IDENTIFICATION AND PRIORITIZATION OF GENOMIC LOCI WITH DISEASE-SPECIFIC METHYLATION

Dissertation zur Erlangung des Grades
„Doktor der Ingenieurwissenschaften“
der Naturwissenschaftlich-Technischen Fakultät I
der Universität des Saarlandes

Author:
Yassen Assenov

Saarbrücken, 2014

UNIVERSITÄT
DES
SAARLANDES

Tag des Kolloquiums	10. Juni 2014
Dekan der Fakultät	Prof. Dr. Markus Bläser
Vorsitzender des Prüfungsausschusses	Prof. Dr. Volkhard Helms
Erstgutachter	Prof. Dr. Dr. Thomas Lengauer
Zweitgutachter	Prof. Dr. Christoph Bock
Akademischer Beisitzer	Dr. Olga Kalinina

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, den

Yassen Assenov

Abstract

Epigenetic systems are an indispensable mechanism in development, they respond to environmental stimuli and are dysregulated in cancer and other diseases. DNA methylation is the best characterized and extensively studied epigenetic mark to date. In the past years, a number of assays have been designed to measure DNA methylation levels genome-wide. This thesis introduces computational techniques for handling DNA methylation data from microarray- and enrichment-based methods. It focuses on disease-oriented studies and addresses the questions of quality control and normalization, inter- and intra-group variability, identification of differentially methylated loci, prioritization of biomarker candidates and prediction of cancer type and other phenotypes.

The presented statistical approaches and heuristics facilitated important discoveries with clinical applications. We showed that neurological and autoimmune disorders can be characterized by their distinct methylation profiles. We observed a strong tissue-specific signal in the methylation profiles of healthy and cancer samples. We were able to accurately predict tumor type of origin of metastatic samples. We showed that neither adenocarcinoma, nor squamous cell carcinoma can be separated into two distinct subtypes with a characteristic global methylation profile. In colon cancer, we identified differentially methylated regions with a potential to be used as biomarkers for predicting microsatellite instability.

Kurzfassung

Epigenetische Systeme sind ein unverzichtbarer Regulationsmechanismus in der Entwicklung von Lebewesen. Sie werden im Rahmen von Krebs und anderen Krankheiten fehlreguliert. DNA-Methylierung ist eine umfassend untersuchte und die am besten charakterisierte epigenetische Markierung. In den vergangenen Jahren wurde eine Reihe von Assays entwickelt, um DNA-Methylierungslevel genomweit zu messen. Diese Arbeit stellt Rechenverfahren für den Umgang mit DNA-Methylierungsdaten von Microarray- und Anreicherungsbasierten Methoden vor, mit dem Fokus auf krankheitsorientierte Studien. Sie befasst sich mit den Fragen der Qualitätskontrolle und Normalisierung, inter- und intra-Gruppen Variabilität, der Identifizierung von differentiell methylierten Regionen, Priorisierung von Biomarker-Kandidaten, sowie der Prognose von Krebstyp und anderen Phänotypen.

Die vorgestellten statistischen Ansätze und Heuristiken ermöglichten wichtigen Entdeckungen mit klinischer Anwendung. Wir könnten zeigen, dass neurologische und Autoimmunerkrankungen durch ihre unterschiedlichen Methylierungsmuster charakterisiert werden. Zudem beobachteten wir ein starkes Gewebe-spezifisches Signal in den Methylierungsprofilen von Krebsproben und gesunden Kontrollen. Dadurch gelang es uns, den ursprünglichen Tumortyp von Metastasen zu identifizieren. In Darmkrebs identifizierten wir differenziell methylierte Regionen, die potenziell als Biomarker zur Vorhersage der Mikrosatelliten-Instabilität verwendet werden können.

Acknowledgements

This work was carried out in the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics* in Saarbrücken. First and foremost, I would like to thank my supervisors Thomas Lengauer and Christoph Bock for introducing me to computational biology in general and computational epigenetics in particular. Their valuable guidance, encouragement and support are the true reason for the findings listed in this work, and many more that I omitted. The other department members helped me advance, I am particularly thankful to Adrian Alexa, André Altmann, Jasmina Bogojeska, Peter Ebert, Lars Feuerbach, Konstantin Halachev, Alexander Junge, Fabian Müller, Nora Speicher and Laura Toloşi for their constant feedback and valuable insights.

I was lucky to have interactions with biologists who are world renouned experts in the field of epigenetics and leaders of large international collaborations. In particular, I would like to thank Manel Esteller, Hendrik Stunnenberg and Jörn Walter for giving me the opportunity to visit their labs and the numerous discussions on the projects. This work would not have been possible without their support. I am also grateful to Olga Bogatyrova for her immense help in improving the introduction of this thesis and for providing Figure 1.2.

Last but certainly not least, I would like to thank Milena, Emma and Mischka for giving me neverending inspiration and aid.

Contents

List of Figures

List of Tables

1 Introduction

This thesis describes important techniques for data quality control, differential methylation and prioritization of genomic loci with respect to their potential relevance to disease. The following sections introduce the biological and mathematical concepts used in later chapters. The biological background is the definition of DNA methylation and other epigenetic mechanisms. Detailed explanations of these concepts are available in the books *Molecular Biology of the Cell* [5] and *Epigenetics* [7]. On the mathematical side, several machine learning approaches are briefly described. The book *The Elements of Statistical Learning* [57] presents these techniques in a comprehensive and well structured form; it also includes insightful comparative analyses.

1.1 Epigenetics and transcriptional regulation

Every living cell in a multicellular organism is programmed and adapts the expression patterns of its DNA through a variety of processes that can be broadly classified as *genetic* and *epigenetic*. The first group consists of very rare adaptive mechanisms that affect irreversibly the genetic material. Examples for genetic changes include mutations, copy number alterations, insertions, deletions, and various forms of recombinations.

Epigenetic mechanisms control the access to the DNA molecules, keeping the underlying sequence of bases intact [94]. Regulating the accessibility to the DNA is achieved through different forms of packaging the double-stranded molecule. The term epigenetics stems from the greek word *ἐπί* (over, above, outer). It was first coined as 'epigenotype' by Conrad Waddington in 1942 in the context of his studies describing the "whole complex of developmental processes" [117, 118].

The degree of DNA accessibility is controlled collectively via processes including DNA methylation, histone modifications and variant replacement, RNA interference and nucleosome positioning. The synchronous workings of these mechanisms determine an *epigenetic state* of a cell. By definition, epigenetic states are maintained during mitosis and are therefore inherited across cell generations. The following paragraphs describe the biochemical processes introduced above, with a pronounced focus on DNA methylation.

1.1.1 DNA Methylation

The methylation of the C5-atom of cytosines is established and maintained by a special family of enzymes – DNA methyltransferases (DNMTs) – using S-adenosyl methionine (SAM) as the methyl group donor (see Figure 1.1). The resulting base is referred to as 5-methylcytosine (5mC).

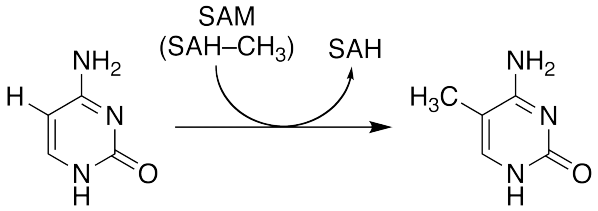


Figure 1.1: Skeletal structures for the methylation of cytosine to form 5-methylcytosine, using S-adenosyl methionine as the source of the methyl group and giving S-adenosyl-L-homocysteine as the by-product.

Methylated cytosines in human are located almost exclusively in the context of a cytosine-guanine dinucleotide (CpG). Note that this dinucleotide sequence is the reverse complement of itself, and therefore the CpG dinucleotides appear in equal number on both DNA strands. The presence of mirrored cytosines in this context allows the establishment of reciprocal methylation patterns during DNA replication. The major methyltransferase DNMT1 is responsible for adding methyl groups to hemimethylated sites, thereby maintaining the methylation patterns across cell generations.

5mC is biochemically less stable than cytosine and is prone to undergo hydrolytic deamination into thymine. Thymine is a valid DNA base which makes the DNA mismatch repair mechanisms less efficient in correcting deaminated 5mCs in comparison to other altered bases [102]. For this reason, CpG dinucleotides are comparatively rare in the genome. They are also very unevenly distributed; a small fraction of them concentrate in short genomic regions termed CpG islands (CGIs) [18, 54]. CGIs are on average 600 bases long and cover less than 1% of the genome. They often co-localize with gene promoters or enhancer elements, and the methylation state of all CpGs in one island tends to be consistent. CGIs associate with approximately half of all annotated human gene promoters[1], often affiliated with constitutively active and highly expressed genes [99, 38].

Role of DNA methylation

Methylation is a global phenomenon which affects between 70 and 80 percent of the CpGs in the human genome and also shows a high degree of tissue- and cell type specificity [44]. The distribution of methylated and unmethylated CpGs is also non-uniform. A small percentage of the CpGs are located in CGIs that are often unmethylated, however, the majority of the CpGs lie in the so-called open sea (areas in the genome outside islands) and are heavily methylated. Due to the general depletion of unmethylated CpGs in the genome, CGIs are prime targets for a variety of activating transcription factors. Examples for such include SP1, E2F and ETS1, all of which contain `CG`s in their consensus binding sequence [70].

[1] Based on the Takai-Jones criteria for CpG islands [106]. Other definitions of a CpG island are published, and the exact values for CGI properties (such as average length, genomic coverage and others) differ based on the definition used.

Methylation plays an important role in many cellular processes, including genomic imprinting [91, 13] and silencing of transposons and other retroviral elements [14]. The latter is critical for the stability of the DNA molecule. On an organismal level, methylation is involved in developmental processes, such as X chromosome inactivation [46] and stem cell differentiation [77].

1.1.2 Chromatin and its constituents

The 1.8 meter long DNA molecules in a cell nucleus are packaged together with *histone* proteins into a highly organized structure called chromatin. The role of chromatin is to keep DNA compact yet accessible, to create transcriptionally active and silent regions, to support DNA replication and to coordinate proper separation of genetic material to daughter cells during cell division. The histones are organized into nucleosomes, each one consisting of an octamer (pairs of histones H2A, H2B, H3 and H4) around which is wrapped 147 base pairs of DNA.

Transcriptional activity is associated with specific variants of the four histone proteins that form a nucleosome. In fact, there are close to 80 genes that encode different histone variants. The expression of major histones is tightly regulated and connected to the life stages of a cell. In addition, less common histone variants are produced by dedicated genes. These specialized histones are deposited preferentially in distinct nuclear domains and contribute to the characteristics of the corresponding genomic regions [97]. H3.3, for example, is a variant of the H3 histone protein and is present at transcriptionally active loci [3].

In addition, histone proteins can be altered by a set of modifications, preferentially at their N-terminal protruding ends, such as mono-, di- and trimethylation, acetylation, phosphorylation and ubiquitination. An encoding convention for these post-translational modifications includes typing the histone protein, followed by letter and position of the affected amino acid, and finally appending the first one or two letters of the modification. For example, H3K9ac stands for acetylation of the lysine residue at position 9 in H3's amino acid sequence. Similarly, trimethylation of residue 27 in H3 (again a lysine) is H3K27me3. Histone modifications are associated with different degrees of chromatin compaction and, consequently, different gene expression levels. Histone marks that coincide with high and low expression are referred to as *activating* and *repressing* marks, respectively. Similarly, one speaks of active and repressed chromatin. The first state coincides with domains in which chromatin is open and easily accessible, known as *euchromatic*, whereas the second state is usually in strongly compacted genomic regions called *heterochromatin*.

1.1.3 Other epigenetic mechanisms

In RNA interference, a short double-stranded RNA (dsRNA) silences its target genes by mRNA degradation or by inhibition of the process of translation [79]. Unlike DNA methylation and the histone modifications, this epigenetic mechanism is highly specific because dsRNAs bind to their targets in sequence-dependent manner [124].

1.1.4 Epigenetic genes

Fine-tuned machinery determines the active or repressed chromatin state of genomic regions through the combination of different epigenetic modifications at the DNA and histones. For example, the methylation of enhancer elements tends to reduce the expression of their targeted genes [51]. Most epigenetic pathways involve enzymes that:

- transfer a modification (writers);
- modify or revert a modification (editors);
- mediate the interactions of protein or protein complexes with the modified DNA and histones (readers).

The DNA methyltransferase (DNMT) protein family forms the writers of DNA methylation. Its most prominent member – DNMT1 – acts on hemimethylated CpG sites and is therefore responsible for maintaining symmetric methylation on both DNA strands, as already mentioned above. The proteins DNMT3a, -3b and -3L can also methylate unmethylated CpGs, a process referred to as *de novo* methylation. They are very active in early embryo development and set up the pattern of methylation.

Another well studied family of epigenetic proteins referred to in this thesis is the histone deacetylases (HDACs) – a class of enzymes that remove acetyl groups from lysine residues in histones and other proteins [100]. There are at least 10 known members of this family and many of them are active only in selected cell types.

1.2 Methylation in disease

Genetic and epigenetic defects have the ability to silence or activate genes. Destabilization of the chromatin facilitates chromosomal breaks and can lead to deletions, translocations and other rearrangements of chromosomes.

Epigenetic aberrations are prominent in cancer and occur in several other diseases, including diabetes, asthma and a variety of neurological [66], autoimmune [92] and cardiovascular [55] disorders. Abnormal low methylation is referred to as *hypomethylation*; similarly, the term *hypermethylation* denotes an anomalous high degree of methylation. Feinberg and Vogelstein discovered in 1983 that one major difference between cancer cells and healthy counterparts was aberrant DNA hypomethylation in tumors [48]. Although they analyzed only three regions in small sample numbers, they speculated about a global hypomethylation trend in cancer – a hypothesis now known to be largely true. In the following years, epigenetic alterations in cancer have been thoroughly investigated, both on the gene level and on the genome-wide level. The overall loss of 5mC often activates pericentromeric satellite DNA and repetitive elements which affects the integrity of the genome, leading to translocations, deletions and other genetic changes.

Methylation of cytosine strongly increases the rate of `C` to `T` transition mutations and is thought to be responsible for about one third of all disease-causing mutations in the

germline [33, 102]. In somatic cells, gene body methylation is a major cause of gene mutations in tumor suppressor genes, such as TP53, which encodes the p53 protein [65]. Also, it was demonstrated for numerous tumor-suppressor genes that hypermethylation of a CGI promoter is associated with gene silencing [34]. Cancer-specific methylation patterns in selected gene promoter sequences were found to correlate with clinical outcomes, and these patterns were named CpG island methylator phenotype (CIMP) [112]. This phenomenon is presented in more details in Chapter 7.

Other epigenetic modifications and the expression levels of epigenetic genes can also be used to stratify disease subtypes, severity, treatment responsiveness or to predict clinical outcomes. For example, H3 acetylation and H3K9me2 can discriminate between cancerous and nonmalignant prostate tissue, also, H3K4me3 can predict the recurrence of prostate-specific antigen accumulation after prostatectomy [45]. EZH2 expression is an independent prognostic marker that is correlated with the aggressiveness of prostate, breast and endometrial cancers [12]. Expression of the DNA repair gene O(6)-methylguanine-DNA methyltransferase (MGMT) antagonizes chemotherapy and radiation treatment [122].

1.2.1 Epigenetic therapies

In addition to the promising role of epigenetics for diagnosis and prognosis, epigenetic mechanisms are currently under investigation as potential targets in the treatment of cancers. In general, one can distinguish between drugs that target regulators of epigenetic patterns (e.g. DNMTs and HDACs) and those tailored to specific mutations in epigenetic genes (e.g. IDH1 R132H).

DNMT inhibitors, such as the cytosine analogs 5-azacytidine and its close derivative decitabine, are hypomethylating agents routinely used in clinical settings. 5-azacytidine is recommended as the first-line treatment of high-risk myelodysplastic syndromes (MDS) [49, 126]. Decitabine is also used in the treatment of MDS [37]; it has recently been approved for treatment of acute myeloid leukemia (AML). HDAC inhibitors form another group of epigenetic drugs with clinical approval. Examples for such include vorinostat [86] and romidepsin; the latter is successfully administered to patients with refractory or relapsed peripheral T-cell lymphoma [31]. Although the clinical efficacy of these compounds is already established, the exact mechanisms by which anti-tumor responses are triggered remains unclear. For example, inhibition of DNA methyltransferases leads to passive demethylation of the genome in treated cells (upon cell division), however, the link between this event and reduced fitness of tumor cells remains elusive. For HDAC treatments, the molecular targets of inhibition are less clearly defined than for the DNMT inhibitors, due to the many subtypes of potentially targeted HDAC enzymes.

An interesting novel approach is the development of an inhibitor for AGI-5198 (a mutant form of the IDH1 gene), which selectively blocks the activity of the protein and leads to growth suppression of cultured cells in soft agar and of mouse xenografts. AGI-5198 treatment does not change the global DNA methylation patterns, however, dimethylation and trimethylation marks of H3K9 are affected, leading to changes in the expression of

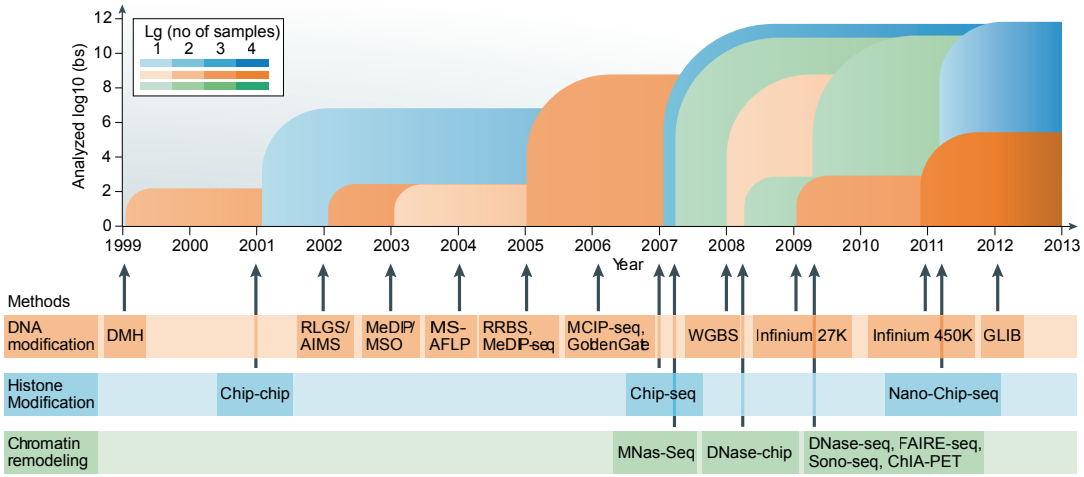


Figure 1.2: Examples of techniques developed for the profiling of DNA modifications, histone modifications and chromatin remodeling. The timeline indicates the year in which a technique was used for the first time in a cancer study. The vertical axis denotes the throughput of an assay (base pairs per sample), and the color intensity indicates the total number of cancer samples and cell lines interrogated by the technique.

genes involved in astroglial differentiation [93].

It is important to note that epigenetic drugs hold a strong potential, but they alone cannot provide a cure for cancer. To date, clinical studies in lung cancer involving only hypomethylating or HDAC inhibition agents show disappointing results [74].

1.2.2 Interplay of mechanisms

Many other lines of evidence lead to the conclusion that focusing on a single epigenetic mark provides very limited understanding and power for clinical application. For example, the "cross-talk" between DNA methylation and histone acetylation, both driven at least partially by environmental stimuli, is likely involved in the process of gene transcription and aberrant gene silencing in tumors [115]. Iorio et al. describe intricate connections between microRNAs, transcription factors, DNA methylation and histone modifications [62]. Shen and Laird provide an excellent overview on the cooperative workings of genetics and epigenetics in cancer, accompanied by many examples of known genetic alterations by epigenetic regulators in different tumor types [101]. Although the focus of this thesis is predominantly on methylation data, the studies presented here should be viewed in the larger context of the symbiotic relationship between genetic and epigenetic mechanisms.

1.3 Technologies for quantifying DNA methylation

Recent advances in microarray and sequencing technologies make the genome-wide profiling of DNA methylation and histone modifications feasible, even in cohorts that contain hundreds or thousands of samples [69]. Figure 1.2 shows the time of introduction and the

rapid development of new methods in the past two decades. Here, the focus is on assays used in the studies described in this work. A few other promising and well established technologies for the quantification of DNA methylation are briefly mentioned. A recent review by Plass et al. provides a comprehensive overview and comparison of all epigenetic assays presented in Figure 1.2 [90].

1.3.1 Affinity enrichment

Affinity enrichment of methylated fragments using antibodies specific for 5mC or using methyl-binding proteins with affinity for methylated native genomic DNA have been used as powerful tools for comprehensive profiling of DNA methylation in the human genome. Affinity purification of methylated DNA was first achieved using the methyl-binding protein MeCP2 [35]. The general technique is referred to as MeDIP [119], mDIP [67] or mCIP [129] and consists of two stages. First, methylated regions are enriched by immunoprecipitation of denatured genomic DNA with an antibody specific for methylated cytosine. As a second step, the captured DNA is hybridized to a microarray. In recent years, the hybridization step is usually replaced by second generation sequencing. In this case, the name of the protocol gains the suffix "-seq", e.g. MeDIP-seq, and these techniques are analyzed in this work.

One improvement over the MeDIP-seq technology is the MethylCap (Methylation Capture) assay developed by Brinkman et al. [24]. The approach consists of capture of methylated DNA by the methyl-binding protein domain (MBD) of MeCP2, and subsequent next-generation sequencing of eluted DNA. As a first step, the isolated genomic material is fragmented using sonication to an average length of 300 base pairs. The DNA fragments are then captured by a GST-MDB fusion protein and paramagnetic beads in a low salt concentation. After removal of the supernatant (the flow-through), NaCl gradient is used to wash and eluate genomic fragments from the immobilized GST-MBD. Retained DNA fragments are sequenced in up to three consecutive steps with increasing salt concentrations.

1.3.2 Quantitative sequencing assays

Treatment of denatured DNA with sodium bisulfite rapidly induces deamination of unmethylated cytosine bases, but has a very weak effect on the methylated residues. This discovery enabled the design of protocols that provide unprecedented single-base resolution of the methylation state in individual DNA strands [53]. Here, we briefly mention the modern bisulfite sequencing assays for genome-wide methylation measurement. For financial reasons, most of these techniques are applied on comparatively small sample sets, however, their high resolution and reproducibility ensures their ever wider application in the near future.

Whole-genome bisulfite sequencing (WGBS) [73] and tagmentation-based whole-genome bisulfite sequencing (TWGBS) [2] are used to measure the methylation level of almost every CpG in the genome. Single-base DNA methylomes are obtained by treating genomic DNA with sodium bisulfite to convert cytosine, but not methylcytosine, to uracil, and subsequent

high-throughput sequencing. The tagmentation-based approach uses adapted transposase-based in vitro shotgun library construction (called *tagmentation*) for whole-genome bisulfite sequencing, which allows the protocol to be applied to very limited amounts of starting material: 30 ng of DNA, compared to 5 μg for WGBS.

The large-scale random approach termed *reduced representation bisulfite sequencing* (RRBS) allows for quantifying the methylation of a substantial subset of CpG sites in the genome [78]. The method is based on size selection of fragments digested by a methylation-insensitive enzyme to generate a 'reduced representation' of the genome of a strain, tissue or cell type. Restriction fragments, 500 to 600 base pairs in length, are equipped with adapters and treated with bisulfite, PCR amplified, cloned and sequenced.

1.3.3 Microarray-based assays

Illumina adapted the genotyping assays implemented on a BeadArray platform to measure methylation state. Similarly to the methods described above, DNA is first bisulfite treated. The converted molecules are used in a whole-genome amplification reaction, before being enzymatically fragmented, precipitated and re-suspended in hybridization buffer. The resulting DNA fragments are hybridized on a genotyping microarray specifically designed for C/T polymorphism. After hybridization, the array is "[...] processed through a primer extension and an immunohistochemistry staining protocol to allow detection of a single-base extension reaction" [16, 17].

The GoldenGate DNA methylation assay interrogates the methylation status of 1,536 CpGs located in the promoters of 808 selected genes [17]. This panel consists of genes involved in various processes, including, among others, imprinting, signaling cascades, DNA repair, differentiation, cell cyle and apoptosis. It includes many known oncogenes and tumor supressors, making the assay particularly useful in studies involving cancer predisposition, progression and metastasis. Two pairs of probes are dedicated to each targeted CpG: an allele-specific oligonucleotide (ASO) and locus-specific oligonucleotide (LSO) probe pair for the methylated state of the cytosine base and a corresponding ASO-LSO pair for the unmethylated state. This probe design is known as Infinium *type I*. A notable number of probes in the GoldenGate assay overlap with single-nucleotide polymorphisms (SNPs), which may interfere with DNA methylation analyses [25].

The Infinium HumanMethylation27 BeadChip (also referred to as Infinium 27k) is a significant improvement over the GoldenGate assay in terms of genomic coverage [16]. This array contains over 27,578 pairs of probes targeting CpG dinucleotides in the proximity of the transcription start sites of 14,475 genes and in 110 miRNA promoters. In contrast to GoldenGate, Infinium 27k includes a probe pair for only one strand of each targeted CpG dinucleotide.

Infinium HumanMethylation450 BeadChip Kit (Infinium 450k) is the third generation of Illumina's genome-wide DNA methylation BeadChip. It facilitates high-throughput methylation profiling of 482,421 CpG sites and 3,091 cytosines in a non-CpG context [15]. This

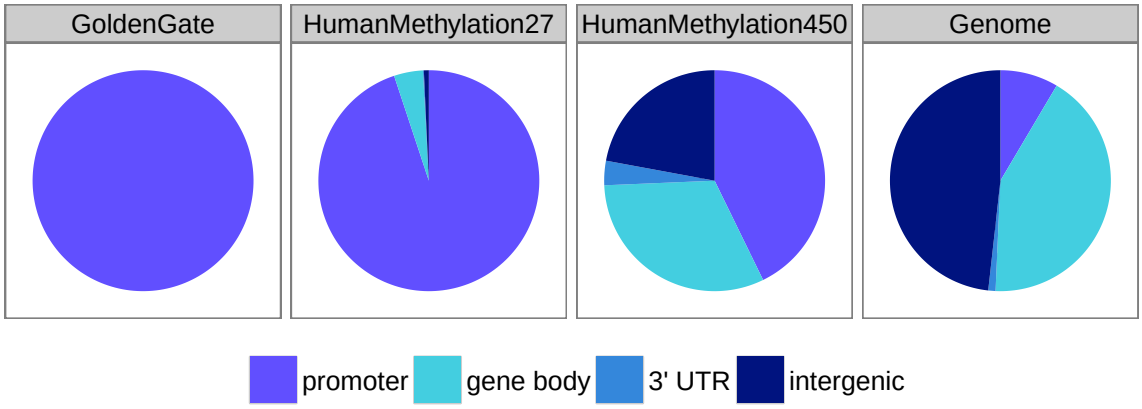


Figure 1.3: Distributions of CpG types covered by Infinium microarrays. The interrogated CpGs are categorized based on the genomic region into which they fall. Genes and 3' UTRs are defined based on RefSeq transcripts. Promoters are the regions spanning 2 Kb upstream and 1 Kb downstream of the transcription start sites of a transcript.

array introduces a new probe design – *type II* – which uses only one probe per locus and is particularly suited to regions of low CpG density. The methylation state of the targeted CpG dinucleotide is determined via single base extension with a labeled nucleotide using red and green color channel for unmethylated and methylated cytosine, respectively. The array also contains probes dedicated to measuring the efficiency of the chemical reactions at different steps of the protocol – bisufite conversion, amplification, hybridization and others. One slide contains 12 microarrays, allowing the interrogation of up to 12 samples in parallel.

Figure 1.3 shows the comparative distributions of targeted CpG dinucleotides by the microarray platforms described above. The promoter-biased coverage of the microarrays inevitably influences the questions addressed in the methylation studies that rely on these assays.

1.4 Statistical methods for the analysis of epigenetic data

The diversity of assays for measuring DNA methylation is a useful toolbox for research on epigenetic mechanisms, but this also presents unique bioinformatic challenges in terms of quality control, data visualization and statistical analyses [21]. This section briefly describes the major facets in the analysis of enrichment- and microarray-based methylation data, and then introduces the machine learning repertoire used in this thesis. More detailed discussions on quality control and data visualization are presented in Chapters 3 and 4, respectively.

1.4.1 Analysis of enrichment-based methods

Sequencing the selected DNA fragments and aligning the reads to a reference genome produces a global map showing differing densities of fragments in different genomic regions. The exact number of fragments found at a specific locus depends on the degree of methylation of this region in the studied cells, but also on the CpG density of the regions of interest, the amount and quality of the input DNA material, the configuration of the sequencing machine, and a host of other factors. For this reason, the count of overlapping fragments alone cannot be used as a reliable measure for the degree of methylation at a given locus. Chapter 3 introduces some commonly applied normalization techniques, as well as approaches for inferring methylation degree as a percentage.

1.4.2 Analysis of DNA methylation microarrays

Like any other microarray technology, the first step in extracting the methylation data is image processing of the array. This is performed by vendor provided software and quantifies the intensities for methylated and unmethylated signals in every interrogated locus. The absolute methylation value in Illumina's microarrays is referred to as β value and is calculated by the formula:

$$\beta = \frac{\max(I_M, 0)}{\max(I_M, 0) + \max(I_U, 0) + 100} \in [0, 1)$$

where I_M and I_U are the signal intensities of the methylated and unmethylated probes (or color channels in type II probes), respectively. A β value close to 0 indicates lack of methylation, whereas β values near 1 stand for full methylation. Another representation is the M values:

$$M = \log_2 \frac{\max(I_M, 0) + 1}{\max(I_U, 0) + 1}$$

Note that M is a continuous variable that can in theory take on any real value. It is not directly interpretable in terms of a methylation percentage, however, it shows more statistical power when identifying differentially methylated loci [42, 131].

In contrast to microarray expression studies, genome-wide DNA methylation measurements are relatively recent, and there is no consensus on statistical methodologies to be applied when working with this type of data. This is partially because of the non-normal distributions of methylation values obtained from microarrays. The Infinium 450k presents an additional challenge for normalization, due to the presence of two probe types, each one with a distinct bias [39]. This issue is discussed in more details in Chapter 3. The following section focuses on working with normalized methylation values. It briefly introduces several statistical methods and comments on their applicability for DNA methylation data.

1.4.3 Machine learning methods

Diagnosis and prognosis are examples of prediction problems, in which an unknown *outcome* (e.g. therapy resistance) needs to be predicted based on measured properties of an object, referred to as *features*. A mathematical model constructed to predict an outcome

is a *prediction model*. The general branch of artificial intelligence that concerns learning and prediction is known as *machine learning* or *statistical learning*. A more precise definition is that prediction models target the problem of *supervised learning*. Given a set of observations in the feature space with known outcomes, a model's parameters are adjusted such that the predictions are close to the true outcomes. This process is referred to as *training*. Later, the trained model can be used for predicting the outcomes of previously unseen observations.

In epigenomic studies, the features are usually numeric in their nature; examples for such include degree of methylation or of a histone modification. When the outcome is also quantitative, e.g. age, the prediction is a *regression*. On the other hand, if the outcome is categorical, e.g. gender, we speak of models for *classification*, in short – classifiers.

A large variety of methods for regression and classification have been developed in the past decades, and the machine learning field continues to expand at an impressive rate. Below, we focus on the machine learning techniques adapted and used in later chapters, and provide a brief introduction to each of them. Much more detailed explanations, including comparisons, discussions on the statistical power, applicability and interpretation are available in [57], as well as in the references given in each section.

Linear models

Given a vector of input features $X^T = (X_1, \ldots, X_p)$, a linear model predicts the outcome Y via the formula:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

One measure for the performance of a regressor is the residual sum of squares (RSS), defined as:

$$RSS(\hat{Y}, Y) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

where $y_1, \ldots, y_N$ are the true outcomes of N observations, and $\hat{y}_1, \ldots, \hat{y}_N$ are the predicted values for the outcome. If $X = (x_{ij})$ is an $N \times p$ matrix storing the feature values of N data points, and $Y = (y_1, \ldots, y_N)$ are their corresponding outcomes, the coefficients of a linear model that minimizes the RSS error metric are defined as:

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\}$$

The value of $\hat{\beta}$ can be analytically derived which guarantees that the training process is fast and efficient.

Lasso method for linear regression

The *lasso* is a coefficient shrinkage method applied on linear models. Using the notation introduced above, we can define the lasso estimate by

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \|\beta_j\| \right\}$$

The formula above is the Lagrangian form of the estimate, and λ is the penalty parameter, that is, the weight of the linear constraint on coefficient values. This constraint is also referred to as L_1-regularization because it effectively places a limit on the L_1 norm of the model coefficients. Computing the lasso solution is a non-linear optimization problem but efficient algorithms exist that calculate simultaneously the entire path of solutions for all applicable values of λ. Sufficiently large values of the parameter cause some of the model coefficients to be set to zero, effectively removing the corresponding features from the model. This property of the lasso makes it a valuable method for feature selection which often leads to noise reduction and therefore more accurate models. Moreover, the resulting models are better interpretable as there are less features on which the outcome depends. In the signal processing literature, this method is known as *basis pursuit* [28].

Lasso linear regression models are applied in Chapter 4 to predict patient's age based on methylation in blood.

L_1-regularized logistic regression for classification

Logistic regression is an approach used for classification tasks. Given K classes, the probability of every class $k \in K$ at a point x is modeled by a sigmoid function:

$$\Pr(G = k|X = x) \;=\; \frac{\exp(\beta_{k0} + x^T\beta_k)}{\sum_{\ell=1}^{K} \exp(\beta_{\ell 0} + x^T\beta_\ell)}$$

Similar to the lasso method, the L_1 shrinkage can be applied to logistic regression. In this case, provided with outcomes of $g_1, \ldots, g_N$ (where $g_i \in \{1, \ldots, K\}$) at N points in the feature space, training an L_1-regularized logistic regression (RLR) model amounts to identifying the coefficients that maximize the following log-likelihood:

$$\max_{\{\beta_{0k}, \beta_k\}_1^K} \left[\sum_{i=1}^{N} \log \Pr(g_i|x_i) \;-\; \frac{\lambda}{2} \sum_{k=1}^{K} \|\beta_j\| \right]$$

An application of RLR using methylation data is presented in Chapter 7, where a model is trained for predicting primary origin of a metastatic sample.

Elastic nets for classification

The elastic net method was designed for (and shown to perform well in) genomic applications, where the number of features greatly exceeds the number of analyzed samples [52]. In these scenarios, L_1-regularized classifiers tend to show good performance by using only some of the most informative features. However, this penalty leads to very unstable solutions in the presense of multiple correlated and informative features. The "choice" which feature is to be included in the model becomes strongly dependent on the training set. In order to overcome this limitation, an elastic net combines the L_1 penalty regularization

and L_2 (ridge) penalty, with a mix parameter α that determines their relative weights. Thus, the regularizer has the form[2]:

$$\lambda \sum_{j=1}^{p} \left(\alpha \left| \beta_j \right| + \frac{1-\alpha}{2} \beta_j^2 \right)$$

Elastic nets were successfully applied in a variety of scenarios, including, among others, studies in the fields of genetic epidemiology [56, 11] and mass spectrometry [128, 108]. Zhuang et al. [131] compared the performance of common feature selection and classification methods on Infinium 27 datasets, and showed that elastic nets show very good performance for classification, along with support vector machines, introduced below.

Support vector machines for regression and classification

The support vector machines (SVMs) are linear classifiers based on the concept of an optimal separating hyperplane between two classes. When the two categories are separable, optimality is defined as maximizing the margin between the decision boundary and the closest data point(s). In the non-separable case, at least one data point lies on the wrong side of the margin for any choice of separating hyperplane and margin width M. Therefore, the linear boundary is chosen such that it maximizes the margin, subject to a limitation on the total distance of points lying on the wrong side.

The optimization problem of a linear SVM is given in the inequality below.

$$y_i \left(x_i^T \beta + \beta_0 \right) \geq M \left(1 - \zeta_i \right)$$

subject to

$$\forall i : \zeta_i \geq 0, \sum \zeta_i \leq constant$$

In the formula above, the value ζ_i quantifies each misclassified point. The value can be thought of as the degree of misclassification, measured in margin units. Therefore, limiting the sum $\sum \zeta_i$ translates into placing a bound on the total amount by which predictions fall on the wrong side of the decision boundary.

The beauty of support vector classifiers lies in their property to efficiently separate classes in a transformed feature space without explicitly defining the transformation itself [23]. This is achieved by using a kernel – a symmetric positive semi-definite function that operates on two data points in feature coordinates and its outcome is equivalent to an inner product of its transformed operands. The transformed space is often of much higher dimensionality, compared to the original feature space. In such a case, a separating hyperplane in the transformed space translates to a non-liner decision bondary in the feature space. This property, along with its good performance, makes SVM one of the widely used machine learning techniques in practice [64, 27].

SVMs are used in Chapters 4 and 7 for predicting age and tumor type, respectively.

[2]The provided equation is the regularization expression implemented in the R package *glmnet*.

