing to a complex individual suggests that the increasing intricacy of the organism must be controlled by biochemical processes and patterns outside the four letter code, probably by epigenome and proteome interaction.

The epigenetic information layer is heritable (Morgan *et al.*, 1999) but is also capable of quickly changing due to the internal programs of cells and/or influences by their environment. The epigenome shows different patterns that do not define or reflect the ”information” stored in the genome of an organism in all its cells, but reflects which part of that information might be inactive or active and thus of relevance in a certain cell, tissue and developmental stage.

The functionality of the epigenome has not been studied to the same extent as the genome. Our basic knowledge about its functionality has been gained mainly from examples covering only parts of it. Those examples al-

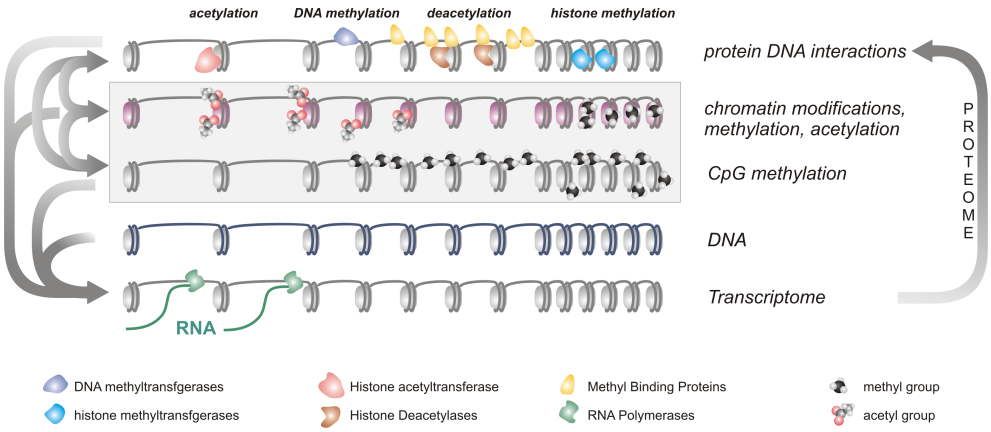

Figure 1.1: **Epigenomics: A simple model of the framework of Genome, Epigenome, Transcriptome and Proteome.** The interaction of DNA methylation, histone acetylation and methylation, chromatin density and structure is influencing the transcription processes in the cell. The local states of the epigenome are correlated with the active coding parts of the genome that are the blueprints for the proteins available in the cell: Histones in stretches of DNA with unmethylated CpGs are acetylated, the chromatin density is relaxed. DNA stretches in opened chromatin and unmethylated promoter regions are associated with transcription. Methylation of CpGs introduced by DNA methyl transferases is followed by binding of methyl binding proteins that allow the docking of protein complexes deacetylating the DNA. The deacetylated DNA leads to higher chromatin density, chromatin becomes methylated. Compacted chromatin and methylated DNA in promoter regions are associated with transcription repression (Fuks, 2005). Within this active framework the DNA itself is a passive, static and unchanged information layer identical in different cells, but controlled differently by the active framework.

ready show that epigenetics has a great influence on inheritance, interacts with the genome, provides a lot of medically relevant information (Yoo & Jones, 2006) - mainly for oncology (Jones, 2002) - and seems to play a major role in the control of transcription (Fuks, 2005). **In order to understand the functionality of cells it is necessary to gain more knowledge of the epigenome.**

1.2 Different layers of epigenetic modifications

The epigenome of vertebrates is mainly defined by chemical modifications of DNA and histones (methylation, acetylation, phosphorylation), which have an influence on the three dimensional structure and density of the chromatin and indirectly influence transcription activity. DNA methylation is reported to directly interact with histone acetylation and histone methylation (see Fig. 1.2). DNA methylation does not alter the code but provides additional information, is stable for years in frozen or prepared tissue samples and purified DNA and is also easily assessed with different methods (Dahl & Guldberg, 2003; Siegmund & Laird, 2002), see also section 1.4. Therefore DNA methylation is the information layer of choice for studying the epigenome - independent of the question of if it is dominating the control of other layers and biological processes or vice versa.

1.3 DNA methylation

Methylation of DNA plays several roles in nature, bacteria protect their DNA from endogenous defense mechanisms by methylating it, thus restriction enzymes, that cut DNA into pieces (for example foreign DNA from phages), are blocked by methyl groups at the cutting sites targeted by the enzymes. In higher organisms methylation is found to play an important role in the individual control of the genome within cells. Methylation of DNA in vertebrate genomes is almost exclusively found in CG base duplets: cytosines followed by guanines, called *CpGs*. Methylation of CpGs in the DNA of humans and other vertebrates distinguishes tissues, imprints parental genes, influences the chromatin structure (Bestor, 1998), and is involved in the regulation of gene expression and cell differentiation (Ehrlich, 2003). DNA methylation patterns change in the process of aging, undergo significant changes in tumorigenesis and allow the differentiation of healthy from malignant tissue samples.

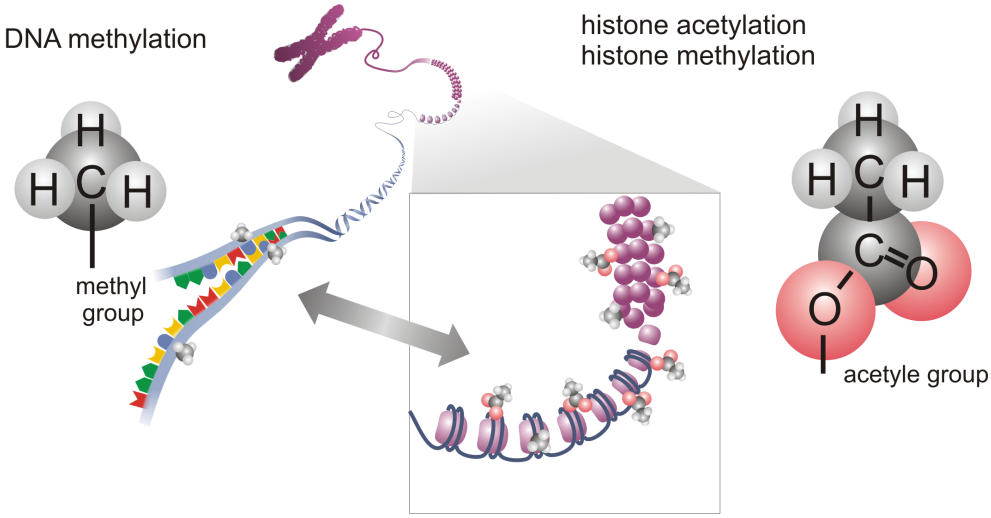

Figure 1.2: **DNA methylation and histone acetylation/methylation in the epigenome**. Two variable layers of information in the epigenome influence each other. The methylation state of CpGs in the DNA that is wound around the histones and packed in the chromatin structure influences the acetylation and methylation state of the histones and vice versa. The modification state of the histones is correlated with the chromatin density and structure (Bestor, 1998). It is not yet proven if there is a dominating layer which provides the leading control and, if so, which one it is. It is also not completely clear how much the one or other assessable chemical modification within the epigenome plays a role as initiation element for further processes or might only be a correlated observation caused by further regulative processes and does not initiate but indicate these processes.

1.3.1 An epigenetic information layer

DNA methylation is a chemical modification, that alters the base cytosine to 5'methyl cytosine. It is sometimes described as the 'fifth base' of the genetic alphabet. This modification provides a variable layer of information on top of the genome that is quasi static within an individual organism. The genomic layer provides identical information in the billions of different cells within an individual, whereas the methylation pattern of cytosines can differ between cells thus distinguishing them from one another, can rapidely change and can be involved in the control of cellular processes. It does this without altering the genetic code itself, but can also be relatively stable and be inherited (Morgan *et al.*, 1999). If the genome of an individual is represented by an alphabet of four letters, the information as to whether a cytosine on a specific DNA molecule is methylated or not, could well be conveyed by case sensitive use of the letter C, which will not change the text but will emphasize certain parts for functional reasons.

In human DNA, methylation of cytosines occurs almost exclusively in the two base palindromic sequence of cytosine followed by guanine, so called *CpG*s[1]. Within a single human cell the methylation of most CpG loci can have three states: 0% homozygote unmethylated, 100% homozygote methylated or 50% heterozygote methylated (except for those loci in an X or Y chromosomal context that have no diploid counterpart). In tissue samples, which are compositions of many cells, methylation becomes quantitative information or a binary mosaic pattern if broken down to single molecules.

Cytosines in the CpG context have a high mutation rate, cytosines in methylated CpGs tend to be deaminated to thymine (Duncan & Miller, 1980). They are about five times less represented in the human genome than expected by the overall base composition, are non randomly distributed and tend to accumulate in relatively CpG dense regions that are in general described as CpG islands (CpGI) [2]. Most CpG islands are reported to be unmethylated in healthy tissue (Grunau *et al.*, 2000; Strichman-Almashanu *et al.*, 2002), but hyper-methylated in cancer (Smiraglia *et al.*, 2001). The fact that such re-

[1] *CpG* means *C*ytosine, *p*hosphate bound, *G*uanine. The sequence is palindromic: it is identical to its reverse complement

[2] There are several different definitions and constraints that identify regions accumulating CpGs as *CpG islands* (Bird, 1986). For the sake of clarity this work will mostly use the term *CpG dense regions*

gions were not lost within evolution is likely to be based on some biological function that exerts evolutionary pressure on them leading to conservation. In fact many such regions show significant sequence conservation between human and mouse (Waterston *et al.*, 2002) and have probably been conserved through evolution due to an important functionality - the possibility to carry a variable additional information, the methylation layer.

1.3.2 The role of DNA methylation

The epigenetic information layer provided by DNA methylation, has been described as playing a role in many different biological pathways, this has lead to increasing interest from the scientific community, mostly with regard to control functions.

1.3.2.1 Gene expression

About half of all known human gene promoters have CpG dense regions in close proximity, which in somatic cells are reported to be mainly unmethylated (Bird, 1986). This is in contrast to most CpGs elsewhere in the genome, which are methylated in about 80% of all cases. It has been demonstrated that methylation of CpGs in promoter regions can suppress the gene expression of the associated gene. Methylation in promoter regions can therefore serve as a switch for gene silencing. The fact that CpG methylation can be recognized by enzymes[3], suggests that they have a direct influence on the binding and activity of transcription factors or co-factors (Yeivin & Razin, 1993; Kass *et al.*, 1997). An indirect way of transcription control via methylation has been described to be based on its influence on the chromatin structure.

1.3.2.2 Chromatin structure

Methylation of CpGs is catalyzed by several DNA methyltransferases (Bestor, 2000) and is suspected to induce a higher density of the chromatin structure leading to transcription repression (Bestor, 1998). Methylated DNA can bind the sequence independent transcriptional repressor MeCP2 followed by a transcriptional corepressor and histonedeacetylase, leading to deacetylated histones (Nan *et al.*, 1997; Nan *et al.*, 1998; Jones *et al.*, 1998). This deacetyla-

[3]Most restriction enzymes that have CpGs in their restriction site are known to be methylation specific whereby in most cases reactivity is blocked by methylation.

tion is associated with repression of transcription due to more densely packed nucleosomes and a more condensed chromatin structure.

1.3.2.3 Genetic imprinting

The mechanism of gene silencing by methylation can play an important role for the imprinting of parental genes. Either the maternal or paternal homologue of approximately 0.1 to 1% of mammal genes is repressed via methylation while the other is expressed mono-allelically (Ferguson-Smith & Surani, 2001). Maternally and paternally imprinted alleles are often located side by side in clusters.

A related mechanism to gene silencing and imprinting is the X chromosome inactivation via methylation, that occurs after fertilization in embryogenesis of females. This results in hyper-methylation and histone hypoacetylation of one of the two X chromosomes and suppresses the activity of its genes (Avner & Heard, 2001) [4]. In healthy somatic cells methylation patterns of imprinted genes or deactivated X chromosomes are very stable, most likely due to a special chromatin structure of the unmethylated allele (Feil & Khosla, 1999). Nevertheless tissue specific exceptions from these methylation states can be expected to be found at certain locations. Alterations for some imprinted genes were reported to be connected to human diseases (Walter & Paulsen, 2003).

1.3.2.4 Development

Within the development of mammals methylation undergoes many different stages (Reik *et al.*, 2001). The genome of germ cells is reprogrammed firstly by demethylation that erases imprints from the previous generation and secondly by de novo methylation that reestablishes imprinting on the mature gametes. After fertilization, the paternal half of the diploid zygote genome is actively demethylated by a currently unknown mechanism, whereas the maternal half is more slowly and passively demethylated by synthesis of unmethylated complement strands in the DNA replication step prior to cell divisions. Within this period the imprints are conserved by a mechanism which is not

[4] X chromosomal genes with 100% methylation in male and 50% methylation in female tissue samples due to X chromosomal silencing are commonly used as known methylation markers for sex to proof new methylation marker detection technologies.

methylation. The imprinted regions are again methylated in the development of inner cells of the blastocyte.

1.3.2.5 Aging

Random changes in methylation of normal somatic cells occur very rarely, but were observed to accumulate with the process of aging (Fraga *et al.*, 2005). It has been reported that during aging CpG dense regions next to genes can show an increase in methylation (Issa *et al.*, 1994; Issa *et al.*, 1996) while global methylation was reported to decline (Wilson *et al.*, 1987).

1.3.2.6 Tumorigenesis

Changes in the methylation pattern, induced during the aging process or otherwise, can contribute to carcinogenesis (Jones, 2002). Methylation of tumor suppressor genes has been suspected to be a possible cause of silencing of transcription and therefore proposed to be a third pathway for loss of function (after intragenic mutation or loss of chromosomal material) (Jones & Laird, 1999) and was found to be as frequent as inactivation by genetic mutations (Jones & Baylin, 2002) Another influence on tumorigenesis might also play a huge role - methylated cytosines have a mutation rate ten times higher than unmethylated cytosines. Inactivation of the human tumor suppressor gene TP53 by point mutation, was reported to be based on methylated cytosine in 50% of all cases (Rideout *et al.*, 1990). In addition methylation seems to change dramatically in cancer development, leading to hyper-methylation of CpG dense regions next to promoters. Probably most of these changes do not initiate carcinogenesis but are an accompanying symptom, that nevertheless can be studied and used for diagnostic purposes.

1.3.2.7 Conclusion

The basic knowledge gained so far about the DNA methylation and the epigenome shows that it is of great importance to get a better understanding of its global functionality within the interactive network of genomics, transcription, proteomics and other fields that help to understand life on a cellular level. Therefore large **DNA methylation studies are fundamental for a better understanding of diseases, cancer development, aging processes, cell differentiation and more.**

1.4 Methylation detection methods

Different methods for methylation measurement like PCR based, DNA chip based and sequencing based methods are described in several reviews (Dahl & Guldberg, 2003; Siegmund & Laird, 2002).

One major group of technologies is based on methylation sensitive enzymatic restriction of the DNA. The methods use restriction enzymes with recognition sites including CpGs, that block a cutting reaction when methylated. Other methods like for example PCR are then used to detect whether a known DNA context containing such a restriction site was cut (was unmethylated) or is for example still amplifiable (was methylated). Though these methods can be very precise and also be able to cover many sites in a genome at once, they are restricted to CpGs located in recognition sites and therefore are not able to provide detailed information about local profiles of CpGs within a close context.

The other group of technologies is based on bisulfite conversion of unmethylated cytosines (Olek *et al.*, 1996). Bisulfite treatment of DNA leads to a chemical conversion of unmethylated cytosine to uracil (see Fig. 1.3). Methylation of cytosines blocks this reaction. In most cases PCR (see Fig. 1.4) is used to amplify regions of interest within the bisulfite converted DNA template whereby positions converted to uracil appear as thymine in the product, which is then measured by different methods (see Fig. 1.5). Typically a tissue sample contains a mixture of different cells, therefore a proper description of methylation at a certain CpG requires quantification of the proportion of the methylated templates at the investigated CpG. This proportion is referred to as the methylation rate of the CpG. After the bisulfite conversion and the PCR reaction, the methylation rate at a CpG can be determined by assessing the proportion of remaining cytosine relative to the thymine. This can be done, for example, by hybridization to oligomer probes on DNA chips (Adorjan *et al.*, 2002) or by DNA sequencing (Frommer *et al.*, 1992). Commonly used sequencing methods include the sequencing of a representative number of subclones of the PCR product or direct PCR sequencing by running independent sequencing reactions for cytosine and thymine using the same dye in different lanes of a sequencing gel (Paul & Clark, 1996). These sequencing methods are costly and labor intensive.

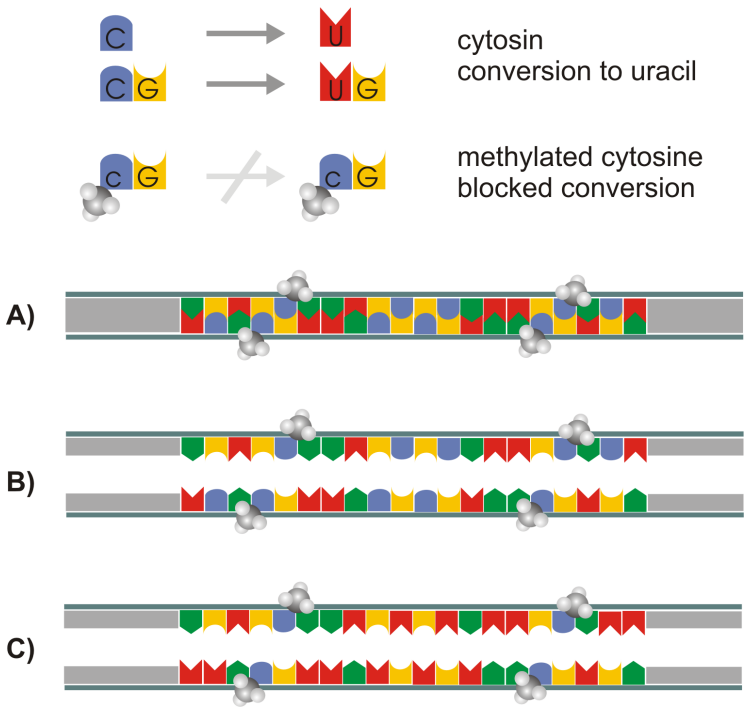

Figure 1.3: **Bisulfite conversion of DNA.** Cytosines in DNA can be deaminated via a chemical reaction that converts them to uracil. Cytosines with a methyl group at 5' position of their carbon ring are protected from this reaction and therefore stay unconverted (In mammal DNA methylation of cytosines is almost exclusively found at CpG positions). **A)** DNA with cytosines of unknown methylation state is **B)** denaturated to its single strands either by applying the melting temperature or by using chemicals. Cytosines that are not methylated are chemically converted to uracil **C)**, destroying the double strand symmetry in a way that does not allow the DNA to renaturate. Methylation information becomes easily detectable as base information.

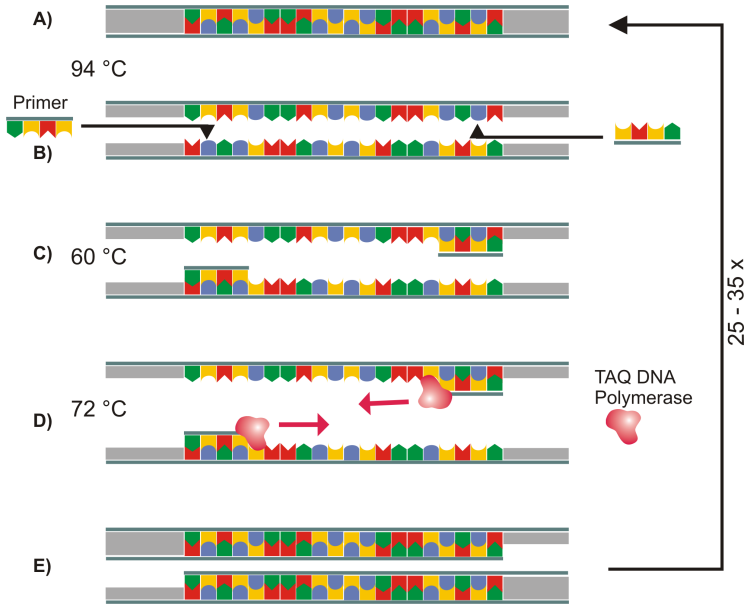

Figure 1.4: **PCR: Polymerase Chain Reaction**. The polymerase chain reaction is used to amplify specific DNA regions, providing billions of identical copies based on some few template molecules. The method uses a repeated cycle with a functional temperature profile (temperatures given in the figure are examples and can vary in setups optimized for specific amplification) and the following components: 1. DNA template, 2. a high temperature stable DNA polymerase (TAQ DNA polymerase: isolated from ıThermus aquaticus, an bacterium that lives in almost boiling water and has a proteome that is adapted to high temperatures), DNA polymerases are enzymes that can synthesize double stranded DNA based on single strands using it as template for the corresponding reverse complement. 3. primers, short starting fragments of synthetic, single stranded DNA that are needed to define short double stranded start positions for the polymerase. Primers are often chosen as unique pairs that flank exactly one desired known sequence in the template. 4. single nucleotides that are used by the polymerase to build the new strands. **A) Denaturation**: Double stranded DNA is denaturated to its single strands by applying a temperature of 94°C. **B) and C) Annealing**: The reaction is cooled down to 60°C to allow the primers to anneal to reverse complement parts in the sequence. **D) Elongation**: The temperature is raised to 72 °C, the ideal working condition for the polymerase that synthesizes double strands from both single strands starting at the primer position and thereby doubles the DNA in the primer defined region. **E) and A)** The process is iteratively repeated, leading to an exponential amplification of the region between the primers until the sources or the enzyme activity are exhausted. One cycle takes about two to six minutes.

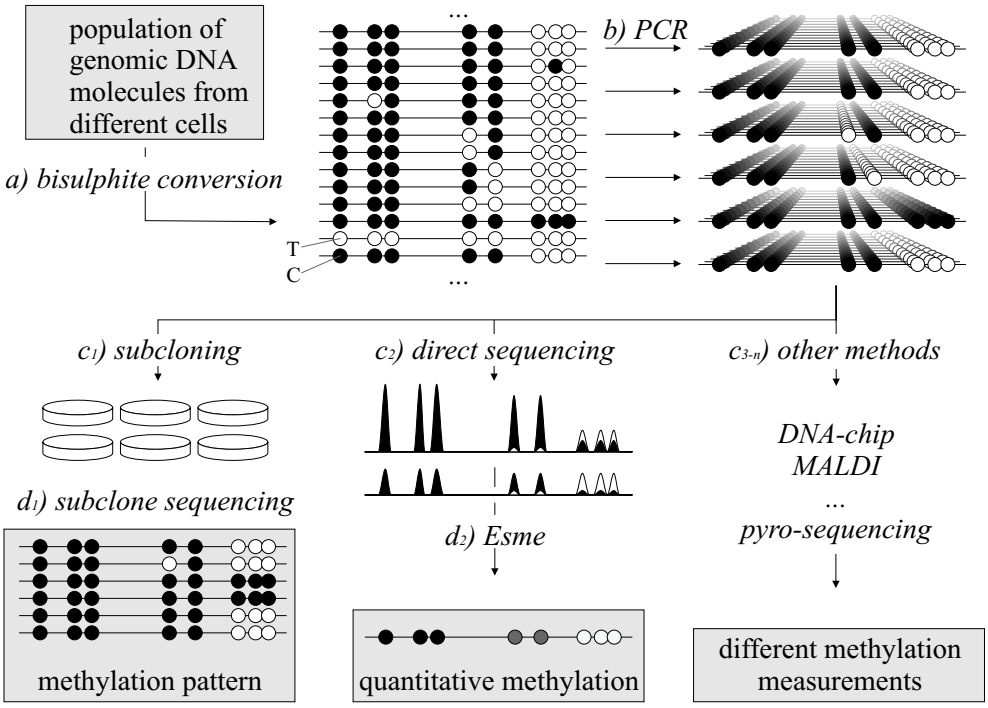

Figure 1.5: **Bisulfite conversion and PCR based methods for methylation detection** (with special respect to sequencing). **a)** Genomic DNA from e.g. tissue samples providing a population of molecules from different cells is treated with sodium bisulfite. This process converts unmethylated cytosines in the DNA molecules to uracil. The reaction is blocked by methylation so that methylated cytosines in CpG context are not chemically converted. **b)** The shown parts match a region of interest for methylation analysis that is amplified with specific primers. There is a possibility to observe amplificate specific and systematical biases or PCR variance. Based on the amount of template and other factors, the PCR more or less representatively amplifies molecule populations differing at CpG positions. Different technologies can be used to characterize the PCR product. To represent the population of different molecules in detail, single copies can be sampled via a subcloning step **c1)** and are characterized via subclone sequencing **c2)**. Other methods like DNA chips, MALDI, or pyrosequencing **c3-n)** can be used to get quantitative measurements for all or a subset of CpG positions. Direct sequencing of the PCR product **c2)** can be used to quantify the average methylation of CpGs in the amplificate either using special sequencing methods or using four dye high throughput sequencing and special algorithms as described in detail in this work.

1.5 Four-dye capillary DNA sequencing

For the Human Epigenome Project direct PCR sequencing on standard sequencing machines was used to achieve the required throughput in a cost effective way. This technology produces four-dye electropherogram data (see Fig. 1.6).

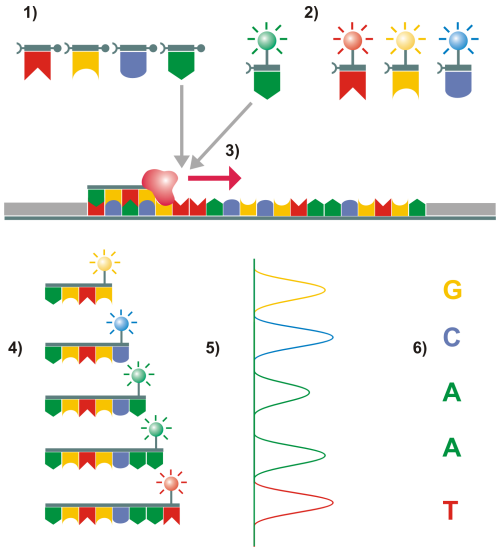

Figure 1.6: **Four dye DNA sequencing** uses millions of identical DNA molecules as templates that are either obtained from PCR or from plasmid preparations. Many single stranded copies from one side of the double strand are synthesized using single nucleotides that are added as a mixture of **1) dNTPs**, which are polymerized and build new DNA single strands. A small fraction of **2) ddNTPs** are added which, if incorporated, do not allow a further elongation. The four base types of ddNTPs are modified and carry different dyes. Their random incorporation within millions of parallel **3) elongation reactions** by a polymerase leads to a statistical distribution of single strands of different (primer to initiated stop position), that are marked by the dye corresponding to their last base. The molecules are separated by length using electrophoresis leading to a **4) ladder of molecules**. Sampling of light emission after excitation with UV light during electrophoresis results in **5) dye signals** that are translated to base **6) sequence information**. In this work proportions of signals are used to quantify base compositions.

Electropherogram data data is stored in trace files and in general represented as time series of signals from the four bases A, C, G and T. The data includes annotations interpreted by basecaller software: maxima of signals and the resulting DNA sequence of the sequencing experiment (see figure 1.7). One very established format for trace data is the scf file format (Dear & Staden, 1992).

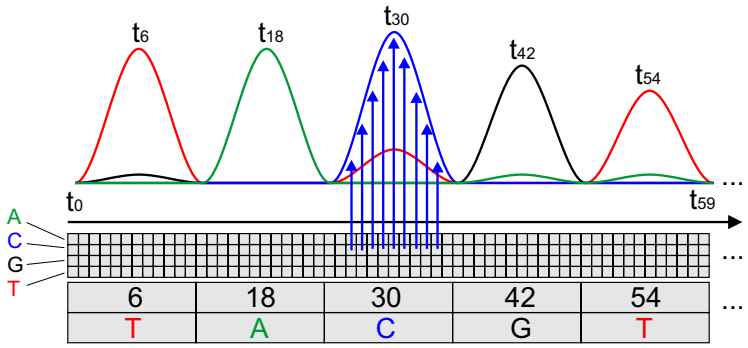

Figure 1.7: **Trace file data** from DNA sequencing (see Fig. 1.6). Trace files mainly contain data fields of two sizes (gray in the figure): 1. preprocessed signal data and 2. the sequence. Four vectors of same size describe time series of signal measurements for the four bases A, C, G and T, two smaller vectors of matching size describe the sequence interpreted from the time series and the the time index of corresponding signals for each base position within the signal data. Beside this general content, different types of trace files might contain machine and provider specific data, annotations and for example raw data from the machine used to generate the time series for base signals.

Though direct sequencing has advantages compared to subclone sequencing, it has one important limitation: Data from direct sequencing of PCR products can describe the average methylation of the CpGs within the amplificate. The averaged methylation for a position over all molecules does not allow assessment of mosaic patterns of the single molecules in case of mixed methylation (see Fig. 1.8).

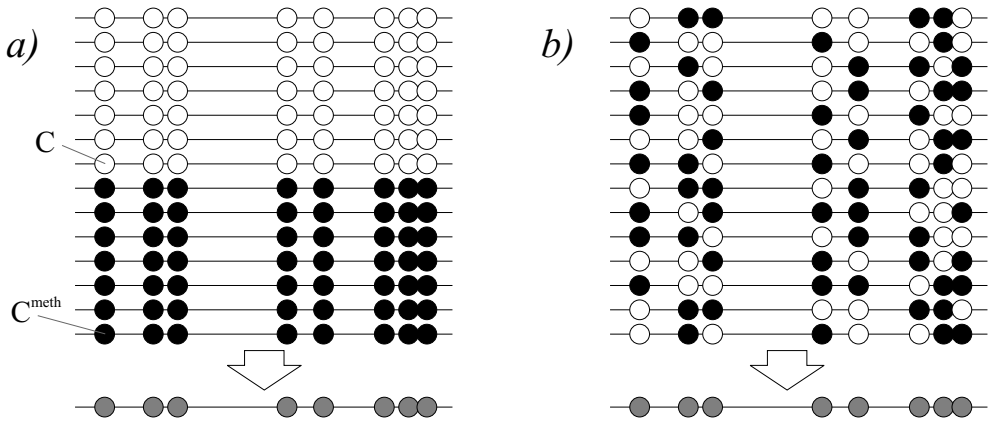

Figure 1.8: Schemes of methylation distributions in a population of molecules with 50% methylation in average at each CpG position. **a)** Within the molecules methylation is homogeneous. **b)** Methylation within molecules is based on a random mosaic pattern based on a certain possibility of a CpG to be either methylated or not. This leads to the average methylation observed at each CpG. The difference of these two possibilities cannot be detected by methods that quantify methylation at CpGs over the whole population but has either to be able to distinguish homogeneous subgroups of molecules or characterize single molecules for example with subcloning.

1.6 Motivation

1.6.1 Why to establish direct bisulfite PCR sequencing for methylation studies

Exploration of the epigenome at a representative level is a great challenge: it is insufficient to assess the DNA methylation patterns of some single individuals. In order to find systematic patterns it is necessary to study several different individuals' epigenomic layer in several different tissues and at different developmental stages or disease states. An analysis to understand methylation in the epigenome has to cover large parts in detail. Such analysis has been limited, because technologies leading to high resolution data and efficient high-throughput studies in an affordable way have not been available.

Methylation in tissue samples which are compositions of different cells is a quantitative information represented by cytosine/thymine proportions after bisulfite conversion of unmethylated cytosines to uracil and polymerase chain reaction (PCR). These PCR products can then be characterized for example by sequencing (Frommer *et al.*, 1992; Olek *et al.*, 1996). In the past high throughput sequencing of one sample used to break down this information to single molecules by subcloning, needed sequencing of a representative number of subclones and later represented the methylation level by averaging the results. This method is costly and labor intensive and therefore is not the first choice if one wants to study DNA methylation within a significant amount of different tissues represented by multiple tissue samples.

A preferable method to circumvent representations of molecules by subcloning that makes large studies affordable is direct sequencing, given that it is possible to quantify methylation using sequencing results from molecule populations with differences at CpG positions.

1.6.2 The goal of this thesis

A main part of this thesis is focused on enabling large DNA methylation studies using the established technology of DNA sequencing by developing a novel quantitative methylation analysis algorithm and workflow based on direct sequencing of PCR products from bisulfite treated DNA. The algorithm gains quantitative methylation information directly from base proportions rep-

resented by different dye signals in four-dye sequencing trace files, handling imbalanced and overscaled signals, incomplete conversion, quality problems and basecaller artifacts (the details of the method are topic of chapter 1.6.2).

The resulting method allows the use of infrastructures provided by the large sequencing facilities all over the world that were installed for sequencing genomes. This technology is a prerequisite for success of the Human Epigenome Project (HEP), the first large genome-wide sequencing study for DNA methylation in many different tissues (Human Epigenome Consortium *et al.*, 2003), initiated in 1999 (Beck *et al.*, 1999).

This thesis provides a closer look at the data of the pilot study and the first work package of the HEP. The pilot study covers genes in the MHC region in chromosome 6 of the human genome in six tissues. The first work package provides data from chromosome 6, 20 and 22 for 12 different tissues. These data provide information about differential methylation, methylation distribution and methylation states of neighbored sites that was not available in this resolution and quantity before the HEP.

Chapter 2

Quantitative analysis of trace data

In this chapter a quantitative methylation analysis algorithm and workflow based on direct DNA sequencing of PCR products from bisulfite-treated DNA with high-throughput sequencing machines is presented and evaluated. I show that the algorithm is useful to replace the alternative procedure of subcloning of PCR products and subclone sequencing, and therefore can extremely reduce costs and efforts for DNA methylation studies. This technology was a prerequisite for the Human Epigenome Project, the first large genome-wide sequencing study for DNA methylation in many different tissues. There is a high overlap between the content of this chapter and the publication describing the algorithm (Lewin *et al.*, 2004).

2.1 Motivation and Theory

Methylation studies by bisulfite sequencing using subcloning and sequencing of multiple subclones to represent a population of molecules are labor intensive and expensive. Direct quantitative measurement of methylation using the PCR product, is an alternative potential method in favor.

The possibility to use trace file electropherogram data for quantitative analysis of base compositions at single sites within pooled DNA was demonstrated for one single nucleotide polymorphism(SNP) (Qiu *et al.*, 2003). The same principle is used here for the measurement of methylation in bisulfite treated DNA product. Quantitative analysis by direct sequencing of PCR products from bisulfite treated DNA implicates several novel challenges: poor signal quality compared to genomic sequencing, overscaled cytosine signals and basecaller artifacts. In combination with the overscaled signals incomplete bisulfite conversion (Grunau *et al.*, 2001; Warnecke *et al.*, 2002) (which is a general problem of all bisulfite based methylation detection methods) influences signal proportions in the trace significantly.

It was therefore necessary to develop a specific algorithm enabling the use of four dye sequencing trace files to gain quantitative methylation information. This newly developed data analysis method allows the use of established high throughput sequencing technology for methylation studies.

2.2 Materials and Methods

2.2.1 Algorithms

The algorithm I present uses four dye electropherogram trace file data (see Fig.1.7) preprocessed by the basecaller of the sequencing machine manufacturer e.g. Applied Biosystems ".abi" files or the well described ".scf" files (Dear & Staden, 1992). The data processing includes the following steps: *(i)* entropy based clipping, *(ii)* signal detection, *(iii)* alignment, *(iv)* trace correction, *(v)* alignment based clipping, *(vi)* equalization of signal intensities signal normalization, *(vii)* signal normalization, *(viii)* compensation of incomplete conversion and *(ix)* methylation estimation (see Fig. 2.2.1). A scheme of the data and the influence of the algorithmic steps *(ii)*, *(iii)*, *(iv)* and *(vii)* is given in Fig. 2.2. Here we present the algorithms for forward sequencing that aims at the estimation of the proportion of cytosine to thymine at the positions of interest. Traces that originate from reverse sequencing and show guanine and adenine signals at corresponding positions can be analyzed by the same algorithm building the reverse complement of the trace files.

(i) Entropy based clipping: We observed that basecallers often generate reads that contain long stretches of called bases with up-scaled background signals after the end of an amplificate. These artifacts are detected by using the normalized Shannon entropy $0 \leq H \leq 1$ of the four trace curves S_b, $b \in \{A,C,G,T\}$ in a sliding window of 200 data points in the time series space of the trace signal data. Flanking sequence stretches with an entropy larger than 0.8 are removed.

$$H = -\sum_{b\in\{A,C,G,T\}} \left(\frac{S_b}{\sum_{B\in\{A,C,G,T\}} S_B} \log_4 \frac{S_b}{\sum_{B\in\{A,C,G,T\}} S_B} \right) \tag{2.1}$$

(ii) Signal detection: For each base position in the trace file we compute corresponding intensities $B^{\text{int}}; B \in \{A,C,G,T\}$ that estimate the base proportions in the molecular mixture. As an appropriate measure we have chosen the areas under the trace corresponding to the respective base for each position in the sequence. By default, the trace segment between neighboring local minima is used for the signal area estimation. If no local minima are present, then the boundaries of the trace segment are estimated as the mid point between two neighboring inflection points.

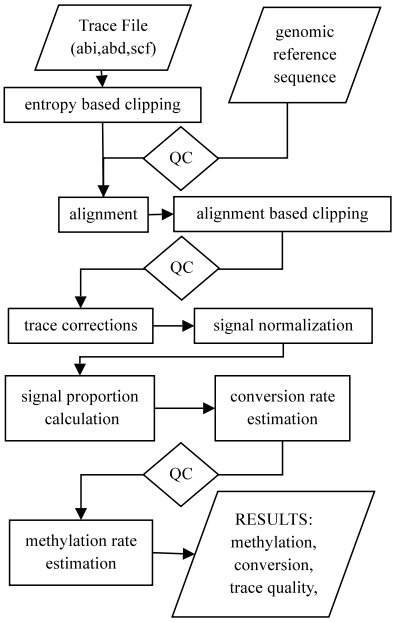

Figure 2.1: Flow chart of all data processing steps of the methylation estimation algorithm. Detailed description of the single steps is given in the text. Between all data processing steps quality control (QC) is performed. The analysis of a single trace file is aborted if the file itself is corrupted or if the genomic reference sequence is missing or if the length of good quality sequence, as determined by the clipping procedure, is below a certain threshold (default is 50 bases) or if the bisulfite conversion rates are below a minimum threshold (default is 65%).

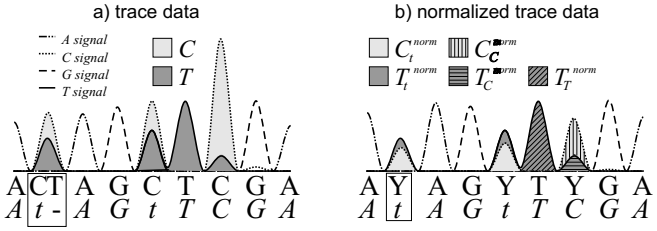

Figure 2.2: Schematic representation of a trace file electropherogram obtained by bisulfite PCR sequencing a) before and b) after signal normalization. The upper sequences below the trace curves in a) represent the sequence called by the standard basecaller and in b) the peak mixture represented using IUPAC code. The sequences at the bottom show the aligned reference sequence whereby t are genomic cytosine positions that are not in CpG context, and expected to be unmethylated and therefore completely convertible. Trace curves are shown for all four bases. For every base position in the reference sequence four base intensities $B^{\text{int}}; B \in \{A,C,G,T\}$ are calculated as the area under the trace curve segment that belongs to the base position (only C^{int} and T^{int} shown in a)). Normalized base intensities for cytosine ($C_b^{\text{norm}}; b \in \{t,T,C\}$) and thymine ($T_b^{\text{norm}}$) seen in b) are used to estimate the bisulfite conversion rate (base intensities at t positions) and the methylation level at each CpG (base intensities at C positions).

(iii) Alignment: The base intensities estimated in the previous step are then mapped to an underlying genomic reference sequence. The a priori availability of the genomic sequence is a prerequisite for our application. To describe the bisulfite converted DNA, the commonly used genomic alphabet (A,C,G,T) is extended by one letter, the lower case t, to distinguish a thymine derived from uracil by bisulfite conversion from a thymine that was present already in the genomic sequence. As an exception, cytosines in a CpG context in the reference sequence are denoted by C because their methylation status and therefore their conversion status is unknown. For the sake of clarity in the notation, these bases should be distinguishable from t which is never methylated and therefore expected to have a complete conversion by the bisulfite treatment. We use the Smith-Waterman algorithm (Smith & Waterman, 1981; Barton, 1993) for optimal local alignments allowing for gaps to align the called sequence of the trace file with the a priori known reference sequence.
Bisulfite treated DNA contains long stretches of T signal. In some cases this is misinterpreted by basecallers by inserting too many T-s into the called sequence. Accounting for this special situation, we have introduced an additional type of gap cost to guarantee proper mapping of CpGs. Assigning costs for gaps between C and G in the reference sequence forces the alignment of CpGs as one functional block to avoid their mismapping. An example of this is given below, whereby costs for gaps (g) are -19 and for special additional gap costs(sg) punishing insertion between C and G in the reference sequence are -20.

```
trace     ATTTTTTTGA            ATTTTTTTGA
reference ATTTTTC-GA            ATTTTT-CGA
          cost(g + sg) = -39    cost(g) = -19
```

(iv) Trace correction: Standard basecallers expect one homogeneous DNA population to be sequenced, therefore they often interpret mixed C and T base intensities at a single position of the reference sequence as two adjacent bases. In contrast to standard sequencing, in our experiments we expect signal mixtures from different DNA populations. It follows that the separation of overlaying intensities belonging to one position into two bases by the basecaller has to be corrected. We identify the separated base intensities by searching adjacent T and C positions in the called sequence from which one is aligned with t or C and the other is introducing a gap into the reference sequence.

These base pairs in the called sequence are then fused into a single base.

(v) Alignment based clipping: The quality of trace files from PCR product sequencing, especially of amplificates from bisulfite treated template containing different molecule populations, is lower than sequences from a homogeneous clone template. Alignment quality as a natural measure to assess sequencing quality is used to identify areas of poor quality. Flanking regions of the sequence are clipped such that the remaining inner part has less than 10% alignment error to the reference sequence.

