Evolutionary Epigenomics – Identifying Functional Genome Elements by Epigenetic Footprints in the DNA

Dissertation

Zur Erlangung des Grades des
Doktors der Naturwissenschaften der
Naturwissenschaftlichen-Technischen Fakultäten der
Universität des Saarlandes

Eingereicht von
Lars Feuerbach

Saarbrücken, 2014

Tag des Kolloquiums:	16.1.2014
Dekan der Fakultät:	Prof. Dr. Mark Groves
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Gerhard Weikum
Gutachter:	Prof. Dr. Dr. Thomas Lengauer
	Prof. Dr. Jotun Hein
Beisitzer:	Dr. Glenn Lawyer

Abstract

Over the last decade, advances in genome sequencing have substantially increased the amount of genomic DNA sequences available. While these rich resources have improved our understanding of genome function, research of the epigenome as a transient but heritable memory system of the cell has only profited from this development indirectly. Although epigenetic information in the form of DNA methylation is not directly encoded in the genomic nucleotide sequence, it increases the mutation rate of cytosine-guanine dinucleotides by the *CpG decay* effect, and thus leaves epigenetic footprints in the DNA. This thesis proposes four approaches to facilitate this information for research. For largely uncharacterized genomes, *CgiHunter* presents an exhaustive algorithm for an unbiased DNA sequence-based annotation of CpG islands as regions that are protected from *CpG decay*. For species with well characterized point mutation frequencies, *EqiScore* identifies regions that evolve under distinct DNA methylation levels. Furthermore, the derived equilibrium distributions for methylated and unmethylated genome regions predict the evolutionary robustness of transcription factor binding site motifs against the *CpG decay* effect. The *AluJudge* annotation and underlying *L-score* provide a method to identify putative active copies of CpG-rich transposable elements within genomes. Additionally, epigenetic footprints in these sequences are applied to predict the germline epigenome of their loci. Moreover, *AluJudge* provides support for the targeted removal of epigenetically silenced repeat copies from CpG island annotations, which are subjected to a methylation-induced erosion process. Finally, the FFK approach enables the prediction of the germline methylome for homologous genome loci.

In a number of case studies on the human genome, I demonstrate how this *evolutionary epigenomics* toolkit can be applied to enhance the epigenomic characterization of the large quantity of currently sequenced vertebrate genomes. Furthermore, these studies show how to improve the identification of novel epigenetic functional genome regions in already well characterized species. Finally, the toolkit opens new avenues for computer-based research of the evolution of genome-wide DNA methylation.

Kurzfassung

In den letzten Jahrzehnten haben Fortschritte in der Genom-Sequenzierung zu einem substanziellen Zuwachs an verfügbaren DNS-Sequenzen geführt. Während diese Ressourcen zu einem verbesserten Verständnis der Funktionsweise von Genomen führten, konnte die Erforschung des Epigenoms als veränderlichem und doch vererbbarem zellulärem Informationsspeicher nur indirekt von dieser Entwicklung profitieren. Obwohl epigenetische Information nicht direkt in Form von genomischen Nukleotid-Sequenzen kodiert wird, sind beide Systeme derart miteinander verflochten, dass gemeinsame evolutionäre Abhängigkeiten einen epigenetischen Fußabdruck in der genomischen DNS erzeugen.
In dieser Arbeit werden vier Ansätze vorgestellt, um diese bisher weitgehend unerforschte Informationsquelle zu erschließen. Gleichsam einem Werkzeugkasten für Probleme der *Evolutionären Epigenomik*, bieten sie für eine Vielzahl verschiedener Szenarien eine Auswahl von einsetzbaren Methoden an.
Für weitgehend uncharakterisierte Genome ermöglicht *CgiHunter*, als kombinatorisch präziser Algorithmus, die auf der DNS-Sequenz basierende Identifikation von CpG Inseln, welche als Zentren von epigenetischer Regulation in Wirbeltier-Genomen bekannt sind.
Für Spezies in denen bereits Modelle der Punktmutationshäufigkeit existieren, können Dinukleotid-Gleichgewichtsverteilungen eingesetzt werden. Sie bieten über den *EqiScore*-Ansatz die Möglichkeit, Genomregionen zu identifizieren, die unter einem erhöhten DNS methylierungs Nieveau evolvieren. Des Weiteren ermöglichen sie eine Vorhersage der evolutionären Robustheit von Transkriptionsfaktor-Bindestellen gegenüber dieser epigenetischen Einflüsse.
Komplementär dazu bietet die *AluJudge* Annotation und der ihr zugrundeliegende *L-Score* für Genome mit CpG-reichen transponierenden Elementen einen Weg, unter ihnen potentiell aktive Kopien zu identifizieren. Darüber hinaus können diese Sequenzen als positions-spezifische Sonden des Keimbahn-Epigenoms eingesetzt werden. Auch unterstützt der *L-Score* die gezielte Entfernung von jenen mehrheitlich epigenetisch inaktiven Regionen aus CpG-Insel-Annotationen, welche einem methylierungs-induziertem Erosions-Prozess unterworfen sind. Zuletzt wird der FFK-Algorithmus, als ein phylogenetischer Ansatz beschrieben, der für nahe verwandte Spezies, wie jene des Primaten-Stammbaums, eine Vorhersage des Keimbahnmethyloms für beliebige Genomregionen ermöglicht.
In einer Reihe von Fallstudien an Hand des menschlichen Genoms, demonstriere ich im Anschluss, die Funktionalität dieser bioinformatischen Werkzeuge. Zum Einen ermöglichen sie die Identifizierung von neuen epigenetisch kontrollierten Regionen im menschlichen Genom. Zum Anderen dienen sie als Beispiel für die epigenomische *in-silico* Charakterisierung der Vielzahl von bald verfügbaren Vertebraten-Genomen. Zuletzt wird das Potential dieser neuen Ansätze für die computerbasierte Erforschung der evolutionären Entwicklung von genomeweiter DNS-Methylierung thematisiert.

Acknowledgments

First, I would like to thank my supervisor Thomas Lengauer for his advice and support during all stages of my PhD studies. I also want to thank Jotun Hein for accompanying me on one of the most important parts of this endeavor as well as for his readiness to act as a reviewer for this thesis. Furthermore, I would like to thank Alice McHardy for her comments on the manuscript.
A special word of thanks goes to my office mates Konstantin Halachev and Yassen Assenov for the countless discussions on epigenetics, bioinformatics and this thesis. Also, I want to thank Christoph Bock for our work together during the development of the *CgiHunter* algorithm. Furthermore, Rune Lyngsø and Glenn Lawyer shared their knowledge of phylogenies and statistics with me, which enabled the solution of some central questions regarding the evolution of the CpG dinucleotide. Moreover, I want to thank Sandra Koser for our joint work on the FFK approach and her assistance in visualizing the associated concepts.
Many thanks also to Barbara Hutter and Jasmina Bogojeska for their helpful comments on the manuscript of this thesis. Furthermore, I would like to thank the members of Jörn Walter's epigenetics lab and my colleagues from the MPI for the lively interaction between the *in vivo* and *in silico* scientists, which makes Saarbrücken a unique place for conducting epigenomic research. Finally, I want to thank my wife Elke and my family for their support.

Table of Contents

List of Figures

List of Tables

Introduction

The release of the human genome at the dawn of the millennium was coupled with the hope that it would elucidate the complex processes in living cells. The goal of the next phase of the Human Genome Project was described as “finding all the functional parts of the genome sequence and using this information to improve the health of individuals and society”(Collins, Green et al. 2003).

However, the genomes of mammalian cells are highly complex and much harder to understand than assumed from pilot studies on smaller systems. This complexity is partly explained by the fact that genomic sequences are interpreted differently based on the information that is not directly encoded in the DNA. Such information is referred to as epigenetic information and it explains, for instance, why cells from different human tissues (e.g., skin, muscles or liver) share identical genomes, but have very distinct phenotypes and functions.

Understanding these complex processes is especially important when they fail. For instance, the causes of complex diseases like cancer, mental disorders or immunodeficiency comprise disrupted DNA repair (Turnbull and Rahman 2008), unconstrained cell growth and proliferation (Jones and Thompson 2009), the failing failure of immune system pathways (Visser, Eichten et al. 2006) and a rearrangement of metabolism (Warburg 1956; Freitag 2006; Linehan, Srinivasan et al. 2010). A growing body of evidence indicates that epigenetic events play a major role in the induction of these diseases (Schanen 2006). This may imply that without an improved understanding of genome and epigenome function, the etiology of many diseases will remain elusive. In order to achieve systematic progress, identification of all the functional parts of the genome sequence involved in epigenetic signaling is critical. Based on such a catalogue, signals in functional genome regions can be discriminated from noise in non-functional genome regions and the relevance of an individual epigenetic signal for a disease can be evaluated.

This thesis contributes a number of algorithmic, statistical and molecular biological ideas that assist in the identification of those functional parts of the genome sequence that are most relevant for epigenetic signaling. All these approaches facilitate the special role of the dinucleotide CpG as an interface between the genome, the epigenome and the interdependencies in their joint evolution. First, this basic concept of the thesis will be explained in a nutshell. In order to take the interdisciplinary nature of this subject into consideration, this summary is intentionally written in a way that does not require advanced understanding of computer science, mathematics, molecular biology or biochemistry. This was done in order to make it understandable for high school students.

A metaphor for epigenome function and evolution

A key feature of cellular life is reproduction. In its essence, this is a process of copying information and applying the reality as data storage. To achieve this, every living cell processes information. This includes external information about its environment (e.g., to guide the uptake of nutrition) and internal information on how to react to external signals, how to regulate the metabolism, and how to calibrate and execute the replication process. Every non-random decision of a cell or complex organism to achieve reproduction is encoded by its biomolecules. They represent the information and cellular program that processes them, and are structured in a tightly interwoven multilayer information system that comprises layers such as the genome, epigenome, transcriptome, proteome, and metabolome. Understanding the nature of each layer of this information system is essential to understanding cell function and generating a hypothesis that can be tested by the scientific method.

For instance, a newspaper is an appealing analogy for the genome of unicellular organisms. This newspaper is full of articles that contain construction blueprints and each article has a headline that encodes to whom the article is of interest. Thus, multiple readers with different interests can browse the document simultaneously and find the required information efficiently. In the genome, this function is fulfilled by genes and their gene promoters. For single cell organisms, like the prokaryotic bacterium *Escherichia coli*, this metaphor is indeed very helpful. The lac-operon, for instance, is a complex of three genes, which encode the blueprints of proteins that are essential for digestion of the sugar lactose. Thinking in terms of its information content, the lac operon is like an article that describes how the cell can survive on a lactose diet. As long as no lactose is available, a repressor protein blocks the promoter of the operon, like someone who is intentionally hiding the headline of a newspaper article. Upon lactose binding, these repressors release the promoter and the information of the genes becomes accessible.

For multicellular organisms, the newspaper metaphor is an oversimplification. Here, the developmental stage of an organism, the type of tissue a cell belongs to, the part of the body in which it is located and a number of other conditions influence the decision of whether a gene should be accessible or not. The human body contains more than a trillion different cells that share the same genome, but which have very different functions and information demands. Therefore, our current understanding of the genome is more like that of a highly organized library made out of DNA. Its information content comprises construction blueprints, administrative regulations, stretches that stabilize the structure and a substantial amount of outdated or experimental material. This library is localized at the core of an industrial complex called the cell. Within the library, a myriad of autonomously acting agents are moving between the ‘bookshelves’, sometimes taking notes in the form of RNA transcripts. Only very few of these notes ever leave the library to go to the cell’s workshops where they act as blueprints for new biomolecules. Many other notes are for internal use only and describe what should or should not be transcribed and exported next. From time to time, incomplete or outdated notes are

produced, which are actively destroyed before they can become effective. For humans, all the information in this library is present in two slightly deviating editions of about 3 billion letters each. One of the editions was derived maternally and the other paternally. Considering that 500 letters constitute one page and 500 pages an entire volume, this makes 24,000 books worth of information. Unlike what we would expect from a human library, the genome is not organized into such compact books, but 46 large volumes called chromosomes. At the same time, no equivalent to a central index or catalogue has yet been discovered.

Instead, the information is organized similar to the newspaper example described above with a few very prominent headlines at the start of each article, *i.e.*, gene. The keywords in these headlines attract the readers (transcription factors). Furthermore, a single reader is not sufficient to initiate the transcription of information from the genome. Instead, a very heterogeneous team is required. Some team members are specialized in directing readers to headlines that they would otherwise not be interested in. Other members ensure that there is always enough writing material available or that the information is easily accessible for the actual scribe. The transcription procedure only starts when the complete teams are assembled. The importance of the different information varies according to the phases in the life of a cell. Thus, the number of experts that attract or repel others from certain headlines changes accordingly. Ultimately, this regulates the amount of each type of biomolecule that is produced at a certain phase.

Up to this point, our description of how cells and genomes interact in unicellular organisms is still in line with the newspaper metaphor. However, in multicultural organisms another layer of complexity is added because the same library is used in cells that have very different tasks. For instance, skin, liver and neuronal cells share the same genome, but differ dramatically in shape and function.
In order to prevent the wrong information from being transcribed, two additional strategies have evolved. First, whole compartments of the library containing temporarily irrelevant information are made less accessible. This way, fewer agents enter these areas and the likelihood of complete teams being assembled is reduced. Thus, the waste of working time and materials is prevented. Furthermore, the production of potentially harmful blueprints, which are only of use in other cell types, is limited. Second, particular parts of the genome are marked by methylation tags that basically state: “Do not start to transcribe here.” Only a small part of the genome remains free of these epigenetic tags, thus adding a further mechanism to the system for focusing on specific information. When two new cells are produced from one parent cell, the whole library is duplicated in a process called replication. The epigenetic marks are maintained during this process, thus conserving the cell type-specific information. Thus, genes which are never relevant for a liver cell only have to be deactivated once by DNA methylation during a process called differentiation and then stay repressed.

This library also contains a substantial amount of information that has no apparent function. Some of it is present in the form of blueprints that have been erroneously duplicated. Others were smuggled into the library from the outside, e.g., by viruses. If these are integrated into areas that are not transcribed, they are not very harmful to the

cell and thus faithfully copied during each library replication cycle. Sometimes errors occur during replication and storage, which results in a mutation of the information. Some of these mutations have systematic causes. For instance, the epigenetic methylation tag that prevents transcription frequently induces a particular type of mutation. Although repair mechanisms have been developed for many other sources of errors, this mutation is not repaired efficiently. Hence, it accumulates over time in areas that are intensively tagged by DNA methylation, thus providing information about the epigenetic history of this genome region.
This particular mutation effect is called *CpG decay* and will be evaluated in-depth in the thesis to gain a better understanding of the genome and epigenome as well as their evolutionary history.

Outline

In chapter 1, key concepts from genomics, epigenomics and evolution are introduced. Chapters 2 to 5 discuss four computational approaches that integrate these three domains. All four approaches are directly or indirectly associated with the *CpG decay* effect. In chapter 2, the enrichment of CpG dinucleotides in epigenetically active regions, so-called CpG islands, is analyzed empirically. To this end, a novel algorithm for the identification of CpG islands is developed and compared to existing approaches. In chapter 3, this analysis is extended to a quantitative level by explicitly modeling the evolutionary forces that cause this enrichment, *i.e.*, the *CpG decay* effect. In chapter 4, this analysis approach is adapted to address the special case of transposable elements, while chapter 5 extends the approach to integrate the information from several species into a joint analysis. All the chapters have the common theme of facilitating the *CpG decay* effect over an evolutionary timescale to gain insight into the genome and epigenome function. Chapter 6 closes the thesis with concluding remarks and the outlook for further development of the evolutionary field of epigenomics.

Chapter 1 – An introduction to evolutionary epigenomics

A central assumption of this thesis is that an improved understanding of genome evolution also leads to an improved understanding of the complex evolutionary interplay between the genome and epigenome. After a brief introduction of the well understood basic genome structure and relevant terminology, this chapter presents an overview of epigenetic genome regulation in order to link the library metaphor presented in the introduction to the scientific facts.
Next, a few essential evolutionary concepts are discussed. This is followed by approaches which utilize this knowledge for deciphering the genome function. The chapter closes by describing the initial research that links genome evolution with epigenetic genome regulation as well as an outline of how this thesis extends this work.

1.1 Basic Genome Function

The term genome comprises all the hereditary information of a cell. In most organisms, this information is encoded by deoxyribonucleic acid (DNA) molecules. These are assembled from two strands of four essential building blocks called nucleotides: adenosine (A), cytosine (C), guanine (G) and thymine (T). Along the strands, nucleotides are covalently linked by phosphate molecules. Across these strands, the bonds are mediated by two hydroxyl bridges between A and T (A/T) and three such bonds between C and G (C/G), respectively. This binding pattern results in the characteristic double helix shape of the DNA molecule, such that one strand encodes the reverse complementary sequence of the other. In some viruses and as temporary information carrier ribonucleic acid (RNA) fulfill a similar function. RNA differs from DNA mainly by the substitution of T by the biochemically very similar uracil (U), and the replacement of deoxyribose by ribose as the sugar component of the nucleotide.
In order to discriminate nucleotide base pairs (bp) across strands from dinucleotide pairs along one strand, the former are denoted by C/G and the latter by CpG, where p symbolizes the phosphate bond linking two individual nucleotides to form a dinucleotide. Thus, CpG/CpG denotes a double-stranded dinucleotide. Applying this four letter alphabet, macromolecules consisting of millions of bps have evolved in many species. These are condensed into compact and highly structured complexes called chromosomes, which are replicated for each cell division by separating both DNA strands and facilitating the complementary base pair principle to synthesize two new double helices in a supervised energy consuming process. In this way, each daughter cell can receive a complete copy of the genome during cell division (Alberts 2002).

To understand how information is encoded in the genomic sequences of these DNA molecules, we can return to the newspaper analogy from the introduction. Each nucleotide is a part of a long text, but to the observer it is unclear which nucleotides form words or sentences. Parts of the text form functional units and are referred to as genes. Teams of molecules called protein complexes produce copies of the DNA in the form of RNA transcripts, which are rapidly degraded within cells. At the start of each gene, there is an administrative area known as promoter, which contains one or more ‘headlines’ where the protein complexes assemble before they start to transcribe the gene. These ‘headlines’ are referred to as transcription factor binding sites (TFBS) and are *cis*-regulatory elements, while the transcription factors (TF) that bind them are called *trans*-regulatory elements. TFBSs are located in close proximity to the transcription start sites (TSS) and only attract specific TFs. The promoters can be co-regulated by more distant *cis*-regulatory elements called enhancers and repressors that either support or hinder the assembly of the protein complexes that initiate the transcription. Furthermore, some TFBSs contain CpGs that can be modified by a process called DNA methylation, which results in an altered affinity of the TFs for this site.
Additionally, the accessibility of bigger text blocks can be temporarily or permanently restricted or encouraged, such that accessibility for proteins is altered. This process is mediated by covalent modifications of the proteins called histones, which organize the DNA structure.

Furthermore, the TF concentration changes over time. Each phase in the cell cycle has its own composition. All these influences vary the probability that a complete transcription complex is assembled at a particular promoter and ultimately, the frequency at which the associated gene is transcribed. Thus, gene regulation is a stochastic process.

Finally, the produced RNA transcripts are exported to other parts of the cell, where they directly regulate cell function (e.g., by regulating the transcription of other genes) or are applied as construction blueprints for new proteins. While the 'text' itself and thus the sequence of all DNA is called the genome, the epigenome as the sum of all epigenetic modifications mainly describes how access to the DNA is regulated to increase or decrease the chances of initiating transcription of a particular gene.

1.2 Epigenetics

Derived from the Latin word *epi*, which means above or on top of, epigenetics describes all the information that is memorized for more than one cell generation without altering the DNA sequence. A more specific definition refers to epigenetics 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). Interestingly enough, the term's definition itself is the subject of an evolutionary process. It was originally coined by Waddington to describe the connection of genetics, the science of heredity, and epigenesis, the process that produces complex organisms from a single cell by differentiation and organ formation (Allis, Jenuwein et al. 2007).

Since the definition of epigenetics is broad, it covers numerous mechanisms. This thesis will focus on two of the major epigenetic systems, *i.e.*, DNA methylation and covalent modifications of histone protein complexes. Both unfold their function mainly by regulating the accessibility of DNA for soluble factors, such as proteins.

1.2.1 DNA methylation

The four DNA building blocks adenosine (A), cytosine (C), guanine (G) and thymine (T) introduced above fall into two classes, in which each of the pyrimidine-based nucleotides (C and T) is paired with the corresponding purin-based nucleotide (G and A). While the A/T pairing is mediated by two hydrogen bonds, the C/G pairing is mediated by three, thus it is the stronger pairing. The length of such double-stranded DNA (dsDNA) is measured by the number of nucleotide base pairs (bp) and reaches up to a few thousand, *i.e.*, kilo base pairs (kbp) in single cell organisms and up to several million *i.e.,* mega base pairs (Mbp) per molecule for complex organisms.

Multiple modifications of the standard nucleotides have been discovered (Low et al. 2001; Guo et al. 2011; Wossidlo, Nakamura et al. 2011); the arguably most important one in vertebrate genomes is the addition of a methyl group to the 5' carbon atom of C that leads to the formation of 5'-methylcytosine (5mC). In principle, this change only affects the DNA strand to which it is applied, thus the information would be lost for the complementary DNA strand after genome replication. In many invertebrates and vertebrates, 5mC is mainly found in a CpG context (Bird 1980). In dsDNA, this dinucleotide forms a palindrome, which is either found in an unmethylated (CpG/CpG) or methylated (5mCpG/5mCpG) form, but rarely hemimethylated (5mCpG/CpG). This phenomenon is caused by special enzymes that copy the methylation information after DNA replication from the original to the newly synthesized DNA strand, which enables the formation of an epigenetic memory (Bird 2002).

CpG methylation is so widespread in mammalian DNA that 5mC is considered the fifth base of the genome (Novik et al. 2002). Recent measurements estimate the genome-wide CpG methylation level in some human tissues to be about 70–80% (Lister et al. 2009; Laurent et al. 2010). Thus, the human genome is subjected to global CpG methylation.
These methylation marks represent information that is written, maintained, read and erased by specific enzymes and proteins (Bird 2002). Since these processes are relevant to an understanding of the dynamics of DNA methylation marks, they are explained in more detail in the following section.

1.2.2 Enzymes related to DNA methylation

Earlier, DNA methylation was compared to a text markup that makes headlines attractive to some readers and unattractive to others. In this analogy, a 'headline' is equivalent to a particular nucleotide sequence that specifies the affinity with which a certain protein can bind to the DNA, e.g., within a promoter region. This binding affinity greatly influences the frequency and duration of protein recruitment, which consequently modulates the secondary effects that the protein induces at this genome locus. Ultimately, it influences whether transcription is initiated or inhibited. Naturally, such a 'headline' can only be affected by CpG methylation if its sequence contains a CpG.

Maintaining and setting methylation marks

DNA methylation at CpGs is mediated by a set of enzymes called *DNA methytransferases* (*DNMT*s). The protein *DNMT1* is also called a maintenance *DNMT* because it preferentially binds hemimethylated CpGs and methylates the remaining cytosine on the unmethylated strand (Bestor and Ingram 1983). In the cell cycle, the protein is highly expressed shortly after the onset of replication; furthermore, it is also associated to the replication machinery (Caiafa and Zampieri 2005). This ensures that the methylation marks are copied to the newly synthesized DNA strand with high fidelity.
De novo methylation of unmethylated CpGs is performed by the enzymes *DNMT3a* and *DNMT3b* (Okano et al. 1999). These are predominantly active during early embryogenesis and cell differentiation. It has been observed that loss of their activity during mouse development is lethal (Reik et al. 2001), thus proving their essentiality for survival.

Erasing methylation marks

In the early development of mouse embryos, two waves of epigenetic reprogramming occur. During these phases, most of the inherited methylation marks are erased. This results in the establishment of totipotent embryonic stem cells (*ES*) (Morgan et al. 2005).

It has not yet been clarified whether this is a consequence of *DNMT* suppression or actively performed by unknown factors (Chahwan et al. 2010). The observation that four transcription factors are sufficient to reset the methylation marks of somatic cells to 'deprogram' differentiated cells into pluripotent stem cells (Takahashi and Yamanaka 2006) supports the second hypothesis. Several mechanisms have been discussed which would allow an active demethylation process, including an oxidative cascade, such as via 5-hydroxyl-C and 5-carboxyl-C, enzymatic removal of the methyl group, direct base excision repair of 5mC and deamination of 5mC to T followed by a base excision repair of the induced T/G mismatch (Wu and Zhang 2010). The last mechanism is actually of greater relevance for this thesis and may play a direct role in the context of genome evolution. It will be revisited in subsection 1.2.3 on *CpG decay*.

Reading methylation marks

The epigenetic information stored in the form of methylated CpGs is read in various ways. Methylated CpGs in *cis*-regulatory elements can either attract or repel the binding of *trans*-factors. These *trans*-factors can have an activating or repressive influence on gene transcription.

Proteins that contain a methyl-CpG-binding domain (*MBD*), for instance, preferentially bind methylated CpGs. Subsequent to binding, they recruit additional factors and often unfold a repressive function, as in the case of the polycomb group proteins (Schwartz and Pirrotta 2007). In contrast, proteins that exclusively bind to unmethylated *cis*-regulatory elements often promote gene transcription, e.g., as has been reported for the transcription factor *Sp1* (Macleod et al. 1994) and the transcription factor *Egr1* (Whang et al. 1998). A special case is the insulator protein *CTCF* (Bird 2002), which has a high affinity for unmethylated DNA. At the same time, however, it has a repressive function for some genes while promoting the transcription of others. *CTCF* plays a major role in genomic imprinting (Reik and Walter 2001).

In vitro experiments have revealed that mouse cells without *DNMT1* show increased expression for 10% and reduced expression for 1–2% of their genes (Jackson-Grusby et al. 2001). Therefore, DNA methylation in mammals is primarily associated with the mediation of repressive signals.

1.2.3 CpG decay

The term *CpG decay* describes the increased mutation rate of cytosine in a CpG context. In many larger vertebrate genomes, CpGs appear at a much lower frequency than expected from the G and C content (CpG suppression). For instance, in the human genome about 1% of all dinucleotides are CpGs, while a frequency of about 6% is expected if all dinucleotides were to appear at the same frequency. CpG is the only dinucleotide that shows such a divergence from the expectation (compare Table 3.1). A number of computational studies have estimated that C in CpG context has an increased mutation rate of 10–50% (Arndt et al. 2003; Lunter and Hein 2004; Siepel and Haussler 2004; Hobolth 2008; Peifer et al. 2008).

This is mainly attributable to DNA methylation. The modified nucleotide 5mC is biochemically less stable than C. It is prone to spontaneous hydrolytic deamination into T. This temperature-dependent process is twice as fast for 5mC than for C, which is deaminated into the nucleotide uracil (U) (Shen, Rideout III et al. 1994). Furthermore, the repair of T/G mismatches is less accurate than the resubstitution of U by C (Holliday and Grigg 1993). While U is not present in a regular DNA sequence and thus easily recognized as the mutated nucleotide, the T/G mismatch is ambiguous and can be repaired in both directions, or simply transmitted to the next cell generation.
Furthermore, the active demethylation during embryogenesis potentially involves targeted deamination of 5mC followed by repair of the induced T/G mismatch as discussed above. The fidelity of this process is uncharacterized, thus opening the possibility that incomplete active demethylation leads to a substitution of C by T. This implies that in each generation, every actively demethylated CpG is exposed to such an additional mutational burden at least once. Only consistently methylated or unmethylated CpG sites are protected from it.
The term *CpG decay* comprises spontaneous deamination as well as the potential risk of error-prone active demethylation.

1.2.4 Evolutionary origins of DNA methylation

The evolutionary history of DNA methylation is an excellent example of a molecular biological mechanism which – once it was developed – acquired a variety of different functions. In prokaryotes, such as the bacteria *Escherichia coli*, DNA methylation primarily has a defensive function. By hosting proteins that methylate particular DNA sequences and produce restriction enzymes that cut unmethylated instances of these sequence patterns, bacteria gained the ability to recognize and destroy foreign DNA (Kobayashi, Nobusato et al. 1999). Thus, the combination of matching restriction and methylation enzymes formed an intracellular immune system against parasitic genome sequences. The ability to cut DNA of competitors or protect the own DNA against their attacks, provided an evolutionary advantage. Similar to an arms race, various types of restriction-methylation (*i.e.* RM gene complexes) evolved. These complexes target different sequence patterns and methylate nucleotides at different positions or are offensively directed to cut DNA at the methylation patterns of other RM gene complexes. This created the great variety of methylation sensitive restriction enzymes, which became an essential tool for modern molecular biology. Among the proteins that performed the DNA methylation were the precursors of 5mC targeted *DNMTs* that mediate DNA methylation in mammalian genomes. According to a comparative study, the last common ancestor of mammals, plants and fungi already possessed genes homolog to *DNMT1* and *DNMT3* (Zemach, McDaniel et al. 2010).
In model organisms from these subgroups of the taxon *Eukarya*, the defensive role of DNA methylation has shifted, as it is primarily suppressing genomic elements called transposons (Goll and Bestor 2005). These 'jumping genes' represent mobile DNA sequences that can change their location in the genome either by a 'cut-and-paste' or 'copy-and-paste' mechanism, depending on the particular subtype. More than 45% of the human genome consists of transposable elements (Goll and Bestor 2005). For most of these transposon copies, a direct contribution to cell function has not yet been characterized, may be very indirect or does not exist. Transposons are transcriptionally inactivated and immobilized by DNA methylation (Kato, Miura et al. 2003; Bourc'his and Bestor 2004). Additionally, the *CpG decay* effect contributes to an accelerated substitution of the contained CpGs. Thus, it initiates long-term inactivation by distorting the DNA sequence (Goll and Bestor 2005).

Early evidence of the regulation of regular genes can be found in some bacteria. In these prokaryotes, DNA methylation of adenine at the palindromic GATC pattern acquired a gene regulatory function (Casadesus and Low 2006). Whether these systems are predecessors of mammalian methylation mediated gene regulation or a case of convergent evolution remains to be clarified.

A central difference between DNA methylation in prokaryotes and eukaryotes is the frequency of the methylation sites. Methylated recognition sites of restriction enzymes and the regulatory GATC motif are more complex patterns than CpG, and thus appear at least two orders of magnitude less frequent in genomic sequence. A transition from the

longer patterns to the shorter ones thus induces the spread of a relative localized modification to a genome-wide phenomenon.
Global 5mC methylation of CpG sites is present in all large genome eukaryotes, but only in a few with small genomes (Goll and Bestor, 2005). This indicates a co-evolution with the increase of genome length. For instance,. *Dipteran* insects (e.g. flies) – but not *Hymenopteran* insects (*i.e.* the honey bee) – lack most of the *DNMTs* required to maintain the full regulatory potential of DNA methylation. Therefore, global DNA methylation may not predate the divergence of these insect species (Goll and Bestor 2005). As mentioned above, plants also possess homologs of these proteins; however, the best characterized model organism *A. thaliana* shows cytosine methylation in an arbitrary sequence context. Moreover, methylation is constrained to repetitive sequences, the coding sequence of highly transcribed genes and to less then 5% of its expressed gene promoters (Henderson and Jacobsen 2007). Until a counterexample is found, it is assumed that cytosine methylation is not a global phenomenon in plants. It is unlikely that global DNA methylation evolved early in the eukaryotic phylogeny and was then lost in all the organisms studied except for vertebrates. This places the most likely origin of global DNA methylation somewhere between the honey bee, which still has the required set of enzymes but shows no evidence of global DNA methylation (Zeng and Yi 2010), and the common ancestor of vertebrates.

Further evidence is provided by a comparative study of the CpG content in gene promoters. It showed that CpGs were notably absent from promoters in the bacterial genomes examined, slowly increased their concentration upstream of the TSS in the worm *C. elegans* and the fruit fly *D. melanogaster* and reached a broad peak in the mosquito *A. gambiae*. In zebrafish and humans, the CpG dense region also finally expanded downstream of the TSS (Khuu et al. 2007). This coincides with the independence of gene regulation from DNA methylation in the worm as well as the fly model organisms and the global presence of DNA methylation in zebrafish and humans. These novel CpG-rich promoter types are also called CpG islands, which are tightly linked to the regulation of transcription by DNA methylation and introduced in detail in the following section.

1.2.5 CpG islands

CpGs as targets of DNA methylation are not evenly distributed along mammalian genomes. In general, they are exceptionally rare. More specifically, CpGs are underrepresented by around five to sixfold compared to the frequency that is expected from the observed single nucleotide frequencies (Antequera 2003).

An exception to this rule are short regions which show an elevated GC content and a CpG frequency that equals the frequency of GpCs (Bird 1980). These regions are usually referred to as *CpG islands* (CGIs) (Gardiner-Garden and Frommer 1987), however, the terms CpG-rich islands (Bird 1986) and *Hpl* tiny fragments are also used. The term *Hpl* refers to the restriction enzyme applied for CGI discovery, which only cuts the unmethylated version of its CpG-containing target motif, and consequently fragments unmethylated CGIs into very small DNA segments (Bird and Taggart 1980). More recently, the term CpG clusters was also proposed (Hackenberg, Previti et al. 2006) to describe regions with high CpG density. The most prominent feature of CGIs is the absence of DNA methylation in most germline tissues, although they contain an exceptionally high concentration of potential targets for methylation. During differentiation, some CGIs become methylated in a tissue-specific manner (Bird 2002). Furthermore, they co-locate with origins of replication, *i.e.*, with those regions that are replicated first during DNA replication (Antequera and Bird 1999). This may contribute to their low methylation level because the concentration of *DNMT1* is very low during early replication. The absence of this enzyme may support the maintenance of the unmethylated state (Caiafa and Zampieri 2005). The unmethylated state is primarily maintained by methylation-determining regions (MDR), which correspond to *cis*-regulatory elements that function as binding sites for proteins (Lienert, Wirbelauer et al. 2011).

CGIs often occur upstream of genes and overlap with their TSSs. Approximately 60% of all promoters in human and mouse fall into this class of CGI promoters (Antequera, 2003). The methylation levels of CpG-rich promoters are reported to be anti-correlated with the expression of the associated genes, while no correlation can be observed for CpG-poor promoters. Furthermore, promoters with an intermediate CpG level can fall into each of these two classes (Weber, Hellmann et al. 2007). How CGIs influence the function of the respective promoters is not fully characterized. The next section briefly introduces the key hypothesis.

CpG islands as promoters

The promoter region of human and mouse promoter CGIs is located between the 5' start of the CGI and the gene's TSS (Cuadrado, Sacristán et al. 2001). CGIs function in a strand-unspecific manner. Thus, a single CGI can activate genes on both strands if the two TSSs are located accordingly (Adachi and Lieber 2002; Carninci, Sandelin et al. 2006).

A variety of transcription factors bind to CGIs (Stapleton, Somma et al. 1993; Tommasi and Pfeifer 1997). For example, *Sp1* is a ubiquitous transcription factor with a GC-rich binding motif (*GC box*), which binds to many CGIs. Its presence is a signal to keep the region unmethylated, and thus functions as an MDR (Macleod. Charlton et al. 1994; Lienert, Wirbelauer et al. 2011). Interestingly, the protein is not necessarily involved in the expression of the associated gene in all tissues (Marin, Karis et al. 1997) and does not influence the expression of many other genes to whose promoters it binds (Saffer, Jackson et al. 1991). This can be explained by the presence of multiple *Sp1* binding sites per promoter (Macleod, Charlton et al. 1994) and other *cis*-regulatory sequences. The observation that a CGI promoter harbors multiple TSSs (Aimée and Bird 2011) supports the assumption that it hosts several independent transcription initiating modules.
Although the interplay between CGIs and *TSS*s is not yet fully understood, this association is strong enough for CGI discovery to be applied in the identification of novel genes (Bird 1987). It was recently demonstrated that although this approach has been used for more than two decades, improvements in this area can still lead to the identification of a substantial number of new functional elements, even for genomes that are investigated in depth, such as those of human and mouse (Illingworth, Kerr et al. 2008; Illingworth, Gruenewald-Schneider et al. 2010). In particular, the prediction of gene promoters for functional transcripts that do not encode proteins but regulate the transcription of other genes – so-called micro RNAs – can be improved by considering CpG dinucleotide distribution (Bhattacharyya, Feuerbach et al. 2012).

Origin of CpG islands

The *CpG decay* effect accounts for the substantial influence that global DNA methylation has had on the genome-wide distribution of CpGs. Therefore, the spread of DNA methylation from a phenomenon that was localized to restriction sites and transposable elements to a genome-wide modification is reflected by a loss of CpGs in the affected genomes (Khuu, Sandor et al. 2007; Yi and Goodisman 2009).

Initial CpG-rich but methylation-free genome regions were observed in vertebrate genomes (Bird 1980). These regions were assumed to be unmethylated in the germline to retain their high CpG frequency (Bird 1980). Since then, several statistical approaches have quantified the impact of DNA methylation on genome evolution (Sved and Bird 1990; Arndt, Burge et al. 2003; Lunter and Hein 2004; Siepel and Haussler 2004; Hobolth 2008; Peifer, Karro et al. 2008) leading to the conclusion that the decay of CpG/CpG into TpG/CpA dinucleotides occurs at a rate that is 10 to 50 times higher than other single nucleotide substitution processes. This provides a plausible explanation for the strong CpG depletion in mammalian genomes.

CpG decay is the only point substitution process outside of the proteincoding region that significantly depends on neighboring nucleotides. The signal produced is strong enough to render the distribution of CpGs indicative of the presence and degree of global DNA methylation in a vertebrate genome (Jabbari, Cacciò et al. 1997; Glass, Thompson et al. 2007). Thus, *CpG decay* translates local epigenetic differences in the germline methylation level by methylation-mediated C to T transitions into genomic differences of CpG density. Following this line of evidence, CpG depletion in the bulk genome is a consequence of global DNA methylation, whereas CGIs reflect the ancient unbiased nucleotide distribution to some extent.

To understand why DNA methylation has spread from a locus-specific modification to the genome-wide default state, it is instructive to correlate this modification with the genome size of the inspected species. In larger vertebrate genomes, CpG methylation is also abundantly present in non-coding regions and repetitive elements (Bird 2002). In these cases the common function is an inhibition of transcription at unfavorable positions. It is possible that the demand for tighter regulation is higher for genomes with large non-coding and putatively neutrally evolving domains that are rich in transposable elements. Thus, global DNA methylation may be one of the mechanisms that enabled vertebrates to increase their genome size without losing too many resources for the production of useless or harmful RNA transcripts. It implemented control over spurious transcription initiation and the spreading of transposable elements (Liu and Schmid 1993; Belancio, Roy-Engel et al. 2010). Furthermore, it enabled the deactivation of parasitic elements such as retroviral DNA (Kato, Ahmed et al. 1996).
As discussed above, in the unmethylated state CpG-rich gene promoters seem to attract ubiquitously expressed transcription factors such as *Sp1* via low-complexity but CpG-rich DNA motifs. These motifs are present at multiple positions per promoter (Aimée and Bird 2011). In certain promoters, an unmethylated germline state protects these motifs from *CpG decay* whereas they are under increased pressure in the remaining genome.

1.2.6 Histones and their modifications

Genomic DNA molecules in the eukaryotic nucleus are highly organized to optimally use the limited space. In eukaryotic cells, the basic unit of organization is the nucleosome. This macromolecular structure consists of a protein octamer of two times four core histone proteins (H2A, H2B, H3 and H4), around which 146 bp of DNA is wrapped (Luger, Mader et al. 1997). With a distance of up to 80 bp until the next nucleosome begins, the diploid human genome is organized into approximately 30 million nucleosomes, each covering an average of 200 bp of DNA (Alberts 2002). Each histone has an N-terminal tail that extends well beyond the nucleosome boundaries, which makes it accessible for soluble factors such as enzymes. These tails are covalently modified in different ways, where each modification amounts to an individual epigenetic signal. These signals attract or repel specific proteins (Luger and Richmond 1998). Acetylation (ac) and mono-, di- or tri-methylation events (me1, me2 and me3) that are targeted to the amino acid lysine (K) at different positions of these tails are of special interest. They are either markers for compressing the DNA into dense, inaccessible heterochromatin, or unfolding it into open and putative actively transcribed euchromatin (Rice and Allis 2001).

Interplay of CGIs, DNA methylation and histone modifications

DNA methylation and histone modifications partially depend on each other. For instance, histone deacetylases (*HDAC*s), which are able to remove the activation signaling acetyl groups from histone tails, are found in protein complexes that also contain the *MBD* domain. Thus, these complexes have an elevated affinity for methylated DNA. Consequently, CpG methylation marks can attract enzymes that initiate the remodeling of the local chromatin structure. Furthermore, unmethylated as well as methylated DNA is bent sharply to form a complex with the histone octamer. This is energetically unfavorable for homopolymers that are either very rich in GC or AT (Rice and Allis 2001). This contributes to a nucleosome-free state of active *CGI* promoters (Caiafa and Zampieri 2005), which explains why these genome regions are hypersensitive for proteins that preferentially cut DNA, which is not bound to nucleosomes (Gross 1988). Additionally, specific histone modifications, such as H3K4me2 and H3K4me3, are strongly correlated with the absence of DNA methylation, and are assumed to repel DNMTs in concert with other factors (Edwards et al. 2010).

1.2.7 Computational epigenetics and epigenetic footprints in the DNA

Epigenetic marks like histone modifications or DNA methylation are not independent of the genome sequence. The sequence-based identification of *CGIs* is arguably one of the first approaches to investigate this connection (Gardiner-Garden and Frommer 1987). A more sophisticated class of computational approaches focus on a fine-grained identification of epigenetic footprints in the DNA. For instance, several studies searched for DNA sequence features that differentiate between two types of CpG-rich regions. The first are regions that are prone to *de novo* methylation during development or are methylated per default in all measured tissues and the second are putatively functional *CGIs* that are unmethylated in most somatic tissues (Feltus, Lee et al. 2003; Bock, Paulsen et al. 2006; Das, Dimitrova et al. 2006; Feltus, Lee et al. 2006; Bock, Walter et al. 2007; Straussman 2009). These approaches apply machine learning methods to identify features of the DNA sequence that best correlate with the epigenetic state of the respective genome regions. While most of these approaches aim at the identification of predictive correlations, a closer examination of the discovered features can enhance our understanding of the exact mechanisms that mediate the interplay between genome and epigenome via the *CGI* function.

The most recent approaches go a step further by taking evolutionary aspects into account. For instance, the conservation of CpGs is applied for the identification of *CGIs* (Cohen, Kenigsberg et al. 2011). Moreover, the estimations of the intensity of deamination were used for a comparative analysis of *CGIs* in human and mouse. This study found similar deamination rates in CpG island promoters, while increased deamination rates are estimated for the remaining mouse CGIs (Hutter, Paulsen et al. 2009).

Improvements in the characterization of epigenetic footprints in the DNA can complement the results obtained from molecular biological experiments.

1.2.8 Summary

The epigenome is an information system that complements the genome. In multicellular organisms, it primarily encodes the relevance of different parts of the genome for the particular cell type. Ultimately, this enables a fine-tuning of gene transcription from the repression of individual binding sites to the silencing of large parts of one X chromosome in female mammalian cells.
DNA methylation as one of the most prominent epigenetic modifications has undergone a remarkable evolutionary process from a rudimentary DNA-based immune system in bacteria to a genome-wide regulatory mechanism and repressor of unfavorable transcription. It has had a deep impact on the evolution of vertebrate genomes via the *CpG decay* effect. In order to elaborate on the evolutionary interaction of genome and epigenome later on, the next section introduces the basic concept of this field.

1.3 Genome Evolution

Charles Darwin postulated two concepts as the main driving forces of evolution, namely variation (mutation) and natural selection. While he described these forces mainly on the macroscopic phenotypic level, he also remarked: “Under nature, the slightest differences of structure or constitution may well turn the nicely balanced scale in the struggle of life, and so be preserved” (Darwin 1864, p. 80). This thesis investigates how the influence of these forces is reflected on the microscopic genetic level, i.e., on the genomic DNA sequence itself.

