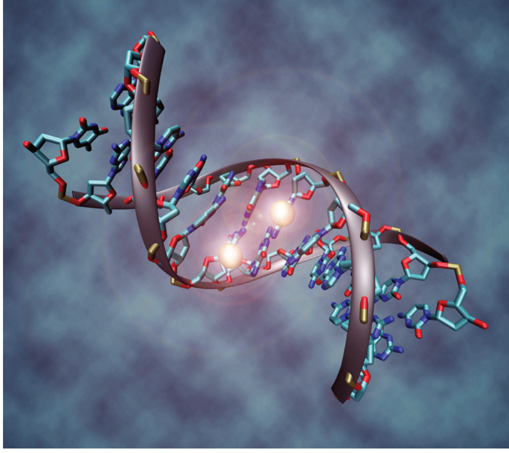
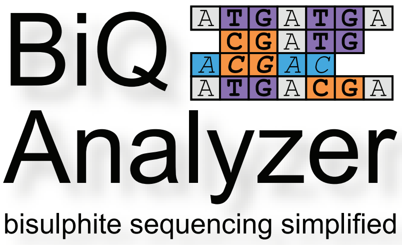
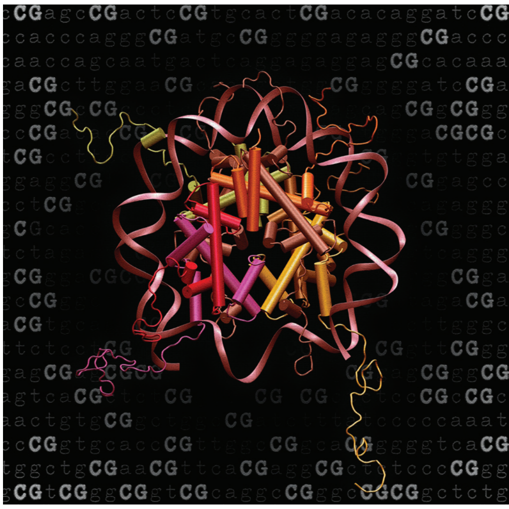
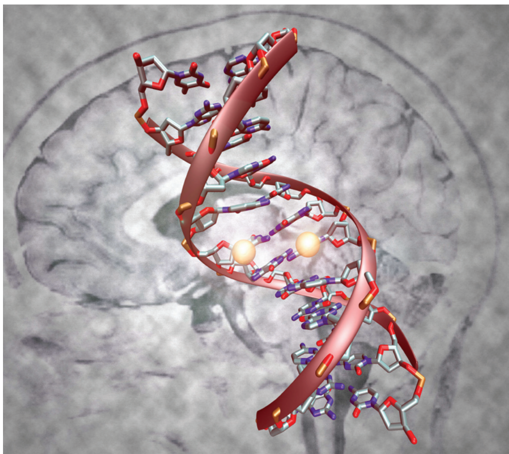








Computational Epigenetics

PhD Thesis by Christoph Bock

Computational Epigenetics
Bioinformatic methods for epigenome prediction, DNA methylation mapping and cancer epigenetics

Dissertation
zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.) im Fach Informatik
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

von
Christoph Bock

eingereicht im Mai 2008

Datum der Einreichung:	31. Mai 2008
Gutachter:	Prof. Dr. Thomas Lengauer, Ph.D.
	Prof. Dr. Jörn Walter
	Prof. Dr. Martin Vingron
Datum des Kolloquiums:	2. Oktober 2008
Dekan der Fakultät:	Prof. Dr. Joachim Weickert
Vorsitzender des Kolloquiums:	Prof. Dr. Gerhard Weikum
Protokollant	Dr. Mario Albrecht

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I would like to thank the advisors of this PhD thesis – Thomas Lengauer and Jörn Walter – for their unconditional support and the freedom they gave me in pursuing my goals and ideas. All work was carried out in the group of Thomas Lengauer, who was an important discussion partner throughout the three-and-a-half years of this PhD project. I am also grateful to Martin Vingron for his willingness to act as an additional reviewer of this thesis.

Collaboration with molecular biologists was essential to the work presented in this thesis. It is my pleasant duty to acknowledge the collaboration partners who performed wet-lab experiments relevant for this thesis, asked challenging questions and provided insightful comments, most notably: Martina Paulsen, Sascha Tierling, Diana Santacruz and Barbara Hutter of Saarland University (Germany), François Fuks, Emmanuelle Viré and Carmen Brenner of Université Libre de Bruxelles (Belgium), Thomas Mikeska of the Peter MacCallum Cancer Centre (Australia), Eivind Hovig and Fang Liu of the Norwegian Radium Hospital (Norway), and Dirk Moser of the University of Trier (Germany).

Furthermore, I thank all current and past students for their invaluable contributions to our ongoing research initiative on computational epigenetics. Konstantin Halachev designed and implemented a substantially enhanced and extended version of the EpiGRAPH backend and contributed important ideas to all aspects of the EpiGRAPH project (described in chapter B-3 of this thesis). Lars Feuerbach contributed important concepts and ideas to the algorithm for CpG island annotation (described in chapter B-4 of this thesis) and is currently designing and implementing a speed-optimized version of this algorithm. Peter Schüffler is currently designing and implementing the MethMarker software, which follows up the results of chapter D-3 and which is briefly outlined in chapter E-3. Yassen Assenov, Martin Kircher, Juliane Perner, Holger Jung, Oliver Frings and Enrico Glaab contributed relevant ideas to different aspects of my work.

Technical support by Joachim Büch is greatly acknowledged. His database expertise was critical for realizing the software architecture of the EpiGRAPH web service. Furthermore, I would like to thank Jörn Rahnenführer, Jochen Maydt, Andre Altmann and Tobias Sing for inspiring discussions about statistics and machine learning and Ruth Schneppen-Christmann for kind support with all kinds of administrative issues.

Finally, I thank Andreas Steffen, Katya Todorova, Kasia Bozek, Susanne Bitzer and Sparte 4 for making Saarbrücken an enjoyable place to live, my parents for support and encouragement, and my friends for being around.

Work on this PhD thesis was supported by basic funding of the Max Planck Society and by the German National Academic Foundation (Studienstiftung des deutschen Volkes). Travel grants by the European Society of Human Genetics and Japanese government funds are gratefully acknowledged. The described work was conducted within the context of the European Union (EU) Network of Excellence “Biosapiens” (LSHGCT-2003–503265) and the EU Network of Excellence “The Epigenome” (LSHG-CT-2004–503433).

ABSTRACT

Epigenetic research aims to understand heritable gene regulation that is not directly encoded in the DNA sequence. Epigenetic mechanisms such as DNA methylation and histone modifications modulate the packaging of the DNA in the nucleus and thereby influence gene expression. Patterns of epigenetic information are faithfully propagated over multiple cell divisions, which makes epigenetic gene regulation a key mechanism for cellular differentiation and cell fate decisions. In addition, incomplete erasure of epigenetic information can lead to complex patterns of non-Mendelian inheritance. Stochastic and environment-induced epigenetic defects are known to play a major role in cancer and ageing, and they may also contribute to mental disorders and autoimmune diseases.

Recent technical advances – such as the development of the ChIP-on-chip and ChIP-seq protocols for genome-wide mapping of epigenetic information – have started to convert epigenetic research into a high-throughput endeavor, to which bioinformatics is expected to make significant contributions. This thesis describes computational work at the intersection of epigenetics and genome research, aiming to address the bioinformatic challenges posed by the human epigenome. While its methods are carried over and adapted from bioinformatics and related fields (including data mining, machine learning, statistics, algorithms, optimization, software engineering and databases), its overarching goal is to contribute to epigenetic research, both directly through analyzing and modeling of epigenetic information, and indirectly through the development of practically useful methods and software toolkits.

This thesis is broadly structured into four parts. The first part gives a brief introduction into epigenetic regulation and inheritance, and reviews the emerging field of computational epigenetics. The second part addresses the question of genome-epigenome interactions using machine learning methods. It is shown that accurate predictions of DNA methylation and other epigenetic modifications can be derived from the genomic DNA sequence. Based on this finding, the EpiGRAPH web service for epigenome analysis and prediction is described, and methods for refined annotation of CpG islands in the human genome are proposed. The third part is dedicated to large-scale analysis of DNA methylation, which is the best-known epigenetic phenomenon. The BiQ Analyzer software toolkit is presented, together with a bioinformatic analysis of the "National Methylome Project for Chromosome 21" dataset, for which BiQ Analyzer had played an enabling role. This part concludes with statistical modeling of DNA methylation variation and an analysis of its implications for DNA methylation mapping in a large number of human individuals. The fourth part describes two pilot projects applying the bioinformatic concepts of this thesis to cancer epigenetics. First, genome-scale datasets are probed for evidence of a link between DNA methylation and Polycomb binding, which is believed to play a role in epigenetic deregulation of cancer cells. Second, a biomarker that tests for cancer-specific DNA methylation is optimized and validated for use in clinical settings.

Arguably the most interesting result of this thesis is the unexpectedly high correlation between genome and epigenome that was found by several methods and based on multiple epigenome datasets. This finding suggests that the role of the genome for epigenetic regulation has been underappreciated, and it underlines the importance of integrated analysis of genome and epigenome. With the EpiGRAPH web service for (epi-) genome analysis and prediction, a research tool is provided to facilitate further investigation of this striking interaction.

KURZFASSUNG

Ziel epigenetischer Forschung ist ein besseres Verständnis der Mechanismen erblicher Gen-Regulation, die nicht direkt in der DNA-Sequenz codiert sind. Epigenetische Veränderungen des Genoms – wie zum Beispiel DNA-Methylierung und Histon-Modifikationen – beeinflussen die räumliche Anordnung der DNA im Zellkern und damit auch die Gen-Expression. Epigenetische Informationen werden über viele Zellteilungen stabil weitergegeben, weswegen die epigenetische Gen-Regulation ein Schlüsselmechanismus für Zell-Differenzierung und Determinierung ist. Darüber hinaus ergeben sich aus dem unvollständigen Löschen von epigenetischen Informationen komplexe nicht-Mendelsche Vererbungsgänge. Stochastische und umweltinduzierte epigenetische Defekte spielen eine wichtige Rolle für Krebs und molekulares Altern, und sie scheinen ebenfalls psychische Störungen und Autoimmun-Erkrankungen zu beeinflussen.

In Folge technischer Fortschritte – wie etwa der Entwicklung der ChIP-on-chip und ChIP-seq Protokolle zur genomweiten Kartierung epigenetischer Informationen – hat eine Transformation der epigenetischen Forschung hin zu Hochdurchsatz-Analysen begonnen, zu der die Bioinformatik einen wichtigen Beitrag leisten muss. Diese Dissertation beschreibt bioinformatische Studien an der Schnittstelle von Epigenetik und Genomforschung, mit dem Ziel einer adäquaten Antwort auf die analytischen Herausforderungen des menschlichen Epigenoms. Während ihre Methoden aus der Bioinformatik und benachbarten Gebieten (Data Mining, maschinelles Lernen, Statistik, Algorithmik, Optimierung, Software Engineering und Datenbanken) entlehnt und adaptiert sind, ist es das übergeordnete Ziel der Arbeit, einen Beitrag zur epigenetischen Forschung zu leisten; und zwar sowohl direkt durch die Analyse und Modellierung epigenetischer Daten, also auch indirekt durch die Entwicklung praktisch verwertbarer Methoden und Software-Werkzeuge.

Diese Dissertation gliedert sich grob in vier Teile. Der erste Teil führt in den Themenkomplex der epigenetischen Vererbung und Gen-Regulation ein und fasst das junge Forschungsgebiet „Computational Epigenetics“ zusammen. Der zweite Teil adressiert die Frage nach Genom-Epigenom-Interaktionen mit Methoden des maschinellen Lernens. Es wird gezeigt, dass aus der genomischen DNA-Sequenz eine akkurate Vorhersage der DNA-Methylierung sowie anderer epigenetischer Modifikationen abgeleitet werden kann. Basierend auf diesem Ergebnis werden der EpiGRAPH-Webservice zur Epigenom-Analyse und Vorhersage beschrieben sowie Methoden für die verbesserte Annotation von CpG-Inseln in Wirbeltier-Genomen ausgearbeitet. Der dritte Teil beschäftigt sich mit der Hochdurchsatzanalyse von DNA-Methylierung, dem bekanntesten epigenetischen Phänomen. Die BiQ Analyzer Software wird vorgestellt, und die Ergebnisse einer bioinformatischen Analyse des „National Methylome Project for Chromosome 21“-Datensatzes werden beschrieben, zu dessen Generierung der BiQ Analyzer einen fundamentalen Beitrag leisten konnte. Den Abschluss dieses Teils bildet die statistische Modellierung von DNA-Methylierungs-Variation und eine Analyse ihrer Bedeutung für die DNA-Methylierungs-Kartierung einer großen Anzahl menschlicher Individuen. Der vierte Teil beschreibt zwei Pilotprojekte, in denen die bioinformatischen Konzepte dieser Arbeit in der Krebs-Epigenetik angewandt werden. Zum einen werden epigenomische Datensätze im Hinblick auf Interaktionen zwischen DNA-Methylierung und Polycomb-Bindestellen untersucht – eine Beziehung, die vermutlich bei der epigenetischen Deregulierung von Krebszellen eine Rolle spielt. Zum anderen wird ein Biomarker für die Ver-

wendung unter klinischen Bedingungen optimiert und validiert, der eine krebsspezifische Veränderung der DNA-Methylierung detektieren kann.

Das vielleicht interessanteste Ergebnis dieser Dissertation ist eine unerwartet hohe Korrelation zwischen Genom und Epigenom, die mit mehreren Methoden und für verschiedenste Epigenom-Datensätze nachgewiesen werden konnte. Dieses Ergebnis legt nahe, dass der regulatorische Einfluss des Genoms auf das Epigenom bisher nicht ausreichend gewürdigt wurde, und es unterstreicht die Wichtigkeit einer integrierten Analyse von Genom und Epigenom. Der EpiGRAPH-Webservice bietet sich als Werkzeug für eine genauere Untersuchung dieser bemerkenswerten Interaktion an.

Part A. Introduction into Computational Epigenetics

"Epigenetics has always been all the weird and wonderful things that can't be explained by genetics" (Denise Barlow)[1]

A-1 Outline

Epigenetic research is currently undergoing a major transformation, from small-scale, hypothesis-driven studies to a genome-scale endeavor with key roles to play for bioinformatics. While the term "epigenetics" dates back to the 1940s (Waddington 1942) and molecular biology research on epigenetic mechanisms has been performed since the 1970s (Holliday 2006), the completion of the human genome sequence has finally provided us with the methods and resources to investigate mammalian epigenomes at a truly genomic scale.

These introductory chapters serve a dual purpose. First, a brief introduction into epigenetics is given, focusing on epigenetic inheritance (chapter A-2) and epigenetic gene regulation (chapter A-3), in order to provide the biological background required for the remainder of this thesis. Second, the emerging field of computational epigenetics is reviewed, in order to sketch the scientific context in which the current work is placed (chapter A-4 to chapter A-7). For consistency of presentation, own published work is included in this review, and the corresponding thesis chapters are cross-referenced (see pages 136f for a list of publications that have arisen from the PhD project described in this thesis). Part A concludes with a brief outline of the remainder of this thesis (chapter A-8).

A-2 Two facets of epigenetic inheritance

Epigenetics is commonly defined as the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Russo et al. 1996). The constitutive property of epigenetic inheritance is that it is encoded in covalent modifications of the DNA and the chromatin proteins attached to it, rather than in the DNA sequence itself (as is the case for genetic inheritance). Since such modifications are more readily altered than the DNA sequence, epigenetic information can be reprogrammed dynamically during cellular differentiation, but is also propagated with substantially lower fidelity than genetic information. An error rate of 10^{-3} has been estimated per site and cell division for DNA methylation (Ushijima et al. 2003), in contrast to values in the order of 10^{-8} per basepair and cell division for genetic mutations (Drake et al. 1998). Epigenetic inheritance occurs both between generations of cells (mitotic inheritance) and between generations of a species (meiotic inheritance).

Epigenetic mitotic inheritance is critically involved in cellular differentiation and cell fate decisions. Recent research has provided a mechanistic understanding of the key phases of epigenetic regulation during development (Reik 2007). To start with, germ cells carry highly specialized and parent-specific epigenetic information. Shortly after fertilization, a fundamental reprogramming step resets most epigenetic information to a default state, which is difficult to analyze in vivo due to the small number of cells available at this stage. However, it seems plausible to assume that specific properties of the DNA sequence play a major role in determining which genomic regions assume which epigenetic state. An epigenome that is repro-

[1] Quoted after: http://epigenome.eu/en/2,9,5

grammed for pluripotency seems to be crucial for the ability of embryonic stem (ES) cells to differentiate into diverse tissue types. During cellular differentiation, ES cells reprogram their epigenetic state once again when tissue-specific transcription factors are activated and pluripotency-specific genes become silenced. In terminally differentiated cells, epigenetic information is faithfully propagated during cell division. However, cellular ageing leads to increasing heterogeneity within a cell population and can also contribute to tumor development (Fraga and Esteller 2007). Finally, the specialized cells of the germline reprogram epigenetic information in a parent-specific way, before it is passed on to the offspring as sperm or egg. In addition to its role in regulating cell-type specific gene expression, epigenetic mitotic inheritance is relevant for X-chromosome inactivation (Heard 2004), the process by which one out of two copies of the X-chromosome in females is randomly selected and constitutively silenced.

Epigenetic meiotic inheritance is caused by incomplete reprogramming in the early embryo, which results in the propagation of epigenetic information from parent to offspring. This phenomenon gives rise to patterns of phenotypic inheritance that are inconsistent with Mendelian rules. First, imprinted genes are inherited and expressed in a parent-specific way, i.e. only the maternal allele is transcribed while the paternal allele is epigenetically silenced (or vice versa). Imprinted genes play a central role in the development of placenta and brain, and they have been linked to several rare neurogenetic disorders as well as to cancer (Solter 2006). Second, acquired traits can be epigenetically transmitted over multiple generations. While this type of inheritance is relatively rare in mammals (Peaston and Whitelaw 2006), for plants it seems to be a common way of adapting gene regulation to a changing environment (Grant-Downton and Dickinson 2006).

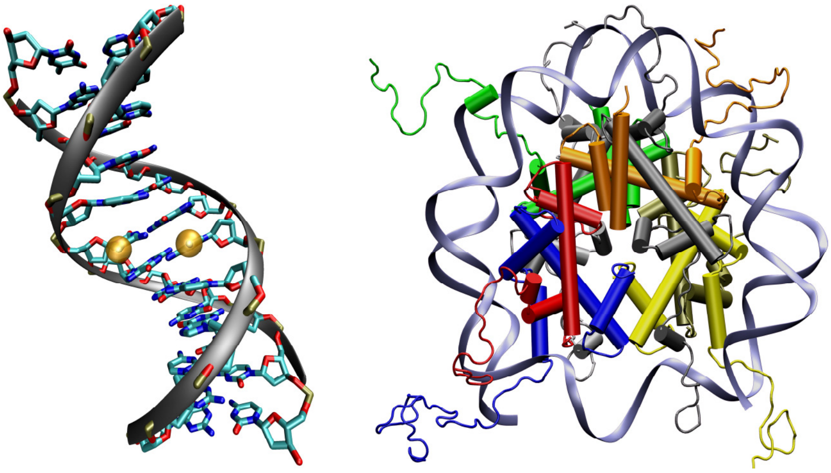


Figure 1. Carriers of epigenetic information: DNA and nucleosome

The left panel shows a DNA double helix that is methylated symmetrically on both strands (orange spheres) at its center CpG (PDB structure: 329d). DNA methylation is the only epigenetic mechanism that directly targets the DNA. The right panel shows a nucleosome spindle consisting of eight histone proteins (center), around which two loops of DNA are wound (PDB structure: 1KX5). The nucleosome is subject to covalent modifications of its histones and to the binding of non-histone proteins.

A-3 Mechanisms of epigenetic regulation

Epigenetic regulation exploits the fact that the packaging of DNA inside the nucleus directly influences gene expression (Dillon 2006). In general, the tighter a gene's DNA is wrapped up, the more likely it is switched off. Conversely, the more accessible it is to the transcription machinery, the more likely it is actively transcribed. Physically, the genome of eukaryotic cells is stored in a highly regulated protein-DNA complex called chromatin, which controls DNA accessibility for cellular processes such as transcription, replication and DNA repair (Woodcock 2006). Epigenetic mechanisms can be both activating (i.e. fostering open chromatin structure, called euchromatin) or repressive (i.e. fostering condensed chromatin structure, called heterochromatin), and different epigenetic mechanisms frequently act synergistically. Three biochemical mechanisms are commonly referred to as epigenetic: (i) DNA methylation, (ii) histone modifications, and (iii) binding of non-histone proteins such as Polycomb and trithorax group complexes.

DNA methylation (Bird 2002; Weber and Schübeler 2007) is the only epigenetic modification that directly affects the DNA. Biochemically, a hydrogen atom of the cytosine base is replaced by a methyl group (Figure 1, left). This does not alter the way in which the cytosine is transcribed into mRNA, but it fosters a locally more compact chromatin structure and affects transcription factor binding. In mammals, DNA methylation is largely confined to cytosines in a CpG context (“CpG” stands for **c**ytidine and **g**uanosine, separated by a **p**hosphate atom), which has two important implications. First, any genomic position that can be methylated is symmetric, i.e. there is a – methylated or unmethylated – cytosine on the forward strand as well as on the reverse strand. Therefore, after DNA replication a specific enzyme can read the DNA methylation pattern of the parent strand and faithfully copy it to the newly synthesized strand, thereby maintaining heritable DNA methylation patterns. Second, in mammalian genomes CpG dinucleotides occur in clusters, and the genomic regions with highest CpG density – the CpG islands – exhibit the lowest levels of DNA methylation. This phenomenon is most likely caused by the fact that mutation rates are substantially higher for methylated CpGs than for unmethylated CpGs, hence absence of DNA methylation at least in the germline seems to be constitutive for long-term maintenance of most CpG islands.

Histone modifications (Kouzarides 2007) are post-translational modifications of the core histone proteins that constitute the nucleosome (Figure 1, right). The long and unstructured N-terminal tails by which histone proteins interact with neighboring nucleosomes are subject to various types of covalent modifications, including lysine and arginine methylation, lysine acetylation and serine phosphorylation (see Kouzarides 2007 for a comprehensive list). Histone modifications influence the nucleosome’s assembly into higher-order packaging structures by moderating its DNA-binding affinity and by recruiting further chromatin remodeling complexes. The concept of the histone code (Turner 2007) suggests that histone modifications are used combinatorially to program genes for activation during subsequent steps of cellular differentiation. While this combinatorial model is consistent with a number of recent observations, including the programmed activation of tissue-specific transcription factors during differentiation of ES cells (Bernstein et al. 2006), it has also been argued that a simpler additive model is often sufficient to explain epigenetic gene regulation by histone modifications (Dion et al. 2005).

Non-histone proteins influence chromatin structure by interacting with nucleosomes and DNA in a number of ways. ATP-dependent chromatin remodeling complexes act like molecular machines and can directly move or displace nucleosomes along the DNA (Gangaraju and Bartholomew 2007). A second group of proteins, which includes heterochromatin protein 1 (HP1) as well as the Polycomb and trithorax group complexes, can be thought of as the readers and writers of the epigenome. They bind to the DNA or to specifically modified histones and catalyze other histone modifications or DNA methylation. The Polycomb group complex 2 (PRC2) for example catalyzes repressive histone methylations and recruits DNA methylation through its interaction with a DNA methyltransferase (Schuettengruber et al. 2007). Transcription factors can also affect chromatin structure, for example through recruitment of histone acetylases. Interestingly, there is evidence that transcription factor binding is sometimes maintained during cell division and would therefore qualify as mitotically heritable (Zhou et al. 2005). Nevertheless, by convention rather than by definition transcription factor binding is not usually regarded as epigenetic. Finally, DNase I hypersensitivity (Boyle et al. 2008), i.e. the presence or absence of sites that are particularly amenable to digestion by the DNase I enzyme, shows some properties that mimic epigenetic mechanisms, but is a technical method that assesses different aspects of chromatin structure rather than a well-defined molecular mechanism.

In summary, a variety of epigenetic mechanisms jointly control the packaging of the DNA, thereby regulating which genes are accessible for transcription. Epigenetic mechanisms are highly interwoven and regulate their target genes (and each other) in a complex network of synergistic and antagonistic interactions. Disentangling this network both biochemically for a small number of representative genes and statistically from a whole-genome perspective, and relating the results to development and disease are important goals of epigenetic research. In the following chapters, we discuss arising bioinformatic challenges, and we show how computational methods have contributed and will continue to contribute to answering important epigenetic questions.

A-4 Generation, low-level processing and quality control of epigenetic data

Various experimental techniques have been developed for genome-wide mapping of epigenetic information (Table 1). These techniques follow a basic three-stage design. First, the epigenetic information is biochemically converted into genetic information, e.g. by enriching genomic regions that carry a particular histone modification in a DNA library. Second, standard DNA techniques such as tiling microarrays or sequencing are applied. Third, computational algorithms are used to infer the epigenetic information from the tiling array data or sequencing output. All experimental methods for epigenome mapping generate large amounts of data and require efficient ways of low-level data processing and quality control.

For ChIP-on-chip (Table 1), the key bioinformatic challenge is to derive a ranked list of over-represented genomic regions from raw probe intensities. Although there are some similarities to the analysis of tiling array data for transcriptome mapping (see Royce et al. 2005 for review), most available algorithms are specifically targeted to peak finding in ChIP-on-chip data. The initial and still widely used solution employs a three-step process (Cawley et al. 2004). First, the microarrays are quantile-normalized and standardized to a common median intensity. Second, a Wilcoxon rank sum test is applied locally on a sliding window to test for differential hybridization and to derive an enrichment score for each probe. Third, significant probes are merged into regions of over-representation if sufficiently close to each other, and these regions are ranked by their combined enrichment. More recently, hidden Markov models were introduced to improve the detection accuracy (first implemented in HMMTiling, Li et al. 2005), linear models were applied to control for differences in probe sensitivity (implemented in MAT (Johnson et al. 2006) for Affymetrix one-color arrays and in MA2C (Song et al. 2007) for NimbleGen two-color arrays), and probabilistic binding models were used to improve spatial resolution (implemented in the JBD algorithm, Qi et al. 2006). Furthermore, several peak finding toolkits have been developed to facilitate routine processing of ChIP-on-chip datasets. TileMap is an easy-to-use peak finder for Affymetrix tiling array data, which has been applied in a number of independent studies (Ji and Wong 2005); Ringo is a Bioconductor package for the analysis of ChIP-on-chip data from the widely used NimbleGen platform (Toedling et al. 2007); ChIPOTle is a basic peak finding macro for Excel, which does not take platform-specific information into account (Buck et al. 2005); and Tilescope is a fully integrated analysis pipeline that is applicable to data from both the Affymetrix and the NimbleGen platform (Zhang et al. 2007b). In spite of the abundance of algorithms published recently, the peak finding problem for ChIP-on-chip data cannot be regarded as being solved. In particular, current peak finders have problems with histone modifications that cover extended genomic regions and they seem to miss a substantial number of weak binding sites. In order to select a biologically meaningful cutoff that distinguishes between significant peaks and random fluctuations, experimental validation of a moderate number of detected peaks continues to be crucial. To guide this process, a framework has been proposed that can help identify most informative regions for validation (Du et al. 2006).

ChIP-on-chip (Buck and Lieb 2004) combines **ch**romatin **i**mmuno**p**recipitation (ChIP) for enriching specific DNA fragments with the power of tiling microarrays for detecting differences between immunoprecipitated and control DNA. Initially, cells are treated with formaldehyde to cross-link any DNA-bound proteins to the DNA. Next, the chromatin is extracted and sheared into small fragments, which are typically around 500 basepairs in length (this step limits the method's resolution). Using an antibody against a histone modification or a chromatin protein the corresponding fragments are enriched. The DNA is then released from these fragments and hybridized to a tiling microarray. Regions that are significantly over-represented in the immunoprecipitated DNA relative to control DNA are regarded as epigenetically modified or protein-bound, depending on the antibody used. In a variant of ChIP-on-chip that is called **me**thyl-**D**NA **i**mmuno**p**recipitation (MeDIP), purified DNA is immunoprecipitated with an antibody against methylated cytosines, giving rise to genomic maps of DNA methylation. While these methods are used in a large number of laboratories world-wide, antibody quality remains a matter of concern and must be monitored carefully. Furthermore, background noise introduced by cross-hybridization and varying oligomer affinities should be accounted for during data processing. To foster data quality and standardization, the Microarray and Gene Expression Data Society has released a checklist of minimum required information about ChIP-on-chip experiments that are to be reported for any dataset (Microarray and Gene Expression Data Society 2005). While ChIP-on-chip is increasingly replaced by ChIP-seq (see below) for genome-wide studies, the former continues to be important for studies of localized genomic regions such as all promoter regions or a specific chromosome.
ChIP-seq (Barski et al. 2007; Mikkelsen et al. 2007) is a variant of ChIP-on-chip that uses high-throughput DNA sequencing rather than tiling arrays for detecting differences between immunoprecipitated and control DNA. This method has two advantages over ChIP-on-chip: (i) data normalization is less of an issue because sequencing results in absolute read counts rather than relative hybridization scores and (ii) recent progress in sequencing-by-synthesis methods (e.g. by Roche/454 and Illumina/Solexa) makes ChIP-seq highly cost-efficient. ChIP-seq shares ChIP-on-chip's dependence on high-quality antibodies and has the additional drawback that extra steps are required to restrict ChIP-seq to specific subregions of the genome. Nevertheless, its unparalleled throughput makes ChIP-seq the prime candidate for comprehensive human epigenome projects.
Bisulfite sequencing exploits the ability of bisulfite to selectively convert unmethylated cytosines into thymines, thereby transforming the DNA methylation state into a methylation-dependent SNP (Hajkova et al. 2002). The application of bisulfite sequencing is restricted to DNA methylation, for which it continues to be the gold standard due to its single-basepair resolution. DNA methylation patterns which are specific to a single cell can be obtained by combining bisulfite treatment with vector cloning and sequencing of a number of clones. However, for cost reasons bisulfite-treated DNA is often subjected to direct sequencing, which destroys any information about co-methylation in a particular cell but is sufficient for deriving profiles of average methylation.

Table 1. Methods for genome-wide mapping of epigenetic information

This table summarized the experimental concept as well as practical considerations of widely used experimental methods for epigenome mapping.

The key bioinformatic step of ChIP-seq (Table 1) is the fast and accurate mapping of short sequence reads to the reference genome. In principle, any seed-based alignment program such as blastn (http://www.ncbi.nlm.nih.gov/BLAST) or BLAT (Kent 2002) is applicable to this task. Nevertheless, seed alignment strategies that are specifically optimized for reads from a particular sequencing platform have been reported to yield substantial increases in speed and coverage (Synamatix Sdn. Bhd. 2007). Two commercial solutions for short ChIP-seq reads are currently available, namely the ELAND tool included in the Solexa analysis pipeline (http://www.solexa.com/) and the SXOligosearch tool (http://www.synamatix.com/). In addition, a customized alignment protocol has been developed at the Broad Institute (Mikkelsen et al. 2007). Unlike relative probe intensities in ChIP-on-chip, each sequence read in a ChIP-seq experiment directly corresponds to a single chromatin fragment that was bound by the antibody during immunoprecipitation. For this reason, it is commonly assumed that ChIP-seq requires almost no normalization and that data analysis can be based directly on sequence read counts (Barski et al. 2007) or sliding window read counts (Mikkelsen et al. 2007). However, an important caveat is that the process of mapping tags to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns. This is because short sequencing reads that (partially) overlap with low-complexity regions or with interspersed repeats stand a higher chance of being discarded for lack of unique genomic alignment.

Bisulfite sequencing (Table 1) requires customized analysis software that accounts for the "fifth base", 5-methyl-cytosine. When bisulfite-treated DNA is sequenced directly (i.e. without vector cloning), the average methylation levels can be estimated using the ESME software (Lewin et al. 2004). This software corrects for systematic bias induced by different molecular weights at methylation-specific SNPs and facilitates quality control. When subclones of bisulfite-treated DNA are sequenced, which is regarded as the gold standard for DNA methylation analysis, methylation patterns are inferred by aligning the clonal sequences to the genomic DNA sequence. The BiQ Analyzer software (Bock et al. 2005, cf. chapter C-2 of this thesis) has been developed to simplify this analysis, to perform stringent quality control and to visualize the results. In addition, specialized primer design programs exist, of which Methyl Primer Express (freely available from http://www.appliedbiosystems.com/) is probably the most widely used. However, manual refinement is often necessary, suggesting that further improvements of primer design programs are needed.

A-5 Epigenome data analysis

Rapid progress of experimental technologies has given rise to several epigenome mapping initiatives (Table 2). These projects have been breaking ground not only in terms of applying and improving large-scale experimental methods, but also in terms of developing bioinformatic methods for analyzing their data.

This is particularly true for the ENCODE project, which has been designed from the onset as a close cooperation between experimental and computational biologists. Although the ENCODE project aims to map functional elements in the human genome rather than to resolve epigenetic questions, the methods and tools that emerged from this project contribute to epigenome data analysis in a number of ways. First, a method for unsupervised segmentation of chromatin data was developed based on wavelet smoothing and hidden Markov models (Thurman et al. 2007). When applied to selected ChIP-on-chip datasets from the ENCODE pilot phase, the algorithm neatly recovered the two main chromatin states: open and transcriptionally competent euchromatin as well as inaccessible and transcriptionally silent heterochromatin. Second, the joint statistical analysis of all 105 ChIP-on-chip datasets from the

ENCODE pilot phase (Zhang et al. 2007a) provides an example of exploratory data analysis on a large and heterogeneous dataset that includes substantial amounts of epigenetic information. Third, several alternative prediction methods for annotating functional promoters were developed and evaluated (Trinklein et al. 2007), indicating that epigenetic data can substantially improve the accuracy of promoter annotation. Fourth, a rigorous statistical test was developed that assesses the significance of overlap between two sets of genomic features, for example between CpG islands and unmethylated genomic regions (ENCODE Project Consortium 2007). The authors show that – under relatively weak assumptions – their Genome Structure Correction method yields realistic *P*-values while other randomization-based methods tend to over-estimate significance. Fifth, the ENCODE project was accompanied by systematic incorporation of epigenome datasets into the UCSC Genome Browser (Thomas et al. 2007), which now provides integrated visualization and standardized retrieval of various genome and epigenome datasets. Finally, the successful collaboration of experimental and bioinformatic researchers in the ENCODE project has raised the awareness of synergies between wet-lab and computational research. The AHEAD task force, for example, acknowledges the critical importance of bioinformatic methods and infrastructure in their proposal for a human epigenome project (Alliance for Human Epigenomics and Disease 2007).

Although the bioinformatic focus of the other large-scale epigenome projects (Table 2) was less pronounced than in the ENCODE project, important bioinformatic progress arose from them as well. The HEROIC project played a catalyzing role for the development of epigenome data storage, visualization and analysis infrastructure in Europe. In fact, in its regulatory builds the Ensembl genome browser (Hubbard et al. 2007) will increasingly incorporate epigenetic information such as genome-wide maps of DNA methylation and histone modifications (P. Flicek, personal communication). The HEP project for the first time explored the challenges and opportunities of high-resolution epigenome analysis in multiple unrelated individuals (Eckhardt et al. 2006; Rakyan et al. 2004), enabling a computational study on inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping (Bock et al. 2008, cf. chapter C-4 of this thesis). And the two large-scale ChIP-seq projects that have been completed recently underline the relevance of analyzing various epigenetic mechanisms simultaneously in a single cell type (Barski et al. 2007) and at multiple stages during cellular differentiation (Mikkelsen et al. 2007). While the general picture emerging from these studies is consistent with mammalian epigenomes being segmented into alternating regions of open and condensed chromatin, many more sophisticated concepts become visible only at high resolution and when analyzing various epigenetic mechanisms simultaneously. For example, it has been shown recently that computational integration of several histone modification maps can be used to predict the locations of enhancers in the human genome, even where these are invisible to phylogenetic methods (Heintzman et al. 2007; Roh et al. 2007).

However, these pioneering epigenome mapping projects also highlight two major impediments to epigenome data analysis: the unsolved problem of public data storage and the lack of experimental standardization. Public data storage in databases such as GenBank and ArrayExpress has played an important role for bioinformatic research, by making primary data available for meta-analysis and benchmarking studies. However, with the advent of ChIP-seq, the central collection of primary data is reaching technical limitations. A typical three-day run on a Solexa sequencer gives rise to hundreds of gigabytes of primary image data and several gigabases of sequence reads, and in less than a year a single Solexa sequencer could generate the equivalent of all sequence data stored in GenBank until 2005. In addition to developing more efficient methods for data processing and storage, it will therefore be necessary to work out policies that regulate how primary data should be archived and how the benefits of publicly available primary data can be maintained when central storage is no longer an option. The second problem, lack of experimental standardization, hampers the computational integration of epigenetic datasets from different studies. Because epigenetic information is tissue-specific and because methods such as ChIP-on-chip are highly sensitive to variation in the experimental protocol, most epigenome datasets that have been published to date are – strictly speaking – incomparable. Nevertheless, several meta-analyses of ChIP-on-chip data have been published and significant correlations have been observed for epigenetic modifications that are associated with an open chromatin structure (Bock et al. 2007; Parisi et al. 2007; Zhang et al. 2007a), while an initial comparison for repressive histone modifications indicated substantially less correlation between different datasets (C. Bock, unpublished data). Although complete standardization is neither realistic nor desirable, it seems advisable to focus different epigenome mapping projects on the same set of cell lines, as is done in the ENCODE project.

Initiator	Summary	Current State	References
AHEAD Task Force (international)	The goal of the "Alliance for Human Epigenomics and Disease" (AHEAD) is to initiate and coordinate a comprehensive human epigenome mapping project. Initially, focus is set on developing a suitable bioinformatic infrastructure and on performing epigenome mapping in a selection of normal tissues, which may provide the reference for subsequent mapping in abnormal cells	In the May 2007 roadmap update, the NIH selected epigenetics as one of two roadmap initiatives to be started immediately. This decision was partially based on a proposal submitted by the AHEAD Task Force	(Alliance for Human Epigenomics and Disease 2007) (Jones and Martienssen 2005)
ENCODE Project Consortium (international)	The NIH-funded "Encyclopedia of DNA Elements" (ENCODE) project aims to map all functional elements in the human genome sequence. Although epigenome mapping is not its main goal, the project includes large-scale mapping of DNA methylation, histone modifications and other epigenetic information	The pilot project comprehensively analyzed 1% of the genome, with results published in June 2007. In the production phase, selected analyses are performed on the entire human genome	(ENCODE Project Consortium 2007) (ENCODE Project Consortium 2004)
HEP Project Consortium (UK/D/F)	The partially EU-funded "Human Epigenome Project" (HEP) analyzed DNA methylation in 43 unrelated individuals at single basepair resolution. Although the analysis was confined to selected regions on three chromosomes, it is the largest high-resolution, multi-individual epigenome dataset published to date	The results of the pilot phase dataset were published in 2004 and the results of the main phase were published in 2006	(Eckhardt et al. 2006) (Rakyan et al. 2004)
HEROIC Project Consortium (EU)	The "High-throughput Epigenetic Regulatory Organisation In Chromatin" (HEROIC) project is a multicenter EU project that applies ChIP-on-chip, chromosome interaction analysis and whole-genome nuclear localization assays to understanding human genome regulation	This EU 'Integrated Project' is funded is funded from 2005 to 2010 and does not involve synchronized pilot or production phases	(HEROIC Project Consortium 2005)
Broad Institute of MIT and Harvard (US)	In a large single-center study, ChIP-seq was used to derive genome-wide maps of chromatin state for mouse ES cells, neural progenitor cells and embryonic fibroblasts	Initial results were published in July 2007	(Mikkelsen et al. 2007)
National Heart, Lung, and Blood Institute of the NIH (US)	In a large single center study, ChIP-seq was used to derive genome-wide maps of chromatin state for human T cells	Initial results were published in June 2007	(Barski et al. 2007)

Table 2. Large-scale epigenome mapping projects

This table lists running and completed initiatives for large-scale epigenome mapping as of March 2008.

A-6 Epigenome prediction: inferring epigenetic states from the DNA sequence

A substantial amount of bioinformatic research has been devoted to the prediction of epigenetic information from characteristics of the genomic DNA sequence. Such predictions serve a dual purpose. First, accurate epigenome predictions can substitute for experimental data, to some degree, which is particularly relevant for newly discovered epigenetic mechanisms and for species other than human and mouse. Second, prediction algorithms build statistical models of epigenetic information from training data and can therefore act as a first step toward quantitative modeling of an epigenetic mechanism.

Promoter prediction – an important topic in bioinformatics since the early 1990s – can be regarded as the first attempt to predict epigenetic states from the DNA sequence. This is because active promoters are characterized by an open and transcriptionally permissive chromatin structure and exhibit specific epigenetic properties such as absence of DNA methylation and enrichment for histone acetylation. A large number of promoter prediction methods have been developed during the last two decades, most of which use DNA sequence characteristics combined with a machine learning algorithm to identify candidate promoters (see Bajic et al. 2004 for a comprehensive overview and benchmarking analysis). In the highly annotated human genome, promoter prediction has lost some of its relevance and researchers are increasingly focusing on advanced questions of transcription control, such as inferring tissue-specific signals (Smith et al. 2007) and reconstructing transcriptional networks (Bulcke et al. 2006).

CpG island prediction has some overlap with promoter prediction since the majority of promoters in mammalian genomes co-localize with CpG islands (Antequera 2003). However, CpG islands play a more general role as mediators of open chromatin structure, and they frequently overlap with enhancers and other regulatory elements. CpG islands were originally discovered by a striking absence of DNA methylation (Cooper et al. 1983), which is regarded as a constitutive feature of CpG islands. The absence of DNA methylation in the germline reduces CpG-to-TpG mutation rates inside CpG islands, leading to over-representation of CpGs relative to the genomic average. CpG islands are often predicted solely based on their GC and CpG frequencies, and multiple variants of the original definition (Gardiner-Garden and Frommer 1987) are in use. However, a recent study showed that these definitions yield high false positive rates, and a refined concept of bona fide CpG islands based on large-scale epigenome prediction was proposed (Bock et al. 2007, cf. chapter B-4 of this thesis).

DNA methylation prediction is conceptually easier than the prediction of more volatile epigenetic mechanisms because DNA methylation patterns exhibit relatively low tissue specificity compared to other epigenetic information. Therefore, it is not surprising that comparable prediction methods applied to DNA methylation data for blood (Bock et al. 2006, cf. section B-2 of this thesis) and brain tissue (Das et al. 2006; Fang et al. 2006) yielded similar results. In all three cases, machine learning methods were used to derive a classifier for presence or absence of DNA methylation in a given region. Prediction accuracies were high, and the most predictive attributes included CpG-rich sequence patterns (Bock et al. 2006; Das et al. 2006; Fang et al. 2006), specific DNA structure properties and repetitive DNA elements (Bock et al. 2006) as well as certain transcription factor binding sites (Fang et al. 2006). Interestingly, a similar method could also predict which genomic regions are prone to becoming methylated in a cell line overexpressing the DNA methyltransferase DNMT1 (Feltus et al. 2003).

Prediction of nucleosome positioning is based on the observation that the sequence composition of DNA molecules strongly affects their nucleosome affinity, i.e. how easily they can be wound around a nucleosome (Satchwell et al. 1986). Several recent papers showed that this in vitro effect has significant impact on the genomic positioning of nucleosomes in vivo (Ioshikhes et al. 2006; Peckham et al. 2007; Segal et al. 2006). Although all three papers focus their analysis on yeast, the highly conserved nature of the nucleosome suggests a general applicability of these results. Indeed, Segal et al. observe that the predictions change little when training is performed on nucleosome positioning data from chicken instead of yeast, and Ioshikhes et al. find that an alignment of multiple yeast species can increase prediction accuracy.

Successful prediction has also been reported for several other epigenetics-related phenomena: DNase I hypersensitive sites could be distinguished from a random control set using support vector machines with *k*-mer sequence motifs as prediction attributes (Noble et al. 2005). Polycomb/trithorax response elements in Drosophila were identified by specific sequence criteria (Ringrose et al. 2003), a finding that may not easily translate to humans since mammalian Polycomb/trithorax response elements exhibit less identifiable sequence patterns. Imprinted genes were predicted using a wide range of genomic features (sequence motifs, CpG islands, repeats, predicted transcription factor binding sites) and a commercial support vector machine-based data mining suite (Luedi et al. 2005). Finally, genes that escape X-chromosome inactivation were predicted by a support vector machine and found to be enriched in Alu repeats and CpG-rich sequence motifs (Wang et al. 2006). However, a conclusive assessment of prediction methods for imprinted genes and for genes that escape inactivation seems problematic due to the small number of affected genes, their clustering in small genomic regions and the difficulty of independent experimental validation.

In summary, a large number of genomic regions exhibit clearly detectable epigenetic footprints in their DNA sequence. This has practical applications for genome annotation and also challenges the notion of *genome* and *epigenome* as two largely independent systems of inheritance working at different time scales. Rather, the genome seems to encode not only genes and cis-regulatory elements, but also a default epigenetic state that becomes active in the absence of other regulatory influences such as the binding of transcription factors or the activity of chromatin remodeling complexes. This interpretation is consistent with the emerging concept of multi-tasking genomes, which simultaneously (and on top of each other) encode genes and their regulation (Kapranov et al. 2007). Furthermore, this model provides an explanation for the fact that only a small subset of suitable consensus binding motifs are actually used by transcription factors in vivo. A new generation of in silico methods for detecting transcription factor binding has already started to benefit from epigenome prediction in order to distinguish functional from non-functional sites (Narlikar et al. 2007).

A-7 Cancer epigenetics: toward improved diagnosis and therapy

It has been known for a long time that mutations and chromosomal deletions can irreversibly destroy tumor suppressor genes and are pivotal events in cancer progression. In contrast, the importance of epigenetic mechanisms for tumor development has been appreciated more recently (see Feinberg and Tycko 2004, for a historical account of cancer epigenetics). It is now clear that a substantial proportion of silenced tumor suppressor genes are lost due to epigenetic deactivation rather than genomic damage (Esteller 2007; Jones and Baylin 2007). Furthermore, a comparison between the epigenetic characteristics of cancer cells and stem cells suggests that epigenetic deregulation may program cells for cancer-like behavior long before they are visually identifiable as tumor cells (Feinberg et al. 2006).

The important role of epigenetic defects for cancer opens up new opportunities for improved diagnosis and therapy. Early diagnosis profits from the fact that epigenetic aberrations occur early during tumorigenesis and are frequently detectable in peripheral blood when destroyed tumor cells leak DNA into the bloodstream (Laird 2003). Epigenetic cancer therapy exploits the fact that – in contrast to genomic damage – epigenetic aberrations are pharmacologically reversible (Yoo and Jones 2006). These active areas of research give rise to two questions which are particularly amenable to bioinformatic analysis. First, given a list of genomic regions exhibiting epigenetic differences between tumor cells and controls (or between different disease subtypes), can we detect common characteristics and infer a functional link between these regions and cancer? Second, can we use bioinformatic methods in order to improve diagnosis and therapy by detecting and classifying important disease subtypes?

Keshet et al. faced a typical instance of the first question, after MeDIP analysis in two cancer cell lines and in a set of primary tumors had detected hundreds of genes whose promoters were selectively methylated in cancer (Keshet et al. 2006). They applied several bioinformatic methods in order to identify common characteristics of these genes, including over-representation analysis of Gene Ontology terms, sequence motif discovery, genomic clustering analysis and comparison with public gene expression data. Based on these computational analyses they concluded that only a small percentage of epigenetically silenced genes in cancer cells are tumor suppressor genes. In contrast, many of the genes that are unlikely to be tumor suppressor genes exhibit certain DNA sequence patterns, which may predispose them for epigenetic silencing – as a side effect rather than cause of tumor development. A recent study elaborated on this finding by applying a more advanced motif discovery pipeline and could identify additional sequence motifs on the same dataset (Eden et al. 2007). The observation that epigenetically silenced genes often share certain sequence motifs in their promoters has also been used in order to detect new candidates for cancer-specific hypermethylation (Goh et al. 2007). To address the substantial class bias – only a small percentage of genes become hypermethylated in a particular cancer – and the lack of an experimental control set, Goh et al. devised an algorithm that iteratively combines unsupervised clustering and supervised prediction. Furthermore, the recent discovery of a link between DNA hypermethylation in cancer and Polycomb binding in ES cells using a combination of bioinformatic comparisons and experimental validation (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007) highlights the synergistic power of computational and experimental methods in cancer epigenetics. Future studies toward understanding the epigenetic characteristics of cancer cells will benefit from the recently launched PubMeth database, which aggregates literature data about which genes have been reported hypermethylated for which cancer (Ongenaert et al. 2007).

The second question is aimed at the discovery and validation of biomarkers for cancer diagnosis, prognosis and therapy optimization (Laird 2003). In an early study on DNA methylation patterns in leukemia, support vector machines applied to DNA methylation microarray data could accurately distinguish between two important disease subtypes, acute lymphoblastic leukemia and acute myeloid leukemia (Model et al. 2001). In a series of papers, Siegmund and coworkers developed (Marjoram et al. 2006; Siegmund et al. 2004) and applied (Weisenberger et al. 2006) clustering methods for unsupervised discovery of epigenetically distinct cancer subtypes. They could show that a well-defined subgroup of colon cancer patients exhibit a substantially elevated frequency of promoter hypermethylation, and they developed a biomarker for diagnosing this disease subtype. Epigenetic biomarkers also play an increasing role for therapy optimization. For example, clinical trials showed that cancer-specific DNA methylation of the MGMT gene promoter can make glioblastomas (brain tumors) more susceptible to chemotherapy with alkylating agents (Hegi et al. 2005). A combination of bioinformatic methods and experiments was recently used to optimize DNA methylation analysis of MGMT and to develop it into a routine clinical biomarker for personalized cancer therapy (Mikeska et al. 2007, cf. chapter D-3 of this thesis). However, in spite of the fast progress in epigenetic cancer diagnosis, few epigenetic cancer biomarkers have yet been validated in large patient cohorts and substantial work remains to be done before epigenetic cancer diagnosis will start having a measurably positive effect on disease burden in the population.

A-8 Outline of the remainder of this thesis

The remainder of this thesis is structured as follows: Part B describes methods development for epigenome prediction using machine learning algorithms and the application of these methods for improving CpG island annotation. Part C focuses on methods and software that support experimental mapping of DNA methylation, as well as on computational analysis of large-scale DNA methylation datasets. Part D describes two case studies applying bioinformatic methods to cancer epigenetics. Finally, Part E concludes the thesis by summarizing its key results, namely a model of the genome and epigenome as two interdependent and tightly correlated carriers of biological information and a bioinformatic pipeline for cancer biomarker discovery, optimization and validation.

Readers who are interested in only one of these two results may want to follow a more targeted reading strategy:

(4) *Identification of a globally high degree of correlation between genome and epigenome*. The reader could skim through the general overview of computational epigenetics (Part A) and its chapter A-6 on epigenome prediction, before moving on to Part B. The introduction of Part B (chapter B-1) motivates why epigenome prediction is arguably the most immediate approach to detecting a globally high degree of interdependence between the human genome and epigenome. The chapters on DNA methylation prediction (chapter B-2) and on the prediction of a wide range of other epigenetic modifications (chapter B-4) provide insights into which aspects of the genomic DNA sequence are most relevant for predicting epigenetic states. Also of interest might be the analysis of high-resolution DNA methylation profiles (chapter C-3), which shows that the genomic basis of DNA methylation is not adequately described by simple consensus binding motifs, and the analysis of inter-individual variation of DNA methylation (chapter C-4), which highlight that the genomic DNA sequence predicts not only the absolute levels of DNA methylation but also their degree of variation. Readers interested in analyzing the interplay between genome and epigenome by themselves are also referred to the description of the EpiGRAPH web service (chapter B-3). Finally, section E-2.1 summarizes the different facets of genome-epigenome interactions analyzed in this thesis.

(5) *A bioinformatic pipeline to facilitate the identification of DNA methylation biomarkers and their conversion into molecular diagnostic tools for clinical use.* The reader could start from Figure 47 and the accompanying text in section E-2.2, which outline how the different methods and software tools developed in this thesis integrate into a workflow for discovery and optimization of DNA methylation biomarkers. Researchers who are interested in the discovery phase of the pipeline could focus their reading on chapter A-7 for a general introduction into cancer epigenetics, on chapter B-3 for an overview of EpiGRAPH and on chapter B-4 for an example of how to apply EpiGRAPH in order to identify regulatory regions that may undergo epigenetic silencing in cancer. On the other hand, researchers who are interested in the optimization phase of the pipeline are recommended to focus on the case study provided in chapter D-3 and to refer to the description of BiQ Analyzer (chapter C-2) as well as the outline of MethMarker and BiomarkerSpace (chapter E-3.1) as supplementary reading.

Part B. Epigenome Prediction

"The major problem, I think, is chromatin. [...] you can inherit something beyond the DNA sequence. That's where the real excitement of genetics is now" (James D. Watson)[1]

B-1 Outline

For practical reasons, genetic and epigenetic mechanisms of gene regulation are frequently studied independently from each other, which has sometimes resulted in what might be considered an over-emphasis on the differences between the two. The goal of the following chapters is to highlight the globally high degree of interdependence between the human genome and epigenome in arguably the most immediate way – by predicting patterns of epigenetic modifications from the genomic DNA sequence.

In chapter B-2, we show that DNA methylation at CpG islands – i.e. in those regions where it directly influences gene expression – can be predicted with high accuracy based on the genomic DNA sequence. To that end, a bioinformatic prediction method is developed and applied to two DNA methylation datasets (Bock et al. 2006). Bioinformatic and experimental validation indicate that prediction accuracies close to 90% are achievable in healthy cells, a finding that has been confirmed by independent studies (Das et al. 2006; Fang et al. 2006). Building upon the bioinformatic methods prototyped for DNA methylation prediction, in chapter B-3 we describe the development of the EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/), which enables biologists to perform complex epigenome predictions without the need to write custom scripts. The practical utility of EpiGRAPH for the analysis of real-world biological problems is highlighted by two case studies, which focus on the epigenetic states of ultraconserved elements and on genes that are preferentially expressed from a single allele. Next, chapter B-4 provides evidence that epigenome prediction is not restricted to DNA methylation, but also feasible for a diverse set of epigenetic modifications that are indicative of an open and transcriptionally active chromatin state. In an attempt to reconcile the epigenetic regulatory function of CpG islands with their sequence-based annotation mode, we predict the epigenetic states of all CpG islands in the human genome and identify a specific subset of CpG islands with a distinctive role in epigenetic gene regulation (Bock et al. 2007). Finally, chapter B-5 addresses two conceptual problems of current CpG island annotations, namely that the definition is underdetermined and that current CpG island finders overlook valid CpG islands. We provide a formal definition of the CpG island annotation problem that resolves the ambiguity problem, and we propose an algorithm for CpG island annotation that finds the optimal solution according to our definition. The correctness of this algorithm is proven.

B-2 Predicting DNA methylation based on the genomic DNA sequence[2]

B-2.1 Motivation

DNA methylation is an important mechanism of epigenetic regulation (cf. section A-3 of this thesis). In vertebrates, DNA methylation is largely confined to cytosines in a CpG context.

[1] Quoted after: Goldberg, A.D., C.D. Allis, and E. Bernstein. 2007. Epigenetics: a landscape takes shape. *Cell* **128:** 635-638.

[2] This chapter describes published work conducted in collaboration with Martina Paulsen, Sascha Tierling, Thomas Mikeska and Jörn Walter (Bock et al. 2006). Sascha Tierling and Thomas Mikeska performed and evaluated the bisulfite sequencing experiments. Martina Paulsen and Jörn Walter contributed to the interpretation of the results.

The classical view is that almost all dispersed CpG dinucleotides in the human genome are methylated by default, whereas CpG dinucleotides inside CpG island promoters are typically unmethylated in normal (i.e. non-neoplasic, non-senescent) tissue (Bird 2002). However, exceptions have been known for a long time, such as de novo methylation during cell differentiation (Arney and Fisher 2004), imprinting (Reik et al. 2003), and X-chromosome inactivation (Heard 2004). Biallelic DNA methylation of CpG island promoters is associated with stable silencing of neighboring or associated genes in most cases and constitutes a frequent event in cancer development (Feinberg and Tycko 2004).

Initial genome-scale studies of CpG island methylation indicate that a sizeable fraction of CpG islands are methylated in normal tissue (Weber et al. 2005; Yamada et al. 2004). However, little is known about the mechanisms that lead to methylation of certain CpG islands while leaving others unmethylated, and it is unclear whether these two groups can be identified by characteristic DNA attributes. Inspired by recent exploratory results pointing toward a significant role of local DNA sequences in determining DNA methylation at individual CpGs (Bhasin et al. 2005; Handa and Jeltsch 2005), as well as for aberrant DNA methylation (Feltus et al. 2003), we performed a comprehensive analysis of the association between DNA-related features and normal CpG island methylation on human chromosome 21.

B-2.2 Methods

Study design

Based on a dataset published by Yamada et al., comprising a substantial number of CpG islands on the non-repetitive parts of human chromosome 21 (Yamada et al. 2004) and a compiled list of 1,184 DNA-related attributes, we quantify the correlation between CpG island methylation and eight attribute classes: (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, (3) CpG island frequency and distribution, (4) predicted DNA structure, (5) gene and exon distribution, (6) predicted transcription factor binding sites, (7) evolutionary conservation, and (8) single nucleotide polymorphisms (SNPs). We identify the attributes that are most predictive in distinguishing between methylated and unmethylated CpG islands and we show that it is possible to predict CpG island methylation from DNA-related features with high accuracy. Finally, we validate our results both experimentally on chromosome 21 and bioinformatically on data from the Human Epigenome Project (Rakyan et al. 2004).

DNA methylation data

This analysis is based on the results of a comprehensive measurement of CpG island methylation on human chromosome 21 (Yamada et al. 2004). Briefly, Yamada et al. repeat-masked the chromosome sequence and computationally identified non-repetitive CpG islands using standard tools and parameters (GC content above 50%, ratio of observed vs. expected number of CpG dinucleotides above 0.6, more than 400 base pairs in length). Next, they designed primers for each identified CpG island and extracted corresponding DNA from samples of human peripheral blood lymphocytes. Finally, they determined the methylation status of each CpG island by methylation-specific restriction enzymes (via HpaII-McrBC-PCR). Yamada et al. validated their method by bisulfite sequencing of selected CpG islands and concluded that it is highly reliable.

Their dataset comprises the methylation status of 149 CpG islands, each belonging to one of the following categories: fully methylated, unmethylated, incompletely methylated, or compositely/differentially methylated. Exploratory analysis using bisulfite sequencing indicated that the latter two classifications were not always unambiguous (T. Mikeska and S. Tierling, personal communication); therefore we focused on the two well-defined categories, fully methylated (31 cases) and unmethylated (103 cases). In order to minimize potential error sources, we re-mapped the boundaries of the CpGs islands that were originally used by Yamada et al. to the hg17 (NCBI35) assembly of the human genome and we excluded two cases (both belonging to the fully methylated class) from the analysis because, in the light of this new mapping, the primers did not pick the intended CpG islands. Therefore, our dataset comprised 132 independent CpG islands, which are distributed relatively evenly across chromosome 21 (data available online: Bock et al. 2006, Table S2).

For validation, we also used data from the HEP pilot study (Rakyan et al. 2004). In this study, Rakyan et al. determined the methylation status of 3,273 unique CpG dinucleotides (belonging to 253 amplicons) across seven tissues and one to eight samples per tissue by means of bisulfite direct sequencing. Out of these 253 amplicons, 210 could be mapped unambiguously to the hg17 (NCBI35) assembly of the human genome and had at least one measurement for each tissue. For these amplicons, we calculated average CpG dinucleotide methylation levels, both separately for individual tissue types and for all tissues combined. Those amplicons below an – arbitrarily chosen – threshold of 60% methylation were marked as unmethylated and those above this threshold were marked as methylated, resulting in a dataset of 163 "methylated" and 47 "unmethylated" amplicons.

DNA-related attributes

In order to identify DNA sequence-related attributes that are correlated with CpG island methylation, we compiled a comprehensive list of attributes that can be linked directly or indirectly to the genomic DNA sequence (the full list is available online: Bock et al. 2006, Table S1). Most attributes take the form of frequencies or numerical scores, averaged over sequence windows and standardized to a default window size of one kilobase. They fall into eight biological classes, namely: (1) DNA sequence properties and patterns (428 attributes), (2) repeat frequency and distribution (494 attributes), (3) CpG island frequency and distribution (16 attributes), (4) predicted DNA structure (28 attributes), (5) gene and exon distribution (60 attributes), (6) predicted transcription factor binding sites (135 attributes), (7) evolutionary conservation (ten attributes), and (8) SNPs (13 attributes). The data for most of these attributes were collected from annotation tracks in the UCSC Genome Browser (Karolchik et al. 2008). However, the attributes for class (1) were directly calculated from DNA sequence and the attributes for class (4) were calculated from DNA sequence by averaging over octamers (Gardiner et al. 2003) and trimers (J. Greenbaum, personal communication), respectively. We calculated these attributes for each CpG island in our dataset, both for the re-mapped CpG island itself and for 11 sequence windows around the CpG island: -20 kb to -10 kb, -10 kb to -5 kb, -5 kb to -2 kb, -2 kb to -1 kb, -1 kb to left boundary of CpG island, CpG island, right boundary of CpG island to 1 kb, 1 kb to 2 kb, 2 kb to 5 kb, 5 kb to 10 kb, 10 kb to 20 kb. Next, we removed those attributes that were zero in (almost) all cases (e.g. binding sites of rare transcription factors), giving us a list of 918 prediction attributes (the Python code used for feature selection is available on request). For the CpG island level statistics (see next section), only the 706 attributes with non-zero values in the CpG island window of at least five methylated and five unmethylated cases were retained. For the sequence neighborhood statistics, only the 833 attributes were retained that had non-zero values in at least five methylated and five unmethylated cases, for at least four out of the 11 sequence windows (it could be argued that this attribute selection step inadequately reduces the *n* used for multiple testing correction. However, this effect is overly compensated by using a highly conservative 1% significance threshold).

Statistics

We performed statistical tests in order to determine attributes that exhibit significantly different values for fully methylated CpG islands compared to unmethylated CpG islands, at two levels. First, we compared all attributes at the CpG island level using the nonparametric Wilcoxon rank sum test (data available online: Bock et al. 2006, Table S1, first worksheet). Second, we compared all attributes across the complete sequence neighborhood of -20 kb to +20 kb around the CpG island (data available online: Bock et al. 2006, Table S1, second worksheet). To that end, quadratic regression functions were fitted over the attribute values in the 11 sequence windows around the CpG island (see previous section) and we used the analysis of variance (ANOVA) statistic to assess whether separate fitting for unmethylated vs. methylated cases resulted in a significantly decreased error compared to combined fitting (quadratic regression functions were chosen to capture symmetry around the CpG island). A caveat of the latter approach is its high sensitivity to violations of the normality assumption, hence we interpret only results that are confirmed by both test statistics.

All significance thresholds were adjusted for multiple testing using the highly conservative Bonferroni method. Technically speaking, we controlled the family-wise error rate to be less than 1%. The very strict correction for multiple testing provides an additional safety margin against false discoveries, which are a common problem in studies with small sample sizes and large numbers of features.

Prediction

Machine learning methodology was used for two tasks: (i) to quantify the correlation between CpG island methylation and several classes of DNA-related attributes and (ii) to predict CpG island methylation from the local genomic neighborhood. The technical procedure (cross-validation) is similar in both cases and is discussed below. However, intention and interpretation differ for the two tasks. Task (ii) is the classical prediction scenario: given a dataset of limited size, we want to train a classifier for predicting CpG island methylation on unknown data and to quantify its expected prediction performance. Therefore, we train the classifier on the full set of 918 attributes, assuming that at least some of these attributes contain information that may be useful for the classifier. In task (i), the goal is not so much to predict new data but to understand existing data. Here, we use a classifier as a tool for quantifying the relationship between an attribute class (e.g. DNA sequence properties or repeats) and CpG island methylation. The rationale behind this approach is straightforward: If a classifier can successfully and reliably predict CpG island methylation using only information from one particular attribute class, then the attributes in this class are interpreted to be biologically associated with CpG island methylation and the prediction performance is used as a measure of the degree of biologically association.

All prediction experiments follow essentially the same procedure. Given the list of CpG islands or amplicons and any selection of attributes from our list, a linear support vector machine (SVM) is repeatedly trained to predict methylation status based on a 90% subset of CpG islands, and its performance is evaluated on the remaining 10% of unseen cases. Technically speaking, we repeat 10-fold stratified cross-validation 20 times with different random partitions and sum the results on the test set (in terms of true negatives, false negatives, false positives, and true positives). The prediction performance is measured as the correlation coefficient between the predictions and the correct values on the test set. This criterion is commonly viewed as superior to comparing prediction accuracies because it is not as strongly affected by unbalanced class distributions (Baldi et al. 2000).

For most prediction experiments (prediction setup A in Table 4 below), we used the linear SVM implementation of the WEKA package (Witten and Frank 2000), which is based on the sequential minimal optimization method (Platt 1999). Additionally, several control experiments were performed that use different algorithms: an SVM with radial basis function kernel (from WEKA package, prediction setup B), AdaBoost M1 with decision tree stumps as the underlying classifier (from WEKA package, prediction setup C), the C4.5 tree generator (from WEKA package, prediction setup D), and a different implementation of a linear SVM (R implementation of LIBSVM (Chang and Lin 2005), prediction setup E). All algorithms were applied with their suggested standard parameters.

Experimental verification

Predictions were performed using a linear SVM that was trained on the full chromosome 21 dataset (132 cases) and all attribute classes. Subsequently, we experimentally determined the methylation status of 12 selected CpG islands by bisulfite sequencing as follows. First, we applied direct sequencing of the PCR product to all 12 CpG islands. In nine cases, this produced unambiguous results (i.e. very high conversion of CpGs = unmethylated, or almost no conversion = methylated). Second, in the three remaining cases with mixed CG/TG sequencing profiles, PCR products were cloned and individual clones were sequenced in order to determine the methylation status. Average methylation was scored from single clone sequences using the BiQ Analyzer software (Bock et al. 2005, cf. chapter C-2 of this thesis). Details of the experimental setting and the primers that we used are available online (Bock et al. 2006, Protocol S1). Human peripheral blood was obtained with the written consent of the donor.

B-2.3 Results

Identification of DNA-related attributes that distinguish methylated CpG islands from their unmethylated counterparts

As a first step toward understanding the relationship between DNA attributes and CpG island methylation, we statistically compared the distributions between methylated and unmethylated CpG islands for all attributes in our list. Using a conservative significance threshold, 41 attributes showed significant differences (Table 3). Of the significant attributes, the majority are frequencies of GC-rich and CpG-rich DNA sequence patterns, which are over-represented in unmethylated CpG islands. Non-strand-specific patterns and patterns that are strand-specific relative to the chromosomal plus-strand occur with similar frequency and composition. Several attributes that refer to repetitive DNA are more frequent in methylated CpG islands (such as segmental duplications, self chain alignments, and tandem repeats).