(vi) Signal intensities in trace data decrease with progression of sampling time. If signals from cytosine in and out of CpG context and thymidine signals are not randomly distributed within an examined region, the proportions of those signals can be over or underinterpreted in normalization based on accumulation at locations with extreme signal intensity. We therefore equalize all signal intensities prior to normalization by dividing all four time series of base signals $B(t); B \in (ACGT)$ at each data point by the average signal intensity within a window of p data points and multiplying with 10,000 (see Fig. 2.3).

$$B'(t) = B(t)\frac{10,000p}{\sum_{t-p/2\ldots t+p/2}^{i} A(i)+C(i)+G(i)+T(i)} \tag{2.2}$$

(vii) Signal normalization: We found that cytosine trace curves often are overscaled in direct bisulfite sequencing traces[1]. Base proportion calculation based on trace curves with different baseline intensities would lead to misleading results. Therefore we normalize the trace curves prior to calculating the proportions of base intensities to determine bisulfite conversion and methylation rate. The normalized base intensities are denoted by B_b^{norm} ; $B \in \{A,C,G,T\}$; $b \in \{C,t,T\}$ that fulfill constraints (2.3) and (2.4) based on average base intensities .

$$\overline{T_T^{\text{norm}}} \equiv \overline{T_C^{\text{norm}} + C_C^{\text{norm}}} \tag{2.3}$$

$$\overline{T_T^{\text{norm}}} \equiv \overline{T_t^{\text{norm}} + C_t^{\text{norm}}} \tag{2.4}$$

Normalization of C^{int} is performed by multiplication of a global factor F_C.

$$C_b^{\text{norm}} = F_C C_b^{\text{int}}, b \in \{C,t,A,G,T\} \tag{2.5}$$

[1]We speculate that this over-scaling is a result of the standard basecaller software compensating for the low frequency of C signals.

Based on the data we use different strategies for normalization. If there are at least three C positions with $C_C^{\text{int}} > T_C^{\text{int}}$ normalization is based on data from these positions (Eq. 2.6 following from Eq. 2.3). Otherwise normalization is based on all t positions (Eq. 2.7 following from Eq. 2.4). In rare cases when all cytosines were unmethylated and converted completely ($C_C^{\text{int}} = 0$) normalization of the cytosine trace curve is impossible and unnecessary.

$$F_C = \frac{\overline{T_T^{\text{int}}} - \overline{T_C^{\text{int}}}}{\overline{C_C^{\text{int}}}} \tag{2.6}$$

$$F_C = \frac{\overline{T_T^{\text{int}}} - \overline{T_t^{\text{int}}}}{\overline{C_t^{\text{int}}}} \tag{2.7}$$

(iix) Compensation of incomplete conversion and *(ix)* methylation estimation: Cytosine base intensity at CpG positions can arise from two sources: from a population of methylated cytosines in the sample DNA and from an incomplete conversion reaction. It follows that the bisulfite conversion rate has to be first estimated to obtain a correct estimation of the methylation rate in the sample DNA. For an individual t the conversion rate R is estimated by

$$R = \frac{T_t^{\text{norm}}}{T_t^{\text{norm}} + C_t^{\text{norm}}}. \tag{2.8}$$

Local R_{loc} and global conversion rates R_{glob} can be determined by averaging over R of individual bases within defined ranges. Then the methylation rate $M, 0 \leq M \leq 1$, at a certain CpG can be estimated by using the following simple linear relationship

$$T_C^{\text{norm}} = R_{\text{glob}}(1-M)(T_C^{\text{norm}} + C_C^{\text{norm}}). \tag{2.9}$$

The equation describes the fact that T base intensity at a C position T_C^{norm} is expected to arise from the unmethylated portion of the sample DNA that is bisulfite converted by rate R. Furthermore the sum of the base intensities $T_C^{\text{norm}} + C_C^{\text{norm}}$ is assumed to be proportional to the total of cytosines in the sample DNA. It follows that the methylation rate then can be estimated by incorporating a correction for the incomplete bisulfite conversion

$$M = 1 - \frac{T_C^{\text{norm}}}{(C_C^{\text{norm}} + T_C^{\text{norm}})R_{\text{glob}}}. \tag{2.10}$$

Signal variance, artifacts or errors in the normalization might lead to negative methylation estimation which is set to 0.

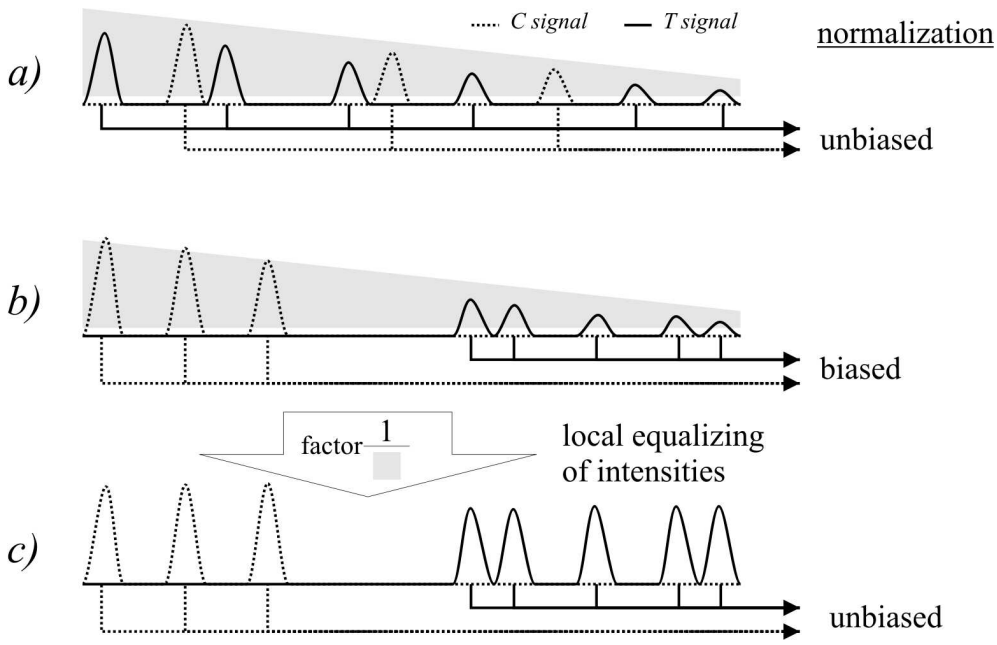

Figure 2.3: Signal equalization: Decrease of signals of other local profiles in trace files data can bias global factors based on base signals that are not evenly distributed (a, b). Local intensity is adjusted by dividing data by local smoothed intensities over all bases, leading to equalized signals.

2.2.2 Implementation

Algorithms described in this work are implemented in C++ as part of the software *esme*. For regular expression handling *boostregex* was used, file compression and decompression was implemented using *(*zlib) tests were implemented using *cppunit*. The software is object oriented and deeply integrated into C++ libraries and other in house software components of the company Epigenomics, allowing corba server and client functionality, file independent use of databases and fully automated use in high throughput. The software *esme*.was and is in use at Epigenomics and at Welcome Trust Sanger Center to analyze data from the Human Epigenome Project. A stand alone version of *esme* that can be used for analysis of the HEP data is freely provided as binary at http://www.epigenome.org. All data interpretation of results in this work was performed using the freely available statistical scripting language *R*. Some characteristics of the free standalone software are:

- C++ binary command line program running on Debian Linux platform
- input analysis of single files or full directory content, detailed data integrity checks and report of corrupted data, use of one or multiple scf, abi and abd trace files
- normalization of data and determination of bisulfite conversion
- estimation of methylation information gained by direct sequencing of PCR products from bisulfite treated DNA taking conversion rate into account
- plotting functionality for traces, modified traces, alignments of traces and reference sequences and results
- tab delimited results in two tables, for trace file quality results and for methylation at CpG levels

2.2.3 Test systems

2.2.3.1 Test systems with known cytosine/thymine ratios

To test how accurate we can measure base proportions in four dye trace data and if our normalization algorithm improves measurements, we created an artificial test system with known cytosine/thymine proportions. A 669 bp long fragment in the promoter region of the gene G6e was amplified by PCR reaction after bisulfite treatment of the template DNA. The bisulfite reaction was set up such that the conversion of cytosines was not perfect. The PCR product was sub-cloned into pCR2.1-Topo vector (invitogen).

96 clones were sequenced. Out of the 96 clones three showing differences at the most positions of genomic cytosine were chosen. The plasmid concentrations of the three stocks were adjusted to the same level. To gain different cytosine/thymine base compositions volumes were mixed in all six permutations of the proportions 1:2:4. These mixtures contain molecules with cytosine and thymine at the original genomic cytosine positions with expected cytosine/(cytosine + thymine) ratios from 0 to 1 in 1/7 steps.

Sense strands of the clone mixtures were sequenced five times using the kit 1.1 on the ABI PRISM 310 (2.4 a). Trace files were analyzed by using the ABI basecaller software 310POP4. Our algorithm was then used to estimate base compositions at each original genomic cytosine position. Estimated values were binned by their expected cytosine/(cytosine + thymine) ratios to assess their distributions and the mean absolute errors.

2.2.3.2 Test system with known methylation

To test our algorithm on data from DNA with defined methylation status, unmethylated human genomic DNA (Molecular Staging) was divided into two equal volumes. DNA in one of the volumes was enzymatically methylated with methylase SssI (NEB) following the manufacturer's protocol. Volumes of methylated and unmethylated DNA were mixed in 20 % steps from 0% to 100%. PCR for 60 amplificates was performed on a Tetrat MJ-research PTC-225.

For cycle sequencing the forward PCR primer was used with ABI kit 1.1 and run on the ABI 3730 DNA analyzer (2.4 b). Trace files were called with

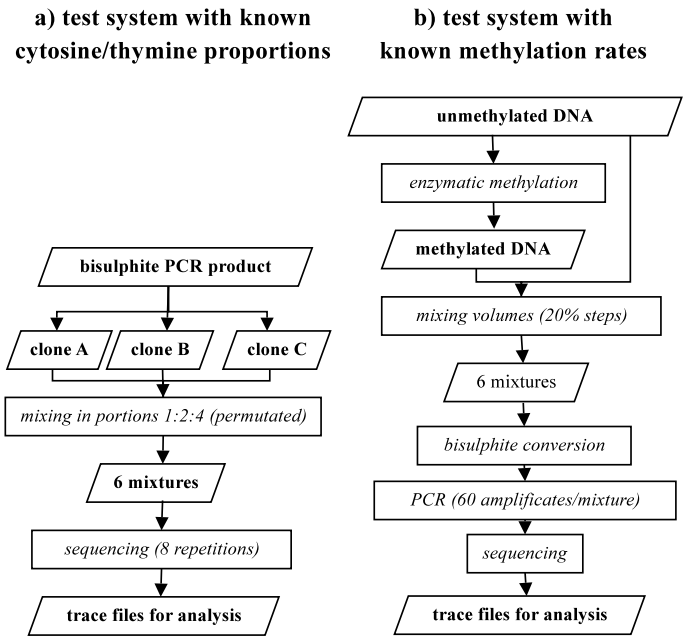

Figure 2.4: Experimental setup of **a)** a test system with known cytosine/thymine proportions **b)** a test system with known methylation rates. Steps that are potential sources for variances or biases in the test systems like mixing steps, incomplete enzymatic methylation, PCR bias (Warnecke *et al.*, 1997) and variance, incomplete bisulfite conversion (Grunau *et al.*, 2001; Warnecke *et al.*, 2002) and variance in the sequencing procedure are typeset in italics.

ABI's basecaller 3730POP7. Our algorithm was then used to estimate the methylation rates at each CpG position. Methylation rates were binned together by their expected methylation levels and variances and mean absolute errors were assessed.

2.2.3.3 Comparison with MALDI-TOFF

Data from direct bisulfite sequencing was compared with data for the same DNAs and genomic sites provided by the group of Ivo G. Gut from the *Centre National de Génotypage* in Paris (CNG): methylation data from MALDI-TOFF mass spectrometry using the GOOD assay (Tost *et al.*, 2003). Data was matched by CpG and tissue sample. Matched data was visualized in congruent plots color coding methylation in matrices describing methylation for each measured site and DNA, with one row per site and one column per site (methylation matrix plots).

Pearson correlation and mean absolute differences were used to describe the data comparison. The comparison was performed for sequencing data from both strands and separately for the data from sequencing different strands, either the cytosine poor forward or the guanine poor reverse strand of the bisulfite PCR product. For visualization of correlation, data from sequencing was binned into 10 bins from 0 to 100% by methylation data gained by MALDI measurements or vice versa.

2.3 Results

2.3.1 Test systems with known cytosine/thymine ratios

To assess the effect of our signal normalization step, we used our algorithms on data from the test system with known cytosine/(cytosine + thymine) ratios. Figs. 2.5 a, b show the distribution of the estimated ratios against the expected ratios in the test system without and with normalization, respectively. The results demonstrate that the normalization step decreases the mean absolute error (represented by the dashed line on the figures) approximately to the half. Sequencing several subclones from a PCR product is an alternative method to measure the cytosine:thymine ratios in bisulfite treated DNA. The measurement error of this method depends mainly on the number of subclones that is sequenced. We benchmarked our direct sequencing method with the subcloning method. We calculated the smallest theoretical measurement error inherent to the subcloning method by simulating the sampling of 10 and 20 subclones based on binomial distributions with a certain C:T ratio. Figs. 2.5 c, d and Table 2.1 show that errors in our estimates are comparable with those that could be obtained by sequencing 20 subclones of a PCR product. From this we can conclude that direct sequencing of bisulfite treated DNA is a viable alternative to subclone sequencing of at least 10 subclones if only the mixture rates are the subject of interest.

	mean SD	mean absolute error
signal proportions	0.110	0.130
normalized signal proportions	0.077	0.055
10 subclones	0.100	0.083
20 subclones	0.072	0.058

Table 2.1: Comparison of mean standard deviations and absolute errors of C/(C+T) signal proportions as estimated in our test system with known cytosine/thymine proportions and simulated representation by subclones.

2.3.2 Test system with known methylation

We have evaluated the performance of our algorithm by using the test system with known methylation rates. Fig. 2.6 shows results for data assessed without use of the algorithm (a), using different aspects of the algorithm only (b to e). Fig. 2.6 f shows the distribution of the estimated methylation rates

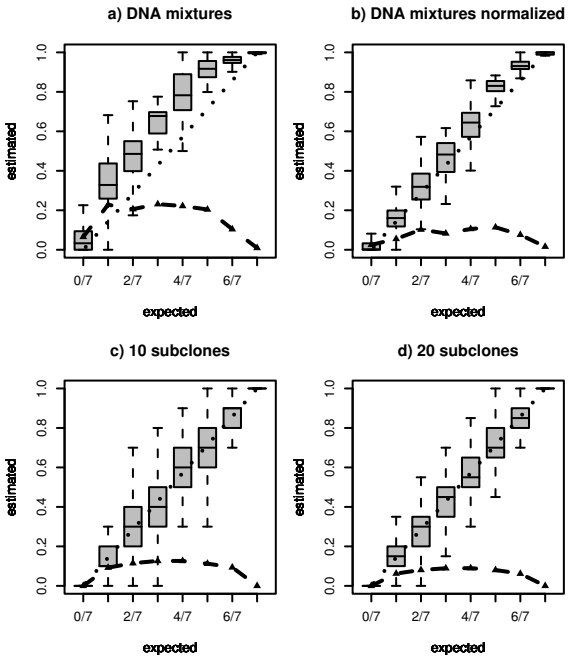

Figure 2.5: **a), b)** Quantitative measurements of C signal proportions in data from single sequencing runs of six clone mixtures with expected C/(C+T) ratios from 0 to 1 in 1/7 steps. The boxplots show the distribution of the estimated values obtained by our algorithm without the normalization and with normalization, respectively. The estimates are plotted against the expected ratios (1039 data points total which means a measurement success rate of 89% given six mixtures, five repetitions and 39 positions). Dashed graphs show the means of absolute errors. **c), d)** Simulated data for representations of mixed DNA in 0 to 1 in 1/7 steps by 10 and 20 subclones based on a binomial distribution.

against the expected methylation rates in the test system using the complete algorithm. For the estimation of methylation rates all steps, normalization of intensities, equalization of signal profiles, and the correction for bisulfite conversion improve the data. Table 2.2 provides corresponding data to Fig. 2.6 summarizing yield of analyzable CpG position (N) errors from expected measurements and variance. The algorithmic step of normalization (b) has a major impact mainly on low methylation rates, where the absence of C signals leads to an overscaling of the C trace and overall improves the yield, reduces error and variance compared to no algorithmic step (a). Correction of incomplete conversion (c) reduces errors and variance, but does not influence the alignability and therefore the yield. If both steps are applied the error is even further decreased as with each of the steps, but variance is still higher than with correction of incomplete conversion only (d). Raising alignability and yield by normalization most probably allows the inclusion and assessment of parts in trace file regions with higher variance and lower quality. If the data is filtered for trace files that relatively stable intensities over their whole length[2], variance of measurements and errors could further be reduced but costed yield. If signal equalization was applied in addition to all other algorithmic steps, the best results were obtained for all parameters.

The methylation rates estimated in this experiment do not show as accurate correlation with the expected rates as was obtained in the previous C:T proportion experiments, where a mixture of subclones was used as a test system. One possible explanation for this is that the real methylation rate in the mixtures of methylated and unmethylated DNA deviates from the expected methylation rate. Systematic biases in the real values of all 60 covered regions can arise from incomplete enzymatic methylation of the DNA or from amplificate specific biases in the PCR reaction itself.

Systematic biases in the test system would lead to deviations from expected values and to a higher variance in the complete data but still allow detection of relative differences in the methylation rates at individual CpG positions. To evaluate the capability of our method to detect differential methylation we paired data from templates with different methylation values for each CpG. Table 2.3 lists the accuracy of classification of higher vs. lower methylated CpGs in the test system.

[2] we used a filter based on coefficients of variances for each base: standard deviations over all signals divided by average signal had to be below 2

	N	MAE	SD
nor normalized neither corrected	1716	0.2675	0.1441
only normalized	1854	0.1744	0.1805
only corrected	1716	0.1876	0.1478
normalized and corrected	1854	0.1417	0.1735
filtered, normalized and corrected	1812	0.1341	0.1535
equalized, normalized and corrected	1998	0.1256	0.1364

Table 2.2: Influence of algorithmic steps. Amount of positions measured (N), mean absolute error (MAE) and standard deviation (SD) in 20% step methylation mixture calibration data.

expected rate	0.2	0.4	0.6	0.8	1
0	91	98	97	99	99
0.2		90	98	99	99
0.4			96	97	98
0.6				79	88
0.8					89

Table 2.3: Test system with known methylation rates: accuracy of sorting paired methylation estimates at identical CpGs in 60 amplificates after normalization and conversion rate correction.

The accuracies for detecting differential methylation in neighboring methylation rates with 20% steps are compared with those that were obtained without normalization and correction for incomplete bisulfite conversion (Table 2.4). The performance clearly improves by using the normalization and the conversion rate correction steps.

Despite the overlap of the distributions of the estimated methylation values (cf. Fig. 2.6 d) we can conclude that the detection of differential methylation is highly accurate. This is in accordance with our hypothesis of having amplificate specific systematic biases in our reference test system.

We have evaluated threshold parameters for quality control. More stringent parameters do not improve the results significantly but lead to lower measurement success rates. For example, raising the threshold for bisulfite conversion from 65% to 80% reduces the mean absolute error by 0.02 % and

	correct sorting[%]	
comparison	raw	norm/corr
0/0.2	84	91
0.2/0.4	71	90
0.4/0.6	86	96
0.6/0.8	77	79
0.8/1	89	89

Table 2.4: Test system with known methylation rates: accuracy of sorting paired methylation estimates at identical CpGs in 60 amplificates with 20% difference with and without using the normalization and correction for incomplete bisulfite conversion.

raises the accuracy by 1.3 % but reduces the number of accessible positions by 16 %.

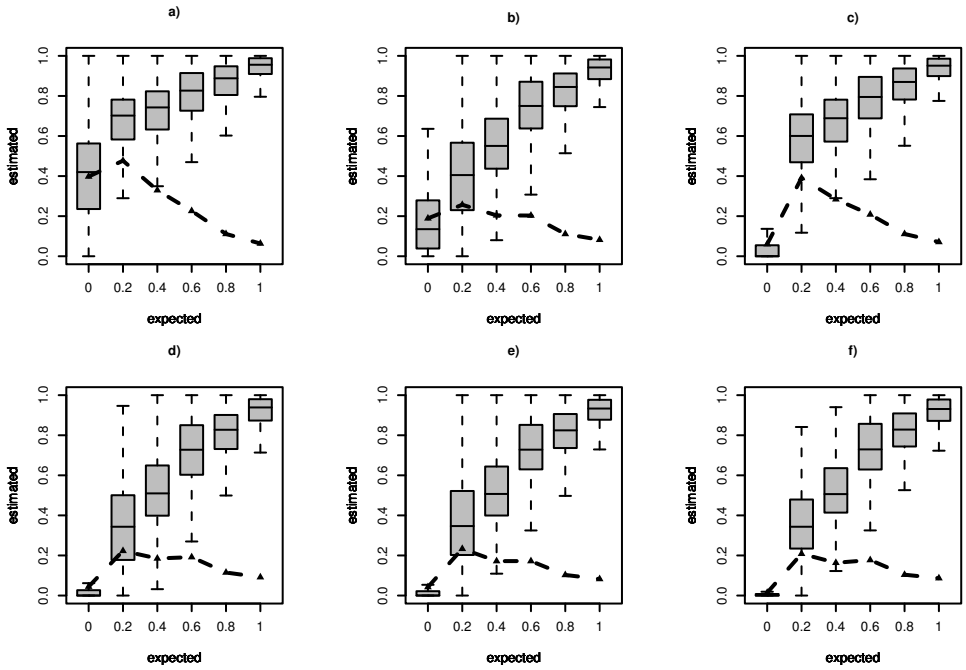

Figure 2.6: Estimation of methylation in the test system with known methylation rates. The boxplots show the distribution of the estimated methylation rates as a function of the expected methylation and the mean absolute error (dashed line). Each box includes data from CpGs of all 60 amplificates measured at the expected methylation rate. Different algorithmic steps were applied: **a)** nor normalized neither corrected **b)** only normalized **c)** only corrected **d)** normalized and corrected **e)** filtered, normalized and corrected **f)** equalized, normalized and corrected .

2.3.3 Comparison with MALDI-TOFF

594 paired CpG/sample methylation measurements from 1. direct sequencing with use of the algorithm and 2. MALDI-TOFF mass spectrometry were compared. Over all the two different methods show similar methylation profiles on CpG level see Fig. 2.7). Despite technical differences, individual biases and variance of the two methods, a correlation of 0.88 could be achieved. Interestingly the data did not only correlate around 0% and 100% but also shows good results measuring mixed methylation. A mean absolute difference of 12% is in the same range as the mean absolute error measured for the

test system with known methylation rates (see table 2.2). Data from forward sequencing performed better than from reverse sequencing (see Fig. 2.8), but the differences found are not significant given the size of the data set.

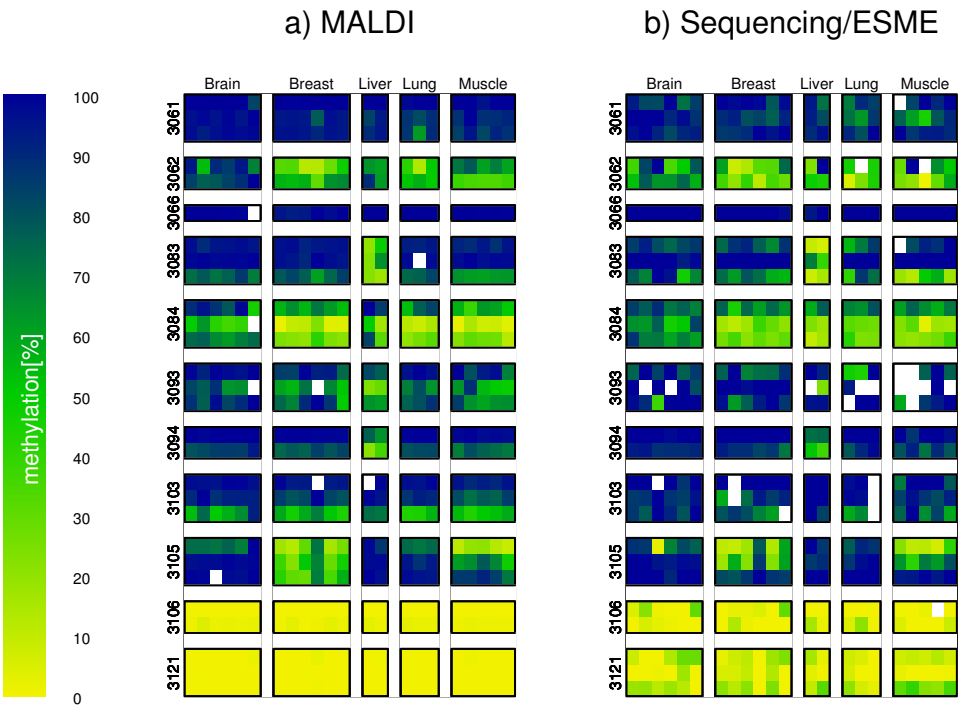

Figure 2.7: Comparison of methylation rates for 28 CpG locations (11 amplificates) in 22 tissue samples (5 tissues) with **a)** MALDI and **b)** Sequencing/Esme. Sequencing data is binned by methylation measured with MALDI. Methylation is color coded from yellow (0%) over green (50%) to blue (100%). White areas lack measurements. A comparison with data binned by sequencing is found in Fig. 7.1 in the appendix.

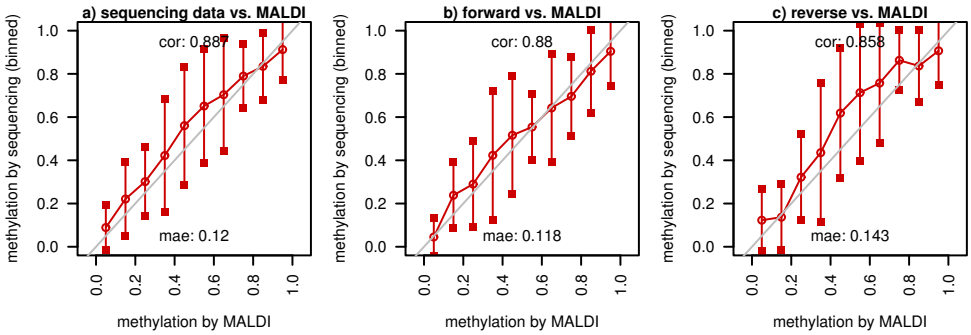

Figure 2.8: **Comparison of methylation measurements with MALDI**. Measurements from the method displayed at the y-axis are binned into ten intervals form 0 to 1 by measurements at the same CpGs and in the same tissue samples with the other method displayed at the x axis and displayed as boxplots. a) Methylation rates at CpGs from forward and reverse sequencing compared to corresponding MALDI measurements. b) Methylation rates at CpGs from forward sequencing compared to corresponding MALDI measurements. c) Methylation rates at CpGs from reverse sequencing compared to corresponding MALDI measurements. Red lines show the means of the binned rates, bars show the standard deviations.

2.4 Conclusion

Results obtained by comparison of data from the algorithm described and reference test systems show that direct PCR sequencing is a viable alternative to estimating methylation rates by sequencing subclones from the PCR product. Replacement of at least 10 times subcloning (see 2.3.1). extremely reduces laboratory work and costs. Furthermore, we have demonstrated that our method can detect differences in methylation rates of at least 20% with high accuracy. Applying our algorithms to bisulfite sequencing data of partially differentially methylated DNA and comparison with MALDI-TOFF data demonstrated that by the aid of the method, CpGs with differential methylation rates between different tissue types can be identified.

The algorithm provides a useful way to analyze big DNA methylation studies like the Human Epigenome Project based on direct sequencing in high throughput facilities. It will help to gain information about differential methylation in many tissue types and increase our understanding of the epigenetic layer in the complex system of gene expression, cell differentiation and tumorigenesis.

Chapter 3

Deconvolution of trace file data

Trace data, especially from non optimized direct bisulfite sequencing processes or from problematic amplificates can show serious quality problems, prohibiting correct base calling and quantitative analysis. This chapter covers a proof of concept study, in which it is shown that in case of systematic problems such as molecular populations with different mobility quality problems can be reduced and overcome with a numerical approach deconvolving the trace data.

3.1 Motivation and Theory

A fraction of trace files gained by direct bisulfite sequencing shows signal echos or shifted signals, especially trace files from reverse sequencing. These echos have an offset of up to three bases to the main signal; multiple echos are observed rarely. We have three main theories about the origin of such phenomena.

1. **Mixtures of primers** with a fraction that is missing bases in the 5' end region and therefore results in populations of different base length after cycle sequencing reaction. Such effect would lead to global echos of same size and offset over the whole trace data.

2. **DNA polymerase slippage** within stretches of the bisulfite converted DNA in the sequencing reaction resulting in insertion or deletion of bases and leading to multiple populations that differ downstream of such stretches. Such effects would lead to echos starting at specific positions in trace file data.

3. **Mobility differences** based on different base compositions, which has been reported before to have an influence (Frank & Koster, 1979), introduced by the bisulfite conversion. Sequencing reaction products resulting from a methylated population in the DNA template could run shifted in comparison to those from unmethylated DNA showing an increasing offset with proceeding sequence length - based on the summation of the effects of different positions. Echos based on such an effect are expected to have identical sizes and raising offset for all base signals except for signals from CpG positions which introduce the effect because of base differences between the populations at these positions. Latter are expected not to show defined echos but different signals that sum up to total signal representation of such positions with identical size and offset as the echos observed at other positions.

In the following we will name the first two effects postulated *echo effects* and the third effect *shift effect*.

Both, echo and shift effects can be observed in trace data, but in most cases we observe shift effects. Echo effects distort the general signal patterns in trace files used for methylation analysis and therefore decrease the accuracy of results from the algorithms used in ESME. Shift effects below a offset

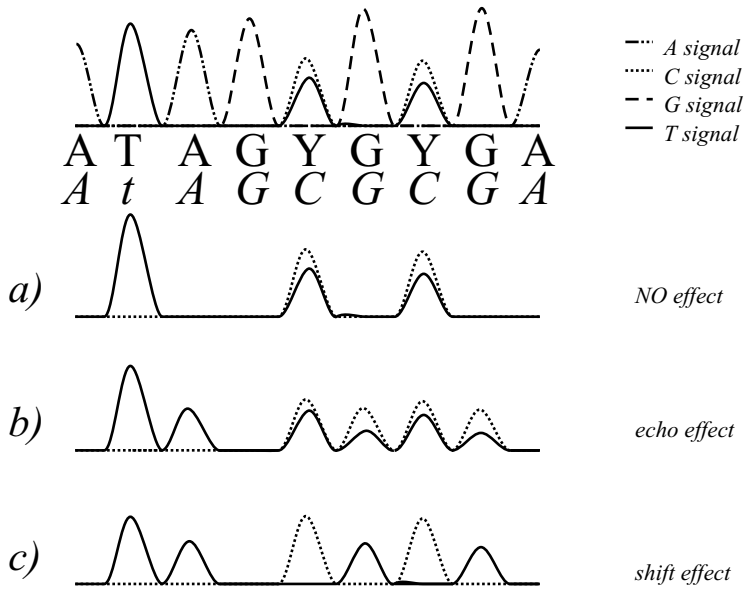

Figure 3.1: **Schemes of thymine and cytosine signals with echo and shift effects at CpG positions and at thymine position:** On the top the complete trace data of all signals is shown with mixed signals at CpG positions with a methylation level of about 60%. **a)** Cytosine and thymine signals only with no effects. **b)** Cytosine and thymine signals in convolved data, both showing an *echo effect*. **c)** Cytosine and thymine signals in data showing a *shift effect* introduced by the different base compositions at CpG positions, which come up as separated signals and sum up to signals belonging to one position that have the same shape as the echo of a thymine position being thymine in both populations.

of half a base position can be compensated by the fact that we do not use peak height to determine signal intensities but their peak area and therefore compute the total signal including slight shoulders and widened peaks. As soon as a shift effect offset reaches a size of half a base position cytosine and thymine signals at CpG positions are separated in a way that make further analysis based on the algorithms used in ESME impossible. It was therefore necessary to find a way to compensate for all such effects.

Deconvolution of trace data was already successfully used to enhanced base calling (Zhang & Allison, 2002; Li, 2001). Data showing the echo effect based on primers could easily be corrected by global deconvolution as used for signal echos in seismography, while data showing echo effects based on slippage could be corrected analogous by local deconvolution used for signal echos in mobile communications. The third effect however needs a specific model and specific kernels for positions that introduce the effect and are due to unknown signal proportions which at the end of the process have to used for quantitative methylation analysis.

3.2 Materials and Methods

3.2.1 Algorithm

The algorithm that deconvolves observed trace file data to compensate for echos, basically optimizes a kernel for the sum of cytosine and thymine signals, and a second kernel for the cytosine signal only, using a special energy function for the latter. The deconvolved cytosine signal is then subtracted from the deconvolved sum signal to get a deconvolved thymine signal. This way the unknown proportions at CpG positions can be circumvented in the model. The variables used in the following are defined in 7.1 in the appendix. The distance δ of the main signal of DNA population $i = 1$ to itself is by definition 0.

$$\delta_1 = 0 \tag{3.1}$$

The sum of the proportions π_i of all k DNA populations is 1.

$$\sum_{i=1}^{k} \pi_i = 1 \tag{3.2}$$

The model of the observed trace signals $O'_B(t)$ in the trace data is the sum of the signals from k DNA populations with different proportions π_i and offset δ_i to the main signal.

$$O'_B(t \mid \pi_1, \ldots, \pi_k, \delta_1, \ldots, \delta_k, F_B(t)) = (1 - \sum_{i=2}^{k} \pi_i) F_B(t) + \sum_{i=2}^{k} \pi_i F_B(t + \delta_i) \tag{3.3}$$

The deconvolved signal for a base F_B is its signal in the observed data O_B normalized with a intensity factor f_B and deconvolved with a kernel H.

$$F_B = (f_B * O_B) \otimes H; B \in \{A, C, G, T, Y\} \tag{3.4}$$

The deconvolution kernel H is dependent on the intensity factors for all bases, and the offsets δ and proportions π of all k molecular populations.

$$H(f_B, \pi_1, \ldots, \pi_k, \delta_1, \ldots, \delta_k); B \in \{A, C, G, T, Y\} \tag{3.5}$$

The minimum energy E_B for kernel optimization is a function of the observed data O_B, the model of idealized data M_B and the kernel H.

$$E_B(O_B, M_B, H) \tag{3.6}$$

(i) The methylation of the cytosines in the CpGs to be measured are unknown by definition. Therefore it is also unknown if they and which proportion were or were not bisulfite converted/amplified to thymidine. This means that there is no a priory expectation for cytosine signals at CpG positions in data from bisulfite converted DNA. This turns out to make the model complicated: if multiple signal 'echos' are expected to be based on base composition differences leading to divergent mobility of DNA populations. Each population/echo might have different proportions of (C or T) at different CpG sites which after signal normalization, convolution and applying π must sum up as the real proportion over all populations/echos. There are two possible ways to deal with this:
1. building a very complex model, that takes individual methylation/conversion rates of each cytosine for each different mobile population into account. This approach would extend the amount of $4 + 2 * k$ parameters describing the model by $n * (k + 1)$ whereby n is the amount of cytosines in the described part of a trace.
2. using an approach that is completely independent of methylation and bisulfite conversion by combination of O_C and O_T signals in both, model and data, and treating it as one signal O_Y describing two bases, that has two signal normalization factors f_C and f_T and a possible C signal shift offset S_C that must be taken into account for each algorithmic step but can be treated as one signal concerning everything else. This second possibility needs an additional algorithmic step that is able to separate the combined signals after deconvolving O_Y.

The second possibility was chosen based on some observations and assumptions:

- Bisulfite conversion rates obtained using recent Epigenomics' technology (or comparable methods), is always almost complete. Given this, the influence of O_C signal remaining from incomplete conversion of cytosines outside of CpGs is negligibly low after signal normalization and after deconvolution relevant F_C signal is to be expected only at CpG positions.

- Local deconvolution of O_C in windows of certain trace data points will mainly cover co-methylated CpGs that therefore can be described with the same deconvolution kernel.

$$O_Y(t, f_C, S_C, f_T) = O_C(t, f_C, S_C) + O_T(t, f_T) \tag{3.7}$$

$$M_Y(t) = M_C(t) + M_T(t) \tag{3.8}$$

(ii) The data is reduced to three signals: O_Y,O_A and O_G. In principle only O_Y is of interest containing all information of O_C and O_T and therefore for methylation. O_A and O_G are only helper signals that
1. map the signals to the reference sequence identifying the CpG positions
2. are expected to show the similar echos/shifts as O_Y and therefore add more data but
3. could for various unknown and trace file individual reason e.g. signal shift effects behave slightly different than O_Y, so that their inclusion might be counter productive.
The algorithm was therefore designed to allow three options: either *a)* the use of O_Y only to find the optimal kernel within a given data window or *b)* the use of all three signals O_A,O_G,O_T, or *c)* first the use of all three signals to find a good initial kernel that is then start for a further optimization based on O_Y only. Kernel optimization is done by finding the set of parameters that minimize the energy E_B of the deconvolved signal. The idealized trace signal model M_B is identical to that used for artificially generated trace data (see section 3.2.4.1 and formula 3.14).

$$E_B\Big(t \mid \pi_1, \ldots, \pi_k, \delta_1, \ldots, \delta_k, F_B(t), O_B(t)\Big) = \Big(M_B(t) - O_B(t) \otimes H\Big)^2 \tag{3.9}$$

$$E = \sum_B E_B; B \in (A, G, Y) \tag{3.10}$$

(iii) The former step provides a kernel that can be used to deconvolve O_Y to F_Y and also provides signal normalization factors f_C and f_T and C signal shift offset S_C, but does only describe the sum of the signals and therefore does not allow deconvolution of the signals of interest C and T separately from each other. The unknown methylation prohibits any meaningful model M_T based on expectation of thymidine signals at thymidine sites and cytosine sites outside of CpG context but without any possible assumption within CpG context. The same applies for M_C: it is impossible to have a model that simulates where C signals are expected to be, but it is possible to have a model defining where C is NOT expected. This allows kernel optimization for O_C

with fixed signal normalization factor f_C and shift S_C. This kernel cannot be optimized with a minimum energy based on differences between F_C and M_C, but with a modified energy function that takes into account that any F_C outside of CpG context and negative F_C within CpG context must be prohibited.

$$E_C = \sum_{t \notin CpG} F_C(t)^2 + \sum_{t \in CpG} \sigma(t) F_C(t)^2 \tag{3.11}$$

whereby

$$\sigma(t) = \begin{cases} 0 & \text{if } F_C(t) \geq 0 \\ 1 & \text{if } F_C(t) < 0 \end{cases} \tag{3.12}$$

(iv) The obtained kernel forces O_C signals (as far as available) into CpG positions. With F_C and F_Y there is no use for a further step finding a kernel for O_T. F_T is simply calculated by subtraction.

$$F_T(t) = F_Y(t) - F_C(t) \tag{3.13}$$

(v) To find and optimize a deconvolution kernel, we used the downhill simplex (Vetterling *et al.*, 2002) method. It was necessary to find a way to optimize the initial kernel of a downhill simplex that
1. results in a kernel with few parameters,
2. does not get lost in local minima.
To fulfill this we started with a more complex kernel allowing many k around a main population and then iteratively repeated kernel optimization with with decreasing number of k, using the δ from the last result for the next initial setting, excluding populations with the lowest π in the last result. In principle except for the optimization of the last kernel, all former steps are only to optimize starting parameters for the kernel in that last step (with few parameters) enhancing the likelihood that it does not get stuck in a local minimum. Models for trace curves $M_B(t)$ were generated as described in section 3.2.4.1, using perfect trace data except for increase in signal peak width with factor 1.3 over the trace.

Due to noise and variance in sequence data we expected deconvolution of trace data without shift or echo effects to lower the data quality. In most of all cases trace data does not show any effects and otherwise such effects tend not to be global but begin at a certain time point in the trace data. Given the former algorithm, such a decision can be done at different times. In order not to correct or harm already correct data, we decided to include an optional

and fully parameterized decision trigger, after step *(ii)* or *(iii)*, whether to deconvolve data within an assessed window or not, which needs at least one echo/shift effect with a certain proportion π of the overall signal and a minimum distance δ to the main signal peaks.

3.2.2 Implementation

Algorithms described in this work were implemented in an object oriented way in C++. Complex algorithms were checked with unit tests using *cppunit*. All algorithms like fft, deconvolution and downhill simplex were based on numerical recipes (Vetterling *et al.*, 2002), altering the iterators used in the FORTRAN based code examples from 1:n to 0:(n-1) and using an object oriented implementation. The algorithms were integrated into Epigenomics software ESME.

3.2.3 Parameter optimization

The optimization of parameters for the algorithm allows infinite possible combinations. The raw data interpretation using the algorithms is extremely costly in computing time and extends trace file analysis time by a factor up to 1000 x (minutes per trace file) compared to simple trace file analysis as described in chapter 1.6.2 (parts of seconds per trace file) based on parameters, e.g. how complex an iteratively optimized kernel is in the beginning, how sizes and steps of shifting windows are and which steps to take.

In order to keep computing time low and to ensure validity on real data, the optimization was performed using real trace file data from bisulfite sequencing. Trace files were used in sets of different sizes, that were hand chosen out of many others and showed obvious shift and echo effects prohibiting meaningful analysis so far. Due to the unknown methylation state of these data, original and results were compared by visual inspection and success was mainly defined by the fact whether C signals from various echos were changed to correctly placed peaks with expected shape and area or if the outcome showed obvious artifacts. Some parameters settings were tested on single trace files only: bad results for one example were counted as valid reason to directly neglect the tested parameter setting instead of extended testing on larger data sets.