Mutations that lead to the exchange of single DNA nucleotides are called point mutations or base substitutions. The frequency of these mutations differs according to their biochemical properties. For instance replacement of the pyrimidine T by the pyrimidine C is more likely than a mutation into the purin A. The previously discussed *CpG decay* effect is another special case that increases the substitution rate. Averaged over larger genome regions, the probability of observing particular types of point mutations is described by base substitution models. The impact of a point mutation on the fitness of a species influences the direction and strength of the force of natural selection on it.

Thus, these substitution models summarize the convolution of mutational and selective forces. To prepare the ground for a substitution model that incorporates *CpG decay*, I will briefly introduce the biological background of point mutations and some of the limitations of such substitution models before we encounter them again in mathematical detail in chapter 3.

1.3.1 Substitution models

Substitution models basically postulate that genomic point mutations of one particular type of nucleotide into another are a stochastic process with fixed rate. The differences between these rates define how fast these substitutions occur relative to each other. Such a model, for instance, may state that transitions (pyrimidine-pyrimidine and purin-purin exchanges) appear at a frequency that is five time higher than transversions (pyrimidine-purin exchanges). Within the framework of such a model, time is defined relatively, *i.e.*, the model can state how many substitutions of type A can be expected during a time span in which one particular substitution of type B is observed. This approach has proved to be very valuable for various applications, e.g., for the reconstruction of the evolutionary tree of life (Delsuc, Brinkmann et al., 2005).

For the methods proposed below, it is of importance that not all mutations in multicellular organisms are passed on to the next generation. In mammals, the germline cells are separated from the remaining somatic cells. In the germline cycle, they comprise a sequence of tissues, which include embryonic stem cells, pluripotent stem cells, primordial germ cells, and ovarian and sperm germ cells in humans. The genome traverses this germline cycle once each generation in order to be passed on to the next individual. Only the DNA changes that occur in these germline cells can be potentially transmitted to the next generation (Johnson, Richardson et al. 2011). Thus, point mutations that are triggered by DNA methylation via the *CpG decay* effect can only affect CpG sites that are methylated in at least one germline tissue.

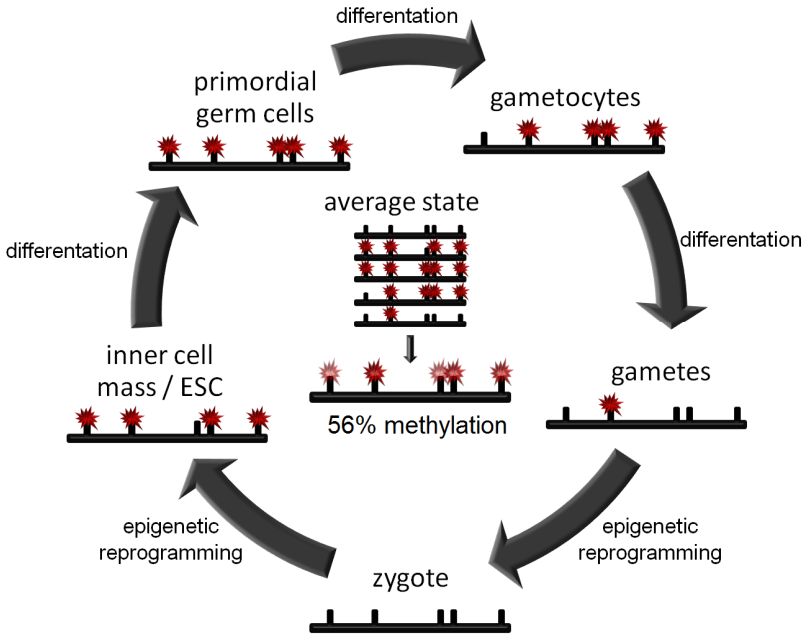


Figure 1.1: Differential methylation in the germline cycle

Differentiation and epigenetic reprogramming induce a dynamic germline methylome. Only CpGs that are methylated within one germline tissue can be affected by the CpG decay effect.

Some factors that influence mutation rates in the germline are species specific. Therefore, "it is well established that rates of substitution naturally vary across species and lineages" (Li and Drummond 2011). For instance, the presence of molecular-biological mechanisms for DNA repair can directly correct a number of mutations. Nutrition and metabolism influences the concentration of intracellular oxygen radicals, which in turn have an impact on the mutation rate (Britten 1986). Finally, biological factors, like the number of inherently error-prone genome replications per generation and the physical time that has elapsed until the genome is passed on to the offspring contribute additional variance to the species-specific mutation rates (Britten 1986). Hence, some of the methods introduced in the thesis are specific for the human genome. Their application to the genomes of other species requires a recalibration.

1.3.2 Selection

Once a point mutation is introduced into the genome, three general types of selective pressure may influence its fate:

(1) Mutations that have no influence on the fitness of the individual are called neutrally evolving. They are not subject to selective pressure.
(2) If a mutation has a negative impact on the individual's fitness, it is subjected to a negative or purifying selection because the corresponding sequence will probably be eliminated from the population over time.
(3) Mutations that improve the individual's fitness are positively selected and can eventually outperform the original state. If a mutation establishes itself as the only version of this genome region in the population, it has reached fixation.

The proportions of the pure point mutation frequencies to each other are best reflected in neutrally evolving sequences.

1.3.3 Comparative Genomics

A key problem in understanding complex genomes is that they are a mixture of functional, potentially functional and nonfunctional parts. To illustrate this, it is instructive to return to the library example. Some books in the library contain the blueprints of products that are of vital importance, for instance, a plan for a protein that can convert starch into sugar. Other books may contain blueprints of products and functions that used to be important or have been outsourced to simians. Parts of such books may potentially be reused in the future or rewritten to fulfill a novel but related purpose.
However, there are also books that accumulate so many random changes that they are beyond repair. Since no one has an overview of the whole library, they are not actively removed. Also, the budget of the library is large enough to even replicate such useless books for the opening of a new branch. From time to time, some of these useless books are degraded beyond recognition. Nevertheless, the library is still crowded with other books of this type. Without a cellular mechanism that has a holistic insight into cell function or extensive selective pressure on saving even a few additional nucleotides, it is unlikely that such a scenario can be prevented.

In other words, it is very likely that every part of a mammalian genome is either functional or a copy of a once functional region with uncertain potential for being beneficial in the future. One approach to deciphering genome function is the attempt to identify those parts that show clear signals of decay or at least rapid changes.

A key method for achieving this is by comparing the genomes of related species. In general, it is expected that the “common features of two organisms will often be encoded within DNA regions that are conserved between the species” (Hardison 2003), while nonfunctional features are subject to neutral evolution. This conservation is not only reflected by pure sequence identity, but also by retaining a sufficient affinity for proteins that bind the functionally conserved DNA (Schmidt, Wilson et al. 2010). In order to understand how such conserved DNA is identified, the *alignment* and *DNA motif* are essential concepts. The following two paragraphs introduce them in more detail.

Alignments

The most common technique for describing the relationship of DNA sequences at different loci that originate from a common ancestral sequence is a sequence alignment. The DNA sequences are represented as text strings over the nucleotide alphabet $\Sigma_A = \{A, C, G, T\}$. "A (global) alignment of two strings S1 and S2 is obtained by first inserting chosen spaces (or dashes), either into or at the ends of S1 and S2, and then placing the two resulting strings one above the other so that every character or space in either string is opposite a unique character or a unique space in the other string." (Gusfield, 1999, p. 216) This mapping can be interpreted as the evolution of a common ancestral sequence into the two observed sequences, which allow the deletion, insertion or substitution of single sequence letters. Hence, such a pairwise alignment represents a hypothetical evolutionary history for two loci and the degree of conservation can be estimated based on the number of documented changes.

Alignments are generated by different algorithms and heuristics with individual trade-off balances between runtime and accuracy for pairwise (Needleman and Wunsch 1970; Smith and Waterman 1981; Lipman and Pearson 1985; Altschul, Gish et al. 1990; Schwartz, Kent et al. 2003), multiple (Thompson, Higgins et al. 1994; Notredame, Higgins et al. 2000) and genome-wide multiple alignments (Blanchette, Kent et al. 2004; Paten, Herrero et al. 2008).

For highly conserved genome regions such as protein coding sequences, the alignments can even be computed for distantly related species. In contrast, regulatory sequences are less well conserved (Farré 2007). Consequently, related sequences (*i.e.*, homolog regulatory sequences) such as promoter sequences are harder to identify by alignment (Margulies, Chen et al. 2006; Margulies and Birney 2008). However, in order to identify the conserved regulatory elements, methods such as phylogenetic footprinting techniques concentrate on "regions which undergo significantly less changes than others" (Miller, Makova et al. 2004). Therefore, these approaches interpret alignments on a quantitative level. Most of them operate under the assumption that the most parsimonious version of the evolutionary history of a locus is the correct one and should be reflected by the alignment. This decreases the runtime of the algorithms and heuristics, but introduces an undetermined bias into the calculation (Hein, Wiuf et al. 2000). Alternatively, statistical alignments take all of the possible evolutionary histories into account. Thus, they compute the joint likelihood of the observed sequences being linked to a common ancestral locus via a series of mutation events. If necessary, the alignments from these alternative histories that most contributed to the total likelihood can be highlighted (Hein, Wiuf et al. 2000).

DNA motifs

While alignments mainly capture similarities in the DNA sequence, DNA motifs are a more flexible format for describing functional DNA elements. They are identified either by top-down or bottom-up approaches (Hannenhalli 2008).
In the first case, a library of sequences representing the binding affinity of a given DNA-binding protein is assembled by isolating the DNA molecules bound by the protein of interest (Elnitski, Jin et al. 2006). These sequences are then aligned and the information obtained is compressed into a representation based on the nucleotide frequencies at each position of the binding site (D'Haeseleer, 2006). The resulting motif can then be applied for localization of the yet unknown protein binding sites in the genome.
In the bottom-up approach, sets of conserved regions are analyzed for recurrent motifs in one or several genomes (Smith, Sumazin et al. 2005).
In contrast to epigenetic modifications, these *cis*-regulatory elements are 'hard-coded' in the DNA sequence and thus present in all cell types as well as during all developmental stages of the organism.

1.4 Comparative epigenomics

As the name implies, comparative epigenomics combines approaches from epigenomics and comparative genomics. This interdisciplinary approach is required because many epigenetic modifications receive their identity via the genome sequence to which they are attached. For instance, a histone complex with an H3K4me3 modification carries little information in itself, but indicates an active promoter region at the genome region that it binds to.

In a pilot study, I demonstrated that pure sequence conservation across mammal genomes is a poor indicator for absence of methylation, but that a conservation of GC and CpG content improves these predictions (Feuerbach, 2007). Very recently, a broader study of conserved epigenetic modifications of pluripotent stem cells from human, mouse and pig were correlated to the conservation of the underlying DNA sequence. This study confirmed that the conservation of epigenetic modifications is not correlated to sequence conservation (Xiao, Xie et al. 2012). Hence, a specific methodology has to be developed to extend comparative genomics methods to comparative epigenomic methods.

A further advantage of comparative epigenomics is that the genome is representative for a whole species and encompasses its full genetic information, while an individual epigenome, such as the methylome of a particular cell type, is a limited representation. To this end, classical comparative genomic approaches can be applied to gain insight into the epigenetic regulation system of a species. The comparative genomic studies of methyltransferase gene conservation discussed above exemplify this statement. For instance, the sequence-based search for genes that encode *trans*-regulatory elements, which are involved in epigenetic regulation such as *DNMTs* (Zemach, McDaniel et al. 2010), provided insight into the conservation of particular signaling pathways across the plant and animal kingdom.

Another class of studies measured the genome-wide distribution of CpG dinucleotides or their concentration around promoters in different species in order to gain insight into changes in the regulation by and the abundance of DNA methylation (Jabbari, Cacciò et al. 1997; Jiang, Han et al. 2007; Khuu, Sandor et al. 2007; Irizarry, Wu et al. 2009; Lechner, Marz et al. 2013).

Up to now, most of these approaches have been rather descriptive and disregard explicit models of genome evolution.

1.5 Synthesis – Evolutionary epigenomics

The first chapter of this thesis introduced the separately studied fields of epigenetics and genome evolution. Comparative epigenomics represents an approach for a combined interpretation. Based on this, I will move on to a more explicit integration of these concepts. With the onset of global DNA methylation, the *CpG decay* effect left a footprint in the genome where selective pressure did not counteract this process (Figure 1.2). Thus, *CpG decay* plays a central role in evolutionary epigenomics as an interface between genome, epigenome and evolution. This thesis pursues four approaches for using the information of the epigenetic footprints left in the DNA to improve our understanding of genome function and evolution.

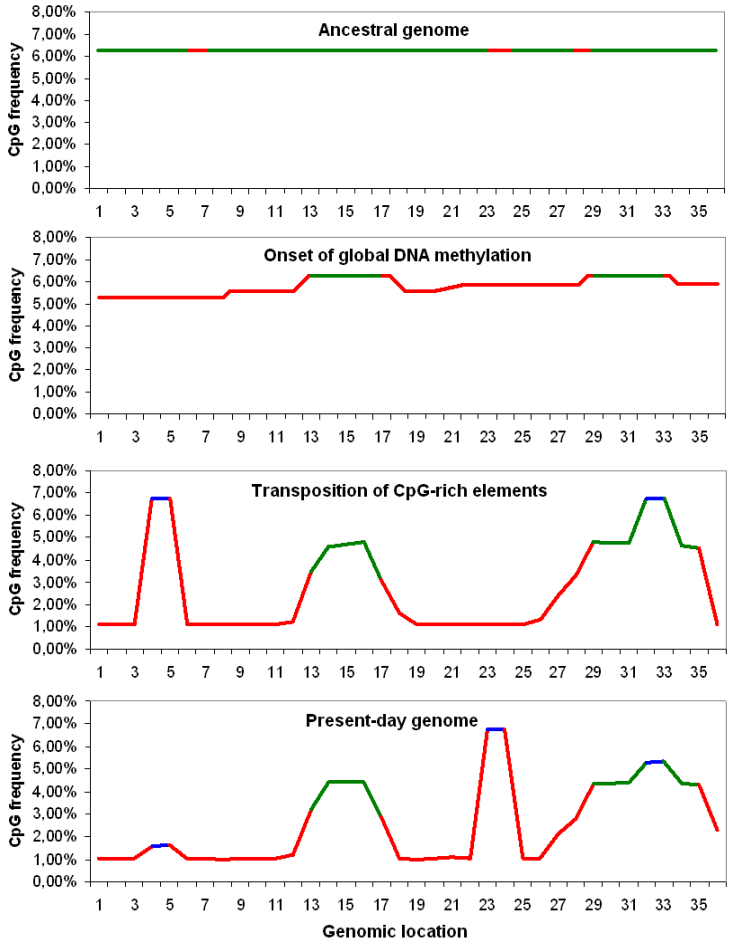


Figure 1.2: Establishment of epigenetic footprints in the DNA
This schematic drawing illustrates how the onset of global DNA methylation (red) in the largely unmethylated (green) ancestral genome leads to a characteristic reduction of CpG content. Selective pressure, unmethylated islands and the spread of CpG-rich elements (blue) such as Alu repeats were the major counteracting forces.

CpG-containing functional genome elements, such as *CGIs*, are assumed to be unmethylated in the germline in order to be protected from *CpG decay*. Without having access to the DNA methylation data, the annotation of CpG-rich regions produces maps of putative epigenetically regulated and functional genome regions (Figure 1.3). Thus, in chapter 2 of this thesis, I revisit the available approaches and discuss the mathematical and computational problems that limit their utility. This is followed by a solution to these problems in the form of the *CgiHunter* algorithm. Subsequently, the added value of *CgiHunter* is validated.

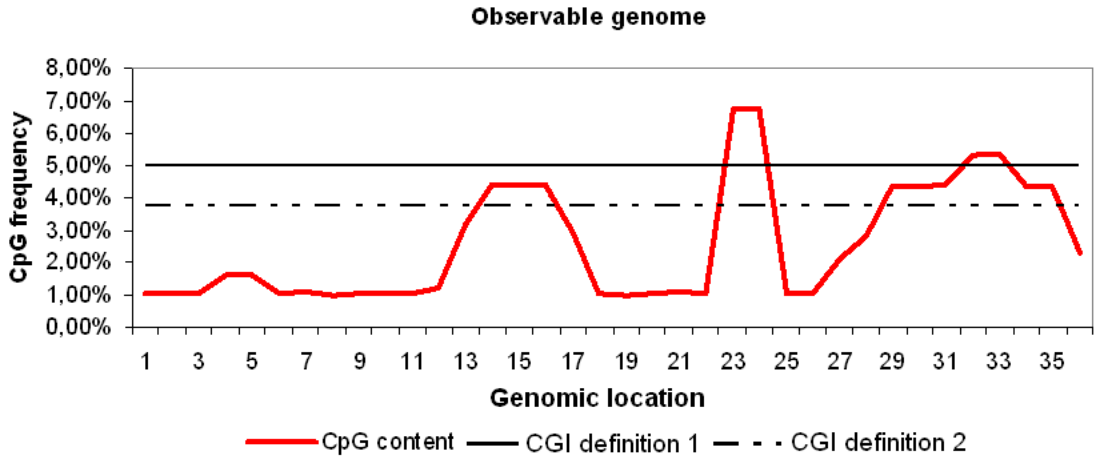


Figure 1.3: DNA sequence-based annotation of CpG islands
Without knowing the DNA methylation state, DNA sequence features can be applied to identify CpG islands. The choice of CGI definition, annotation algorithm and annotation parameters determines the size of the annotated regions and whether or not CpG-rich transposable elements are avoided or annotated.

Mutation rates in neutrally evolving DNA can be used to determine the dynamics and the equilibrium distribution of nucleotides and dinucleotides reached in methylated and unmethylated genome regions (Figure 1.4). Depending on their CpG and nucleotide content, functional elements such as transcription factor binding sites are subjected to different amounts of mutational pressure from *CpG decay*. Thus, the degree to which they profit from colocalization with the unmethylated genome regions varies. Explicit models of methylation-constraint genome evolution can quantify this pressure (chapter 3).

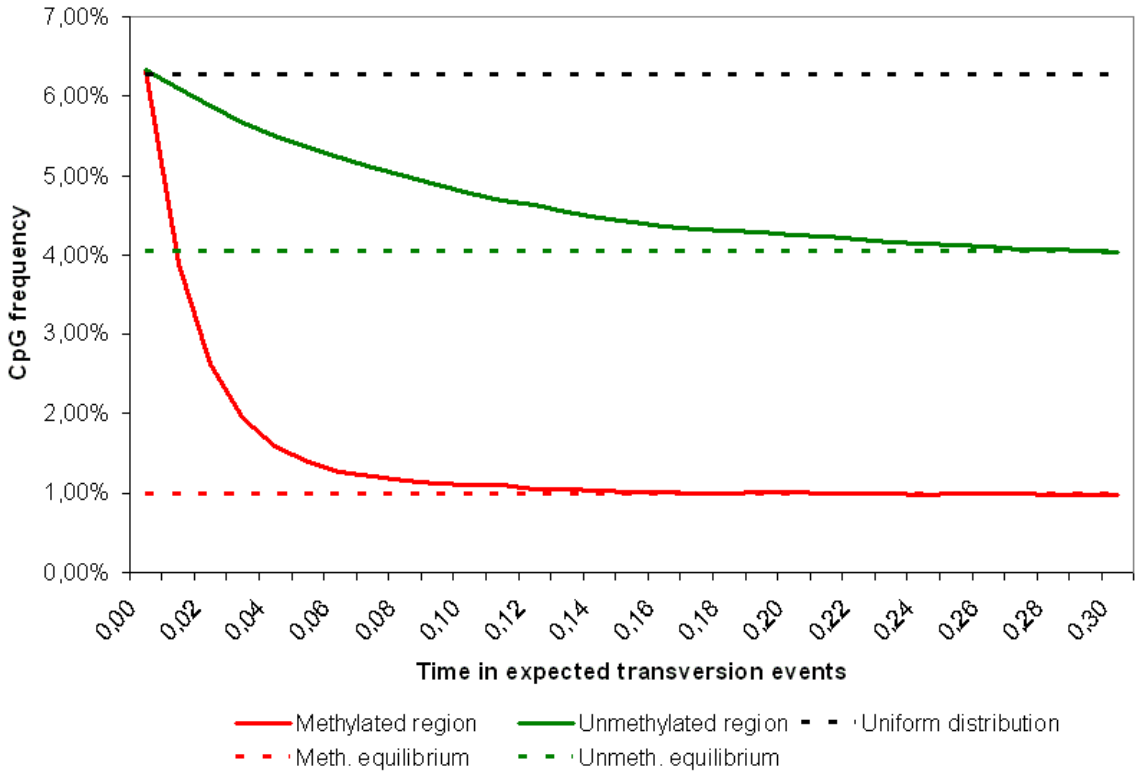


Figure 1.4: CpG frequency reaches different equilibrium distributions in methylated and unmethylated genome regions
In neutrally evolving DNA, the mutation rates determine the dynamic equilibrium around which every nucleotide and dinucleotide fluctuates. The rate at which equilibrium is reached depends on the nucleotide distribution of the ancestral DNA sequence.

CGI identification approaches are counteracted by CpG-rich transposable elements which, in evolutionary terms, are too young to be eroded by *CpG decay* to the CpG content level of the genomic background. Furthermore, some of these elements are suspected to have acquired novel functions, e.g., as an alternative promoter. A targeted approach to sort out the protected from the eroding transposon copies can help to solve this problem (Figure 1.5).

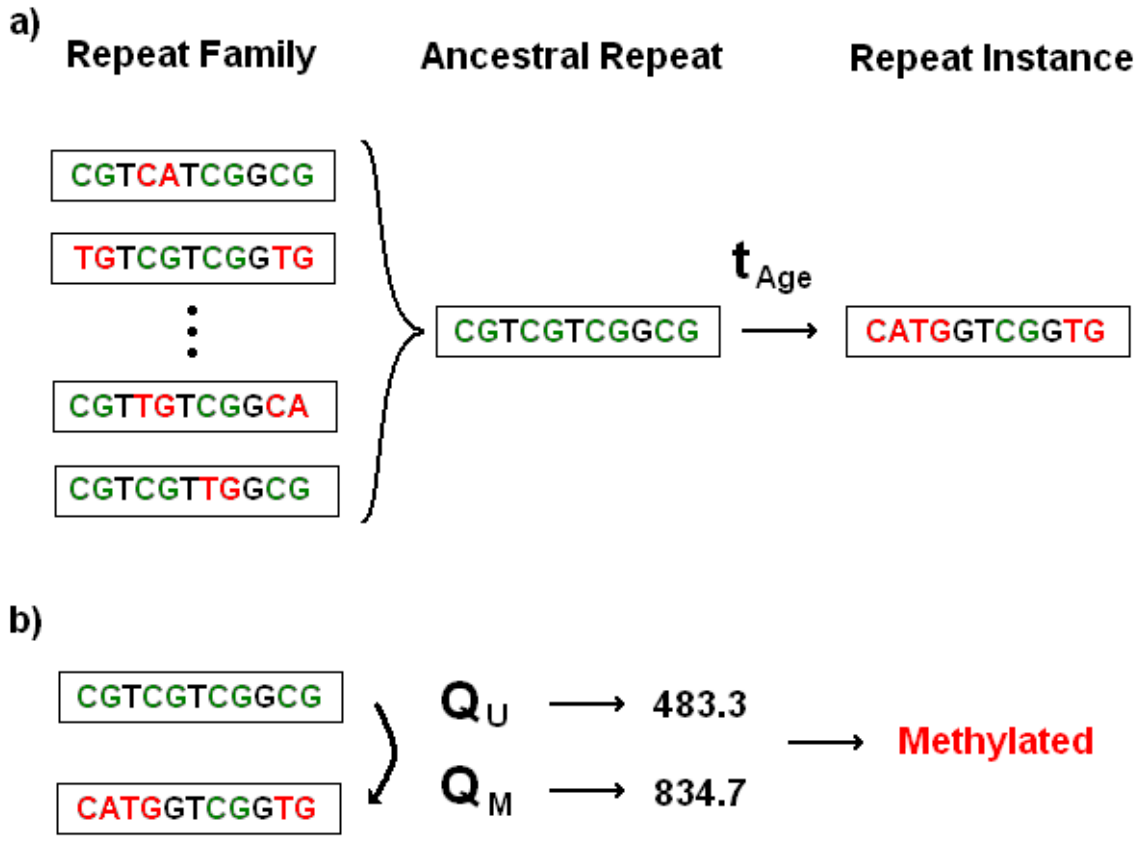


Figure 1.5: Prediction of transposon methylation level
(a) The numerous instances of a repeat/transposon family are applied to reconstruct the sequence of their last common ancestor, i.e., the sequence of the ancestral repeat. (b) By using mathematical models of sequence evolution for methylated (Q_M) and unmethylated (Q_U) DNA, the most likely germline methylation state of a single repeat instance can be predicted.

Chapter 5

Germline methylation levels have a great influence on the CpG content of a particular genome locus (Figure 1.6). By considering several genomes in parallel, these footprints can be applied to reconstruct the evolutionary history of the local germline methylome. In chapter 5, an algorithm is developed and gradually refined to perform such predictions. To this end, simulation studies are applied to characterize the properties of the method in detail. Finally, the algorithm is validated on the methylome data of human and chimpanzee germline tissue.

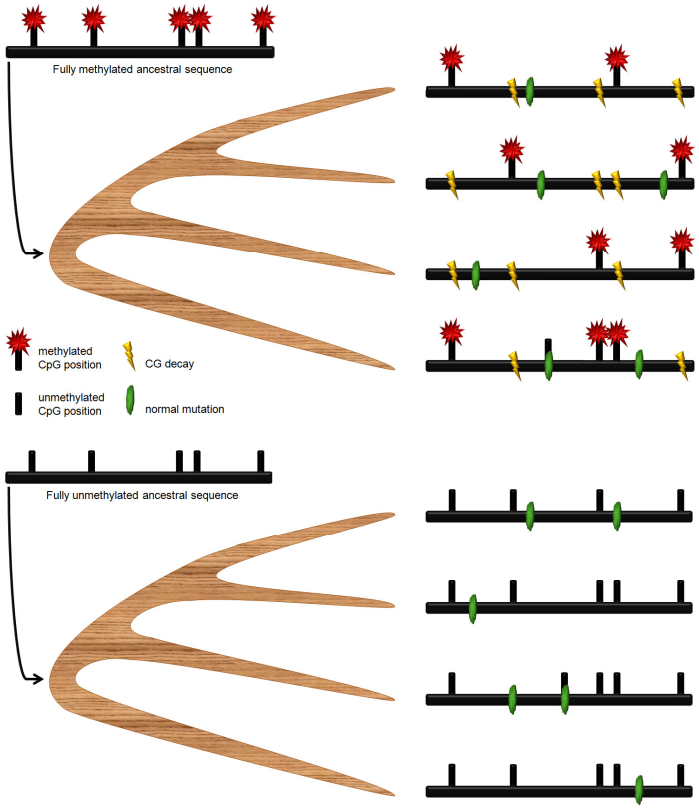


Figure 1.6: Epigenetic footprints in the DNA of homologous genome regions
This figure shows two versions of how the DNA sequence of a last common ancestor evolved into four descendants. In the upper panel, the ancestral sequence was fully methylated, whereas in the lower panel it was fully unmethylated. The number of mutated CpGs is strongly influenced by the methylation state of the DNA sequence.

Chapter 6

Finally, in chapter 6 a brief review of the methods introduced in this paper is given and discussed in the context of the annotation of novel vertebrate genomes. Furthermore, a summary is provided of the scientific insight gained from studying the evolution of genomes under the influence of DNA methylation. The concluding discussion of this thesis is based on these analyses.

Chapter 2 - Identification of CpG islands

DNA methylation marks are nearly exclusively detectable at Cs in CpG context, in the genomes of most mammalian tissues. Hereby, each CpG represents the epigenetic equivalent to the concept of a bit from information theory. Its information content is stably maintained, if it takes the states methylated or unmethylated. By default, most CpGs are methylated in the human genome, while the minority of unmethylated CpGs is primarily found in dense clusters called CpG islands (CGIs).
Molecular biological studies have revealed the correlation of CpG density with tissue-specific lack of methylation, activating histone modifications and transcriptional activity. More specifically, 40-60% of all human gene promoters overlap with CGIs. Thus, CGIs are implied as carrier of biological function especially in context of gene regulation. An accurate annotation of CGIs in the human, as well as in any other mammalian or higher vertebrate genome, is of great interest for the identification of genes and other genomic elements regulated by DNA methylation. Furthermore, such annotations enable a comparative analysis of *CGI* evolution.

Additionally, the human genome contains numerous regions of high CpG density that show none of the other characteristic indicators of biological function. On the level of the DNA sequence, the dividing line between CGIs and non-CGI regions in general is not clearly defined as well as the exact boundaries of individual CGIs. This complicates a systematic analysis. A large initiative is currently sequences 10,000 genomes of different vertebrate species (10K-Genomes-Scientists 2009). For a systematic characterization of this large resource a sound definition of a CGI and a reliable algorithm, which annotate these CGIs in genomic sequences, is required. This basis then enables a comparative analysis of CGI evolution. Especially, to decide if the function and mode of operation of homologous regulatory genome regions is conserved, these algorithms have to guarantee that no valid CGI is missed. To determine the best software for this project, the benchmark, as a comparison of different approaches on a common gold standard dataset, is the method of choice in computer science.

The CpG island annotation problem in a nutshell

To identify CpG islands by means of computational DNA sequence analysis, three elements are required: A definition of a functional CpG island, a sequence-based definition of a CpG island, which translate the functional definition into criteria that can be evaluated by a computer program, and an algorithm that identifies these sequence-based CGIs in genomic sequence.

As depicted in Figure 1.2, genome regions that are stably protected from *CpG decay* in the germline of globally methylated genomes form CGIs. These regions are of interest, as they frequently collocate with transcription start sites (TSS), *cis*-regulatory elements and maintain marks of active chromatin, such as the absence of methylation in somatic tissue and histone modification that are associated with transcriptional and regulatory activity. Such regions are biological functional CpG islands. In contrast, recently transposed CpG-rich elements that undergo rapid *CpG decay* or regions that are slightly enriched in CpG dinucleotide content without showing additional evidence for biologic function are just CpG-rich regions. A good CGI annotation preferentially annotated genome regions that are highly enriched in functional marks. The CGI annotation problem comprises the generation of such annotations, the demonstration of their quality and the selection of a good tradeoff between sensitivity and specificity for functional marks.

To solve this problem, sequence-based CpG island definitions follow three main strategies: First, CGIs are defined as regions with elevate GC content, an elevated ration of observed CpGs over the statistically expected number of CpGs and a minimal length (*classic CGI definition*). Second, CGIs are defined as regions of elevated CpG density if compared to the genomic background. Third, CGIs are defined via hidden markov models (HMM). These HMMs are calibrated by inferred transition probabilities from one nucleotide to the following nucleotide along the genome sequence *i.e.* the dinucleotide frequencies. Alternatively, the GC and CpG content of small genome regions are applied to define the states of the markov chain. There is no general agreement on a definition, the correct choice of the definition's parameter or an algorithm for the generation of CGI annotations.

During my work on the CGI annotation problem, I identified two mathematical pitfalls in the *classic CGI definition*. Solving them directly is computational expensive. Therefore, in cooperation with Christoph Bock the *CgiHunter* algorithm was developed. In the following, I describe this algorithm and validate its theoretical advantages compared with a heuristic that does not address both pitfalls of the *classic CGI definition*. Then, a comparison of CGI annotations generated by programs of all three strategies is conducted on a gold standard dataset showing that the theoretical advantages translate into an improved annotation quality. This chapter concludes with the discussion of the *CpG Mountain annotation*, which integrates several CpG island annotations by their specificity for unmethylated regions into a heatmap annotation of CpG island strength, thus presenting a solution for the CGI annotation problem.

2.1 Computational approaches for CpG island identification

Throughout this thesis I use the term *functional CGI definition* to describe an abstract biological concept. In contrast, the term *sequence-based CGI definition* is used to describe a set of DNA sequence-based rules that define regions in the genome, which have the potential to be instances of this concept. Where this distinction is neglected in the literature, *procedural CGI definitions* are applied that mix *sequenced-based CGI definitions* with the algorithms that are applied for the CGI annotation, and define the outcome of this procedure as functional CGIs.
In the following, the history of the CpG island term is introduced in more detail and a number of prominent CGI annotation procedures and definitions are discussed to illustrate the problems that arise from this procedure.

2.1.1 Classical CGI definition and sliding-window based algorithms

The term "CpG island" was actually coined in a study (Gardiner-Garden and Frommer 1987) that undertook the first systematic attempt to DNA sequence-based identification of regions that were formally known as HTF islands (Bird and Taggart 1980), CpG-rich islands (Bird 1986) or methylation free islands. Hereby, the term HTF stands for Hpa-tiny fraction and refers to the restriction enzymes *HpaI* and *HpaII*. These only cut unmethylated CpGs, and thus, enabled the observation of *CGIs* by a characteristic outcome of DNA restriction experiments with this enzyme.
To identify these regions by a computational approach, Gardiner-Garden and Frommer computed the *GC content* and the O/E_{CpG} (the number of CpGs observed in a region divided by the number of CpGs expected from its C and G content) in a 100 bp long window, which was moved along the DNA sequence with a step width of 1 bp. Regions were only reported if they exceeded thresholds of 50% for CG content and 0.6 for O/E_{CpG} over a length of at least 200 bp. Others have later used the same *sequence-based CGI definition*, but applied different thresholds. To discriminate this choice of parameters it is referenced as *GGF definition* with t_a=50%, t_b=0.6, t_c=200 bp (compare Box 1). The general approach to use these three sequence characteristics to annotate CGIs is called *classic CpG island definition*. In their publication the authors emphasized the ad-hoc character of the exact threshold choices as a working definition.

Box 1 – The classical CGI definition

Given a DNA sequence $R = \{A, C, G, T\}^n$ of length n, with R_i being the nucleotide at position i, we define a region R_{ij} with $0 \leq i < j \leq n$ and length $j - i$ as a subsequence of R such that R_i is the first nucleotide from the left that is contained in the region and R_j is the first nucleotide from the right that is not contained in the region. We call a set of three constraints a *classic CpG island definition* if they have the following form:

$g_{ij} := (\frac{\#C_{ij} + \#G_{ij}}{j - i} \geq t_a)$ (GC content criterion),

$o_{ij} := (\frac{\#CpG_{ij} \cdot (j - i)}{\#C_{ij} \cdot \#G_{ij}} \geq t_b)$ (observed vs. expected CpG frequency criterion),

$l_{ij} := (j - i \geq t_c)$ (minimal length criterion).

The variables $\#C_{ij}$, $\#G_{ij}$ and $\#CpG_{ij}$ refer to the numbers of cytosines, guanines and CpG dinucleotides, respectively, that are contained in R_{ij}. For regions with no CpGs, o_{ij} is set to False, and thus, exception for which the denominator becomes zero cause no problems. Individual *classic CpG island definition* are either referenced by the choices for the three thresholds t_a, t_b and t_c, or by introduced abbreviation, such as *GGF* or *TJ*.
R_{ij} is called a *CGI* according to such a definition if the Boolean variables g_{ij}, o_{ij} and l_{ij} are evaluated as True.

This computational approach was later modified in various ways, for instance by adapting the length of the moving window to fit the minimal length criterion t_c (Matsuo, Clay et al. 1993). As indicated in Figure 2.1, this change results not simply in a reduced runtime of the algorithm, but has a deeper impact on the results of the annotation procedure. In the actual publication the effect on the consistency of the algorithm was not investigated.

A second modification of Matsuo and colleagues was the raising of t_c to 500 bp to avoid the annotation of CpG-rich ALU repeats. This was a reaction to the observation that approximately two thirds of the *CGIs* derived by the *GGF definition* overlapped with these CpG-rich repetitive elements. This exemplifies the problems that arise from mixing the *functional* and *sequenced-based CGI definition*. Instead of explicitly formulating the expectations a CGI should fulfill, the parameters of the annotation procedure are altered until the resulting annotation is less offending to some implicit expectations.

Later a more complex algorithm was proposed (Takai and Jones 2002). It applies a three step strategy consisting of a seed steep, an extend steep and a pruning step. In the seed step a sliding window that takes the size of the minimal length threshold t_c is moved along the DNA sequence until t_a and t_b are satisfied. In the extend step the seed window is extended in both directions by windows of size t_c until either the *GC content* drops below t_a or the O/E_{CpG} drops below t_b. In the pruning step the window is then pruned in 1 bp steps until all constraints are again satisfied. The thereby identified region is reported as *CGI*.

The resulting *CpG Island Searcher* software was tested with eight different parameter sets for the underlying *classical CGI definition*s. These were then evaluated with respect to their overlap with the first exons/5'UTRs of genes, other exons, ALU repeats and regions without any of these annotations. Thus, implicitly *CGIs* were functionally defined as CpG rich regions that show great overlap with 5'UTRs of genes, but lesser overlap with intragenic, repetitive or 'unknown' region. The study conclude that the strictest definition tested (*TJ definition*: t_a=55%, t_b=0.65, t_c=500 bp) shows the best performance (Takai and Jones 2002). Thus, rather than being applied as a tool the *CGI* definition itself became a research topic.

The applied objective function is a very rudimental *functional CGI definition*. Therefore, this approach was a considerable advancement over previous approaches and the updated definition was readily accepted by the scientific community.
Still it had two major flaws. First, all three selected threshold were the most stringent criteria tested in their respective category. Thus, it is possible that better parameter sets can be found.
Second, the algorithm overlooks in the seed step a substantial number of regions that meet all three constraints (compare Figure 2.1). If the distance between two CpG rich genome regions is just a few bp longer then the search window, the extend step is never evoked. This problem can not simply be solved by applying a larger search window, as at other locations this greater length dilutes the *GC content* or O/E_{CpG} to a point were t_b or t_c are violated. I call this problem the *Single-sliding-window bias*. Box 2 introduces it in more detail. In section 2.3.1 the resulting deficits in *CGI* annotations are quantified in context of the human genome. As a consequence of this bias, a *CGI* definition that is applied with a single-sliding-window algorithm such as the *CpG Island Searcher* appears stricter than it is, as only a fraction of the regions that meet the criteria are reported.

Box 2: Single-sliding-window bias

If only a single window is moved along the DNA sequence, valid *CGIs* can be overlooked.
Given the sequence CGATC and the thresholds t_a=55%, t_b=0.6 and t_c=4. The properties of the three subsequences that meet t_c are given below.

Sequence	GC content	O/E_{CpG}	Length	CGI
CGAT	50%	4.0	4	No
GATC	50%	0.0	4	No
CGATC	60%	2.0	5	Yes
TCGATC	50%	3.0	6	No
CGATCA	50%	3.0	6	No
TCGATCA	42.9%	3.5	7	No

Table 2.1: Example of Single-sliding-window bias
Only a sliding-window of length 5 can identify a CGI with t_a=55%, t_b=0.6 and t_c=4 in the sequence TCGATCA.

Although no window of length t_c reaches the required GC content, the sequence contains a valid *CGI*. If we extend the example to the sequence TCGATCA, we can observe that also longer windows fail to detect the valid *CGI*.
But 5 is not a perfect choice for the window size. For instance, the sequence A*CGAAACC*A contains no subsequence of length 5 with more the 40% CG content but a valid CGI of length 7 (shown in italic). In section 2.3.1 a benchmark is performed to estimate how severely this bias impacts *CGI* annotations in the human genome.

A step towards solving the *Single-sliding-window bias* was made by the design of a nearly exhaustive search heuristic for *CGIs* (Hsieh, Chen et al. 2009). The authors explicitly identified another drawback of the general sliding-window method namely that multiple genome regions can overlap, satisfying all three constraints, but their union does not (*ambiguity problem,* compare Box 3 and Figure 2.1). Hence, the *classical CGI definition* is ambiguous and only heuristic elements in previously proposed algorithms lead to unambiguous annotations. Still the approach by Hsieh and colleagues cannot guarantee to identify every sequence that fulfills a given *classical CGI definition*.

Box 3 : Ambiguity problem

On the one hand, this problem is caused by the multiplicative influence of *#C* and *#G* on the O/E_{CpG}, while their relationship in the GC content is linear. This difference becomes relevant whenever the distribution of C and G is unbalanced within a region.

Length	#C	#G	#CpG	GC content	O/E_{CpG}
200	1	100	1	50.25 %	2
200	100	1	1	50.25 %	2
200	50	51	1	50.25 %	0.078
400	101	101	2	50.25%	0.078

Table 2.2: Unbalanced C and G distribution leads to ambiguous annotations

Thus, C-rich and G-rich DNA sequence can be located close to each other in such a way that they harbor partly overlapping regions that satisfy a given *classical CGI definition*. Therefore, it is unclear which candidate *CGI* is to be reported, although the example above demonstrated that the resulting offset can be large.

On the other hand, the GC-content threshold also leads to ambiguity, as visualized in Figure 2.1. Here, three equivalent annotations are possible for one CGI.

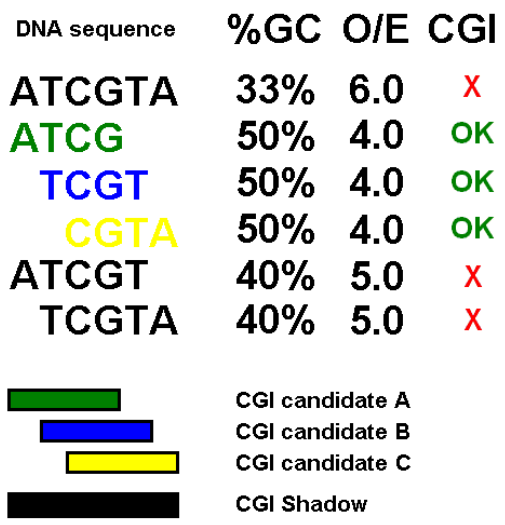


Figure 2.1: Ambiguity problem of overlapping CGIs

2.1.2 Approaches based on CpG-density

Half a decade after the publication of the *CpG Island Searcher*, its core concept, the sliding-window approach, was pointed out as the source of its unreliability. Under the assumption that CpG density is the DNA sequence property that correlates with absence of methylation, and thus, with biological function, the group around Hackenberg removed the layer of subjectivity introduced by this algorithmic technique and directly searched for CpG clusters. Their method measures the distance between all neighboring CpGs in a given sequence, estimates a characteristic threshold value from the distribution of these distances and subsequently constructs clusters of CpG dinucleotides, which lie closer to each other than this threshold (Hackenberg, Previti et al. 2006). In consequence, also very short segments are annotated. The method has the advantage of being efficient and more readily adjustable to the genomes of new species, but in direct comparison the *TJ definition* in combination with the *CpG Island Searcher* software produces annotations that more closely resembled those of unmethylated gene promoters (Han and Zhao 2009). A refined method that is based on the cumulative mutual information (*CMI*) of distances between CpG dinucleotides was later proposed by Su and colleagues under the name CpG MI (Su, Zhang et al. 2009). In addition to the density of the CpGs that is used by the *CpG cluster* algorithm, this method exploits more subtle information, which pertains to the exact spacing of the dinucleotides. Therefore, a number of additional parameters have to be chosen or learned from test data. As the method was compared with other approaches on the data from which these parameters were obtained, it remains to be clarified how this method perform in an unbiased benchmark.

2.1.3 Annotations based on Hidden Markov models

Hidden markov models (HMM) were another class of statistical models with great promises of improving *CGI* annotation (Durbin, Eddy et al. 1998). The transition probabilities from one nucleotide to the next in *CGI* and non-*CGI* regions yield an elegant eight state HMM. By inferring these transition probabilities from training sequences, the approach was expected to automatically produce a meaningful *CGI* definition. Unfortunately, this assumption was too simplistic, as the base composition of the genome also varies independently from the CpG island property. Thus, the number of hidden states is much larger. In consequence, this model never gained particular relevance for genome research (Wu, Caffo et al. 2010).
The authors that formulated this critique, furthermore proposed an alternative approach. Therefore, the scale of the HMM was changed from states in the range of single nucleotides to small genome regions (8-32 bp length) that are characterized by their G-, C- and CpG-content (Wu, Caffo et al. 2010). To avoid over-representation of repetitive elements, these were masked in the genome sequence. The authors demonstrated that the resulting model can be effectively fitted to genomes of different species (Irizarry, Wu et al. 2009) and reported an improved annotation of differentially methylated regions (*DMR*s) for an independently published human in-house dataset (Irizarry, Ladd-Acosta et al. 2009). I will refer to this approach as the *complex HMM*.