Rank	Attribute Name	Attribute Description	Attribute Class	Higher Value for	Single Test Significance
1	SAI_len	Total length of self-alignments (alignments of the human genome against itself)	(2)	Methylated CpG Islands	2.62×10^{-11}
2	SAI_no	Total number of self-alignments	(2)	Methylated CpG Islands	3.23×10^{-10}
3	Pat_CCGC	Chromosome plus-strand pattern frequency of CCGC	(1)	Unmethylated CpG Islands	5.18×10^{-10}
4	Pat_CCCC	Chromosome plus-strand pattern frequency of CCCC	(1)	Unmethylated CpG Islands	1.39×10^{-9}
5	SAI_std	Standard deviation of self-alignment lengths	(2)	Methylated CpG Islands	1.96×10^{-9}
6	Uni_AAAG	Non-strand-specific pattern frequency of AAAG/CTTT	(1)	Unmethylated CpG Islands	8.87×10^{-9}
7	fC_std	Standard deviation of C content distribution	(1)	Unmethylated CpG Islands	9.13×10^{-9}
8	Rise_avg	Average DNA structure rise (as predicted from sequence)	(4)	Methylated CpG Islands	3.82×10^{-8}
9	Pat_CGCC	Chromosome plus-strand pattern frequency of CGCC	(1)	Unmethylated CpG Islands	5.05×10^{-8}
10	Pat_AAAG	Chromosome plus-strand pattern frequency of AAAG	(1)	Unmethylated CpG Islands	7.72×10^{-8}
11	Roll_skew	Skewness of DNA structure roll distribution (as predicted from sequence)	(4)	Unmethylated CpG Islands	1.15×10^{-7}
12	Pat_CTCC	Chromosome plus-strand pattern frequency of CTCC	(1)	Unmethylated CpG Islands	1.46×10^{-7}
13	fCG_std	Standard deviation of CpG content distribution	(1)	Unmethylated CpG Islands	2.15×10^{-7}
14	Pat_TCCC	Chromosome plus-strand pattern frequency of TCCC	(1)	Unmethylated CpG Islands	2.57×10^{-7}
15	SDu_no	Total number of sequential duplications	(2)	Methylated CpG Islands	3.49×10^{-7}
16	SAI_sco	Average self-alignment score	(2)	Methylated CpG Islands	4.19×10^{-7}
17	Pat_CTTT	Chromosome plus-strand pattern frequency of CTTT	(1)	Unmethylated CpG Islands	4.23×10^{-7}
18	Uni_CGGA	Non-strand-specific pattern frequency of CGGA/TCCG	(1)	Unmethylated CpG Islands	5.15×10^{-7}
19	Uni_CCGC	Non-strand-specific pattern frequency of CCGC/GCGG	(1)	Unmethylated CpG Islands	9.08×10^{-7}
20	Pat_CGGA	Chromosome plus-strand pattern frequency of CGGA	(1)	Unmethylated CpG Islands	1.16×10^{-6}
21	Pat_GCCG	Chromosome plus-strand pattern frequency of GCCG	(1)	Unmethylated CpG Islands	1.46×10^{-6}
22	Uni_AAGG	Non-strand-specific pattern frequency of AAGG/CCTT	(1)	Unmethylated CpG Islands	1.58×10^{-6}
23	Pat_CCCG	Chromosome plus-strand pattern frequency of CCCG	(1)	Unmethylated CpG Islands	1.86×10^{-6}
24	SAI_avg	Average length of self-alignments	(2)	Methylated CpG Islands	1.91×10^{-6}
25	Pat_TCCG	Chromosome plus-strand pattern frequency of TCCG	(1)	Unmethylated CpG Islands	2.60×10^{-6}
26	Pat_CGCG	Chromosome plus-strand pattern frequency of CGCG	(1)	Unmethylated CpG Islands	2.65×10^{-6}
27	Uni_CGCG	Non-strand-specific pattern frequency of CGCG/CGCG	(1)	Unmethylated CpG Islands	2.65×10^{-6}
28	Pat_ACCC	Chromosome plus-strand pattern frequency of ACCC	(1)	Unmethylated CpG Islands	2.87×10^{-6}
29	Uni_CAAA	Non-strand-specific pattern frequency of CAAA/TTTG	(1)	Unmethylated CpG Islands	2.90×10^{-6}
30	Pat_CAAA	Chromosome plus-strand pattern frequency of CAAA	(1)	Unmethylated CpG Islands	3.01×10^{-6}
31	Uni_CGGC	Non-strand-specific pattern frequency of CGGC/GCCG	(1)	Unmethylated CpG Islands	3.46×10^{-6}
32	Pat_GCCC	Chromosome plus-strand pattern frequency of GCCC	(1)	Unmethylated CpG Islands	3.95×10^{-6}
33	Pat_GGAA	Chromosome plus-strand pattern frequency of GGAA	(1)	Unmethylated CpG Islands	5.93×10^{-6}
34	Pat_TATT	Chromosome plus-strand pattern frequency of TATT	(1)	Unmethylated CpG Islands	6.43×10^{-6}
35	Pat_CCGG	Chromosome plus-strand pattern frequency of CCGG	(1)	Unmethylated CpG Islands	7.21×10^{-6}
36	Uni_CCGG	Non-strand-specific pattern frequency of CCGG/CCGG	(1)	Unmethylated CpG Islands	7.21×10^{-6}
37	Tan_sco	Goodness of fit score of tandem repeats	(2)	Methylated CpG Islands	9.34×10^{-6}
38	Uni_CACC	Non-strand-specific pattern frequency of CACC/GGTG	(1)	Methylated CpG Islands	9.55×10^{-6}
39	Tan_avg	Average lengths of tandem repeats	(2)	Methylated CpG Islands	9.60×10^{-6}
40	RC1_Low_	Alignment score of low complexity class repeats	(2)	Unmethylated CpG Islands	1.37×10^{-5}
41	RF1_Low_	Alignment score of low complexity family repeats	(2)	Unmethylated CpG Islands	1.37×10^{-5}

Table 3. DNA-related attributes differ significantly between methylated and unmethylated CpG islands

This table lists all attributes with significantly different distribution among methylated and unmethylated CpG islands, respectively, according to a Wilcoxon test with Bonferroni correction for multiple testing and an overall significance threshold of 1% (data for non-significant attributes is available online: Bock et al. 2006, Table S1). The rightmost column displays single-test *P*-values, the significance threshold after multiple testing correction is 0.01/706 = 1.42×10^{-5}. Attributes with significantly higher values in fully methylated CpG islands are in green. Attributes in red are significantly higher in unmethylated CpG islands. Detailed information on attribute definitions is available online (Bock et al. 2006, Table S3).

Interestingly, two aspects of predicted DNA structure, most prominently the average rise of the DNA helix, also show different distributions for methylated and unmethylated CpG islands (see Olson et al. 2001 for an overview of DNA structure nomenclature). The role of predicted DNA structure becomes even more pronounced when considering not only the CpG island itself, but also the -20-kb to +20-kb sequence windows surrounding it. In that case, the predicted average rise and the predicted average twist are the second and third most significant among all attributes (data available online: Bock et al. 2006, Table S1, second worksheet). An inspection of the corresponding boxplots (Figure 2) shows that the predicted DNA rise increases on average within CpG islands relative to the genomic neighborhood, whereas the twist decreases. However, this effect is much stronger for methylated than for unmethylated CpG islands. Hence, methylated CpG islands tend to co-locate with areas of unusual predicted DNA structure.

Furthermore, it is apparent from Table 3 that a single pattern is over-represented in methylated CpG islands, namely the non-strand-specific CACC/GGTG pattern. Because this pattern contains a TpG, in contrast to the CpG-rich patterns that are frequent in unmethylated CpG islands, it is tempting to argue that this pattern may be the result of sporadic deamination of original GG^{M}CG patterns, as such mutations are less likely to be repaired for methylated CpGs (Caiafa and Zampieri 2005). In order to test whether differential CpG → TpG mutation rates may be a source of differential pattern frequencies between methylated and unmethylated CpG islands, we compared the palindromic pattern CGCG with the non-strand-specific pattern TGTG/CACA, which can evolve from the former pattern by two subsequent deamination mutations.

In agreement with our hypothesis, the CGCG pattern is found more frequently in unmethylated CpG islands (mean of 12.61 occurrences per kb) than in methylated CpG islands (7.15 occurrences per kb) and the TGTG / CACA pattern more frequently in methylated CpG islands (10.92 occurrences per kb) than in unmethylated CpG islands (2.93 occurrences per kb). In both cases, *P*-values were below 0.001 according to a Wilcoxon test. These results suggest that during evolution, higher rates of germline CpG → TpG mutation occurred in those CpG islands that are methylated in human lymphocytes compared to those that are unmethylated.

Finally, we analyzed the dataset for evidence of experimental bias. Because restriction enzyme digestion was used to discriminate between methylated and unmethylated CpG islands (Yamada et al. 2004), incomplete digestion is a potential error source. In this case, we would expect the HpaII recognition site (CCGG) to behave differently from patterns that are never cut (e.g. being more strongly enriched). However, we observe no indication of this in our attribute statistics (data available online: Bock et al. 2006, Table S1, first worksheet). Five out of ten GC-rich and CpG-containing sequence patterns have higher *P*-values than the CCGG pattern (CCGC, CGCC, GCCG, CCCG, and CGCG), while the same number of patterns have a lower *P*-value (GCGC, CGGC, GCGG, CGGG, GGCG). We conclude that the experimental method that was applied by Yamada et al. is sufficiently unbiased for our analysis.

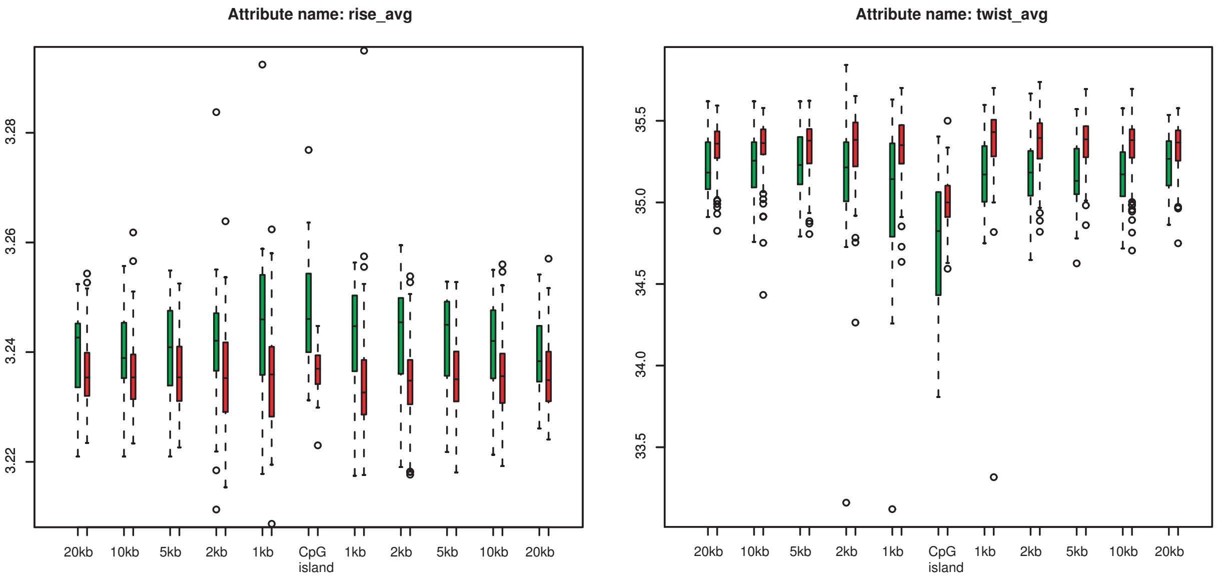


Figure 2. Predicted DNA structure differs in the neighborhood of methylated CpG islands compared with their unmethylated counterparts

The diagram on the left shows boxplots of the predicted DNA rise distribution over the CpG island and the ten sequence windows from −20 kb to 20 kb surrounding the CpG island (averaged over all 132 CpG islands in the chromosome 21 dataset). Green bars correspond to methylated CpG islands, red bars to unmethylated CpG islands. The diagram on the right shows similar information for the predicted DNA twist.

Quantification of the association between DNA-related attributes and CpG island methylation
Strikingly, all attributes that were significantly different between methylated and unmethylated CpG islands (Table 3) fall into three (out of eight) attribute classes: (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, and (4) predicted DNA structure. In order to investigate this observation more systematically, we calculated the group-wise correlation between CpG island methylation and each of the eight attribute classes.

In contrast to single-attribute correlation coefficients, group-wise correlations are able to capture combined effects of interacting attributes (e.g. when neither A nor B has any significant impact on methylation alone, but the combined presence of both is highly associated with a certain methylation status). Support vector machines (SVMs) have been successfully employed to detect such joint effects. Therefore, we trained a (linear) SVM to predict CpG island methylation and tested its performance on unseen data (10-fold cross-validation). Then, we calculated the correlation coefficient between the SVM's predictions on unseen data and the correct values, averaging over 20 independent cross-validation runs. This measure gives us a conservative estimate for the group-wise correlation between the attribute group and CpG island methylation – conservative because it may well be that the SVM does not capture all information on CpG island methylation that is present in the attribute group, while it is highly unlikely to predict methylation correctly over multiple runs if not enough information is contained in the attributes.

Our results substantiate the observation that three classes of DNA-related attributes are distinctly associated with CpG island methylation status (Table 4, experiments 1 to 8). (1) DNA sequence properties and patterns as well as (2) repeat frequency and distribution are correlated with CpG island methylation at medium to high rates (correlation coefficient of 0.635 and 0.657, respectively), whereas (4) predicted DNA structure falls behind (0.486). Three of the remaining attribute classes exhibit weak correlation with CpG island methylation, namely (5) gene and exon distribution (0.300), (8) SNPs (0.286), and (3) CpG island frequency and distribution (0.045). (6) predicted transcription factor binding sites and (7) evolutionary conservation are uncorrelated with CpG island methylation (-0.021 and 0.000, respectively). Furthermore, the combination of all eight attribute classes results in a higher correlation value than any single class (0.740), indicating that at least some attribute classes capture complementary information.

To quantify the degree of complementarity and to find out which attribute classes are positively correlated with DNA methylation only due to indirect or secondary effects, we applied the following strategy. Given two attribute classes, we calculate the correlation for both classes separately and for the combination of both. If the latter is higher than any of the former, we can conclude that the attributes complement each other. Comparing DNA sequence with all other attribute classes reveals that only (2) repeat frequency and distribution and (4) predicted DNA structure give rise to an increased correlation when combined with (1) DNA sequence properties and patterns, by 18.4% and 8.3%, respectively (Table 4, experiments 10 to 16). However, among these three classes, all combinations significantly increase the correlation (Table 4, experiments 10, 12, and 17).

Therefore, we conclude that three attribute classes, namely (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, and (4) predicted DNA structure are correlated with CpG island methylation on their own right (primary effect). The remaining attribute classes are either not correlated with CpG island methylation at all (class 7 evolutionary conservation and class 8 predicted transcription factor binding sites), or their correlations are secondary, explainable by their co-location with certain DNA sequence patterns alone (class 3 CpG island frequency and distribution, class 5 gene and exon distribution, and class 8 SNPs).

ID	Attribute Set	Number of Attributes	Prediction Method	Correlation	Accuracy	TN	FN	FP	TP
1	DNA sequence properties and patterns	426	A (linear SVM)	0.635	0.884	1,994	241	66	339
2	Repeat frequency and distribution	311	A (linear SVM)	0.657	0.890	1,995	225	65	355
3	CpG island frequency and distribution	13	A (linear SVM)	0.045	0.755	1,994	532	116	48
4	Predicted DNA structure	28	A (linear SVM)	0.486	0.844	2,053	406	7	174
5	Gene and exon distribution	52	A (linear SVM)	0.300	0.806	2,044	495	16	85
6	Predicted transcription factor binding sites	68	A (linear SVM)	−0.021	0.779	2,056	580	4	0
7	Evolutionary conservation	10	A (linear SVM)	0.000	0.780	2,060	580	0	0
8	Single nucleotide polymorphisms	10	A (linear SVM)	0.286	0.804	2,030	487	30	93
9	All attributes	918	A (linear SVM)	0.740	0.915	2,027	191	33	389
10	Class 1 (sequence) and class 2 (repeats)	737	A (linear SVM)	0.752	0.919	2,037	191	23	389
11	Class 1 and class 3 (CpG islands)	439	A (linear SVM)	0.626	0.880	1,977	233	83	347
12	Class 1 and class 4 (DNA structure)	454	A (linear SVM)	0.688	0.900	2,024	229	36	351
13	Class 1 and class 5 (genes)	478	A (linear SVM)	0.614	0.877	1,980	244	80	336
14	Class 1 and class 6 (TFBS)	494	A (linear SVM)	0.655	0.890	2,007	238	53	342
15	Class 1 and class 7 (conservation)	436	A (linear SVM)	0.626	0.881	1,989	243	71	337
16	Class 1 and class 8 (SNPs)	436	A (linear SVM)	0.618	0.879	1,988	248	72	332
17	Class 2 (repeats) and class 4 (DNA structure)	339	A (linear SVM)	0.713	0.907	2,020	205	40	375
18	DNA sequence properties and patterns	426	B (RBF-kernel SVM)	0.580	0.869	2,040	327	20	253
19	DNA sequence properties and patterns	426	C (AdaBoost)	0.664	0.892	2,009	233	51	347
20	DNA sequence properties and patterns	426	D (C4.5 trees)	0.566	0.852	1,869	200	191	380
21	DNA sequence properties and patterns	426	E (linear SVM using LIBSVM in R)	0.684	0.898	2,018	226	42	354
22	Transcription start site overlap	1	Heuristic (if TSS overlap: unmethylated, otherwise: throw coin)	0.358	0.788	1,810	310	250	270
23	Empty set	0	Trivial (predict everything as unmethylated)	0.000	0.780	2,060	580	0	0

Table 4. The predictive power of attribute classes differs remarkably; control experiments confirm the choice of the prediction method

This table summarizes the prediction experiments that were performed in order to analyze the association between DNA-related attributes and CpG island methylation (1 to 17), plus several control experiments (18 to 23). Each row corresponds to one prediction experiment. The column "Attribute Set" specifies the attributes that were used for prediction, "Number of Attributes" gives the size of the attribute set, and "Prediction Method" summarizes the algorithm used (see Methods section for details). The columns "TN," "FN," "FP," and "TP" give the test-set results for true negatives, false negatives, false positives, and true positives over a 10-fold stratified cross-validation that was repeated 20 times. Correlation and accuracy (the remaining two columns) are calculated in the usual way (Baldi et al. 2000) with the modification that, in the case of correlation, we add 0.0001 to TN, FN, FP, and TP to prevent the correlation from being undefined when an algorithm always predicts the same class.

Prediction of CpG island methylation status from DNA-related attributes

While the previous section was concerned with quantifying the relative contribution of different attribute classes to explaining CpG island methylation, the same methodology can be used to predict the methylation status of new CpG islands. Here we report the prediction performance of our method and we address potential limitations.

Without prior knowledge it is sensible to include all 918 non-zero attributes simultaneously in order to achieve best prediction results. In a 10-fold stratified cross-validation of a linear SVM, which we repeated 20 times with different random partitions, this setup resulted in an average correlation of 0.74, a test set accuracy of 91.5%, a specificity of 98.4%, and a sensitivity of 67.1% (Table 4, experiment 9).

In order to test the appropriateness of the prediction method that we used (SVM with linear kernel), we performed several control experiments employing other state-of-the-art machine learning algorithms (Hastie et al. 2001), namely SVM with radial basis function kernel, AdaBoost using tree stumps, C4.5 tree generator, and a second widely used implementation of a linear SVM. The results show that performances of all methods lie within the same range (Table 4, experiments 18 to 21).

Next, we investigated how prediction accuracies vary between CpG islands that are located at different positions relative to their closest annotated gene. For this analysis we regard a single CpG island as reliably predicted if its prediction is correct in at least 15 out of 20 randomized cross-validations, and we manually assigned each of the 132 CpG islands of the chromosome 21 dataset to one of the following categories (data available online: Bock et al. 2006, Table S2):

- Category 1: Promoter CpG islands, defined as overlapping with the transcription start site of an annotated gene: 80 cases fall into this category, of which 78 are unmethylated.
- Category 2: Intragenic CpG islands, defined as overlapping introns and/or exons of an annotated gene, but not the transcription start site: 24 cases fall into this category, of which 12 are unmethylated.
- Category 3: Gene-terminal CpG islands, defined as overlapping mainly the last exon and/or the 3' UTR of an annotated gene: six cases fall into this category, of which one is unmethylated.
- Category 4: Intergenic CpG islands, defined as not showing any overlap with an annotated human gene: 22 cases fall into this category, of which 12 are unmethylated.

Our results show that prediction accuracy is highest for promoter CpG islands, for which 77 unmethylated cases and one methylated case are predicted correctly in more than 15 out of 20 runs (98% accuracy); the second methylated case is predicted correctly in seven out of 20 runs and the one remaining unmethylated case is correctly predicted in only three runs. In categories 2, 3, and 4, the number of methylated and unmethylated CpG islands is almost balanced, thus prediction is much more difficult. Nevertheless, prediction accuracies stay high: For intragenic CpG islands, 20 cases are predicted correctly in more than 15 runs (83% accuracy). Among the gene-terminal CpG islands, four cases are predicted correctly in more than 15 runs (67% accuracy), and of all intergenic CpG islands, 18 are correctly predicted in more than 15 runs (82% accuracy).

In conclusion, our method achieves high prediction accuracy for CpG islands from all four categories. Finally, we note that the method significantly outperforms a heuristic prediction which relies on transcriptional start site overlap alone (Table 4, experiment 22), and that the very high specificity of the method (98.4%) facilitates chromosome-wide screening for methylated CpG islands, giving rise to a low number of false-positives.

Experimental validation by bisulfite sequencing

In order to further substantiate the reliability of our method, we experimentally validated its predictions for 12 CpG islands. To that end, we first predicted the methylation state of all CpG islands on chromosome 21 that were not part of the original dataset (Yamada et al. 2004); e.g. because they did not match the strict CpG island criteria imposed by Yamada et al. or because they (marginally) overlap with repetitive DNA. Next, we selected eight CpG islands that were predicted as unmethylated and four CpG islands that were predicted as methylated, and we experimentally determined their methylation status in human peripheral blood by bisulfite sequencing.

Hence, while keeping species (human) and chromosomes (21) identical, we varied experimental technique (bisulfite sequencing instead of restriction enzyme digestion), cell type (peripheral blood instead of lymphocytes), sample origin (healthy European female instead of healthy unspecified), and – of course – the CpG island. In the selection of validation CpG islands, we did not stratify for CpG island categories (see previous section) because we wanted to assess the method's overall performance across all categories of CpG islands.

The experimental results (Table 5) show that our prediction was correct in ten out of 11 cases (*P*-value below 0.01). The 12th case, predicted as methylated, showed incomplete yet significant methylation of 54%. Hence, our method can predict CpG island methylation with high accuracy on a previously unknown test set.

CpG Island Position (NCBI35)	Closest Gene	Method	Number of CpGs	Number of CpGs Analyzed	Methylation	Experimental Result	Prediction
Chr 21, 13331442–13331790	*C21orf 99*	Direct sequencing	14	11	91%	Methylated	Methylated
Chr 21, 13904631–13904830	*ANKRD21*	Direct sequencing	11	6	100%	Methylated	Methylated
Chr 21, 14676951–14678040	*STCH*	Direct sequencing	21	15	0%	Unmethylated	Unmethylated
Chr 21, 18538786–18539754	*CHODL*	Cloning and sequencing (nine clones)	26	26	7%	Unmethylated	Unmethylated
Chr 21, 26866818–26867612	*CYYR1*	Direct sequencing	21	14	0%	Unmethylated	Unmethylated
Chr 21, 29318596–29319405	*USP16*	Direct sequencing	18	13	0%	Unmethylated	Unmethylated
Chr 21, 30892864–30893090	*KRTAP6–2*	Direct sequencing	10	8	100%	Methylated	Methylated
Chr 21, 33836092–33837874	*GART*	Direct sequencing	18	14	0%	Unmethylated	Unmethylated
Chr 21, 38209756–38211197	*KCNJ6*	Direct sequencing	25	12	0%	Unmethylated	Unmethylated
Chr 21, 43461259–43461636	*CRYAA*	Cloning and sequencing (nine clones)	15	15	54%	Incomplete	Methylated
Chr 21, 45117025–45119447	*PTTG1IP*	Cloning and sequencing (five clones)	19	19	2%	Unmethylated	Unmethylated
Chr 21, 45669125–45669487	*C21orf 123*	Direct sequencing	10	7	100%	Methylated	Unmethylated

Table 5. Twelve CpG islands were analyzed experimentally to validate our predictions

This table summarizes the results of bisulfite sequencing of 12 selected CpG islands together with our prediction that was based on all attribute sets. In nine cases, bisulfite direct sequencing produced unambiguous results. In the three remaining cases, PCR products were cloned and individual clones were sequenced in order to determine the methylation status.

Comparison with the HEP dataset

The DNA methylation dataset originating from the HEP pilot study (Rakyan et al. 2004) gives us the opportunity to assess the generality of our method and the transferability of the predictions that we obtain from the chromosome 21 dataset. A priori, one would not expect a high degree of transferability because the HEP data vary from the chromosome 21 data that were used to develop the method in several important aspects. First, almost 90% of amplicons for which DNA methylation profiles were established do not fulfill CpG island properties. Second, the HEP did not analyze lymphocytes but a variety of other tissues (adipose, brain, breast, liver, lung, muscle, and prostate). Third, all analyzed sequences belong to the relatively small and exceptional major histocompatibility complex region on chromosome 6.

In order to make the HEP dataset accessible to our method, which works on CpG islands (or on DNA stretches of comparable length), we calculated the average DNA methylation level for every HEP amplicon and we defined a threshold to distinguish methylated from unmethylated amplicons (see Methods section for details). Next, we trained our method on the chromosome 21 dataset and predicted the methylation status of all HEP amplicons, in a similar way as was done for the experimental validation in the previous section. The results show a prediction accuracy that is low but still better than random (correlation = 0.15, accuracy = 74.7%, true negatives = 10, false negatives = 16, false positives = 37, true positives = 147). Hence, there seems to be a core association between DNA-related features and CpG island methylation that is similar or identical across tissues and genomic locations. This association can be specified further by a comparison of prediction error rates. First, we observe a remarkably low false negative rate of 10%. In other words, the characteristics that were learned to predict CpG islands as methylated in lymphocytes are to some extent transferable across tissues and genomic locations, giving rise to a low false negative rate on the HEP dataset. Second, the false positive rate was 8-fold higher than the corresponding false negative rate (79%), indicating that it is difficult to transfer the DNA-related characteristics of unmethylated cases between the two datasets.

Next, we analyzed to what degree the prediction performance improves when the method is provided with a more adequate training dataset, i.e. when it is permitted to learn the characteristics that are unique to the HEP dataset. To that end, we trained and evaluated our prediction method in a cross-validation on the HEP dataset, using all eight attribute classes. Taking into account all HEP amplicons, this resulted in a sharp increase in prediction performance, with a correlation of 0.47 and an accuracy of 82.4% (true negatives = 25.7, false negatives = 15.6, false positives = 21.3, true positives = 147.4, averaged over 20 independent cross-validation runs). A further performance increase was observed when we repeated the analysis on amplicons that do not deviate too strongly from the CpG island characteristic, for which the prediction method was developed. We sorted all amplicons by the ratio of observed vs. expected CpG dinucleotide frequency, and ran a separate training and prediction analysis for the top, middle, and bottom 70 cases. Results show a correlation of 0.59 for the top group and 0.73 for the middle group (one third of the amplicons in the top group and none in the middle group fulfill CpG island properties). In contrast, predictions fail for the bottom group (correlation = -0.02), in which unmethylated cases are rare (six out of 70), possibly because sample size is too small or because these cases behave more randomly.

These results indicate that our prediction method is also well-suited for predicting the average methylation status for sequences that are not necessarily CpG islands, at least when a suitable training set is provided and CpG dinucleotide frequency is not too low.

Finally, because the HEP dataset contains methylation information for seven different tissues it should be possible, in principle, to detect evidence of tissue-specific methylation regulation, e.g. binding site patterns of tissue-specific transcription factors. Therefore, one would expect that the prediction performance of our method was higher if trained on data from only one tissue, compared to the combination of all tissues, at least when focusing only on the most tissue-specific amplicons. However, we find no evidence for this in our dataset. Instead, prediction performances for individual tissues closely resemble the average case (data not shown). There are several possible explanations for the method's failure to learn tissue-specific methylation information from the HEP dataset. On the one hand, tissue-specific methylation may be largely uncorrelated with the sequence-related attributes that we analyzed. On the other hand, the dataset may simply be too small. In fact, only between five and 19 out of 210 amplicons per tissue deviate from the "default" state calculated as the consensus methylation over all tissues.

B-2.4 Discussion

We have shown that CpG island methylation can be predicted from the genomic DNA sequence, suggesting that predictive bioinformatic analysis may contribute to our understanding of the biology that controls methylation in vivo. We initially identified DNA-related attributes that discriminate significantly between methylated and unmethylated CpG islands in human lymphocytes. Next, we quantified the correlation of CpG island methylation with eight groups of DNA-related attributes and found DNA sequence patterns, repeat frequencies, and predicted DNA structure to be the key contributors. Finally, we developed a machine learning method that can predict the methylation status of unknown CpG islands and we validated the accuracy and reliability of this method both statistically and experimentally.

A number of observations are worthwhile to comment on. First, in line with earlier observations we find almost all promoter CpG islands unmethylated, but also a significant number of intergenic CpG islands, which are often distant from any annotated gene. Little is known about the functional role of these intergenic CpG islands. However, it has been observed that unmethylated CpG islands often co-localize with DNA replication origins (Antequera 2003), and we believe that it would be worthwhile to perform a systematic experimental study analyzing the functional role of unmethylated intergenic CpG islands. DNA methylation predictions may help to speed up and guide such work.

Second, we found evidence that the default DNA methylation status of many CpG islands may be relatively stable during evolution. By comparing frequencies of the CGCG pattern to its mutated counterpart TGTG/CACA (the former is over-represented in unmethylated CpG islands of our dataset whereas the latter is over-represented in methylated CpG islands), we concluded that higher CpG → TpG mutation rates have applied to the CpG islands that we find methylated in human lymphocytes, than to those that we find unmethylated. Given that methylated CpG dinucleotides are more prone to CpG → TpG mutations (Caiafa and Zampieri 2005), a straightforward explanation would be that the methylation status of lymphocytes is not only similar to that found in the germline where mutations become fixed, but has also been stable over evolutionary time, such that the observed mutations could accumulate.

Third, our results show that certain aspects of DNA sequence and (predicted) DNA structure, such as a high DNA rise and a low DNA twist, seem to be associated with methylated CpG islands in vivo. It would be interesting to analyze how these sequence and structure attributes correlate with the in vitro recognition and methylation potential of CpG-rich sequences by mammalian DNA methyltransferases. Some reports suggest that unusual DNA structures, e.g. repeats and cruciform structures (Chen et al. 1998), can lead to increased methylation activity by DNA methyltransferases. Moreover, local transitions between DNA in A-form, B-form, or Z-form may influence the methylation potential of the DNA, and it is tempting to speculate that some of our observed parameters may reflect such local differences in DNA structure formation.

Fourth, differences in error rates when training on the chromosome 21 dataset and testing on the HEP dataset suggest that DNA-related characteristics identifying consistently methylated CpG islands are robust across tissues and genomic locations while those identifying unmethylated CpG islands are not – and have to be learned specifically for each tissue or genomic location. This interpretation is consistent with the hypothesis that most CpG islands in the human genome can become methylated, and do so if they are not preserved in the unmethylated state by specific (and tissue-dependent) influences, for example by transcription factor binding.

B-3 EpiGRAPH: A user-friendly tool for advanced (epi-) genome analysis and prediction[1]

B-3.1 Motivation

Having shown the utility of large-scale statistical analysis and prediction for DNA methylation, it became apparent that a similar approach could be used to address other topics of epigenetic and genome research as well. We thus decided to extend our method into a software

[1] This chapter describes work conducted in collaboration with Konstantin Halachev and Joachim Büch. Konstantin Halachev designed and implemented a substantially enhanced and extended version of the EpiGRAPH backend and contributed important ideas to all aspects of the project. Joachim Büch set up and maintained the technical infrastructure.

toolkit that is specifically targeted toward the computational analysis of genomic regions in the context of complex mammalian genomes. This toolkit addresses the growing need for genome analysis tools that focus on genomic regions rather than on protein-coding genes (Bernstein et al. 2007; Chen and Rajewsky 2007; Kapranov et al. 2007).

With recent experimental innovations such as tiling microarrays and next-generation sequencing (Mardis 2008; Schones and Zhao 2008; van Steensel 2005), methods are now in hand for mapping the genetic and epigenetic characteristics of all functional elements in mammalian genomes (ENCODE Project Consortium 2004; ENCODE Project Consortium 2007). However, the lack of bioinformatic tools that support the analysis and interpretation of the resulting large-scale datasets poses a major bottleneck for the discovery of novel biological insights. This is particularly true as even simple statistical analysis of genome-scale datasets requires significant bioinformatic skills, while the use of advanced computational methods is currently beyond the reach of most biologists.

To set the context for presenting the EpiGRAPH toolkit for genome-scale analysis of interactions between genome, epigenome and transcriptome, we briefly review related bioinformatic software tools and we outline potential limitations, which EpiGRAPH aims to address. (i) Genome browsers such as the UCSC Genome Browser (Karolchik et al. 2008) and Ensembl (Flicek et al. 2008) do a great job storing, integrating and visualizing a wide range of datasets that can be linked to specific positions in the genome. However, in their current versions genome browsers do not provide support for statistical analysis or data mining that would enable the user to test for significant co-localization between two sets of genomic regions. (ii) Gene-centered analysis tools, including GSEA (Subramanian et al. 2007), DAVID (Huang et al. 2007) and parts of the Bioconductor library (Gentleman et al. 2004), are highly useful for analyzing gene expression data, but their approach is difficult to generalize to the analysis of genomic regions that are not directly associated with genes. (iii) Workflow management systems such as Taverna (Hull et al. 2006) and Kepler (http://kepler-project.org/) have great potential to provide unified solutions for all kinds of data analysis problems in bioinformatics and beyond, but cannot yet replace specialized genome analysis tools that solve a specific range of tasks at high performance and for large datasets. (iv) The Galaxy web service (Blankenberg et al. 2007; Giardine et al. 2005) prototypes what one could describe as a powerful pocket calculator for genomes. It provides an online genome workbench for performing calculations on sets of genomic regions, including intersecting, joining and merging sets of regions, without having to download any interim results. Through its focus on relatively simple operations and an efficient batch mode, Galaxy is highly useful even for large datasets. However, Galaxy currently does not provide support for statistical testing, data mining or prediction.

In the absence of dedicated software tools for quantitative, genome-scale analysis of the interactions between genome, epigenome and transcriptome, many bioinformaticians have addressed such questions manually, downloading all relevant datasets from existing repositories and writing one-time-use scripts for data integration, statistical analysis and prediction (Berry et al. 2006; Bock et al. 2006; Cohen et al. 2006; Das et al. 2006; Fang et al. 2006; Luedi et al. 2007; Luedi et al. 2005; Montgomery et al. 2007; Wang et al. 2006). Unfortunately, such studies are time-consuming to perform, difficult to reproduce and require bioinformatic skills that are beyond the reach of most biologists. We therefore decided to develop EpiGRAPH, pulling together our experience and established workflows from a number of studies (Bock et al. 2006; Bock et al. 2007; Bock et al. 2008; Liu et al. 2007) and incorporating them into a powerful and easy-to-use web service. In the remainder of this chapter, we outline the basic concepts of EpiGRAPH, we showcase its practical use and utility in two biologically relevant case studies, we outline how EpiGRAPH can be adapted and extended for custom analysis scenarios, and we describe the methods and software paradigms that provide the foundations of EpiGRAPH.

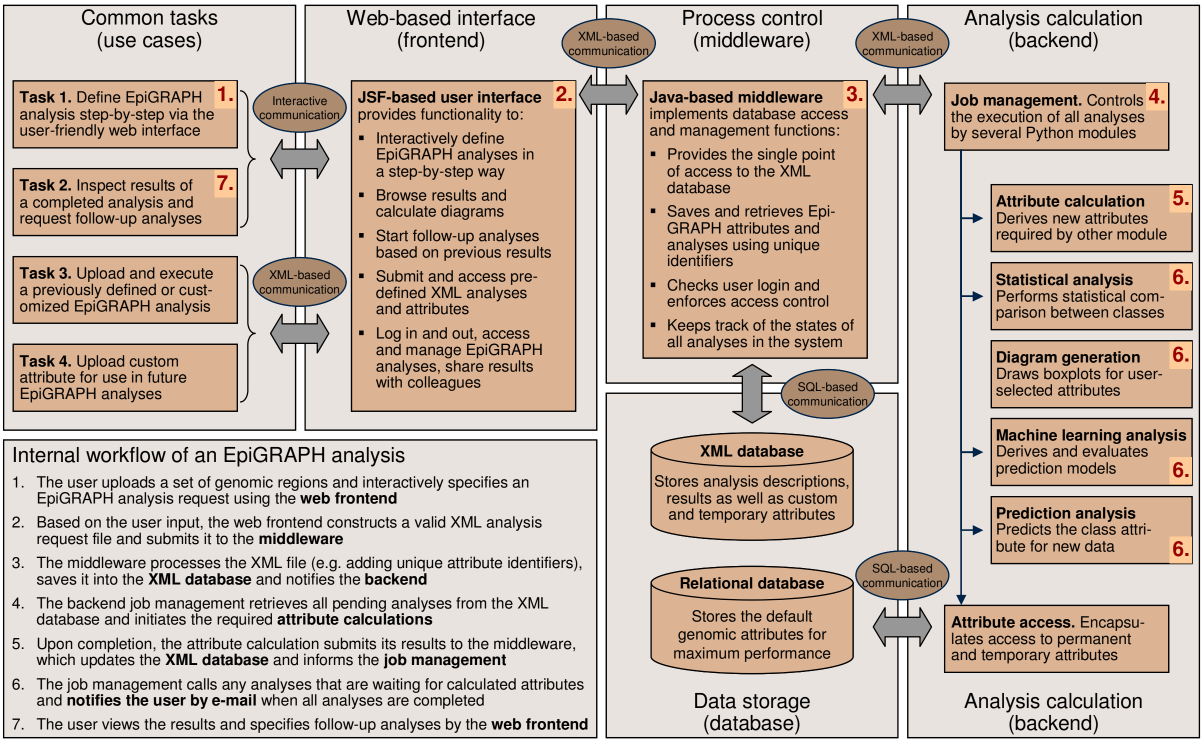
B-3.2 Methods

EpiGRAPH software architecture and analysis workflow

The key design decision underlying EpiGRAPH's software architecture is to store each EpiGRAPH analysis in a single XML file. This XML file contains not only a detailed specification of the analysis and its supplementary attributes, but also its current processing status and, upon completion, its results. All XML files processed by EpiGRAPH conform to the standardized X-GRAF format (discussed in more detail below) and are stored in an XML database.

EpiGRAPH's XML-based, analysis-centric design offers a number of advantages over alternative architectures: (i) *Reproducibility*: All information relevant to an analysis, including its specifications and results, are bundled in a single file, which provides a complete documentation of the analysis. The same analysis can be rerun at any time simply by uploading its XML file into the EpiGRAPH web service. (ii) *Parallel processing*: Because the different analysis modules operate on different parts of the XML tree, they can work in parallel without generating write-write conflicts. (iii) *Interoperability and error checking*: The use of XML files facilitates data exchange with other software systems, and the X-GRAF format provides error checking when XML files are constructed manually or exchanged between different software systems.

Technically, the EpiGRAPH web service consists of three software components and two logical databases (Figure 3). (i) The *web-based frontend* provides convenient access to EpiGRAPH's functionality over the internet. The frontend is implemented in Java (http://www.java.com/), utilizing the JavaServer Faces framework for its user interface and Java servlets as well as JavaServer Pages for operating as a web application. (ii) The *process control middleware* provides a single point of access to the analyses and custom attributes stored in the XML database and it enforces compliance with the X-GRAF XML format. The middleware is implemented as a Java servlet and makes its services available via XML-RPC (http://www.xmlrpc.com/). (iii) The *analysis calculation backend* performs all attribute calculations and bioinformatic analyses required to fulfill an EpiGRAPH analysis request and submits its results to the middleware, which stores them in the XML database. The backend is implemented in Python (http://www.python.org/), using the R package (http://www.r-project.org/) for statistical analysis and diagram generation as well as the Weka package (http://www.cs.waikato.ac.nz/~ml/weka/) for machine learning and prediction analysis. (iv) The *relational database* stores EpiGRAPH's default attributes. Oracle Database 10g (http://www.oracle.com/database/) is used with pre-calculated indices in order to achieve high-performance database retrieval. (v) The *XML database* provides central storage of all XML files and enables parallelized access to the XML files as a whole as well as to specific subnodes. Oracle XML DB (http://www.oracle.com/technology/tech/xml/xmldb/index.html) is used, which is an XML database extension of the Oracle database. Technically, Oracle XML DB decomposes all XML files into relational database tables, making use of the X-GRAF schema definition and object-relational mapping. Hence, while the relational database and the XML database of EpiGRAPH are logically distinct and used for different types of data (default attributes vs. analysis requests and custom attributes), both types of data are ultimately stored in the same database management system.


Common tasks
(use cases)
Web-based interface
(frontend)
Process control
(middleware)
Analysis calculation
(backend)
XML-based communication
XML-based communication
Interactive communication
XML-based communication
SQL-based communication
SQL-based communication
Task 1. Define EpiGRAPH analysis step-by-step via the user-friendly web interface 1.
Task 2. Inspect results of a completed analysis and request follow-up analyses 7.
Task 3. Upload and execute a previously defined or customized EpiGRAPH analysis
Task 4. Upload custom attribute for use in future EpiGRAPH analyses
JSF-based user interface 2.
provides functionality to:
▪ Interactively define EpiGRAPH analyses in a step-by-step way
▪ Browse results and calculate diagrams
▪ Start follow-up analyses based on previous results
▪ Submit and access predefined XML analyses and attributes
▪ Log in and out, access and manage EpiGRAPH analyses, share results with colleagues
Java-based middleware 3.
implements database access and management functions:
▪ Provides the single point of access to the XML database
▪ Saves and retrieves EpiGRAPH attributes and analyses using unique identifiers
▪ Checks user login and enforces access control
▪ Keeps track of the states of all analyses in the system
Job management. Controls the execution of all analyses by several Python modules 4.
Attribute calculation 5.
Derives new attributes required by other module
Statistical analysis 6.
Performs statistical comparison between classes
Diagram generation 6.
Draws boxplots for user-selected attributes
Machine learning analysis
Derives and evaluates prediction models 6.
Prediction analysis
Predicts the class attribute for new data 6.
Attribute access. Encapsulates access to permanent and temporary attributes
XML database
Stores analysis descriptions, results as well as custom and temporary attributes
Relational database
Stores the default genomic attributes for maximum performance
Data storage
(database)
Analysis calculation
(backend)
Internal workflow of an EpiGRAPH analysis

1. The user uploads a set of genomic regions and interactively specifies an EpiGRAPH analysis request using the **web frontend**
2. Based on the user input, the web frontend constructs a valid XML analysis request file and submits it to the **middleware**
3. The middleware processes the XML file (e.g. adding unique attribute identifiers), saves it into the **XML database** and notifies the **backend**
4. The backend job management retrieves all pending analyses from the XML database and initiates the required **attribute calculations**
5. Upon completion, the attribute calculation submits its results to the middleware, which updates the **XML database** and informs the **job management**
6. The job management calls any analyses that are waiting for calculated attributes and **notifies the user by e-mail** when all analyses are completed
7. The user views the results and specifies follow-up analyses by the **web frontend**

Figure 3. Outline of EpiGRAPH's software architecture

This figure shows a schematic overview of EpiGRAPH's core components, and it describes their interaction in a typical analysis workflow. The red numbers indicate the key component(s) for each step of the workflow description (bottom left). Abbreviation: JSF – Java Server Faces (a Java-based web application framework).

Importantly, the choice of technologies for each component reflects the specific requirements of the tasks it performs. The frontend has to provide a user-friendly interface in a variety of web browsers, which is best achieved using a web application framework such as JavaServer Faces. The middleware makes connections with the XML database and performs extensive XML processing, hence the use of Java with its high-quality library support for Oracle XML DB (http://www.oracle.com/technology/tech/xml/xmldb/index.html), StAX (http://jcp.org/en/jsr/detail?id=173) and JAXB (https://jaxb.dev.java.net/) is an appropriate choice. The backend implements most of EpiGRAPH's application logic and is likely to be extended by other researchers, therefore Python (http://www.python.org/) was selected due to its proven track record for fast and robust software engineering in scientific applications, platform independence and its wide acceptance within the bioinformatics community.

The internal workflow of an EpiGRAPH analysis is depicted in Figure 3, showing how the different components interact when fulfilling an EpiGRAPH analysis request.

Genomes and (epi-) genomic attributes included in EpiGRAPH

EpiGRAPH currently includes five genome assemblies of four species: (i) *hg18*: the latest assembly of the human genome (NCBI36.1); (ii) *hg17*: the genome assembly used for the ENCODE project pilot phase (NCBI35); (iii) *mm9*: the latest assembly of the mouse genome (NCBI37); (iv) *panTro2*: the latest assembly of the chimp genome; (v) *galGal3*: the latest assembly of the chicken genome. For each of these genomes, we manually selected a large number of genomic attributes that are likely to be predictive of interesting genomic phenomena (see Table 6 and http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html for details). When calculated for a specific genomic region, most of these attributes take the form of overlap frequencies (e.g. how many exons overlap with the genomic region?), overlap lengths (e.g. how many basepairs of exonic DNA overlap with the genomic region?) or DNA sequence pattern frequencies (e.g. how many times does the pattern "TATA" occur in the genomic region?). All of these attributes are standardized to a default region size of one kilobase in order to be comparable between regions of different size. In addition, EpiGRAPH uses score attributes, which are averaged over all overlapping regions of a specific type (e.g. what is the average exon number of all genes overlapping with the genomic region?), and category attributes, which split up an attribute into subattributes (e.g. how many coding vs. non-coding SNPs overlap with the genomic region?).

The data for most of these attributes were collected from annotation tracks in the UCSC Genome Browser, using an automated data retrieval pipeline. In addition, published genomic datasets that appear to be of particular interest are imported into the database on a regular basis. Currently, this includes data on histone modifications (Barski et al. 2007), DNA methylation (Rollins et al. 2006), regulatory CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis), DNA helix structure (Gardiner et al. 2003), DNA solvent accessibility (Greenbaum et al. 2007), tissue-specific gene expression (Su et al. 2004), isochores (Costantini et al. 2006) and transcription initiation events (Carninci et al. 2006). Finally, users can upload custom datasets into the database, which will be available in further analyses by the same user just like EpiGRAPH's default attributes.

Attribute Groups	Total Number of Attributes					Attributes (Examples)
	hg18	*hg17*	*mm9*	*panTro2*	*galGal3*	
DNA Sequence	178	178	178	178	178	Frequency of "TATA" pattern, cytosine content or CpG frequency
DNA Structure	24	24	24	24	24	Predicted DNA helix twist, predicted solvent accessibility
Repetitive DNA	95	92	73	91	49	Overlap with Alus, LINEs and tandem repeats
Chromosome Organization	18	29	15	-	-	Overlap with chromosomal bands and isochores
Evolutionary History	94	101	-	-	44	Overlap with evolutionary conserved regions
Population Variation	75	69	-	-	-	SNP density and overlap with specific SNP types (e.g. non-synonymous exonic or splice site)
Genes	37	60	20	10	10	Overlap with annotated genes, pseudogenes and predicted microRNA genes
Regulatory Regions	249	125	5	5	5	Overlap with CpG islands and predicted transcription factor binding sites
Transcriptome	49	65	9	9	9	Overlap with ESTs and mRNA sequences
Epigenome and Chromatin Structure	80	17	-	-	-	Overlap with regions exhibiting DNA methylation or specific histone modifications
ENCODE Transcriptome	-	19	-	-	-	Overlap with transcription fragments from the ENCODE 1% pilot project
ENCODE Epigenome and Chromatin Structure	-	323	-	-	-	Overlap with regions of known epigenetic state from the ENCODE 1% pilot project
Sum	899	1102	324	317	319	

Table 6. List of default attributes included in EpiGRAPH

This table summarizes the attribute groups that are currently included in EpiGRAPH. Due to different degrees of annotation, numbers differ between the genomes of human (*hg18* and *hg17*), mouse (*mm9*), chimp (*panTro2*) and chicken (*galGal3*).

Attribute calculation

The basic functionality of EpiGRAPH's attribute calculation module is to calculate a large number of genomic attributes (such as frequency and length of overlap with EpiGRAPH's default attributes) for any set of genomic regions submitted to the web service. This step is a prerequisite for all further analyses, and it is typically the most computationally intensive and time-consuming part of an EpiGRAPH analysis. The attribute calculation makes extensive use of multithreading in order to increase performance.

Beyond its core task of deriving hundreds or even thousands of different attributes for each genomic region in the input dataset, the attribute calculation module provides three additional features that add to its value as a general genome calculator. First, the user can define derived attributes, thus augmenting genomic attributes that are already contained in the database (e.g. deriving a set of putative promoter regions from a gene attribute). Second, random control regions can be calculated such that they match their source regions in terms of chromosome and length distribution, GC content, repeat content and/or exon overlap. Technically, this is achieved by repeatedly sampling random genomic regions of given length from a given chromosome and retaining a region only if its GC content, repeat content and/or exon overlap are within a user-specified interval around the corresponding value of the source region. Third, attributes can be calculated not only for the genomic regions provided in the input dataset, but also for fixed windows left and right of these regions, in order to capture significant differences in the upstream or downstream neighborhood of a specific set of regions. All results calculated by the attribute calculation module can be used as basis for further EpiGRAPH analyses or downloaded in tab-separated value format for analysis outside EpiGRAPH.

Statistical analysis and diagram generation

Two of EpiGRAPH's four analytical modules – statistical analysis and diagram generation – help the user identify individual attributes that differ between sets of genomic regions that fall into two classes, which we denote as "positives" and "negatives". The statistical analysis module calculates pairwise statistical tests between the sets of positives and negatives, separately for each genomic attribute. The nonparametric Wilcoxon rank-sum test is used for numeric attributes and Fisher's exact test is used for discrete attributes. *P*-values are adjusted for multiple testing by the highly conservative Bonferroni method, which controls the family-wise error rate, and by a more recent and usually preferred method that controls the false discovery rate (Benjamini and Hochberg 1995). While EpiGRAPH suggests using an overall significance threshold of 5%, the user is free to select different values. If multiple windows around the genomic regions of interest are taken into account and tested simultaneously, the user can specify weights to control how the *P*-value threshold is distributed when testing for significant attributes in each of these windows. A typical choice is to use a relatively high *P*-value of, say, 3% for the central window (i.e. the regions provided by the input dataset) while distributing the remaining 2% equally among the upstream and downstream windows. This way, the additional testing for strong effects in the neighborhood comes at the cost of only a limited decrease in statistical power at the genomic regions of interest.

While the statistical analysis module focuses on the question whether or not a specific attribute differs significantly between the sets of positives and negatives, the diagram generation module can help assess the effect size, i.e. the quantitative difference between positives and negatives. For any selected attribute, this module derives boxplots contrasting the attribute's distribution among the positives with that among the negatives.

Machine learning analysis and prediction analysis

In contrast to the statistical analysis module focusing on individual attributes, the machine learning analysis module assesses how well attribute groups collectively differentiate between the sets of positives and negatives. We treat this question as a machine learning task, predicting for each genomic region whether it is likely to belong to the set of positives or to the set of negatives and interpreting the prediction performance achieved for a specific attribute group as a measure of how well this group discriminates between positives and negatives.

Technically, a machine learning algorithm (e.g. a support vector machine) is repeatedly trained and tested on partitions of the training dataset following a four-step procedure (all parameters mentioned below are default values and can be changed by the user): (i) If the set of positives contains more than twice as many genomic regions as the set of negatives (or vice versa), the larger set is randomly downsampled such that the class imbalance never exceeds 67% vs. 33%, thus limiting prediction bias toward the majority class. (ii) Using 10-fold cross-validation, the machine learning algorithm is repeatedly trained on 90% of the genomic regions and tested on the remaining 10%. (iii) Cross-validation is repeated ten times with random partition assignments. (iv) The overall prediction performance is measured by the correlation coefficient between the predictions and the correct values on the cross-validation test sets, as well as by the corresponding values for percent accuracy, sensitivity and specificity, averaged over all cross-validation runs.

During prediction analysis, a machine learning algorithm is trained as described above, but now a bootstrapped sample drawn from the full training dataset (if necessary downsampling is used to enforce a maximum class imbalance of 67% vs. 33%). The trained prediction model is then applied to predict the likelihood of belonging to the set of positives for all genomic regions in an additional, user-supplied set of genomic regions. The resulting quantitative predictions for each region can assume values between zero and one, with a value of zero corresponding to a high-confidence negative prediction, a value of 0.5 to a borderline case, and a value of one to a high-confidence positive prediction. This process is repeated ten times with different bootstrapped samples in order to obtain an additional criterion for the reliability of the predictions, and a consensus prediction, a mean confidence value as well as a standard deviation of the confidence values is reported for each genomic region and each prediction setup.

For both machine learning analysis and prediction analysis, EpiGRAPH currently supports the use of eight different machine learning methods/configurations: (i) Support vector machine with linear kernel; (ii) support vector machine with RBF kernel, (iii) AdaBoost on tree stumps, (iv) logistic regression, (v) random forest, (vi) C4.5 tree generator, (vii) Bayesian network, and (viii) naïve Bayes, all of which are implemented using functions from the Weka package (http://www.cs.waikato.ac.nz/~ml/weka/) with default parameters. For comparison and to give a baseline for the expected accuracy, we also include a trivial algorithm that always predicts the majority class.

A. Overview of the XML schema definition specifying the X-GRAF format

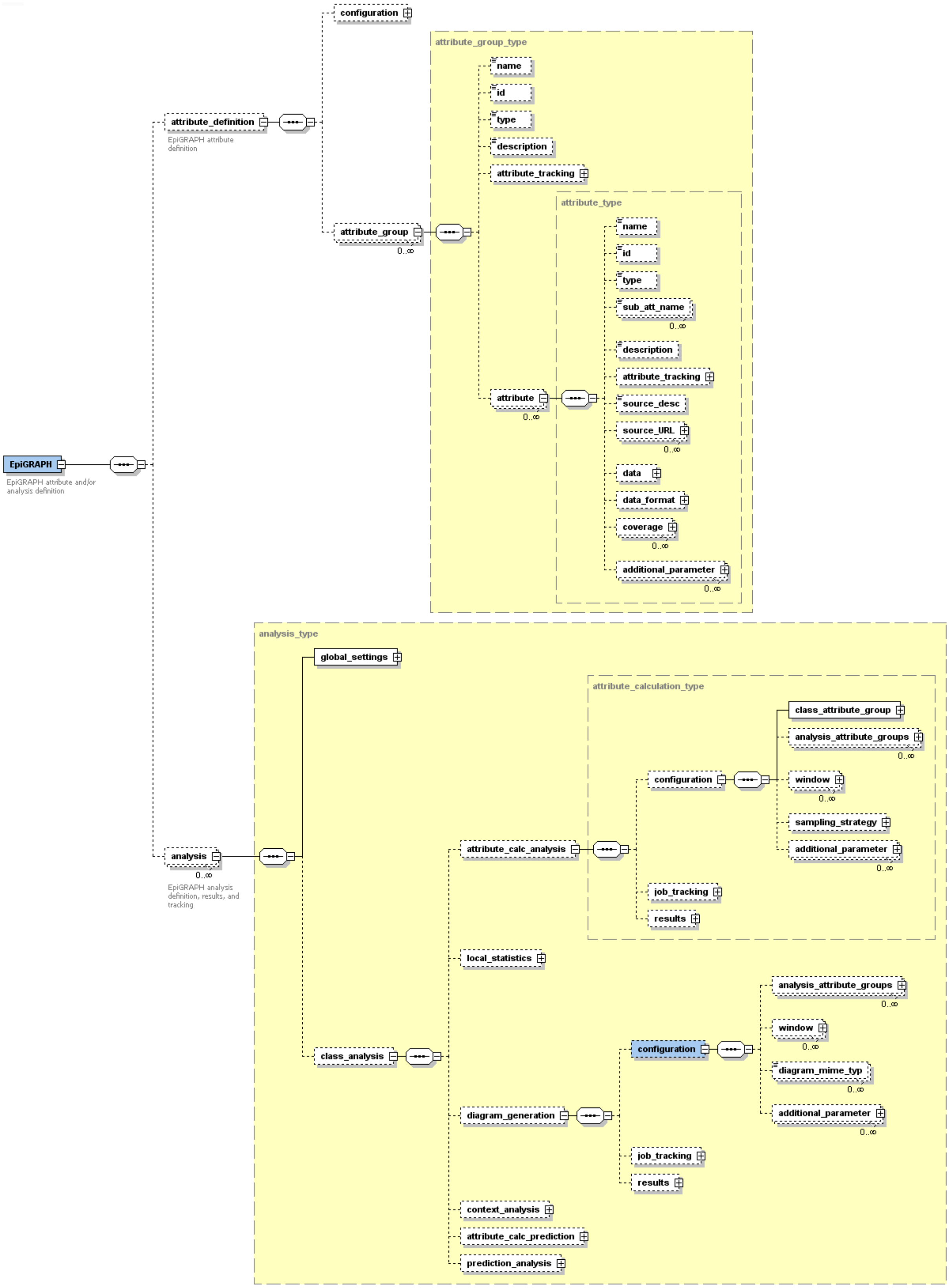
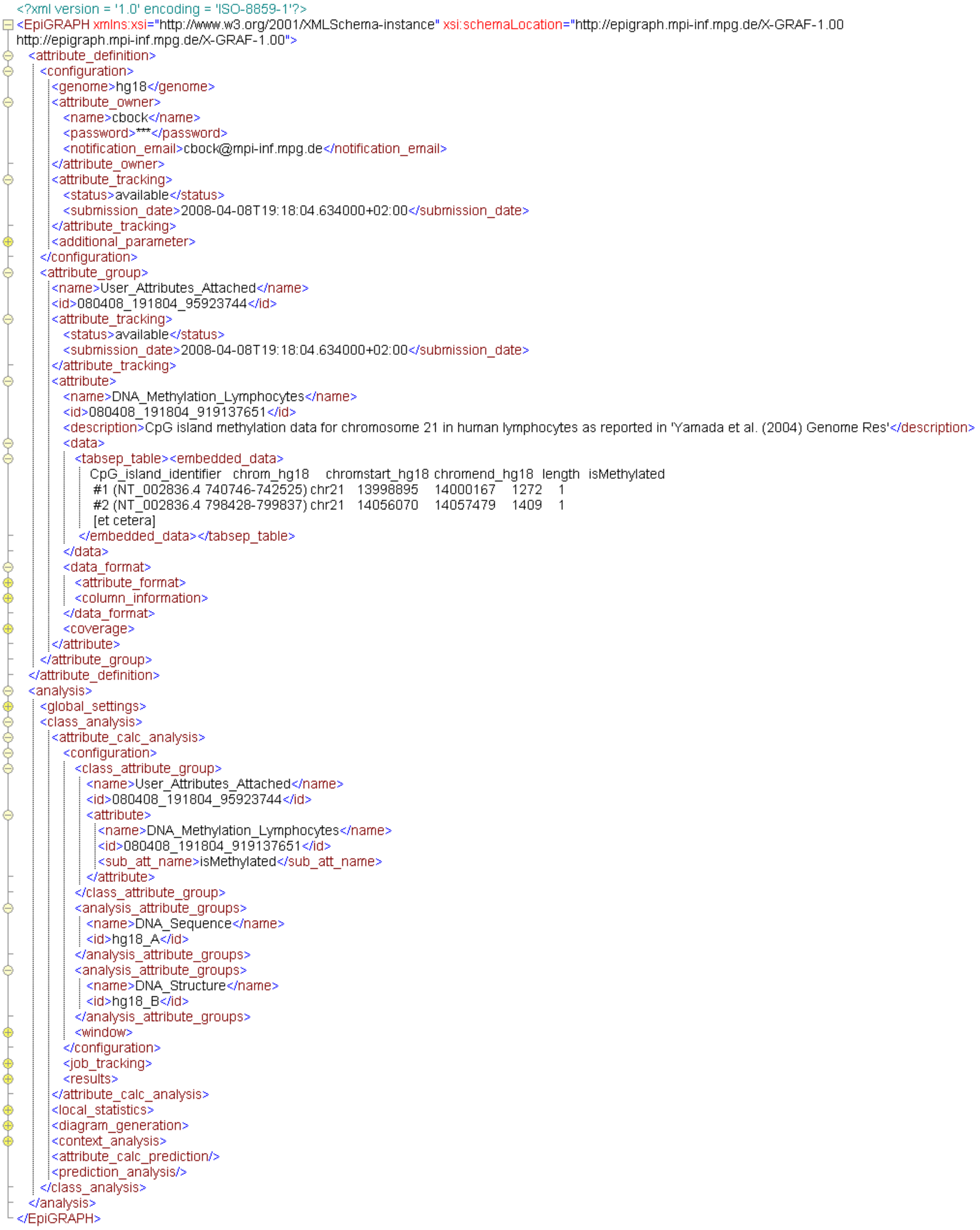


Figure 4. Documentation of EpiGRAPH analyses in the X-GRAF format

This figure illustrates the **X**ML **G**enomic **R**elationship **A**nalysis **F**ormat (X-GRAF), which is the format used by EpiGRAPH to keep track of all analyses and attributes. Panel A displays an outline of the XML schema definition that defines the format and which is used to validate every submitted EpiGRAPH analysis request. Panel B displays an excerpt of an XML analysis documentation conforming to the X-GRAF format. This XML file was calculated with the EpiGRAPH web service (cf. EpiGRAPH tutorial 1 online) and downloaded from the results page.

B. Example of an X-GRAF-compatible XML file documenting an EpiGRAPH analysis



```
<?xml version = '1.0' encoding = 'ISO-8859-1'?>
<EpiGRAPH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://epigraph.mpi-inf.mpg.de/X-GRAF-1.00
http://epigraph.mpi-inf.mpg.de/X-GRAF-1.00">
  <attribute_definition>
    <configuration>
      <genome>hg18</genome>
      <attribute_owner>
        <name>cbock</name>
        <password>***</password>
        <notification_email>cbock@mpi-inf.mpg.de</notification_email>
      </attribute_owner>
      <attribute_tracking>
        <status>available</status>
        <submission_date>2008-04-08T19:18:04.634000+02:00</submission_date>
      </attribute_tracking>
      <additional_parameter>
    </configuration>
    <attribute_group>
      <name>User_Attributes_Attached</name>
      <id>080408_191804_95923744</id>
      <attribute_tracking>
        <status>available</status>
        <submission_date>2008-04-08T19:18:04.634000+02:00</submission_date>
      </attribute_tracking>
      <attribute>
        <name>DNA_Methylation_Lymphocytes</name>
        <id>080408_191804_919137651</id>
        <description>CpG island methylation data for chromosome 21 in human lymphocytes as reported in 'Yamada et al. (2004) Genome Res'</description>
        <data>
          <tabsep_table><embedded_data>
            CpG_island_identifier   chrom_hg18    chromstart_hg18 chromend_hg18  length  isMethylated
             #1 (NT_002836.4 740746-742525) chr21   13998895    14000167    1272    1
             #2 (NT_002836.4 798428-799837) chr21   14056070    14057479    1409    1
             [et cetera]
           </embedded_data></tabsep_table>
        </data>
        <data_format>
          <attribute_format>
          <column_information>
        </data_format>
        <coverage>
      </attribute>
    </attribute_group>
  </attribute_definition>
  <analysis>
    <global_settings>
    <class_analysis>
      <attribute_calc_analysis>
        <configuration>
          <class_attribute_group>
            <name>User_Attributes_Attached</name>
            <id>080408_191804_95923744</id>
            <attribute>
              <name>DNA_Methylation_Lymphocytes</name>
              <id>080408_191804_919137651</id>
              <sub_att_name>isMethylated</sub_att_name>
            </attribute>
          </class_attribute_group>
          <analysis_attribute_groups>
            <name>DNA_Sequence</name>
            <id>hg18_A</id>
          </analysis_attribute_groups>
          <analysis_attribute_groups>
            <name>DNA_Structure</name>
            <id>hg18_B</id>
          </analysis_attribute_groups>
          <window>
        </configuration>
        <job_tracking>
        <results>
      </attribute_calc_analysis>
      <local_statistics>
      <diagram_generation>
      <context_analysis>
      <attribute_calc_prediction/>
      <prediction_analysis/>
    </class_analysis>
  </analysis>
</EpiGRAPH>
```

Figure 4 (continued).

X-GRAF format

Throughout EpiGRAPH's workflow (Figure 3), analyses and custom attributes are stored in XML files. In order to standardize the format of these XML files and to facilitate interoperability between the frontend, middleware and backend components, we defined the **X**ML **G**enomic **R**elationship **A**nalysis **F**ormat (X-GRAF). X-GRAF consists of an XML schema, against which each X-GRAF-compatible XML file has to validate in order to be regarded as syntactically correct, and a set of rules that describe the semantic interpretation of X-GRAF-compliant XML files (see http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html for details). X-GRAF-compatible XML files can incorporate two major subtrees, "attribute definition" and "analysis" (see Figure 4 for illustration). The attribute definition section keeps track of genomic attributes, which are organized in attribute groups and can be defined by embedded tab-separated tables or by referring to external data sources such as a database or a URL. The analysis section documents all analysis steps, including attribute calculation, statistical analysis, diagram generation, machine learning analysis and prediction analysis. Each of these subsections comprises an analysis configuration (a description of what is to be calculated), analysis tracking information (e.g. submission data, current state and error messages) and the results of the analysis (in the form of tables and diagrams directly embedded in the XML file).

Although X-GRAF was created for EpiGRAPH, it is designed with additional applications in mind. Being both formalized and sufficiently easy-to-understand, X-GRAF may provide a suitable basis for analysis specification, results documentation and data exchange of future genome analysis tools and statistical genome browsers.

B-3.3 Results

The EpiGRAPH web service – concepts and applications

EpiGRAPH (http://epigraph.mpi-inf.mpg.de/) is designed to facilitate advanced bioinformatic analysis of genome and epigenome datasets. Such datasets frequently consist of sets of genomic regions that share certain characteristics, e.g. being bound by a specific chromatin protein or having undergone significant levels of selection in the human lineage. Typically, these genomic regions fall into opposing classes, e.g. Polycomb-bound vs. unbound promoter regions or significantly conserved vs. nonconserved regulatory elements. Even when this is not the case, it is possible to generate a randomized set of control regions to complement a given set of genomic regions. EpiGRAPH thus focuses on the analysis of sets of genomic regions that fall into two classes, which we denote as "positives" and "negatives".

EpiGRAPH offers four major analytical modules (see Figure 3 for an overview of EpiGRAPH's software architecture and Figure 5 to Figure 9 for screenshots of exemplary results): (i) The statistical analysis module identifies attributes that differ significantly between the sets of positives and negatives, based on a large attribute database comprising hundreds of genome and epigenome datasets (Figure 6); (ii) the diagram generation module draws boxplots visualizing the distribution of a selected attribute among the sets of positives vs. negatives (Figure 8); (iii) the machine learning analysis module evaluates how well machine learning algorithms such as support vector machines can classify the genomic regions of the input dataset into positives vs. negatives, based on different combinations of prediction attributes (Figure 5); and (vi) the prediction analysis module predicts whether a genomic region that is not contained in the input dataset belongs to the positives or to the negatives, thus exploiting any correlations detected by the machine learning analysis module for the prediction of new data (Figure 9). As an additional option, all modules allow for taking into account adjacent windows upstream and downstream of the regions of interest, which is often useful when analyzing promoter regions.

A typical EpiGRAPH analysis follows a pre-defined workflow. Initially, the user uploads an input dataset to the web server, which may be derived from wet-lab analysis (e.g. ChIP-on-chip experiments) or prior bioinformatic calculations (e.g. computational screening for regions that are under selective pressure). This input dataset takes the form of a table of genomic regions (i.e. containing columns for chromosome, start position and end position), with binary class values specifying for each region whether it belongs to the positives or negatives. When no class value is given, EpiGRAPH regards all genomic regions of the input dataset as positives and assists the user with calculating a set of random control regions to be used as negatives. After submission of the analysis request, EpiGRAPH calculates a large number of potentially relevant attributes for each genomic region in the input dataset. Most of these attributes take the form of overlap frequencies or score values, quantifying the co-localization of the genomic regions in the input dataset with publicly available annotation data for the respective genome. Upon completion of the attribute calculation (which can take hours or even days for large input datasets), EpiGRAPH's statistical and machine learning modules calculate an initial assessment of significant differences between the positives and negatives and an assessment of whether or not these differences are sufficient for bioinformatic prediction. Beyond the initial analysis, the user can specify follow-up analyses based on the pre-calculated attributes. In particular, the diagram generation module can be used to visualize the most interesting differences between positives and negatives as detected by the statistical analysis, and the prediction analysis module lets the user predict the class value of new regions, for example in order to extrapolate experimental data to genomic regions that were not covered by the experiment.

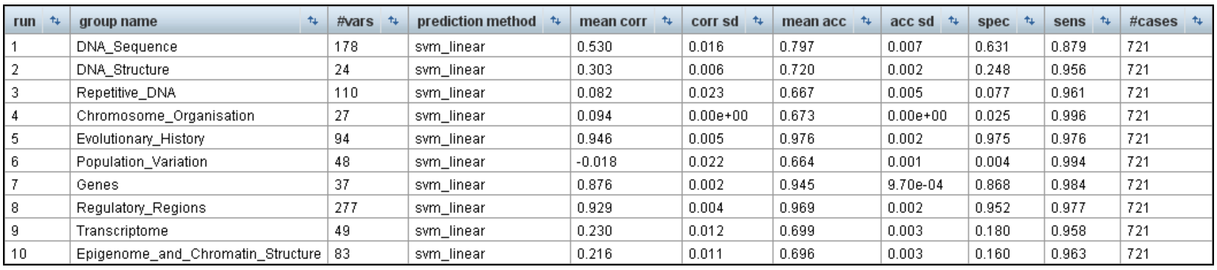
The key to EpiGRAPH's practical utility is its database, for which we collected a large number of attributes that are likely to play a role in genome function and epigenetic regulation. For the most thoroughly annotated human genome, EpiGRAPH currently includes more than a thousand attributes, falling into twelve attribute groups (see Table 6 for an overview and http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html for details): (i) DNA sequence, (ii) DNA structure, (iii) repetitive DNA, (iv) chromosome organization, (v) evolutionary history, (vi) population variation, (vii) genes, (viii) regulatory regions, (ix) transcriptome, (x) epigenome and chromatin structure, (xi) ENCODE transcriptome and (xii) ENCODE epigenome and chromatin structure. EpiGRAPH also incorporates the genomes of chimp, mouse and chicken, with slightly lower numbers of attributes (further genomes will be added according to user demand). In addition to using EpiGRAPH's default attributes, users can upload new datasets and define custom attributes for use inside EpiGRAPH, which is particularly useful when relevant experimental data are available for a specific analysis.

To demonstrate the practical use and utility of EpiGRAPH, the following subsections describe two case studies in which EpiGRAPH has been applied to real-world biological problems. Furthermore, several video tutorials are available online (http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html), which provide a step-by-step introduction into using EpiGRAPH for genome analysis and epigenome prediction.

Case study 1: EpiGRAPH identifies epigenetic and gene regulatory properties of ultraconserved elements

In genome research, evolutionary conservation is considered a major predictor of functional relevance (Gomez-Skarmeta et al. 2006). It has thus puzzled researchers that some of the most conserved genomic regions are located in gene deserts, rather than in protein-coding genes or other genomic regions with well-established biological function (Bejerano et al. 2004). Recent results confirm that these “ultraconserved elements” are indeed subject to strong selective pressure and are not just genomic coldspots with low mutation rates (Katzman et al. 2007). On the other hand, it has been shown that genetically engineered mice lacking selected ultraconserved elements are viable and phenotypically normal under lab conditions (Ahituv et al. 2007).