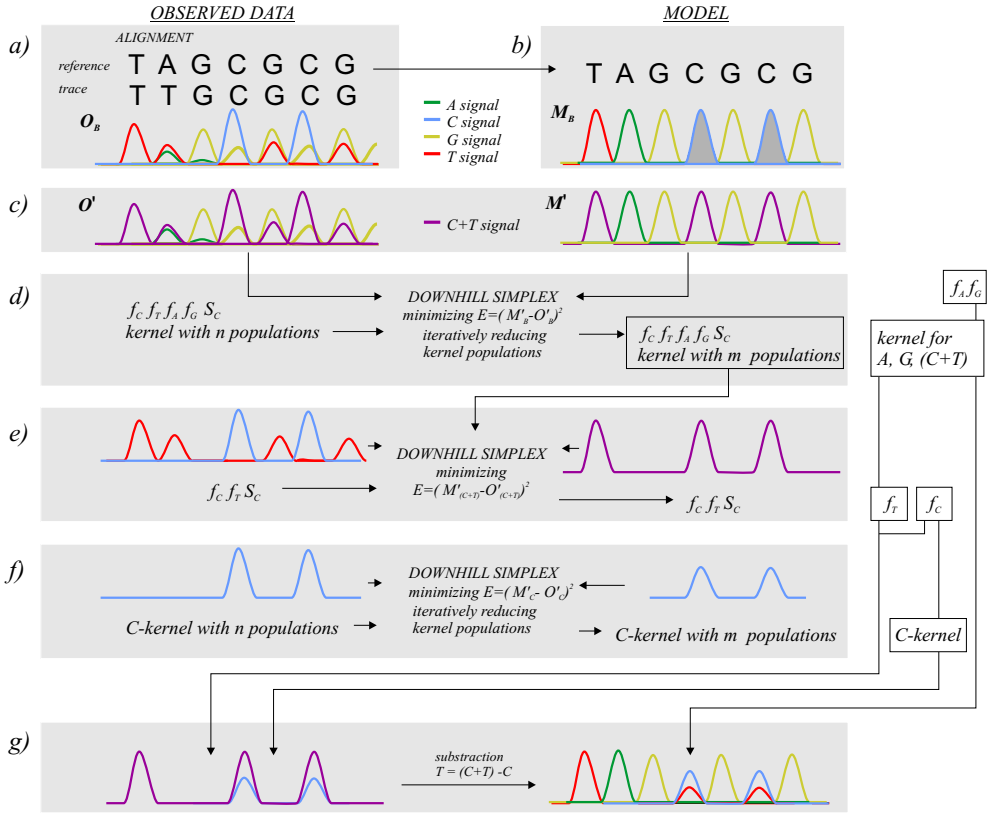

Figure 3.2: **Deconvolution algorithm**. Based on an alignment of observed trace data O in a trace file to a reference sequence (**a**), a trace data model M is built, using the base positions in the trace file and expected bases in the aligned reference sequence (**b**). To allow a methylation and conversion independent model, C and T signals are summed up in a simplified observation O' and model M'. In the unnormalized observed data O' the unknown C and T signal proportions resulting from converted or unconverted cytosines are equivalently treated as one signal, whereby base signal intensity factors f_C, f_T and a possible shift offset S_C for cytosine signal are included (**c**). A downhill simplex algorithm is used to find the optimal kernel for deconvolving O' to fit M' (**d**) including optimization of normalization factors F_B for all four signals. A parameterized step by step reduction from n to m signal echos in the kernel helps to overcome local minima. Afterwards a more detailed optimization of the very sensitive factors used for signal intensity normalization of C and T is performed (**e**), that allows the calculation of a deconvolved C+T signal. A kernel for the separated C signal, that is exclusively expected at CpG positions, is locally determined to provide a deconvolved cytosine signal (**f**). Again local minima are overcome by a reduction from n to m signal echos in the kernel. T signal is now calculated by subtraction of the deconvolved C from the deconvolved C+T signal (**g**).

First **different sliding window sizes** were tested, from 1000 to 100 data points, using different step sizes, from full window size, down to single data point for windows smaller 250. For kernel optimization **different starting kernels** with and without iterative reduction of kernel complexity were used, within a range of $k = 11$ at a maximum and $k = 1$ at minimum, with different π_1 ranging from 0.25 to 0.75 for the main and $(1-\pi_1)/(k-1)$ for the other populations, with different sets of $\delta_i \in [-30, 30]$ (δ_i was in general restricted to this range also within the kernel optimization with downhill simplex). All **three options for step** *(ii)* were tested with different settings of other parameters.

The **influence of a decision trigger**, whether trace data should be deconvolved or not, was tested on a data set with known methylation and without shift- or echo effects (see section 2.2.3.2) and on real data of 50% methylated DNA showing shift effects.

3.2.4 Test systems

To assess and optimize the deconvolution algorithm, three test systems were used: Generated trace files (see section 3.2.4.1), data from defined mixtures of methylated and unmethylated DNA in 20% steps, as already used for assessing the basic algorithm (see section 2.2.3.2) and trace files obtained from known mixtures of 50% methylated DNA that showed quality problems based on echos in their signal profiles.

3.2.4.1 Generated trace files

For basic tests of algorithms and for the possibility to generate models used in algorithms, it was necessary to implement a trace data generator. The generator allows the inclusion of different observed characteristics of data obtained from direct bisulfite sequencing, conversion and export as trace file (scf 3.0).The basis of all trace data provided by the generator are four base signals built by sums of peaks. Peaks are intensities from a signals S calculated in an interval a cosine function positioned around different positions p, horizontally stretched by their width w (in signal sampling positions) and

multiplied by their maximum intensity I.

$$S(p+i) = \frac{I}{2} + \cos\left(\frac{2\pi i}{w}\right)\frac{I}{2}; i \in \left[-\frac{w}{2} \ldots \frac{w}{2}\right] \tag{3.14}$$

Thereby the internal handling of such built trace data allows shifting complete four base traces or single base trace information, multiply them by global factors or intensity profiles, add, subtract, randomize[1] or convolve them. This functionality is used to simulate the following observed trace data characteristics:

- Different signal intensities and extremely overscaled signals.
- In silico bisulfite conversion including methylation simulation at CpG positions.
- Decreasing intensity and increasing peak width within proceeding trace sampling time.
- Pseudo-random variation of height and width of signal peaks.
- Addition of pseudo-random noise from other base signals based on a linear distribution and a maximum intensity.
- Shifts and/or echos in single or complete signals.

For generation of defined traces based on genomic sequences of amplificates a specific batch mode was integrated into ESME allowing parameterized definitions of characteristics. Examples for generated trace data can be seen in Fig. 3.3.

[1] To be reproducible, randomization within trace data generation is performed deterministically based on the simple use of *rand* in *C++*.

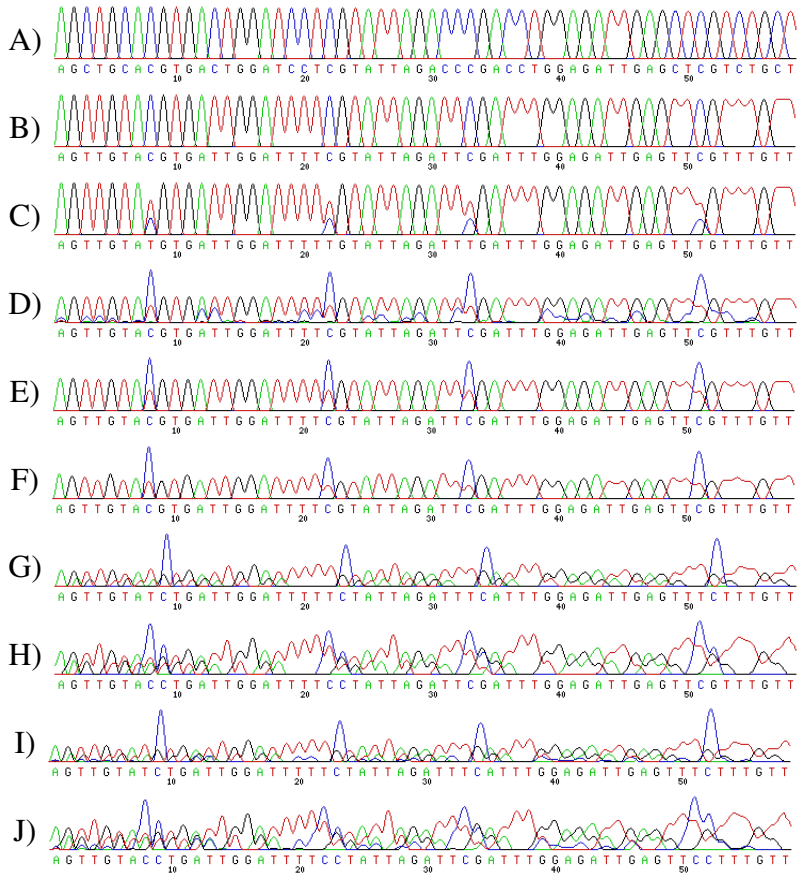

Figure 3.3: Examples for generated trace files based on the sequence AGCTGCACGTGACTGGATCCTCGTATTAGACCCGACCTGGAGATTGAGCTCGTCTGCT with 10 % intensity decrease: **A)**. conversion: 0. methylation: 1. C signal factor: 1. **B)**. conversion: 1. methylation: 1. C signal factor: 1. **C)**. conversion: 1. methylation: 0.33. C signal factor: 1. **D)**. conversion: 1. methylation: 0.33. C signal factor: 6. noise: 10%. **E)**. conversion: 1. methylation: 0.33. C signal factor: 5. **F)**. conversion: 1. methylation: 0.33. C signal factor: 6. variance (h): $\pm$0.2. variance (w): $\pm$0.05. **G)**. conversion: 1. methylation: 0.33. C signal factor: 6. variance (h): $\pm$0.2. variance (w): $\pm$0.05. *shifted* methylation: 15. **H)**. conversion: 1. methylation: 0.33. C signal factor: 6. variance (h): $\pm$0.2. variance (w): $\pm$0.05. convolution kernel: 0δ, 0.5π; 12δ, 0.3π; -6δ, 0.2π. **I)**. conversion: 1. methylation: 0.33. C signal factor: 6. noise: 5%. variance (h): $\pm$0.2. variance (w): $\pm$0.05. *shifted* methylation: 15. **J)**. conversion: 1. methylation: 0.33. C signal factor: 6. noise: 5%. variance (h): $\pm$0.2. variance (w): $\pm$0.05. convolution kernel: 0δ, 0.5π; 12δ, 0.3π; -6δ, 0.2π.

3.3 Results and Discussion

Though convoluted trace data occurs in parts of real data sets and sometimes biases and destroys data for some sites of interest, the frequency is still so low, that is is not easy to get useful data sets based on calibration data with known methylation that shows such phenomena. The main data set for testing and optimization was therefore artificially generated data.

3.3.1 Parameter optimization

Sampling rates in trace data were on average about 12 data points per base. Window sizes for local deconvolution turned out to be trade-offs: the shorter a window, the more likely it is to cover insufficient data, the longer the window, the more likely is it to cover CpGs that are either not co-methylated or are already located in regions that would need different kernels. The step sizes are mainly a trade-off of quality and computing time. We found that a window size of 240 data points (covering approximately 20 bp) and a shift of 200 (resulting in 40 bp overlap) lead to overall stable results.

Kernel optimization with downhill simplex lead into local minima with high frequency, when starting with a kernel allowing for $k = 3$ only, independent of other parameters. Kernels with $k > 4$ did not lead to stable results, which is consistent with the following observation: in most cases the number of additional echos beside the main signal observed in trace data were 1 or 2, in rare cases 3. The use of the iterative process lead to better results. The best setting tested used a starting kernel with

$k_{start} = 11$, $\pi_1 = 0.5$, $\delta_1 = 0$
$\pi_j = 0.05$ and $\delta_j \in (-30, -24, -18, -12, -6, 6, 12, 18, 24, 30)$ for $j \in (2...11)$

and iteratively reduced k by 2 in four steps down to $k_{end} = 3$. The starting kernel covers any echo in distances of half a base around $\pm$ 30 data points of the main signal, the final kernel allows for two additional signal echos.

The third option *c)* for finding a kernel in the algorithmic step *ii* turned out to lead to the best results: first using all signals O_A, O_G and O_Y, and then using the result as initial kernel of a kernel optimization for O_Y only.

3.3.2 Deconvolution of generated data

Generated data to simulate different trace quality for 2,142 CpGs at 6 different methylation levels in 20% steps was used to assess the deconvolution algorithm and compare it to simple analysis with and without signal normalization (see table 3.1 and Fig. 7.2 to 7.8.) In trace data that except for overscaled cytosine signals shows an idealized profile deconvolution leads to comparable results as signal normalization. Adding variance in peak height to the simulation does not significantly changes the outcome, but random noise by small false signals at all positions leads to worse results using the deconvolution algorithm, which suggests, that data without signal echos or shift effects should better be not use deconvolution. As soon as shift effects or echo effects are simulated, the yield of assessable CpG sites, that can still be aligned/associated with reference sequences drops. In these cases deconvolution leads to higher yields and smaller errors when measuring methylation. In the worst case scenarios for shift or echo effects with overscaled cytosine signal, signal variance and noise, deconvolution is still able to enhance methylation measurements significantly. In case of shift effects it is the only method that leads to meaningful data in the range of expected methylation levels of 20 to 80%.

3.3.3 Deconvolution of real data

The use of the deconvolution algorithm on **real data from 50% methylated DNA mixtures** with partially observed **echo and shift effects** enhanced the data quality. The fact that parts of the data were unaffected by shifts/echos (e.g. the first half of the trace file) were reason to test the algorithm with a trigger used in each data window deciding whether to locally deconvolve or not (based on thresholds for a minimum neighbor signal population besides the main signal, either in a kernel found for A, G, Y or for C). The error rate was lowered by applying deconvolution, and became better with use of the trigger (see Fig. 3.4). Signal normalization with local absolute error around 23% using signal normalization only could be reduced to around 13% using the C kernel based trigger. Examples of trace data fore 50% methylation before and after deconvolution are available in Fig. 7.12 to 7.14 and examples demonstrating the influence of deconvolution on shift effects (50% methylated DNA) at data level are shown in Fig. 7.9 to 7.11 in the appendix.

The use of real data from the **test system with known methylation** that

	no normalization			normalization			deconvolution		
simulated	yield	MAE	SD	yield	MAE	SD	yield	MAE	SD
C=5	100	20.0	0.8	100	3.1	1.8	100	2.3	2.3
C=2,v	100	21.0	1.0	100	3.1	2.3	100	3.3	4.3
C=2,n	86	24.0	2.1	88	2.5	3.0	88	4.7	6.0
C=2,v,e	79	18.0	6.8	98	8.3	8.5	100	5.6	6.5
C=2,v,s	67	15.0	6.8	67	15.3	6.6	98	6.2	6.5
C=2,v,n,e	50	20.0	5.9	83	9.8	9.3	86	7.5	7.7
C=2,v,n,s	53	19.0	4.1	72	20.0	5.0	86	7.6	9.0

Table 3.1: Generated trace data from 6 x 100 amplificates (6 x 2,142 CpG sites) with methylation rates from 0% to 100% in 20% steps, influence of different simulation parameters: **C** cytosine signal overscaled in dependence of simulated methylation level; 40 x for 0%, 10 x for 20%, 5 x for 40%, 3.3x for 60%, 2.5 x for 80% and 2 x for 100%, **v** variance in signal peak height $\pm$ 20% and width $\pm$5%, **e** signal echos with $\pi_2 = 0.3$, $\delta_2 = 12$, $\pi_3 = 0.2$, $\delta_3 = -6$, **s** C signal 15 data points shifted to the right, **n** random noise 0-5% false base signals for each base, whereby C signal noise multiplied with factor. The data was analyzed 1. without signal normalization, 2. with signal normalization 3. with data deconvolution. Data in the table shows yield of assessable CpGs [%], mean absolute error from expected measurement MAE [% methylation] and average standard deviation for measurements binned by expectation value [% methylation]. Corresponding boxplots and histograms showing results in dependence of simulated methylation rate are found in Fig. 7.2 to 7.8 in the appendix.

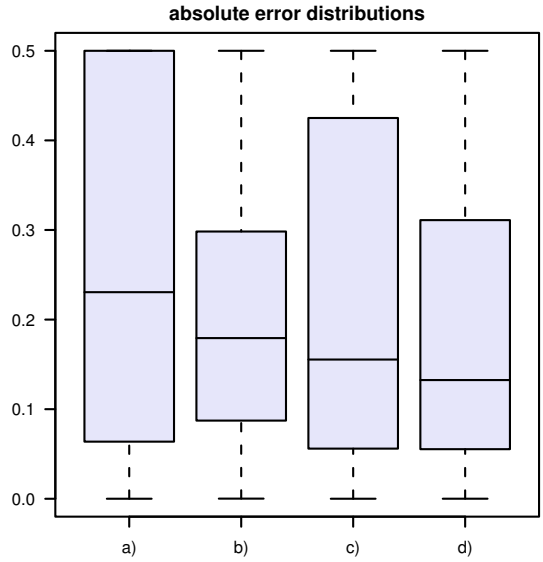

Figure 3.4: Boxplot of absolute Errors in Phi DNA 50% methylation mixtures in G rich sequencing trace data. **a)** not deconvolved, **b)** deconvolved, **c)** partially deconvolved (A, G, Y kernel based decision), **d)** partially deconvolved (C kernel based decision).

showed **no shift or echo effects** but relatively high noise demonstrated that for such data the general use of deconvolution reduces data yield and quality (see table 3.2 and Fig. 3.5). The use of local deconvolution with formally tested parameters lead to better error, but still to high loss of the data, which shows that there might be either more parameter optimization necessary or a completely different approach to a priory detect and deconvolve data with echos and shift effects only.

	N	MAE	SD
equalized, normalized and corrected	1998	0.1256	0.1364
deconvolved	954	0.1723	0.1871
partially deconvolved	1320	0.1358	0.1890

Table 3.2: Influence of algorithmic steps. Amount of positions measured (N), mean absolute error (MAE) and standard deviation (SD) in 20% step methylation mixture calibration data. The first

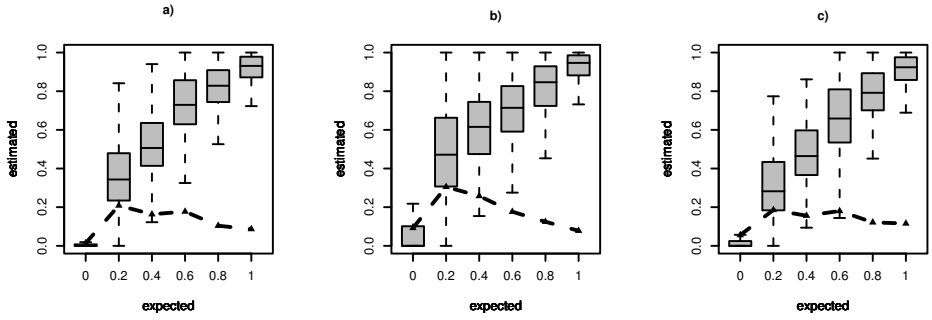

Figure 3.5: Estimation of methylation in the test system with known methylation rates. The boxplots show the distribution of the estimated methylation rates as a function of the expected methylation and the mean absolute error (dashed line). Each box includes data from CpGs of all 60 amplificates measured at the expected methylation rate. Data from different algorithmic steps: **a)** equalized, normalized and corrected (without deconvolution) **b)** deconvolved **c)** partially deconvolved (decision based on *C* kernel).

3.4 Conclusions and Outlook

Only a small portion of trace files from direct bisulfite sequencing show systematical echos or shift effects due to molecule populations with different mobility in the electrophoresis. For these data the deconvolution algorithm presented in this proof of principle is a method that can significantly enhance the data quality - even for data that any basecaller or visual inspector would discard for quality reasons. In the case of data that has no need to be deconvolved, the application of the algorithm does not improve the data and reduces the yield. Therefore, further improvement of the algorithm to better detect data in need of deconvolution would be required if one wanted to apply it more generally. At the time being four facts lead to the decision to not enhance the algorithm further and include it as a default step when analyzing trace data: **1.** the fact that the observation of the cured phenomena in the data is rare and therefore plays a minor role; **2.** the success of projects on the the wet lab side increasing data quality (not part of this work); **3.** the very restricted amount of real data available to train/improve the algorithm to have a reliable and stable test system that allows a release for big or even commercial projects; **4.** the highly increased computing time.

Despite this decision, the method could well be adapted and enhanced in case of new big data sets of trace files that have a high portion of shift/echo effects. This might be the fact for technical reasons e.g. due to specific tissue sample treatment and/or choice of genomic sites with a bias to DNA sequences tending to these effects.

Chapter 4

Methylation data analysis for the HEP

In this chapter the Human Epigenome Project (HEP) data, the first worldwide large scale high resolution data set for methylation at CpG level with 1,885,000 measurements profiling 12 healthy tissues on three chromosomes, and a pilot study data set based on 6 healthy tissues, are assessed to address many questions about DNA methylation and to find answers and theories for: association with genomic functionality, spacial profiles, influence of CpG density and evolutionary conservation, differential methylation, comethylation behavior of CpGs and more. Several content of this chapter will partially overlap with publications of results of the HEP (Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004), which were mainly based on data analysis performed in this work.

4.1 Motivation and Theory

The previous two chapters covered technical aspects that enable high resolution DNA methylation studies by direct bisulfite sequencing, and the optimization thereof. The motivation behind the development of these methods was their use in large-scale projects concerning methylation data, and addressing biological questions related to methylation profiling and the behavior of the epigenetic methylation layer. This chapter will address the following questions using data based on 43 healthy tissue samples and unique DNA sequences in regions of interest (ROIs):

- **Functional methylation:** The methylation of genomic sites is associated with chromatin density, accessibility of the DNA for proteins and transcription regulation. Is CpG methylation different in exons, introns, or around transcription start sites? What is the spatial profile of methylation within these functional regions? What role does the promixity transcription factor binding sites play? Are there differences between known and predicted genes? What are the chromosomal methylation profiles of measured ROIs?

- **CpG density/islands:** About half of all promoter neighborhoods contain CpG dense regions. How does CpG density influence methylation profiles? To what extent is it correlated and does it influence functional methylation?

- **Tissue specific methylation:** Different tissues have different functions, proteomes and transcription profiles. How frequently is differential methylation observed between different tissue types? Is it affected by including or excluding not yet fully differentiated and/or specific tissue types such as fetal tissue, sperm and placenta? Where are differentially methylated sites found?

- **Global methylation changes with proceeding age:** How does age influence global methylation in non repeat covering regions? Does it increase or even decrease significantly?

- **Autosomal sex specific differential methylation:** Is there an autosomal global methylation difference between males and females ? Do we find sites with significant differential methylation between males and females?

- **Sequence homology and methylation:** Evolutionary conserved regions with higher CpG content found by homology between mouse and human might be conserved due to functional reasons that are related to their CpGs, especially outside of the gene context. If differential methylation could be regarded as evidence for functional methylation, do conserved sites show more or less differential methylation between tissues?

- **Cell differentiation:** Methylation is suspected to play an important role in cell differentiation. Is there differential methylation and, if so, how strong is it between closely related differentiated tissues and their fetal successors?

- **Co-methylation:** Within a certain number of bases and between certain boundaries, CpGs might be organized in co-methylated blocks (ComBs) and show similar methylation behavior due to functional reasons or the mechanisms controlling their methylation state (see 4.1.1). How large are these blocks and the distances between CpGs to be expected to be co-methylated? In what content does CpG density influence co-methylation?

4.1.1 Co-Methylation

The methylation profiles of CpGs within a certain distance in bases or in proximity based on the DNA three dimensional structure are most probably not independent. In most cases adjacent CpGs will have a correlated methylation state; they are co-methylated. Two main factors will influence the methylation behavior within a short distance: 1. association of CpGs to the same biologically functional group, 2. mechanisms for methylation and demethylation addressing whole stretches, e.g. SS1 methylase binding the DNA at a certain position, wandering along the strand, processing a part and then detaching. The data obtained by bisulfite sequencing allows the assessment of the co-methylation structure of CpG methylation over short and long distances.

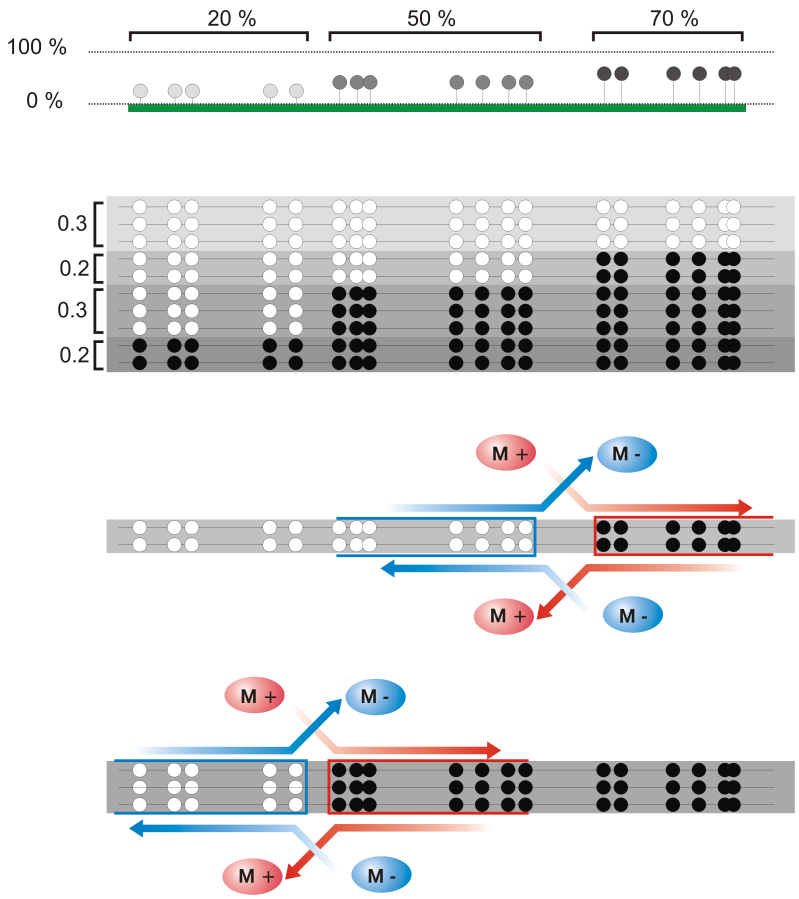

Figure 4.1: **Simplified model for co-methylated blocks (ComBs)** given no mosaic patterns: CpGs observed in co-methylated blocks showing e.g. 20 (light gray lollipops), 50 (gray) and 70% methylation (dark grey) in average over all molecules. The blocks are based on different molecule populations from different cells with unmethylated (white circles) and methylated CpGs (black). The molecules or their ancestors from parental cells must at one time have been supplied with the pattern by a methylating mechanism (methylase) or a demethylating mechanism, that worked along a specific stretch of DNA and docked/started or undocked/stopped at a given position within the boundary of two ComBs. The blocks are defined by the mechanisms and the limiting boundaries, whereby the boundaries might be based on DNA sequence patterns or complex factors like spacial behavior of the secondary structure.

4.2 Materials and Methods

4.2.1 HEP pilot study

It was part of this study to find a simple strategy for choosing locations within the MHC region of the human chromosome 6 to be used within the pilot study of the HEP. Locations for methylation profiling in the human MHC were chosen in 255 regions that are in the context of genes, based on a corrected and annotated draft of the MHC region based on IHGSC, 2001. These location, called *Regions Of Interest* (ROIs), are 2.5 kb in size and can be divided into two groups:

1. 5'-UTR promoter related ROIs located from 2 kb upstream to 500 bp downstream of the transcription start site (TSS). Genes that shared promoter regions were represented together only once. For genes with multiple annotated TSS the first was used (based on 5' position).

2. Intragenic ROIs covering 2.5 kb fragments with the highest CpG density located from 500 bp after TSS down to the end of the gene. For longer genes more than one of these ROIs was designed. For each gene the amplificate with the highest CpG density was always used, the others were optional and ranked by their CpG density.

The following steps after study design and before trace file data analysis were not the author's work and either performed at Epigenomics AG in Berlin using proprietary methods and protocols or at Welcome Trust Sanger Institute (WTSI) in Cambridge UK. Amplificate design within the ROIs was performed with *Epigenomics'* proprietary in house software (Rujan et al.)[1] allowing a maximum amplificate length of 500 bases. For each ROI where it was possible we used the amplificate with the most CpGs that lead to PCR products on bisulfite converted human Lymphocyte DNA (Promega) but did not lead to products using unconverted DNA. This selection criteria resulted in 253 amplificates that were used in the study.

DNA was extracted from 30 tissue samples representing six tissue groups (number of samples per group given in brackets): brain (5), breast(6), liver(2), lung (5), muscle (4), prostate (8). Bisulfite conversion of DNA and PCR with

[1] patent number DE 102 36 406; "Verfahren zur Amplifikation von Nukleinsäuren mit geringer Komplexität"; Inventors: T.Rujan, Ch.Piepenbrock, A.Schmitt, P.Adorjan

conversion product specific primer pairs were performed at Epigenomics. For each PCR product two direct sequencing reactions and runs from both sides of the amplificate using one of the PCR primers each were performed at WTSI.

Trace files resulting from sequencing were used as raw data to gain methylation information using Epigenomics' software ESME that uses an implementation of the algorithms described in this work (see section 2.2.1). Trace file analysis was done in parallel at WTSI and at Epigenomics. All IT steps, conception, realization and application for this pilot study, except for the amplificate design, were performed by the author. Data interpretation presented in this work used ESME results in tab delimited tables, and in one case MALDI-TOFF data provided from external partners (CNG, Paris) and the statistical script language *R*.

4.2.2 HEP study on chromosome 6, 20 and 22

For the first workpackage of the HEP, amplificates were designed for 2.5 kb regions of interest (ROIs). Choice of more than 2/3 of the ROIs used the same strategy as in the pilot study (see section 4.2.1). Additional ROIs in *evolutionary conserved regions* (ECR) were chosen by a minimum of 70% DNA sequence similarity between mouse and human, preferably but not exclusively in non coding intergenic or intronic sites (see table 4.5). Amplificate design was performed as in the pilot study. The source and handling of cells and tissue samples, the amplicon selection and classification, DNA extraction, PCR amplification and sequencing that lead to the data assessed in this work is described in (Eckhardt *et al.*, 2006). Raw data processing of trace files to gain DNA methylation data was performed with Epigenomics' software ESME a C++ implementation that uses the algorithms previously described in this work (see section 2.2.1). All sequencing raw data processing, data classification by mapping to CpG islands, exons, introns, TSS, and all data interpretation shown in this work and in (Eckhardt *et al.*, 2006) was performed by the author.

4.2.3 Data Interpretation

Methylation data was assessed at different levels of aggregation: CpG wise, amplificate wise, tissue sample wise, tissue wise. The basic data set with the highest resolution contains methylation rates for each CpG in every tissue

sample and is calculated by averaging (median) over all technical repetitions for each CpG/tissue sample. CpG wise methylation for tissues is calculated by averaging (mean) CpG data from the basic data set over all tissue samples from identical tissue origin. Amplificate wise methylation for tissue samples is calculated by averaging (mean) CpG/sample data from the basic data set over all CpGs from identical amplificates. Amplificate wise methylation for tissues is calculated by first averaging (mean) CpG data from the basic data set over all tissue samples from identical tissue origin and second averaging (mean) over all tissue samples from identical tissue origin. Data was mapped to genomic annotations obtained from the ENSEMBL database (Curwen *et al.*, 2004) by chromosomal coordinates: genes, tss, exons, CpG islands. CpG densities associated to regions were calculated within 500 bp windows around the center of amplificates.

4.2.3.1 Data quality

In the HEP work package 1 methylation data quality was assessed by analysis of repetitions. There were three different kinds of repetitions available:

- technical repetitions of the same amplificate/sample/sequencing strand.
- repetitions of the same amplificate/sample PCR product but sequenced on the reverse complement strand.
- CpG measurements for the same CpG and sample, based on different overlapping amplificates.

For all three kind of repetitions correlation and differences in measurement were assessed.

4.2.3.2 Function associated methylation behavior

Methylation data is associated with annotated functions based on distances of CpGs to specific genomic positions, positions of CpGs between start and end of an annotation and in case of average amplificate methylation overlap of the amplificates with annotated regions were used. The annotations came from ENSEMBL (Curwen *et al.*, 2004) and for gene types from VEGA annotations (Ashurst *et al.*, 2005).

Most genes lack a clearly defined and biologically confirmed location of their promoter. Transcription start sites (TSS) on the other hand can be predicted with high accuracy (Down & Hubbard, 2002). We therefore used regions around the located 2000 bases upstream and 500 bases downstream of 5' UTR to define **promoter associated regions**.

The same definitions as used in the choice of regions of interest (ROIs) for promoter sites (see section 4.2.1). For detailed assessment of different **introns and exons**, data was grouped for exon/intron 1,2,3 and as all numbered 4 or higher, for detailed assessment those groups were binned into thirds to assess 5' anterior, middle and 3' rear part. Methylation and average methylation associated with different functional groups were displayed in profiles (see also 4.2.3.6), histograms, and described in tables. Bimodal distributions of methylation measured in healthy tissue (mainly distributed around 0 and 100%) were also assessed for different functional groups looking only for **strong hypo-methylation** (<= 10%) and **strong hyper-methylation** (>= 90%).

For **CpG island definitions**, we used a slightly modified version of the definition by (Bird, 1986): a GC content of 50%, a ratio of observed to expected CpGs of 0.6 b and a minimum length of 400 bp (instead of 200 bp as in the cited definition).

4.2.3.3 Differential methylation

Kruskall-Wallis tests on amplificate wise methylation data were used to determine differential methylation between tissues. Some sites found to be differentially methylated between tissues represented by only few tissue samples were experimentally validated by sequencing of independent DNA samples. For global estimations of differential methylation occurrence (not for statistical significance of single sites), we defined the amount of p-values smaller than 0.05 as a good estimate for the amount of differential methylation and checked whether one would find similar results with resampled data. To do so, the ratio of p-values from Kruskall-Wallis tests for all amplificates below a certain threshold (0.05) was compared to the distribution of 1000 analogous ratios obtained by resampling of the data set (based on sampling of the tissue annotation). Data obtained from tests and resampling over the complete data set were split into subgroups based on the genetic context of amplificates to assess, if occurrence of tissue specific differential methylation is correlated

by DNA functionality.

Data for two group comparisons (e.g. separated by age or sex) was filtered by samples: related differences were calculated intra-tissue wise only for tissues that were covered by enough samples to provide groups. In case of age, data was used only for patients up to age 35 (young) and starting at age 60 (old). Differential methylation by tests for equality (null hypothesis to be rejected) of two groups were performed using Wilcoxon tests. CpG wise methylation differences between tissues or groups were calculated as mean of all differences between the group averages (mean) for all CpGs. Amplificate wise methylation differences between tissues or groups were calculated as mean of all differences between the group averages (mean) for all amplificates. Unsupervised clustering of methylation data is based on manhattan distances.

4.2.3.4 Homology between mouse and human DNA

The epigenome of mouse and human is reported to be even conserved on histone level (Bernstein *et al.*, 2005). We therefore assessed our findings for differential methylation for all amplificates in different regions with respect to human mouse homology. Homology between mouse and human DNA sequences was calculated as the amount of identical bases in a optimal local alignment of the human amplificate and the best blast result on the mouse genome (extended on both sides to fit the amplificate length) divided by the full length of the amplificate (not only the aligned part). Homology and p-values found for differential methylation (see 4.2.3.3) were binned by associated annotations and compared.

4.2.3.5 Co-Methylation

For analysis of co-methylation, the data set was filtered to exclude technical outliers. Medians of CpGs methylation measurements for the same CpG, tissue sample and sequencing strand were restricted to those available from both strands, with a maximum absolute difference of 10% between measurements from the two strands. Based on this criterion, 38% of CpGs were excluded from the analysis. After filtering the average methylation between both strands were sued for further analysis. Methylation changes were calculated as absolute differences between all available pairs of different CpG positions in the same sample within a given window of 20.000 bases and within

identical amplificate. For some analyzes the resulting data set was either restricted to differences from adjacent CpG position pairs (no other CpGs in the sequence in between) and/or to those gained within the same amplificate (restricting the maximum distance to maximum amplificate length).

Co-Methylation of CpGs was described as function of the distance in bases displaying either the observed ratio of equal methylation behavior (defined by an absolute difference <= 10 %), as ratio of observed methylation changes (defined as an absolute difference >= 25%) or as observed average absolute methylation difference. Long range co-methylation was assessed between amplificates using identical methods as for CpGs after averaging measurements for each amplificate and sample.

Results were compared with two resampled data sets: 1. Complete resampling: after filtering methylation methylation values within identical samples were randomly resampled. 2. Resampling of chromosomal start positions of amplificates: preserving amplificate internal distances and methylation values (not used for co-methylation between or within amplificates).

4.2.3.6 Profiles of methylation or co-methylation

In this work there are many calculations describing and visualizations displaying tendencies in profiles. The profiles are based on

- **absolute positions**i n bases in a coordinate system like chromosomal position or distances to a TSS.
- **relative positions** between 0 and 1 to describe the location from begin to end of for example an exon/intron which allows overviews over multiple entities with different sizes.

In case of boxplots (and histograms), data was binned by coordinate intervals of same size containing different number of measurements. In cases of dotplots the data was binned into intervals of different sizes but with equal amount of measurements in each bin: data was first numerical ordered by the x-axis values, sorted into bins of the desired size (e.g. 1000 data points), and then represented by means over x and y data. For methylation, these values in most cases provide the ratio of unmethylated and methylated CpGs, which represent bimodal distributions with values that are mainly around 0 and 100%.

4.3 Results and Discussion

4.3.1 Data set overview

4.3.1.1 HEP pilot study data set

A statistical summary for the pilot project is given in the following table 4.1.

	n	sd
Amplicons analyzed	253	
Amplicons in 5'-UTR region	72	
Amplicons in intragenic regions	181	
Mean amplicon length [bp]	438	65
Maximum amplicon length [bp]	500	
Minimum amplicon length [bp]	171	
Mean Number of CpGs/amplicon	13	8
Mean G+C content of amplicons	0.56	0.0703
Amplicons with 200 bp window fulfilling CpG island definition	82	
Amplicons fulfilling CpG island definitions over the whole sequence	23	
Ratio of assessable trace files [%]	81	
Mean alignable part of trace files [bp]	339	122
Ratio CpG sites observed/CpG sites expected per amplicon	0.824	0.11
CpG measurements total	134065	
Unique CpG sites analyzed	3273	
Ratio of unique CpG sites with methylation differences > 20%	0.81	
Ratio of unique CpG sites with methylation differences > 50%	0.45	
Ratio of CpGs with methylation differences > 20% between tissues	0.41	

Table 4.1: Overview over data in the **pilot project of the Human Epigenome Project (HEP)** focusing on the Major Histone Complex (MHC) located in chromosome 6.

4.3.1.2 HEP work package 1 data set

A short summary for the dimensions of the data is given in table 4.2. The coverage of functional annotations in the genome is described in table 4.3. Overall methylation measured, amplificate lengths and CpG content are displayed in Fig. 4.2. More details of annotations associated with amplificates is given in Fig. 4.3 to 4.5 and the corresponding tables 4.4 to 4.6.

	n	sd
Amplicons analyzed	2,524	
Mean amplicon length [bp]	411	77
Mean Number of CpGs/amplicon	16	10
CpGs measured (exclusive unassessable CpGs)	40,386	
Tissues assessed	12	
Tissue samples used	43	
trace files assessed	217,243	
CpG raw data measurements	1,885,003	
CpG/sample measurements	1,271,004	

Table 4.2: Overview over data in the **HEP work package 1 of the Human Epigenome Project (HEP)** providing data for chromosome 6, 20 and 22. Initially the data set included two more tissue samples, which were removed from the study due to DNA amount problems.

Table 4.3: HEP work package 1: Overlap of measurable CpG sites and functional groups within chromosomes 6, 20 and 22. A site is counted as overlapped/covered if there is at least one measurable CpG within the location.

	6	20	22	All
CpG islands in Chromosome(s)	1070	662	547	2279
CpG Islands covered	256	29	226	511
CpG Islands percentage covered	24	4	41	14
CpGs in CpG islands	7392	1016	8425	16833
Measurements in CpG islands	279167	39974	372431	691572
genes covered	383	401	89	873
CpGs in Genes	14071	8042	2376	24489
Measurements in genes	710062	303212	117275	1130549
exons covered	454	376	23	853
CpGs in exons	6682	3974	176	10832
Measurements in exons	352494	150129	7398	510021
introns covered	465	337	118	920
CpGs in introns	7389	4068	2200	13657
Measurements in introns	357568	153083	109877	620528
TSS sites 2500 upstream covered	186	230	31	447
CpGs in TSS 2500 upstream	4743	4171	467	9381
Measurements in TSS 2500 upstream	214775	157799	19823	392397
CpGs in **complete data set**	21672	13072	5642	40386
Measurements in complete data set	1075938	524227	284838	1885003

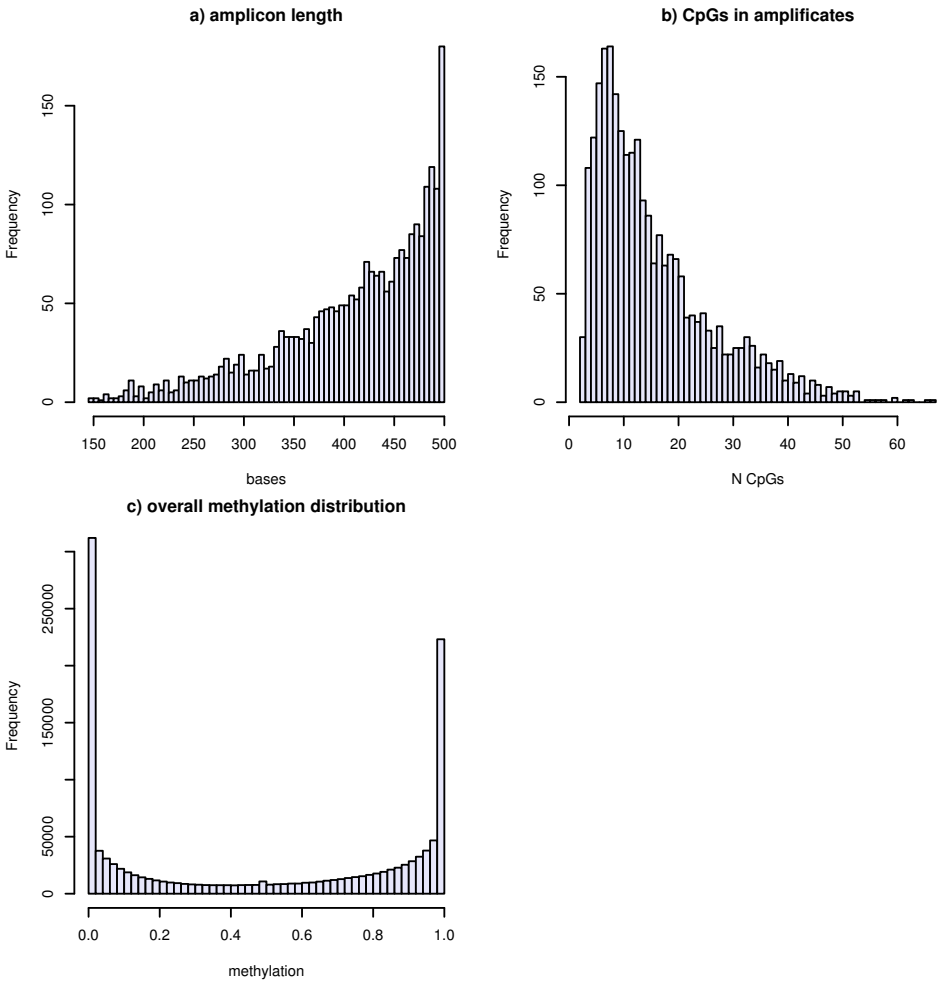

Figure 4.2: HEP work package 1 data characteristics. a) Length of the 2,524 amplificates used in the HEP (411 $\pm$ 77 sd in average.) b) Distribution of 40,386 CpGs in the amplificates (16 $\pm$ 10.8) c) Distribution of all 1,271,044 CpG/sample methylation measurements in chromosome 6, 20 and 22 based on 1,885,003 raw data measurements that were generated with 217,243 trace files used with DNA from 45 tissue samples.