2.1.4 Critical summary

Three major problems prevent a generally accepted procedure for a *CGI* annotation based on the three classic constraints: (i) the lack of a *functional CGI* definition, (ii) the lack of an algorithm that guarantees the identification of all genome regions that meet a given *classical CGI definition* and (iii) the lack of a common gold standard for a comparative benchmark of all approaches.

Due to issue (i) and (iii) it is not possible to proof the quality of an individual annotation tool. While, for instance, the *CpG cluster* captures all CpG rich elements in a genome, *CpG Island Searcher* shows a better performance in annotating gene promoters, and yet another program may outperform the alternatives in annotating unmethylated sequence. Even if a program is selected, its annotations can greatly vary according to the parameters chosen for the annotation.
Problem (ii) is more readily addressable, as the ability of programs to identify genome regions that meet a certain *sequence-based CGI definition* can be estimated by mathematical proofs (compare 2.2.1) and with empirical studies (compare 2.3.1).
Ultimately, such methodically considerations are of minor interest if a program displays outstanding performance on a biological gold standard dataset. Only recently such dataset became available including genome-wide DNA methylation maps, catalogues of marks of open chromatin structure and hotspots of transcription initiation. Thus, for the first time a genome-wide benchmark is made possible.

To address problem (i), I propose to define a biological functional CGI (BF-CGI) as genome region with three properties:

(a) A BF-CGI is protected against *CpG decay* in the germline (*evolutionary property*)

The *evolutionary property* accounts for the fact that CGIs withstand the depletion of CpGs in the genome. In addition to the absence of methylation in the germline, strong selective pressure and biased gene conversion (Duret and Arndt 2008) are mechanisms that enable a region to meet criterion (a). In contrast, an ALU repeat with high CpG content that shows signs of rapid *CpG decay* (Feuerbach, Lyngsø et al. 2011) does not satisfy (a).

(b) A BF-CGI is unmethylated and associated with histone modifications that are marks for transcriptionally active chromatin in at least one tissue (*epigenetic property*).

The *epigenetic property* requires a BF-CGI to carry marks of active chromatin. A CpG-rich exon, for instance, may withstand *CpG decay*, but is probably methylated in all tissues. Thus, such an exon does not meet (b).

(c) The removal of a BF-CGI significantly changes the expression of at least one gene in at least one tissue (*regulatory property*).

The *regulatory property* excludes elements that autonomously maintain an unmethylated state, but are unable to influence transcription. In the benchmark below, the proximity to RNA polymerase binding sites is applied as an approximation for this functional aspect. Experimentally removing a CGI and measuring a significant change in the expression of at least one transcript in comparison to an unaltered control, would proof this feature.

Applying this definition, the quality of a *sequence-based CGI* definition is assessed in terms of its ability to predict regions that overlap loci with properties (a)-(c), while omitting region that do not show any of these properties. Finally, the quality of a CGI annotation software is evaluated by its ability to identify all regions that fulfill a given *sequence-based CGI* definition.

An empirical conducted comparison of five sequence-based CGI annotation programs came to the conclusion that no approach showed a clear advantage over the traditional sliding window-based annotation technique of the *classical CGI definition* (Hutter, Paulsen et al. 2009). Also on the theoretical level the *classical CGI definition* appears to be superior to the other methods for purely *sequence-based CGI* annotation approaches.

Unlike the simple HMM, *CpG cluster* and *CpG MI* the *classical CGI definition* considers the influence of varying GC content on the probability of observing CpGs by chance. Furthermore, it uses fewer free parameters than, for instance, the *complex HMM* and *CpG MI*. Moreover, these parameters can be more readily interpreted.
The drawbacks of the *classical CGI definition* are the *Ambiguity problem* described above and the *Single-sliding-window bias*. Both introduce a yet uncharacterized amount of bias and variance into the produced CGI annotations, *i.e.* small changes in the sequence can lead to complete skipping or significantly shifting the position of a CGI. This property biases comparative studies.

Hence, a sliding-window-based algorithm that solves the *Ambiguity problem* and the *Single-sliding-window bias* is desirable. Furthermore, the parameters for this algorithm have to be selected in such a way that the produced *sequence-based CGI annotations* have a high specificity for the three criteria of a BF-CGI.

2.2 The CgiHunter algorithm

The considerations at the end of the previous section lead to the conclusion that a reliable method for the identification of *CGIs* that meet a *classical CGI definition*, is a promising strategy for improving *CGI* annotations.

Therefore, the three problems named at the end of the last section have to be solved:

P1 – *Single-sliding-window bias*: a DNA sequence of length m that satisfies all three constraints of a *classical CGI definition* not necessarily contains for each window length l, with $t_c \leq l < m$ a subsequence that meets these criteria as well.

P2 – *Ambiguity problem*: two DNA sequences that overlap can individually meet all three constraints of the *classical CGI definition*, while the sequence that spans both sequences does not.

P3 – *Optimal choice of thresholds*: the optimal choices for t_a, t_b and t_c are unknown and may vary with respect to the interpretation of a *functional CGI definition* and the genome of interest.

2.2.1 A solution for the single-sliding-window bias

A direct solution of the single-sliding-window bias is a *brute force* approach, ehich individually investigates all subsequences of a genomic DNA sequence for their compliance with t_a, t_b and t_c. A sequence of length n contains $n(n+1)/2$ subsequences. For mammalian chromosomes, which reach a lengths in the order of 10^8 bp, such an approach is rather inefficient.
Therefore, we derived a *divide-and-conquer* algorithm that identify those subsequences that violate at least one of the three constrains. In consequence, the long DNA sequence is iteratively subdivided in a number of shorter, non-overlapping fragments until a *brute-force* approach becomes feasible. We call this the *filter step* of the *CgiHunter* algorithm. Its conceptual development was performed in close cooperation with Christoph Bock and a preliminary version has been reported before (Bock 2008).

The central idea is derived from the observation that upper and lower bounds on the C, G and GC content of a genome sequence are often sufficient to demine that it is not a *CGI*. Let $R_{ij'}$, R_{ij} and $R_{ij''}$, with $i < j' < j < j''$, be three genome regions that we call the red, the blue and the green window respectively.

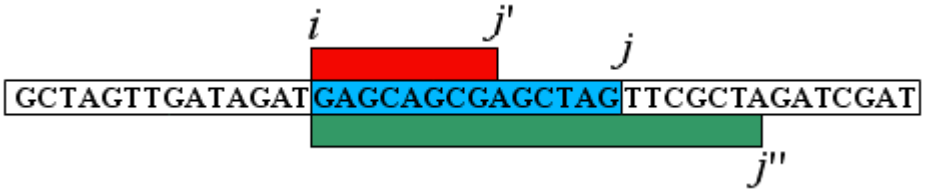


Figure 2.2: Illustration of CgiHunter filter interval
To deduce if any blue sequence between i and j with j'<j<j" is not a CGI it is in most cases sufficient to determine the sequence composition of the red and the green combing windows. Thus, many genome regions can be excluded from a computation intensive analysis with all window sizes by the filter step.

Instead of evaluating for any blue window R_{ij} if it is a *CGI*, we assess the G, C and CpG content of the shorter red ($R_{ij'}$) and longer green ($R_{ij''}$) windows (see Figure 2.2). It is obvious that all blue windows have the same or a lower nucleotide/dinucleotide count than the green window and the same or a higher nucleotide/dinucleotide counts than the red window. The GC content and observed vs. expected CpG frequency criterion are modified such a way that only values from the red and the green windows are used. At the same time it is guaranteed that the obtained values are always greater then the exact values.

We yield $\frac{\#G_{ij''}+\#C_{ij''}}{j'-i} \geq \frac{\#G_{ij}+\#C_{ij}}{j-i}$ and $\frac{\#CpG_{ij''}\cdot(j''-i)}{\#G_{ij'}\cdot\#C_{ij'}} \geq \frac{\#CpG_{ij}\cdot(j-i)}{\#G_{ij}\cdot\#C_{ij}}$.

Substituting these terms into g_{ij} and o_{ij} leads to the filter conditions:

$$g_{ij'j''} := \left(\frac{\#C_{ij''} + \#G_{ij''}}{j' - i} \geq t_a \right) \quad \text{(GC content filter)},$$

$$o_{ij'j''} := \left(\frac{\#CpG_{ij''} \cdot (j'' - i)}{\#C_{ij'} \cdot \#G_{ij'}} \geq t_b \right) \quad \text{(observed vs. expected CpG frequency filter)},$$

If already one of these terms is smaller than the respective threshold t_a or t_b, *i.e.* $g_{ij'j''}$ or $o_{ij'j''}$ are evaluated as False, all blue windows are bound to fail t_a and t_b as well and thus are excluded from the analysis.
The effort to determine the C, G and CpG content in the red and green window is proportional to their length and thus linear in contrast to the quadratic number of operations required to determine these values for all blue windows. Consecutively, the sequence position *i*+1 on is analyzed. Thus, the two bases that enter and leave the red and green window have to be determined. Once the content of both windows is assessed, the computational cost for shifting them by one base pair is independent of the window lengths, *i.e.* constant. For a genome sequence of length *n*, the window has to be shifted at most n times, and thus the complexity of this operation is O (n).

The efficiency of the analysis is increased by applying an array of windows. This array starts from size t_c, where the next bigger window is *cw* times the size of the current window. The largest window has the size *n*. We call this array of windows a *comb* and *cw* the combwidth (Figure 2.3). This parallel procedure enables each of the inner windows of the *comb* to once providing the lower bound values (red window) and once the upper bound values (green window). Thus, the efficiency of the algorithm is approximately doubled. The number of windows in the comb is proportional to $\log_{cw}(n)$. Thus, the overall complexity of the filter step is $O(n \cdot \log n)$.

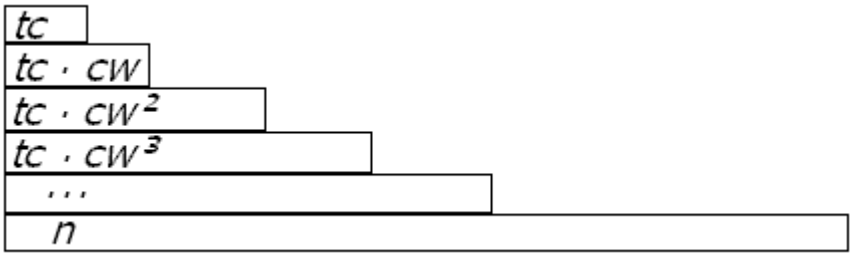


Figure 2.3: Schema of the comb data structure
The comb is constructed by combining several filter intervals. The smallest window in the comb has the length t_c. The window lengths grow with factor cw. The largest window has length n. For each comb window the C, G and GC content is stored during the combing procedure.

The choice of the combwidth cw

The efficiency of the filter step, *i.e.* the number of regions that are excluded over the number of regions that do not meet the three constrains, scales with *cw*. If *cw* is selected such a way that the *comb* contains all values between t_c and n, the filter conditions become the classical *CGI* constraints. Runtime becomes maximal and every retained region is a *sequence-based CGI* according to the applied definition. An increase of *cw* enables to increase the number of windows that are simultaneously analyzed, but as the upper bound for the approximated value is rising soon the conditions are always true. In consequence, the choice of *cw* is a tradeoff between the runtime of the algorithm and the number of non-*CGI* regions that are excluded from the analysis (Figure 2.4).
Figure 2.5 illustrates how filtering divides the genome in subsequences, by excluding the possibility that the regions in-between are members of any *CGI*.
As the execution of the filter algorithm scales with $O(n \cdot \log n)$ it is beneficial to first segment the genome with a large *cw* in smaller segments and than repeat the filtering with a smaller *cw* on each of these segments.
Hereby, the original DNA sequence is fragmented into smaller segments, thus resulting in a considerable reduction of n and creating a divide-and-conquer scenario (Figure 2.5). To enumerate all *CGIs* in such a segment, all possible window sizes between t_c and n have to be applied for combing. For the human chromosome 22, for instance, an initial filtering with *cw=1.2* already reduces the remaining sequence from ~50 Mbp to ~14 Mbp that are distributed over many individual segments. A subsequent filtering of these with *cw=1.05* takes 16 minutes. Thus, the combined runtime of 1.7 h is more than a threefold reduction compared to the 5.5 h CPU time for the analysis of the full chromosome with *cw=1.05*. The computation was performed on a desktop computer with a 3 GHz Duo Core processor and 4 GB main memory.

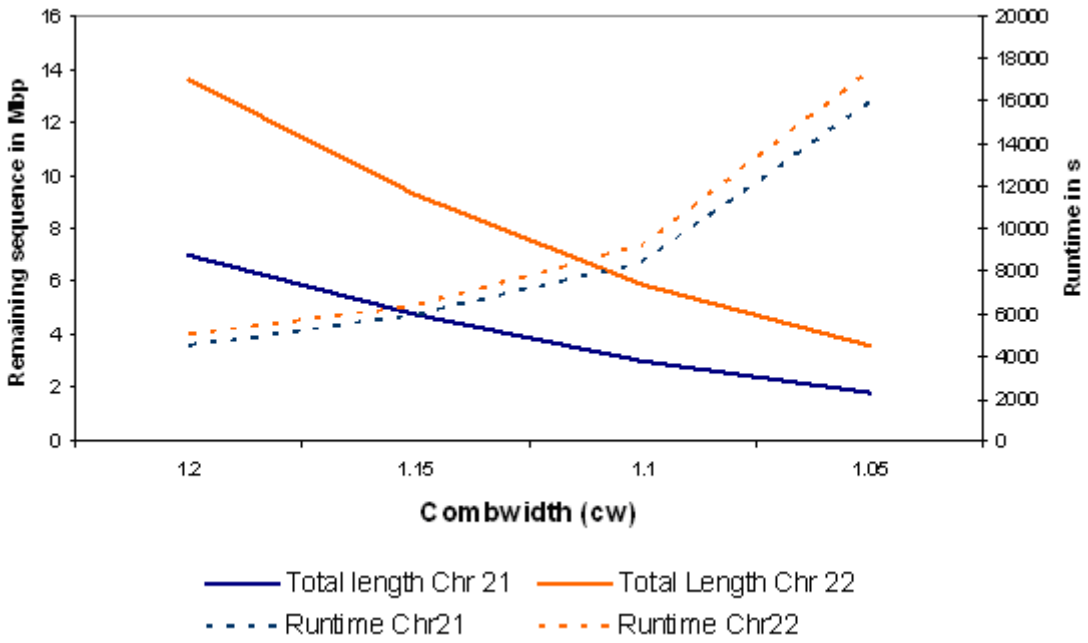


Figure 2.4: Tradeoff between runtime and efficiency of CgiHunter filter step
The filter step was applied with different values of cw for the CpG-poor chromosome 21 (46.9 Mbp) and the CpG-rich chromosome 22 (49.7 Mbp). Both solid lines show the decrease of the total length of those sequence segments that remain after filtering with decreasing values of cw. The dashed lines show the increase of runtime.

Subsequently, the output of filter is analyzed by a *brute force* approach in which all candidate *CGIs* are reported. This solves P1.

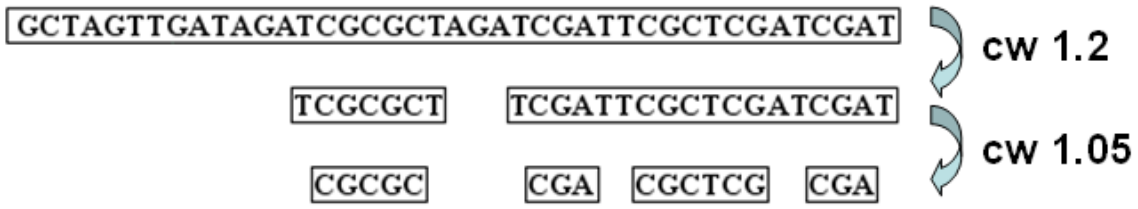


Figure 2.5: Successive filter steps implement a divide-and-conquer strategy
The genomic DNA sequence on top is filtered with combing intervals, which are subject to a larger cw. The resulting sequence segments in the middle are considerably smaller and can be processed independently by a computationally more expensive, but filter step smaller combwidth cw.

Filter algorithm and correctness proof

The function *filter* takes two arguments: The string *sequence* contains the DNA sequence in which *CGIs* should be annotated, and the numeric variable *cw* defines relative width of the comb intervals. For instance t_c=500 and *cw*=1.1 define the sequence 500, 550, 605, 666…, n.

```
Function filter(sequence, cw):
1       n := length(sequence)
2       t_c := minimal length of a CpG island
3       c[0,..,m] := array of comb values initialized recursively as follows:
        c[0] := t_c
        c[a+1] := c[a] * cw
        c[m] := n
4       o_ij'j'' := filtering condition for CpG observed vs. expected ratio (see above)
5       g_ij'j'' := filtering condition for GC content (see above)
6       for k in 0 to n-t_c   (iterate through all potential start positions of a CGI)
7               update C,G and CpG counts for all comb windows
8               for l in 0 to m-1   (iterate through all adjacent pairs of comb values)
9                       if k + c[l+1] < n and o_k(k+c[l])(k+c[l+1]) and g_k(k+c[l])(k+c[l+1])
10                              store region R_k(k+c[l+1]) (Region may contain CGI)
11      merge all stored regions that overlap
12      return the list of merged regions
```

Algorithm 2.1: Pseudocode of filter step of CgiHunter

Lemma 1: The function filter never overlooks a valid *CGI*
Proof: Let R_{ij} be a valid CpG island with i being its start index and j its end index. It follows from the definition of a *CGI* that $t_c \le j-i \le n$ and that $0 \le i \le n-t_c$. Hence it follows from line 3 that a value $p \in \{0,..,m\}$ exists such that $c[p] \le j-i \le c[p+1]$. Because R_{ij} is a *CGI* it follows that the filtering conditions $o_{k(k+c[p])(k+c[p+1])}$ and $g_{k(k+c[p])(k+c[p+1])}$ hold for $k=i$. Therefore, the region $R_{i(i+c[p+1])}$ is stored during the call of the *filter* function and R_{ij} is contained in this region. In consequence, R_{ij} is covered by the returned list of regions. *q.e.d.*

The *filter* function has two loops. The outer loop is running for approximately n iterations. Given the exponential increase of comb-window lengths defined by *cw*, the inner loop runs for $log(n)$ iterations. The total runtime of *filter* is bounded by $O(n\ log(n))$.

After the genome sequence is subdivided in sufficiently small fragments the algorithm can be evoked with a comb that contains all values between t_c and n. This directly returns the *CGI Shadow* as the union of all valid *CGIs*.

2.2.2 A solution to the ambiguity problem

In the last subsection I described an algorithm that reported all regions in a DNA sequence that meet a given *classical CGI definition*. As these frequently overlap with each other the question remains which region should be reported. A simple computation shows that it is not feasible to report every candidate region. For instance a 1000 bp long region that consists of CpG rich repeats would easily results in about a million candidates, as nearly every of its subsequences meets all constraints.

Reporting the *CGI Shadow* instead of individual *CGIs* solves this problem. This is a valid solution, as all nucleotide positions covered by a *CGI Shadow* belong to at least one valid *CGI*. If the applied *sequence-based CGI definition* is the best available representation of a *functional CGI definition*, reporting not the entire *CGI Shadow* would actually discard valid regions. If in contrast selected parameters of the *sequence-based CGI definition* itself are suboptimal, it is advisable to select better parameters rather than applying complicated post-processing steps. Therefore, the *CGI Shadow* solves P2.

Previously, it has been proposed to store every identified CGI separately and then applying a scoring function to select the optimal CGI for representing a cluster of overlapping CGIs (Bock 2008). As the total number of CGIs is only bounded by n^2 and already reaches for frequently occurring CpG-rich regions of length 10 kbp the order of 10^8, this is an I/O intensive step. Additionally, it requires the choice of a scoring function with low complexity, but good performance. Until now no according scoring function has been identified.
In contrast, the here proposed approach has the advantage that overlapping CGIs are directly merged into *CGI Shadows*. As there are at most $n/(t_c+1)$ separate *CGI Shadows* per DNA sequence, for which only start and end position have to be memorized, this information can reside in the computer main memory during computation. Furthermore, the three classical constraints remain the only free parameters that influence the results of the annotation.
To make the results of the *CGI Shadow* approach more transparent, a proof track is generated that separately annotates the longest leading and tailing *CGIs* within each *CGI shadow*.

Box 4: ***Generalization of Filter approach***

Sliding window approaches are applied to solve different problems in genome research. Often the size of the search window is a free parameter that influences the results by a *Single-sliding-window bias*. In specific cases, the hierarchical filter approach of *CgiHunter* is general interest to eliminate this parameter. The algorithm exploits that nucleotide and dinucleotide counts only grow monotonously when the analyzed sequence is elongated. Wherever decision functions are assembled from atomic elements that are based on such nucleotide counts, they have the potential to be reformulated such that that a *comb* can be constructed.

2.2.3 An optimal choice of thresholds

The aim of this subsection is to find a solution for P3. To this end, values for t_a, t_b and t_c have to be selected that, when applied in combination with *CgiHunter,* annotate regions, which comply with the three properties of the BF-CGI definition.
By annotating genome regions that have a higher O/E_{CpG} than the remaining genome, regions are enriched that are protected from *CpG decay* (a). For a set of parameters, the correlation with absence of DNA methylation and marks of open chromatin (b), is determined by comparison with according experimental data. Finally, the influence of a *CGI* on gene transcription (c) can be inferred indirectly from the co-localization with RNA-Polymerase occupancy sites.

By benchmarking a number of CGI annotations against an objective function, the strategy of Takei and Jones is adopted and extended by using a larger set of CGI annotations and applying a wider range of gold standard data as objective function. The details of this benchmark are described in the section 2.3.2.

2.3 A benchmark for genome-wide CpG island annotations

In this section, first *CgiHunter* is compared with a single-sliding-window-based approach for *CGI* annotation. Then, based on biological gold standard data an exhaustive benchmark is applied to characterize different *sequenced-based CGI definition* derived by *CgiHunter* and other state-of-the-art annotation programs.

2.3.1 Methodical improvements of sliding-window-based CGI annotation algorithms

The *CgiHunter* algorithm was implemented, including a XML-based management system for the administration of the parallelized computation and a graphical user interface. As a benchmark, annotations for the *TJ* and *GGF criteria based CGI definition* were computed for the human genome (*hg18*) and the mouse genome (*mm9*). While for the annotation of the *TJ definition* a custom desktop computer with 1 GB hard disk space was sufficient, the *GGF definition* was applied on 20 nodes computer cluster.

Annotation type	**Human genome**	**Mouse genome**
Shadow annotation TJ	32.9 h	17.2 h
Shadow annotation GGF	536.2 h	284.1 h

Table 2.3: Runtime performance of CgiHunter
For both annotation types and CGI definitions the runtime in CPU hours over all parallel processes is summed.

A runtime analysis showed that for the stricter *TJ definition* the annotation was generated in less then two days for the human as well as the mouse genome. The runtime of the *GGF definition*-based annotation was considerably higher (Table 2.3). A detailed study of the annotations produced for the human chromosome 21 showed that the *GGF* annotation covered 3.56 Mbp, which corresponds roughly 10 % of the chromosomes known genome sequence. This 10 times more than expected from functional studies. This indicates that for an exhaustive annotation based on the classical CGI constrains, the annotation parameters of the *GGF definition* are too lenient for whole genome annotations. A direct comparison with the *CpG Island Searcher* software demonstrates the influence of the *Ambiguity problem* and the *Single-sliding-window bias* on the annotation length (Table 2.4).

Definition and software	**#Islands**	**Length**
GGF CpG Island Searcher	5086	1,804,392 bp
GGF CGI Shadow	4845	3,565,341 bp
TJ CpG Island Searcher	447	480,566 bp
TJ CGI Shadow	464	1,046,554 bp

Table 2.4: CGI annotations of human chromosome 21

To estimate the methodical improvement of the *CgiHunter*'s *CGI Shadow* annotation over the single-sliding-window approach as implemented in *CpG Island Searcher*, I designed a self-consistency benchmark. An algorithm for *CGI* annotation is called self-consistent, if a lowering of one or several of the applied thresholds only leads to the annotation of novel genome regions, either by extending existing *CGIs* or adding new ones. This behavior is mathematically expected from the *classical CGI definition*, *i.e.* a non-heuristic algorithm produces a nested set of annotations (*self-consistency*). The deviation from this *self-consistency* indicates the inherent bias of a method.

For the benchmark, I annotated human chromosome 21 and 22 with three *classical CGI definitions*. These only differed in the minimal length threshold (t_c=200, 500 and 1000 respectively), while GC content (t_a = 50%) and O/E_{CpG} (t_b = 0.6) were fixed. Hence, all *CGIs* in the t_c= 1000 annotation fulfill the constraints of the t_c=500 annotation and both of them also fulfill the constraints in the t_c=200 annotation. For *CgiHunter*'s *CGI Shadow* annotation this property holds and no bias was observed. Hence, *CgiHunter* is *self-consistent*. For *CpG Island Searcher*, the ~1.8 Mb long t_c=200 annotation misses at least 16.35 % of valid *CGI* regions according to its own stricter annotations. Of these, 73 kbp and 86 kbp are exclusively annotated by the t_c=500 and t_c=1000 annotations respectively, on top of 135 kbp that are annotated by both.
16.35 % is only a lower bound for the heuristics bias, as additional annotations with different t_c thresholds identify further missed regions. This observation confirmed the combined impact of the *Ambiguity problem* and the *Single-sliding-window bias* on practically applied CGI annotations.

Next, I compared the algorithms performance on the whole genomes of human and mouse, applying the *GGF* and *TJ classical CGI definitions* (Table 2.5). Therefore, I counted the total number of *CGIs* in one annotation that do not overlap with any CGI in the other annotation. 4% and 16% of individual *CGIs* are missed completely by the *CpG Island Searcher*, while the *CgiHunter* missed no *CGI* identified by the other program.
Thereby I conclude that the *CgiHunter* algorithm is on the algorithmically level a significant improvement over the single-sliding-window-based method *CpG Island Searcher*.

Genome	**CGI definition**	**Missed CGIs (abs.)**	**Missed CGIs (rel.)**
Human	GGF	33,754	10 %
Human	TJ	2,190	6 %
Mouse	GGF	28,283	16 %
Mouse	TJ	868	4 %

Table 2.5: Absolute and relative number of CGIs missed by CpG Island Searcher

Impact on comparative studies

To assess if the observed bias influences the practical use of CpG island annotation in genome research, I performed a comparative study on promoter *CGIs* in human and mouse following a previously proposed workflow (Jiang, Han et al. 2007). The research question was: How many orthologous genes switched their promoter type from CGI promoter to non-CGI promoter?

Therefore, I generated two CGI annotations based on the *TJ definition* (t_a=55%, t_b=0.65, t_c=500 bp). The single-sliding-window approach was again represented by *CpG Island Searcher*. Each gene out of the 2918 considered orthologous pairs that had a CGI overlapping with a 2 kbp upstream to 500 bp downstream region around its transcription start site (*TSS*) was labeled to have a *CGI* promoter.

According to the single-sliding-window-based approach, 313 human genes lost their promoter CGI at the orthologous mouse loci, but the *CGI Shadow* annotation only detect 290 losses (7.3 % less). In the opposite case, mouse genes that lost their promoter CGI in human, the differences are 156 to 136 (12.8 % less).
Assuming that a selected *sequence-based CGI definition* is the best available representation of the abstract CGI concept, an unbiased approach clearly improves the systematic study of the posed research question.

Considering that similar studies may be performed in future on datasets of hundreds of vertebrate genomes underlines the importance of a robust annotation algorithm, such as *CgiHunter*.

2.3.2 Biological performance benchmark for CGI annotation software

While the last section demonstrated that *CgiHunter* is a considerable algorithmically improvement within the class of sliding-window-based *CGI* annotations, in this section a broader selection of annotation software for CpG-rich genome regions is evaluated with respect to their correlation with biological features.
This benchmark is based on gene annotations, genome-wide methylation data, chromatin immunoprecipitation sequencing (ChIP-seq) experiments on polymerase binding events and histone modifications.
To determine the influence of all three thresholds on the annotation performance, I generated 100 *CGI Shadow* annotations from all combinations defined by a parameter grid with $t_a \in \{50\%, 55\%, 60\%, 65\%, 70\%\}$, $t_b \in \{0.6, 0.65, 0.7, 0.75, 0.8\}$ and $t_c \in \{500, 600, 700, 800\}$.
As follows from our *functional CGI definition*, an annotation is better the fewer methylated sites overlap with it, the more activating chromatin marks are observed and the higher the measured level of occupancy with RNA polymerase II is.

Annotation of unmethylated genome regions

It is expected that BF-CGIs are preferentially co-located with unmethylated DNA. Two methylome datasets obtained by genome-wide bisulfite conversion experiments were applied (Laurent, Wong et al. 2010). The data was derived from a neonatal fibroblast cell line (*Fibro*) and the human embryonic stem cell line H1 (*H1*). The average methylation level for each annotation was compared the total annotation length (Figure 2.6a). This analysis strongly indicates that *CgiHunter* yields consistently a better trade-off between annotation length and a low average methylation level.

Identification of promoter CGIs

Another application of *CGI* annotations is their ability to mark the 5' end of genes, and thus they support gene identification even in well researched genomes (Illingworth, Gruenewald-Schneider et al. 2010). Therefore, it is beneficial if many CGIs overlap with gene TSSs, while the total number of CGIs remains small, thereby maintaining a small number of false positives. Figure 2.6b shows that the relationship of the number of CGIs per annotation to the number of CGIs overlapping with known protein-coding genes is partitioned into two trends. For low CGI numbers, a decreasing strictness of the annotation leads to a constant increase of discovered genes, were as a rule of thumb of each 2 additional *CGIs* one overlaps a new 5'UTR. Around eleven thousand discovered genes, or 25 k *CGIs*, this ascend reaches a plateau, where presumably most *CGI*-promoter genes are annotated. All non-*CgiHunter* annotations in the benchmark start beyond this point, and thus contain more *CGIs* than necessary for the purpose of gene finding. Furthermore, with few exceptions, the *CgiHunter* annotations identify more *TSS*s, while requiring a smaller number of annotated regions.

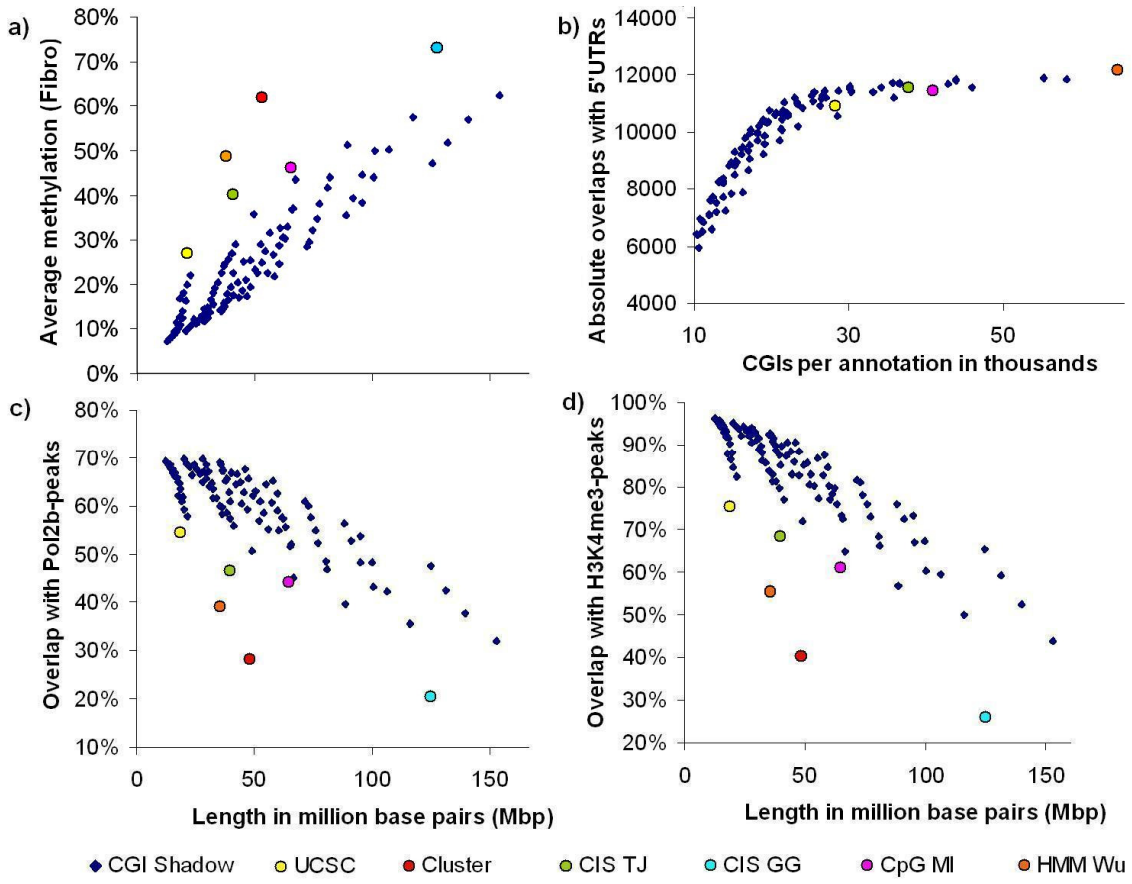


Figure 2.6: Correlation of CGI annotations with genomic and epigenomic features
In the four diagrams, the CGI annotations' total length, respectively, absolute number of annotated regions is compared to (a) their average methylation level in fibroblast cells, (b) the absolute overlap with 5'UTRs of protein coding genes, (c) the overlap with RNA Polymerase II binding peaks and (d) the presence of H3K4me3 peaks in any tissue . The 100 CGI Shadow annotations generated by CgiHunter are shown in blue. The individual annotations generated by other programs are displayed by colored circles: the CpG island track of the UCSC Genome Browser (yellow), the standard CpG Cluster annotation (red), the TJ (light green) and GGF annotations (light blue) computed by the CpG island searcher, the CpG MI annotation (violet) and the hidden markov model approach by Wu et al. (orange).

Co-localization with polymerase binding sites and histone modifications

Epigenetically active CGIs are assumed to be co-located with sites of active transcription and histone modifications that correlate with open chromatin states. Applying ChIP-seq measurements to test the distribution of these modifications around the CGI annotations (Figure 2.6c and 2.6d) confirms that *CgiHunter* annotations show without exception higher enrichment of these marks than the non-*CgiHunter* annotations.

2.3.3 Discussion

CgiHunter is a methodical and practical improvement in comparison to other existing CGI annotation software. The algorithm is sufficiently fast, self-consistent, misses no base pair that belongs to a valid sequence-based CGI and annotates no base pair that dose not. The results of the biological benchmark show that it outperforms all other annotations tools.
This benchmark also demonstrated that of the four tested characteristics, only the overlap with promoter regions clearly indicate a set of thresholds. Remarkably, a number of quite different parameter choices yield very similar results. For instance, (t_a=65, t_b=0.7, t_c=800) and (t_a=65, t_b=0.75, t_c=500) only differ by 110 kbp in total length (Table 2.6 and Appendix A).

Considering that CGIs also overlap other functional elements than promoters of protein coding genes, it is not sufficient to select an optimal set of thresholds only based on this characteristic. Moreover, Table 2.6 and Figure 2.6 also indicate that characteristics such as average DNA methylation vary considerable for annotations of equal total length. Therefore no clear hierarchy between the three thresholds can be established.

GC content t_a	**O/ECpG t_b**	**Min. Length t_c**	**Total length**	**Avg. methylation**
65 %	0.7	800 bp	28.75 Mbp	13.3 %
65 %	0.75	500 bp	28.64 Mbp	14.6 %
60 %	0.8	800 bp	28.38 Mbp	11.7 %
65 %	0.75	600 bp	27.41 Mbp	13.0 %

Table 2.6: Examples for CGI Shadow annotations with good tradeoff between promoter overlap and total length

For the characteristics of open chromatin, transcription initiation and absence of methylation the correlation of annotation size to signal enrichment is more gradual. For instance whole genome methylation levels rise smoothly from very strict annotations to very lenient ones as shown in the surface plot of Figure 2.7.

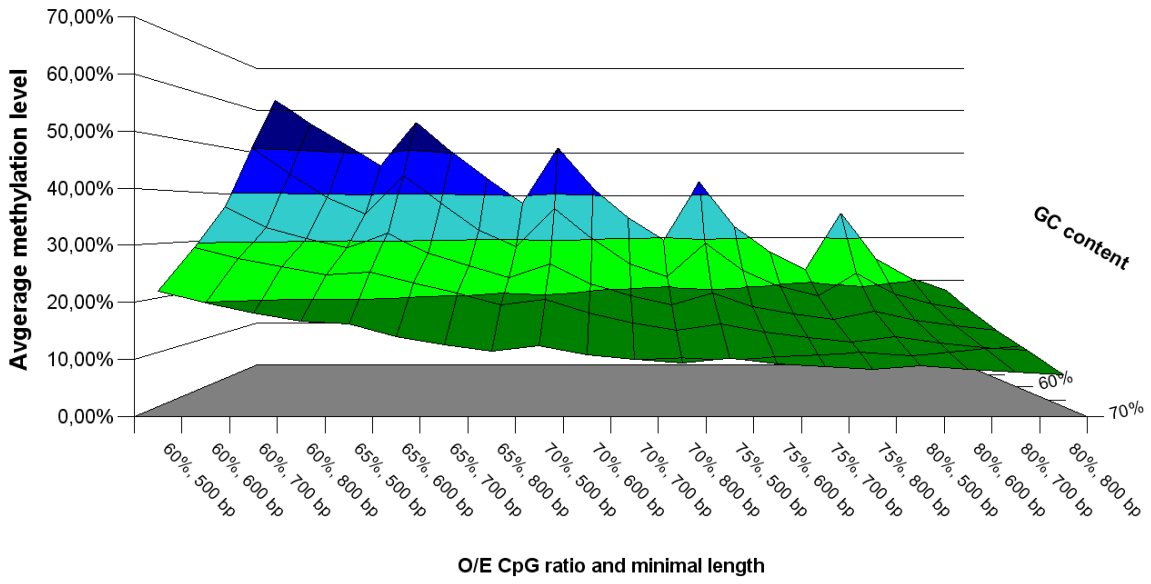


Figure 2.7: Average methylation level drops continuously with increasing strictness
For a whole-genome methylome datasets the average methylation level of CGIs is plotted against the applied CGI definitions of 100 CGI Shadow annotations.

The term CpG island is associated with the idea that a waterline separates two domains. The CpG-rich unmethylated, euchromatic and actively transcribed genome regions are located above this imaginary waterline, while the CpG-poor methylated genome regions are located below. The benchmark with 100 *CGI Shadow* annotations that was performed in this section found no clear evidence for such a separation. A possible explanation for this is that CGIs of different strength and different degree of methylation exist. Furthermore, the genome region around the core CGI, which is refereed to as the CGI shore, shows an enrichment of functional marks above the genomic background, but below those of the core. A second major conclusion is therefore that the binary nature of the CpG island metaphor is misleading, *i.e.* a more gradual classification is closer to the biological truth (Figure 2.8).

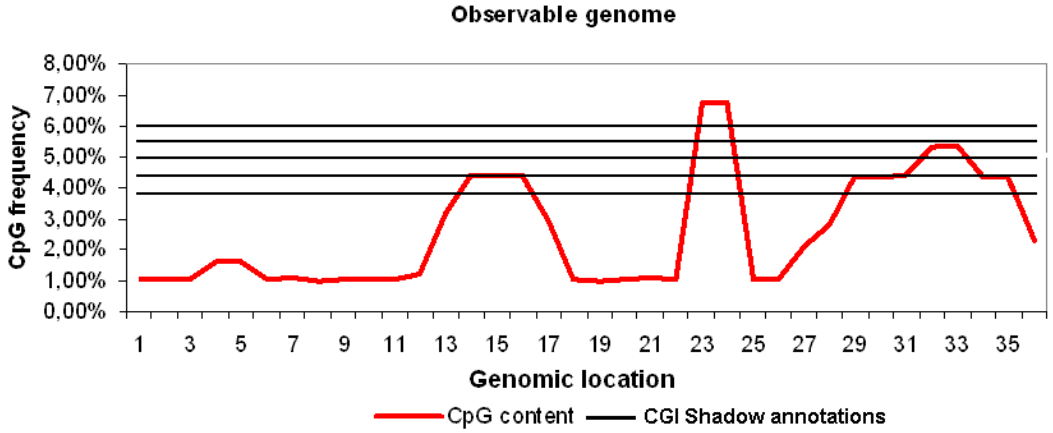


Figure 2.8: CGI characteristic is not binary but continues

2.4 Beyond the binary CGI concept

In the last section the limited usefulness of the binary CpG island metaphor was discussed. As alternative to the old island metaphor, I propose a more gradual analogy in which CpG rich genome regions are treated like mountain tops of different height. As height correlates with absence of methylation, open chromatin and active transcription, the valleys between the peaks and the flatland that represents most of the genome are inactive and methylated. As depicted in Figure 2.8, the height of the tops can be probed by different CGI annotations to produce the equivalent of a topographical map. To generate an annotation that represents this metaphor, the 100 *CGI Shadow* annotations are integrated into the *CGI Mountain* annotation. As each of these annotations is characterized by three parameters it is not possible to directly establish an order. More specifically, it is not clear if an annotation with t_a=60% and t_b=0.7 is stricter than an annotation with t_a=65% and t_b=0.65.

If available, epigenome data can be applied to establish such a hierarchy. To this end, to each of the *CGI Shadow* annotations a *strictness* value that represents the strength of its *CGI* characteristics is attributed. The stricter an annotation is the more it is concentrated on genome regions with enriched marks for activation, and thus applied to establish a hierarchy between the different *CGI* annotations.
To demonstrate the utility of this approach, I applied the median methylation level in fibroblast (*Fibro*) for all CGIs as representative of an annotation's *strictness*. This value was rounded to full percent and all annotations with equal *strictness* were merged. Then the resulting annotations were merged in a procedure that retains for each base pair the highest *strictness* score of all annotations that overlap with it (Figure 2.9). Finally, the score was encoded by a color code that ranged from highest *strictness* in dark green to lowest *strictness* in dark blue. Thus, the *CGI Mountain* annotation is a heatmap representation of a combination of GC content and O/E_{CpG} *i.e.* of *CGI* strength. The annotation was made available for the research community at *http://cgihunter.bioinf.mpi-inf.mpg.de/*.

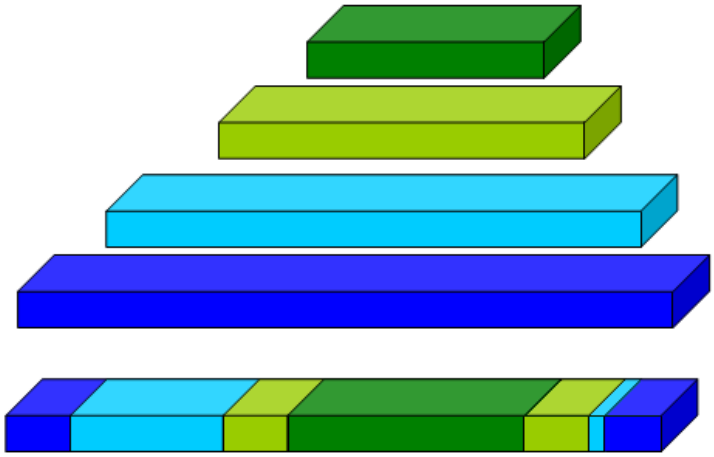


Figure 2.9: Schematic drawing of CpG Mountain annotation
The figure illustrates the construction of CpG Mountain annotation from individual CGI shadow annotations. The colors indicate the strictness of each annotation wherey dark green indicates the highest value. By retaining the highest strictness score in the integrated CGM track, the CpG-rich cores of extended CGIs are highlighted, while also the regions with only slight CpG enrichment are distinguished from the genomic background.