To help elucidate potential biological functions of ultraconserved elements on a molecular level, EpiGRAPH analyses were performed on three published sets of ultraconserved elements (Derti et al. 2006): (i) human-rodent ultraconserved elements (present in human, mouse and rat), (ii) mammalian ultraconserved elements (present in human, mouse and dog) and (iii) mammalian-avian ultraconserved elements (present in human and chicken). Initially, we let EpiGRAPH compare the set of human-rodent ultraconserved elements with a set of random control regions, which were derived according to EpiGRAPH’s default parameters (same chromosome and length distribution as input dataset, strictly limited deviation in terms of GC content, repeat content and exon overlap). The results of the machine learning analysis – which is typically a good starting point for interpreting EpiGRAPH’s results – provide us with a global and quantitative assessment of the degree to which different attribute groups are predictive of ultraconserved elements. Not surprisingly, the attribute group “evolutionary history” is most highly correlated with the class attribute (Figure 5). In other words, the attributes in this group are collectively most predictive of whether a genomic region is an ultraconserved element or whether it belongs to the random control set. Furthermore, above-random prediction accuracy (i.e. non-zero correlations) were also observed for “DNA sequence”, “DNA structure”, “genes”, “regulatory regions”, “transcriptome”, and “epigenome and chromatin structure” (Figure 5).

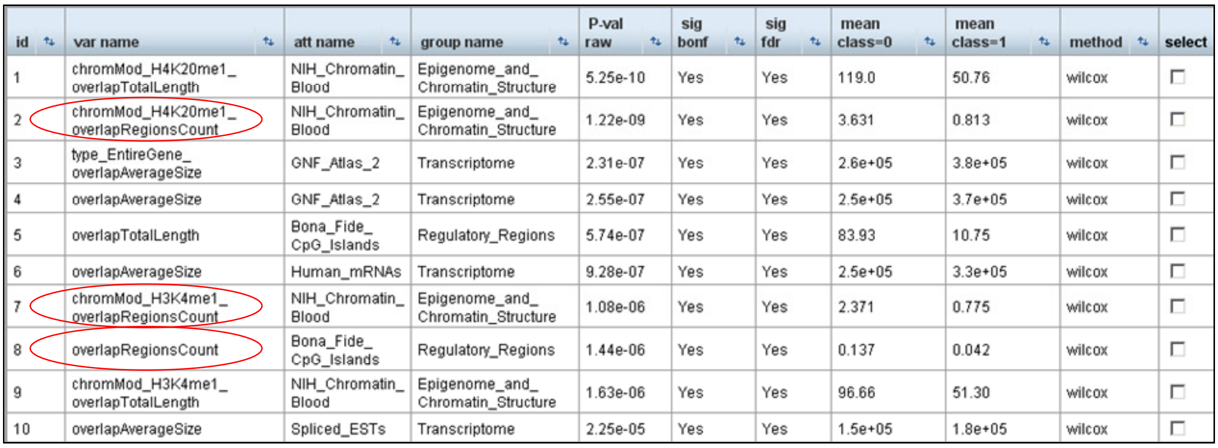


run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	spec	sens	#cases
1	DNA_Sequence	178	svm_linear	0.530	0.016	0.797	0.007	0.631	0.879	721
2	DNA_Structure	24	svm_linear	0.303	0.006	0.720	0.002	0.248	0.956	721
3	Repetitive_DNA	110	svm_linear	0.082	0.023	0.667	0.005	0.077	0.961	721
4	Chromosome_Organisation	27	svm_linear	0.094	0.00e+00	0.673	0.00e+00	0.025	0.996	721
5	Evolutionary_History	94	svm_linear	0.946	0.005	0.976	0.002	0.975	0.976	721
6	Population_Variation	48	svm_linear	-0.018	0.022	0.664	0.001	0.004	0.994	721
7	Genes	37	svm_linear	0.876	0.002	0.945	9.70e-04	0.868	0.984	721
8	Regulatory_Regions	277	svm_linear	0.929	0.004	0.969	0.002	0.952	0.977	721
9	Transcriptome	49	svm_linear	0.230	0.012	0.699	0.003	0.180	0.958	721
10	Epigenome_and_Chromatin_Structure	83	svm_linear	0.216	0.011	0.696	0.003	0.160	0.963	721

Figure 5. Results screenshot of EpiGRAPH’s machine learning module quantifying the predictability of ultraconserved elements based on different groups of (epi-) genomic attributes

This screenshot displays the results of a machine learning analysis of human-rodent ultraconserved elements vs. a randomly drawn control set with matched chromosome and length distribution as well as similar values for GC content, repeat content and exon overlap (EpiGRAPH’s default parameters were used). The values in the table correspond to the average performance of a linear support vector machine that was trained and evaluated in ten repetitions of a tenfold cross-validation, and they include mean correlation (mean corr), prediction accuracy (mean acc), sensitivity (sens) and specificity (spec). Additional columns display standard deviations observed among the repeated cross-validations with random partition assignment (corr sd and acc sd), the number of variables in each attribute group (#vars) and the total number of genomic regions included in the analysis (#cases).

We next asked whether epigenetic or regulatory differences exist between the three different groups of ultraconserved elements, which could provide hints on their molecular function. A pairwise EpiGRAPH comparison between human-rodent and mammalian ultraconserved elements did not identify a single significantly different attribute, suggesting that no clear-cut functional differences exist. In contrast, comparison between mammalian and mammalian-avian ultraconserved elements identified 30 significant attributes from the groups “regulatory regions”, “transcriptome”, and “epigenome and chromatin structure” (Figure 6), with a false discovery rate threshold of 0.05 (EpiGRAPH's default significance threshold). The significant attributes include the frequency with which these ultraconserved elements overlap with the repressive histone modification H4K20me1 (P = 1.2 · 10^{-9}, 4.5 times enriched in mammalian as compared to mammalian-avian ultraconserved elements), the frequency of overlap with the activating histone modification H3K4me1 (P = 1.1 · 10^{-6}, 3.1 times enriched) and the frequency of overlap with bona fide CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis), which are indicative of an open and transcriptionally accessible chromatin structure (P = 1.4 · 10^{-6}, 3.3 times enriched). Hence, mammalian ultraconserved elements seem to be more highly regulated by both activating and repressive mechanisms than their mammalian-avian counterparts.



id	var name	att name	group name	P-val raw	sig bonf	sig fdr	mean class=0	mean class=1	method	select
1	chromMod_H4K20me1_ overlapTotalLength	NIH_Chromatin_ Blood	Epigenome_and_ Chromatin_Structure	5.25e-10	Yes	Yes	119.0	50.76	wilcox	□
2	chromMod_H4K20me1_ overlapRegionsCount	NIH_Chromatin_ Blood	Epigenome_and_ Chromatin_Structure	1.22e-09	Yes	Yes	3.631	0.813	wilcox	□
3	type_EntireGene_ overlapAverageSize	GNF_Atlas_2	Transcriptome	2.31e-07	Yes	Yes	2.6e+05	3.8e+05	wilcox	□
4	overlapAverageSize	GNF_Atlas_2	Transcriptome	2.55e-07	Yes	Yes	2.5e+05	3.7e+05	wilcox	□
5	overlapTotalLength	Bona_Fide_ CpG_Islands	Regulatory_Regions	5.74e-07	Yes	Yes	83.93	10.75	wilcox	□
6	overlapAverageSize	Human_mRNAs	Transcriptome	9.28e-07	Yes	Yes	2.5e+05	3.3e+05	wilcox	□
7	chromMod_H3K4me1_ overlapRegionsCount	NIH_Chromatin_ Blood	Epigenome_and_ Chromatin_Structure	1.08e-06	Yes	Yes	2.371	0.775	wilcox	□
8	overlapRegionsCount	Bona_Fide_ CpG_Islands	Regulatory_Regions	1.44e-06	Yes	Yes	0.137	0.042	wilcox	□
9	chromMod_H3K4me1_ overlapTotalLength	NIH_Chromatin_ Blood	Epigenome_and_ Chromatin_Structure	1.63e-06	Yes	Yes	96.66	51.30	wilcox	□
10	overlapAverageSize	Spliced_ESTs	Transcriptome	2.25e-05	Yes	Yes	1.5e+05	1.8e+05	wilcox	□

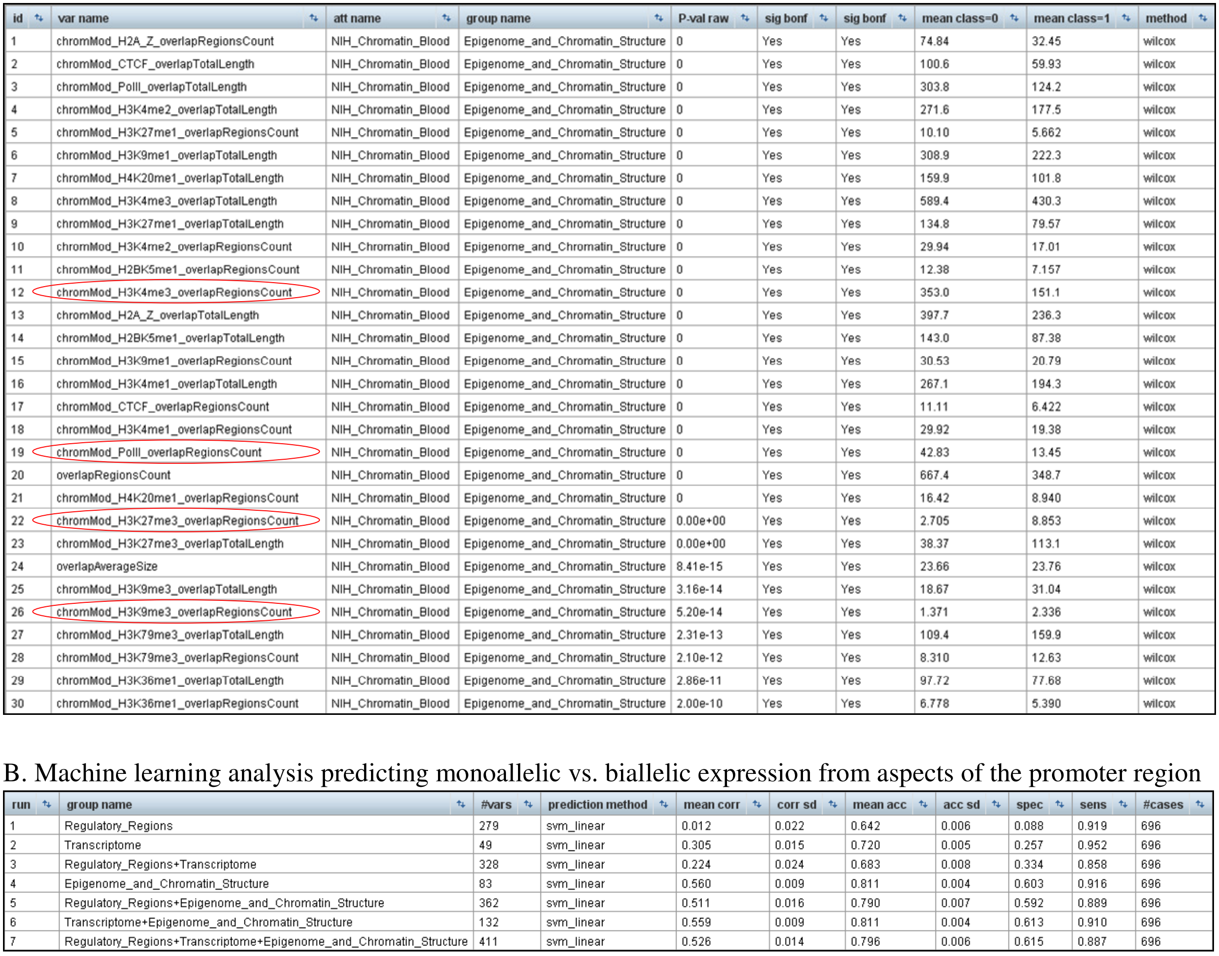
Figure 6. Results screenshot of EpiGRAPH's statistical analysis module identifying significant gene regulatory differences between ultraconserved elements that are restricted to mammals and those that are also present in birds

This screenshot displays the results of a statistical analysis comparing ultraconserved elements identified for human, mouse and rat (class = 0) with those identified for human and chicken (class = 1), in terms of three attribute groups: "regulatory regions", "transcriptome", and "epigenome and chromatin structure". Statistical testing was performed using the nonparametric Wilcoxon rank-sum test and P-values were adjusted for multiple testing using the highly conservative Bonferroni method (sig bonf) as well as the false discovery rate method (sig fdr). A global significance threshold of 0.05 was used for both methods. With the "select" column on the right, EpiGRAPH provides the option of requesting boxplots visualizing any attribute's distribution among the two classes (see Figure 8 for an example). Highlighted attributes are discussed in the text, and an explanation of the attribute names is available from the EpiGRAPH website: http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html.

Case study 2: EpiGRAPH predicts monoallelic gene expression based on its characteristic pattern of histone modifications

While the majority of human genes are expressed from both alleles, a sizable proportion is expressed exclusively from a single allele, with important biological consequences: (i) genomic imprinting, i.e. parent-specific mono-allelic gene expression, plays a critical role in normal development and gives rise to non-Mendelian patterns of inheritance (Reik 2007); (ii) X-chromosome inactivation in females leads to mitotically heritable silencing of one randomly selected X chromosome (Heard 2004); and (iii) random monoallelic gene expression, e.g. of odorant receptor genes and immune-system related genes, increases the phenotypic diversity among clonal cells (Gimelbrant et al. 2007). While it is evident that epigenetic regulation plays a role in monoallelic gene expression and attempts have been made to predict genomic imprinting and X-chromosome inactivation from the DNA sequence (Luedi et al. 2007; Luedi et al. 2005; Wang et al. 2006), a genome-wide analysis of the determinants of monoallelic expression is currently lacking.

A. Statistical analysis comparing promoter regions of monoallelically vs. biallelically expressed genes



id	var name	att name	group name	P-val raw	sig bonf	sig bonf	mean class=0	mean class=1	method
1	chromMod_H2A_Z_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	74.84	32.45	wilcox
2	chromMod_CTCF_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	100.6	59.93	wilcox
3	chromMod_PolII_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	303.8	124.2	wilcox
4	chromMod_H3K4me2_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	271.6	177.5	wilcox
5	chromMod_H3K27me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	10.10	5.662	wilcox
6	chromMod_H3K9me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	308.9	222.3	wilcox
7	chromMod_H4K20me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	159.9	101.8	wilcox
8	chromMod_H3K4me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	589.4	430.3	wilcox
9	chromMod_H3K27me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	134.8	79.57	wilcox
10	chromMod_H3K4me2_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	29.94	17.01	wilcox
11	chromMod_H2BK5me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	12.38	7.157	wilcox
12	chromMod_H3K4me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	353.0	151.1	wilcox
13	chromMod_H2A_Z_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	397.7	236.3	wilcox
14	chromMod_H2BK5me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	143.0	87.38	wilcox
15	chromMod_H3K9me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	30.53	20.79	wilcox
16	chromMod_H3K4me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	267.1	194.3	wilcox
17	chromMod_CTCF_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	11.11	6.422	wilcox
18	chromMod_H3K4me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	29.92	19.38	wilcox
19	chromMod_PolII_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	42.83	13.45	wilcox
20	overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	667.4	348.7	wilcox
21	chromMod_H4K20me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	16.42	8.940	wilcox
22	chromMod_H3K27me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0.00e+00	Yes	Yes	2.705	8.853	wilcox
23	chromMod_H3K27me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0.00e+00	Yes	Yes	38.37	113.1	wilcox
24	overlapAverageSize	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	8.41e-15	Yes	Yes	23.66	23.76	wilcox
25	chromMod_H3K9me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	3.16e-14	Yes	Yes	18.67	31.04	wilcox
26	chromMod_H3K9me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	5.20e-14	Yes	Yes	1.371	2.336	wilcox
27	chromMod_H3K79me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	2.31e-13	Yes	Yes	109.4	159.9	wilcox
28	chromMod_H3K79me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	2.10e-12	Yes	Yes	8.310	12.63	wilcox
29	chromMod_H3K36me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	2.86e-11	Yes	Yes	97.72	77.68	wilcox
30	chromMod_H3K36me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	2.00e-10	Yes	Yes	6.778	5.390	wilcox

B. Machine learning analysis predicting monoallelic vs. biallelic expression from aspects of the promoter region

run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	spec	sens	#cases
1	Regulatory_Regions	279	svm_linear	0.012	0.022	0.642	0.006	0.088	0.919	696
2	Transcriptome	49	svm_linear	0.305	0.015	0.720	0.005	0.257	0.952	696
3	Regulatory_Regions+Transcriptome	328	svm_linear	0.224	0.024	0.683	0.008	0.334	0.858	696
4	Epigenome_and_Chromatin_Structure	83	svm_linear	0.560	0.009	0.811	0.004	0.603	0.916	696
5	Regulatory_Regions+Epigenome_and_Chromatin_Structure	362	svm_linear	0.511	0.016	0.790	0.007	0.592	0.889	696
6	Transcriptome+Epigenome_and_Chromatin_Structure	132	svm_linear	0.559	0.009	0.811	0.004	0.613	0.910	696
7	Regulatory_Regions+Transcriptome+Epigenome_and_Chromatin_Structure	411	svm_linear	0.526	0.014	0.796	0.006	0.615	0.887	696

Figure 7. EpiGRAPH results screenshots indicating that promoters of monoallelically expressed genes are enriched with repressive histone modifications and can be predicted bioinformatically

These screenshots display the results of an EpiGRAPH analysis comparing the promoter regions of genes exhibiting monoallelic (class = 1) vs. biallelic (class = 0) gene expression. The results of the statistical analysis module (panel A) show that the promoters of monoallelically expressed genes are more likely to exhibit repressive histone modifications (such as H3K27 trimethylation) and less likely to exhibit activating histone modifications (such as H3K4 trimethylation) than those of biallelically expressed genes. According to the results of the machine learning analysis module (panel B), these differences are sufficient to predict monoallelic gene expression with significant accuracy, using a linear support vector machine and epigenetic characteristics of the promoter region as prediction attributes. The figure format is identical to Figure 5 and Figure 6.

We applied EpiGRAPH to an experimentally derived dataset of monoallelic vs. biallelic expression for 4,000 human genes (Gimelbrant et al. 2007), in order to identify characteristic patterns at the promoter regions of monoallelically expressed genes and to predict their location genome-wide. We focused our analysis on three attribute groups for which a relation to monoallelic gene expression is most plausible biologically, namely “regulatory regions”, “transcriptome”, and “epigenome and chromatin structure”. (It is generally a good idea when working with EpiGRAPH to focus on promising attribute groups rather than always running a full-blown analysis, because this will reduce the overall computation time as well as the multiple-testing penalty incurred in the statistical analysis.) EpiGRAPH’s statistical analysis identified highly significant differences between the promoter regions ranging from 1,250 bp upstream to 250 bp downstream of the annotated transcription start site, depending on the expression status of the associated gene (Figure 7A). Promoter regions of monoallelically expressed genes are enriched with repressive histone modifications such as H3K27 trimethylation ($P < 10^{-20}$, 3.3 times enriched) and H3K9 trimethylation ($P = 5.2 \cdot 10^{-14}$, 1.7 times enriched), but depleted in terms of the frequency of polymerase II binding ($P < 10^{-20}$, 3.2 times depleted) and activating histone modifications such as H3K4 trimethylation ($P < 10^{-20}$, 2.3 times depleted). To visualize these differences, boxplots were created using EpiGRAPH's diagram generation module (Figure 8).

Machine learning analysis shows that the differences between the two classes are sufficient to predict with more than 80% accuracy whether or not a gene is monoallelically expressed, based epigenome data (Figure 7B). We thus applied EpiGRAPH's prediction analysis module (Figure 9) to generate a genome-wide map of monoallelically expressed genes. This map constitutes the first comprehensive annotation of allele-specific gene expression in the human genome, and – given its estimated 92% sensitivity (Figure 7B) – it should provide a useful resource for biologists trying to elucidate the many biological roles of monoallelic gene expression.

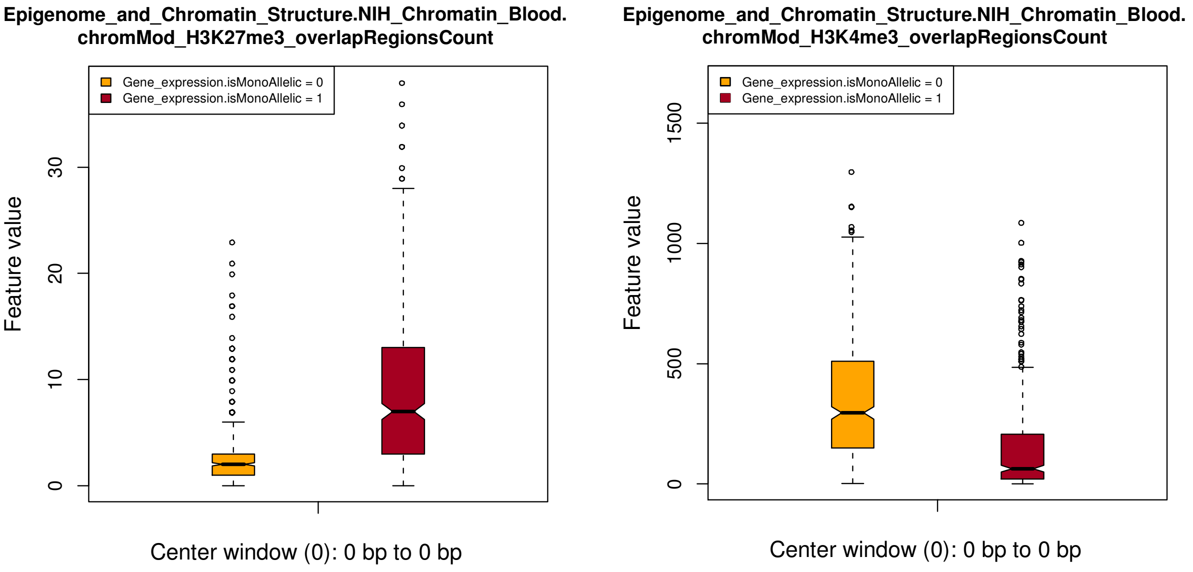
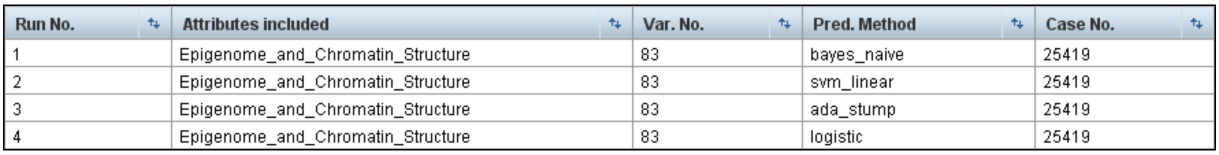


Figure 8. Results screenshot of EpiGRAPH's diagram generation module highlighting differential histone modification patterns for the promoters of monoallelically vs. biallelically expressed genes

These screenshots display EpiGRAPH-generated boxplots comparing the promoter regions of genes exhibiting monoallelic (class = 1) vs. biallelic gene expression (class = 0) in terms of their frequency of overlap with histone H3K27 trimethylation (left) and histone H3K4 trimethylation (right). The boxplots are in standard format (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box) and outliers are shown as circles.



Run No.	Attributes included	Var. No.	Pred. Method	Case No.
1	Epigenome_and_Chromatin_Structure	83	bayes_naive	25419
2	Epigenome_and_Chromatin_Structure	83	svm_linear	25419
3	Epigenome_and_Chromatin_Structure	83	ada_stump	25419
4	Epigenome_and_Chromatin_Structure	83	logistic	25419

Figure 9. Results screenshot of EpiGRAPH's prediction analysis identifying candidates for monoallelic expression among all human genes

This screenshot summarizes an EpiGRAPH prediction analysis performed on the promoter regions (-1,250 bp to 250 bp surrounding the annotated transcription start site) of all RefSeq-annotated genes. Briefly, four machine learning methods – a support vector machine with linear kernel, a boosting algorithm, logistic regression and the naïve Bayes algorithm – are trained on the input dataset of promoter regions with known expression status, and the trained models are used to predict the status of all RefSeq-annotated genes. To obtain a confidence criterion, EpiGRAPH repeats this procedure ten times using bootstrapped samples of the training dataset.

Adapting and extending EpiGRAPH

The case studies described above highlight the diversity of options that the EpiGRAPH web service provides even without customization. However, EpiGRAPH can be much more powerful when the user exploits its options for configuration and extension. We here describe five ways in which EpiGRAPH can be customized, in increasing order of complexity and power.

First, it is possible to use EpiGRAPH for attribute calculation only, thus profiting from EpiGRAPH's large and carefully selected set of default attributes, while performing follow-up analyses offline (e.g. by the R statistics package). To that end, the user performs a normal EpiGRAPH analysis and presses the "Download Data Table" button on the results page to obtain a large, tab-separated data file containing all attribute values for all genomic regions in the input dataset.

Second, the user can add custom genomic attributes to EpiGRAPH, using the "Upload Custom Attribute Dataset" button on the overview page. A new custom attribute can be defined in three ways: (i) by uploading a set of genomic regions; (ii) by specifying how the attribute can be calculated from other attributes that are already present in the database (e.g. filtering rows that match a specific condition or defining additional columns); and (iii) by deriving a randomized control attribute that matches an existing attribute in terms of its GC content, repeat content and/or exon overlap. Custom attributes can be included in EpiGRAPH analyses just like any of the default attributes, but they are exclusively accessible to the user who has defined them.

Third, the user can specify advanced analysis requests and attribute calculations directly in EpiGRAPH's XML format. Internally, the EpiGRAPH web frontend encodes all analysis requests in the standardized X-GRAF XML format, before they are transmitted to EpiGRAPH's middleware and backend (see Methods and Figure 3 for details). Through the "Execute Analysis Based on Existing XML File" button on the overview page, the user can upload XML files conforming to the X-GRAF format directly, thus bypassing the web frontend. This can be useful for several reasons: (i) When running the same analysis on different datasets, it is often convenient to initially design the analysis using the web frontend, then download the X-GRAF file and use a text editor or a custom script to produce separate versions for each dataset. (ii) Sharing X-GRAF files with other researchers (e.g. by inclusion in the supplementary data of a paper) will enable them to reproduce the analysis by simply submitting the X-GRAF files back to the EpiGRAPH web service, thus contributing to reproducible research (Gentleman 2005). (iii) Some of the more advanced features (e.g. calculated attributes with multiple new columns) are supported by the calculation engine but cannot be specified easily using the web frontend.

Fourth, the user can download a "light" version of the EpiGRAPH calculation engine for local installation, which runs on any computer with recent versions of Python (http://www.python.org/), R (for statistical analysis, http://www.r-project.org/) and Weka (for machine learning analysis, http://www.cs.waikato.ac.nz/~ml/weka/), after a few additional libraries have been installed. The "light" version (source code available from http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html) is particularly useful for researchers developing new bioinformatic methods for genome analysis, such as new flavors of the statistical analysis, diagram generation, machine learning analysis and prediction analysis, but who do not want to spend their time writing code for attribute calculation. The main disadvantage of the "light" version is that in the absence of a relational database all genomic attributes have to be stored in flat files. However, the "light" version is code-compatible with the full version of EpiGRAPH. Hence it is possible to develop and test new modules using the "light" version and to incorporate the completed modules into the EpiGRAPH web service.

Fifth, the user can download and install the full version of EpiGRAPH (source code available on request), which includes the process control middleware and the web frontend components as well as a version of the calculation engine that provides full database support. While running a full-blown EpiGRAPH server locally is a non-trivial task and requires both a Java application server (e.g. Tomcat, http://tomcat.apache.org/) and an Oracle database server with XML DB support (http://www.oracle.com/technology/tech/xml/xmldb/index.html), this setting gives the user full flexibility for customizing EpiGRAPH and a powerful infrastructure for genome analysis.

B-3.4 Discussion

EpiGRAPH contributes to a new generation of powerful and easy-to-use genome analysis tools that enable biologists to perform complex bioinformatic analyses online – without having to learn a programming language or downloading and manually processing large datasets. With its focus on statistical and machine learning methods, EpiGRAPH goes substantially beyond existing tools, providing a workflow that helps to uncover biologically meaningful associations among genome-scale datasets. We highlighted EpiGRAPH's utility with two case studies. First, we showed that mammalian-specific ultraconserved elements exhibit a distinct epigenetic profile when compared to more widely detectable ultraconserved elements. Second, we identified patterns of histone modifications that are significantly associated with monoallelic gene expression, and we exploited this finding for predicting allele-specific expression for all annotated genes in the human genome.

EpiGRAPH integrates well with existing bioinformatics resources and infrastructure. It can be regarded as part of a three-step data analysis pipeline involving genome browsers, genome calculators and genome data analysis tools (Figure 10): (i) Researchers typically start the analysis of new genome-scale datasets by uploading a pre-processed and quality-controlled data file into a genome browser, where it can be visualized and manually inspected. The UCSC Genome Browser (Karolchik et al. 2008) is popular for this task, due to the ease with which custom data tracks can be displayed alongside public genome annotations. (ii) Based on initial observations, it is usually necessary to pick a subset of genomic regions for further analysis, e.g. all promoter regions that are bound by a specific protein. The Galaxy web service (Blankenberg et al. 2007; Giardine et al. 2005) is well-suited to performing the necessary calculations and filtering, and the UCSC Genome Browser as well as EpiGRAPH also implement features that facilitate the selection of interesting regions for further analysis. (iii) Finally, it is often desirable to perform sophisticated statistical analysis and data mining on the potentially large set of interesting regions, in order to discover, test and interpret correlations and interactions with biologically interesting phenomena. For this step, a comprehensive and easy-to-use toolkit has been lacking. We developed EpiGRAPH to fill this gap, thereby enabling biologists to perform advanced bioinformatic analysis and prediction with little need for bioinformatic support. We demonstrate the interplay of UCSC Genome Browser, Galaxy and EpiGRAPH by a case study focusing on the (epi-) genomic characteristics of highly polymorphic promoter regions in the human genome, which is available as a step-by-step video tutorial from http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html.

In the future, we anticipate that the three layers of genome browsing, calculation and analysis tools will increasingly merge into a single application, for which "statistical genome browser" might be an appropriate term. To that end, it will be neither necessary nor beneficial to integrate all functionality and underlying databases into a single monolithic tool. Instead, a distributed network of interoperable genome analysis web services is likely to emerge. A genome browser could act as a single point of entry, from which the user initiates a complex analysis. The analysis is then split into logical blocks, encoded in an XML-based analysis description language (such as the X-GRAF format prototyped in EpiGRAPH) and distributed over the internet to those calculation servers at which all required datasets and analysis modules are available. Finally, the decentrally calculated results are merged and displayed to the user at the central genome browser frontend. EpiGRAPH was developed with this scenario in mind and prototypes software paradigms required for distributed genome analysis by concerted action of specialized tools.

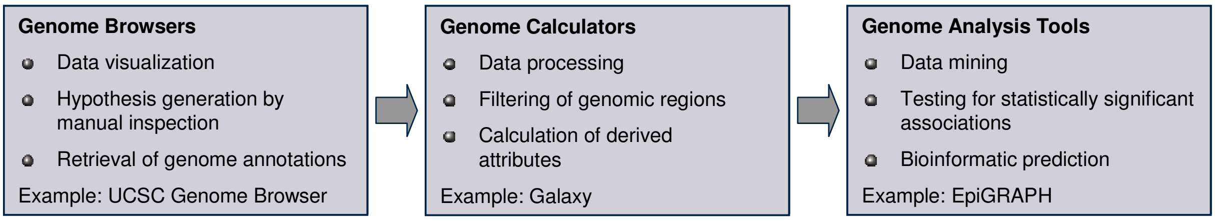

Genome Browsers
Data visualization
Hypothesis generation by manual inspection
Retrieval of genome annotations
Example: UCSC Genome Browser
Genome Calculators
Data processing
Filtering of genomic regions
Calculation of derived attributes
Example: Galaxy
Genome Analysis Tools
Data mining
Testing for statistically significant associations
Bioinformatic prediction
Example: EpiGRAPH

Figure 10. Exemplary workflow for web-based analysis of large genome and epigenome datasets

This figure outlines a workflow of (epi-) genome data analysis using publicly available web services. Initially, the user uploads a newly generated dataset into a genome browser for visualization and hypothesis generation by visual inspection (left box). Next, he or she processes the data with a genome calculator such as Galaxy, in order to extract and prepare interesting regions for in-depth analysis (center box). Finally, genome analysis tools such as EpiGRAPH can be used to test for significant associations with genome annotation data and to perform bioinformatic prediction (right box).

B-4 CpG island mapping by epigenome prediction[1]

B-4.1 Motivation

Having established the EpiGRAPH tool for epigenome analysis and prediction, we were able to embark on a project aimed at reconciling the two facets of CpG islands, namely their epigenetic function and their sequence-based definition. CpG islands are genomic regions characterized by an exceptionally high CpG dinucleotide frequency (Bird 2002; Bird 1986; Caiafa and Zampieri 2005). They are among the most important regulatory regions in vertebrate genomes, with functional roles for both normal and disease-related gene expression (Antequera 2003; Laird 2005).

Originally, CpG islands were discovered by virtue of an epigenetic property, namely the absence of DNA methylation: when the human genome was experimentally digested with methylation-sensitive restriction enzymes, some genomic regions were cut into small fragments, while the bulk of the genome remained uncut (Cooper et al. 1983). Since the restriction enzyme used (HpaII) cuts DNA only at unmethylated CpG dinucleotides, it was concluded that a small but significant fraction of the genome is reproducibly unmethylated.

[1] This chapter describes published work conducted in collaboration with Martina Paulsen and Jörn Walter (Bock et al. 2007), who contributed to the interpretation of the results.

After a sample of these so-called HpaII tiny fragments had been sequenced, it became obvious that they were highly GC-rich and CpG-rich (Bird 1986). In an early computational analysis, this observation was utilized to define such regions as CpG islands, and a simple set of criteria was suggested in order to identify CpG islands based on their DNA sequence alone (Gardiner-Garden and Frommer 1987). According to these Gardiner-Garden sequence criteria a genomic region has to fulfill three conditions in order to classify as a CpG island: (1) GC content above 50%, (2) ratio of observed to expected number of CpG dinucleotides above 0.6, and (3) length greater than 200 basepairs (bp). Because the amount of sequence data strongly outnumbered the experimental data available for DNA methylation, this definition quickly replaced the original methylation-based concept.

Since their initial discovery, CpG islands have been subject to extensive research. Today it is known that CpG islands (according to the DNA sequence criteria mentioned above or slightly modified variants) associate with more than three-quarters of all known transcription start sites (Bajic et al. 2006) and with 88% of active promoters identified in primary fibroblasts (Kim et al. 2005), indicating that they bear important regulatory functions. Furthermore, they are hotspots of specific histone modifications (Bernstein et al. 2005; Roh et al. 2005), they frequently bind ubiquitous transcription factors such as SP1 (Cawley et al. 2004), and they exhibit particularly accessible chromatin structures (Crawford et al. 2006). For these reasons, CpG islands are routinely used for a wide range of tasks in genome analysis and annotation. For example, they play a fundamental role for promoter prediction (Bajic et al. 2004), and their use as candidate regions of aberrant DNA methylation has contributed significantly to our understanding of epigenetics (Ushijima 2005).

However, the current sequence-based definitions of CpG islands (Gardiner-Garden and Frommer 1987; Takai and Jones 2002) incur several disadvantages, which hamper both their theoretical and practical value. First, they are based on three threshold parameters that lack a clear biological justification. For example, it is unclear why 200 bp should be the most appropriate minimum length to define CpG islands, especially since even a random permutation of the genome sequence would give rise to a substantial number of CpG islands with this minimum length according to the Gardiner-Garden criteria. A minimum length of 500 bp is also widely used, and its use was motivated by its ability to exclude most Alu-repeat-associated regions (Takai and Jones 2002), but again, no systematic analysis or parameter selection method has been performed to justify this particular value.

Second, current definitions are purely dichotomic, i.e. a particular region either qualifies as a CpG island or it does not. This approach not only fails to account for the fact that CpG islands can differ considerably in terms of their sequence composition and epigenetic states. It can also lead to non-intuitive special cases. For example, even if a short CpG-rich region fails to fulfill CpG island criteria on its own, the same region may well fulfill the criteria after small and seemingly unrelated changes of a few neighboring nucleotides. Thus, the mapping of CpG islands is inherently unstable and depends not only on the definition used but also on the exact implementation of the mapping software. In contrast, the introduction of a numerical score for CpG island strength would allow for distinguishing weak, intermediate, and strong CpG islands, without the necessity of a fixed all-or-nothing threshold.

Third, and most critically, sequence-based CpG island criteria fail to distinguish between "bona fide" CpG islands on the one hand, i.e. CpG islands that are consistent with the original notion of CpG islands as unmethylated genomic regions that serve as transcription regulators and exhibit an open and transcriptionally competent chromatin structure (Bird 1986), and

CpG-rich regions lacking these characteristics on the other hand. More precisely, current CpG island criteria seem to be sufficiently sensitive in the sense that they detect most bona fide CpG islands in the human genome, but their specificity is low, i.e. they give rise to a substantial number of false positive classifications. For example, Yamada et al. observed that almost a third of the putative CpG islands analyzed showed significant DNA methylation (Yamada et al. 2004), in contradiction with the original concept of CpG islands as unmethylated regions.

In order to resolve the significant drawbacks of current sequence-based CpG island criteria, it was suggested to abandon the concept of CpG islands altogether and to replace it by direct counting of CpG dinucleotides (Saxonov et al. 2006). In this study, we propose a less radical but arguably more viable strategy. Our approach maintains the high sensitivity of current CpG island criteria, but substantially improves their specificity, it introduces a more biologically meaningful way of selecting thresholds, and it accounts for the fact that CpG islands quantitatively differ in their strength.

The fundamental concept of this study is to combine an initial, sequence-based mapping of CpG islands with subsequent prediction of CpG island strength. CpG island strength is expressed as a single quantitative score per CpG island, summarizing its inherent tendency – across different cell types and tissues – to exhibit an unmethylated, open, and transcriptionally competent chromatin structure. It is calculated as a combination of epigenome predictions and provides a measure for discrimination between bona fide CpG islands on the one hand and regions that are just CpG-rich but show no evidence of the epigenetic and functional characteristics of bona fide CpG islands on the other hand. We evaluate the predicted CpG island scores by comparison with large-scale experimental datasets on DNA methylation and transcription initiation sites, since absence of DNA methylation and presence of promoter activity are regarded as characteristic of bona fide CpG islands.

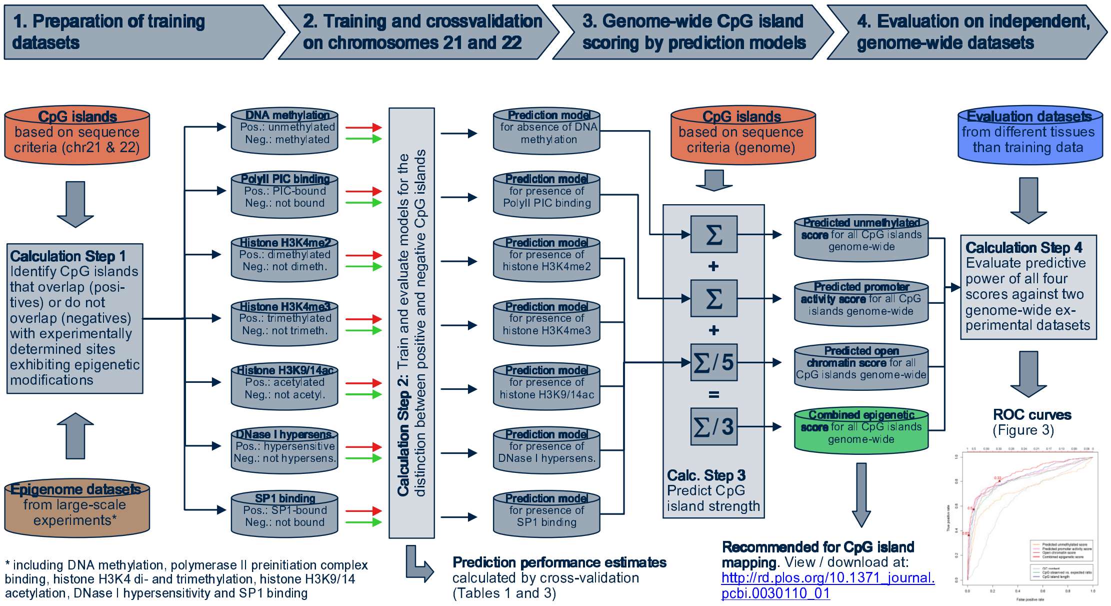

1. Preparation of training datasets
2. Training and crossvalidation on chromosomes 21 and 22
3. Genome-wide CpG island scoring by prediction models
4. Evaluation on independent, genome-wide datasets
CpG islands based on sequence criteria (chr21 & 22)
Calculation Step 1 Identify CpG islands that overlap (positives) or do not overlap (negatives) with experimentally determined sites exhibiting epigenetic modifications
Epigenome datasets from large-scale experiments*
DNA methylation Pos.: unmethylated Neg.: methylated
PolyII PIC binding Pos.: PIC-bound Neg.: not bound
Histone H3K4me2 Pos.: dimethylated Neg.: not dimeth.
Histone H3K4me3 Pos.: trimethylated Neg.: not trimeth.
Histone H3K9/14ac Pos.: acetylated Neg.: not acetyl.
DNase I hypersens. Pos.: hypersensitive Neg.: not hypersens.
SP1 binding Pos.: SP1-bound Neg.: not bound
Calculation Step 2: Train and evaluate models for the distinction between positive and negative CpG islands
Prediction model for absence of DNA methylation
Prediction model for presence of PolyII PIC binding
Prediction model for presence of histone H3K4me2
Prediction model for presence of histone H3K4me3
Prediction model for presence of histone H3K9/14ac
Prediction model for presence of DNase I hypersens.
Prediction model for presence of SP1 binding
* including DNA methylation, polymerase II preinitiation complex binding, histone H3K4 di- and trimethylation, histone H3K9/14 acetylation, DNase I hypersensitivity and SP1 binding
Prediction performance estimates calculated by cross-validation (Tables 1 and 3)
CpG islands based on sequence criteria (genome)
Σ + Σ + $\Sigma/5$ = $\Sigma/3$
Calc. Step 3 Predict CpG island strength
Predicted unmethylated score for all CpG islands genome-wide
Predicted promoter activity score for all CpG islands genome-wide
Predicted open chromatin score for all CpG islands genome-wide
Combined epigenetic score for all CpG islands genome-wide
Recommended for CpG island mapping. View / download at: http://rd.plos.org/10.1371_journal.pcbi.0030110_01
Evaluation datasets from different tissues than training data
Calculation Step 4 Evaluate predictive power of all four scores against two genome-wide experimental datasets
ROC curves (Figure 3)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Predicted unmethylated score
Predicted promoter activity score
Open chromatin score
Combined epigenetic score
GC content
CpG observed vs. expected ratio
CpG island length

Figure 11. Conceptual overview of CpG island mapping by epigenome prediction

This figure outlines the workflow used in this study to derive quantitative scores of CpG island strength, and to evaluate their performance as predictors of bona fide CpG islands. The arrows at the top describe the phases of the analysis, the cylinders correspond to input datasets (orange, blue, and brown cylinders) and results datasets (grey and teal cylinders), and the rectangular boxes represent major computational steps. The sigmas in the calculation step 3 box stand for summation over the input. The figure is slightly simplified and focuses on a single CpG island map. In fact, the entire workflow was performed separately for three CpG island maps that differ in the repeat-exclusion strategy used (TJU, GGF, and GGM), with subsequent benchmarking of their performances (Figure 15).

Figure 11 provides a schematic overview of our approach, which is necessarily complex since we derive and benchmark four different scores of CpG island strength using combinations of large-scale epigenome datasets. From left to right, the first step comprises preparation of seven training datasets, based on pairwise overlaps between CpG island maps and epigenome datasets. In the second step, a prediction model is trained and its performance is estimated for each training dataset. The resulting prediction models are then used to score all CpG islands genome-wide. From these scores – in step three – four CpG island scores are calculated. In step four, a performance comparison on two large-scale evaluation datasets shows that the "combined epigenetic score" is the best indicator of CpG island strength and most predictive of bona fide CpG islands. All training and testing in this study is performed on chromosomes 21 and 22 for reasons of data availability. Predictions are calculated and validated on the entire genome. The entire workflow as outlined in Figure 11 was repeated three times, for three widely used CpG island maps. By comparing the results, we show that CpG island strength predictions provide an improvement over each map, and we are able to select the most appropriate setup for the final maps of predicted bona fide CpG islands.

B-4.2 Methods

CpG island maps

In order to calculate genome-wide CpG island maps according to the traditional sequence-based definition, we downloaded both the unmasked and the repeat-masked versions of the hg17 (NCBI35) human genome assembly from the UCSC Genome Browser (Karolchik et al. 2008), and we ran a slightly modified version of the CpG Island Searcher script (Takai and Jones 2002) with the following parameters. Calculation of the TJU map: GC content above 55%, CpG observed vs. expected ratio above 0.65, length above 500 bp, based on the unmasked genome. Calculation of the GGF map: GC content above 50%, CpG observed vs. expected ratio above 0.6, length above 200 bp, based on the unmasked genome. Calculation of the GGM map: GC content above 50%, CpG observed vs. expected ratio above 0.6, length above 200 bp, based on the repeat-masked genome. Finally, for GGF we determined the number of non-repetitive basepairs by comparison with the repeat-masked genome version and discarded all CpG islands for which this value was below 200 bp.

Epigenome prediction

EpiGRAPH (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis) was applied to statistically analyze and predict DNA methylation, promoter activity, and the five components of the open chromatin score, using the following attribute groups: (i) DNA sequence patterns and properties (426 attributes), (ii) repeat attributes, frequency, and distribution (311 attributes), (iii) predicted DNA helix structure (28 attributes), (iv) predicted transcription factor binding sites (68 attributes), (v) evolutionary conservation and single nucleotide polymorphisms (ten attributes), and (vi) CpG island attributes (four attributes). EpiGRAPH's prediction analysis module was used to derive a score for all CpG islands in the human genome. This quantitative prediction can then be used directly as a CpG island score or it can be subjected to further calculations as described below.

Prediction scores for CpG island strength

The calculation of all four CpG island scores made use of EpiGRAPH, combined with appropriate training data. Calculations were performed on the hg17 (NCBI35) genome assembly. Where necessary, data were remapped using the UCSC Genome Browser LiftOver tool (Karolchik et al. 2008).

The predicted unmethylated score is based on training data from an experimental analysis of CpG island methylation in human lymphocytes (Yamada et al. 2004, dataset obtained from the supplementary material). Using a methylation-specific restriction enzyme and PCR, Yamada et al. measured DNA methylation states for 149 CpG-rich regions on chromosome 21q, of which 132 cases showed an unambiguous methylation pattern and could be mapped to the current genome assembly. All CpG islands that overlap (by at least 1 bp) with one of the 103 unmethylated regions were combined into the positive training set and all CpG islands that overlap with one of the 29 methylated cases were combined into the negative training set. The resulting training dataset was then processed with EpiGRAPH in order to derive predicted unmethylated scores for all CpG islands according to TJU, GGF, and GGM.

The predicted promoter activity score is based on training data from an experimental analysis of polymerase II preinitiation complex binding in human fibroblasts (Kim et al. 2005, dataset obtained from the supplementary material). Using the ChIP-on-chip protocol and a highly conservative method for identifying regions of over-representation from the raw data, Kim et al. derived a genome-wide map of the most likely binding sites. All CpG islands on chromosome 21 and 22 that overlap by at least 1 bp with one of these binding sites were combined into the positive training set. The negative training set was constructed from those CpG islands on chromosome 21 and 22 that are at least 500 bp away from the nearest binding site. The resulting training dataset was then processed with EpiGRAPH in order to derive predicted promoter activity scores for all CpG islands according to TJU, GGF, and GGM.

The open chromatin score is based on training data from several large-scale analyses. (1) Using the ChIP-on-chip protocol, Bernstein et al. (Bernstein et al. 2005) derived histone modification data for the HepG2 cell line, including H3K4 di- and trimethylation and H3K9/14 acetylation (dataset obtained from http://www.broad.mit.edu/cell/chromatin_study). Their analysis comprised the non-repetitive parts of chromosomes 21 and 22, for which they calculated sites of significant over-representation. (2) Using DNase I digestion and high-throughput tag sequencing, Crawford et al. (Crawford et al. 2006) derived a genome-wide profile of DNase I hypersensitive sites in CD4+ T cells (dataset obtained from the UCSC Genome Browser). (3) Using the ChIP-on-chip protocol, Cawley et al. (Cawley et al. 2004) derived binding data for the ubiquitous transcription factor SP1 in the Jurkat cell line (dataset obtained from http://transcriptome.affymetrix.com/publication/tfbs). Their data comprise the non-repetitive parts of chromosomes 21 and 22, for which they calculated sites of significant over-representation. For each of the five epigenetic modifications, respectively, we constructed a training dataset as follows. All CpG islands on chromosome 21 and 22 that overlap with the most significant sites for the respective epigenetic modification (as reported by the original authors) were included in the positive training set, and all CpG islands on chromosome 21 and 22 that were at least 500 bp away from the nearest site were included in the negative training set. All five resulting training datasets were then processed with EpiGRAPH, and the five predictions for each CpG island were averaged, in order to derive open chromatin scores for all CpG islands according to TJU, GGF, and GGM.

The combined epigenetic prediction score is calculated for each CpG island as the (unweighted) average of its predicted unmethylated score, its predicted promoter activity score, and its open chromatin score. Since all three components can assume values from zero to one, the same is true for their average.

Evaluation on experimental datasets of DNA methylation and promoter activity

For evaluation based on DNA methylation data, we used a dataset by Rollins et al. (Rollins et al. 2006), who identified 3,073 unmethylated and 2,565 methylated domains in human brain tissue (dataset obtained from http://epigenomics.cu-genome.org/html/meth_landscape). Their data are based on paired-end sequencing from two DNA libraries that were constructed by digestion with methylation-sensitive restriction enzymes, such that one library is highly enriched with unmethylated regions while the other contains almost exclusively methylated regions. We regarded a CpG island as unmethylated if it overlapped by at least 25% with an unmethylated domain and as methylated if it overlapped by at least 25% with a methylated domain. No cases were observed in which a single CpG island overlapped with an unmethylated and a methylated domain simultaneously.

For evaluation based on promoter activity, we used a dataset from the FANTOM3 consortium (Carninci et al. 2006), who performed large-scale CAGE analysis (i.e. tag sequencing of 5’ ends of full-length mRNA) on cDNA libraries derived from a wide range of tissues and cell types (dataset obtained from http://gerg01.gsc.riken.jp/cage_analysis/export/hg17prmtr). All CpG islands that contained at least three tags (i.e. experimental evidences of independent transcription initiation events) were regarded as CpG islands with promoter activity, while all other cases were regarded as CpG islands that show either no or only spurious promoter activity.

ROC curves were constructed in the usual way (Fawcett 2004), using the ROCR library (Sing et al. 2005) and the R statistical package (http://www.r-project.org). The diagrams illustrating the comparison of the different repeat-exclusion strategies were constructed similarly, with some customizing to ensure that every unmethylated domain is counted only once for the true positive rate, even if it overlaps with several CpG islands simultaneously. All R scripts are available on request.

Co-localization analysis

In order to show that the five components of the open chromatin score exhibit significant overlap with each other and with the three CpG island maps (TJU, GGF, and GGM), we performed a co-localization analysis of these eight datasets on chromosomes 21 and 22. To this end, a custom script was written that counts the number of sites of one type that overlap with a second type, for all pairs of site types (i.e. epigenetically modified regions and CpG islands). From these values, overlap percentages were calculated and plotted as a heatmap (Figure 12A).

However, frequent and long regions are obviously more likely to overlap than rare and short regions. We therefore normalized the observed frequency of overlap by the expected frequency for a uniform distribution, using the following procedure. (1) For each site type, we derived a random control set with similar set size, length distribution, and repeat overlap. Technically, for each record in the corresponding dataset, a random site of identical length was drawn from the entire length of chromosomes 21 and 22. If this random site was within five percentage points of its corresponding record in terms of repeat content it was retained;

otherwise, a new random site was drawn. (2) Pairwise frequencies of overlap between all control regions were counted. (3) Steps 1 and 2 were repeated 20 times, and frequencies of overlap were averaged. (4) The observed frequencies of overlap for the real data were divided by the averaged random overlap frequencies, giving rise to *n*-fold over- and under-representations. Figure 12B reports base-2 log scores of these over-representations (positive sign) and under-representations (negative sign).

B-4.3 Results

Preparation of traditional CpG island maps as the basis for prediction

Our prediction of CpG island strength and mapping of bona fide CpG islands started from traditional CpG island maps, which we derived by means of widely used sequence-based CpG island criteria. This is unlikely to significantly reduce the completeness of our mapping since the original CpG island criteria (Gardiner-Garden and Frommer 1987) are regarded as highly sensitive and there is no evidence that they miss a substantial number of bona fide CpG islands.

The application of traditional CpG island finder algorithms faces the problem of repetitive DNA in the genome. Some evolutionarily recent repeat insertions are CpG-rich (e.g. Alu elements) and could erroneously be identified as CpG islands even though they most likely bear little regulatory function (Takai and Jones 2002). Several methods have been suggested to address this problem, but their efficacy has not been systematically investigated. We therefore applied and compared three widely used calculation methods: (1) repeat exclusion by using strict thresholds for GC content (55%), CpG observed vs. expected ratio (0.65), and CpG island length (500 bp) as suggested by Takai and Jones (Takai and Jones 2002); (2) repeat exclusion by combining the standard Gardiner-Garden thresholds (Gardiner-Garden and Frommer 1987) with subsequent removal of all CpG islands that comprise less than 200 bp of non-repetitive DNA; and (3) repeat exclusion by applying the standard thresholds (Gardiner-Garden and Frommer 1987) to the repeat-masked genome.

Using each of these methods, we derived a genome-wide map of CpG islands. Method 1, which we refer to as TJU (for "Takai Jones unmasked") for the remainder of this chapter, gave rise to 37,531 CpG islands genome-wide. Method 2, which we refer to as GGF (for "Gardiner-Garden filtered"), gave rise to 94,450 CpG islands genome-wide. And method 3, which we refer to as GGM ("Gardiner-Garden masked"), gave rise to 109,600 CpG islands genome-wide. All three maps were processed in parallel through most of this study.

Establishment of training datasets for CpG island strength prediction

Absence of DNA methylation and presence of promoter activity are regarded as characteristic of bona fide CpG islands. Therefore, we hypothesized that computational predictions of DNA methylation and promoter activity might provide suitable scores of CpG island strength and thus indicators for the genome-wide mapping of bona fide CpG islands. In previous work focusing on human lymphocytes, we showed that prediction of CpG island methylation is possible with high accuracy based on the DNA sequence plus additional information such as the DNA helix structure and the distribution of repetitive DNA elements (Bock et al. 2006, cf. chapter B-2 of this thesis). Our finding has recently been independently confirmed for brain tissue (Das et al. 2006; Fang et al. 2006) and is expected to hold for a wide range of cell types and tissues. Computational promoter prediction is a well-studied topic and is also feasible with high accuracy across different cell types and tissues (see Bajic et al. 2004, and references therein).

We therefore prepared training datasets for DNA methylation and promoter activity (calculation step 1 in Figure 11), to be processed with EpiGRAPH (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis). Each training dataset was constructed by identifying pairwise overlaps between the three CpG island maps (Figure 11, orange cylinder) and experimental epigenome datasets on DNA methylation and promoter activity (Figure 11, brown cylinder), giving rise to a set of positives (i.e. regions that exhibit characteristics of bona fide CpG islands) as well as a set of negatives (i.e. regions that do not) for both DNA methylation and promoter activity (Figure 11, grey cylinders between calculation steps 1 and 2). For the prediction of unmethylated vs. methylated CpG islands, training datasets were constructed using DNA methylation data that Yamada et al. established for chromosome 21q (Yamada et al. 2004). Similarly, for the prediction of CpG islands that show evidence of promoter activity vs. those that do not, training datasets were constructed using the genome-wide list of polymerase II preinitiation complex binding sites that Kim et al. established for primary fibroblasts (Kim et al. 2005) (for consistency with additional predictions that we report below, we restricted the latter dataset to chromosomes 21 and 22).

CpG island strength estimated by predicted DNA methylation and promoter activity

Processing the training data for DNA methylation and promoter activity with EpiGRAPH showed that accurate distinction was possible between unmethylated and methylated CpG islands and, similarly, between CpG islands that exhibit evidence of promoter activity (namely polymerase II preinitiation complex binding sites) and those that do not (Table 9, full data available online: Bock et al. 2007, tables S1 and S2). The analysis of most predictive attributes helps to understand how this prediction performance is achieved (data available online: Bock et al. 2007, tables S3 and S4). First, unmethylated CpG islands contain significantly fewer tandem repeats and segmental duplications than their methylated counterparts. Second, polymerase II preinitiation complex-bound CpG islands overlap more frequently with highly conserved regions than do unbound CpG islands. And third, both unmethylated and polymerase II preinitiation complex-bound CpG islands are highly enriched with CpG-rich sequence patterns and regions of low predicted DNA rise (which is an important aspect of DNA helix structure, discussed in Olson et al. 2001). These results support the hypothesis that the prediction score for DNA methylation at CpG islands as well as the prediction score for polymerase II preinitiation complex binding at CpG islands are both suitable indicators of CpG island strength. We denote their genome-wide prediction values derived by the epigenome prediction pipeline as the "predicted unmethylated score" and the "predicted promoter activity score," respectively, and evaluate their predictiveness for CpG island strength below.

CpG Island Map	Overlap Prediction for Unmethylated versus Methylated		Overlap Prediction for Polymerase II PIC Binding	
	Correlation	Accuracy	Correlation	Accuracy
TJU	0.661	85.3%	0.416	74.0%
GGF	0.573	81.7%	0.665	84.2%
GGM	0.561	81.1%	0.608	81.7%

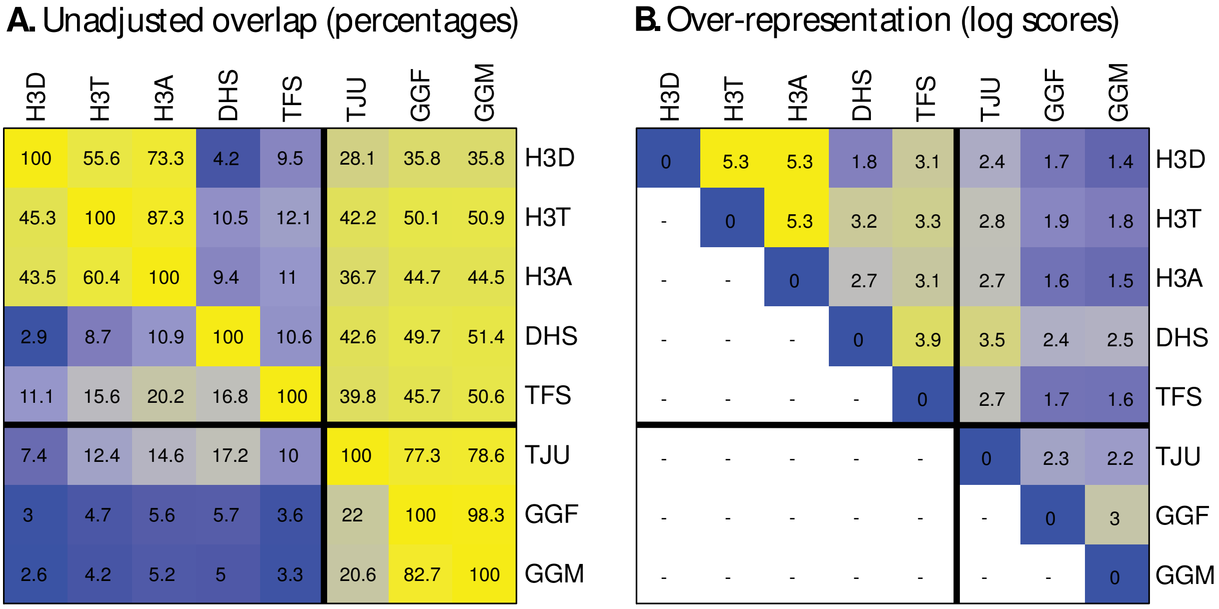
Table 7. Prediction performance for DNA methylation and promoter activity at CpG islands

This table shows the performance that EpiGRAPH achieves for the distinction between CpG islands that overlap with unmethylated regions and those that overlap with methylated regions (left), and similarly for the distinction between CpG islands that overlap with experimentally determined sites of polymerase II preinitiation complex (PIC) binding and those that do not (right). All values are calculated over a 10-fold cross-validation that was repeated ten times with random partitioning.

CpG island strength estimated by predicted epigenetic state and chromatin structure

CpG island scores that focus exclusively on the absence of DNA methylation or on evidence of promoter activity may be insufficient for capturing all aspects of the complex epigenetic and functional states that characterizes bona fide CpG islands. To construct a more comprehensive epigenetic scoring of CpG island strength, we collected five additional large-scale epigenome datasets from the literature, each one describing a different aspect of an open and transcriptionally competent chromatin structure: histone H3K4 di- and trimethylation (Bernstein et al. 2005), histone H3K9/14 acetylation (Bernstein et al. 2005), DNase I hypersensitivity (Crawford et al. 2006) and SP1 transcription factor binding (Cawley et al. 2004). All these datasets cover the non-repetitive parts of human chromosomes 21 and 22, to which we confine our analysis.

A genomic co-localization analysis that we performed for these five datasets showed that epigenetically modified regions indeed exhibit significant overlap with all three CpG island maps (Figure 12). Briefly, this analysis involved two steps. First, the absolute number of pairwise overlaps along chromosomes 21 and 22 was counted for each pairwise combination of epigenetic modification and CpG island definition (Figure 12A). Second, these numbers were normalized by the expected frequency of overlap under the assumption of CpG islands and epigenetically modified regions being uniformly distributed (Figure 12B), in order to correct for length and frequency differences (see Methods section for details).



A. Unadjusted overlap (percentages)

	H3D	H3T	H3A	DHS	TFS	TJU	GGF	GGM
H3D	100	55.6	73.3	4.2	9.5	28.1	35.8	35.8
H3T	45.3	100	87.3	10.5	12.1	42.2	50.1	50.9
H3A	43.5	60.4	100	9.4	11	36.7	44.7	44.5
DHS	2.9	8.7	10.9	100	10.6	42.6	49.7	51.4
TFS	11.1	15.6	20.2	16.8	100	39.8	45.7	50.6
TJU	7.4	12.4	14.6	17.2	10	100	77.3	78.6
GGF	3	4.7	5.6	5.7	3.6	22	100	98.3
GGM	2.6	4.2	5.2	5	3.3	20.6	82.7	100

B. Over-representation (log scores)

	H3D	H3T	H3A	DHS	TFS	TJU	GGF	GGM
H3D	0	5.3	5.3	1.8	3.1	2.4	1.7	1.4
H3T	-	0	5.3	3.2	3.3	2.8	1.9	1.8
H3A	-	-	0	2.7	3.1	2.7	1.6	1.5
DHS	-	-	-	0	3.9	3.5	2.4	2.5
TFS	-	-	-	-	0	2.7	1.7	1.6
TJU	-	-	-	-	-	0	2.3	2.2
GGF	-	-	-	-	-	-	0	3
GGM	-	-	-	-	-	-	-	0

Figure 12. Co-localization between the five components of the open chromatin score and the three CpG Island maps

Panel A displays the relative frequencies of overlap between epigenetically modified sites and CpG islands (percentage values). Panel B displays the degree of over-representation relative to a simulated case in which sites are uniformly distributed over the chromosomes (base-2 log scores). Yellow boxes correspond to frequent overlap, blue boxes to rare overlap. Abbreviations are as follows: H3D, histone H3K4 dimethylation; H3T, histone H3K4 trimethylation; H3A, histone H3K9/14 acetylation; DHS, DNase I hypersensitive sites; TFS, SP1 transcription factor binding, plus the CpG island abbreviations used throughout this study (TJU, GGF, and GGM). The diagram in Panel B is symmetrical as the result of averaging, therefore only the upper right triangular matrix is reported. (A) is not symmetrical, as is obvious from an example: 51.4% of all 578 known DNase I hypersensitive sites on chromosomes 21 and 22 overlap with a GGM CpG island, while only 5.0% of all 5,913 GGM CpG islands overlap with an experimentally determined DNase I hypersensitive site.

Intriguingly, for all datasets the enrichment observed in the genomic co-localization analysis is highly skewed toward the same specific set of CpG islands, which frequently overlap with several epigenetic modifications simultaneously (Table 8). For example, CpG islands that show evidence of two out of five epigenetic modifications simultaneously are observed 10-fold to 20-fold more frequently than expected under a uniform distribution. We therefore concluded that all five epigenetic modifications do in fact capture different epigenetic indicators of a single concept, namely, whether or not a particular CpG island fosters an open and transcriptionally competent chromatin structure.

CpG Island Map	**Observed/Expected Frequency of Overlap with *n* out of Five Epigenetic Modifications**					
	***n* = 0**	***n* = 1**	***n* = 2**	***n* = 3**	***n* = 4**	***n* = 5**
TJU	949/1,238.5 = 0.8	180/113.3 = 1.6	99/6.2 = 16.0	71/0.2 = 355.0	50/0.0	9/0.0
GGF	4,290/4,545.8 = 0.9	284/301.6 = 0.9	117/10.6 = 11.0	97/0.0	63/0.0	7/0.0
GGM	5,260/5,549.7 = 0.9	345/351.9 = 1.0	130/11.2 = 11.6	115/0.3 = 383.3	56/0.0	7/0.0

Table 8. A subset of CpG islands exhibit highly significant overlap with multiple epigenetic modifications simultaneously

This table contrasts the observed and the expected frequencies with which CpG islands overlap with a certain number (zero to five) of the five epigenetic modifications that contribute to the open chromatin score (i.e. histone H3K4 di- and trimethylation, histone H3K9/14 acetylation, DNase I hypersensitivity, and SP1 binding). The format of the table entries is as follows: observed frequency/expected frequency = over-representation ratio. Expected frequencies were calculated by simulation under the assumption of uniform distribution. Overlap with four or more epigenetic modifications was too rare to occur in these simulations. Hence, no degrees of over-representation were calculated for the two rightmost columns.

In order to convert this observation into a method for scoring CpG island strength, we prepared training datasets and applied EpiGRAPH separately for each of the five epigenetic modifications (calculation steps 1 and 2 in Figure 11). In all cases, a linear support vector machine was able to discriminate with significant accuracy between CpG islands that overlap with the particular epigenetic modification and those that do not (Table 11, full data available online: Bock et al. 2007, Table S5). Analysis of the most predictive attributes showed that the former are more likely to contain CpG-rich patterns, are more conserved, and exhibit a characteristic predicted helix structure (a comprehensive list of significant differences is available online: Bock et al. 2007, Table S6). Furthermore, we observed high correlations between the prediction scores for all five epigenetic modifications (data available online: Bock et al. 2007, Table S7), which provided additional support for the conclusion that they represent aspects of a single concept. Therefore, for each CpG island we calculated the average over all five predictions and thereby derived a single "open chromatin score" (calculation step 3 in Figure 11). Finally, since the predicted unmethylated score, the predicted promoter activity score, and the open chromatin score can be assumed to capture complementary aspects of a CpG island's epigenetic and functional state, we combined these three scores into an additional consensus score that we call the "combined epigenetic score" of CpG island strength.

CpG Island Map	**Overlap Prediction for**									
	Histone H3K4me2		**Histone H3K4me3**		**Histone H3K9ac/H3K14ac**		**DNase I Hypersensitivity**		**SP1 Binding**	
	Correlation	**Accuracy**	**Correlation**	**Accuracy**	**Correlation**	**Accuracy**	**Correlation**	**Accuracy**	**Correlation**	**Accuracy**
TJU	0.294	67.5%	0.302	68.5%	0.380	72.5%	0.368	71.4%	0.374	72.2%
GGF	0.308	68.2%	0.431	73.8%	0.398	72.6%	0.548	79.3%	0.433	73.5%
GGM	0.324	68.8%	0.397	72.7%	0.407	73.3%	0.54	79.1%	0.417	73.5%

Table 9. Prediction performance for the distinction between CpG islands that overlap with a particular epigenetic modification and those that do not

For each component of the open chromatin score, this table shows the performance that EpiGRAPH achieves for the distinction between CpG islands that overlap with that particular epigenetic modification and those that do not. All values are calculated over a tenfold cross-validation that was repeated ten times with random partitioning. Abbreviations are as follows: H3K4me2, H3K4 dimethylation; H3K4me3, H3K4 trimethylation; H3K9ac/H3K14ac, H3K9/14 acetylation.

Independent evaluation of CpG island strength predictions

For each of the predictions described above, the performance was assessed by means of cross-validation. While this procedure can provide an accurate estimate of the prediction performance expected on new data of the same type, it is not sufficient for establishing the prediction scores as a quantitative indicator of CpG island strength. First, all training and testing was restricted to chromosomes 21 and 22, therefore it could not be assessed how well the predictions generalize to the entire genome. Second, cross-validation on a single dataset cannot exclude the risk of overfitting to the special properties of this particular dataset, which can include both biological factors (such as tissue-specific and cell-type-specific effects) and technical problems (such as experimental bias toward specific genome regions).

Therefore, we performed an additional evaluation, based on two large-scale datasets (Figure 11, blue cylinder): (1) a random sample of unmethylated and methylated regions in the human genome derived from brain tissue by means of large-scale tag sequencing of DNA fragments generated by methylation-sensitive restriction enzymes (Rollins et al. 2006), and (2) a genome-wide map of experimentally determined transcription start sites obtained for a wide range of tissues by the FANTOM3 project (Carninci et al. 2006). Independent evaluation (without retraining) on these datasets can overcome both limitations of the previously described cross-validations. First, the two datasets cover the entire (non-repetitive) human genome, not only two chromosomes like the training data. Second, both datasets deviate significantly in terms of tissue type, cell type, and experimental protocol from all training datasets used in this study. Hence, any significant prediction performance that the CpG island scores achieve on these evaluation datasets can be attributed to inherent and robust properties of the CpG islands themselves.

The first evaluation dataset was constructed by identifying overlap between CpG islands and regions of known methylation state, giving rise to experimentally positive CpG islands (i.e. overlapping with unmethylated regions) and experimentally negative CpG islands (i.e. overlapping with methylated regions). The second evaluation dataset was constructed by identifying overlap between CpG islands and experimentally determined transcription start sites. CpG islands that harbor at least three independent transcription initiation events were included in the set of positives, while all remaining CpG islands were included in the set of negatives.