1.5 Outline

This work outlines analysis steps in several projects studying the DNA methylation profiles of thousands of samples in total. Although every project is unique in its combination of data source, heterogeneity, quality, time frame and goals, they all follow similar workflows. First, Chapter 2 introduces these studies, as well as the datasets which are referred to in subsequent chapters. The role of this chapter is also to reveal the common context in the analyses presented later. Starting from Chapter 3, the structure of this thesis mimics the steps performed in each of these analyses, whereby one chapter is dedicated to an important stage (or milestone) in the studies. Every chapter is designed to highlight the similarities between the different studies and also to present distinct techniques that target methylation data from a specific platform.

Not surprisingly, the sequence of chapters partially overlaps with the workflow presented in a recent review on the analysis and interpretation of methylation data [20]. Chapter 3 addresses the important question of quality control beyond the assay-specific procedures performed in the laboratory. Chapter 4 discusses the observed variability in methylation and shows simple techniques applied to reduce the complexity of the data and provide a global overview of a cohort. Chapters 5 and 6 change the direction to data mining and present approaches that guide the search for differentially methylated and phenotype-associated genomic sites or loci. Chapter 7 illustrates the predictive power of DNA methylation in the context of tumor diagnosis. Finally, Chapter 8 reiterates the major findings and presents interesting directions for future research.

Linguistic style

For reasons of consistency, this thesis follows the first person plural form ("we"). Unless otherwise noted, algorithm design and all analyses are performed by the author himself. General approach to a problem, parameter adjustments and interpretation of results are always the outcome of discussions with collaborators.

2 Projects and datasets

This thesis presents details from several separate studies on DNA methylation, each operating on a different set of DNA samples. The analyzed datasets are examined in the following chapters and are referenced in multiple sections. For ease of reading, these datasets are referred to by using short phrases: the GoldenGate, the colon cancer, the lung cancer, and the TCGA dataset. Each of them is represented by a matrix that contains CpG sites (or probes) as rows, and interrogated samples as columns. The methylation studies and these sets are briefly introduced below. Their usage in this thesis is summarized in Figure 2.2.

2.1 GoldenGate

In a collaboration with Agustín Fernández, Jose Ignacio Martin-Subero and others, we performed extensive analysis on the methylation profiles of a large collection of DNA samples from normal tissues, primary and metastatic tumors, cell lines and samples from patients with different non-cancerous disorders [50]. The methylation was interrogated using Illumina's GoldenGate assay. The filtered dataset containing all 1,628 analyzed samples is available at the NCBI Gene Expression Omnibus (ncbi.nlm.nih.gov/geo), under accession number GSE28094. We refer to this dataset as the *GoldenGate dataset* later in this thesis.

This large and heterogeneous collection of samples enabled us to study the methylation patterns in different contexts – methylation in normal (healthy) tissues, tumor-specific methylation, cancer cell lines and non-cancerous diseases. In the following paragraph, we briefly outline some of the findings.

The dataset contains 424 samples from normal tissues, and we observed strikingly consistent methylation patterns across individuals in almost all primary tissues. Indeed, if we group the samples by tissue type, the between-group distances are markedly larger than the within group distances (see Figure 2.1). Inter-invidual methylation differences at CGI-associated loci were significantly lower than the differences at CpGs outside islands. We also studied the impact of aging on the methylation profiles of leukocytes and colon tissue. We identified sets of probes that gain and lose methylation with age and confirmed previous findings on this topic. Chapter 4 discusses inter-individual and age-specific methylation in details.

The largest fraction of samples in this dataset is cancer-related, including 855 primary tumors, 50 metastatic samples, 25 premalignant lesions, 82 cancer cell lines and 42 cancers of unknown primary origin. Interestingly, a tumor-type methylation signature is present even when the tissue-type specific probes are removed. In general, tumors are characterized by higher methylation variability. Tumor progression is accompanied by gain of methyla-

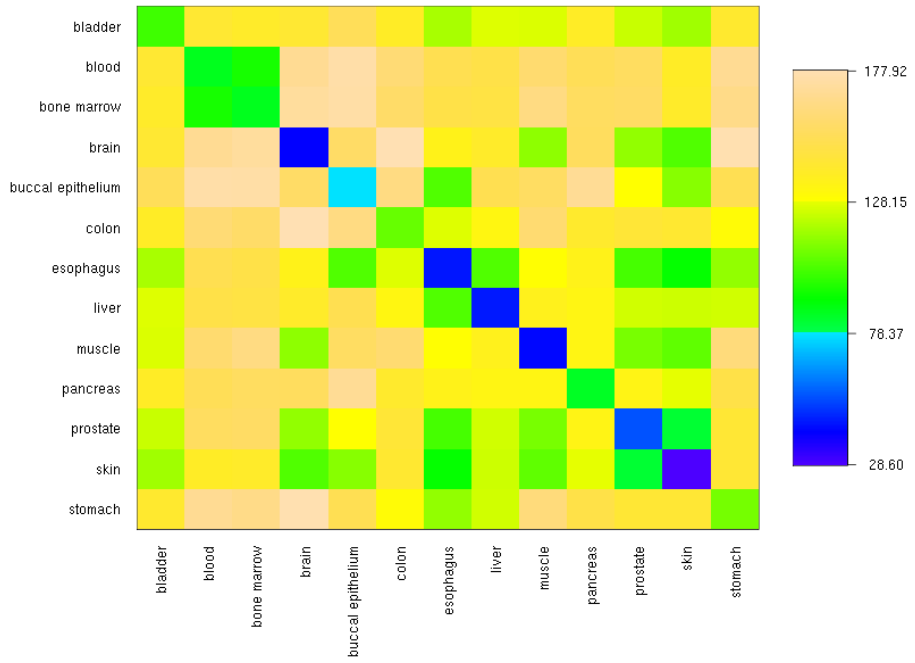


Figure 2.1: Heatmap showing mean within- and between-group Manhattan distances the normal tissue samples from the GoldenGate dataset.

tion within CGI-associated promoters and loss of methylation outside CGIs. Examples for such observations are presented in Chapter 4.

Overall, disruption of DNA methylation patters appears to be a global phenomenon not only in cancer, but also in other diseases in human. The techniques we applied to define differential methylation, quantify differences, identify and characterize loci with disease-specific methylation are very similar to the approach taken in our study on tumor-specific methylation. Since this thesis focuses on specific bioinformatic approaches, we do not cover the analysis of non-cancerous diseases in details.

2.2 Colon cancer

The *colon cancer* dataset is a collection of 50 MethylCap-seq samples comprising 25 tumor-normal pairs from colorectal cancer patients [104]. One sample pair was ignored due to potential mislabeling, details are provided in Chapter 3. Clinical annotation of the analyzed patients is available in Supplementary Table S1. This study was a collaboration with Femke Simmer, Arjen Brinkman and others. We aimed at and identified potential biomarkers for colorectal cancer diagnosis, and made important observations related to the genome-wide methylome of colon cancer.

We identified 184 frequently differentially methylated regions, among them were novel hypermethylated gene promoters, some of which we validated using pyrosequencing [1].

[1] Selection of fragments for validation and the validation procedure itself was performed by Femke Simmer

2.3 Lung cancer

The *lung cancer* dataset comprises over 450 lung cancer samples and 25 normal lung controls, all interrogated by Infinium 450k [96]. This study is the result of a collaboration with Fabian Müller and several members of Manel Esteller's lab at the Bellvitge Biomedical Research Institute in Spain. We attempted to stratify non-small cell lung carcinoma into methylation-specific subtypes; we also identified and validated a prognostic panel on relapse free survival, consisting of five gene promoters. The dataset with normalized methylation β values is available at the NCBI GENE Expression Omnibus under accession number GSE39279.

This thesis includes analyses that are not presented (and nicely complement the findings) in the publication. For example, Chapter 7 presents an approach to the identification of methylome-specific tumor subtypes and studies the bimodality properties of the interrogated CpGs.

2.4 RnBeads

Almost every section in this thesis is backed by analysis performed using *RnBeads* – an R package for comprehensive analysis of DNA methylation data [10]. Being the initiator of this software tool and one of its core developers, the author incorporated in its design many of the aspects discussed in this thesis.

2.5 External datasets

In addition to the samples analyzed in each study, we used external datasets to validate our findings, enrich the annotation of identified differentially methylated regions, or to apply machine learning techniques and test their potential for clinical relevance. The largest datasets from public repositories referred to in this thesis are listed below.

2.5.1 TCGA

The Cancer Genome Atlas (cancergenome.nih.gov) is a large consortium that was initiated with the aim to systematically catalogue all genetic and epigenetic aberrations in dozens of tumor types [30]. The molecular profiling performed at the participating institutions includes measurements of copy number aberrations using SNP- and CGH-arrays; DNA methylation using Illumina microarrays; exome, mRNA, protein and miRNA transcription using microarray and sequencing technologies, and other forms molecular characterization. As of January 2014, TCGA collects samples and characterizes 29 tumor types, the data for 16 of them can be used freely. The table below lists these tumor types. The study codes shown below are used later in this thesis.

Code	Tumor Type
LAML	Acute Myeloid Leukemia
BRCA	Breast cancer
KICH	Chromophobe renal cell carcinoma
KIRC	Clear cell kidney carcinoma
COAD	Colon and rectal adenocarcinoma
SKCM	Cutaneous melanoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
OV	Ovarian serous cystadenocarcinoma
THCA	Papillary thyroid carcinoma
READ	Rectal adenocarcinoma
STAD	Stomach adenocarcinoma
BLCA	Urothelial bladder cancer
UCEC	Uterine corpus endometrial carcinoma

2.5.2 ENCODE

The Encyclopedia of DNA Elements (ENCODE) consortium is an international collaboration of research groups with the goal to build a comprehensive list of functional elements in the human genome. These include elements that act at the protein and RNA levels, and "regulatory elements that control cells and circumstances in which a gene is active" (encodeproject.org). The ENCODE consortium focuses on cell lines only. It initiates successful collaborations between investigators with diverse backgrounds and expertise in the generation and analysis of data, which has resulted in revolutionary discoveries on transcription regulation [40], DNA accessibility [111], and chromatin states [58], among others. In this thesis, we use ENCODE in Chapter 6, when we test for association between differentially methylated regions and histone marks.

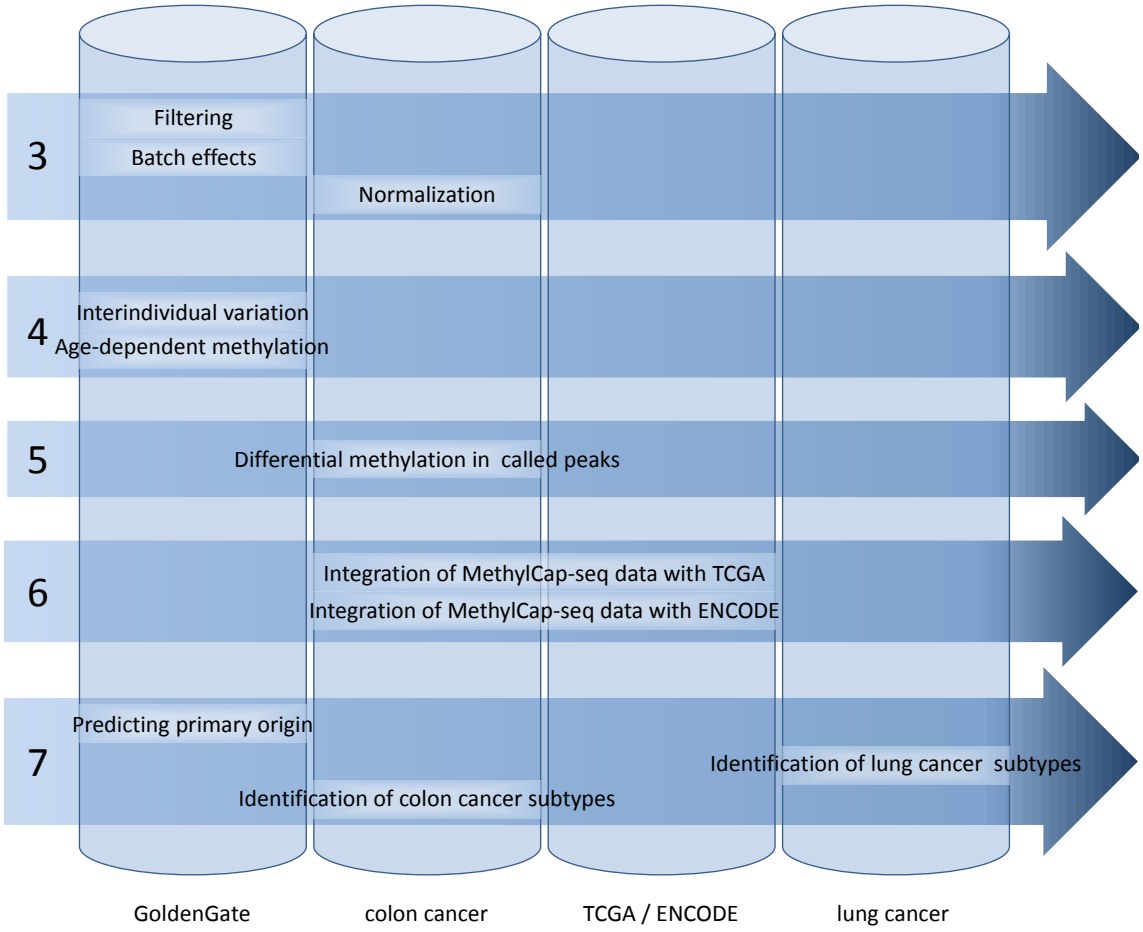


Figure 2.2: Usage of datasets in the chapters of this thesis. The chapters are represented by horizontal arrows.

3 Quality control and normalization of methylation data

This chapter introduces part of the functionality of RnBeads in its first two sections and accompanies the descriptions with examples from microarray-based DNA methylation data. Work on quality control of methylation data started in collaborations within the CANCERDIP consortium [50], and was further extended during the development of RnBeads [10]. The Chapter begins by listing available algorithms for normalization and follow with a description of Greedycut – our approach to filtering low quality CpG sites and samples. We then briefly introduce the aspect of confounding factors, an integral part of every epigenome-wide association study. Section 3.4 is not related to RnBeads. It addresses enrichment-based methylation data, and compares several approaches to normalization. These are the initial steps in the analysis of the colon cancer study [104] – another collaboration within CANCERDIP. The last section presents a brief summary.

3.1 Probe types and DNA regions

Like any other microarray technology, the Infinium methylation data is prone to multiple sources of bias and requires careful quality control and normalization. As mentioned in Chapter 1, Infinium 450k includes control probes that are used as indicators of many aspects of the reactions. The signal intensities of these probes aid the processes of quality control by quantifying the efficiency of bisulfite conversion, staining, hybridization, polymorphism, base extension and target removal. Cross-hybridization in particular is potentially a major issue in Infinium arrays [29], and is a possible explanation for misleading results in published studies [19]. In addition to control probes, the Infinium 450k assay includes 65 SNP-specific probes that can be used to estimate genetic similarity and identify potential sample mislabeling.

The standalone application GenomeStudio, developed by Illumina, provides a simple algorithm for background correction and normalization, which is also implemented in the R packages methylumi [114] and minfi [75]. More sophisticated normalization methods that take into account the different chemistry of the two probe types, include the peak based correction [39], subset-quantile within array normalization [75], beta mixture quantile dilation [109] and others [89]. Published systematic evaluations of these techniques reach slightly different conclusions, but they agree on the need for advanced normalization algorithms to be applied before data exploration or targeted search for differentially methylated sites and regions [76, 89].

The control probe intensity signals can be studied using our R package RnBeads, along

with the values of the SNP-encoding probes. Furthermore, RnBeads supports all normalization methods described above, mostly through integration with existing R packages[1].

3.2 Filtering probes and samples

In addition to normalization and quality control through visual inspection, RnBeads includes a flexible approach to removing potentially biased or undesired sites and samples from the studied dataset. Filtering out uncertain measurements is the most common approach for quality control in microarray-based methylation studies [20]. In its filtering modules, RnBeads applies a configurable sequence of filtering steps, each targeting a specific quality metric or bias. Some examples for filtering rules include: (1) removing sites that overlap with known SNPs, (2) removing low quality sites and samples (the definition of quality is described below), (3) ignoring sex chromosomes, (4) removing sites with many missing values or (5) low methylation variance. Step 1 is implemented by a procedure suggested in a recent study on pediatric acute lymphoblastic leukemia [83]. The remainder of this section focuses on the notion of unreliable measurements and the RnBeads implementation of Step 2.

Every measurement in Illumina methylation microarrays is accompanied by a detection p-value that quantifies the reliability of the methylation signal. High p-values are indicative of bad measurements, for example, due to incomplete hybridizations. In the GoldenGate dataset, we examined two aspects of filtering out probes and samples based on the detection p-values – selecting a p-value threshold to define an unreliable β value measurement, and a cutoff (fraction of unreliable measurements) to define low quality probe or sample. We applied a threshold value of 0.01 as it leads to a clear distinction between reliable and unreliable β values, and a cutoff of 5%. Thus, we first removed all probes that contain detection p-values above 0.01 in 5% or more of the samples. As a second step, we removed all samples that contain detection p-values above 0.01 in 5% or more of their (remaining) probes. The result of this procedure is visually depicted in Figure 3.1.

The motivation behind the steps described above is to achieve a dataset with a required purity (maximum fractions of unreliable measurements per probe and per sample). Let us formally define the problem of filtering a dataset based on the discussed notion of reliability of its measurements.

Problem P3.1 Given an indicator matrix A:

$$\underset{m\times n}{A} = (a_{ij})\,|_{i=\{1,\ldots,m\},j=\{1,\ldots,n\}} \in \{0,1\} \tag{3.1}$$

where $a_{ij} = 1$ denotes an unreliable measurement of probe i in sample j, find a maximum induced submatrix $B_{m'\times n'} = (b_{ij})$ of A such that every row and every column of B have an impurity at most a threshold $t \in [0,1]$:

[1] Control probe intensity plots, as well as work on integrating most of the normalization methods into RnBeads has been performed by Pavlo Lutsik.

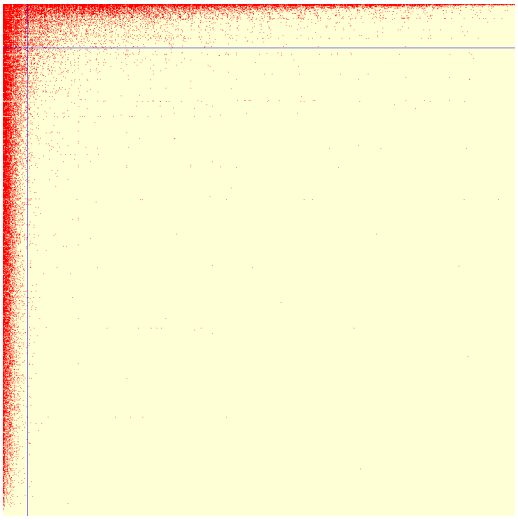


Figure 3.1: Color-coded representation of the matrix of all detection p-values in the GoldenGate dataset. Bad p-values (above the threshold of 0.01) are depicted as red points, all other measurements are yellow. Both probes (rows) and samples (columns) are listed in descending order with respect to the number of bad p-values they contain. The blue lines show the approximate sections of the matrix that are ignored after applying the selected cutoff of 5%.

$$\forall i = 1, \ldots, m' : \sum_{j=1}^{n'} b_{ij} \leq t$$

$$\forall j = 1, \ldots, n' : \sum_{i=1}^{m'} b_{ij} \leq t$$

It is important to note that there is not necessarily a unique submatrix B that is a solution to the above problem. Finding all solutions is a computationally intensive task. Moreover, additional criteria need to be defined in order to present one of the solutions as the filtered matrix to be operated on in downstream analysis. The case of potential multiple solutions is best illustrated through an example, as the one given in Figure 3.2. Before showing the computational complexity of the problem, we are going to reformulate it as a problem in graph theory. The indicator matrix $A_{m \times n}$ in Equation 3.1 can represent a bipartite graph $G = (U, W, E)$, where $U = \{u_1, \ldots, u_m\}$, $W = \{w_1, \ldots, w_n\}$, and $u_i w_j \in E \Leftrightarrow a_{ij} = 0$. We are then searching for a maximum induced subgraph $B = (U' \subseteq U, W' \subseteq W, E')$ of A, such that: $\forall u \in U' : degree(u) \geq t\,|W'|$ and $\forall w \in W' : degree(w) \geq t\,|U'|$ for a given $t \in [0, 1]$. A special case of this problem for $t = 1$ is well studied in the field of discrete mathematics. It concerns finding a maximum induced biclique in a bipartite graph, and is shown to be NP-complete [87]. Moreover, finding the number of maximum induced bicliques is a #P-complete problem [68]. Zhang et al. recently presented a fast algorithm for finding all maximum bicliques in a bipartite graph, however, it is applicable to graphs of sizes corresponding only up to a small fraction of the Infinium 450k probes [130].

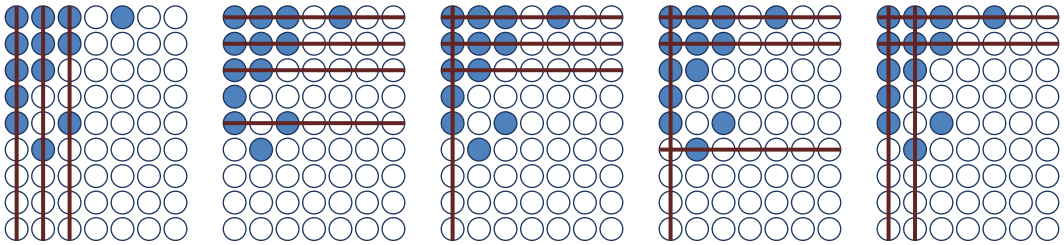


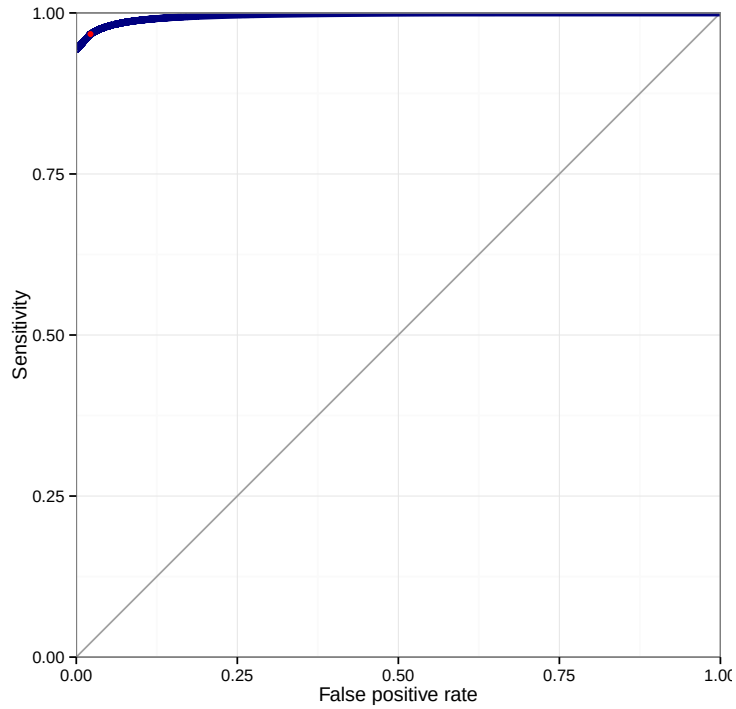
Figure 3.2: Maximum induced submatrices B, solutions to Problem P3.1 for an example indicator matrix A of size 9 rows $\times$ 7 columns. Values of 1 are represented by filled blue circles, whereas empty circles denote values of 0. The threshold t in the problem is set to 0.25. Induced submatrices are formed by removing rows and/or columns of A, as depicted by horizontal and vertical red lines, respectively.

For the reasons listed above, RnBeads does not search for a submatrix of pre-defined purity when filtering a dataset based on the reliability of its measurements. Instead, we devised an algorithm that is motivated by viewing the process of filtering probes and samples as a prediction problem, and minimizes an error metric[2]. Briefly, the algorithm iteratively removes the probe or sample in the given dataset that contains the largest fraction of unreliable measurements. The input of every step is a binary matrix storing the state of every measurement, and the result is a matrix with smaller dimensions (one row or one column less) than the input. The steps are performed until all remaining measurements are reliable, or the resulting matrix is of size zero. The full run of the algorithm produces a sequence of nested matrices, each storing the retained measurements after the corresponding step. Of note, the algorithm often executes over a hundred thousand steps before termination.

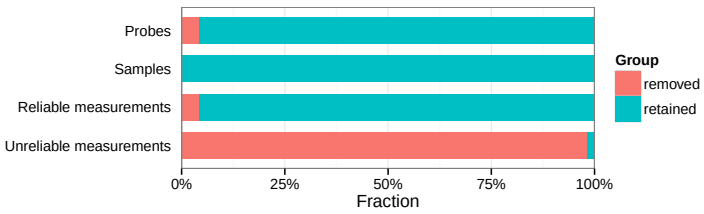
The outcome of every step can be put in a prediction framework – the retained measurements are predicted to be reliable, and the removed ones are assumed to be unreliable. We can therefore quantitatively define the efficiency of every step, for example, by calculating the type I and type II errors. Greedycut computes sensitivity and specificity at every executed step, and then selects the outcome of the first step that maximizes a given criterion on these metrics for prediction quality. By default, sensitivity and specificity are given equal weight, which amounts to selecting the point furthest away from the diagonal on Figure 3.3. In principle, applying any criterion within the framework described here produces a unique solution.

It is worth noting that Greedycut's usage is not restricted to microarray data only. The algorithm can be applied to any methylation dataset in which measurement reliability can be defined. In bisulfite sequencing data, for example, the coverage at every interrogated CpG site is often used as an indicator of reliability.

[2]Collaboration with Jing Cui, who should also be credited with naming the procedure Greedycut.



(a)



(b)

Figure 3.3: (a) Sensitivity and false positive rate calculated for every step of Greedycut in a dataset of 765 breast cancer samples, obtained from TCGA. The red circle marks the iteration that maximizes distance to the diagonal. (b) Summary of retained and removed probes and samples after applying Greedycut.

3.3 Batch effects

RnBeads uses both visual and statistical means to identify associations between the methylation landscape of the dataset and the provided sample annotation. As a first step, principal component analysis and multi-dimensional scaling are performed and the inter-sample distances are juxtaposed with the available annotation in an interactive manner through visual properties, such as point color and shape. A more analytical approach performs tests for association between a major principal component on one side and a sample annotation column on the other side. Significant results of these tests can be informative in two respects. First, they can spot signals in the space of methylation values that are informative of a sample phenotype. Second, they can identify sources of strong technical bias, such as, for example, processing date, slide, or technician responsible for sample preparation. Such undesired associations are referred to as batch effects and have potentially detrimental effect on the power of the study when they remain undetected.

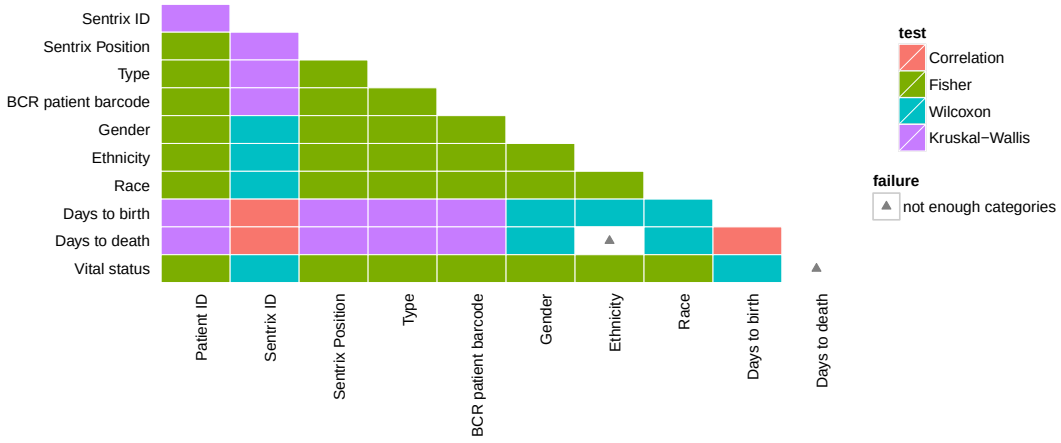
In addition, RnBeads checks for the presence of potential confounding factors by testing every pair of sample properties (columns in the annotation table) for a significant association. The tests used to calculate a p-value given two properties A and B depend on their data types:

- If both properties contain categorical data (e.g. tissue type and sample processing date), the test of choice is a two-sided Fisher's exact test.
- If both properties contain numerical data (e.g. amount of starting genomic material and age of individual), the correlation coefficient between the traits is computed. A p-value is estimated using permutation tests.
- If property A is categorical and property B contains numeric data, p-value for association is calculated by comparing the values of B for the different categories in A. The test of choice is a two-sided Wilcoxon rank sum test (when A defines two categories) or a Kruskal-Wallis one-way analysis of variance (when A separates the samples into three or more categories).

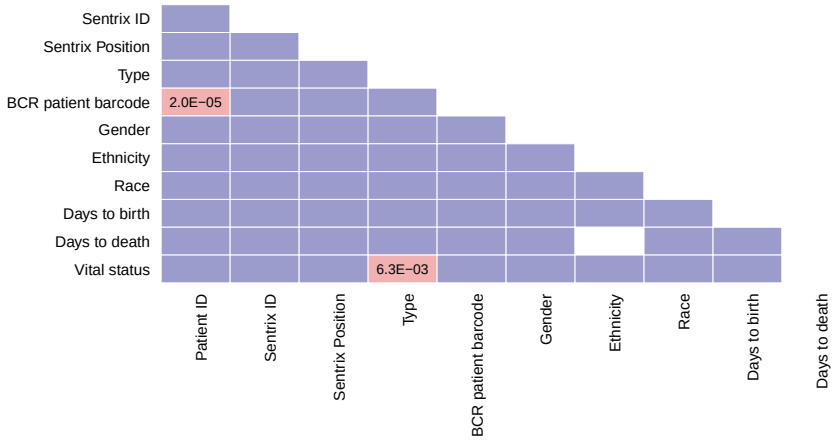
Figure 3.4 shows the results of the pairwise annotation tests, performed on the glioblastoma samples from the TCGA dataset. All results are visualized in a coherent report, providing systematic assessment of the associations present in the dataset.

It is important to note that the p-values calculated with RnBeads by the tests described above are not corrected for multiple testing. We made this decision due to the specific goal of this particular analysis, which is supposed to warn the users of confounders and batch effects that could potentially affect downstream analysis. Hence it is the more the conservative approach for RnBeads to omit multiple testing correction and to report all potential confounders that reach statistical significance in an uncorrected test. Also, this systematic analysis is designed to produce warnings for possible overlooked associations and it does not correct for batch effects[3].

[3] Confounders can be incorporated in the differential methylation module, implemented by Fabian Müller.



(a)



(b)

Figure 3.4: (a) Table of performed tests on pairs of traits in Infinium450k GBM samples from TCGA. Test names (correlation + permutation test, Fisher's exact test, Wilcoxon rank sum test and Kruskal-Wallis one-way analysis of variance) are color-coded according to the legend given on the right-hand side. (b) Table of resulting p-values from the performed tests on pairs of traits. Significant p-values (less than 0.01) are printed in pink boxes. Non-significant values are represented by blue boxes. White cells denote missing values.

3.4 Quality control and normalization of enrichment-based methylation data

This section focuses on methylation data obtained using a protocol for the enrichment of methylated DNA fragments followed by sequencing. There are two main strategies to undertake in the analysis of these data.

The first approach is to estimate methylation degrees using DNA methylation inference software. The MEDME [88] and BATMAN [41] tools, among others, provide this functionality. In general, methylation inference requires careful calibration of model parameters and/or significant computational resources. In addition, a resolution of β values comparable to the microarray-based approaches can be achieved only by deep coverage and high quality alignment. Once absolute methylation is estimated, the analysis could follow the steps for filtering and batch effect identification outlined in the previous section.

An alternative approach is to perform peak calling, that is, identify the genomic regions with significantly high number of fragments. The result of applying this procedure on a dataset is a matrix containing the number of overlapping fragments with every peak in every sample. For example, peak calling on the colon cancer dataset yielded 329,613 regions with a median length of approximately 2 Kb[4]. The following sections introduce approaches to normalizing this matrix and apply them on the colon cancer dataset.

3.4.1 Normalization of tag enrichment

Here, we introduce four approaches to defining a measure of peak enrichments (heights). All of them define scores that are based on the read count (also referred to as tag count) per peak. The tag count itself is considered to be the non-normalized quantification for peak height. There are two factors for normalization – peak width and inter-experiment variability. Figure 3.5 shows a schematic representation of reads, aligned to the human genome, which was used to calculate tag count. This figure also introduces the notations for tag count and total tag count used later.

We applied different approaches to normalizing with respect to peak width and inter-experiment variability separately and in concert. The resulting normalized scores – tag occupancy, scaled tag count, scaled tag occupancy and tag density – are introduced in the following sections.

Corrections for peak width

We applied two simple techniques to normalizing tag counts with respect to peak widths. The resulting scores are introduced below.

Tag occupancy

[4] Alignment, peak calling and merging of overlapping peaks was performed by Arjen Brinkman.

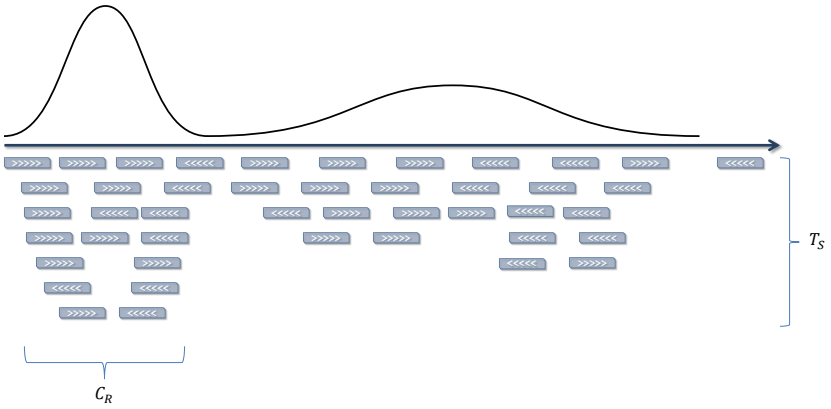


Figure 3.5: Schematic representation of a genomic region and all sequencing reads aligned to it. Two peaks were called in this region. The horizontal arrow depicts the DNA segment, and reads are represented by filled rectangles. Every read is extended to a fragment of length approximately four times the read length. The directionality of the reads is shown within the rectangles. C_R denotes the number of reads aligned to a peak, that is, the tag count. T_S quantifies the total number of aligned reads in an experiment. The curve above the DNA arrow interpolates the histogram of aligned reads and thus gives an indication of the tag counts per peak.

The measure *tag occupancy* is defined as the tag count normalized by the width of the peak, measured in units of fragment lengths. Thus, the tag occupancy of region R in sample S is defined as:

$$O_R(S) = C_R(S) \times L_M / L_R$$

where $C_R(S)$ is the absolute read count in R, L_M is the median read length, and L_R is the length of R. In the colon cancer study, for example, the median fragment length is 269 bp, which translates to occupancy of $0.269 \times C_R(S)/L_R$.

Tag density

The relative read count, or *tag density*, of a region R in sample S is defined as:

$$D_R(S) = C_R(S)/(T_S \times L_R)$$

where $C_R(S)$ is the absolute read count in R, T_S is the total number of reads in sample S, and L_R is the length of region R in kilobases.

As can be seen from the definition above, tag density is identical[5] to reads per kilobase per million mapped reads (RPKM). RPKM is a widely used measure, introduced by Mortazavi et al. for RNA-seq, which "facilitates transparent comparison of transcript levels both within and between samples" [81].

[5] up to a constant factor

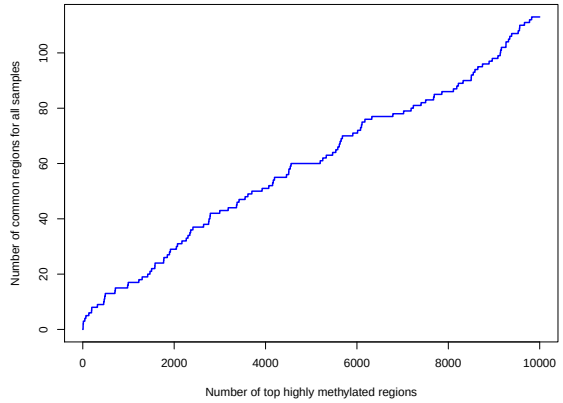


Figure 3.6: Number of common peaks across all samples among the top K methylated peaks per sample, as a function of K.

Corrections for experimental conditions

Including the total number of reads in the formula for tag density ensures that this score is normalized also for inter-experimental variability. Essentially, correction for experimental conditions involves the multiplication of a score by a sample-specific scaling factor. Tag density sets this factor to T_S^{-1}.