Next, I tested if the tissue-specific methylation data that was applied to rank the *CGI Shadow annotations*, generalizes well to the methylome of other tissues. To this end, the *CGI Mountain* track was intersected with methylation data from a human embryonic stem cell line (*hESC*). The series of box-and-whisker diagrams in Figure 2.10 visualize the achieved correlation of 0.52. Although terminally differentiated fibroblasts and totipotent embryonic stem cells represent different poles in the differentiation spectrum, the annotation well generalizes in this case.

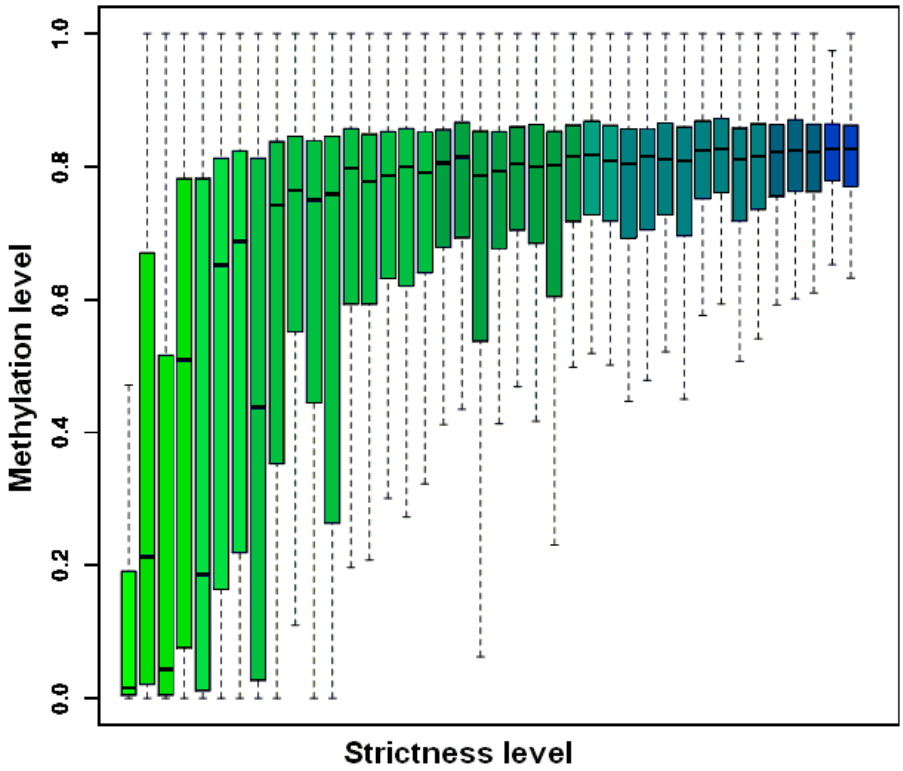


Figure 2.10: CpG Mountain levels correlate with hESC methylation level
For each strictness level of the CpG Mountain annotation a box-and-whisker diagram visualizes the methylation level of all associated regions in the human embryonic stem cell line hESC. Hereby, the bold line denotes the mean methylation, the ends of the colored bars the first, respectively, third quartile, and the end points of the dashed lines the maximal and minimal values. The chosen colors the colors for each strictness level correspond to the according color in genome browser track.

Sequence-based CpG Mountain annotation

For genomes for which no methylome or comparable epigenetic data is available, the total length of the annotations can be applied as proxy for the related strictness level. As visualized in Figure 2.6a, the thus introduced bias is limited by the good correlation between annotation length and methylation level in case of the human genome. To test this hypothesis, I generated a *CGI Mountain* annotation in which the *strictness* score was based on annotation length. Hereby, the choice of the total number of *strictness* levels critically influenced the correlation with the methylome data. Of the three tested parameters 100, 50 and 33 the latter produced results that were comparable to those derived by the methylation-based strictness order (Appendix B).

2.4.1 CpG island shores and CGIs of intermediate strength

The *CpG Mountain annotation* captures the continuous nature of the CGI characteristic. Previously computational epigenetic study already introduced the idea of CGI strength as byproduct of the statistical classifier that were applied to differentiate methylated from unmethylated CGIs (Bock, Walter et al. 2007). This concept of CGI strength represents an alternative to a binary classification, but also raises the questions if it is a biologically justified. Regarding *CpG decay* as strong contributor to *CGI* formation, there are three factors that influence the formation of CGIs with intermediate CpG content:

First, these *CGIs* of intermediate strength are in a transient state (formation or erosion), in which they constantly lose or gain CpGs over a longer timescale.

Second, *CGIs* fluctuate around an individual equilibrium of the CpG distribution that is distinct from that of the stronger *CGIs*. Differential methylation within the germline cycle is one explanation for such a diverging equilibrium (Figure 1.1). Another explaination is that DNA methylation is a stochastic process with variable mean and variance for CGIs of different strength. Furthermore, the distance to the center of the CGI may influence the probability that methylation takes place. These difference then translate into altered mutation frequencies.

Third, there may be a variable degree of selective pressure for or against functional CpG containing motifs, thus raising or decreasing the CpG density in particular regions over time independently from the germline methylation state.

A recent study indeed found evidence that CGIs can be classified into distinct groups by the rate with which they gain, lose and conserve individual CpG dinucleotides (Cohen, Kenigsberg et al. 2011). This supports the first factor as it shows that CGIs are dynamic entities that grow or shrink, and get as well strengthened or weakened in their CGI characteristics by evolutionary forces.

The same study presents evidence against the third hypothesis by observing that individual CpGs in primate promoter CGIs are only under weak selective pressure (Cohen, Kenigsberg et al. 2011).

The hypothesis of differential methylation within the germline cycle, is in line with the observation that CpG island shores show high variability in their methylation levels between tissues (Molaro, Hodges et al. 2011) and within cancer (Irizarry, Ladd-Acosta et al. 2009).

To gain a better understanding of this question weak CGIs are compared with the shores or edges of strong *CGIs* to investigate if both show comparable characteristics. To this end I applied the *EpiExplorer* tool developed by Konstantin Halachev and Hannah Bast (Halachev, Bast et al. 2012).

The neighborhood plots produced by this software were applied to measure the prevalence with which epigenetic marks are observed in the neighborhood of *CGI Shadow annotations* of different strictness. As displayed in Figure 2.11, the enrichment of these epigenetic regulation marks correlates in the direct local neighborhood with the strictness of the CGI annotation. For the stricter annotations, the enrichment levels drop between 1 kbp and 3 kbp to the maximal value of the lenient annotations, before it approaches the genomic average. This observation indicates that on average enrichment of epigenetic modifications in the neighborhood of strong CGIs is similar to the average enrichment in weaker CGIs.

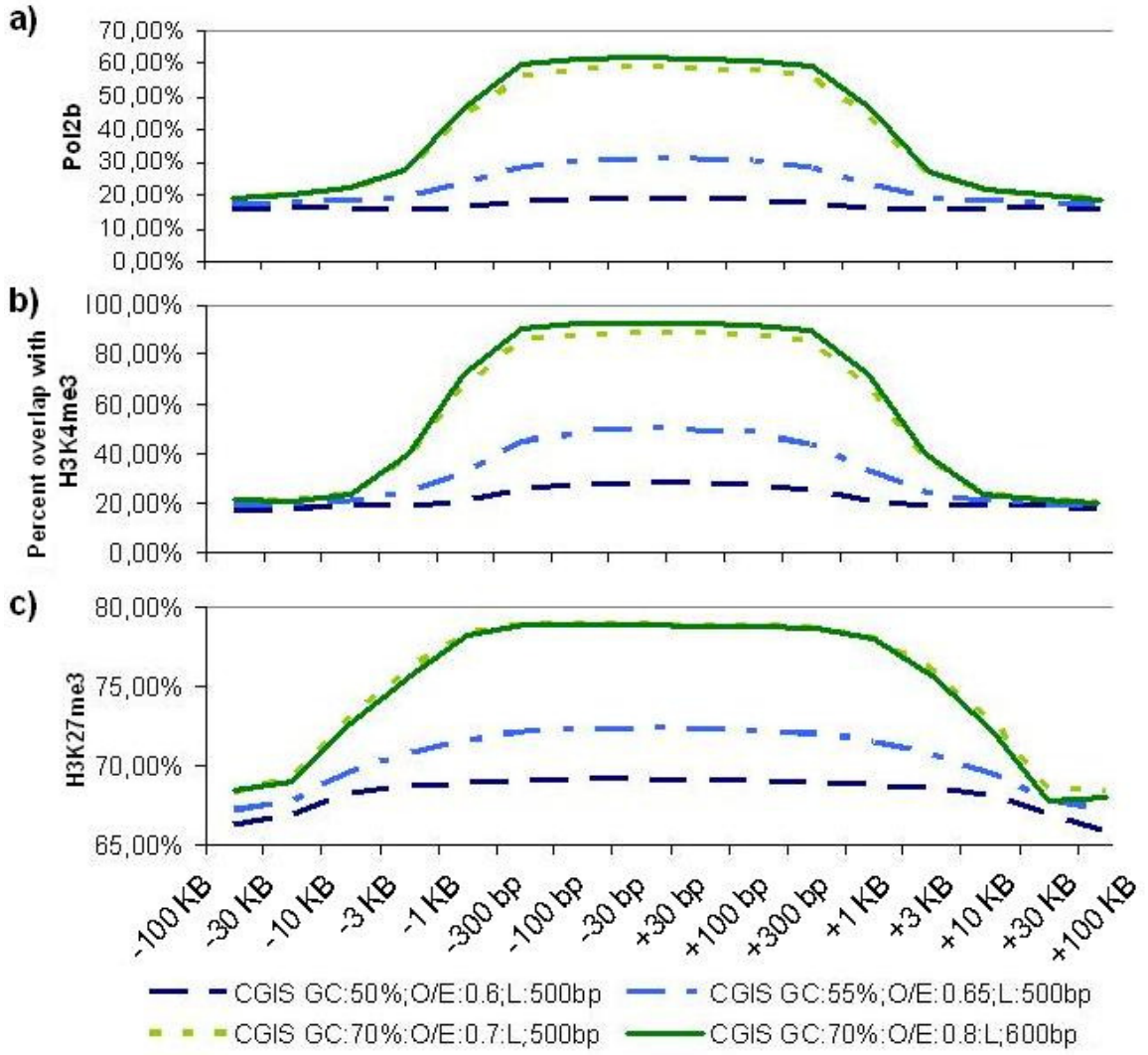


Figure 2.11: Histone modification and polymerase binding close to CGIs
For four CGI Shadow annotations with varying strictness the presence of RNA polymerase II binding events (Pol2b) (a) and H3K4me3 (b) and H3K27me3 (c) histone modification peaks relative to their location are displayed. For all three datasets the presence of the modification has a higher local correlation with the stricter annotations, although the effect strength varies between differences of 10% and 80%.

2.5 Discussion

This chapter discussed the limitations of previous establish software for DNA sequence-based *CGI* annotations. For clarification of the annotation approaches objective function, the concept of a biological functional CpG island was introduced. To enable the annotation of these BF-CGIs the *CgiHunter* algorithm was proposed. Then, the theoretical as well as practical advantages of the algorithm were demonstrated in a comprehensive benchmark. Furthermore, the methodical search for a precise *CGI* definition showed that the binary concept of the CpG island itself is questionable. Therefore, the *CpG Mountain annotation* was designed, as a more continuously concept for the annotation of CpG-rich genome regions. For the human genome a version of it was constructed, to optimally reflect methylome data from human fibroblasts. This correlation generalizes well to methylome data from embryonic stem cells.
Finding a smooth correlation of CpG island characteristic with several markers of open chromatin raises the question on how CGIs of intermediate strength are formed and maintained. To explore the three factors that may contribute to this process in more detail, the next chapter analysis the dynamics with which genome regions converge towards individual dinucleotide equilibrium distributions according to the rate of *CpG decay* they are exposed to.

Chapter 3 - The influence of DNA methylation on DNA sequence composition - A quantitative model of methylation-constrained genome evolution

Searching for CGIs with *sequence-based CGI definitions*, as conducted in the last chapter, is an empirical approach to identify putative targets of epigenetic regulation. The exact choice of the applied thresholds has no intrinsic justification. In this chapter, the evolutionary forces, which lead to the creation CGIs, are quantified directly. To this end, I construct a mathematical model of genome evolution that is aware of the *CpG decay* process. This model is then applied to pursue three research questions:

First, what is the long term impact of DNA methylation and the *CpG decay* effect on neutrally evolving genomes? More precisely, biochemical and molecular biological processes, such as spontaneous deamination and biased DNA repair, apply pressure on the DNA sequence composition. My means of a simulation, I characterize the different equilibria these processes approach in methylated and unmethylated DNA.

Second, CpG-rich *cis*-regulatory elements in unmethylated CGIs benefit from the protection against *CpG decay* by absence of methylation. What can be learned from quantifying this benefit?

Third, has this quantitative model the potential to substitute the empirical CGI annotations or the *CpG Mountatin annotation*?

3.1 Mathematical models of genome evolution

For studying the influence of methylation-mediated *CpG decay* on the genome composition, substitution processes, which contribute either to the degradation or formation of CpGs, have to be considered as well. For instance, a C to G transversion, for instance, can either create novel CpGs in case of CpC/GpG to CpG/CpG substitutions or degrade them in the reverse process CpG/CpG to CpC/GpG. To this end, a mathematical model of genome evolution is required to capture the balance between CpG degradation and CpG creation.

To construct a model of genome evolution that comprises the influence of DNA methylation a brief overview of the available methods is of assistance, as point substitutions in genome sequences can be described at different degrees of detail. For instance, the Jukes-Cantor model postulates that every nucleotide has equal probability to be replaced by any other (Jukes and Cantor 1969). The slightly more detailed Kimura-2-parameter model assumes individual rates for transitions (pyrimidine to pyrimidine and purin to purin substitutions) and transversions (pyrimidine to purin and *vice versa* substitutions) (Kimura 1980). A detailed strand-specific model requires 12 parameters to represent all possible substitutions.
A strand-unspecific model requires only six parameters, as a substitution on one strand is always accompanied by a mirror-substitution on the other strand, *i.e.* an A to G substitution is a T to C substitutions on the complementary strand a. In case of neutral evolution, substitutions appear with equal likelihood on either of the two strands, such that the respective rates for the mirror-events can be expected to be equal (Lobry and Lobry 1999). Such a model comprises four transversion rates and two transition rates. In some species the transversion rates are rather similar and can be represented by an average transversion rate to reduce the number of parameters from six to three (Arndt, Burge et al. 2002; Peifer, Karro et al. 2008).
Additionally, the influence of neighboring nucleotides on the point substitution rate of an individual nucleotide can be considered. This adds one free parameter to the model for each presumed interaction. In the human genome, *CpG decay*, as substitution of CpG to TpG and its mirror-substitution CpG to CpA, is the only significant neighbor-dependent substitution processes in neutrally evolving DNA (Arndt, Burge et al. 2003; Lunter and Hein 2004; Siepel and Haussler 2004; Hobolth 2008; Peifer, Karro et al. 2008). Including this process as well, leads to a four substitution rate model comprising the average transversion rate r_1-r_4 = r_{tr}, two transitions - A/T to G/C r_5 and G/C to T/A r_6 - and the *CpG decay* r_7. As the rates are relative to each other the fourth rate is defined by the remaining three, and thus, the model has three degrees of freedom.
For the human and the mouse genome estimations for these three parameters were derived previously (Peifer, Karro et al. 2008). To this end alignments of 38 DNA repeat families were applied, facilitating that their ancestral sequences can be well reconstructed from their numerous copies in present genomes (Jurka 1994) . Moreover, the majority of all repeats evolves free of selective pressure, thus reducing the bias in the rate estimates. To simplify the interpretation of the model, the rates are scaled such that one unit of time corresponds to the interval in which for a single nucleotide one transversion of each type

is expected. As each nucleotide can undergo two different transversions, one unit is equivalent to two transversions per site. This results for the human in $r_5/r_{tr} = 3.02$, $r_6/r_{tr} = 5.08$ and $r_7/r_{tr} = 48.3$, and indicates the dominance of the *CpG decay* rate over the other substitution processes.

In the remaining chapter, I examine the hypothesis that methylated genome regions evolve under the influence of all four substitution processes, while unmethylated genome regions are only affected by the context independent processes, and thus, are immune against *CpG decay*.
Formally, these models are described as the stochastic processes f_M and f_U. These take a genome region R and an age t as input and map them to an accordingly mutated region R^t.

3.2 Simulated evolution

A direct approach to characterize the difference between methylated and unmethylated genome evolution is the simulation. Thereby, I implemented a simulation engine for the stochastic processes f_M and f_U.
The resulting software iterates recursively over two steps. Based on a given DNA sequence and the corresponding substitution rates, it computes the waiting time until the next substitution event. Then, the exact nature of the substitution is determined and the DNA sequence is updated. The process starts at t=0 and terminates when a given time limit t_l is reached.

The waiting time t_w for an individual substitution is described by an exponential decay process with the respective rate. The joint waiting time for multiple substitution processes is derived from an exponential decay process with the summed rate of all individual processes ((Karlin and Taylor 1975) p.133). To this end, the software counts the number of nucleotides and CpG dinucleotides in the sequence, multiplies them with the respective rates and sums the products to yield the joint substitution rate r_j .Then, t_w is drawn from an exponential distribution with mean r_j^{-1}.

Next, t is advanced to $t + t_w$. If the termination condition $t > t_l$ is reached, the simulated mutation occurred after the observed time interval. Hence, the current DNA sequence is reported without an additional change.
Otherwise, the interval [0,1] is segmented into regions that represent the possible substitution events by the size of their respective contributions to r_j. Then, by drawing a uniformly distributed random number, the nucleotide that is to be substituted is determined and replaced. Finally, r_j is updated according to the changes in the nucleotide and dinucleotide distribution to prepare the next cycle. The simulation engine was implemented in the scripting language Python Version 2.4.

3.3 Equilibrium distributions of neutrally evolving genomes

Evolutionary processes acting on DNA sequences in which the point substitution rates are unequally distributed, can lead to unbalanced distribution of the nucleotides, dinucleotides and oligo-nucleotides. In the human genome, some of these patterns are rather transient (in methylated DNA for instance the CpG dinucleotides), while others are relatively stable (for instance the thymine and the adenosine nucleotides).
Over longer time spans the stable states will be enriched and the transient states will be depleted, but because transitions between all patterns are possible, none will be indefinitely lost (Sved and Bird 1990). Moreover, in larger DNA sequences a dynamic equilibrium is approached.
The simulation engine described in the previous subsection is an appropriate tool to asses whether such equilibria are formed under the methylated or the unmethylated model, how fast they are reached and also to quantify their difference.

In formal terms following computation is made: Let g be a function that derives the dinucleotide distribution G^t from R^t. Here I approximate the limit $G^\infty := \lim_{t\to\infty} g(f(R,t))$ that is expected to be similar for sequences of infinite length ($n \to \infty$).

3.3.1 Numerical derivation of sequence equilibrium distribution

To approximate the equilibrium distributions, I applied the previously described simulation engine 100 times to artificial DNA sequences with uniform base distribution of length 10 kbp. Applying r_5/r_{tr} = 3.02, r_6/r_{tr} = 5.08 and r_7/r_{tr} = 48.3 (100% methylated equilibrium), the simulations were performed for both models for 100 time units, *i.e.* until each site underwent on average 200 transversions. As it is unclear if the *CpG decay* rate that was derived from the alignments of transposable elements generalizes to the whole genome, the equilibrium for r_7/r_{tr} = 24.15 was additionally computed (50% methylated equilibrium). For comparison, I determined the nucleotide and dinucleotide distribution within the human genome (assembly hg18) by counting their frequency in the genome sequence. The results are summarized in Table 3.1. The frequency of CpGs at the 100% methylated equilibrium (0.55 %) is about half of the frequency observed in the whole genome (0.99 %) and those of the methylated equilibrium 50% (0.98 %). This either indicates that the human genome is not in equilibrium or that indeed the rate estimates derived from the repetitive sequences are elevated compared to the remaining genome. For the analysis in this chapter, a conservative estimation of the *CpG decay* rate appears more reasonable, as it gives a lower bound on the differences between methylated and unmethylated genome regions. Therefore, I apply the 50% methylated equilibrium determined by r_7/r_{tr} = 24.15 for the subsequent analysis.

Dinucleotide and Nucleotide	Unmethylated equilibrium	50% methylated equilibrium	100% methylated equilibrium	Human Genome (hg18)
AA	9.05 %	9.97 %	10.06 %	9.77 %
AC	5.98 %	5.70 %	5.63 %	5.03 %
AG	5.99 %	6.26 %	6.29 %	6.99 %
AT	9.06 %	10.51 %	10.77 %	7.72 %
CA	5.99 %	6.97 %	7.13 %	7.25 %
CC	3.96 %	3.35 %	3.27 %	5.21 %
CG	**3.96 %**	**0.98 %**	**0.55 %**	**0.99 %**
CT	5.99 %	6.26 %	6.28 %	7.00 %
GA	5.99 %	5.43 %	5.32 %	5.93 %
GC	3.96 %	3.08 %	2.98 %	4.27 %
GG	3.97 %	3.36 %	3.3 %	5.22 %
GT	5.99 %	5.69 %	5.65 %	5.05 %
TA	9.05 %	10.09 %	10.24 %	6.56 %
TC	5.99 %	5.43 %	5.35 %	5.94 %
TG	5.99 %	6.95 %	7.11 %	7.27 %
TT	9.07 %	9.96 %	10.06 %	9.80 %
A	30.08 %	32.35 %	32.75 %	29.52 %
C	19.94 %	17.58 %	17.24 %	20.45 %
G	19.88 %	17.54 %	17.25 %	20.47 %
T	30.10 %	32.53 %	32.76 %	29.56 %

Table 3.1: Dinucleotide and nucleotide frequencies at different epigenetic equilibria

For the applied methylated equilibrium and the unmethylated equilibrium the standard deviation over all runs lay below $2 \cdot 10^{-4}$ for all dinucleotides. This indicates that indeed the sequence distribution fluctuates around a dynamic equilibrium (Table 3.2).
CpGs show the largest differences between the methylated and the unmethylated equilibrium. In absolute percentage CpGs appear 2.98% more often in unmethylated than in methylated genome regions. On closer examination it can be observed that the dinucleotide ApT, with an underrepresentation of 1.46%, alone is already half as informative as CpG. Aggregation over the four dinucleotides CpA, TpG, ApT and TpA leads to a joint frequency difference of 4.43%. These results indicate that in genomes, which reached their equilibrium state, the search for islands with decreased frequency of these four dinucleotides potentially detect signatures of methylation-free genome evolution more efficiently then CpG centered approaches.

Dinuc.	Unmeth.	Unmeth. std.	Meth.	Meth. std.	Difference
AA	9.053253	+- 0.001352	9.970225	+- 0.006837	-0.916972
AC	5.982184	+- 0.010744	5.698340	+- 0.007465	0.283843
AG	5.986750	+- 0.002798	6.263422	+- 0.001555	-0.276672
AT	9.056424	+- 0.010879	10.519951	+- 0.009351	-1.463527
CA	5.987135	+- 0.003461	6.967452	+- 0.003334	-0.980316
CC	3.959960	+- 0.005844	3.354588	+- 0.000418	0.605373
CG	3.957400	+- 0.005954	0.982524	+- 0.002780	2.974876
CT	5.994753	+- 0.014380	6.259831	+- 0.003250	-0.265078
GA	5.985175	+- 0.008126	5.428309	+- 0.001141	0.556865
GC	3.963677	+- 0.011462	3.078965	+- 0.009246	0.884712
GG	3.974045	+- 0.004760	3.355838	+- 0.001723	0.618206
GT	5.988586	+- 0.016849	5.691327	+- 0.001297	0.297258
TA	9.052783	+- 0.007606	10.085932	+- 0.017771	-1.033150
TC	5.993492	+- 0.005108	5.432881	+- 0.007571	0.560611
TG	5.993697	+- 0.000406	6.952575	+- 0.005070	-0.958878
TT	9.070688	+- 0.004074	9.957840	+- 0.010636	-0.887151

Table 3.2: Equilibrium frequencies of dinucleotides under unmethylated and methylated constraints
This table shows for each dinucleotide the computed mean relative frequency at the equilibrium and the corresponding standard deviation. Furthermore, the difference between both equilibria is displayed. The data is averaged over the 100 simulation runs.

The CpG frequency approaches its equilibrium very rapidly (Figure 1.4). After 0.3 expected transversion per site the distance to the equilibrium is below one per mille. Moreover, the spread between the unmethylated and the methylated evolution is growing fast. This indicates that after the onset of global DNA methylation the epigenetic footprints of germline methylation manifested rapidly. Furthermore, this implies a strong pressure on functional DNA sequences that contain CpGs.

3.4 Evaluating substitution pressure on sequence motifs

The equilibrium distributions, derived in the previous section, represent the two endpoints towards which DNA sequence evolves in methylated and unmethylated genome regions. Within the model there are three explanations for sequences that diverge from the equilibria defined by their germline methylation state. These sequences either occur rarely, are under selective pressure, or have changed their methylation state recently and not yet converged to the new equilibrium. The explanation that a certain sequence has not yet converged to an equilibrium will be discussed in detail in chapter 4 and 5. The balance between the first and the second explanation will be examined in detail in this section.
More specifically, the likelihood that a certain sequence appears under both equilibria by chance is computed. This also quantifies the selective pressure that is required to maintain it under each methylation state. If the likelihoods are unbalanced, a preference for one of the two states can be inferred. Furthermore, if it occurs more frequently than expected in one of the two methylation states, it is subjected to selective pressure.

To conduct this estimation and number of likelihood model are required. These are then calibrated with the equilibrium distributions computed in the last section. Then they are applied in three different scenarios: to evaluate the occurrences of simple DNA sequences, of collections of DNA sequences, and of binding motifs of transcription factors.

3.4.1 Likelihood of DNA sequences

A dinucleotide inherently contains information on the spatial relationship of single nucleotides *i.e.* in case of ApT the information that A is followed by T. Thus, in case of a dinucleotide frequency distributions D, with D_{XY} denoting the frequency of XpY, this spatial information can be easily transformed into a transition matrix P^D. The transition probability from X to Y is computed by:

$$P^D_{X,Y} = D_{XY} / \sum_{i \in \Sigma_A} D_{Xi} ,$$

whereby Σ_A denotes the nucleotide alphabet. P^D defines a first-order Markov chain, which can directly be applied to computing sequence likelihoods.

Simple DNA sequences – k-mers

A DNA sequence (or oligo-nucleotide) is a string of length n over the alphabet Σ_A, which is indexed like the genome region R in section 2.2.1, *i.e.* R_i denotes the i-th nucleotide in the genome region. The probability of $Pr(R|D,S)$ is computed by $S_{R0} \cdot \prod_{i=0}^{n-1} P^D_{Ri,Ri+1}$, where the prior probability of the first nucleotide S_{R0} is given by the nucleotide distribution S.

Collections of DNA sequences - cis-regulatory sequences

By appropriate experiments samples of DNA sequences that are bound by specific transcription factors can be determined. In this way the binding properties of transcription factors (TFs) are characterized by a selection of *cis*-regulatory sequences $[R^1,\ldots,R^m]$ and their corresponding relative observation frequencies $[f_O^1, \ldots, f_O^m]$.
To quantify if a particular evolutionary pressure favors their formation and maintenance, each of these sequences is treated as a simple DNA sequence and the derived probabilities are aggregated as weighted sum over the observation frequencies:

$$\Pr(TF \mid D,S) = \sum_{i=1}^{m} f_O^m P(R^m \mid D,S)\,.$$

The derived likelihoods are small and depend on the lengths of the binding sequences. To enable the interpretation of these values, I compare them between different methylation constraints. For the dinucleotide distributions U (unmethylated) and M (methylated), with nucleotide priors S_U and S_M, respectively, I define the likelihood ratio of TF as:

$$L(TF) = \frac{\Pr(TF \mid U, S_U)}{\Pr(TF \mid M, S_M)}\,.$$

This odds ratio is larger than 1 if the corresponding *cis*-regulatory sequences occur more frequently in sequences that evolve unconstrained by DNA methylation, *i.e.* in unmethylated genome regions.

Position-specific scoring matrix

Transcription factor binding sites or similar sequence patterns are often not directly described by individual sequences, but in summarized form. In this approach, the relative nucleotide frequency at each position of the sequence is counted to compute a position-specific scoring matrix (PSSM) that encodes a sequence motif. Such PSSMs or position weight matrices (PWMs) are an established notation and reviewed in (Hannenhalli 2008). It is noteworthy, that by summarizing the exact binding sequences, information is lost. For instance, if R_1=AC, with f_O^1=0.5, and R_2=GT with f_O^2=0.5, the respective PSSM $\wp_E$ takes the form visualized in Table 3.3.

$\wp_E$	Position 1	Position 2
A	0.5	0
C	0	0.5
G	0.5	0
T	0	0.5

Table 3.3: PSSM representation loses information on dinucleotide content
Although in the original sequences only two types of dinucleotides are observed, the corresponding profile can generate four different dinucleotides with equal probability.

Also, information on interdependencies of more distant positions, the so-called phase information, is lost (Stormo 2000; Bulyk, Johnson et al. 2002). Many TFBS are only reported in PSSM format. This has the advantage that they can be visualized as Sequence Logos, which represents the information content of each position (Schneider and Stephens 1990). Furthermore, they can be efficiently applied to *in silico* screening of genome databases via profile alignments (Vlieghe, Sandelin et al. 2006). Although, PSSMs are less informative for the analysis of context-dependent substitution processes than collections of *cis*-regulatory sequences, they were evaluated where no access to the primary experimental data was obtainable.

To determine the consequences for the research question, the likelihood of a PSSM under different evolutionary background models is computed and compared to the model based on the full sequence information. To this end, a variant of the forward algorithm is applied(Durbin, Eddy et al. 1998), to enable the analysis of these PSSMs with regard to methylation-induced evolutionary pressure.
The matrix *Z*, which contains the intermediate results of this dynamic programming approach, has the same size as the PSSM $\wp$. The entries in the first column are initialized as :

$$Z(Y,1 \mid D,\pi) = \wp(Y,1) \sum_{X \in \Sigma_A} \pi_X P^D_{X,Y} \text{ , with } X \in \Sigma_A .$$

The remaining matrix is computed recursively by:

$$Z(Y,i \mid D,\pi) = \wp(Y,i) \sum_{X \in \Sigma_A} Z(X,i-1 \mid D,\pi) P^D_{X,Y} .$$

The sum over the last column of *Z* then yields the likelihood of the motif under the given constrains $\Pr(\wp \mid D,\pi)$. Again, a score based on an odds ratio enables the interpretation of these likelihoods:

$$L(\wp) = \frac{\Pr(\wp \mid U,\pi_U)}{\Pr(\wp \mid M,\pi_M)} .$$

3.4.2 Evaluating selective pressure on DNA composition

Molecular biological studies and computational epigenetic approaches discovered several DNA sequence patterns that are either enriched in methylation-resistant or methylation-prone genome loci. This association can have different reasons. First, these patterns are functionally involved in maintaining the methylated or unmethylated state of a region (*active association*). Second, these patterns are a consequence of the presence or absence of the *CpG decay* effect (*passive association*). Third, the patterns are co-located with functional elements that are preferentially methylated or unmethylated (*indirect association*). Fourth, the function of these patterns is regulated by DNA methylation (*regulatory association*).

For instance, a binding site for a protein complex that actively recruits a DNA methyltransferase is an example for an *active association*. Without this binding-site the methylation state of the locus would change.
The CpG dinucleotide itself is an example for a *passive association* as are all oligo-nucleotide patterns with a high O/E_{CpG}. They are depleted over time from methylated genome regions via *CpG decay*, but are protected from this effect in unmethylated domains.
Binding sites of general transcription factors are enriched in functional promoters, which in turn often co-localize with CGIs. Thus, they are candidates for *indirect association* to the unmethylated state of these promoters.
Binding sites of tissue- or developmental stage-specific proteins or elements that are regulated by imprinting are activated (*CTCF-binding sites*) or deactivated (*STAT* binding sites) by DNA methylation. Therefore, they have to be located in genome regions that can be methylated and unmethylated depending on the epigenetic state of the cell. Furthermore, these sites have to be protected from *CpG decay* in the germline. Such sites are examples for *regulatory association*.

Estimating to which degree each of these association types are causal for an observed correlation of, for instance, a binding motif $\wp$ to a methylation state, is not trivial.
I approach this problem by comparing the likelihood of $\wp$ in DNA sequences that obey the equilibrium distributions M or U. The resulting odds-ratio indicates if a pattern is favored under a certain methylation regime or has to resist additional mutational pressure to stay conserved. This quantifies the amount of *passive association* for the particular motif. Furthermore, I study the distribution of binding motifs within and outside of CGIs to infer if they are enriched or depleted in CGIs. Motifs that are neutral with respect to the methylation state, but are anyway enriched in CGIs are candidates for *indirect association*.

In the following subsections, the location of TFs with respect to CGIs is assessed. Then the *passive association* of their PSSMs is quantified by the above introduced technique. Subsequently, the introduced methodology is applied for a novel interpretation of several previously performed studies.

3.4.3 CpG decay induced pressure on TFBS

First, the location of transcription factor binding sites (TFBS) that are conserved in the human, mouse and rat genome is analyzed ('TFBS Conserved' track of the UCSC Genome Browser based on Transfac Matrix Database version 7.0). Following the comparative genomics assumption that conservation correlates to function, this approach enables us to focus on putative functional TFBSs. The applied genome annotation reports all motifs that exceed the background probability of observing the reported motif by 2.33 standard deviations. This is equivalent to a significance level of 0.01. Hereby, our null hypothesis on the distribution of conserved TFBSs in the human genome is a uniform distribution. The frequency with which these TFBS sites overlap with CGIs, characterizes their empirical distribution within the human genome. For each type of TFBS the ratio of instances that overlap the *CGI Mountain annotation* (compare chapter 2) over the total number of observed TFBS (*CGI fraction*) indicates the prevalence of the underlying motif in CGIs. The full *CGI Mountain annotation* covers 156 Mbp, *i.e.* 5 % of the human genome. Any significant divergence of the *CGI fraction* from this value leads to the rejection of the null hypothesis. Significance is tested by a Chi-Square test through comparison of the observed and expected counts of TFBS in the *CGI fraction*. Moreover, the ration *CGI fraction* over 0.05 is an indicative value for the strength of this divergence (*CGI Overrepresentation*). Finally, the mean *CGI Mountain* strictness level for the TFBS in the *CGI fraction* yields insights into the preferred location of TFBS within the *CGIs*. A high mean correlates to locations at the cores of CGIs, while a low mean documents locations towards the edges *i.e.* shores of the CGIs. Table 3.4 shows the 20 TFBS with the highest *CGI Overrepresentation*. A full overview of all 246 TFBS can be found in Appendix C.

TFBS ID	In CGIs	All TFBS	CGI fraction	Mean CGM strict.	CGI Overrep.	p-value
SP1_Q6	837	941	88.95	91.70	17.79	0.00E-99
SP1_01	683	769	88.82	91.24	17.764	0.00E-99
AP2_Q6	623	709	87.87	92.01	17.574	0.00E-99
NFY_01	97	128	75.78	90.02	15.156	0.00E-99
PAX4_01	441	608	72.53	91.24	14.506	0.00E-99
PAX5_01	406	576	70.49	91.06	14.098	0.00E-99
CETS1P54_01	238	339	70.21	89.37	14.042	0.00E-99
EGR3_01	999	1509	66.2	90.65	13.24	0.00E-99
NRF2_01	455	717	63.46	89.72	12.692	0.00E-99
ELK1_02	386	617	62.56	90.02	12.512	0.00E-99
USF_C	191	306	62.42	90.05	12.484	0.00E-99
E2F_03	149	240	62.08	90.28	12.416	0.00E-99
MAZR_01	526	859	61.23	89.92	12.246	0.00E-99
E2F_02	538	895	60.11	89.02	12.022	0.00E-99
NMYC_01	702	1196	58.7	89.36	11.74	0.00E-99
EGR1_01	953	1693	56.29	90.48	11.258	0.00E-99
PAX5_02	352	645	54.57	90.42	10.914	0.00E-99
GATA2_01	76	145	52.41	89.21	10.482	0.00E-99
NGFIC_01	990	1986	49.85	89.93	9.97	0.00E-99
CREB_02	219	442	49.55	89.57	9.91	0.00E-99

Table 3.4: Top-20 TFBS overrepresented in CGIs

The majority of the conserved TFBS deviate significantly from a uniform distribution. More specific, for 201 of 246 TFs the Chi-Square test detects a p-value below the conservative threshold of 10^{-4}. Only 31 TFBS are overrepresented in the *non-CGI fraction*, while 128 show an at least 2 fold overrepresentation in the *CGI fraction*. The mean *CGI Mountain* strictness correlates with the *CGI Overrepresentation*. This indicats that TFBS that are overrepresented in CGIs tend co-locate with the CpG-dense CGI cores rather then the less CpG dense CGI shores. The strongest *CGI Overrepresentation* of 17.8 can be detected for the two motifs of the TF *Sp1*.

To infer in how far the *CpG decay* effect can explain this overrepresentation, the above introduced methods are applied. Starting with *Sp1* as a showcase, the analysis is expanded to all TFBS in the JASPAR database (Vlieghe, Sandelin et al. 2006), and then to motifs that have been selected for their ability to discriminate between methylated and unmethylated genome regions.

Sp1 as showcase

Sp1 was discovered as the first representative of a family of *Sp1*-like transcription factors, which contains eight additional members (Zhong and Meng 2005).The *Sp1* transcription factor was implied early as mediator of CGI function (Macleod, Charlton et al. 1994), as it binds the *GC-box* element (GGGGCGGGG) and is present in many CGIs. I tested in how far the protection from *CpG decay* contributes to this overrepresentation.
The JASPAR database reports 35 *cis*-regulatory sequences for *Sp1*. The $\log_2$-odds scores for these sequences varies between -0.28 and 6.29 with an average of 2.75. Thus, *Sp1* binding sites appear on average with a 6.7 fold higher likelihood (2.75 log-likelihood) in unmethylated sequences.

An analysis of the PSSM reported in *JASPAR* only predicts an overrepresentation of 1.54, respectively a $\log_2$-odds score of 0.62. Although the results show a similar trend, this indicates that the loss of information by converting the explicit sequences into PSSM form is substantial. Especially the CpGs that are present in many *Sp1 cis*-regulatory sequences are only weakly represented by the PSSM. On basis of the direct computation, I conclude that *Sp1* binding sites are either strongly favored by an unmethylated neutral substitution process or were adapted by evolution for unmethylated genome regions.

PSSM based analysis of the JASPAR TFBS database

Being aware of the potential bias, the biding sites of many TFs are only available in form of PSSMs. To obtain a broad overview of the spectrum of odds ratios, $L(\wp)$ was computed for 549 PSSMs of TFBSs, obtained from the vertebrate-core set of the JASPAR database (Table 3.5 and Appendix D). Odds-Ratios ranged from 67.6 for *DAL81* (Figure 3.1) to 0.5 for *Pou5f1* (Figure 3.2). 186 PSSMs showed higher likelihood over the unmethylated background model *U*. Of these 93 showed at least a two-fold higher preference for formation by chance in unmethylated DNA. No PSSM displayed such a strong preference for methylated DNA.

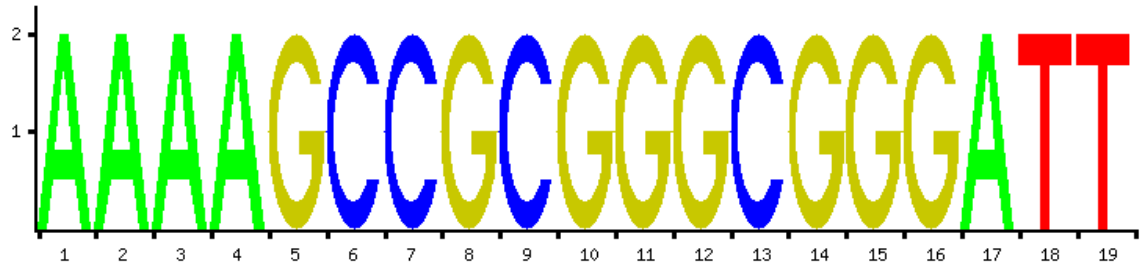


Figure 3.1: Sequence logo DAL81

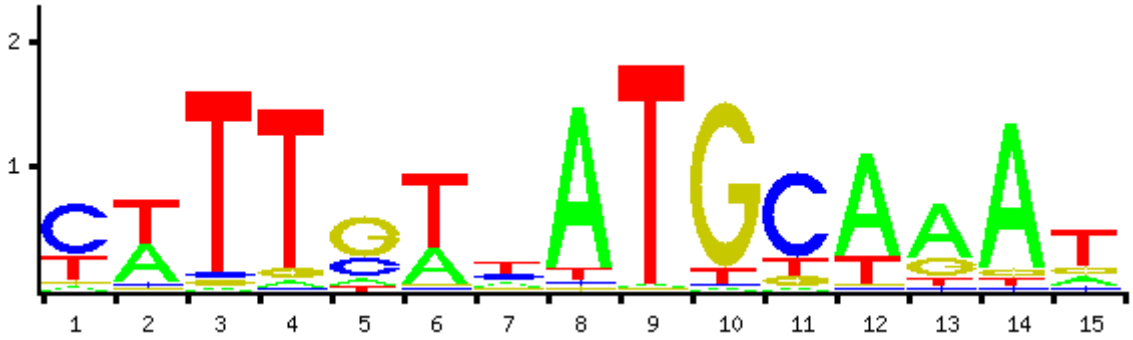


Figure 3.2: Sequence logo Pou5f1

This indicates that no TFBS is overrepresented in methylated DNA by chance beyond a 2:1 ratio, while numerous TFBS exist, which exceed this ratio in favor of unmethylated genome regions. This observation is explained by a mathematical argument. The absence of CpGs in binding motifs only slightly increases the preference for *M*, while the presence of CpGs introduces a strong preference for *U* (*asymmetric preference*). In other words, motif that is highly specific for *U* can be easily constructed by including many high probability CpG sites. In contrast, it is not possible to construct a motif that strongly favors methylated genome regions over unmethylated ones.

Rank	TFBS Name	Odds-Ratio	Log Odds
1	DAL81	67.59394	6.078822
2	RSC30	24.19466	4.596617
3	PDR3	17.12353	4.097909
4	RSC3	14.12888	3.820575
5	IME1	13.82657	3.789372
6	RDS1	10.41102	3.380039
7	SWI4	8.482207	3.08444
8	MBP1::SWI6	7.578759	2.921962
9	MIZF	7.335728	2.87494
10	UGA3	6.71728	2.747877
11	STP1	6.605268	2.723617
12	LEU3	6.503181	2.701146
13	YLL054C	6.424421	2.683566
14	STP2	5.754297	2.52464
15	PDR1	5.477768	2.453588
16	SUT1	5.361134	2.422538
17	E2F1	5.333781	2.415158
18	MBP1	5.22171	2.384522
19	NHP10	4.881334	2.287276
20	GAL4	4.745383	2.246525
21	UME6	4.665303	2.221971
22	CHA4	4.633145	2.211992
23	TEA1	4.578067	2.194738
24	PUT3	4.526733	2.17847
25	CAT8	4.398683	2.137072
26	SNT2	4.36613	2.126355
27	RDS2	4.319054	2.110715
28	YER184C	4.219052	2.076919
29	SIP4	4.215525	2.075712
30	YJL103C	4.199741	2.0703
31	HAL9	4.156235	2.055277
32	YBR239C	4.147621	2.052284
33	TBS1	4.071869	2.025691
34	RDR1	4.06493	2.023231
35	ASG1	4.056344	2.02018
36	CEP3	4.004818	2.001737
37	XBP1	3.959635	1.985368
38	STB4	3.94848	1.981298
39	YLR278C	3.910654	1.96741
40	PDR8	3.900859	1.963792

Table 3.5: Top-40 PSSM motifs ranked by odds ratio

Summary

Conserved transcription factor binding sites (TFBSs) show a higher prevalence for CGIs than for the remaining genome. A similar tendency is predicted by the likelihood ratios computed for the position-specific scoring matrices (PSSMs) of the TFs, which in general favor sequences closer to the unmethylated equilibrium. The ranked lists produced from both comparisons are resources for judging the association type of an individual TF to CpG-rich sequences. A high rank in the likelihood list indicates a strong *passive association*. A high rank in the conserved TFBS list despite a low rank in the likelihood list indicates an *active* or *indirect* association. In the next sections these resources are used for the interpretation of computational epigenetics study results.