All four CpG island scores were then evaluated against these two evaluation datasets using receiver operating characteristic (ROC) curves, which is the standard method for benchmarking classifiers in machine learning (Fawcett 2004) (calculation step 4 in Figure 11). These ROC curves interpret the score of any one CpG island as its predicted likelihood of being a bona fide CpG island. For all possible thresholds on the CpG island score, they describe the trade-off between the true positive rate (i.e. the percentage of bona fide CpG islands that are detected, also called sensitivity) and the false positive rate (i.e. the percentage of negatives that are erroneously classified as bona fide CpG islands, which is equal to one minus specificity) and thereby assess how well the particular CpG island score predicts the evaluation datasets. A purely random score would on average result in a ROC curve that is a straight line from (0,0) to (1,1); the closer the curve bends toward the top left corner, the better is the performance of the CpG island score.

The ROC curves show that all four CpG island scores that we constructed (i.e. the predicted unmethylated score, the predicted promoter activity score, the open chromatin score, and the combined epigenetic score) perform significantly better than random (Figure 13) and can therefore be used to improve the accuracy of CpG island mapping. Nevertheless, we observe several differences. On both evaluation datasets, the predicted unmethylated score performs worst of all four scores. This contrasts with the high accuracy of the methylation prediction itself (Table 7) and points to high divergence between the training dataset and the evaluation datasets, possibly arising from tissue specificity of DNA methylation as well as from experimental biases. The predicted promoter activity score performs well for both evaluation datasets, which is also the case for the open chromatin score. Finally, the combined epigenetic score, i.e. the consensus prediction of all three individual CpG island scores, performs better than each individual score. This result shows that the three individual scores – each derived from data for different cell types and for different aspects of CpG island strength – do provide complementary information that can be combined to increase prediction performance.

For comparison, we also plotted the performance of the GC content, the CpG observed vs. expected ratio, and the length of CpG islands, interpreting them as indicators of CpG island strength (Figure 13), and we observed a surprising result. On the one hand, GC content performs only slightly better than random, and the CpG observed vs. expected ratio – arguably the most natural sequence-based indicator of CpG island strength – performs substantially worse than the promoter activity score, the open chromatin score, and the combined epigenetic score. On the other hand, CpG island length (which one might have dismissed as a rather technical aspect of the sequence-based CpG island definition, designed to exclude short and insignificant CpG islands) turns out to perform very well, second only to the combined epigenetic score in terms of overall prediction performance (i.e. area under the ROC curve (Fawcett 2004), averaged over Figure 13A to F). Although this finding contributes little to the main impetus of this paper, which is to reconcile CpG island mapping with the epigenetic and functional concept of bona fide CpG islands, it can help design a simple heuristic to approximate the combined epigenetic score. We discuss this point in more detail in a separate section below.

In addition to the analysis by ROC curves, we performed a second evaluation, in order to assess whether the combined epigenetic score predicts not only the likelihood that a particular CpG island exhibits promoter activity (as shown by the ROC curves), but also the strength of its promoter activity. To that end, we plotted the number of transcription start site tags (as an indicator of promoter strength) for all CpG islands that harbor experimentally determined transcription start sites at all against the combined epigenetic score (Figure 14). The results show that promoter CpG islands with a high combined epigenetic score indeed exhibit substantially stronger promoter activity than promoter CpG islands with a low combined epigenetic score.

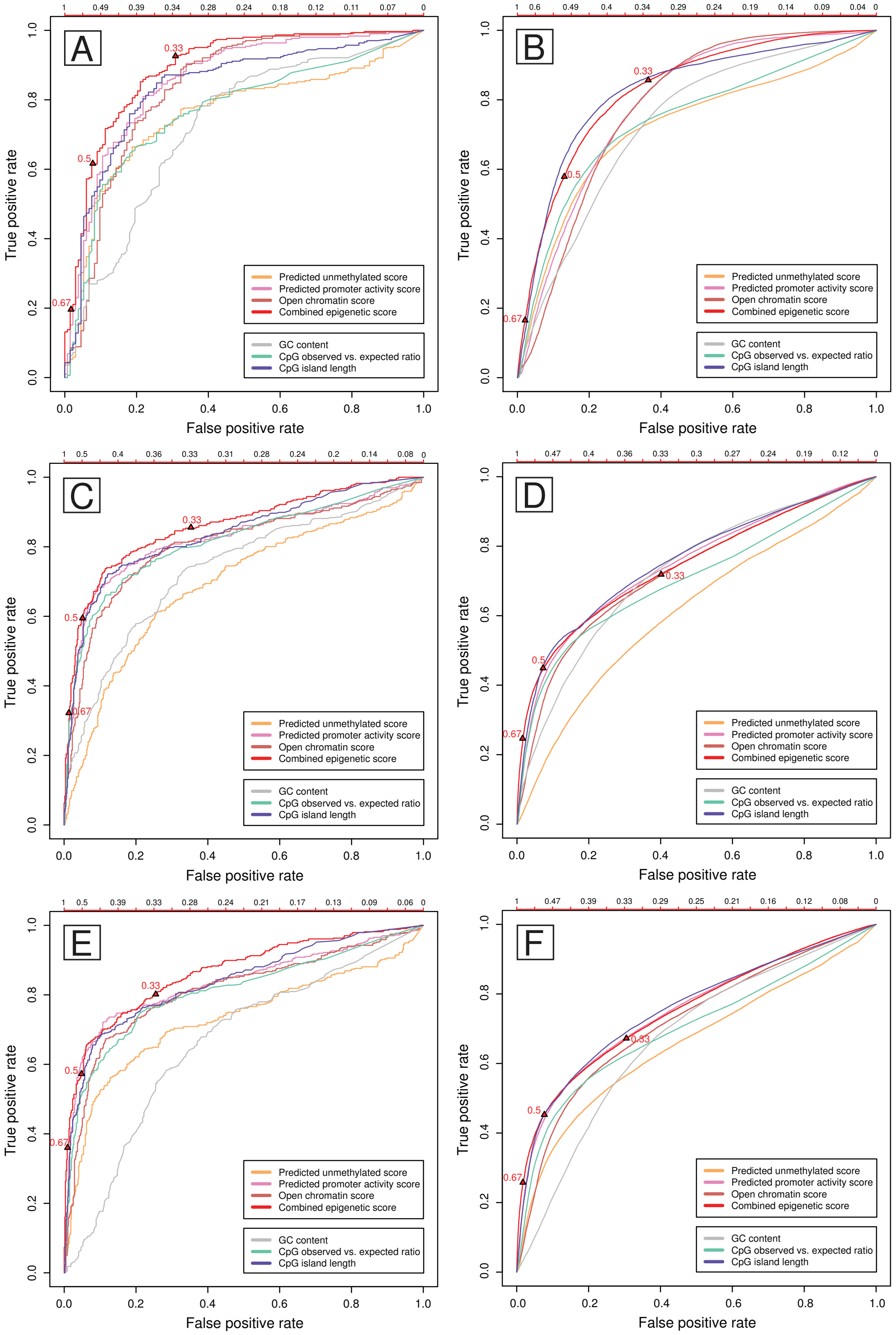


Figure 13. ROC curves comparing the performance of four prediction scores and three sequence criteria against DNA methylation and promoter activity

This figure compares the prediction performance of four CpG island scores that are based on epigenome prediction (upper legend box) and of three simple sequence criteria (lower legend box). In panels A, C and E, overlap with unmethylated regions is used for evaluation, and in panels B, D and F, overlap with experimentally determined transcription start sites (as an indicator of promoter activity) is used instead. All graphs plot the true positive rate against the false positive rate in the form of ROC curves (Fawcett 2004). The scales on top of the plots display the threshold values for the combined epigenetic score that correspond to the trade-off between false positive rate and true positive rate at any one position. The epigenetically motivated thresholds for the combined epigenetic score are highlighted by triangles: 0.5 (balance between sensitivity and specificity), 0.33 (high sensitivity), and 0.67 (high specificity). Averaged across all six graphs, the ROC area-under-curve performance measure (i.e. the percentage of the unit square that lies below the ROC curve) amounts to the following values: predicted unmethylated score, 65.4%; predicted promoter activity score, 74.8%; open chromatin score, 72.2%; combined epigenetic score, 75.8%, GC content, 67.1%; CpG observed vs. expected score, 70.6%; and CpG island length, 75.5%.

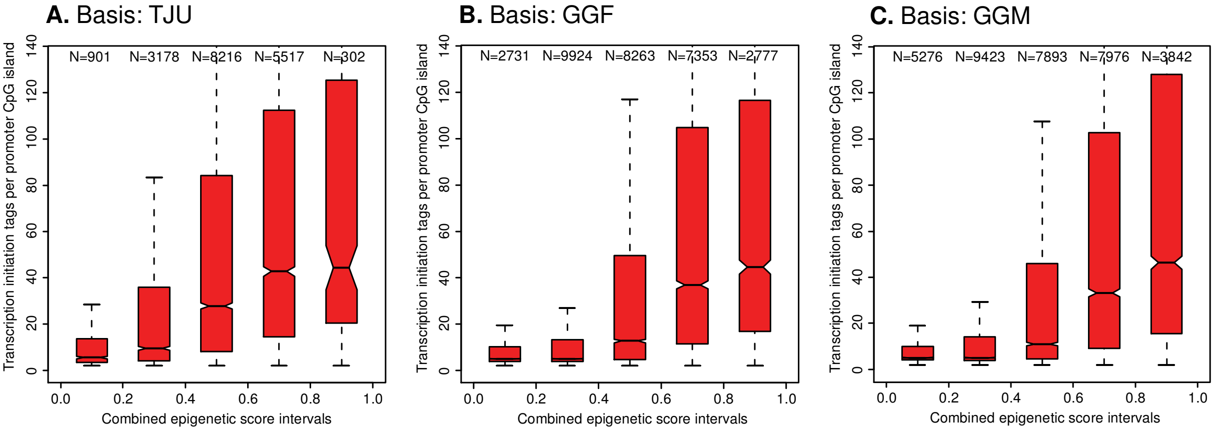


Figure 14. Boxplots comparing the promoter strength between high-scoring and low-scoring promoter CpG islands

This figure shows boxplots of the average number of transcription start site tags per CpG island (as an indicator of promoter strength), restricted to those CpG islands that show experimental evidence of promoter activity at all (i.e. at least three transcription start site tags fall within the CpG island). Separate boxplots are drawn for CpG islands that fall into different intervals in terms of their combined epigenetic score (i.e. 0 to 0.2, 0.2 to 0.4, etc.). The standard boxplot format is used (boxes show center quartiles, whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box, and non-overlapping notches provide evidence of significantly different medians), and outliers are hidden.

Selection of the most appropriate CpG island map as the basis for prediction

Up to this point, we carried out all analyses in parallel for the three CpG island maps that we derived using different repeat-exclusion strategies (TJU, GGF, and GGM). In order to select the most appropriate setup for the final map of predicted bona fide CpG islands, we benchmarked these strategies on both evaluation datasets. Since ROC curves cannot easily account for the different number of CpG islands in each of the three maps, we constructed an alternative type of diagram for this purpose (Figure 15). This diagram plots the precision of the classification (i.e. the percentage of predicted bona fide CpG islands that are supported either by the DNA methylation dataset or by the transcription start site dataset) and the true positive rate (i.e. the percentage of unmethylated CpG islands or CpG islands harboring transcription start sites that are correctly predicted, respectively) against the total number of CpG islands that are selected for any particular threshold.

The results show that there is generally high agreement between the performance of the combined epigenetic score on each of the three CpG island maps (Figure 15), apart from the trivial fact that the overall sizes of the three maps differ. Nevertheless, the combined epigenetic score performs slightly better on the GGM map (i.e. repeat exclusion using RepeatMasker, with subsequent application of the Gardiner-Garden criteria for CpG island detection) than on the two alternative maps, and this setup was therefore chosen. The GGM map has two additional advantages. First, in contrast to the GGF map, it does not require choosing a cutoff for the maximum repeat content that is permitted per CpG island. Second, in contrast to the TJU map, the DNA sequence parameters used to derive the GGM map are so permissive that virtually every non-repetitive, CpG-rich region that exceeds 200 bp is selected and scored. Thus, scores are also calculated for regions that show little potential to be bona fide CpG islands but which may be of interest for comprehensive scans of particular genomic regions.

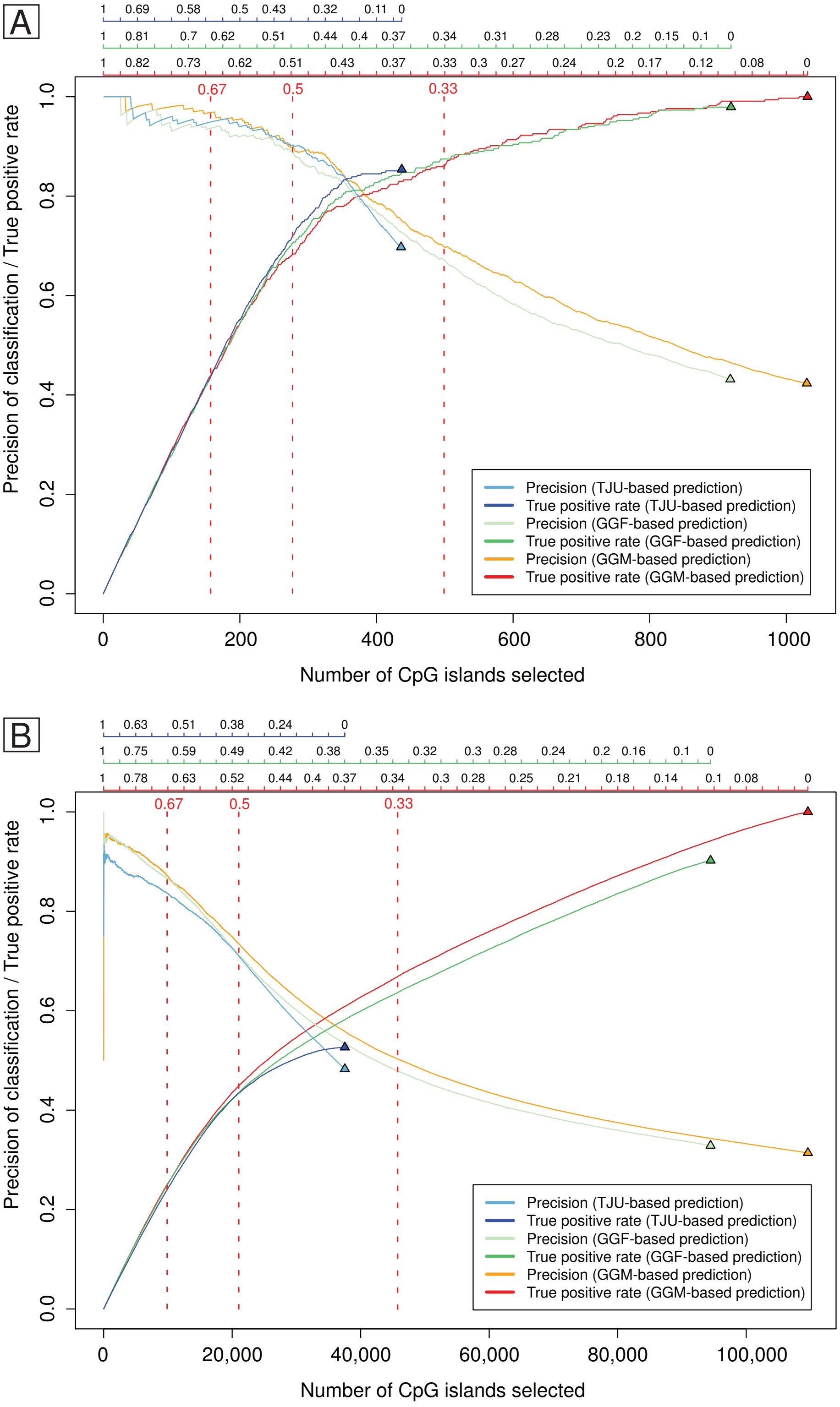


Figure 15. Performance of the combined epigenetic score compared between CpG island maps that use different repeat-exclusion strategies

This figure plots the precision (i.e. the percentage of experimentally supported bona fide CpG islands among all selected CpG islands) and the true positive rate (i.e. the percentage of experimentally supported bona fide CpG islands that are selected) over the total number of cases predicted as bona fide CpG islands, for any valid threshold on the combined epigenetic score. Evaluation criteria are absence of DNA methylation (panel A) and presence of promoter activity as indicated by experimentally determined transcription start sites (panel B). The three scales on top of each plot display the score thresholds that correspond to the number of CpG islands selected. Dashed lines show the three thresholds that were used to derive the final bona fide CpG island maps on the basis of the GGM dataset. Numbers on the *x*-axis are significantly lower for the diagram in panel A than in panel B because of the fact that the DNA methylation dataset covers only a random sample of unmethylated and methylated CpG islands, while the promoter activity dataset covers essentially all non-repetitive CpG islands genome-wide.

At http://rd.plos.org/10.1371_journal.pcbi.0030110_01, we report the combined epigenetic score for all CpG islands that fulfill the Gardiner-Garden criteria on the repeat-masked genome (GGM). Since our evaluations showed that the combined epigenetic score provides an accurate and robust estimate of CpG island strength (i.e. of a CpG island's inherent tendency to exhibit an open and transcriptionally competent chromatin structure), these scores can be directly used for a number of applications. For example, they add important quantitative information to support functional genome annotation as well as the interpretation of experimental epigenome data, and they can be used to prioritize candidate regions, e.g. when selecting a fixed number of most promising regulatory CpG islands for experimental follow-up.

Mapping of predicted bona fide CpG islands using the combined epigenetic score

Although our analysis emphasizes the importance of quantitative information on CpG island strength, in order to distinguish gradually between bona fide CpG islands and those CpG-rich regions that show no evidence of a regulatory role (Figure 13 and Figure 14), we acknowledge that certain applications would benefit from a fixed threshold on the combined epigenetic score. For example, in order to derive a genome-wide list of predicted bona fide CpG islands or for selecting regions to be spotted on a CpG island microarray, it is necessary to make a trade-off between thresholds that are low enough to achieve high sensitivity (i.e. most bona fide CpG islands are included) and high enough to maintain high specificity (i.e. few CpG-rich regions that show no evidence of a regulatory role are selected).

Fortunately, the way the combined epigenetic score is defined immediately suggests a threshold that balances sensitivity and specificity and carries a biologically meaningful interpretation. Because the combined epigenetic score is the average of the confidence values (or predicted likelihoods) with which a particular CpG island is classified (i) as unmethylated, (ii) as exhibiting promoter activity, and (iii) as fostering open chromatin structure, it can itself be interpreted as a likelihood value. It assigns a score between zero and one to each CpG island that reflects both the likelihood (Figure 13) and the strength (Figure 14) with which the CpG island exhibits the open and transcriptionally competent chromatin state that is characteristic of bona fide CpG islands. A value of zero thus corresponds to a completely silenced, inactive, and inaccessibly buried CpG island, while a value of one corresponds to an unmethylated, highly accessible CpG island with strong promoter activity. Between these two extremes, a value of 0.5 corresponds to CpG islands that are equally likely to be bona fide CpG islands or not. This value therefore provides a suitable threshold for CpG island mapping, as it balances sensitivity and specificity. We would recommend this threshold for most applications.

Nevertheless, certain tasks (e.g. genome annotation) may require increased sensitivity in order to annotate as many bona fide CpG islands as possible and would therefore profit from a less stringent threshold, such as 0.33. Conversely, a highly conservative threshold of 0.67 is useful when selecting candidate regulatory regions for experimental follow-up, in order to minimize the risk of wasting resources on false positives. To support decision-making about the most appropriate map to use for a particular application, Table 10A (column "combined epigenetic score") provides quantitative data on true positive rates and false positive rates calculated for both evaluation criteria, DNA methylation and promoter activity.

Using the GGM map as the basis (109,600 CpG islands for the entire human genome) and the combined epigenetic score as the indicator of CpG island strength, we calculated maps of predicted bona fide CpG islands. Using the balanced 0.5 threshold, 21,631 genomic regions are predicted as bona fide CpG islands (19.7%); for the highly sensitive 0.33 threshold, this value is 46,182 (42.1%); and for the highly specific 0.67 threshold, we predict 10,281 bona fide CpG islands genome-wide (9.4%).

All CpG island maps are available for download and as UCSC Genome Browser tracks (http://rd.plos.org/10.1371_journal.pcbi.0030110_01). A summary of their genomic distribution is provided online (Bock et al. 2007, table S8). Furthermore, we assessed how frequently bona fide CpG islands associate with genes, exons, annotated transcription start sites, and highly conserved regions (data available online: Bock et al. 2007, table S9). As expected, predicted bona fide CpG islands are highly associated with annotated transcription start sites and evolutionarily conserved regions, and this effect is stronger for the specific threshold than for the balanced and the sensitive thresholds. However, even of the 10,281 strongest CpG islands in the human genome, i.e. those whose scores exceed the highly specific 0.67 threshold, more than 40% do not overlap with an Ensembl-annotated transcription start site. Thus, we conclude that our prediction of CpG island strength identifies a significant number of regions with open and transcriptionally competent chromatin structure that are not known promoters of protein-coding genes.

Evaluation of CpG island length as a heuristic for the combined epigenetic score

As outlined above, the combined epigenetic score has a conceptual advantage over more conventional ways of predicting CpG island strength because it directly links CpG island maps to the epigenetic and functional role that CpG islands are assumed to play in the human genome. However, it bears one significant disadvantage: the calculation of the combined epigenetic score is complex and computationally demanding. While we alleviate this issue by providing pre-calculated maps for the current assemblies of the human genome, it would be helpful to have a second estimate of CpG island strength available that is significantly simpler to calculate, even at the cost of a somewhat reduced performance. As suggested above and supported by Figure 13, CpG island length can be used in this way. It is substantially, though not perfectly, correlated with the combined epigenetic score (Pearson's $r = 0.59$), and it gives rise to a ROC area-under-curve (Fawcett 2004) performance that is not dramatically lower than that of the combined epigenetic score (Figure 13).

However, it is unclear what might be suitable thresholds in order to map bona fide CpG islands on the basis of their length, since – in contrast to the combined epigenetic score – CpG island length does not reflect any specific epigenetic concept. We propose that the most appropriate solution is to select thresholds such that the resulting maps resemble those calculated from the combined epigenetic score in terms of the false positive rate. That is, the length heuristic should not make more errors when detecting bona fide CpG islands than the combined epigenetic score, but it may well detect fewer (worse) or more (better) bona fide CpG islands, as measured by the true positive rate. Table 10A provides a performance comparison of bona fide CpG island maps derived from the combined epigenetic score vs. maps derived using the CpG island length heuristic, with thresholds selected such that the false positive rate is as close as possible to that of the maps derived from the combined epigenetic score (Table 10A). Taking the results for both evaluation datasets into account and rounding to the closest hundred, we concluded that a minimum length of 700 bp is the most appropriate threshold for the balanced case. For sensitive mapping the most appropriate minimum length is 300 bp, and for the specific mapping the most appropriate minimum length is 1,400 bp. Direct performance comparison with the maps derived from the combined epigenetic score (Table 10B) shows that this length-based heuristic performs equally well for sensitive mapping (slightly worse for DNA methylation, slightly better for promoter activity), but falls short for both the balanced and the specific maps. Differences are particularly strong for the specific case, in which the map based on the combined epigenetic score predicts 65% (DNA methylation: true positive rate of 36.2% vs. 22.0%) and 56% (promoter activity: 21.7% vs. 13.9%) more bona fide CpG islands than the heuristic when false positive rates are fixed to 1.2% for both maps.

Evaluation Dataset	Type of Mapping	CpG Island Scoring Method	Comparison 1			Comparison 2		
			Threshold	False Positive Rate	True Positive Rate	Threshold	False Positive Rate	True Positive Rate
DNA methylation	Sensitive	Combined epigenetic score	**0.33**	25.5%	**80.0%**	0.31	28.1%	**82.1%**
DNA methylation	Sensitive	CpG island length	315 bp	25.4%	77.1%	**300 bp**	28.1%	78.0%
DNA methylation	Balanced	Combined epigenetic score	**0.5**	5.2%	**57.3%**	0.48	5.7%	**61.7%**
DNA methylation	Balanced	CpG island length	759 bp	5.2%	53.4%	**700 bp**	5.7%	56.4%
DNA methylation	Specific	Combined epigenetic score	**0.67**	0.8%	**36.0%**	0.67	1.2%	**36.2%**
DNA methylation	Specific	CpG island length	1,496 bp	0.8%	17.2%	**1,400 bp**	1.2%	22.0%
Promoter activity	Sensitive	Combined epigenetic score	**0.33**	30.8%	67.3%	0.33	30.7%	67.2%
Promoter activity	Sensitive	CpG island length	300 bp	30.7%	**69.2%**	**300 bp**	30.7%	**69.2%**
Promoter activity	Balanced	Combined epigenetic score	**0.5**	7.9%	**45.7%**	0.52	6.5%	**43.1%**
Promoter activity	Balanced	CpG island length	624 bp	7.9%	45.6%	**700 bp**	6.5%	42.7%
Promoter activity	Specific	Combined epigenetic score	**0.67**	1.8%	**25.9%**	0.71	1.2%	**21.7%**
Promoter activity	Specific	CpG island length	1,225 bp	1.8%	19.3%	**1,400 bp**	1.2%	13.9%

Table 10. Performance comparison between the combined epigenetic score and the CpG island length

This table compares the performance of bona fide CpG island mapping using the combined epigenetic score with a simple length-based mapping heuristic. Comparison 1 indicates the performance of the three standard thresholds of the combined epigenetic score (sensitive, 0.33; balanced, 0.5; and specific, 0.67), as well as the performance of corresponding maps derived using the highest CpG island length thresholds that lead to lesser or equal false positive rates. Comparison 2 is a similar comparison, in which the CpG island length thresholds are fixed (sensitive, 300 bp; balanced, 700 bp; and specific, 1,400 bp), while the thresholds for the combined epigenetic score are selected such that the false positive rates of the corresponding maps are less than or equal to the length-based false positive rate. All results are based on the GGM map and are reported separately for the two evaluation criteria, DNA methylation and promoter activity. In the "Threshold" columns the fixed thresholds are in bold; in the "True Positive Rate" columns the higher scores are in bold.

We conclude that the length-based heuristic can be used for a general mapping of bona fide CpG islands, preferably with a minimum length threshold of 300 bp. However, as soon as high specificity is desirable, we strongly recommend using the maps of predicted bona fide CpG islands that are based on the combined epigenetic score. This conclusion is consistent with the observation that exclusively sequence-based CpG island maps achieve high sensitivity but lack specificity, i.e. they include many regions that fail to exhibit the epigenetic and functional characteristics of bona fide CpG islands.

B-4.4 Discussion

The CpG island strength as a theoretical concept captures the inherent tendency of a particular CpG island to exhibit the characteristic epigenetic and functional state of bona fide CpG islands. This includes, but is not limited to, absence of DNA methylation as well as presence and strength of promoter activity. The concept of CpG island strength is abstracted from any tissue-specific or cell-type-specific variation of the epigenetic states, and should be viewed as a description of the default state that is encoded in the DNA sequence of a particular CpG island, and which the CpG island will assume in the absence of any strong influences toward deviation (such as imprinting-related differential methylation or cancer-related epigenetic silencing). Since we observed clear-cut quantitative differences among CpG islands (Figure 14) and a highly significant clustering of epigenetic modifications in a subset of CpG islands (Table 8), we conclude that this concept adds important information to traditional CpG island maps. Furthermore, it provides a straightforward solution for the lack of specificity of these maps.

In order to predict CpG island strength for each CpG island in the human genome, we initially predicted multiple epigenetic modifications independently. These genome-wide predictions were highly correlated, hence we could combine them into a consensus prediction of CpG island strength. The predictive power of this combined epigenetic score (and of several alternative CpG island scores) was evaluated on large-scale experimental datasets of DNA methylation and promoter activity. We also selected and motivated biologically plausible thresholds on the combined epigenetic score, leading to maps of predicted bona fide CpG islands that are more accurate than current sequence-based maps. For example, even the most restrictive definition (Takai and Jones 2002) of CpG islands (TJU) gives rise to a set of CpG islands of which approximately one third are methylated, i.e. CpG-rich regions that fail to exhibit the characteristics of bona fide CpG islands according to our evaluation dataset. Using a sensitive threshold of 0.33 on the combined epigenetic score, this fraction can be reduced by two-thirds, while losing less than 8% of the unmethylated, potentially bona fide CpG islands (Figure 13A). Similar improvements were observed for evaluation of promoter activity and for two additional CpG island maps (GGF and GGM). We therefore conclude that a post-processing step utilizing epigenome prediction significantly increases the accuracy of CpG island mapping and can help overcome the weaknesses of current CpG island definitions. We also showed that a simple length-based mapping heuristic that selects only CpG islands with a minimum length of 300 bp on the repeat-masked genome is suitable for sensitive mapping of bona fide CpG islands but performs substantially worse than the combined epigenetic score when high specificity is desired.

The key concept of our analysis was to move beyond a purely sequence-based definition of CpG islands (which many researchers have been optimizing in the past, cf. Larsen et al. 1992; Li et al. 2002; Luque-Escamilla et al. 2005; Ponger and Mouchiroud 2002; Wang and Leung 2004) and to incorporate epigenome and chromatin data. This approach is consistent with the common notion of CpG islands being functionally and epigenetically exceptional regions, but gave rise to two conceptual difficulties. First, epigenome and chromatin data are tissue-specific and cell-type-specific. It is thus necessary to abstract information from these variations in order to derive a single CpG island map for the human genome (instead of specific maps for all major tissues and cell types). Second, comprehensive epigenome data were available only for chromosomes 21 and 22, not for the entire genome. We addressed both issues by introducing epigenome prediction as the method for scoring CpG island strength, instead of using epigenome data directly.

Potential limitations of this study arise from the epigenome datasets that were employed for training and evaluation. First, two out of the five ChIP-on-chip datasets that we used are based on ligation-mediated PCR amplification (Cawley et al. 2004; Kim et al. 2005), which creates an experimental bias toward GC-rich regions (the other three are based on a more appropriate linear DNA amplification method). Second, the lists of over-represented regions from the ChIP-on-chip studies that we used are most likely overly conservative (Ji and Wong 2005). However, in spite of these shortcomings of the underlying datasets, we observed consistent results across multiple datasets, which were obtained from different cell types, in different labs, and with different experimental protocols. Therefore, such error sources are highly unlikely to invalidate our main results. An additional limitation refers to our ability to exhaustively evaluate the performance of the predictions: because the concepts of CpG island strength and of bona fide CpG islands describe inherent properties of CpG islands, which abstract from their epigenetic state in a particular tissue or cell type, they are difficult to measure experimentally. We therefore performed our evaluations on datasets that significantly deviate in their experimental and biological characteristics from all training data that was used, and we paid as much attention to deriving consistent and biologically plausible predictions of CpG island strength as to achieving the highest performance on the evaluation criteria. Furthermore, for reasons of data availability we focused on epigenetic modifications that are associated with open and transcriptionally competent chromatin. Future extensions of this work should include repressive epigenetic modifications as well, such as histone H3K9 methylation and H3K27 methylation. On this basis, combined with larger datasets, it may be possible to deconstruct the predicted CpG island strength into individual components for all major epigenetic modifications.

B-5 An optimization-based approach to CpG island annotation[1]

While the method for CpG island mapping by epigenome prediction outlined in the previous section resolves several key issues of CpG island annotation, it still relies on traditional CpG island finders for identification of candidate CpG islands. In this chapter, we critically assess potential shortcomings of current CpG island definitions and search algorithms and propose improvements.

Historically, CpG islands were first defined in a rather *ad hoc* way, based on a small set of DNA sequences surrounding the transcription start sites of well-known genes (Gardiner-Garden and Frommer 1987). As the tentativeness of the original definition is often overlooked, it is worth quoting verbatim from the paper of Gardiner-Garden and Frommer (Gardiner-Garden and Frommer 1987): “For the purpose of this survey, regions of DNA with a moving average of %G+C over 50 and Obs/Exp CpG over 0.6 have been classified as CpG-rich regions. CpG-rich regions over 200 bp in length are unlikely to have occurred by chance alone, so, as a working definition, have been labeled as CpG islands”. We will refer to this definition throughout this chapter, but we will use the more common terms “GC content” rather than “%G+C” and “CpG observed vs. expected ratio” rather than “Obs/Exp CpG”.

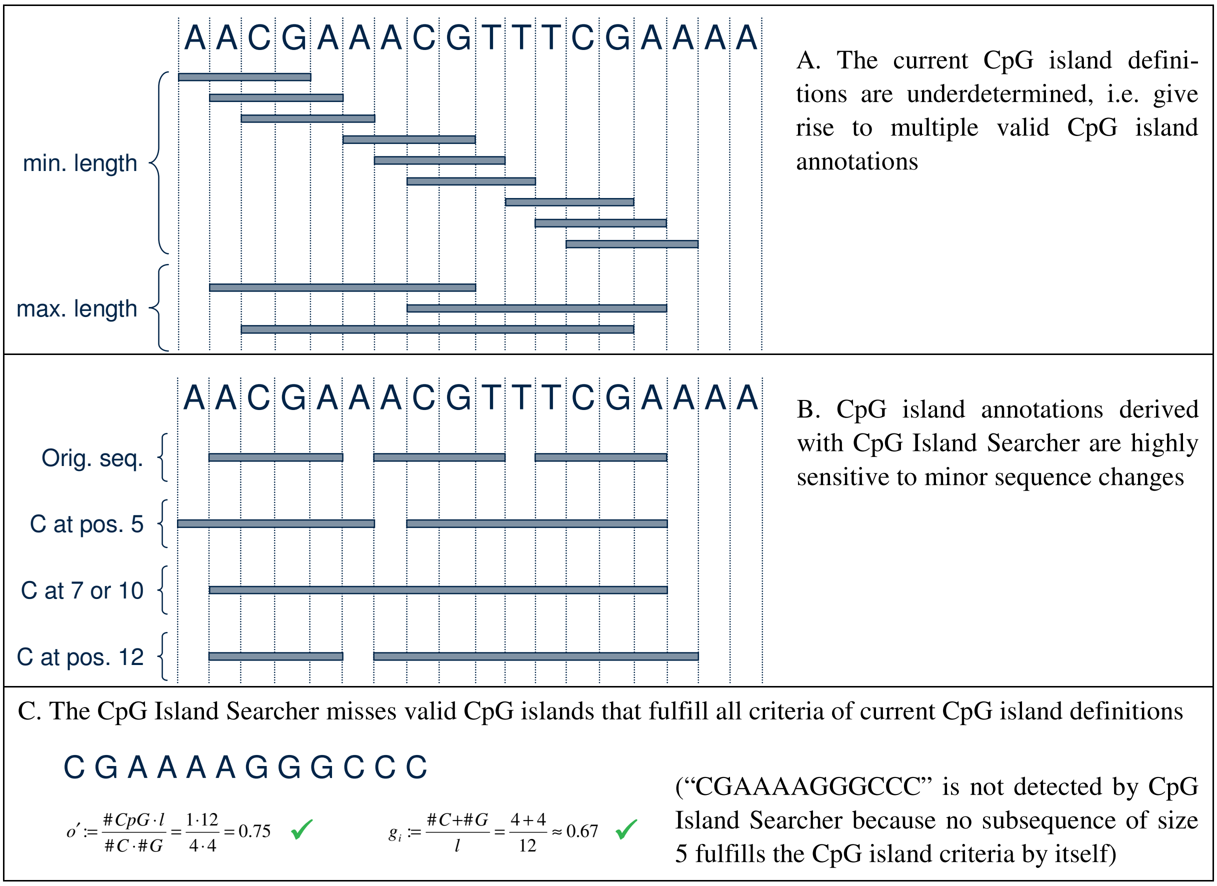
Over the years, multiple modifications and improvements of this definition have been proposed (Aissani and Bernardi 1991; Bock et al. 2007; Hannenhalli and Levy 2001; Ioshikhes and Zhang 2000; Larsen et al. 1992; Li et al. 2002; Luque-Escamilla et al. 2005; Matsuo et al. 1993; Ponger et al. 2001; Ponger and Mouchiroud 2002; Takai and Jones 2002; Wang and Leung 2004). In particular, it was observed that the majority of CpG islands according to the Gardiner-Garden definition lack the regulatory and functional roles considered constitutive of CpG islands, either due to colocalization with repetitive elements (Takai and Jones 2002) or because they exhibit condensed chromatin structure across multiple tissues (Bock et al. 2007, cf. chapter B-4 of this thesis). Furthermore, several attempts have been made to abandon the concept of CpG islands altogether and to replace it by clustering or direct counting of CpG dinucleotides (Glass et al. 2007; Hackenberg et al. 2006; Saxonov et al. 2006). However, given that no convincing case has been made why the latter approaches should be practically or conceptually superior, it is unlikely that the time-tested, flexible and widely used concept of CpG islands is soon to be replaced by a radically new approach.

[1] This chapter describes work conducted in collaboration with Lars Feuerbach, who contributed important concepts and ideas to the algorithm for CpG island annotation and who is currently designing and implementing a speed-optimized version of the proposed algorithm.

In spite of the wide relevance of CpG islands in genome research, currently used definitions and software tools exhibit several shortcomings that hamper reliable annotation of all CpG islands in mammalian genomes. First, current definitions are underdetermined and ambiguous in terms of the CpG island annotation they specify. For a given DNA sequence and set of parameters, the definitions give rise to multiple different CpG island annotations, all of which are valid but which are often inconsistent with each other (Figure 16A). Second, current software tools for CpG island finding exploit – rather than limit – the ambiguity of the definitions, using algorithms that are highly unstable relative to minor differences in the DNA sequence. We demonstrate the latter point for the CpG Island Searcher (Takai and Jones 2002; Takai and Jones 2003), which is arguably the most popular software for CpG island finding, noting that similar arguments apply to other CpG island finders as well. Figure 16B displays a CpG island annotation calculated with CpG Island Searcher for a short DNA sequence (top row) as well as for four variants in which a single A is replaced by a C. From a theoretical point of view, two types of CpG island annotations seem plausible, one that is locally minimal (i.e. CpG islands are shortened until all basepairs are essential for fulfillment of the CpG island criteria) and one that is locally maximal (i.e. CpG islands are extended until any further extension will no longer fulfill the CpG island criteria). For the sequences in Figure 16B, the former alternative gives rise to three small CpG islands (as in row 1), and the latter alternative gives rise to one large CpG island spanning almost the entire sequence (as in row 3). Both solutions are feasible independent of whether and where a single A is replaced by a C (as described in Figure 16B). The CpG Island Searcher, however, is critically sensitive to this seemingly irrelevant sequence change and produces four qualitatively different annotations with no apparent correlation between the position of the replaced A and the number of CpG islands reported. (Technically, this problem is caused by a convoluted method for extending and shrinking CpG islands that is implemented in CpG Island Searcher.) Third, the CpG Island Searcher (as well as other software toolkits for CpG island finding) frequently overlooks valid CpG islands that fulfill all conditions of the definition, due to liberal use of heuristics (an example is given in Figure 16C). Based on these observations, we conclude that conceptual shortcomings of established CpG island definitions as well as practical issues of existing CpG islands finders currently preclude stable, accurate and reliable CpG island annotation of vertebrate genomes.

The goal of this chapter is to show that the shortcomings of current methods for CpG island annotation in mammalian genomes can be addressed by a more formal, informatics-based approach. As basis for our work, we propose a mathematically stringent formulation of the CpG island annotation problem, which resolves the ambiguity of the original definition by introducing an optimality condition. Next, we describe an algorithm for CpG island annotation that finds the correct CpG island annotation for a given DNA sequence according to the proposed definition, a fact that is formally proven. Furthermore, we illustrate that the formulation of CpG island annotation as an optimization problem provides a natural way of adapting the CpG island definition to different genomes and epigenomes.

While the work described in this chapter is somewhat exploratory, our medium-term goal is to develop a practically feasible and computationally efficient method for calculating complete and correct CpG island annotations of all mammalian genomes. Toward this goal, Lars Feuerbach currently designs and implements a speed-optimized version of the proposed algorithm, which is not part of this thesis.



A. The current CpG island definitions are underdetermined, i.e. give rise to multiple valid CpG island annotations

B. CpG island annotations derived with CpG Island Searcher are highly sensitive to minor sequence changes

C. The CpG Island Searcher misses valid CpG islands that fulfill all criteria of current CpG island definitions

$o' := \frac{\#CpG \cdot l}{\#C \cdot \#G} = \frac{1 \cdot 12}{4 \cdot 4} = 0.75$ ✓ $g_i := \frac{\#C + \#G}{l} = \frac{4+4}{12} \approx 0.67$ ✓

(“CGAAAAGGGCCC” is not detected by CpG Island Searcher because no subsequence of size 5 fulfills the CpG island criteria by itself)

Figure 16. Problems of current CpG island definitions and CpG island finders

Using small demonstration examples, this figure illustrates four problems that hamper stable, accurate and reliable CpG island annotation of vertebrate genomes. Panel A displays all locally minimal (top) and locally maximal (bottom) CpG islands according to a simplified CpG island definition requiring a minimum GC content of 50%, a CpG observed vs. expected ratio of at least 0.6 and a minimum length of 4 basepairs. It highlights that a single DNA sequence usually gives rise to multiple CpG island annotations, which are inconsistent with each other but all fulfill the same CpG island definition. Panel B depicts the CpG island annotations reported by the CpG Island Searcher (Takai and Jones 2002; Takai and Jones 2003) for the displayed DNA sequence as well as for four variants in which a single A is replaced by a C. Although these sequence changes were designed such that they need not affect the CpG island annotation, the results of CpG Island Searcher are altered quite dramatically and in non-obvious ways. Panel C gives an example of a valid CpG island that CpG island searcher misses because it contains no minimal-length region that fulfills CpG island criteria (here, we require a minimum CpG island length of 5 basepairs, in order to prevent a single CpG from fulfilling CpG island criteria on its own right).

B-5.1 Methods

Mathematical formulation of the CpG island annotation problem

Multiple variations of the initial CpG island definition by Gardiner-Garden and Frommer have been used (referenced above), usually in order to tweak their sensitivity and specificity to a specific application. These variants differ not only in the thresholds they impose on the GC content, CpG observed vs. expected ratio and length, but also in their handling of repetitive elements, the way in which the genome is scanned and the use of post-processing steps for merging neighboring CpG islands or excluding inappropriate CpG islands. The following definition is a moderate generalization of the Gardiner-Garden definition. It is flexible enough to incorporate multiple variants of the original definition but also provides a natural way to define a single best CpG island annotation for any given sequence.

Let $D = \{A, C, G, T\}^m$ be a DNA sequence with start position (index) zero and length *m*. We call a subsequence $D^{[s;\,e)}$ with start position *s* (inclusive) and end position *e* (exclusive) a CpG island and $A = \{D^{[s_i;\,e_i)}\}$ with $i \in [1;\, n]$ a CpG island annotation of *D* if the following conditions for separation, correctness, completeness and optimality are fulfilled.

(1) Separation (any two CpG islands are separated by at least one basepair): For all $i, j \in [1;\, n],\ i \neq j:\ s_i > e_j$ or $e_i < s_j$.

(2) Correctness (all CpG islands fulfill CpG island criteria): For all $i \in [1;\, n]$:

a. $g_i := \dfrac{\#C_i + \#G_i}{e_i - s_i} \geq t_a$ (GC content criterion)

b. $o_i := \dfrac{\#CpG_i \cdot (e_i - s_i)}{\#C_i \cdot \#G_i} \geq t_b$ (CpG observed vs. expected criterion)

c. $l_i := e_i - s_i \geq t_c$ (length criterion)

d. $z_i := f(D^{[s_i;\,e_i)}) \geq t_d$ (minimum score criterion, optional)

(3) Completeness (all CpG island candidates, i.e. regions that fulfill CpG island criteria, overlap with an element of the CpG island annotation): For any subsequence $D^{[s;\,e)}$ of *D* that fulfills the correctness condition (2), a $D^{[s_i;\,e_i)} \in A$ exists such that $s_i \leq e$ and $e_i \geq s$.

(4) Optimality (high-scoring CpG island candidates are included with higher priority than low-scoring CpG island candidates): Let $A = \{D^{[s_i;e_i)}\}$ be sorted in descending order of score values $f(D^{[s_i;\,e_i)})$. Then, for any subset *S* of *A* with $S = \bigcup_{j=1}^{i} \{D^{[s_j;e_j)}\}$ and size $i \in [1;\, n]$, any $D^{[s;e)}$ that fulfills the correctness condition (2) will either overlap with a higher-scoring CpG island that is already contained in the set *S* (i.e. $\exists D^{[s_j;\,e_j)} \in S : s \leq e_j \wedge e \geq s_j \wedge f(D^{[s;\,e)}) \leq f(D^{[s_j;\,e_j)})$) or it will have a lower score than any CpG island in *S* (i.e. $\forall D^{[s_j;\,e_j)} \in S : f(D^{[s;\,e)}) \leq f(D^{[s_j;\,e_j)})$.

Here, the variables g_i, o_i, l_i and z_i stand for the GC content, CpG observed vs. expected ratio, length and score, respectively, of the CpG island $D^{[s_i;\,e_i)}$. The parameter values t_a, t_b, t_c and t_d are constant thresholds that are chosen based on biological considerations (discussed below), $\#C_i$ and $\#G_i$ stand for the number of C and G nucleotides, respectively, in the region $D^{[s_i;\,e_i)}$, and $\#CpG_i$ stands for the number of CpG dinucleotides (i.e. "CG" patterns) in $D^{[s_i;\,e_i)}$.

We note that the CpG observed vs. expected criterion first introduced by Gardiner-Garden and Frommer (Gardiner-Garden and Frommer 1987) is exact for all lengths of $D^{[s_i;\,e_i)}$, which is not entirely obvious (see Proofs section). We also introduce a scoring function *f*, which unambiguously selects a single optimal CpG island annotation of *D* from the potentially large set of annotations that fulfill conditions (1), (2) and (3), and which can also be used to exclude weak CpG islands via the optional minimum score criterion. This scoring function can be selected according to biological considerations but has to fulfill two conditions. First, it must be defined for all regions $D^{[s;\,e)}$ in *D*. Second, it must be injective, i.e. two non-identical regions in D must be assigned non-identical score value (if this condition is violated the CpG island annotation is no longer guaranteed to be unambiguous).

An algorithm for exhaustive CpG island annotation of mammalian genomes

CpG island annotation of a sequence D of length m can be split into two consecutive steps, *search* and *annotation*: First, identify all subsequences of D that meet CpG island criteria (i.e. fulfill condition (2) of the above definition); second, find the single subset of these candidate CpG islands that constitutes a valid CpG island annotation of D (i.e. fulfills conditions (1) to (4) of the above definition). In principle, the search step can be performed by testing every possible subsequence of D for fulfillment of the CpG island criteria. However, this brute-force strategy is computationally infeasible for mammalian genomes, because it scales quadratically with the sequence length. We therefore introduce a pre-filtering strategy that discards a substantial proportion of genomic regions that cannot harbor any CpG islands and, thereby, greatly reduces the number of subsequence of D that have to be processed by subsequent exhaustive search. The annotation step is solved by a greedy algorithm, iteratively selecting the most high-scoring CpG island that does not overlap with any previously selected CpG island. Both steps are briefly described below, and the corresponding pseudocode is given in Figure 18.

During the *search step* (see Figure 17 for illustration), D is scanned with a “comb” of sliding windows with fixed sizes ranging from the minimum CpG island length t_c to the length of D plus one (we use $W = [t_c, 1.2 \cdot t_c, 1.2^2 \cdot t_c, 1.2^3 \cdot t_c, \ldots, m + 1]$, all values being rounded to the closest integer). For each sequence position in D and each window size w in W, the frequencies of Gs, Cs and CpGs in the subsequence $D^{[p-w;\, p)}$ are determined. If for a given position p and two consecutive window lengths w_i and w_{i+1} in W, the two corresponding sequence windows $D^{[p-w_i;\, p)}$ and $D^{[p-w_{i+1};\, p)}$ fall substantially short of the CpG island criteria, it is often possible to exclude all subsequences $D^{[p-w;\, p)}$ with $w_i \leq w < w_{i+1}$ from further analysis because under no circumstances can they be candidate CpG islands. More specifically, $D^{[p-w;\, p)}$ cannot fulfill CpG island criteria if at least one of the following conditions is violated:

$$(1) \quad \overline{g_i} := \frac{\#C_{i+1} + \#G_{i+1}}{w_i} \geq t_a$$

$$(2) \quad \overline{o_i} := \frac{\#CpG_{i+1} \cdot w_{i+1}}{\#C_i \cdot \#G_i} \geq t_b$$

Here, $\#C_i$, $\#G_i$ and $\#CpG_i$ as well as $\#C_{i+1}$, $\#G_{i+1}$ and $\#CpG_{i+1}$ denote the nucleotide counts in $D^{[p-w_i;\, p)}$ and $D^{[p-w_{i+1};\, p)}$, respectively, and $\overline{g_i}$ as well as $\overline{o_i}$ denote upper bounds on the GC content and observed vs. expected ratio of all possible subsequences $D^{[p;\, p+w)}$. A formal proof that the combing step does not overlook any valid CpG island is given in the Proofs section below.

For all subsequences of D that cannot be excluded by the combing conditions, the values for g_i, o_i, l_i and z_i are calculated and the region is accepted as a candidate CpG island into a list L if the corresponding threshold parameters t_a, t_b, t_c and t_d are met. In contrast to existing CpG island finding algorithms, the search step does not require any extending or merging of CpG islands. Rather, if several CpG islands at a particular position fulfill CpG island criteria, all of them are added to the list of candidate CpG islands and the selection which of them to include in the final CpG island annotation is left to the annotation step. Furthermore, we note that the choice of combing windows W influences the runtime performance but has no effect on the correctness of the algorithm as long as t_c and $m + 1$ are included as bottom and top window sizes in W.

During the *annotation step*, a greedy algorithm is used to select a subset of candidate CpG islands such that the resulting CpG island annotation A fulfills all conditions (1) to (4) of the above definition. First, the list of candidate CpG islands L is sorted by decreasing score values $f(D^{[s_i;\, e_i)})$. Because L can be large, we use external mergesort (Zheng and Larson 1996) to sort L in place on an external disk. Second, the candidate CpG islands in L are processed in decreasing order of their score values and selected if they do not overlap with higher-scoring CpG islands, safeguarding condition (1) of the definition. This step is efficiently performed by growing a binary search tree (or a B+ tree when the size of the tree exceeds available memory) of the genomic regions of all selected CpG islands. Finally, the search tree is traversed in depth-first mode and a list A containing the CpG island annotation of sequence D is constructed. We note that the optimality condition (4) is automatically taken care of by processing the list of candidate CpG islands in descending order of score values and that the completeness condition (3) is fulfilled because L contains all candidate CpG islands in D. A formal proof of correctness and termination of this algorithm is given in the Proofs section below.

Overall, the algorithm has a worst-case complexity of $O(m^2 \cdot \log(m^2))$, due to the need for sorting the list of candidate CpG islands L, which scales quadratically with the length m of D. However, in mammalian genomes CpGs are substantially depleted and CpG islands are rare, hence a large fraction of regions are already discarded in the filtering step, which can be performed in $O(m \cdot \log(m))$. Initial empirical results based on a Python prototype that implements the algorithm outlined in Figure 18 (with some modifications, designed and programmed by Lars Feuerbach) suggest that exhaustive CpG island annotation of the human genome is feasible within a runtime in the order of days to weeks on a single CPU (L. Feuerbach and C. Bock, unpublished observation). While further optimization is clearly possible and desirable, a runtime of approximately 24 hours on standard hardware may well be acceptable given that CpG island annotations have to be calculated only once for a given genome assembly.

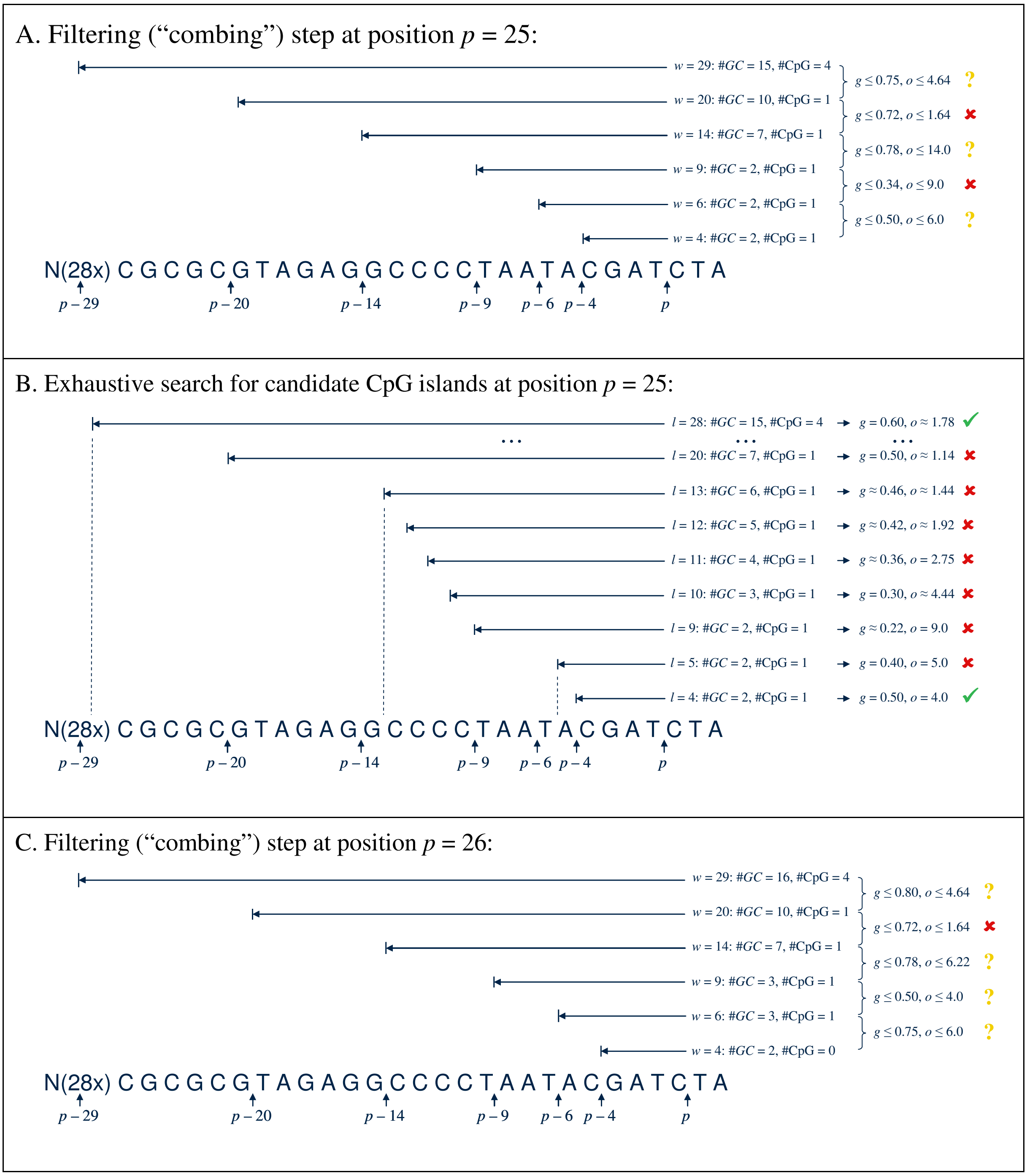

A. Filtering (“combing”) step at position p = 25:
w = 29: #GC = 15, #CpG = 4
w = 20: #GC = 10, #CpG = 1
w = 14: #GC = 7, #CpG = 1
w = 9: #GC = 2, #CpG = 1
w = 6: #GC = 2, #CpG = 1
w = 4: #GC = 2, #CpG = 1
$g \leq 0.75$, $o \leq 4.64$?
$g \leq 0.72$, $o \leq 1.64$ ✗
$g \leq 0.78$, $o \leq 14.0$?
$g \leq 0.34$, $o \leq 9.0$ ✗
$g \leq 0.50$, $o \leq 6.0$?
N(28x) C G C G C G T A G A G G C C C C T A A T A C G A T C T A
$p - 29$
$p - 20$
$p - 14$
$p - 9$
$p - 6$
$p - 4$
p
B. Exhaustive search for candidate CpG islands at position p = 25:
l = 28: #GC = 15, #CpG = 4 → g = 0.60, $o \approx 1.78$ ✓
…
l = 20: #GC = 7, #CpG = 1 → g = 0.50, $o \approx 1.14$ ✗
l = 13: #GC = 6, #CpG = 1 → $g \approx 0.46$, $o \approx 1.44$ ✗
l = 12: #GC = 5, #CpG = 1 → $g \approx 0.42$, $o \approx 1.92$ ✗
l = 11: #GC = 4, #CpG = 1 → $g \approx 0.36$, o = 2.75 ✗
l = 10: #GC = 3, #CpG = 1 → g = 0.30, $o \approx 4.44$ ✗
l = 9: #GC = 2, #CpG = 1 → $g \approx 0.22$, o = 9.0 ✗
l = 5: #GC = 2, #CpG = 1 → g = 0.40, o = 5.0 ✗
l = 4: #GC = 2, #CpG = 1 → g = 0.50, o = 4.0 ✓
N(28x) C G C G C G T A G A G G C C C C T A A T A C G A T C T A
$p - 29$
$p - 20$
$p - 14$
$p - 9$
$p - 6$
$p - 4$
p
C. Filtering (“combing”) step at position p = 26:
w = 29: #GC = 16, #CpG = 4
w = 20: #GC = 10, #CpG = 1
w = 14: #GC = 7, #CpG = 1
w = 9: #GC = 3, #CpG = 1
w = 6: #GC = 3, #CpG = 1
w = 4: #GC = 2, #CpG = 0
$g \leq 0.80$, $o \leq 4.64$?
$g \leq 0.72$, $o \leq 1.64$ ✗
$g \leq 0.78$, $o \leq 6.22$?
$g \leq 0.50$, $o \leq 4.0$?
$g \leq 0.75$, $o \leq 6.0$?
N(28x) C G C G C G T A G A G G C C C C T A A T A C G A T C T A
$p - 29$
$p - 20$
$p - 14$
$p - 9$
$p - 6$
$p - 4$
p

Figure 17. Simple search-step example of the CpG island annotation algorithm

This figure illustrates the search step of the CpG island annotation algorithm on a short DNA sequence D of length 29, using simplified CpG island criteria (t_a= 0.5, t_b = 1.75 and t_c = 4) and no scoring function. For position p = 25, panel A exemplarily shows how “combing” with six sliding windows reduces the number of subsequences that can potentially contain CpG islands. Panel B depicts the follow-up exhaustive search performed on the selected length intervals. Panel C illustrates how the combing results change when p is incremented by one and the sliding windows moved by one position to the right. All diagrams are best read bottom to top, i.e. from the DNA sequence upward. The symbols carry the following meaning: “?” – window can potentially contain CpG islands; “✗” – window cannot contain any CpG island (for panels A and C) or region does not fulfill CpG island criteria (for panel B) ; “✓” – region fulfills CpG island criteria and is thus a candidate CpG island.

Adapting the definition and annotation of CpG islands to specific genomes

It has been shown that the original CpG island definition gives rise to a large number of apparent false positives, i.e. genomic regions that show no evidence of the regulatory and functional roles that are considered constitutive of CpG islands (Bock et al. 2007; Takai and Jones 2002). Furthermore, it has been reported that CpG islands are apparently rarer in the mouse genome than in the human genome, at least when the same CpG island definition is used (Antequera and Bird 1993; Waterston et al. 2002). These observations highlight the need for adapting the CpG island definition to the properties of specific genomes.

Our formulation of CpG island annotation as an optimization problem and the use of a scoring function *f* to prioritize the selection of candidate CpG islands provides a natural way of incorporating such considerations into the CpG island definition. For example, we can use the weighted sum of GC content, CpG observed vs. expected ratio and CpG island length as a predictor of CpG island strength (Bock et al. 2007, cf. chapter B-4 of this thesis), in order to distinguish bona fide CpG islands from false positives. To that end, we define a new scoring function $f_{weighted}(D^{[s_i;\, e_i)}) := w_1 \cdot g_i + w_2 \cdot o_i + w_3 \cdot l_i + c \cdot [m \cdot (e - s) + s]$. The first three terms define the weighted sum, while the sole purpose of the last term is to ensure that $f_{weighted}$ is injective (a requirement of our definition). The parameter *c* is chosen such that – for all candidate CpG islands – this term is (i) greater than zero and (ii) smaller than the smallest difference between the sums of the first three terms for any pair of candidate CpG islands for which this difference is non-zero. In other words, the last term is used only for breaking ties between CpG islands that would otherwise carry exactly identical scores. The weight parameters w_1, w_2 and w_3 can either be selected based on biological knowledge or they can be learnt from a training dataset comprising bona fide CpG islands and false positives, using an optimization method such as simulated annealing or evolutionary algorithms.

B-5.2 Proofs

Formula for the CpG observed vs. expected ratio

The CpG observed vs. expected ratio is defined as the ratio between the number of observed CpGs (*#CpG*) and its expected value based on the number of Cs (*#C*) and Gs (*#G*), for a given DNA sequence $D = \{A, C, G, T\}^m$ with start index zero and length *m*.

Theorem. Under the assumption of independence between its positions (i.e. assuming that *D* is a zero-order Markov chain), the expected value for the number of CpGs in *D* is:

$$E(\#CpG) = \frac{\#C \cdot \#G}{m}.$$

Proof. This formula for the expected value is not entirely obvious, given the fact that CpGs can start at any position between zero and *m* – 2, but not at the last position of the sequence, and that a CpG starting at position *p* precludes a CpG at position *p* + 1. However, it can be derived easily, using the linearity of the expected value *E*, as well as the observations that no CpG can start at position *m* – 1 and that the probability of a G occurring at position *p* + 1 increases from $\frac{\#G}{m}$ to $\frac{\#G}{m-1}$ when $D^{[p;\, p+1)}$ is known to be a C:

$$E(\#CpG) = \sum_{i=0}^{m-1} E\left[prob(D^{[p;\ p+2)} = "CG")\right] = \sum_{i=0}^{m-2} prob(D^{[p;\ p+1)} = "C") \cdot prob(D^{[p;\ p+1)} = "G") = (m-1) \cdot \frac{\#C}{m} \cdot \frac{\#G}{m-1} = \frac{\#C \cdot \#G}{m}.$$

The formula for the CpG observed vs. expected ratio follows directly:

$$o_i := \frac{\#CpG_i \cdot (e_i - s_i)}{\#C_i \cdot \#G_i}.$$

Correctness proof for the CpG island annotation algorithm

Theorem: For any DNA sequence *D*, threshold parameters t_a, t_b, t_c and t_d greater than zero and an injective scoring function *f* (according to the definition of the CpG island annotation problem outlined above), the algorithm calculates the correct CpG island annotation *A*.

Proof: The four conditions of the definition and the question of termination are treated separately. All line numbers refer to the pseudocode given in Figure 18.

(1) Separation condition

For *A* to violate the separation condition, two non-separated regions $D^{[s_i;e_i)}$ and $D^{[s_j;e_j)}$ with $s_i \le e_j$ and $e_i \ge s_j$ must be inserted into the binary tree *B* (line 42), from which *A* is constructed (line 43). Because *B* is filled iteratively, without loss of generality we can assume that $D^{[s_i;e_i)}$ is already contained in *B* when $D^{[s_j;e_j)}$ is inserted. Because *B* is a sorted tree (with *compareRegions* defining a total order on all regions in *D*), the insertion will ultimately lead to direct comparison of $D^{[s_i;e_i)}$ and $D^{[s_j;e_j)}$ (unless it fails earlier due to $D^{[s_j;e_j)}$ overlapping with another CpG island in the tree). In that case, however, *compareRegions* will return zero (indicating overlap) because of $s_i \le e_j$ and $e_i \ge s_j$, and $D^{[s_j;e_j)}$ will thus not be inserted into *B* (line 20).

(2) Correctness condition

First, we show that after execution of line 23 the count variables #*C*[*w*], #*G*[*w*] and #*CpG*[*w*] are equal to the number of Cs, Gs and CpGs, respectively, present in the subsequence *D*[*p* – *w, p*), for all values of *p* and *w* of the for-loops in lines 21 and 22. In the first iteration (*p* = 1), this assertion is obviously true for all *w* in *W* because *D*[*p* – *w* , *p*) overlaps with only a single non-NULL sequence character, *D*[*p* – 1] = *D*[0], and the count variables are updated accordingly by the *slideWindow* function (lines 12 to 15). Next, assume that the assertion is correct after the *p*-th loop iteration ($p \in [1, ..., m-1]$). Then it will also be correct for the subsequent loop iteration (*p*' = *p* + 1) because the *slideWindow* function decrements the corresponding count variables by one for any C, G or CpG leaving the sliding window at position *p*' – *w* – 1 and increments it by one for any C, G or CpG entering the sliding window at position *p*' – 1. By induction over *p*, it follows that the count variables are correct for all values of *p* and *w* of the for-loops in lines 21 and 22.

Second, we use a similar argument to show that before the execution of line 30, the count variables *#C*', *#G*', and *#CpG*' are equal to the number of Cs, Gs and CpGs, respectively, present in the subsequence *D*[*p* – *l* , *p*), for all values of *p* and *l* of the for-loops in lines 21 and 29. In the first iteration of the inner loop, this assertion follows directly from the previous argument because *l* = *w* and the correctness of the count values has already been shown for all values of p and w. Next, the induction step from l to $l' = l + 1$ is similar as above, with the only difference that the window is extended rather than slid, such that it is sufficient to increment the count variables for positions entering the window.

Third, we note that the correctness of the count values *#C'*, *#G'* and *#CpG'* together with the test for fulfillment of the CpG island criteria (lines 30 to 33) ensure that only subsequences that meet condition (2) of the definition are added to the list of candidate CpG islands L (line 34). Because the CpG island annotation A is constructed as a subset of L (lines 41 to 42), it follows that A contains only regions that fulfill CpG island criteria.

(3) Completeness condition

Let $D[p - l\ ,\ p)$ be a subsequence of D with length l and (di-) nucleotide frequencies *#C*, *#G*, and *#CpG*, which fulfills CpG island criteria (i.e. condition (2) of the definition). First, we show that this region is added to L during the search step. By definition, p cannot exceed the length of D, and because $D[p - l\ ,\ p)$ is a CpG island, it is no shorter than t_c; in other words: $t_c \leq l \leq m$. Therefore, a j exists such that l falls between two consecutive combing window sizes $W[j]$ and $W[j + 1]$, i.e. $W[j] \leq l < W[j + 1]$. The corresponding subsequences $D[p - W[j]\ ,\ p)$ and $D[p - W[j + 1]\ ,\ p)$ are processed during the combing step (lines 21 to 28) and result in the region size interval [W[j]; $W[j + 1]$) being marked for exhaustive search (lines 25 to 28), as can be seen from the following estimate:

$$(1)\quad \max GC = \frac{\#C[j+1] + \#G[j+1]}{W[j]} \geq \frac{\#C + \#G}{l} \geq t_a$$

$$(2)\quad \max OE = \frac{\#CpG[j+1] \cdot W[j+1]}{\#C[j] \cdot \#G[j]} \geq \frac{\#CpG \cdot l}{\#C \cdot \#G} \geq t_b$$

The important point of this estimate is that $D[p - W[j + 1]\ ,\ p)$ contains at least as many Cs, Gs and CpGs, respectively, as $D[p - l\ ,\ p)$ because the latter is fully contained in the former. Similarly, $D[p - W[j],\ p)$ can never exceed $D[p - l\ ,\ p)$ in length because the former is fully contained in the latter. Therefore, $D[p - l\ ,\ p)$ is processed by the exhaustive search (lines 29 to 38) and added to the list of candidate CpG islands L (line 34) because it fulfills CpG island criteria.

Second, because $D[p - l\ ,\ p)$ is an element of L and all elements of L are processed during the annotation step (lines 41 to 42), two cases can occur. On the one hand, $D[p - l\ ,\ p)$ may be inserted into B, in which case the condition (3) is fulfilled because $D[p - l\ ,\ p)$ overlaps with itself and is a CpG island. On the other hand, the insertion may fail because the tree search terminates on a region already contained in B, with *compareRegions* returning zero. In that case it follows that a region $D[p' - l'\ ,\ p')$ exists in B, such that $p \leq p' + l'$ and $p + l \geq p'$. Because $D[p' - l'\ ,\ p')$ is a CpG island (cf. correctness condition), condition (3) is fulfilled.

(4) Optimality condition

Let S be a subset of A containing the i CpG islands with highest score values for f. Assume that a region $D[p - l , p)$ exists such that: (i) $D[p - l , p)$ fulfills CpG island criteria (i.e. condition (2) of the definition is met), (ii) $D[p - l , p)$ does not overlap with any $D[p_1 - l_1 , p_1)$ in S with $f(D[p - l , p)) \leq f(D[p_1 - l_1, p_1))$ and (iii) a $D[p_2 - l_2 , p_2)$ exists in S with $f(D[p - l , p)) \geq f(D[p_2 - l_2, p_2))$. Because $D[p - l , p)$ is a candidate CpG island (first assumption), it is added to L during the search step (cf. correctness condition). Then, two cases can occur: Either $D[p - l , p)$ has been added to B during the annotation step, which leads to a contradiction with the second assumption because $D[p - l , p)$ overlaps with itself and would be among the i CpG islands with highest scores (third assumption), or its insertion into B has failed (line 42). The insertion fails only when *compareRegions* returns zero, hence a $D[p_3 - l_3 , p_3)$ exists that overlaps with $D[p - l , p)$, i.e. $p \leq p_3 + l_3$ and $p + l \geq p_3$. This other region must have a higher score than $D[p - l , p)$ because it was inserted into B prior to $D[p - l , p)$. In that case, however, $D[p_3 - l_3 , p_3)$ would be among the i CpG islands with highest scores, thus contradicting the third assumption. The optimality condition follows.