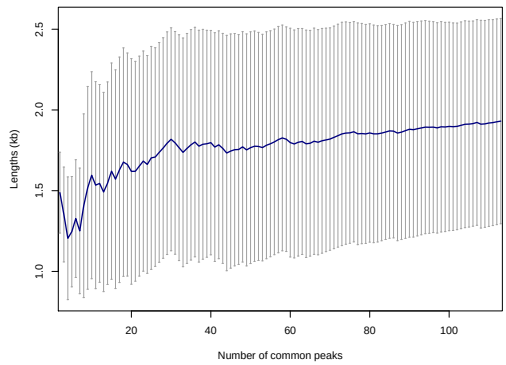
We applied another method for inter-sample normalization which calculates scaling factors based on fully methylated regions. Briefly, we first identify a set of peaks that are very densely methylated in all experiments. Next, we calculate sample-specific scaling factors based on the tag counts in these peaks. The scaling factors modify the scores in a way that minimizes the observed inter-sample differences. In the final step, we apply each scaling factor to the scores for all peaks in the respective experiment.

Common peaks

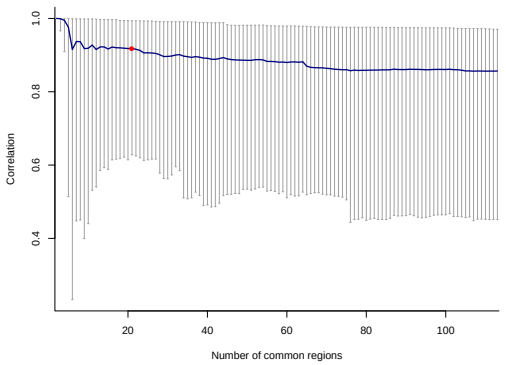
We first identified a set regions whose dense methylation is of vital importance to every somatic human cell. The procedure used to identify this set is described below.

We used tag occupancy as a measure of methylation degree and identified the top K most densely methylated peaks in every sample. We focused only on peaks of length up to ten fragments located on autosomes. The rationale behind the exclusion of long regions is that they are often the result of merging multiple overlapping peaks, and thus tend to have a very heterogenous read coverage. Sex chromosomes were ignored because our dataset consists of samples from both female and male patients. 203,818 regions were left for examination after applying these filtering criteria. We next counted the number of common peaks in all samples among the top K densely methylated peaks, for all values of K from 1 to 10,000. The results are shown in the Figure 3.6.

Our aim was to identify regions that are consistently fully methylated and exclude the ones that exhibit cell type-specific methylation. The set of common peaks among the top



(a) Distributions of region lengths for all tested sets of common peaks. Whiskers denote standard deviation.



(b) Ranges of pairwise sample correlations on the vectors of methylation of common peaks. Whiskers depict minimum and maximum correlations.

Figure 3.7: Statistics on common peaks. Blue line shows mean values. The set of common regions pointed to by the correlation-based suggestion is denoted by a red circle.

K methylated peaks is the estimation for these consistently methylated regions. It is important to achieve high sensitivity (specificity is of lesser concern), therefore, we selected a value for K after which the number of common peaks remains unchanged for the longest range. We use the term *k-based suggestion* for this estimation. The suggested value for K is 4,558 and the corresponding size of common peaks is 60.

The choice of K introduced above may seem somewhat arbitrary, especially since the shape of the curve in Figure 3.6 does not reflect the assumption of the applied heuristics. For this reason, we examined the properties of the common peaks in general, and the selected ones in particular. We next defined an alternative strategy for identifying regions with ubiquitous methylation (described below). First, we studied the distribution of region lengths of all nested sets of common peaks. As shown in Figure 3.7(a), there is no obvious bias within the set of common peaks in terms of region length.

In case the set of common peaks identifies consistently methylated regions, the pairwise correlations between the samples are expected to be close to 1. Indeed, the distributions of observed correlations in most tested sets of common peaks are tailed towards 1. In many cases, all pairwise sample correlations are above 0.5. The correlation ranges are shown in Figure 3.7(b).

This observation was used to generate another suggestion for the choice of number of common peaks. We selected the set of size at least 10 with the most consistently high correlations; in other words, the set with the highest value for minimum pairwise sample correlation. We use the term *correlation-based suggestion* for this estimation. The suggested value for K is 1,471 and the correponding size of common peaks is 21.

Scaling factors

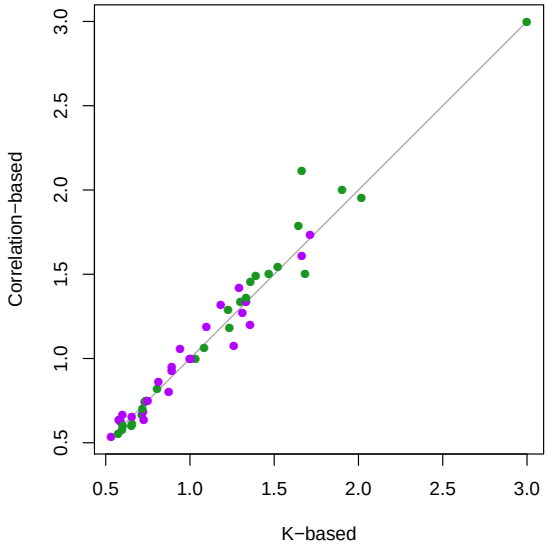


Figure 3.8: Agreement in scaling factors between suggestion strategies for a set of common regions. Healthy samples are represented by green circles, and colon cancer samples - by purple circles.

Given a set of common densely methylated regions, we computed sample-specific scaling as follows: First, we calculated the median tag count for each of the common peaks. Next, we examined the corresponding tag counts for these peaks at every sample individually and calculated the coefficients for every peak that rescale its count to the corresponding peak median. In the final step, we selected the median scaling coefficient for every sample as its scaling factor. The resulting factors are in the range 0.5 to 3 (see Figure 3.8) and are in a strong agreement with the range of values observed for total number of reads per sample. Moreover, the two suggestion strategies produce remarkably concordant sample specific scaling factors (Figure 3.8). We used the k-based suggestion to define the scores *scaled tag count* and *scaled tag occupancy* as follows:

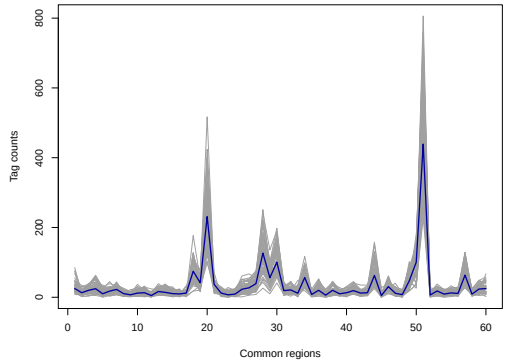
$$SC_R(S) = C_R(S) \times F_S$$
$$SO_R(S) = O_R(S) \times F_S$$

where $C_R(S)$ is the absolute read count in R, $O_R(S)$ is the occupancy of R, and F_S is the scaling factor for sample S.

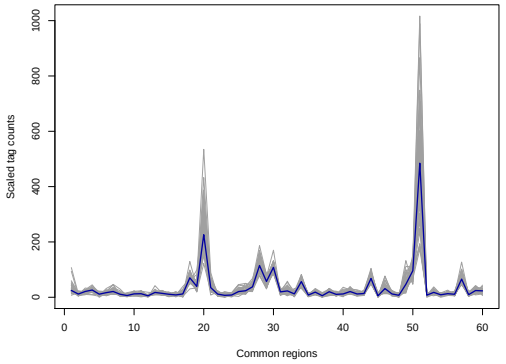
Score comparison

We compared the different scores by plotting the values at the common densely methylated peaks in all samples. These values are shown in Figure 3.9. Not surprisingly, the inter-sample agreement is best for the scaled scores at these regions. Note that scaled tag occupancy and tag density show similar behaviour, despite the fact that they operate in very different ranges.

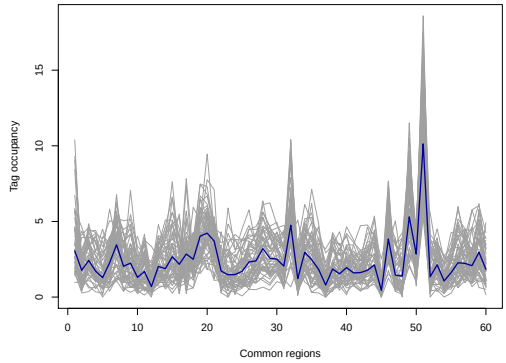
We also compared the distributions of pairwise correlation coefficients at the common methylated peaks (Figure 3.10). The very high correlations in tag counts visibly decrease



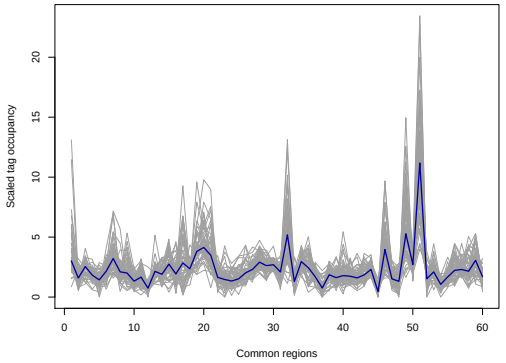
(a) Tag counts



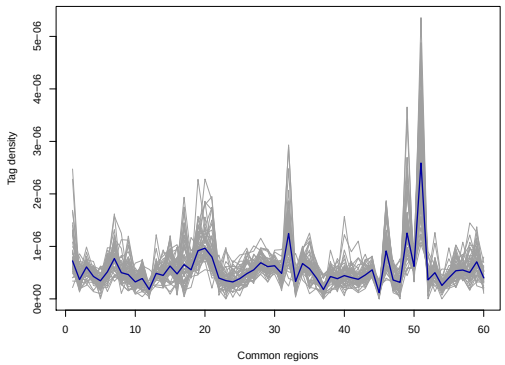
(b) Scaled tag counts



(c) Tag occupancy

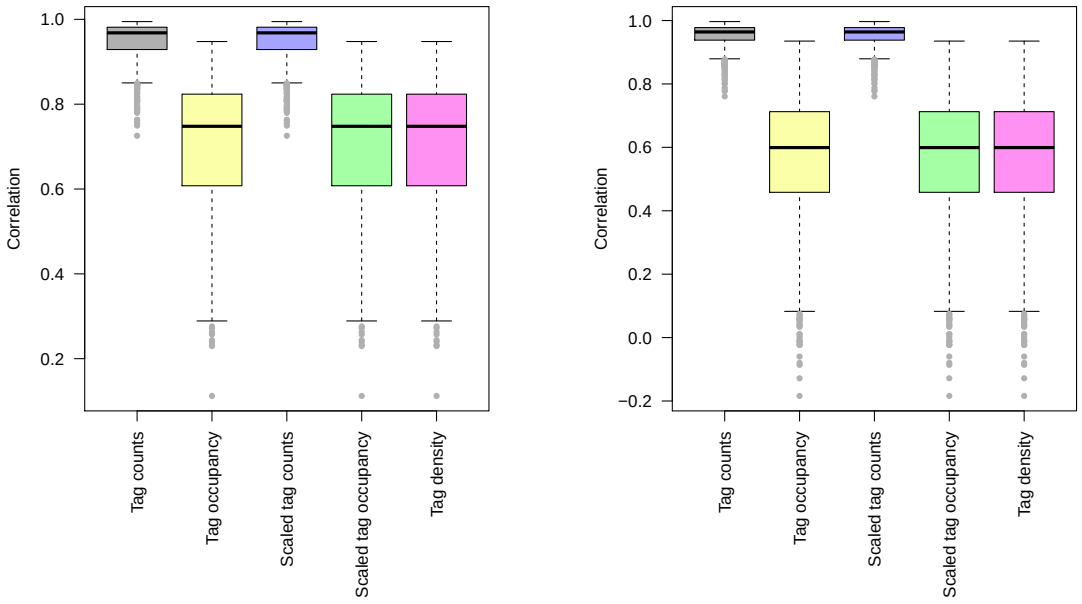


(d) Scaled tag occupancy



(e) Tag density

Figure 3.9: Line plot showing the score values at the common densely methylated peaks. The peaks are listed on the horizontal axis and are given in no specific order. Grey thin lines denote score values at individual samples. The peak median value is visualized by a blue line.



(a) K-based suggestion

(b) correlation-based suggestion

Figure 3.10: Pairwise sample correlations of scores at the common densely methylated peaks. Distributions are visualized by box plots.

in measures that correct for peak width. Scaling does not change the correlation values because it essentially multiplies the methylation vector (of a sample) by a constant.

It is important to note that tumor samples have in general a better read coverage than the corresponding healthy ones. However, when the score is corrected for both peak width and inter-sample variability, the densely methylated peaks in normal samples tend to have higher scores than the "leading" peaks in cancer samples. Since the vast majority of the regions are unmethylated and tag densities are relative scores within a sample, the discrepancy at the high methylation scores is compensated by assigning slightly lower values for the unmethylated peaks in normal samples.

3.5 Summary

Different aspects of quality control and filtering were presented in this chapter. The design of the Infinium 450k array incorporates a selection of control probes that aid the process of quality checks for the efficiency of bisulfite conversion, hybridization and other chemical reactions performed in the assay. RnBeads visualizes these readings and builds on top by providing a highly configurable filtering pipeline. A notable step in this pipeline is the Greedycut algorithm for automatic removal of low quality samples and sites. We also presented the comprehensive approach of testing for associations implemented in RnBeads. This approach allows its users to identify or confirm potential batch effects, confounding factors, and other signals encoded in the methylation profiles of the analysed samples.

Finally, four normalized scores for enrichment-based methylation datasets were introduced. These metrics are further compared in the following chapters of this work.

4 DNA methylation profiles

This chapter continues the exploratory analysis style of Chapter 3 and uses the GoldenGate dataset in its examples. Most of the results presented here are also included in the publication that resulted from this collaboration [50]. The first section presents an approach to comparing the methylation variability of sample groups. The following one addresses the question of methylation changes with age and the power of DNA methylation to predict patient's age. The question of age-dependent methylation might seem unrelated to this thesis. However, a growing body of evidence suggests that DNA methylation in human exhibits a drift with age that could contribute to phenotype, including cancer. Moreover, aging pathologies could accelerate the methylation drift, e.g. by chronic inflammation, thereby creating a "vicious cycle" [63]. Identification and characterization of the genomic loci affected by aging is of immense importance to understanding the link between cancer and age, and could be a resource to use in designing preventive strategies.

4.1 Interindividual variation

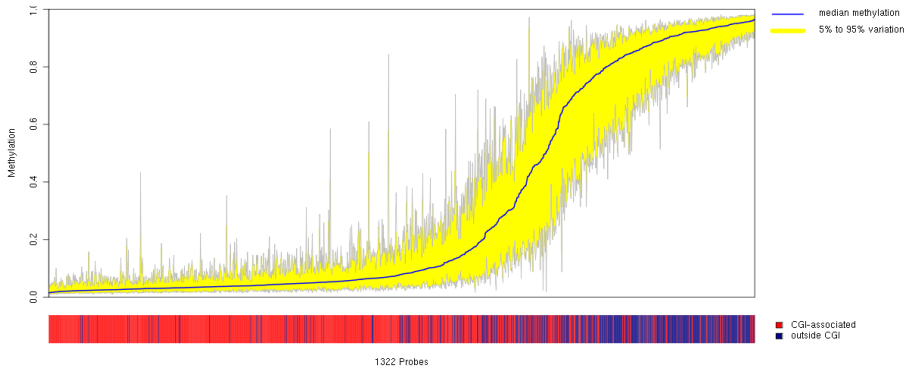
In this section, we attempt to quantify and visualize the variability of methylation in a given tissue across sample population. For this purpose, we constructed the so-called *deviation plot* that depicts the variability of methylation values for a set of samples. Figure 4.1 shows examples of such plots based on the GoldenGate dataset.

The amount of variation in the methylation profiles can be quantified as the relative area of deviation (yellow bars) in a deviation plot, which is a number between 0 and 1. An area of zero indicates no variation, whereas the value of 1 depicts that all possible degrees of methylation are observed for every probe. The deviation areas are also referred to as *profile areas* in this chapter.

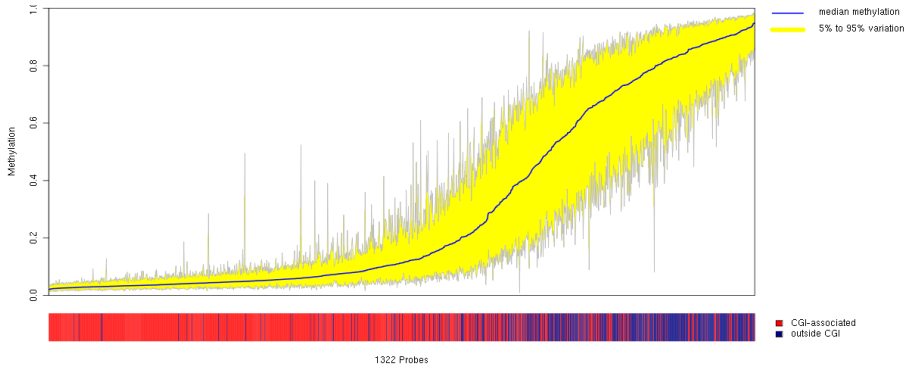
One clear observation from the deviation plots is that there is a strong correlation between median methylation and probe CGI status. This property holds for all sample groups studied in the dataset, and is also the case for mean methylation. The significance of the correlations was estimated by permutation tests using 10^6 repetitions and all p-values were highly significant (data not shown). In almost all sample groups, we observed that CGI-associated probes exhibit significantly smaller variability than the probes that do not lie in a CpG island. Notable exceptions include the profiles of lymphoblastomas, lymphoid neoplasias and cell lines.

4.1.1 Comparisons of sample sets

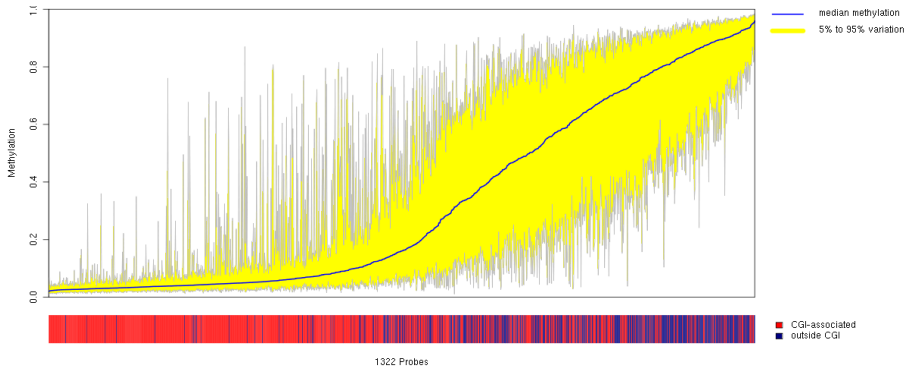
The deviation plots are a very informative tool for visualizing probe and sample variability in methylation. However, direct comparison of deviation plots could be misleading due



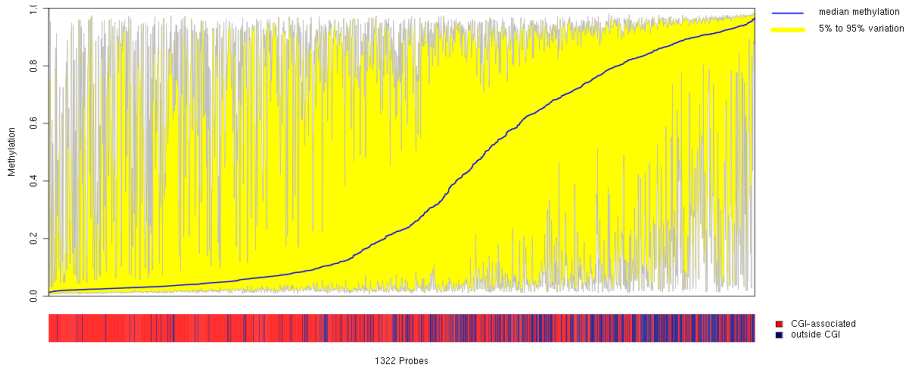
(a) Healthy blood (180 samples)



(b) Healthy colon (97 samples)



(c) Primary tumor colon (110 samples)



(d) Cell lines (107 samples)

Figure 4.1: Deviation plots of selected sample groups. Probes are ordered on the x axis and are sorted in increasing order with respect to their median methylation, as visualized by the blue curve. The yellow area enclosed with a grey border depicts the 5th and 95th percentile among the methylation values for each probe.

to differences in the sizes of the sample sets depicted. As an example, the profile area in the stem cells set is slightly larger than the one for normal blood samples, however, the number of samples underlying the stem cells plot is only 15% of the sample set size for normal blood. While the set of 180 blood samples is likely to be an adequate reflection of the population variability in methylation, we cannot claim this for the small and heterogeneous set of 27 stem cell types. A more unbiased view would involve equal or comparative sample set sizes. The following paragraph presents the approach taken to compare the methylation variability of sample sets based on profile areas.

Within a single sample, methylation measurements across the genome are not pairwise independent, and their inter-dependencies are likely to be very complex. This effectively prohibits deriving an analytical expression for the expected total methylation variability within a population (e.g. whole blood of all female individuals of European descent) based on an observed cohort, even if we assume that patients were selected in an unbiased manner[1]. We thus decided to compare the methylation variability of two sample sets by equalizing them in size, that is, randomly downsampling the larger set, and comparing the observed profile areas. This approach assumes that the compared sample sets are full populations, or are representatives of populations and were drawn using the same bias. An example case of the latter scenario is comparing the methylation variabilities of different tissues obtained from the same (or similar group of) patients. We also repeat the downsampling procedure many times in order to reduce the effect of the additional variance.

The need to work with sets of similar sizes, as well as the effects of downsampling can be seen by studying the large collection of normal blood samples. We can assume that these samples are randomly drawn from a population of interest. We selected 100 random subsets of every possible size from the available 180 samples. The corresponding profile areas are shown in Figure 4.2. The variability of ≈ 40 samples is a value similar to the one computed on the whole set of 180, approximating the variability in the population. If we were to consider a subset of smaller size, the profile area would very likely be a strong underestimation of the full variability.

This technique can be used for comparing the variability of sample groups of different sizes. The variability is quantified by the profile area of the deviation plot for a sample set of a fixed (small) size. This profile area can be estimated for the sample sets of larger sizes using the sampling procedure described above. Figure 4.3 shows the result of the comparison between the groups of healthy colon, primary tumor colon and colon metastases. The healthy tissue shows very coherent methylation profiles across patients, whereas cancer cell populations exhibit much higher methylation variability compared to the cells of origin.

[1] The statistical term is to have a random sample of i.i.d. variables. It is not used in the text because this thesis assigns a different meaning to the word *sample*.

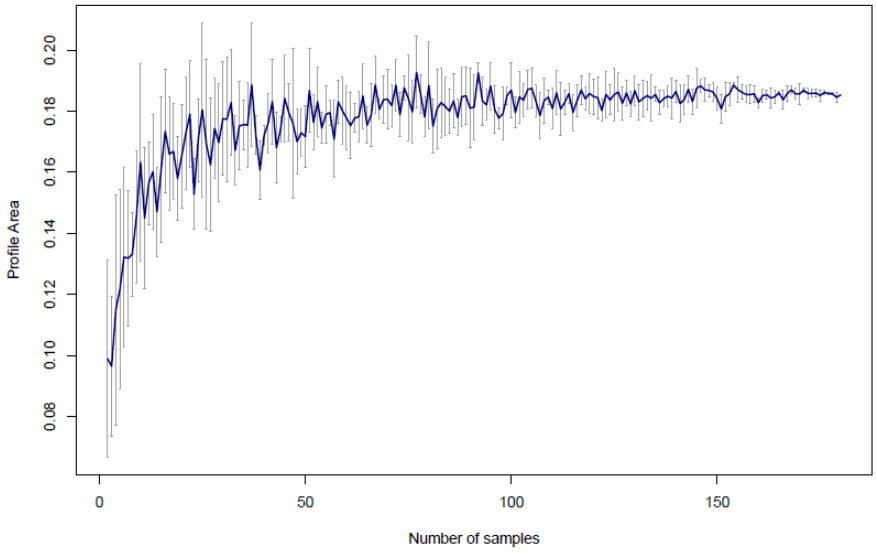


Figure 4.2: Relationship between profile area and sample group size estimated using subsampling of 180 healthy peripheral blood samples. Whiskers depict the full range of observed profile areas in 100 repetitions.

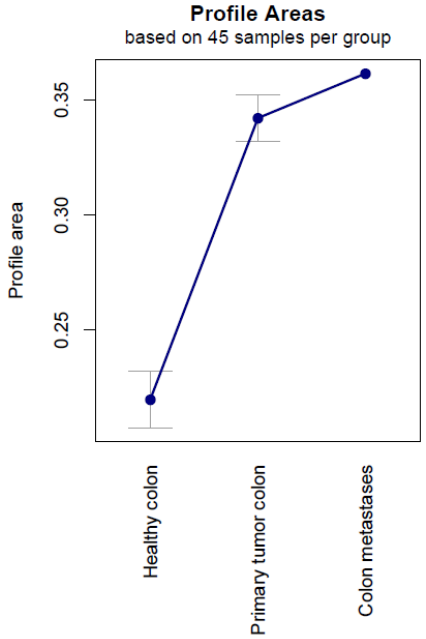


Figure 4.3: Profile areas of subsampled healthy colon, primary colorectal tumor and metastatic tissues. Whiskers show standard deviations based on 100 repetitions.

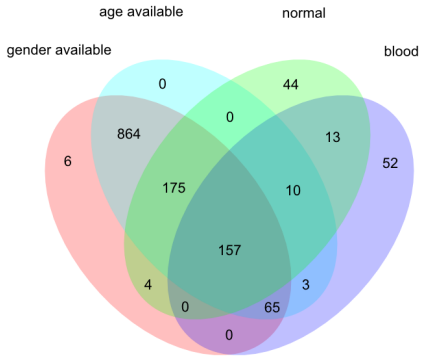


Figure 4.4: Venn diagram listing the number of samples from healthy solid tissues and blood in the GoldenGate dataset.

4.2 Age-dependent methylation

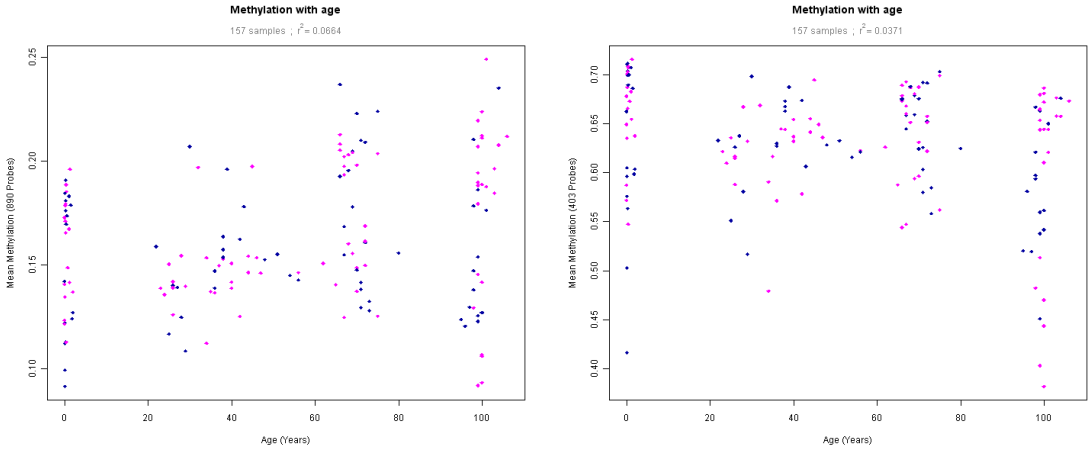
We analyzed the methylation differences of normal samples in the GoldenGate dataset with respect to age. The ages and genders of the originators of 332 of the normal samples were known. The other 71 normal samples were not included in the following analysis. The Venn diagram in Figure 4.4 shows the number of samples from healthy tissues and blood in the dataset. The healthy blood samples are used later for predicting age based on the methylation profile.

The first step in the analysis is to see whether the total (that is, the average) degree of methylation changes with age. In addition to studying globally all probes, we focused on selected subsets. As an example, Figure 4.5 shows mean methylation of the CGI- and non-CGI-associated probes in blood, correlated with the age of all included individuals. The groups of very young (less than 2 years) and very old (over 90 years) individuals show markedly variable inter-individual methylation. This suggests higher variability in cell type composition of blood in those target groups. The observed methylation changes in a large pediatric cohort also indicate a shift in the composition of multiple cell types in peripheral blood [6]. Among centenarians, the wide range of T cell subpopulation frequencies is a phenomenon that has been extensively studied in the past two decades [116, 47].

Despite visible trends of mean methylation changing with age, the correlations between age and average methylation in the studied probe groups and tissue types are rather weak. A notable exception is bone marrow where $r^2 = 0.64$, however, the limited set of (only 11) samples from this tissue does not allow an extrapolation claim that the observed high correlation is a global trend.

4.2.1 Age-associated probes

The sample groups of healthy blood and colon present the opportunity to investigate relationship between age and methylation at the level of specific probes. Similar studies



(a) CGI probes (b) non-CGI probes

Figure 4.5: Scatter plot of mean methylation of (a) CGI-associated and (b) non-CGI-associated probes in healthy blood samples. Probes on chromosome X are excluded. Every point represents a sample; pink and blue colors denote female and male gender, respectively.

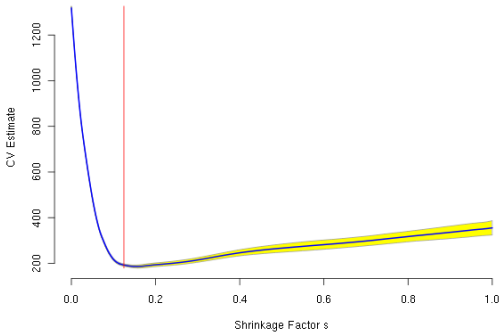
have been performed on healthy, as well as on cancer samples [110]. For each of these two sample groups, we measured the correlation between every probe and the respective age of the individual. In addition, we used the methylation state of a probe as a single covariate in a generalized linear model to predict age. Applying analysis of variance (ANOVA) on the models then enabled us to quantify the predictive power of every probe through a p-value. The p-values were corrected for multiple testing using the Bonferroni method. After applying a threshold of 0.01, 342 probes in total exhibit significant association between age and their methylation state in blood. 78 of these probes have a absolute correlation with age of over 0.5; this list of probes is presented in [50]. Similarly, the ANOVA procedure revealed that methylation state in colon of 10 probes is significantly associated with age. These probes are listed in Supplementary Table S2.

4.2.2 Predicting age

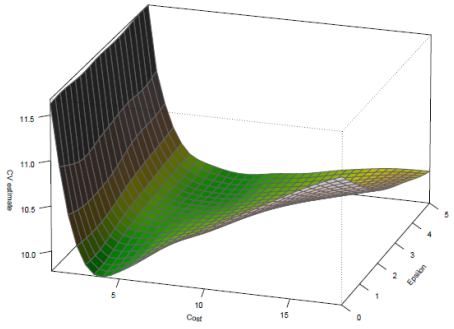
As we have seen from the analyses above, global methylation level is not a good indicator of age due to high inter-individual and inter-tissue variances. We decided to circumvent the influence of tissue-specific methylation by selecting blood samples only, and then we tested two methods for predicting age based on the methylation of the available CpG loci. In this setting, we used either all available samples, or individuals of age between 20 and 80 years only. The number of all healthy blood samples annotated with age is 157. 85 of them have an age between 20 and 80 years. Later in this report, we refer to the aforementioned sample sets as *unrestricted* and *restricted* scenarios, respectively.

Prediction accuracy

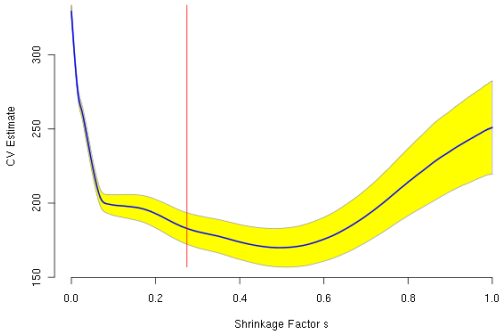
In terms of statistical learning, age prediction in our case is a regression problem in a high dimensional feature space with a small training set. These properties restrict us to



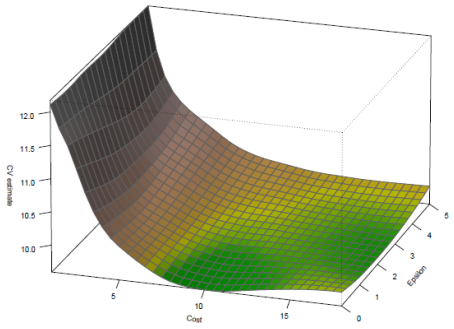
(a) Lasso regression, unrestricted set



(b) SVM, unrestricted set



(c) Lasso regression, restricted set

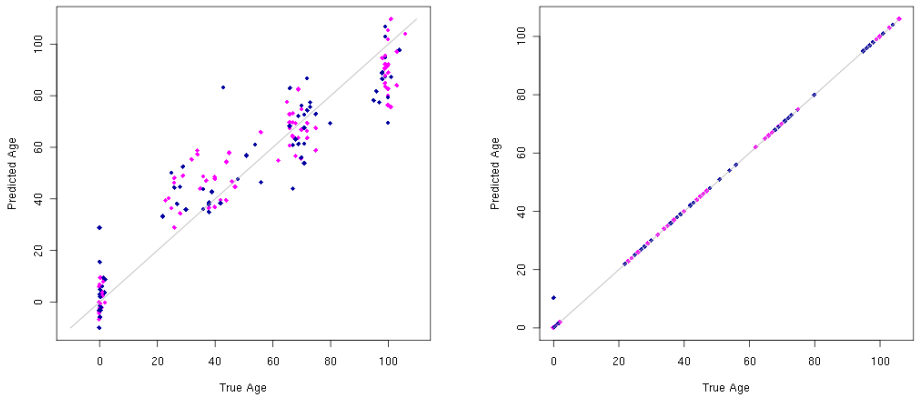


(d) SVM, restricted set

Figure 4.6: CV estimates of the error, used in training lasso regression on age based on the unrestricted (a) and restricted (c) sample sets, as well as SVM on the same sample sets (b) and (d). CV estimates in the lasso method are mean squared error, and for SVM are absolute error in years. The visualized values were used in parameter estimation. The selected values for the lasso models are depicted by red vertical lines.

apply strong assumptions – feature independence among others – and use linear models. The first model we tested is lasso method for regression. We estimated the best value for the shrinkage factor s with 10-fold cross validation (CV). Due to the limited number of available samples, CV estimates were highly unstable, showing large dependence on the (random) splitting of the entire set into folds. Therefore, we repeated the 10-fold CV estimation procedure 100 times and averaged the outcomes. The second model we tested is linear support vector machine (SVM), in the form using ε-insensitive loss. We experimented with values for ε from 0 to 5, and values for the cost parameter C from 1 to 18. CV estimates were computed in complete analogy to the lasso method. Taking into account the common source for all probes, we did not normalize the inputs before fitting a model.

Figure 4.6 shows the calculated CV estimates for absolute error. We applied the orthodox strategy for selecting the shrinkage factor in the lasso models. More precisely, we selected the simplest model with a CV estimate within one standard deviation of the



(a) Lasso regression (b) SVM

Figure 4.7: True and predicted age of the unrestricted sample set modeled by lasso regression (a) and SVM (b).

minimal observed. The models with the selected shrinkage factors trained on the full sets, apply non-zero coefficients to 22 and 28 probes for the unrestricted and restricted scenario, respectively. The methylation states of all other probes do not play a role in age prediction. Not surprisingly, the two sets of selected probes have a significantly large overlap (12 probes in common, p-value $= 3.1 \times 10^{-16}$).

Contrary to our expectations, restricting the age span to the range between 20 and 80 years does not enable the training of more accurate models. Both methods show comparable results, SVM slightly outperforming the lasso method. The estimated mean error of a lasso model is approximately 10.5 years; the corresponding error for an SVM is ≈ 9.5 years.

Model variance

In comparison to the lasso method, support vector machines seem more unstable. The unrestricted and restricted scenarios lead to very different landscapes of estimated absolute prediction over the parameter space. The CV estimates in the restricted scenario suggest that applying higher values to the cost parameter might lead to more accurate models. Another indication of the instability of the SVM models is given in the paragraphs below.

In the following analysis, we compared the models trained on the full sample sets with parameters estimated by the procedures described above. Therefore, we show the performance of age prediction models on their training sets. This could provide hints to what extend the linearity assumption of the methods hold, and also if the classifiers tend to under- or overfit the data. The scatter plots in Figure 4.7 show true versus predicted age for blood samples. In a perfect predictor, all points lie along the diagonal grey line.

Clearly, SVM prediction errors are extremely low. Focusing on the lasso models, we can

see that predictions of samples with true age below 55 years tend to have an upward bias. Contrarily, the predictions for samples of ages more than 55 years are downward biased. The plots also nicely visualize the insecurity of the compared predictors. The training errors for the SVM models are unrealistically low, as we have shown by the CV estimates in the previous section. This tendency to overtrain, as well as the instability on the SVM models is mostly explained by the fact that they use the methylation values of all probes in age prediction. We have tested values for the cost parameter only up to 18. Higher values for the cost parameter might lead to more stable models, however, prediction accuracy unlikely to improve drastically.