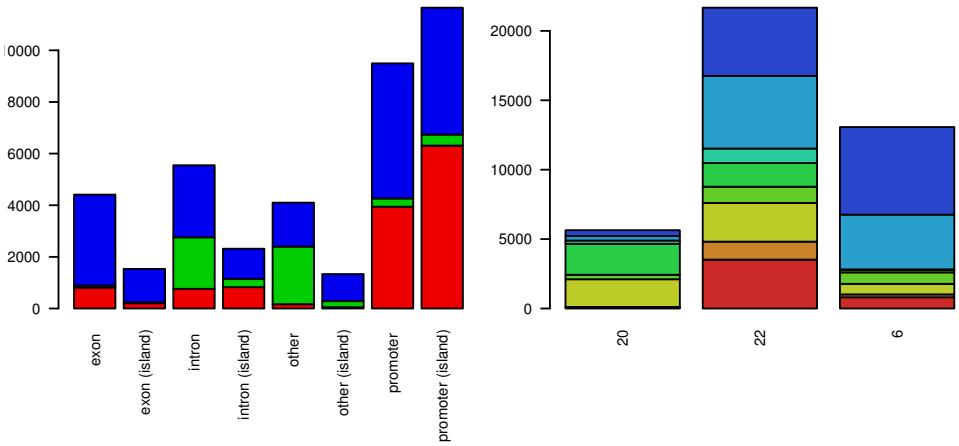

Figure 4.3: HEP work package 1, chromosome wise distribution of CpG sites in exons, introns and potential promoter regions (TSS associated) in and out of CpG islands. Left plot: Chromosomes ■ 6 ■ 20 ■ 22. Right plot: Annotated functions ■ exon ■ exon (island) ■ intron ■ intron (island) ■ other ■ other (island) ■ promoter ■ promoter (island). Data corresponds to table 4.4.

Table 4.4: HEP work package 1, chromosome wise distribution of CpG sites in exons, introns and potential promoter regions (TSS associated) in and out of CpG islands. Corresponding data to Fig. 4.3.

	20	22	6	**Sum**
exon	79	3522	810	4411
exon (island)	33	1295	205	1533
intron	1996	2790	761	5547
intron (island)	321	1167	827	2315
other	2231	1703	167	4101
other (island)	236	1047	50	1333
promoter	320	5232	3942	9494
promoter (island)	426	4916	6310	11652
Total	5642	21672	13072	40386

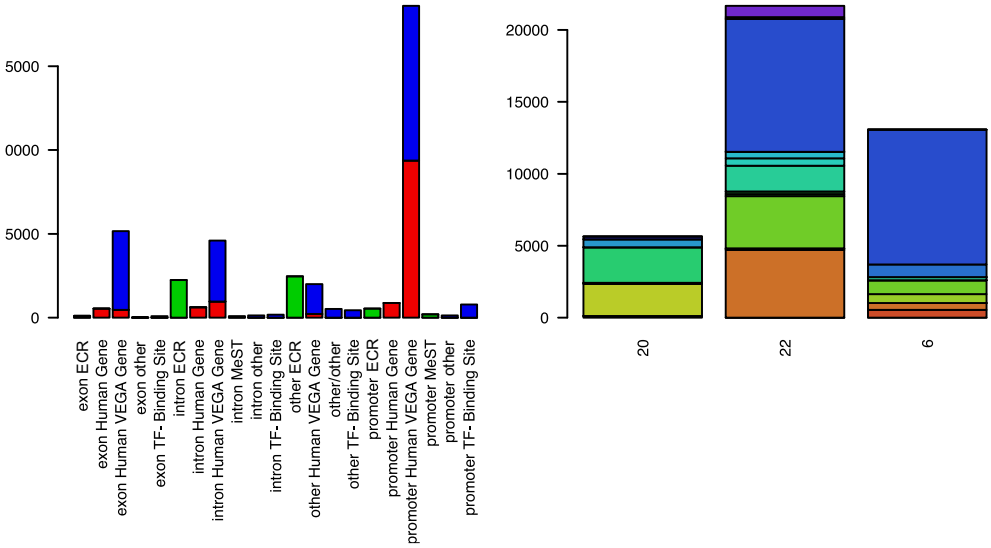

Figure 4.4: HEP work package 1, chromosome wise distribution of CpG sites in exons, introns and potential promoter regions (TSS associated) sorted by reason of their choice: gene association, evolutionary conserved regions (ECR), methylation sensitive tag (MeST), transcription factor binding site or other. Left plot: 6 20 22. Right plot: exon ECR exon Human Gene exon Human VEGA Gene exon other exon TF-Binding Site intron ECR intron Human Gene intron Human VEGA Gene intron MeST intron other intron TF-Binding Site intergenic ECR intergenic Human VEGA Gene intergenic other intergenic TF-Binding Site promoter ECR promoter Human Gene promoter Human VEGA Gene promoter MeST promoter other promoter TF-Binding Site. Data corresponds to table 4.5.

Table 4.5: HEP work package 1, chromosome wise distribution of CpG sites in exons, introns and potential promoter regions (TSS associated) sorted by reason of their choice: gene association, evolutionary conserved regions (ECR), methylation sensitive tag (MeST), transcription factor binding site or other. Corresponding data to Fig. 4.4.

	20	22	6	**Sum**
exon ECR	112	0	0	112
exon Human Gene	0	0	548	548
exon Human VEGA Gene	0	4700	467	5167

Table 4.5: HEP work package 1, chromosome wise distribution of CpG sites in exons, introns and potential promoter regions (TSS associated) sorted by reason of their choice: gene association, evolutionary conserved regions (ECR), methylation sensitive tag (MeST), transcription factor binding site or other. Corresponding data to Fig. 4.4.

	20	22	6	**Sum**
exon other	0	29	0	29
exon TF-Binding Site	0	88	0	88
intron ECR	2242	0	0	2242
intron Human Gene	0	0	624	624
intron Human VEGA Gene	0	3638	964	4602
intron MeST	75	0	0	75
intron other	0	138	0	138
intron TF-Binding Site	0	181	0	181
other ECR	2467	0	0	2467
other Human VEGA Gene	0	1783	217	2000
other other	0	521	0	521
other TF-Binding Site	0	446	0	446
promoter ECR	547	0	0	547
promoter Human Gene	0	0	879	879
promoter Human VEGA Gene	0	9233	9373	18606
promoter MeST	199	0	0	199
promoter other	0	130	0	130
promoter TF-Binding Site	0	785	0	785
Total	5642	21672	13072	40386

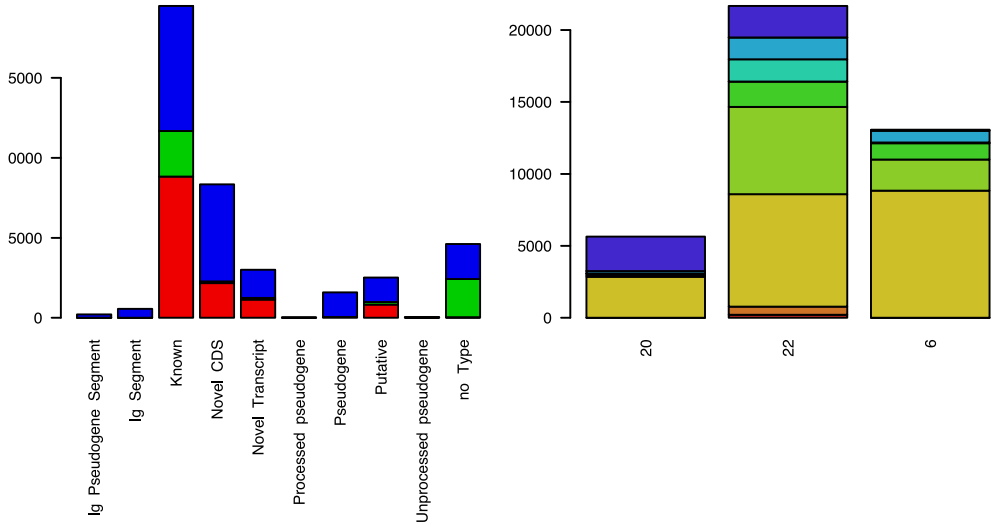

Figure 4.5: HEP work package 1, chromosome wise distribution of CpG sites in different types of gene annotations. Left plot: 6 20 22. Right plot: Ig Pseudogene Segment Ig Segment Known Novel CDS Novel Transcript Processed pseudogene Pseudogene Putative Unprocessed pseudogene no Type. Data corresponds to table 4.6.

Table 4.6: HEP work package 1, chromosome wise distribution of CpG sites in different types of gene annotations. Corresponding data to Fig. 4.5.

	20	22	6	**Sum**
Ig Pseudogene Segment	0	209	0	209
Ig Segment	0	564	0	564
Known	2854	7817	8830	19501
Novel CDS	109	6066	2164	8339
Novel Transcript	107	1767	1134	3008
Processed pseudogene	5	0	13	18
Pseudogene	0	1540	48	1588
Putative	169	1525	822	2516
Unprocessed pseudogene	6	0	25	31
no Type	2392	2184	36	4612
Total	5642	21672	13072	40386

4.3.2 HEP work package 1 data quality

The number of purely technical repetitions in this study (15,655 data pairs) was comparably small to the number of repetitions by reverse strand sequencing (557,837 data pairs): technical repetitions were only done at the beginning of the study (working on chromosome 22) and were not originally planned, reverse sequencing was planned for all amplificates throughout the study. Overlapping amplificates (91,528 data pairs) mainly occurred due to overlap of biological functions (e.g. genes) that were both covered by amplificates.

The correlation of the three types of repetitions are: 0.9 for technical repetitions, 0.87 for for reverse strand sequencing and 0.85 for repeated measurements for CpGs in overlapping amplificates. Correlation scatterplots, correlated data binned in boxplots, distributions of differences and median methylation measured for the three repetition data subsets can be found in Fig. 4.7, 4.8 and 4.9.

The observed correlations fulfill our expectations: purely technical sequencing of identical amplificates on the same strand is expected to lead to trace files with identical systematical technical biases and artifacts (as far as existing) and is therefore expected to correlate best. Sequencing results from the reverse strand were slightly better correlated than we expected: CpGs are assessed by G/A signals on the reverse strand of bisulfite converted DNA instead of C/T signals on the sense strand, trace files contain positions assessed in reverse order showing opposite decreasing signal intensity and different individual systematical trace profiles. Despite these technical differences still 62% of the data pairs showed methylation differences <= 10%. Sequencing results from overlapping amplificates do not only come from a different region sequenced with different primers but also are based on another PCR product with different primers and potentially a slightly different PCR bias (as far as existing).

Methylation data in this study shows bimodal distributions around 0 and 100%, which could in theory be based on biases in the technical method. Comparison of methylation data from sequencing with comparable data from other technical methods shows identical profiles. An example from the pilot study is seen in Fig. 4.6, where data from direct sequencing is compared with methylation data measured on a MALDI-TOFF. Though other technolo-

gies might themselves not be a golden standard, we expect bimodal data most likely to be based on biology and not on artifacts in our method.

The overall performance and technical variation of methylation measurement in this study lead to a reliable data set to answer most of our questions. In cases were light changes in methylation played an important role in data interpretation (co-methylation), we used medians of data measured on both strands with less than 10% methylation difference.

A compact graphical overview over the complete data of the work package over all amplificates, tissues and chromosomes is given in Fig. 4.10, including chromosomal profiles for CpG density, fragment coverage and gene density. Extended versions of these figures are found in the appendix: Fig. 7.15 to 7.17. Spacial methylation profiles on chromosomes for methylation data measured for different tissue types are found in Fig. 7.18 but are most likely biased by our design criteria and not representative for the chromosomes.

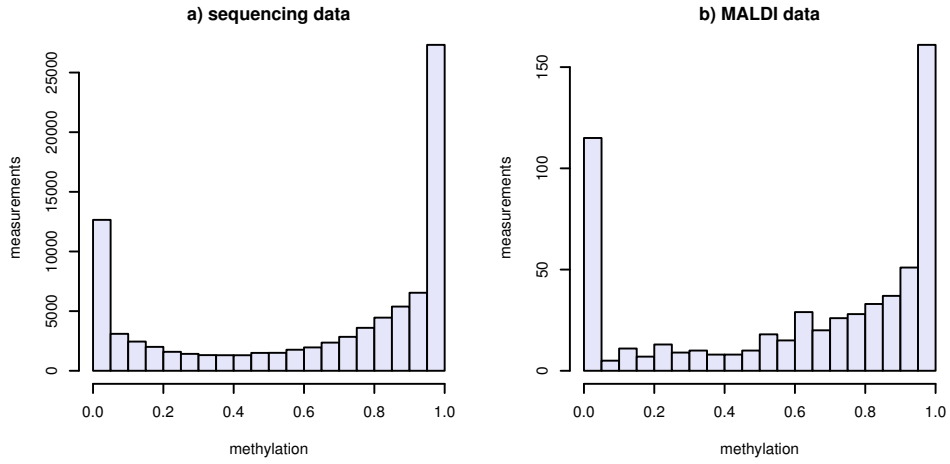

Figure 4.6: Bimodal distributions of **CpG methylation** measurements from **a) sequencing/Esme analysis** and **b) MALDI-TOFF**. a) Methylation based on 86,374 single CpGs in different tissue samples (median for technical repetitions). b) Methylation based on 614 MALDI-TOFF measurements.

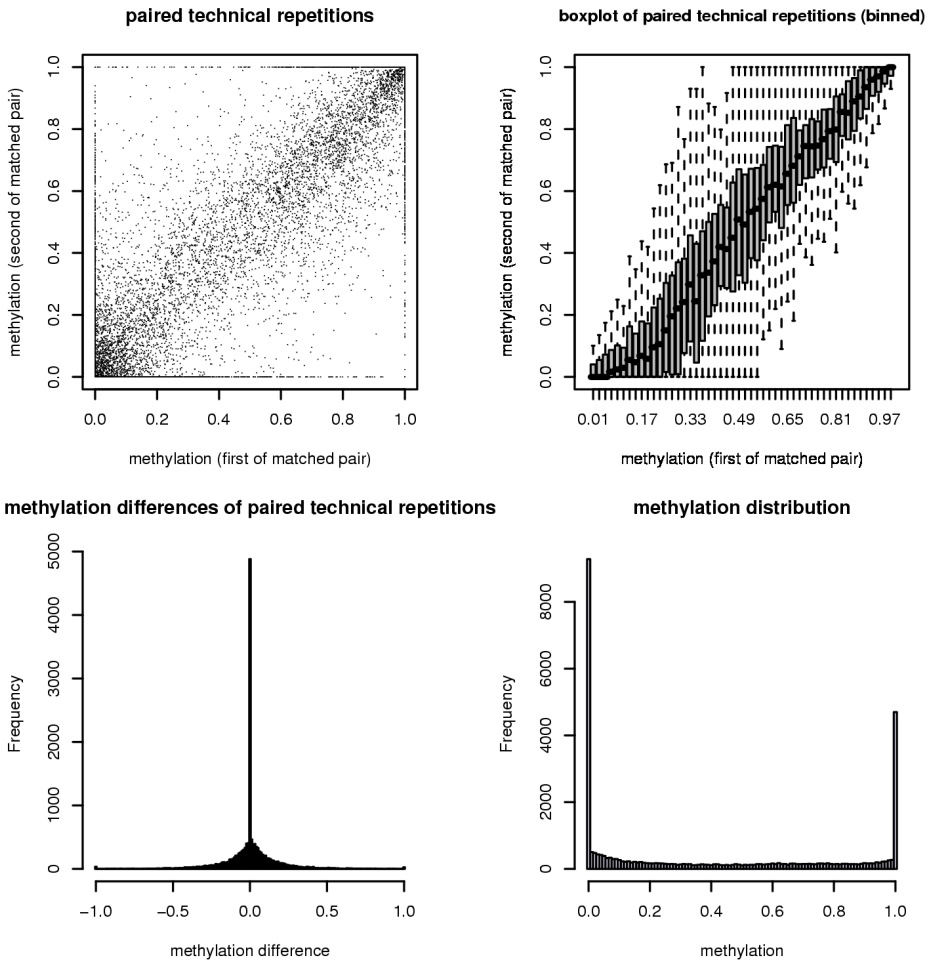

Figure 4.7: **Technical repetitions**, based on 15,655 data pairs which have identical sequencing strand, sample DNA, amplificate and CpG but were measured in independent technical repetitions. The correlation was 0.9. **Top left: scatter plot** of repetitions. **Top right: boxplots** of correlated methylation data pairs binned into 50 groups. **Bottom left: Differences** of methylation measurements of pairs. **Bottom right: Methylation distribution** of the subset of the data covered by the repetitions used in this analysis.

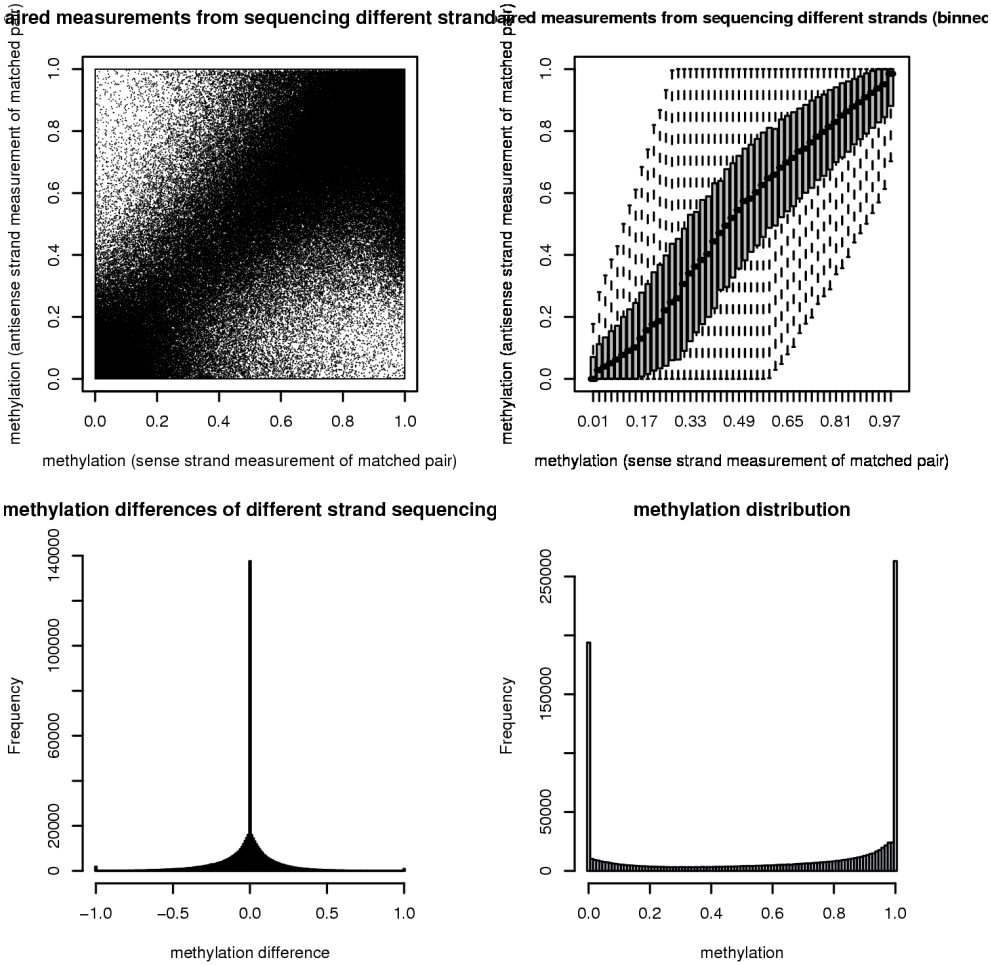

Figure 4.8: **Comparison of sequencing strand based data pairs**, based on 557,837 data pairs, which have identical sample DNA, amplificate and CpG but were measured either by sequencing the cytosine poor strand from PCR after bisulfite conversion or the guanine poor reverse strand. The correlation was 0.867. **Top left: scatter plot** of repetitions. **Top right: boxplots** of correlated methylation data pairs binned into 50 groups. **Bottom left: Differences** of methylation measurements of pairs. **Bottom right: Methylation distribution** of the subset of the data covered by the repetitions used in this analysis.

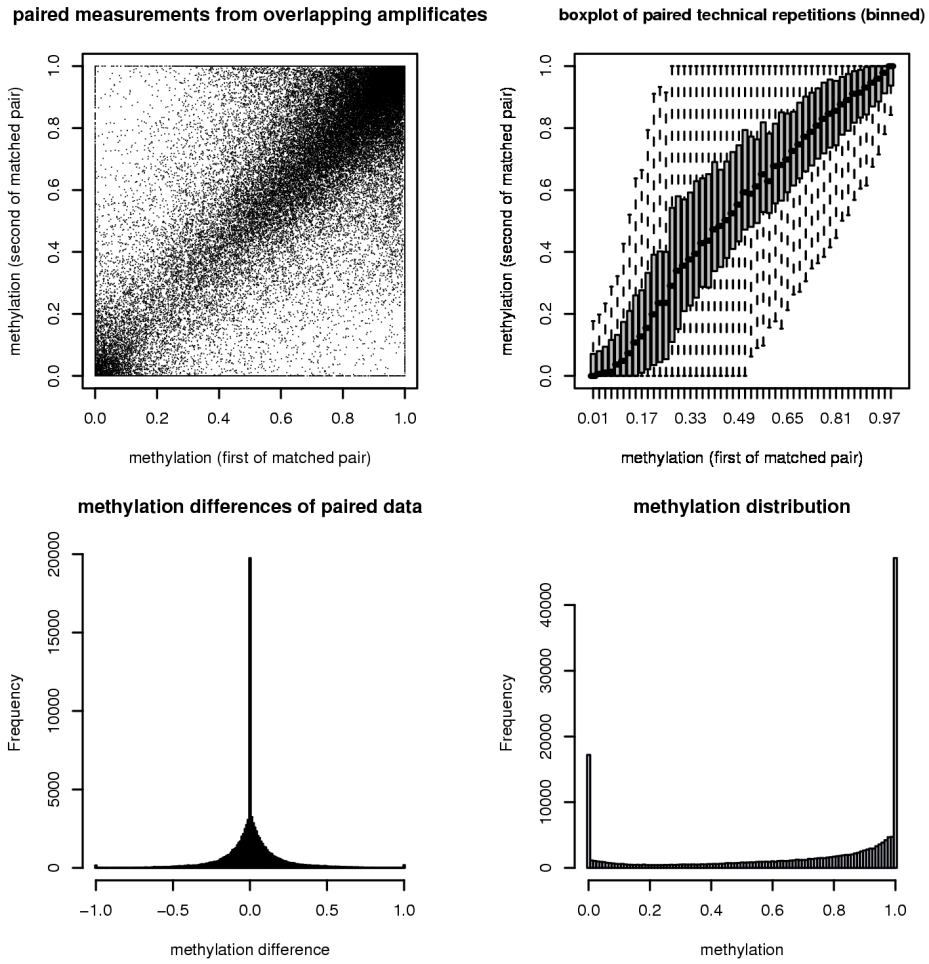

Figure 4.9: **Measurement repetitions by overlapping amplificates**, based on 91,528 data pairs which have identical sequencing strand, sample DNA and CpG but were measured in different overlapping amplificates that covered the identical CpG. The correlation was 0.85. **Top left: scatter plot** of repetitions. **Top right: boxplots** of correlated methylation data pairs binned into 50 groups. **Bottom left: Differences** of methylation measurements of pairs. **Bottom right: Methylation distribution** of the subset of the data covered by the repetitions used in this analysis.

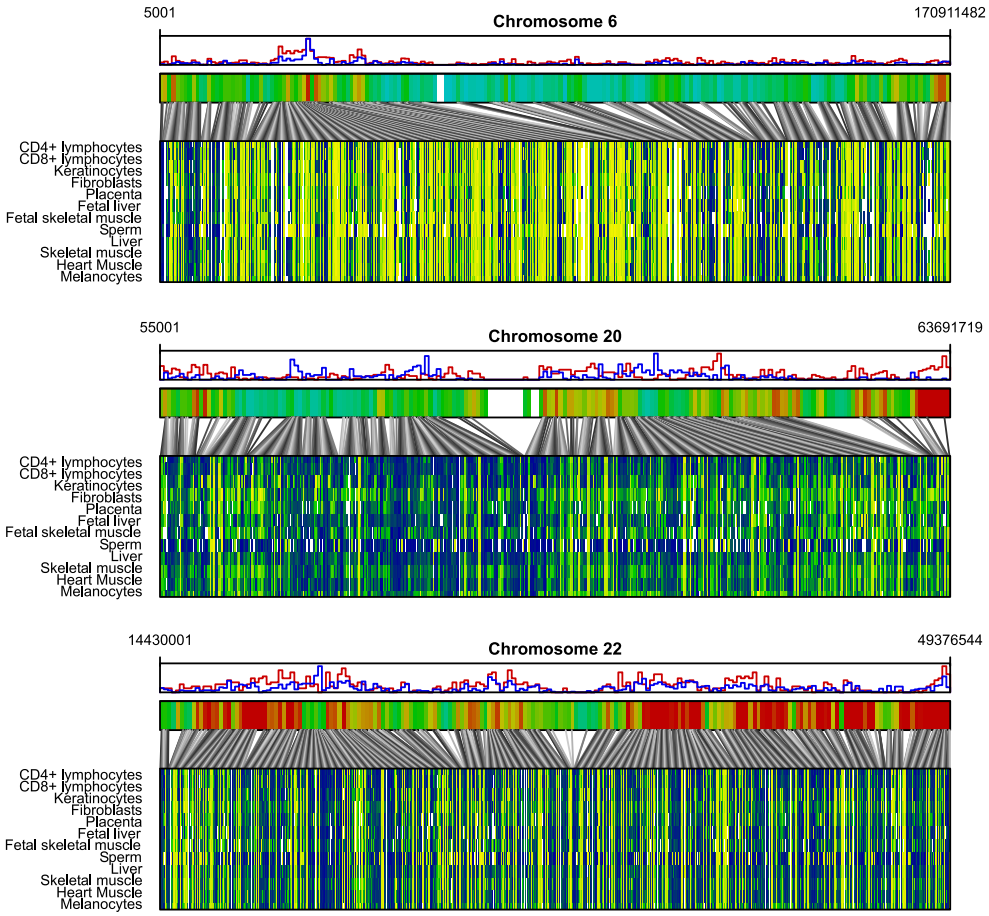

Figure 4.10: Chromosome 6, 20 and 22: On top of each chromosome the relative density of annotated genes (red) and fragments assessed (blue) are given. CpG densities are color coded in the bar below from ■ 0, ■ 0.005, ■ 0.01, ■ 0.015 to ■ 0.02. Associated methylation matrices for averaged methylation for tissue and amplificate are given from ■ 0% over ■ 50% to ■ 100%. Extended versions of these plots including a higher resolution map of the chromosome are found in Fig. 7.15 to 7.17 in the appendix.

4.3.3 Function associated methylation

4.3.3.1 Distribution of methylation

Methylation measured in healthy tissue shows bimodal distributions, this was found independent in pilot study (Fig. 4.11) and in the work package 1 of the HEP (chromosome 22 data as example in Fig. 7.29). The fact that not all CpGs in an amplificate are co-methylated leads to smoother data, when averaging all CpGs over amplificates, but preserves the bimodal distribution. The majority of CpGs tends either to be highly hypo- or hypermethylated. In this work a more stringent definition is used than in (Eckhardt *et al.*, 2006) (see section 4.2.3.2). In intragenic regions 62% of all measured CpGs fall into the extremes, this will be discussed in more detail in section 4.3.3.2.

Both, pilot study and work package 1 show, that tendencies to and direction of methylation are highly dependent on CpG density and biological function of the region. In general CpGs in CpG islands outside of repeats (as assessed in the HEP) tend to be hypomethylated, especially in promoter associated regions (71% hypermethylation versus 3% hypomethylation around TSS $\pm$ 1000bp), where almost no hypermethylation is observed. Regions with a CpG density > 7% show almost no hypermethylation and those > 10% are all hypomethylated (see Fig. 4.13). The high mutation rate of methylated CpGs that tend to spontaneous deamination (Duncan & Miller, 1980) probably puts a high pressure on regions with such high CpG density: at least in germ line cells they will have to be unmethylated in order to be preserved. The mechanism that prohibits methylation at such sites might be strong enough to influence the methylation state also within all somatic tissues.

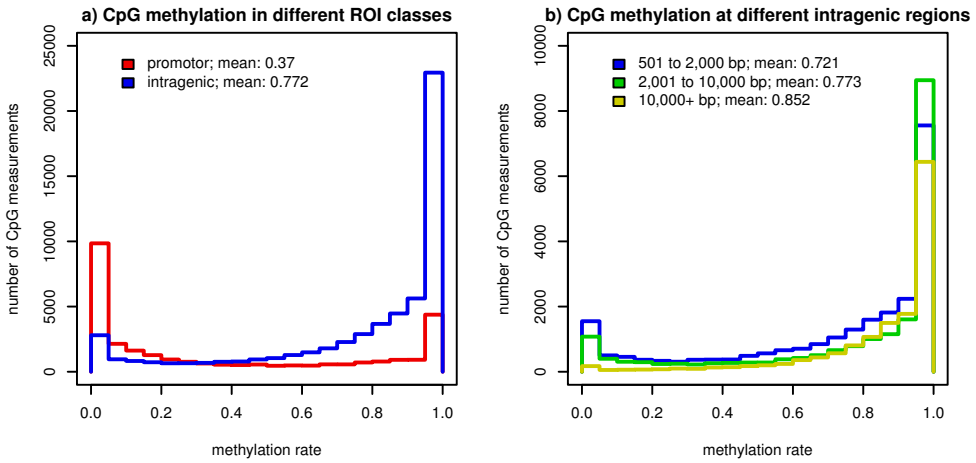

Figure 4.11: **CpG methylation distributions by region**. **a)** The bimodal distributions of methylation in promoter related regions (red) and intragenic regions (blue) are compared. **b)** Bimodal distributions of measurements from intragenic regions are split into groups by their distance to the gene start.

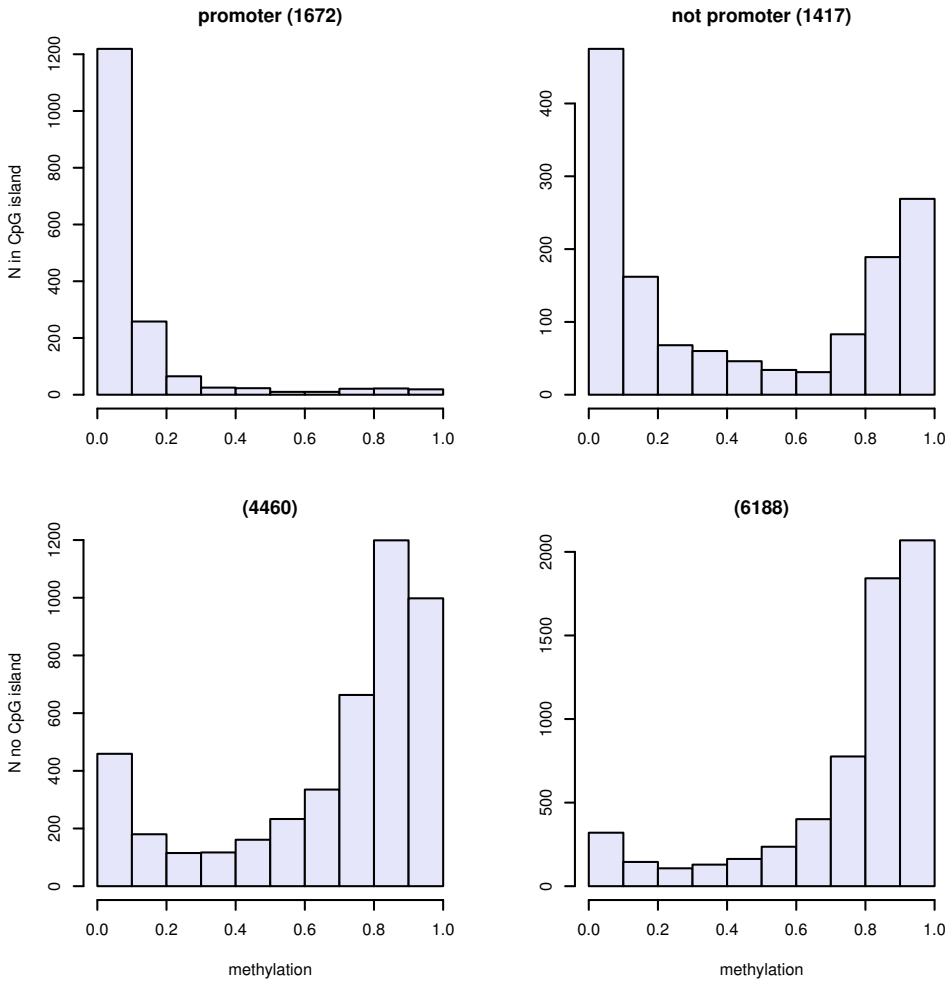

Figure 4.12: Histogram of amplificate methylation measured in **Chromosome 22**. The data is averaged amplificate methylation per tissue. Data is shown grouped by **CpG island** and **promoter** association. Top left: promoter and CpG island associated amplificates. Top right: CpG island associated but not in promoter. Bottom left: not CpG island but promoter associated. Bottom right: neither CpG island nor promoter associated. Amount of data per histogram given in brackets. Similar distributions for data from chromosome 6, 20 and for the complete data set are found in Fig. 7.29, 7.30 and 7.31.

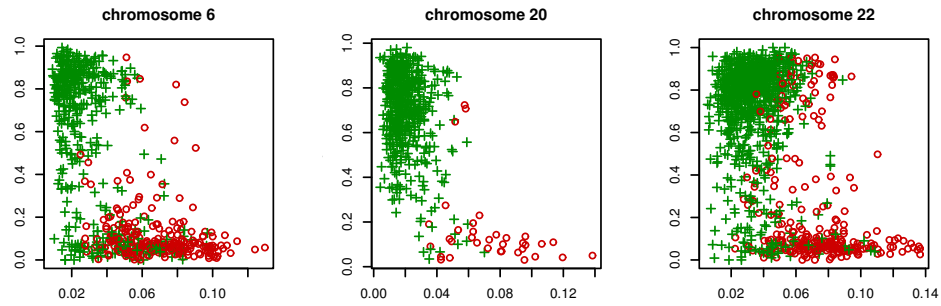

Figure 4.13: Methylation in dependency of CpG density. Measurement points are average amplificate methylation over all Tissues. Red circles are based on measurements that are CpG island associated. Data displayed per chromosome 6: 677, chromosome 20: 551, chromosome 22: 1189.

4.3.3.2 Methylation in promoters, exons and introns

Methylation of CpGs and amplificates that are located in promoter associated regions (from -2000 to 500 bp around TSS), inside exons or introns showed clear differences. In promoter associated regions hypomethylation dominated. When assessing methylation on CpG level around TSS we found that there is a strong symmetrical hypomethylation around the TSS beginning within ± 2000 bp (see Fig. 4.14 top left). This methylation profile reaches a bottom around **TSS ± 1000 bp** with 62% hypomethylation, 8% hypermethylation. The effect was even stronger in case of verified Sp1 transcription factor binding sites (Cawley *et al.*, 2004) within amplificate range (*Sp1*, see Fig. 7.25) and/or CpG island (*CpGI*) association (see Fig. 7.26); **TSS ± 1000 *Sp1***: 76%/1%, **TSS ± 1000 *CpGI***: 71%/3%, **TSS ± 1000 *Sp1/CpGI***: 78%, 0%. The very small effect found for Sp1 neither confirms former reports of the influence of methylation on those transcription factor binding sites (Mancini *et al.*, 1999; Clark *et al.*, 1997) nor confirms the articles stating the opposite (Holler *et al.*, 1988; Harrington *et al.*, 1988). It could be a coincidence: TSS annotation used for the analysis might well be more accurate if transcription factor binding sites are present.

Exons and introns after exon 1 and intron 1 which were in clear spacial proximity to the TSS and tended to be hypomethylated (**exon 1**: 54% hypo-, 16% hypermethylated, **intron 1**: 54%/13%), showed a growing tendency to be hypermethylated, especially in exons, with raising number and distance to the TSS, which was stronger for exons from **exon 2**: 23% hypo-, 43% hypermethylated, **intron 2**: 14%/34%, **exon 3**: 4%/60%, **intron 3**: 13%/44% to **exon 4+**: 3%/58%, **intron 4+**: 15%/43%. The differences for 3 and 4+ are small and it is most likely that exons as well as introns after a certain distance from TSS show in average similar profiles. Interestingly exons and introns show different profiles and statistically highly significant differences (for statistics see table 4.7). Introns contain a 4 to 5 times higher percentage of hypomethylated CpGs than exons and less hypermethylated CpGs, and the methylation profile tends to decrease a bit with increased distance to the flanking exons while the spacial methylation profile in exons stays pretty much at the same level (see Fig. 4.3.3.2 top right and bottom).

Hypomethylation is often associated with open chromatin structure, which provides access to the DNA, allowing regulation, transcription and more. The

strong tendency of regions around TSS and statistical evident more frequent hypomethylation in introns compared to exons suggests that it might play a role to allow access to these sites, to primarily allow transcription, whereby sites in introns might hint to alternative transcription start sites. Exons on the other hand are mainly conserved for their coding functionality and might therefore contain less elements playing a role in regulative processes. The high mutation rate of methylated CpGs by spontaneous deamination seems contradictory to the fact that coding parts in genes show the highest ratio of hypermethylated methylated sites, inclusive the triplet CGA that might well mutate to a stop codon. Though the following hypothesis is speculative: It might well be that stability gained by dense packed chromatin structure for non regulative regions might be of higher importance for the overall chromosomal organization than the influence of locally higher mutation rates might cause problems.

Sorting the data into groups of predicted/novel and known genes leads to methylation profiles that with stronger tendencies in known genes but in principle identical findings (see Fig. 7.27 and table 7.3) in the appendix. This might indicate that a majority of novel/predicted genes could be functional. More detailed information about hypo- and hypermethylation around TSS and in genes associated with different annotated genetic functionality/properties is listed in table 7.1 to 7.10 in the appendix.