(5) Termination

The algorithm contains five loops. The first loop (line 21) iterates over the length of D, which is finite by definition; the second loop (line 22) and third loop (line 24) iterate over the number of combing window sizes, which comprises a subset of integer values between zero and $m + 1$; the fourth loop (line 29) iterates over a number of subsequences of D and the fifth loop (line 41) iterates over a list that can at maximum consistent of all subsequences of D. Hence, all loops involve a limited number of iterations and the algorithm thus terminates in finite time.

Input

$D[0, \ldots, m)$ *String of length m representing the DNA sequence, with* $D[i] \in \{A, C, G, T\}$

f *Scoring function, must be injective (i.e. no two genomic regions receive the same score) and defined for all subsequences of D*

t_a *Minimum threshold on the GC content*

t_b *Minimum threshold on the CpG observed vs. expected ratio*

t_c *Minimum threshold on the CpG island length*

t_d *Minimum threshold on the CpG island score*

Output

$A[0, \ldots, n)$ *List of genomic regions constituting the single valid CpG island annotation of D for the given parameters and scoring function*

Initialization and definition of helper functions

```
$D[-m, \ldots, 0) \leftarrow$ [NULL, …, NULL]        // Initialize positions outside the sequence with missing values (which simplifies the pseudocode and obviates the need for handling end-of-sequence cases)
$W[1, \ldots, v] \leftarrow \mathrm{int}([t_c, 1.2 \cdot t_c, 1.2^2 \cdot t_c, 1.2^3 \cdot t_c, \ldots, m + 1])$        // Define sliding window sizes for the combing step (all values are rounded to the closest integer value)
$\#C[1, \ldots, v] \leftarrow [0,\ldots,0]$        // Initialize nucleotide counts for each combing window size
$\#G[1, \ldots, v] \leftarrow [0,\ldots,0]$
$\#CpG[1, \ldots, v] \leftarrow [0,\ldots,0]$
$L \leftarrow []$        // Initialize list of candidate CpG islands (i.e. regions fulfilling CpG island criteria)

slideWindow ← function ($p, w, \#C, \#G, \#CpG$):        // Function for shifting the sliding window right by one position
    if $D[p - w - 1]$ = "C" then:
        $\#C[w] \leftarrow \#C[w] - 1$        // Decrement the counts of Cs, Gs and CpGs for nucleotides leaving the sliding window
        if $D[p - w]$ = "G" then: $\#CpG[w] \leftarrow \#CpG[w] - 1$
    if $D[p - w - 1]$ = "G" then: $\#G[w] \leftarrow \#G[w] - 1$
    if $D[p - 1]$ = "G" then:
        $\#G[w] \leftarrow \#G[w] + 1$        // Increment the counts of Cs, Gs and CpGs for nucleotides entering the sliding window
        if $D[p - 2]$ = "C" then: $\#CpG[w] \leftarrow \#CpG[w] + 1$
    if $D[p - 1]$ = "C" then: $\#C[w] \leftarrow \#C[w] + 1$
    return ($\#C, \#G, \#CpG$)        // Return the updated nucleotide counts

compareRegions ← function ($s, l, s', l'$):        // Compare function used by the binary search tree to determine whether a candidate CpG island is overlapping with (or directly adjacent to) an already selected CpG island in the tree (result = 0), is left of it (result = -1) or is right of it (result = 1)
    if $s + 1 < s'$: return -1
    if $s > s' + l'$: return 1
    if $s \leq s' + l'$ and $s + l \geq s'$: return 0
```

Figure 18. Pseudocode for the CpG island annotation algorithm

The pseudocode of the CpG island annotation algorithm comprises two major steps, *search* and *annotation*. During the search step, all CpG island candidates are identified, first by fast "combing", which retains only those genomic regions that can potentially contain CpG islands, and second, by exhaustive search on these target regions. During the annotation step, a subset of high-scoring and non-overlapping CpG islands is selected from the list of all candidate CpG islands, giving rise to a CpG island annotation that fulfills all criteria of the proposed definition (see Proofs section for a correctness proof for this algorithm).

Search step

```
for p ← 1 to m do:                                        For each nucleotide in D, acting as potential end
                                                          position of a CpG island D[p – l , p):
    for w in W do:                                        Calculate the C, G and CpG frequency in all com-
        (#C, #G, #CpG) ← slideWindow(p, w, #C, #G, #CpG)  bing windows by maintaining sliding window
                                                          counts
    for j ← 1 to v – 1 do:                                For each pair of combing windows j, j + 1:
        maxGC ← (#C[j + 1] + #G[j + 1]) / W[j]            Calculate an upper bound for the GC content and
        maxOE ← (#CpG[j + 1] · W[j + 1]) / (#C[j] · #G[j]) observed vs. expected ratio of all subsequences of
                                                          D ending at position p with length w, W[j] ≤ w ≤
                                                          W[j+1]
        if maxGC ≥ $t_a$ and maxOE ≥ $t_b$ then:          Perform exhaustive search on all regions that
                                                          could contain a candidate CpG island
            (#C', #G', #CpG') ← (#C[j], #G[j], #CpG[j])   Store nucleotide counts of the shorter window j in
                                                          local variables
            for l ← W[j] to W[j + 1] – 1 do:              For each length l in the combing interval:
                g ← (#C' + #G') / l
                o ← (#CpG' · l) / (#C' · #G')             Calculate CpG island properties for the subse-
                z ← f(D[p – l , p))                       quence of D starting at position (p – l) with length l
                if g ≥ $t_a$ and o ≥ $t_a$ and l ≥ $t_a$ and z ≥ $t_a$:   If the current region fulfills CpG island criteria:
                    L = L + {(p – l, l, z)}               Add its start position, length and score to the list of
                                                          candidate CpG islands
                if D[p – l – 1] = "C" then:
                    #C' ← #C' + 1                         Extend the current region by one nucleotide to the
                    if D[p – l] = "G" then:               left and increment the counts of Cs, Gs and CpGs
                        #CpG' ← #CpG' + 1                 accordingly
                if D[p – l – 1] = "G" then: #G' ← #G' + 1
```

Annotation step

```
L ← performExternalMergeSort(L)              Sorts the candidate CpG islands in descending order of f-scores. An
                                             external sorting algorithm is used because the size of L may exceed
                                             available memory
B ← initializeBinaryTree(compareRegions)     Initialize a binary search tree for keeping track of selected CpG isl-
                                             ands, with compareRegions used as the compare function of the search
                                             tree
for (p – l, l, z) in L:                      For each candidate CpG in descending order of scores:
    findOrInsert(B, (p – l, l, z))           Insert it into the tree only if it does not overlap with already selected
                                             CpG islands
A ← performDepthFirstTraversal(B)            'Flatten' the tree into a list of CpG islands
```

Figure 18 (continued).

Part C. DNA Methylation Mapping

Epigenetics offers us a […] kind of map. One where we can zoom in and zoom out. A map of many colors, with street signs so we can navigate, routes that we can choose, destinations that we can change (Jill Neimark)[1]

C-1 Outline

DNA methylation – often considered the showcase example of epigenetic regulation (cf. section A-3 of this thesis) – adds a layer of regulatory information to the genome that has broad relevance for normal development and disease. Therefore, it is not surprising that DNA methylation mapping of the entire human genome was suggested early on as a logical next step to follow the Human Genome Project (Evans 2000). After several years of rapid technological advances, this goal seems now in reach, and several international projects aim to map DNA methylation genome-wide, at high resolution and in multiple tissues, cell types and individuals simultaneously (reviewed in Bernstein et al. 2007; Bock and Lengauer 2008).

In the following chapters, three pilot studies addressing challenges of large-scale DNA methylation mapping are described. In chapter C-2, we present BiQ Analyzer, a bioinformatic software tool for visual analysis and quality control of DNA methylation data obtained by bisulfite sequencing (Bock et al. 2005). In chapter C-3, we summarize the results of bioinformatic analysis performed on the NAME21 (National Methylome Project for Chromosome 21) dataset. The NAME21 project made extensive use of bisulfite sequencing and the BiQ Analyzer software in order to map DNA methylation at single-basepair and single-cell resolution for a sizable fraction of human promoter regions (see Jeltsch et al. 2006 for project announcement; the results publication is in preparation). Finally, in chapter C-4, computational analysis of inter-individual variation of DNA methylation highlights how the combination of classical statistics, machine learning and computational simulation can help identify a cost-efficient strategy for genome-wide DNA methylation mapping in a large number of individuals (Bock et al. 2008).

C-2 BiQ Analyzer: Visualization and quality control for DNA methylation data from bisulfite sequencing[2]

C-2.1 Motivation

The most accurate and probably the most widely used experimental protocol for analyzing DNA methylation makes use of selective conversion of unmethylated cytosines to uracils, induced by bisulfite treatment (Frommer et al. 1992; Hajkova et al. 2002). Subsequent amplification, cloning, sequencing, and comparison with the genomic sequence allows for identification of unmethylated cytosines, which appear as thymines in a multiple sequence alignment. Although bisulfite sequencing protocol is generally reliable, the necessary data processing steps are tedious to perform manually, and several potential error sources have to be addressed. We developed BiQ Analyzer, an interactive software tool that provides start-to-end support for this process. In an easy-to-use manner, the tool helps the user to import the se-

[1] Quoted after: http://www.edge.org/q2007/q07_14.html

[2] This chapter describes published work conducted in collaboration with Sabine Reither, Thomas Mikeska, Martina Paulsen and Jörn Walter (Bock et al. 2005). Martina Paulsen suggested to develop a software for automated analysis of DNA methylation data from bisulfite sequencing. All collaboration partners contributed their expertise with manual data analysis and acted as pilot users of BiQ Analyzer.

quence files from the sequencer, to align them, to exclude or correct critical sequences, to document the experiment, to perform basic statistical analysis and to produce publication-quality diagrams.

C-2.2 Methods

Potential error sources in bisulfite sequencing arise from three phases of the experimental protocol: bisulfite conversion, PCR, and sequencing. Each of these steps can give rise to characteristic errors in the sequences, which the experimenter must address before deriving DNA methylation profiles. Here we describe these error types, their impact on methylation data, and the measures of quality control that BiQ Analyzer applies to identify the critical sequences.

(1) **Incomplete conversion**. In bisulfite sequencing we assume that all unconverted Cs were originally methylated. Therefore, when the bisulfite treatment fails to convert unmethylated Cs, DNA methylation will be overestimated. Fortunately, for vertebrates it is possible to identify those sequences with low conversion rates, assuming that Cs outside a CpG context are always unmethylated (Reik et al. 2003). BiQ Analyzer calculates the conversion rate of a sequence as the ratio between the number of correctly converted Cs outside a CpG context divided by the sum of converted and unconverted Cs outside a CpG context. By default, BiQ Analyzer marks all sequences with a conversion rate below 90% as critical (the default values for all parameters were selected based on expert opinion by researchers with extensive experience in bisulfite sequencing and can be modified according to user preferences).

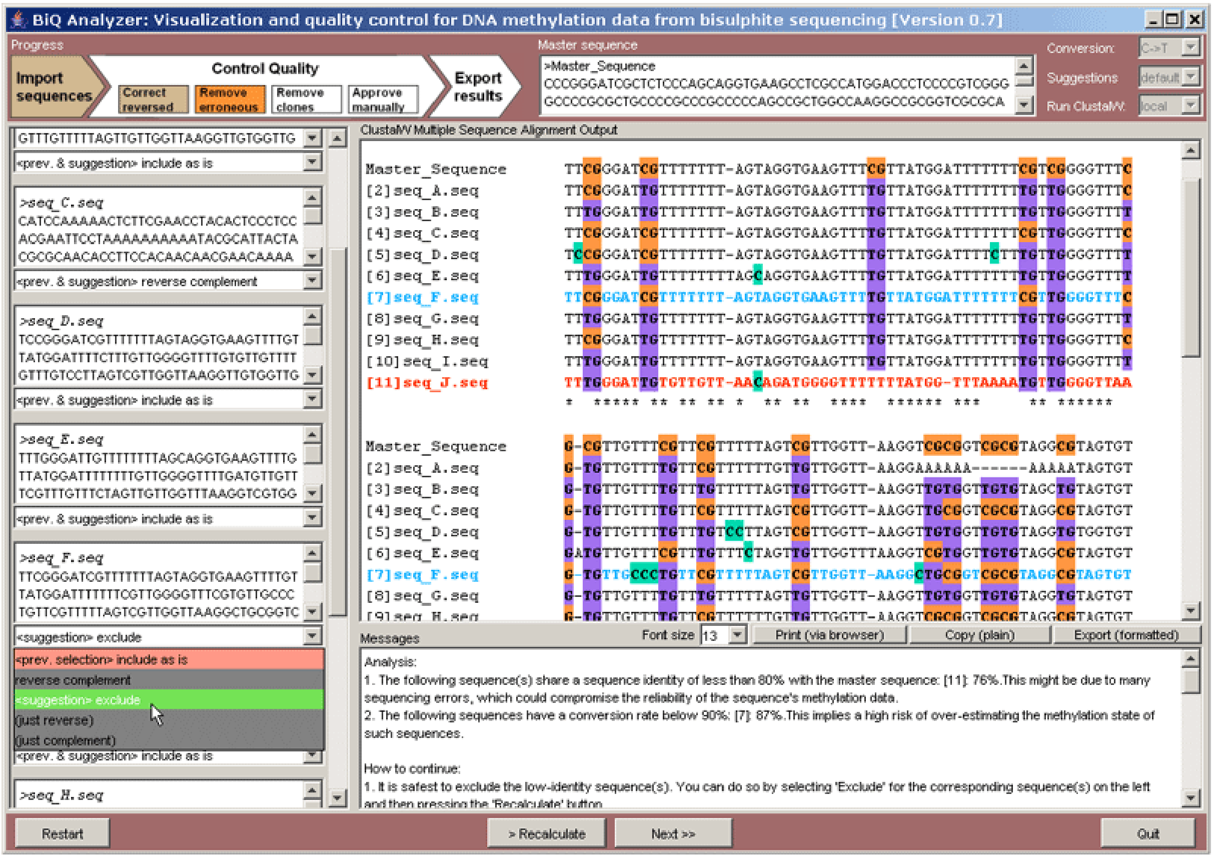
(2) **Clone sequences**. PCR amplification bias can result in vast over-representation of sequences from a single cell or from a small number of cells. Regarding the resulting identical clones as independent sources of DNA methylation data would result in biased estimation of the overall variability of DNA methylation within a sample. BiQ Analyzer thus implements a heuristic clone detection method. It marks those sequences as critical that are identical in all correctly aligned C positions. The advantage of this method over simple sequence comparison is that it is insensitive to sequence truncations and sequencing errors at non-C positions. However, it can lead to discarding of a significant number of valid clones when conversion rates are close to 100% and all cells in a sample are highly similar in their DNA methylation patterns.

(3) **Sequencing errors**. Sequencing errors changing C to T and vice versa can lead to errors in the DNA methylation patterns derived from the sequences. Therefore, BiQ Analyzer suggests to exclude all sequences that fall below a local sequence identity level of 80% as compared to the genome sequence (C-to-T conversions and truncations are ignored). Furthermore, in our experiments we regularly observe ambiguous base insertions within a CpG context (i.e. CG → CTG or CG → TCG). In these cases, BiQ Analyzer reports the methylation state of the CpG dinucleotide as unknown.

As a Java application, BiQ Analyzer runs on almost any platform, requiring only a recent version of the Java virtual machine (which can be downloaded from http://www.javasoft.com/) and a screen resolution of at least 1024 · 768 pixels. The multiple sequence alignment makes use of a local version of ClustalW (Thompson et al. 1994), which is included in the standard download package. The alignment step is computationally expensive and can be slow on older computers. Therefore, the program provides an option to calculate the alignment over the internet on a high-performance computer at the Max Planck Institute for Informatics.

C-2.3 Results and Discussion

BiQ Analyzer is a software tool designed to mimic the manual process of DNA methylation analysis (Figure 19). In several steps, the user is guided from the import of sequences, across several phases of quality control and multiple sequence alignment, to a questionnaire documenting the experiment. In each of the quality control steps, the program makes suggestions on how to handle critical sequences, but the ultimate decision to include or exclude a sequence always stays with the user. Based on the user decisions during that process, the program finally generates a single-file HTML documentation (including publication-quality methylation diagrams in the widely-used “lollipop” style) and saves the derived methylation data to the system clipboard, ready for subsequent analysis with spreadsheet software or a statistics package.



Conversion: C->T

Suggestions: default

Run ClustalW: local

Font size 13

```
>Master_Sequence
CCCGGGATCGCTCTCCCAGCAGGTGAAGCCTCGCCATGGACCCTCCCCGTCGGG
GCCCCGCGCTGCCCCGCCCGCCCCCAGCCGCTGGCCAAGGCCGCGGTCGCGCA
```

```
GTTTGTTTTTAGTTGTTGGTTAAGGTTGTGGTTG
<prev. & suggestion> include as is

>seq_C.seq
CATCCAAAAACTCTTCGAACCTACACTCCCTCC
ACGAATTCCTAAAAAAAAAAATACGCATTACTA
CGCGCAACACCTTCCACAACAACGAACAAAA
<prev. & suggestion> reverse complement

>seq_D.seq
TCCGGGATCGTTTTTTTTAGTAGGTGAAGTTTTGT
TATGGATTTTCTTTGTTGGGGTTTTGTGTTGTTTT
GTTTGTCCTTAGTCGTTGGTTAAGGTTGTGGTTG
<prev. & suggestion> include as is

>seq_E.seq
TTTGGGATTGTTTTTTTTTAGCAGGTGAAGTTTTG
TTATGGATTTTTTTTGTTGGGGTTTTGATGTTGTT
TCGTTTGTTTCTAGTTGTTGGTTTAAGGTCGTGG
<prev. & suggestion> include as is

>seq_F.seq
TTCGGGATCGTTTTTTTTAGTAGGTGAAGTTTTGT
TATGGATTTTTTTCGTTGGGGTTTCGTGTTGCCC
TGTTCGTTTTTAGTCGTTGGTTAAGGCTGCGGTC
<suggestion> exclude
<prev. selection> include as is
reverse complement
<suggestion> exclude
(just reverse)
(just complement)
<prev. & suggestion> include as is

>seq_H.seq
```

```
Master_Sequence    TTCGGGATCGTTTTTTTT-AGTAGGTGAAGTTTCGTTATGGATTTTTTTTCGTCGGGGTTTC
[2]seq_A.seq       TTCGGGATTGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTC
[3]seq_B.seq       TTTGGGATTGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTT
[4]seq_C.seq       TTCGGGATCGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTCGTTGGGGTTTC
[5]seq_D.seq       TCCGGGATCGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTCTTTGTTGGGGTTTT
[6]seq_E.seq       TTTGGGATTGTTTTTTTTTAGCAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTT
[7]seq_F.seq       TTCGGGATCGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTCGTTGGGGTTTC
[8]seq_G.seq       TTTGGGATTGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTT
[9]seq_H.seq       TTCGGGATTGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTC
[10]seq_I.seq      TTTGGGATTGTTTTTTTT-AGTAGGTGAAGTTTTGTTATGGATTTTTTTTTGTTGGGGTTTT
[11]seq_J.seq      TTTGGGATTGTGTTGTT-AACAGATGGGGTTTTTTTATGG-TTTAAAATGTTGGGGTTAA
                   *  ***** ** ** ** *  ** **  ****  ****** ***      ** ******

Master_Sequence    G-CGTTGTTTCGTTCGTTTTTAGTCGTTGGTT-AAGGTCGCGGTCGCGTAGGCGTAGTGT
[2]seq_A.seq       G-TGTTGTTTTGTTCGTTTTTTTGTTGTTGGTT-AAGGAAAAAA------AAAAATAGTGT
[3]seq_B.seq       G-TGTTGTTTTGTTTGTTTTTAGTTGTTGGTT-AAGGTTGTGGTTGTGTAGCTGTAGTGT
[4]seq_C.seq       G-TGTTGTTTTGTTCGTTTTTAGTCGTTGGTT-AAGGTTGCGGTCGCGTAGGCGTAGTGT
[5]seq_D.seq       G-TGTTGTTTTGTTTGTCCTTAGTCGTTGGTT-AAGGTTGTGGTTGTGTAGGTGTGGTGT
[6]seq_E.seq       GATGTTGTTTCGTTTGTTTCTAGTTGTTGGTTTAAGGTCGTGGTTGTGTAGGCGTAGTGT
[7]seq_F.seq       G-TGTTGCCCTGTTCGTTTTTAGTCGTTGGTT-AAGGCTGCGGTCGCGTAGGCGTAGTGT
[8]seq_G.seq       G-TGTTGTTTTGTTTGTTTTTAGTTGTTGGTT-AAGGTTGTGGTTGTGTAGGTGTAGTGT
[9]seq_H.seq       G-TGTTGTTTTGTTCGTTTTTTTGTTGTTGGTT-AAGGTCGCGGTCGCGTAGGCGTAGTGT
```

Analysis:

1. The following sequence(s) share a sequence identity of less than 80% with the master sequence: [11]: 76%.This might be due to many sequencing errors, which could compromise the reliability of the sequence's methylation data.
2. The following sequences have a conversion rate below 90%: [7]: 87%.This implies a high risk of over-estimating the methylation state of such sequences.

How to continue:

1. It is safest to exclude the low-identity sequence(s). You can do so by selecting 'Exclude' for the corresponding sequence(s) on the left and then pressing the 'Recalculate' button

Figure 19. BiQ Analyzer provides a user-friendly interface and automatic expert advice to support analysis and quality control of DNA methylation data obtained by bisulfite sequencing

This figure shows a typical BiQ Analyzer screenshot. The text boxes on the left contain the raw sequences. The main window on the right displays a ClustalW multiple sequence alignment with CpGs, unconverted Cs, and critical sequences being highlighted. The text box below guides the user through the program and provides hints regarding sequences with quality problems. The bar at the top displays the current status and accepts some general settings. Finally, the button bar at the bottom enables the user to navigate through the different steps.

In summary, BiQ Analyzer provides start-to-end support for visualization and quality control of DNA methylation data from bisulfite sequencing. For the frequent user of bisulfite sequencing it will lead to significant speed-up of the data analysis process. The occasional user will benefit from extensive hints that help to perform rigorous quality control. Beyond that, BiQ Analyzer promises to be a first step toward standardization in quality control and documentation. Non-commercial users can download BiQ Analyzer free of charge from http://biq-analyzer.bioinf.mpi-inf.mpg.de/, commercial licenses are available through Max-Planck-Innovation (http://www.max-planck-innovation.de/de/industrie/technologieangebote/software/article.php?id=2662).

C-3 Insights from computational analysis of high-resolution DNA methylation data[1]

Although several DNA methylation maps covering significant parts of the human genome have been published previously (Eckhardt et al. 2006; Rakyan et al. 2004; Rollins et al. 2006; Weber et al. 2005; Weber et al. 2007; Yamada et al. 2004), each study made major simplifications. Weber et al. used MeDIP analysis, which is principally limited to a resolution of approximately 100 bp; Rollins et al. and Yamada et al. used restriction enzymes, which can assess DNA methylation only at a small subset of CpG dinucleotides fulfilling specific DNA sequence constraints; and Rakyan et al. as well as Eckhardt et al. used direct sequencing of bisulfite-converted DNA, which experimentally averages out DNA methylation patterns that are specific to individual cells or alleles (see chapter C-4 of this thesis for a more detailed discussion of different experimental methods for DNA methylation mapping).

The only experimental method that can overcome all of these limitations is clonal bisulfite sequencing (Frommer et al. 1992; Hajkova et al. 2002), which many researchers regard as the gold standard for DNA methylation analysis. However, clonal bisulfite sequencing is costly and labor-intensive, which is why its use has so far been restricted to small-scale studies. The goal of the German National Methylome Project for Chromosome 21 (NAME 21) is to show that genome-scale analysis of DNA methylation by clonal bisulfite sequencing is feasible and leads to new insights into epigenetic gene regulation in normal and diseased cells. Funded as part of the German National Genome Research Network (NGFN-2), the NAME-21 project comprises four collaborating groups: Albert Jeltsch's group at Jacobs University (Bremen, Germany), Jörn Walter's group at Saarland University (Saarbrücken, Germany), Richard Reinhardt's group at the Max Planck Institute for Molecular Genetics (Berlin, Germany) and Matthias Platzer's group at the Leibniz Institute for Age Research – Fritz Lipmann Institute (Jena, Germany). Within the scope of the NAME-21 project, the promoter regions of all protein-coding genes on chromosome 21 were analyzed in five different cell types, including two types of primary tissue, two cancer cell lines and a trisomic-21 fibroblast cell line (derived from a Down syndrome patient). Here we report statistical and bioinformatic analysis of the resulting dataset of high-resolution DNA methylation profiles for chromosome 21.

C-3.1 Methods

DNA methylation dataset

The DNA methylation dataset analyzed in this study covers the promoter regions of all 189 protein-coding genes on chromosome 21, according to annotation data for the hg17 assembly

[1] This chapter describes work conducted in collaboration with Sascha Tierling, Christian Rohde, Yingying Zhang, Diana Santacruz, Martina Paulsen and Jörn Walter (Tierling et al., in preparation). Sascha Tierling, Christian Rohde, Yingying Zhang and Diana Santacruz performed and evaluated the bisulfite sequencing experiments. Martina Paulsen and Jörn Walter contributed to the interpretation of the results.

of the human genome (NCBI35). Where possible, amplicons were placed such that they overlap with the gene's annotated transcription start site, and for many genes several amplicons were analyzed. In total, the German National Methylome Project for Chromosome 21 dataset contains DNA methylation profiles derived by clonal bisulfite sequencing of 289 amplicons, each analyzed in five different cell types: (i) primary blood and (ii) primary fibroblasts, both obtained from healthy adults, (iii) the human hepatocellular liver carcinoma cell line HepG2, (iv) the human embryonic kidney cell line HEK293 and (v) a fibroblast cell line that was derived from a trisomy-21 patient. All raw data were initially processed with BiQ Analyzer (Bock et al. 2005, cf. chapter C-2 of this thesis) by the person who performed the experiment, in order to enforce consistent quality control and to derive DNA methylation patterns. A total number of 27,069 clones passed quality control and were included in the analysis.

Bioinformatic analysis

Statistical analysis was performed using the R statistics software (www.r-project.org/) and included the use of hierarchical clustering, linear models, several curve-fitting algorithms and boxplots for visualization. Specifically, we applied hierarchical clustering with complete linkage and Euclidian distance to derive Figure 25, R's *loess* function with default parameters to fit local polynomial regression curves in Figure 21 to Figure 23, and a custom script to calculate unweighted moving averages based on a window of size 11 bp that is centered on the selected position in Figure 21. The direction of transcription was taken into account when plotting DNA methylation profiles around annotated transcription start sites (Figure 22) as well as around experimentally determined transcription initiation events (data not shown).

DNA methylation data was compared to several public datasets: (i) Manually curated RefSeq gene annotations – maintained by the National Center for Biotechnology Information (Pruitt et al. 2007) – were obtained from the UCSC Genome Browser (Karolchik et al. 2008). (ii) Experimentally determined transcription initiation events – based on a recent study sequencing 5' ends of full-length cDNAs and mapping these "CAGE tags" back to the human genome – were downloaded from the supplementary website of the paper describing this dataset (Carninci et al. 2006); (iii) Genome-wide binding site predictions for the NRSF transcription factor – calculated using the TRANSFAC 7.0 database and a custom software developed at UCSC (see http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=tfbsConsSites for details) – were obtained from the UCSC Genome Browser (Karolchik et al. 2008); (iv) Genome-wide predictions of CTCF binding sites – based on a refined consensus binding motif for the transcription factor CTCF that has been reported recently – were downloaded from the supplementary website of the corresponding paper (Kim et al. 2007).

Potential correlations between DNA methylation and the genomic DNA sequence were assessed with WebLogo (Crooks et al. 2004) and EpiGRAPH (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis). To prepare for WebLogo analysis, DNA methylation levels for each CpG position were averaged over all tissues and clones, and the CpGs were classified into high methylation (≥80%), moderate methylation (20% to 80%) and low methylation (≤20%). Next, we used EpiGRAPH to retrieve 22 bp of genomic DNA sequence centered on each CpG position. The resulting sets of DNA sequences were submitted to the WebLogo web server, which generated DNA sequence logos visualizing any position-specific bias in the DNA sequence left and right of CpGs that exhibit high, moderate or low DNA methylation levels. For EpiGRAPH analysis, average levels of DNA methylation were calculated for each amplicon, and a list containing the top-20% most highly methylated amplicons (positives) as well as the bottom-20% amplicons with lowest levels of DNA methylation (negatives) was submitted to the EpiGRAPH web service (hg18 version, with all default attributes included). Furthermore, we defined a measure of tissue-specific DNA methylation, which is calculated for each amplicon as the average root-mean-square deviation between all pairs of DNA methylation profiles originating from different tissues. Again, the top-20% amplicons with highest values were regarded as positives and the bottom-20% amplicons with lowest values as negatives, and the resulting list was submitted to the EpiGRAPH web service for statistical analysis and prediction.

C-3.2 Results

The key advantage of the current dataset over existing DNA methylation maps lies in its use of clonal bisulfite sequencing. Because all bisulfite-modified DNA fragments were cloned into vectors prior to PCR and sequencing, each sequence represents a single allele from a single cell, rather than an average over large and potentially heterogeneous cell populations, as is the case for other commonly used methods for DNA methylation mapping.

We were thus able to test and confirm that DNA methylation is distributed bimodally not only at the amplicon level and CpG dinucleotide level (which has been shown previously by Rakyan et al. 2004), but also at the level of individual clones (Figure 20). More specifically, in each single cell promoter regions of genes are likely to be either fully unmethylated (less than 20% methylation) or fully methylated (more than 80% methylation), while each intermediate step is less likely than the two extremes. When summing over the frequencies of all intermediate levels of DNA methylation (20% to 80%), intermediate promoter methylation becomes more frequent than full methylation, but absence of DNA methylation remains the most frequent state.

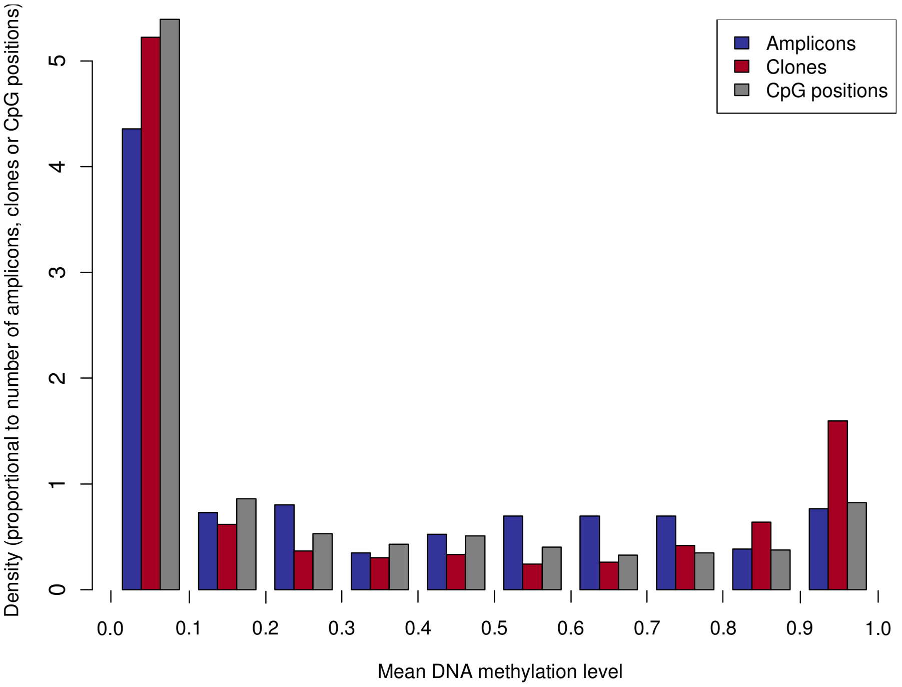


Figure 20. DNA methylation is distributed bimodally at the levels of amplicons, clones and CpG positions

This figure displays histograms of average DNA methylation levels for all analyzed amplicons (blue), clones (red) and individual CpG positions (grey), based on DNA methylation data for five cell types included in the dataset. To display these distributions within a single diagram, the *y*-axis plots normalized densities rather than frequency count values.

Next, we assessed how well the DNA methylation states correlate between any two CpGs within the same clone (i.e. originating from the same allele in the same cell). This question has high practical relevance because several methods for DNA methylation mapping – including all restriction-enzyme based protocols – rely upon the casual observation that neighboring CpGs are frequently co-methylated. We here report the first large-scale confirmation of this hypothesis, observing that two CpGs within the same clone are co-methylated with frequencies close to 90% over distances of up to 300 bp (Figure 21).

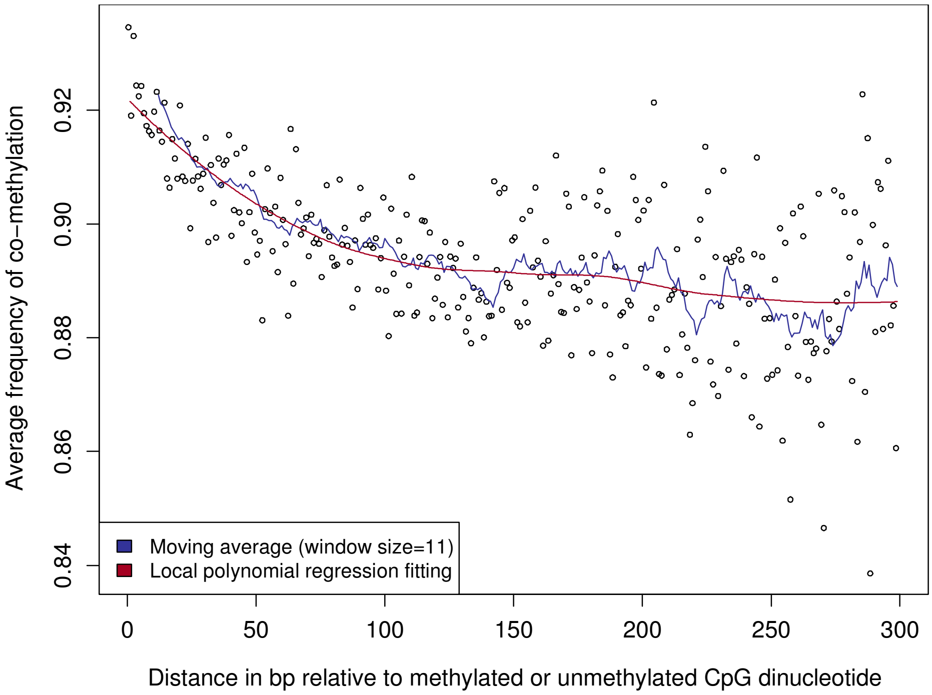
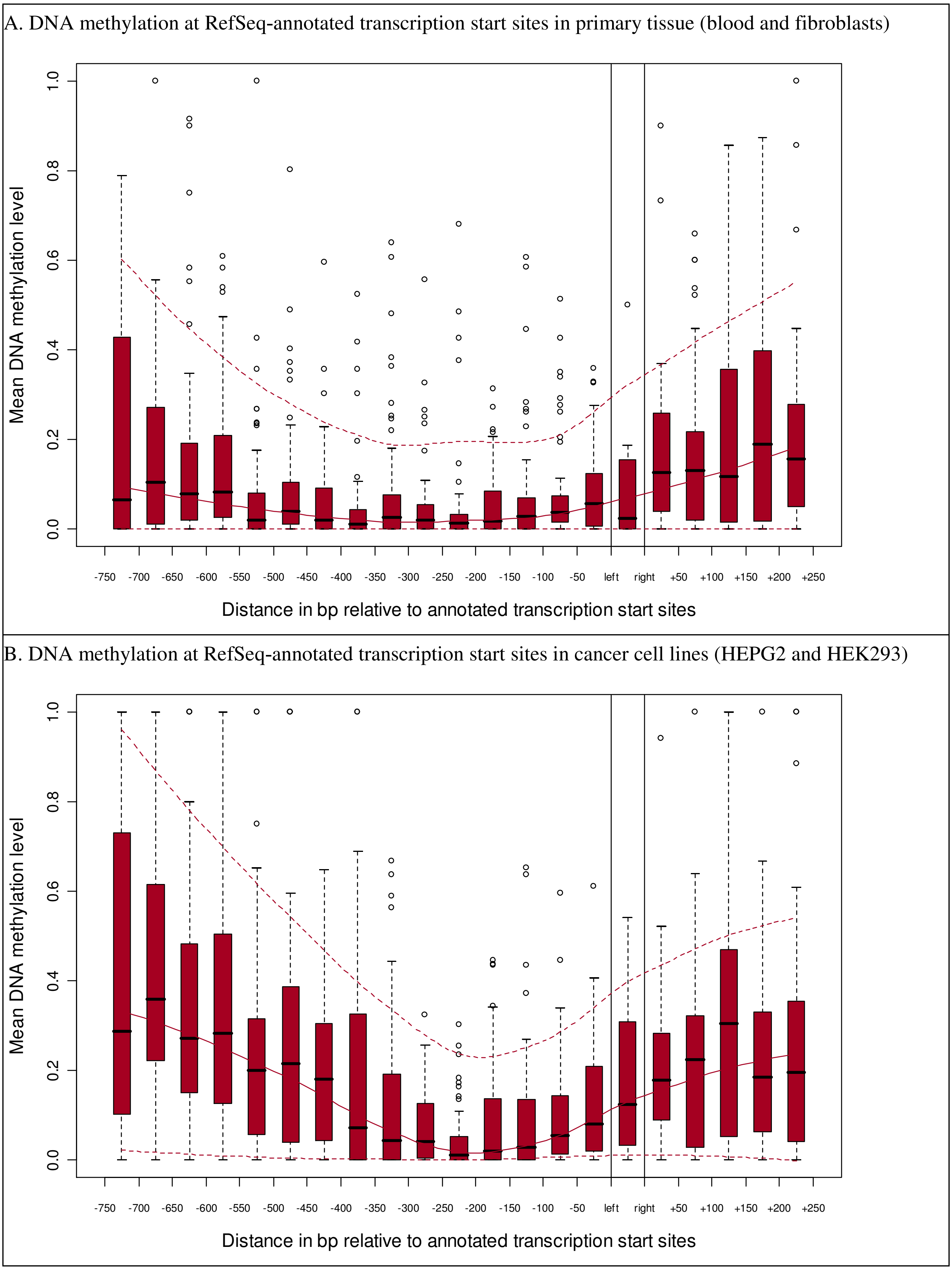


Figure 21. CpGs within the same clone are frequently co-methylated

This figure displays the frequency of co-methylation between two CpGs situated in the same clone at a specific distance, averaging over all pairs of CpGs in all clones included in the dataset. Similar results were obtained when comparing average DNA methylation levels within the same amplicon, rather than binary methylation states within the same clone (data not shown).

Absence of DNA methylation is a well-known hallmark of core promoters at active and temporarily silent genes, while high levels of DNA methylation at the transcription start site are linked to mitotically heritable silencing of the corresponding gene (Bird 2002). To investigate the spatial distribution of methylated and unmethylated CpGs around transcription start sites in normal tissue, we overlaid all DNA methylation patterns from blood and fibroblasts according to the location of the transcription start site inside the amplicons (Figure 22A). Technically, this procedure was based on RefSeq-annotated transcription start sites, which were mapped on top of each other in a strand-specific manner, i.e. with the same direction of transcription.



A. DNA methylation at RefSeq-annotated transcription start sites in primary tissue (blood and fibroblasts)

B. DNA methylation at RefSeq-annotated transcription start sites in cancer cell lines (HEPG2 and HEK293)

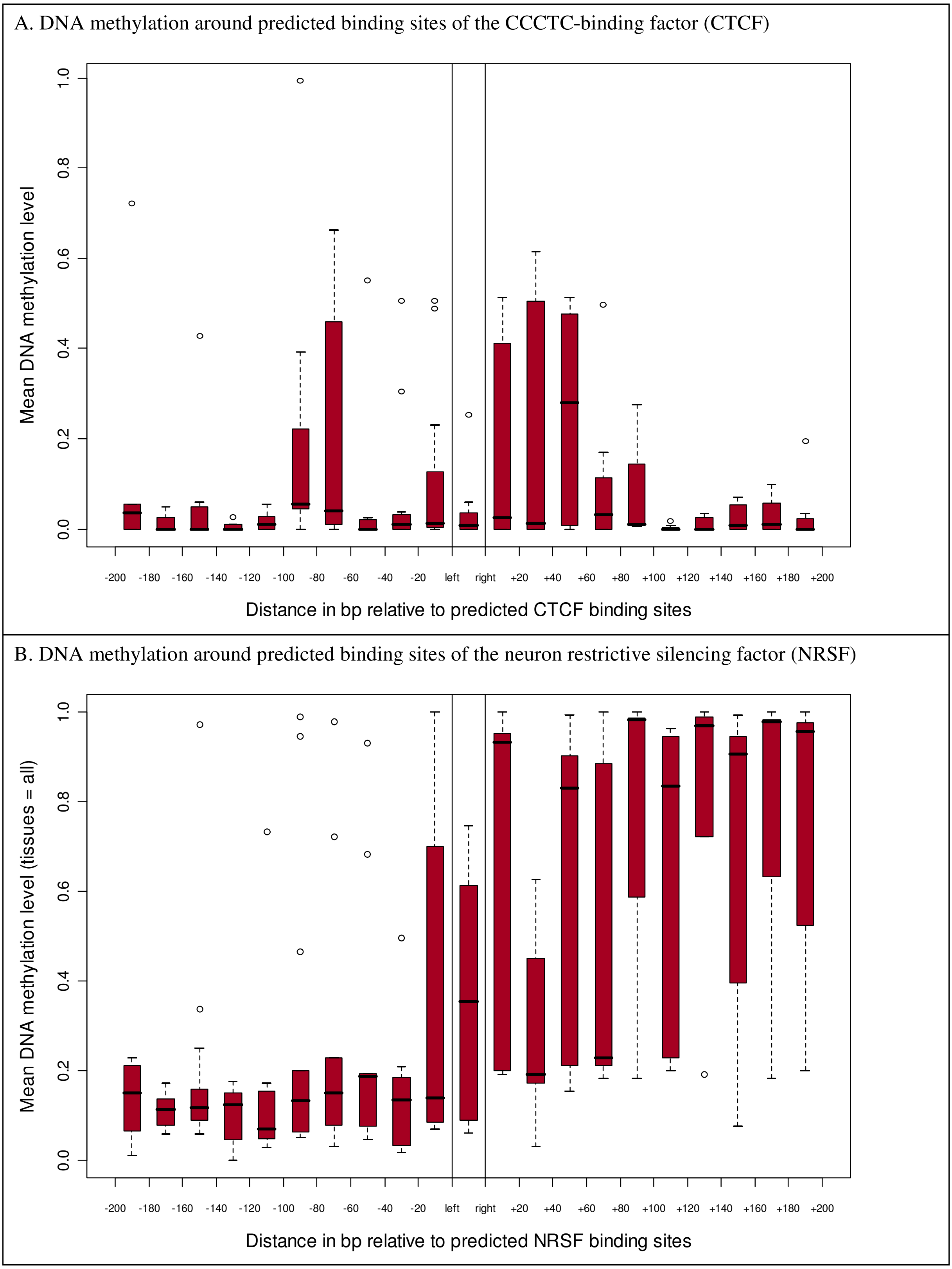
Figure 22. RefSeq-annotated genes exhibit a window of unmethylated DNA upstream of the transcription start site, which is substantially smaller in cancer cell lines than in normal tissue

This figure displays the average DNA methylation level in the genomic neighborhood of annotated transcription start sites, separately for the two normal cell types (panel A) and for the two cancer cell lines (panel B) included in the dataset. DNA methylation levels surrounding a core promoter of -5 bp (left) to +5 bp (right) – relative to the transcription start site of all RefSeq genes on chromosome 21 – were overlaid in a strand-specific manner. The boxplots are in standard format (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box), the solid curve is fitted to the median values and the dashed curves are fitted to the 90% and 10% percentiles, respectively, using local polynomial regression.

We observe low levels of DNA methylation in the entire genomic region surrounding transcription start sites and a particularly pronounced window of unmethylated DNA upstream of protein-coding genes, peaking around -200 bp to -250 bp relative to the RefSeq-annotated transcription start site. Higher levels of DNA methylation are observed further upstream and downstream, but also directly at the annotated transcription start site. For cancer cell lines, a similar distribution of DNA methylation is observed, with the exception that the average level of DNA methylation is substantially elevated (Figure 22B) and that the unmethylated window is restricted to a core region ranging from -200 bp to -250 bp relative to the annotated transcription start site.

While it is known that relatively small windows of unmethylated DNA can be sufficient to maintain active transcription (Brinkman et al. 2007; Wong et al. 2006), finding the most distinctively unmethylated region consistently upstream of the transcription start site, rather than overlapping with it, was unexpected. A potential explanation for this finding is based on the observation that current gene annotations frequently omit, or inappropriately truncate, 5' UTRs (Carninci et al. 2006; ENCODE Project Consortium 2007; Kapranov et al. 2005), which is due to the fact that experimental methods for fast and accurate mapping of transcription start sites have emerged only recently. If this explanation is true and the asymmetrical distribution of DNA methylation is an artifact of inaccurate genome annotation, we would expect a significantly different distribution of DNA methylation around unbiased transcription start sites. To test this hypothesis, we obtained a large dataset of in vivo transcription initiation events that were experimentally determined by 5' end sequencing of full-length cDNA sequences (Carninci et al. 2006), and we compared their genomic location to our DNA methylation data. Initially, we performed a similar analysis as reported in Figure 22 on the genomic location of empirical transcription initiation events. The results indicate that DNA methylation profiles are symmetrical around experimentally determined transcription start sites (data not shown), which provides some support for the hypothesis that the asymmetrical pattern observed in Figure 22 is an artifact of gene annotation bias. Next, we tested whether the distribution of transcription initiation events on chromosome 21 is similarly biased toward upstream regions relative to annotated transcription start sites, which would confirm an explanation based on gene annotation bias. However, we observed a highly symmetrical distribution of experimentally determined transcription start sites relative to the RefSeq-annotated transcription start sites on chromosome 21 (data not shown). Hence, a conclusive explanation of the observed asymmetrical distribution of DNA methylation around annotated transcription start sites will require further analysis, preferably on genome-wide DNA methylation datasets.

Next, we searched for correlative evidence of cross-talk between DNA methylation and the binding patterns of transcription factors with a known role in epigenetic gene regulation. We focused our analysis on two transcription factors for which highly predictive consensus binding motifs are available, for two reasons. First, empirical binding data (e.g. from ChIP-on-chip experiments) do not provide the single-basepair resolution that is required for high-resolution comparison with DNA methylation patterns. Second, the majority of known consensus binding motifs lack specificity for in vivo binding (Gomez-Skarmeta et al. 2006). Due to these practical limitations, we restricted our analysis to two specific transcription factors – the CCCTC-binding factor (CTCF) and the neuron restrictive silencing factor (NRSF) – for which genome-wide experimental analysis recently confirmed consensus binding motifs that are highly predictive of in vivo binding: GTGGCCACCAGGGGGCGCCG for CTCF (Kim et al. 2007) and TTCAGCACCACGGACAGCGCC for NRSF (Johnson et al. 2007).



A. DNA methylation around predicted binding sites of the CCCTC-binding factor (CTCF)

B. DNA methylation around predicted binding sites of the neuron restrictive silencing factor (NRSF)

Figure 23. Predicted binding sites of two selected transcription factors with a role in epigenetic gene regulation show characteristic DNA methylation profiles

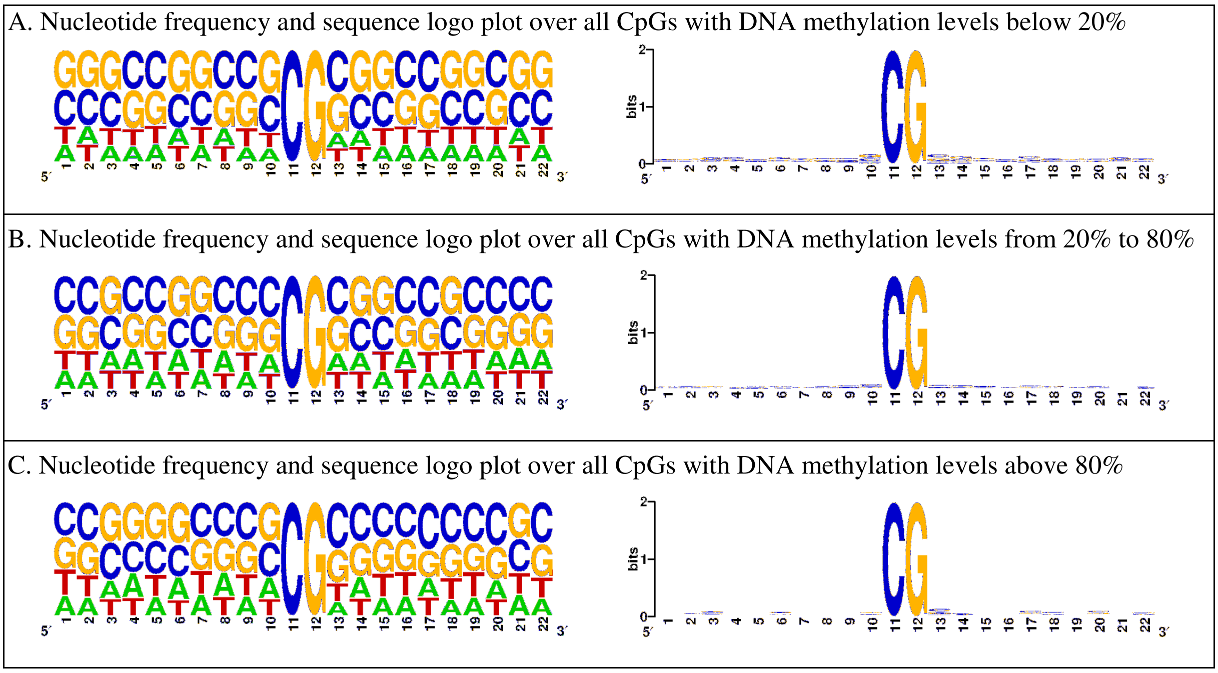
This figure displays average DNA methylation levels in the genomic neighborhood of transcription factor binding sites identified by bioinformatic scanning for known consensus motifs. Putative binding sites for CTCF (panel A) were obtained from a recent paper reporting an improved consensus motif based on genome-wide mapping (Kim et al. 2007). For NRSF (panel B), which is also known as repressor element-1 silencing transcription factor (REST), binding site predictions were obtained from the UCSC Genome Browser. The diagram format follows Figure 22.

Furthermore, both CTCF and NRSF are known to be involved in epigenetic gene regulation. CTCF is believed to function as an insulator, separating genomic compartments that are regulated differently and limiting the spreading of heterochromatin. Comparison with DNA methylation data shows that predicted CTCF binding sites are largely unmethylated (Figure 23A), consistent with an earlier observation that DNA methylation abolishes CTCF binding (Hark et al. 2000). However, from our dataset we observe no evidence that CTCF binding sites would preferentially locate at boundaries between genomic regions with high and low levels of DNA methylation. NRSF, which is also known as repressor element-1 silencing transcription factor (REST), is a transcriptional repressor that induces epigenetic gene silencing by interaction with the corepressor complexes CoREST and mSin3 (Ooi and Wood 2007). Comparison with DNA methylation data shows a striking difference between the DNA methylation patterns upstream and downstream of the NRSF binding site (Figure 23B), i.e. the kind of patterns that one might expect from an epigenetic insulator element.

Having assessed the potential role of two specific types of genomic elements – transcription start sites and transcription factor binding sites – for the distribution of DNA methylation, we next focused on the impact of the genomic DNA sequence itself. There are (at least) two different ways in which DNA methylation could be correlated with the DNA sequence. First, specific consensus sequences that incorporate CpGs at specific positions could make these CpGs more or less prone to DNA methylation, acting in a position-specific fashion ("position-specific" implies that shuffling of the consensus sequence would disable the effect). Second, specific sequence motifs could contribute cumulatively to an unmethylated state of several CpGs in their vicinity, without requiring specific positioning relative to any particular CpG.

To test for a position-specific effect, which has been reported in a recent study (Handa and Jeltsch 2005), we first classified all CpGs in our dataset into low, medium or high levels of DNA methylation and retrieved their surrounding genomic DNA sequences. Next, nucleotide frequency plots (Figure 24, left column) and DNA sequence logos (Figure 24, right column) were generated separately for the three classes of CpGs using the WebLogo software (Crooks et al. 2004). These plots show that the nucleotide frequency distribution is similar at all positions left and right of the central CpG (Figure 24, left column) and that its information content is close to zero (Figure 24, right column), arguing against a major in vivo role of position-specific effects on the DNA methylation status of individual CpGs. In particular, we could not replicate the DNA sequence motifs for high (CTTGCGCAAG) and low (TGTTCGGTGG) DNA methylation levels that have been reported previously (Handa and Jeltsch 2005).

In contrast, there is a clear tendency toward higher levels of GC content in the vicinity of unmethylated CpGs (Figure 24A), as compared to methylated CpGs (Figure 24C), consistent with a cumulative, non-position-specific effect of GC content on DNA methylation propensity. To assess cumulative effects more systematically, we derived a list of confidently methylated and confidently unmethylated amplicons (see Methods section for details) and used the EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis) to test which groups of genomic attributes are predictive of whether or not an amplicon is confidently unmethylated. While the prediction performance based on non-position-specific 1-mer, 2-mer and 4-mer DNA sequence motives is relatively moderate (correlation: 0.41, accuracy: 74.2%), the inclusion of other genomic attributes results in high prediction accuracies. In particular, bona fide CpG island predictions – derived previously to predict open vs. condensed chromatin structure (Bock et al. 2007, cf. chapter B-4 of this thesis) – and ChIP-seq data for multiple histone modifications (Barski et al. 2007) are highly discriminatory between amplicons with high vs. low levels of DNA methylation. Based on attributes relating to sequence-based annotations of regulatory regions (such as bona fide CpG island predictions and transcription factor binding site predictions), EpiGRAPH achieves similar prediction performance (correlation: 0.82, accuracy: 91.8%) as reported previously (Bock et al. 2006, cf. chapter B-2 of this thesis), and based on experimental data on chromatin structure in unrelated human blood samples, EpiGRAPH's prediction performance is close to optimal (correlation: 0.91, accuracy: 96.0%).


A. Nucleotide frequency and sequence logo plot over all CpGs with DNA methylation levels below 20%

B. Nucleotide frequency and sequence logo plot over all CpGs with DNA methylation levels from 20% to 80%

C. Nucleotide frequency and sequence logo plot over all CpGs with DNA methylation levels above 80%

Figure 24. Methylated and unmethylated CpGs are not preferentially located in position-specific consensus motifs

This figure displays nucleotide frequency plots (left) as well as DNA sequence logos (right) of the DNA sequence neighborhood of preferentially unmethylated CpGs (panel A), intermediate CpGs (panel B) and preferentially methylated CpGs (panel C). While the frequency plots visualize the percentage with which a specific base occurs at a specific position left or right of a CpG with known methylation level, the sequence logos plot the information content in bits, which measures how strongly the base distribution at a specific position deviates from equal base frequency. Plots were generated using WebLogo (Crooks et al. 2004) and DNA methylation data from all five cell types were included in the analysis.

In the next step, we analyzed the dataset for evidence of tissue-specific DNA methylation. We clustered all five cell types according to their average DNA methylation levels in all amplicons (Figure 25). The resulting cluster tree shows that promoter DNA methylation levels are highly correlated between related cell types, in particular between the two cancer cell lines and between the two types of primary tissue, blood and fibroblasts. Interestingly, DNA methylation in normal fibroblasts and in a trisomic-21 fibroblast cell line (derived from a Down syndrome patient) are highly similar, arguing against a major role of DNA methylation in Down syndrome. To corroborate this finding, we investigated whether other chromatin modifications might play a role in determining which genes are expressed at the expected 1.5-fold level in trisomic-21 cells and which genes undergo compensatory effects. We obtained a list of 49 unique transcripts showing significant compensation of gene expression and 28 unique transcripts showing approximately 1.5-fold or higher expression in trisomic-21 cells, based on a recent study that analyzed gene expression in ten Down syndrome patients and in eleven controls (Yahya-Graison et al. 2007). The promoter regions of these genes were analyzed with

EpiGRAPH, and although EpiGRAPH reported above-random performance when predicting which genes are subject to compensation and which are not (correlation: 0.30, accuracy: 68.1%, based on all default attributes), no single attribute was significantly different between the two classes after correction for multiple testing (data not shown).

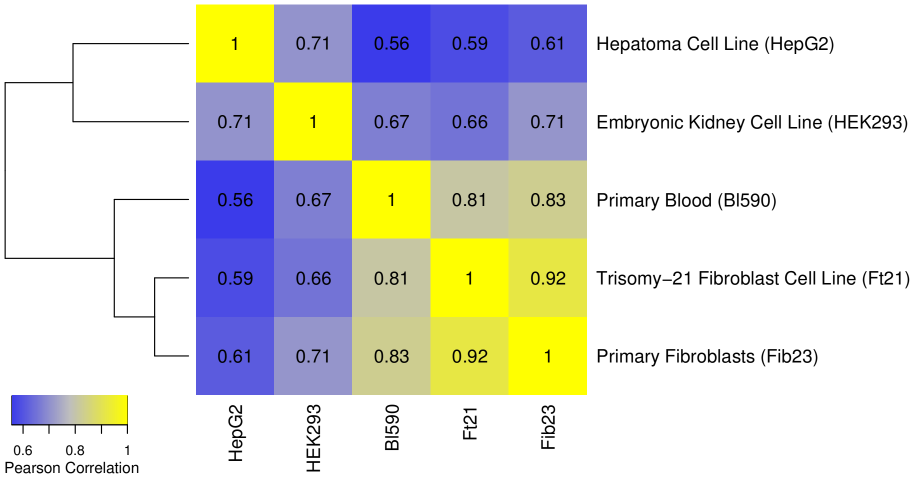


Figure 25. DNA methylation is highly correlated between related tissues

This figure displays a clustered heatmap summarizing the similarity between amplicon methylation levels of all five cell types included in the analysis. The numerical values displayed in the heatmap are pairwise Pearson correlation coefficients between each pair of cell types, calculated on the average DNA methylation levels for all amplicons with sufficient data.

Finally, we asked whether we could predict from characteristics of the genome sequence which amplicons are prone to tissue-specific DNA methylation and which are not. To that end, we calculated the degree of tissue-specific methylation for each amplicon in the dataset and derived a list of confidently tissue-specific and confidently tissue-invariant amplicons (see Methods for details). This list was analyzed with EpiGRAPH, and the results show that amplicons exhibiting highly tissue-specific DNA methylation are less CpG-rich, depleted in transcription-related chromatin modifications (e.g. histone H3K4 methylation and RNA polymerase II binding) and less likely to be associated with expressed genes than tissue-invariant amplicons. In summary, EpiGRAPH could predict tissue specificity of DNA methylation with high accuracy based on its default attributes (correlation: 0.76, accuracy: 89.3%).

C-3.3 Discussion

This chapter described the results of bioinformatic analysis performed on a large-scale DNA methylation dataset, which originates from the NAME-21 project. The novel feature of this dataset is that it provides single-cell resolution as well as single-basepair resolution for a sizable fraction of human promoter regions, allowing us to test several hypotheses that are difficult to address on DNA methylation maps derived with experimental methods other than clonal bisulfite sequencing.

First, we could show that the DNA methylation state of single CpGs is highly predictive of presence or absence of DNA methylation at nearby CpGs on the same allele in the same cell. This observation has important practical implications because it confirms the underlying assumption of restriction-enzyme based methods for DNA methylation analysis (Huang et al. 1999; Khulan et al. 2006; Pfister et al. 2007; Rollins et al. 2006): When these methods are used, it is usually assumed that the subset of CpGs accessible by restriction enzymes is sufficiently representative of the overall DNA methylation states in a given genomic region.

Second, overlaying DNA methylation patterns relative to annotated transcription start sites gave rise to high-resolution profiles of DNA methylation around an average transcription start site. Consistent with current knowledge (Bird 2002), we found that DNA methylation levels are low close to the transcription start site (in normal cells more so than in cancer cells) and start to increase several hundred basepairs upstream and downstream of the transcription start site, respectively. Unexpectedly, however, we observed that the region of minimal methylation did not directly overlap with the transcription start site, but was located on average 200 bp to 250 bp upstream (Figure 22). As the distribution of DNA methylation relative to the location of empirical transcription initiation events was found to be symmetrical (data not shown), a potential explanation for our observation could be an inherent tendency of the RefSeq gene annotations to place the transcription start site too far downstream. However, a direct comparison between annotated transcription start sites and empirical transcription initiation events did not corroborate this explanation, potentially because genes with low levels of expression were under-represented in the empirical dataset. Further research is clearly warranted given the potential utility of DNA methylation for improving gene annotation in the human genome (transcription start site annotation based on absence of DNA methylation could overcome two limitations of current methods: First, it is applicable to genes with low expression rates, which mRNA sequencing often misses; second, in contrast to histone methylation and acetylation (Trinklein et al. 2007), DNA methylation can be measured at single-basepair resolution).

Third, we overlaid DNA methylation patterns relative to predicted transcription factor binding sites for CTCF and NRSF and obtained results that are somewhat difficult to interpret. In contrast to expectations, we observed a pronounced boundary between regions of low and high methylation coincident with NRSF binding sites, while no such difference was apparent at binding sites of the putative insulator protein CTCF. This result awaits validation by genome-wide DNA methylation datasets, but could potentially indicate the need to reappraise the roles of CTCF and NRSF for the genomic distribution of DNA methylation.

Fourth, we reassessed the predictiveness of DNA-related attributes for DNA methylation at the amplicon level and could confirm previous results (cf. chapter B-2 and chapter B-4 of this thesis). In contrast, our results did not provide evidence for the existence of a position-specific consensus DNA sequence motif that induces or abolishes DNA methylation at individual CpGs. These observations are consistent with a model in which DNA sequence patterns contribute to the DNA methylation state in ways that are non-position-specific and cumulative, rather position-specific and combinatorial (cf. section E-2.1 for further discussion).

Finally, because the current dataset comprises DNA methylation profiles for five different tissues we were able to address several aspects of tissue specificity. In particular, we found that EpiGRAPH can accurately distinguish between amplicons exhibiting highly tissue-specific levels of DNA methylation on the one hand and amplicons exhibiting highly stable levels of DNA methylation on the other hand. However, EpiGRAPH was not able to predict with sufficient accuracy in which tissue(s) each tissue-specific region exhibits high vs. low levels of DNA methylation. Furthermore, we tested the hypothesis that DNA methylation might play a role as a compensatory mechanism of over-expression in trisomic-21 cells, thus contributing to observed heterogeneity among Down syndrome phenotypes. However, no clear-cut differences were apparent when comparing DNA methylation profiles of fibroblasts from a Down syndrome patient with those of a healthy control sample. While we cannot exclude subtle effects that become apparent only when analyzing large patient cohorts, it seems unlikely that DNA methylation plays a similarly prominent role in Down syndrome as it does for cancer.

C-4 Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping[1]

C-4.1 Motivation

Initial DNA methylation mapping projects have not only revealed a highly complex distribution of DNA methylation in the human genome, but also highlight the prevalence of inter-individual variation among DNA methylation profiles from different individuals (Murrell et al. 2005). As the impact of epigenetic polymorphisms on gene expression and phenotypic traits is well-established (Peaston and Whitelaw 2006; Wong et al. 2005), DNA methylation variation is potentially a major contributor to phenotypic variation in humans.

Better understanding of inter-individual variation of DNA methylation is desirable also from a practical point of view. First, it is critical to know the range of DNA methylation variation in healthy individuals in order to confidently detect aberrant methylation in diseased patients. This point is exemplified by the optimization of cancer biomarkers for robustness and precision, during which it is critical to select CpG dinucleotides that exhibit small amounts of inter-individual variation within the groups of cancer patients and controls, respectively, but strong variation between the two groups (Mikeska et al. 2007, cf. chapter D-3 of this thesis). Second, large-scale DNA methylation mapping initiatives profit from robust estimates of DNA methylation variation, because such estimates provide a basis for rational choice of required sample sizes, selection of appropriate experimental methods and identification of genomic regions that require particular depth of analysis.

In this chapter, we quantitatively analyze the characteristics and determinants of DNA methylation variation among healthy individuals, based on a large-scale and high-resolution dataset originating from the Human Epigenome Project of the Sanger Institute and Epigenomics AG (Eckhardt et al. 2006). Furthermore, we use a combination of computational modeling and simulation in order to benchmark current experimental methods for DNA methylation mapping and to guide the strategy for establishing mammalian reference methylomes. In particular, we ask whether the use of high-resolution mapping methods is required and informative for all parts of the human genome or whether cheaper medium-resolution methods may be sufficient at least for certain parts of the genome.

C-4.2 Methods

A dataset of high-resolution DNA methylation profiles from multiple individuals

As the basis for this study, we selected the Human Epigenome Project (HEP) dataset reported by Eckhardt et al. (Eckhardt et al. 2006), because it is the largest high-resolution, multi-individual dataset of DNA methylation profiles that has been published to date. Briefly, Eckhardt et al. combined direct Sanger sequencing of bisulfite-converted DNA with a bioinfor-

[1] This chapter describes published work conducted in collaboration with Martina Paulsen and Jörn Walter (Bock et al. 2008), who contributed to the interpretation of the results.

matic method for deriving quantitative and high-resolution DNA methylation profiles from chromatograms. With this strategy, they analyzed 2,524 amplicons on human chromosomes 6, 20 and 22 in 12 different tissues and 43 different samples. Ten samples belonging to three tissues (CD4+ T lymphocytes, CD8+ T lymphocytes and melanocytes) originate from single donors and were selected for our analysis, while the remaining samples are pooled DNA from several individuals, which makes them less suitable for analyzing inter-individual variation. For clarity, all results presented in this paper are based on CD4+ T lymphocyte data. Comparable results were obtained for the other two tissues as well as for the cross-tissue comparison of all ten samples (data not shown).

Raw data were downloaded from http://www.sanger.ac.uk/PostGenomics/epigenome/. To prepare for further analysis, the different record types ("analysis", "trace" and "variation") were merged by their identifiers, and records corresponding to technical controls were discarded. The analysis described here is based on the second technical replicate, which contains more valid data than the first technical replicate. Averaging of the two replicates was not an option because of incomplete overlap, but the analyses described in this paper were repeated on the first replicate and comparable results were obtained. Amplicons with insufficient data were removed, giving rise to a dataset of 1,705 amplicons. These amplicons are on average 287 basepairs long (first to last assessed CpG) with a standard deviation of 85 basepairs, and they comprise 16 CpGs, on average. The majority (58%) of the amplicons overlap with CpG islands according to the Gardiner-Garden criteria with repeat-masking and a quarter (25%) overlap with more stringent bona fide CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis).

Statistical analysis of DNA methylation variation among healthy individuals

Three complementary measures of inter-individual variation between DNA methylation profiles are used in this study (defined below): (i) pairwise deviation between high-resolution profiles, (ii) deviation between mean and high-resolution profile and (iii) pairwise deviation between means. These measures are calculated from pairwise comparisons between DNA methylation profiles of non-identical individuals, which are averaged separately for each amplicon in the dataset. In the pairwise comparisons, the root-mean-square deviation (RMSD) is used to assess deviations between DNA methylation profiles of different individuals. The RMSD is more appropriate for this purpose than the Pearson correlation coefficient (which is the other popular measure of similarity/deviation) because DNA methylation levels are frequently constant within an amplicon, in which case their standard deviation becomes zero and the correlation coefficient is undefined.