4.3 Results and discussion

This chapter focused on methylation microarrays and discussed problems in studying methylation profiles of sample groups. All analyses were performed on the GoldenGate dataset, however, the conclusion drawn transcend this technology. First, we presented an approach to visualize and quantify overall methylation variability in a population. We showed a clear distinction between CGI- and non-CGI-associated probes with respect to their inter-individual variability. Of note, the deviation plots introduced here are also implemented in RnBeads.

We next studied the relationship between locus-specific (and also genome-wide) methylation and age of an individual. We identified a group of CpGs as candidates for loci with age-dependent methylation, some of them mentioned in previous studies. Overall, linear SVMs show us that relatively accurate age prediction of linear models based on DNA methylation in blood is theoretically possible. However, a representative set of much more than the available 157 samples is required. Lasso models, on the other hand, can be created based on the currently available data without risk of underfitting or overtraining. Moreover, a lasso predictor does not require whole methylation profile in order to predict the age of new samples; it focuses on a small number of CpG loci. Unfortunately, the accuracy of this model family does not seem to be remarkable which is partially due to the biased genomic coverage of the GoldenGate assay. The difference between predicted and true age is often more than 10 years. Recently, our hypothesis was confirmed by Horvath, who trained a linear model with elastic net penalty to predict (a transformed) chronological age based on Infinium probe methylation [59]. The model selected a set of 353 CpGs that are informative of age and showed remarkable accuracy across a wide spectrum of tissues.

5 Differentially methylated regions

CpG-specific analysis of differential DNA methylation provides very high resolution maps of epigenetic alterations. It is a valuable resource with a wide spectrum of applications, for example, identifying change of affinity in (unknown) transcription factor binding motifs. However, the single-site approach presents considerable challenges due to its high multiple testing burden and also because single cytosines could be strongly affected by biological and technical noise.

5.1 Differential methylation in microarrays

In the GoldenGate dataset, we trained elastic nets not only for the purpose of classification, but also to be used as feature selection methods. In addition, we defined heuristic rules to identify differentially methylated probes when small sample sets were compared. In general, probes were considered independently and two criteria were applied – statistical testing (usually Wilcoxon test or Kruskal Wallis test) and a threshold for minimal difference between the mean methylations of the compared groups. RnBeads implements an elegant rank-based approach to address the question of differential methylation but it is not discussed in this work[1]. Instead, we present strategies for the identification of differentially methylated peaks in enrichment-based methylation data, using the colon cancer dataset as their application.

5.2 Differential methylation in called peaks

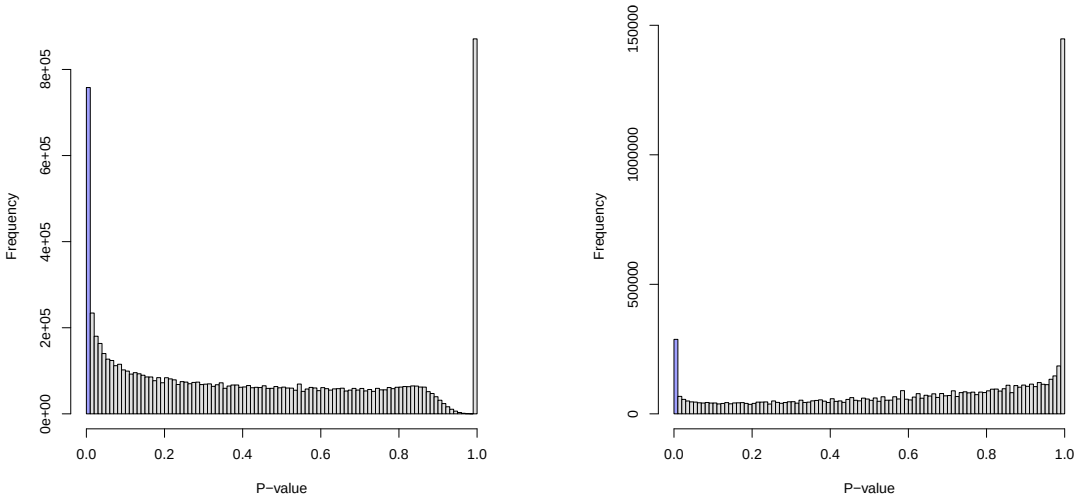
This section reflects the work on the colon cancer study within the CANCERDIP consortium [104] . It focuses on calling differentially methylated peaks (or regions, DMRs) given enrichment data on paired samples. In the colon cancer dataset, these sample pairs are tumor vs corresponding healthy tissue of the same individual. As already discussed in Chapter 3, raw enrichment data is comprised of read counts in preselected genomic regions (identified peaks). In the following, we use the normalized scores defined earlier in order to facilitate comparisons between samples. We introduce the concept of a differential methylation event, and conclude by commenting on the properties of the regions for which hypo- or hypermethylation is a frequently observed event.

5.2.1 Criteria for differential methylation

Association and odds ratio

Two measurements are available for every region and patient – one from the normal and one from the tumor sample. We tested for correlation between the observed read counts

[1] The differential methylation module of RnBeads has been implemented by Fabian Müller.



(a) Before correction

(b) After correction

Figure 5.1: Histograms of p-values (a) and q-values (b) obtained by testing for association between absolute tag count and cancer. The left-most bar, colored in purple, shows the frequency of values below the threshold of 0.01.

in both samples using Fisher's exact test. An example for a region on chromosome 21 is given in the table below.

	8N	**8T**	**Total**
21:46,574,663-46,667,956	3,451	3,915	7,366
Rest of the genome	26,247,134	12,993,096	39,240,230
Total	26,250,585	12,997,011	39,247,596

The p-value for this test on patient 8 is extremely low: 3.2×10^{-273}. Using this procedure, we constructed a table of p-values for all genomic regions and sample pairs. Every row in this table denotes a peak, and each column – a patient. The dimensions of the table are 329,613 rows $\times$ 24 columns. Next, we adjusted the p-values for multiple testing using the Benjamini-Hochberg algorithm. The distribution of all obtained p-values before and after adjustment is given in Figure 5.1. The table of adjusted p-values is used as the first criterion to establish if a region R is differentially methylated in a given patient. The second criterion is referred to as the *odds ratio* and is defined below.

Given a region R and a matched pair of samples S_N and S_T, we used the score of R as an indication of hyper- or hypomethylation in tumor. Large difference in the two values implies differential methylation. We quantify the term large difference by the constant c, as shown in the inequalities below:

$$\text{Implication for hypomethylation:} \quad D_R(S_N) > 0 \ \cap \ c \times D_R(S_T) \leq D_R(S_N)$$
$$\text{Implication for hypermethylation:} \quad D_R(S_T) > 0 \ \cap \ c \times D_R(S_N) \leq D_R(S_T)$$

Empirical studies suggest that a value of $c = 2$ achieves a good compromise between specificity and sensitivity in differential methylation. We refer to this requirement for a

minimum of two-fold enrichment or depletion as odds ratio. Note that scores of the same region in different samples are compared and thus normalizations on the region length have no effect on the outcome. Consequently, we compare the DMRs identified using tag count, scaled tag count and tag density only. Using tag occupancy and scaled tag occupancy leads to identical results with tag count and scaled tag count, respectively; therefore, the occupancy measures are omitted in the discussions that follow.

Combining criteria for differential methylation

The strategy we applied for the identification of differential methylation involves both criteria described above. In particular, region R is *hypomethylated* in a given patient when the odds ratio of the respective sample pair implies hypomethylation *and* there is a significant association between the tag counts and cancer status of this region, quantified by adjusted p-value < 0.01. Similarly, R is *hypermethylated* in a given patient only when the odds ratio implies hypermethylation and the corresponding adjusted p-value is below the significance threshold shown above. In the following, we present a comprehensive analysis of our approach to identifying DMRs. One of the aspects covered is the properties of the criteria both individually and acting in concert. In the latter cases, the strategy described in this paragraph is referred to as the *combined strategy*.

Region length bias

An inherent property of Fisher's exact test is that large values in the contingency table have the potential to (and also tend to) produce lower p-values for significant association. Applying multiple testing correction – a necessary step when millions of tests are performed – exacerbates this effect. As a consequence, a peak with low number of tag counts across all samples is unlikely to be identified as a DMR, even if it shows very high odds ratios. To quantify this potential bias, we inspected the peaks that pass the significance threshold for association between tag count and cancer in at least one patient. We studied their lengths and median number of tags. Figure 5.2 shows the distribution of the inspected values, compared to the corresponding values for the full set of peaks, as well as the regions that show implication for differential methylation.

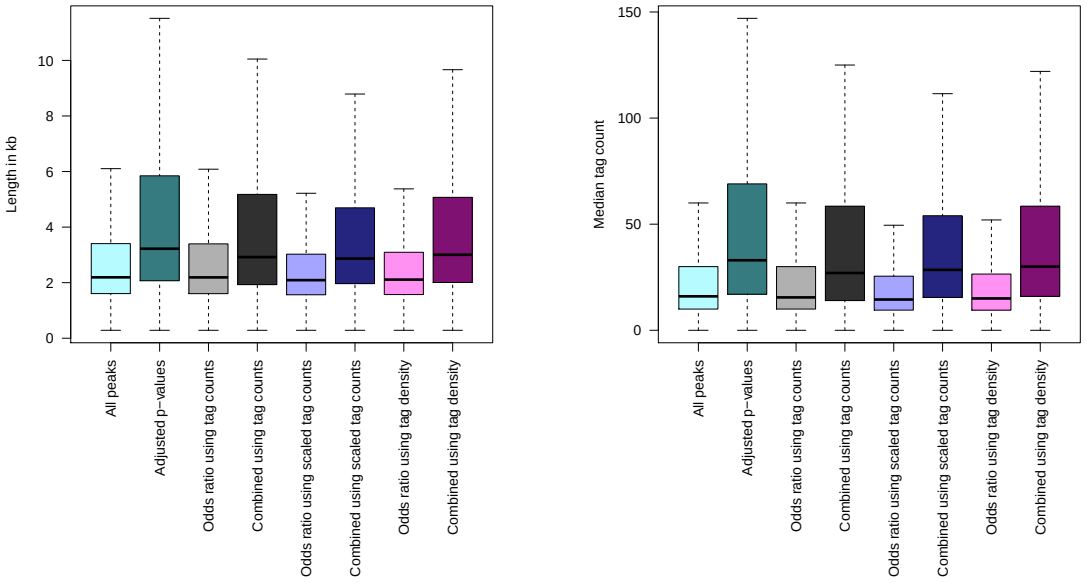
Clearly, the approach involving Fisher's exact test introduces a strong bias towards regions that are long and heavily populated with tags. This property is not shared by the procedure using the odds ratio, suggesting that Fisher test p-values alone might be characterized by a high type II error (false negative rate) and are therefore a potentially unreliable approach.

5.2.2 Differential methylation events

We categorized every region R into one of the following four types:

hypomethylated indication for hypomethylation of R is observed in at least one patient, but not for hypermethylation;

hypermethylated indication for hypermethylation of R is observed in at least one patient, but not for hypomethylation;



(a) Peak lengths

(b) Tag counts per peak

Figure 5.2: Distributions of widths and median tag counts for: all peaks, peaks that show significant association between tag count and cancer in at least one patient; peaks that show implication for differential methylation in at least one patient; peaks that are identified as DMRs in at least one patient.

hypo- and hypermethlyated indications for hypo- and hypermthylation of R are observed;

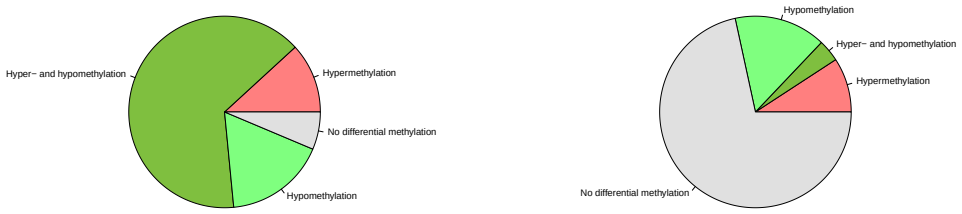
not differentially methylated indication for differential methylation of R is not observed in any of the patients.

Figure 5.3 shows the distribution of region types across the whole genome when tag density is used for calculating odds ratio. The other scores exhibit very similar behaviour.

If we examine the number of non-differentially methylated regions, it is clear that the specificity of odds ratio as a single criterion is very low. Combining the two criteria introduced earlier in this chapter substantially reduces the number of observed hypo- and hypermethylation events. The combined strategy is therefore the preferred approach towards identifying a small set of functionally relevant differentially methylated regions.

Number of DMRs per patient

The patients show very inconsistent behaviour in terms of hypo- and hypermethylation events observed. Multiple factors related to sample preparation and DNA processing, in addition to tumor heterogeneity, seem to contribute to this effect. The barplot in Figure 5.4 below show the number of hypo- and hypermethylated peaks in every sample pair based on tag density. While the absolute number of hyper- and hypomethylation events is strongly influenced by the choice of peak score, the inter-patient differences remain remarkably stable.



(a) Odds ratio only (b) Combined strategy

Figure 5.3: Relative frequencies of region categories genome-wide. The score considered is tag density.

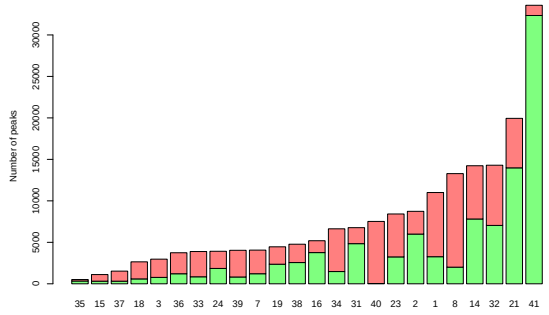


Figure 5.4: Number of hypo- and hypermethylation events per patient. Hypomethylation is denoted by green color, and hypermethylation - by red.

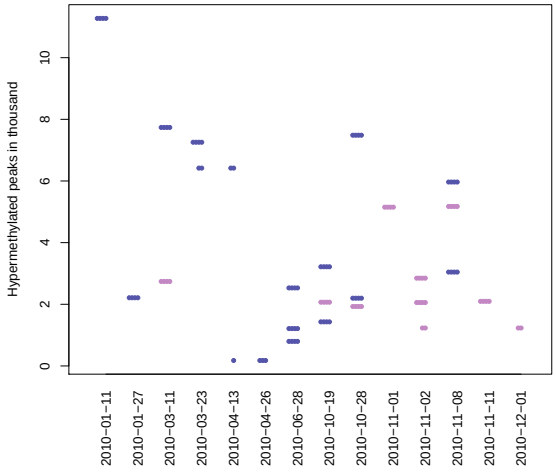


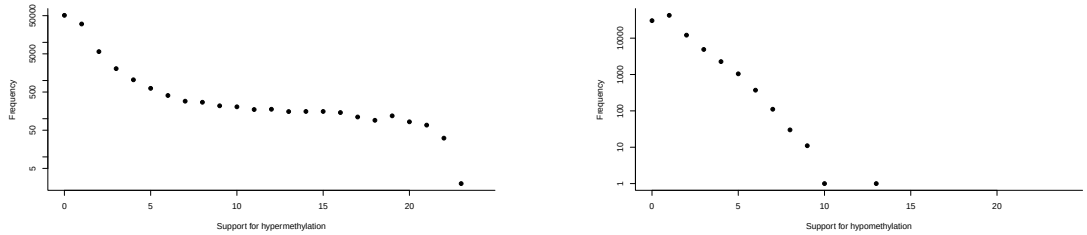
Figure 5.5: Relationship between a batch processing factor (date of MethylCap) and the number of hypermethylated regions found for the respective patient. Color denotes gender; pink points are samples from female patients, and blue points – from males.

In order to find potential sources for the observed interindividual discrepancies, we checked for associations between number of differentially methylated regions and batch processing or clinical features. We selected the properties introduced in Chapter 2. The results are presented in Supplementary Table S3. A detailed discussion on the observed p-values is beyond the scope of this thesis; as an example, Figure 5.5 shows a non-trivial, albeit a strong dependency between MethylCap processing date and number of hypermethylated regions.

5.2.3 Support for hypo- and hypermethylation

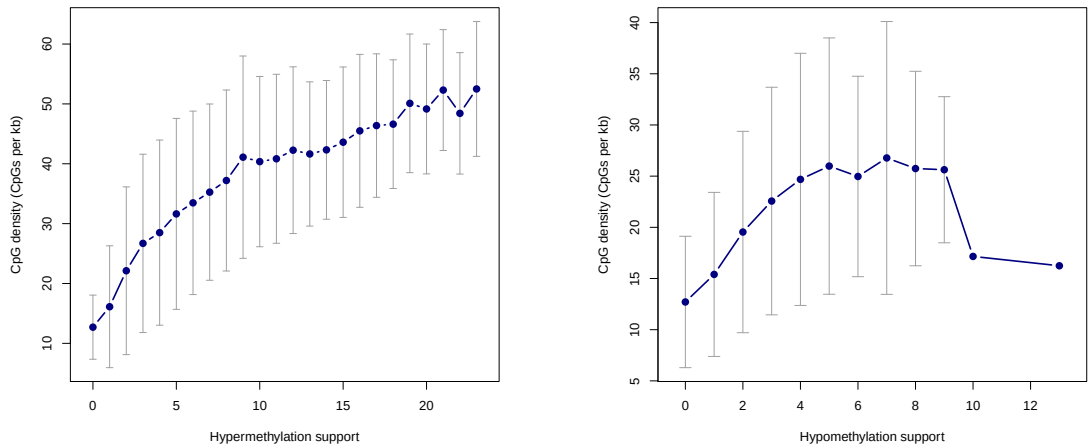
We focused on the combined strategy and inspected the evidence for hypo- and hypermethylation observed for each peak. We define the *support for hypomethylation* of region R as the number of patients in whom R is found to be hypomethylated. The term *support for hypermetylation* is defined analogously. Since we have data on 24 patients, the supports of hypo- and hypermethylation for a given region R are integer numbers between 0 and 24. Note that support is defined only in the context of a given peak score. Indeed, the values for support vary (albeit not dramatically) among the measurements inspected in this study. Figure 5.6 shows the distribution of values for hypo- and hypermethylation supports based on tag density as a score. Using other scores leads to almost identical distributions.

As can be seen in Figure 5.6, the support for hypomethylation exhibits a small variance, whereas the support for hypermethylation spans the interval from 0 to 23. The observed inconsistency in hypomethylation could be explained by two aspects. First, advantageous hypomethylation of specific DNA region in cancer is more difficult to identify due to the genome-scale loss of methylation during tumor progression. Second, genomic regions undergoing significant hypomethylation in colon cancer could possess sequence characteristics disfavored by the MethylCap-seq technology. In addition, a large fraction of the regions



(a) Hypermethylation (b) Hypomethylation

Figure 5.6: Histogram of values for support for hypo- and hypermethylation. Frequencies are depicted as points.



(a) Hypermethylation (b) Hypomethylation

Figure 5.7: Relationship between support for (a) hyper- / (b) hypomethylation on one side, and CpG density on the other side. Mean values of CpG density are depicted by blue points; whiskers denote one standard deviation.

with a hypermethylation support above 8 are also characterized by a positive support for hypomethylation and vice versa.

We found that hypermethylation support of a peak is positively correlated with CpG density. Not surprisingly, it is also associated with [distance to the nearest] CpG island. However, the dependency between hypomethylation support and these properties of the DNA sequence is unclear. As an example, Figure 5.7 visualizes the joint distributions of support and CpG density.

High confidence differentially methylated regions

In this section, we treat hypomethylation and hypermethylation as independent events. After applying the combined criteria for detection of differential methylation, we constructed tables of events. The rows in these tables correspond to all peaks, and columns represent patients. Every cell in a table is labeled as hypomethylation, hypermethylation

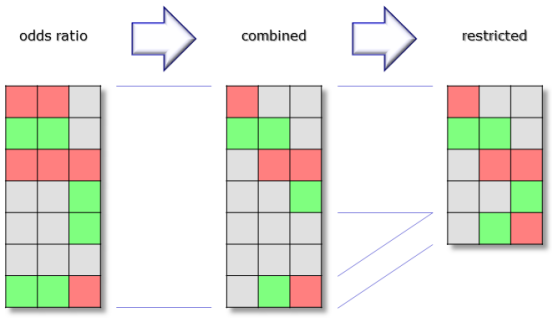


Figure 5.8: Schematic representation of the event tables constructed in the differential methylation analysis of the colon cancer dataset. Rows in these tables correspond to peaks, and columns represent patients. Green cells denote hypomethylation events, whereas red cells indicate hypermethylation events. Grey depicts lack of evidence for differential methylation.

or no differential methylation. The *odds ratio table* contains the indications of hypo- and hypermethylation based on the odds ratios defined above. Similarly, the *combined table* stores the events from the combined strategy introduced in this chapter. If we focus only on those regions that exhibit hypo- or hypermethylation in at least one patient, we obtain a smaller table, which we refer to as *restricted*. The restricted table is used later in the estimation of thresholds for defining high-confidence DMRs. A schematic representation of the tables introduced in this paragraph is shown in Figure 5.8.

Undoubtedly, a large value for support for an event in region R implies high confidence in the statement that the event is correctly identified and is recurrent in colon cancer. In order to assign a p-value for confidence to every region based on its observed support, we model the construction of events by a generator that uses a finite sequence of Bernoulli trials. In other words, according to our hypermethylation model, a biased coin is flipped for each cell in the event table. The outcome determines if the cell is labeled as evidence for hypermethylation. The probability for success of a Bernoulli trial is estimated as the fraction of hypermethylation events in the resulting table. A similar generator is used to label some of the cells as evidence for hypomethylation. We refer to the models describing the combined table as *liberal*, whereas the models describing the restricted table are *conservative*. The estimated probabilities are shown in Figure 5.9.

Given the models described above, we can compute the probability that a region R is provided support S for a certain event. Since the event labels of R are the results of 24 Bernoulli trials, the support values follow a binomial distribution. Figure 5.10 shows the complementary cumulative distribution functions for the hypomethylation and hypermethylation events.

We used the conservative models and applied a significance threshold of 10^{-3}. In most cases, this translated the requirement for a minimum hypo- or hypermethylation support of 6. Finally, we labeled a region hypermethylated in colon cancer with high confidence

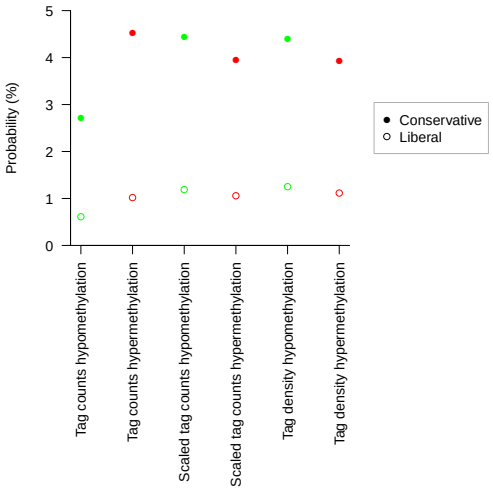
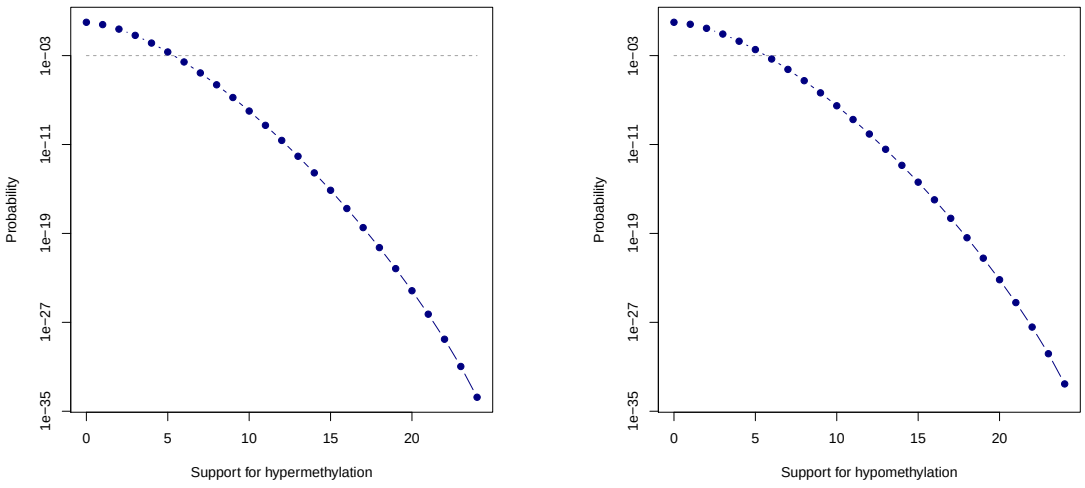
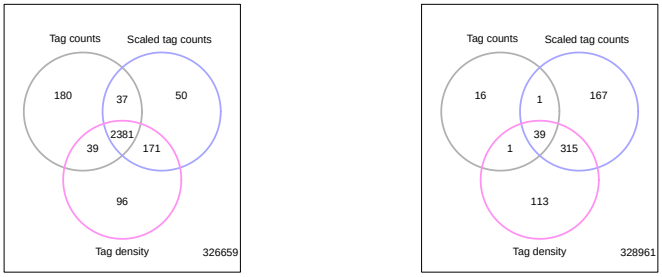


Figure 5.9: Scatter plot of inferred probabilities for hypo- and hypermethylation. Empty and filled circles represent liberal and conservative probabilities, respectively. Hypomethylation is indicated by green color, whereas hypermethylation probabilities appear in red.



(a) Hypermethylation (b) Hypomethylation

Figure 5.10: Complementary cumulative distribution function (tail distribution) of a binomial random variable. Success probability is calculated as the fraction of hyper- (a) or hypomethylated (b) events. Number of trials is total number of events. The applied threshold of 0.001 is depicted by a horizontal grey dashed line.



(a) Hypermethylation (b) Hypomethylation

Figure 5.11: Venn diagram showing the overlaps between the sets of high confidence DMRs obtained using different peak scores.

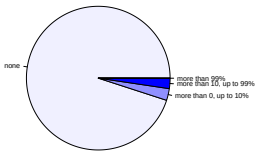
when both of the following conditions are met: (1) the support for hypermethylation of this region is at least the minimum support required, and (2) the support for hypomethylation of this region is at most 1. The property hypomethylated in colon cancer with high confidence is defined analogously. In total, we identified between 57 and 522 hypomethylated, and over 2,600 hypermethylated regions using the definition above.

Figure 5.11 below shows the overlaps between the discussed sets of peaks. The lists of high confidence hypermethylated regions obtained using different peak scores show a very strong agreement. The sets of hypomethylated regions are less consistent.

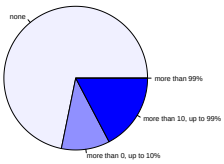
Next, we examined the CpG-related properties of the high confidence DMR sets. For every peak region, we counted the fraction of its sequence that lies in a CpG island (CGI). The vast majority of peak regions are located outside CGIs; less than 0.04% of all regions are fully occupied by an island. Figure 5.12 divides the regions into four groups based on the fraction of region identified as CpG island. It also shows the distribution of CGI fractions for those regions that have a positive overlap.

Note that the sets of hypomethylated regions are enriched in peaks that overlap with CpG islands. Since there is no association between support for hypomethylation and distance to closest CGI (Figure 5.7), there could be two or more distinct categories of genomic elements that are hypomethylated in colon cancer.

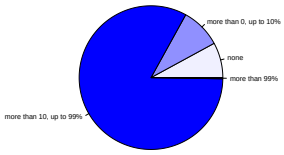
We also inspected the properties described in Figure 5.7 for the high confidence regions. The results are visualized in Figure 5.13. As expected, hypomethylated regions display a high variability in the distance to the closest CpG island, whereas the overwhelming majority of the hypermethylated peaks are overlapping at least partially with a CGI. In addition, we observe a significantly higher CpG density in the hypomethylated regions, and even a higher one in the hypermethylated ones.



(a) All regions

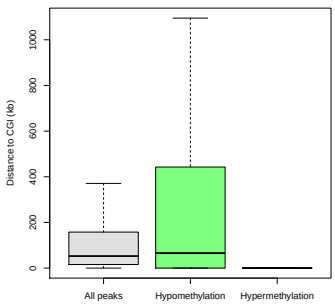


(b) High-confidence hypomethylated regions

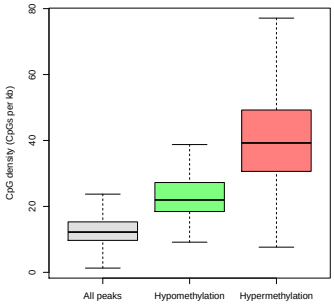


(c) High-confidence hypermethylated regions

Figure 5.12: Separation of peaks based on their overlap with CpG islands.



(a) Distance to CGI



(b) CpG density

Figure 5.13: Distribution of the values for distance to closest CGI / CpG density for three groups of peaks – all regions, high confidence hypomethylated regions, and high confidence hypermethylated ones.

5.3 Summary

We inspected two criteria for differential methylation and argued that their combination is a sensible approach to identifying a reliable set of DMRs. In consideration of the results presented in the last section, we can conclude that the procedure outlined in this chapter produces relatively consistent results regardless of the peak score used. Generally, applying normalization to the score leads to the identification of a larger number DMRs in comparison to using raw values for score, i.e. tag counts. Approximately 7% of all peaks are consistently identified as high confidence hypermethylated regions. However, the set of high confidence hypomethylated regions is an order of magnitude smaller; it is also more sensitive to the peak score in use.

6 Prioritization of epigenetics biomarker candidates

6.1 The need for data integration

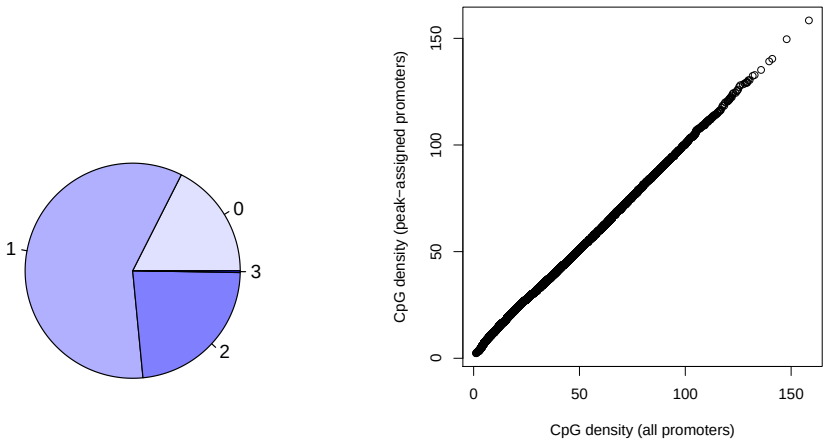
Several large-scale epigenome mapping initiatives are currently ongoing and show promising results. Chapter 2 already mentioned the large consortia TCGA and ENCODE. Another example is the International Human Epigenome Consortium (IHEC), having a goal to produce comprehensive epigenome maps for 1,000 biomedically relevant human cell populations [98]. The European BLUEPRINT project focuses on hematopoietic cells and their associated diseases [1]. The DEEP project investigates cell types relevant for metabolic and inflammatory diseases. The International Cancer Genome Consortium (ICGC, icgc.org) is the international collaboration with goals and approach almost identical to TCGA. ICGC produces genomic, epigenomic and transcriptomic profiles of samples from 50 different cancer types.

All initiatives described above stress the need for statistical and software tools for integrative data analysis that guides biologically relevant findings. A considerable progress is made in this direction [43], however, we still lack a clear explanation for the causes and direct consequences of aberrant DNA methylation in diseases. This chapter addresses the issue of integrating additional data sources in DNA methylation studies in order to validate, filter and prioritize identified differentially methylated regions, potentially indicative of cancer progression. Similarly to Chapter 5, the analysis steps presented here are the result of a collaboration on the colon cancer dataset within the CANCERDIP consortium [104].

6.2 Integration of MethylCap-seq data with TCGA

This section briefly studies the colon cancer methylation and expression datasets available for download from the TCGA project. It investigates the agreement between the conclusions that can be drawn from these datasets and the regions we have identified in the chapter on differential methylation (Chapter 5).

The technology for quantifying DNA methylation used in the data downloaded from TCGA is the Infinium 27k assay developed by Illumina [16]. This platform was introduced in Chapter 1. Briefly, bisulfite-converted DNA is hybridized on an array that covers 27,578 CpGs in the human genome. The large majority of the studied dinucleotides are located in the promoter regions of protein- and RNA-coding genes. The assay defines the promoter area of a gene as the region starting 1.5 Kb upstream and ending 1 Kb downstream of



(a) (b)

Figure 6.1: (a) Pie chart classifying the Infinium 27k promoters based on the number peaks in the colon cancer dataset that overlap them. (b) CpG density of all Infinium 27k promoters compared to the density of only those promoters that overlap with peaks.

its transcription start site (TSS). Only 622 probes in the assay are not assigned to a gene. Following a close inspection of the probe's coordinates, and respecting the promoter definition given above, we unambiguously assigned 164 of these probes to a gene. After this update, there are in total 27,093 CpGs that can be assigned to the promoter area of a gene with known symbol. The Infinium assay includes probes for 14,470 gene promoters in total, which we refer to as Infinium promoters (and Infinium genes, respectively) later in this chapter. Every promoter is covered by 1 to 12 probes. We inspected the overlaps of the Infinium promoter regions with the peaks from the colon cancer dataset. Approximately 80% of the promoters overlap at least partially with one or more peaks. We refer to these as *peak-assigned promoters*. Figure 6.1(a) shows the relative sizes of different promoter groups, classified based on how many peaks they overlap with. The overlap statistics depicted there are a good indication that the set of Infinium promoters can be used as a validation for the MethylCap-seq profiles. Moreover, the Infinium promoters that can be assigned to peaks do not introduce a strong CpG density bias[1], as can be seen in the Q-Q plot on Figure 6.1(b).

6.2.1 Methylation dataset

We downloaded Infinium 27k methylation profiles for 15 normal (healthy) and 165 tumor colon samples from the colon adenocarcinoma dataset of TCGA. Every normal sample had a matched tumor sample from the same patient. For the remaining analyses, we focused on these 15 sample pairs and ignored the remaining unmatched tumors. Detection p-values were used only in cases when duplicated samples were merged. In this scenario, probes were considered independently and the combined β value for a probe P was calculated as

[1] Although applying Wilcoxon or Kolmogorov-Smirnov test to compare the distributions of CpG densities for promoters overlapping and not overlapping with peaks yields a significantly low p-value.

the median β value in all duplicates that have a detection p-value < 0.01 for P.

The strategy we applied for identifying differentially methylated CpG loci and Infinium promoters is motivated largely by the approach described in the chapter dedicated on this topic, Chapter 5. We considered every probe or promoter independently, and used two criteria for differential methylation: (1) absolute increase in median methylation of least 0.25, and (2) a p-value < 0.01 obtained after applying Wilcoxon rank sum test and multiple testing correction using the Benjamini-Hochberg method.

As an example, let us assume that two probes are associated with the promoter of gene G. Therefore, there are 30 β values for G in normal colon samples and 30 – in colon cancer samples. If we denote these sets of numbers as $N_G^{(\beta)}$ and $T_G^{(\beta)}$, respectively, we classify G as hypomethylated if and only if

$$M\left(N_G^{(\beta)}\right) - M\left(T_G^{(\beta)}\right) \geq 0.25 \;\;\&\;\; P_{\text{Wilcoxon}}\left(N_G^{(\beta)}, T_G^{(\beta)}\right) < 0.01$$

where M denotes median and P_{Wilcoxon} is the FDR-corrected p-value after applying Wilcoxon rank sum test. The criteria for hypermethylation are constructed in an analogous fashion.

Using the procedure outlined above, we identified 161 hypomethylated and 583 hypermethylated Infinium promoters. The first criterion turned out to be stricter; a summary of the number selected probes and promoters is avaialable in Supplementary Table S4. The heatmap in Figure 6.2 shows the methylation degrees of these promoters in all sample pairs.

6.2.2 Expression dataset

At the time of these analyses, the gene expression platform of choice in TCGA was Agilent G4502A – a custom gene expression microarray in high density format [72]. It includes 244 thousand features that correspond to approximately 111,000 unique probes. The majority of these probes target mRNA transcripts, each of which can be mapped to a unique location in the genome. We focused only on the transcripts that are unambiguously mapped to a genomic location and refer to them as *Agilent transcripts*. We first constructed a mapping between Infinium genes and Agilent transcripts. More precisely, we associated transcript T to gene G when T is located downstream of G's transcription start site, lies at a distance of maximum 10 Kb from it, and there is no other TSS closer to T. Note that the Infinium genes are only a subset of all genes in the human genome, therefore, a fraction of the Agilent transcripts are expected to not be assigned to an Infinium gene. Indeed, the overwhelming majority of the transcripts are not located in the vicinity of any Infinium promoter. The closest TSS to a transcript sometimes lies at more than 40 Mb. Figure 6.3(a) shows part of the distribution of distances to the closest TSS of Infinium gene.