3.4.4 Correlation of tetranucleotides to CGI methylation state

Two consecutive studies assessed the overrepresentation of 4-meres (tetranucleotides) in unmethylated CGIs in comparison to methylated CGIs. Both studies were descriptive and primarily focused on identifying a predictive correlation between different properties of CpG rich genome regions and their methylation state. Applying our knowledge of the influence of DNA methylation on genome evolution now enables a more refined evaluation of the results.

The pilot study was performed on a previously published dataset of 706 CGIs with known methylation state in lymphocytes (Bock, Paulsen et al. 2006). The follow-up study applied a larger methylation dataset from brain tissue and facilitated three different *sequence-based CGI definitions* (Bock, Walter et al. 2007). Based on these definitions, the authors applied the *CpG Island Searcher* software to produce three different *sequence-based CGI annotations*. The *TJU* annotation followed the *TJ* criteria, whereas the *GGF'* and *GGM* followed each the *GGF* criteria. Hereby, the *GGF'* was filtered by removing all CGIs that contained less than 200 bp non-repetitive sequence and the *GGM* was computed on a repeat masked genome (compare section 2.1.1). This approach resulted in 37531, 94450 and 10600 CGIs. In both studies a non-parametric Wilcoxon Rank-sum test determined those tetranucleotides, *i.e.* 4-meres that showed a significant overrepresentation in either methylated or unmethylated CGIs. The obtained p-values were then corrected for multiple testing. Table 3.6 displays the p-values of all four cohorts with the likelihood ratio obtained from the DNA word model.

All detected patterns showed an overrepresentation in unmethylated genome regions. This is in line with the *asymmetric preference* observed in the previous subsection. The tetranucleotide CGCC was reported in all four cohorts as being significantly overrepresented. Surprisingly enough the enrichment of the pattern CCGC was less pronounced in the *GFF'* and *TJU* sets, while its overrepresentation in *GGM* and the *Pilot* set was highly significant. As the likelihood ratios of both patterns are equal, this indicates a bias in the data, which is most likely related to the different ways repetitive sequences are treated.
Potentially, one of both tetranucleotides is very abundant in CpG-rich repetitive sequences. An inspection of the consensus sequence of the ALU repeats (Appendix F) verifies this assumption. The pattern CGCC and its reverse complement GGCG is contained 8 times in the consensus sequence, while CCGC/GCGG is only contained 2 times. ALU repeats are often associated to methylated CGIs thus counterbalancing the natural enrichment of CCGC in unmethylated CGIs. In the repeat masked *GGM* set, this balancing effect is neutralized.

The palindromic pattern CGCG is expected to be 16.9 enriched in unmethylated *CGIs*. Surprisingly, only the *Pilot* cohort and the *GGM* set detect a significant correlation. An inspection of the ALU consensus sequence shows four occurrences of the pattern of which two overlap with each other. This again indicates a bias in the data.

The pattern CCCC and CTCC are not overrepresented in ALU repeats. CCCC shows small but significant correlation with the unmethylated state in the *GGM*, *TJU* and *Pilot* cohorts. This follows the slightly increased likelihood ratio. Therefore, this pattern may be explained by *passive association*. The association of CTCC from the pilot study is not confirmed in any of the other cohorts and most probably was an artifact of the small sample size.
The best candidate for an *active* or *indirect association* is the tetranucleotide AAAG. Although it is expected to be underrepresented in unmethylated CGIs, it is significantly enriched in all cohorts, but especially in *TJU* and *Pilot*. The benchmark performed in the chapter 2, assists us in evaluating this observation. The *TJU* is equivalent to the *CIS TJ* annotation, thus represents CGIs that show good overlap with promoters and is relative compact. Hence, the high significance reached in the TJU set, indicates a relationship to promoter function and transcription initiation, thus arguing for an *indirect association*.

Pattern	**Pilot**	**TJU**	**GGF'**	**GGM**	**U/M ratio**
CCGC	**3.7×10^{-7}**	**1.6×10^{-5}**	**1.7×10^{-3}**	**3.9×10^{-6}**	6.17
CCCC	**9.8×10^{-7}**	1.1×10^{-3}	9.9×10^{-2}	6.1×10^{-3}	1.64
AAAG/CTTT	6.3×10^{-6}	**8.5×10^{-5}**	**1.7×10^{-2}**	1.6×10^{-3}	0.8
CGCC	3.6×10^{-5}	5.1×10^{-3}	3.2×10^{-2}	**8.0×10^{-5}**	6.17
CTCC	1.0×10^{-4}	-	-	-	1.25
CGCG/CGCG	1.8×10^{-2}	-	-	1.1×10^{-3}	21.36
TATT	-	2.4×10^{-3}	-	-	0.7
GGAA	-	2.8×10^{-3}	2.1×10^{-2}	-	1.19
GAAA	-	3.3×10^{-3}	-	-	0.92
TCCT	-	3.9×10^{-3}	-	-	1.25
CAAA	-	6.0×10^{-3}	-	-	0.72
CCCG	-	7.5×10^{-3}	-	-	5.66
TTCT	-	8.1×10^{-3}	-	-	0.96
AAGG	-	8.1×10^{-3}	-	-	1.03
GTTC	-	1.0×10^{-2}	-	-	1.06
CGGA	-	1.3×10^{-2}	4.2×10^{-2}	-	5.32
GCCG	-	-	-	1.9×10^{-3}	6.17

Table 3.6: Comparison of overrepresented 4-mers in different CGI annotations
The table shows the 4-mers, which were significantly overrepresented two studies. For four different region sets these p-values are displayed, which are corrected with the Bonferroni method for multiple testing. The last column displays the likelihood ratio that these 4-mers reach under the unmethylated over the methylated background models. The two smallest p-values in each cohort are denoted in green.

Overall the *GGF'* observations appear to be a diluted version of the *TJU* and only the *GGM* displays diverging patterns. The observation that many of the significant tetranucleotide patterns are associated to ALU repeats, underline the necessity for a more explicit treatment of these sequences. I will return to this observation in chapter 4.

3.4.5 Correlation of Tetranucleotides to promoter regions

In the second correlation study experimentally validated promoter sites and regions with open chromatin structure were applied as marker for *CGI* function, instead of the tissue specific methylation levels of the first correlation study (Bock, Walter et al. 2007). The tetranucleotide CACA/TGTG was strongly enriched in transcriptionally inactive regions in two datasets, while the pattern CGCG was strongly enriched in regions that showed promoter activity (Table 3.7). This correlation perfectly reflects the expectations from the likelihood ratios. While the first pattern is a putative product of a double deamination event, the second pattern is a not deaminated double CpG. This general tendency was strongest for the *GGF'* and *GGM* annotations where all tetranucluotides among the 30 most significant features contained at least one CpG. In the *TJU* set five out of nine tetranucleotides contained a CpG.

Pattern	TJU	GGF'	GGM	U/M ratio
CACA/TGTG	-	*$< 1.0\times10^{-55}$*	*$< 1.0\times10^{-55}$*	0.78
CGCG/CGCG	**1.1×10^{-10}**	**8.7×10^{-45}**	**6.3×10^{-46}**	21.36
CGCC/GGCG	-	**1.2×10^{-43}**	**3.0×10^{-46}**	6.17
CCGC/GCGG	-	**1.1×10^{-40}**	**1.6×10^{-42}**	6.17
GCGC/GCGC	-	**7.3×10^{-35}**	**5.2×10^{-33}**	6.73
CCGG/CCGG	-	**2.7×10^{-30}**	**2.6×10^{-31}**	5.66
CCCG/CGGG	-	**1.4×10^{-29}**	**1.1×10^{-34}**	5.66
CGGC/GCCG	-	**1.9×10^{-28}**	**9.1×10^{-26}**	6.17

Table 3.7: Tetranuclotides that correlate with promoter function
Negative correlation is highlighted in red italic and positive correlation in bold green writing. P-values are displayed for all features that ranked among the 30 most significant ones in the respective CGI annotation.

To summarize this finding, the tetranucleotide composition of genome regions correlate far better with promoter function than with tissue specific methylation, with respect to the expectations from the equilibriums based evolutionary model. This indicates that promoters are unmethylated in the germline, while tissue-specific methylation in lymphocytes correlates less strictly to the germline methylation state.

Moreover, this suggests that the observed nucleotide patterns are not *actively associated* with the tissue-specific methylation state. The observation that all these tetranucleotides are under strong mutational pressure in methylated sequences, indicate that they require protection from *CpG decay* in the germline. Whether this protection is mediated by a germline-specific lack of methylation or by strong selective pressure remains to be clarified.

3.4.6 Correlation of sequence motifs to DNA methylation and promoter activity

A third study attempted to discriminate methylated from unmethylated genome regions by their DNA sequence. In contrast to the studies in the previous section, it focused on transcription factor binding sites from a database and the identification of novel motifs (Straussman 2009). The applied machine learning setup identified six motifs, which were enriched in unmethylated regions. This section applies the techniques described above, to build hypothesis for the explanation of the observed correlations.

Two of the six motifs were previously known – *Sp1*-like and *STAT1* binding sites -, and four were classified as novel. A closer examination of these novel motifs during the preparation of this analysis showed that three of these four novel motifs are fragments or degenerated versions of known TFBS.

In the second step of the Straussman *et al.* study, different algorithms were applied to derive weights for each motif's *PSSM,* and thus, a ranking of their importance for the prediction of the methylation state of arbitrary genome regions was established. Table 3.8 reports for each motif these algorithm-specific weights (as approximation for a motif's importance), the odds ratios derived by the equilibriums-based approach and the TFBS it represents.

Motif names	**Algorithm 1**	**Algorithm 2+3**	**U/M odds ratio**	**TFBS**
NGGGGGCGGGGYN	0.31	148.12	2.59	Sp1-like
CANTTCCS	0.24	53.59	1.57	STAT1
CGCGC	0.28	158.85	31.71	Frag. Egr1
CCGCSCC	0.21	121.46	10.70	GC-Box
CGCNNNCGC	0.27	150.35	41.50	Deg. Egr1
CTAR	0.19	56.3	1.22	unknown

Table 3.8: Predictive motifs are favored by methylation-free neutral evolution
Those TFBS motifs marked as fragments (Frag.) or degenerated (Deg.) were not recognized as belonging to the respective transcription factor in the original study

The odds ratios clearly indicate that all motifs are favored by neutral evolution in a methylation-free setting. Especially, the GC-Box and the Egr1-associated motifs are under strong mutational pressure in methylated DNA.

The *PSSM* of the *Sp1*-like element (NGGGGGCGGGGYN), which the authors derived from the *TRANSFAC* database (Wingender, Dietze et al. 1996), slightly deviates from the *Sp1* TFBS motif in the *JASPAR* database. According to all three algorithms it was ranked to be among the top-3 weighted features. Upon close examination, the reverse complement of the directly recovered motif CCGCSCC, where S represents C or G, is GG(C/G)GCGG and thus closely resembles the GC-box, which is the core element of *Sp1* and *Sp1*-like TFBS. Interestingly enough, due to the reduced number of wild card nucleotides, this motif version is considerably stronger favored by the unmethylated background model than the general *Sp1* motif. It represents a specialized sequence tag that is bound by multiple factors. Speaking in terms of the library metaphor, it is like a keyword that many different readers recognize.
It is noteworthy that the already discussed 4-mer CCGC/GCGG is part of this second variant of this *cis*-regulatory sequence, while CGCC/GGCG can also be recovered if the wildcard S is chosen accordingly. Thus, *Sp1* binding sites account for two of the recovered motifs, which indicate unmethylated genome regions, and contribute to the predictive power of two 4-mers discussed in the previous section. This is in line with previous reports and the expectations from the likelihood model.

Both of the other motifs in the top-3 list (CGCGC and CGCNNNCGC) were substantially stronger favored by the unmethylated background model (31.71 and 41.5 fold higher likelihoods). CGCGC does not correspond directly to a known binding motif, but is also conditionally contained in the GC-Box CCGCSCC. As reported in the last subsection the CGCG as well as the GCGC tetranucleotides strongly correlate with promoter activity. Apparently the 5-mer that joins both tetranucleotides is also predictive for the absence of methylation.

A search in the *JASPAR* database revealed that the reverse complement of CGCNNNCGC (GCGNNNGCG) resembles a degenerated version of transcription factor *Egr1* binding motif GCG(T/G)GGGCG (U/M odds ratio 3.7), and thus is called *Egr1*-like motif in the following. It is much stronger favored by the unmethylated equilibrium then the *Egr1* motif in the JASPAR database and takes the second respectively the third rank among the motif weights in the Straussman *et al.* study.

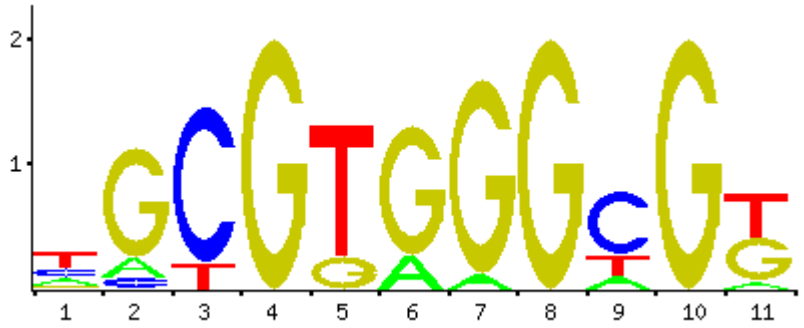


Figure 3.3: Sequence logo of Egr1

Egr1 is also known as *Zif268*, *NGFI-A*, *TIS8*, *Krox-24* and *ZENK* and involved in mammalian brain development (Knapska and Kaczmarek 2004). Furthermore, it directly regulates multiple tumor suppressor genes including *TGFβ1*, *PTEN*, *p53* and *fibronectin* (Baron et al., 2008). It takes rank 16 among the TFBS overrepresented in *CGIs* (compare Table 3.3). The promoters, and thus the binding sites for *Egr1*, are reported to be hypermethylated in certain cancer types (Whang, Wu et al. 1998). The direct prevention of *Egr1* binding by DNA methylation detected in those studies indicates that the CpG sites in the motif play a role in the tissue-specific regulation of *Egr1* binding (*regulatory association*). Furthermore, the equilibrium-based approach predicts that over longer time spans DNA methylation induces a high substitution pressure on the *cis*-regulatory sequence. This indicates that the motif benefits from co-location with unmethylated regions by improved conservation (*passive association*), rather than directly repelling *de novo* methylation (*active association*).

Egr1 belongs to the zinc finger protein family. Each tri-nucleotide in the motif is recognized by a particular domain of the protein. In a targeted mutation experiment it was demonstrated that an altered form of the protein recognizes variants of the central trinucleotide (Wu, Yang et al. 1995). Furthermore, three other members of the EGR-family *Egr2* (rank 28 in Appendix D), *Egr3* (rank 8 in Table 3.3) (Patwardhan, Gashler et al. 1991) and *Egr4* (Zipfel, Decker et al. 1998) are known to have nearly identical protein structure and binding motifs. If these homologous factors have slightly different preferences for the central trinucleotide, this may explain that the reverse complement of the more unspecific motif GCGNNNGCG was discovered instead of GCG(T/G)GGGCG.

The remaining motif with previously known function (CANTTCCS) is recognized by transcription factor *STAT1*, which plays an important role in inferon-mediated immune response and tumor suppression by regulating cell growth and apoptosis induction (Hartman et al., 2005). Furthermore, it is known that CpG methylation can prevent *STAT1* binding (Chen, He et al. 2000). The corresponding motif is only slightly favored by neutral evolution under the unmethylated background model and not among the top-3 weighted features for any of the algorithms, thus a *regulatory association* to regions that are unmethylated in the analyzed tissue is the most likely explanation for the predictive power of the motif.

The final motif CTAR ranks least and second least among the weights. It benefits the least from the protection against CpG decay and no association to a binding protein could be established. The reason for its co-location with unmethylated genome regions remains elusive.

Summary

Applying the equilibrium distributions of dinucleotides in methylated and unmethylated DNA proved to be supportive for the interpretation of DNA patterns that are associated with the absence of DNA methylation. Furthermore, the comparison of three studies showed that the *Sp1* binding motif *i.e.* the GC-box in parts or in complete form is a recurring feature that distinguishes genome regions by their methylation state. The motif was encountered in different forms, which all preferentially occur in unmethylated DNA. Thus, our findings support previous claims that *Sp1* is the strongest candidate for directly influencing the methylation state of a genome region.

A second interesting binding motif is that of *Egr1*. It is under strong pressure from *CpG decay* and binding of its *trans* factor is influence by the methylation state of the DNA sequence. This makes *Egr1*-binding sites into *passively associated* markers of the germline methylation state that require protection from *CpG decay* to maintain the potential for a direct tissue-specific regulation by DNA methylation (*regulatory association*).

Furthermore, within the study evidence accumulated that statistical learning methods are biased by the strong correlation of repetitive elements with DNA methylation, and thus instead of identifying epigenetic footprints in the DNA, and not surprisingly, partially derive their predictive power from identifying this latent variable. Moreover, tissue-specific methylation levels showed a less pronounced correlation to DNA sequence features, than other markers of transcriptional activity and open chromatin structure. This may reflect differences between the somatic methylome and germline methylation levels. While the former are tissue-specific and cannot cause heritable changes in the DNA, the later provide protection for all CpGs that have a regulatory function in any tissue, and thus are required to preserve the regulatory potential of a genome region.

3.5 Searching unmethylated regions by equilibrium distribution

The previous chapter focused on sequence-based CpG island annotations. They were primarily motivated by an observed correlation between CpG content and certain genome functions, such as the absence of DNA methylation and the initiation of transcription. Secondarily, they were explained by the spontaneous deamination mechanism and its methylation-mediated influence on point substitution rates, but did not directly apply this knowledge to select the parameters for the annotation procedures. In this section, the genome-shaping force of *CpG decay* is considered in an explicitly quantified form. Previously, the ratio of $(\#TpG+\#CpA)/2\cdot\#CpG$ was applied for a similar purpose (Hutter, Paulsen et al. 2009). The here discussed method can be interpreted as a systematic extension of this approach, which also facilitates the remaining dinucleotide frequencies.

To this end, a distance measure is derived to determine if a genome region is closer to the methylated or the unmethylated equilibrium. This *EqiScore* is then applied to predict the methylation state of CpG islands. Finally, the *EqiScore* distribution in the human genome is characterized.

3.5.1 Definition of EqiScore

To quantify similarity of the dinucleotide distribution G within an arbitrary genome region R to the dinucleotide distribution at the equilibrium (U unmethylated / M methylated equilibrium), both can be represented as vectors in a 16-dimensional Euclidean space. The similarity of such vectors is described by the angle between them, which is computed by the cosine similarity:

$$\cos(\varpi) = \frac{G_1 \cdot G_2}{\| G_1 \| * \| G_2 \|}$$

Hereby, the cosine of the angle between two vectors equals their dot product over the product of their lengths. Identical vectors have an angle of 0 degrees (cosine similarity of 1). All vector components are dinucleotide frequencies from the interval [0,1], such that the greatest possible difference between two dinucleotide vectors is an angle of 90 degrees (cosine similarity of 0). Thus, as long as G_1 and G_2 have at least one co-occurring dinucleotide, $\cos(\varpi) > 0$ always holds.

To translate these similarities into an informative score, the ratio between the cosines similarity of the dinucleotide distribution of the given region to the unmethylated equilibrium over its similarity to the methylated equilibrium is computed:

$$\frac{\cos(\varpi_U)}{\cos(\varpi_M)} = \frac{G \cdot U}{G \cdot M} * \frac{\| M \|}{\| U \|} = \frac{G \cdot U}{G \cdot M} * \delta_{Norm}.$$

The term δ_{Norm}, as the ratio of $\| M \|$ over $\| U \|$, is constant for fixed equilibrium distributions. By multiplying with 100 and rounding to the next integer, an easily interpretable score is created. It is compatible with the integrated filter functions of Genome Browsers. Therefore, the *EqiScore* of a region is defined as:

$$EqiScore(R) \leftarrow 100 * \frac{G \cdot U}{G \cdot M}.$$

3.5.2 Characterization of EqiScore annotations

The predictive power of *EqiScore* depends on the amount of time a genome region was subjected to a constant condition in the germline. Figure 3.4 shows how different levels of GC content in the ancestral genome sequence influence the time interval in which genome regions under different epigenetic constraints have a similar *EqiScore*.
This is of importance bearing in mind that the human lineage has not yet reached equilibrium (Sved and Bird 1990). Furthermore, occurring rearrangements and transposon activity can integrate genome sequences into domains with different methylation levels.

.

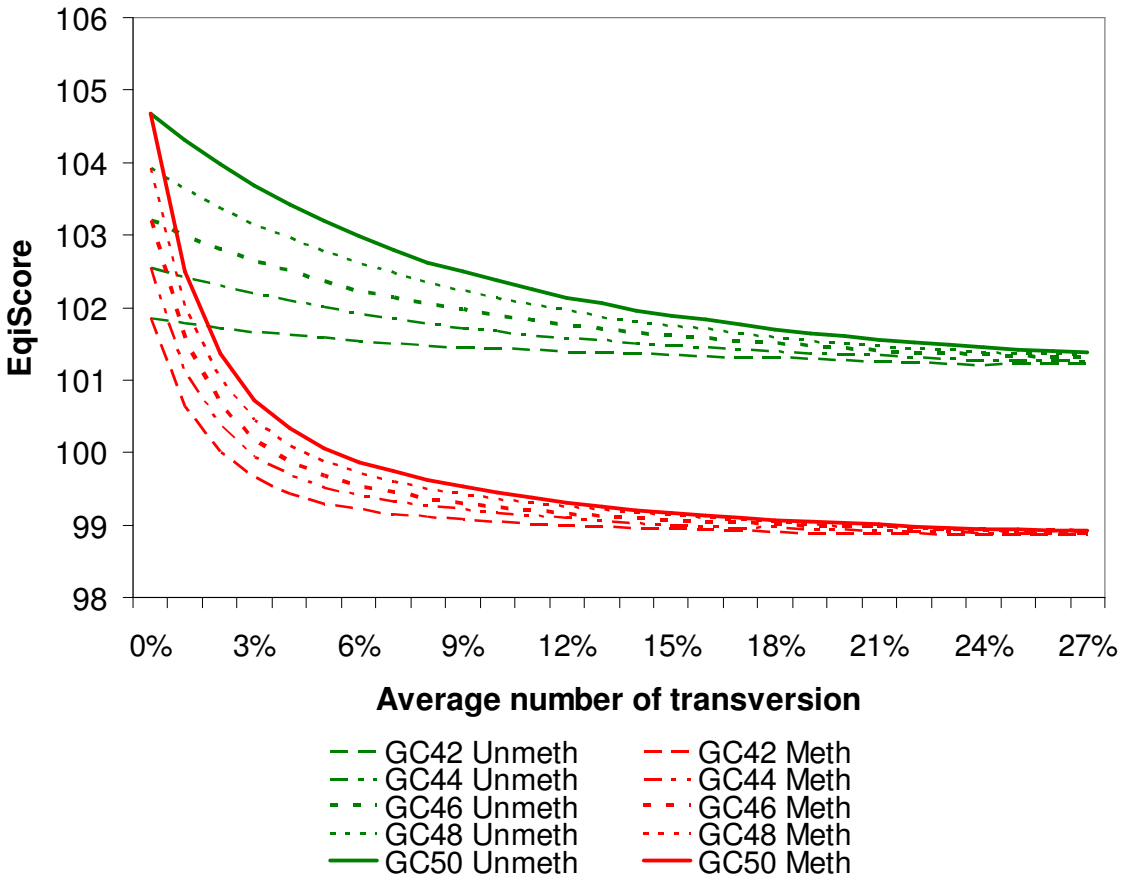


Figure 3.4: Relationship of EqiScore and region age

Change of the EqiScore is shown in regions that evolved under stable methylation constraints. The diagram shows five different starting conditions ranging from sequence with 42% GC-content to uniformly distributed sequence with 50% GC-content. The trajectories are averaged over 100 simulated evolutions for each methylation constraint. Scores are computed including the scaling factor.

Formally speaking, an *EqiScore* analysis requires that M, U and the genomic dinucleotide frequencies at time point t, G^t, are known, while the ancestral sequence R and the time interval t until R evolved into G^t are unknown. Based on the *EqiScore* of G^t, it is estimated whether it is more likely that methylated or unmethylated neutral evolution generated G^t. A simulation study verified that such a classification performs best for $t \to \infty$ (Figure 3.4). At starting time the *EqiScore* is heavily influenced by the original sequence composition, but after a time equal to 4% transversions the score can well discriminate both classes

Figure 3.5 shows the portion of the genome that has at least a certain *EqiScore*. It is remarkable how strongly large parts of the genome exceed the predicted equilibria of unmethylated DNA at an *EqiScore* of 102. This indicates that the genome is locally more skewed towards the unmethylated equilibrium than the uniformly distributed genome sequence applied in the simulation study. In other words, large parts of the genome are not yet at the equilibrium or under selective pressure.

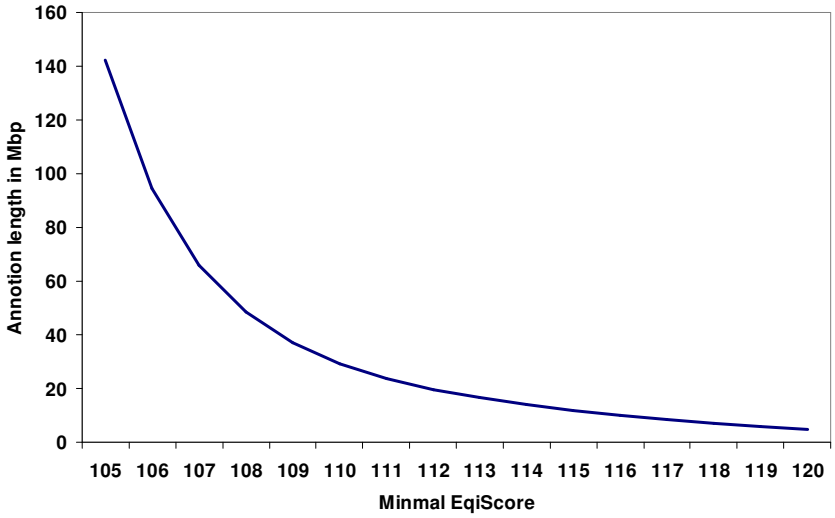


Figure 3.5: Size of EqiScore based annotation
The minimal EqiScore that qualifies a bin to be retained in the annotation is plotted against the total annotation length.

3.5.3 EqiScore as predictor of tissue-specific DNA methylation

Next, the ability of *EqiScore* to predict the local methylation state in the human genome was evaluated. As an initial test, I benchmarked *EqiScore* against the *CGIs* dataset (Yamada, Watanabe et al. 2004) that was applied in the earlier discussed *Pilot* study on the prediction of CGI methylation states (Bock, Paulsen et al. 2006). The average *EqiScore* of the unmethylated *CGIs* was 110.4, while the methylated *CGIs* achieved 108.2. These values are heavily influenced by the algorithm that was applied to annotate the CGIs, as it determines how far CGI is extended in both directions around its core. To asses the prediction performance independently from the *CGI* annotation procedure of the original study, I then moved windows of different sizes over the sequences and recorded the *EqiScore* for each position. Predictions were performed once based on the maximum and once based on the median of these values. The prediction performances of both approaches were assessed by computing the area under curve (*AUC*) of the respective receiver operating characteristic (*ROC*) curves (Table 3.9), using the *ROCR* package (Sing, Sander et al. 2005). All procedures produced an *AUC* above 70%. Windows of length 500 performed best with *AUC*s around 80% for the maximum as well as the median approach.

Window size	**AUC – Max**	**AUC - Median**
100 bp	74.3 %	71.4 %
200 bp	70.8 %	75.0 %
300 bp	73.2 %	77.9 %
400 bp	80.0 %	79.9 %
500 bp	80.2 %	80.9 %
Full region	70.8 %	70.8 %

Table 3.9: EqiScore as predictor of CGI methylation
For different sizes of the sliding EqiScore window predictions were performed according to the maximal score (AUC-Max) or median score (AUC-Median) and the AUC of the resulting ROC curve reported. For the Full region only one value exists, thus maximal and median score are identical.

Figure 3.6 displays the *ROC* curve that compares the best and the worst performing procedures. It is noteworthy that the *EqiScore* computed on the full *CGI* is the worst predictor of methylation. This is most likely a consequence of the originally applied algorithm for the *CGI* annotation, which extended the region as much as its constraints allow, and thus diluted the signal that *EqiScore* detects.

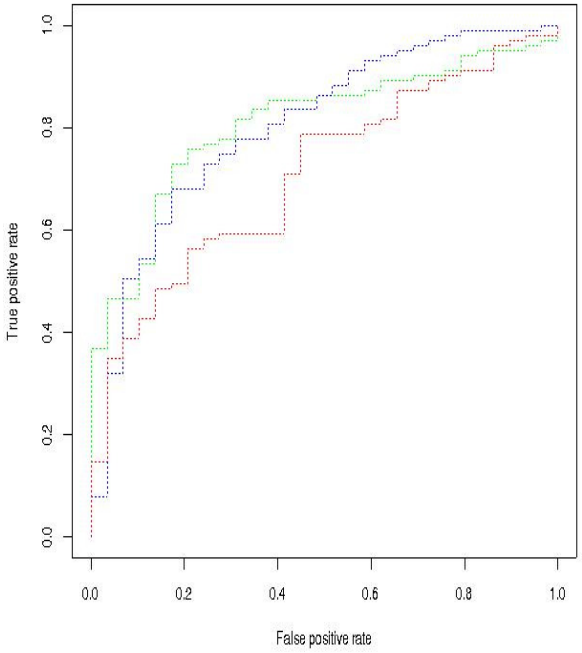


Figure 3.6: EqiScore as predictor of CGI methylation
The median (green) and the maximum (blue) EqiScore computed by 500 bp long sliding windows on all CGIs outperforms the EqiScore computed on the complete CGI as predictor of methylation state.

3.5.4 EqiScore as predictor of local chromatin state

Next, I computed genome-wide *EqiScore* annotations for bins with the size of size 50 bp (runtime: 2 h 8 min). As the *EqiScore* of a single nucleotide position is not very informative, the local sequence neighborhood is taken into account. Thus, the *EqiScore* of each bin was computed by considering the neighboring 200 bp (100 bp on each side).

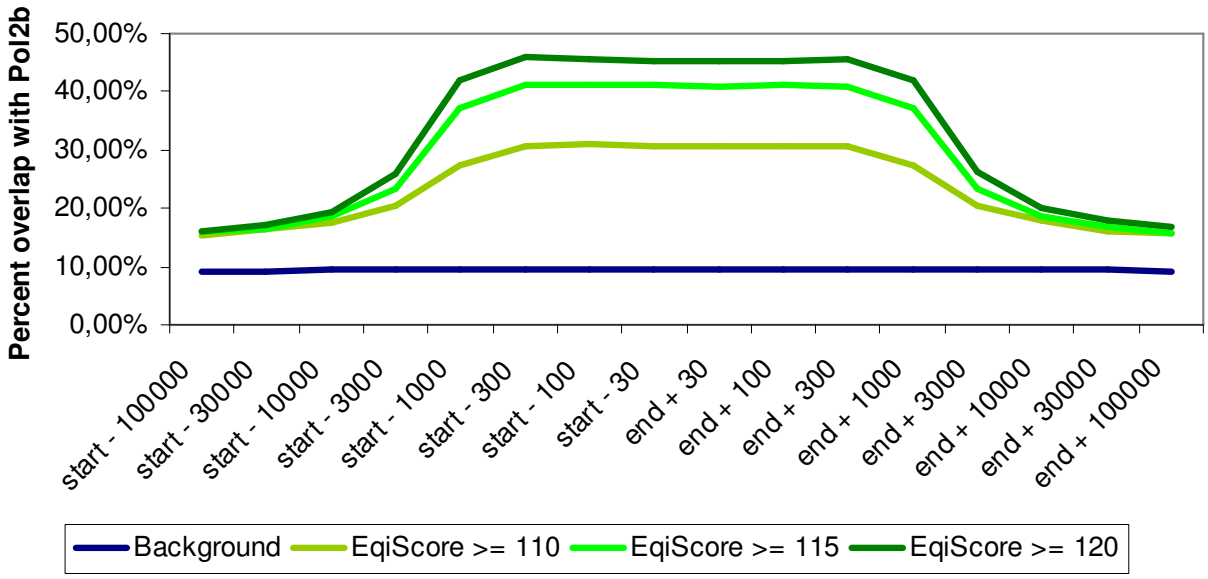


Figure 3.7: The neighborhood of genome regions with high EqiScore is enriched with RNA Polymerase II binding sites

I then tested if a correlation between *EqiScore*, Polymerase occupancy and histone modification can be observed. As illustrated in Figure 3.7 and Appendix F, RNA Polymerase II binding sites and histone modifications are observed more frequently in the local neighborhood of regions with higher *EqiScore*. This enrichment was measured for activating histone modifications as well as for repressive ones, indicating that higher *EqiScore* correlates not exclusively with marks of open chromatin, but with stronger epigenetic control across tissues in general. A high *EqiScore* predicts a region's low methylation level in the germline. Possibly a considerable number of such regions are tightly epigenetically regulated during differentiation in somatic tissue, thus potentially also gaining heterochromatic features such as repressive histone marks and DNA methylation.

Summary and discussion

The *EqiScore* approach can be directly applied to translate knowledge on the methylated and unmethylated dinucleotide equilibrium distributions into an equivalent of a CGI annotation. The produced annotations correlate with CGI annotations, polymerase binding sites, activating and repressive chromatin marks. The method shows parallels to the HMM-based approach proposed by Wu et al., which was introduced in the last chapter. Both approaches use sequence features of small genome regions to quantify the contained epigenetic footprints in the DNA. Where *EqiScore* derives its parameters directly from the equilibrium distributions, the approach by Wu et al. infers them from the sequence data. The discrepancy between the empirically measured and the simulated *EqiScore* at equilibrium strongly indicates that large parts of the genome are not in equilibrium. Nonetheless, *EqiScore* predicts the somatic methylation state of CGIs in human lymphocytes with a remarkable accuracy. This indicates that the epigenetic footprints in the DNA alone can predict a substantial fraction of tissue-specific DNA methylation. The neighborhood plots furthermore show a quantitative relationship between the *EqiScore* and the enrichment of epigenetic marks. This leads to the conclusion that *EqiScore* as an equilibriums-based method is an effective approach to choose parameters for the annotation of epigenetic footprints in the DNA.

3.6 Discussion

In this chapter nucleotide substitution rates in the human genome were applied to numerically approximate the equilibrium distributions of fully methylated and fully unmethylated DNA. These differ profoundly in their CpG, CpA and TpG content and all dinucleotides that exclusively contain A and T.
These equilibrium distributions were then applied to reevaluate descriptive results on DNA patterns that were previously found to be enriched in unmethylated regions. This analysis confirms that *CpG decay* explains the predictive power of most tetranucleotides. During this process evidence accumulated that repetitive elements introduced bias into analysis of epigenetic footprints in the DNA by containing CpG-rich tetranucleoties at very different frequencies.
Furthermore, within the previously reported DNA motifs a degenerated form of the CpG-rich *Egr1* binding site was identified, which is under high neutral mutation pressure outside of unmethylated regions, and thus strongly profits from the protection of promoters from DNA methylation. Furthermore, the *Sp1*-binding site and its functional core element the GC-box was recurrent in all three analyzed studies, although it was not always directly recognized. This supports that hypothesis that the GC-box, as recognition site of several transcription factors, is putatively the most relevant functional element in functional CGIs. It is under pressure from *CpG decay*, is implied in the active protection from DNA methylation and co-located with many CGI promoters. Hence, it shows evidence for *active*, *passive* and *indirect association*.
Finally, I successfully tested the ability of the twin-equilibrium-based *EqiScore* to predict unmethylated CGIs and to annotate regions under active epigenetic regulation. As *EqiScore* is purely based on the differences in neutral substitution rate, it is complementary to *CGI*-based annotations that have empirically chosen parameters.

In response to the detected bias by CpG-rich repetitive sequences, the next chapter introduces a framework to explicitly analyze repeat-specific epigenetic footprints in the DNA to discriminate methylated from unmethylated repetitive sequences.

Chapter 4 - Inferring methylation-induced evolutionary pressure by pairwise alignments of ancestral-descendant sequences

In the attempt to understand the influence of DNA methylation on genome evolution, any additional information potentially improves our reconstruction of the evolutionary process. In this chapter knowledge of the ancestral DNA sequence of an individual genome region is applied for this purpose. To this end, I introduce a statistical framework that compares which of two such methylation models provides the better interpretation of how a specific ancestral sequence evolved into its descendant. This work is an extension of previously reported results (Feuerbach, Lyngsø et al. 2011).

4.1 The statistical framework

The framework is constructed from two components: An analytical model of *CpG decay*-aware DNA nucleotide substitution and a Bayesian model of CpG conservation.

4.1.1 Analytical model of CpG decay

An analytical model of sequence evolution has certain advantages over the simulation approach. Instead of running many repetitions of an experiment to archive a robust characterization of the underlying process, the analytical form can be directly evaluated. Where each simulation represents one possible evolutionary path, the analytical model summarizes all possible evolutionary paths.

To construct such a model, I integrate the substitution rates that underlie the methylated f_M and unmethylated f_U evolutionary processes into a finite-state continuous-time Markov chain model (properties reviewed in Appendix G).
Therefore, the substitution rates are translated into the rate matrix Q. For any neighborhood-dependent substitution model Q is derived for oligo-nucleotides of arbitrary length n_Q that is greater or equal to that of the largest neighborhood-effect included in the model. Here n_Q=2 to account for the *CpG decay* process.
Q is constructed by considering the rows as ancestral sequences that evolve into the descendant sequences denoted by the columns (Figure 4.1). If such a substitution can be explained by a single point mutation, the cell is set to the respective rate. From the definition of the finite-state continuous-time Markov chain follows that the probability of observing two substitution events at the same time is zero. In consequence, all cells that correspond to sequence pairs with more than one substitution are set to zero. Finally, the rate of conservation is set to the negative sum of all values in the corresponding row.

The probability that oligo-nucleotide a is substituted by an oligo-nucleotide b after t units of time have passed is obtained by computing $P(t) = e^{Qt}$ and selecting the corresponding cell $P_{a,b}(t)$. Alternatively, $P(t)$ can be approximated by a Taylor series:

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{t^n Q^n}{n!}$$ (Karlin and Taylor 1975).

To test the influence of n_Q on the computation, a software that construct Q for arbitrary n_Q was implemented.

Q	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	*	r_{tr}	r_5	r_{tr}	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0	0	0
AC	r_{tr}	*	r_{tr}	r_6	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0	0
AG	r_6	r_{tr}	*	r_{tr}	0	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0
AT	r_{tr}	r_5	r_{tr}	*	0	0	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}
CA	r_{tr}	0	0	0	*	r_{tr}	r_5	r_{tr}	r_{tr}	0	0	0	r_6	0	0	0
CC	0	r_{tr}	0	0	r_{tr}	*	r_{tr}	r_6	0	r_{tr}	0	0	0	r_6	0	0
CG	0	0	r_{tr}	0	!	r_{tr}	!	r_{tr}	0	0	r_{tr}	0	0	0	!	0
CT	0	0	0	r_{tr}	r_{tr}	r_5	r_{tr}	*	0	0	0	r_{tr}	0	0	0	r_6
GA	r_6	0	0	0	r_{tr}	0	0	0	*	r_{tr}	r_5	r_{tr}	r_{tr}	0	0	0
GC	0	r_6	0	0	0	r_{tr}	0	0	r_{tr}	*	r_{tr}	r_6	0	r_{tr}	0	0
GG	0	0	r_6	0	0	0	r_{tr}	0	r_6	r_{tr}	*	r_{tr}	0	0	r_{tr}	0
GT	0	0	0	r_6	0	0	0	r_{tr}	r_{tr}	r_5	r_{tr}	*	0	0	0	r_{tr}
TA	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0	0	0	*	r_{tr}	r_5	r_{tr}
TC	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0	0	r_{tr}	*	r_{tr}	r_6
TG	0	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	0	r_6	r_{tr}	*	r_{tr}
TT	0	0	0	r_{tr}	0	0	0	r_5	0	0	0	r_{tr}	r_{tr}	r_5	r_{tr}	*

Figure 4.1: Construction schema for rate matrix Q with $n_Q = 2$

*Consistent with Peifer et al. 2008 the transversion rates r_1-r_4 are summarized by the average transversion rate r_{tr}. Furthermore the A/T to G/C and G/C to A/T transition rates are denoted by r_5 and r_6, while the cells that are affected by the CpG decay effect and the corresponding rate r_7 are indicated by exclamation mark. The * symbolizes the negative value of the sum over the remaining row.*

Propagation of context-dependencies

In the general case, the Markov chain based on Q that comprises a neighborhood dependent substitution process can only produce analytically correct results if n_Q equals the length of the analyzed DNA sequence. For larger sequences this becomes quickly infeasible, although some algebraic shortcuts exist (Lunter and Hein 2004). This phenomenon this caused by a border effect. Let n_Q be large but the sequence under consideration is at least one nucleotide longer. For any oligo-nucleotide in Q that starts with a G it is unclear if a C precedes it. Therefore, it is unclear if the *CpG decay* process affects the substitution of G, and thus, the probability that the nucleotide is substituted into an A is modeled incorrectly. To capture both possibilities n_Q has to be extended, but this introduces the same problem for the novel leading position. Although, the introduced error is small and rapidly decreases for growing n_Q, it exists.
An empirical test of our model for n_Q =2 and n_Q =4, indicated that actually for the four dinucleotides CpG, CpA, TpG and TpA, the model is exact for all $n_Q \geq 2$. This effect is caused by the limited impact of the *CpG decay* effect on CpG dinucleotides and by having a single shared transversion rate.
In Appendix H I proof that under our model assumptions the probability of CpG, CpA, TpG and TpA be either conserved or mutate back into their original state after an arbitrary time span is independent from their neighborhood. This implies that a Q with n_Q=*2* is sufficient to predict the probability that a CpG is conserved at a specific locus.

4.1.2 A Bayesian model of CpG conservation

Given the ancestral and the descendant DNA sequence are known for a specific loci (Figure 1.5b). First, the DNA sequence s_a is aligned to its descendant sequence s_d. If the time t that has passed while s_a evolved into s_d is unknown, the number of observed transversion events is counted in the thus produced pairwise alignment A. Bearing in mind that each nucleotide can underwent two different transversion events, the count of transversion events in the alignment divided by two approximates t. This approximation improves with growing sequence length and for t smaller than 1. To compute the likelihood of an evolutionary model M_L considering the observed alignment A, I apply Bayes' theorem:

$$P(M_L \mid A) = \frac{P(A \mid M_L) \cdot P(M_L)}{P(A)}.$$

Under the assumption that DNA is either fully methylated (M_M) or fully unmethylated (M_U) in the germline, the $P(A)$ term is canceled, while the odds ratio is computed:

$$L = \frac{P(M_U \mid A)}{P(M_M \mid A)} = \frac{P(A \mid M_U) \cdot P(M_U)}{P(A \mid M_M) \cdot P(M_M)}.$$

The term $\frac{P(M_U)}{P(M_M)}$ describes the prior probability to observe germline methylation and is a free parameter of the model that is calibrated with empirical data.
This leaves us with $P(A \mid M)$, which is approximated by the most informative difference between M_U and M_M, namely, the conservation of CpG dinucleotides. The probability p that an individual CpG in s_a is still or again a CpG in s_d after time t is described by $P_{CpG,CpG}(t)$ which can be computed by the model-specific rate matrices Q. Furthermore, this process is independent from neighboring sites even over longer time scales. Thus, CpGs cannot influence each other's back-mutation rate. In consequence, we have a number of independent Bernoulli experiments that follow a binomial distribution, with $nCpG$ being the number of CpGs in s_a and $kCpG$ the number of retained CpGs in s_d:

$$P(A \mid M) = \binom{nCpG}{kCpG} p^{kCpG} (1-p)^{nCpG-kCpG}.$$

I finally define the *L-score* as the logarithm to the base 2 of L. The higher this value, the greater is the likelihood that the ancestral sequence was never methylated during its evolution into the descendant sequence.