The *pairwise deviation between high-resolution profiles (v_1)* compares DNA methylation levels at every single CpG, thereby assessing how similar or different DNA methylation profiles from unrelated individuals are at a given amplicon. This measure is defined by the following formula (calculated separately for each amplicon), in which m is the number of samples from different individuals, P_i and P_j are sets of CpG positions with valid measurements for samples i and j, respectively, $x_{i,k}$ and $x_{j,k}$ are the methylation levels measured at position k in samples i and j, respectively, and n is the number of positions in common between P_i and P_j:

$$v_1 = \tfrac{1}{m\cdot(m-1)} \sum_{i=1}^{m} \sum_{j=1,j\neq i}^{m} \sqrt{\tfrac{1}{n} \sum_{k\in P_i \cap P_j} (x_{i,k} - x_{j,k})^2} \; .$$

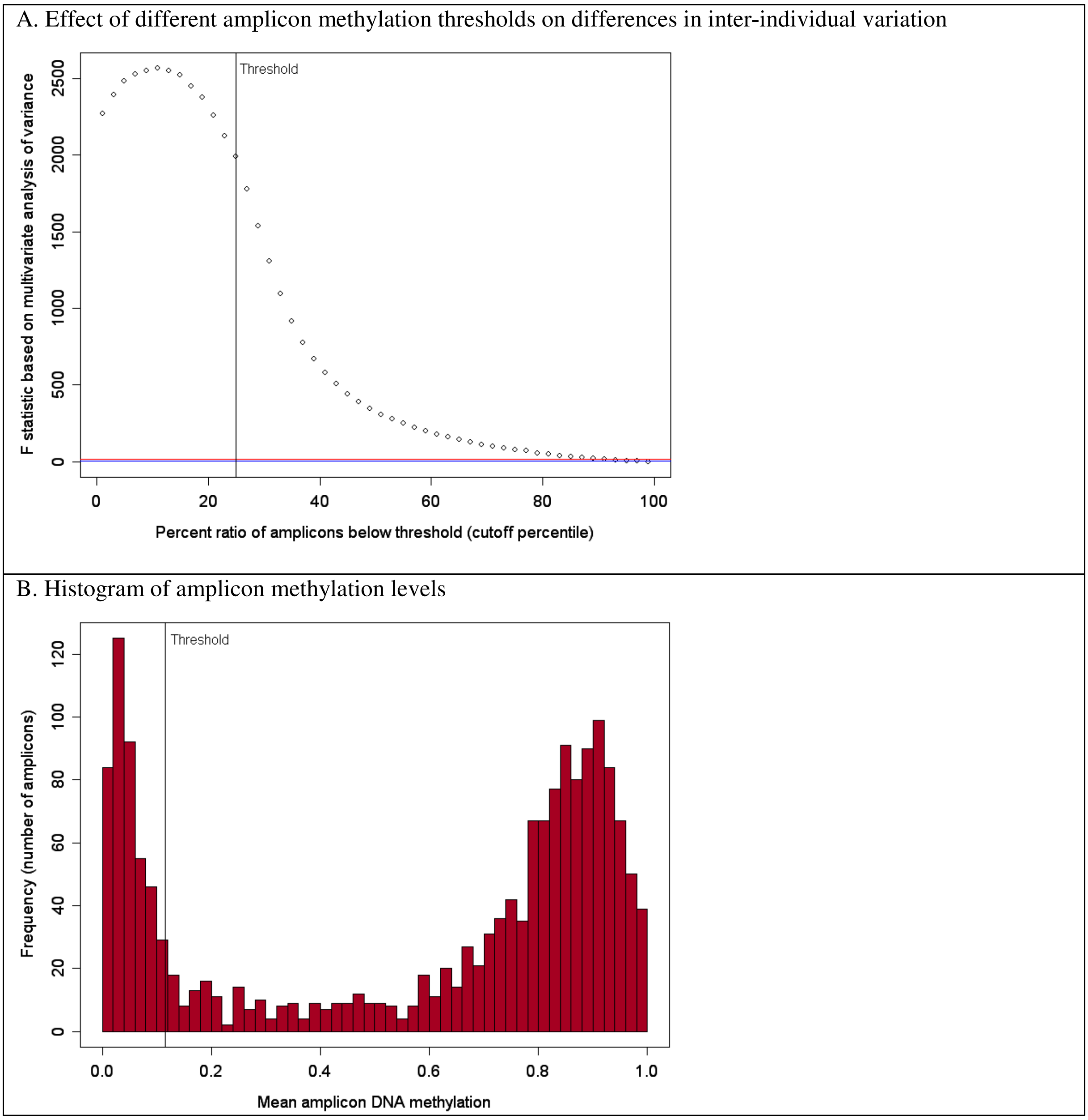
The *deviation between mean and high-resolution profile (v_2)* compares the mean methylation level of one individual to the high-resolution DNA methylation profile of other individuals, thereby assessing how predictive the mean amplicon methylation of one individual is for the DNA methylation profile of unrelated individuals. This measure is defined by the following formula, in which one DNA methylation profile is replaced by its mean ($\bar{x}_j = \tfrac{1}{|P_j|} \sum_{k\in P_j} x_{j,k}$):

$$v_2 = \tfrac{1}{m\cdot(m-1)} \sum_{i=1}^{m} \sum_{j=1,j\neq i}^{m} \sqrt{\tfrac{1}{n} \sum_{k\in P_i \cap P_j} (x_{i,k} - \bar{x}_j)^2} \; .$$

The *pairwise deviation between means (v_3)* compares the mean amplicon methylation levels between a set of individuals, ignoring the sequential order of methylated and unmethylated CpGs. It is defined by replacing the remaining DNA methylation profile in formula v_2 by its mean ($\bar{x}_i = \tfrac{1}{|P_i|} \sum_{k\in P_i} x_{i,k}$), giving rise to mean absolute differences between individuals:

$$v_3 = \tfrac{1}{m\cdot(m-1)} \sum_{i=1}^{m} \sum_{j=1,j\neq i}^{m} \sqrt{\tfrac{1}{n} \sum_{k\in P_i \cap P_j} (\bar{x}_i - \bar{x}_j)^2} = \tfrac{1}{m\cdot(m-1)} \sum_{i=1}^{m} \sum_{j=1,j\neq i}^{m} \left|\bar{x}_i - \bar{x}_j\right| .$$

These three measures of inter-individual variation were calculated for each valid amplicon in the HEP dataset. Based on these data, we sought to derive a threshold on DNA methylation that separates two distinct groups of amplicons – high vs. low methylation – such that the between-group differences in terms of the three measures of inter-individual variation are high. To that end, 99 potential thresholds were assessed (splitting the dataset after each integer percentile of DNA methylation levels) and the suitability of each was tested using multivariate analysis of variance (MANOVA). Briefly, MANOVA (Tabachnick and Fidell 2007) finds the linear combination of the three measures that is most discriminative between the two amplicon groups (i.e. between those amplicons exceeding the DNA methylation threshold and those falling below the threshold, respectively), and it assesses the significance of this difference. Figure 26A shows a plot of the corresponding *F* statistic and Figure 26B shows the DNA methylation histogram. Based on these two diagrams, it seems plausible to split the groups at the 25th percentile, corresponding to an 11.5% threshold on the amplicon's DNA methylation level. In terms of their characteristics of inter-individual variation, the amplicons below this threshold (first quartile or top-25% of most unmethylated amplicons) are strikingly distinct from the remaining amplicons ($P << 10^{-10}$).



A. Effect of different amplicon methylation thresholds on differences in inter-individual variation

B. Histogram of amplicon methylation levels

Figure 26. Informed selection of a threshold on amplicon DNA methylation

This figure summarizes a statistical analysis showing that the top-25% of most unmethylated amplicons (which corresponds to all amplicons with a mean methylation level of less than 11.5%) form a distinct group not only in terms of their DNA methylation levels (panel B), but also in terms of their characteristics of inter-individual variation (panel A). Panel A plots *F* statistic values for 99 different multivariate analyses of variance (MANOVA), each corresponding to a threshold on the mean amplicon methylation that splits the dataset at an integer percentile. MANOVA maximizes the differences in terms of inter-individual variation (measured by the dependent variables v_1, v_2 and v_3) between the two groups, i.e. between amplicons above and below the threshold. The blue and red lines at the bottom of this diagram correspond to *P*-values of 0.01 and 10^{-10}, respectively, indicating that all reasonable thresholds lead to highly significant differences between the two groups of amplicons. Panel B displays a histogram of mean amplicon DNA methylation. In both diagrams, the selected threshold (top-25% unmethylated amplicons, equivalent to an amplicon methylation level of less than 11.5%) is highlighted by a vertical line.

In silico benchmarking of experimental methods for DNA methylation mapping

To benchmark experimental methods for DNA methylation mapping, we introduce the following modification of the v_1 measure of inter-individual variation. The first DNA methylation profile is pre-processed by a function that simulates experimental analysis, and this pre-processed profile is compared to the second profile (all simulation functions are defined and explained in Table 11, and an illustrative example is given in Figure 27). This way, we can assess how well the simulated measurement for a particular experimental method predicts the high-resolution DNA methylation profile of other individuals. Formally, these new criteria v_{method} (one per simulation function) are defined by the following formula, in which x_i is the vector of methylation levels of sample i at positions P_i, f_{method} is the simulation function and *method* is any of the identifiers in Table 11, rightmost column (i.e. A1 to G7):

$$v_{method} = \frac{1}{m \cdot (m-1)} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} \sqrt{\frac{1}{n} \sum_{k \in P_i \cap P_j} (f_{method}(x_i)_k - x_{j,k})^2}$$

While this formula is a straightforward extension of the described measures of inter-individual variation, the choice of simulation functions in Table 11 may warrant further discussion. Each simulation function computes the expected DNA methylation measurement for a specific experimental method on an amplicon with known methylation profile. The simulation functions are modeled after the known mechanisms and experimental constraints of the underlying experimental methods. Different variants of the same method are included in order to assess the sensitivity to different experimental conditions. When constructing these simulation functions, we relied not only on our own practical experience and on literature research, but also consulted with domain experts within the EU Network of Excellence "The Epigenome", in order to verify that the design of the rules and the choice of parameters were appropriately modeling the experiment under optimal conditions. In addition, we validated that the sensitivity of our results to the choice of parameters was generally low (see Results section). Therefore, while a comprehensive empirical evaluation is not feasible in the absence of a large-scale experimental benchmarking dataset, we conclude that the rules in Table 11 are sufficiently accurate and reliable for the purposes of this study.

Input: A set of high-resolution DNA methylation profiles for the same region / amplicon, derived from unrelated individuals:

Sample_1 = (0.6, 0.8, 0.7, 0.9)

Sample_2 = (0.2, 0.7, 0.7, 0.6)

Sample_3 = (0.3, 0.8, 0.9, 0.5)

Method: For each of the experimental protocols considered, compare the rule-derived profile with the high-resolution profile:

- First example: Quantitative bisulfite sequencing (method F1) maintains high-resolution information

$Sample\_1_{F1}$ = (0.6, 0.8, 0.7, 0.9)

$Sample\_2_{F1}$ = (0.2, 0.7, 0.7, 0.6)

$Sample\_3_{F1}$ = (0.3, 0.8, 0.9, 0.4)

⇨

	$Sample\_1_{F1}$	$Sample\_2_{F1}$	$Sample\_3_{F1}$
Sample_1	-	-	-
Sample_2	0.255 (*)	-	-
Sample_3	0.269	0.132	-

(*) Calculated as follows: $RMSD = \sqrt{\frac{1}{4}[(0.6-0.2)^2+(0.8-0.7)^2+(0.7-0.7)^2+(0.9-0.6)]} = \sqrt{0.065} \approx 0.255$

- Second example: Quantitative immunoprecipitation (method D1) measures average DNA methylation levels

$Sample\_1_{D1}$ = (0.75, 0.75, 0.75, 0.75)

$Sample\_2_{D1}$ = (0.55, 0.55, 0.55, 0.55)

$Sample\_3_{D1}$ = (0.6, 0.6, 0.6, 0.6)

⇨

	$Sample\_1_{D1}$	$Sample\_2_{D1}$	$Sample\_3_{D1}$
Sample_1	-	-	-
Sample_2	0.287	-	-
Sample_3	0.269	0.250	-

Output: Average differences (RMSDs) over all amplicons in the HEP dataset, separately for each method and tissue type

Figure 27. Illustrative example of the computational benchmarking method

This figure displays exemplary in silico benchmarking for two methods (F1 and D1, see Table 11 for details) on a short amplicon with known methylation profile from three unrelated individuals. Computational rules are used to simulate which measurements the methods would report if applied experimentally. Scaled up to multiple amplicons, this analytical strategy can be used to benchmark how well different methods capture inter-individually stable patterns of DNA methylation.

Method name	References	Method type	Comment	Simulation function
Differential methylation hybridization (DMH)	(Huang et al. 1999) (Khulan et al. 2006) (Pfister et al. 2007)	Methylation-specific digestion, qualitative	Quantification is difficult due to different oligomer affinities and melting temperatures	A1: HiMeth if $\#CpG_{pattern* \,\&\, meth \geq 50\%} \geq 3$ A2: HiMeth if $\#CpG_{pattern* \,\&\, meth \geq 50\%} \geq 2$ A3: HiMeth if $\#CpG_{pattern* \,\&\, meth \geq 50\%} \geq 1$ * pattern in {ACGT, CCGC, CCGG, GCGC}
Sequencing of methylation-specific digestion products	(Rollins et al. 2006)	Methylation-specific digestion, quantitative	Quantification is possible if sequencing depth is high	B1: Profile(all CpGs in ACGT patterns) B2: Profile(all CpGs in CCGC patterns) B3: Profile(all CpGs in CCGG patterns) B4: Profile(all CpGs in GCGC patterns) B5: Profile(all CpGs in all four patterns)
Methyl-DNA immunoprecipitation plus tiling microarrays (MeDIP-chip)	(Weber et al. 2005) (Weber et al. 2007) (Zhang et al. 2006) (Zilberman et al. 2007)	Immunoprecipitation, qualitative	Quantification is difficult due to different oligomer affinities and melting temperatures	C1: HiMeth if $\#CpG_{meth \geq 67\%} \geq 4$* C2: HiMeth if $\#CpG_{meth \geq 50\%} \geq 3$* C3: HiMeth if $\#CpG_{meth \geq 33\%} \geq 2$* * minimum value per 200 bp
Sequencing of MeDIP-generated DNA libraries (MeDIP-seq)	Established at several labs, e.g. at the Max Planck Institute for Molecular Genetics (H. Lehrach, personal communication)	Immunoprecipitation, quantitative	Quantification is possible if the enrichment scores are statistically corrected for local differences in CpG density	D1: Value(Mean(all CpGs)) D2: Value(Median(all CpGs))
Microarray hybridization of bisulfite-converted DNA	(Adorjan et al. 2002) (Gitan et al. 2002) (Kimura et al. 2005) (Yan et al. 2004)	Bisulfite conversion, qualitative	Quantification has been attempted but is often unreliable	E1: HiMeth if mean(all CpGs) ≥ 67% E2: HiMeth if mean(all CpGs) ≥ 50% E3: HiMeth if mean(all CpGs) ≥ 33%
Direct sequencing of bisulfite-converted DNA	(Eckhardt et al. 2006) (Lewin et al. 2004) (Rakyan et al. 2004)	Bisulfite conversion, quantitative	Quantitative and applicable to either all CpGs of an amplicon (by Sanger sequencing) or to a subset (by primer extension or pyrosequencing)	F1: Profile(all CpGs) F2 to F5: Profile(1 to 4 random CpGs) F6: Profile(center CpG) F7: Profile(first and last CpG) F8: Profile(CpGs at positions ⅓ and ⅔) F9: Profile(first, center and last CpG) F10 to F20: Profile (0%, 10%, …, 100% of CpGs, rounded to the closest integer and randomly selected)
Rule-based guess (for comparison as a negative control)	None	No DNA methylation data is taken into account	Worst-case baseline that any method should compare favorably to	G1: Value(0% methylated) G2: Value(50% methylated) G3: Value(100% methylated) G4: Value(LowMeth) G5: Value(MeanMeth) G6: Value(HiMeth) G7: Profile(random methylation values)

Table 11. Functions for computational simulation of experimental methods for DNA methylation mapping

This table summarizes the experimental methods for DNA methylation mapping that are covered in this study, and it describes the functions that were constructed to simulate them in silico (rightmost column). The simulation functions are written in an abbreviated notation, as if-clauses, as profile statements or as value assignments. (i) For if-clause rules, a methylation constant named HiMeth is assigned to all CpGs in amplicons identified as high-methylation and a constant named LowMeth is assigned to all CpGs in low-methylation amplicons. We set HiMeth = 80.39% and LowMeth = 13.13%, which are the mean methylation levels of all amplicon that exceed or fall below 50% methylation, respectively, in the HEP dataset. (ii) For profile statements, a subset of CpGs that fulfill the condition in brackets are selected and the methylation values of all unselected CpGs are determined by interpolation or extrapolation. (iii) Value assignments are a special case of profile statements, in which no CpGs are selected and the methylation values of all CpGs are set to a constant value (MeanMeth = 56.91% for the HEP dataset). $\#CpG_{condition}$ stands for the number of CpGs in the amplicon that fulfill the condition. The source code implementing each of these rules is available on request (written in the Python programming language).

Bioinformatic analysis and prediction of inter-individual variation of DNA methylation

For quantitative analysis and prediction of the improvement in accuracy achievable by high-resolution methylation mapping of a specific amplicon, we defined the high-resolution improvement h as the difference between the inter-individual deviation calculated for medium-resolution MeDIP (v_{D1}) and high-resolution bisulfite sequencing (v_{F1}), i.e. $h = v_{D1} - v_{F1}$. Amplicons with high values exhibit inter-individually similar DNA methylation *patterns* while amplicons with low values do not (but may still exhibit similar *average* methylation levels across individuals).

The EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis) was used to test a large number of genomic attributes for their ability to distinguish between amplicons with high vs. low high-resolution improvement. To that end, two lists of amplicons (the top and the bottom quartile in terms of the high-resolution improvement) were uploaded into EpiGRAPH, and EpiGRAPH was used to test 845 genomic attributes for significant differences between the two lists. Significance was assessed by pairwise Wilcoxon tests, a global significance threshold of 5% was chosen and the highly conservative Bonferroni method was applied to correct for multiple testing.

Linear regression models were constructed to predict the dependent variable *h* (high-resolution improvement) from different subsets of independent variables. Both forward selection and backward selection were used to identify the most appropriate combination of independent variables. The following independent variables were included in the analysis: mean and standard deviation of amplicon methylation; GC content and CpG observed vs. expected ratio, calculated as in the definition of CpG islands (Gardiner-Garden and Frommer 1987); the relative frequency of three DNA sequence patterns (CG, CA and GC); the percent overlap with traditional and bona fide CpG islands, based on our work toward an improved annotation of CpG islands for the human genome (Bock et al. 2007, cf. chapter B-4 of this thesis); the degree of promoter activity derived from large-scale experimental data on transcription initiation events (Carninci et al. 2006); and the degree of transcriptional activity derived from the frequency of overlap with human ESTs from GenBank. The first four of these attributes were calculated directly from the HEP dataset and all other attributes were calculated by the EpiGRAPH web service. All statistical analyses were performed using the R statistics software (www.r-project.org/).

C-4.3 Results

DNA methylation profiles show complex patterns of variation among healthy individuals

Toward a better understanding of DNA methylation variation in healthy individuals, we defined three measures of inter-individual variation and applied them to the HEP dataset (see Methods section for details). Each measure captures a different aspect of inter-individual variation. The *pairwise deviation between high-resolution profiles* assesses the deviation between DNA methylation profiles from different individuals by summing over methylation differences of individual CpGs. Its value is low when all profiles for an amplicon are similar in terms of both their overall DNA methylation levels and their DNA methylation patterns. The *pairwise deviation between means* measures inter-individual differences of the average amplicon methylation. Its value is low when all DNA profiles share a similar mean, irrespective of the exact distribution and sequential order of methylated and unmethylated CpGs. The *deviation between mean and high-resolution profile* is a hybrid of the other two measures. Its value is low when DNA methylation profiles show little deviation from the mean of other profiles. Figure 28 illustrates the different behavior of these measures for two amplicons with artificially designed DNA methylation profiles.

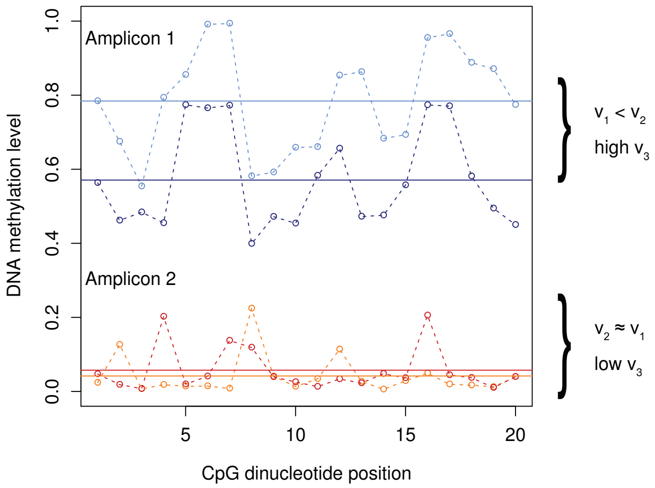


Figure 28. DNA methylation variation among healthy individuals (schematic figure)

This figure displays artificial DNA methylation data for two amplicons with two unrelated samples/profiles each, which were designed to illustrate the effect of the three measures of inter-individual variation used in this study. The typical amplicon with high overall methylation (blue profiles, top) has a relatively high *pairwise deviation between means (v_3)* and a *pairwise deviation between high-resolution profiles (v_1)* that is substantially lower than the *deviation between mean and high-resolution profile (v_2)*, which is reflected in a substantial correlation between the rising and falling of the DNA methylation profile curves over the length of the amplicon. In contrast, the typical amplicon with low overall methylation (red profiles, bottom) has a low *pairwise deviation between means (v_3)* and similar values for *pairwise deviation between high-resolution profiles (v_1)* and *deviation between mean and high-resolution profile (v_2)*, indicating that the fluctuations in the profiles are not inter-individually conserved and presumably random.

Using these three measures we analyzed whether the characteristics of inter-individual variation differ between amplicons with low vs. high DNA methylation levels (Figure 29, left). The results show that – by all three measures – the inter-individual variation of DNA methylation is lower for unmethylated amplicons than for methylated amplicons. This effect is strongest for the *pairwise deviation between means* (61% reduction), but also highly significant for the other measures ($P < 10^{-20}$ in all cases). This observation was not unexpected and could be explained by overlap with CpG islands, which are well-known to be stably unmethylated in a wide range of tissues. To test the role that CpG islands may play for this effect, we grouped amplicons by their overlap with bona fide CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis) and repeated the analysis (Figure 29, right). The results were similar, although the reduction of variance was less pronounced in the amplicons overlapping with bona fide CpG islands than in the experimentally unmethylated CpG islands. We also repeated this analysis using the Gardiner-Garden definition of CpG islands (Gardiner-Garden and Frommer 1987) and observe further dilution of the effect (data not shown), consistent with previous reports suggesting that traditional CpG island criteria give rise to a large number of false positives (Bock et al. 2007; Shen et al. 2007).

Beyond these results, which confirm and quantify previous observations, we found a second major difference between methylated and unmethylated amplicons. For methylated amplicons (and also for amplicons outside CpG islands), the *pairwise deviation between high-resolution profiles* is substantially lower than the *deviation between mean and high-resolution profile*. In contrast, differences are small for unmethylated amplicons and for amplicons overlapping with bona fide CpG islands (Figure 29). Importantly, this is not a side effect of the smaller scope for variation available to amplicons that were selected by their low DNA methylation averages, which is shown by plotting the top-25% most highly unmethylated amplicons vs. the top-25% most highly methylated amplicons (to which similar constraints apply) or the amplicons with an average methylation below 25% vs. the amplicons with an average methylation above 75% (Figure 30). In both cases, the difference of *deviation between mean and high-resolution profile* minus *pairwise deviation between high-resolution profiles* is consistently higher in methylated amplicons than in unmethylated amplicons.

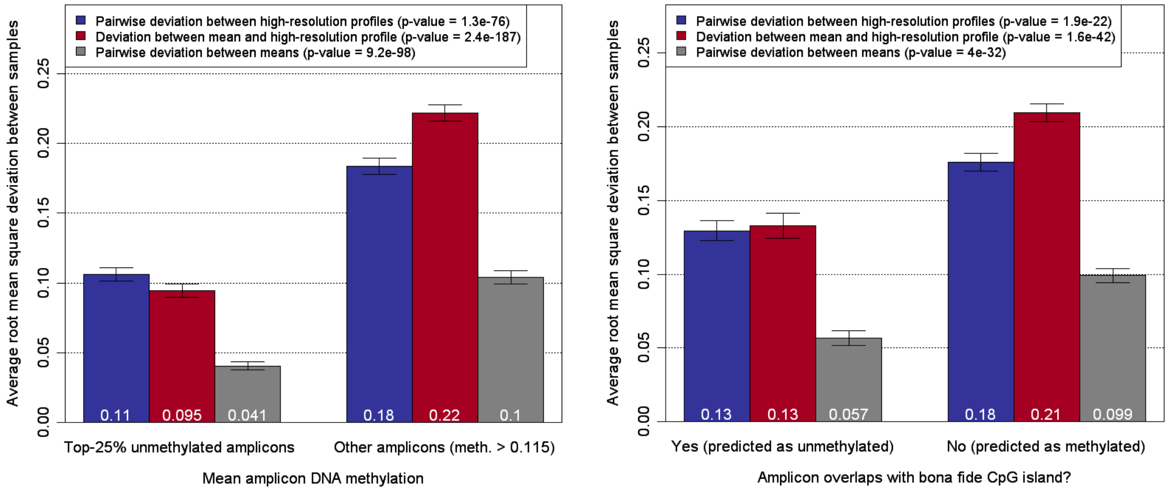


Figure 29. Effect of average amplicon methylation (left) and overlap with bona fide CpG islands (right) on inter-individual variation of DNA methylation

This figure shows the means of the three measures of DNA methylation variation as bar plots. In the left panel, values are reported separately for the top-25% most unmethylated amplicons with an average amplicon methylation of less than 11.5% (this threshold is motivated in the Methods section) and for the remaining 75% of amplicons. In the right panel, distinction is made between amplicons that overlap with a bona fide CpG island (Bock et al. 2007, cf. chapter B-4 of this thesis) and those that do not. In both cases, error bars represent 95% confidence intervals under the assumption of normal distribution and the *P*-values in the legends are based on two-sample, two-sided, *t*-tests between the group means for each measure.

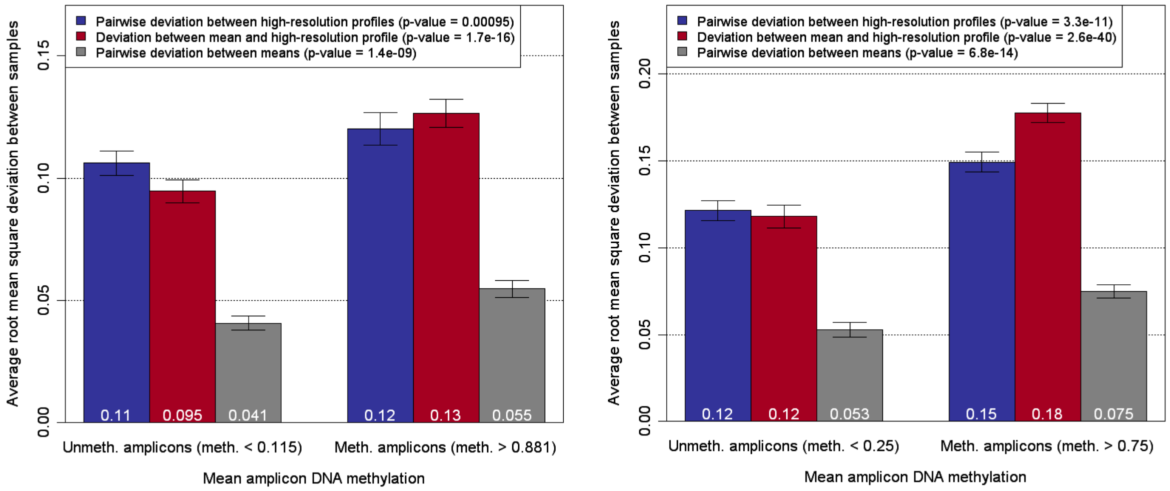


Figure 30. Different characteristics of inter-individual variation between amplicons with low and high methylation levels are not a side effect of smaller scope for variation among the former

Comparison between the top-25% most unmethylated amplicons (having an average amplicon methylation of less than 11.5%) and the remaining 75% of amplicons shows that both the *pairwise deviation between means* (v_3) and the difference of *deviation between mean and high-resolution profile* (v_2) minus *pairwise deviation between high-resolution profiles* (v_1) are smaller in unmethylated amplicons (see Figure 29). Importantly, this is not a side effect of the smaller scope for variation available to amplicons that have been pre-selected by their DNA methylation level being close to zero, as can be seen from this figure. Both diagrams plot the three measures of inter-individual variation for amplicons with DNA methylation levels close to 0% and, separately, of amplicons with DNA methylation levels close to 100%, to which similar scope-for-variation constraints apply. Specifically, in the left panel a comparison between the top-25% most unmethylated amplicons and the top-25% most methylated amplicons is shown, and in the right panel the comparison is made between amplicons with an average methylation below 25% and those above 75%. In both cases, error bars represent 95% confidence intervals under the assumption of normal distribution and the *P*-values in the legends are based on two-sample, two-sided t-tests between the group means for each measure.

These results indicate a qualitative difference in the characteristics of inter-individual variation of DNA methylation depending on the average level of DNA methylation and the CpG density (see Figure 28 for illustration). Methylated amplicons and amplicons outside CpG islands exhibit a high degree of inter-individual variation, but they also exhibit significant conservation of specific DNA methylation patterns between individuals, which is evident from the fact that the predictiveness across individuals increases when high-resolution profiles are compared. In contrast, in unmethylated amplicons and CpG islands the overall degree of inter-individual variation is substantially lower, but the high-resolution profile of one individual is not more predictive of other individuals' DNA methylation than its mean methylation level. Hence, we can regard the DNA methylation patterns of methylated and CpG-poor amplicons as informative at high resolution, while the average methylation level may suffice to characterize DNA methylation at unmethylated amplicons and CpG islands.

In silico benchmarking shows that high-resolution methylation mapping is most informative outside CpG islands

The different characteristics of inter-individual variation have important implications for a rational choice of experimental methods for DNA methylation mapping: High-resolution mapping (e.g. by bisulfite sequencing) would be required outside CpG islands, while methods that assess average methylation levels (e.g. MeDIP) were sufficient for CpG islands. To substantiate this conclusion, we conducted a comprehensive in silico benchmarking study of six widely used experimental methods for DNA methylation mapping. This benchmarking is based on the assumption that DNA methylation profiles are only informative to the degree to which they are conserved between individuals (see Discussion for critical assessment), and it adheres to a straightforward protocol: For each amplicon in the HEP dataset and all pairs of (non-identical) DNA methylation profiles derived from different individuals, we compared how well a simulated measurement – calculated from the first profile – predicts the second DNA methylation profile. The key point is that the measurement derived from the first profile is calculated in a way that models the experimental characteristics of different methods for DNA methylation mapping (Table 11). For example, for method C2 (qualitative immunoprecipitation), an amplicon is considered methylated if more than three CpGs per 200 basepairs exhibit DNA methylation levels above 50%. For method B3 (quantitative analysis of HpaII methylation-sensitive restriction libraries), the simulated measurement is calculated by assigning the known methylation levels to all CpGs that overlap with the enzyme's recognition sites (CCGG), while the methylation levels of all remaining CpG dinucleotides are determined by interpolation or extrapolation. This way, the benchmarking assesses how accurately different experimental methods map inter-individually stable DNA methylation.

The results calculated over all amplicons (Figure 31) show that high-resolution mapping by bisulfite sequencing (method F1) gives rise to the lowest inter-individual deviation (v_{F1} = 0.164). Therefore, a substantial number of CpGs in the genome must exhibit inter-individually stable DNA methylation patterns, which can be detected only by high-resolution bisulfite sequencing. However, quantitative immunoprecipitation (method D1) follows relatively closely behind, with an average deviation that is 16% higher than that of bisulfite sequencing. Qualitative methods – which test whether the DNA methylation in a genomic region exceeds a specific threshold rather than measuring its exact value – tend to perform worse than quantitative methods. The best one (method E3) results in an average deviation that is 36% worse than that of bisulfite sequencing. This, however, is still substantially better than random guessing (method G7), which would lead to average deviations that are more than three times worse than for bisulfite sequencing. Two additional observations are worth highlighting. First, alternative rules for the same experimental method (listed in the same rows in Table 11) perform similarly. This is particularly evident for the method groups A1 to A3, C1 to C3, D1 to D2 and E1 to E3, indicating that our results are robust regarding the choice of parameters for these rules. Second, for those methods that interrogate several individual CpGs to assess an amplicon's methylation status, careful selection of representative CpGs can increase performance. For example, selecting the first and last CpG of an amplicon (method F7) or randomly selecting two CpGs (method F3) performs worse than the more representative selection of the two CpGs that are located most closely to positions one third and two thirds relative to the amplicon length (method F8).

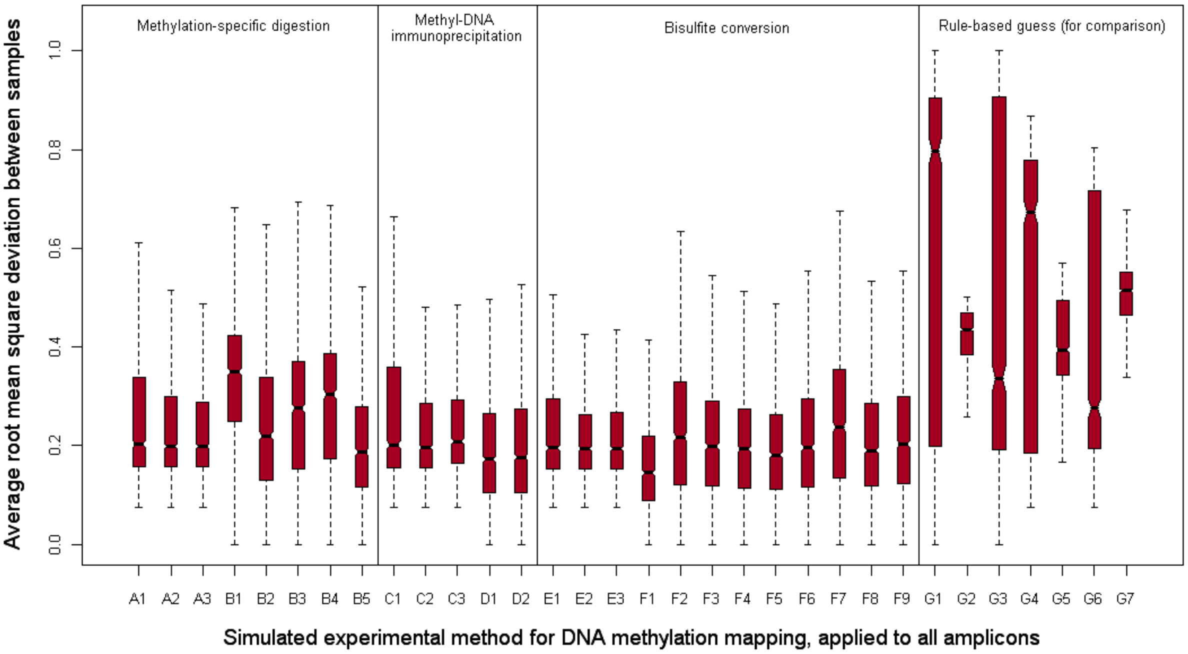
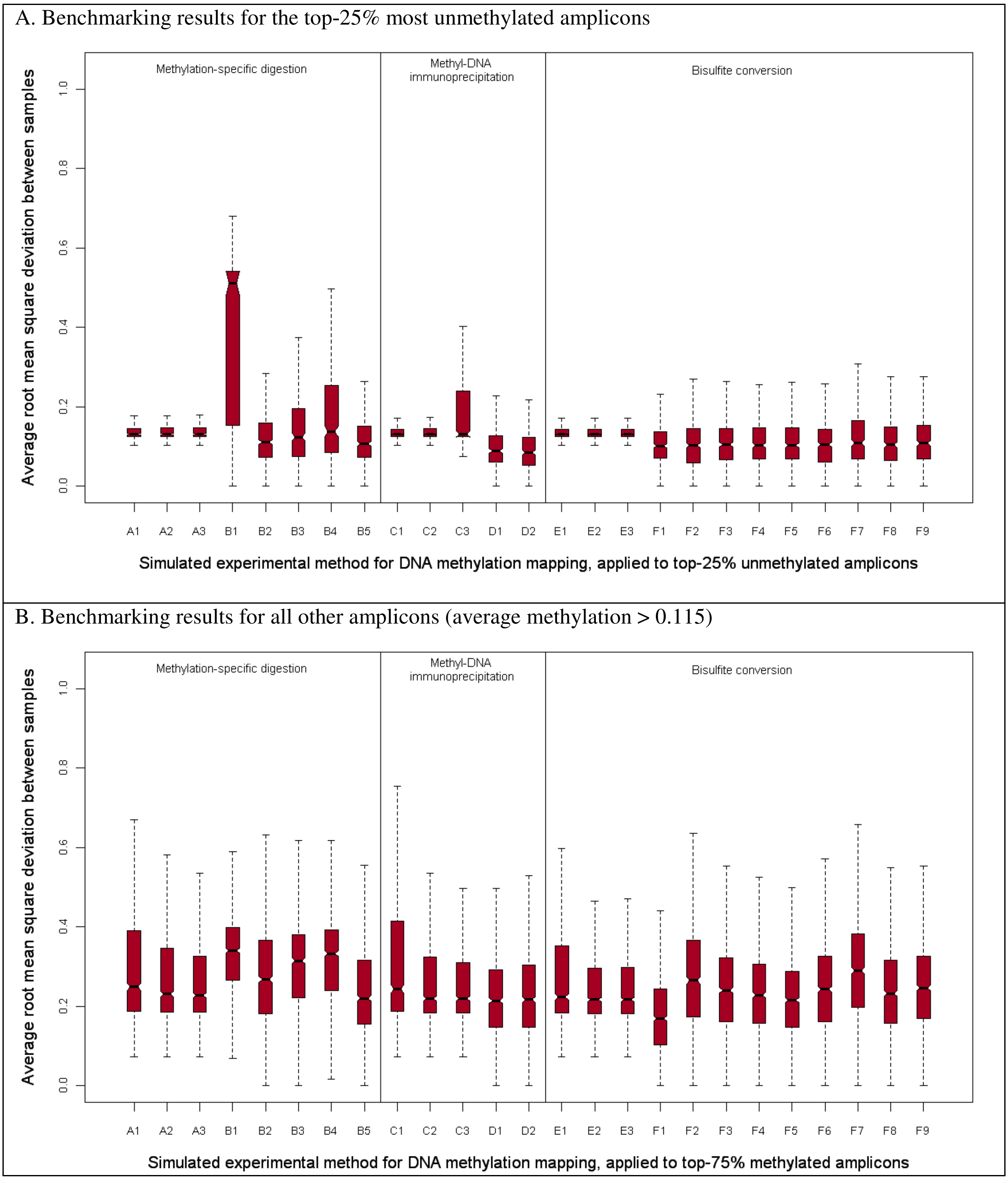


Figure 31. Benchmarking results for experimental mapping of DNA methylation

This figure displays the results of in silico benchmarking of different DNA methylation mapping methods for all amplicons. The *y*-axis shows v_{method} values for all experimental methods included in this study (A1 to F9, described in Table 11) and for seven controls, which are based on guessing rules rather than on experimental data (G1 to G7, described in Table 11). The standard boxplot format is used (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box) and outliers are hidden.

Having established this in silico prediction setup, we could test our initial hypothesis that bisulfite sequencing (method F1) performs better than, for example, MeDIP (method D1), but does so only for amplicons that exhibit moderate to high levels of DNA methylation and do not overlap with CpG islands. The results strongly support our hypothesis (Figure 32 and Figure 33). For amplicons with moderate to high levels of DNA methylation (the same threshold is used as in Figure 29), as well as for amplicons that do not overlap with a bona fide CpG island, MeDIP performs almost 20% worse than bisulfite sequencing. In contrast, for amplicons overlapping with a bona fide CpG island the difference is less than 3%, and for the most unmethylated amplicons MeDIP performs even better than bisulfite sequencing (by 11%) – arguably because it averages out uninformative fluctuations.

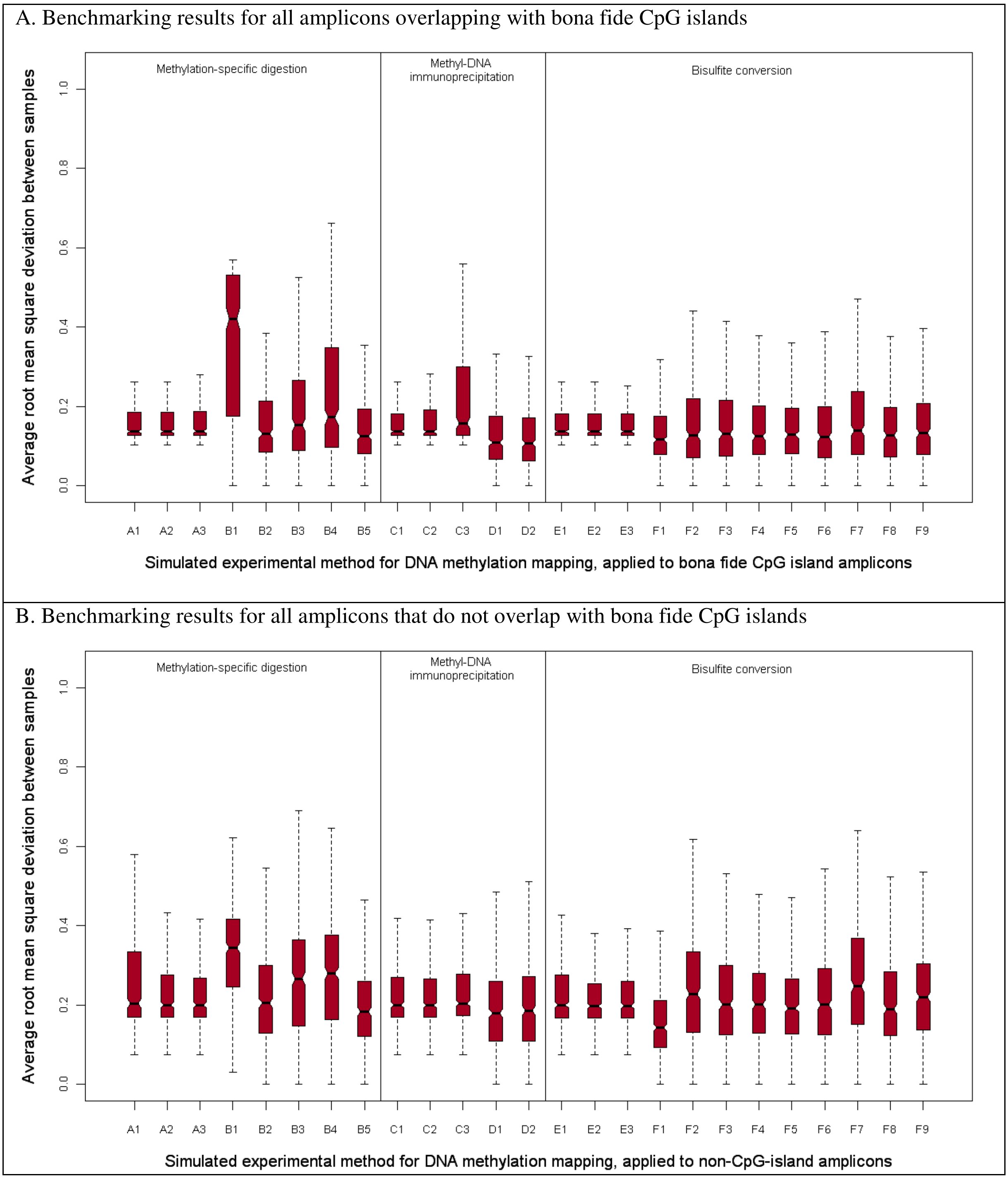


A. Benchmarking results for the top-25% most unmethylated amplicons

B. Benchmarking results for all other amplicons (average methylation > 0.115)

Figure 32. Effect of average amplicon methylation on benchmarking results for experimental mapping of DNA methylation

This figure displays the results of in silico benchmarking of different methods for DNA methylation mapping. Distinction is made between amplicons belonging to the top-25% most unmethylated amplicons with an average amplicon methylation of less than 11.5% (panel A) and the remaining 75% of amplicons (panel B). The y-axis plots v_{method} values for all experimental methods included in this study (A1 to F9, described in Table 11). The standard boxplot format is used (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box) and outliers are hidden.



A. Benchmarking results for all amplicons overlapping with bona fide CpG islands

B. Benchmarking results for all amplicons that do not overlap with bona fide CpG islands

Figure 33. Effect of overlap with bona fide CpG islands on benchmarking results for experimental mapping of DNA methylation

This figure displays the results of in silico benchmarking of different methods for DNA methylation mapping. Distinction is made between amplicons that overlap with a bona fide CpG island (panel A) and those that do not (panel B). The *y*-axis plots v_{method} values for all experimental methods included in this study (A1 to F9, described in Table 11). The standard boxplot format is used (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box) and outliers are hidden.

The accuracy improvement of high-resolution methylation mapping can be predicted from the DNA sequence

Up to this point, we have used the overlap with bona fide CpG islands as a sequence-based criterion to discriminate amplicons for which measuring average methylation levels is sufficient from those requiring high-resolution mapping. However, a priori it is not clear that this criterion provides the most accurate discrimination. To put the identification of regions that benefit from high-resolution mapping onto a more systematic basis, we applied the following two-step process. First, we used the EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis) to obtain a broad basis of potentially predictive attributes. Second, for a selection of highly significant attributes from the EpiGRAPH analysis, we constructed linear regression models that quantitatively predict the high-resolution improvement, which we define as the difference between the inter-individual deviation for simulated medium-resolution MeDIP and simulated high-resolution bisulfite sequencing.

A total of 845 genomic attributes were included in the EpiGRAPH analysis, each belonging to one of the following attribute groups: DNA sequence, DNA structure, repetitive DNA, chromosome organization, evolutionary history, population variation, genes, regulatory regions, transcriptome, epigenome and chromatin structure. Of these attributes, 96 were found to be significantly different between amplicons with high vs. low high-resolution improvement. We selected seven highly significant attributes for in-depth analysis, namely the relative frequency of the DNA sequence patterns CG, CA and GC, the percent overlap with traditional and bona fide CpG islands, a quantitative measure of promoter activity (CAGE tag frequency) and a quantitative measure of transcriptional activity (EST overlap frequency). Together with the mean and standard deviation of amplicon methylation as well as the GC content and CpG observed vs. expected ratio, this gave rise to a list of eleven independent variables, which we assessed for their potential as predictors of high-resolution improvement (the dependent variable). Initially, we calculated pairwise Pearson correlation coefficients between the independent and dependent variables (Table 12), highlighting substantial correlation not only between the independent and dependent variables, but also among the independent variables.

Pearson correlation	High-resolution improvement	Mean amplicon methylation	SD of amplicon methylation	GC content	CpG observed versus expected ratio	Frequency of CG pattern	Frequency of CA pattern	Frequency of GC pattern	Overlap with traditional CpG islands	Overlap with *bona fide* CpG islands	Frequency of transcription initiation	Frequency of overlap with human ESTs
High-resolution improvement	1.00	0.18	0.47	−0.17	−0.22	−0.25	0.14	−0.19	−0.19	−0.20	−0.14	−0.03
Mean amplicon methylation		1.00	0.27	−0.37	−0.51	−0.58	0.58	−0.35	−0.55	−0.68	−0.35	−0.03
SD of amplicon methylation			1.00	−0.27	−0.39	−0.42	0.24	−0.29	−0.37	−0.36	−0.21	−0.01
GC content				1.00	0.26	0.74	−0.28	0.83	0.54	0.40	0.27	0.04
CpG observed versus expected ratio					1.00	0.80	−0.65	0.36	0.71	0.30	0.30	0.01
Frequency of CG pattern						1.00	−0.62	0.74	0.78	0.63	0.40	0.02
Frequency of CA pattern							1.00	−0.20	−0.54	−0.30	−0.30	−0.01
Frequency of GC pattern								1.00	0.51	0.36	0.30	0.00
Overlap with traditional CpG islands									1.00	0.60	0.26	0.03
Overlap with *bona fide* CpG islands										1.00	0.35	0.02
Frequency of transcription initiation											1.00	0.06
Frequency of overlap with human ESTs												1.00

Table 12. Correlation between high-resolution improvement and its potential predictors

This table displays pairwise Pearson correlation coefficients for the accuracy improvement of high-resolution methylation mapping (first row) and several potential factors of influence. Orange boxes mark strong positive correlation and blue boxes mark strong negative correlation.

We therefore used the statistical framework of linear regression to control for correlations among the independent variables and to derive a prediction model for the high-resolution improvement. Using feature selection (both forward and backward selection gave the same results), we determined the optimal combination of independent variables, and the following regression model was calculated as an optimal linear predictor of high resolution improvement: $h_i = 0.3202 \cdot a_i - 0.0463 \cdot b_i - 0.0264 \cdot c_i + 0.0102 \cdot d_i$. In this formula, h_i stands for the high-resolution improvement of amplicon *i*, a_i is the standard deviation of amplicon methylation, b_i the GC content, c_i the CpG observed vs. expected ratio and d_i the percent overlap with traditional CpG islands (calculated on the repeat-masked genome). This regression model gives rise to a residual standard error of $s_E = 0.0611$ and an adjusted correlation coefficient of $r = 0.48$, and is highly significant ($P < 10^{-10}$). Figure 34 shows a scatter plot comparing the model's prediction with the observed high-resolution improvement, indicating that the prediction accuracy is high for low values of h_i and decreases substantially for high values of h_i, due to high variance among the observed values.

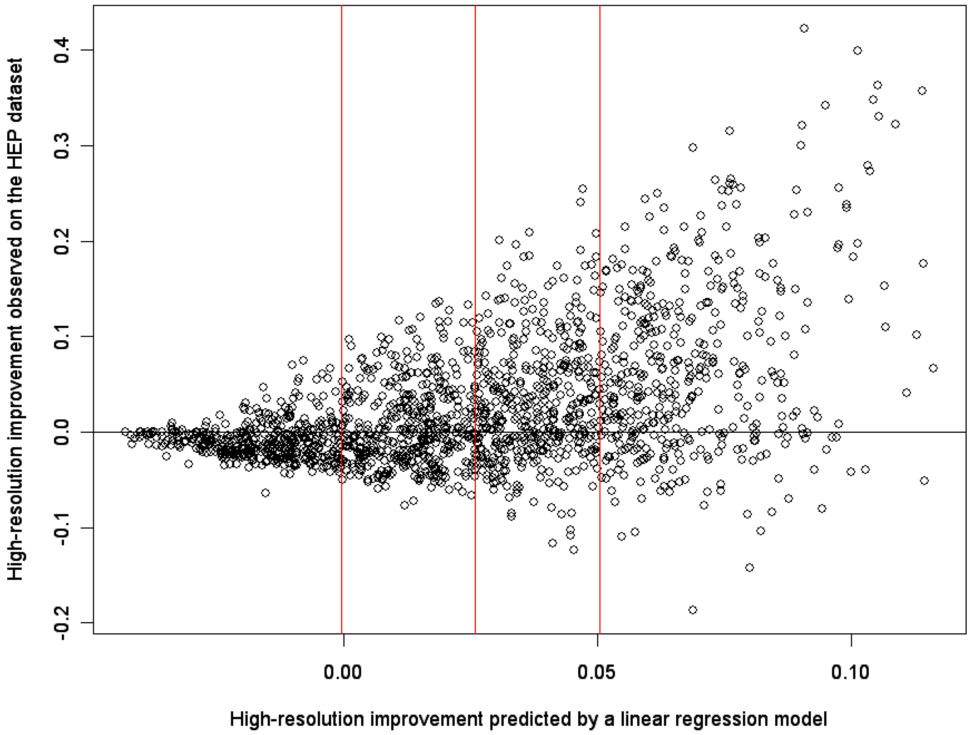


Figure 34. Correlation between high-resolution improvement predicted by a linear regression model and its observed values

This figure displays a scatter plot of the high-resolution improvement that is predicted by a linear regression model based on four attributes (standard deviation of amplicon methylation, GC content, CpG observed vs. expected ratio and overlap with traditional CpG islands) and the high-resolution improvement observed on the HEP dataset. The vertical lines indicate quartiles on the predicted high-resolution improvement. The Pearson correlation coefficient between the two variables is 0.48.

Prediction of high-resolution improvement facilitates cost-efficient DNA methylation mapping

Finally, we asked whether prediction models could be used prospectively, to help decide which amplicons require high-resolution analysis (e.g. by bisulfite sequencing) and for which amplicons it would be sufficient to measure their average methylation level (e.g. by cost-efficient MeDIP analysis). We stipulated that the second alternative would be acceptable and sufficient only if the risk is less than 5% that a substantial loss of accuracy is incurred for a specific amplicon. Based on Lewin et al. (Lewin et al. 2004), who report a mean absolute error of 14 percentage points for CpG methylation levels determined by bisulfite sequencing, we speak of a substantial loss of accuracy if the high-resolution improvement exceeds 0.14.

The goal then was to predictively identify as many amplicons as possible that exhibit a low high-resolution improvement, while not exceeding a false positive rate of 5%. To that end, we derived a new linear regression model that does not include the standard deviation of amplicon methylation (this value is typically unknown when planning experimental mapping of DNA methylation). After feature selection, the following regression model was obtained: $h_i = 0.1406 - 0.0796 \cdot b_i - 0.0646 \cdot c_i - 0.1894 \cdot e_i - 0.0167 \cdot f_i$, in which e_i stands for the relative frequency of sequence pattern CA, f_i for the percent overlap with bona fide CpG islands and the other variables are as above. The accuracy of this model is lower than that of the previous model ($s_E = 0.0671$ and $r = 0.26$), but it is still highly significant ($P < 10^{-10}$). Next, the threshold on the predicted high-resolution improvement was chosen such that no more than 5% of amplicons below this threshold exhibit an observed high-resolution improvement of 0.14. This calculation resulted in a threshold value of 0.0358, selecting 1118 out of 1705 amplicons (65.6%) for which high-resolution analysis is highly unlikely to provide substantially improved accuracy over (cheaper) analysis of average methylation levels.

C-4.4 Discussion

This study analyzed inter-individual stability and variation of DNA methylation profiles among healthy individuals. Using statistical methods we could show that the DNA methylation state of CpG-rich regions is exhaustively characterized by their average methylation levels, while high-resolution DNA methylation patterns are informative only in regions with low CpG density (high-resolution patterns are considered informative if they improve prediction accuracy when comparing DNA methylation across individuals). A plausible biological explanation would be that above a critical CpG density, neighboring CpGs influence each other's DNA methylation states so strongly that individual CpGs cannot stably maintain DNA methylation states deviating from those of their neighbors. Biochemically, this could be due to methylation-specific enhancement or repression of DNA methyltransferase activity, a mechanism that has been proposed to contribute to spreading of DNA methylation (Turker 2002). In contrast, individual CpGs in CpG-poor regions lack this pressure from neighboring CpGs, and other effects – such as the local DNA sequence environment (Handa and Jeltsch 2005) or transcription factor binding (Xu et al. 2007) – are likely to determine their DNA methylation states. From a systems point of view, we suggest that CpG islands may act as emergent and bistable epigenetic switches, in which multiple CpGs collectively maintain a CpG-island-wide "on" or "off" state. This concept is consistent with experimental data, including the bimodal distribution of promoter methylation observed in normal cells (Weber et al. 2007) and the fact that entire CpG islands, rather than single CpGs, become aberrantly methylated in cancer (Laird 2005). It is also supported by two recent in silico studies showing that cooperativity among neighboring CpGs (Sontag et al. 2006) and spatially close nucleosomes (Dodd et al. 2007) is required for a genomic region to function as a bistable epigenetic switch.

Based on our statistical results, we also considered the practical implications for experimental analysis of DNA methylation. Through the combination of computational simulation and benchmarking across unrelated individuals, we could show that in CpG-poor genomic regions, high-resolution methods such as bisulfite sequencing perform substantially better than medium-resolution methods such as MeDIP. In contrast, both methods perform similarly for CpG-rich genomic regions and CpG islands, owing to high fluctuation of the sequential order of methylated and unmethylated CpGs in these regions. We derived a linear classifier for predicting which genomic regions benefit from bisulfite sequencing and for which regions MeDIP is sufficient. To highlight the potential cost savings arising from these results, we briefly sketch how the classifier could influence a HEP-like project planned today: Assume that half of the total project costs are variable and proportional to the number of amplicons analyzed. Furthermore, assume that it is four times as expensive to assess all CpGs in an amplicon (by Sanger sequencing of bisulfite-converted DNA) than to assess an amplicon's average methylation level (e.g. by MeDIP). Our predictions would enable us to apply cheap methods to roughly two thirds (65.6%) of all HEP amplicons (see Results section), such that only 5% of these would have benefited significantly from costly high-resolution analysis (i.e. the high-resolution improvement would be less than 0.14 for 95% of all amplicons). This would give rise to savings of 25% in terms of overall project costs, compared to the indiscriminate high-resolution strategy used in the HEP. Alternatively, these savings would permit the analysis of 50% more samples at the same overall project costs (assuming that non-proportional costs are unaffected by the increased throughput).

It is, however, important to keep several limitations of our analysis in mind. First, all results are currently based on a single, albeit large, dataset and should be further validated on data obtained in different labs, potentially with different methods and for a larger number of samples. Second, because the HEP dataset was generated using direct sequencing rather than sequencing of clones, we were unable to assess the degree of variation within a single sample. Third, because the HEP dataset uses only a single sample per individual, we cannot exclude that a substantial percentage of the observed inter-individual differences may also be present between different samples of the same individual, e.g. as a result of tissue heterogeneity. Fourth, our simulation rules compute the expected DNA methylation measurement under optimal conditions, ignoring aspects such as robustness with respect to varying DNA quality or minor variation in the experimental protocol. Hence, the benchmarking results describe an inherent property of the different methods rather than their actual performance under a specific set of actual conditions (i.e. a specific protocol used by a specific researcher in a specific lab). Fifth, our assumption that those high-resolution DNA methylation patterns fluctuating randomly between unrelated individuals are uninformative and can be replaced by their mean holds true only when the goal is to make generalizable claims about DNA methylation patterns in a particular genomic region. This is obviously the case in large-scale epigenome projects aimed at the establishment of reference maps of DNA methylation, and also for most cancer epigenetics and biomarker discovery projects. However, this assumption is less appropriate when analyzing epigenetic regulation in a single cell: a single methylated CpG may well be functional, e.g. preventing a transcription factor from binding to the DNA, even if it is not inter-individually conserved.

These limitations notwithstanding, our results provide a first quantitative basis for strategic decision making in large-scale DNA methylation mapping. Combining all of our observations, we propose the following cost-optimized two-track strategy for mammalian methylome projects: On the one hand, DNA methylation at all CpG islands (or more accurately: at all CpG-rich regions predicted by the classifier that is described in the Results section) should be analyzed in a large number of individuals, in order to quantify the degree of epigenetic variation within human populations. For these experiments, a cost-efficient medium-resolution method such as MeDIP is sufficient, since our results show that the methylation state of CpG islands is exhaustively characterized by their average methylation levels. On the other hand, in a smaller number of individuals the entire genome – consisting mostly of CpG-poor regions – should be analyzed by high-throughput bisulfite sequencing (Meissner et al. 2005), in order to provide a basis for assessing which of these CpGs play a functional role in gene regulation or chromatin structure formation. This two-track strategy contrasts with the naïve approach of mapping DNA methylation at high resolution where CpG density is high and at low resolution where their density is low, which underlines the relevance of computational analysis for informed planning of epigenome projects.

Part D. Cancer Epigenetics

"Much [of cancer epigenetics] is still unknown, but the unfolding scenario shows great promise for a better understanding of cancer biology and for improvement in the management of human tumors" (Manel Esteller)[1]

D-1 Outline

Given the fundamental role of epigenetic regulation in the context of development, cell differentiation, tissue specificity and stem cell identity, it is not surprising that epigenetic errors are major contributors to tumorigenesis (Baylin and Ohm 2006; Feinberg and Tycko 2004). The link between epigenetic regulation and cancer is most established for DNA methylation, which becomes altered in two opposing ways. First, aberrant hypermethylation (i.e. increased DNA methylation) of promoter regions frequently results in cancer-specific silencing of tumor suppressor genes (Esteller 2007). Second, simultaneous hypomethylation (i.e. decreased DNA methylation) of repetitive genomic regions can contribute to genome instability (Feinberg and Tycko 2004; Laird 2005; Ting et al. 2006). Furthermore, an exciting set of papers has reported a potential connection between aberrant DNA methylation in cancer and epigenetic regulation by Polycomb group proteins in embryonic stem cells (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007), and it was hypothesized that epigenetic deregulation can program stem cells for malignancy long before they are histologically identifiable as tumor cells (Feinberg et al. 2006). Recent advances in high-throughput epigenome mapping are likely to provide unprecedented insights into the etiology of human cancers and could give rise to new approaches for early diagnosis, prognosis and therapy optimization.

In the following chapters, two case studies are presented in which we apply bioinformatic methods to cancer epigenetics. In the first study (chapter D-2), we show that a biochemically identified interaction between DNA methylation and Polycomb group proteins is likely to be functionally relevant in cancer cells, which could contribute to our mechanistic understanding of epigenetic deregulation in cancer (Viré, et al., *submitted*). In the second study (chapter D-3), which is more immediately targeted toward improving cancer therapy, we optimize a well-established biomarker of chemotherapy resistance for cheap and robust application in routine clinical diagnosis (Mikeska et al. 2007).

Although these two case studies address specialized topics, we aimed to develop bioinformatic approaches that generalize to a broader class of problems in cancer epigenetics. The first case study can be regarded as an exploratory example of how to assess the genome-wide functional relevance of an experimentally identified interaction of chromatin proteins, using a combination of ChIP-on-chip experiments, computational analysis and bioinformatics-guided experimental validation. The second case study prototypes a systematic approach for optimizing DNA methylation biomarkers, a topic that is discussed in more detail in section E-2.2 below.

[1] Quoted after: Esteller, M. 2005. DNA methylation: approaches, methods, and applications. CRC Press, Boca Raton; London.

D-2 Relevance of the methyl-CpG binding protein MeCP2 for Polycomb recruitment in cancer cells[1]

D-2.1 Motivation

Polycomb group proteins are transcriptional repressors that play a role in biological processes such as embryogenesis, cell differentiation and cancer development (Buszczak and Spradling 2006; Schuettengruber et al. 2007). They act as parts of large protein complexes that are conserved from Drosophila to mammals. The Polycomb repressive complex PRC2 is composed, among other proteins, of EZH2, EED and SUZ12 (Schwartz and Pirrotta 2007). EZH2 is the catalytic component of this complex and can confer histone methylation to lysine 27 of histone H3, thereby inducing silencing of neighboring genes. Several studies suggest a functional link between transcriptional repression by Polycomb group proteins and DNA methylation in the context of cancer (see Ohm and Baylin 2007 and references therein). In particular, Viré et al. showed for a cancer cell line that EZH2 can induce de novo DNA methylation through its association with DNA methyltransferases (Viré et al. 2006).

In an attempt to improve our understanding of PRC2 recruitment in mammals and its potential relevance for aberrant gene silencing in cancer, we asked whether the reverse link might also be true, i.e. that DNA methylation might foster PRC2 recruitment. A plausible candidate for moderating such a relationship is the methyl-CpG binding protein 2 (MeCP2), a transcription factor that binds specifically to methylated cytosines (Shahbazian and Zoghbi 2002). Several biochemical assays indicated that MeCP2 can indeed target EZH2 to specific promoters. In the U2OS cancer cell line, EZH2 co-immunoprecipitated with MeCP2, and GST pull-down assays confirmed a protein-protein interaction between these two proteins (Viré, et al., submitted).

These results suggest that MeCP2 binding might play a role in guiding PRC2 binding in mammalian cells, which would constitute an intriguing mechanism of reciprocal feedback between two major epigenetic repressors, DNA methylation and binding by Polycomb group proteins. However, before such conclusions can be drawn, it is essential to confirm the colocalization and functional interaction of MeCP2 and EZH2 at a large number of target genes. ChIP-on-chip analysis was therefore performed for MeCP2 and EZH2 in the U2OS cancer cell line and in the WI38 fibroblast cell line, in order to test for significant association of MeCP2 and EZH2 on a truly genomic scale. In the following, we focus on the bioinformatic analysis of the resulting datasets, describing experimental details only insofar as they are indispensable for proper understanding of the computational part.

D-2.2 Methods

ChIP-on-chip microarray design

ChIP-on-chip analysis was performed on promoter tiling arrays manufactured by NimbleGen Systems, Inc. (Madison, WI). Each microarray comprised approx. 385,000 probes, with a median probe spacing of 100 basepairs and probe lengths between 50 and 75 basepairs. Two microarrays were combined to tile the putative promoter regions of all high-confidence tran-

[1] This chapter describes work conducted in collaboration with Emmanuelle Viré, Hélène Denis, Carmen Brenner and François Fuks (Viré et al, submitted). Emmanuelle Viré, Hélène Denis and Carmen Brenner performed the wet-lab experiments, while Emmanuelle Viré and François Fuks contributed to the interpretation of the results.

scripts annotated for the human genome. Specifically, transcripts from the following gene annotation databases were incorporated: RefSeq genes, the Mammalian Gene Collection, and UCSC Known Genes. A total number of 59,357 transcripts were included in the promoter tiling array design. For each high-confidence transcript, the region from -3,500 basepairs (upstream) to +750 basepairs (downstream) relative to the annotated transcription start site was tiled, and where neighboring genes had overlapping promoter regions, these were merged into joint regions. Repetitive regions such as retrotransposons and tandem repeats were included only where they were degenerate enough to place unique probes. In summary, the total genomic coverage of the two-slide promoter tiling array used for all ChIP-on-chip hybridizations was approx. 110 megabases, with a total number of 23,047 distinct genomic regions and a mean region length of 4,758 basepairs.

ChIP-on-chip quality control and analysis of probe intensity data

Initial quality control was performed with the help of the NimbleScan software package (http://www.nimblegen.com/products/software/nimblescan.html), following the recommended procedures. Additional quality control steps were performed with Bioconductor (Du et al. 2006) and Ringo (Bracken et al. 2006; Lee et al. 2006; Squazzo et al. 2006). Bioconductor is a microarray analysis package for the R statistics software (http://www.r-project.org) and the most widely used open-source tool for microarray analysis. Ringo is a data processing and quality control package that extends Bioconductor's functions to the analysis of NimbleGen ChIP-on-chip data.

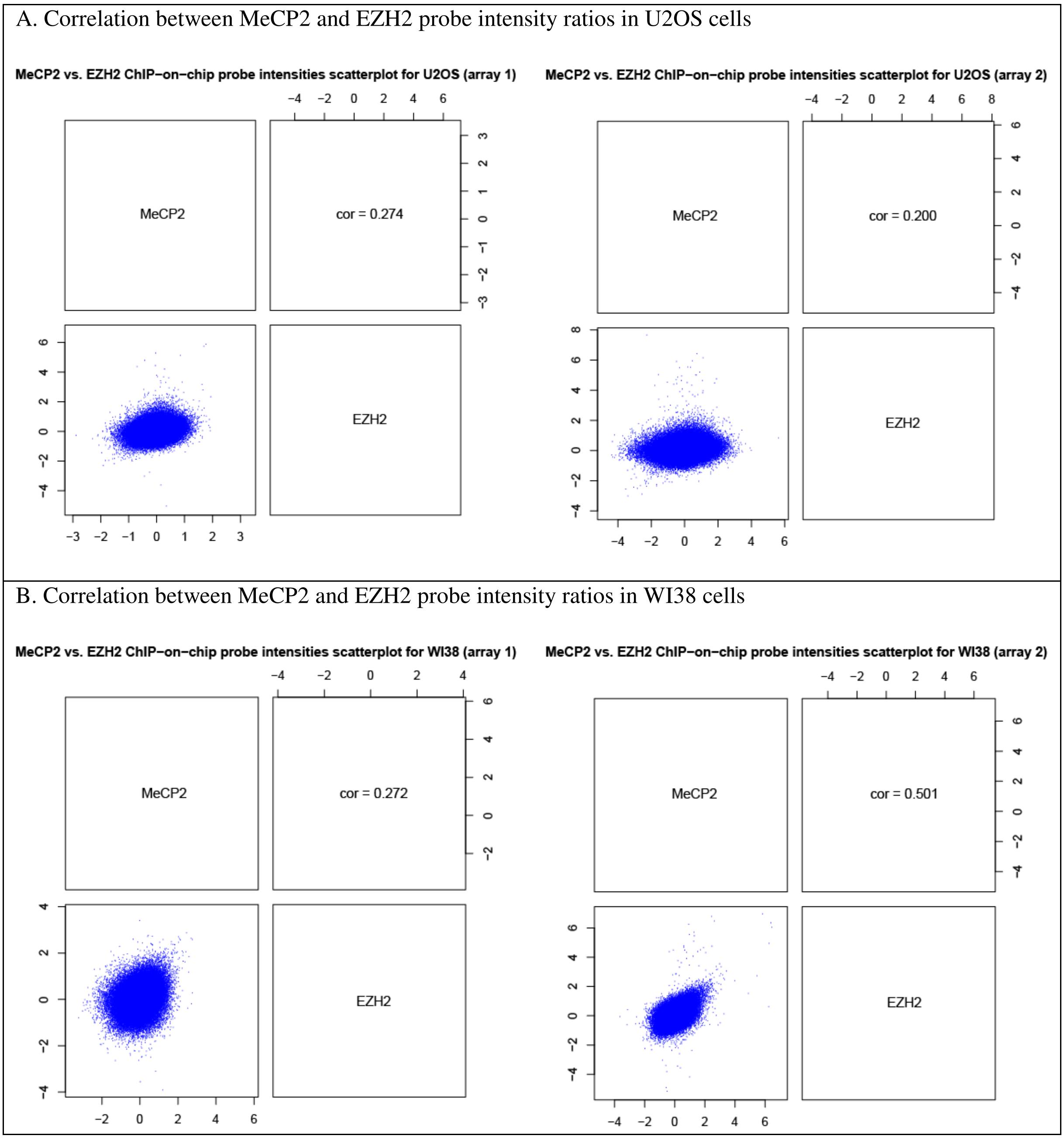
First, the raw microarray scanner data for Cy3 and Cy5 intensities were imported using Ringo's *readNimblegen* function, and diagrams were generated to visualize the spatial distribution of raw probe intensities on the microarray surface. Manual inspection showed no obvious scratches, uneven distribution of probe intensities or other artifacts that would be indicative of experimental problems during microarray hybridization (data not shown).

Second, raw probe intensity values were normalized by the variance-stabilization normalization method (Huber et al. 2002), and log scores were calculated for the probe intensity ratios observed for immunoprecipitated DNA vs. control DNA. These log scores were then visualized as scatterplots (Figure 35), resulting in a low to moderate correlation between MeCP2 and EZH2, as would be expected for two chromatin modifications that show significant overlap only in specific genomic regions. Pearson correlation coefficients ranged from 0.2 to 0.5 and significantly different from zero in all cases ($P < 10^{-15}$).

ChIP-on-chip peak detection, threshold selection and experimental validation

Having passed quality control, raw probe intensities were normalized with NimbleScan in order to adjust for different probe characteristics, and scaled probe intensity log-2 ratios were calculated. To derive high-confidence binding sites for MeCP2 and EZH2 from the probe intensity ratios, a two-step peak detection and threshold selection strategy was applied.

Peak detection was performed with the peak finder implemented in NimbleScan. Briefly, a sliding window method was applied to the scaled probe intensity log-2 ratios to identify sets of neighboring probes that exceed specific length and intensity thresholds. Next, random permutation was used in order to assess the significance of these putative peaks and to assign false discovery rate (FDR) estimates (Benjamini and Hochberg 1995). Default parameters were used throughout, giving rise to a single list of peaks per ChIP-on-chip experiment. In U2OS cells, a total number of 7,009 distinct peaks were detected for MeCP2 and 6,378 distinct peaks were detected for EZH2. In WI38 cells, the number of distinct peaks was slightly lower: 6,101 for MeCP2 and 4,246 for EZH2.