Using the procedure outlined above, we mapped 16,609 transcripts to Infinium genes. Note that not all probes contained measurements in the TCGA study on colon adenocarcinoma, therefore, expression values were not available for all the Agilent transcripts that

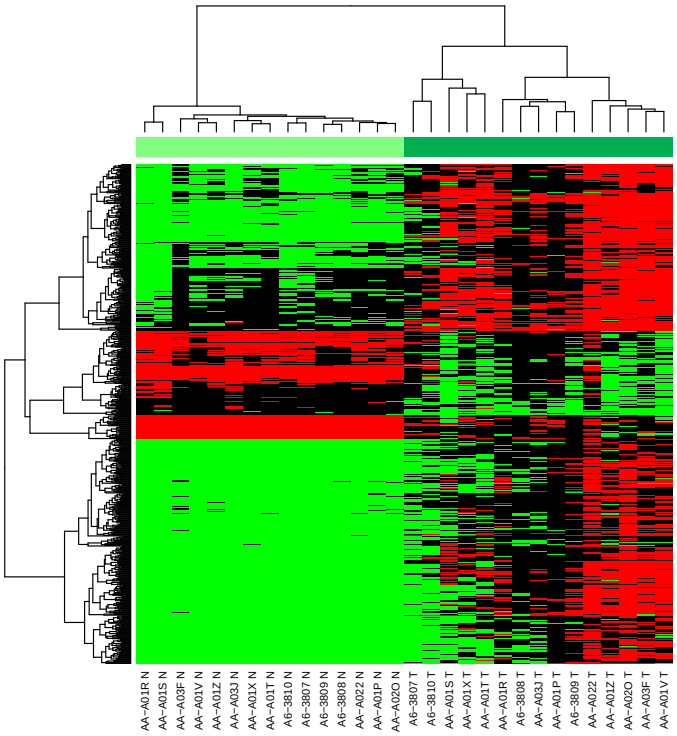
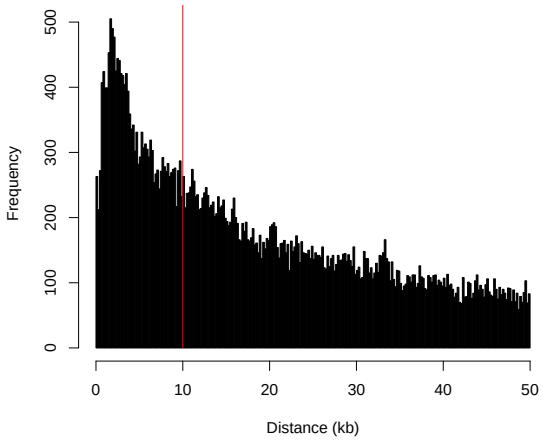
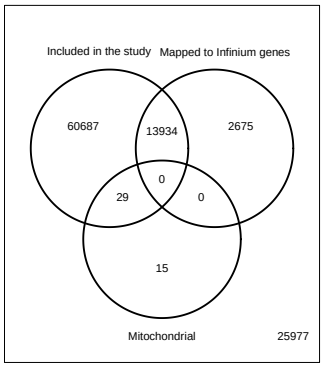


Figure 6.2: Heatmap of methylation degrees for selected Infinium 27k promoters. Every row denotes a gene promoter, and columns are samples. In the color palette used to represent methylation values, bright green denotes 0 (unmethylated), black denotes 0.5 and bright red – 1 (fully methylated). The heatmap includes measurements for 161 hypomethylated and 583 hypermethylated gene promoters in colon cancer. Column labels and patient identifiers appended with one letter that encodes the sample type, N for normal and T for tumor. The sample type is also encoded by a color bar on the columns. Hierarchical clustering is performed using Manhattan distance and complete linkage.



(a)



(b)

Figure 6.3: (a) Histogram of distances from Agilent transcripts to their closest transcription start sites of an Infinium gene. Only distances up to 50 Kb are shown. The red vertical line marks a threshold of 10 Kb, beyond which Agilent transcripts are not considered for associating with a gene. (b) Venn diagram showing number of Agilent transcripts included in the study and associated to Infinium genes.

were associated to Infinium genes. The Venn diagram in Figure 6.5(b) provides a summary on how many transcripts were associated and were used in the analyses that follow. Transcripts associated to Infinium genes tend to have a slightly higher CpG density (Figure 6.4); it is, however, significantly lower compared to corresponding distributions for hypo- and hypermethylated peaks (see Figure 5.13 in the chapter on differential methylation).

As suggested in Figure 6.5(b), 13,934 transcripts provide relative expression values for 5,547 of the Infinium genes. A substantial part of these transcripts (10,387) also overlap with the peaks identified in our MethylCap-seq study on colon cancer. Due to the discrepancy in lengths (data not shown), a transcript might overlap with more than one peak. Although the majority of 10 thousand Agilent transcripts map to a single gene, there are individual cases of transcripts overlapping with up to 29 genes. Similarly, a given Infinium gene is associated with between 0 and 12 transcripts. Later in this section, we use the mean expression values of all transcripts (more precisely, probes that map to the transcripts) associated with a gene to estimate gene's expression[2]. Level 2 (probe-level) mRNA expression values of 155 colon cancer samples were downloaded from TCGA.

6.2.3 Comparison to MethylCap-seq results

We focused on the gene level, and checked how many of the Infinium promoters were identified as differentially methylated based on the β values obtained from TCGA. The exact procedure is described earlier in this section. Note that many of the Infinium genes also

[2] Alternatively, we experimented using the maximum expression instead of the mean value. This strategy does not significantly change the results and conclusions in this chapter.

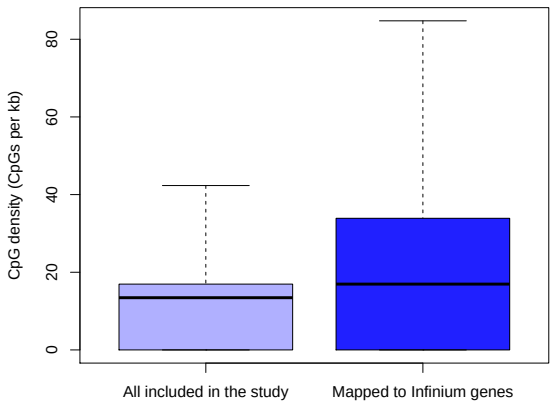


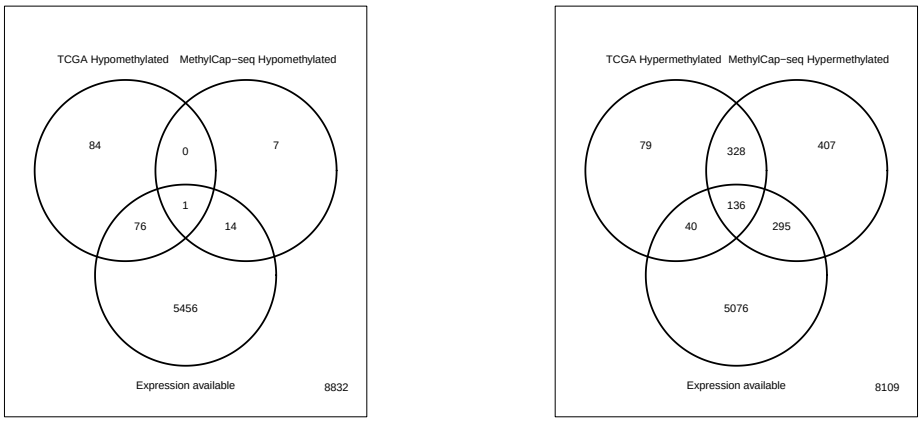
Figure 6.4: Distributions of CpG densities of Agilent transcripts.

overlap with one or more peaks. We consider a promoter to be hypomethylated in the colon cancer dataset if it overlaps with at least one hypomethylated peak. Similarly, a promoter that overlaps with hypermethylated peak is considered [MethylCap-seq] hypermethylated. It is important to note that none of the examined promoters was found to overlap with both a hypo- and a hypermethylated peak. Figure 6.5 shows the classification of Infinium promoters based on the criteria for differential methylation described here.

Figure 6.5 shows that, even if we ignore the TCGA-based definition of differential methylation, the overlap between the Infinium genes for which mRNA expression is available, and which can be classified as hypo- or hypermethylated in the colon cancer dataset, is relatively low. We can, however, use the methylation increase and expression values as a validation tool for the hypo- and hypermethylation support defined in Chapter 5.

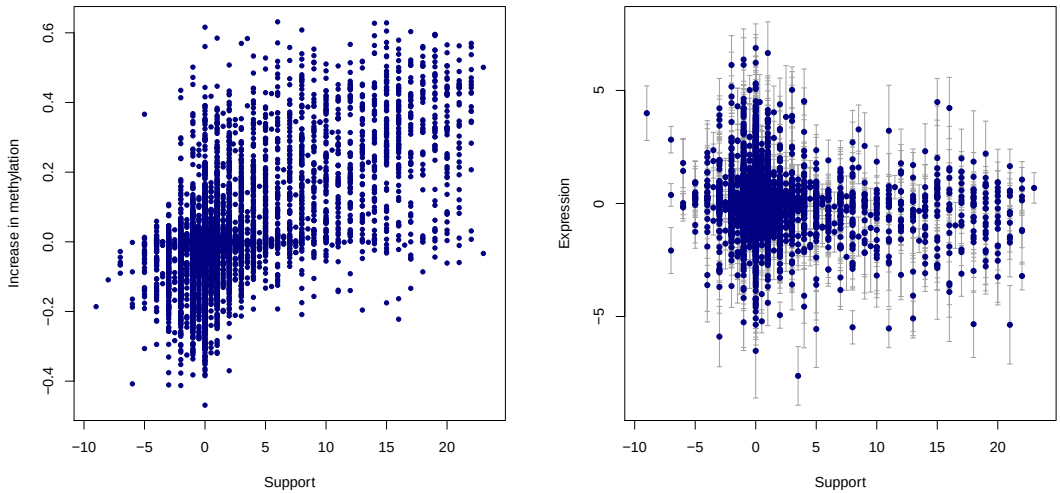
We define the support for hypomethylation of an Infinium promoter to be the average hypomethylation support of the peaks that overlap with this promoter. The support for hypermethylation of a promoter is defined in an analogous fashion. The *support for Infinium promoter* G is the difference between the support for hyper- and support for hypomethylation of G. Note that, unlike peaks, promoters do not necessarily have integer values for support. Figure 6.6 shows the correlation between support and TCGA-based measurements of Infinium genes. The correlation coefficients are provided in Supplementary Table S5.

Not surprisingly, increase in methylation shows a very strong positive correlation with support. Note that support reflects the frequency of an event of differential methylation, and mean increase in methylation is the average change in the methylation degree in observed tumor. The agreement between two colon cancer datasets using different technologies suggests that the concept of DMR's support introduced in the chapter on differential methylation can be transferred to unseen data. There is also a negative correlation between support and average expression but the relationship is less clear due to outliers, i.e. the cases in which there is a single gene with a given support, the observed expression deviates



(a) Hypomethylation (b) Hypermethylation

Figure 6.5: Venn diagram of differentially methylated Infinium promoters in the colon cancer dataset and in TCGA.



(a) Increase in methylation (b) Mean expression

Figure 6.6: Scatter/point-and-whisker plot showing the relationship between support (x axis) and a TCGA score (y axis) of Infinium genes. Blue points indicate mean score; standard deviation is visualized by gray whiskers.

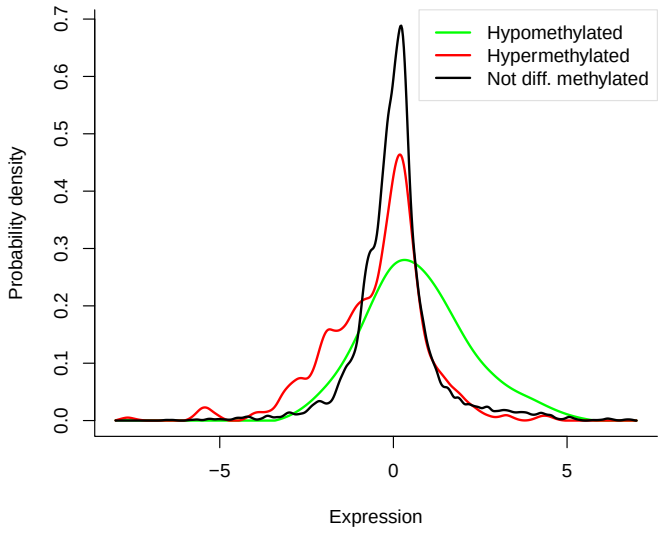


Figure 6.7: Approximate densities for expression of genes with hypomethylated, hypermethylated and not differentially methylated Infinium promoters. The curves shown are absolute values of smoothing splines fitted on histograms with equispaced bins of width 0.1.

from the trend. Moreover, the mean expression values fall almost exclusively in a short range around 0, which is consistent with other epigenetic studies in the sense that changes in promoter methylation do not always have a strong impact on mRNA expression levels.

We used the lists of identified differentially methylated regions in our study and compared their increase in methylation and expression values (at the regions, for which these measurements are available). Estimated probability density functions are shown in Figure 6.7. These analyses clearly show the low correlation between promoter methylation and gene's expression indicating that many other factors influence the level of transcription. Although we observe a shift between expression distributions of genes with hypo- or hypermethylated promoters, the vast majority of these genes have expression levels that are low and indistinguishable from the remaining transcripts in the genome.

6.3 Integration of MethylCap-seq data with ENCODE

The ENCyclopedia Of DNA Elements (ENCODE) was introduced in Chapter 2 as an ongoing program funded by the National Human Genome Research Institute that aims at identifying all functional elements in the human genome [32]. Being well into its production phase, this consortium provides the data from a variety of high quality epigenome-wide annotation studies. The table below lists all ChIP-seq datasets that were downloaded for the purpose of the analyses presented in this section.

Dataset name	Description
H1-H3K4me3	H3K4me3 levels in the H1 embryonic stem cell line
H1-H3K27me3	H3K27me3 levels in the H1 embryonic stem cell line
H1-H3K36me3	H3K36me3 levels in the H1 embryonic stem cell line
H1-DNase	Deoxyribonuclease I (DNase I)-binding regions in the H1 embryonic stem cell line
H1-Pol2	Polymerase II-binding regions in the H1 embryonic stem cell line
H1-CTCF	CTCF-binding regions in the H1 embryonic stem cell line
HCT116-H3K4me3	H3K4me3 levels in the HCT116 colon cancer cell line
HCT116-H3K27me3	H3K27me3 levels in the HCT116 colon cancer cell line
HCT116-DNase	Deoxyribonuclease I (DNase I)-binding regions in the HCT116 colon cancer cell line
HCT116-Pol2	Polymerase II-binding regions in the HCT116 colon cancer cell line
HCT116-CTCF	CTCF-binding regions in the HCT116 colon cancer cell line

We downloaded the reads for the datasets listed above and performed batch coordinate conversion (liftOver) to HG18 whenever necessary[3]. We then counted the number of fragments that overlap with each of the peaks called in our MethylCap-seq study. In addition to tag count, we calculated the normalized scores tag occupancy and tag density for each of the ENCODE element datasets.

Properties of the DNA elements

We calculated the correlations between each pair of elements, using the raw and normalized scores. As expected, we observed high correlations for the scores of identical elements in different cell types; an exception being H3K27me3. Furthermore, H3K4me3, DNase and Pol2 show strong pairwise correlations in ES cells. This is also observed in colon cancer, albeit to a lesser extent. DNase and CTCF tag counts are exceptionally highly correlated in H1, however, this relationship diminishes after applying score normalization. For each ENCODE dataset, we computed the correlations between the score of a peak and its CpG density. The observed correlations are presented in Figure 6.8. If we focus on the tag occupancy measure, we can see that only H3K4me3, H3K27me3 and DNase in embryonic stem cells are strongly (positively) correlated with CpG density.

DNA elements and methylation

We also computed the correlations between the scores for methylation and the DNA elements for each sample. We found epigenetic modifications and other functional genomic elements that are not related to methylation in healthy cells but are strongly correlated to methylation in cancer. The datasets that exhibit this property are H1-H3K4me3, H1-H3K27me3, H1-DNase and H1-Pol2. A similar trend, but somewhat more obscure, can be seen for H1-CTCF, HCT116-DNase and HCT116-CTCF. In the last step of this analysis, we checked for correlation between support for hypo- or hypermethylation and the respec-

[3]This work has been performed by Arjen Brinkman.

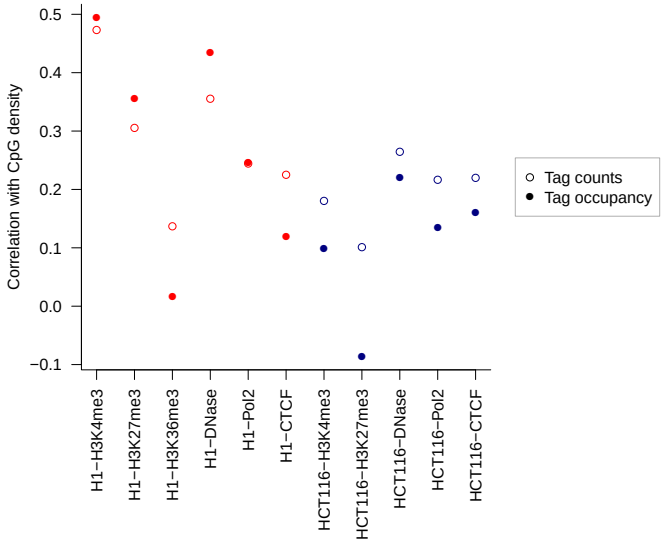


Figure 6.8: Scatter plot showing the correlations between scores for ENCODE elements and CpG density of the peaks. The embryonic stem cell line H1 is shown in red, and the colon cancer cell ine HCT116 - in blue. Correlations of tag counts are depicted by empty circles; filled circles denote correlations of tag occupancy scores.

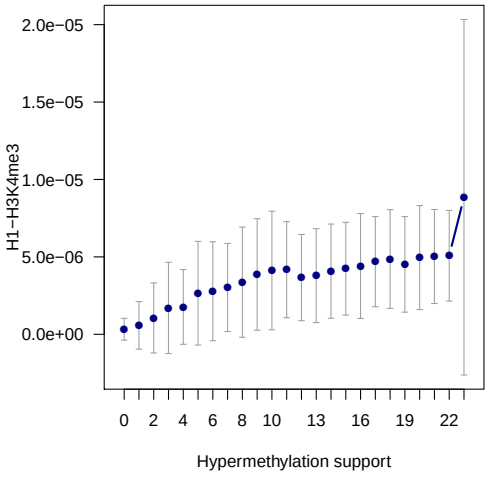
tive scores for DNA elements obtained from ENCODE. H1-H3K4me3, H1-H3K27me3 and H1-DNase showed strong positive correlation with support for hypermethylation. This relationship is visualized in Figure 6.9. Hypomethylation events did not show association with any of the ENCODE scores.

6.4 Prioritization of epigenetic biomarkers

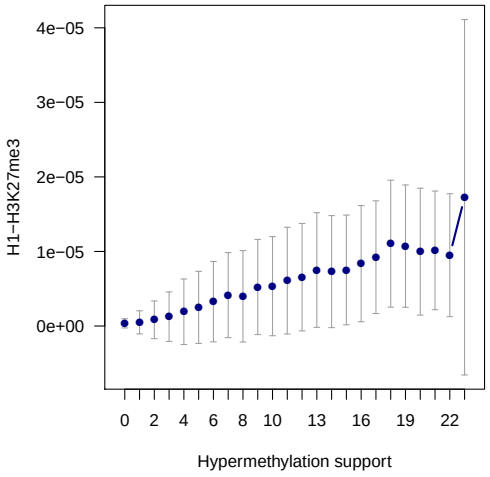
Having identified lists of differentially methylated regions (DMRs), we devised a strategy to order them based on their potential to be elements that are informative of tumor progression. We used several characteristics of the elements to estimate their potential, i.e. their functional relevance for colon cancer. In addition to support (as described in Chapter 5), we considered CpG density, the scores for H3K4me3, H3K27me3 and DNase I hypersensitivity (obtained from ENCODE and described in the previous section), as well as relative expression of overlapping mRNA transcripts (obtained from The Cancer Genome Atlas project and described earlier in this chapter).

Since the results for hypomethylation in cancer are less consistent across different scoring strategies, we focused on hypermethylated regions only. Note that expression values are available for a fraction of the DMRs. Table 6.1 summarizes the number of regions prioritized in this study.

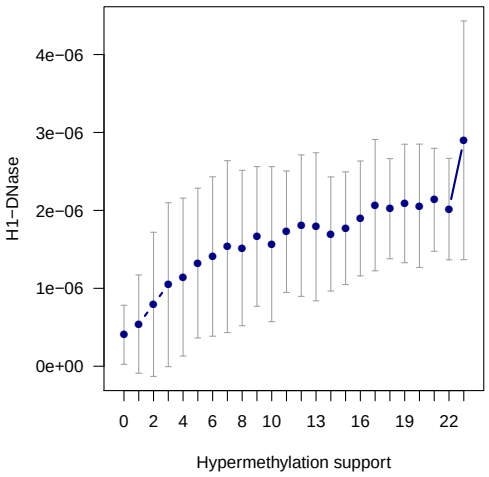
The procedure applied for biomarker prioritization uses the scores described in Chapter 3 and it involves two simple techniques – rank transformation and aggregation. The first



(a) H1-H3K4me3



(b) H1-H3K27me3



(c) H1-DNase

Figure 6.9: Correlation between support for hypermethylation and selected genome-wide ChIP-seq datasets downloaded from ENCODE. Mean values of CpG density are depicted by blue points; whiskers denote one standard deviation.

Score \ List	**Full list**	**Expression available**
Tag occupancy	2,634	565
Tag density	2,684	552

Table 6.1: Number of DMRs used as input in the prioritization scheme.

manipulation transforms any sequence of N scores into ranks, e.g. numbers from 1 to N if the original scores are all unique. Duplicated scores are assigned identical ranks – the average value of the rank range they span. The second manipulation involves combining two or more lines of evidence into an aggregated score. In case all indications are to be considered, we combine the scores using the average. If every single indication has the potential to characterize functional relevance, we consider the best score/rank as the aggregated score. As a side note, the same techniques for rank transformation and aggregation were later successfully applied in RnBeads for the prioritization of differentially methylated regions.

It is important to note that rank transformation and aggregation are general techniques, applied for combining different lines of evidence for functional relevance of a DMR. The steps described below, however, are not universally applicable to all projects involving biomarker prioritization. Rather, these are the heuristic rules applied in the colon cancer project, taking into consideration the properties of the MethylCap-seq technology, prior knowledge about cancer-specific events, as well as (lack of) availability of independent cohorts. Figure 6.10 illustrates the two pipelines for prioritization using a toy example of four fictional regions, named A, B, C and D. The input is a table listing all identified high-confidence DMRs as rows, and their collected properties as columns. The first algorithm (the left branch in Figure 6.10) ignores the expression scores assigned to the regions, and considers the five remaining characteristics. The second algorithm (the right branch in Figure 6.10) focuses on regions with known expression values, effectively shrinking the list of candidates to a fraction of its original size. The steps described in the figure are as follows:

Step	*Pipeline 1*	*Pipeline 2*
Step I	Ignoring the column Expression, followed by rank transformation. Higher numbers are given better (smaller) ranks.	Focusing only on regions with available expression values, followed by rank transformation. Higher numbers are given better (smaller) ranks.
Step II	Rank combination of all columns except Support. The best (smallest) rank is selected for every region.	
Step III	Rank combination of the columns Support and Other evidence. The average rank is computed for every region.	
Step IV	Rank transformation of the aggregated score. Lower numbers are given better (smaller) ranks.	

6.4.1 Results

Partial results from this prioritization strategy, more precisely, the top 20 hypermethylated peaks, are listed in Supplementary Table S7. Regions are sorted based on their aggregated score.

We examined the relative importance of the different indicators in the final ranking of hypermethylated regions. By studying the mean rank of the top K candidates (see Figure 6.11), it is easy to notice that support is the dominant factor in the candidate

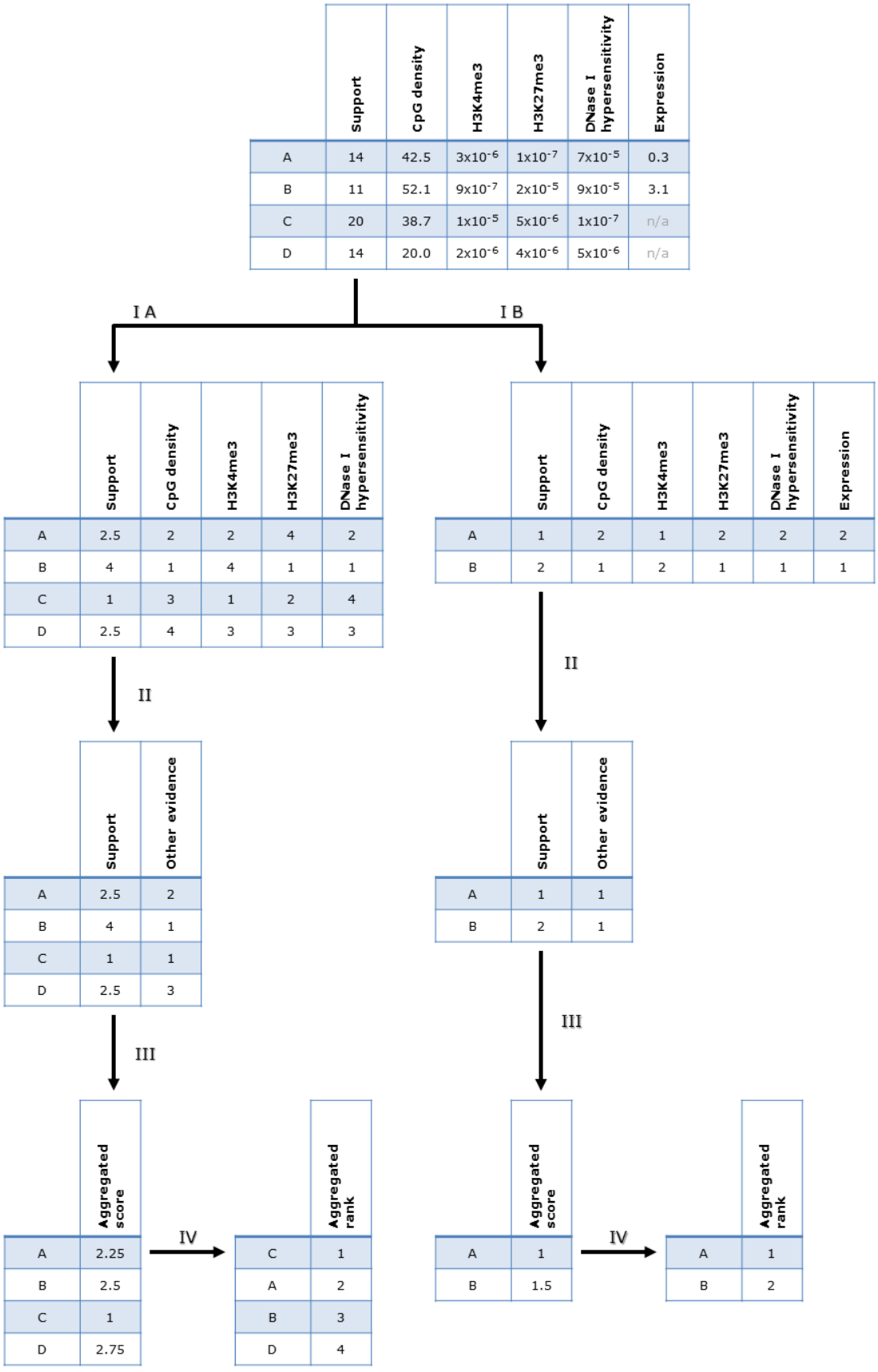


Figure 6.10: Two strategies for DMR candidate biomarker prioritization applied on a toy example set consisting of four regions: A, B, C and D. Each strategy consists of applying four consecutive steps (transformations on the table of candidates) denoted by Roman numerals; its intermediate and final results are displayed in a dedicated branch.

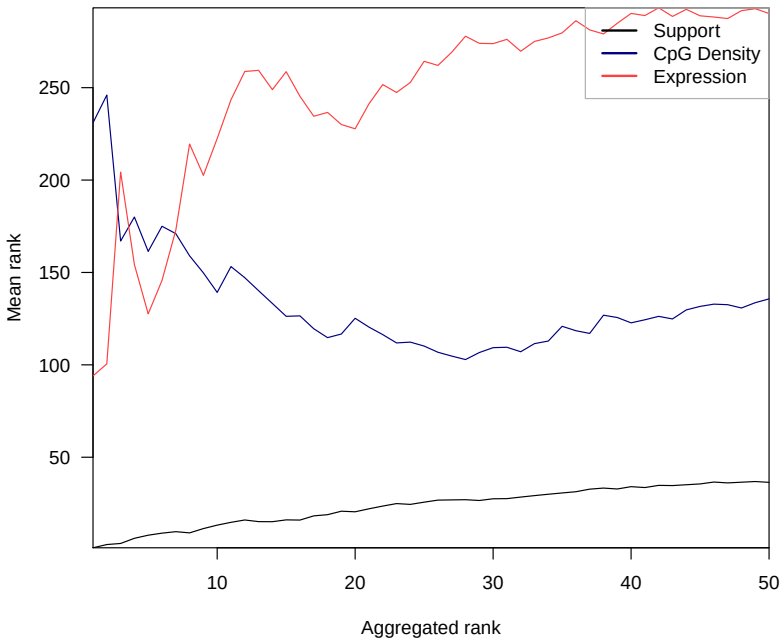


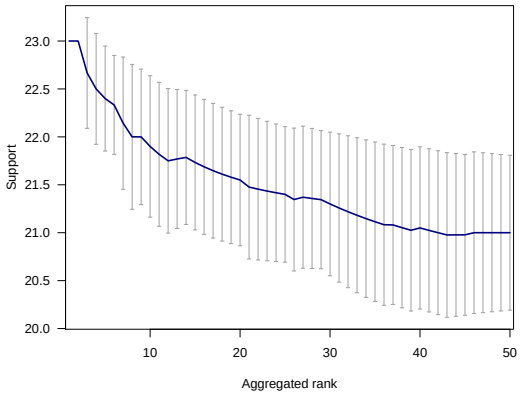
Figure 6.11: Mean rank of selected indications among the top K biomarker candidates. The horizontal axis lists the tested values for K. Indications are denoted by colors.

prioritization. CpG density and expression level seem to be more influential than histone marks and DNase I hypersensitivity (data not shown).

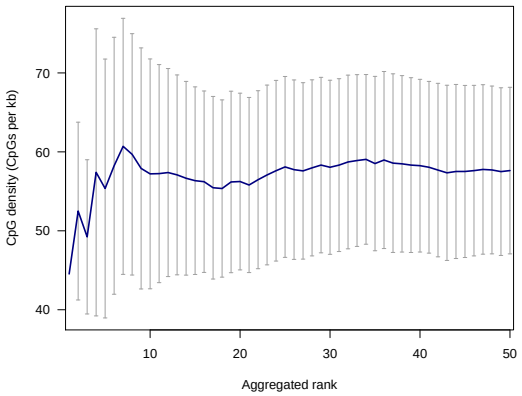
We also tested for bias in the ranking with respect to several factors, including, among others, region length and CpG density. We computed the mean value of the studied property for the top K biomarker candidates, and inspected the change in the mean value as K varies from 1 to 50. Selected results are shown in Figure 6.12. Surprisingly, only support shows a strong correlation with the aggregated rank, indicating that the other factors have a subtle effect on the prioritization results. Keep in mind that these factors include CpG density and expression, that are used in calculating the final ranks.

6.5 Results and discussion

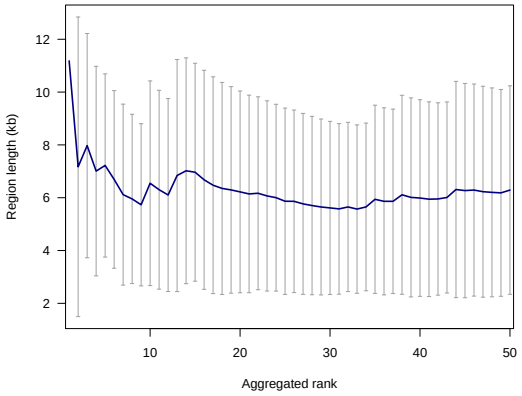
In this Chapter, we proposed methods for integrating datasets of the same type from heterogeneous sources, and of different data types. We take an approach focusing on predefined regions – gene promoters and gene bodies in this case – that enables us to normalize data to a common baseline when the outcomes of different technologies need to be combined. Despite the low overlap in targeted regions between MethylCap-seq and microarray data, we successfully used methylation and expression microarray studies to confirm our findings and thereby gain confidence in the power of the support for differential methylation introduced in Chapter 5. When closer technologies are integrated, as is the case with histone marks from ENCODE, we applied the same steps for data preprocessing and normalization



(a) Support



(b) CpG density



(c) Region length

Figure 6.12: Distributions of selected scores in the top K biomarker candidates. K varies from 1 to 50, as depicted in the horizontal axes. Score distributions are visualized by line-and-whisker plot, where the line shows variation in the mean value and whiskers measure standard deviations.

which permitted more rigorous comparison and identification of strongly correlated signals – a necessary (but not sufficient) condition for association.

Finally, we propose a simple and flexible prioritization strategy that allows for ranking candidate differentially methylated regions based on their association with another metric (e.g. expression) and/or evidence observed in an independent dataset.

7 Methylation-based cancer type classification

7.1 Epigenetic signatures and fingerprints

The analyses in this chapter are partially inspired by the ongoing scientific debate on the phenomenon of CpG Island Methylator Phenotype (CIMP). This term was first coined by Toyota at el., based on the interrogation of 30 newly cloned differentially methylated DNA sequences in 43 colorectal cancer samples [112]. The authors conjecture that the CGI hypermethylation events observed in the majority of tumor samples occur at loci that become progressively methylated with age in healthy tissue. In contrast, the CIMP positive tumors are a subset of carcinomas and adenomas in which tumor suppressor gene hypermethylation has resulted in damaged DNA repair mechanisms. For this reason, the cases of CIMP positive tumors are often associated with microsatellite instability (MSI) but form a clinically defined subtype. Four years later, Yamashita and colleagues challenged the existence of a CIMP phenotype when examining the methylation state of 6 CGIs at tumor suppressor genes and 30 other locations in a cohort of over 200 colorectal cancer patients [125]. Their results indicate that tumor-specific somatic hypermethylation is a widespread age-dependent process that follows a Gaussian distribution. Moreover, MSI is a better indicator of the observed phenotype (e.g. relapse-free survival) in colon cancer. In the past years, several publications appeared with comprehensive studies that support the hypothesis for a prognostic value of CIMP [95, 85, 121], of MSI over epigenetic changes [9, 61], and also showing lack of support for both claims [8].

It is important to note that essentially every publication characterizing CIMP uses a distinct panel of islands or genes, along with a specific set of rules that define the methylator phenotype. Moreover, it becomes progressively clearer that CIMP is not a unifying phenotype but is rather specific for different tumor types. Hypermethylation events that are associated with gene mutations and patient survival have been described in at least 8 cancers [60], including, among others, glioblastoma (termed G-CIMP) [84] and lung adenocarcinoma [103]. Hughes et al. provide a systematic review of the publications on CIMP-related studies in colorectal cancer, and conclude that the debate surrounding this issue "will likely continue until a biological cause for CIMP has been determined" [60]. Identifying associations between epigenetic genes and methylation patterns seems to be a promising direction for future research [36].

7.2 Tumor types and subtypes

Much like the discussion on CIMP above, this Chapter presents three approaches taken to identify epigenetic patterns indicative of a tumor type, subtype or a similar clinically relevant phenotype based on methylation data. The first section focuses on using methylation data to predict tumor of origin. This analysis was performed on the GoldenGate dataset, in the context of the CANCERDIP consortium, and the resulting model is one of the major contributions of the associated publication [50].

The second section discusses bimodality in the context of identifying two subgroups in the lung cancer dataset. As mentioned in Chapter 2, the focus of the study on the lung cancer dataset was finding a prognostic panel for relapse free survival. The discussions presented here were not included in the publication, as they address the more general question of whether lung adenocarcinoma or squamous cell carcinoma can be stratified into methylation-specific subtypes.