Table 4.7: Wilcoxon tests for CpG methylation in introns versus exons.

groups	N1	mean1	sd1	N2	mean2	sd2	p-value
exon 1 vs. intron 1	58753	0.296	0.375	74307	0.282	0.36	8.5e-05
exon 2 vs. intron 2	18791	0.639	0.397	38243	0.652	0.344	2.22e-16
exon 3 vs. intron 3	10892	0.843	0.235	11070	0.708	0.342	0
exon 4 vs. intron 4	33146	0.84	0.231	41708	0.685	0.353	0

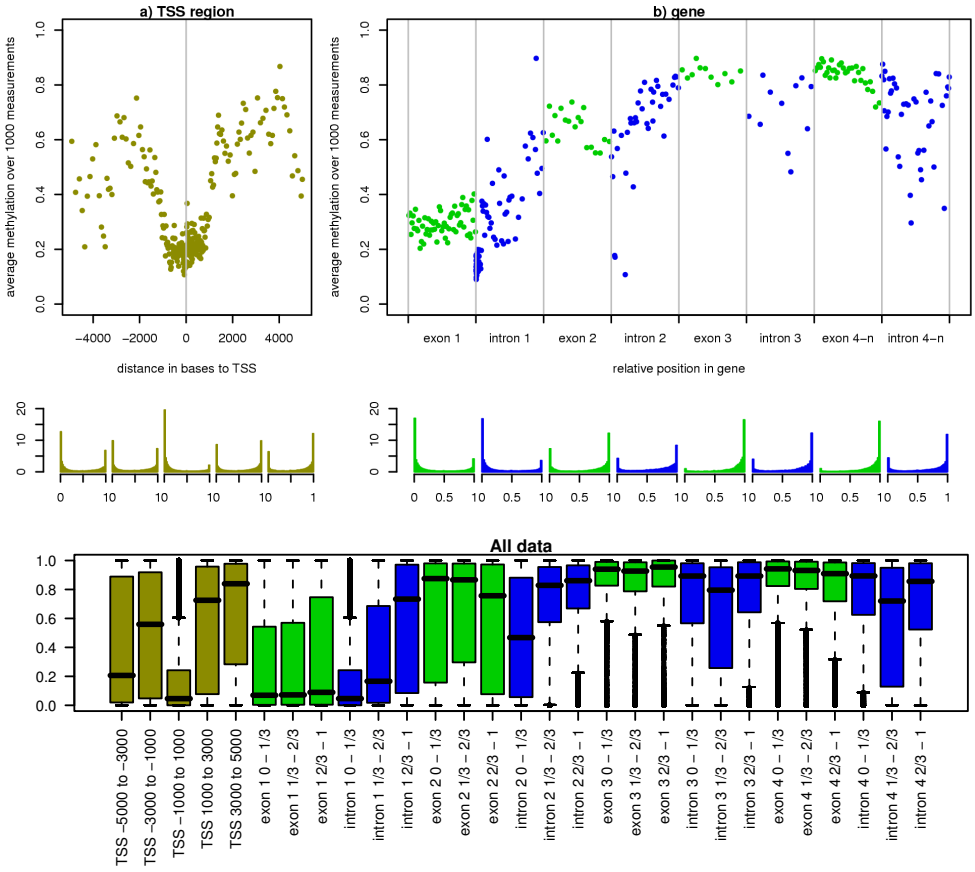

Figure 4.14: Methylation profiles in genes based on CpG methylation for all tissues, gene/TSS associated data only. **Top**: profiles binning sorted data into bins of 1000 measurements. Data around TSS (yellow, top left) is plotted by base distance to TSS. Data within gene (top right) shows average methylation of 1000 measurement in numbered exons (green) and introns (blue) given in *relative* positions (from 5' start to 3' end) to allow a summary of data over all genes though exon and intron sizes differ. Each point in the plots on top is the mean over the 1000 measurements. Due to the bimodal nature of the data, each point represents the tendency for the data to either be hypo- or hypermethylated, which is shown by the corresponding histograms in the middle. **Middle**: Corresponding methylation distributions to above, the data around the TSS (middle left) is binned into 2000 bp intervals with borders at -5000, -3000, -1000, 1000, 3000 and 5000 bp. Data in the gene is shown for each numbered exon/intron. **Bottom**: Boxplots of data binned by base positions around TSS and by functional regions within gene, thereby dividing each exon/intron into three parts. 284,141 measurements around TSS and 286,920 measurements within genes. Proportions of hypo- and hypermethylation can be found in tables 7.1 and 7.2 in the appendix.

4.3.4 Tissue specific differential methylation

The loci for methylation measurements in this study were either chosen by gene context or evolutionary conservation combined with a preferably high CpG density. Repeats are not covered. Intergenic sites are only chosen by conservation and are underrepresented in the data. The data set is therefore biased which has to be considered for all statements and findings in this section.

The **average tissue methylation** over all CpGs of the HEP work package 1 ('global' methylation for functional sites) was at minimum 43.8% for fibroblasts to 50.9% at maximum for CD8+ lymphocytes (see right labels of Fig. 4.15). Though sperm DNA showed the second lowest methylation (45.6%) neither sperm, nor placenta or fetal tissues showed strong tendencies to be more or less globally unmethylated or methylated. Within this range, at the lower end fibroblasts are outstandingly low methylated with a difference of 1.8% methylation to sperm. The highest average methylation was measured in lymphocytes and liver (including fetal liver) and there is a gap of 1.3% methylation to the next tissue. Strong similarity of related tissues like CD4+ lymphocytes and CD8+ lymphocytes or liver and fetal liver, show that these findings are unlikely to be random. For CD4+ and CD8+ these findings are consistent with the very close relation of their expression profiles (Zeng *et al.*, 2004). These differences hint that there might be strong tissue specific differences in methylation of functional sites that can be observed even at global level.

The **effect size of differential methylation (d.m.) between tissue**, was assessed by absolute methylation differences of paired data from CpGs in different tissues were averaged (see Fig. 4.16). It shows clearly that sperm has the most different methylation patterns compared to all others (ranging from 17 to 20% methylation difference). It also confirms that related tissues tend to show the lowest differences: different lymphocytes show the smallest differences, followed by heart muscle and skeletal muscle. Interestingly placenta and liver tend to show little differences to muscles.

The **efficiency of methylation to differentiate tissues** can be seen, using unsupervised clustering (see Fig. 4.17 for data from chromosome 22, from HEP work package 1). Within the data set, the single sperm sample is

the outlier. Other samples clearly cluster in blocks of their tissues (except for melanocytes). Fetal samples are close to their parental tissues. Similar differential methylation behavior can be observed in the data from the pilot study (see Fig. 4.18), demonstrating the same on a different sample/tissue set. These results clearly suggest that methylation has tissue specific profiles that can well distinguish and identify different healthy tissues.

To measure the **frequency of differential methylation** in our data set we used CpG wise kruskal-wallis test over the whole data set and used different p-values as thresholds to estimate the ratio of differentially methylated sites. This was done for the whole data set and for a reduced data set that excluded sperm, placenta and fetal tissues (see Fig. 4.19). Comparison of the found ratios with comparable data from 1,000 resampling show that the findings cannot be random: the found ratios are far distant from the resampling data. A p-value threshold of 0.05 would suggests that up to half of all CpGs assessed show d.m. This is of course only a valid statement for a global overview. The study dimensions with very high number of features (CpGs) and a comparably low number of samples do not allow the assignment of meaningful statistics to defined single CpGs.

A similar approach with data subsets assessing the **frequency of differential methylation in dependence of genomic functions** (see Fig. 4.20) shows that non coding, mainly intronic evolutionary conserved regions (ECRs) clearly have the highest ratio of dm (around 75%). CpG island associated regions in and outside of promoters show the lowest rate within the data set (around 25%) far from non conserved exons and introns (around 44%) and non CpG island associated sites in promoter and intergenic regions (around 56%). This finding is consistent with other studies using restriction landmark genomic scanning (Shiota *et al.*, 2002; Costello *et al.*, 2002).

First these findings show clearly that tissue specific differential methylation is no longer a phenomena observed only at few sites and between few tissues so far (Shiota, 2004) but is very frequent and strong. It therefore must either play an important role in the programming of the different cell types biology or at least be correlated with it, which is consistent with the fact that tissue specific differential methylation is reported to be associated with gene expression (Song *et al.*, 2005). Second it raises the question if methylation as mechanism for transcriptional control might not primarily be located in the

CpG dense regions around TSS and promoter sites, where hypomethylation could well have the more general function to keep the chromatin structure open and the region accessible - independent of up or down regulation of transcription. Third it strongly suggests that CpGs in conserved non coding regions are conserved due to variable tissue specific methylation based functionality that needs.

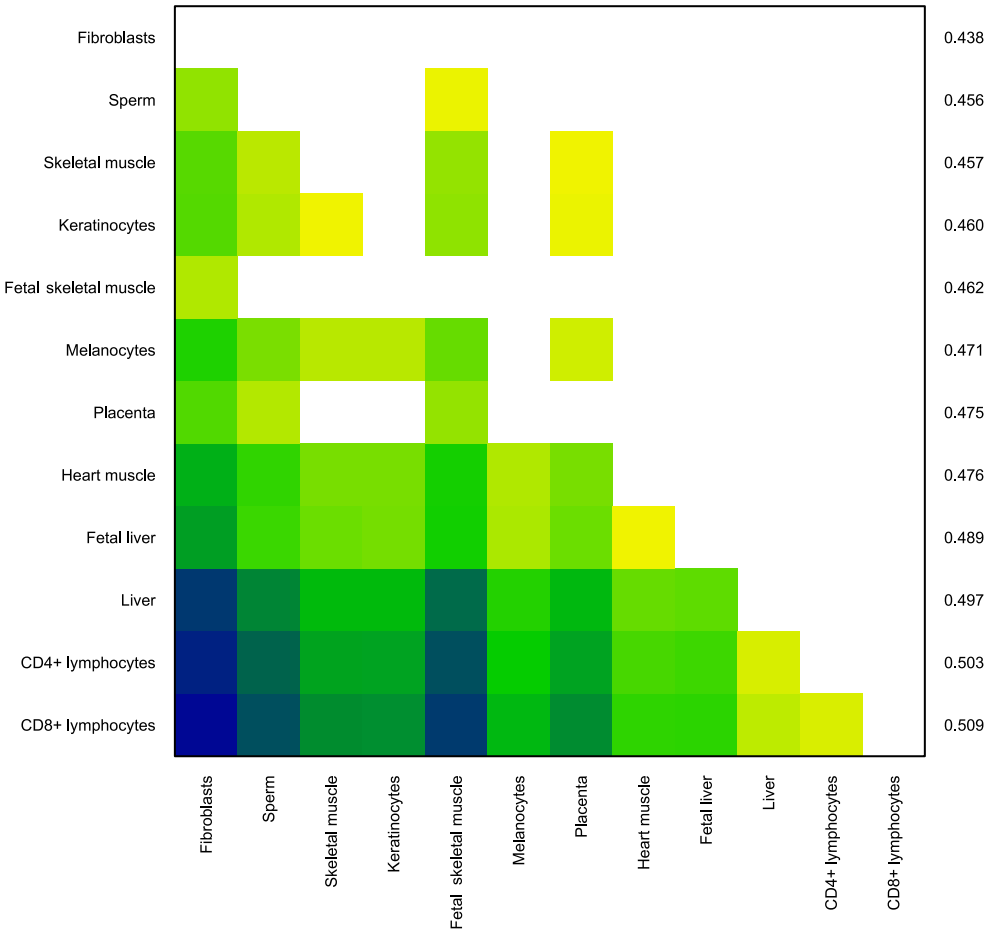

Figure 4.15: Average over all methylation differences between tissues based on matched CpGs. The color codes from yellow (0) to blue (7.5 %). Average CpG methylation per tissue over the whole data set is given on the right.

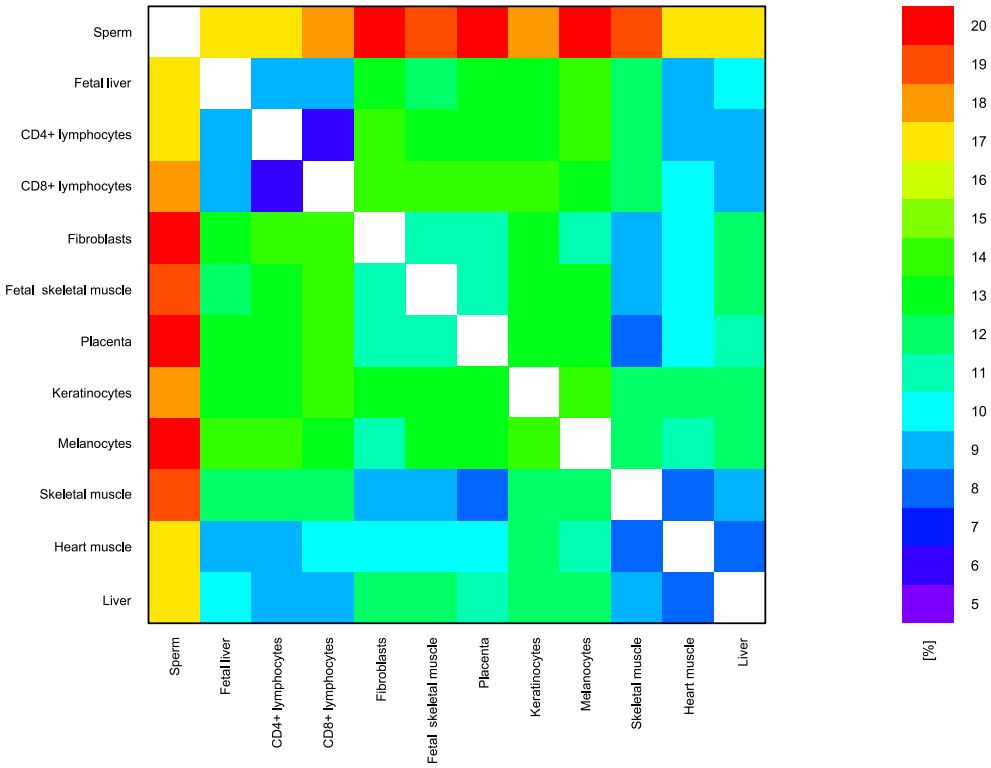

Figure 4.16: Average absolute methylation differences between tissues, based on matched CpGs. The color codes from blue (5%) to red (20%). Average CpG methylation per tissue over the whole data set is given on the right.

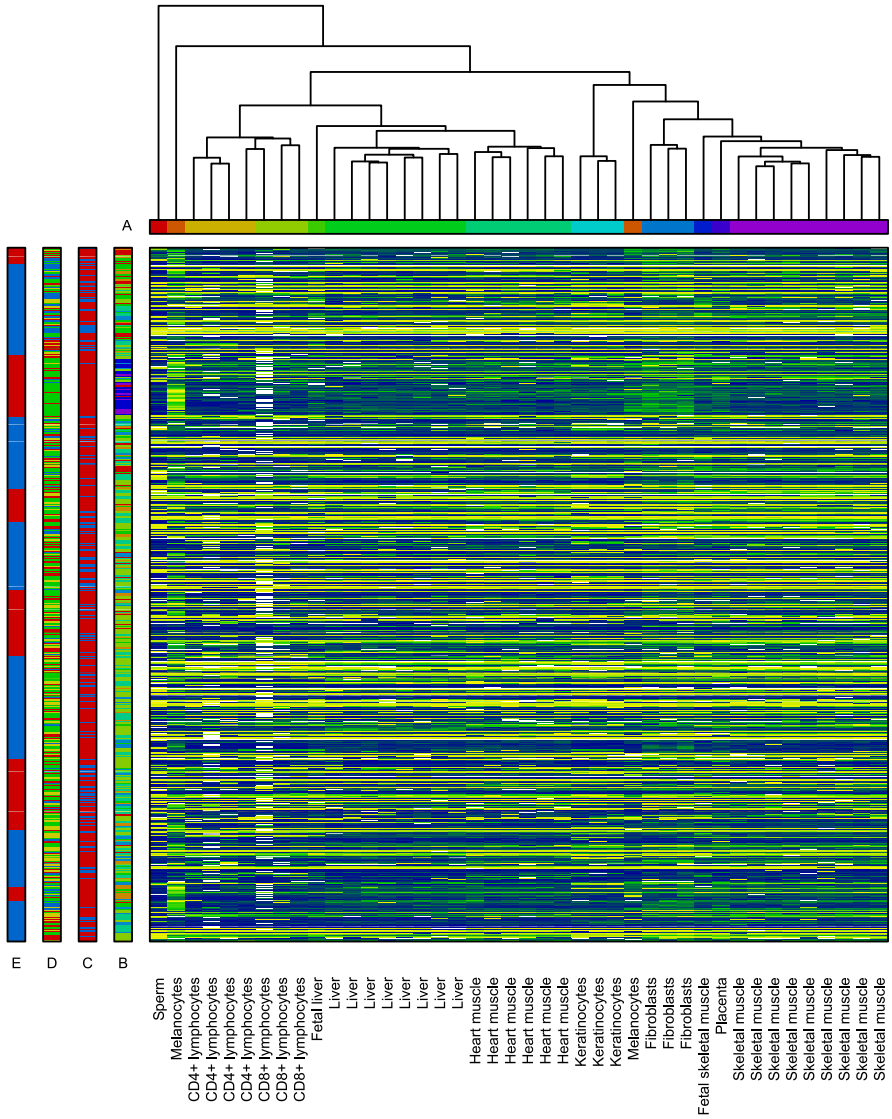

Figure 4.17: **Colored methylation plot, chromosome 22**. Color codes from 0% over 50% to 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by chromosomal position. X axis: **Tissue type**. Horizontal dimension group descriptors: **A) Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **B) Amplificate type** Pseudogene Putative Known no Type Novel CDS Novel Transcript Ig Segment Ig Pseudogene Segment . **C) CpG island state** NOT island is island . **D) Amplificate subtype** exon other promoter intron . **E) Amplificate karyotype** G-band R-band .

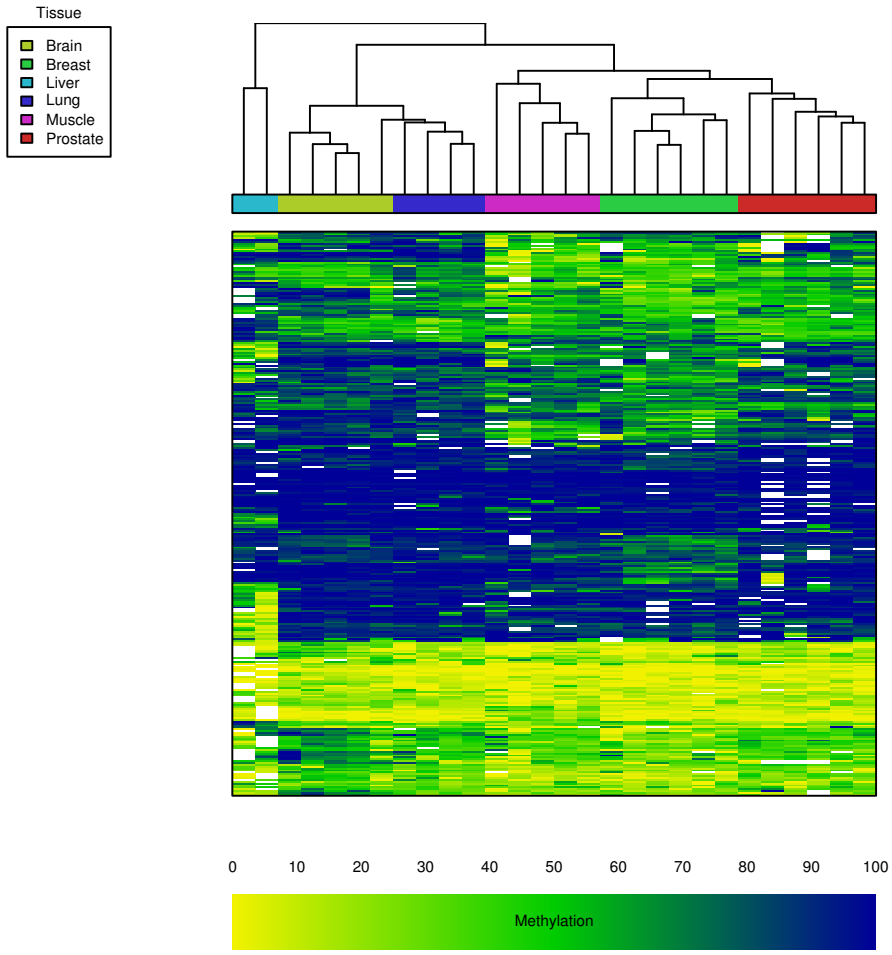

Figure 4.18: Hierarchical clustering of methylation data from **HEP pilot project**. Each column is a tissue sample, each row a CpG position. 301 CpGs are plotted. The 24 best marker regions showing differential methylation were used. Clustering: Positions without measurements were reduced in 25 steps with decreasing thresholds alternating deletion of rows and columns with too many empty positions. The final maximum of empty positions was 16%. Matrix positions still lacking measurements (white positions) were filled with average rates over all samples for that CpG. Clustering was performed in both dimensions based on euclidean distances. The color bar on the top identifies the tissue type given in the legend left of it, the color bar on the left identifies the sequences region given in the legend on its bottom.

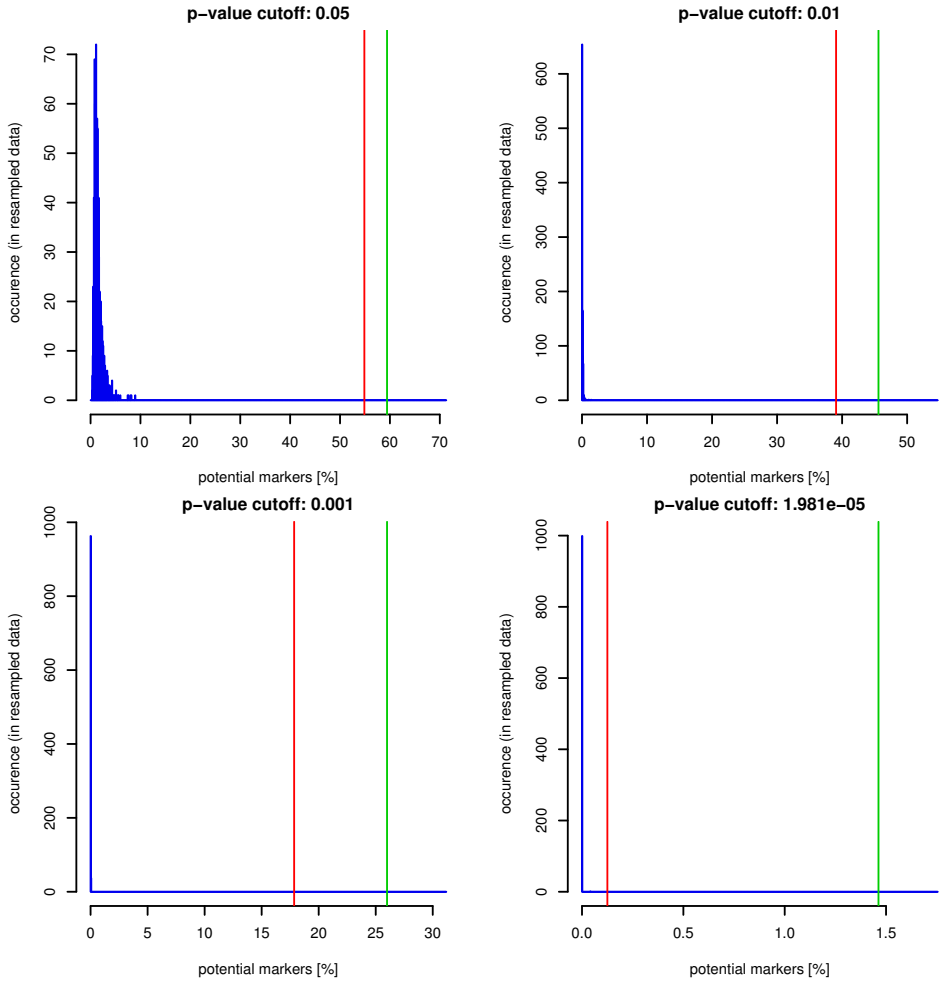

Figure 4.19: Kruskal Wallis test based ratios of marker candidates [%] for tissue based differential methylation in 2,524 amplificates (CpG data averaged over amplificates). The blue histograms show occurrence of ratios found by 1,000 x resampling, the red line shows the percentage of potential markers found in the original data (full data set), the green line shows the percentage based on a reduced data set excluding all low represented and special tissue groups (fetal tissues, placenta and sperm removed). Four p-value thresholds were used: 0.05, 0.01, 0.005 and 0.05 corrected for multiple testing by 2,524 amplificates in the data set.

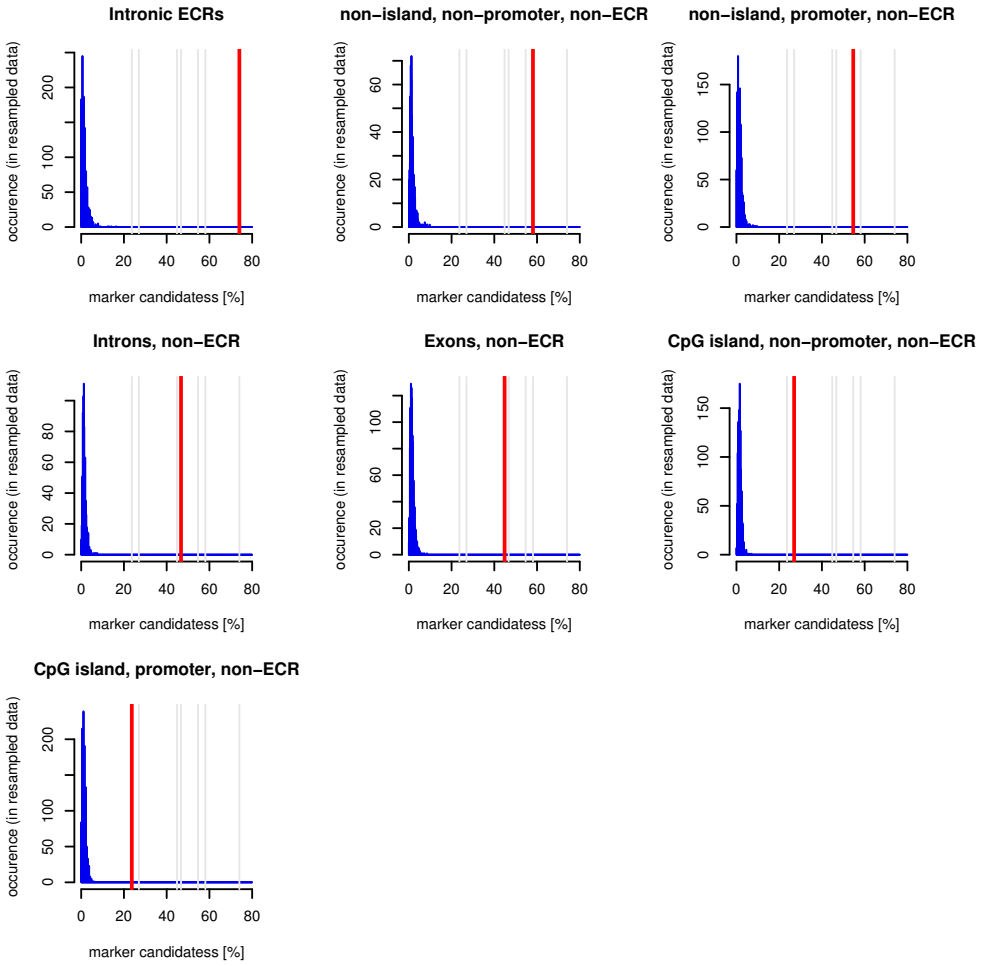

Figure 4.20: Rates of tissue specific differential methylation markers candidates on amplificate averaged methylation data. Marker candidates were defined as amplificates where a Kruskal Wallis test lead to uncorrected p-values <= 0.05. The data was split into different groups. Plots were sorted descending by percentage of marker candidates. Red lines show the percentage of markers found for the group of interest. Grey lines show the percentage of markers found for other groups. Blue histograms are distributions of rates found by 1,000 x resampling of tissue definitions of the data.

4.3.5 Differential methylation and mouse homology

One important group of sites chosen and analyzed in the HEP were non coding evolutionary conserved regions (ECRs) based on human/mouse homology. They turned out to be the sites with the highest frequency of tissue specific methylation observed in this study (see Fig. fig:samplings in section 4.3.4). For more than 70% of all fragments in ECRs an uncorrected p-value <= 0.05 was found. The human/mouse homology of amplificates was compared with p-values from Kruskal Wallis tests using methylation data on CpG level split into subsets based on different function relations of the amplificates (see Fig. 4.21).

We found that promoter associated sites and introns have a comparable level of homology but show more differential methylation (d.m.) in introns. Exons show in average more homology than promoter and intron associated sites, but a similar level of d.m. as introns. The higher grade of conservation in exons is based on the coding functionality and does not seem to influence the intragenic level of d.m. The ECRs show the highest grade of conservation (which was criterion for their choice) and also the highest rate of differential methylation. ECRs were almost exclusively chosen in non coding regions, in introns and intergenic sites that in some cases were hundreds of megabases apart from the next annotated gene. This suggests that the most direct and less biased group for comparison with ECRs possible are intronic sites. A separate comparison of d.m. and conservation of ECRs and intronic regions on CpG data level is displayed in Fig. 4.22. This data supports the theory that evolutionary conservation (of non coding regions with high CpG density) is highly correlated to d.m. and that such sites probably play a role that is important enough to be conserved.

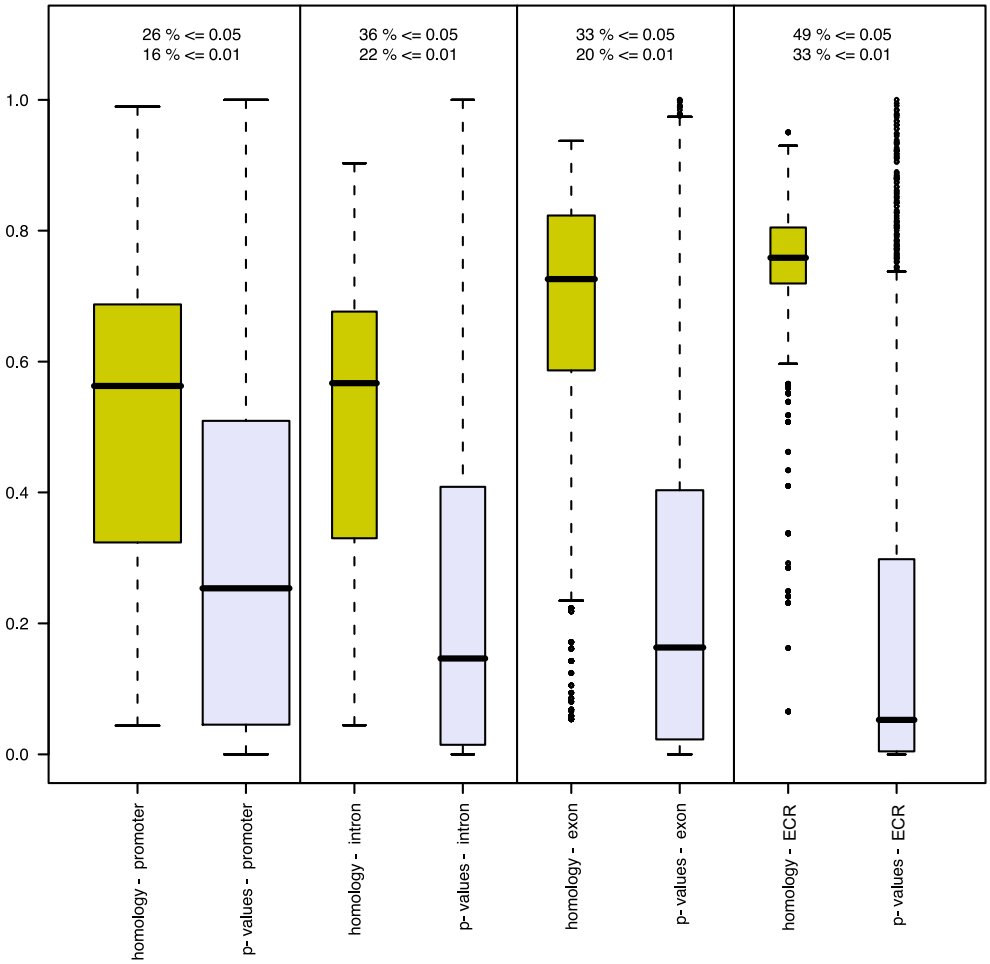

Figure 4.21: **Comparison of homology and differential methylation on CpG level.** Corresponding boxplots of amplificate human/mouse homology (yellow) and corresponding p-values from Kruskal Wallis tests (grey) for tissue specific methylation (data on CpG level) are plot side by side. The data was split into four groups containing data associated with promoters, introns, exons and evolutionary conserved regions (ECRs). The percentage of CpGs below a certain p-value threshold are given on top of each of the four groups.

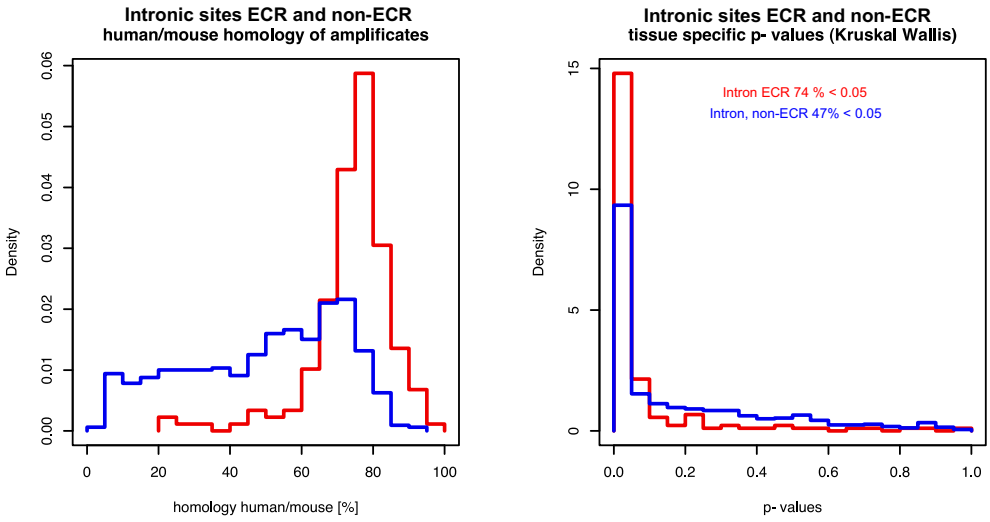

Figure 4.22: **Histograms of homology and differential methylation in evolutionary conserved regions and intronic sites.** The left plot shows the distribution of the human/mouse homology, the right plot shows the distribution of p-values from Kruskal Wallis tests, assessing the tissue specific d.m. on CpG level. Data from intronic amplificates is displayed in red, data from ECRs is displayed in blue.

4.3.6 Cell differentiation and age

Unsupervised clustering of the full data set from HEP work package 1 (see section 4.3.4) already showed that fetal tissues cluster close but not within their adult tissues. This finding is supported by clustering of two data subsets reduced to liver respectively muscle tissue. Fetal skeletal muscle is closer to adult skeletal muscle than to heart muscle (see Fig. 4.23). Within the liver data set the fetal tissue is clearly the outlier, though within the data set for chromosome 22 it was right next to liver and heart muscle (see Fig. 4.17). Fetal tissues tend to already show similar but not identical methylation profiles as their parental tissues. This observation supports the theory that methylation plays an important role in cell differentiation: the methylation profile and its correlated biological functions are on their way to the final state but partially need to be different to allow biological functionality specific for their developmental status.

Samples in different age groups or from patients with different sex did not cluster. In addition there is no evidence that global DNA methylation tends to systematically raise or get lower with proceeding age (see Fig. 4.24). The amount of 130,904 paired measurements and 10,000 x resampling of the data to simulate a distribution of random findings clearly shows that the very low average methylation difference between old and young is within a very tight random distribution and might well be found by chance. Paired data of old and young is highly correlated. Wilcoxon tests on amplificate level did not find any statistical significant data for age group specific differential methylation, neither in liver nor in muscle. Within these tissues and the locations assessed we can not find any systematical increase in methylation next to genes as previously reported (Issa *et al.*, 1994) (Issa *et al.*, 1996). For global methylation which was reported to decline (Wilson *et al.*, 1987) the missing coverage of regions with repeats does not allow any statements.

This does not mean that global methylation alterations even with tendency to hyper- or hypomethylation with progressing age are impossible: First, the tissues we used for the analysis might simply not be affected by such changes, other tissues might well behave completely different. Second, there might be chromosome specific different methylation behavior in other chromosomes that 6, 20 and 22. Third, regions we did not asses in our study, like for example repeats, especially those with high CpG density, might behave com-

pletely different. But it is most unlikely that some kind of mechanism or loss of control for methylating mechanisms leads to a significantly raised or lowered level of methylation in any tissue and at random CpG sites genome wide. More likely potential methylation changes in progressing age introduce more variance and more unspecific patterns as reported from comparisons of methylation patterns for aging monozygotic twins (Fraga *et al.*, 2005). But the average level of methylation seems to be stable.

Sex specific methylation was neither found as global methylation tendency in the data (no X chromosomal data), nor by statistical testing for differential methylation on CpG or amplificate level within our data set, despite the power to find differential methylation as demonstrated for tissues specific differential methylation. Sex specific methylation in autosomes might therefore in general be rare.

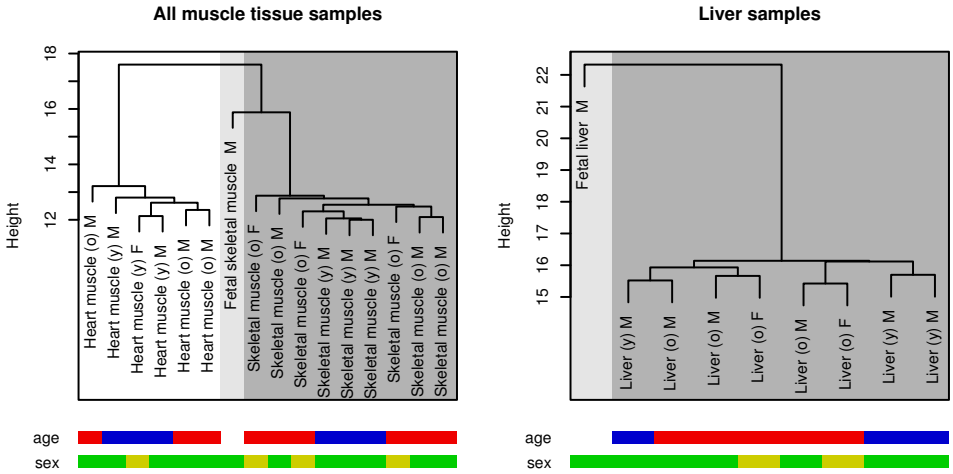

Figure 4.23: Unsupervised clustering on muscle and liver tissues. Age groups are color coded as ■ old ■ young, sex is color coded as ■ female ■ male. In addition age group and sex are at the end of sample names: (o) for old, (y) four young, M for male and F for female. Clustering based on euclidean distances between methylation of 11,001 (muscle) and 16,104 CpGs (liver). The different data set sizes are caused by filtering of CpGs that lacked measurements for one or more of the samples.

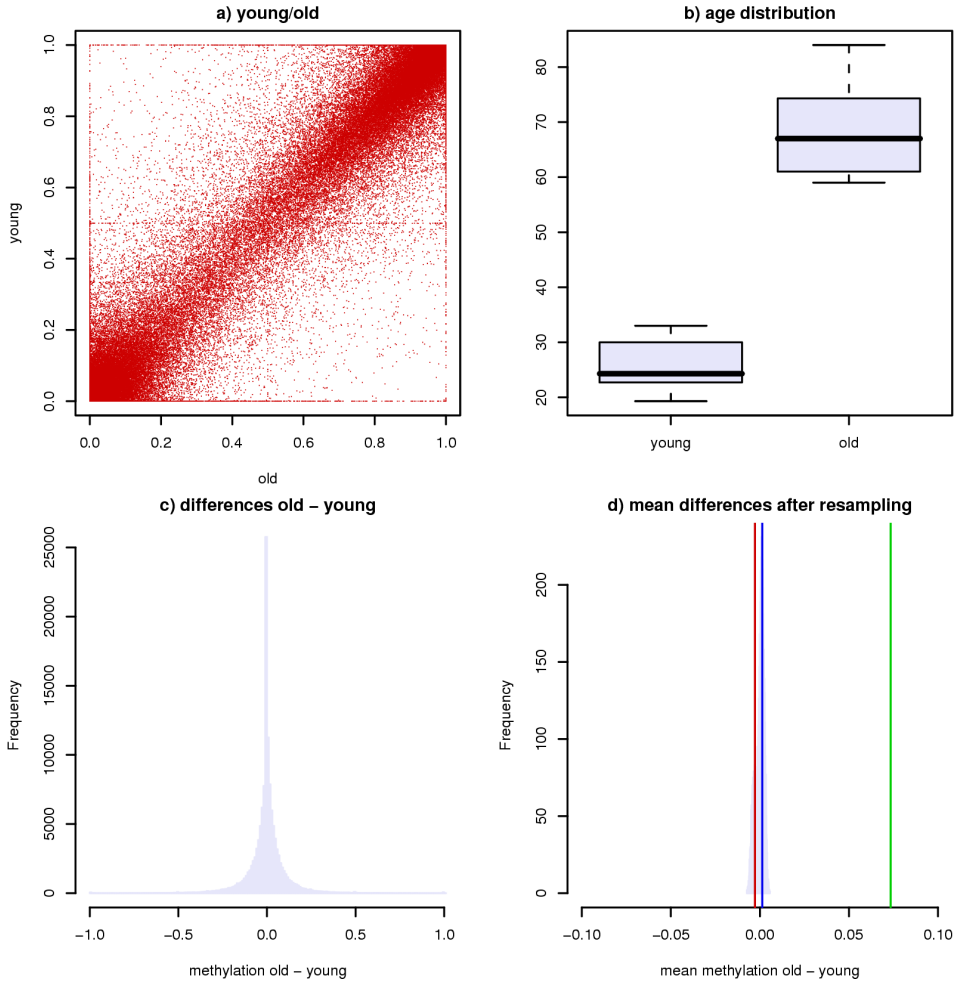

Figure 4.24: Influence of age on global DNA methylation. a) Scatter plot of 130,943 paired mean methylation, each point is an averaged methylation of an individual CpG measured in a specific tissue that was present in samples from old and young persons. Correlation: 0.959. b) Distribution of age associated with samples sorted in groups defined as old: 68.2 (8.17SD) and young: 25.8 (4.38SD). Intermediate age was discarded from this subset. In 57,455 cases methylation was higher in older group, in 58,893 cases the difference was vice versa. c) Distribution of differences of paired measurements. d) Distribution of mean methylation differences between old and young. 10,000 x intra tissue age group resampling . The red line marks the average methylation of -0.00275 found in the original data. In 17% of all cases the differences found by resampling were smaller than the original average, in 83% they were bigger. The blue line shows the differences between the sexes (male - female) in the same data set. The green line is the biggest measured tissue specific difference of CD8+ lymphocytes and fibroblasts.

4.3.7 Co-Methylation

Within the data of the HEP, methylation was often observed to be organized in co-methylated blocks of CpGs (ComBs) that could be identified by methylation changes at borders between ComBs and that often were differentially methylated ComBs. An example is given in Fig. 4.25.

The data from the pilot study suggested that within a range of 500 bp and therefore within amplificates used in the HEP there is a very high probability to find identical methylation behavior of CpGs. Data from the work package 1 with overlapping and neighbor amplificates allows analysis on CpG and on amplificate level (see Fig. 4.26). Blocks of co-methylated CpGs seem to reach maximum sizes of about 2000 bp, but rarely. In most cases blocks sizes will be below 500 bp. In cancer much longer parts have been observed to be co-methylated (Xu *et al.*, 1999; Frigola *et al.*, 2006). The difference is probably due to the loss of the fine controls regulating the methylation profiles within cancerogenesis.