A. Correlation between MeCP2 and EZH2 probe intensity ratios in U2OS cells

B. Correlation between MeCP2 and EZH2 probe intensity ratios in WI38 cells

Figure 35. Scatterplots of ChIP-on-chip probe intensity ratios

This figure displays a scatterplot of log scores of probe intensity ratios between immunoprecipitated DNA and control DNA, based on ChIP-on-chip experiments for MeCP2 and EZH2 (bottom left square in each diagram). In addition, the corresponding Pearson correlation coefficients are reported (top right square in each diagram).

Obviously, a large percentage of these peaks are due to random fluctuations in the experiment, rather than due to reproducible binding sites. Because no experimental replicates of the ChIP-on-chip experiments were available for comparison, we applied a combination of threshold selection and small-scale validation in order to distinguish reproducible binding sites from statistical and biological artifacts. In theory, the FDR estimate derived by random permutation testing provides a statistical basis for threshold selection and it would be an obvious choice to use a 5% cutoff on the FDR. However, it has been frequently observed that significance estimates that are based purely on random permutation of ChIP-on-chip data are insufficient for biologically meaningful threshold selection and that a moderate number of putative peaks should be experimentally validated in an independent assay. Therefore, from each of the four lists of ChIP-on-chip peaks, a substantial number of putative binding sites were selected for independent validation by conventional ChIP. The selection procedure was random and a knowledge-based stratification strategy was used in order to increase the resolution (i.e. the density of targets for validation) at positions in the lists where – on the basis of prior knowledge – the validation rates might be expected to drop. Using this protocol, conventional ChIP was performed for 32 putative MeCP2 binding sites and 25 putative EZH2 binding sites in U2OS cells. Furthermore, conventional ChIP was performed for 25 putative MeCP2 binding sites and 22 putative EZH2 binding sites in WI38 cells.

For the MeCP2 datasets, the widely used five percent threshold on the FDR is well-supported by experimental validation. Conventional ChIP gave rise to independent validation rates of 71% (U2OS) and 69% (WI38), respectively, for putative peaks with an FDR below five percent. This threshold was therefore selected for all further analyses, resulting in 503 high-confidence MeCP2 binding sites for the U2OS cancer cell line and 218 high-confidence MeCP2 binding sites for the WI38 fibroblast cell line. On the basis of these data, a conservative estimate can be calculated for the total number of promoter-associated MeCP2 binding sites in the human genome. For the U2OS cancer cell line, this value is $503 \cdot 0.71 = 357$ and for the WI38 fibroblast cell line it amounts to $218 \cdot 0.69 = 150$.

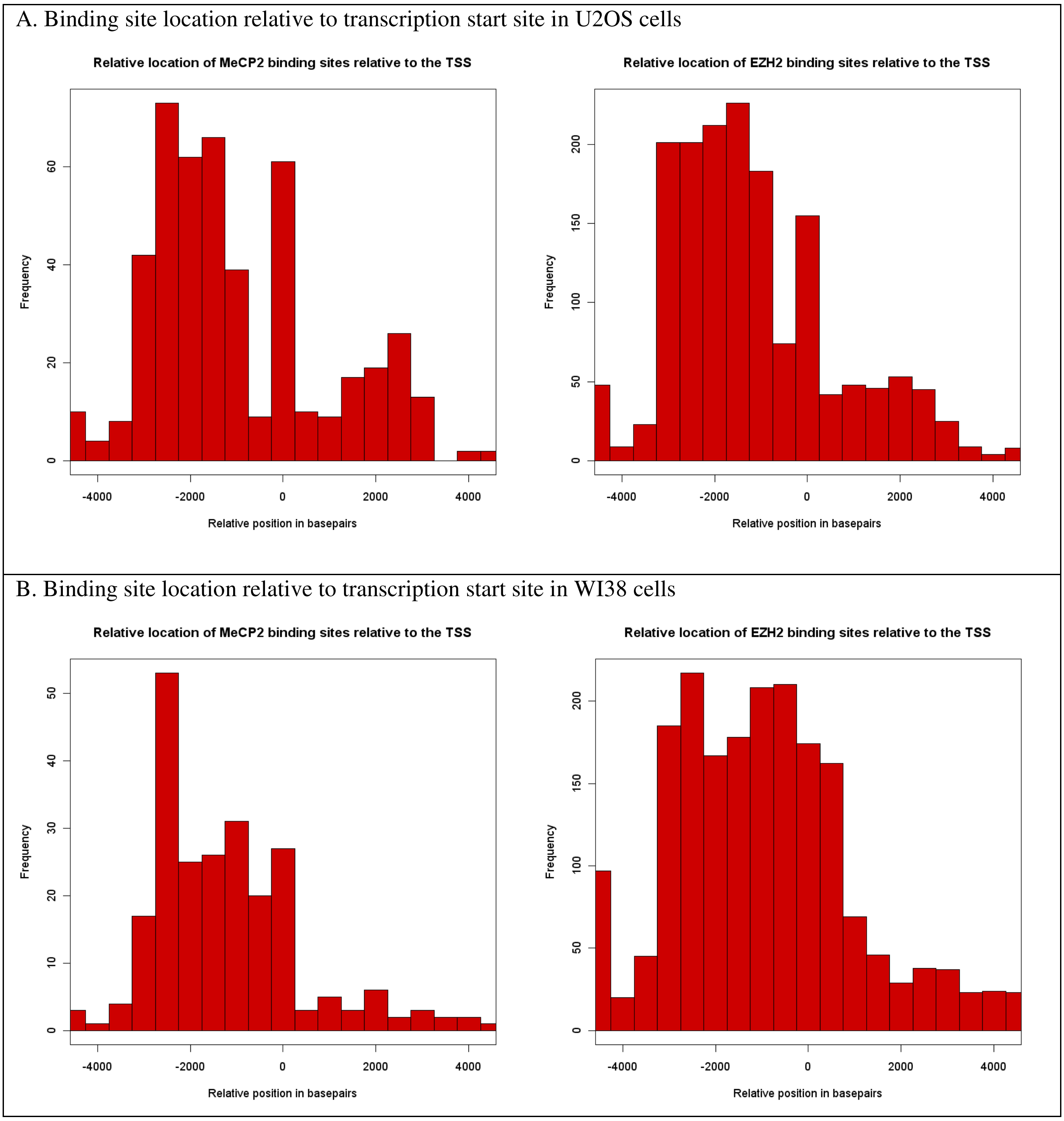
For the EZH2 datasets, the five percent threshold appeared to be too conservative. First, it would give rise to only 438 (U2OS) and 29 (WI38) significant peaks, respectively. These values substantially deviate from those reported previously for promoter binding by the SUZ12 protein (Pruitt et al. 2007), which forms the Polycomb repressive complex PRC2 together with EZH2 and other proteins. Second, when this criterion is applied, one fails to detect several known PRC2-bound genes that score highly but do not quite pass the five-percent threshold. Therefore, conventional ChIP was used to determine how many of the top-ranking peaks detected by ChIP-on-chip analysis should be regarded as EZH2 binding sites. The results indicated that for both cell lines the 2,000 top-ranking peaks qualify as high-confidence binding sites, with independent validation rates of 92% (U2OS) and 73% (WI38), respectively.

D-2.3 Results

Localization of MeCP2 and EZH2 binding sites relative to genes

To prepare for the analysis of co-binding by MeCP2 and EZH2 in the promoter regions of specific genes, the lists of high-confidence MeCP2 binding sites derived from ChIP-on-chip data were merged with gene annotation data as follows. First, a list of 23,001 distinct promoter regions was created that are covered by the promoter microarray and which map to regions on assembled nuclear chromosomes of the hg18 assembly of the human genome (NCBI36). Promoter regions mapping to the mitochondrial genome or to unassembled (“random”) chromosomal regions were discarded. Second, by pairwise comparison between this list and (i) the UCSC Known Genes annotation, (ii) the RefSeq gene annotation and (iii) the list of high-confidence MeCP2 binding sites (FDR < 5%), merged lists were created for U2OS and WI38, respectively.

On the basis of these lists and similar lists for EZH2, the location of MeCP2 and EZH2 binding sites relative to the closest transcription start site was investigated (Figure 36). As expected, both MeCP2 and EZH2 show a high tendency to bind upstream rather than downstream of the transcription start site. Furthermore, MeCP2 binding is most frequent further upstream (around 2,500 basepairs upstream of the transcription start site) than EZH2 binding, while the latter is observed equally frequently over a large region from 3,000 basepairs upstream down to the transcription start site. A second peak is observed for MeCP2 binding directly overlapping with the transcription start site.


A. Binding site location relative to transcription start site in U2OS cells

B. Binding site location relative to transcription start site in WI38 cells

Figure 36. Location of binding sites for MeCP2 and EZH2 relative to the transcription start site

This figure displays histograms for the distance of binding sites for MeCP2 (left column) and EZH2 (right column) relative to the closest transcription start site. Distance is measured from the center of the binding site as reported by peak detection, and orientation is relative to the direction of transcription of the corresponding gene, i.e. negative values indicate that the binding site is upstream of the transcription start site and positive values indicate that it is downstream. A value of zero refers to binding sites that overlap with the transcription start site.

Statistical and experimental analysis of co-binding by MeCP2 and EZH2

Two independent lines of evidence were used to assess the significance of co-binding by MeCP2 and EZH2 at the promoter regions of annotated genes:

(1) For the subset of promoter regions confirmed as MeCP2 binding sites during validation/threshold selection, co-binding by EZH2 was analyzed experimentally by conventional ChIP. In U2OS cells, co-binding by EZH2 was observed at 12 out of 20 validated MeCP2 targets (60%). Given an estimated number of 2000 EZH2 binding sites among all 23,001 promoters, this observation is highly significant ($P < 1.2 \cdot 10^{-8}$). In WI38 cells, co-binding by EZH2 was observed at 6 out of 11 validated MeCP2 targets (55%), which is also highly significant ($P < 1.3 \cdot 10^{-4}$). *P*-values were calculated by an exact binomial test for the probability of success in a Bernoulli experiment.

(2) In order to confirm the significance of overlap between MeCP2 and EZH2 on a genome-wide scale, the number of promoter regions was counted which are high-confidence binding sites of (i) both MeCP2 and EZH2, (ii) MeCP2 only, (iii) EZH2 only and (iv) neither. The overlap between MeCP2 and EZH2 was found to be highly significant for the U2OS cancer cell line as well as for the WI38 fibroblast cell line ($P < 10^{-15}$ in both cases). For U2OS, 40.1% of all MeCP2-bound promoters exhibited co-binding by EZH2, and the odds ratio measuring the over-representation of co-binding compared to random expectation was estimated at 9.62, with a 95% confidence interval ranging from 7.90 to 11.70 (Figure 37A). For WI38, 34.6% of all MeCP2-bound promoters exhibited co-binding by EZH2, and the odds ratio was estimated at 6.26, with a 95% confidence interval ranging from 4.63 to 8.39 (Figure 37B). Statistical significance was assessed using Fisher's exact test, which is based on the hypergeometric distribution. To exclude potential bias from promoter regions of different lengths, the analysis was repeated on a subset of 9,875 promoter regions which were all of size 4,250 bp, with similar results.

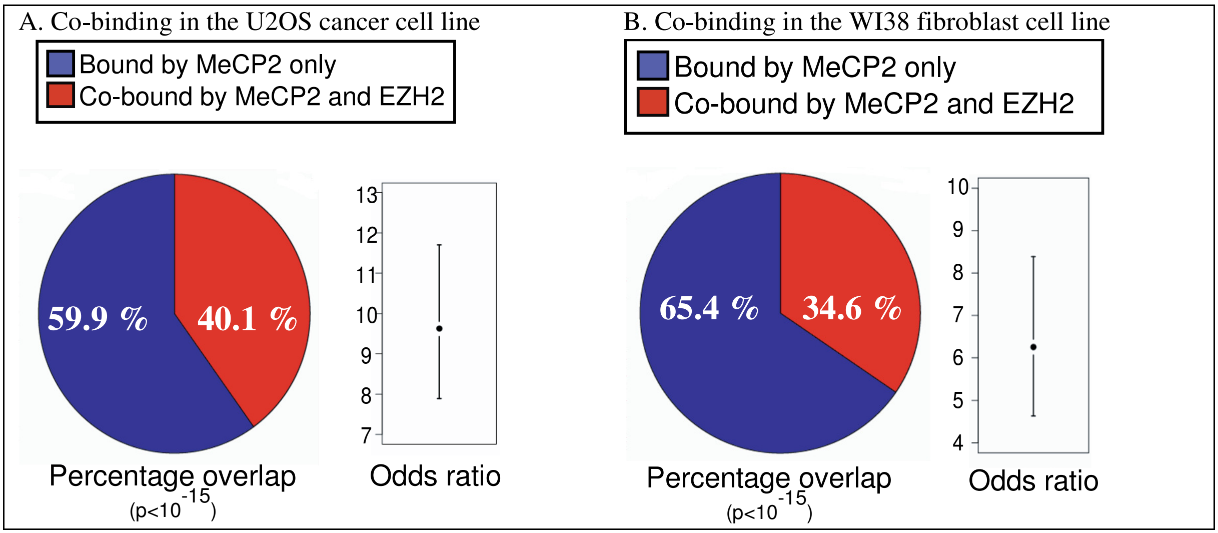


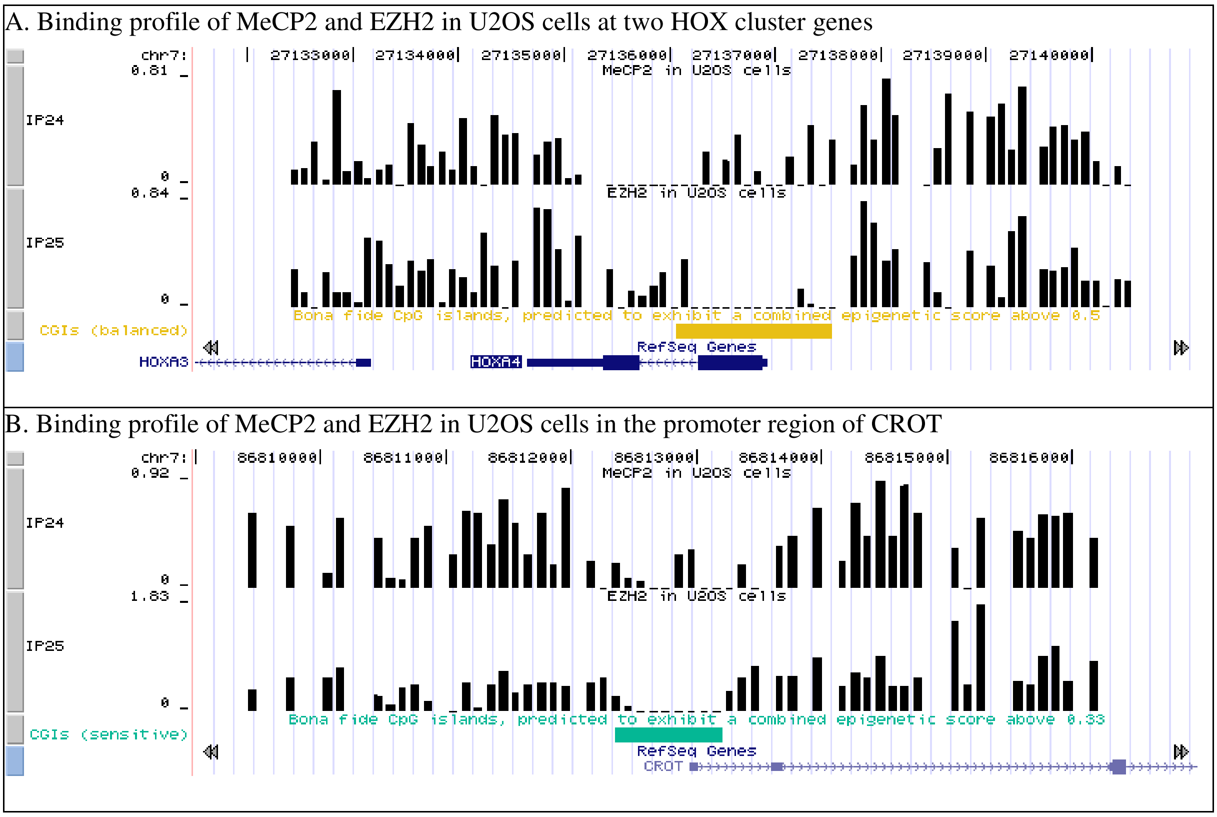
Figure 37. ChIP-on-chip experiments support a genome-wide link between MeCP2 and EZH2 binding

This figure summarizes the results of genome-wide ChIP-on-chip analysis for MeCP2 and EZH2 in the U2OS cancer cell line (panel A) and in the WI38 fibroblast cell line (panel B). The figure comprises pie charts of the percentage of MeCP2-bound promoters that are co-bound by EZH2 and error bar plots with the means and 95% confidence intervals for the odds ratio of EZH2 binding at MeCP-bound promoters. The results show that the binding frequency of EZH2 at MeCP2-bound promoters is highly significant and 6-fold to 10-fold increased over random expectation.

Characteristic binding schemes of MeCP2 and EZH2

The ChIP-on-chip datasets for MeCP2 and EZH2 binding in U2OS and WI38 can be used not only to infer significant co-binding of MeCP2 and EZH2 at a sizable fraction of human promoters, but also to investigate common binding schemes of these two proteins.

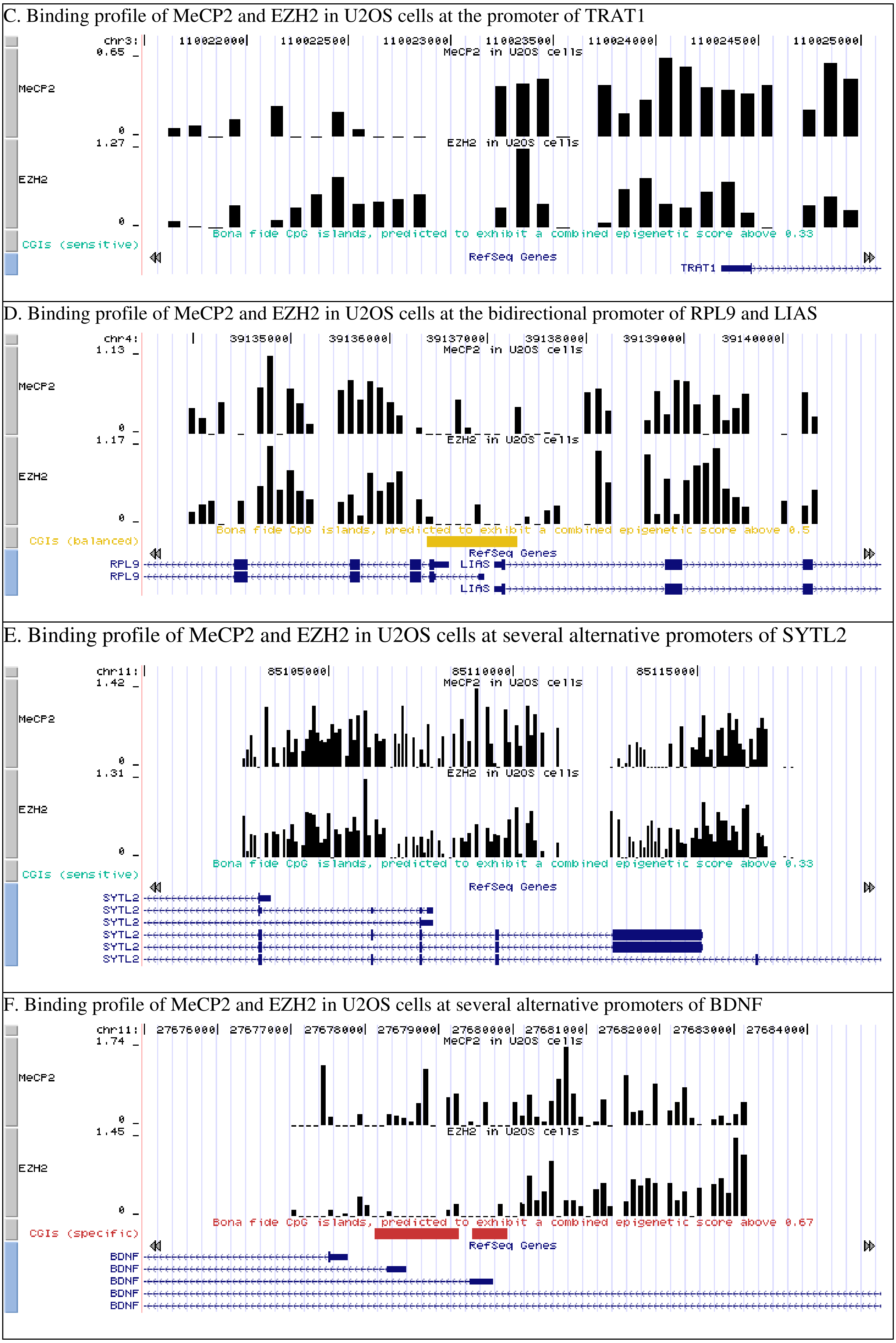
To that end, the ChIP-on-chip profiles for U2OS cells were visually inspected in the UCSC Genome Browser (Shahbazian and Zoghbi 2002), and exemplary cases were selected. Co-binding of MeCP2 and EZH2 frequently covered large parts of the promoter region, not only for single promoters (Figure 38A, B and C) but also for bidirectional promoters (Figure 38D) and for alternative promoters of a single gene (Figure 38E and F). However, both MeCP2 and EZH2 binding were excluded from CpG islands overlapping the transcription start site (Figure 38A, B, D and F), while binding was frequently observed at the start site of transcripts that did not exhibit a CpG island promoter (Figure 38C and E). Interestingly, the alternative promoters of SYTL2 (Figure 38E) differed in their levels of MeCP2 and EZH2 binding, indicating that selective binding of MeCP2 and EZH2 might regulate gene expression by influencing alternative promoter usage. Co-binding by MeCP2 and EZH2 was observed at a number of genes that have previously been related to MeCP2 binding or regulation by Polycomb group proteins. Examples include the HOX cluster genes HOXA3 and HOXA4 (Figure 38A) as well as the well-known MeCP2 target gene BDNF (Figure 38F).


A. Binding profile of MeCP2 and EZH2 in U2OS cells at two HOX cluster genes

B. Binding profile of MeCP2 and EZH2 in U2OS cells in the promoter region of CROT

Figure 38. Exemplary binding schemes of MeCP2 and EZH2 in the U2OS cancer cell line

This figure displays ChIP-on-chip profiles derived for MeCP2 and EZH2 binding in U2OS cells. Data visualization is based on the custom track feature of the UCSC Genome Browser (Karolchik et al. 2008). Negative values are set to zero. The CpG island track is based on CpG island mapping and classification into strong (red), moderate (yellow) and weak bona fide CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis). The gene track is based on RefSeq gene annotations (Pruitt et al. 2007).



C. Binding profile of MeCP2 and EZH2 in U2OS cells at the promoter of TRAT1

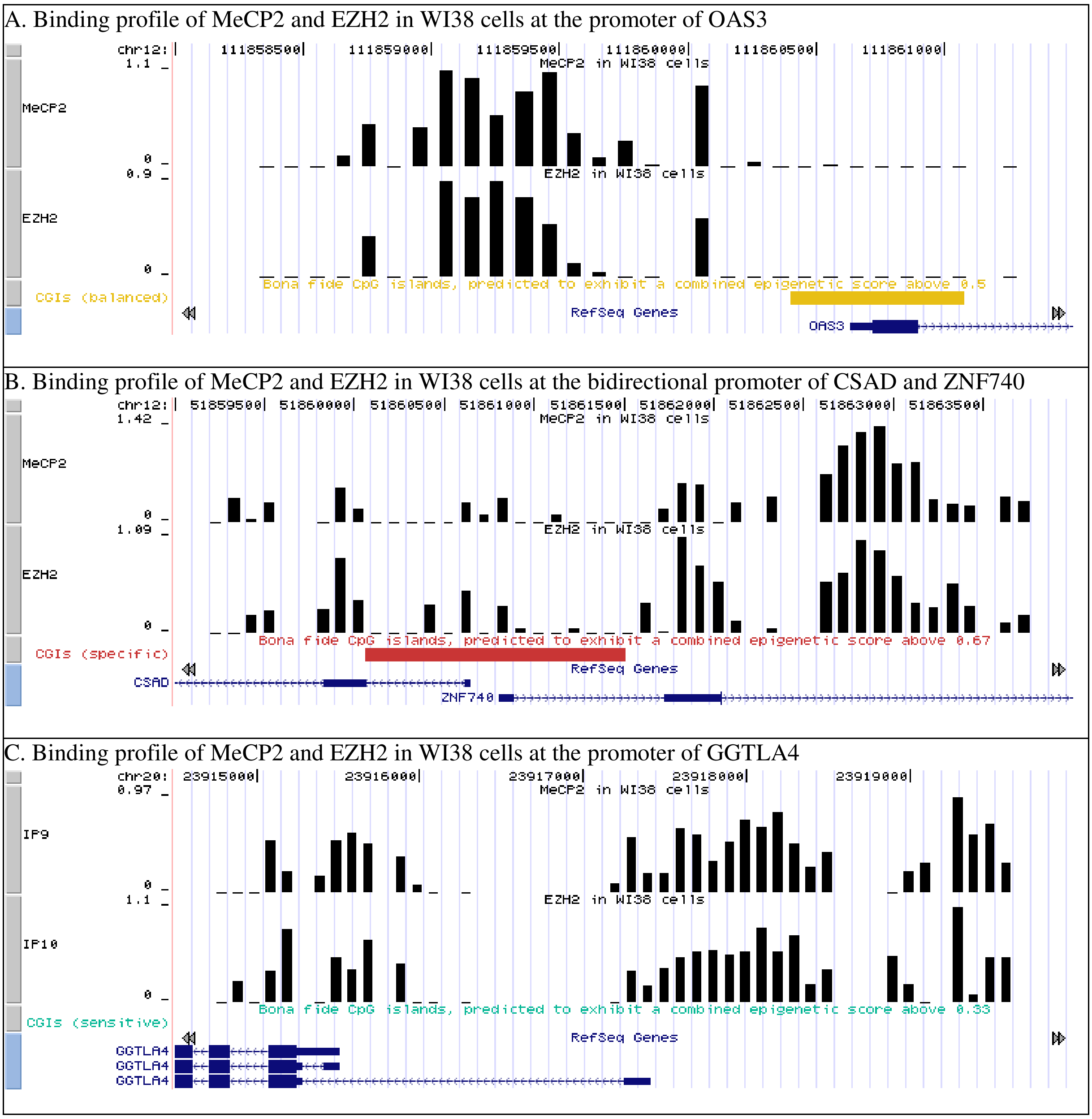
D. Binding profile of MeCP2 and EZH2 in U2OS cells at the bidirectional promoter of RPL9 and LIAS

E. Binding profile of MeCP2 and EZH2 in U2OS cells at several alternative promoters of SYTL2

F. Binding profile of MeCP2 and EZH2 in U2OS cells at several alternative promoters of BDNF

Figure 38 (continued).

Manual inspection was also conducted on the ChIP-on-chip profiles corresponding to WI38 cells, and exemplary cases are highlighted in Figure 39. First, while there was clear evidence of cell-type-specific MeCP2 and EZH2 binding, the common binding schemes observed in WI38 cells were similar to those in U2OS cells. Second, the overall degree of MeCP2 binding appeared to be slightly lower in the WI38 ChIP-on-chip profile than in the U2OS dataset, as is exemplified by SYTL2 (Figure 39E) and the well-known MeCP2 target gene BDNF (Figure 39F). Third, OAS3 (Figure 39A) provides an example of a gene showing binding depletion at a promoter CpG island together with a strong MeCP2 and EZH2 binding peak upstream of the transcription start site, consistent with the histograms reported in Figure 36.

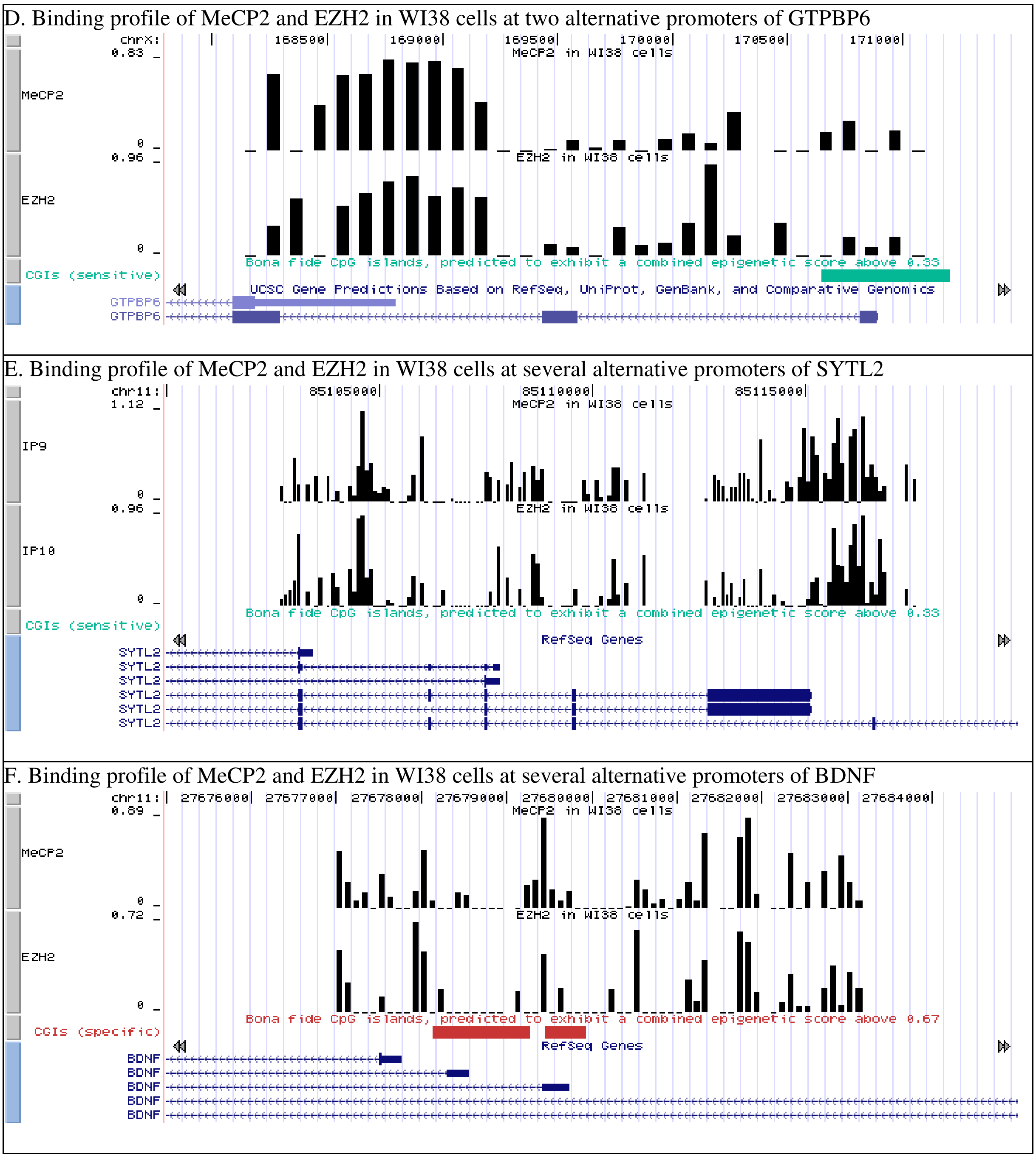

A. Binding profile of MeCP2 and EZH2 in WI38 cells at the promoter of OAS3

B. Binding profile of MeCP2 and EZH2 in WI38 cells at the bidirectional promoter of CSAD and ZNF740

C. Binding profile of MeCP2 and EZH2 in WI38 cells at the promoter of GGTLA4

Figure 39. Exemplary binding schemes of MeCP2 and EZH2 in the WI38 fibroblast cell line.

This figure displays ChIP-on-chip profiles derived for MeCP2 and EZH2 binding in WI38 cells. Data visualization is based on the custom track feature of the UCSC Genome Browser (Karolchik et al. 2008). Negative values are set to zero. The CpG island track is based on CpG island mapping and classification into strong (red), moderate (yellow) and weak bona fide CpG islands (Bock et al. 2007, cf. chapter B-4 of this thesis). The gene track is based on RefSeq gene annotations (Pruitt et al. 2007), and in one case (D), for which the RefSeq annotation misses a plausible alternative transcript, on the UCSC Known Genes annotation (Hsu et al. 2006).



D. Binding profile of MeCP2 and EZH2 in WI38 cells at two alternative promoters of GTPBP6

E. Binding profile of MeCP2 and EZH2 in WI38 cells at several alternative promoters of SYTL2

F. Binding profile of MeCP2 and EZH2 in WI38 cells at several alternative promoters of BDNF

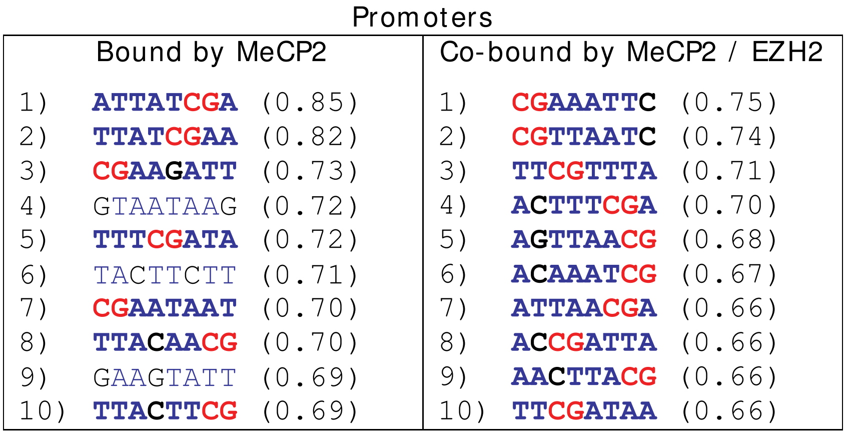
Figure 39 (continued).

Characteristic DNA sequences associated with MeCP2 and EZH2 binding

MeCP2 has been shown to bind specifically to methylated cytosines, rather than to a particular DNA sequence motif (Shahbazian and Zoghbi 2002). Nevertheless, a recent report provides in vitro evidence for preferential MeCP2 binding to specific DNA sequences containing at least four A/T nucleotides in direct vicinity of the methylated CpG (Klose et al. 2005). Klose et al. also reported moderate enrichment of such DNA sequence motifs when cloning and sequencing DNA from ChIP experiments for MeCP2 in embryonic fibroblasts. To confirm the in vivo relevance of the preference of MeCP2 for binding to A/T-rich genomic regions, and also to assess whether it might play a role in MeCP2-targeted EZH2 binding, bioinformatic motif discovery was performed on promoters that were bound by MeCP2 and EZH2 according to the ChIP-on-chip data for U2OS cells.

Briefly, DNA sequences were obtained for the 1 kilobase region directly upstream of the transcription start site of genes whose promoters are bound by (i) MeCP2 or (ii) MeCP2 and EZH2, giving rise to two lists of DNA sequences. These lists were randomly down-sampled to fifty sequences in order to reduce computational demand and were then analyzed by the Weeder motif discovery algorithm (Pavesi et al. 2004). Weeder is an enumerative method for pattern discovery that can accommodate degenerate motifs and performed well in a recent benchmarking study (Tompa et al. 2005). In order to assess the variance introduced by down-sampling, the analysis was repeated three times with random sampling. The results were qualitatively comparable, although the top-scoring motifs differed, as is the rule rather than the exception for motif discovery. The results reported in Figure 40 are based on the first run of the analysis.

Among the ten top-scoring motifs for MeCP2 binding were seven motifs that meet the criteria of Klose et al., i.e. containing at least one CpG and at least four A/T nucleotides (Figure 40). In contrast, fewer than 1.5 motifs meeting these criteria can be expected in a random selection of patterns. Intriguingly, preference for AT-rich DNA sequences surrounding a single CpG dinucleotide was also characteristic of promoters that are co-bound by MeCP2 and EZH2. In fact, all ten top-scoring motifs for promoters co-bound by MeCP2 and EZH2 are A/T-rich and contain a CpG dinucleotide.



Promoters	
Bound by MeCP2	Co-bound by MeCP2 / EZH2
1) **ATTATCGA** (0.85)	1) **CGAAATTC** (0.75)
2) **TTATCGAA** (0.82)	2) **CGTTAATC** (0.74)
3) **CGAAGATT** (0.73)	3) **TTCGTTTA** (0.71)
4) GTAATAAG (0.72)	4) **ACTTTCGA** (0.70)
5) **TTTCGATA** (0.72)	5) **AGTTAACG** (0.68)
6) TACTTCTT (0.71)	6) **ACAAATCG** (0.67)
7) **CGAATAAT** (0.70)	7) **ATTAACGA** (0.66)
8) **TTACAACG** (0.70)	8) **ACCGATTA** (0.66)
9) GAAGTATT (0.69)	9) **AACTTACG** (0.66)
10) **TTACTTCG** (0.69)	10) **TTCGATAA** (0.66)

Figure 40. Enriched DNA sequence patterns in promoters bound by MeCP2 and EZH2

This figure depicts the ten most significant DNA motifs detected in the upstream sequence (1 kb directly upstream from the transcription start site) of promoter regions bound by MeCP2 (left column) or co-bound by MeCP2 and EZH2 (right column). All motifs were discovered by the Weeder algorithm (Pavesi et al. 2004) and are highly statistically significant ($P < 0.001$, based on permutation testing). DNA motifs containing at least four A/T nucleotides and one CpG dinucleotide are highlighted in bold print. Such sequences were reported to be particularly amenable to MeCP2 binding in vitro (Klose et al. 2005).

D-2.4 Discussion

Through genome-wide ChIP-on-chip analysis of MeCP2 and EZH2 binding in the U2OS cancer cell line and in the WI38 fibroblast cell line, as well as through siRNA knockdown experiments at specific loci (data not shown), we could confirm that MeCP2 plays a functional role in recruiting EZH2 for a significant subset of its target genes. Although this protein interaction can explain only a small percentage of Polycomb binding sites in the human genome, our results shed light on the poorly understood mechanisms by which mammalian Polycomb repressive complexes are targeted to specific promoter regions. In the context of our model of the epigenome being partially determined by the underlying DNA sequence (cf. Part B and Part C of this thesis for empirical evidence and section E-2.1 for discussion), it is interesting to observe that a subset of EZH2 binding sites are enriched for specific DNA sequence motifs that are preferentially bound by MeCP2 (Figure 40). Hence, we may be observing a mechanism by which DNA sequence specificity is conferred to a protein complex that does not exhibit DNA sequence specificity by itself (PRC2), via a transiently bound recruiting factor that is sequence-specific (MeCP2). Given that MeCP2 is unlikely to be the only transcription factor being able to recruit PRC2 to specific promoter regions, this additional level of complexity might be a reason why Polycomb response elements, i.e. DNA sequence motifs that result in reproducible binding by Polycomb repressive complexes (Sparmann and van Lohuizen 2006), have so far eluded detection in mammals.

Furthermore, our data support a model of epigenetic repression that may help explain how specific promoter regions become aberrantly silenced in cancer cells: In previous research it was shown that Polycomb binding can lead to recruitment of DNA methyltransferases and de novo DNA methylation at specific promoter (Viré et al. 2006), and our current results indicate that the methyl-binding protein MeCP2 can, in turn, recruit Polycomb binding. Hence, these two mechanisms could give rise to a self-propagating feedback loop enforcing long-term transcriptional repression of specific genes. According to our data, this feedback loop seems to be more dominant in the U2OS cancer cell line than in normal fibroblasts, consistent with previous papers reporting high correlation between aberrant DNA methylation and Polycomb binding specifically for cancer cells (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007) and with a recent study confirming the finding of Viré et al. (Viré et al. 2006) in leukemic cells (Villa et al. 2007). Hence, defects and de-regulation of the proteins involved in these two mechanisms of induced epigenetic repression are prime candidates for a causal role in aberrant DNA methylation.

In conclusion, our study highlights the relevance of bioinformatic analysis for elucidating the mechanisms of epigenetic gene regulation in cancer cells. While biochemical methods such as protein-interaction assays and knock-down experiments are critical for establishing causal effects, these methods are technically cumbersome and are only for a small number of target genes. Bioinformatic methods in connection with genome-wide ChIP-on-chip data can supplement functional evidence for selected target genes by correlative evidence at a genomic scale, thus confirming or refuting genome-wide biological relevance for newly discovered mechanisms of epigenetic gene regulation.

D-3 Optimizing a DNA-methylation-based biomarker of chemotherapy resistance for use in clinical settings[1]

D-3.1 Motivation

Chemotherapy is an important treatment option for most cancers. Alkylating agents, which are the most widely used class of chemotherapeutic drugs, induce extensive DNA damage and can kill cancer cells during various phases of the cell cycle. However, a significant percentage of tumors are resistant to alkylating chemotherapy, which has been related to the activity of

[1] This chapter describes work conducted in collaboration with Thomas Mikeska and Andreas Waha (Mikeska et al. 2007). Thomas Mikeska performed and evaluated the wet-lab experiments, while both Thomas Mikeska and Andreas Waha contributed to the interpretation of the results.

DNA repair genes such as MGMT (Gerson 2004). The human O6-methylguanine DNA methyltransferase (MGMT) gene encodes a protein that can remove alkyl groups from the O6-position of guanine, which significantly reduces the amount of DNA damage induced by alkylating agents. Because the DNA repair reaction consumes the protein, high levels of MGMT expression are instrumental for chemotherapy resistance against alkylating agents. In contrast, low levels of MGMT expression, e.g. as the result of aberrant hypermethylation of the MGMT's promoter region, can make tumors susceptible to alkylating chemotherapy (Gerson 2004).

Recent clinical trials convincingly confirmed the role of MGMT expression for chemotherapy resistance in gliomas (Esteller et al. 2000) as well as glioblastomas (Hegi et al. 2004; Hegi et al. 2005), and they could establish strong correlation between promoter hypermethylation of MGMT and resistance to alkylating agents. Specifically, Hegi et al. observed that treatment with temozolomide significantly increased the median survival (from 15.3 months to 21.7 months) in the 45% subset of patients in which the promoter of MGMT was found to be hypermethylated, while no significant difference was observed for those patients in which the promoter was unmethylated and MGMT was presumably expressed at normal rate (Hegi et al. 2005). Strong clinical evidence therefore supports the routine use of MGMT promoter hypermethylation as a biomarker predicting resistance to alkylating chemotherapy in glioblastomas, which would lead to more informed treatment decisions and improved therapy.

However, the clinical trials by Esteller et al. and Hegi et al. used methylation-specific PCR (MSP) to assess the DNA methylation status of the MGMT promoter region, which significantly impedes routine use in clinical settings. First, MSP is a qualitative method that can only detect presence or absence of one specific DNA methylation pattern, which poses a significant risk of false-positive and false-negative results. Second, MSP is not robust with respect to low, and highly variable, levels of DNA quality, which is the rule rather than the exception in routine clinical diagnosis. Third, MSP does not perform well on formalin-fixed, paraffin-embedded (FFPE) specimens, which increases the cost of sample processing and storage and impedes quick testing of MGMT's potential as a biomarker for other cancers because most archival tumor samples with known clinical history are stored as FFPE specimens.

Therefore, we sought to develop an alternative method for analyzing MGMT promoter hypermethylation that overcomes all of these limitations, and which is cheap, fast and robust enough to be used in routine clinical diagnosis. Based on three experimental methods that are well-suited for clinical settings, several dozen candidate biomarkers were constructed and statistically evaluated on high-quality DNA methylation data obtained by clonal bisulfite sequencing. Next, for each of the three methods, the optimal candidate biomarker was experimentally tested. Finally, we constructed and validated logistic regression models that predict the status of MGMT promoter hypermethylation based on the results from experimental assays that are fully adequate for clinical settings. In the following, we focus on the statistical and bioinformatic aspects of this study, describing experimental details only insofar as they are indispensable for proper understanding of the computational part.

D-3.2 Methods

A four-step process was devised to optimize the analysis of MGMT promoter methylation for routine clinical use as a predictor of chemotherapy resistance. First, DNA was extracted from 22 snap-frozen primary glioblastoma samples and from three snap-frozen normal brain controls, and it was subjected to bisulfite sequencing. Based on DNA methylation patterns for an extensive region covering the transcription start site, the first exon and parts of the first intron of the MGMT gene (Figure 41), tumors were clustered into methylated and unmethylated cases. Second, we determined a list of candidate biomarkers that are feasible with on of the following methods: COBRA (**Co**mbined **B**isulfite **R**estriction **A**nalysis) (Xiong and Laird 1997), SIRPH (**S**NuPE **IP**-**RP** **H**PLC) (El-Maarri et al. 2002) and bisulfite pyrosequencing (Colella et al. 2003). Third, each candidate biomarker was statistically evaluated on all DNA methylation profiles and experimentally confirmed by performing the assay on samples with known DNA methylation patterns. Fourth, we statistically optimized each biomarker and assessed its accuracy and robustness. For details of the experimental methods, we refer to the published paper (Mikeska et al. 2007).

Statistical analysis was performed using the SPSS statistics package (SPSS for Windows, Chicago: SPSS Inc). Hierarchical clustering was based on the DNA methylation averages and standard deviations of all CpG positions 1 to 25, calculated over all sequenced clones. Between-groups average linkage was used with squared Euclidean distance as interval measure. Logistic regression models were calculated with the WEKA machine learning toolkit (Frank et al. 2004) using default parameters. Prediction accuracy was estimated using leave-one-out cross-validation, i.e. by repeatedly training a logistic regression model on 12 out of the 13 available cases and testing it on the single remaining case.

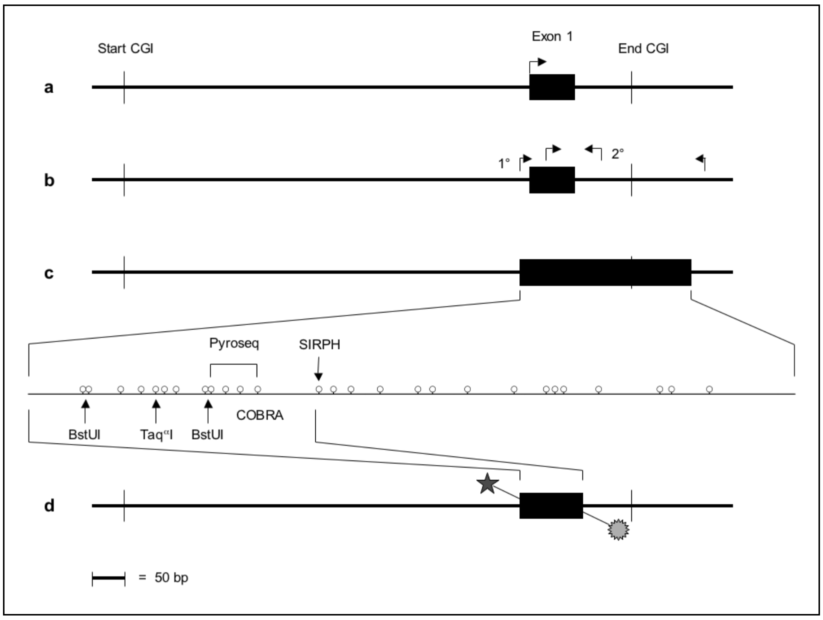


Figure 41. Overview of the promoter region of the MGMT gene

This schematic figure displays the structure of the promoter region of MGMT as well as the location of several candidate biomarkers. Panel A displays the location of a CpG island (CGI) that spans not only the promoter region of MGMT but also the first exon and parts of the first intron. Panel B displays the location of the primers used by the nested PCR approach of Hegi et al. (Hegi et al. 2005). Panel C displays the location of the genomic region analyzed by clonal bisulfite sequencing (each circle corresponds to one CpG dinucleotide), highlighting those CpGs that are accessible to COBRA, SIRPH and bisulfite pyrosequencing. Panel D displays the location of the PCR product used for the COBRA assay. This figure was prepared by Thomas Mikeska.

D-3.3 Results

Bisulfite sequencing of the MGMT promoter region

Our goal is to select a clinically applicable DNA methylation assay for the MGMT promoter region that is highly correlated with the overall state of promoter methylation. To obtain a solid basis for the necessary selection and optimization step, we first established DNA methylation patterns for 22 glioblastomas and three normal brain controls using clonal bisulfite sequencing (due to its costly and time-consuming protocol, clonal bisulfite sequencing is not suitable for clinical settings, but its accuracy and resolution make it the gold standard for DNA methylation mapping). The genomic region analyzed by clonal bisulfite sequencing spans 266 bp and 27 CpG positions, overlapping with the transcription start site, the first exon as well as parts of the first intron of MGMT (Figure 41). This region also comprises the CpGs used by Hegi et al. in their MSP approach (CpG positions 5 to 9 and 13 to 16). A representative set of DNA methylation patterns for tumor and control samples – processed using BiQ Analyzer (Bock et al. 2005, cf. chapter C-2 of this thesis) – is shown in Figure 44. For further analysis we calculated the average DNA methylation profile for each of the 25 samples (i.e. the vector of DNA methylation levels in percent for all CpGs in the promoter region of MGMT).

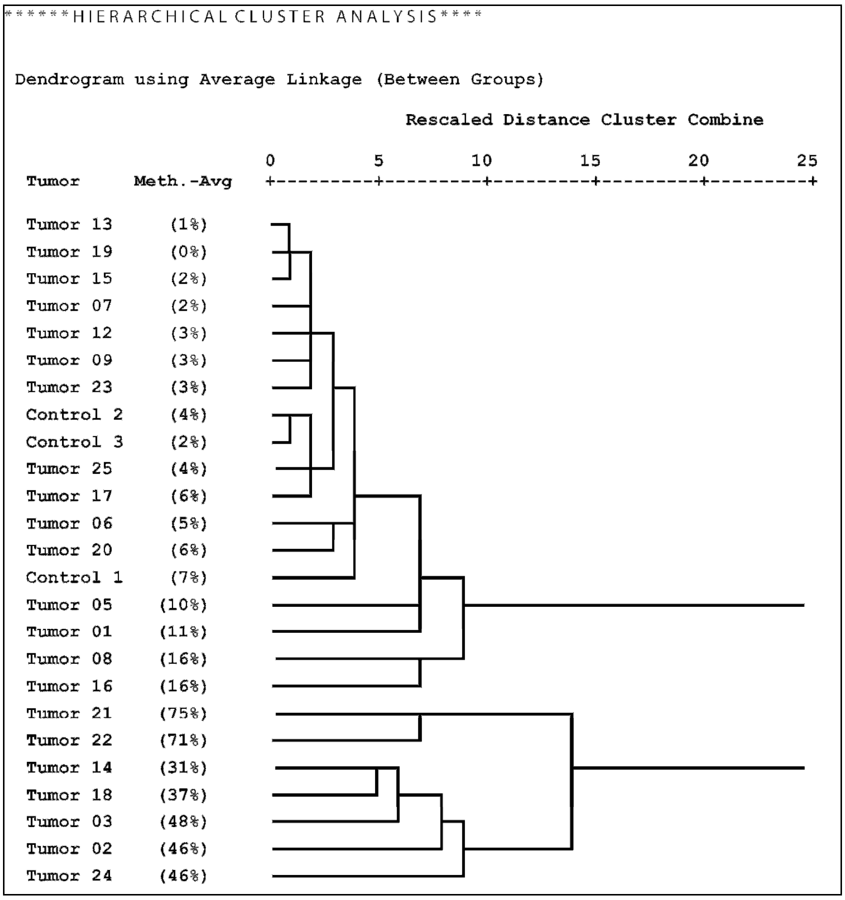
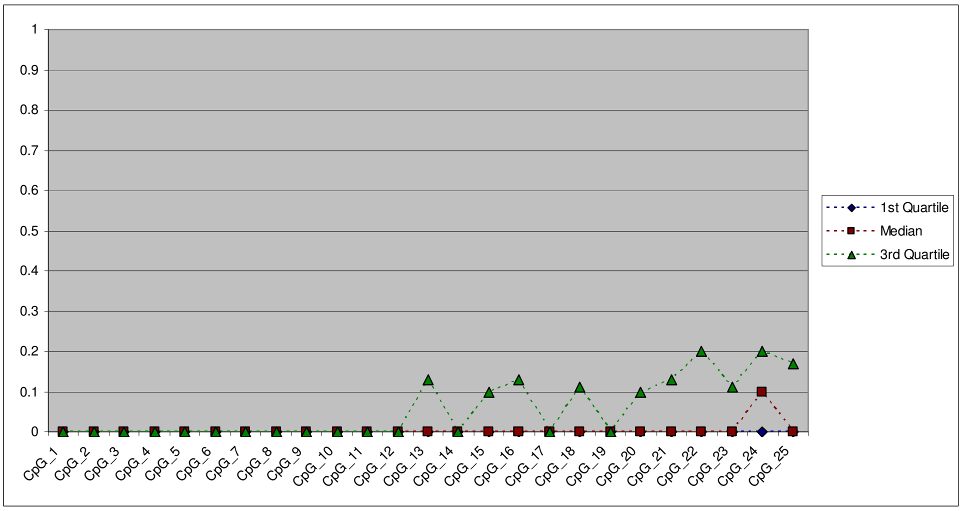

```
******HIERARCHICAL CLUSTER ANALYSIS****

Dendrogram using Average Linkage (Between Groups)

                              Rescaled Distance Cluster Combine

                       0         5        10        15        20        25
Tumor      Meth.-Avg   +---------+---------+---------+---------+---------+

Tumor 13     (1%)
Tumor 19     (0%)
Tumor 15     (2%)
Tumor 07     (2%)
Tumor 12     (3%)
Tumor 09     (3%)
Tumor 23     (3%)
Control 2    (4%)
Control 3    (2%)
Tumor 25     (4%)
Tumor 17     (6%)
Tumor 06     (5%)
Tumor 20     (6%)
Control 1    (7%)
Tumor 05    (10%)
Tumor 01    (11%)
Tumor 08    (16%)
Tumor 16    (16%)
Tumor 21    (75%)
Tumor 22    (71%)
Tumor 14    (31%)
Tumor 18    (37%)
Tumor 03    (48%)
Tumor 02    (46%)
Tumor 24    (46%)
```

Figure 42. Hierarchical clustering of average DNA methylation profiles

This figure displays a clustering tree for 22 glioblastoma samples and three normal brain controls, based on profiles of MGMT promoter methylation. Hierarchical clustering was performed on vectors combining the DNA methylation means and standard deviations over all individual clones at CpG positions 1 to 25 for each sample.

As expected, hierarchical clustering of the 25 DNA methylation profiles gives rise to two well-separated clusters of samples with low vs. high levels of DNA methylation (Figure 42). The first cluster consists of tumor samples 01, 05, 06, 07 08, 09, 12, 13, 15, 16, 17, 19, 20, 23 and 25, as well as the largely unmethylated normal brain controls 1, 2 and 3. The second cluster contains tumor samples 02, 03, 14, 18, 21, 22 and 24, which exhibit significant methylation levels. Tumor 08 and 16, members of the first cluster, are most accurately described as borderline cases. They exhibit few heavily methylated alleles among a large number of unmethylated alleles (displayed in Figure 44A and C, respectively). Therefore they are qualitatively different from the other tumor samples in the first cluster, which exhibit only sporadic DNA methylation (e.g. Tumor 13, displayed in Figure 44H) that is also observed among the control samples (e.g. Control 3, displayed in Figure 44J). This behavior is also apparent from the dendrogram, in which Tumor 08 and Tumor 16 are localized between the clearly unmethylated and the clearly methylated samples. Based on these peculiarities and the fact that higher variation is present in the methylated cluster, we decided to assign these two intermediate cases to the methylated cluster when calculating DNA methylation profiles. Consequently, tumor samples 01, 05, 06, 07, 09, 12, 13, 15, 17, 19, 20, 23, and 25 are classified as unmethylated samples, while samples 02, 03, 08, 14, 16, 18, 21, 22, and 24 are classified as methylated samples. Average methylation profiles for both clusters (excluding normal brain controls) are displayed in Figure 43.

A. DNA methylation profile of the MGMT promoter region of tumors classified as unmethylated



B. DNA methylation profile of the MGMT promoter region of tumors classified as methylated

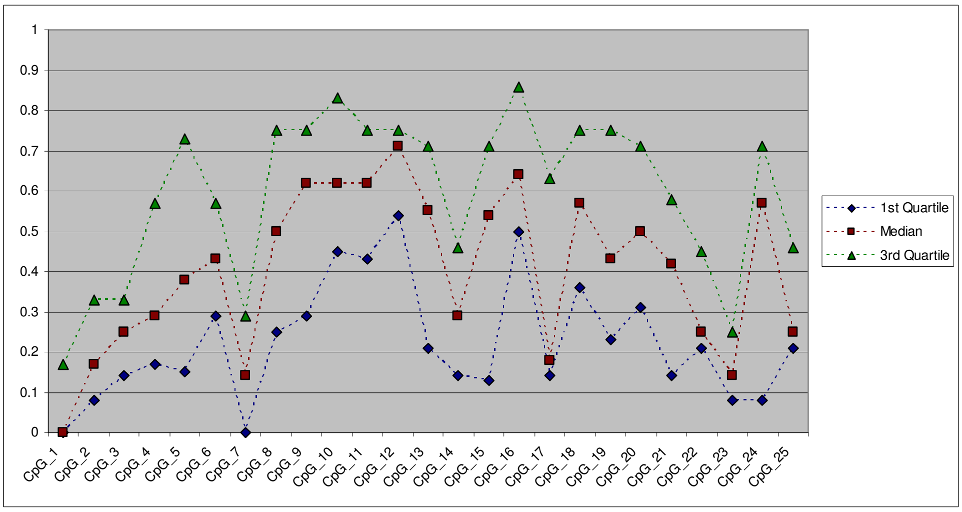


Figure 43. DNA methylation profiles of tumors according to their clustering into cases with unmethylated (panel A) and methylated (panel B) MGMT promoter regions

This figure displays DNA methylation profiles averaged over all tumor samples in the low-methylation cluster (panel A) and in the high-methylation cluster (panel B), respectively.

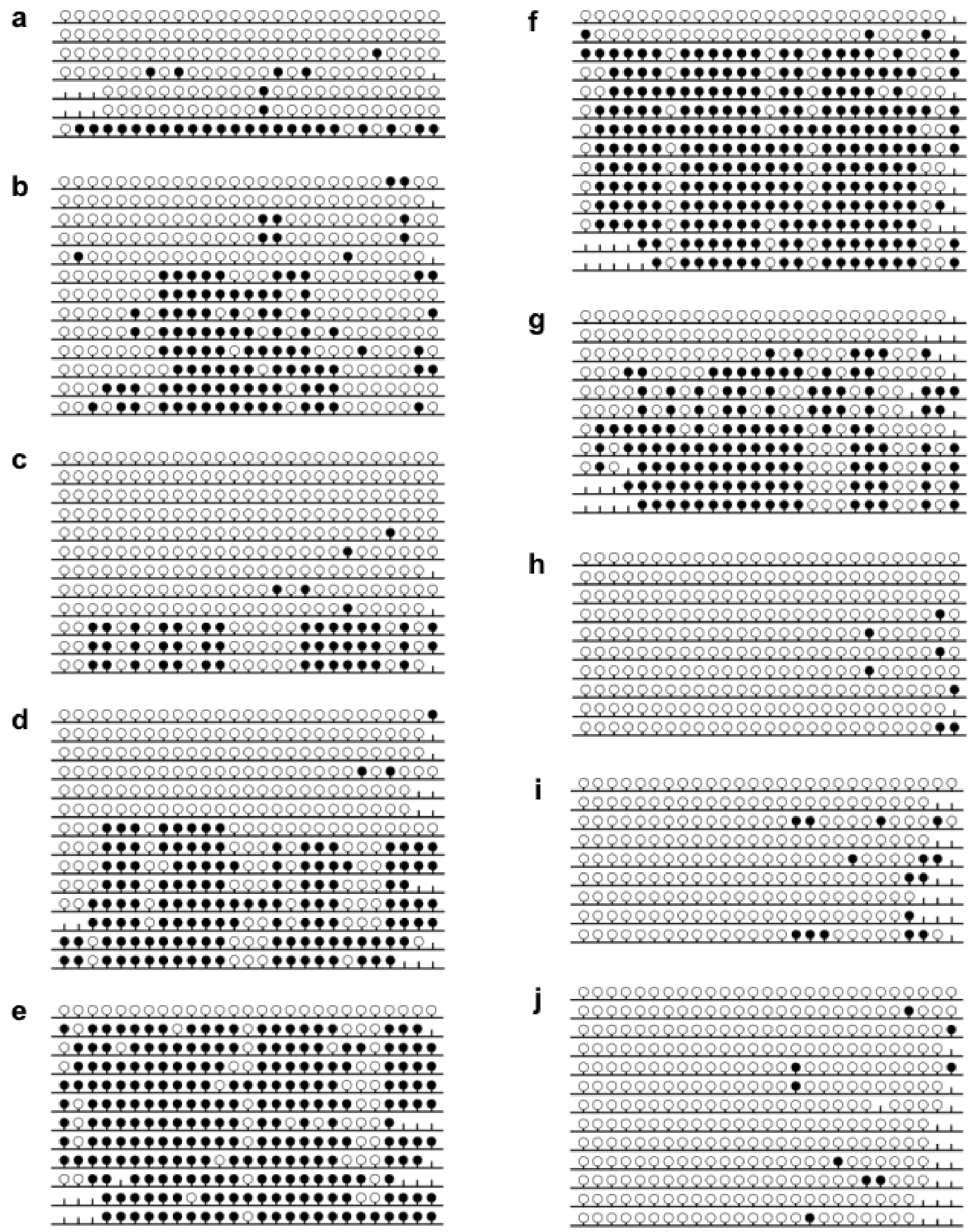


Figure 44. DNA methylation patterns of the MGMT promoter region obtained by clonal bisulfite sequencing

This figure displays a representative set of DNA methylation patterns obtained by clonal bisulfite sequencing (a: Tumor 08, b: Tumor 14, c: Tumor 16, d: Tumor 18, e: Tumor 21, f: Tumor 22, g: Tumor 24, h: Tumor 13, i: Tumor 20 and j: Control 3). Filled circles correspond to methylated CpGs, unfilled circles correspond to unmethylated CpGs and the vertical lines without a circle correspond to missing values. The diagrams were generated with the BiQ Analyzer software (Bock et al. 2005, cf. chapter C-2 of this thesis) and the figure was prepared by Thomas Mikeska.

Construction of candidate biomarkers for COBRA, SIRPH and pyrosequencing

From the DNA methylation profiles of the unmethylated and methylated tumor clusters (Figure 43), we concluded that the DNA methylation states of all CpGs at positions 2 to 6 and positions 8 to 13 are likely to be accurate predictors of the average level of amplicon methylation (a Pearson correlation coefficient above 0.8 was observed between the CpG methylation level and the amplicon methylation level for all of these CpGs). We therefore focused on these CpGs and determined for each position whether it can be readily analyzed by at least one of the three clinically applicable assays included in this study. For COBRA, we used a combination of two restriction enzymes, Taq^{a}I and BstUI, which enables us to assess DNA methylation at positions 1 and 2 simultaneously (median DNA methylation in methylated samples: 0% and 17%, respectively), at position 5 (median DNA methylation in methylated samples: 38%) and at positions 8 and 9 simultaneously (median DNA methylation in methylated samples: 50% and 62%, respectively). With SIRPH, only position 13 could be targeted (median DNA methylation in methylated samples: 55%). Pyrosequencing enabled us to assess DNA methylation at position 9, 10, 11 and 12 simultaneously (median DNA methylation in methylated samples in the range of 62% to 71%). By combining the positions accessible to each method is several ways, we obtained a total number of 23 candidate biomarkers for further analysis (Table 13).

			Correlation for score calculated from bisulfite data				Correlation for score calculated from experimental evaluation			
Experimental method	Marker ID	Analyzed CpG positions	Pearson's *r*	Perfor-mance	Spearman's rho	Perfor-mance	Pearson's *r*	Perfor-mance	Spearman's rho	Perfor-mance
COBRA	CO1	1 to 2 (comb.)	0.696	–	0.554	–	0.433	–	0.670	–
	CO2	8 to 9 (comb.)	0.887	–	0.804	–	0.938	+	0.879	+
	CO3	5	0.928	–	0.857	–	0.933	+	0.856	+
	CO4	1 to 2 (comb.), 8 to 9 (comb.)	0.927	–	0.881	Ø	0.916	+	0.875	+
	CO5	1 to 2 (comb.), 5	0.937	Ø	0.869	Ø	0.926	+	0.857	+
	CO6	8 to 9 (comb.), 5	0.949	Ø	0.908	+	0.947	+	0.856	+
	CO7*	1 to 2 (comb.), 8 to 9 (comb.), 5	0.961	+	0.929	+	0.942	+	0.856	+
SIRPH	SI01*	13	0.955	Ø	0.888	Ø	0.908	Ø	0.844	Ø
Pyrosequencing	Py01	9	0.877	–	0.734	–	0.901	Ø	0.838	Ø
	Py02	10	0.894	–	0.817	–	0.886	Ø	0.792	Ø
	Py03	11	0.962	+	0.898	+	0.867	–	0.740	–
	Py04	12	0.959	Ø	0.910	+	0.872	–	0.825	Ø
	Py05	9, 10	0.950	Ø	0.868	Ø	0.903	Ø	0.792	Ø
	Py06	9, 11	0.935	Ø	0.858	–	0.886	–	0.774	–
	Py07	9, 12	0.952	Ø	0.867	Ø	0.888	Ø	0.825	Ø
	Py08	10, 11	0.954	Ø	0.886	Ø	0.895	Ø	0.748	–
	Py09	10, 12	0.943	Ø	0.888	Ø	0.887	Ø	0.796	Ø
	Py10	11, 12	0.967	+	0.898	+	0.873	–	0.758	–
	Py11	9, 10, 11	0.958	Ø	0.872	Ø	0.898	Ø	0.787	Ø
	Py12	9, 10, 12	0.962	+	0.874	Ø	0.895	Ø	0.792	Ø
	Py13	9, 11, 12	0.957	Ø	0.863	Ø	0.883	–	0.793	Ø
	Py14	10, 11, 12	0.961	+	0.887	Ø	0.888	Ø	0.769	–
	Py15*	9, 10, 11, 12	0.964	+	0.873	Ø	0.892	Ø	0.784	Ø
Minimum			0.696		0.554		0.433		0.670	
1st quartile			0.932		0.861		0.886		0.779	
Median			0.952		0.873		0.895		0.793	
3rd quartile			0.960		0.888		0.912		0.850	
Maximum			0.967		0.929		0.947		0.879	

Table 13. Statistical evaluation of candidate biomarkers assessing MGMT promoter methylation

This table summarizes the correlation between the overall DNA methylation level of the MGMT promoter region on the one hand and the candidate biomarker scores (before optimization) on the other hand. In columns four to seven, scores are calculated based on bisulfite sequencing data for all tumor samples, whereas the scores in the four rightmost columns are based on the experimental results of the candidate biomarkers applied on tumor samples 12 to 25. All correlation coefficients but one (0.433) are significantly different from zero ($P < 0.01$ in each individual test). In the *Performance* columns, a minus (–) indicates that the correlation is among the bottom 25% of the column, a plus (+) indicates that it is among the top 25%, and an average sign (Ø) indicates that it falls in between. The asterisks in the second column from the left highlight the candidate biomarkers that were selected for each experimental method.

Statistical selection of the most accurate candidate biomarkers

We pursued two complementary strategies to assess how well each of the 23 candidate biomarkers predicts the DNA methylation state of a sample (unmethylated or methylated, designated by zero or one, respectively), which is known from clustering of the bisulfite sequencing data (Figure 42). First, we simulated the measurements of each of the candidate biomarkers in silico, based on the DNA methylation profiles available from clonal bisulfite sequencing, and we calculated the correlation between the simulated values and the (binary) sample methylation class (Table 13, left columns). Second, for a representative subset of 14 tumor samples (sample numbers 14 to 25), we re-analyzed all selected CpGs with the respective experimental methods, and we calculated correlations between the experimentally derived scores and the sample methylation class (Table 13, right columns). To maintain a fair comparison no biomarker optimization was performed at this stage and the scores of biomarkers that include several CpG positions were calculated as unweighted averages of the DNA methylation values of each position.

The correlation was measured both by the (linear) Pearson correlation coefficient and the (rank-based) Spearman correlation coefficient. For most biomarker candidates in both the in silico and experimental analysis, we observed high correlation between the biomarker score and the sample methylation class derived from the bisulfite sequencing data. Based on these performance evaluations (Table 13) as well as on the number of CpG positions per biomarker (as discussed below, larger numbers confer greater robustness against unknown SNPs), we selected one high-scoring biomarker candidate for each of the methods, namely CO7 for COBRA, SI01 for SIRPH and Py15 for pyrosequencing.

Biomarker optimization and performance evaluation

In the final step, for each of the three selected biomarkers (CO7, SI01 and Py15) we constructed optimized logistic regression models, validated them by cross-validation and derived safe decision boundaries by re-applying the classification formulae to 14 tumor samples with full experimental data and known sample methylation class.

First, to test whether logistic regression can accurately predict the sample methylation classes, we trained logistic regression models for each of the biomarkers and validated them by leave-one-out cross-validation against the sample methylation class values (during this analysis, Tumor 16 was excluded because it constituted an outlier for all experimental methods used; it is however included in the final validation diagrams described below). For both the CO7 and the Py15 candidate biomarker, logistic regression led to correct classification of all 13 tumor samples (100% test set accuracy as determined by leave-one-out cross-validation). For the SI01 candidate biomarker, 12 out of 13 tumor samples were classified correctly (92% test set accuracy). This result is consistent with our observation that all biomarker candidates perform well, but indicates that the SIRPH assay is less predictive than the other two assays.