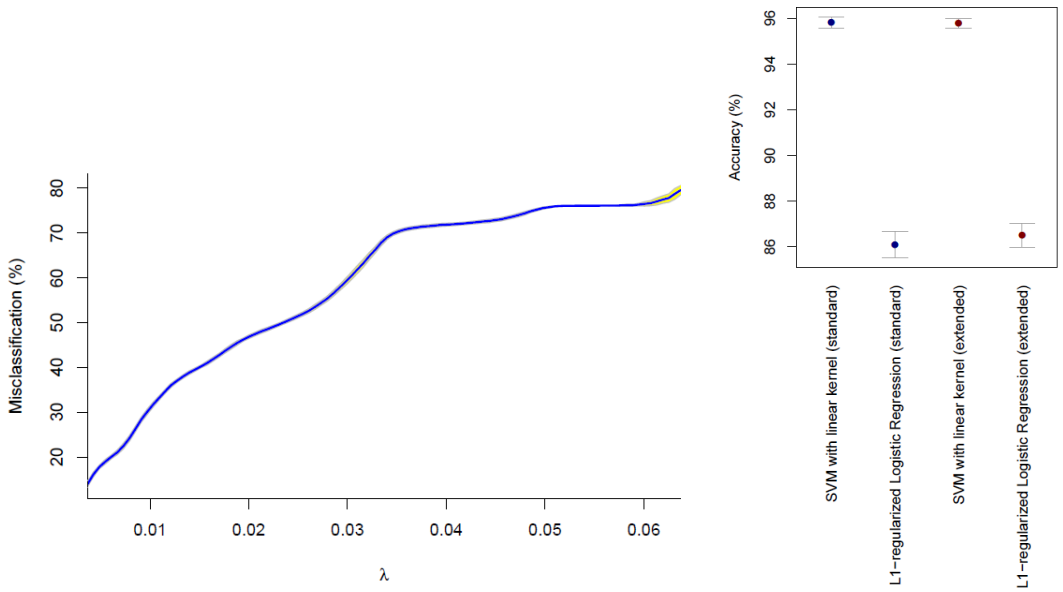
The last section presents an analysis of the colon cancer dataset, another collaboration conducted within the CANCERDIP consortium. It studies the stability of established model families trained on high-dimensional data. We did not find strong associations between differential methylation events and MSI or clinical properties of the tumor samples [104]. Results on the existence of methylation-specific subtype (such as CIMP) are inconclusive due to the very limited number of patients.

7.2.1 Predicting primary origin

The GoldenGate methylation dataset contains 42[1] carcinomas of unknown primary origin (CUPs). These are metastatic samples for which the primary tumor type is not known. The large set of available primary tumors can be used to train a classifier. We considered only the cancer types represented by at least 10 primary samples. There are 827 samples from 24 tumor types that meet this criterion. We trained linear support vector machines (SVMs) and L_1-regularized logistic regression (RLR) models on the primary tumor samples. Given an unseen sample, both methods are able to produce probabilities for the sample to belong to each of the 24 cancer types. We then inspected the performance of the methods on classifying metastatic samples (for which the primary origin is known), before using them to predict the origin of the CUP samples.

Linear support vector machines were trained using the "one versus all" approach. More precisely, 24 classifiers were trained, each of them labeling a unique cancer type as case and all other types as control. In classifying a new data point, the predicted case probabilities of each model were extracted and rescaled to sum up to 1. Note that the dimensionality of the feature space exceeds the number of training points. Also, the sample groups of tumor type tend to be clearly separated. For this reason, the SVM classifier produced identical solutions for the whole range of tested values for the cost parameter C (1 to 10). However, the selection for the value of the parameter λ in L_1-regularized logistic regression models is

[1] The answer to the ultimate question of life, the universe and everything.



(a) (b)

Figure 7.1: 10-fold cross-validation (CV) estimates of the misclassification error rate, used in estimating the λ parameter of L_1-regularized logistic regression model (a). Comparison of the CV estimates for the accuracy of SVM and RLR (b).

non-trivial. For this purpose, 10-fold cross validation (CV) was performed on the training set of 827 samples, and test error was estimated for a sequence of one hundred positive values for λ. In order to test the stability of the CV estimates, this randomized process was repeated 50 times. For the sake of error comparison, SVM classifiers were trained on the same folds used in training logistic regression models.

Each of the methods described above was applied to the samples using the 1,322 probes that pass all filtering criteria. In addition, we applied the methods using an extended set of 1,366 probes, including 44 probes that are hypermethylated in females. We refer to the smaller and larger probe sets – and consequently, to the models using them – as *standard* and *extended*, respectively.

Parameter selection and training error

Figure 7.1(a) shows the estimated misclassification error for the logistic regression models. The selected λ in both cases is the lowest value in the sequence: approximately 0.0037. The number of probes with nonzero coefficients are 196 and 198 for the standard and extended models, respectively. In fact, the addition of 44 gender-specific probes does not affect the prediction accuracies of the models (see Figure 7.1(b)). SVM models clearly outperform RLR in this scenario. This observation was later confirmed for Infinium 27k data as well [131].

Note that the training set of samples is highly imbalanced. The best represented tumor

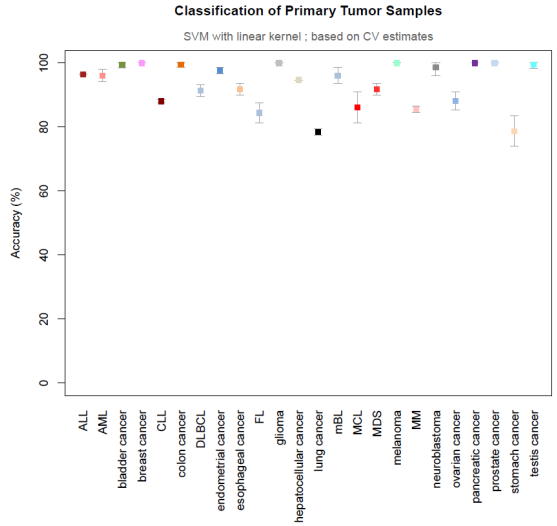


Figure 7.2: Cross-validation estimates of the accuracy of an SVM model with linear kernel, reported separately for each of its possible outcomes.

types are colon cancer and glioma, with 110 and 97 samples, respectively. For some cancer types, such as follicular lymphoma and multiple myeloma, only 14 samples are available. We decided to test if and to what extend the trained models could show bias toward the numerously represented tumors. As a first step, we segregated the errors per tissue type. Figure 7.2 shows the estimated errors of the full models trained on the complete set of 827 primary tumor samples. The correlations between mean accuracy and sample size per tumor type are positive ($r^2 = 0.2$ for SVM and 0.31 for RLR), which is to be expected, but the values are not large enough to suggest a strong bias towards better represented cancer types.

Classification of metastatic samples

Note that all calculations in the previous sections are based on the primary tumor samples only. We used the trained models to predict the origin of metastatic samples. The Golden-Gate dataset contains 50 metastatic samples, including 32 colon cancer metastases in liver, 13 colon cancer metastases in brain and 5 renal tumor metastases in brain. Our training set of primary tumors includes colon cancer but not kidney cancer. Therefore, a perfect model would predict very high probability for colon as a primary origin of the colon cancer metastatic samples, and would estimate low probabilities for all known tumor types as an origin of the kidney cancer metastases. The results of all methods are available in Supplementary Table S8, and the prediction procedure is described in the following paragraph.

We used a probability threshold P for predicted origin. More precisely, tissue of origin T was predicted for a metastatic sample if and only if the following conditions are met: (1) the tumor type that corresponds to T received the highest probability $Pr(T)$ among all tumor types for the sample, and (2) $Pr(T) \geq P$. In case all the tumor types are predicted

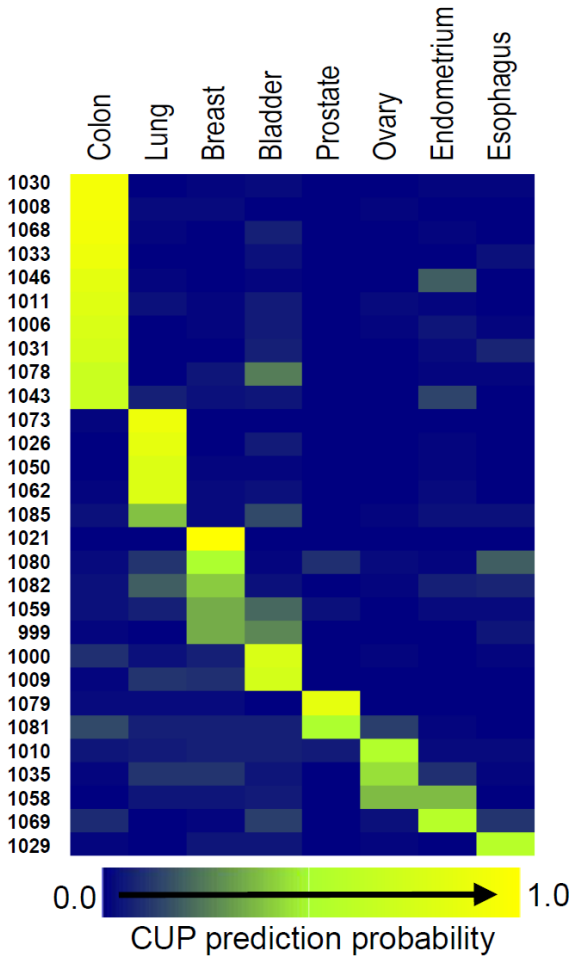


Figure 7.3: Heatmap visualizing predicted probabilities of origin (columns) for the CUP samples (rows) in the GoldenGate dataset. Only tumor types that have a highest probability for being the origin of at least one metastatic sample are included.

as origin with probabilities lower than the threshold, we define the origin of the metastatic sample as uncertain. The probability decision threshold for the families of prediction models was set in a somewhat arbitrary fashion to 10/24 – 10 times more than random. In fact, any choice for a threshold between 0.3 and 0.55 has a negligible impact on the results and no effect on the summary presented below (see Supplementary Figure S1).

Next, we computed the accuracies of the models based the 50 metastatic samples. A classification was correct only if the predicted origin for colon cancer metastasis is colon, and for renal metastasis is uncertain. The accuracies computed using the above procedure were approximately 90% for all models. Finally, Figure 7.3 summarizes the predictions of SVM standard model on the available CUP samples. Later validation yielded accuracy of $\approx 80\%$. The results for the other tested models were almost identical (data not shown).

Of note, we recently repeated the training procedure described above on a collection over 4,600 Infinium 450k samples from 16 solid tumor types from TCGA, and obtained

comparable results. We also experimented with different machine learning techniques, such as logistic regression model with elastic net penalty. CV estimates of this model are presented in Supplementary Figure S2.

Summary

Here we compared the applicability and performance of two linear classifiers when predicting tumor type of GoldenGate samples. A large collection of primary tumors samples was used as a training set. The classifiers were then evaluated on metastatic samples. Both models show remarkable accuracy and were successfully applied for the prediction of origin of CUPs.

7.2.2 Identification of lung cancer subtypes

This section focuses on the two lung cancer types in the lung cancer datasets that are represented by large sample groups – adenocarcinoma and squamous cell carcinoma. All steps outlined below were performed for both tumor types independently. Unless stated otherwise, the figures that accompany the description depict results on the squamous cell carcinoma samples. Two approaches are presented: cluster analysis and locus analysis.

Clustering and variability analysis

We attempted to subcategorize each group into subtypes based exclusively on the methylation profiles of the samples. We achieved this task using the following strategy:

We first applied several clustering approaches that identified distinct sample subgroups. We also tested for strong correlations between the clusters and the directions of largest variance, as defined by principal component analysis on the high-dimensional methylation data. Once the subgroups are defined, we inspected each CpG dinucleotide individually and found the ones whose methylation state is informative of a particular subgroup.

Locus analysis

We also identified CpGs and promoters that show bimodal behaviour in each of the studied tumor subtypes. More precisely, we searched for loci that (1) are unmethylated or methylated in normal lung tissue, and (2) show bimodal distribution of their β values in the tumor of interest.

This analysis is based on minfi processing with control normalization and background substraction. Probe and sample filtering was done as described in Chapter 3. The following sections describe in details the cluster analysis steps introduced here. Justification for the selected methods is provided, as well as their underlying assumptions and the results. The later sections describe the simulations and the results of the locus analysis introduced above.

Clustering

We applied five clustering techniques on the transformed data of a sample group – agglomerative hierarchical clustering using three different linkage methods, spectral clustering

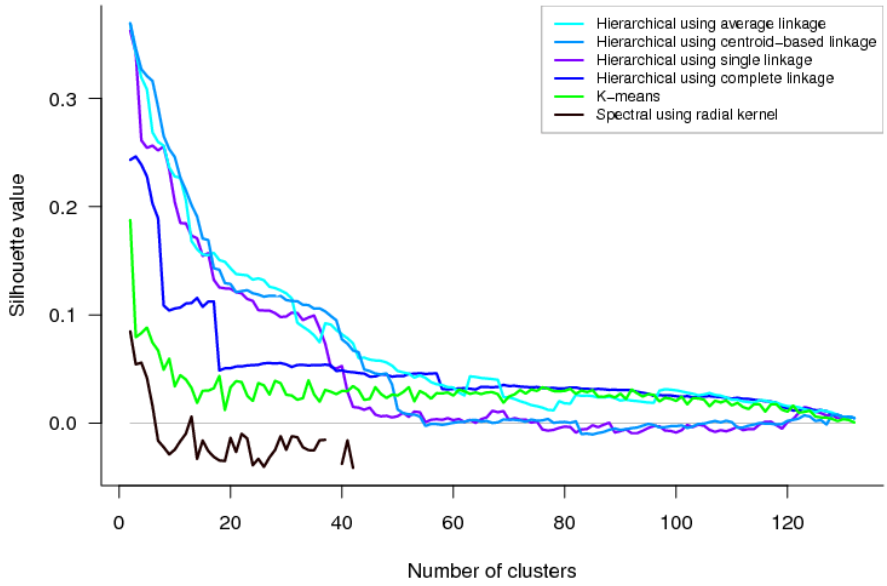


Figure 7.4: Line plot of the silhouette values of the clustering algorithm outcomes for each applicable value of K (number of clusters) in squamous cell carcinoma samples.

with radial basis (RBF) kernel, as well as k-means clustering. In all approaches, the number of clusters K that produced the best sample separation was determined by calculating a silhouette value for each possible outcome. The resulting silhouette values were markedly low in all cases. Figure 7.4 visualizes them for the squamous cell carcinoma samples; the adenocarcinoma set produced a very similar outcome. All algorithms showed very consistent results with respect to number of clusters K, but not when it comes to cluster assignments for individual samples.

Hierarchical clustering using average, centroid-based or single linkage method cannot separate the tumor samples into subtypes in a convincing manner because one of the subtypes is usually represented by a single sample. Hierarchical clustering with complete linkage, k-means and spectral clustering with radial basis kernel, on the other hand, do not produce compact and dense clusters, as indicated by the comparatively low silhouette values. Moreover, the methods are inconsistent with respect to the subgroups identified, as already noted above.

Clusters and variability

We performed principal component analysis on the same β value matrix that was used for clustering. This is the matrix constructed from the methylation degrees of the samples in the group of interest at the $418,612$ probes. In addition to dimensionality reduction, this analysis also showed us the main directions and the span of the variability of the studied tumor samples in the high-dimensional Infinium 450k probe space. We conjecture that every identifiable subtype based on methylation patterns is represented by a point cloud within our dataset. Moreover, this point cloud occupies distinct value ranges in one or more of the principal components of the analyzed type.

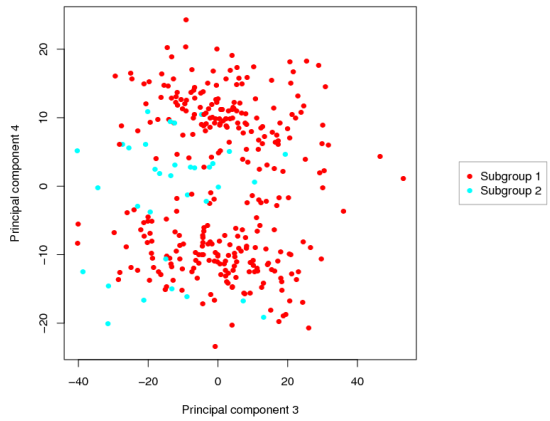


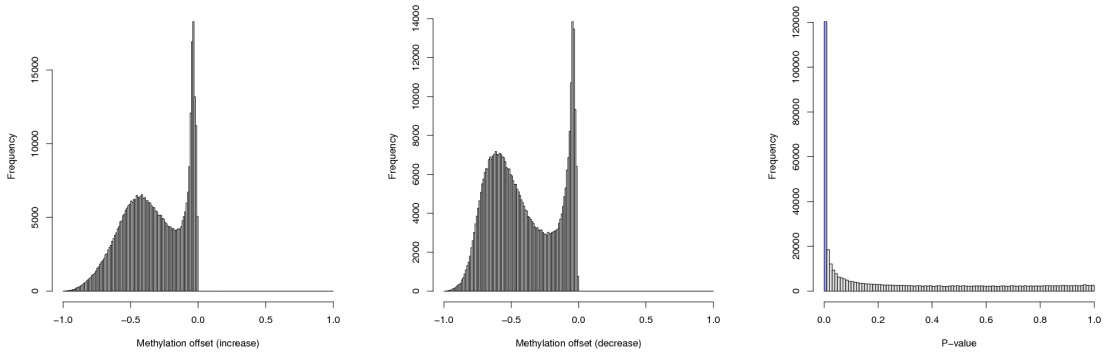
Figure 7.5: Scatter plot of all samples in the adenocarcinoma group in the third and fourth principal components. The color of a point depicts its cluster membership as determined by hierarchical clustering using complete linkage as an agglomeration method.

We inspected the first eight principal components. Cumulatively, they explain over 50% of the total variance in each of the tumor types. Judging by the spread of the samples, no distinct division into point clouds appear. The only hint of separation into groups can be seen by visualizing the fourth principal component coordinates (see Figure 7.5), however, this representation does not seem to reflect true intersample distances. Overall, no strong correlations could be identified between sample values at any pair of a principal component (among the first eight) and a clustering outcome. The apparent lack of dense coherent clusters suggests that, when considering the genome-wide methylation patterns, no distinct subtypes can be identified.

Locus analysis

Using the clustering algorithms described above, we defined distinct subgroups in each of the studied sample groups. In order to identify CpGs that are informative of a subgroup, we used an approach similar to the one described in Chapter 5. More precisely, we compared the samples in subgroup i ($i \in 1, \ldots, K$) to all other samples in the group, and checked for each probe separately if it is differentially methylated between the two sample sets. Differential methylation was quantified by four measures – offset, difference in means, relative difference, and a p-value. Formally, let A and B be two non-empty sets of sample indices corresponding to two different clusters identified in a tumor type, and the corresponding sample methylation values for a given probe P is $M_{P,A} = \{\beta_i\}, i \in A$ and $M_{P,B} = \{\beta_i\}, i \in B$. The measures we calculated are then defined as follows:

1. methylation offset for increase: $\min(M_{P,A}) - \max(M_{P,B})$
2. methylation offset for decrease: $\min(M_{P,B}) - \max(M_{P,A})$
3. difference in means: $\mu(M_{P,A}) - \mu(M_{P,B})$
4. relative differential methylation: $\log_2(\mu(M_{P,A})/\mu(M_{P,B}))$



(a) Offset for increase (b) Offset for decrease (c) P-value

Figure 7.6: Histogram of observed values of offset for increase (a), offset for decrease (b) and p-values after correction (c).

5. p-value for differential methylation (Student's t-test on the samples from the two sets), adjusted for multiple testing using the Benjamini-Hochberg approach.

Figure 7.6 shows example distributions of three of the measures defined above for one of two clusters identified using hierarchical clustering with complete linkage on the squamous cell carcinoma samples. Note that every probe with a positive offset is very informative of a subtype – its methylation status can be used to accurately separate the samples belonging to a specific subtype from the rest of the samples in the examined group. Unfortunately, we almost never observe probes that have positive offsets, despite the fact that the majority of the differential methylation p-values are very low. Remarkably, in all cases in which k-means or hierarchical clustering with complete linkage was applied, the distribution of methylation offsets can be well approximated by a mixture of two normal distributions. We exploited this property further and fitted a Gaussian mixture model in each of these scenarios. We defined probe P in cluster c to be *informative by high methylation* when all of the following conditions are met:

- P is more likely to belong to the Gaussian with the higher mean (that the one with the lower mean) in the mixture model of its offset for increase in methylation;
- P has a difference in means of at least 0.2;
- P has a relative differential methylation of at least 0.6;
- The adjusted p-value for P is less than 0.01.

The definition of a probe being informative by low methylation is analogous.

Note that some of these thresholds seem ad-hoc, nonetheless, justification can be provided. For example, the value of 0.6 (-0.6 for informative by low methylation) applied to the quotient corresponds to approximately 1.5-fold increase (decrease) in mean methylation between two sample sets. In general, we label a probe informative in a sample group if it is informative by high or low methylation for at least one of the identified subgroups. The table below provides summary of the number of informative probes found in the studied carcinomas.

Algorithm \Sample Group	Adenocarcinoma	Squamous cell carcinoma
Hierarchical using complete linkage	0	2
K-means	0	0
Spectral using radial kernel	0	8

Four of the eight informative probes (based on spectral clustering) in squamous cell carcinoma are associated with genes. We give a short description of three of these genes below. The fourth one is C20orf186.

- HECW2's product is E3 ubiquitin-protein ligase that mediates ubiquitination of TP73. It acts to stabilize TP73 and enhance activation of transcription by TP73 [80]. The probe `cg20197814` is located in the gene's 5' UTR.
- The protein encoded by SCGB1D1 is a member of the lipophilin subfamily, part of the uteroglobin superfamily, and is an ortholog of prostatein, the major secretory glycoprotein of the rat ventral prostate gland. The protein may bind androgens and other steroids; it may also bind estramustine – a chemotherapeutic agent used for prostate cancer [71]. The probe `cg01772980` is located near the gene's transcription start site.
- Aquaporin 8 (AQP8) is a water channel protein. Aquaporins are a family of small integral membrane proteins related to the major intrinsic protein (MIP or AQP0). Aquaporin 8 was found to be ubiquitously expressed in cervical squamous cell carcinoma [127]. The probe `cg02916147` is associated with the promoter of AQP8.

Figure 7.7 displays all methylation values of the identified informative probes in a heatmap.

Notes on bimodality

A critical aspect of this analysis is the identification of bimodality. More precisely, we need to identify loci (e.g. probes) whose β values are best described by a bimodal distribution – one mode of low methylation and one of high methylation. In contrast, a reasonable assumption for a locus with consistent methylation is a single normal distribution; its mean shows the average methylation degree within the studied population, and the deviation quantifies the observed variability. Note that Gaussian mixture is a more complex model than a single Gaussian. Therefore, a better fit for the mixture model on a limited set of values should not be blindly extrapolated. We attempted to quantify the better fit of a mixture model in our framework using simulations described in the paragraph below.

We performed simulation tests in which a specified number of values $x_1, \ldots, x_n \in [0, 1]$ were randomly drawn from a normal distribution with mean $\mu = 0.5$ and standard deviation $\sigma = 0.1, 0.15, \ldots, 0.5$. The goodness of fit of a normal distribution and Gaussian mixture model to the drawn values was quantified using the log-likelihood measure:

$$\begin{aligned} L_{EM} &= \sum log(L(x_i)) \text{ in the fitted Gaussian mixture model} \\ L_N &= \sum log(L(x_i)) \text{ in the fitted single Gaussian model} \end{aligned}$$

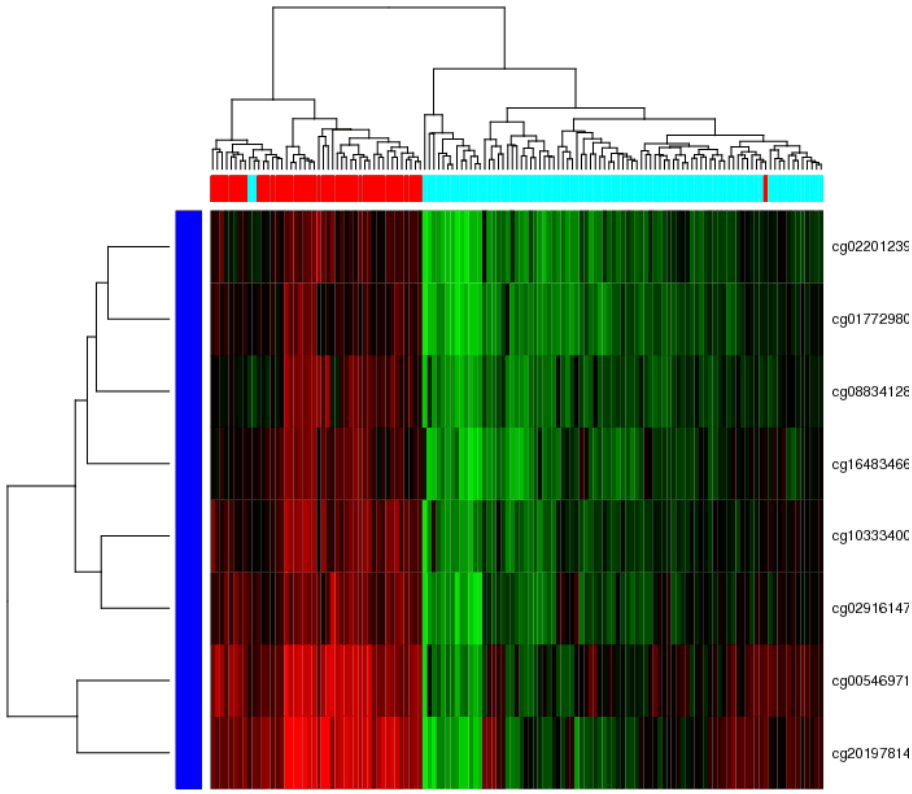


Figure 7.7: Heatmap of methylation values at informative CpGs. Every row in this heatmap corresponds to an Infinium probe, and every column – to a sample. Methylation is color-coded using a palette from bright green (no methylation), through black (50% methylation) to bright red (close to 100% methylation). Sample colors denote cluster assignment, and probe blue color legend shows that all probes lie outside CpG islands.

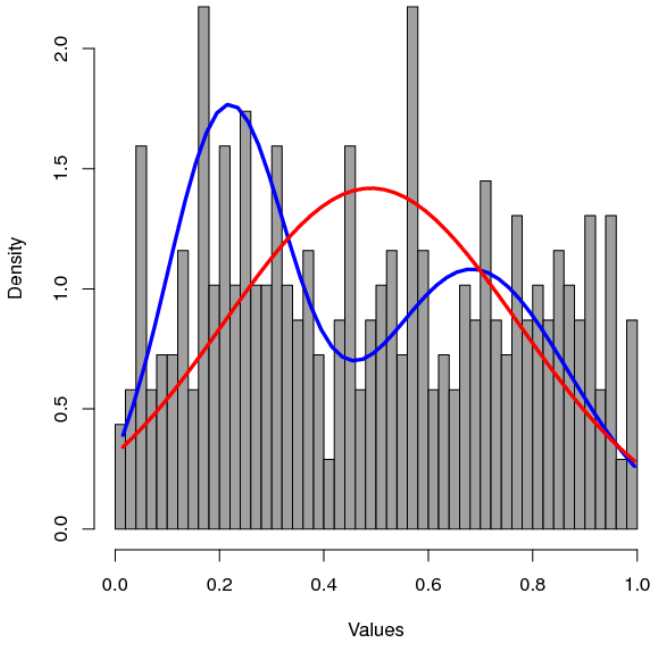


Figure 7.8: Histogram showing example simulation of methylation β values. The values were drawn from a distribution with mean 0.5. The density function of the fitted Gaussian is depicted by a red line, and the density of the Gaussian mixture – by a blue line.

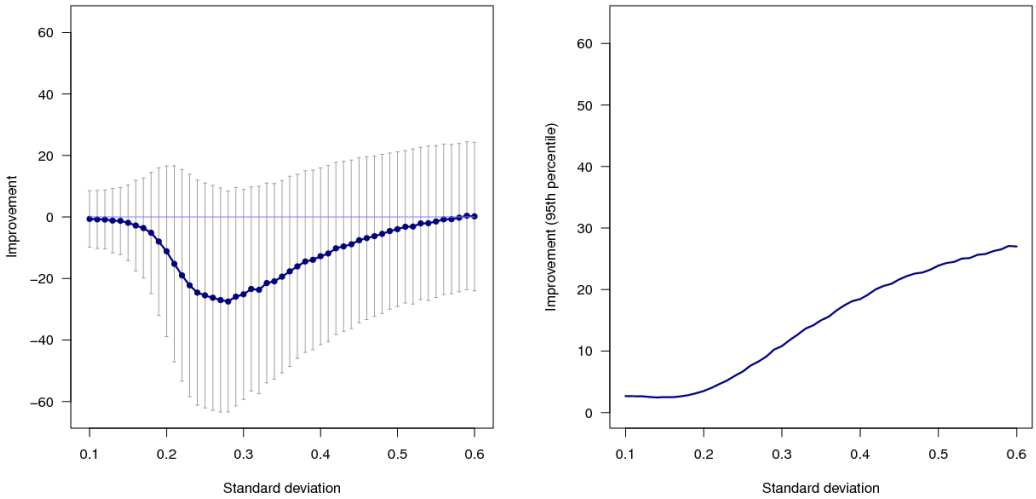
Note that log-likelihoods are negative values. Also, $L_{EM} > L_N$ indicates a better fit of the mixture model. The Gaussian mixture models were fitted using the expectation maximization algorithm. For the number of values drawn, we decided to mimic the number of samples in a cancer subtype. We therefore performed simulations using $n = 345$ (corresponding to the adeno group) and $n = 133$ (squamous group) values. Figure 7.8 shows an example in which the Gaussian mixture has a better fit than the normal distribution approximation, although the values were generated from a single Gaussian.

We performed 10,000 simulations for each pair (μ, σ) of normal distribution parameters. We defined the improvement as $L_{EM} - L_N$. Positive improvement indicates that a Gaussian mixture model fits the generated values better than a single Gaussian. Figure 7.9 shows the distributions of the resulting improvement values.

We selected as a threshold the largest among the 95th percentiles, rounded up to the closest integer value. This gives us $T = 28$ for improvement in the adeno group, and $T = 14$ for the squamous carcinoma samples. Using these thresholds, we can safely estimate that the false discovery rate is less than 5%.

Bimodal probes in lung cancer

In this concluding step, we identify probes that are unmethylated or methylated in normal tissue and show bimodal distributions in a cancer subtype. We define a probe to be



(a) Mean and standard deviation

(b) 95th percentile

Figure 7.9: Distributions of log-likelihood improvements for the simulations based on the adenocarcinoma sample group. The underlying value sets were drawn from Gaussian distributions with mean $\mu = 0.5$ and different values for standard deviation σ, shown on the x axis. (a) Blue points denote mean values of the calculated improvements, and whiskers show standard deviations. (b) The blue line depicts the 95th percentile of each distribution of improvements.

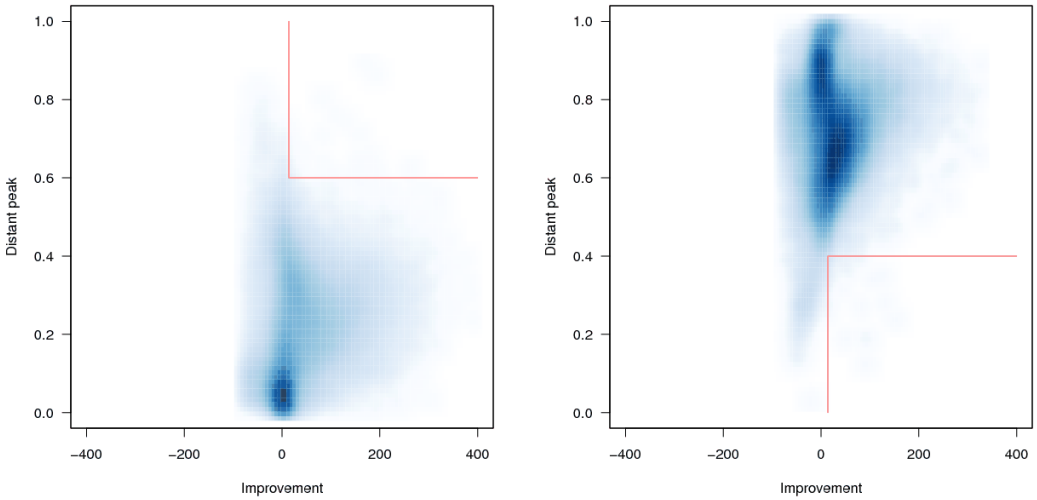
consistently unmethylated in normal lung when its β values is at most 0.4 in all samples and the mean β is at most 0.2. Similarly, a probe is consistently methylated in the group of normal samples when its methylation is at least 0.6 in all normal samples and the average methylation is at least 0.8[2]. The table below summarizes the number of consistent probes we found in the examined normal samples.

Regions	Probes	Promoters
Unmethylated	149,446	3,240
Methylated	106,624	1,600
Total	256,070	4,840

Each of the consistent loci was tested for bimodality in a group of cancer samples using the approach described earlier in this section. As already discussed, an improvement of at least T suggests that the methylation of a locus is bimodal. However, bimodality on its own does not necessarily imply heterogeneous methylation. We therefore imposed additional criterion, targeting probes and promoters that can divide samples in categories of low and high methylation. This criterion is described in the paragraph below.

Fitting a Gaussian mixture model gives two modes (peaks) – one of low and one of high methylation. For the loci that are consistently unmethylated in normal lung, we define the distant peak to be the mode of high methylation in the disease sample group. We consider

[2]When stricter filtering criteria for consistent low and high methylation are applied, the subsequent steps fail to identify bimodal probes in the studied tumor types.



(a) Unmethylated in normal (b) Methylated in normal

Figure 7.10: Density plot of improvement and distant peak β value. A distant peak is the peak further away from the methylation state in normal tissue, e.g. the high methylation peak for loci that are unmethylated in normal. Red lines show the thresholds applied.

the locus of interest to categorize samples into ones of low and of high methylation when the distant peak is at a value of at least 0.6. Similarly, the distant peak in case of loci that are consistently methylated in normal lung is the lower mode of the mixture model fitted on disease samples. The locus is then used in subsequent analysis only if its distant peak is at a value lower than 0.4. Figure 7.10 visualizes the bivariate distributions of improvement values and distance peak methylation. One can clearly see that the selected thresholds are conservative, leading to the identification of a small number of high-confidence bimodal probes in tumor.

We refer to loci that pass both criteria (improvement and distant peak value) as strongly bimodal. The table below summarizes the number of strong bimodal loci we identified in each of the studied scenarios.

Regions	probes	probes	promoters	promoters
Sample group	adeno	squamous	adeno	squamous
Threshold for improvement	28	14	28	14
Bimodal loci (unmethylated in normal)	21	38	0	0
Bimodal loci (methylated in normal)	21	25	1	0
Bimodal loci (total)	42	63	1	0

Listings of the identifiers are available in Supplementary Table S9. Clustering based on the methylation values of all disease samples at the selected loci does not reveal dense clusters in the adenocarcinoma dataset (data not shown). In the case for squamous cell carcinoma, however, two separate groups might be identifiable (Supplementary Figure S3).

Summary

In this section, we investigated the possibility for methylation-specific tumor subtypes in lung adenocarcinoma and squamous cell carcinoma. When observing the genome-wide methylation patterns, the clustering results suggest that there are no discrete set of categories. A plausible explanation would be that we observe a continuous spectrum of hypo- and hypermethylation events that contribute to tumor proliferation.

In addition, we identified CpGs and promoters that show bimodal behaviour in each of the studied tumor subtypes. Empirical studies suggest that the methylation values at the selected bimodal probes in adenocarcinoma cannot distinguish between subgroups of samples, but in the case of squamous cell carcinoma, subgroup-specific profiles might exist.

7.2.3 Identification of colon cancer subtypes

In this section we focus on the colon cancer dataset and study the relationship between the sets of high confidence differentially methylated regions and several clinical properties of the patients. The differentially methylated peaks, along with the procedure for obtaining them, are described in Chapter 5. We used the information encoded in the patient-specific events to predict the following clinical attributes:

Tumor grade 17 patients presented with Grade 2 (G2) colon cancer and the other 7 : with G3.

Duke's stage 16 of the tumors are classified as Duke's stage A or B (considered a single category AB in this analysis), another 5 tumors are stage C and the remaining 3 tumors are advanced at stage D.

KRAS state 13 of the tumor samples contain a wild type (WT) allele of the KRAS gene, whereas mutations (MUT) in this gene are found in the remaining 11 cases.

Microsatellite stability 7 of the tumors are considered microsatellite instable (MSI), the other 17 are microsatellite stable (MSS).

In the following, we refer to a combination of peak score and clinical attribute as a *scenario*. The input data for the prediction methods consists of the events of the differentially methylated regions. A hypomethylation event was denoted by the value -1, a hypermethylation event by 1, and 0 signifies no differential methylation. The response variables considered are the four aforementioned clinical properties.

In some cases two or more regions provide identical data, that is, their events match for each patient. In mathematical terms such regions are represented by the same event vector. The presence of indistinguishable features may exert a negative effect on the stability of a classifier. We therefore ignored duplicated data by randomly selecting only one region within a set of peaks with identical event vectors. The remaining peaks are summarized in the table below:

Score	Regions
Tag counts	2,034
Scaled tag counts	2,566
Tag density	2,566

Region Selection Methods

The simplest approach to identifying regions whose differential methylation events are correlated with a clinical property, is through a statistical test. Since there are up to three events per peak, and the clinical properties are also categorical variables, Fisher's exact test would be most appropriate. The initial set of results presented in this section concerns this test for significant correlation. When it comes to prediction, we first applied elastic nets to select a subset of the differentially methylated peaks that are related to the attribute of interest. The elastic net method is applied in cases where the number of features greatly exceeds the number of analyzed samples. As a comparison, we also applied a simpler procedure for feature selection, namely forward selection on logistic regression. These machine learning techniques are introduced in Chapter 1.