The probability to observe the same methylation behavior of two CpGs is not only dependent on distance, but probably based on the fact whether CpGs share a functional domain with certain properties especially properties. There is strong evidence, that CpG density is one important parameter that influences co-methylation (see Fig. 4.27), which could mean that CpG islands often are single functional blocks of co-methylated CoMBs and that smaller blocks share CpGs that are locally spacial organized in blocks (as found in many examples in this study when examining intra-amplificate methylation on CpG level). Longer distances of two neighbor CpGs (which means without any other CpG in between) leads to a much stronger increasing possibility to have a methylation change >= 25% (which is per definition the end of any ComB) than longer distances with CpGs in between.

In general the high density data gained by sequencing shows that in many cases the measurement of a single CpG might well be used to characterize whole regions - given the prior knowledge of co-methylation or a probability estimate. Functional/regulatory and structure influencing methylation is not methylation of single CpGs, CpG islands or other arbitrary regions, but the smallest meaningful entity is the ComB.

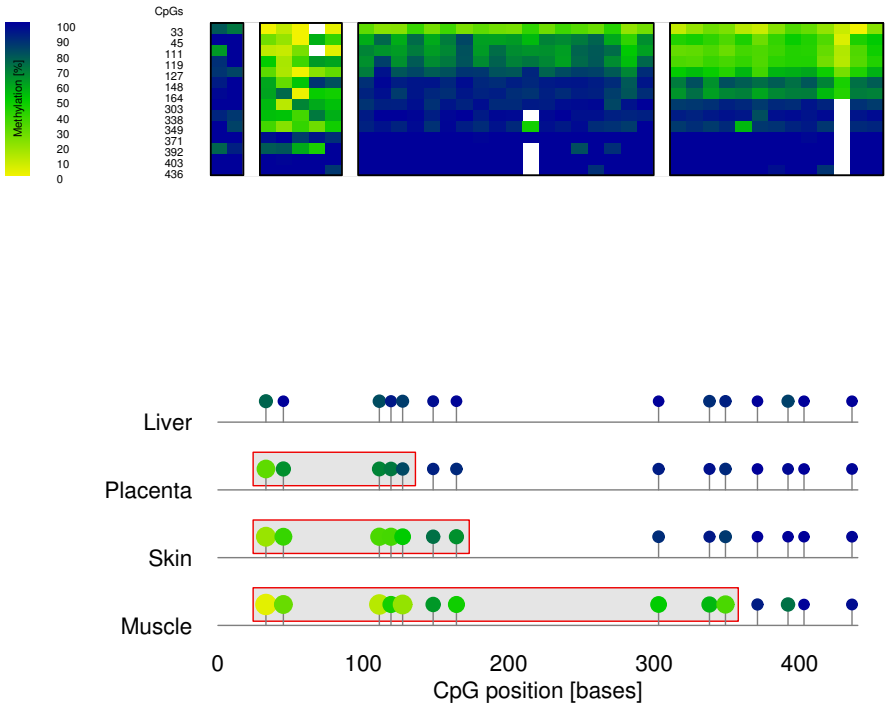

Figure 4.25: Co-Methylated blocks of CpGs (ComBs) showing tissue specific differential methylation. Local groups of CpGs show identical methylation behavior that at the borders of the ComBs change. The data within this example could be described by the methylation state of four ComBs: CpG 1-5, 6-7, 8-10, 11-14. This example was chosen because of the high number of ComBs in a short stretch of DNA sequence. In most examples data within one amplificate contained only one border between blocks.

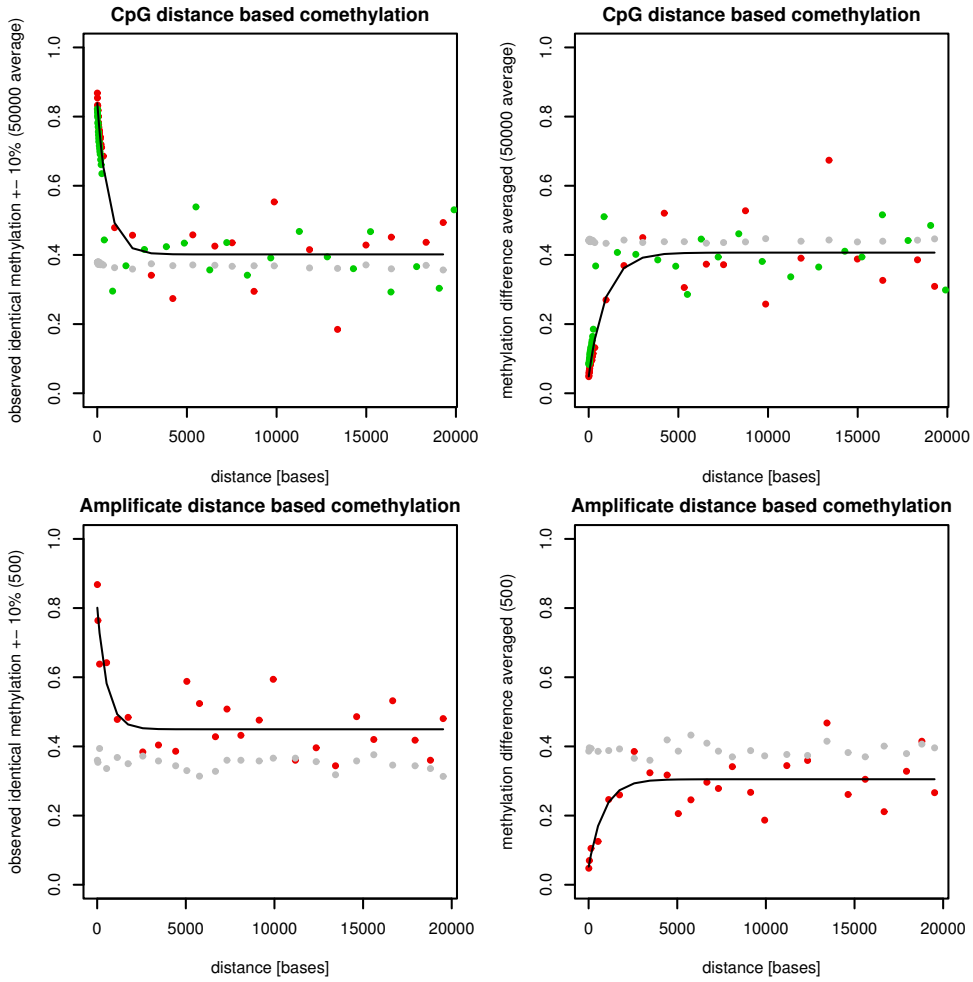

Figure 4.26: Co-Methylation. Red: original data, grey: data based on resampling, green: data based on resampling of amplificate positions in the chromosomes.

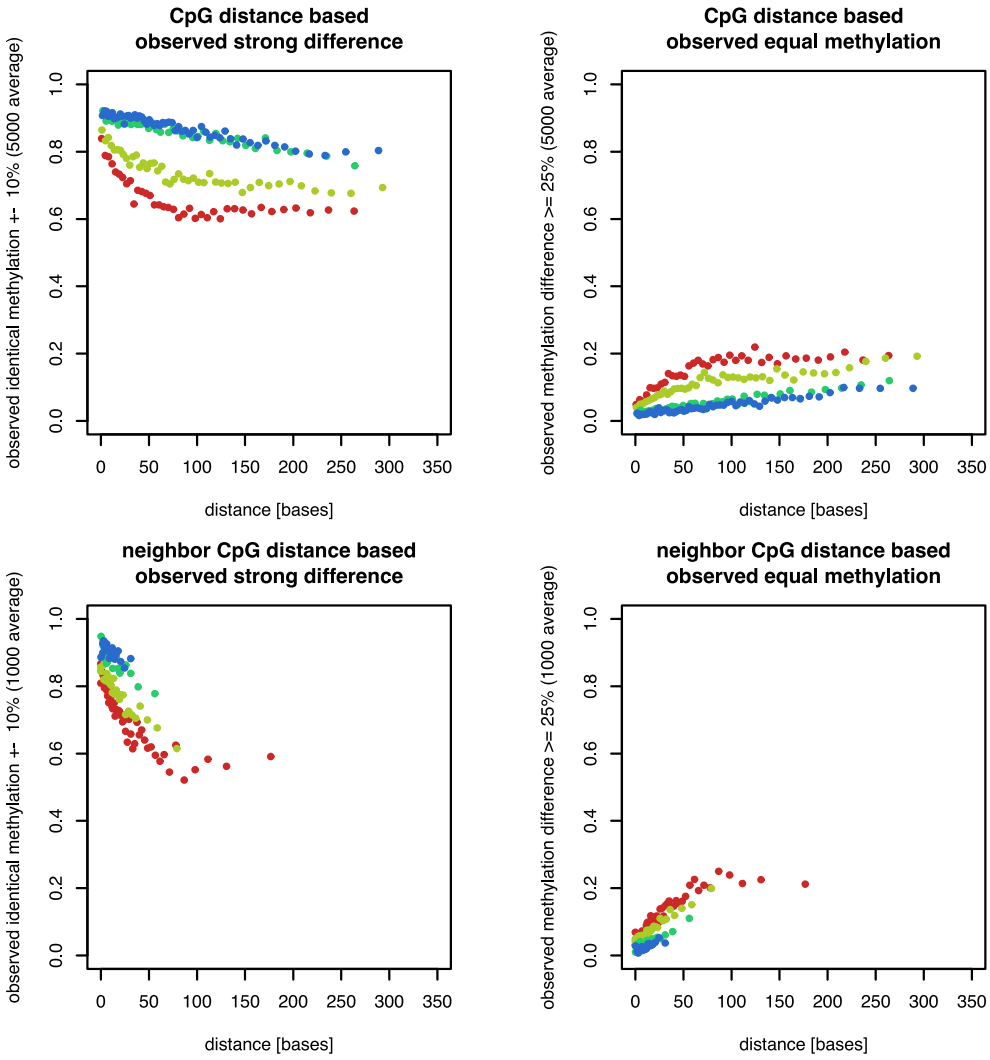

Figure 4.27: Co-Methylation of CpGs and neighbor CpGs. Observation of equal methylation and methylation changes. Data is binned by CpG density quantiles from red (low CpG density) over yellow and green to blue (high CpG density).

Chapter 5

Conclusions and outlook

So far the HEP has been a successful and very interesting study that provided data in an amount and resolution formerly not available. The development of the algorithm enabling the HEP gave us a stable and reliable tool to analyze raw data. The large amount of high-resolution data enabled us to describe methylation in human DNA in a general but also detailed way. There is more information hidden in the data already available, answering questions that were not the topic of this work, about methylation and about its role as one mechanism interacting with others. Integration of the HEP data with other biological data, proteomics data, expression data, gene network information, and data from other epigenetic layers will provide new insights into cell biology. Ongoing activities in the HEP (Eckhardt *et al.*, 2004) will add more data.

The high frequency of tissue specific differential methylation, especially at evolutionary conserved non coding sites - sometimes far away from known genes, and the characteristic methylation profiles around the TSS, suggest that methylation is even more important and more interesting than expected. Though the results from the HEP so far are important, it is still just one short episode amongst others at the beginning of the detailed understanding of this epigenetic layer. Many unanswered questions remain, that will need further data and analysis.

Besides high throughput sequencing based on the Sanger method, new and more powerful technologies have become and will become available, enabling the rapid and concurrent genome wide methylation profiling of many sites, such as established platforms for differentially methylated hybridiza-

tion (DMH). New and other Human Epigenome Projects will be able to add further data (Jones & Martienssen, 2005). However so far few of these technologies provide data at CpG level resolution and without gaps over complete regions. New sequencing technologies might soon fill the gap and generate huge amounts of data. Until then, the HEP data might for some time be an important data set, that can help to answer questions that need such level of data. At some point in the future it might be enough to measure single CpGs or complete blocks of CpGs at once, given the knowledge that they represent one co-methylated and functional block.

Currently still few information is available. Upcoming high-resolution data might enable better analysis and description of co-methylated blocks (ComBs), which could be used to train and verify ComB predictors based on DNA sequence or other correlated properties. After such or similar approaches it will be possible to map measurements from other technologies to such blocks and to describe genome wide profiles that cover all functional sites. Such a step might enable us to understand the machinery controlling the methylation of DNA in a better way. For example, what property identifies the borders of a ComB that needs to be either methylated or demethylated at some time point in cell differentiation?

The author is in pleasant anticipation of future findings in this field of cell biology and regrets that there is no more time to answer all these questions by himself or better yet in co-operation with others in the world wide scientific community. Nevertheless the author encourages everybody to use the HEP data, to profit from its findings and ideas mentioned in this work, or to contact him personally to discuss related topics.

Chapter 6

Acknowledgements

First I want to thank Professor Jörn Walter for his supervision and for providing me good ideas and detailed constructive feedback.

Special thanks goes to Florian Eckhart and Stephan Beck and all people at the Wellcome Trust Sanger Institute, at Epigenomics and at the CNG involved in the HEP - this work on the data interpretation side would not have been possible without them initiating, organizing and working within the project. In this context I would like to thank Alexander Olek for being one of the main initiators and Kurt Berlin for his continuous support of the HEP. I thank Karen Novik, Roger Horton and Vardhman Rakyan at WTSI for their intensive work and for good discussions and Jörg Tost at CNG for sharing his MALDI data.

I also would like to thank the great colleagues at Epigenomics that helped me to learn a lot, especially Péter Adorján for support in paper writing, Fabian Model for discussing deconvolution algorithms, Klaus Jünemann, Thomas König and Dirk Habighorst for insights into C++, Christian Piepenbrock and Armin Schmitt for discussing basic methods and ideas, and Matthias Burger, Tamas Rujan and many others for supporting me over the years.

Finally I would like to thank Gunter Weiss, Debjani Roy and Margit Kalcklösch for critically reading my thesis.

Chapter 7

APPENDIX

7.1 Variable definitions

Variables defined for deconvolution of trace data:

O_B	observed trace signals (with echos) for base B
$O_B(t)$	observed trace signals (with echos) for base B at time t
O'_B	model for observed trace signals for base B
$O'_B(t)$	model for observed trace signals for base B at time t
F_B	deconvolved and normalized signal of base B
$F_B(t)$	deconvolved and normalized signal of base B at time t
M_B	model for ideal trace signal of base B
$M_B(t)$	model for ideal trace signal of base B at time t
$B \in (A,C,G,T,Y)$	bases of DNA
f_B	unknown signal intensity factor of the signal of base B before normalization
k	number of molecule populations with different mobility
$\pi_i \in [0,1]$	proportion of DNA population i, $i \in (1...k)$
δ_i	shift population of i, $i \in (1...k)$
H_B	kernel for deconvolution of O_B
E_B	energy describing discrepancy between model and data, for base B
E	energy describing discrepancy between model and data, to be minimized

7.2 Plots

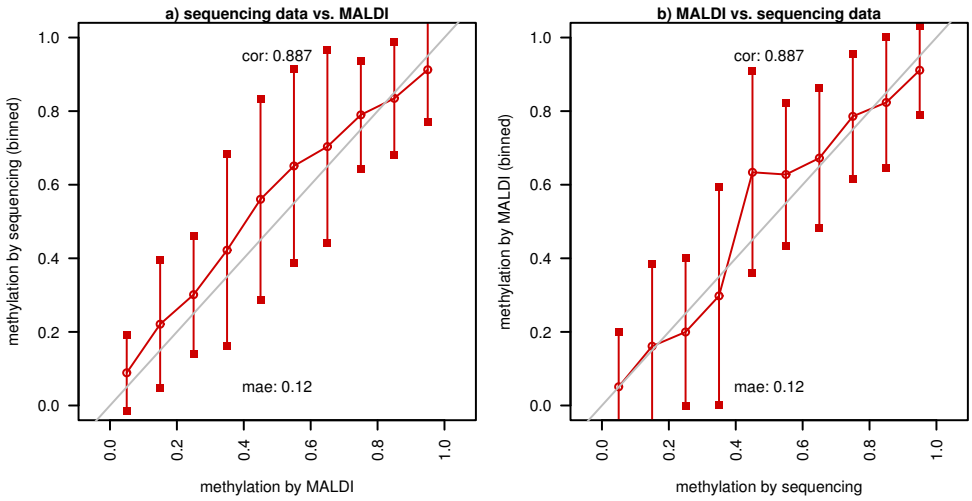

Figure 7.1: **Comparison of methylation measurements obtained from direct bisulfite sequencing with MALDI**. a) Methylation rates at CpGs from forward and reverse sequencing compared to corresponding MALDI measurements, by binning the sequencing based rates into 10 bins from 0 to 1 based on the corresponding MALDI based methylation measurements. b) Methylation rates from MALDI compared to corresponding measurements from forward and reverse sequencing, by binning the MALDI data into 10 bins from 0 to 1 based on the corresponding sequencing based methylation measurements. Red lines show the means of the binned rates, bars show the standard deviations.

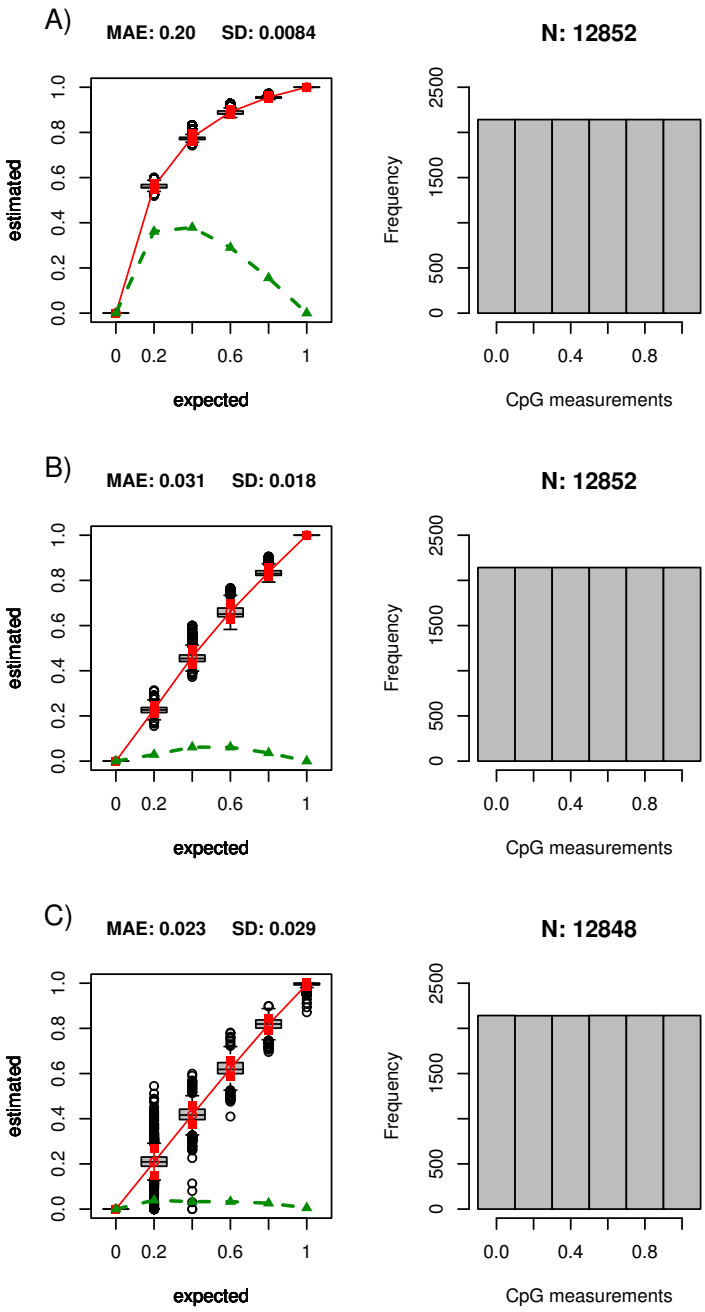

Figure 7.2: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, linear C signal overscaling): A) unnormalized, B) normalized, C) deconvolved

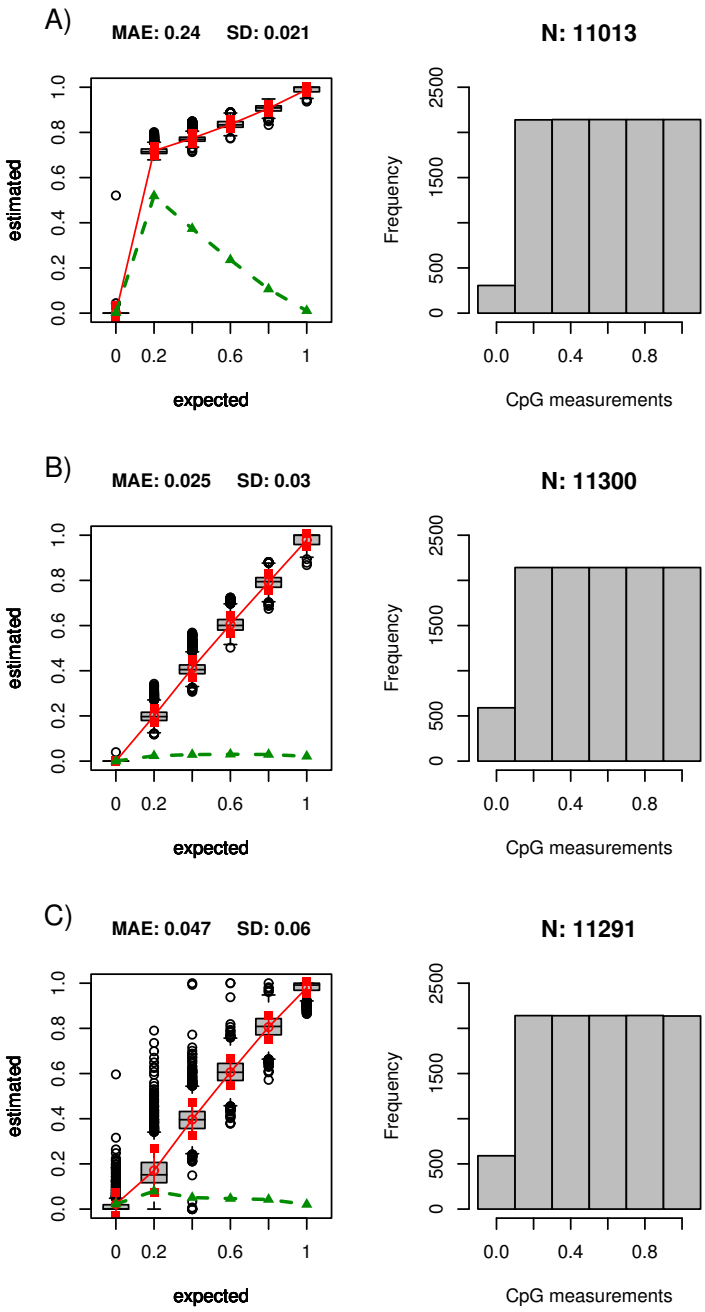

Figure 7.3: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, noise): A) unnormalized, B) normalized, C) deconvolved

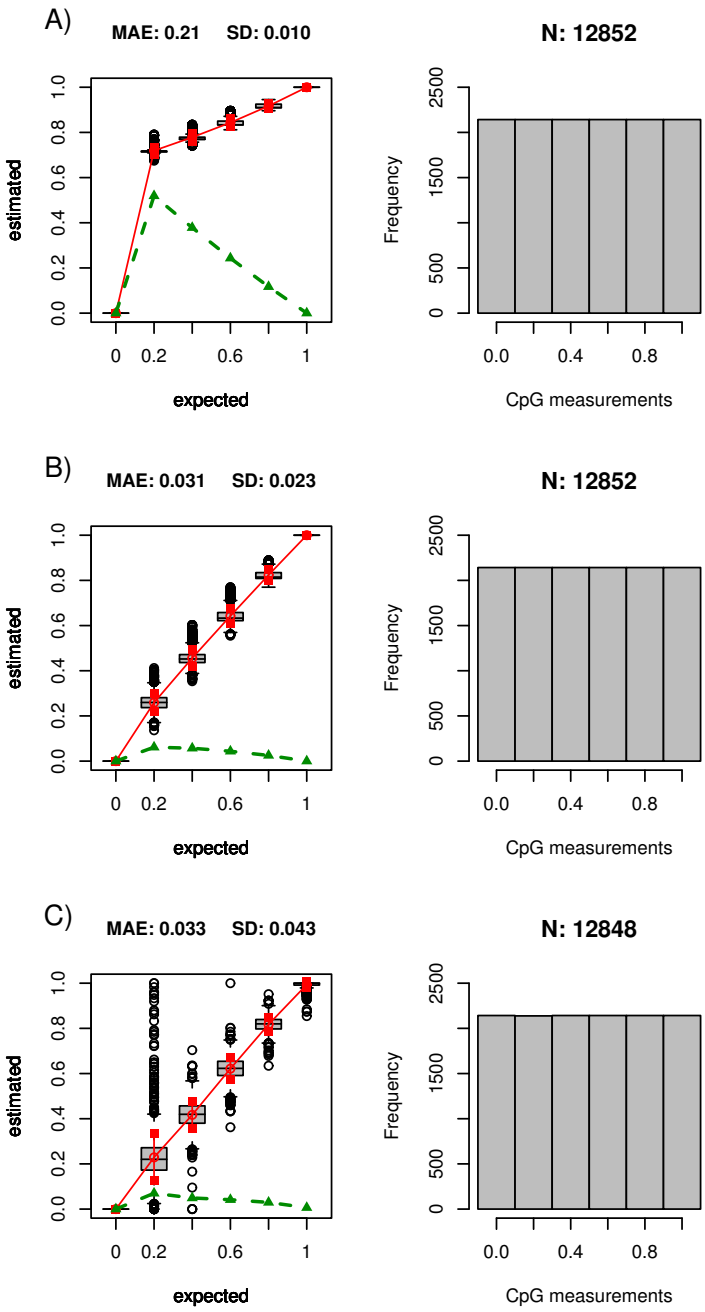

Figure 7.4: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, random intensities): A) unnormalized, B) normalized, C) deconvolved

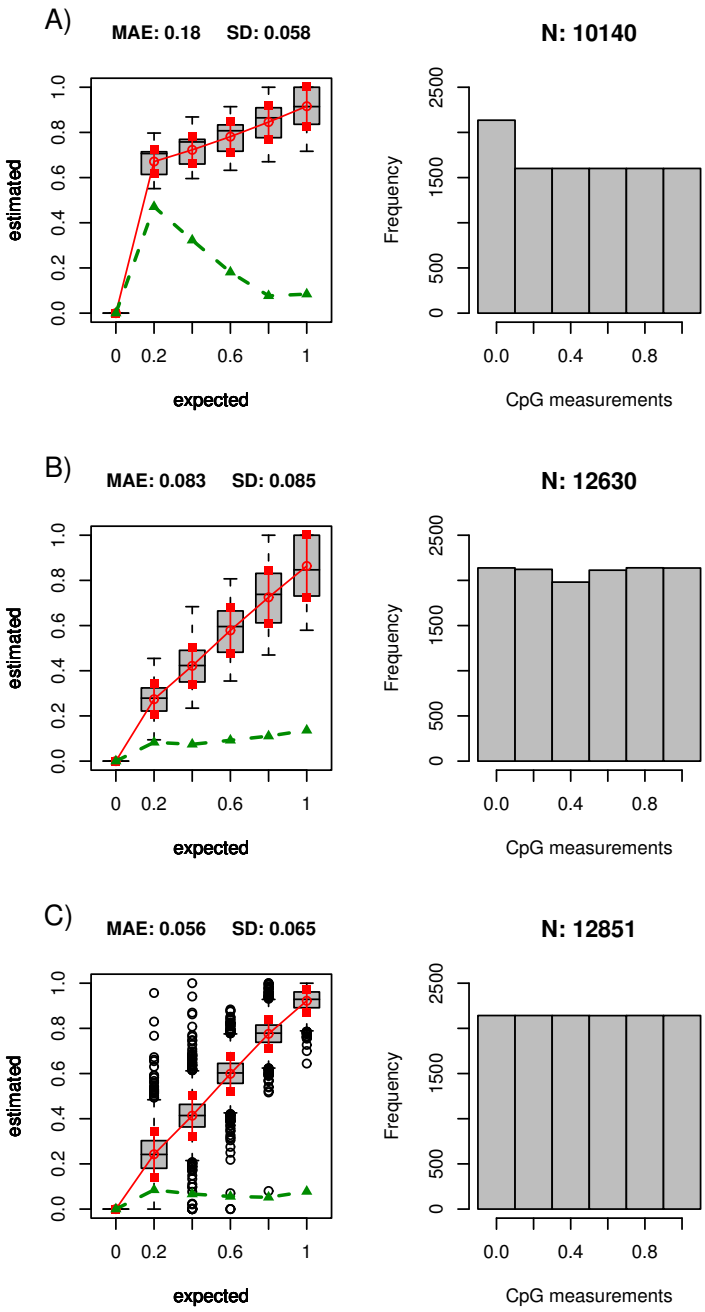

Figure 7.5: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, random intensities, convolved): A) unnormalized, B) normalized, C) deconvolved

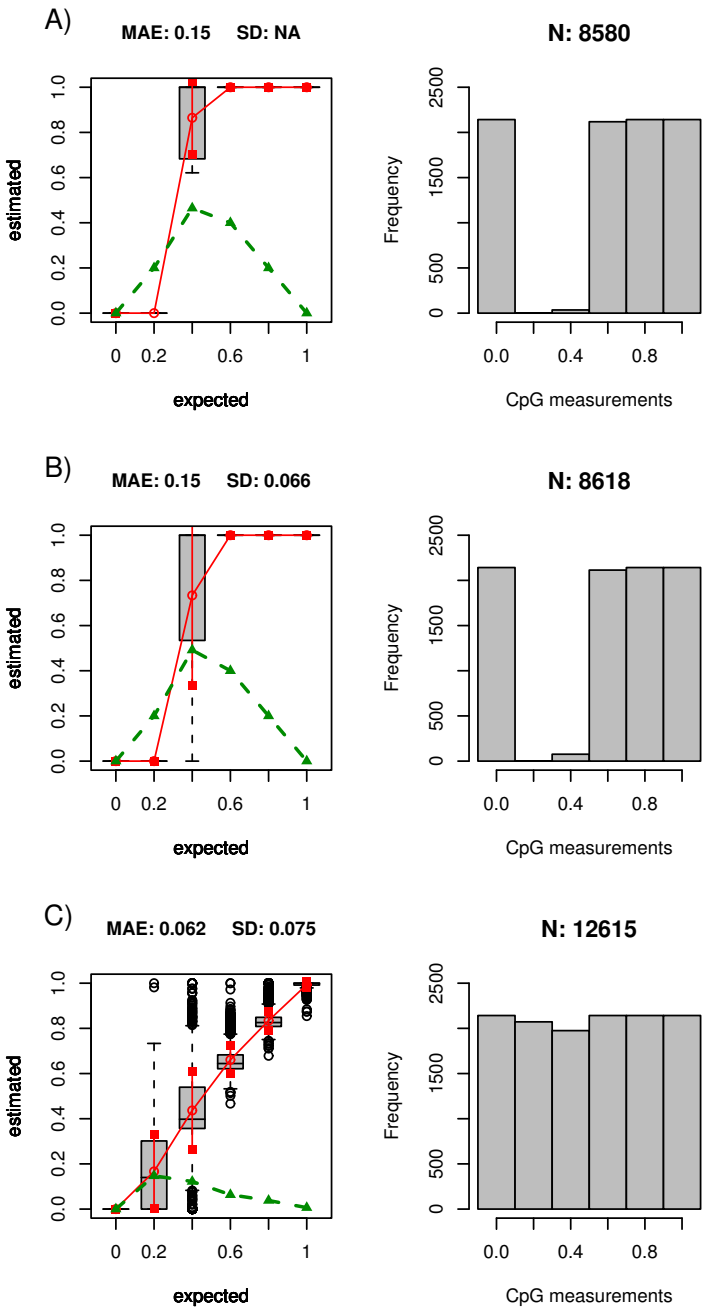

Figure 7.6: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, random intensities, shifted): A) unnormalized, B) normalized, C) deconvolved

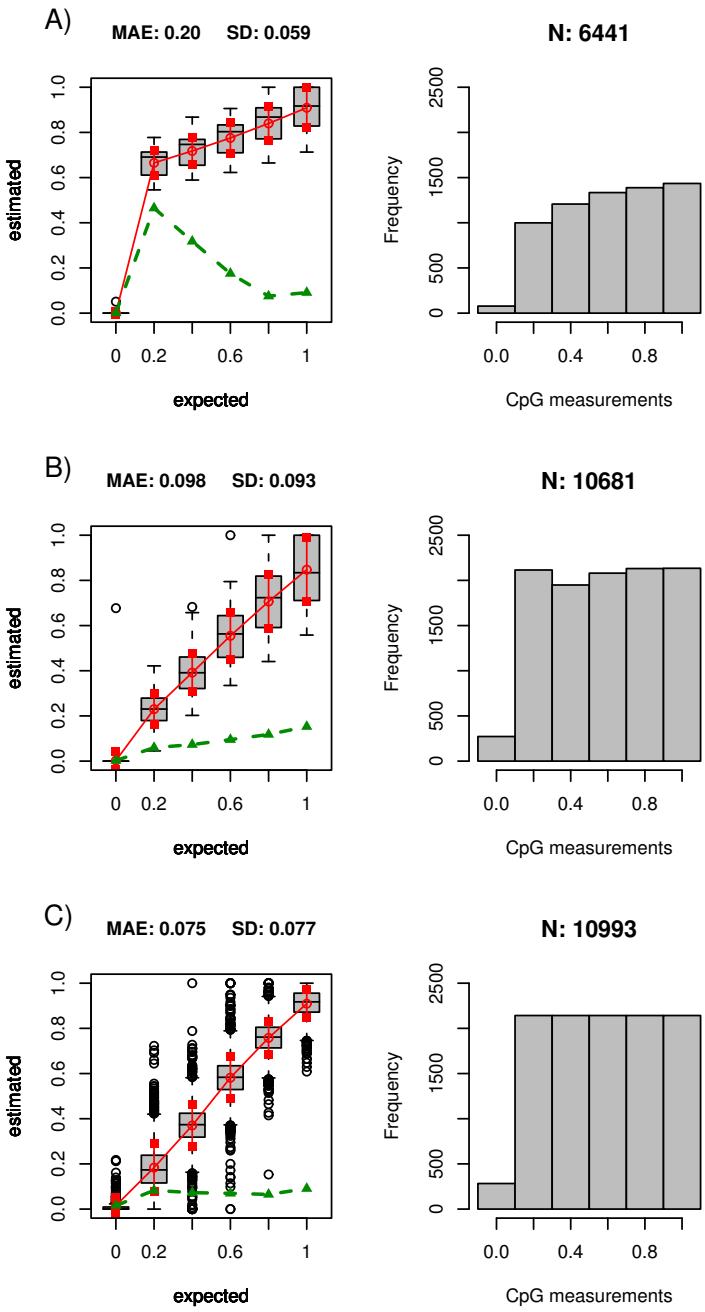

Figure 7.7: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, random intensities, convolved, noise): A) unnormalized, B) normalized, C) deconvolved

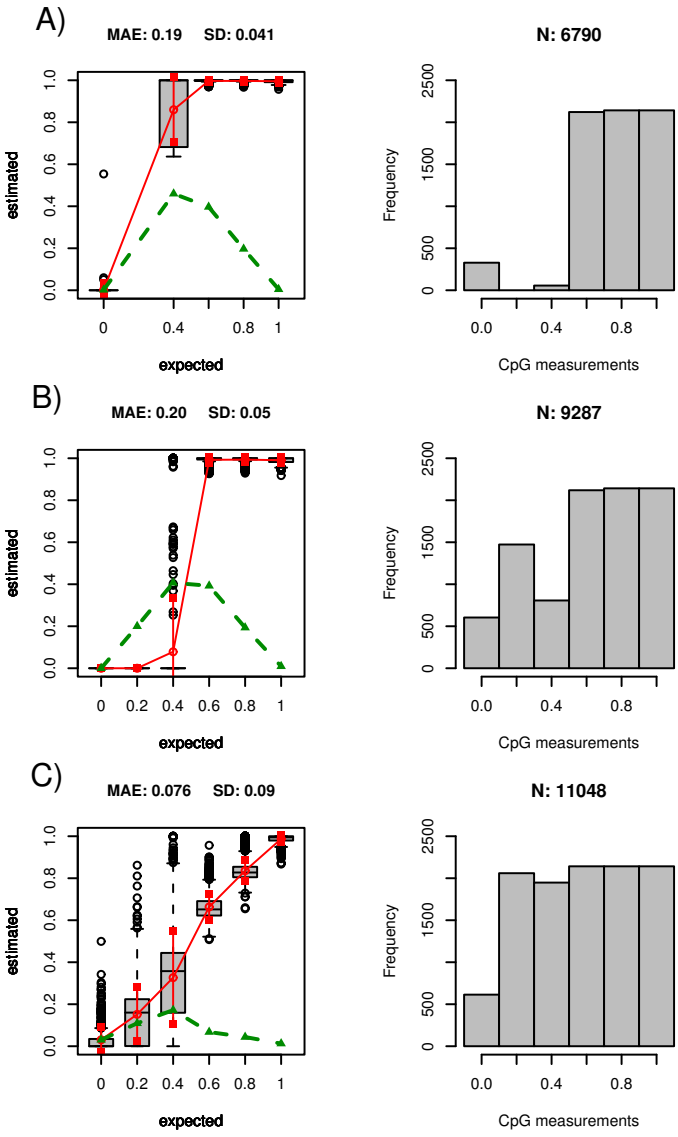

Figure 7.8: Generated data (600 traces from 100 amplificates with six different methylation rates simulated, non linear C signal overscaling, random intensities, shift, noise): A) unnormalized, B) normalized, C) deconvolved

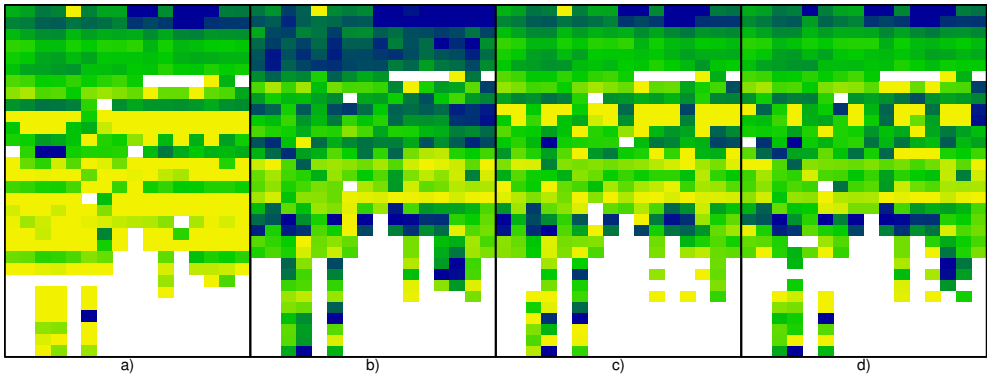

Figure 7.9: Methylation matrix plot example 1 of measurements from sequencing 50% methyalted DNA. **a)** not deconvolved, **b)** deconvolved, **c)** partially deconvolved (A, G, Y kernel based decision), **d)** partially deconvolved (C kernel based decision).

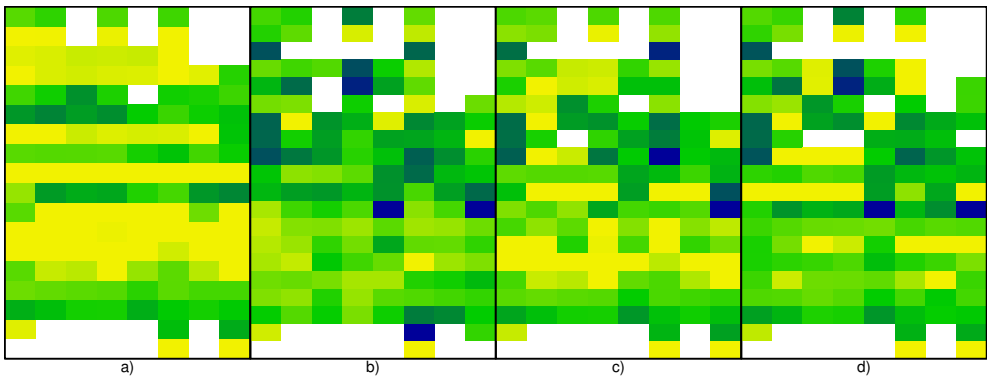

Figure 7.10: Methylation matrix plot example 2 of measurements from sequencing 50% methyalted DNA. **a)** not deconvolved, **b)** deconvolved, **c)** partially deconvolved (A, G, Y kernel based decision), **d)** partially deconvolved (C kernel based decision).

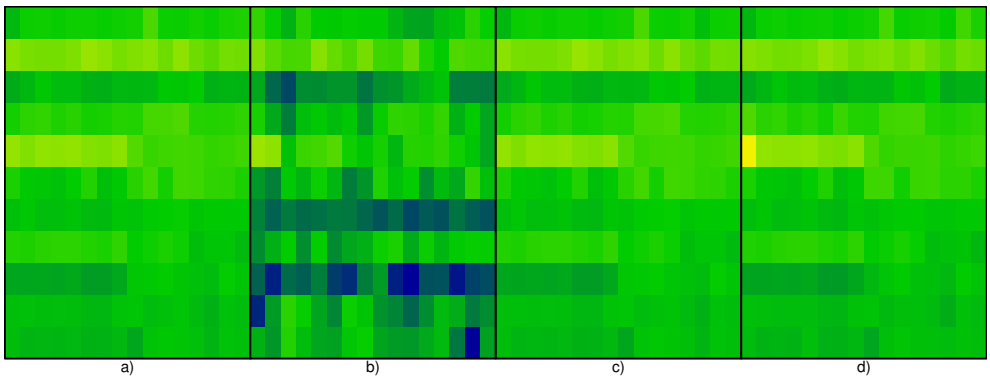

Figure 7.11: Methylation matrix plot example 3 of measurements from sequencing 50% methyalted DNA. **a)** not deconvolved, **b)** deconvolved, **c)** partially deconvolved (A, G, Y kernel based decision), **d)** partially deconvolved (C kernel based decision).