Second, we trained logistic regression models on all 13 validation cases (again excluding Tumor 16), giving rise to the following three formulae predicting whether or not a tumor sample should be considered as MGMT-promoter hypermethylated and therefore likely to be sensitive to alkylating chemotherapy:

- COBRA: $Score_{CO7} = -339.385 \cdot CpG_{1/2} + 196.192 \cdot CpG_{8/9} + 137.296 \cdot CpG_{5} - 21.803$
- SIRPH: $Score_{SI01} = 306.601 \cdot CpG_{13} - 36.792$;
- Pyrosequencing: $Score_{Py15} = 21.330 \cdot CpG_{9} + 24.806 \cdot CpG_{10} + 18.637 \cdot CpG_{11} + 21.503 \cdot CpG_{12} - 20.197$

In these formulae, the CpG_x variables refer to the measured DNA methylation score at each position, overall positive scores predict the presence of significant promoter methylation and overall negative scores predict absence of promoter methylation. For COBRA, we unexpectedly observed a highly negative coefficient for CpG position 1/2. Closer inspection showed that this CpG position was unmethylated in almost all tumors, including those that exhibit significant levels of DNA methylation elsewhere in the MGMT promoter. Hence it does not provide a reliable indicator of MGMT promoter methylation and may be susceptible to errors caused by random noise. We therefore recalculated the COBRA classification formula without CpG position 1/2 and obtained the following formula for the CO7_revised biomarker: $Score_{CO7\_revised} = 75.574 \cdot CpG_{8/9} + 63.821 \cdot CpG_5 - 21.071$. We recommend using this formula in practical applications as it is likely to be more robust toward DNA methylation variation.

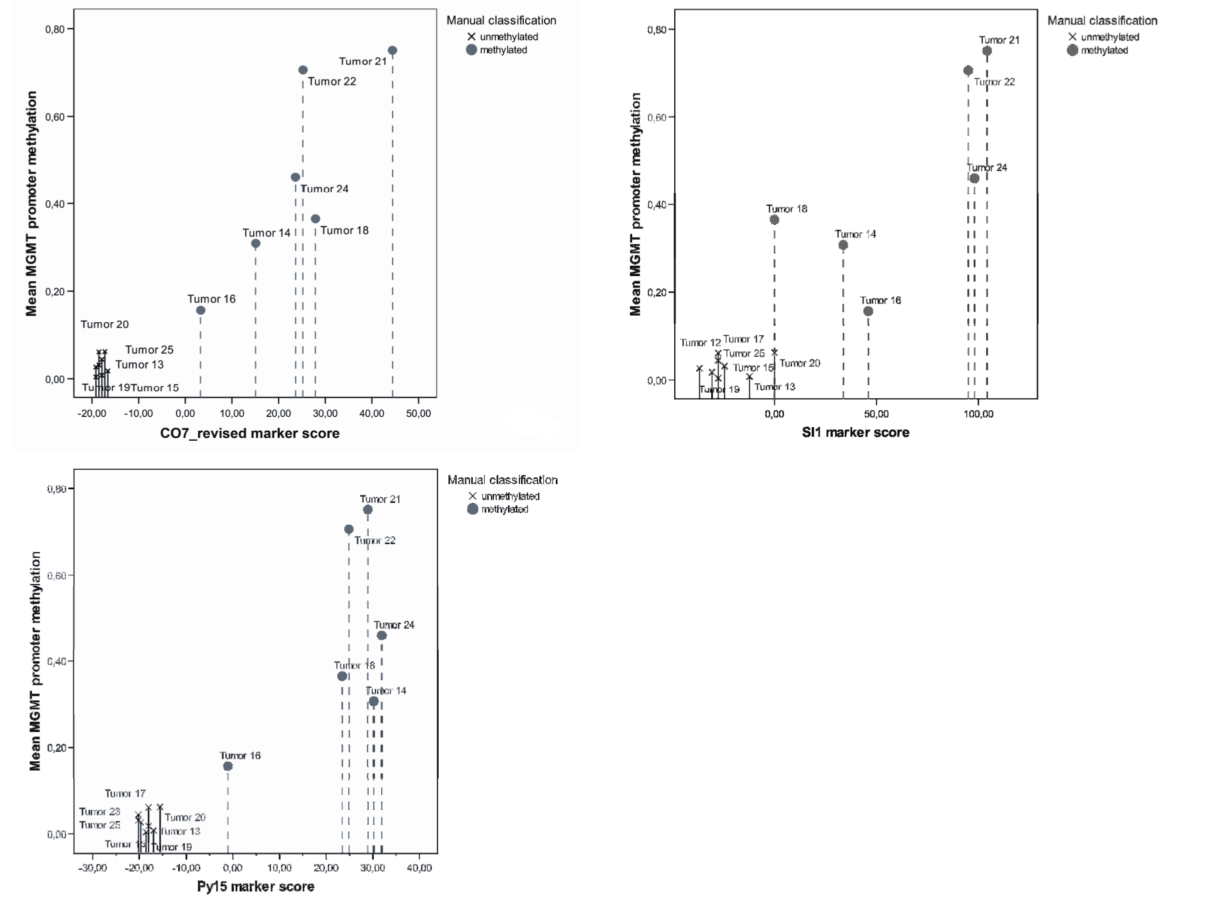


Figure 45. Performance of optimized MGMT biomarkers for COBRA, SIRPH and pyrosequencing

This figure displays a performance evaluation of the optimized COBRA (CO7_revised, top left), SIRPH (SI1, top right) and pyrosequencing (Py15, bottom left) biomarkers predicting the average level of MGMT promoter methylation in seven largely unmethylated tumor samples (13, 15, 17, 19, 20, 23 and 25), five highly methylated tumor samples (14, 18, 21, 22, 24) and one borderline case (16).

While it is straightforward to apply these formulae to new tumor samples (experimentally determine the values for the CpG_x variables, plug these values into the formula, calculate the overall score and compare this value with a threshold), the choice of appropriate thresholds distinguishing unmethylated cases from methylated and borderline cases requires some consideration. In the absence of a large patient cohort with known clinical history, we decided to re-apply the classification formulae to the full validation dataset, now including the borderline case 16, and to perform threshold selection by visual inspection. While this strategy is certainly not optimal, the selection of highly conservative thresholds and the fact that methylated and unmethylated tumor samples are separated by a large margin support the validity of our approach. For CO7_revised (Figure 45, top left), we observed that scores below -10 were highly indicative of overall absence of MGMT promoter methylation, while scores above 10 were consistently associated with the presence of promoter methylation. Tumor 16 fell between these two thresholds, indicating that no clear conclusion is possible in the region between -10 and 10. For SI01 (Figure 45, top right), tumors with scores below -10 should be classified as unmethylated and tumors with scores above 80 can safely be regarded as methylated. However, SI01 gives rise to a large interval for which no clear conclusion is possible, due to high score variance within the two clusters. For Py15 (Figure 45, bottom left), this uncertainty interval is substantially smaller, due to lower score variance within each of the two clusters. Scores below -10 provide strong evidence of an unmethylated MGMT promoter, while scores above 10 indicate substantial MGMT promoter methylation. The diagrams in Figure 45 visually confirm our previous observation that the pyrosequencing biomarker (Py15) is superior to both CO7/CO7_revised and SI01, and that CO7/CO7_revised is superior to SI01.

Since the Py15 biomarker incorporates four different CpG positions, it is relatively tolerant toward biological or experimental noise. Simulation showed that changing the measurement of any single CpG to either zero or to one (which could happen due rare C-T-SNPs at the analyzed CpG dinucleotide or due to incomplete bisulfite conversion) can only convert a clearly methylated or a clearly unmethylated sample into a borderline case (or vice versa), but cannot – for none of the samples we analyzed – convert a clearly methylated case into a clearly unmethylated case (or vice versa). Therefore, we conclude that the Py15 marker and, to a lesser extent, the CO7_revised marker provide the necessary statistical robustness to cope with borderline cases.

D-3.4 Discussion

The goal of this study was to develop a robust and cost-efficient assay for measuring hypermethylation of the promoter region of the MGMT gene. While the relevance of MGMT promoter hypermethylation as a predictor of chemotherapy resistance in glioblastomas has been established previously (Hegi et al. 2004; Hegi et al. 2005), the applicability of the existing, MSP-based, assay is limited by technical constraints and lack of robustness. Therefore, we developed and applied a bioinformatic workflow that helped us optimize alternative assays, which make use of highly robust experimental protocols such as bisulfite pyrosequencing and COBRA.

Based on statistical validations (cross-validation and simulated introduction of single-nucleotide polymorphisms at measured CpGs), we conclude that the described pyrosequencing assay is suitable for application in clinical settings and enables accurate and robust identification of MGMT promoter hypermethylation, thus guiding personalized treatment of glioblastomas. The biomarker's robustness could also be confirmed on FFPE specimens (data not shown, cf. Mikeska et al. 2007), which makes it possible to cost-efficiently investigate MGMT promoter hypermethylation in archival tissues.

While the focus of this study was the MGMT gene, the bioinformatic workflow described herein can be applied more generally to optimize epigenetic biomarkers for use in clinical settings, which is a recurring problem in translational research. Following up on the results described in this chapter, we currently develop software tools that facilitate the optimization of biomarker candidates for clinical use and their validation in large patient cohorts (see section E-2.2 below for further discussion).

Part E. Conclusion and Outlook

E-1 Outline

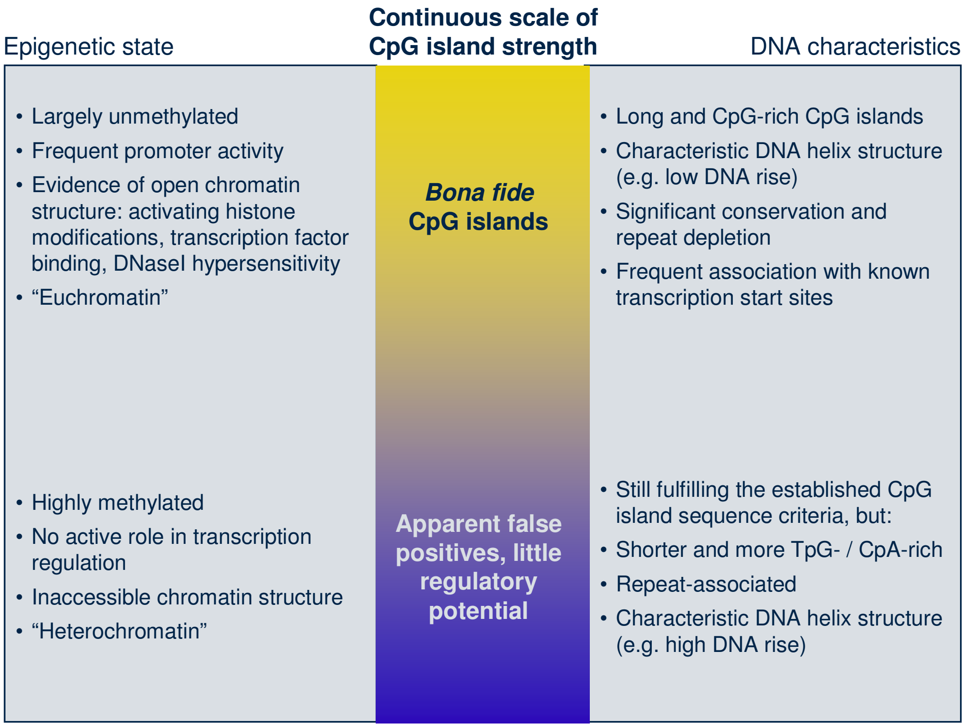
The purpose of the following sections is to highlight how the results of the preceding chapters fit together to support a model of the genome and epigenome as two interdependent and tightly correlated carriers of biological information. Furthermore, we sketch how the bioinformatic methods and tools developed in this thesis assemble into an integrated pipeline for discovery, optimization and validation of DNA methylation biomarkers. We hope that this pipeline will facilitate the discovery of biomarker candidates for use in cancer diagnosis and therapy optimization, and contribute to an increased speed of translation from biomarker candidates into clinically validated molecular diagnostic assays. The thesis concludes with an outline of three follow-up projects in which we currently exploit the practical utility of several results outlined in previous chapters, followed by a brief sketch of emerging trends in the field of computational epigenetics.

E-2 Conclusion

E-2.1 Genome and epigenome: complex dependencies

Throughout this thesis, a recurring theme was the globally high degree of correlation that we observed between the human genome and epigenome. Using a spectrum of quantitative methods including parametric and non-parametric statistical testing, linear regression models, support vector machines and other machine learning algorithms, we found that specific aspects of the genomic DNA sequence (such as DNA sequence patterns, structural motifs and the distribution of repetitive DNA) correlate strongly with a broad range of epigenetic modifications. Thus, epigenetic information could be predicted with significant accuracy from the genomic DNA sequence. Highest prediction accuracies were obtained for DNA methylation (chapters B-2, C-3 and C-4), but we also found that activating histone modifications (such as H3K4 methylation and H3 acetylation) and several hallmarks of active transcription initiation sites (such as DNase I hypersensitivity, SP1 binding, polymerase II pre-initiation complex binding and CAGE tag density) are significantly correlated with specific characteristics of DNA sequence and structure (chapter B-4).

While this globally high degree of correlation between genome and epigenome was observed both inside and outside CpG islands (chapters B-2, B-4, C-3 and C-4), we will focus our further discussion on the former because of the paramount role of CpG islands for gene regulation (Antequera 2003; Bajic et al. 2006). Figure 46 illustrates schematically how CpG islands differ in terms of their epigenetic states, and how these differences are mirrored by differences in their DNA characteristics: CpG islands that are frequently unmethylated, exhibit promoter activity, and/or foster open chromatin structure also exhibit exceptional DNA characteristics, including high levels of CpG enrichment, high conservation, significant repeat depletion, and a specific predicted helix structure. On the other hand, methylated and transcriptionally inactive regions (that still fulfill the traditional CpG island criteria) exhibit converse DNA characteristics. Importantly, these differences are gradual and quantitative in nature, i.e. many CpG islands fall between the two extremes in terms of both their genomic and epigenomic characteristics. We did not find evidence of a combinatorial “DNA sequence code”, in which a specific combination of DNA sequence motifs or transcription factor binding sites would be required and sufficient for a particular epigenetic state. Rather, it seems that the presence of multiple CpG-rich sequence patterns, specific structural properties of the DNA and the degree of repeat depletion cumulatively contribute to an open and transcriptionally accessible epigenetic state.



Epigenetic state

Continuous scale of CpG island strength

DNA characteristics

- Largely unmethylated
- Frequent promoter activity
- Evidence of open chromatin structure: activating histone modifications, transcription factor binding, DNaseI hypersensitivity
- “Euchromatin”

Bona fide CpG islands

- Long and CpG-rich CpG islands
- Characteristic DNA helix structure (e.g. low DNA rise)
- Significant conservation and repeat depletion
- Frequent association with known transcription start sites

- Highly methylated
- No active role in transcription regulation
- Inaccessible chromatin structure
- “Heterochromatin”

Apparent false positives, little regulatory potential

- Still fulfilling the established CpG island sequence criteria, but:
- Shorter and more TpG- / CpA-rich
- Repeat-associated
- Characteristic DNA helix structure (e.g. high DNA rise)

Figure 46. Correlation between characteristics of the genomic DNA sequence and the epigenetic and functional states of CpG islands

This figure illustrates the link between genome (left) and epigenome (right) at CpG islands, which was discovered by statistical analysis and epigenome prediction on a broad range of epigenetic modifications (chapters B-2, B-4, C-3 and C-4). CpG islands in the human genome can apparently be ordered on a scale of increasingly open and transcriptionally competent chromatin structure (left) and simultaneously on a scale of characteristic DNA attributes (right), with high correlation between both scales.

Collectively, these results indicate that it is overly reductionistic to view the epigenome as an additional layer of regulatory function unconstrained by the underlying genomic DNA sequence. Rather, our results suggest a model in which the genome encodes a preprogrammed epigenetic state for each CpG island, which will be realized in any particular cell unless specific mechanisms of epigenetic regulation (such as X-chromosome inactivation, imprinting or cancer-specific hypermethylation) overrule this default state. We thus propose that each CpG island in the human genome can be assigned a propensity toward either an open and transcriptionally accessible or a condensed and silenced chromatin structure that is encoded in its DNA, and we believe that this propensity is the biological correlate that we capture with our epigenome prediction scores.

This model is consistent with the observation that tissue-specific regulation of epigenetic modifications such as DNA methylation and activating histone modifications is less widespread than one might have anticipated (ENCODE Project Consortium 2007; Song et al. 2005), and it explains why we observed higher prediction accuracies for DNA methylation, which is highly stable between tissues (Eckhardt et al. 2006), than for more volatile histone modifications (Kouzarides 2007; Trojer and Reinberg 2006). Furthermore, our model seems biologically plausible given that the genome is likely to act as a blueprint constituting the well-defined “ground state” of the epigenome that is established in two waves of epigenetic reprogramming in the germline and in the early embryo (Reik 2007).

E-2.2 A bioinformatic pipeline for cancer biomarker discovery, optimization and validation

DNA methylation biomarkers hold great promise for improving cancer therapy. For many cancers aberrant methylation is detectable already in early-stage and pre-malignant tumors (Esteller 2008; Feinberg et al. 2006; Laird 2003), when surgical treatment can be highly effective. Furthermore, specific DNA methylation patterns often correlate with clinical parameters such as cancer stage, survival time and chemotherapy resistance, which gives rise to exciting opportunities for accurate prognosis and informed treatment decisions, thus enabling a more personalized cancer therapy.

However, in spite of a number of recent successes (Hegi et al. 2005; Lofton-Day et al. 2007; Shames et al. 2006; Weisenberger et al. 2006), so far DNA methylation biomarkers have failed to fulfill their promise of significantly improving routine cancer therapy. This has a number of reasons, some of which are common to all cancer biomarkers (Ludwig and Weinstein 2005; Pepe et al. 2001): (i) high cost of validating biomarkers in a large number of patients, (ii) reproducibility problems across different patient cohorts, (iii) lack of specificity and cost issues when using biomarkers for population screening, (iv) reluctance to incorporate molecular biomarkers into time-tested staging systems, and (v) economic disincentives such as reduced scope for blockbuster drugs when cancer staging becomes more accurate. Additional obstacles apply specifically to DNA methylation biomarkers: (i) the experimental methods most commonly used for DNA methylation analysis in research settings are not applicable in clinical settings, due to high cost or lack of robustness (Mikeska et al. 2007), and (ii) sufficiently accurate DNA methylation biomarkers often require a biostatistical model that integrates several measurements (Laird 2003), often necessitating the use of complex mathematical calculations during routine diagnosis.

By combining the methods developed in this thesis, we believe that it may be possible to improve on these issues and to facilitate the development of clinically useful DNA methylation biomarkers. To that end, we propose a bioinformatics-driven pipeline for cancer biomarker discovery, optimization and validation (Figure 47). Its phase 1 (top row in Figure 47) describes the process leading from samples of cases and controls (typically tumor tissue vs. healthy tissue) to a set of candidate biomarkers, i.e. regions that are differentially methylated between the cases and the controls. Phase 2 (bottom row in Figure 47) describes the translation of a candidate biomarker into a validated molecular diagnostic assay that is readily usable in clinical settings. The phases of our pipeline are consistent with those proposed by Pepe et al. (Pepe et al. 2001), with phase 1 mapping to their “Preclinical exploratory phase 1” and phase 2 to their “Clinical assay and validation phase 2”. Their phases 3 to 5 correspond to multiple iterations of step 6 in our pipeline.

During phase 1 (top row in Figure 47), genome-scale methods for DNA methylation mapping are applied to screen the genome for differential DNA methylation in a relatively small number of cases and controls. In order to increase cost efficiency, it is often desirable to restrict the experimental analysis to promising candidate regions of cancer-specific hypermethylation (step 1), e.g. by using custom CpG island microarrays for MeDIP analysis. Based on our work on DNA methylation prediction and improved CpG island annotation (chapters B-2, B-4 and B-5 of this thesis), it is possible to bioinformatically identify and exclude consistently methylated CpG islands, which carry little potential for cancer-specific hypermethylation and need not be assayed. Next, DNA methylation mapping is performed experimentally for the selected genomic regions (step 2), using genome-scale methods such as MeDIP (Weber et al. 2005) or high-throughput bisulfite sequencing (Eckhardt et al. 2006; Meissner et al. 2005), and the datasets are pre-processed with well-established tools (see chapters C-3 and D-2 for exploratory examples). Based on the resulting DNA methylation maps, genomic regions are identified that exhibit differential DNA methylation patterns among cases and controls, and these candidate biomarkers are then prioritized according to their predicted potential for discrimination between cases and controls (step 3). Importantly, this prediction step should not rely on DNA methylation levels alone, but also take genome annotation data into account. For example, genomic regions with low levels of inter-individual variation (cf. chapter C-4 of this thesis) may be more robust than highly variable regions when tested in different patient cohorts, and genomic regions predicted as unmethylated in blood (cf. chapter B-2 of this thesis) may be less sensitive to tumor sample contamination by non-tumor cells. EpiGRAPH (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis) is a versatile tool to support the prioritization of candidate biomarkers.

During phase 2 (bottom row in Figure 47), selected candidate biomarkers are optimized for robust and cost-efficient experimental analysis, and their predictiveness is validated in large patient cohorts. The key step of this phase is to move from costly genome-scale methods with low sample throughput to highly targeted methods that can cost-efficiently assess DNA methylation at specific regions in a large number of samples. As an alternative to ad hoc solutions – such as selecting CpGs from the vicinity of transcription start sites – we have prototyped a more systematic approach (Mikeska et al. 2007, cf. chapter D-3 of this thesis): First, a high-resolution profile of DNA methylation at the region of interest is obtained, using clonal bisulfite sequencing of a representative subset of cases and controls (step 4). Next, robust and cost-efficient DNA methylation assays are designed for protocols such as COBRA, MSP or bisulfite pyrosequencing, and these assays are optimized such that they measure DNA methylation at CpG positions that are highly informative for the overall state of DNA methylation (step 5). We provide software toolkits to support these two steps. BiQ Analyzer (Bock et al. 2005, cf. chapter C-2 of this thesis) facilitates the analysis of bisulfite sequencing results, and MethMarker (cf. section E-3.1 and Figure 48) implements expert rules for design, selection and optimization of the most appropriate assays for use in clinical settings. Finally, the optimized DNA methylation assays are validated in a large patient cohort (step 6). Based on the validation results, the biomarker's prediction parameters are trained and high-confidence decision thresholds are estimated. We currently develop the BiomarkerSpace web service to support validation as well as secure use of validated clinical biomarkers over an internet portal application (cf. section E-3.1).

In summary, the pipeline outlined in Figure 47 is an attempt to formulate and systematize the key steps required for discovery, optimization and validation of novel DNA methylation biomarkers, and to highlight good-practice methods for each step. Importantly, the pipeline is backed by user-friendly bioinformatic tools (developed partially within this PhD project), which automate repetitive tasks and support key decisions with rule-based advice and predictive statistics. For two software tools – EpiGRAPH and BiQ Analyzer – stable production versions have been released. The MethMarker software is currently in beta testing stage and available on request, while BiomarkerSpace is still under development. All software tools are and will be available free of charge to academic users. We believe that this pipeline – if widely adopted – can significantly reduce the development time of novel biomarkers, thereby helping to fulfill the promise of DNA methylation biomarkers for improving cancer therapy.

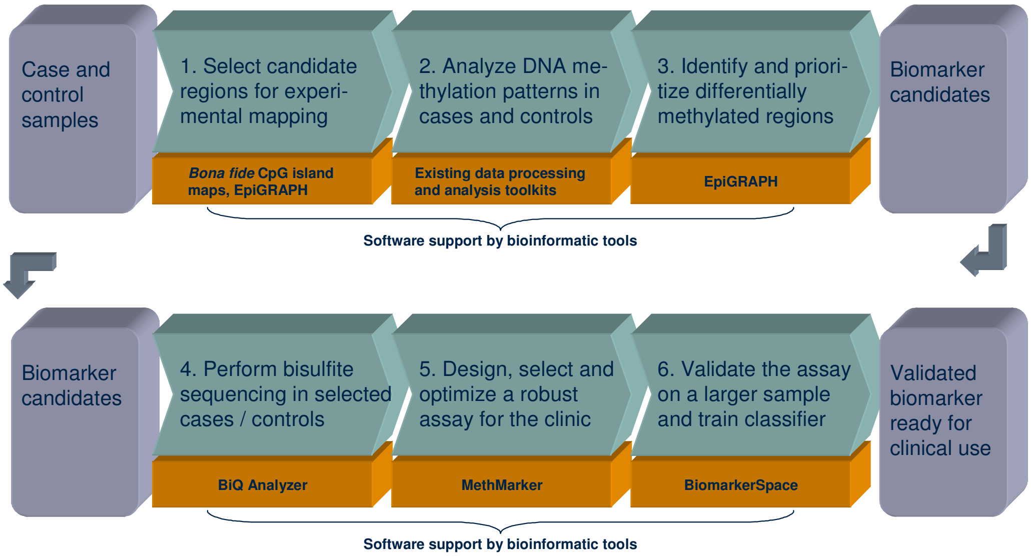

Case and control samples
1. Select candidate regions for experimental mapping
2. Analyze DNA methylation patterns in cases and controls
3. Identify and prioritize differentially methylated regions
Biomarker candidates
Bona fide CpG island maps, EpiGRAPH
Existing data processing and analysis toolkits
EpiGRAPH
Software support by bioinformatic tools
Biomarker candidates
4. Perform bisulfite sequencing in selected cases / controls
5. Design, select and optimize a robust assay for the clinic
6. Validate the assay on a larger sample and train classifier
Validated biomarker ready for clinical use
BiQ Analyzer
MethMarker
BiomarkerSpace
Software support by bioinformatic tools

Figure 47. Bioinformatic pipeline for cancer biomarker discovery, optimization and validation

This diagram outlines the key steps of a systematic pipeline designed to facilitate the discovery of DNA methylation biomarkers and their translation into clinically validated molecular diagnostic assays. Bioinformatic tools are highlighted that support – and partially automate – the key tasks of each phase. The bona fide CpG island maps (step 1) are described in chapter B-4, EpiGRAPH (step 3) in chapter B-3, BiQ Analyzer (step 4) in chapter C-2 and MethMarker (step 5) as well as BiomarkerSpace (step 6, still in development) are briefly outlined in section E-3.1.

E-3 Outlook

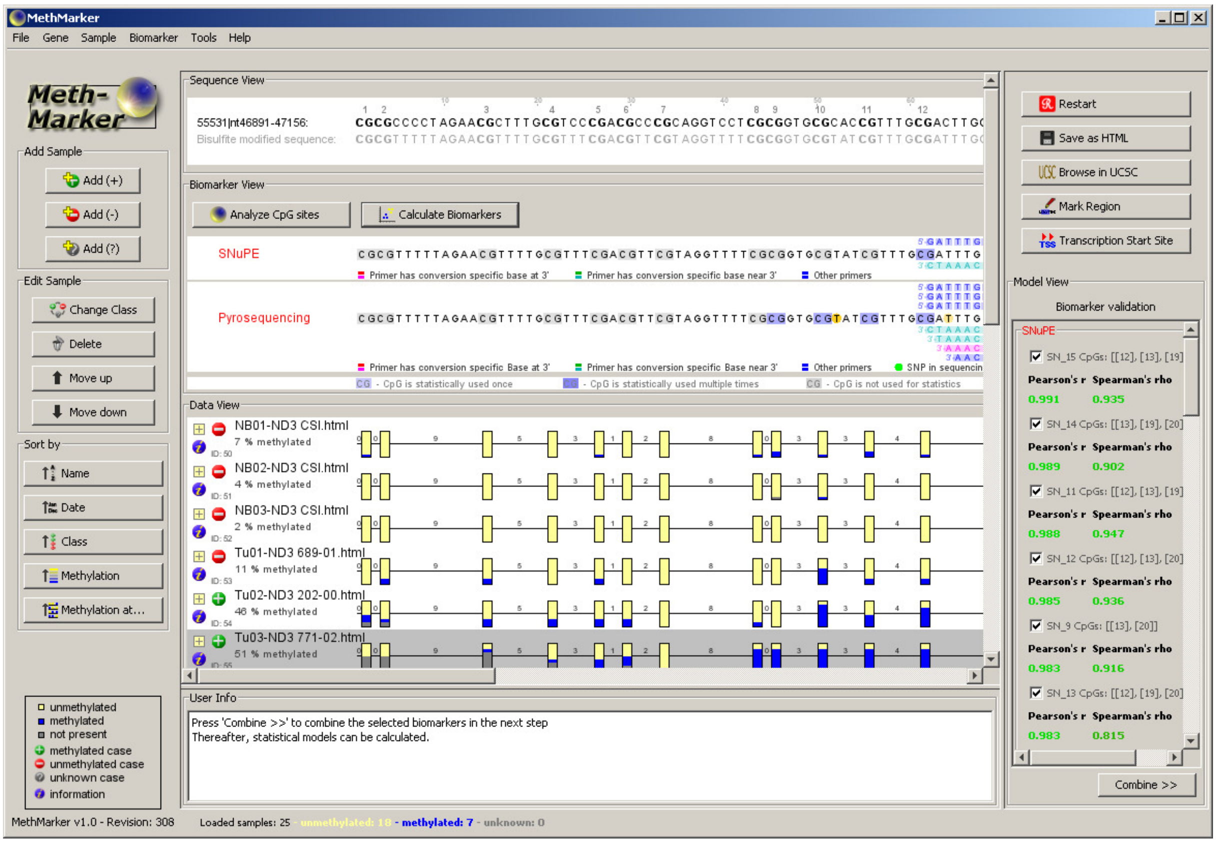
E-3.1 Ongoing projects following up results of this thesis

In several ongoing research projects we aim to exploit the practical utility of results described in this thesis, in the context of both cancer epigenetics and genome annotation.

- Two key software components of the DNA methylation biomarker pipeline outlined in Figure 47 – MethMarker and BiomarkerSpace – are currently being designed and implemented (in collaboration with Peter Schüffler and Thomas Mikeska). Figure 48 shows a screenshot of the current MethMarker beta version (programmed by Peter Schüffler), which provides bioinformatic support for the key steps that were performed manually in the MGMT biomarker optimization project (Mikeska et al. 2007, cf. chapter D-3 of this thesis).
- The EU-funded CANCERDIP project (in cooperation with five biological and clinical partners across Europe, started on January 1st, 2008) aims to discover novel DNA methylation biomarkers for colon cancer and leukemia by means of genome-wide DNA methylation mapping. Our role is to coordinate bioinformatic data analysis throughout the project, to make predictions about promising target genes and to model epigenetic regulatory mechanisms related to cancer development. These tasks rely heavily on EpiGRAPH (cf. chapter B-3 of this thesis) and on the bioinformatic pipeline for cancer biomarker discovery outlined in Figure 47.

- The CpG island annotation algorithm described in chapter B-5 is currently being developed into a CpG island annotation toolkit that is sufficiently fast, self-contained and configurable to be integrated into genome annotation pipelines (in collaboration with Lars Feuerbach). The goal of this toolkit is to provide a more accurate alternative to the CpG island finders currently in use by genome browsers.

Two themes are common to these ongoing research projects. On the one hand, we are joining forces with cancer researcher in order to exploit and extend the relevance of our research in the context of cancer epigenetics. On the other hand, we work toward making additional methods developed within this PhD project available as software packages and/or web servers, believing that user-friendly software will be important for progress in all areas of epigenetic research.



Press 'Combine >>' to combine the selected biomarkers in the next step
Thereafter, statistical models can be calculated.

☑ SN_15 CpGs: [[12], [13], [19]		
	Pearson's r	Spearman's rho
	0.991	0.935
☑ SN_14 CpGs: [[13], [19], [20]		
	Pearson's r	Spearman's rho
	0.989	0.902
☑ SN_11 CpGs: [[12], [13], [19]		
	Pearson's r	Spearman's rho
	0.988	0.947
☑ SN_12 CpGs: [[12], [13], [20]		
	Pearson's r	Spearman's rho
	0.985	0.936
☑ SN_9 CpGs: [[13], [20]]		
	Pearson's r	Spearman's rho
	0.983	0.916
☑ SN_13 CpGs: [[12], [19], [20]		
	Pearson's r	Spearman's rho
	0.983	0.815

MethMarker v1.0 - Revision: 308 Loaded samples: 25 - unmethylated: 18 - methylated: 7 - unknown: 0

Figure 48. The MethMarker software facilitates optimization of candidate DNA methylation biomarkers for validation and routine use in clinical settings

This screenshot shows a typical use case in which MethMarker is being applied to selecting a cheap and robust DNA methylation assay for a differentially methylated region that has been linked to cancer in previous work, but which needs to be validated in clinical settings. As a next step following assay design with MethMarker, it will be possible export and upload a digital description of the optimized biomarker onto the BiomarkerSpace web server, which will provide a central gateway for biomarker validation, routine use and performance monitoring (MethMarker and BiomarkerSpace are collaboration projects with Peter Schüffler and Thomas Mikeska).

E-3.2 A wider perspective on computational epigenetics

Research in computational epigenetics has progressed substantially during the last four years, and the current speed of primary data generation suggests that no calmer waters are in sight. In this last section, we briefly consider trends and developments that may influence computational epigenetics in the coming years.

(1) The mere quantity of epigenetic data arising from epigenome projects will pose a paramount bioinformatic challenge along all segments of the scientific value chain, from storage and management of raw data over data analysis and biological discovery toward the construction and integration of quantitative models.

(2) Epigenome data analysis will increasingly take the proteins into account that read and write epigenetic information, as well as their interaction partners and regulatory networks. Such reverse engineering of epigenetic regulation could lead to a quantitative model and, ultimately, to rational manipulation of the core circuitry that controls cell fate and pluripotency (Boyer et al. 2005).

(3) The decreasing cost of epigenome mapping will enable quantitative analysis of epigenetic variation in human populations. Recent twin studies suggest that both environmental influences (Fraga et al. 2005) and genetic variation (Heijmans et al. 2007) contribute to epigenetic variation. It will be a daunting bioinformatic task to distill putative functional connections from the integration of epigenome data with gene expression profiles and haplotype maps for a large sample from a heterogeneous population.

(4) Epigenome mapping in multiple species will add an evolutionary perspective to computational epigenetics. Initial results suggest that orthologous regions in different mammals carry similar epigenetic information (Bernstein et al. 2005; Enard et al. 2004), which is expected since the DNA encodes parts of its epigenetic state (Bock et al. 2007; Segal et al. 2006). It will be interesting to see whether comparative epigenomics can significantly improve our ability to identify functionally important sites in the human genome, as is the case for comparative genomics.

(5) Theoretical modeling will provide a way to fathom our mechanistic and quantitative understanding of epigenetic mechanisms. For example, two recent studies could show that cooperativity among the proteins that write epigenetic information is required for stably maintaining the state of an epigenetic switch in the presence of highly dynamic fluctuations at the molecular level (Dodd et al. 2007; Sontag et al. 2006). Modeling studies can thus help explain how the high-level phenomena that we observe for epigenetic regulation emerge from the dynamic interplay of various epigenetic mechanisms.

(6) The development of powerful and easy-to-use “statistical genome browsers” will enable biologists to perform complex epigenome data analysis online without requiring strong statistical or programming skills. Tools like Galaxy (Blankenberg et al. 2007; Giardine et al. 2005) and EpiGRAPH (http://epigraph.mpi-inf.mpg.de/, cf. chapter B-3 of this thesis), which let their users design and execute genome analyses through an intuitive web front-end, are first steps in this direction, and further tools are likely to follow.

(7) Epigenetic mechanisms could turn out to play a role in diseases other than cancer, as there is strong circumstantial evidence for epigenetic regulation being involved in mental disorders, autoimmune diseases and other complex diseases (Bjornsson et al. 2004; Feinberg 2007). Bioinformatic methods such as text mining and exploratory data mining may play a role in identifying and prioritizing concrete hypotheses for experimental validation.

In conclusion, exciting times are ahead for research in computational epigenetics!

Part F. List of Publications

Papers in peer-reviewed journals (first author)

*Bock, C., Halachev, K., Büch, J. and Lengauer, T. (2008) EpiGRAPH: A user-friendly software for advanced (epi-) genome analysis and prediction, *in revision*.

*Bock, C., Walter, J., Paulsen, M. and Lengauer, T. (2008) Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research* 36: e55.

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

*Bock, C., Walter, J., Paulsen, M. and Lengauer, T. (2007) CpG Island Mapping by Epigenome Prediction, *PLoS Computational Biology*, **3**, e110. [2]

*Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T. and Walter, J. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure, *PLoS Genetics*, **2**, e26. [3]

*Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J. and Lengauer, T. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing, *Bioinformatics*, **21**, 4067-4068. [4]

Papers in peer-reviewed journals (contributing author)

Viré, E., Bock, C., Denis, H., Brenner, C., Dedeurwaerder, S., Ballestar, E., Jacinto, F.V., Alaminos, M., Driessen, M., Bollen, M., Di Croce, L., de Launoit, Y., Lengauer, T., Shiekhattar, R., Esteller, M. and *Fuks, F. (2008) MeCP2 associates with the Polycomb Group protein EZH2 to repress transcription, *in revision*.

Kircher, M., Bock, C. and *Paulsen, M. (2008) Structural conservation versus functional divergence of maternally expressed microRNAs in the *Dlk1*/*Gtl2* imprinting region, BMC Genomics, 9, 346.

*Moser, D., Ekawardhani, S., Kumsta, R., Palmason, H., Bock, C., Athanassiadou, Z., Lesch, K.-P., and Meyer, J. (2008) Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site, *Neuropsychopharmacology*, http://dx.doi.org/10.1038/npp.2008.77.

Mikeska, T., Bock, C., El-Maarri, O., Hübner, A., Ehrentraut, D., Schramm, J., Felsberg, J., Kahl, P., Büttner, R., Pietsch, T. and *Waha, A. (2007) Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis, *Journal of Molecular Diagnostics*, **9**, 368-381. [5]

[1] Most frequently accessed *Bioinformatics* paper in winter 2007/2008 (>5,000 downloads between November and March).

[2] The Max Planck Society issued a press release (http://tinyurl.com/2br4hc).

[3] On the cover of the March 2006 issue of PLoS Genetics and among the top ten of most frequently accessed papers for several months. The Max Planck Society and Saarland University issued press releases (http://tinyurl.com/2xz3qo).

[4] BiQ Analyzer has been the first non-commercial software to become part of the Applied Biosystems Software Community Program. Dozens of analyses are performed every week by researchers from all over the world.

[5] The optimized biomarker is in use at the German Brain Tumor Center, University of Bonn, Germany.

* Corresponding author

Liu, F., Tostesen, E., Sundet, J.K., Jenssen, T.K., Bock, C., Jerstad, G.I., Thilly, W.G. and *Hovig, E. (2007) The human genomic melting map, *PLoS Computational Biology*, **3**, e93.

Papers in conference proceedings, books, etc.

Bock, C. and Lengauer, T. (2007) Computational Epigenetics: Bioinformatik für neue Wege in der Krebsforschung, *Jahrbuch 2007*. Max-Planck-Gesellschaft.

Bock, C. (2006) Bioinformatik: Neue Strategien gegen Krebs, *Deutsches Ärzteblatt*, **103**, Supplement: PRAXiS, 10-11.

Bock, C., Halachev, K. and Lengauer, T. (2006) Bioinformatisches Data Mining erschließt neue Strategien gegen Krebs. In Haasis, K., Heinzl, A. and Klumpp, D. (eds), *Aktuelle Trends in der Softwareforschung*. dpunkt.verlag, Heidelberg.

Bock, C. and Lengauer, T. (2006) Computational epigenetics: bioinformatics prediction for new approaches to cancer treatment, *Spotlight 2006 (Featured Research Story of the Max Planck Institute for Informatics)*. Max Planck Institute for Informatics, Saarbrücken.

Part G. References

Adorjan, P., J. Distler, E. Lipscher, F. Model, J. Müller, C. Pelet et al. 2002. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res* **30:** e21.

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal, L.A. Pennacchio et al. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5:** e234.

Aissani, B. and G. Bernardi. 1991. CpG islands: features and distribution in the genomes of vertebrates. *Gene* **106:** 173-183.

Alliance for Human Epigenomics and Disease. 2007. Proposal for an International AHEAD Pilot Project. Available: http://www.aacr.org/Uploads/DocumentRepository/TaskForces/ahead_pilot_project_proposal_may2007.pdf. Accessed 31 March 2008.

Antequera, F. 2003. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* **60:** 1647-1658.

Antequera, F. and A. Bird. 1993. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90:** 11995-11999.

Arney, K.L. and A.G. Fisher. 2004. Epigenetic aspects of differentiation. *J Cell Sci* **117:** 4355-4363.

Bajic, V.B., S.L. Tan, A. Christoffels, C. Schonbach, L. Lipovich, L. Yang et al. 2006. Mice and men: their promoter properties. *PLoS Genet* **2:** e54.

Bajic, V.B., S.L. Tan, Y. Suzuki, and S. Sugano. 2004. Promoter prediction analysis on the whole human genome. *Nat Biotechnol* **22:** 1467-1473.

Baldi, P., S. Brunak, Y. Chauvin, C.A. Andersen, and H. Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16:** 412-424.

Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823-837.

Baylin, S.B. and J.E. Ohm. 2006. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer* **6:** 107-116.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick et al. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321-1325.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57:** 289-300.

Bernstein, B.E., M. Kamal, K. Lindblad-Toh, S. Bekiranov, D.K. Bailey, D.J. Huebert et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169-181.

Bernstein, B.E., A. Meissner, and E.S. Lander. 2007. The mammalian epigenome. *Cell* **128:** 669-681.

Bernstein, B.E., T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315-326.

Berry, C., S. Hannenhalli, J. Leipzig, and F.D. Bushman. 2006. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol* **2:** e157.

Bhasin, M., H. Zhang, E.L. Reinherz, and P.A. Reche. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett* **579:** 4302-4308.

Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16:** 6-21.

Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209-213.

Bjornsson, H.T., M.D. Fallin, and A.P. Feinberg. 2004. An integrated epigenetic and genetic approach to common human disease. *Trends Genet* **20:** 350-358.

Blankenberg, D., J. Taylor, I. Schenck, J. He, Y. Zhang, M. Ghent et al. 2007. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res* **17:** 960-964.

Bock, C. and T. Lengauer. 2008. Computational epigenetics. *Bioinformatics* **24:** 1-10.

Bock, C., M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* **2:** e26.

Bock, C., S. Reither, T. Mikeska, M. Paulsen, J. Walter, and T. Lengauer. 2005. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* **21:** 4067-4068.

Bock, C., J. Walter, M. Paulsen, and T. Lengauer. 2007. CpG island mapping by epigenome prediction. *PLoS Comput Biol* **3:** e110.

Bock, C., J. Walter, M. Paulsen, and T. Lengauer. 2008. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping [http://dx.doi.org/10.1093/nar/gkn122]. *Nucleic Acids Res*.

Boyer, L.A., T.I. Lee, M.F. Cole, S.E. Johnstone, S.S. Levine, J.P. Zucker et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947-956.

Boyle, A.P., S. Davis, H.P. Shulha, P. Meltzer, E.H. Margulies, Z. Weng et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311-322.

Bracken, A.P., N. Dietrich, D. Pasini, K.H. Hansen, and K. Helin. 2006. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev* **20:** 1123-1136.

Brinkman, A.B., S.W. Pennings, G.G. Braliou, L.E. Rietveld, and H.G. Stunnenberg. 2007. DNA methylation immediately adjacent to active histone marking does not silence transcription. *Nucleic Acids Res* **35:** 801-811.

Buck, M.J. and J.D. Lieb. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83:** 349-360.

Buck, M.J., A.B. Nobel, and J.D. Lieb. 2005. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* **6:** R97.

Bulcke, D., T. Van, Lemmens, Karen, V. de Peer, Yves et al. 2006. Inferring transcriptional networks by mining 'omics' data. *Current Bioinformatics* **1:** 313.

Buszczak, M. and A.C. Spradling. 2006. Searching chromatin for stem cell identity. *Cell* **125:** 233-236.

Caiafa, P. and M. Zampieri. 2005. DNA methylation and chromatin structure: the puzzling CpG islands. *J Cell Biochem* **94:** 257-265.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626-635.

Cawley, S., S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499-509.

Chang, C.-C. and C.-J. Lin. 2005. LIBSVM: a library for Support Vector Machines. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed 31 March 2008.

Chen, K. and N. Rajewsky. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8:** 93-103.

Chen, X., S.V. Mariappan, R.K. Moyzis, E.M. Bradbury, and G. Gupta. 1998. Hairpin induced slippage and hyper-methylation of the fragile X DNA triplets. *J Biomol Struct Dyn* **15:** 745-756.

Cohen, S.M., T.S. Furey, N.A. Doggett, and D.G. Kaufman. 2006. Genome-wide sequence and functional analysis of early replicating DNA in normal human fibroblasts. *BMC Genomics* **7:** 301.

Colella, S., L. Shen, K.A. Baggerly, J.P. Issa, and R. Krahe. 2003. Sensitive and quantitative universal Pyrosequencing methylation analysis of CpG sites. *Biotechniques* **35:** 146-150.

Cooper, D.N., M.H. Taggart, and A.P. Bird. 1983. Unmethylated domains in vertebrate DNA. *Nucleic Acids Res* **11:** 647-658.

Costantini, M., O. Clay, F. Auletta, and G. Bernardi. 2006. An isochore map of human chromosomes. *Genome Res* **16:** 536-541.

Crawford, G.E., I.E. Holt, J. Whittle, B.D. Webb, D. Tai, S. Davis et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16:** 123-131.

Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Das, R., N. Dimitrova, Z. Xuan, R.A. Rollins, F. Haghighi, J.R. Edwards et al. 2006. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A* **103:** 10713-10716.

Derti, A., F.P. Roth, G.M. Church, and C.T. Wu. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38:** 1216-1220.

Dillon, N. 2006. Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* **14:** 117-126.

Dion, M.F., S.J. Altschuler, L.F. Wu, and O.J. Rando. 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A* **102:** 5501-5506.

Dodd, I.B., M.A. Micheelsen, K. Sneppen, and G. Thon. 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* **129:** 813-822.

Drake, J.W., B. Charlesworth, D. Charlesworth, and J.F. Crow. 1998. Rates of spontaneous mutation. *Genetics* **148:** 1667-1686.

Du, J., J.S. Rozowsky, J.O. Korbel, Z.D. Zhang, T.E. Royce, M.H. Schultz et al. 2006. A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics* **22:** 3016-3024.

Eckhardt, F., J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38:** 1378-1385.

Eden, E., D. Lipson, S. Yogev, and Z. Yakhini. 2007. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **3:** e39.

El-Maarri, O., U. Herbiniaux, J. Walter, and J. Oldenburg. 2002. A rapid, quantitative, non-radioactive bisulfite-SNuPE- IP RP HPLC assay for methylation analysis at specific CpG sites. *Nucleic Acids Res* **30:** e25.

Enard, W., A. Fassbender, F. Model, P. Adorjan, S. Paabo, and A. Olek. 2004. Differences in DNA methylation patterns between humans and chimpanzees. *Curr Biol* **14:** R148-149.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799-816.

Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8:** 286-298.

Esteller, M. 2008. Epigenetics in cancer. *N Engl J Med* **358:** 1148-1159.

Esteller, M., J. Garcia-Foncillas, E. Andion, S.N. Goodman, O.F. Hidalgo, V. Vanaclocha et al. 2000. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med* **343:** 1350-1354.

Evans, G.A. 2000. Designer science and the "omic" revolution. *Nat Biotechnol* **18:** 127.

Fang, F., S. Fan, X. Zhang, and M.Q. Zhang. 2006. Predicting methylation status of CpG islands in the human brain. *Bioinformatics* **22:** 2204-2209.

Fawcett, T. 2004. ROC graphs: notes and practical considerations for researchers. Technical Report HPL-2003-4. HP Labs, Palo Alto.

Feinberg, A.P. 2007. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447:** 433-440.

Feinberg, A.P., R. Ohlsson, and S. Henikoff. 2006. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* **7:** 21-33.

Feinberg, A.P. and B. Tycko. 2004. The history of cancer epigenetics. *Nat Rev Cancer* **4:** 143-153.

Feltus, F.A., E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. 2003. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci U S A* **100:** 12253-12258.

Flicek, P., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen et al. 2008. Ensembl 2008. *Nucleic Acids Res* **36:** D707-714.

Fraga, M.F., E. Ballestar, M.F. Paz, S. Ropero, F. Setien, M.L. Ballestar et al. 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* **102:** 10604-10609.

Fraga, M.F. and M. Esteller. 2007. Epigenetics and aging: the targets and the marks. *Trends Genet* **23:** 413-418.

Frank, E., M. Hall, L. Trigg, G. Holmes, and I.H. Witten. 2004. Data mining in bioinformatics using Weka. *Bioinformatics* **20:** 2479-2481.

Frommer, M., L.E. McDonald, D.S. Millar, C.M. Collis, F. Watt, G.W. Grigg et al. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89:** 1827-1831.

Gangaraju, V.K. and B. Bartholomew. 2007. Mechanisms of ATP dependent chromatin remodeling. *Mutat Res* **618:** 3-17.

Gardiner-Garden, M. and M. Frommer. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196:** 261-282.

Gardiner, E.J., C.A. Hunter, M.J. Packer, D.S. Palmer, and P. Willett. 2003. Sequence-dependent DNA structure: a database of octamer structural parameters. *J Mol Biol* **332:** 1025-1035.

Gentleman, R. 2005. Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology* **4**.

Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80.

Gerson, S.L. 2004. MGMT: its role in cancer aetiology and cancer therapeutics. *Nat Rev Cancer* **4:** 296-307.

Giardine, B., C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15:** 1451-1455.

Gimelbrant, A., J.N. Hutchinson, B.R. Thompson, and A. Chess. 2007. Widespread monoallelic expression on human autosomes. *Science* **318:** 1136-1140.

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

Glass, J.L., R.F. Thompson, B. Khulan, M.E. Figueroa, E.N. Olivier, E.J. Oakley et al. 2007. CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res* **35:** 6798-6807.

Goh, L., S.K. Murphy, S. Muhkerjee, and T.S. Furey. 2007. Genomic sweeping for hypermethylated genes. *Bioinformatics* **23:** 281-288.

Goldberg, A.D., C.D. Allis, and E. Bernstein. 2007. Epigenetics: a landscape takes shape. *Cell* **128:** 635-638.

Gomez-Skarmeta, J.L., B. Lenhard, and T.S. Becker. 2006. New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences. *Dev Dyn* **235:** 870-885.

Grant-Downton, R.T. and H.G. Dickinson. 2006. Epigenetics and its implications for plant biology 2. The 'epigenetic epiphany': epigenetics, evolution and beyond. *Ann Bot (Lond)* **97:** 11-27.

Greenbaum, J.A., B. Pang, and T.D. Tullius. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res* **17:** 947-953.

Hackenberg, M., C. Previti, P.L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza, and J.L. Oliver. 2006. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7:** 446.

Hajkova, P., O. el-Maarri, S. Engemann, J. Oswald, A. Olek, and J. Walter. 2002. DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol Biol* **200:** 143-154.

Handa, V. and A. Jeltsch. 2005. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol* **348:** 1103-1112.

Hannenhalli, S. and S. Levy. 2001. Promoter prediction in the human genome. *Bioinformatics* **17 Suppl 1:** S90-96.

Hark, A.T., C.J. Schoenherr, D.J. Katz, R.S. Ingram, J.M. Levorse, and S.M. Tilghman. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405:** 486-489.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2001. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York.

Heard, E. 2004. Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* **16:** 247-255.

Hegi, M.E., A.C. Diserens, S. Godard, P.Y. Dietrich, L. Regli, S. Ostermann et al. 2004. Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clin Cancer Res* **10:** 1871-1874.

Hegi, M.E., A.C. Diserens, T. Gorlia, M.F. Hamou, N. de Tribolet, M. Weller et al. 2005. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* **352:** 997-1003.

Heijmans, B.T., D. Kremer, E.W. Tobi, D.I. Boomsma, and P.E. Slagboom. 2007. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum Mol Genet* **16:** 547-554.

Heintzman, N.D., R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311-318.

HEROIC Project Consortium. 2005. High-throughput Epigenetic Regulatory Organisation In Chromatin - Project Fact Sheet. Available: http://cordis.europa.eu/fetch?CALLER=FP6_PROJ&ACTION=D&DOC=1&CAT=PROJ&QUERY=1183993108794&RCN=78439. Accessed 31 March 2008.

Holliday, R. 2006. Epigenetics: a historical overview. *Epigenetics* **1:** 76-80.

Hsu, F., W.J. Kent, H. Clawson, R.M. Kuhn, M. Diekhans, and D. Haussler. 2006. The UCSC Known Genes. *Bioinformatics* **22:** 1036-1046.

Huang, D.W., B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* **35:** W169-175.

Huang, T.H., M.R. Perry, and D.E. Laux. 1999. Methylation profiling of CpG islands in human breast cancer cells. *Hum Mol Genet* **8:** 459-470.

Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen et al. 2007. Ensembl 2007. *Nucleic Acids Res* **35:** D610-617.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1:** S96-104.

Hull, D., K. Wolstencroft, R. Stevens, C. Goble, M.R. Pocock, P. Li et al. 2006. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* **34:** W729-732.

Ioshikhes, I.P., I. Albert, S.J. Zanton, and B.F. Pugh. 2006. Nucleosome positions predicted through comparative genomics. *Nat Genet* **38:** 1210-1215.

Ioshikhes, I.P. and M.Q. Zhang. 2000. Large-scale human promoter mapping using CpG islands. *Nat Genet* **26:** 61-63.

Jeltsch, A., J. Walter, R. Reinhardt, and M. Platzer. 2006. German human methylome project started. *Cancer Res* **66:** 7378.

Ji, H. and W.H. Wong. 2005. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21:** 3629-3636.

Johnson, D.S., A. Mortazavi, R.M. Myers, and B. Wold. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497-1502.

Johnson, W.E., W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown et al. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A* **103:** 12457-12462.

Jones, P.A. and S.B. Baylin. 2007. The epigenomics of cancer. *Cell* **128:** 683-692.

Jones, P.A. and R. Martienssen. 2005. A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res* **65:** 11241-11246.

Kapranov, P., J. Drenkow, J. Cheng, J. Long, G. Helt, S. Dike et al. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* **15:** 987-997.

Kapranov, P., A.T. Willingham, and T.R. Gingeras. 2007. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8:** 413-423.

Karolchik, D., R.M. Kuhn, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773-779.

Katzman, S., A.D. Kern, G. Bejerano, G. Fewell, L. Fulton, R.K. Wilson et al. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317:** 915.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Keshet, I., Y. Schlesinger, S. Farkash, E. Rand, M. Hecht, E. Segal et al. 2006. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* **38:** 149-153.

Khulan, B., R.F. Thompson, K. Ye, M.J. Fazzari, M. Suzuki, E. Stasiek et al. 2006. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res* **16:** 1046-1055.

Kim, T.H., Z.K. Abdullaev, A.D. Smith, K.A. Ching, D.I. Loukinov, R.D. Green et al. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231-1245.

Kim, T.H., L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond et al. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876-880.

Kimura, N., T. Nagasaka, J. Murakami, H. Sasamoto, M. Murakami, N. Tanaka et al. 2005. Methylation profiles of genes utilizing newly developed CpG island methylation microarray on colorectal cancer patients. *Nucleic Acids Res* **33:** e46.

Klose, R.J., S.A. Sarraf, L. Schmiedeberg, S.M. McDermott, I. Stancheva, and A.P. Bird. 2005. DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell* **19:** 667-678.

Kouzarides, T. 2007. Chromatin modifications and their function. *Cell* **128:** 693-705.

Laird, P.W. 2003. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **3:** 253-266.

Laird, P.W. 2005. Cancer epigenetics. *Hum Mol Genet* **14 Spec No 1:** R65-76.

Larsen, F., G. Gundersen, R. Lopez, and H. Prydz. 1992. CpG islands as gene markers in the human genome. *Genomics* **13:** 1095-1107.

Lee, T.I., R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125:** 301-313.

Lewin, J., A.O. Schmitt, P. Adorjan, T. Hildmann, and C. Piepenbrock. 2004. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics* **20:** 3005-3012.

Li, W., P. Bernaola-Galvan, F. Haghighi, and I. Grosse. 2002. Applications of recursive segmentation to the analysis of DNA sequences. *Computers & Chemistry* **26:** 491-510.

Li, W., C.A. Meyer, and X.S. Liu. 2005. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21 Suppl 1:** i274-282.

Liu, F., E. Tostesen, J.K. Sundet, T.K. Jenssen, C. Bock, G.I. Jerstad et al. 2007. The human genomic melting map. *PLoS Comput Biol* **3:** e93.

Lofton-Day, C., F. Model, T. Devos, R. Tetzner, J. Distler, M. Schuster et al. 2007. DNA-Methylation Biomarkers for Blood-Based Colorectal Cancer Screening. *Clin Chem.*

Ludwig, J.A. and J.N. Weinstein. 2005. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* **5:** 845-856.

Luedi, P.P., F.S. Dietrich, J.R. Weidman, J.M. Bosko, R.L. Jirtle, and A.J. Hartemink. 2007. Computational and experimental identification of novel human imprinted genes. *Genome Res* **17:** 1723-1730.

Luedi, P.P., A.J. Hartemink, and R.L. Jirtle. 2005. Genome-wide prediction of imprinted murine genes. *Genome Res* **15:** 875-884.

Luque-Escamilla, P.L., J. Martinez-Aroza, J.L. Oliver, J.F. Gomez-Lopera, and R. Roman-Roldan. 2005. Compositional searching of CpG islands in the human genome. *Physical Review E. Statistical, Nonlinear, & Soft Matter Physics* **71:** 061925.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24:** 133-141.

Marjoram, P., J. Chang, P.W. Laird, and K.D. Siegmund. 2006. Cluster analysis for DNA methylation profiles having a detection threshold. *BMC Bioinformatics* **7:** 361.

Matsuo, K., O. Clay, T. Takahashi, J. Silke, and W. Schaffner. 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat Cell Mol Genet* **19:** 543-555.

Meissner, A., A. Gnirke, G.W. Bell, B. Ramsahoye, E.S. Lander, and R. Jaenisch. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33:** 5868-5877.

Microarray and Gene Expression Data Society. 2005. The MIAME Checklist – update January 2005. Available: http://www.mged.org/Workgroups/MIAME/MIAMEchecklist_chipchip.pdf. Accessed 31 March 2008.

Mikeska, T., C. Bock, O. El-Maarri, A. Hübner, D. Ehrentraut, J. Schramm et al. 2007. Optimization of Quantitative MGMT Promoter Methylation Analysis Using Pyrosequencing and Combined Bisulfite Restriction Analysis. *J Mol Diagn* **9:** 368-381.

Mikkelsen, T.S., M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553-560.

Model, F., P. Adorjan, A. Olek, and C. Piepenbrock. 2001. Feature selection for DNA methylation based cancer classification. *Bioinformatics* **17 Suppl 1:** S157-164.

Montgomery, S.B., O.L. Griffith, J.M. Schuetz, A. Brooks-Wilson, and S.J. Jones. 2007. A Survey of Genomic Properties for the Detection of Regulatory Polymorphisms. *PLoS Comput Biol* **3:** e106.

Murrell, A., V.K. Rakyan, and S. Beck. 2005. From genome to epigenome. *Hum Mol Genet* **14 Spec No 1:** R3-R10.

Narlikar, L., R. Gordân, and A. Hartemink. 2007. Nucleosome occupancy information improves de novo motif discovery. In *Research in Computational Molecular Biology, 11th Annual International Conference, RECOMB 2007, Oakland, CA, USA, April 21-25, 2007, Proceedings* (eds. T.P. Speed and H. Huang). Springer-Verlag, New York.

Noble, W.S., S. Kuehn, R. Thurman, M. Yu, and J. Stamatoyannopoulos. 2005. Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics* **21 Suppl 1:** i338-i343.

Ohm, J.E. and S.B. Baylin. 2007. Stem cell chromatin patterns: an instructive mechanism for DNA hypermethylation? *Cell Cycle* **6:** 1040-1043.

Ohm, J.E., K.M. McGarvey, X. Yu, L. Cheng, K.E. Schuebel, L. Cope et al. 2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* **39:** 237-242.

Olson, W.K., M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey et al. 2001. A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol* **313:** 229-237.

Ongenaert, M., L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge. 2007. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*

Ooi, L. and I.C. Wood. 2007. Chromatin crosstalk in development and disease: lessons from REST. *Nat Rev Genet* **8:** 544-554.

Parisi, F., P. Wirapati, and F. Naef. 2007. Identifying synergistic regulation involving c-Myc and sp1 in human tissues. *Nucleic Acids Res* **35:** 1098-1107.

Pavesi, G., P. Mereghetti, G. Mauri, and G. Pesole. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32:** W199-203.

Peaston, A.E. and E. Whitelaw. 2006. Epigenetics and phenotypic variation in mammals. *Mamm Genome* **17:** 365-374.

Peckham, H.E., R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl et al. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res* **17:** 1170-1177.

Pepe, M.S., R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist et al. 2001. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* **93:** 1054-1061.

Pfister, S., C. Schlaeger, F. Mendrzyk, A. Wittmann, A. Benner, A. Kulozik et al. 2007. Array-based profiling of reference-independent methylation status (aPRIMES) identifies frequent promoter methylation and consecutive downregulation of ZIC2 in pediatric medulloblastoma. *Nucleic Acids Res* **35:** e51.

Platt, J. 1999. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning* (eds. B. Schölkopf C.J.C. Burges, and A.J. Smola), pp. 185-208. MIT Press, Cambridge, MA.

Ponger, L., L. Duret, and D. Mouchiroud. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* **11:** 1854-1860.

Ponger, L. and D. Mouchiroud. 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18:** 631-633.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61-65.

Qi, Y., A. Rolfe, K.D. MacIsaac, G.K. Gerber, D. Pokholok, J. Zeitlinger et al. 2006. High-resolution computational models of genome binding events. *Nat Biotechnol* **24:** 963-970.

Rakyan, V.K., T. Hildmann, K.L. Novik, J. Lewin, J. Tost, A.V. Cox et al. 2004. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol* **2:** e405.

Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447:** 425-432.

Reik, W., F. Santos, and W. Dean. 2003. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology* **59:** 21-32.

Ringrose, L., M. Rehmsmeier, J.M. Dura, and R. Paro. 2003. Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Dev Cell* **5:** 759-771.

Roh, T.Y., S. Cuddapah, and K. Zhao. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19:** 542-552.

Roh, T.Y., G. Wei, C.M. Farrell, and K. Zhao. 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res* **17:** 74-81.

Rollins, R.A., F. Haghighi, J.R. Edwards, R. Das, M.Q. Zhang, J. Ju et al. 2006. Large-scale structure of genomic methylation patterns. *Genome Res* **16:** 157-163.

Royce, T.E., J.S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman et al. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* **21:** 466-475.

Russo, V.E.A., R.A. Martienssen, and A.D. Riggs. 1996. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Satchwell, S.C., H.R. Drew, and A.A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191:** 659-675.

Saxonov, S., P. Berg, and D.L. Brutlag. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103:** 1412-1417.

Schlesinger, Y., R. Straussman, I. Keshet, S. Farkash, M. Hecht, J. Zimmerman et al. 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39:** 232-236.

Schones, D.E. and K. Zhao. 2008. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9:** 179-191.

Schuettengruber, B., D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli. 2007. Genome regulation by polycomb and trithorax proteins. *Cell* **128:** 735-745.

Schwartz, Y.B. and V. Pirrotta. 2007. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet* **8:** 9-22.

Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore et al. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772-778.

Shahbazian, M.D. and H.Y. Zoghbi. 2002. Rett syndrome and MeCP2: linking epigenetics and neuronal function. *Am J Hum Genet* **71:** 1259-1272.

Shames, D.S., L. Girard, B. Gao, M. Sato, C.M. Lewis, N. Shivapurkar et al. 2006. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med* **3:** e486.

Shen, L., Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed et al. 2007. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet* **3:** 2023-2036.

Siegmund, K.D., P.W. Laird, and I.A. Laird-Offringa. 2004. A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* **20:** 1896-1904.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* **21:** 3940-3941.

Smith, A.D., P. Sumazin, and M.Q. Zhang. 2007. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol* **3:** 73.

Solter, D. 2006. Imprinting today: end of the beginning or beginning of the end? *Cytogenet Genome Res* **113:** 12-16.

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

Song, J.S., W.E. Johnson, X. Zhu, X. Zhang, W. Li, A.K. Manrai et al. 2007. Model-based analysis of two-color arrays (MA2C). *Genome Biol* **8:** R178.

Sontag, L.B., M.C. Lorincz, and E.G. Luebeck. 2006. Dynamics, stability and inheritance of somatic DNA methylation imprints. *J Theor Biol* **242:** 890-899.

Sparmann, A. and M. van Lohuizen. 2006. Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer* **6:** 846-856.

Squazzo, S.L., H. O'Geen, V.M. Komashko, S.R. Krig, V.X. Jin, S.W. Jang et al. 2006. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16:** 890-900.

Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Subramanian, A., H. Kuehn, J. Gould, P. Tamayo, and J.P. Mesirov. 2007. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23:** 3251-3253.

Synamatix Sdn. Bhd. 2007. SXOligoSearch Supporting Document. Available: http://synasite.mgrc.com.my:8080/sxog/files/SXOligoSearch_benchmark.pdf. Accessed 31 March 2008.

Tabachnick, B.G. and L.S. Fidell. 2007. *Using multivariate statistics*. Pearson/Allyn & Bacon, Boston.

Takai, D. and P.A. Jones. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99:** 3740-3745.

Takai, D. and P.A. Jones. 2003. The CpG island searcher: a new WWW resource. *In Silico Biol* **3:** 235-240.

Thomas, D.J., K.R. Rosenbloom, H. Clawson, A.S. Hinrichs, H. Trumbower, B.J. Raney et al. 2007. The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res* **35:** D663-667.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Thurman, R.E., N. Day, W.S. Noble, and J.A. Stamatoyannopoulos. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* **17:** 917-927.

Ting, A.H., K.M. McGarvey, and S.B. Baylin. 2006. The cancer epigenome--components and functional correlates. *Genes Dev* **20:** 3215-3231.

Toedling, J., O. Sklyar, and W. Huber. 2007. Ringo - an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8:** 221.

Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137-144.

Trinklein, N.D., U. Karaoz, J. Wu, A. Halees, S. Force Aldred, P.J. Collins et al. 2007. Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res* **17:** 720-731.

Trojer, P. and D. Reinberg. 2006. Histone lysine demethylases and their impact on epigenetics. *Cell* **125:** 213-217.

Turker, M.S. 2002. Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene* **21:** 5388-5393.

Turner, B.M. 2007. Defining an epigenetic code. *Nat Cell Biol* **9:** 2-6.

Ushijima, T. 2005. Detection and interpretation of altered methylation patterns in cancer cells. *Nat Rev Cancer* **5:** 223-231.

Ushijima, T., N. Watanabe, E. Okochi, A. Kaneda, T. Sugimura, and K. Miyamoto. 2003. Fidelity of the methylation pattern and its variation in the genome. *Genome Res* **13:** 868-874.

van Steensel, B. 2005. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat Genet* **37 Suppl:** S18-24.

Villa, R., D. Pasini, A. Gutierrez, L. Morey, M. Occhionorelli, E. Viré et al. 2007. Role of the polycomb repressive complex 2 in acute promyelocytic leukemia. *Cancer Cell* **11:** 513-525.

Viré, E., C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot et al. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439:** 871-874.

Waddington, C.H. 1942. The epigenotype. *Endeavour* **1:** 18-20.

Wang, Y. and F.C. Leung. 2004. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics* **20:** 1170-1177.

Wang, Z., H.F. Willard, S. Mukherjee, and T.S. Furey. 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput Biol* **2:** e113.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Weber, M., J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam et al. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37:** 853-862.

Weber, M., I. Hellmann, M.B. Stadler, L. Ramos, S. Pääbo, M. Rebhan et al. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39:** 457-466.

Weber, M. and D. Schübeler. 2007. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol* **19:** 273-280.

Weisenberger, D.J., K.D. Siegmund, M. Campan, J. Young, T.I. Long, M.A. Faasse et al. 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* **38:** 787-793.

Widschwendter, M., H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth et al. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* **39:** 157-158.

Witten, I.H. and E. Frank. 2000. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, Calif.

Wong, A.H., Gottesman, II, and A. Petronis. 2005. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum Mol Genet* **14 Spec No 1:** R11-18.

Wong, N.C., L.H. Wong, J.M. Quach, P. Canham, J.M. Craig, J.Z. Song et al. 2006. Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet* **2:** e17.

Woodcock, C.L. 2006. Chromatin architecture. *Curr Opin Struct Biol* **16:** 213-220.

Xiong, Z. and P.W. Laird. 1997. COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res* **25:** 2532-2534.

Xu, J., S.D. Pope, A.R. Jazirehi, J.L. Attema, P. Papathanasiou, J.A. Watts et al. 2007. Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proc Natl Acad Sci U S A* **104:** 12377-12382.

Yahya-Graison, E.A., J. Aubert, L. Dauphinot, I. Rivals, M. Prieur, G. Golfier et al. 2007. Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. *Am J Hum Genet* **81:** 475-491.

Yamada, Y., H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka et al. 2004. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res* **14:** 247-266.

Yan, P.S., S.H. Wei, and T.H. Huang. 2004. Methylation-specific oligonucleotide microarray. *Methods Mol Biol* **287:** 251-260.

Yoo, C.B. and P.A. Jones. 2006. Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov* **5:** 37-50.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S.W. Chan, H. Chen et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126:** 1189-1201.

Zhang, Z.D., A. Paccanaro, Y. Fu, S. Weissman, Z. Weng, J. Chang et al. 2007a. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res* **17:** 787-797.

Zhang, Z.D., J. Rozowsky, H.Y. Lam, J. Du, M. Snyder, and M. Gerstein. 2007b. Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol* **8:** R81.

Zheng, L.Q. and P.A. Larson. 1996. Speeding up external mergesort. *IEEE Transactions On Knowledge And Data Engineering* **8:** 322-332.

Zhou, G.L., D.P. Liu, and C.C. Liang. 2005. Memory mechanisms of active transcription during cell division. *Bioessays* **27:** 1239-1245.

Zilberman, D., M. Gehring, R.K. Tran, T. Ballinger, and S. Henikoff. 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39:** 61-69.