4.2 Reconstruction of local germline methylation state from ancestor-descendent alignments of transposable elements

The *EqiScore* method introduced in the previous chapter is able to compare a snapshot of an ongoing neutral evolutionary process with two possible endpoints. The longer this process went on, the closer it approaches one of these endpoints, and hence, the better *EqiScore* performs. Unfortunately, the unknown age and starting point of the process introduce a factor of uncertainty into this procedure. Here this problem is solved, by concentrating on a special case in which good approximations for both of these parameters are available. The ancestral sequence of transposable elements can be very accurately reconstructed from its numerous copies. The age of an individual copy can be inferred from the number of mutations it acquired in comparison to its ancestral sequence.

I first characterize the proposed method in a simulation study, to asses its theoretical performance and limitations. In a small scale pilot study, it is then applied to the sequences of transposable elements in the human genome to asses in how far the produced predictions correlate with genomic and epigenomic features. Then, a genome-wide annotation for all repeats of the ALU family is produced and its quality with respect to epigenetic annotations is assessed.

4.2.1 Simulation study

This section was adapted from the original publication (Feuerbach, Lyngsø et al. 2011)

For this analysis, I generated three libraries with each 100 sequences of length 10 kbp. The *Uniform* library contained uniformly distributed sequences, while the *Methylated* and *Unmethylated* libraries were derived from the *Uniform* library by simulated evolution under methylated or unmethylated constrains, respectively, until they reach approximately their equilibrium distributions.

Next, in three individual runs sequences of length n = 200, 500 and 1000 bp were drawn from all three libraries and then separately evolved 10 times under both methylation regimes until a transversion distance of r_{tr}=0.04 was reached. The likelihood ratios of both models were recorded in time steps of 0.001, resulting in 2000 time series per sequence library. A time series consist of 41 time points from r_{tr}=0.0 until r_{tr}=0.04. Subsequently, the evolutionary model was predicted from the signs of the likelihood ratios and compared to the true model. The objective of the classification was to correctly predict which sequence evolved under the unmethylated model.
A true positive (TP) was defined as a correctly predicted unmethylated sequence. A true negative (TN) was defined as a correctly predicted methylated sequence. Hence, runs under the unmethylated model produced either true positive (TP) or false negative (FN) classifications, while runs under the methylated model produced either true negative (TN) or false positive (FP) results. These counts were used to compute the three performance measures:

$$\text{Sensitivity: } \frac{TP}{TP+FN},$$

$$\text{Specificity: } \frac{TN}{TN+FP},$$

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN}.$$

.

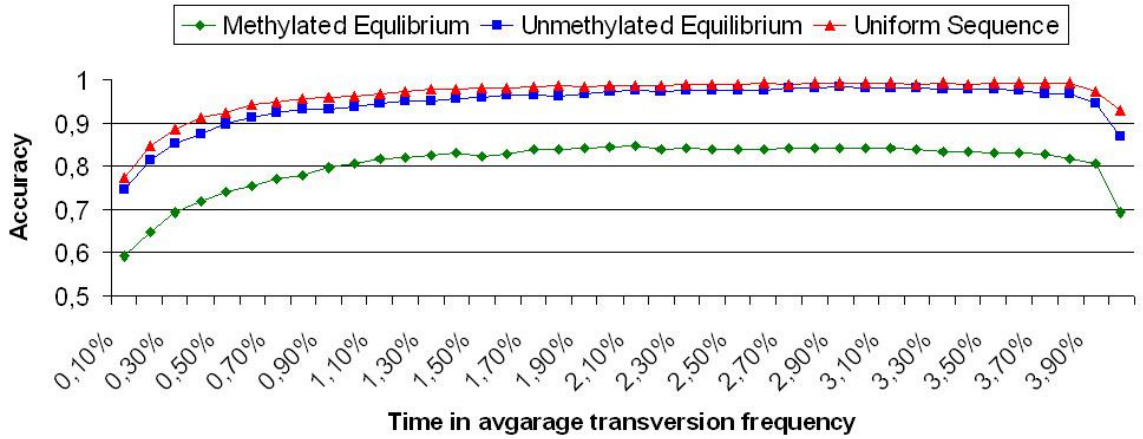


Figure 4.2: Prediction accuracy from simulated evolution of 500 bp sequences *Ancestral sequences were drawn from different equilibrium distributions and performance is plotted over increasing divergence time. Each trajectory is computed over 1000 runs under methylated and 1000 runs under the unmethylated model.*

The prediction accuracy is influenced by two factors. On the one hand by the average number of CpGs per genome region that is determined by its length and epigenetic history. The prediction accuracy is lowest for short regions that are close to the CpG-poor equilibrium distribution of methylated sequences, and highest in long CpG-rich uniformly generated sequences. On the other hand the prediction accuracy is influenced by the evolutionary distance between the ancestral and the descendant sequence. For very short distances the effect on the sequence is not strong enough for a good prediction. The accuracy grows with increasing distance and soon reaches a plateau. When the point is reached were nearly all CpGs are expected to be decayed the model begins to fail. Figure 4.2 shows for sequences of length 500 bp the averaged prediction accuracy over the whole simulation period. It shows that he model becomes unstable around 4% transversions, which corresponds to 8 % transversion events per site, which leads to a sharp decline of the prediction accuracy.

Prediction accuracy on biological sequence

Next, this stochastic experiment was repeats for a natural occurring sequence, namely the consensuses sequence of the *AluSx* transposon family. The high number of 23 CpG dinucleotides that are contained in this approximately 300 bp long sequence resulted in very accurate predictions in the age range of 1% to 4% transversion rate (Figure 4.3). Furthermore, I detected a bias towards unmethylated predictions that resulted in an increased sensitivity.

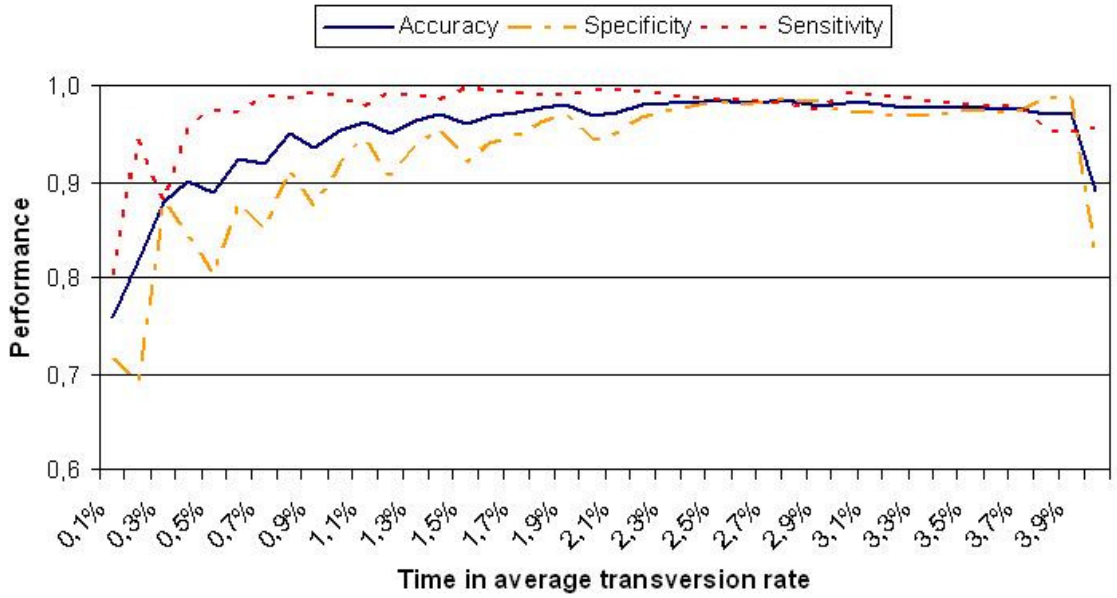


Figure 4.3: Prediction accuracy from simulated evolution of AluSx consensus sequence
The trajectory is computed over 1000 runs under methylated and 1000 runs under the unmethylated model.

Prediction accuracy under noisy substitution rates

To assess how robust the model is with respect to noise in the data, I repeated the experiment and used a normally distributed error. Here, I varied each of the seven rates individually. The noisy mutation rates were computed as $r_{new} = r_{old} \cdot N(1, \sigma^2)$ with the standard deviation σ being a number smaller than one. Negative rates were not permitted. The evaluation was performed using the original rates. As expected, the introduced noise had a negative effect on the performance. Still, even for relatively strong perturbations with a standard deviation of 0.4, the model reaches an accuracy level of 90% around one percent average transversion rate (Figure 4.4).

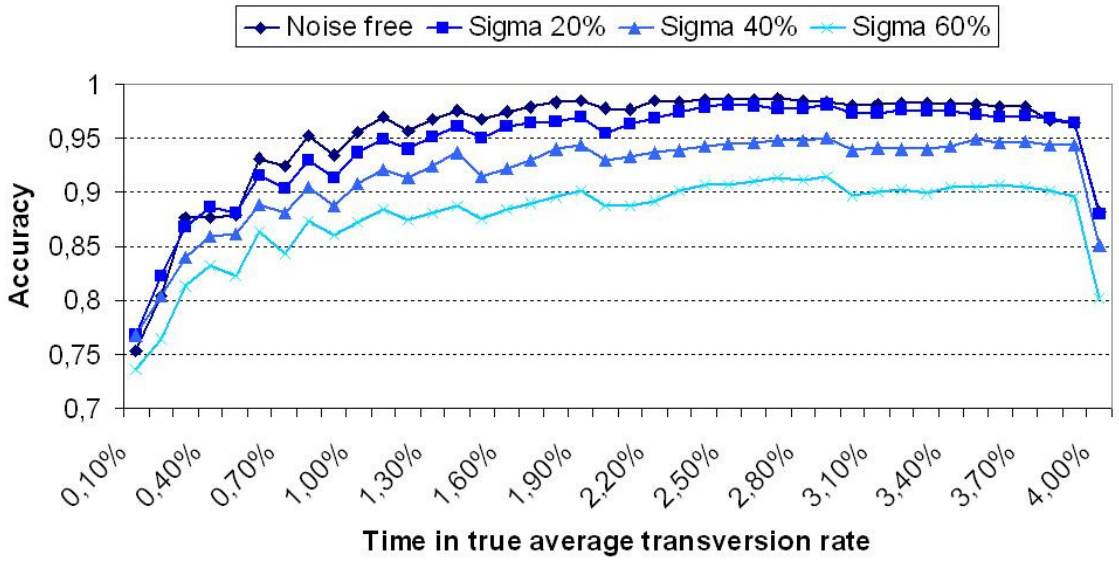


Figure 4.4: Prediction performance from simulated evolution of AluSx consensus sequence with Gaussian noise on all substitution rates

Prediction accuracy under noisy methylation levels

Furthermore, the model was tested for robustness against noisy methylation states. Assuming that fully methylated or fully unmethylated regions are an exception, I either reduced or increased the rate r_7. In case of methylated sequences $r_7 = 48.3 \cdot |N(1,\sigma^2)|$ and for methylated sequences $r_7 = 48.3 \cdot |N(0,\sigma^2)|$ was applied to simulate moderately methylated or nearly fully unmethylated sequences. Variations of the epigenetic state up to 5 % are nearly undetectable. The model achieves for standard deviations up to 15 % from 1% average transversion on 90% accuracy and above. This implies that the model also provides good results, when the methylation state of regions in not binary, but follows a bimodal distribution (Figure 4.5).

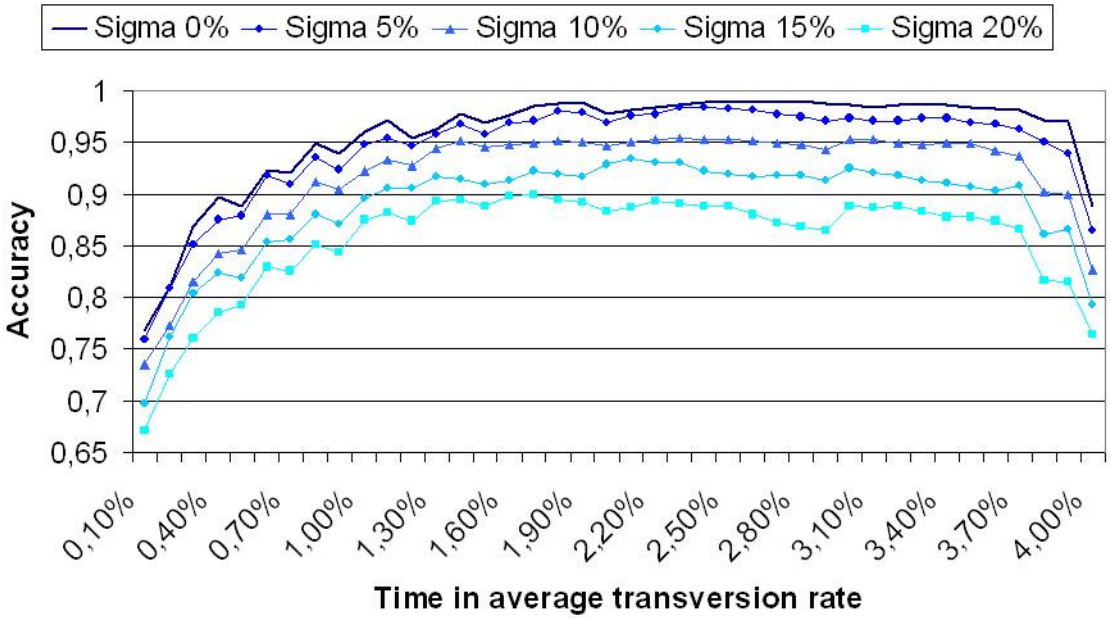


Figure 4.5: Prediction performance from simulated evolution of AluSx consensus sequence with Gaussian noise on methylation level

4.2.2 Validation in the human genome

Pilot study on human AluSx repetitive elements

The coordinates of 104,346 AluSx repeat copies that are longer than 200 bp (RepeatMasker V327 from UCSC genome browser) were obtained and aligned to their reconstructed ancestral sequence (taken from RepBase14.05.fasta at http://www.giri.org). Ancestral repeat sequences in RepBase are consensus sequences derived by majority vote on each position in the multiple alignments of all known members of a repeat subfamily. Special treatment is only given to TpG and CpA dinucleotides, which are adjusted to CpG in the consensus sequence, if observed in a 1:1 relation in an alignment (Jurka 1994; Jurka, Kapitonov et al. 2005). For each repeat instance, I counted the average number of observed transversions over all optimal pairwise alignments to estimate its individual age. The computational procedures were implemented in the python programming language version 2.4 (http://www.python.org/). For the pairwise alignments I applied the *pairwise2* function from the Biopython library (http://biopython.org). The age of a repeat instance, as the time interval since the divergence from the ancestral sequence, was estimated individually for each sequence. Therefore, the number of transversions between the ancestral repeat and the repeat instance were counted. Then, the L-score was computed.

As reported in the original publication (Feuerbach, Lyngsø et al. 2011), higher *L-scores* correlated with closeness to *CGIs* and also with unmethylated regions taken from a published dataset (Illingworth, Kerr et al. 2008) (Figure 4.6).

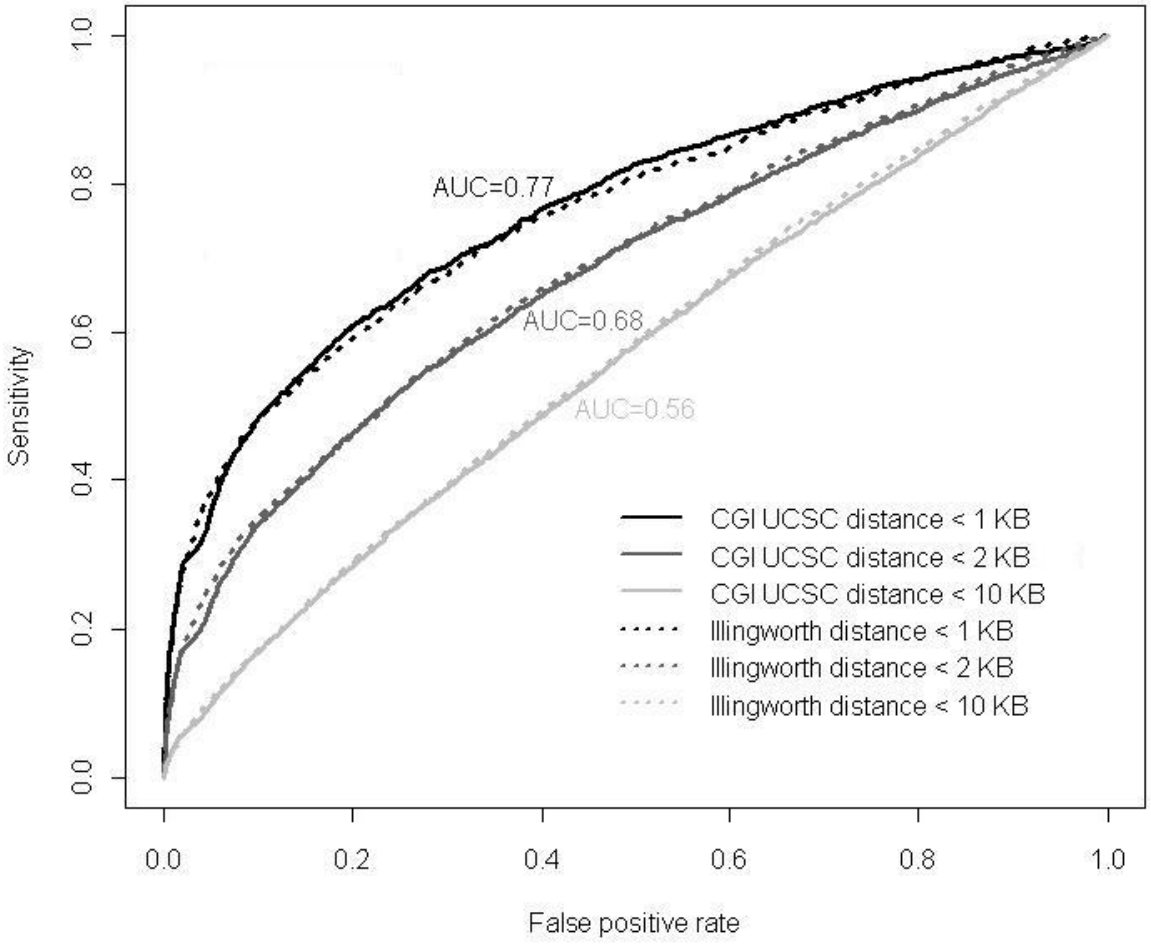


Figure 4.6: High scoring repeats are collocated with CpG islands and unmethylated regions

The displayed ROC curve treats the L-scores as predictors of either CGIs from the stringent UCSC annotation or unmethylated regions from the Illingworth et al. dataset in the vicinity of the respective repeat. With a relaxation of the considered radius from 1 kbp (1 KB) to 10 kbp (10 KB) the prediction performance drops rapidly.

The AluJudge Track and the correlation of L-score to epigenetic marks

Following up the pilot study, I applied the above described procedure to compute *L-scores* for all ALU elements in the RepBase14.05 release, which contained at least 10 CpGs in their ancestral sequence (~1.2 million repeat instances). Based on the results of the simulation study, *L-scores* for repeat instances with transversion rates below 1% or above 4% were set to 0 with respect to the limited prediction accuracy beyond these boundaries (710,484 *cases*). The remaining elements were grouped into four categories: 409,599 cases of strongly methylated (*L-score* < -5), 18,858 cases lightly methylated (-5 < *L-score* <0), 12,089 cases lightly unmethylated (0 < *L-score* < 5), and 25,845 cases strongly unmethylated (5 < *L-score*) repeat instances (Figure 4.7).

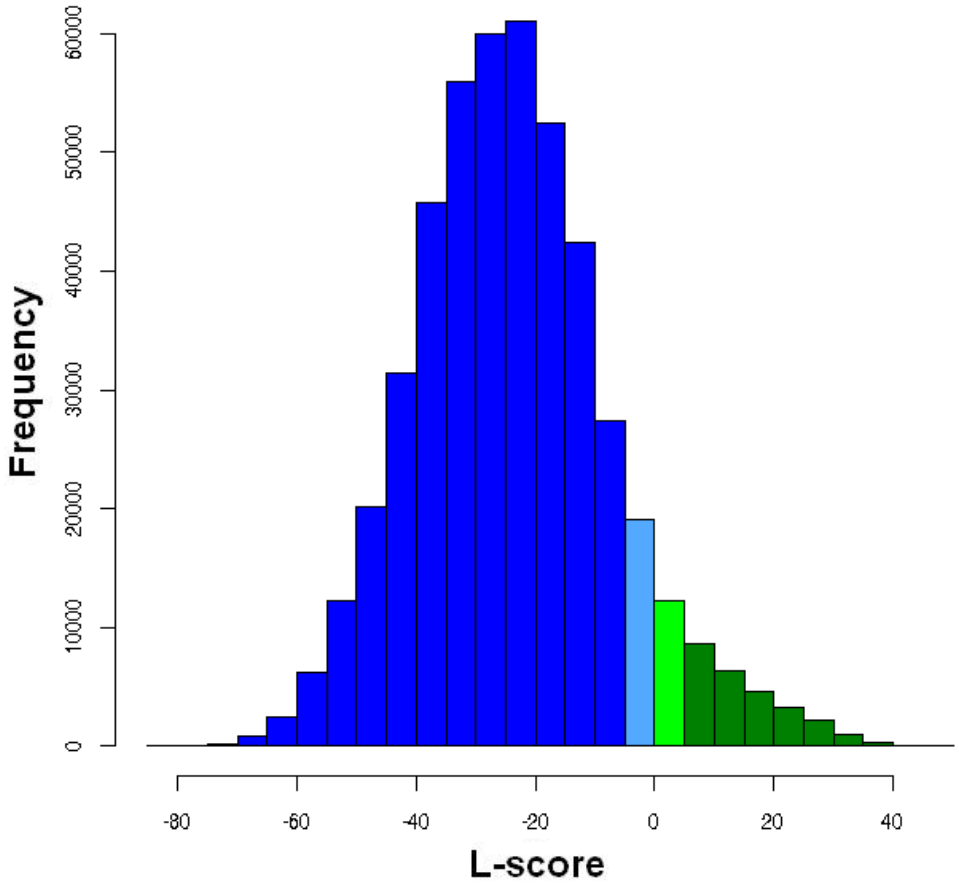


Figure 4.7: Histogram of AluJudge scores
Histogram bars are colored according to the four L-score classes: strongly methylated in dark blue, lightly methylated in light blue, lightly unmethylated in light green, and strongly unmethylated in dark green.

Next, I analyzed the local neighborhood of the scored ALU repeats. To this end, I applied the *EpiExplorer* tool to generate neighborhood plots based on histone modification and polymerase occupancy data.

As expected from the strand unspecific setup of the analysis, all results were symmetric with regard to upstream and downstream locations.
For a number of histone modifications (H3K9ac, H3K27ac, H3K4me1, H3K4me2 and H3K4me3) and the RNA polymerase II occupancy measurements, I observed local correlation effects that were most pronounced in a radius of 3 kbp around the repeats (compare Figure 4.8 and Appendix I). The remaining histone modifications (H3K9me1, H3K27me3, H3K36me3, H4K20me1) show long ranging differences (up to 100 kbp) between groups (compare Appendix I).
In all cases, effect sizes are limited to differences of 5 % to 15 %. Due to the higher number of observations already differences of 2 % are highly significant (point-wise chi-square test with Yates correction yield p-values < 0.001).

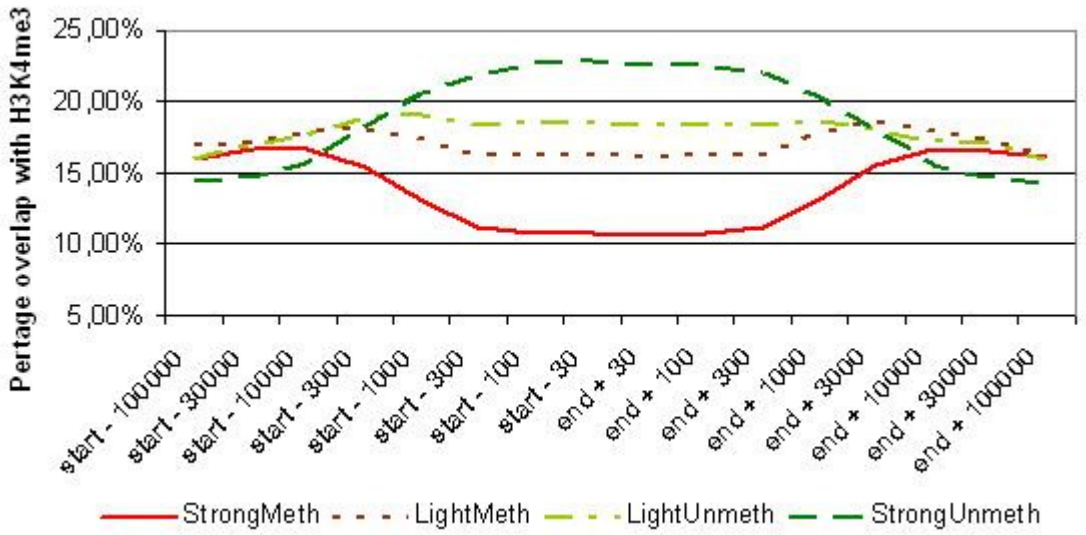


Figure 4.8: H3K4me3 in neighborhood of ALU repeats from different AluJudge classes
H3K4me3 is an epigenetic marker of gene promoters. It is enriched in the direct neighborhood of transposons with high L-score and depleted in areas close to low scoring ALU instances.

Local effects

For most histone modifications with local effects, the repeats with higher L-scores were correlated with higher modification enrichment, while lower scorning repeats were depleted in signal peaks (compare Figure 4.8). A remarkable exception from this rule was H3K4me1, for which depletion was observed for all repeats. This modification is a marker for enhancer elements (see Appendix I). While transposable elements have been discussed before as potential mechanism for the establishment of alternative promoters of existing gens or as promoters for novel non-coding DNA, this observation indicates that they do not play this role for enhancer elements.

Long ranging effects

The modifications H3K9me1, H3K27ac, H3K36me3 and H4K20me1 showed long ranging differences between groups. H3K36me3 is known for marking RNA polymerase II elongation areas. The three remaining modifications are markers of large active or inactive domains. Thus, these results indicate different probabilities for repeats of the different classes to be located in such domains. In all four cases lower L-scores correlated with enriched overlap with these modifications.

4.3 Discussion

By comparing a reconstructed ancestral sequence with its descendants, a direct quantification of the impact of DNA methylation on genome evolution is enabled.
The good performance of the method in the simulation study was validated by the collocation of high scoring repeats with CGIs, unmethylated regions, transcriptional activity and marks of epigenetic regulation through histone modifications. Not surprisingly, the observed correlations were imperfect, as only rough estimates of the repeat age were available, and additional confounding factors like regional differences in substitution frequency and potential selective pressure were not considered (Cohen, Kenigsberg et al. 2011). I conclude that the proposed method facilitate CpG-rich repeats as an information rich, sequence-based predictor of their local germline-epigenome neighborhood. Moreover, the correlation of high scoring repeat instance with histone-modification that mark promoter activity suggests that the *AluJudge* annotation can support the identification of ALU repeats that acquired a novel function, for instance, as alternative promoter.

Chapter 5 – A phylogenetic approach for germline methylation reconstruction

Often the ancestral sequence of a genome region is not available to enable a precise characterization of the influence of DNA methylation on the evolutionary process, but instead several descendent sequences are known.
For instance, as consequence of segmental duplications multiple copies of the same region are present in form of paralogous loci within a single genome. Moreover, closely related species contain orthologous loci that originated from the same genome region in the last common ancestor. This source of information is used in numerous comparative genomic applications, and especially, for the purpose of phylogeny reconstruction. Here I adapt it to reconstruct the germline methylome, using the substitution model introduced in the previous chapter.

To this end, a statistical model is introduced, which computes the likelihood that a given alignment of genome sequences evolved from any possible common ancestor sequence. Hereby, the evolutionary history of the sequences is represented by a phylogenetic tree. The model is sensitive to the methylation level the sequence was subjected to (compare Figure 1.6). By maximizing the likelihood of the model through slowly adapting the methylation level of individual branches in the phylogenetic tree, the true evolutionary history is approximated.

Then, I characterize the performance of the model in a simulation study. In a first scenario, a setting in which orthologous sequences evolve under a stable methylation level is adopted. In a second scenario, the case of paralogous sequences that underwent rapid changes of the methylation state is characterized.

This chapter concludes with a validation study in which the model was applied to promoter sequences in the primate lineage. The model performance was then assessed by methylome data for the male germline of human and chimpanzee.

5.1 Statistical framework

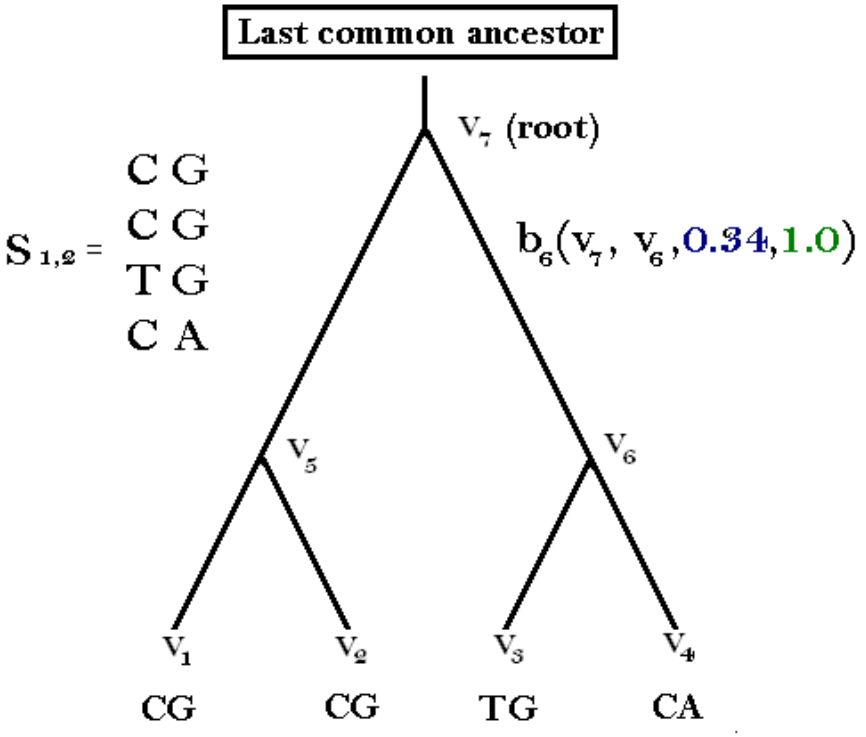


Figure 5.1: Example Phylogeny

The evolutionary history that connects the loci of the multiple alignment S to the last common ancestor is modeled by a phylogenic tree. Here, the first two columns of the alignment are assigned to the leaf nodes v_1-v_4. The branches define the topology of the tree. The branch 6 that connects node v_6 and v_7 is displayed in detail. The time interval that this branch represents is encoded by variable t, which is depicted in blue. The average methylation level of the genome region was subjected to during this time interval is displayed in green. It can take any value between 0, i.e. completely unmethylated, or 1, i.e. completely methylated. This section describes a statistical model that computes for any setting of the time intervals and methylation levels the likelihood that the observed sequences evolved from any common ancestor.

Let S be a set of m aligned DNA sequences of equal length n over the nucleotide alphabet $\Sigma_A = \{A, C, G, T\}$, and $\Pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ their prior probabilities in the last common ancestor. Then, $S_{i,j}$ denotes the alignment columns from position i to j, with for instance $i=1$ and $j=2$ representing the first dinucleotide in the alignment.
The true evolutionary history of the sequences in S is approximated by the rooted phylogenetic tree $T(S,V,B)$. Here V denotes the set of tree nodes v_1 to v_{2m-1}, assigning the first m ids to the leaves, with leaf k containing the k-th row of the alignment denoted by $S_{i,j,k}$. B denotes the set of branches in the tree. Each branch of the branches b_1 to b_{2m-2} is a tuple of length four that contains the parent node $p \in V$, the child node $c \in V$, the branch length $t > 0$ and its methylation level $\lambda \in [0;1]$, such that b_z is represented by $(p_z, c_z, t_z, \lambda_z)$. The methylation level is attributed to the branches, rather than the nodes, as it affects the evolution along the branches. Therefore, its effect can only be detected over time intervals and not for time points. Figure 5.1 shows an overview of the applied notations.

To model the evolution along the branches of the phylogeny, an adapted version of the substitution model introduced in chapter 4. The *CpG decay* rate r_7 is multiplied by the methylation level λ to account for incomplete methylation. The resulting rate $r_7\lambda$ is then applied like r_7 to obtain an altered rate matrix Q_λ, were $\lambda=1$ results in Q_M and $\lambda=0$ in Q_U.

Approximations of the evolutionary history T are described by variations of branch lengths or methylation levels in B. Thus, an approximated phylogenetic tree is denoted by $T_x(S,V,B_x)$ with $x = 0$ being the starting configuration and $x = i$ being the i-th iteration in the approximation. The likelihood of S under T_x and Q is denoted by $L(S|T_x,Q)$, whereby sequence k in S corresponds to the leaf with node id k.

To compute $L(S|T_x,Q)$, I adapted Felsensteins "pruning" algorithm (Felsenstein 1981) based on its representation in (Siepel and Haussler 2004). The basic idea of this approach is to compute, for a given substitution model and tree topology, the summed likelihood of all possible evolutionary explanations on the how the sequences in S evolved from a common ancestor. The algorithm makes use of dynamic programming to solve this task efficiently.

For a leaf v_k the likelihood to represent a certain dinucleotide x is exactly 1 for $S_{i,j,k.}=x$, and 0 for the remaining sequences of length $j-i+1$.
In case of binary phylogenetic trees, the likelihood that an inner node p was dinucleotide x, is recursively computed from the likelihoods that it evolved into any possible sequence state of the two child nodes c1 and c2:

$$L(p = x \mid T_x, Q) = \sum_{y1}[L(c1 = y1 \mid T_x, Q) P_{x,y1}(t_{c1}, \lambda_{c1})] \cdot \sum_{y2}[L(c2 = y2 \mid T_x, Q) P_{x,y2}(t_{c2}, \lambda_{c2})]$$

.

By adding additional child nodes to the equation, it can be generalized for non-binary trees. For an individual column in S the likelihood of the whole tree is inferred from the root node by summing all partial dinucleotide likelihoods weighted by their prior probabilities. To obtain a total likelihood for S, the product of the dinucleotide column likelihoods is computed. In the following this algorithm is referred to as FF-algorithm.

The influence of the CpG decay effect on the aggregation of the total likelihood

During the likelihood computation, we reencounter the problem that the context dependent *CpG decay* rate propagates dependencies along the sequences in methylated DNA (compare chapter 4). I considered three strategies on how to encounter the problem:

1) Compute the rate matrix for the full sequence length
2) Compute the likelihood for each of the *n*-1 dinucleotide positions
3) Compute the likelihood for those alignment columns with high information content, namely columns that contain CpGs, TpGs or CpAs.

For long alignments, *i.e.* large *n,* strategy 1) increases the size of rate matrix *Q*, such that numerical stability of accurately computing *P* by a Taylor series is not guaranteed an computation becomes infeasible. For small *n* the information content of the observable sequences is relative low. Therefore, this strategy is of very limited applicability. Strategy 2) can be applied straightforwardly, although the obtained likelihoods for the individual columns are not independent anymore, and the result is an approximation of the true likelihood. Strategy 3) reduces the computational effort of the algorithm and reduces the dependency-bias of the likelihood computation, but introduces a new bias towards CpGs and CpG decay positions into the approach. For the pilot studies below strategy 2) is tested for its applicability. Later strategies 2) and 3) are compared in more detail.

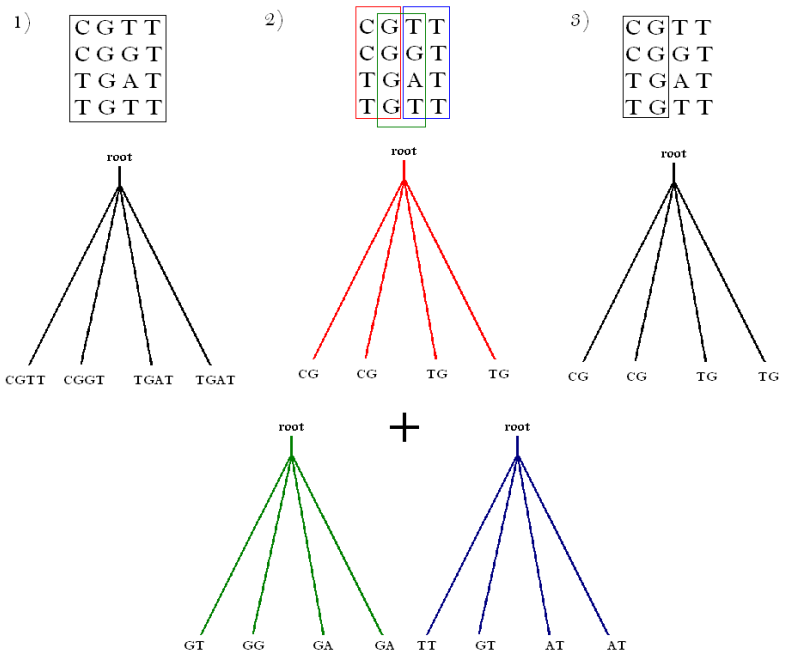


Figure 5.2: Strategies to encounter propagation of neighborhood effect
The first strategy applies rate matrices over the full length of the sequence alignment. The second strategy considers each dinucleotide independently and aggregates the resulting likelihood. The third strategy focuses on the positions that discriminates methylated and unmethylated sequences most efficiently.

Choosing a topology for the pilot study

Sequence evolution was performed using the previously described simulation engine. To determine the potential of the approach, I exclusively applied star phylogenies, in which all descendants have the same distance to the last common ancestor and no further branching occurred. Although, such phylogenies are rarely encountered in biological settings, they eliminate the influence of the tree topology on the results of the pilot study (compare Figure 5.3).

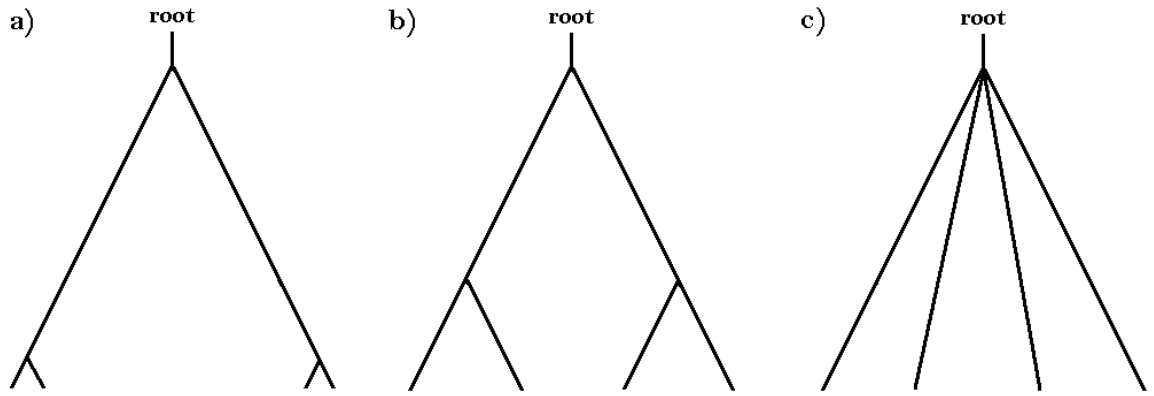


Figure 5.3: Phylogeny topologies
Here three possible topologies for phylogenies with four observable species are displayed. In topology a) the last two speciation events occurred shortly before the time point at which the sequences were observed. In this short period the chance is small that a significant amount of substitutions occured. Thus, the information content of a) is very similar to that of a star topology with two branches. Topology b) is more complex, as the difference in the distances between the last common ancestor, the second speciation event and the observable sequence influence the performance of the predictions. Finally, the topology c) is a four branches star topology, in which each branch fully contributes to the prediction.

5.2 Classification of phylogenies with uniform methylation states

This analysis is performed under the assumption that the germline methylation state of functional genome regions changes slowly over the course of evolution, implying that all branches in the phylogeny have the same methylation state. This is a reasonable assumption, as significant changes in the methylation state lead to epigenetic deregulation. In functional genome regions, like active promoters or enhancers, such changes are rarely advantageous to the individual and thus subjected to selective pressure. The following approach is most efficiently applied, if among related species orthologous functional elements share a comparable germline methylation level. The assay assesses how well the method can predict the true methylation state in this scenario depending on the distance to the last common ancestor and the number of observed loci.

To this end, the prediction accuracy of the model is measured under the assumption that all observed sequences evolved either under fully methylated or fully unmethylated conditions (Figure 5.4). Therefore, T_0^U is defined as the tree topology for which all methylation levels λ are set to 0, and T_0^M as the tree topology for which all λ are set to 1. Then, the log-likelihood ratio of $L(S|T_0^U,Q)$ over $L(S|T_0^M,Q)$ is computed and S is classified according to the sign of the result.

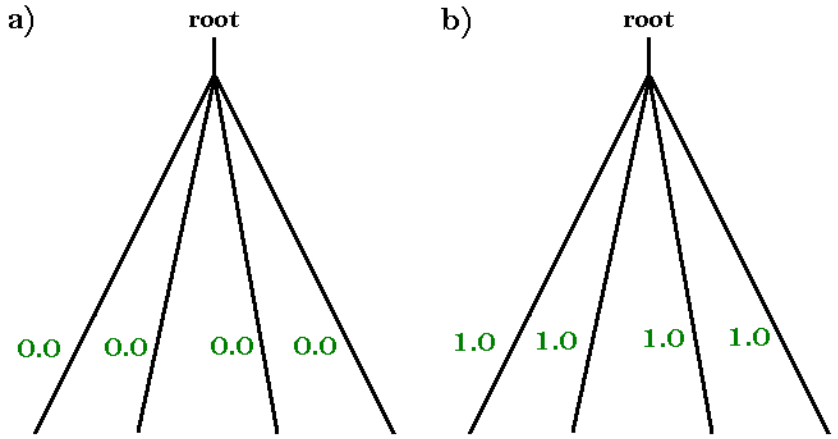


Figure 5.4: Phylogenies with uniform methylation states
For all branches the methylation label λ was set either to 0 or 1. Simulations were performed for phylogenies that differed by the length of their branches (parameter t).

The best performance of the model was reached for evolutionary distances that were comparable to those observed for the ancestral sequence approach in chapter 4. Furthermore, up to a count of three each additionally observed orthologous locus considerably improved the prediction performance (Figure 5.5).