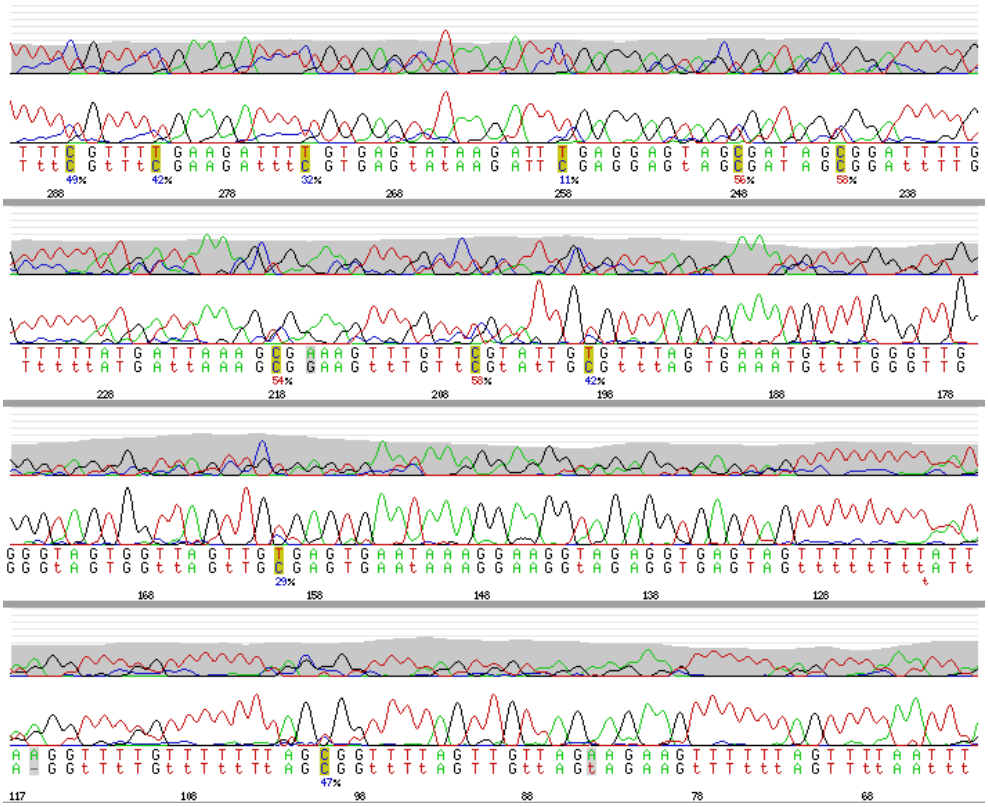

Figure 7.12: Deconvolved trace data example 1 for 50% methylated DNA.

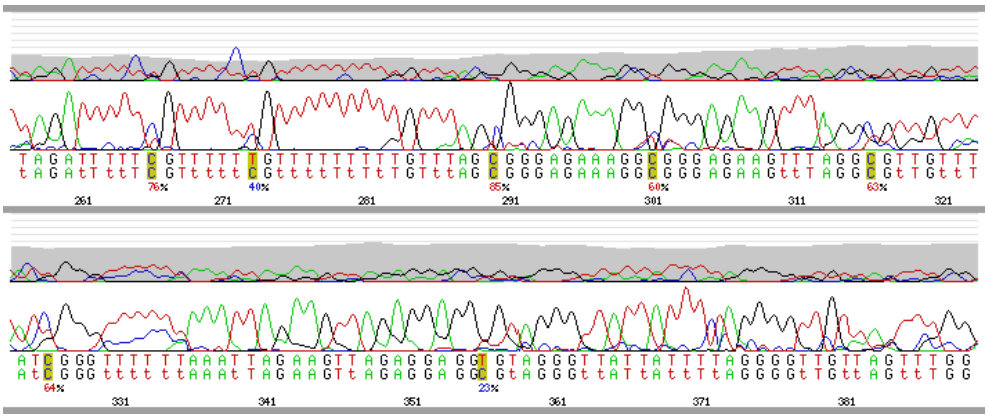

Figure 7.13: Deconvolved trace data example 2 for 50% methylated DNA.

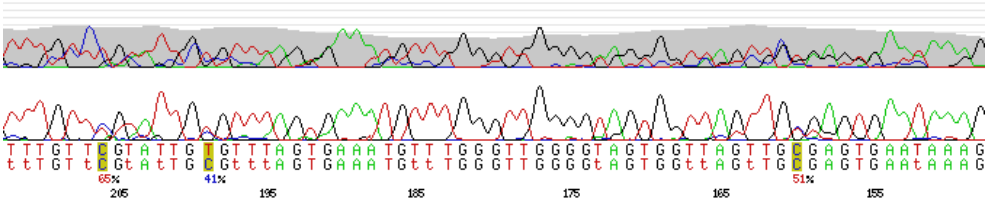

Figure 7.14: Deconvolved trace data example 3 for 50% methylated DNA.

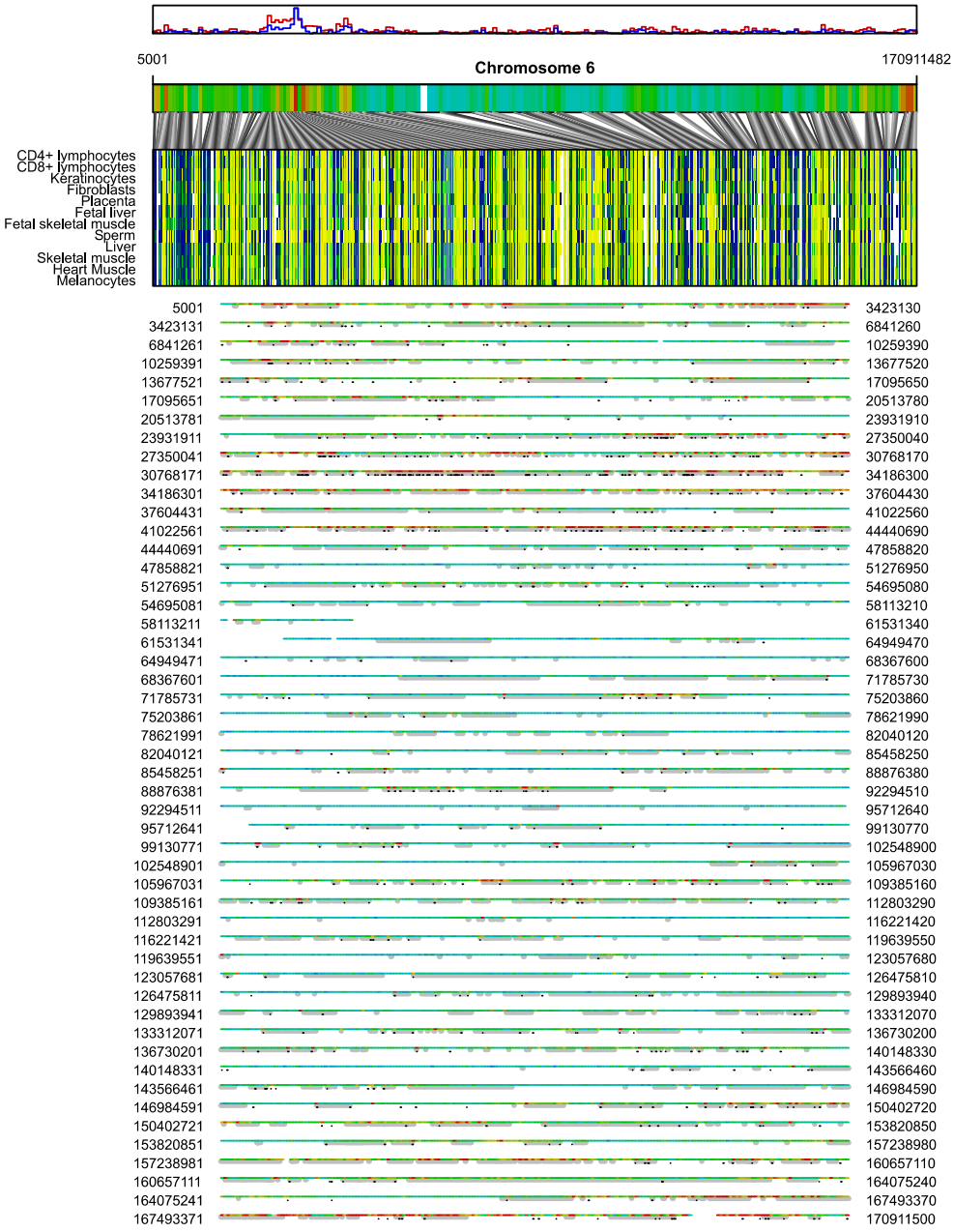

Figure 7.15: **Chromosome 6**: On top of each chromosome the relative density of annotated genes (red) and fragments assessed (blue) are given. CpG densities are color coded in the bar below from ■ 0, ■ 0.005, ■ 0.01, ■ 0.015 to ■ 0.02. Associated methylation matrices for averaged methylation for tissue and amplificate are given from ■ 0% over ■ 50% to ■ 100%. The detail map below the methylation matrix shows CpG density profiles as described above but and highlights regions with genes (grey) and fragments assessed (black) that were part of the data described in this work.

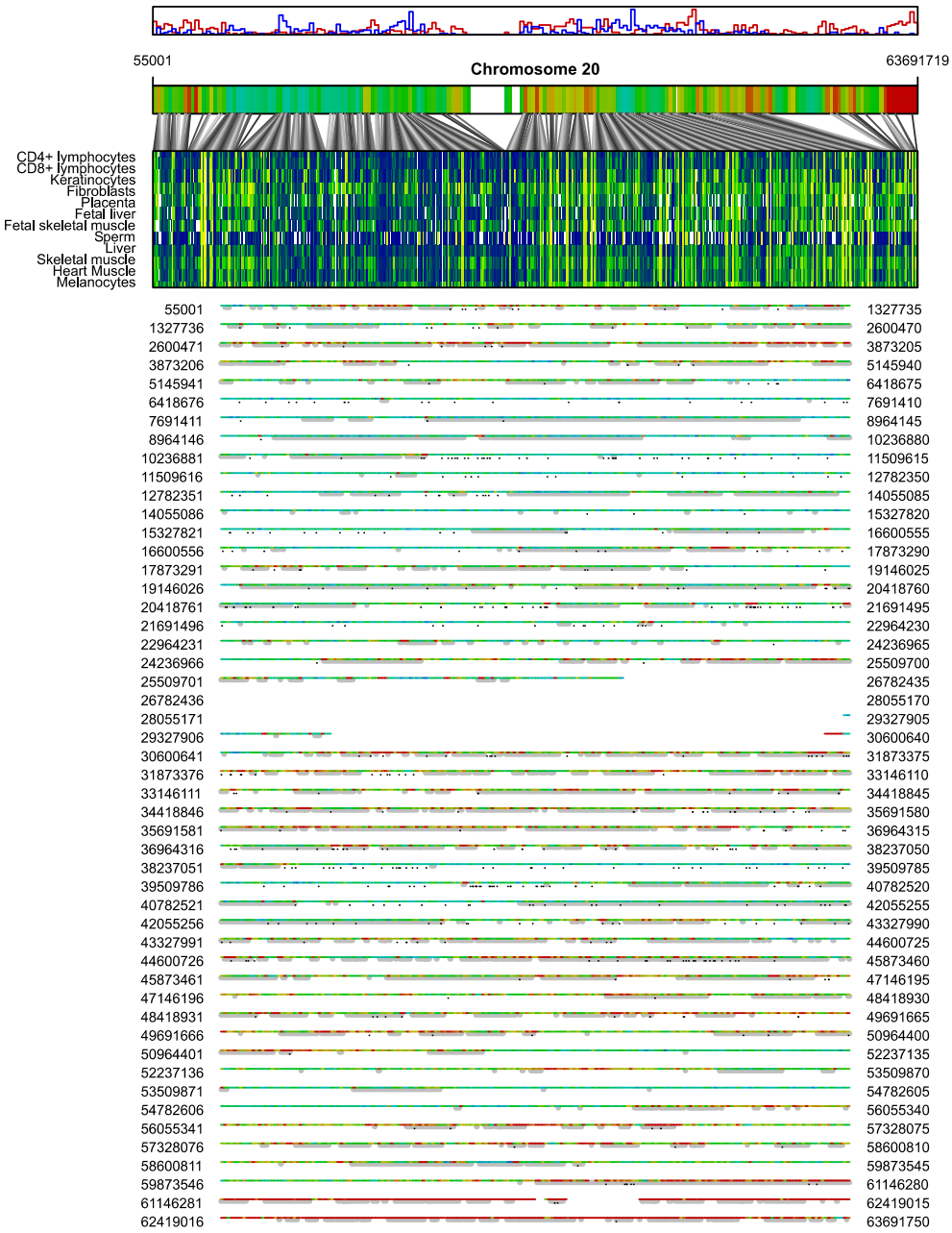

Figure 7.16: **Chromosome 20**: On top of each chromosome the relative density of annotated genes (red) and fragments assessed (blue) are given. CpG densities are color coded in the bar below from ■ 0, ■ 0.005, ■ 0.01, ■ 0.015 to ■ 0.02. Associated methylation matrices for averaged methylation for tissue and amplificate are given from ■ 0% over ■ 50% to ■ 100%. The detail map below the methylation matrix shows CpG density profiles as described above but and highlights regions with genes (grey) and fragments assessed (black) that were part of the data described in this work.

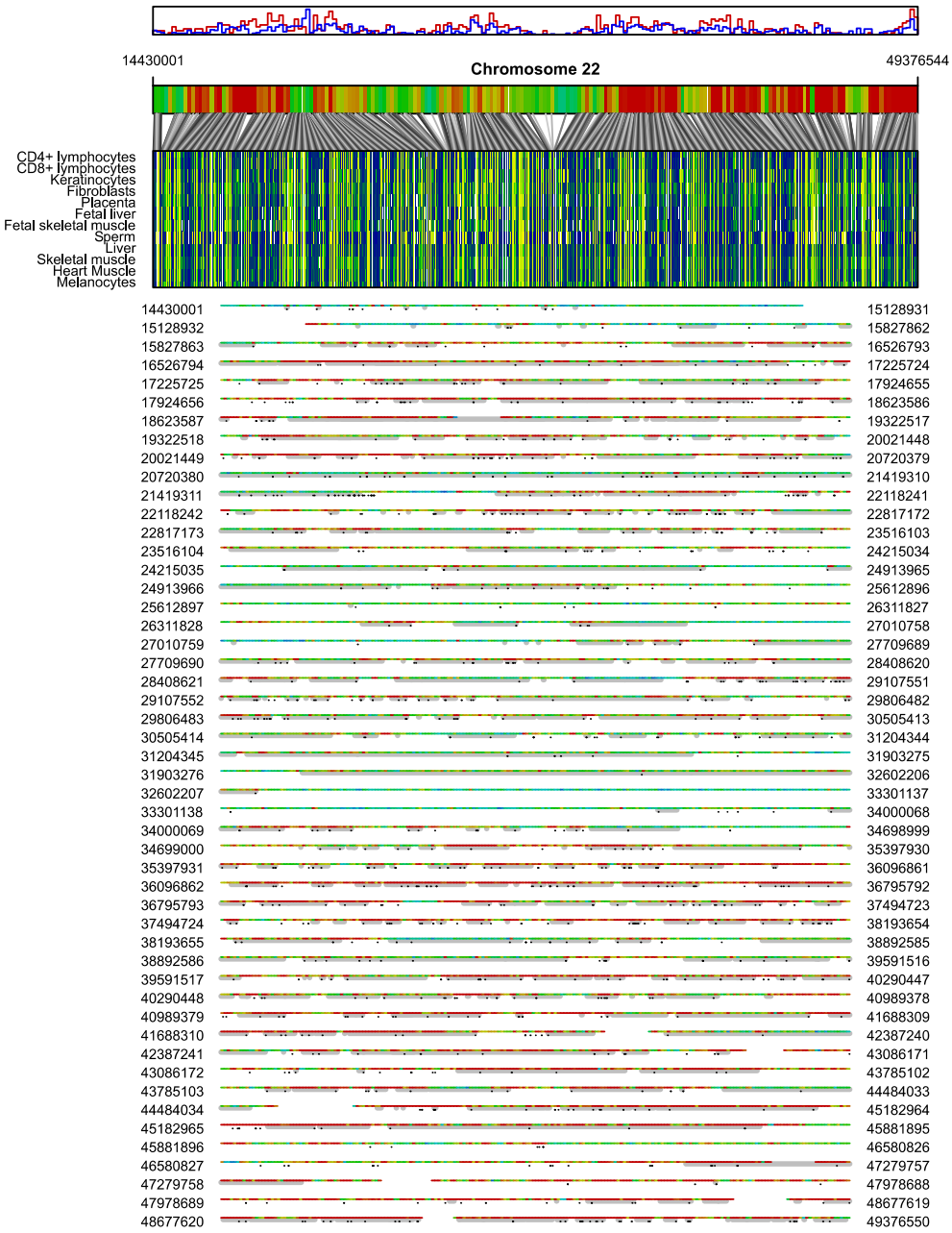

Figure 7.17: **Chromosome 22**: On top of each chromosome the relative density of annotated genes (red) and fragments assessed (blue) are given. CpG densities are color coded in the bar below from ■ 0, ■ 0.005, ■ 0.01, ■ 0.015 to ■ 0.02. Associated methylation matrices for averaged methylation for tissue and amplificate are given from ■ 0% over ■ 50% to ■ 100%. The detail map below the methylation matrix shows CpG density profiles as described above but and highlights regions with genes (grey) and fragments assessed (black) that were part of the data described in this work.

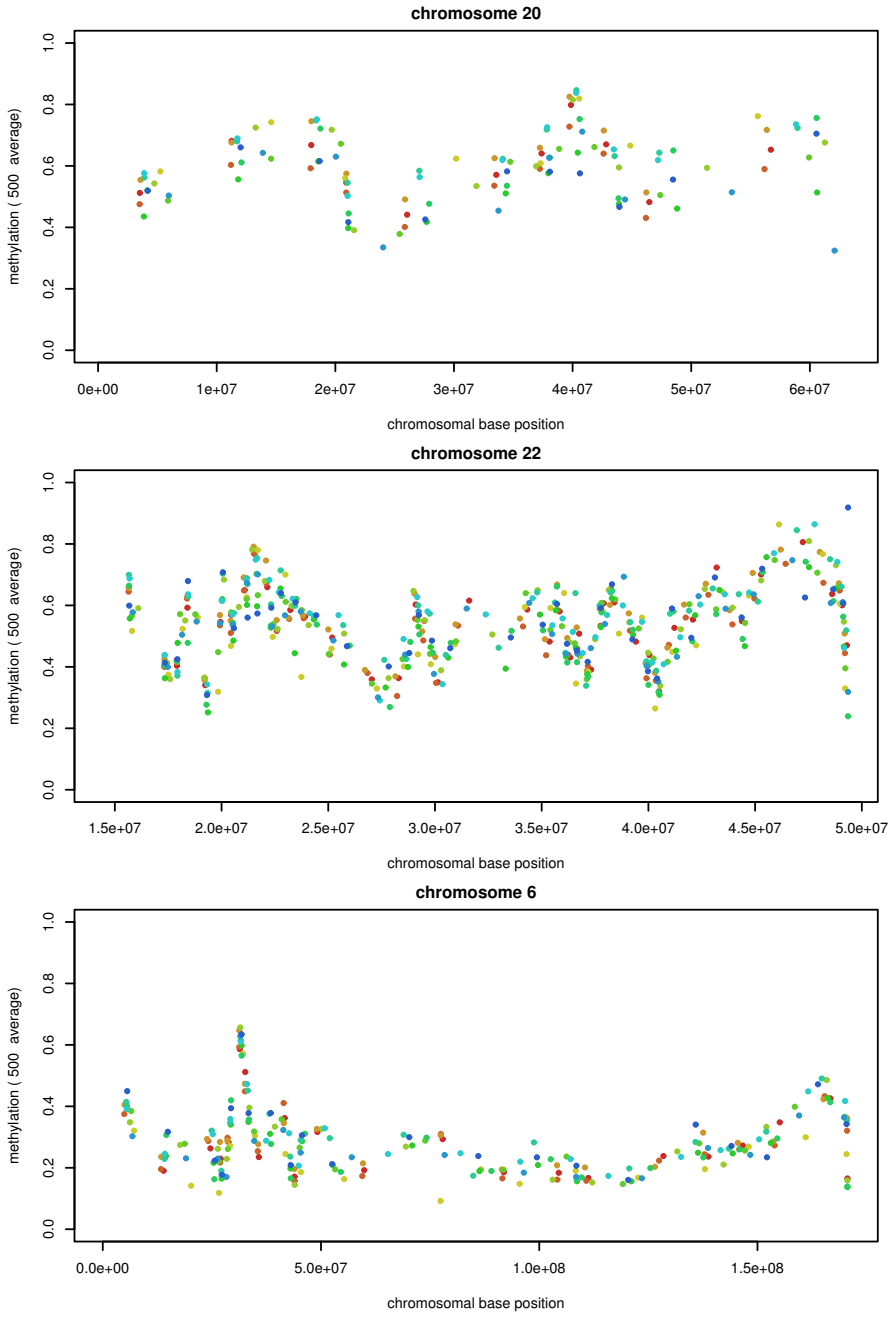

Figure 7.18: Chromosomal methylation profiles in spacial coordiante system differentiating different tissue types. These profiles are most likely biased by our design criteria and not representative for the chromosomes. Color coding: heart muscle skeletal muscle liver sperm fetal liver placenta fibroblasts keratinocytes CD8+ lymphocytes CD4+ lymphocytes fetal skeletal muscle melanocytes

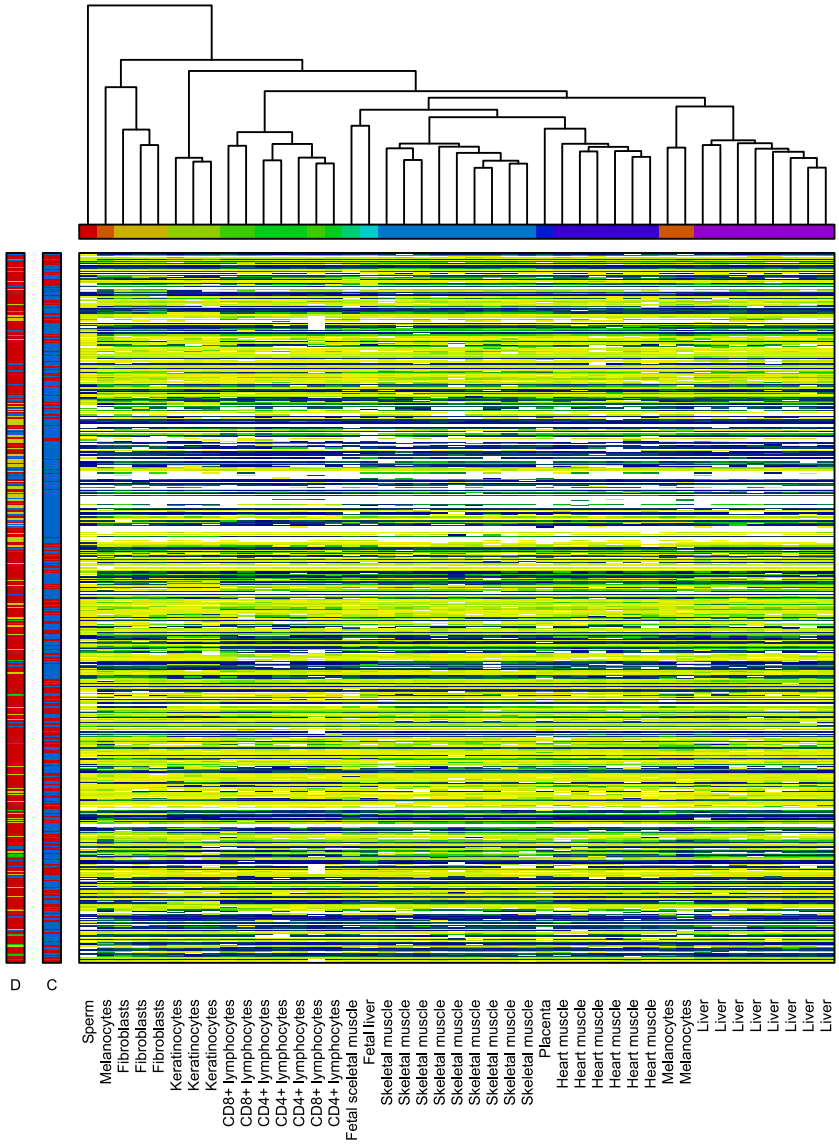

Figure 7.19: **Colored methylation plot, chromosome 6**. Color codes from ■ 0% over ■ 50% to ■ 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by chromosomal position. X axis: **Tissue type**. Horizontal dimension group descriptors: **Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **C) CpG island** ■ NO island ■ is island. **D) Amplificate subtype** ■ promoter ■ exon ■ other ■ intron.

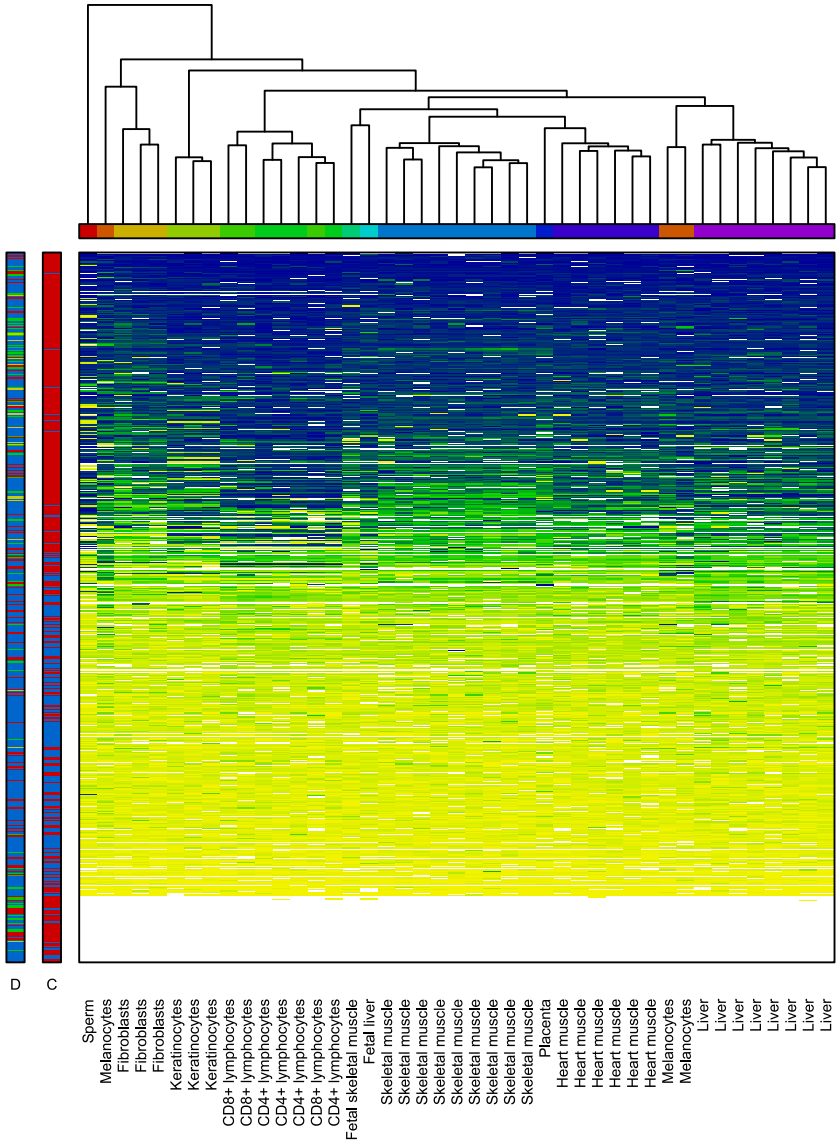

Figure 7.20: **Colored methylation plot, chromosome 6**. Color codes from ■ 0% over ■ 50% to ■ 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by average methylation. X axis: **Tissue type**. Horizontal dimension group descriptors: **Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **C) CpG island state** ■ NO island ■ is island. **D) Amplificate subtype** ■ intron ■ promoter ■ exon ■ other.

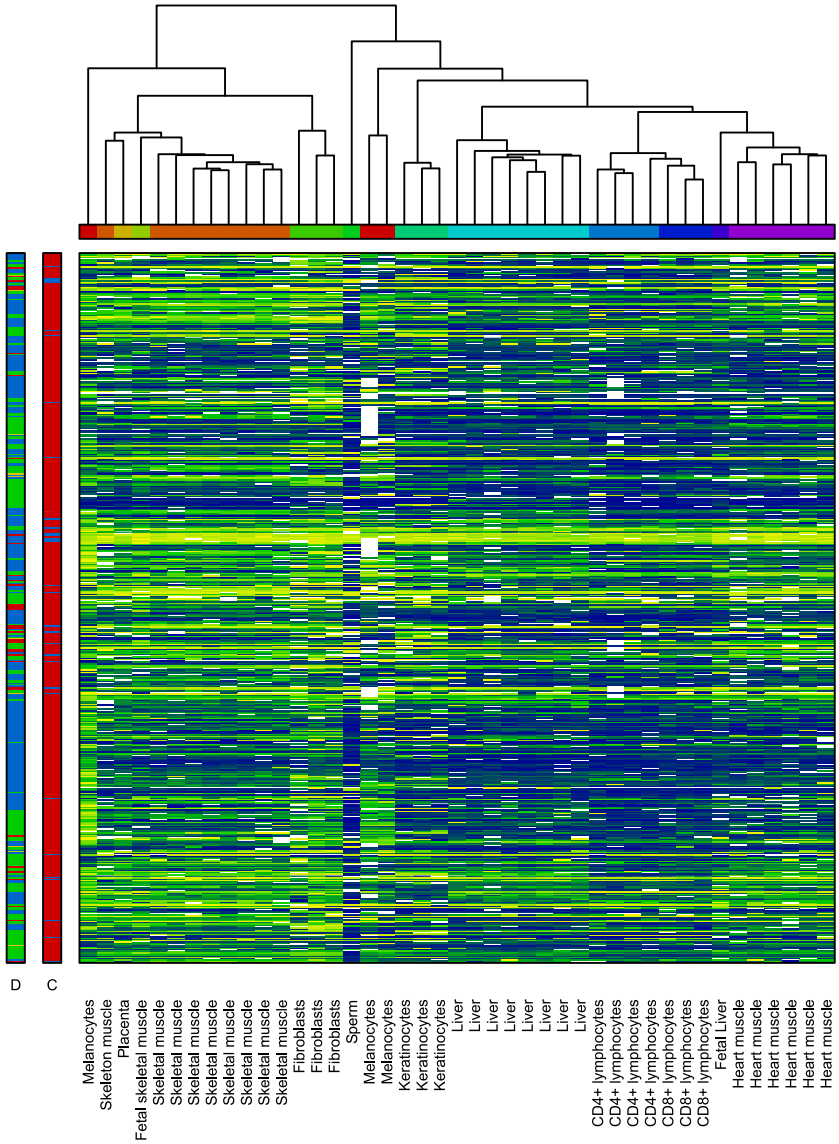

Figure 7.21: **Colored methylation plot, chromosome 20**. Color codes from ■ 0% over ■ 50% to ■ 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by chromosomal position. X axis: **Tissue type**. Horizontal dimension group descriptors: **Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **C) CpG island state** ■ NO island ■ is island. **D) Amplificate subtype** ■ promoter ■ other ■ intron ■ exon.

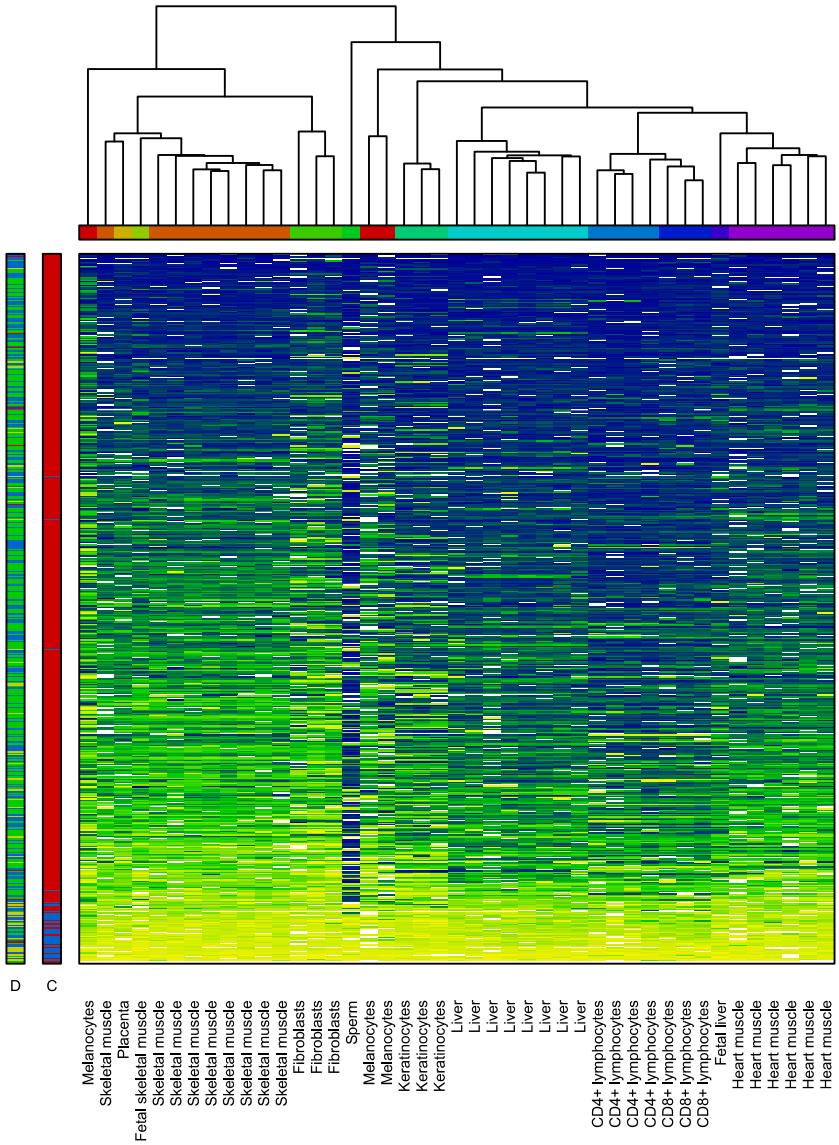

Figure 7.22: **Colored methylation plot, chromosome 20**. Color codes from ■ 0% over ■ 50% to ■ 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by average methylation. X axis: **Tissue type**. Horizontal dimension group descriptors: **Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **C) CpG island state** ■ NO island ■ is island. **D) Amplificate subtype** ■ exon ■ intron ■ other ■ promoter.

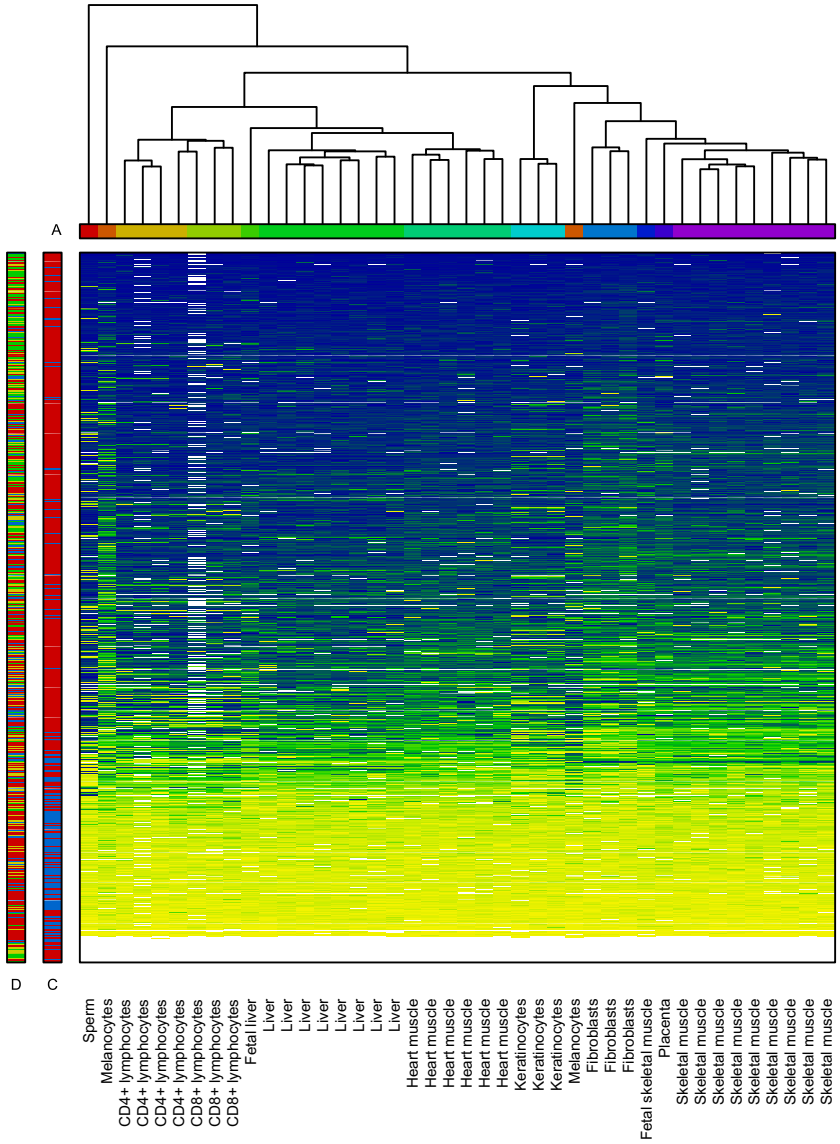

Figure 7.23: **Colored methylation plot, chromosome 22**. Color codes from 0% over 50% to 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by average methylation. X axis: **Tissue type**. Horizontal dimension group descriptors: **A) Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **C) CpG island state** No island is island. **D) Amplificate subtype** promoter other exon intron.

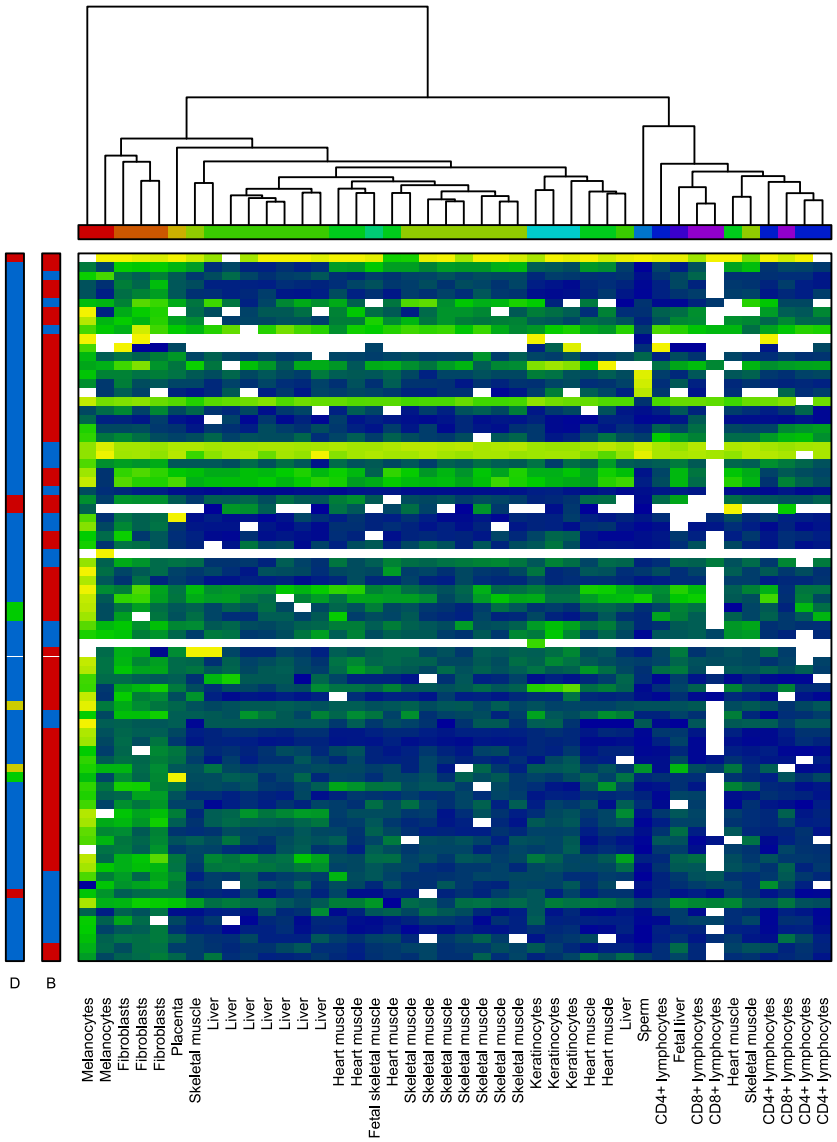

Figure 7.24: **Colored methylation plot, Ig segments in chromosome 22 only**. Color codes from ■ 0% over ■ 50% to ■ 100% methylation. White parts lack measurement data. Y axis: genomic sites ordered by chromosomal position. X axis: **Tissue type**. Horizontal dimension group descriptors: **Tissue** Color codes correspond to the tissue types given at the bottom of the plot. Vertical dimension group descriptors: **B) Amplificate type** ■ Ig Segment ■ Ig Pseudogene Segment. **D) Amplificate subtype** ■ other ■ promoter ■ exon ■ intron.