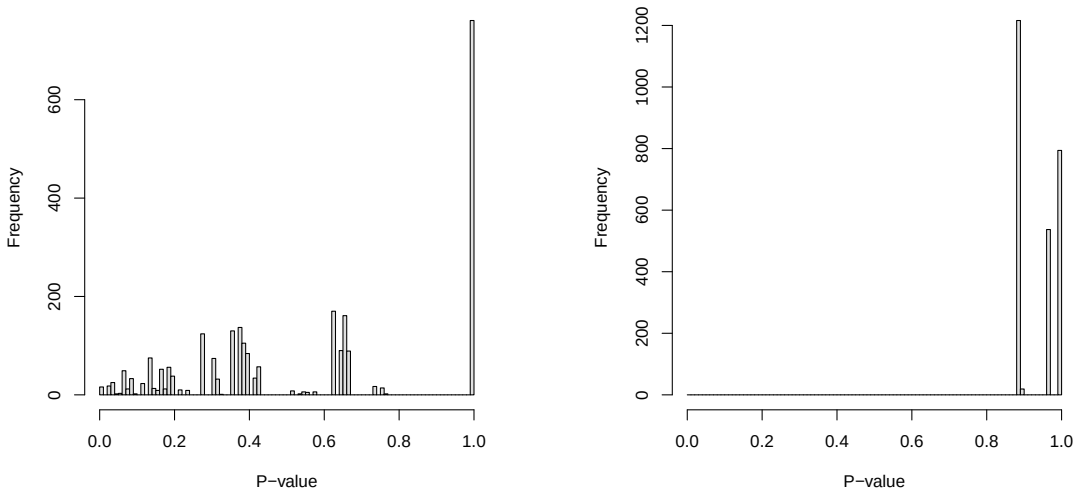
Fisher's exact test

We applied Fisher's exact test to each peak in each scenario. Note that we are comparing two properties in a set of 24 patients. The small size of the tested population invariably leads to p-values on a coarse scale. Figure 7.11 shows the distributions of p-values obtained using tag density for score and tumor grade as an outcome. The other scenarios show very similar distributions. Clearly, none of the p-values appears significant after false discovery rate (FDR) correction. This phenomenon is mainly due to the fact that we rely on 24 observations. This places a restrictive limit on the minimal possible p-value that may result from the statistical test. Unfortunately, the problem of multiple testing cannot be alleviated at this stage, and therefore, no lists of informative peaks can be obtained using this simple method.

Elastic nets

Elastic net classifiers contain two parameters – regularization parameter λ and L_1 penalty weight α. Training an elastic net involves parameter estimation, which was implemented using 4-fold cross validation. More precisely, we first split the set of 24 samples in 4 folds (subsets) of 6 samples each. We then trained elastic nets four times on a grid of 100×100 parameter values supplying three of the folds as training data and using the remaining fold to estimate the test error of the classifier. This error is referred to as *CV estimate* later in this section. In order to assess the stability of the method, the procedure described above was repeated 100 times using different splits of the samples.

Not surprisingly, the small number of observations leads to highly uncertain estimates of the cross-entropy (see Figure 7.12) and the misclassification error (data not shown). The shape of the error landscape, however, seems to be similar across repetitions. This is an indication that the elastic net might be a stable method for region selection. The selected parameter values in each of the 100 cross validation repetitions are plotted in Figure 7.13.



(a) Before multiple testing correction

(b) After multiple testing correction

Figure 7.11: Histograms of p-values obtained by association for correlation between differential methylation events and a clinical outcome using Fisher's exact test.

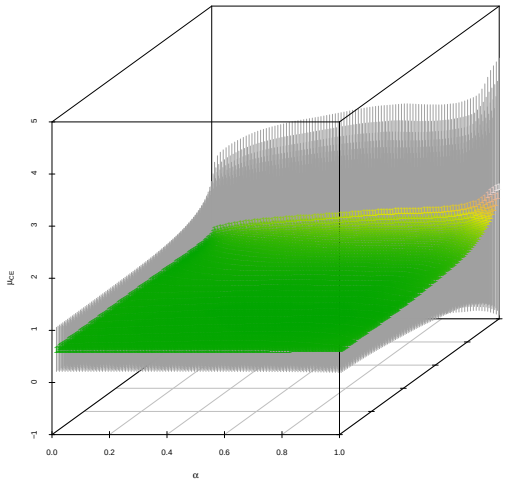


Figure 7.12: Cross validation estimates of elastic net's cross-entropy error at different parameter values. The horizontal axes show the values for the α parameter and the λ parameter ranks (the exact sequence of λ values differ depending on the selected α). The vertical axis show the cross-entropy error. The color of a point is a topographical encoding of its cross-entropy value: the lowest values observed are denoted by dark green, and the highest ones - by bright red. Grey vertical lines show the standard deviation of the CV estimate observed at a particular parameter combination.

The parameter estimates in the case of tumor grade are fairly stable, whereas the minima in the error landscape when predicting Dukes stage, KRAS mutations and microsatellite instability are very inconsistent across repetitions. In the latter case, an elastic net classifier is unable to identify a predictive pattern for the corresponding clinical properties. In fact, the misclassification errors of the classifiers in most scenarios is not better than random[3].

The features with non-zero coefficients in a classifier trained on all available patients using the optimal parameter combination can be considered as selected regions by the respective model. The number of selected regions also tends to fluctuate across repetitions. Figure 7.14 visualizes this phenomenon when predicting Dukes stage based on tag density values; the trained elastic net models in other scenarios show a very similar behavior. For each scenario, we computed the median number of non-zero features and focused on the classifier that is closest to this value. In case of ties, one repetition was selected at random. The selected classifiers are denoted by bright red dots in Figure 7.13. Later in this section, we refer to these peaks as *regions selected by elastic nets*.

Forward selection

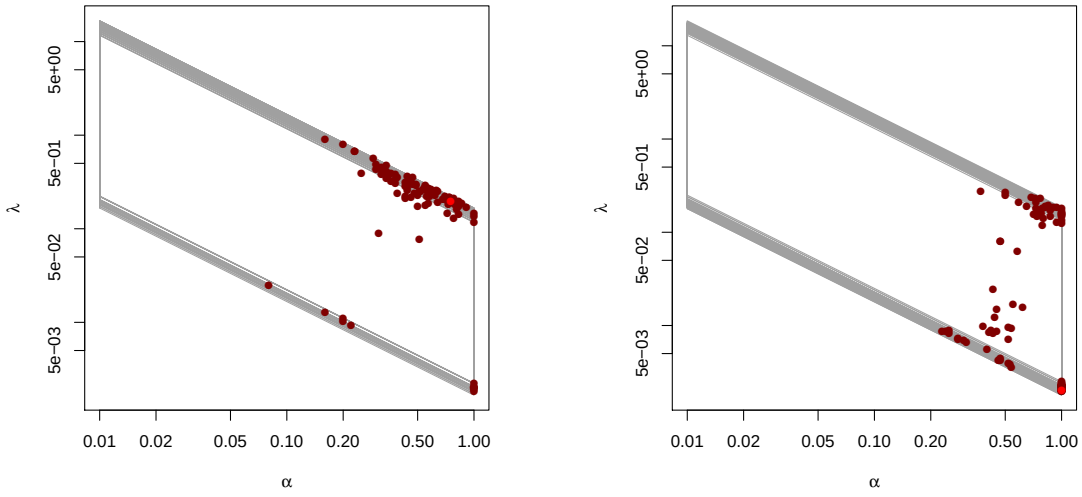
We compared the performance of elastic net classifiers with forward [feature] selection, applied on logistic regression models. The usage of this simple method is motivated by the inadequately small set of patients in comparison to the number of the peaks in consideration, coupled with the indication of instability of a classifier that operates on a large parameter space. When predicting tumor grade, KRAS mutations and microsatellite instability, the underlying models were binary logistic regression. In the case of Dukes stage, proportional odds logistic regression models were trained.

After training a sequence of models by successively adding features in a manner that improves the prediction most, we selected the model with the lowest deviance. The numbers of selected regions by this approach are very similar to the corresponding numbers obtained using elastic net. Figure 7.15 shows the values of variance for the different models using tag density as a score. The plots include models with feature space dimensionality only up to the number of training points (patients), because the deviance remains unchanged when the complexity increases above this value.

Similarly to elastic nets, the predictive power of the selected logistic regression models is weak in all studied scenarios. In addition, the two model families compared in this section tend to adopt different sets of regions in their predictions. Supplementary Table S10 provides the numbers of regions selected by these models in every scenario.

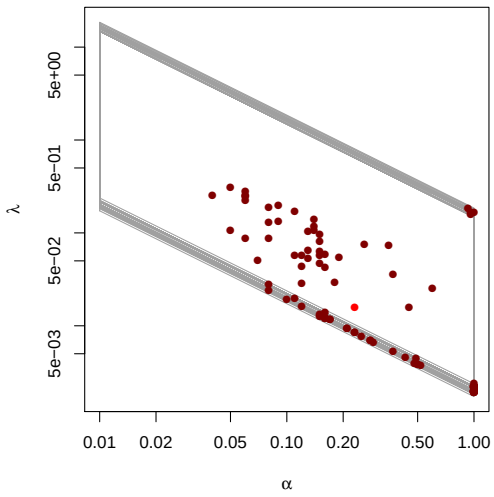
Despite its instability and low CV estimate, the elastic net model targeting microsatellite instability seems to select a set of regions that has the potential of forming a diagnostic panel. The differential methylation events in these regions are displayed in Figure 7.16. The heatmap suggests that a set of three or four simple rules could predict the state of an unknown sample based on its differential methylation events in the targeted regions. If colorectal tumors displaying microsatellite instability prove to be of high clinical value,

[3] Here, the studied model is compared to a dummy classifier that always predicts the most common class.

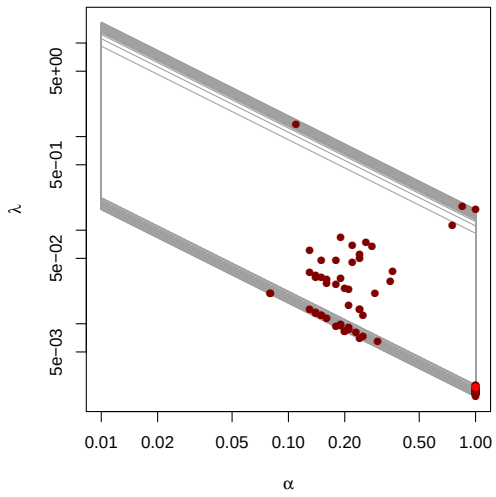


(a) Tumor grade

(b) Dukes stage



(c) KRAS state



(d) Microsatellite stability

Figure 7.13: Parameter space in training elastic net on a clinical property. The two axes show the values of the parameters α and λ. The set of covered values by cross validation is depicted by a grey parallelogram. The parameter combination that minimizes cross-entropy error is encircled in red. Since the parameter selection procedure was repeated 100 times, there are 100 parallelograms and 100 red circles in every plot. The model used for the extraction of significant regions is marked in bright red.

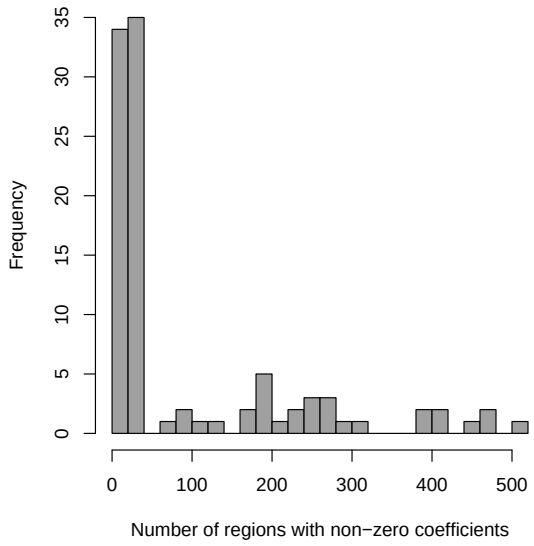


Figure 7.14: Histogram of number of selected regions by elastic net classifiers predicting Dukes stage based on tag density values.

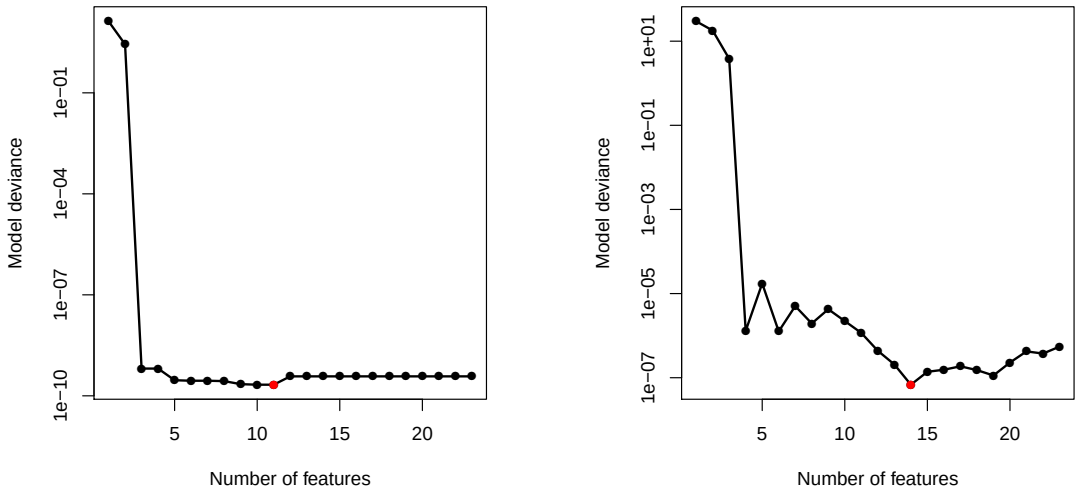
this candidate for a panel could be investigated further and needs to be validated in a larger cohort of patients.

7.3 Summary

This chapter addressed a broad spectrum of bioinformatic challenges related to cancer diagnosis. We first compared two linear models for predicting tumor type based on GoldenGate data. Both models showed impressive accuracies, indicating that the sample sizes of several dozen or a hundred samples per tumor type are large enough for a linear to capture a signal of (tumor) type-specific methylation. The clinical applicability of the trained models was validated by testing them on carcinomas of unknown primary origin.

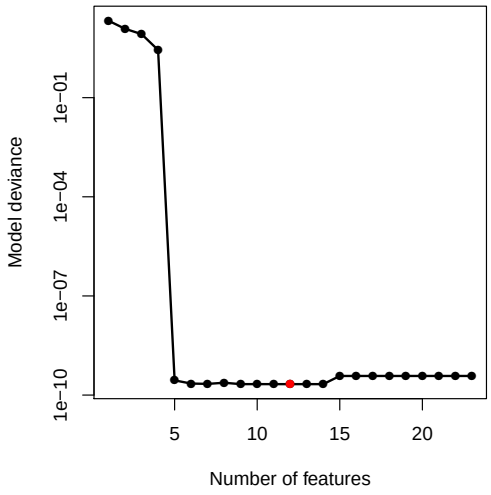
In the lung cancer dataset, we presented a strategy for identifying subtypes in lung adenocarcinoma and squamous cell carcinoma. The results obtained lead us to the conjecture that, in lung cancer, we observe a continuous spectrum of hypo- and hypermethylation events that contribute to tumor proliferation. In addition, we identified CpGs and promoters that show bimodal behaviour in each of the studied tumor subtypes. Empirical studies suggest that the methylation values at the selected bimodal probes in adenocarcinoma cannot distinguish between subgroups of samples, but in the case of squamous cell carcinoma, at least two subgroup profiles can be identified.

In the color cancer dataset, we studied the differential methylation events and searched for predictive markers of patient's clinical attributes. The tested methods include simple statistical tests, as well as linear models with elastic net penalty and forward feature selection. The limited cohort, coupled with very high data dimensionality did not allow us to train stable models or pinpoint strong candidates. Elastic net selected promising

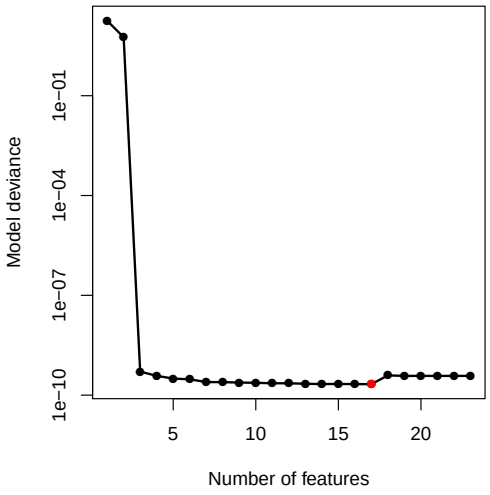


(a) Tumor grade

(b) Dukes stage



(c) KRAS state



(d) Microsatellite stability

Figure 7.15: Deviances of logistic regression models obtained using forward selection. The model with the lowest deviance in the sequence is marked in red.

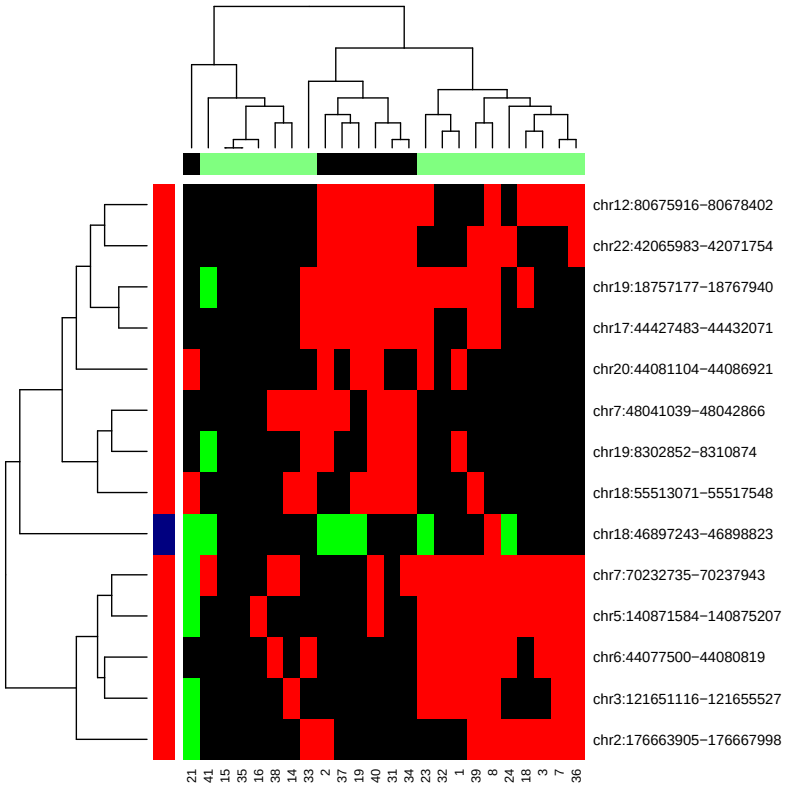


Figure 7.16: Heatmap of hyper- and hypomethylation status of regions selected by elastic net. The regions are represented by rows in the heatmap; the columns show all available patients. Patient identifiers are given at the bottom row. Hypermethylation is denoted by red color, hypomethylation - by green color, and no differential methylation - by black color in the heatmap. Regions are color-coded based on their CGI status - regions marked in red overlap with a CpG island, whereas those marked by a blue rectange lie outside CGIs. Columns in the heatmap are color-coded based on microsatellite stability status of the cancer in the corresponding patient.

candidates for indication of microsatellite instability, however, clinical application requires their further validation in a large cohort.

8 Conclusions and outlook

This work presents a collection of techniques for quality control, differential methylation and candidate biomarker prioritization in methylation-based datasets. These techniques were successfully applied is several collaboration projects and facilitated the comparison and analysis of methylation-based datasets some of which consist of heterogeneous samples. Some of the major findings of these studies are recapitulated below, accompanied by short notes on the methods applied.

8.1 Methylation profiling and differential methylation

We were first to analyze a collection of over 1,800 samples interrogated using the first generation of microarray-based methylation measurement technology, namely, the GoldenGate assay by Illumina. We experimented with a set of heuristic methods for probe and sample filtering and the identification of differentially methylated sites. In addition, we quantified the inter-sample methylation variability of a given population using selected quantiles. Last but not least, we compared several machine learning methods for phenotype prediction based on a vector of methylation values.

Our study confirmed and extended many previously known or suspected properties of gene promoter methylation. For example, we identified 68 genes that show age-dependent methylation, some of which have been mentioned in previous reports. We also confirmed that hypermethylation with age tends to occur in Polycomb-occupied promoters. By comparing the methylation patterns of embryonic and adult stem cells to differentiated ones, we show that induction of differentiation of these stem cell types leads to a methylation profile strongly resembling the one of the corresponding primary tissue, however, some loci gain less methylation. Furthermore, we observed that inter-individual methylation differences occur predominantly in CpG poor promoters. The DNA methylation map that emerges when clustering tumor samples shows a type-specific profile. Our analysis of a large collection of cancer cell lines revealed that their methylation profiles are distinct from all primary and metastatic tissues examined. In general, cell lines keep many of the methylation traits of their primary tumor ancestors, however, they also contain a large fraction of hypermethylated sites, predominantly located in CGIs.

By identifying tissue type- and tumor type-specific probes, we showed that elastic nets can be used as a feature selection method for methylation data when the ratio between number of features and number of samples is fairly small (up to ≈ 3). However, the classifier becomes very unstable as this number increases. Our set of heuristic definitions for differential methylation, dependent on the number of samples in the compared group, showed promising results that were validated by pyrosequencing. We identified novel deregulated

genes in non-cancerous conditions, including dementia, lupus and myopathy.

In colon cancer, our definition of differential methylation events based on MethylCap-seq methylation profiles proved very useful for identifying frequently hypermethylated regions, and less so in the case of hypomethylation. We showed that the hypermethylated regions with a strong support coincide with bivalent loci in human embryonic stem cells. Importantly, we estimated that the majority of the genes with hypermethylated promoters do not have significantly reduced expression levels in color cancer, because they are lowly expressed in the corresponding normal tissue. By designing a simple and flexible prioritization strategy, we identified high-confidence differentially methylated regions. The elastic net model showed a potential to guide the identification of a diagnostic panel, however, its instability inevitably poses a question on the reliability of the selected regions.

8.2 Tumor types and subtypes

We compared two models for predicting tumor type based on microarray methylation profiles – SVM with linear kernel and L_1-regularized logistic regression. Both models showed high accuracy (based on cross-validation estimates) during training on samples from primary solid tumors. Moreover, both models were highly efficient in predicting origin of metastatic samples, and were successfully applied on carcinomas of unknown primary origin.

We show that non-small-cell lung cancer cannot be unambiguously stratified into two distinct groups based on CpG methylation profiling by Infinium 450k. Rather, the examined patient cohort seems to show a continuous spectrum of hypo- and hypermethylation events that contribute to tumor proliferation.

8.3 Outlook

After the successful initiation of several large-scale international collaborations, the size of epigenetic repositories is growing at an unprecedented rate. One serious bioinformatics challenge not addressed in this thesis, is the ability to process ever larger datasets. Tamborero et al. recently made use of the available mutation data for 12 tumor types in TCGA, represented by over 3 thousand samples in total [107]. They identified novel driver genes by systematically cataloging all high confidence candidates for drivers using a voting of available methods and additional rules for inclusion. Their study shows, above all, the power that can be gained by simultaneous analysis of multiple datasets. Similar pan-cancer studies with a focus on the epigenetic mechanisms are currently being conducted [120].

Another major challenge is understanding the biological processes and functional mechanisms that underlie the observed and validated associations between epigenetic changes and phenotype. Tackling this problem by individual researchers or groups seems infeasible; it involves the careful integration of multidimensional data from genetic and other sources, as well as collaboration and sharing of expert knowledge. We already mentioned this need for integration in Chapter 6. Similarly, deciphering the signatures of CIMP tumors can

be achieved only by advancing our understanding on the biochemical processes specific for these phenotypes, as we argued in Chapter 7. An example for upcoming efforts for such integration is the suggestion by Stunnenberg and Hubner for the incorporation of proteomic data into genome-wide and epigenome-wide studies [105].

Bringing epigenetic diagnostic markers and medication to clinical practice is another attractive direction in current and future research. Large efforts focus on developing HDAC inhibitors for cancer treatment [123], as well as demethylating agents for hematological malignancies [82] and solid tumors [4]. Importantly, response to epigenetic drugs might take months after treatment initiation [113]. Development of reliable (computationally assisted) means of predicting response are therefore critical for the selection of adequate therapies. Drawing a parallel between the cellular heterogeneity in cancers and the accelerated evolution of HIV, Bock and Lengauer suggest an approach to cancer treatment that is highly personalized and assisted by computational means, similar to the successful application of HIV drug combination therapies [22]. Despite the serious economic and regulatory obstructions to such strategy, epigenetic combination therapies for cancer are recognized as the best tool in cases of resistance to current treatments or refractory states [26].

Bibliography

[1] Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. Blueprint to decode the epigenetic signature written in blood. *Nature biotechnology 30*, 3 (2012), 224–226.

[2] Adey, A., and Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome research 22*, 6 (2012), 1139–1143.

[3] Ahmad, K., and Henikoff, S. The histone variant h3. 3 marks active chromatin by replication-independent nucleosome assembly. *Molecular cell 9*, 6 (2002), 1191–1200.

[4] Ahuja, N., Easwaren, H., and Baylin, S. B. Harnessing the potential of epigenetic therapy to target solid tumors. *The Journal of clinical investigation 124*, 1 (2014), 56–63.

[5] Alberts, B., Johson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. 2002.

[6] Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N., and Warren, S. T. Age-associated dna methylation in pediatric populations. *Genome research 22*, 4 (2012), 623–632.

[7] Allis, C. D., Jenuwein, T., and Reinberg, D. *Epigenetics*. 2006.

[8] An, C., Choi, I.-S., Yao, J. C., Worah, S., Xie, K., Mansfield, P. F., Ajani, J. A., Rashid, A., Hamilton, S. R., and Wu, T.-T. Prognostic significance of cpg island methylator phenotype and microsatellite instability in gastric carcinoma. *Clinical cancer research 11*, 2 (2005), 656–663.

[9] Anacleto, C., Leopoldino, A. M., Rossi, B., Soares, F. A., Lopes, A., Rocha, J. C. C., Caballero, O., Camargo, A. A., Simpson, A. J., and Pena, S. D. Colorectal cancer "methylator phenotype": fact or artifact? *Neoplasia (New York, NY) 7*, 4 (2005), 331.

[10] Assenov, Y., Müller, F., Lutsik, P., Walter, J., and Lengauer, T. Compehensive analysis of dna methylation data with rnbeads. *Nature Methods* (2014 in press).

[11] Ayers, K. L., and Cordell, H. J. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic epidemiology 34*, 8 (2010), 879–891.

[12] Bachmann, I. M., Halvorsen, O. J., Collett, K., Stefansson, I. M., Straume, O., Haukaas, S. A., Salvesen, H. B., Otte, A. P., and Akslen, L. A. Ezh2 expression is associated with high proliferation rate and aggressive tumor subgroups in cutaneous melanoma and cancers of the endometrium, prostate, and breast. *Journal of Clinical Oncology 24*, 2 (2006), 268–273.

[13] Barlow, D. P. Genomic imprinting: a mammalian epigenetic discovery model. *Annual review of genetics 45* (2011), 379–403.

[14] Bestor, T. H. The host defence function of genomic methylation patterns. In *Novartis Found. Symp* (1998), vol. 214, pp. 187–195.

[15] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. High density dna methylation array with single cpg site resolution. *Genomics 98*, 4 (2011), 288–295.

[16] Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. Genome-wide dna methylation profiling using infinium® assay.

[17] Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., et al. High-throughput dna methylation profiling using universal bead arrays. *Genome research 16*, 3 (2006), 383–393.

[18] Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell 40*, 1 (1985), 91–99.

[19] Blair, J. D., and Price, E. M. Illuminating potential technical artifacts of dna-methylation array probes. *American journal of human genetics 91*, 4 (2012), 760.

[20] Bock, C. Analysing and interpreting dna methylation data. *Nature Reviews Genetics 13*, 10 (2012), 705–719.

[21] Bock, C., and Lengauer, T. Computational epigenetics. *Bioinformatics 24*, 1 (2008), 1–10.

[22] Bock, C., and Lengauer, T. Managing drug resistance in cancer: lessons from hiv therapy. *Nature Reviews Cancer 12*, 7 (2012), 494–501.

[23] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (1992), ACM, pp. 144–152.

[24] Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., and Stunnenberg, H. G. Whole-genome dna methylation profiling using methylcap-seq. *Methods 52*, 3 (2010), 232–236.

[25] Byun, H.-M., Siegmund, K. D., Pan, F., Weisenberger, D. J., Kanel, G., Laird, P. W., and Yang, A. S. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue-and individual-specific dna methylation patterns. *Human molecular genetics 18*, 24 (2009), 4808–4817.

[26] Campbell, R. M., and Tummino, P. J. Cancer epigenetics drug discovery and development: the challenge of hitting the mark. *The Journal of clinical investigation 124*, 1 (2014), 64–69.

[27] Chang, C.-C., and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 3 (2011), 27.

[28] Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing 20*, 1 (1998), 33–61.

[29] Chen, Y.-a., Choufani, S., Ferreira, J. C., Grafodatskaya, D., Butcher, D. T., and Weksberg, R. Sequence overlap between autosomal and sex-linked probes on the illumina humanmethylation27 microarray. *Genomics 97*, 4 (2011), 214–222.

[30] Chin, L., Andersen, J. N., and Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature medicine 17*, 3 (2011), 297–303.

[31] Coiffier, B., Pro, B., Prince, H. M., Foss, F., Sokol, L., Greenwood, M., Caballero, D., Borchmann, P., Morschhauser, F., Wilhelm, M., et al. Results from a pivotal, open-label, phase ii study of romidepsin in relapsed or refractory peripheral t-cell lymphoma after prior systemic therapy. *Journal of Clinical Oncology 30*, 6 (2012), 631–636.

[32] Consortium, E. P., et al. A userŠs guide to the encyclopedia of dna elements (encode). *PLoS Biol 9*, 4 (2011), e1001046.

[33] Cooper, D. N., and Youssoufian, H. The cpg dinucleotide and human genetic disease. *Human genetics 78*, 2 (1988), 151–155.

[34] Costello, J. F., and Plass, C. Methylation matters. *Journal of Medical Genetics 38*, 5 (2001), 285–303.

[35] Cross, S. H., Charlton, J. A., Nan, X., and Bird, A. P. Purification of cpg islands using a methylated dna binding column. *Nature genetics 6*, 3 (1994), 236–244.

[36] Curtin, K., Slattery, M. L., and Samowitz, W. S. Cpg island methylation in colorectal cancer: past, present and future. *Pathology research international 2011* (2011).

[37] Daskalakis, M., Nguyen, T. T., Nguyen, C., Guldberg, P., Köhler, G., Wijermans, P., Jones, P. A., and Lübbert, M. Demethylation of a hypermethylated p15/ink4b gene in patients with myelodysplastic syndrome by 5-aza-2'-deoxycytidine (decitabine) treatment. *Blood 100*, 8 (2002), 2957–2964.

[38] Deaton, A. M., and Bird, A. Cpg islands and the regulation of transcription. *Genes & development 25*, 10 (2011), 1010–1022.

[39] Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. Evaluation of the infinium methylation 450k technology. *Epigenomics 3*, 6 (2011), 771–784.

[40] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. Landscape of transcription in human cells. *Nature 489*, 7414 (2012), 101–108.

[41] Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graef, S., Johnson, N., Herrero, J., Tomazou, E. M., et al. A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis. *Nature biotechnology 26*, 7 (2008), 779–785.

[42] Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W., Hou, L., and Lin, S. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics 11*, 1 (2010), 587.

[43] Dunham, I., Birney, E., Lajoie, B. R., Sanyal, A., Dong, X., Greven, M., Lin, X., Wang, J., Whitfield, T. W., Zhuang, J., et al. An integrated encyclopedia of dna elements in the human genome.

[44] Ehrlich, M., Gama-Sosa, M. A., Huang, L.-H., Midgett, R. M., Kuo, K. C., McCune, R. A., and Gehrke, C. Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research 10*, 8 (1982), 2709–2721.

[45] Ellinger, J., Kahl, P., von der Gathen, J., Rogenhofer, S., Heukamp, L. C., Gütgemann, I., Walter, B., Hofstädter, F., Büttner, R., Müller, S. C., et al. Global levels of histone modifications predict prostate cancer recurrence. *The Prostate 70*, 1 (2010), 61–69.

[46] Epstein, C. J., Smith, S., Travis, B., and Tucker, G. Both x chromosomes function before visible x-chromosome inactivation in female mouse embryos.

[47] Fagnoni, F. F., Vescovini, R., Passeri, G., Bologna, G., Pedrazzoni, M., Lavagetto, G., Casti, A., Franceschi, C., Passeri, M., and Sansoni, P. Shortage of circulating naive cd8+ t cells provides new insights on immunodeficiency in aging. *Blood 95*, 9 (2000), 2860–2868.

[48] Feinberg, A. P., and Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts.

[49] Fenaux, P., Mufti, G. J., Hellstrom-Lindberg, E., Santini, V., Finelli, C., Giagounidis, A., Schoch, R., Gattermann, N., Sanz, G., List, A., et al. Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase iii study. *The lancet oncology 10*, 3 (2009), 223–232.

[50] Fernandez, A. F., Assenov, Y., Martin-Subero, J. I., Balint, B., Siebert, R., Taniguchi, H., Yamamoto, H., Hidalgo, M., Tan, A.-C., Galm, O., et al. A dna methylation fingerprint of 1628 human samples. *Genome research 22*, 2 (2012), 407–419.

[51] Forrester, W. C., Fernández, L. A., and Grosschedl, R. Nuclear matrix attachment regions antagonize methylation-dependent repression of long-range enhancer–promoter interactions. *Genes & development 13*, 22 (1999), 3003–3014.

[52] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*, 1 (2010), 1.

[53] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences 89*, 5 (1992), 1827–1831.

[54] Gardiner-Garden, M., and Frommer, M. Cpg islands in vertebrate genomes. *Journal of molecular biology 196*, 2 (1987), 261–282.

[55] Gluckman, P. D., Hanson, M. A., Buklijas, T., Low, F. M., and Beedle, A. S. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *Nature Reviews Endocrinology 5*, 7 (2009), 401–408.

[56] Guzzetta, G., Jurman, G., and Furlanello, C. A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC bioinformatics 11*, Suppl 8 (2010), S3.

[57] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*, second edition ed. 2009.

[58] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods 9*, 5 (2012), 473–476.

[59] Horvath, S. Dna methylation age of human tissues and cell types. *Genome biology 14*, 10 (2013), R115.

[60] Hughes, L. A., Khalid-de Bakker, C. A., Smits, K. M., van den Brandt, P. A., Jonkers, D., Ahuja, N., Herman, J. G., Weijenberg, M. P., and van Engeland, M. The cpg island methylator phenotype in colorectal cancer: progress and problems. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1825*, 1 (2012), 77–85.

[61] Imai, K., and Yamamoto, H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. *Carcinogenesis 29*, 4 (2008), 673–680.

[62] Iorio, M. V., Piovan, C., and Croce, C. M. Interplay between microrna and the epigenetic machinery: an intricate network. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1799*, 10 (2010), 694–701.

[63] Issa, J.-p. Aging and epigenetic drift: a vicious cycle. *The Journal of clinical investigation 124*, 1 (2014), 24–29.

[64] Joachims, T. Making large scale svm learning practical.

[65] Jones, P. A., and Laird, P. W. Cancer-epigenetics comes of age. *Nature genetics 21*, 2 (1999), 163–167.

[66] Kelly, T. K., De Carvalho, D. D., and Jones, P. A. Epigenetic modifications as therapeutic targets. *Nature biotechnology 28*, 10 (2010), 1069–1078.

[67] Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. A., Niveleau, A., Cedar, H., et al. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature genetics 38*, 2 (2006), 149–153.

[68] Kuznetsov, S. O. On computing the size of a lattice and related decision problems. *Order 18*, 4 (2001), 313–321.

[69] Laird, P. W. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics 11*, 3 (2010), 191–203.

[70] Landolin, J. M., Johnson, D. S., Trinklein, N. D., Aldred, S. F., Medina, C., Shulha, H., Weng, Z., and Myers, R. M. Sequence features that drive human promoter function and tissue specificity. *Genome research 20*, 7 (2010), 890–898.

[71] Lehrer, R. I., Xu, G., Abduragimov, A., Dinh, N. N., Qu, X.-D., Martin, D., and Glasgow, B. J. Lipophilin, a novel heterodimeric protein of human tears. *FEBS letters 432*, 3 (1998), 163–167.

[72] LeProust, E. AgilentŠs microarray platform: How high-fidelity dna synthesis maximizes the dynamic range of gene expression measurements.