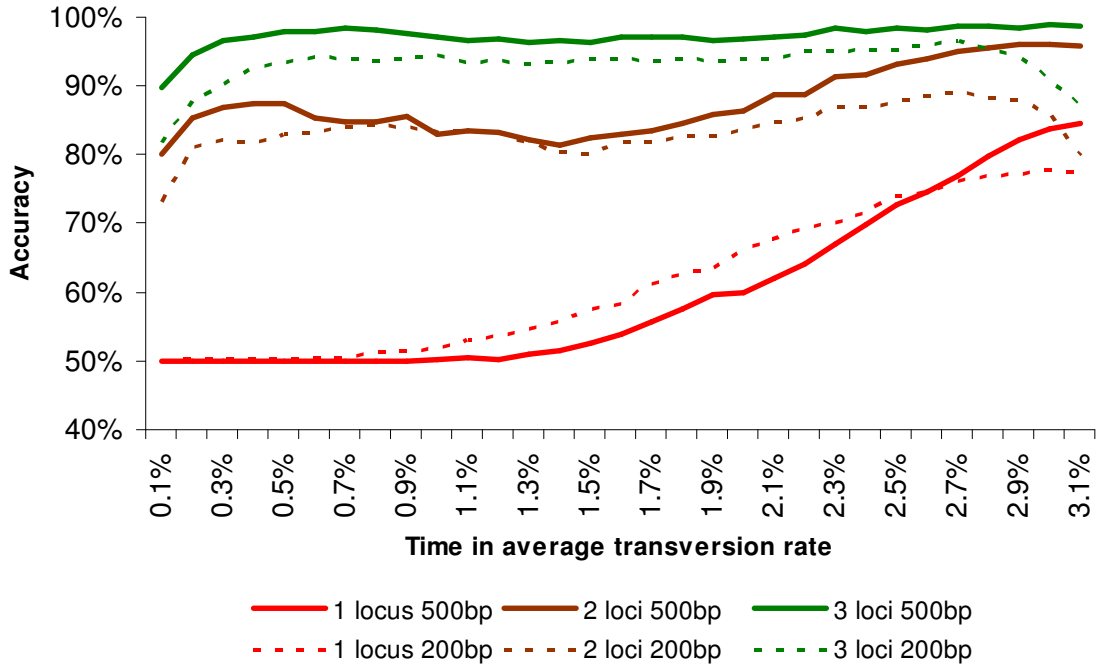


Figure 5.5: Prediction accuracy correlates with number and length of loci

Uniformly distributed sequences of length 200 bp and 500 bp are evolved according to a one branch, two branches or three branches star phylogeny. This simulation was performed 500 times under methylated and 500 times under unmethylated substitution constrains. The FF-algorithm is applied to predict the methylation state and the prediction accuracy is displayed.

It is important to note that additional information in form of an extra locus improved the prediction accuracy stronger than an increased locus length. For example a prediction based on three loci of length 200 bp, which is equivalent to 600 bp of total sequence information, had a consistently higher accuracy than one based on two loci with length 500 bp that together cover 1 kbp. This observation holds until sequences reach an age of 2.7% average transversion rate. At this point the prediction accuracy drops for all simulation runs that were based on 200 bp long loci. For sequences of length 500 bp no significant drop was detected in the simulated time frame.
The results of the one branch case need a separate discussion. The amount of information available in this special case is comparable to that applied in the *EqiScore* approach of chapter 3, with the difference that an estimate for the age parameter t is available. Hence, these classifications are implicitly performed based on closeness to the particular equilibrium distribution.

The initial simulation indicated that sequences with length 500 bp are sufficient for robust predictions. To validate this assumption, I extended the simulation to time intervals with 4.1 % average transversion rate and 5 branch phylogenies (Figure 5.6).

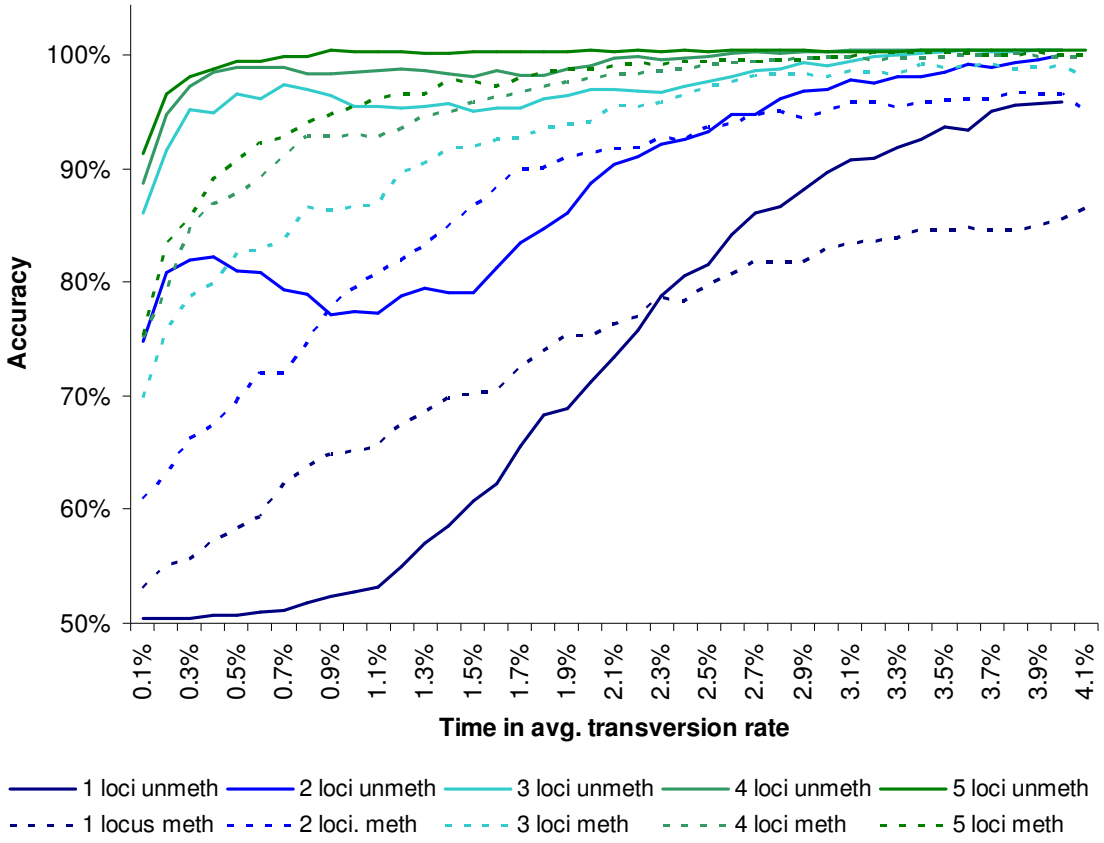


Figure 5.6: Prediction accuracy depends on nucleotide distribution and loci number
Sequences with length 500 bp with nucleotide distribution according to the methylated and unmethylated sequence equilibrium are evolved according to star phylogenies with one to five branches.

These results show that the locus length of 500 bp is sufficient. A lasting reduction in accuracy can be earliest detected beyond the 4 % time point. As explained before, the fact that each nucleotide can undergo two types of transversions, this accounts for approximately 8% observable transversions. A local drop in accuracy of the two, three and four loci phylogenies roughly between 0.5% to 1.5% time units is observable. This effect was also observed in the first simulation. It is caused by a local loss of sensitivity, *i.e.* the ability to detect unmethylated branches, whereas specificity reaches its maximum in this range. This phenomenon will be discussed in more detail in section 5.3.2-5.3.4.

5.2.1 Summary

The simulations for phylogenies with uniform methylation state show that the approach correctly predicts the methylation state within a certain time interval with high accuracy. The number of observable sequences m, as well as the length of these sequences n and the length of the branches t influence the prediction performance. While an increase of m has a greater impact then an increase of n, most parameters positively correlate with the performance. In contrast, the parameter t has an optimal range in which prediction performs best, as previously observed in chapter 4.
It is noteworthy that this rage is strongly influenced by m and n. Hence, by increasing the length and number of sequences in the alignment, the time period in which the model produces reliable predictions is extended.

Overall the model's performance is excellent whenever sequences in three or more species are aligned.

5.3 Classification of phylogenies with mixed methylation states

In some scenarios the assumption that all loci evolve under the same methylation regime is less realistic. Pseudogenes for example represent copies of functional genes that putatively lost or reduced their potential to contribute to cell function. In consequence, the selective pressure on their promoter is reduced, and thus, these evolve neutrally. Furthermore, the duplicated gene is relocated in the genome and possibly inserted into heterochromatic genome domains *i.e.* into a genomic environment that is unfavorable for maintaining an unmethylated state. For these regions changes in the methylation state can occur with a much higher probability than for orthologous functional elements. In contrast, other genes duplicated by retroposition maintained their transcriptional activity and evolved into *bona fide* genes (Vinckenbosch, Dupanloup et al. 2006). In these cases a better conservation of the epigenetic promoter regulation is expected.
To address this scenario and discriminate the functional gene copies from the degenerated ones, different methylation labels are allowed for each individual branch of the phylogeny. Thus, a multiclass classification problem has to be solved. For binary methylation states *i.e.* λ=0 or λ=1, a phylogeny with x branches results in up to 2^x possible configurations, of which the one with the highest likelihood is predicted to be the true one.

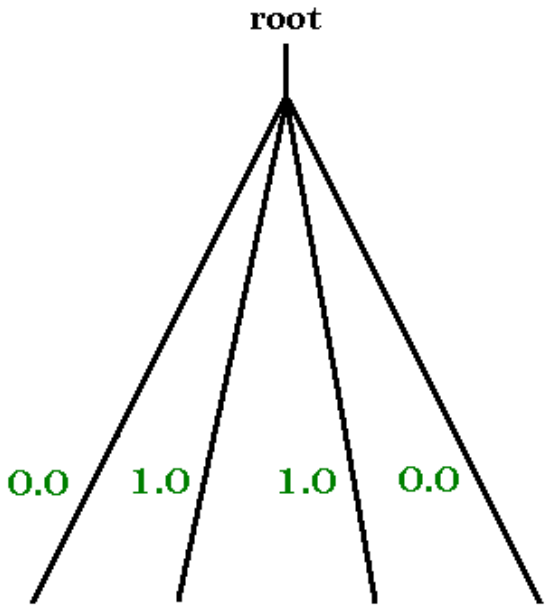


Figure 5.7: Phylogenies with mixed methylation states

In this approach sequences were generated as in the uniform methylation state approach, but predictions were performed on each possible permutation of the labels. The figure shows the 0110 permutation of a four branch star phylogeny, were the true permutation is either 0000 or 1111.

5.3.1 Strategies to determine the methylation state of the branches

For the pilot study, I addressed the multiclass classification problem by an *exhaustive* reconstruction that enumerates all possible label permutations and computes the corresponding likelihoods. Then, the different permutations were ranked according to their likelihoods. Next, the rank of the true permutation is determined and the Hamming distance of the methylation levels of the top ranking topology to the true topology is computed. The performance of the method was then characterized in different setups. As this approach is only feasible for a limited number of sequences and discrete methylation levels, I considered further methods. To this end, I adapted standard text book approaches (Hastie, Tibshirani et al. 2001).

In the *Monte Carlo* reconstruction approach the methylation labels of all branches are initialized with either a specific methylation level (all methylated, all unmethylated) or randomly. In each iteration, a randomly selected branch label is switched, *i.e.* a methylated branch is altered to unmethylatated and *vice versa*. If this increases the overall likelihood ($L(S|T_{i+1},Q) > L(S|T_i,Q)$) of the evolutionary history the change is accepted, otherwise the original configuration is restored ($L(S|T_{i+1},Q) < L(S|T_i,Q) \rightarrow L(S|T_{i+1},Q) := L(S|T_i,Q)$). This procedure is iterated until convergence of the likelihood.

A more refined version is the *Monte Carlo* approach with *simulated annealing*, which allows to escape local minima. Hereby, also changes that decrease the likelihood are accepted if the ratio $L(S|T_i,Q)/L(S|T_{i+1},Q)$ is smaller than an annealing threshold Δ_A. The annealing threshold is lowered after each iteration until it reaches 1. Then, the strategy basically is switched to the standard *Monte Carlo* approach.

5.3.2 Model performance

In the simulation study of the previous sections, simulation results were true or false, thus the accuracy of the prediction was easily determined. To adapt the performance evaluation for the *exhaustive* classifier, the number of disagreeing methylation labels per phylogeny is counted. This is basically the Hamming distance of the true methylation label permutation and the predicted methylation label permutation. This value is then aggregated over 100 repetitions of the simulation into a score and divided by the highest possible score (*m* times 100). Again, I applied star topologies and analyze the influence of the number of branches on the prediction performance (Figure 5.8). For a comprehensive assessment only the extreme cases are simulated in which all branches are methylated or unmethylated, while the model explores all permutations.

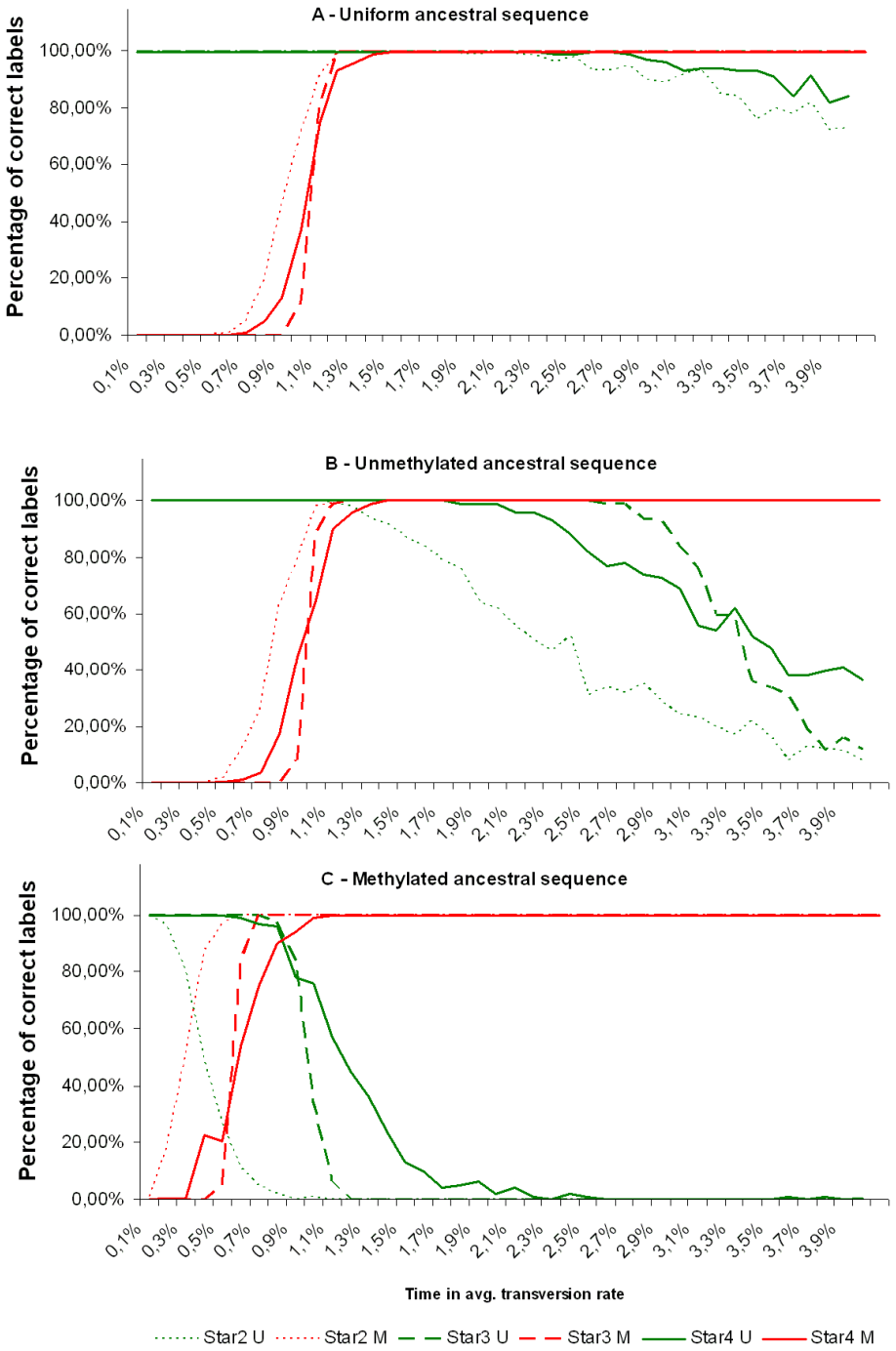


Figure 5.8: **Performance of mixed methylation level prediction**
For sequences of 1 kbp length drawn from three different sequence distributions (panel A-C) the prediction performance on phylogenies with star topology and two to four branches is displayed. Prediction performance is measured by the number of mislabeled branches divided by the number of branches in the topology. Green lines show the average over 100 simulated evolutions under the unmethylated (permutation 0000) and red lines under the methylated (permutation 1111) models.

The results in Figure 5.8 show that the performance of the model depends strongly on the sequence composition of the original sequence. For uniformly distributed sequences (panel A) already for tree stumps (m = 2) most of the branches are correctly predicted in the age interval from 1% to 4% average transversion. For sequences drawn from the unmethylated equilibrium (panel B) the performance peaks for those 2 branch trees between 1.0% and 1.5% average transversion. It considerably improves for each added branch. For CpG-poor sequences from the methylated equilibrium (panel C) the performance is worst and only for the 3 and 4 branch models good results are achieved in a very narrow age window between 0.8% and 1% average transversion rate.
It can also be observed that briefly after the simulated speciation, and thus at small average transversion rates, the model over predicts the unmethylated state in all scenarios (*fade-in bias*). With increasing age of the independent sequences this trend is inverted and all sequences are predicted to be methylated (*fade-out bias*). The onset of these trends depends on the original sequence composition and the number of branches in the phylogeny. Both parameters influence the window in which this bias is cancelled by the strength of the true signal. As these biases were already encountered in chapter 4 and in the uniform methylation setup, I will briefly discuss their putative causes.

5.3.3 Fade-in bias

The explanation for the *fade-in bias* is most likely the discreet nature of the *CpG decay* event. On limited data (short sequences, few copies and low-CpG content) the required time interval until the number of expected CpG decays approaches 1 is relatively large. In consequence, in most simulation instances early on no CpG decay is observed. As our model is continuous, this supports the likelihood that a single branch or the whole phylogeny is unmethylated. Concretely, if we expect to observe ~0.01 *CpG decays* under M_U and ~0.2 events under M_M, but actually observe 0 events, the unmethylated model will always be favored. By increasing the number of CpGs per branch, either through elongation of the sequence, increasing the CpG density, or by adding additional branches, the expected time until the first *CpG decay* is shortened (and thus also to the second, third or *n*-th event).

Therefore, t_α as the time point at which the number of expected CpG decay events equals 1 can function as characteristic value to evaluate the lower bound at which distance between the species the model can produce reliable results for a particular phylogeny. It is determined by $1 = \#CpGs \cdot (P_{CpG \to CpA}(t_\alpha) + P_{CpG \to TpG}(t_\alpha))$, where P is computed under M_M. Hereby, #CpGs is the number of CpGs under evolution and usually unknown, but can be estimated from the observable sequences.

This equation simplifies under the assumption of symmetry to $1 = \#CpGs \cdot (P_{CpG \to CpA/TpG}(t_\alpha))$, with $2 \cdot P_{CpG \to CpA}(t_\alpha) = 2 \cdot P_{CpG \to TpG}(t_\alpha) = P_{CpG \to CpA/TpG}(t_\alpha)$. The value of t_α can be approximated by numerical approaches, and furthermore, the derivation of an analytical solution appears possible.

5.3.4 Fade-out bias

The explanation for the *fade-out bias* lies in the limited number of CpG sites. During methylation-free evolution the M_U model derives the major part of its additional likelihood mass from the number of CpG sites that did not undergo *CpG decay*. At a certain age most of the original CpG sites are affected in at least one branch by any disruptive mutation. Especially for 2 branch models this decreases the distance to the likelihood under the M_M model. On the other hand, each conserved TpG and CpA position can also be explained by multiple *CpG decay* events, and thus, improves the likelihood of the M_M model stronger than that of the M_U model. This also applies to positions at which by chance TpG/CpA pairings are formed by random mutations at corresponding positions in the phylogeny. The probability that TpG/TpG, CpA/CpA or TpG/CpA pairs are formed after the sequences became largely independent from their common ancestor is larger than the probability that a CpG/CpG pair evolves. Hence, over long time spans M_M will always dominate M_U. This especially applies for genome regions that are naturally rich in TpGs and CpAs such as sequences close to the methylated equilibrium. An increase in the number of branches can counteract this trend for a while as the time span in which the common ancestor is reconstructed robustly is prolonged. These analytical considerations are well supported by the simulation results.

Fortunately, in biological scenarios the question weather a locus as recently lost its unmethylated status is asked more frequently than the reverse case, such that the natural particular bad performance of the model in CpG-poor, but CpA/TpG-rich sequences has little practical relevance.

5.4 Comparison of uniform and mixed methylation labels

The *uniform* methylation labels have the advantage that only two different tree configurations have to be tested. This can be done rather efficiently, but the assumption of stable methylation states is rather strong and biases the results if it is violated. The *mixed* methylation labels require the computation of likelihoods for a larger number of branch label permutations. This is feasible for small trees, but introduces a higher variance into the model. Both models show good results within a particular time interval of the distance to the last common ancestor that is defined by the number of species and CpGs in the ancestral sequence.

Moreover, this time interval is influenced by two kinds of biases, the *fade-in* and the *fade-out bias*. Both biases are not a consequence of the model, but of the data. By increasing the amount of informative observations *i.e.* by including more CpGs into the alignment, either via longer sequence, sequence of higher CpG density or the addition of more sequences, theses biases is reduced.

5.4.1 On the necessity of regularization of methylation level changes

In this chapter two ways to model the methylation levels within the evolutionary history T were introduced. In the *uniform* model all branches have the same methylation state, and in the *mixed* model switching of the methylation state was possible at every branch.
Both approaches represent extremes. In a biological scenario changes are possible, but rare. If we consider a phylogeny with many short branches, it would be unlikely to observe an alternating series of methylation labels along the branches, although this may be the explanation that maximizes the likelihood. The introduction of a regularization term ξ can solve this problem. ξ penalizes changes of the methylation level at internal nodes by reducing the likelihood based on the difference between the methylation label to the parent node and to the child node. Given a node v with a parental branch (p,v,t_p,λ_p) and a child branch (v,c,t_c,λ_c) the penalty factor φ is then computed as:

$$\varphi(\lambda_p, \lambda_c, \xi) = \frac{1}{1 + \xi(|\, \lambda_p - \lambda_c \,|)}$$

The likelihood computation for a parent child pair is updated by:

$$L(v = x \mid T_x, Q) = \varphi(\lambda_p, \lambda_c, \xi) \sum\nolimits_y [L(c = y \mid T_x, Q) P_{x,y}(t_c, \lambda_c)]$$

In this scenario the *uniform* approach is equivalent to a ξ that is set to infinity, while in the *mixed* classification approach ξ is set to zero. By scaling ξ to an appropriate value between both extremes the theoretical advantages and drawbacks of both methods can be traded against each other. On the one hand, in the case of short locus length, the uniform classification is robust with respect to statistical outliers at individual leaves. On the other hand, it is heavily biased if the underlying assumption does not hold.
In contrast, the mixed classification is able to identify changes of methylation levels during evolution, but increases variance of the prediction, as already small changes in one of the sequences may tip the scale for the label prediction of individual branches.
An intermediate value for ξ appears appropriate to balance bias against variance. For now this feature is not implemented into the algorithm, as the prior knowledge on the variability of methylation states is very limited.

5.5 The FFK-algorithm

This study was conducted in cooperation with Sandra Koser, who improved the implementation of the algorithm, and conducted the computations under my supervision.

For the proof-of-principle study of the last section the *exhaustive* approach proved to be sufficiently efficient to sample the full search space. For phylogenies with higher numbers of branches this soon becomes infeasible. Similarly, the *exhaustive* approach is hindering the extension of the model to a more continuous choice of λ.

In both cases the *Monte carlo*-based strategies efficiently overcome the computational bottleneck. As both variants of the *Monte carlo*-based approach require the repeated execution of the FF-algorithm, it was optimized by Sandra Koser under my supervision in the framework of a master thesis (Koser 2012). The resulting FFK-algorithm optimized the likelihood computation by reducing the number of tree-traversals for strategy 2) from n-1 to 1. Furthermore, it implements both *Monte carlo*-based approaches. A detailed characterization of the FFK-algorithm confirmed that the results obtained from the pilot study on the FF-algorithm, generalizes to the extended version of the algorithm (Koser 2012). In particular, the influence on alignment length, number of observed species and CpG content in the last common ancestor sequence on the time-interval of optimal performance was confirmed and characterized in detail.

To validate the applicability of the FFK-algorithm on a biological dataset, I designed a study based on homologous promoter sequences of primates. This data was complemented by two publically available methylome datasets for the human and chimpanzee male germline represented my sperm tissue.

5.5.1 Generalization to continuous methylation levels

For simplicity the FF-algorithm focused on binary methylation states, *i.e.* fully methylated or fully unmethylated DNA sequences. The FFK algorithm extended the approach to continuous methylation levels. Therefore, the rate matrix Q_λ is introduced in which the *CpG decay*-describing rate is scaled by a tuning parameter λ (Hobolth 2008). For λ=1.0 we have Q_λ=Q_M and for λ=0 the equality of Q_λ=Q_U is established. The framework is defined such a way that the transition from the binary to the continuous case could be performed by minor adaptation of the software implementation. For reasons of computational efficiency, the interval of possible methylation values split into discrete values ranging from 0% to 100% in 10% steps.

5.5.2 Validation study design

To demonstrate the applicability of the FFK-algorithm in a biological meaningful context, we derived a set of genes that is conserved between mouse and human(Iwama and Gojobori 2004), and has been applied before in context of a comparative epigenomics study (Jiang, Han et al. 2007; Feuerbach, Halachev et al. 2012). The gene coordinates for the human genome assembly hg18 were then derived via the BioMart resource (Kasprzyk 2011). The coordinates were then mapped to the chimpanzee (panTro2), orangutan (panAbe2), rhesus macaque (rheMac2) and marmoset (calJac3) genomes via the liftover resource of the Galaxy tool (Giardine, Riemer et al. 2005; Blankenberg, Kuster et al. 2010). Finally, multiple DNA sequence alignments for the resulting set of 556 promoters, which are conserved across several primate species, were derived by ClustalW V2.0.12 applying the standard alignment parameters (Thompson, Higgins et al. 1994). Additionally, two previously published methylome dataset of the male germline of humans and chimpanzee were acquired (Molaro, Hodges et al. 2011). In contrast to the simulation study, these multiple alignments contained gaps. As input for the FFK algorithm, only alignment columns without gaps were considered in the computation (Figure 5.9).

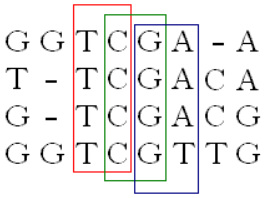


Figure 5.9: Treatment of gaps in the multiple alignment
Only dinucleotide positions without gaps were included into the computation.

In a first assay we applied the FFK algorithm to the multiple alignment of each promoter considering every gap-free dinucleotide column for the likelihood computation (strategy 2). The methylation label of the branch to the human species was then compared with the average methylation level in human sperm. The resulting predictions for each promoter are reported in Appendix J.

5.5.3 Training of mutation rates

To calibrate the applied mutation rates to primate promoters, the methylome data was used to compute the mean methylation level of each homolog promoter in human and chimpanzee. This value was applied to set fixed values of λ in the human-chimpanzee phylogeny. Then, the pairwise alignments of all promoters were used to set the leaf nodes. Then, a variant of the FFK-algorithm was applied in which the likelihood computation was applied to assess r_5, r_6 and r_7 instead of the methylation levels. The rates for transversions were preset to one. The resulting rates for promoter sequences are: $\frac{r_{tr}}{r_{tr}} = 1$, $\frac{r_5}{r_{tr}} = 8.8$, $\frac{r_6}{r_{tr}} = 7.7$, $\frac{r_7}{r_{tr}} = 41.8$, thus showing a slightly less pronounced CpG decay effect.

5.5.4 Results of validation study

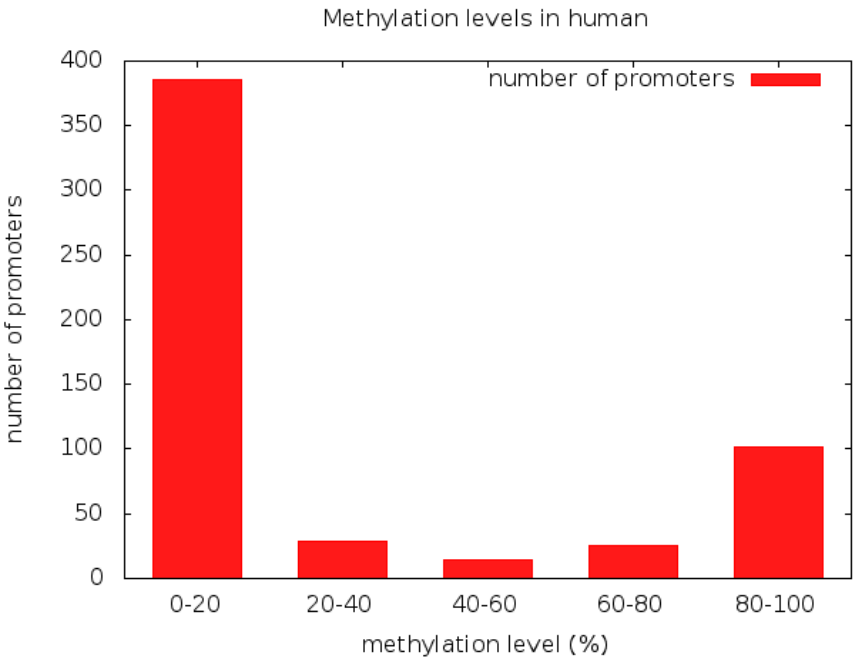


Figure 5.10: Methylation state of human promoters
The number of promoters that fall into each of the five depicted ranges of methylation levels are displayed as histogram.

The methylation levels of the promoters are largely bimodal (Figure 5.10). Therefore, we defined a divergence of 30% from the true methylation level as sufficient to correctly predict the methylation mode of a promoter.

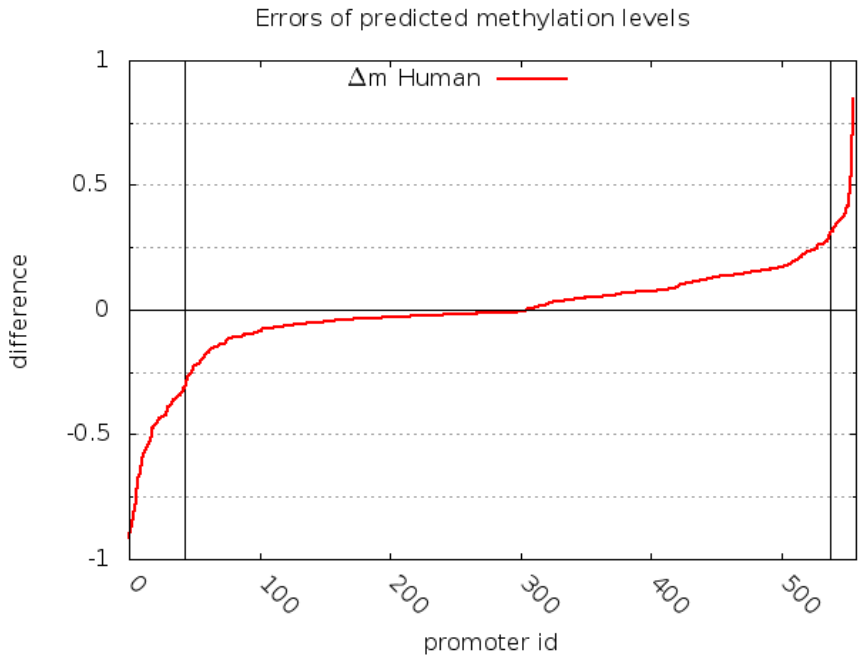


Figure 5.11 Difference between predicted and measured promoter methylation
The difference between the predicted and measured promoter methylation level is displayed for the human promoters. The vertical lines mark where the absolute differenced is higher than 30 %.

For 496 (89%) promoters the difference between predicted and observed value (Δ_m) was smaller than 30% (Figure 5.11). Next, we tested if the CpG positions and the products of *CpG decay* (CpA and TpG) alone are sufficient for a reliable prediction *i.e.* we tested strategy 3). Therefore, we computed the predictions once on the full sequence and once only on positions that contained in a least one of the sequences a CpG, TpG or CpA. Both predictions yield similar results, although the runtime of the second approach was significantly reduced (Table 5.1).

Alignment length (bp)	**Runtime (min)**	
	All positions	CpG, TpG and CpA
1000	22.2	3.9
2000	42.6	7.7
3000	60.7	13.1
5000	79.0	22.8

Table 5.1: **Runtime of FFK for all position vs. CpG, TpG and CpA positions**

5.6 Discussion

The presented simulation results show that the footprint of *CpG decay* in the human genome is strong enough to enable a local reconstruction of the germline methylome. Furthermore, the influence of parameters such as the length of the homologous locus, the initial sequence composition, the number of observable descendant sequences and the distance to the last common ancestor on the prediction quality was quantified.

The first of two major findings is that the number of observable descendant sequences has a higher priority than the locus length for an improvement of the classification. This indicates that the method can be well applied for the study of short sequences such as gene promoters (1-2 kbp length), if enough homologous regions are considered, while the method is inadequate for more extended areas that underwent a single segmental duplication event or are only observed in two species.

The second major finding is the limited performance of the approach for regions that were close to the methylated equilibrium in the last common ancestor. Here I identified a systematic bias in favor of predicting all descendants as being methylated. Hence, I propose to apply the method primarily to decide if a CpG-rich ancestral sequence gained DNA methylation in the germline or maintained an unmethylated state.

Finally, the model performance strongly depends on the evolutionary distances between the considered genomes. The range in which optimal predictions are possible is constraint by the fade-in and the fade-out bias. Both are influenced by the number of CpGs in the ancestral sequence and the number of sequences in the alignment. Before the model is applied in a specific scenario, the tools introduced in this chapter can be applied to estimate the reliability of the produced predictions.

Chapter 6 - Conclusions and perspectives

Evolution as a concept is one of the corner stones of modern science. In this thesis I traced its influence onto the interdependent development of genomes and epigenomes over long time spans. To this end, I developed an array of DNA sequence-based methods and characterized their applicability. Methods that directly proved to be of practical relevance were then applied to improve the annotation of the human genome. Specifically, the *CGI Mountain* annotation and the *AluJudge* track were generated. Moreover, for a number of additional approaches such as the *EqiScore* or the FFK-algorithm proof-of-principle and validation studies were performed. Furthermore, steps for their further development were described.
Additionally, the developed tools were applied to derive general insights into the influence of DNA methylation on the evolution of regulatory sequences.

6.1 A toolbox for evolutionary epigenomics

On a qualitative level, the CpG dinucleotide was previously known to be one of the interfaces at which the evolutionary interplay of genome and epigenome most prominently manifests. Four of the here proposed novel methodologies, namely *CgiHunter*, *EqiScore*, the *L-score* based *AluJudge* annotation and the FFK-algorithm are systematic approaches to characterize this interaction on a quantitative level. The presented simulation studies demonstrated the theoretical advantages of all methods. Furthermore, their validation on independent epigenetic datasets demonstrated their practical applicability. The here assembled evolutionary epigenomics toolkit is highly complementary.

We may consider the scenario of a novel whole-genome sequence for a largely uncharacterized vertebrate species. For *CgiHunter* this genomic sequence as input is sufficient. By using a parameter grid, a number of *CGI Shadow* annotations can be produced and either analyzed individually or in the merged form of a *CGI Mountain* annotation. Like a topographic map, this will yield a number of insights into the spatial distribution of CpG-rich regions, their length and the density of their cores.

If an empirical choice of the parameter grid is to be circumvented additional information in form of nucleotide substitution models can be applied. Already from estimations of four substitution rates, *EqiScore* infers its remaining model parameters by numerical simulation and apply these dinucleotide equilibrium distributions to identity regions with well conserved DNA methylation levels. Especially, for genome regions, which stably evolved under a particular methylation level, this method is very effective.

CgiHunter and *EqiScore* both implicitly assume equal age for all input sequences and are thus biased in case of younger genome elements such as the primate-specific CpG-rich transposable elements from the ALU family. *AluJudge* and its statistical model corrects for this source of bias, by quantifying the intensity of the *CpG decay* effect each repeat instance is subjected to. Thus, the germline methylation state of each repeat instance is inferred.

Finally, the genome-wide applicable FFK-algorithm combines the advantages of *EqiScore* and *L-Score* for cases in which multiple sequence alignments for species of fitting evolutionary distances are available. It is most effective to decide if genome regions recently became methylated in the germline, and thus complements the timescale at which the equilibrium-based *EqiScore* is effective.

The here presented results from the benchmarks and simulation studies are, likewise a calibration curve, a valuable resource to select the appropriate method from this toolbox to answer a specific research questions.

6.2 CpG islands buffer selective pressure on CpG-rich binding-sites

Another insight of this thesis is that selective pressure on the DNA is not exclusively directed on specific nucleotide sequences or binding-motifs. It has been proposed earlier that two subtypes of TFBS exist: 'highly selected sites that rarely occur by chance and auxiliary sites that are available by convenience'(Wasserman and Sandelin 2004). The authors furthermore argue that TFs such as Sp1 are associated to the second subtype. The equilibrium-based analysis in chapter 3 showed that protecting specifically the dinucleotide CpG in small fraction of the genome from methylation-mediated decay has a strong influence on the likelihood that such TFBS are formed and maintained.

The development of this mechanism appears very beneficial as otherwise the genome is encountered with two unfavorable extremes. Its *cis*-regulatory sequences may be very complex, *i.e*. relative long and very specific in the nucleotide composition. This results in a small likelihood for their random formation, but is requiring a strong selective pressure against the smallest change in these sequences. Furthermore, a TFBS that is specific enough for a small genome may be too unspecific for a large genome. Thus, with increasing genome size the DNA binding domains of TFs would be subjected to strong pressure to co-evolve.
Alternatively, low complexity motifs are easier to conserve, but in the vast mammalian genomes they can frequently form by random mutations at unfavorable positions. The suppression of these low complexity motifs requires strong selective pressure, or else result in uncontrolled transcription, which wastes resources, may lead to deregulation of gene expression and reduces the concentration of free transcription factors.

Genome-wide DNA methylation presents a third option. It enabled the formation of protected methylation-free pockets in which low-complexity CpG containing motifs such as the GC-box are maintained with high probability. These motifs are frequent enough that several are detectable within each CGI. This leads to the characteristic multiple TSS sites per CGI promoter, provides a sufficient robustness against random mutations and presents a potential mechanism for calibrating the exact transcription level via mutation and selection of individual binding sites.
Outside of these pockets these CpG-rich motifs are disabled immediately by DNA methylation and, speaking in evolutionary time scales, rapidly erode via *CpG decay*. This mechanism is functionally similar to the suppression of transposable elements by DNA methylation, which takes also placed in species without genome-wide DNA methylation.

In consequence, the neutral mutation rate is calibrated by the local DNA methylation level, and selective pressure is imposed to conserve a beneficial methylation distribution in the germline *i.e.* the formation and maintenance of unmethylated pockets that result in CGI formation, rather then exclusively acting on individual *cis*-regulatory sequences. This conclusion is supported by recent comparative studies on primate *CGI* promoters that found no evidence for selective pressure on individual CpGs (Cohen, Kenigsberg et al. 2011).

With the unbiased bottom-up approach *CgiHunter* as well as the evolutionary motivated top-down attempt *EqiScore*, this thesis improved the annotation of such regions with elevated CpG content, and demonstrated that the intensity, with which they withstand *CpG decay*, correlates with their regulatory activity in terms of histone marks and transcription initialization via polymerase occupancy (compare chapter 2 and 3).
A key problem of the field is that such observed correlations make no statement about the direction of the causality. More precise, two possible scenarios are possible. First, certain regulatory sequences contain more CpG dinucleotides because they were preferentially located in unmethylated domains. Second, the maintenance of unmethylated domains during the spread of DNA methylation to a genome-wide phenomenon was necessary to protect specific regulatory sequences from *CpG decay*.

Complementarily, TFs that bind CpG-poor motifs such as the TATA-box exist and have been available as building blocks for the evolutionary processes during the creation of higher-vertebrate and mammalian genomes. But in those species that gained genome-wide DNA methylation, promoter regions gained CpG dinucleotides instead of relying stronger on CpG-less elements (Khuu, Sandor et al. 2007). Moreover, the tetranucleotides that predict promoter function of *CGIs* by co-location with RNA polymerase binding sites were rich in CpGs (compare chapter 3). This indicates that *CpG decay* is not an inherent problem, but a feature of the genetic-epigenetic-evolutionary system.

CpG decay fosters the concentration of functional, CpG-rich, low-complexity motifs at specific locations that remain unmethylated in the germline. In the remaining genome, global DNA methylation actively suppresses the formation of sites that promote transcription at random positions. With the equilibriums based method developed in this thesis, we have now the ability to quantify this effect. For instance for the binding motives of *Sp1* and *Egr1* it accounts for a 6-fold and 42-fold overrepresentation in unmethylated regions.
In a 3 billion nucleotide long genome this already reduces the number of randomly form sites significantly. If nonetheless a useless or harmful *cis*-regulatory site is formed or inserted by transposition, it is neutralized by DNA methylation as the binding affinity of many *trans*-factor, such as TFs, is reduced by this covalent modification. Conveniently enough this system also leads to an accelerated erosion of such a transposed site. For the repeats of the ALU family the *AluJudge* annotation confirms previous reports that this process affects the majority of the repeat copies.

6.3 The special role of CpG island edges and weak CpG islands

The study of regions with intermediate CpG content indicates that these areas also display intermediate enrichment of epigenetic signals. These regions are a mixture of weaker CGIs and the edges of strong CGIs. According to our understanding of the *CpG decay* effect this implies that they are less frequently methylated than the remaining genome, but more frequently than strong CGIs. As they are still epigenetically active and partially unmethylated, a too recent transposition to detect footprints of CpG decay, can not explain these observations. Furthermore, selective pressure on individual sequence motifs that counter acts the CpG decay effect could not be detected (Cohen, Kenigsberg et al. 2011). This leaves difference in the temporal distribution of the methylation level as a possible explanation.

Three concepts are plausible. First, the methylation of an individual CpG has a stochastic component with a high probability of methylation outside of CGIs which gradually reduces towards the core of strong CGIs. With probabilities in-between the extremes of full methylation and complete absence of methylation the chance of *CpG decay* is also moderate.
Second, recent evolutionary changes in the size of the CpG islands may have triggered the gain or loss of CpGs at their edges (Matsuo, Clay et al. 1993).
Third, the degree of methylation differs among the tissues of the germline (Figure 1.1). Thus, not only the time of elevated probability to undergo spontaneous deamination is variable, but also the error-prone reprogramming of methylation marks occur with varying site-specific frequencies. The difference between methylation levels in sperm and somatic tissue indicates that such a scenario realistic (Molaro, Hodges et al. 2011). Apparently, unmethylated domains are much larger at this point in the male germline than in somatic tissue. Therefore, similar changes between the different cell types of the germline cycle and especially between the female and male ‘dialects’ of the associated epigenetic reprogramming are likely.

6.4 Perspectives

Evolutionary epigenomics research will be fueled through the rapid gain of vertebrate genome sequences (10K-Genomes-Scientists 2009). Without the direct requirement of additional epigenomic experiments, novel genomes can be analyzed for the distribution of regions that are protected from *CpG decay* in the germline.
The next step to continue along this avenue is the application of the here presented toolbox for the characterization of other mammalian species, for which high quality DNA sequences are already available. This endeavor promises insights into the adaptation of the epigenetic regulation of these relative closely related organisms to the different environmental niches.
In parallel these methods can be used to improve the discovery of epigenetically regulated loci outside of CGI promoter regions. The non-coding areas of the human genome are devoid of clear landmarks that separate functional from non-functional domains. The here presented tools can assist in characterizing these areas, by annotating loci that were protected from *CpG decay*.
Such identification of functional genome regions is of high practical relevance for medical research. This information is crucial for the prioritization research on of differentially methylated or mutated regions in the genomes of diseased cells for in-depth analysis.
For instance, a number of complex human diseases that are associated to yet unknown heritable risk factors such undiscovered elements may play a key role. The prioritization of candidates by the conservation of their methylation-free germline state may contribute to focusing on the most relevant loci among the high number of potential targets.