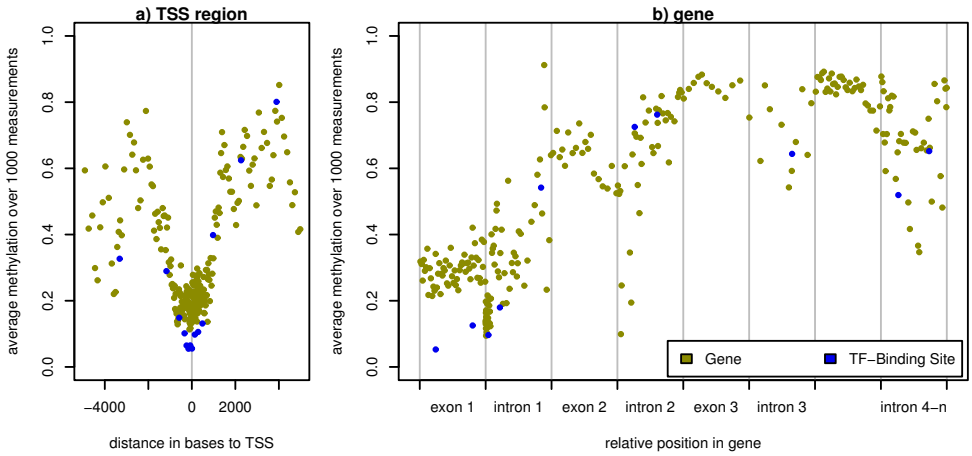

Figure 7.25: Methylation distribution based on CpG methylation in different gene types and all tissue types. Data was grouped by TF binding site classification: Within TSS regions 13,344 measurements were associated with TF binding sites (263,593 were not), in genes 9,994 measurements could be associated with binding sites, (272320 not). Details and numbers of high and low methylated proportions are available in table 7.5 and 7.6.

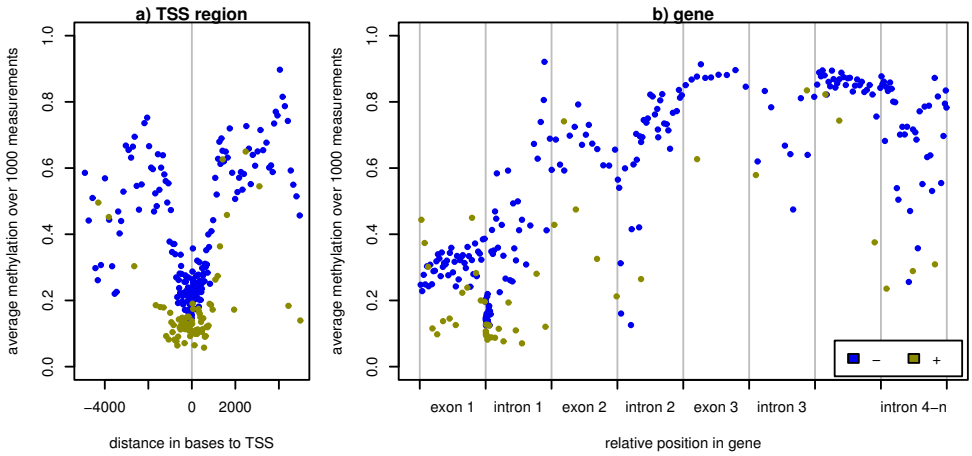

Figure 7.26: Methylation distribution based on CpG methylation in different gene types and all tissue types. Data was grouped by CpG island classification, whereby withinin genes 74,063 measurements were available inside and 232,962 outside of CpG islands, in TSS regions it was 46,662 in islands and 232,962 outside. Details and numbers of high and low methylated proportions are available in table 7.7 and 7.8.

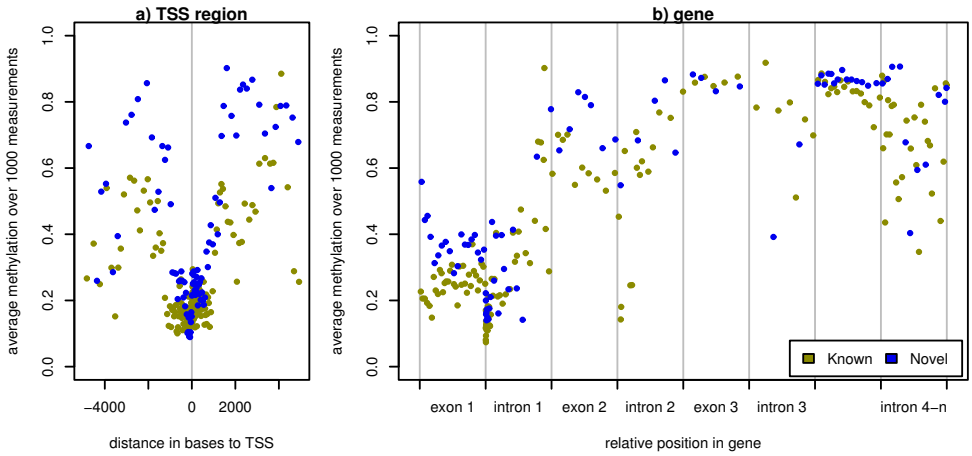

Figure 7.27: Methylation distribution based on CpG methylation in different gene types and all tissue types. Data was grouped by known genes and novel/predicted genes: 157,995 measurements for known genes around TSS (91,740 for novel), and 160,324 for known genes within the gene body (83,317 for novel). Details and numbers of high and low methylated proportions are available in table 7.3 and 7.4.

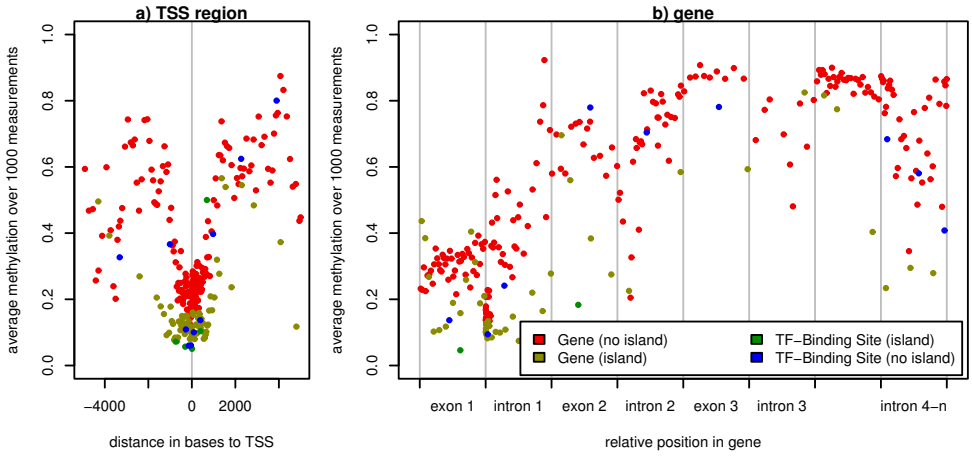

Figure 7.28: Methylation distribution based on CpG methylation in different gene types and all tissue types. DAta was grouped by TF binding site and CpG island association. Data within TSS regions: 69,360 in islands, 9,342 TF associated, 4,002 both, 192,103 neither island, nor TF associated. Data within genes: 44,744 in islands, 8,288 TF associated, 1,706 both, 220,280 neither island, nor TF associated. Data within genes: Details and numbers of high and low methylated proportions are available in table 7.9 and 7.10 in the appendix. Individual profiles grouping the data only by CpG island or TF binding site are found in Fig. 7.25 and 7.26.

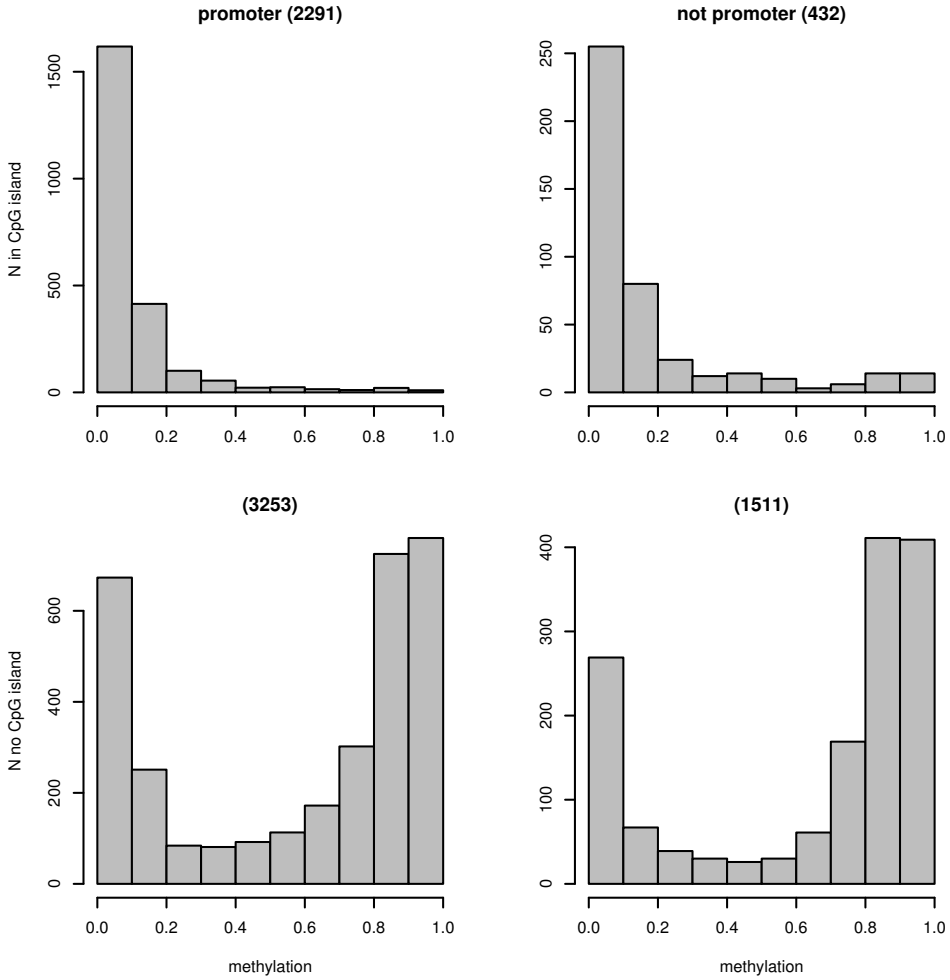

Figure 7.29: Methylation distribution in **Chromosome 6**. Average amplificate methylation levels per tissue (7487 total). Data is shown grouped by **CpG island** and **promoter** association.

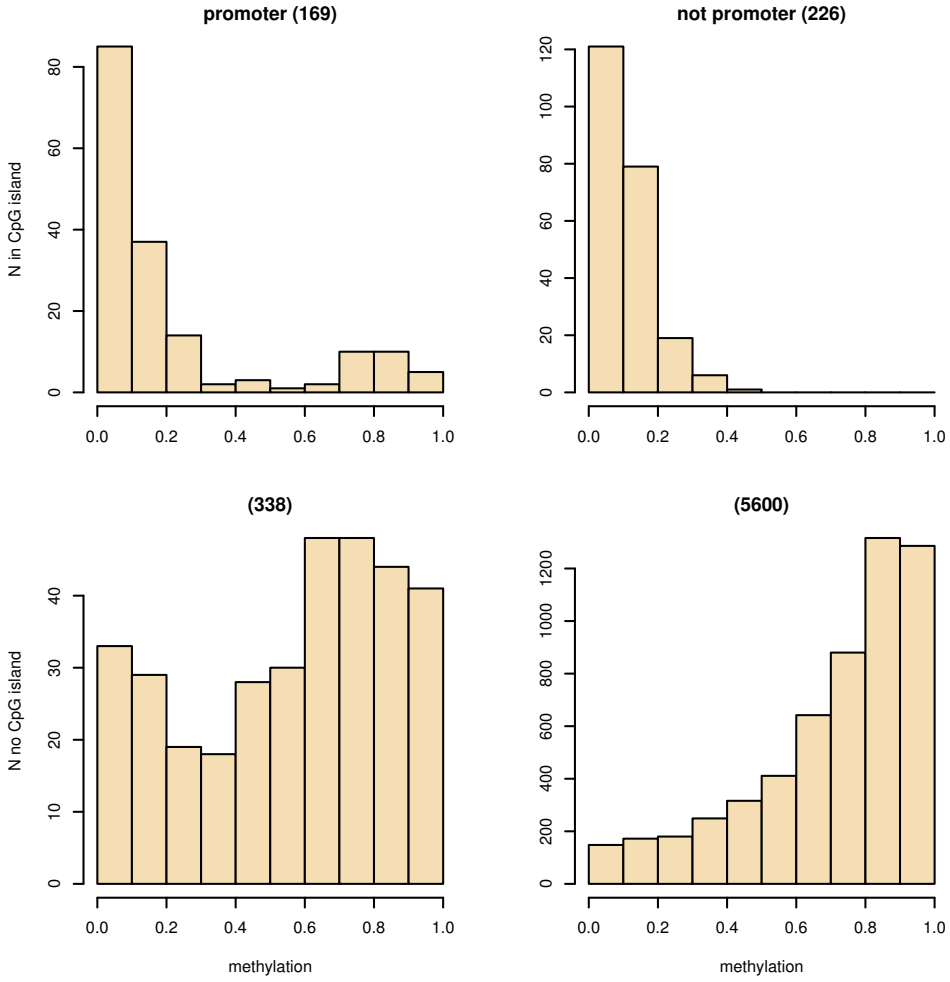

Figure 7.30: Methylation distribution in **Chromosome 20**. Average amplificate methylation levels per tissue. Data is shown grouped by **CpG island** and **promoter** association.

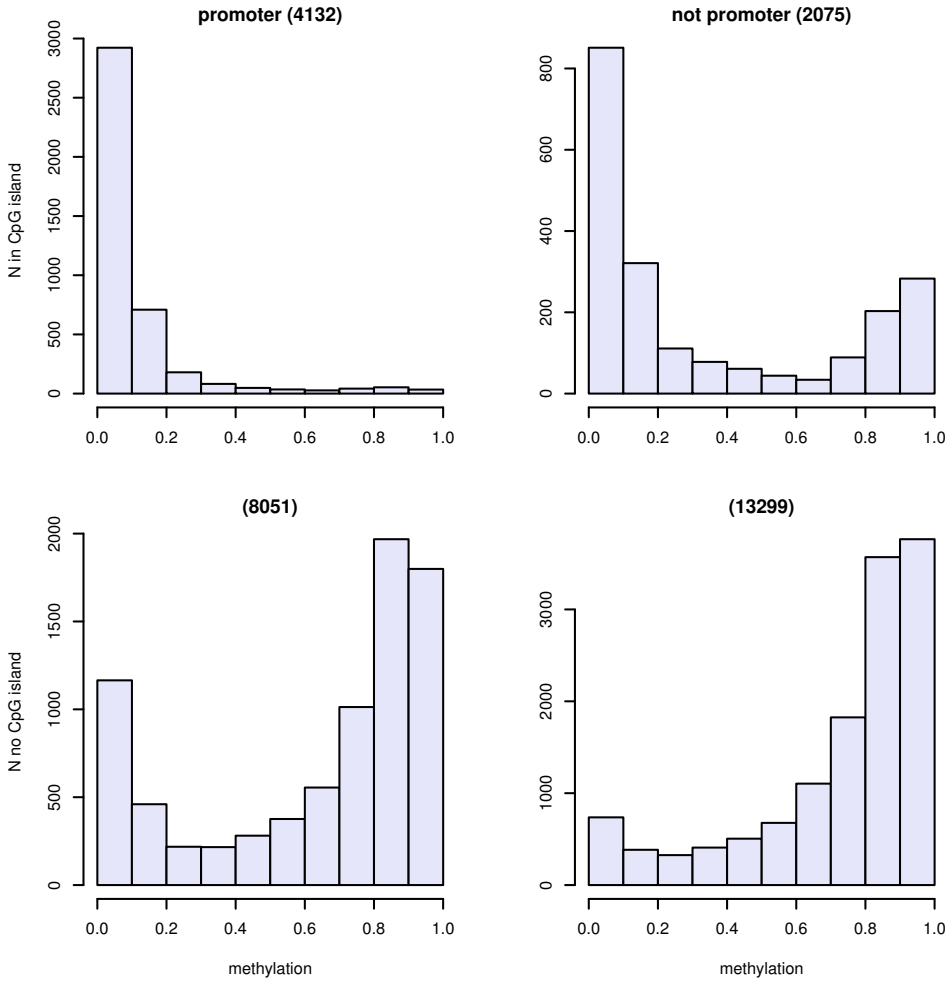

Figure 7.31: Methylation distribution (all data). Data is shown grouped by **CpG island** and **promoter** association.

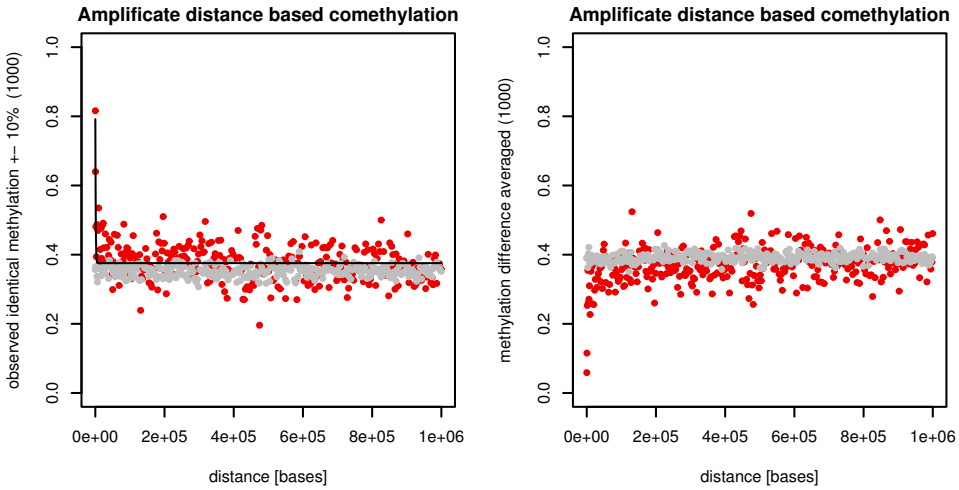

Figure 7.32: Co-Methylation, long distance between amplificates.

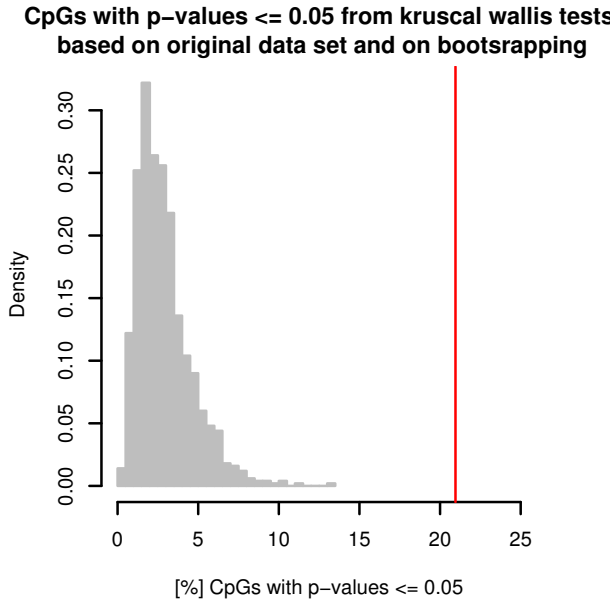

Figure 7.33: Ratio of tissue specific methylated CpGs found to be significant (p value $\leq$ 0.05) with Kruskal Wallis tests in MHC data from Esme 3.0.0. Tests were performed CpG wise. Bootstrapping was performed 1000 times conserving correlation structures and the size and amount of groups by sampling index vectors assigning original data to samples. The histogram is based on the bootstrapping values, the vertical line is at the position of the real ratio of significant CpGs in the data.

7.3 Tables

Table 7.1: Table for gene data. Complete data.

methylation	10%	90%
all gene data	31	31
1 exon	54	16
1 exon 0 - 1/3	55	15
1 exon 1/3 - 2/3	55	15
1 exon 2/3 - 1	52	18
1 intron	54	13
1 intron 0 - 1/3	62	8
1 intron 1/3 - 2/3	43	13
1 intron 2/3 - 1	26	37
2 exon	23	43
2 exon 0 - 1/3	23	46
2 exon 1/3 - 2/3	20	45
2 exon 2/3 - 1	28	37
2 intron	14	34
2 intron 0 - 1/3	30	23
2 intron 1/3 - 2/3	7	38
2 intron 2/3 - 1	4	42
3 exon	4	60
3 exon 0 - 1/3	2	62
3 exon 1/3 - 2/3	4	57
3 exon 2/3 - 1	6	62
3 intron	13	44
3 intron 0 - 1/3	11	49
3 intron 1/3 - 2/3	19	34
3 intron 2/3 - 1	9	49
4 exon	3	58
4 exon 0 - 1/3	3	62
4 exon 1/3 - 2/3	2	58
4 exon 2/3 - 1	7	52
4 intron	15	43
4 intron 0 - 1/3	11	49
4 intron 1/3 - 2/3	23	33
4 intron 2/3 - 1	12	44

Table 7.2: Table for tss data. Complete data.

methylation	<10%	>90%
all TSS data	51	17
-5000 to -3000	42	24

Table 7.2: em (continued)

methylation	<10%	>90%
-3000 to -1000	31	27
-1000 to 1000	62	8
1000 to 3000	27	35
3000 to 5000	19	42

Table 7.3: Table for gene data. Known and novel classification.

group	Known gene		new/predicted	
group	Known gene		new/predicted	
methylation	< 10%	> 90%	< 10%	> 90%
all gene data	36	28	29	36
1 exon	58	13	46	22
1 exon 0 - 1/3	60	10	43	25
1 exon 1/3 - 2/3	57	13	50	20
1 exon 2/3 - 1	56	15	46	22
1 intron	56	12	55	14
1 intron 0 - 1/3	67	5	60	11
1 intron 1/3 - 2/3	41	13	52	10
1 intron 2/3 - 1	29	37	11	47
2 exon	25	41	13	51
2 exon 0 - 1/3	22	48	21	49
2 exon 1/3 - 2/3	29	36	6	55
2 exon 2/3 - 1	25	35	19	46
2 intron	26	28	12	42
2 intron 0 - 1/3	36	20	20	34
2 intron 1/3 - 2/3	15	34	3	51
2 intron 2/3 - 1	4	50	9	44
3 exon	3	62	1	63
3 exon 0 - 1/3	2	62	1	66
3 exon 1/3 - 2/3	4	62	2	54
3 exon 2/3 - 1	3	63	2	69
3 intron	11	47	30	33
3 intron 0 - 1/3	4	63	35	28
3 intron 1/3 - 2/3	16	40	41	18
3 intron 2/3 - 1	12	40	9	58
4 exon	3	55	2	64
4 exon 0 - 1/3	2	60	3	66
4 exon 1/3 - 2/3	1	54	2	65
4 exon 2/3 - 1	8	49	2	60
4 intron	15	42	10	51
4 intron 0 - 1/3	14	44	1	66
4 intron 1/3 - 2/3	16	40	27	28
4 intron 2/3 - 1	16	41	4	56

Table 7.4: Table for tss data. Known and novel classification.

group	Known gene		new/predicted	
group	Known gene		new/predicted	
methylation	< 10%	> 90%	< 10%	> 90%
all TSS data	57	12	43	23

Table 7.4: em (continued)

group	Known gene		new/predicted	
-5000 to -3000	51	19	37	30
-3000 to -1000	43	20	12	39
-1000 to 1000	66	6	60	10
1000 to 3000	37	24	14	49
3000 to 5000	26	36	13	50

Table 7.5: Table for gene data. TF binding site classification.

group	no TF		TF	
methylation	< 10%	> 90%	< 10%	> 90%
all gene data	31	31	38	22
1 exon	54	16	76	1
1 exon 0 - 1/3	55	14	85	0
1 exon 1/3 - 2/3	55	15	75	1
1 exon 2/3 - 1	52	18	61	3
1 intron	53	13	62	7
1 intron 0 - 1/3	62	8	74	1
1 intron 1/3 - 2/3	43	13	40	13
1 intron 2/3 - 1	27	38	18	38
2 exon	23	43	10	62
2 exon 0 - 1/3	23	46	13	71
2 exon 1/3 - 2/3	20	44	12	55
2 exon 2/3 - 1	28	36	0	53
2 intron	15	34	4	37
2 intron 0 - 1/3	31	23	1	38
2 intron 1/3 - 2/3	8	38	6	37
2 intron 2/3 - 1	4	42	1	37
3 exon	3	61	29	32
3 exon 0 - 1/3	2	63	27	37
3 exon 1/3 - 2/3	4	58	21	34
3 exon 2/3 - 1	4	63	38	24
3 intron	13	45	10	13
3 intron 0 - 1/3	11	50	9	8
3 intron 1/3 - 2/3	19	34	11	44
3 intron 2/3 - 1	9	49		
4 exon	3	58	12	48
4 exon 0 - 1/3	2	62	14	40
4 exon 1/3 - 2/3	2	58	4	61
4 exon 2/3 - 1	7	52	20	36
4 intron	15	43	19	35
4 intron 0 - 1/3	11	49	10	35
4 intron 1/3 - 2/3	24	33	28	32

Table 7.5: em (continued)

group	no TF		TF	
4 intron 2/3 - 1	12	44	17	39

Table 7.6: Table for tss data. TF binding site classification.

group	no TF		TF	
methylation	< 10%	> 90%	< 10%	> 90%
all TSS data	51	17	63	8
-5000 to -3000	42	24	54	10
-3000 to -1000	31	28	35	18
-1000 to 1000	62	8	76	1
1000 to 3000	28	35	13	45
3000 to 5000	20	42	6	35

Table 7.7: Table for gene data. CpG island classification.

group	no island		CpG island	
methylation	< 10%	> 90%	< 10%	> 90%
all gene data	27	34	57	15
1 exon	52	16	62	14
1 exon 0 - 1/3	55	14	58	14
1 exon 1/3 - 2/3	52	17	72	6
1 exon 2/3 - 1	49	17	60	20
1 intron	49	17	69	2
1 intron 0 - 1/3	60	10	72	1
1 intron 1/3 - 2/3	38	16	64	1
1 intron 2/3 - 1	18	46	56	6
2 exon	20	44	38	37
2 exon 0 - 1/3	23	42	24	53
2 exon 1/3 - 2/3	16	47	45	25
2 exon 2/3 - 1	21	41	58	19
2 intron	13	34	52	22
2 intron 0 - 1/3	31	22	55	11
2 intron 1/3 - 2/3	5	39	92	0
2 intron 2/3 - 1	4	41	7	76
3 exon	1	64	30	32
3 exon 0 - 1/3	1	66	21	25
3 exon 1/3 - 2/3	1	62	32	33
3 exon 2/3 - 1	1	66	36	38
3 intron	14	43	9	57
3 intron 0 - 1/3	12	51	11	32
3 intron 1/3 - 2/3	21	34	13	41

Table 7.7: em (continued)

group	no island		CpG island	
3 intron 2/3 - 1	10	43	8	72
4 exon	2	60	16	46
4 exon 0 - 1/3	2	63	9	60
4 exon 1/3 - 2/3	1	59	11	49
4 exon 2/3 - 1	4	56	33	17
4 intron	12	45	56	10
4 intron 0 - 1/3	7	52	52	7
4 intron 1/3 - 2/3	20	34	63	14
4 intron 2/3 - 1	11	45	52	8

Table 7.8: Table for tss data. CpG island classification.

group	no island		CpG island	
methylation	< 10%	> 90%	< 10%	> 90%
all TSS data	46	20	66	6
-5000 to -3000	43	24	37	25
-3000 to -1000	24	34	62	3
-1000 to 1000	59	10	71	3
1000 to 3000	23	38	42	23
3000 to 5000	17	44	40	13

Table 7.9: Table for gene data. TF binding site assessment, CpG island grouping.

group	in island		NO island		TF in island		TF outside island	
methylation	< 10%	> 90%	< 10%	> 90%	< 10%	> 90%	< 10%	> 90%
all gene data	56	15	27	34	71	0	31	27
1 exon	62	14	52	16	88	0	68	1
1 exon 0 - 1/3	57	14	55	14	89	0	82	0
1 exon 1/3 - 2/3	70	7	52	17	92	0	61	1
1 exon 2/3 - 1	61	19	49	17	72	3	56	3
1 intron	69	2	49	17	66	0	60	10
1 intron 0 - 1/3	72	1	59	11	78	0	73	1
1 intron 1/3 - 2/3	65	1	38	16	48	1	32	26
1 intron 2/3 - 1	56	6	18	47	0	0	18	40
2 exon	38	38	20	43	77	0	6	65
2 exon 0 - 1/3	24	53	23	41			13	71
2 exon 1/3 - 2/3	44	26	17	46	77	0	0	65
2 exon 2/3 - 1	58	19	21	41			0	53
2 intron	52	22	13	34			4	37
2 intron 0 - 1/3	55	11	32	21			1	38

Table 7.9: em (continued)

group	in island		NO island		TF in island		TF outside island	
2 intron 1/3 - 2/3	92	0	5	40			6	37
2 intron 2/3 - 1	7	76	4	41			1	37
3 exon	26	38	1	64	50	0	0	77
3 exon 0 - 1/3	14	30	1	66	56	0	0	72
3 exon 1/3 - 2/3	31	37	1	62	39	0	0	75
3 exon 2/3 - 1	32	48	1	65	52	0	0	92
3 intron	10	59	15	43	5	4	13	20
3 intron 0 - 1/3	12	37	12	52	5	4	14	12
3 intron 1/3 - 2/3	13	41	21	34			11	44
3 intron 2/3 - 1	8	72	10	43				
4 exon	16	46	2	60			12	48
4 exon 0 - 1/3	9	60	2	63			14	40
4 exon 1/3 - 2/3	11	49	1	59			4	61
4 exon 2/3 - 1	33	17	4	57			20	36
4 intron	56	10	12	46			19	35
4 intron 0 - 1/3	52	7	7	53			10	35
4 intron 1/3 - 2/3	63	14	20	34			28	32
4 intron 2/3 - 1	52	8	10	45			17	39

Table 7.10: Table for tss data. TF binding site assessment, CpG island grouping.

group	in island		NO island		TF in island		TF outside island	
methylation	< 10%	> 90%	< 10%	> 90%	< 10%	> 90%	< 10%	> 90%
all TSS data	65	6	46	20	78	0	56	12
-5000 to -3000	39	23	42	25			54	10
-3000 to -1000	62	3	23	35	73	2	30	20
-1000 to 1000	71	3	59	10	78	0	74	1
1000 to 3000	42	23	23	38			13	45
3000 to 5000	40	13	18	44			6	35

Bibliography

Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A. R., Gutig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Muller, V., Otto, T., Scholz, C., Schulz, W. A., Seifert, H. H., Schwope, I., Ziebarth, H., Berlin, K., Piepenbrock, C. & Olek, A. (2002) Tumour class prediction and discovery by microarray-based dna methylation analysis. *Nucleic Acids Res,* **30** (5), e21. 16

Ashurst, J. L., Chen, C.-K., Gilbert, J. G. R., Jekosch, K., Keenan, S., Meidl, P., Searle, S. M., Stalker, J., Storey, R., Trevanion, S., Wilming, L. & Hubbard, T. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res,* **33** (Database issue), 459–465. 71

Avner, P. & Heard, E. (2001) X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet,* **2** (1), 59–67. 14

Barton, G. J. (1993) An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput Appl Biosci,* **9** (6), 729–34. 29

Beck, S., Olek, A. & Walter, J. (1999) From genomics to epigenomics: a loftier view of life. *Nat Biotechnol,* **17** (12), 1144. Comment. 24

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J. r., Gingeras, T. R., Schreiber, S. L. & Lander, E. S. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell,* **120** (2), 169–181. Comparative Study. 73

Bestor, T. H. (1998) Gene silencing. methylation meets acetylation. *Nature,* **393** (6683), 311–312. 10, 11, 13

Bestor, T. H. (2000) The dna methyltransferases of mammals. *Hum Mol Genet,* **9** (16), 2395–402. 13

Bird, A. P. (1986) Cpg-rich islands and the function of dna methylation. *Nature,* **321** (6067), 209–13. 12, 13, 72

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. & Gingeras, T. R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell,* **116** (4), 499–509. 93

Clark, S. J., Harrison, J. & Molloy, P. L. (1997) Sp1 binding is inhibited by (m)Cp(m)CpG methylation. *Gene,* **195** (1), 67–71. 93

Consortium, I. H. G. S. (2004) Finishing the euchromatic sequence of the human genome. *Nature,* **431** (7011), 931–945. 8

Costello, J. F., Smiraglia, D. J. & Plass, C. (2002) Restriction landmark genome scanning. *Methods,* **27** (2), 144–149. 97

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J. & Clamp, M. (2004) The Ensembl automatic gene annotation system. *Genome Res,* **14** (5), 942–950. Comparative Study. 8, 71

Dahl, C. & Guldberg, P. (2003) Dna methylation analysis techniques. *Biogerontology,* **4** (4), 233–50. 10, 16

Dear, S. & Staden, R. (1992) A standard file format for data from dna sequencing instruments. *DNA Seq,* **3** (2), 107–10. 21, 27

Down, T. A. & Hubbard, T. J. P. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res,* **12** (3), 458–461. 72

Duncan, B. K. & Miller, J. H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature,* **287** (5782), 560–561. 12, 89

Eckhardt, F., Beck, S., Gut, I. G. & Berlin, K. (2004) Future potential of the Human Epigenome Project. *Expert Rev Mol Diagn,* **4** (5), 609–618. 114

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. & Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet,* **38** (12), 1378–1385. Comparative Study. 1, 2, 4, 65, 70, 89

Ehrlich, M. (2003) Expression of various genes is controlled by dna methylation during mammalian development. *J Cell Biochem,* **88** (5), 899–910. 10

Feil, R. & Khosla, S. (1999) Genomic imprinting in mammals: an interplay between chromatin and dna methylation? *Trends Genet,* **15** (11), 431–435. 14

Ferguson-Smith, A. C. & Surani, M. A. (2001) Imprinting and the epigenetic asymmetry between parental genomes. *Science,* **293** (5532), 1086–1089. 14

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suner, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., Plass, C. & Esteller, M. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A,* **102** (30), 10604–10609. Comparative Study. 15, 108

Frank, R. & Koster, H. (1979) DNA chain length markers and the influence of base composition on electrophoretic mobility of oligodeoxyribonucleotides in polyacrylamide-gels. *Nucleic Acids Res,* **6** (6), 2069–2087. 47

Frigola, J., Song, J., Stirzaker, C., Hinshelwood, R. A., Peinado, M. A. & Clark, S. J. (2006) Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat Genet,* **38** (5), 540–549. 110

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. & Paul, C. L. (1992) A genomic sequencing protocol

that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proc Natl Acad Sci U S A,* **89** (5), 1827–31. 16, 23

Fuks, F. (2005) DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev,* **15** (5), 490–495. 9, 10

Grunau, C., Clark, S. J. & Rosenthal, A. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res,* **29** (13), E65–5. 26, 35

Grunau, C., Hindermann, W. & Rosenthal, A. (2000) Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes. *Hum Mol Genet,* **9** (18), 2651–2663. 12

Harrington, M. A., Jones, P. A., Imagawa, M. & Karin, M. (1988) Cytosine methylation does not affect binding of transcription factor Sp1. *Proc Natl Acad Sci U S A,* **85** (7), 2066–2070. 93

Holler, M., Westin, G., Jiricny, J. & Schaffner, W. (1988) Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. *Genes Dev,* **2** (9), 1127–1135. 93

Human Epigenome Consortium, Epigenomics AG, The Welcome Trust Sanger Institute & Centre National de Genotypage (2003). Human Epigenome Project. http://www.epigenome.org/. 24

Issa, J. P., Ottaviano, Y. L., Celano, P., Hamilton, S. R., Davidson, N. E. & Baylin, S. B. (1994) Methylation of the oestrogen receptor cpg island links ageing and neoplasia in human colon. *Nat Genet,* **7** (4), 536–40. 15, 107

Issa, J. P., Vertino, P. M., Boehm, C. D., Newsham, I. F. & Baylin, S. B. (1996) Switch from monoallelic to biallelic human igf2 promoter methylation during aging and carcinogenesis. *Proc Natl Acad Sci U S A,* **93** (21), 11757–62. 15, 107

Jones, P. A. (2002) Dna methylation and cancer. *Oncogene,* **21** (35), 5358–60. 10, 15

Jones, P. A. & Baylin, S. B. (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet,* **3** (6), 415–28. 15

Jones, P. A. & Laird, P. W. (1999) Cancer epigenetics comes of age. *Nat Genet,* **21** (2), 163–167. 15

Jones, P. A. & Martienssen, R. (2005) A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res,* **65** (24), 11241–11246. Congresses. 115

Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J. & Wolffe, A. P. (1998) Methylated dna and mecp2 recruit histone deacetylase to repress transcription. *Nat Genet,* **19** (2), 187–91. 13

Kass, S. U., Pruss, D. & Wolffe, A. P. (1997) How does dna methylation repress transcription? *Trends Genet,* **13** (11), 444–449. 13

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm,

L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S. & Chen, Y. J. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409** (6822), 860–921. 8

Lewin, B. (2003) *Genes*. 8th edition edition,, Benjamin Cummings. 8

Lewin, J., Schmitt, A. O., Adorjan, P., Hildmann, T. & Piepenbrock, C. (2004) Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics,* **20** (17), 3005–3012. Comparative Study. 1, 2, 4, 25

Li, L. (2001). Dna sequencing and parametric deconvolution. Florida State University. 49

Mancini, D. N., Singh, S. M., Archer, T. K. & Rodenhiser, D. I. (1999) Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. *Oncogene,* **18** (28), 4108–4119. Comparative Study. 93

Morgan, H. D., Sutherland, H. G., Martin, D. I. & Whitelaw, E. (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet,* **23** (3), 314–318. 8, 12

Nan, X., Campoy, F. J. & Bird, A. (1997) Mecp2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell,* **88** (4), 471–81. 13

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N. & Bird, A. (1998) Transcriptional repression by the methyl-cpg-binding protein mecp2 involves a histone deacetylase complex. *Nature,* **393** (6683), 386–389. 13

Olek, A., Oswald, J. & Walter, J. (1996) A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res,* **24** (24), 5064–5066. 16, 23

Paul, C. L. & Clark, S. J. (1996) Cytosine methylation: quantitation by automated genomic sequencing and GENESCAN™ analysis. *Biotechniques,* **21** (1), 126–33. 16

Qiu, P., Soder, G. J., Sanfiorenzo, V. J., Wang, L., Greene, J. R., Fritz, M. A. & Cai, X. Y. (2003) Quantification of single nucleotide polymorphisms by automated dna sequencing. *Biochem Biophys Res Commun,* **309** (2), 331–338. 26

Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K. & Beck, S. (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol,* **2** (12), 1–14. 1, 2, 4, 65

Reik, W., Dean, W. & Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science,* **293** (5532), 1089–93. 14

Rideout, W. M. r., Coetzee, G. A., Olumi, A. F. & Jones, P. A. (1990) 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science,* **249** (4974), 1288–1290. 15

Shiota, K. (2004) DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. *Cytogenet Genome Res,* **105** (2-4), 325–334. 97

Shiota, K., Kogo, Y., Ohgane, J., Imamura, T., Urano, A., Nishino, K., Tanaka, S. & Hattori, N. (2002) Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. *Genes Cells,* **7** (9), 961–969. 97

Siegmund, K. D. & Laird, P. W. (2002) Analysis of complex methylation data. *Methods,* **27** (2), 170–178. 10, 16

Smiraglia, D. J., Rush, L. J., Fruhwald, M. C., Dai, Z., Held, W. A., Costello, J. F., Lang, J. C., Eng, C., Li, B., Wright, F. A., Caligiuri, M. A. & Plass, C. (2001) Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum Mol Genet,* **10** (13), 1413–1419. Comparative Study. 12

Smith, T. F. & Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol,* **147** (1), 195–197. 29

Song, F., Smith, J. F., Kimura, M. T., Morrow, A. D., Matsuyama, T., Nagase, H. & Held, W. A. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc Natl Acad Sci U S A,* **102** (9), 3336–3341. 97

Strichman-Almashanu, L. Z., Lee, R. S., Onyango, P. O., Perlman, E., Flam, F., Frieman, M. B. & Feinberg, A. P. (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res,* **12** (4), 543–554. 12

Tost, J., Schatz, P., Schuster, M., Berlin, K. & Gut, I. G. (2003) Analysis and accurate quantification of CpG methylation by MALDI mass spectrometry. *Nucleic Acids Res,* **31** (9), e50. 36

Vetterling, W. T., Teukolsky, S. A., Press, W. H. & Flannery, B. P. (2002) *Numerical Recipes in C++, The Art of Scientific Computing*. 2nd edition edition,, Cambridge University Press, Cambridge. 53, 54

Walter, J. & Paulsen, M. (2003) Imprinting and disease. *Semin Cell Dev Biol,* **14** (1), 101–110. 14

Warnecke, P. M., Stirzaker, C., Melki, J. R., Millar, D. S., Paul, C. L. & Clark, S. J. (1997) Detection and measurement of pcr bias in quantitative methylation analysis of bisulphite-treated dna. *Nucleic Acids Res,* **25** (21), 4422–4426. 35

Warnecke, P. M., Stirzaker, C., Song, J., Grunau, C., Melki, J. R. & Clark, S. J. (2002) Identification and resolution of artifacts in bisulfite sequencing. *Methods,* **27** (2), 101–107. 26, 35

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau,

A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C. & Lander, E. S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature,* **420** (6915), 520–62. 13

Wilson, V. L., Smith, R. A., Ma, S. & Cutler, R. G. (1987) Genomic 5-methyldeoxycytidine decreases with age. *J Biol Chem,* **262** (21), 9948–51. 15, 107

Xu, G. L., Bestor, T. H., Bourc'his, D., Hsieh, C. L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J. J. & Viegas-Pequignot, E. (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature,* **402** (6758), 187–191. 110

Yeivin, A. & Razin, A. (1993) Gene methylation patterns and expression. *EXS,* **64**, 523–68. 13

Yoo, C. B. & Jones, P. A. (2006) Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov,* **5** (1), 37–50. 10

Zeng, W., Kajigaya, S., Chen, G., Risitano, A. M., Nunez, O. & Young, N. S. (2004) Transcript profile of CD4+ and CD8+ T cells from the bone marrow of acquired aplastic anemia patients. *Exp Hematol,* **32** (9), 806–814. 96

Zhang, X.-P. & Allison, D. B. (2002). Iterative deconvolution for automatic basecalling in sequencing analysis. Technical report Applied Biosystems. 49