[73] Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature 462*, 7271 (2009), 315–322.

[74] Liu, S. V., Fabbri, M., Gitlitz, B. J., and Laird-Offringa, I. A. Epigenetic therapy in lung cancer. *Frontiers in oncology 3* (2013).

[75] Maksimovic, J., Gordon, L., Oshlack, A., et al. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biol 13*, 6 (2012), R44.

[76] Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C. J., Ekström, T. J., Teschendorff, A. E., Tegnér, J., et al. An evaluation of analysis pipelines for dna methylation profiling using the illumina humanmethylation450 beadchip platform. *Epigenetics 8*, 3 (2013), 333–346.

[77] Meissner, A. Epigenetic modifications in pluripotent and differentiated cells. *Nature biotechnology 28*, 10 (2010), 1079–1088.

[78] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research 33*, 18 (2005), 5868–5877.

[79] Mittal, V. Improving the efficiency of rna interference in mammals. *Nature reviews genetics 5*, 5 (2004), 355–365.

[80] Miyazaki, K., Ozaki, T., Kato, C., Hanamoto, T., Fujita, T., Irino, S., Watanabe, K.-i., Nakagawa, T., and Nakagawara, A. A novel hect-type e3 ubiquitin ligase, nedl2, stabilizes p73 and enhances its transcriptional activity. *Biochemical and biophysical research communications 308*, 1 (2003), 106–113.

[81] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods 5*, 7 (2008), 621–628.

[82] Navada, S. C., Steinmann, J., Lübbert, M., and Silverman, L. R. Clinical development of demethylating agents in hematology. *The Journal of clinical investigation 124*, 1 (2014), 40–46.

[83] Nordlund, J., Bäcklin, C. L., Wahlberg, P., Busche, S., Berglund, E. C., Eloranta, M.-L., Flaegstad, T., Forestier, E., Frost, B.-M., Harila-Saari, A., et al. Genome-wide signatures of differential dna methylation in pediatric acute lymphoblastic leukemia. *Genome biology 14*, 9 (2013), r105.

[84] Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., et al. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell 17*, 5 (2010), 510–522.

[85] Ogino, S., Cantor, M., Kawasaki, T., Brahmandam, M., Kirkner, G. J., Weisenberger, D. J., Campan, M., Laird, P. W., Loda, M., and Fuchs, C. S. Cpg island methylator phenotype (cimp) of colorectal cancer is best characterised by quantitative dna methylation analysis and prospective cohort studies. *Gut 55*, 7 (2006), 1000–1006.

[86] Olsen, E. A., Kim, Y. H., Kuzel, T. M., Pacheco, T. R., Foss, F. M., Parker, S., Frankel, S. R., Chen, C., Ricker, J. L., Arduino, J. M., et al. Phase iib multicenter trial of vorinostat in patients with persistent, progressive, or

treatment refractory cutaneous t-cell lymphoma. *Journal of Clinical Oncology 25*, 21 (2007), 3109–3115.

[87] Peeters, R. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics 131*, 3 (2003), 651–654.

[88] Pelizzola, M., Koga, Y., Urban, A. E., Krauthammer, M., Weissman, S., Halaban, R., and Molinaro, A. M. Medme: an experimental and analytical methodology for the estimation of dna methylation levels based on microarray derived medip-enrichment. *Genome research 18*, 10 (2008), 1652–1659.

[89] Pidsley, R., Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. A data-driven approach to preprocessing illumina 450k methylation array data. *BMC genomics 14*, 1 (2013), 293.

[90] Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R., and Lichter, P. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature Reviews Genetics 14*, 11 (2013), 765–780.

[91] Reik, W., and Walter, J. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics 2*, 1 (2001), 21–32.

[92] Richardson, B. Primer: epigenetics of autoimmunity. *Nature Clinical Practice Rheumatology 3*, 9 (2007), 521–527.

[93] Rohle, D., Popovici-Muller, J., Palaskas, N., Turcan, S., Grommes, C., Campos, C., Tsoi, J., Clark, O., Oldrini, B., Komisopoulou, E., et al. An inhibitor of mutant idh1 delays growth and promotes differentiation of glioma cells. *Science 340*, 6132 (2013), 626–630.

[94] Russo, V. E., Martienssen, R. A., Riggs, A. D., et al. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

[95] Samowitz, W. S., Albertsen, H., Herrick, J., Levin, T. R., Sweeney, C., Murtaugh, M. A., Wolff, R. K., and Slattery, M. L. Evaluation of a large, population-based sample supports a cpg island methylator phenotype in colon cancer. *Gastroenterology 129*, 3 (2005), 837–845.

[96] Sandoval, J., Mendez-Gonzalez, J., Nadal, E., Chen, G., Carmona, F. J., Sayols, S., Moran, S., Heyn, H., Vizoso, M., Gomez, A., et al. A prognostic dna methylation signature for stage i non–small-cell lung cancer. *Journal of Clinical Oncology 31*, 32 (2013), 4140–4147.

[97] Sarma, K., and Reinberg, D. Histone variants meet their match. *Nature reviews Molecular cell biology 6*, 2 (2005), 139–149.

[98] Satterlee, J. S., Schübeler, D., and Ng, H.-H. Tackling the epigenome: challenges and opportunities for collaboration. *Nature biotechnology 28*, 10 (2010), 1039–1044.

[99] Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America 103*, 5 (2006), 1412–1417.

[100] Shahbazian, M. D., and Grunstein, M. Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem. 76* (2007), 75–100.

[101] Shen, H., and Laird, P. W. Interplay between the cancer genome and epigenome. *Cell 153*, 1 (2013), 38–55.

[102] Shen, J.-C., Rideout, W. M., and Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded dna. *Nucleic acids research 22*, 6 (1994), 972–976.

[103] Shinjo, K., Okamoto, Y., An, B., Yokoyama, T., Takeuchi, I., Fujii, M., Osada, H., Usami, N., Hasegawa, Y., Ito, H., et al. Integrated analysis of genetic and epigenetic alterations reveals cpg island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. *Carcinogenesis 33*, 7 (2012), 1277–1285.

[104] Simmer, F., Brinkman, A. B., Assenov, Y., Matarese, F., Kaan, A., Sabatino, L., Villanueva, A., Huertas, D., Esteller, M., Lengauer, T., et al. Comparative genome-wide dna methylation analysis of colorectal tumor and matched normal tissues. *Epigenetics 7*, 12 (2012), 1355–1367.

[105] Stunnenberg, H. G., and Hubner, N. C. Genomics meets proteomics: identifying the culprits in disease. *Human genetics* (2013), 1–12.

[106] Takai, D., and Jones, P. A. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences 99*, 6 (2002), 3740–3745.

[107] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports 3* (2013).

[108] Tan, G., Lou, Z., Liao, W., Zhu, Z., Dong, X., Zhang, W., Li, W., and Chai, Y. Potential biomarkers in mouse myocardium of doxorubicin-induced cardiomyopathy: a metabonomic method and its application. *PloS one 6*, 11 (2011), e27683.

[109] Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics 29*, 2 (2013), 189–196.

[110] Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G.,

Maxwell, A. P., et al. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research 20*, 4 (2010), 440–446.

[111] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. The accessible chromatin landscape of the human genome. *Nature 489*, 7414 (2012), 75–82.

[112] Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J.-P. J. Cpg island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences 96*, 15 (1999), 8681–8686.

[113] Treppendahl, M. B., Kristensen, L. S., and Grnønbæ æk, K. Predicting response to epigenetic therapy. *The Journal of clinical investigation 124*, 1 (2014), 47–55.

[114] Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D. Low-level processing of illumina infinium dna methylation beadarrays. *Nucleic acids research 41*, 7 (2013), e90–e90.

[115] Vaissière, T., Sawan, C., and Herceg, Z. Epigenetic interplay between histone modifications and dna methylation in gene silencing. *Mutation Research/Reviews in Mutation Research 659*, 1 (2008), 40–48.

[116] Wack, A., Cossarizza, A., Heltai, S., Barbieri, D., D'Addato, S., Fransceschi, C., Dellabona, P., and Casorati, G. Age-related modifications of the human alphabeta t cell repertoire due to different clonal expansions in the cd4+ and cd8+ subsets. *International immunology 10*, 9 (1998), 1281–1288.

[117] Waddington, C. The pupal contraction as an epigenetic crisis in drosophila. In *Proceedings of the Zoological Society of London* (1942), vol. 111, Wiley Online Library, pp. 181–188.

[118] Waddington, C. H. The epigenotype. *International journal of epidemiology 41*, 1 (2012), 10–13.

[119] Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nature genetics 39*, 4 (2007), 457–466.

[120] Weisenberger, D. J. Characterizing dna methylation alterations from the cancer genome atlas. *The Journal of clinical investigation 124*, 1 (2014), 17–23.

[121] Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., Kang, G. H., Widschwendter, M., Weener, D., Buchanan, D., et al. Cpg island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with braf mutation in colorectal cancer. *Nature genetics 38*, 7 (2006), 787–793.

[122] Weller, M., Stupp, R., Reifenberger, G., Brandes, A. A., van den Bent, M. J., Wick, W., and Hegi, M. E. Mgmt promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Reviews Neurology 6*, 1 (2009), 39–51.

[123] West, A. C., and Johnstone, R. W. New and emerging hdac inhibitors for cancer treatment. *The Journal of clinical investigation 124*, 1 (2014), 30–39.

[124] Wilson, R. C., and Doudna, J. A. Molecular mechanisms of rna interference. *Annual review of biophysics 42* (2013), 217–239.

[125] Yamashita, K., Dai, T., Dai, Y., Yamamoto, F., and Perucho, M. Genetics supersedes epigenetics in colon cancer phenotype. *Cancer cell 4*, 2 (2003), 121–131.

[126] Yang, X., Lay, F., Han, H., and Jones, P. A. Targeting dna methylation for epigenetic therapy. *Trends in pharmacological sciences 31*, 11 (2010), 536–546.

[127] Yao, J., Zhou, C., Wei, L., Wang, S., and Shi, Y. Expression of aquaporin-8 and bcl-2 protein in human cervical carcinoma and their correlations]. *Zhonghua fu chan ke za zhi 43*, 3 (2008), 205.

[128] Zhang, F. Z., and Hong, D. Elastic net-based framework for imaging mass spectrometry data biomarker selection and classification. *Statistics in medicine 30*, 7 (2011), 753–768.

[129] Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., et al. Genome-wide high-resolution mapping and functional analysis of dna methylation in< i> arabidopsis</i>. *Cell 126*, 6 (2006), 1189–1201.

[130] Zhang, Y., Chesler, E. J., and Langston, M. A. On finding bicliques in bipartite graphs: a novel algorithm with application to the integration of diverse biological data types. In *2013 46th Hawaii International Conference on System Sciences* (2008), IEEE Computer Society, pp. 473–473.

[131] Zhuang, J., Widschwendter, M., and Teschendorff, A. E. A comparison of feature selection and classification methods in dna methylation studies using the illumina infinium platform. *BMC bioinformatics 13*, 1 (2012), 59.

Appendices

List of publications

1. Fernandez, A.F., Assenov, Y., Martin-Subero J.I., Balint B., Siebert R., Taniguchi H., Yamamoto H., Hidalgo M., Tan A.C., Galm O., Ferrer I., Sanchez-Cespedes M., Villanueva A., Carmona J., Sanchez-Mut J.V., Berdasco M., Moreno V., Capella G., Monk D., Ballestar E., Ropero S., Martinez R., Sanchez-Carbayo M., Prosper F., Agirre X., Fraga M.F., Grana O., Perez-Jurado L., Mora J., Puig S., Prat J., Badimon L., Puca A.A., Meltzer S.J., Lengauer T., Bridgewater J., Bock C., Esteller M. A DNA methylation fingerprint of 1628 human samples. Genome Research 22, 2 (2012), 407-419.

2. Calvanese, V., Fernández, A.F., Urdinguio, R.G., Suárez-Alvarez, B., Mangas, C., Pérez-García, V., Bueno, C., Montes, R., Ramos-Mejía, V., Martínez-Camblor, P., Ferrero, C., Assenov, Y., Bock, C., Menendez, P., Carrera, A.C., Lopez-Larrea, C., Fraga, M.F. A promoter DNA demethylation landscape of human hematopoietic differentiation. Nucleic Acids Research 40, 1 (2012), 116-131.

3. Feuerbach, L., Halachev, K., Assenov, Y., Müller, F., Bock, C., Lengauer, T. Analyzing epigenome data in context of genome evolution and human diseases. Methods in Molecular Biology 856, 4 (2012), 431-467.

4. Simmer, F., Brinkman, A.B., Assenov, Y., Matarese, F., Kaan, A., Sabatino, L., Villanueva, A., Huertas, D., Esteller, M., Lengauer, T., Bock, C., Colantuoni, V., Altucci, L., Stunnenberg, H.G. Comparative genome-wide DNA methylation analysis of colorectal tumor and matched normal tissues. Epigenetics 7, 12 (2012), 1355-1367.

5. Sandoval, J., Mendez-Gonzalez, J., Nadal, E., Chen, G., Carmona, F.J., Sayols, S., Moran, S., Heyn, H., Vizoso, M., Gomez, A., Sanchez-Cespedes, M., Assenov, Y., Müller, F., Bock, C., Taron, M., Mora, J., Muscarella, L.A., Liloglou, T., Davies, M., Pollan, M., Pajares, M.J., Torre, W., Montuenga, L.M., Brambilla, E., Field, J.K., Roz, L., Lo Iacono, M., Scagliotti, G.V., Rosell, R., Beer, D.G., Esteller, M. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. Journal of Clinical Oncology 31, 32 (2013), 4140-4147.

6. Assenov, Y., Müller, F., Lutsik, P., Walter, J., and Lengauer, T. Compehensive analysis of DNA methylation data with RnBeads. Nature Methods (2014), in press.

Supplementary figures

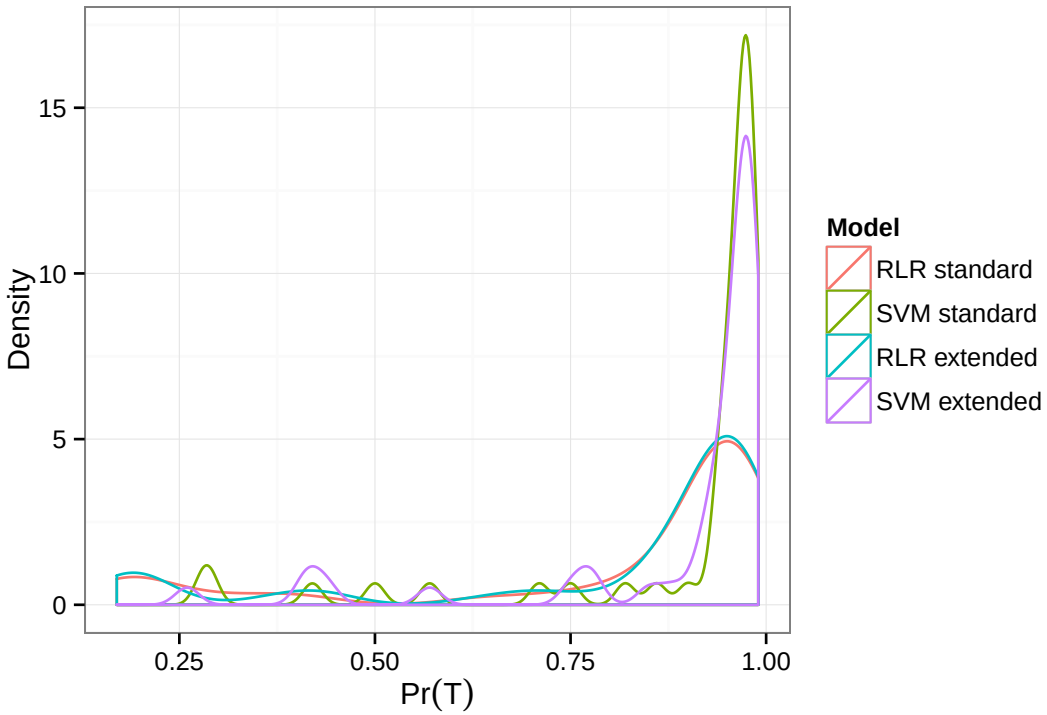


Figure S1: Probabilities for the dominant class of the models predicting tumor of primary origin. The distributions are shown as densities estimated based on a test set of 50 metastatic samples.

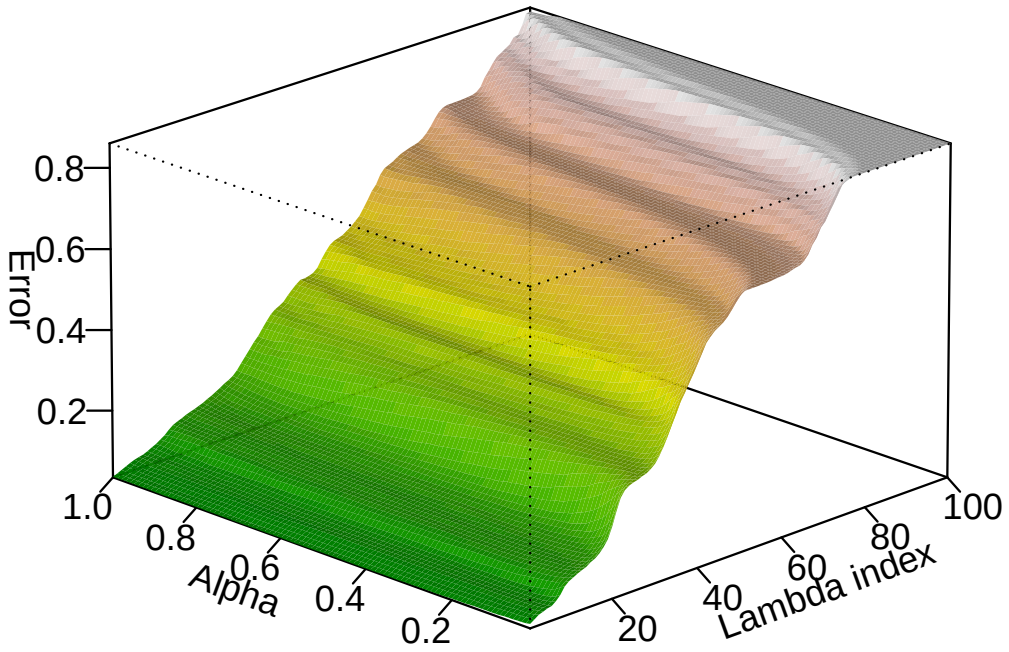


Figure S2: CV estimates of misclassification error obtained after training logistic regression model with elastic net penalty. The training set consists of over 4,600 Infinium 450k samples from 16 different solid tumor types, downloaded from TCGA.

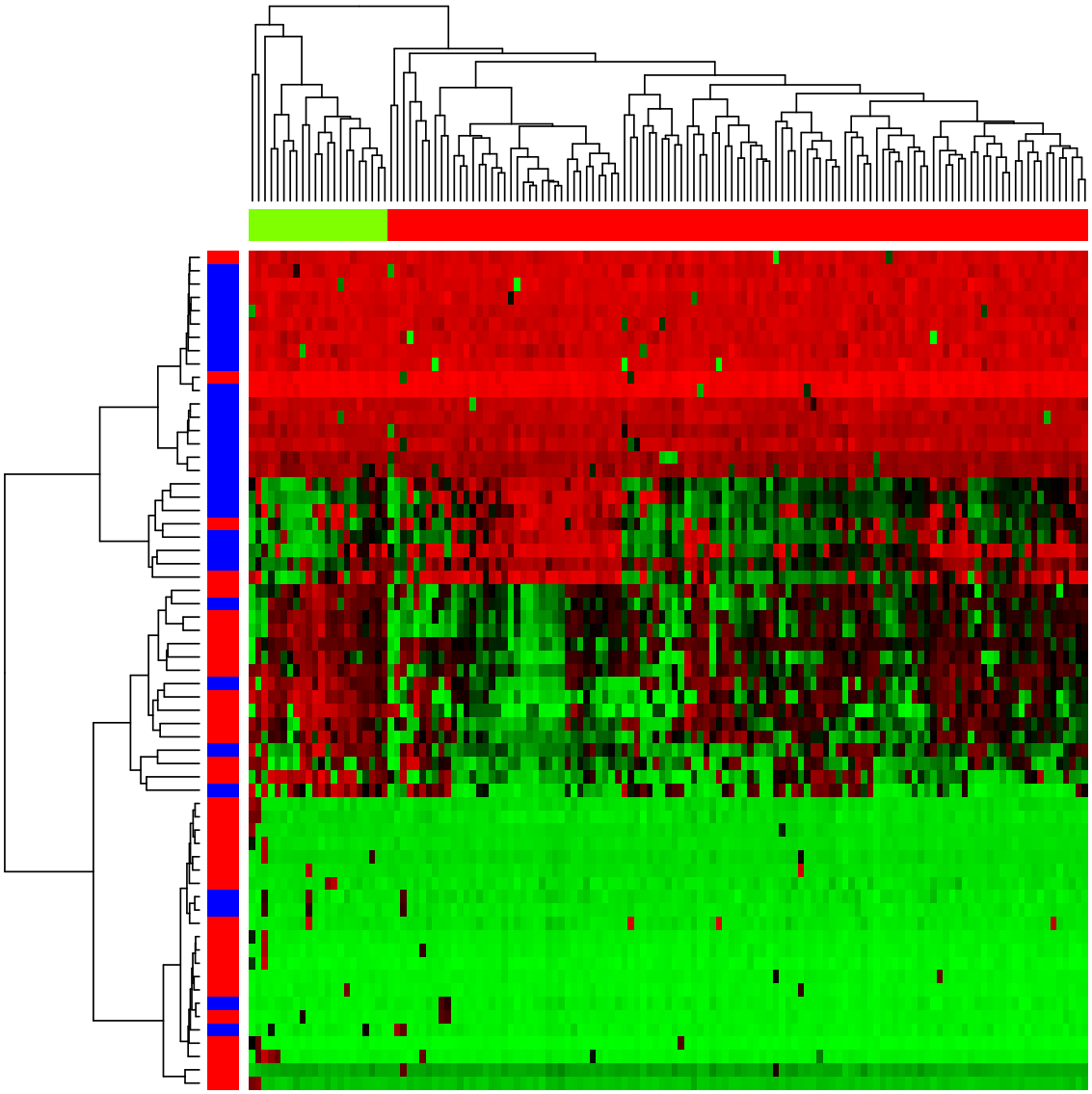


Figure S3: Heatmap of methylation values at CpGs that were identified as strongly bimodal. Every row in this heatmap corresponds to an Infinium probe, and every column – to a sample. Values are color-coded using a palette from bright green (no methylation), through black (50% methylation) to bright red (close to 100% methylation). Row color denotes probe's relation to a CpG island: the probes within a CpG island are marked by a red stripe, whereas the others are blue. Column colors denote sample subgroup association.

Supplementary tables

Table S1: Clinical annotation of the patients in the colon cancer dataset. *Dukes* = Dukes Stage; *KRAS* = KRAS Mutation Status; *MS* = Microsatellite Instability Status; *WT* = wild type

Patient	**Age**	**Gender**	**Grade**	**Dukes**	**KRAS**	**MS**
9	78	M	G2	D	WT	MSS
4	44	M	G2	B	G13D	MSI
5	66	M	G3	D	G12D	MSS
12	71	F	G2	D	WT	MSI
14	62	M	G2	B	WT	MSS
15	57	M	G2	B	WT	MSS
21	36	M	G3	D	G13D	MSI
22	63	F	G1	A	G12V	MSI
23	76	F	G2	D	WT	MSS
32	81	M	G2	C	WT	MSS
33	53	M	G3	D	WT	MSS
35	78	M	G2	B	WT	MSS
36	81	M	G2	B	G13D	MSS
37	74	M	G2	C	WT	MSI
38	85	M	G2	B	G12D	MSS
1	60	M	G2	B	WT	MSS
2	81	F	G3	C	WT	MSI
3	81	M	G3	C	G13D	MSS
7	51	F	G3	B	G12D	MSS
8	83	M	G3	B	G12A	MSS
13	83	M	G2	B	G12C	MSS
16	70	M	G2	A	WT	MSS
17	77	M	G2	D	WT	MSS
18	50	F	G2	B	G12C	MSS
19	87	F	G2	B	G12D	MSI
24	81	F	G2	B	G12V	MSS
31	75	F	G3	C	WT	MSI
34	79	F	G2	B	G12D	MSI
39	66	M	G2	B	WT	MSS
40	63	M	G2	B	G12V	MSI
41	65	F	G2	B	WT	MSS

Table S2: Probes in the GoldenGate dataset showing significant associations between methylation degree in healthy colon and age of individual. *Chr* = Chromosome; *Cor* = Correlation with age.

ID	Chr	Location	CGI	Gene	Cor	P-value
ADCYAP1_P398_F	18	894989	yes	ADCYAP1	0.45	6.05E-03
CCNA1_E7_F	13	35904640	yes	CCNA1	0.45	6.30E-03
CHGA_E52_F	14	92459297	yes	CHGA	0.51	1.17E-04
HCK_P858_F	20	30102860	yes	HCK	0.44	9.47E-03
KDR_E79_F	4	55686440	yes	KDR	0.44	8.70E-03
KDR_P445_R	4	55686964	yes	KDR	0.44	9.51E-03
MOS_E60_R	8	57189035	yes	MOS	0.47	2.33E-03
MYOD1_E156_F	11	17697891	yes	MYOD1	0.46	3.83E-03
NPY_P295_F	7	24290039	yes	NPY	0.44	9.07E-03
PENK_E26_F	8	57521117	yes	PENK	0.44	7.62E-03

Table S3: Table of associations between number of differentially methylated regions (columns) and sample processing and patient clinical information (rows).

	Hypomethylated	Hypermethylated	Total DMRs
Arrival Date	8.15E-01	7.24E-01	1.92E-01
Sonication Date	3.50E-01	1.00E+00	6.45E-01
Sonication Researcher	5.88E-01	8.59E-01	2.05E-01
Auto-MethylCap Date	4.12E-06	5.41E-08	6.17E-08
Library Preparation Date	1.42E-05	1.13E-06	2.43E-06
Solexa Run Date	4.85E-05	3.90E-05	2.12E-06
Solexa Operator	8.66E-02	1.03E-03	9.61E-04
Flowcell	4.85E-05	3.90E-05	2.12E-06
Lane	5.40E-03	1.67E-01	5.43E-02
Age	3.71E-01	7.92E-01	3.34E-01
Gender	2.83E-01	3.47E-01	6.40E-01
Grade	9.68E-01	4.98E-01	7.86E-01
Dukes	7.14E-01	6.03E-01	5.75E-01
Stage-T	8.34E-01	1.74E-01	4.66E-01
Stage-N	9.81E-01	7.72E-01	8.96E-01
KRAS	6.85E-01	5.56E-01	7.28E-01
MS	8.74E-01	9.51E-01	4.55E-01

Table S4: Table of associations between number of differentially methylated regions (columns) and sample processing and patient clinical information (rows).

Increase in Median Methylation	Probes	Promoters
Less than -0.25	412	191
Between -0.25 and 0.25	25814	13613
At least 0.25	1352	666

Table S5: Correlations between Infinium promoter support (defined in Chapter 6) and methylation and metrics in TCGA datasets.

Support based on	Increase in Methylation	Mean Expression
Tag counts	0.62	-0.13
Scaled tag counts	0.63	-0.12
Tag density	0.63	-0.13

Table S9: Strong bimodal Infinium 450k probes in the two cancer types of the lung cancer dataset. *Chr* = Chromosome; *Enh* = Enhancer; *SG* = Size of Smaller Group.

ID	Chr	Location	Strand	Gene	Enh	SG
Adenocarcinoma, unmethylated probes						
cg13064658	1	212003989	-	LPGAT1		3
cg15936121	1	153700431	-	INTS3		5
cg02628050	2	70418124	+	C2orf42		5
cg26347887	2	88927196	+	EIF2AK3		4
cg08175029	3	23848512	+	UBE2E1		3
cg13677120	4	52709175	-	DCUN1D4		2
cg23521603	7	130353913	-	TSGA13;COPG2		3
cg26597982	8	143484414	+	TSNARE1		2
cg13483474	9	139001555	-		yes	2
cg14204415	9	131464988	-	PKN3		2
cg24745327	11	65190198	-	NEAT1		3
cg05728786	15	40401260	-	BMF		3
cg06135755	15	86338090	+	KLHL25		3
cg06401851	16	31190843	+	FUS		2
cg07562135	19	39340654	+	HNRNPL		4
cg22310976	19	49468491	+	FTL		4
cg04234465	1	28974730	-	RNU11		2
cg13273097	3	152879313	-	RAP2B		6
cg16027376	8	144653975	+	C8orf73	yes	2
cg16068096	10	89102268	+	LOC728190;LOC439994		2
cg25296314	12	122277851	-	HPD		2

continues on next page ...

... continued from previous page

Adenocarcinoma, methylated probes						
cg08975528	6	31867700	+	ZBTB12		20
cg00176066	7	6196591	-	USP42		6
cg16331358	10	1147675	-	WDR37		4
cg07859478	19	808116	-	PTBP1		3
cg15344716	20	60889058	-	LAMA5		4
cg24448565	X	37404614	+			5
cg24048263	2	242659698	-	ING5		3
cg18235278	5	124228992	-			2
cg04752818	7	1536622	-	INTS1		3
cg13533616	9	131889368	-	PPP2R4	yes	2
cg21244116	9	140729241	-	EHMT1		3
cg24517875	10	134538559	+	INPP5A		2
cg15896447	14	24683287	-	MDP1;CHMP4A		3
cg02447095	16	70748321	+	VAC14	yes	7
cg03287527	16	10843654	+	NUBP1	yes	3
cg08241115	16	722688	+	RHOT2		4
cg27516925	16	90095949	-	GAS8;C16orf3		2
cg19357094	17	79260670	+	SLC38A10		2
cg05505294	19	1442231	+			3
cg26648488	19	5711704	-	LONP1		3
cg27000503	19	1950427	+	CSNK1G2		3
Squamous cell carcinoma, umethylated probes						
cg16955726	X	134125259	+	LOC644538	yes	60
cg01799338	1	36348937	+	EIF2C1		2
cg26841013	1	228248013	-	WNT3A	yes	48
cg13699355	2	468179	-		yes	55
cg26347887	2	88927196	+	EIF2AK3		2
cg05278650	3	13009132	+	IQSEC1		3
cg10748086	3	13009316	-	IQSEC1		2
cg15709989	3	185912227	+	DGKG	yes	59
cg17641046	3	37034473	+	MLH1;EPM2AIP1		3
cg12654349	5	56205094	-	C5orf35		2
cg02563952	6	31707803	+	MSH5		2
cg03302738	6	31707502	+	MSH5		2
cg04880558	6	31707613	+	MSH5		2
cg08312215	6	33266943	+	RGL2		4
cg18488157	6	29521598	+		yes	38
cg15617814	11	131780492	-	NTM		65
cg19717586	11	131781257	+	NTM		54
cg24745327	11	65190198	-	NEAT1		2

continues on next page ...

... continued from previous page

cg07915921	12	54321502	+			42
cg09670128	12	52627047	+	KRT7		51
cg13879483	12	95942907	+	USP44		51
cg01870456	13	111267991	-	CARKD		2
cg05362517	13	37393368	+	RFXAP		2
cg02308192	14	75593334	-	NEK9		3
cg14087806	17	73030732	-			7
cg03502002	18	74962133	-	GALR1	yes	48
cg12019614	19	11353996	-	DOCK6	yes	51
cg13021192	20	57582213	+	CTSZ		2
cg21120539	20	57582241	+	CTSZ		2
cg01660911	22	17082772	+	psiTPTE22		63
cg12003064	22	46263834	+			32
cg01352705	2	216979551	+	XRCC5		3
cg05037927	2	61372117	-	C2orf74		66
cg13821577	2	216979737	+	XRCC5		3
cg12706983	6	28092239	+	ZSCAN16		4
cg14454942	6	8064764	+	MUTED		2
cg12559208	19	54668230	+	TMC4		53
cg19310786	20	57582371	+	CTSZ		2

Squamous cell carcinoma, methylated probes

cg02365596	2	24413782	-	C2orf84		3
cg07486474	4	628649	-	PDE6B		41
cg08754725	6	3293098	-	SLC22A23		2
cg22198853	6	1594411	+			40
cg07333231	8	1051728	+			62
cg27649037	8	53322510	+	ST18		30
cg14339778	10	134359611	-	INPP5A		2
cg10508127	11	67462816	+			63
cg19052829	17	80809619	+	TBCD		2
cg20558091	17	79122298	+	AATK		40
cg27341866	19	2278618	-	C19orf35		56
cg24448565	X	37404614	+			2
cg15688683	1	1425815	+	ATAD3B		2
cg18367529	1	1161866	-	SDF4		5
cg24048263	2	242659698	-	ING5		3
cg22575656	7	1539157	+	INTS1		2
cg02230133	8	143424799	+	TSNARE1		53
cg11792616	8	30018355	+	DCTN6		2
cg00257187	10	91401349	-	PANK1		2
cg09070101	10	1130179	+	WDR37		4
cg27591016	13	47259150	-	LRCH1	yes	2

continues on next page ...

... continued from previous page

cg00929655	15	92459909	-	SLCO3A1	2
cg13289118	15	75255681	-		3
cg27516925	16	90095949	-	GAS8;C16orf3	2
cg06090161	17	79244943	-	SLC38A10	8

Table S6: Correlation coefficients between hypo- and hypermethylation support and average scores for ENCODE datasets in the studied peaks.

Correlation	H1-H3K4me3	H1-H3K27me3	H1-H3K36me3	H1-DNase	H1-Pol2	H1-CTCF	HCT116-H3K4me3	HCT116-H3K27me3	HCT116-DNase	HCT116-Pol2	HCT116-CTCF
Tag counts hypomethylation support	0.00	0.02	-0.06	-0.01	-0.03	-0.01	-0.03	-0.02	-0.02	-0.03	-0.02
Tag counts hypermethylation support	0.35	0.47	0.01	0.31	0.15	0.05	0.01	-0.01	0.06	0.03	0.05
Scaled tag counts hypomethylation support	0.01	0.02	-0.05	0.03	-0.01	0.00	-0.03	-0.05	0.00	-0.03	-0.01
Scaled tag counts hypermethylation support	0.36	0.47	0.00	0.31	0.16	0.05	0.01	-0.01	0.07	0.04	0.05
Tag density hypomethylation support	0.01	0.02	-0.05	0.04	-0.01	0.00	-0.03	-0.05	0.00	-0.03	-0.01
Tag density hypermethylation support	0.36	0.47	0.00	0.31	0.16	0.05	0.01	-0.01	0.06	0.04	0.06

Table S7: Top 20 hypermethylated regions in the analyzed colon cancer dataset.

Region	Aggregated Score	Support	CpG Density
chr10:102880421-102891603	2.5	23	44.5
chr8:97574251-97577410	13.5	23	60.5
chr13:27393400-27402978	16.0	22	42.7
chr13:94160632-94164746	24.5	22	81.9
chr21:36997726-37005801	41.0	22	47.2
chr2:47649523-47653570	47.0	22	72.6
chr4:154927352-154934414	52.5	21	45.5
chr6:150326470-150329134	52.5	21	75.5
chr13:92675712-92679677	53.5	22	43.6
chr7:154933063-154937874	54.5	21	52.6
chr7:157176446-157180286	56.5	22	70.3
chr8:24868480-24871916	59.5	21	62.9
chr8:25952330-25966210	60.0	21	51.0
chr2:119318398-119334083	62.0	22	53.6
chr12:103373785-103377642	62.5	21	57.6
chr14:50628796-50632703	62.5	21	58.9
chr6:108591209-108600567	63.5	22	51.1
chr5:127901440-127903817	69.0	21	54.3
chr8:11594639-11600842	69.0	21	52.1
chr1:50651808-50667073	69.5	21	55.2

Table S8: Prediction of origin of metastatic samples, GoldenGate dataset. Numer of samples considered to be correctly predicted are shown in **bold**.

True origin \ Prediction	**Correct type**	**Uncertain**	**Wrong type**	**Accuracy**
RLR standard				
colon cancer	**43**	1	1	0.96
kidney cancer	n.a.	**3**	2	0.60
total	**46**		4	0.92
SVM standard				
colon cancer	**44**	0	1	0.98
kidney cancer	n.a.	**1**	4	0.20
total	**45**		5	0.90
RLR extended				
colon cancer	**43**	1	1	0.96
kidney cancer	n.a.	**4**	1	0.80
total	**47**		3	0.94
SVM extended				
colon cancer	**43**	0	2	0.96
kidney cancer	n.a.	**1**	4	0.20
total	**44**		6	0.88

Table S10: Number of regions selected by different models predicting colon cancer subtypes.

Regions \ Score	**Tag counts**	**Scaled tag counts**	**Tag density**
All differentially methylated	2034	2566	2566
Elastic net			
Predictive of Grade	1	3	1
Predictive of Dukes	25	23	22
Predictive of KRAS	261	37	257
Predictive of MS	26	18	14
Forward selection			
Predictive of Grade	13	14	11
Predictive of Dukes	19	13	14
Predictive of KRAS	18	16	12
Predictive of MS	12	13	17