The methods themselves do not represent endpoints. Beside the optimization of runtime, usability and interpretability of the results, they can be integrated more tightly with other comparative genomics approaches such as motif discovery or the annotation of transposable elements. Most importantly, the interaction between genome sequence, histone-complexes and DNA methylation ought to be addressed directly.

Finally, insights into the fascinating patterns of genomic and epigenomic co-evolution are a value in itself. The intriguing question is how small genomes of the complexity of a newspaper have evolved into the huge autonomously organized library that we call the human genome? The steeping stones this thesis provides may support others on the long road to a deep understanding of this subject.

References

10K-Genomes-Scientists (2009). "Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species." HEREDITY **100**(6): 659-674.

Adachi, N. and M. R. Lieber (2002). "Bidirectional Gene Organization: A Common Architectural Feature of the Human Genome." Cell **109**(7): 807-809.

Aimée, M. and A. Bird (2011). "CpG islands and the regulation of transcription." Genes & Development **24**: 1010-1022.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Allis, C. D., T. Jenuwein, et al. (2007). Epigenetics. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Altschul, S. F., W. Gish, et al. (1990). "Basic Local Alignment Search Tool." J. Mol. Biol. **215**: 403-410.

Antequera, F. (2003). "Structure, function and evolution of CpG island promoters." Cell Mol Life Sci **60**(8): 1647-58.

Antequera, F. and A. Bird (1999). "CpG islands as genomic footprints of promoters that are associated with replication origins." Current biology : CB **9**(17): R661-R667.

Arndt, P. F., C. B. Burge, et al. (2002). DNA sequence evolution with neighbor-dependent mutation. Sixth Annual International Conference on Computational Biology, New York.

Arndt, P. F., C. B. Burge, et al. (2003). "DNA sequence evolution with neighbor-dependent mutation." Journal of Computational Biology **10**(3-4): 313–322-313–322.

Belancio, V. P., A. M. Roy-Engel, et al. (2010). "All y'all need to know 'bout retroelements in cancer." Semin. Cancer. Biol. **20**(4): 200-210.

Bestor, T. and V. Ingram (1983). "Two DNA methyltransferases from murin erythroleukemia cells: Purification, sequence specificity, and mode of interaction with DNA." Proc. Natl Acad. Sci. USA **80**: 5559-5563.

Bhattacharyya, M., L. Feuerbach, et al. (2012). "MicroRNA transcription start site prediction with multi-objective feature selection." Stat. Appl. Genet. Mol. Biol. **11**(1).

Bird, A. (1987). "CpG islands as gene markers in the vertebrate nucleus." Trends in Genetics **3**: 342-347.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Res **8**(7): 1499-504.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Research **8**(7): 1499 -1504-1499 -1504.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**(6067): 209-13.

Bird, A. P. and M. H. Taggart (1980). "Variable patterns of total DNA and rDNA methylation in animals." Nucleic Acids Research **8**(7): 1485-1497.

Blanchette, M., W. J. Kent, et al. (2004). "Aligning Multiple Genomics Sequences With the Threaded Blockset Aligner." Genome Res. **14**(4): 708-715.
Blankenberg, D., G. V. Kuster, et al. (2010). Galaxy: A Web-Based Genome Analysis Tool for Experimentalists, John Wiley & Sons, Inc.
Bock, C. (2008). Computational Epigenetics - Bioinformatic methods for epigenome prediction, DNA methylation mapping and cancer epigenetics. Computational Biology & Applied Algorithmics. Saarbrücken, Universität des Saarlandes. **PhD**.
Bock, C., M. Paulsen, et al. (2006). "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure." PLoS Genet **2**(3): e26.
Bock, C., J. Walter, et al. (2007). "CpG island mapping by epigenome prediction." PLoS Comput Biol **3**(6): e110.
Bourc'his, D. and T. H. Bestor (2004). "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L." Nature **431**(7004): 96-99.
Britten, R. J. (1986). "Rates of DNA Sequence Evolution Differ Between Taxonomie Groups." Science **231**: 1393-1398.
Bulyk, M. L., P. L. F. Johnson, et al. (2002). "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors." Nucleic Acids Res. **30**(5): 1255-1261.
Caiafa, P. and M. Zampieri (2005). "DNA methylation and chromatin structure: the puzzling CpG islands." J Cell Biochem **94**(2): 257-65.
Carninci, P., A. Sandelin, et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-35.
Casadesus, J. and D. Low (2006). "Epigenetic Gene Regulation in the Bacterial World." Microbiology and Molecular Biology Reviews **70**(3): 830-856.
Chahwan, R., S. N. Wontakal, et al. (2010). "Crosstalk between genetic and epigenetic information through cytosine deamination." Trends in Genetics **26**(10): 443-448.
Chen, B., L. He, et al. (2000). "Inhibition of the Interferon-γ/Signal Transducers and Activators of Transcription (STAT) Pathway by Hypermethylation at a STAT-binding Site in the p21WAF1 Promoter Region." Cancer Research **60**(12): 3290 - 3298-3290 -3298.
Cohen, Netta M., E. Kenigsberg, et al. (2011). "Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection." Cell **145**(5): 773-786.
Collins, F. S., E. D. Green, et al. (2003). "A vision for the futute of genomics research." Nature **442**: 835-847.
Cuadrado, M., M. Sacristán, et al. (2001). "Species-specific organization of CpG island promoters at mammalian homologous genes." EMBO Reports **2**(7): 586-592.
D'Haeseleer, P. (2006). "What are DNA sequence motifs?" Nat Biotech **24**(4): 423-425.
Darwin, C. (1864). On the origin of species by means of natural selection. Preservation of favoured races in the struggle for life. New York, D. Appleton and company.
Das, R., N. Dimitrova, et al. (2006). "Computational prediction of methylation status in human genomic sequences." Proc Natl Acad Sci U S A **103**(28): 10713-6.
Delsuc, F., H. Brinkmann, et al. (2005). "Phylogenomics and the reconstruction of the tree of life." Nature Reviews Genetics **6**: 361-375.

Durbin, R., S. Eddy, et al. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, Cambridge University Press.
Duret, L. and P. F. Arndt (2008). "The Impact of Recombination on Nucleotide Substitutions in the Human Genome." PLoS Genet **4**(5): e1000071-e1000071.
Edwards, J., A. O'Donnell, et al. (2010). "Chromation and sequence features that define the fine and gross structure of genomic methylation patterns." Genome Research.
Elnitski, L., V. X. Jin, et al. (2006). "Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques." Genome Res. **16**: 1455-1464.
Farré D, B. N., Mularoni L, Messeguer X, Albà MM (2007). "Housekeeping genes tend to show reduced upstream conservation." Genome Biology **8**(R 140).
Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J. Mol Evol **17**(6): 368-76.
Feltus, F. A., E. K. Lee, et al. (2003). "Predicting aberrant CpG island methylation." Proc Natl Acad Sci U S A **100**(21): 12253-8.
Feltus, F. A., E. K. Lee, et al. (2006). "DNA motifs associated with aberrant CpG island methylation." Genomics **87**(5): 572-9.
Feuerbach, L. (2007). Towards Comparative Epigenomics: A Pilot Study on DNA Methylation of CpG Islands on Human Chromosome 21 (Master Thesis). Mathematics and Informatics. Saarbrücken, Universität des Saarlandes.
Feuerbach, L., K. Halachev, et al. (2012). Analyzing epigenome datra in context of genome evolution and human diseases. Evolutionary Genomics. M. Anisimova. New York, Springer**:** 431-468.
Feuerbach, L., R. B. Lyngsø, et al. (2011). "Reconstructing the Ancestral Germ Line Methylation State of Young Repeats." Molecular Biology and Evolution **28**(6): 1777-1784.
Freitag, C. M. (2006). "The genetics of autistic disorders and its clinical relevance: a review of the literature." Molecular Psychiatry **12**: 2-22.
Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.
Giardine, B., C. Riemer, et al. (2005). "Galaxy: a platform for interactive large-scale genome analysis." Genome Res **15**(10): 1451-5.
Glass, J. L., R. F. Thompson, et al. (2007). "CG dinucleotide clustering is a species-specific property of the genome." Nucleic Acids Res **35**(20): 6798-807.
Goll, M. G. and T. H. Bestor (2005). "Eukaryotic cytosine methyltransferases." Biochemistry **74**: 481-514.
Gross, D. S. (1988). "Nuclease Hypersensitivity Sites in Chromatin." Ann. Rev. Biochem. **57**: 159-197.
Guo, J. U., Y. Su, et al. (2011). "Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain." Cell **145**(3): 423-434.
Hackenberg, M., C. Previti, et al. (2006). "CpGcluster: a distance-based algorithm for CpG-island detection." BMC Bioinformatics **7**: 446.
Halachev, K., H. Bast, et al. (2012). "EpiExplorer: live exploration and global analysis of large epigenomic datasets." Genome Biology **13**(R96).
Han, L. and Z. Zhao (2009). "CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome?" BMC Bioinformatics **10**: 65.

Hannenhalli, S. (2008). "Eukaryotic transcription factor binding sites - modeling and integrative search." Bioinformatics **24**(11): 1325-1331.
Hardison, R. C. (2003). "Comparative Genomics." PLoS Biology **1**(2): 150-160.
Hastie, T., R. Tibshirani, et al. (2001). The elements of statistical learning : data mining, inference, and prediction. New York, Springer.
Hein, J., C. Wiuf, et al. (2000). "Statistical Alignment: Computational Properties Homology Testing and Goodness-of-Fit." J. Mol. Biol. **302**: 265-279.
Henderson, I. R. and S. E. Jacobsen (2007). "Epigenetic inheritance in plants." Nature **447**: 418-424.
Hobolth, A. (2008). "A Markov chain Monte Carlo Expectation Maximization Algorithm for Statistical Analysis of DNA Sequence Evolution with Neighbor-Dependent Substitution Rates." Journal of Computational and Graphical Statistics **17**(1): 138-162.
Holliday, R. and G. W. Grigg (1993). "DNA methylation and mutation." Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **285**(1): 61-67.
Hsieh, F., S. C. Chen, et al. (2009). "A Nearly Exhaustive Search for CpG Islands on Whole Chromosomes." *IJB* **5**(1).
Hutter, B., M. Paulsen, et al. (2009). "Identifying CpG Islands by Differnt Computational Techniques." OMICS **13**(2).
Illingworth, R., A. Kerr, et al. (2008). "A novel CpG island set identifies tissue-specific methylation at developmental gene loci." PLoS Biol **6**(1): e22.
Illingworth, R. S., U. Gruenewald-Schneider, et al. (2010). "Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome." PLoS Genet **6**(9): e1001134.
Irizarry, R. A., C. Ladd-Acosta, et al. (2009). "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores." Nat Genet **41**(2): 178-186.
Irizarry, R. A., H. Wu, et al. (2009). "A species-generalized probabilistic model-based definition of CpG islands." Mamm Genome.
Iwama, H. and T. Gojobori (2004). "Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network." Proc. Natl Acad. Sci. USA **101**: 17156-17161.
Jabbari, K., S. Cacciò, et al. (1997). "Evolutionary changes in CpG and methylation levels in the genome of vertebrates." Gene **205**(1-2): 109-118.
Jackson-Grusby, L., C. Beard, et al. (2001). "Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation." Nat Genet **27**(1): 31-39.
Jiang, C., L. Han, et al. (2007). "Features and Trend of Loss of Promoter-Associated CpG Islands in the Human and Mouse Genomes." Molecular Biology and Evolution **24**(9): 1991-2000.
Johnson, A. D., E. Richardson, et al. (2011). "Evolution of the germ line-soma relationship in vertebrate embryos." Reproduction **141**: 291-300.
Jones, R. G. and C. B. Thompson (2009). "Tumor suppressors and cell metabolism: a recipe for cancer growth." Genes & Development **23**: 537-548.
Jukes, T. H. and C. R. Cantor (1969). "Evolution of protein molecules." Mammalian Protein Metabolisn: 21-132.

Jurka, J. (1994). Approaches to Identificatrion and Analysis of Interspersed Repetitive DNA Sequences. Automated DNA sequencing and analysis. M. D. Adams, C. Fields and J. C. Venter. London, Academic Press Limited**:** 294-298.
Jurka, J., V. V. Kapitonov, et al. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and Genome Research **110**(1-4): 462-467.
Karlin, S. and H. M. Taylor (1975). A First Course in Stocastic Processes. New York-London, Academic Press.
Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database **2011**.
Kato, M., A. Miura, et al. (2003). "Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis." Curr. Biol. **13**(5): 421-426.
Kato, T., M. Ahmed, et al. (1996). "Inactivation of hepatitis virus cDNA transgene by hypermethylation in transgenic mice." Arch. Virol. **141**(5): 951-958.
Khuu, P., M. Sandor, et al. (2007). "Phylogenomic analysis of the emergence of GC-rich transcription elements." Proceedings of the National Academy of Sciences **104**(42): 16528-16533.
Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." Journal of Molecular Evolution **16**(2): 111-120.
Knapska, E. and L. Kaczmarek (2004). "A gene for neuronal plasticity in the mammalian brain: Zif268/Egr-1/NGFI-A/Krox-24/TIS8/ZENK?" Progress in Neurobiology **74**(4): 183-211.
Kobayashi, I., A. Nobusato, et al. (1999). "Shaping the genome - restriction-modification systems as mobile genetic elements." Curr. Opin. Genet. Dev. **9**(6): 640-656.
Koser, S. (2012). Reconstructing Ancestral Methylation States in Phylogenetic Trees. Center for Bioinformatics. Saarbrücken, Saarland University. **Master of Sience, Bioinformatics:** 70.
Laurent, L., E. Wong, et al. (2010). "Dynamic changes in the human methylome during differentiation." Genome Research **20**(3): 320-331.
Lechner, M., M. Marz, et al. (2013). "The correlation of genome size and DNA methylation rate in metazoans." Theory Biosci, **132**(1).
Li, W. L. S. and A. J. Drummond (2011). "Model Averaging and Bayes Factor Calculation of Releaxed Molecular Clocks in Bayesian Phylogenetics." Mol Biol Evol **Advanced Access**.
Lienert, F., C. Wirbelauer, et al. (2011). "Identification of genetic elements that autonomously determine DNA methylation states." Nature Genetics **43**(11): 1091-U78.
Linehan, W. M., R. Srinivasan, et al. (2010). "The genetic basis of kidney cancer: a metabolic disease." Nature Reviews Urology **7**: 277-285.
Lipman, D. J. and W. R. Pearson (1985). "Rapid and sensitive protein similarity searches." Science **227**: 1435-1441.
Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(7271): 315-322.
Liu, W. M. and C. W. Schmid (1993). "Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation." Nucleic Acids Res. **21**(6): 1351-1359.

Lobry, J. R. and C. Lobry (1999). "Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant." Molecular Biology and Evolution **16**(6): 719-719.
Low, D. A., N. J. Weyand, et al. (2001). "Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence." Infection and Ummunity **69**(12): 7197-7204.
Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8[thinsp]A resolution." Nature **389**(6648): 251-260.
Luger, K. and T. J. Richmond (1998). "The histone tails of the nucleosome." Curr. Opin. Genet. Dev. **8**(2): 140-142.
Lunter, G. and J. Hein (2004). "A nucleotide substitution model with nearest-neighbour interactions." Bioinformatics **20**(suppl_1): i216-223-i216-223.
Macleod, D., J. Charlton, et al. (1994). "Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island." Genes & Development **8**(19): 2282-2292.
Margulies, E. H. and E. Birney (2008). "Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes." Nat Rev Genet **9**(4): 303-13.
Margulies, E. H., C. W. Chen, et al. (2006). "Differences between pair-wise and multiple-sequence alignment methods affect vertebrate genome comparisons." Trends in Genetics **22**(4): 187-193.
Marin, M., A. Karis, et al. (1997). "Transcription Factor Sp1 Is Essential for Early Embryonic Development but Dispensable for Cell Growth and Differentiation." Cell **89**(4): 619-628.
Matsuo, K., O. Clay, et al. (1993). "Evidence for erosion of mouse CpG islands during mammalian evolution." Somat Cell Mol Genet **19**(6): 543-55.
Miller, W., K. D. Makova, et al. (2004). "Comparative genomics." Annu Rev Genomics Hum Genet **5**: 15-56.
Molaro, A., E. Hodges, et al. (2011). "Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates." Cell **146**(6): 1029-1041.
Morgan, H. D., F. Santos, et al. (2005). "Epigenetic reprogramming in mammals." Human Molecular Genetics **14**(suppl 1): R47-R58-R47-R58.
Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.
Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A Noval Method for Fast and Accurate Multiple Sequence Alignment." J. Mol. Biol. **48**: 443-453.
Novik, K. L., I. Nimmrich, et al. (2002). "Epigenomics: genome-wide study of methylation phenomena." Current issues in molecular biology **4**: 111–128-111–128.
Okano, M., D. W. Bell, et al. (1999). "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." Cell **99**: 247-257.
Paten, B., J. Herrero, et al. (2008). "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignments with paralogs." Genome Res. **18**: 1814-1828.

Patwardhan, S., A. Gashler, et al. (1991). "EGR3, a novel member of the Egr family of genes encoding immediate-early transcription factors." Oncogene **6**(6): 917-928.
Peifer, M., J. E. Karro, et al. (2008). "Is there an acceleration of the CpG transition rate during the mammalian radiation?" Bioinformatics **24**(19): 2157-2164.
Reik, W., W. Dean, et al. (2001). "Epigenetic reprogramming in mammalian development." Science **293**(5532): 1089-93.
Reik, W. and J. Walter (2001). "Genomic imprinting: parental influence on the genome." Nat Rev Genet **2**(1): 21-32.
Rice, J. C. and C. D. Allis (2001). "Histone methylation versus histone acetylation: new insights into epigenetic regulation." Current Opinion in Cell Biology **13**(3): 263-273.
Russo, V. E. A., R. A. Martienssen, et al. (1996). Epigenetic mechanisms of gene regulation. Plainview, N.Y., Cold Spring Harbor Laboratory Press.
Saffer, J., S. Jackson, et al. (1991). "Developmental Expression of Sp1 in the Mouse." Mol. Cell. Biol. **11**(4): 2189-2199.
Schanen, N. C. (2006). "Epigenetics of autism spectrum disorder." Human Molecular Genetics **15**(Review Issue 2): R138-R150.
Schmidt, D., D. W. Wilson, et al. (2010). "Five-Vertebrate ChIP-seq Reveals the Evolutionray Dynamics of Transcription Factor Binding." Science **328**(5981): 1036-1040.
Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Research **18**(20): 6097-6100.
Schwartz, S., W. J. Kent, et al. (2003). "Human-Mouse Alignments with BLASTZ." Genome Res. **13**: 103-107.
Schwartz, Y. B. and V. Pirrotta (2007). "Polycomb silencing mechanisms and the management of genomic programmes." Nat Rev Genet **8**(1): 9-22.
Shen, J. C., W. M. Rideout III, et al. (1994). "The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA." Nucleic Acids Research **22**(6): 972-972.
Siepel, A. and D. Haussler (2004). "Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood." Mol Biol Evol **21**(3): 468-488.
Sing, T., O. Sander, et al. (2005). "ROCR: visualizing classifier performance in R." Bioinformatics **21**(20): 3940-1.
Smith, A. D., P. Sumazin, et al. (2005). "Identifying tissue-selective transcription factor binding sites in vertebrate promoters." Proceedings of the National Academy of Sciences of the United States of America **102**(5): 1560-1565.
Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J. Mol. Biol. **147**: 195-197.
Stapleton, G., M. Patrizia Somma, et al. (1993). "Cell type-specific interactions of transcription factors with a housekeeping promoter in vivo." Nucleic Acids Research **21**(10): 2465-2471.
Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.
Straussman, R. (2009). "Developmental programming of CpG island methylation profiles in the human genome." Nature Struct. Mol. Biol. **16**: 564-571.

Su, J., Y. Zhang, et al. (2009). "CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes." Nucleic Acids Research **38**(1): e6-e6.
Sved, J. and A. Bird (1990). "The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model." Proceedings of the National Academy of Sciences of the United States of America **87**(12): 4692-4692.
Takahashi, K. and S. Yamanaka (2006). "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." Cell **126**(4): 663-676.
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**(6): 3740-5.
Thompson, J. D., D. G. Higgins, et al. (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**(22): 4673-80.
Tommasi, S. and G. P. Pfeifer (1997). "Constitutive Protection of E2F Recognition Sequences in the Human Thymidine Kinase Promoter during Cell Cycle Progression." Journal of Biological Chemistry **272**(48): 30483-30490.
Turnbull, C. and N. Rahman (2008). "Genetic Predisposition to Breast Cancer: Past, Present, and Future." Annual Review of Genomics and Human Genetics **9**: 321-345.
Vinckenbosch, N., I. Dupanloup, et al. (2006). "Evolutionary fate of retroposed gene copies in the human genome." Proc. Natl Acad. Sci. USA **103**(p): 3220-3225.
Visser, K. E., A. Eichten, et al. (2006). "Paradoxical roles of the immune system during cancer development." Nature Rev. Cancer **6**: 24-37.
Vlieghe, D., A. Sandelin, et al. (2006). "A new generation of JASPAR, the open-access repository for transcription factor binding site profiles." Nucleic Acids Res **34**(Database issue): D95-7.
Warburg, O. (1956). "On the Origin of Cancer Cells." Science **123**(3191): 309-314.
Wasserman, W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nature Reviews Genetics **5**: 276-287.
Weber, M., I. Hellmann, et al. (2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." Nat Genet **39**(4): 457-466.
Whang, Y. E., X. Wu, et al. (1998). "Inactivation of the tumor suppressor PTEN/MMAC1 in advanced human prostate cancer through loss of expression." Proceedings of the National Academy of Sciences **95**(9): 5246-5250.
Wingender, E., P. Dietze, et al. (1996). "TRANSFAC: a database on transcription factors and their DNA binding sites." Nucleic Acids Research **24**(1): 238-241.
Wossidlo, M., T. Nakamura, et al. (2011). "5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming." Nat. Commun. **2**.
Wu, H., B. Caffo, et al. (2010). "Redefining CpG islands using hidden Markov models." Biostatistics **11**(3): 499 -514-499 -514.
Wu, H., W. P. Yang, et al. (1995). "Building zinc fingers by selection: toward a therapeutic application." Proceedings of the National Academy of Sciences **92**(2): 344-348.

Wu, S. C. and Y. Zhang (2010). "Active DNA demethylation: many roads lead to Rome." Nature Reviews Molecular Cell Biology **11**: 607-620.
Xiao, S., D. Xie, et al. (2012). "Comparative Epigenomics Annotation of Regulatory DNA." Cell **149**: 1381-1392.
Yamada, Y., H. Watanabe, et al. (2004). "A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q." Genome Res **14**(2): 247-66.
Yi, S. V. and M. A. D. Goodisman (2009). "Computational approaches for understanding the evolution of DNA methylation in animals." epigenetics **4**(1559-2294): 551-556.
Zemach, A., I. E. McDaniel, et al. (2010). "Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation." Science **328**(5980): 916-919.
Zeng, J. and S. V. Yi (2010). "DNA Methylation and Genome Evolution in Honeybee: Gene Length, Expression, Functional Enrichment Covary with Evolutionary Signature of DNA Methylation." Genome Biol Evol **2**: 770-780.
Zhong, C. and A. Meng (2005). "Sp1-like transcription factors are regulators of embryonic development in vetebrates." Development, Growth and Diferentiation **47**(4): 201-211.
Zipfel, P. F., E. L. Decker, et al. (1998). "The human zinc finger protein EGR-4 acts as autoregulatory transcriptional repressor." BBA **1354**(2): 134-144.

Appendix A – Methylation level of human CGI Shadow annotations

CGI Shadow annotations produced by CgiHunter:

CG-content	Obs/Exp ratio	min length	TotalLength	avgLength	TotalCgis	avgMethylation
50	60	500	153797775	2067	74387	0.62499
50	65	500	116922512	2041	57264	0.57396
50	70	500	89328911	1978	45154	0.513163
50	75	500	67090293	1905	35206	0.434378
50	80	500	49509208	1769	27983	0.3579
55	60	500	100900488	2330	43292	0.500802
55	65	500	81681369	2268	36008	0.440318
55	70	500	65910468	2206	29867	0.368785
55	75	500	52353904	2097	24965	0.290839
55	80	500	40494819	1918	21108	0.224823
60	60	500	66369000	2212	29997	0.370497
60	65	500	56461756	2143	26337	0.315517
60	70	500	47847656	2077	23034	0.253055
60	75	500	39931087	1973	20233	0.193458
60	80	500	32434081	1799	18027	0.15616
65	60	500	41772466	1796	23248	0.290875
65	65	500	37229012	1761	21132	0.245257
65	70	500	32878666	1735	18944	0.19244
65	75	500	28640844	1681	17030	0.146253
65	80	500	24211802	1579	15324	0.121733
70	60	500	22396563	1384	16173	0.219159
70	65	500	20578643	1401	14682	0.162798
70	70	500	19185288	1412	13584	0.124488
70	75	500	17736893	1393	12729	0.102266
70	80	500	15804734	1347	11728	0.088633
50	60	600	140501506	2584	54359	0.56929
50	65	600	106878638	2536	42139	0.502579
50	70	600	81049765	2486	32597	0.416546
50	75	600	60953157	2362	25798	0.325893
50	80	600	44961328	2150	20904	0.250999
55	60	600	95762438	2714	35279	0.444471
55	65	600	77653721	2610	29752	0.380466
55	70	600	62732768	2513	24956	0.302679
55	75	600	49896106	2352	21206	0.232293
55	80	600	38465378	2127	18078	0.179778
60	60	600	63827719	2444	26108	0.328315
60	65	600	54404630	2343	23212	0.274575
60	70	600	46084710	2264	20354	0.210936
60	75	600	38498347	2124	18125	0.165173
60	80	600	31028633	1938	16008	0.137227
65	60	600	40148681	1929	20805	0.269753
65	65	600	35771695	1887	18949	0.223993

65	70	600	31483751	1871	16827	0.166601
65	75	600	27407336	1804	15190	0.129718
65	80	600	22985233	1698	13531	0.109075
70	60	600	20969650	1500	13974	0.198929
70	65	600	19267992	1522	12653	0.140548
70	70	600	18044433	1525	11830	0.107523
70	75	600	16679614	1501	11109	0.092345
70	80	600	14763097	1455	10140	0.082056
50	60	700	131790162	3049	43218	0.518553
50	65	700	100330506	2976	33709	0.439659
50	70	700	76486901	2871	26640	0.347551
50	75	700	57836154	2658	21757	0.267411
50	80	700	42629544	2376	17939	0.204758
55	60	700	91697275	3071	29855	0.392426
55	65	700	74295554	2936	25301	0.320886
55	70	700	60123255	2798	21482	0.24558
55	75	700	47998145	2556	18775	0.194542
55	80	700	36886949	2283	16156	0.157106
60	60	700	61993688	2606	23786	0.305121
60	65	700	52734850	2498	21109	0.248507
60	70	700	44631824	2407	18535	0.187536
60	75	700	37169748	2254	16485	0.149746
60	80	700	29695305	2065	14380	0.125283
65	60	700	38559540	2053	18774	0.255076
65	65	700	34227502	2017	16963	0.204236
65	70	700	30137097	1996	15093	0.14817
65	75	700	26183438	1920	13636	0.119549
65	80	700	21778930	1813	12009	0.101533
70	60	700	19527005	1619	12060	0.180909
70	65	700	18018037	1637	11003	0.125905
70	70	700	16941361	1631	10381	0.1002
70	75	700	15648278	1603	9759	0.088181
70	80	700	13770416	1555	8852	0.078058
50	60	800	125326958	3479	36019	0.470946
50	65	800	95524449	3375	28302	0.382499
50	70	800	73226185	3188	22964	0.294514
50	75	800	55506915	2895	19168	0.225277
50	80	800	40770286	2557	15943	0.176487
55	60	800	88790024	3338	26594	0.355986
55	65	800	71977562	3169	22710	0.283502
55	70	800	58391681	2980	19588	0.216861
55	75	800	46492353	2708	17164	0.174696
55	80	800	35425395	2421	14627	0.142858
60	60	800	60287274	2747	21944	0.286281
60	65	800	51099127	2642	19335	0.224611
60	70	800	43260459	2537	17050	0.169783
60	75	800	35928435	2369	15164	0.13931
60	80	800	28377680	2185	12983	0.117172
65	60	800	36899183	2179	16927	0.240111
65	65	800	32596039	2155	15125	0.182152

65	70	800	28755933	2122	13547	0.133283
65	75	800	24927163	2034	12253	0.111371
65	80	800	20588779	1922	10710	0.095532
70	60	800	18182268	1728	10520	0.16734
70	65	800	16829000	1744	9649	0.115097
70	70	800	15845029	1733	9140	0.09362
70	75	800	14586825	1702	8566	0.082511
70	80	800	12744108	1655	7697	0.073781

Appendix B – CpG Mountain annotations computed by region length

CpG Mountain annotations generated with Cgi Shadow annotation length as approximation of CpG island strength.

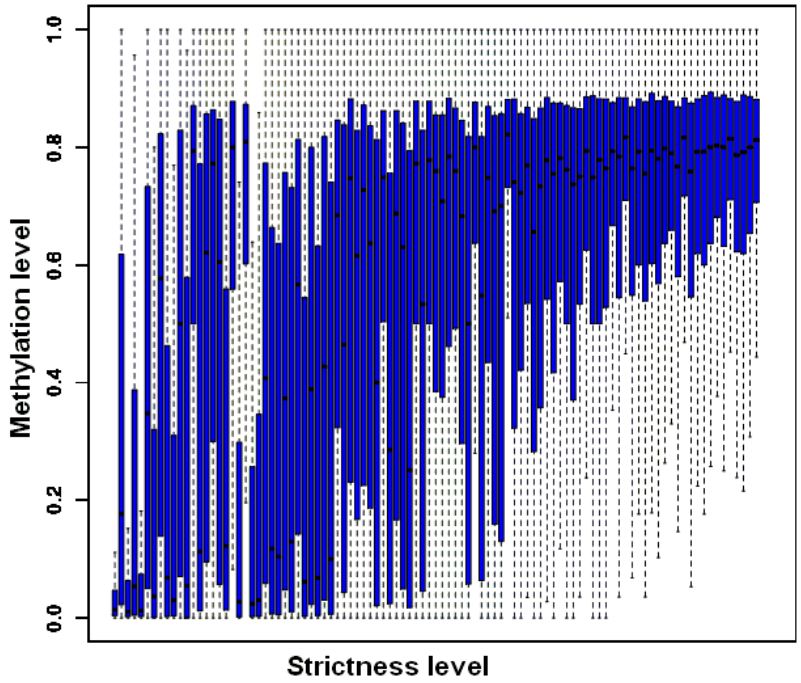


Figure: CGM annotation with 100 Strictness levels based on Cgi Shadow total length

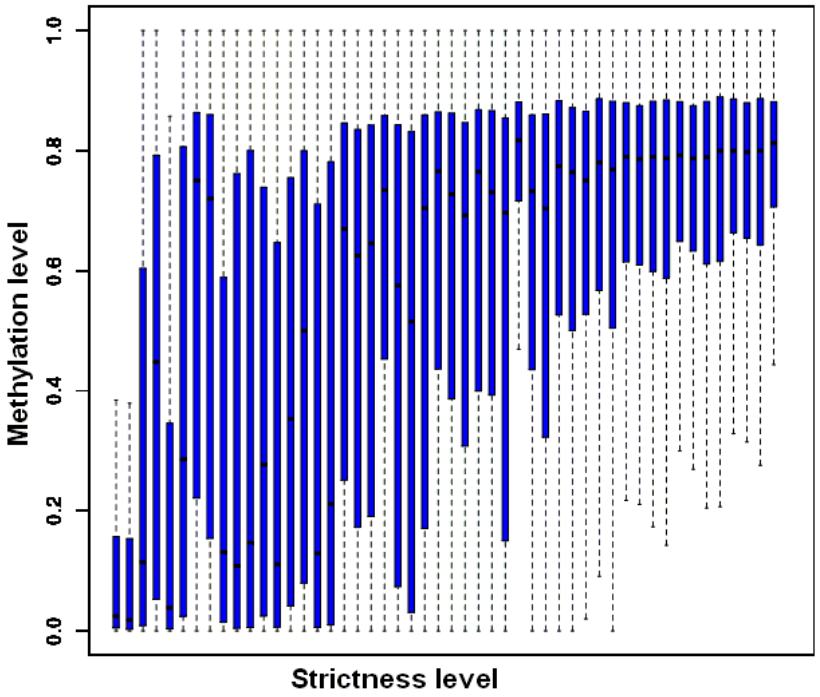


Figure: CGM annotation with 50 Strictness levels based on Cgi Shadow total length

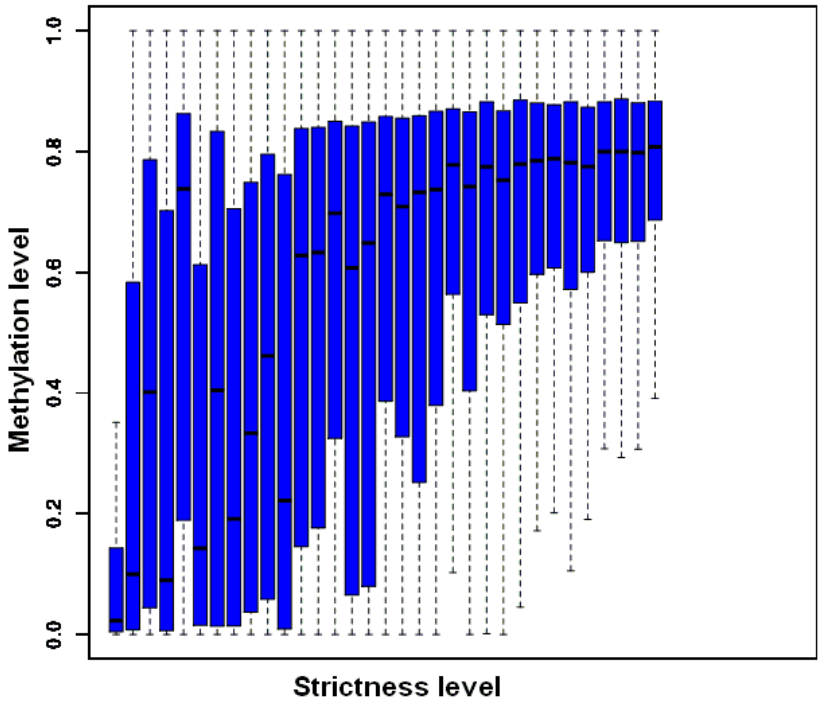


Figure: CGM annotation with 33 Strictness levels based on Cgi Shadow total length

Appendix C – Overlap of conserved TFBS with CGIs

The table shows for each TFBS the number of predicted sites in comparison to the number of sites that overlap the CGI Mountain annotation. The CGM annotation covers 5% of the genome. Thus, dividing that ratio of CGM to all sites by 0.05 yields the relative overrepresentation of the TFBS in CGIs. Furthermore, for the fraction of sites that are located in CGIs the mean CGI Mountain score is reported.

TFBS	In CGIs	All TFBS	CGI fraction	Mean CGM strict.	CGI Overrep.	p-value
SP1_Q6	837	941	88.95	91.70	17.79	0.00E+00
SP1_01	683	769	88.82	91.24	17.764	0.00E+00
AP2_Q6	623	709	87.87	92.01	17.574	0.00E+00
NFY_01	97	128	75.78	90.02	15.156	0.00E+00
PAX4_01	441	608	72.53	91.24	14.506	0.00E+00
PAX5_01	406	576	70.49	91.06	14.098	0.00E+00
CETS1P54_01	238	339	70.21	89.37	14.042	0.00E+00
EGR3_01	999	1509	66.2	90.65	13.24	0.00E+00
NRF2_01	455	717	63.46	89.72	12.692	0.00E+00
ELK1_02	386	617	62.56	90.02	12.512	0.00E+00
USF_C	191	306	62.42	90.05	12.484	0.00E+00
E2F_03	149	240	62.08	90.28	12.416	0.00E+00
MAZR_01	526	859	61.23	89.92	12.246	0.00E+00
E2F_02	538	895	60.11	89.02	12.022	0.00E+00
NMYC_01	702	1196	58.7	89.36	11.74	0.00E+00
EGR1_01	953	1693	56.29	90.48	11.258	0.00E+00
PAX5_02	352	645	54.57	90.42	10.914	0.00E+00
GATA2_01	76	145	52.41	89.21	10.482	0.00E+00
NGFIC_01	990	1986	49.85	89.93	9.97	0.00E+00
CREB_02	219	442	49.55	89.57	9.91	0.00E+00
AHRARNT_01	558	1144	48.78	89.10	9.756	0.00E+00
MYCMAX_03	839	1731	48.47	89.30	9.694	0.00E+00
ATF_01	736	1537	47.89	89.03	9.578	0.00E+00
CREBP1_Q2	685	1433	47.8	89.26	9.56	0.00E+00
USF_Q6	434	924	46.97	88.97	9.394	0.00E+00
AHRARNT_02	627	1355	46.27	89.77	9.254	0.00E+00
ARNT_01	486	1061	45.81	88.68	9.162	0.00E+00
EGR2_01	768	1766	43.49	89.32	8.698	0.00E+00
SPZ1_01	366	843	43.42	88.57	8.684	0.00E+00
AHR_01	413	958	43.11	88.35	8.622	0.00E+00
MAX_01	288	686	41.98	90.00	8.396	0.00E+00
MZF1_02	998	2398	41.62	87.93	8.324	0.00E+00
CREB_01	434	1060	40.94	88.36	8.188	0.00E+00
TAXCREB_01	470	1165	40.34	89.20	8.068	0.00E+00
PAX4_03	714	1774	40.25	87.82	8.05	0.00E+00
CREBP1CJUN_01	499	1255	39.76	88.62	7.952	0.00E+00
AP4_Q5	9	23	39.13	83.67	7.826	5.90E-14

LMO2COM_01	179	483	37.06	87.44	7.412	0.00E+00
ARNT_02	487	1407	34.61	88.17	6.922	0.00E+00
GATA1_01	102	297	34.34	88.21	6.868	0.00E+00
CREB_Q4	41	122	33.61	87.88	6.722	1.25E-47
MYCMAX_01	407	1218	33.42	88.25	6.684	0.00E+00
NFKAPPAB50_01	112	343	32.65	89.27	6.53	0.00E+00
CMYB_01	501	1564	32.03	87.81	6.406	0.00E+00
RREB1_01	843	2662	31.67	87.28	6.334	0.00E+00
TAXCREB_02	481	1567	30.7	87.88	6.14	0.00E+00
E47_01	319	1057	30.18	85.89	6.036	0.00E+00
NFY_Q6	237	793	29.89	87.03	5.978	0.00E+00
E2F_01	414	1408	29.4	87.73	5.88	0.00E+00
USF_01	185	647	28.59	87.22	5.718	6.54E-167
P300_01	461	1715	26.88	86.70	5.376	0.00E+00
STAT1_01	281	1073	26.19	88.50	5.238	0.00E+00
ELK1_01	336	1299	25.87	86.12	5.174	0.00E+00
CREB_Q2	67	260	25.77	85.40	5.154	2.77E-53
SREBP1_01	545	2127	25.62	87.90	5.124	0.00E+00
ATF6_01	293	1213	24.15	86.01	4.83	0.00E+00
OLF1_01	443	1858	23.84	86.79	4.768	0.00E+00
XBP1_01	440	1868	23.55	86.03	4.71	0.00E+00
HEN1_02	246	1082	22.74	87.30	4.548	0.00E+00
MYOD_01	378	1679	22.51	86.01	4.502	0.00E+00
NRSF_01	429	1913	22.43	85.36	4.486	0.00E+00
YY1_02	326	1501	21.72	87.34	4.344	0.00E+00
AREB6_03	438	2055	21.31	87.06	4.262	0.00E+00
CP2_01	299	1430	20.91	84.82	4.182	0.00E+00
AP4_01	276	1341	20.58	86.92	4.116	0.00E+00
PAX2_01	378	1865	20.27	87.33	4.054	0.00E+00
P53_01	479	2425	19.75	87.12	3.95	0.00E+00
PAX3_01	441	2294	19.22	86.86	3.844	0.00E+00
HMX1_01	429	2278	18.83	85.55	3.766	0.00E+00
PAX2_02	37	197	18.78	85.11	3.756	6.97E-19
MYOGNF1_01	251	1341	18.72	86.15	3.744	0.00E+00
CREL_01	508	2741	18.53	85.61	3.706	0.00E+00
MYCMAX_02	313	1702	18.39	85.13	3.678	0.00E+00
MIF1_01	434	2376	18.27	86.26	3.654	0.00E+00
HEN1_01	281	1540	18.25	87.17	3.65	0.00E+00
HOX13_01	277	1531	18.09	85.38	3.618	0.00E+00
NFKAPPAB_01	432	2529	17.08	85.94	3.416	0.00E+00
ZID_01	292	1745	16.73	85.43	3.346	0.00E+00
NFKB_Q6	61	377	16.18	85.95	3.236	2.27E-23
ROAZ_01	465	2885	16.12	85.16	3.224	0.00E+00
SREBP1_02	510	3171	16.08	83.98	3.216	0.00E+00
NFKB_C	452	2812	16.07	85.17	3.214	0.00E+00
PAX4_04	305	1929	15.81	82.60	3.162	0.00E+00
YY1_01	311	2005	15.51	84.22	3.102	0.00E+00
ZIC2_01	277	1811	15.3	85.19	3.06	6.98E-90
ZIC1_01	298	2012	14.81	84.12	2.962	1.14E-90
HNF4_01_B	113	782	14.45	82.23	2.89	7.75E-34

ZIC3_01	309	2168	14.25	83.61	2.85	5.66E-87
HTF_01	428	3076	13.91	82.80	2.782	0.00E+00
COMP1_01	267	1937	13.78	84.39	2.756	2.11E-70
NFKAPPAB65_01	187	1359	13.76	86.08	2.752	1.13E-49
AP1FJ_Q2	111	807	13.75	86.16	2.75	3.68E-30
STAT3_01	305	2285	13.35	84.72	2.67	6.96E-75
NFY_C	150	1130	13.27	82.60	2.654	2.67E-37
AREB6_02	151	1157	13.05	84.30	2.61	3.28E-36
GATA3_01	129	989	13.04	83.27	2.608	3.83E-31
RFX1_01	485	3789	12.8	84.44	2.56	0.00E+00
RFX1_02	710	5613	12.65	82.82	2.53	0.00E+00
AP1_Q6	9	72	12.5	82.67	2.5	3.50E-03
NF1_Q6	181	1450	12.48	81.44	2.496	4.65E-39
ER_Q6	458	3685	12.43	84.45	2.486	4.15E-95
AREB6_01	322	2622	12.28	84.83	2.456	1.34E-65
CDPCR1_01	402	3311	12.14	82.45	2.428	2.70E-79
MEIS1_01	353	2913	12.12	81.67	2.424	1.52E-69
PPARG_03	410	3466	11.83	82.18	2.366	5.46E-76
HNF4_01	160	1353	11.83	82.95	2.366	1.05E-30
EN1_01	104	883	11.78	82.73	2.356	2.43E-20
BACH2_01	543	4835	11.23	82.55	2.246	6.26E-88
IK2_01	62	563	11.01	83.45	2.202	5.92E-11
IK1_01	378	3469	10.9	83.47	2.18	3.63E-57
STAT5A_02	418	3840	10.89	82.31	2.178	7.43E-63
TATA_01	470	4335	10.84	82.01	2.168	1.04E-69
NCX_01	381	3527	10.8	82.56	2.16	2.61E-56
COUP_01	596	5523	10.79	81.04	2.158	8.43E-87
GATA1_03	369	3452	10.69	83.26	2.138	4.28E-53
SRF_Q6	304	2849	10.67	82.50	2.134	7.58E-44
SEF1_C	336	3164	10.62	82.52	2.124	1.15E-47
CREBP1_01	483	4550	10.62	81.84	2.124	1.18E-67
MYB_Q6	256	2416	10.6	81.32	2.12	1.62E-36
LUN1_01	419	3965	10.57	81.95	2.114	3.23E-58
NFE2_01	525	4968	10.57	82.48	2.114	1.75E-72
AP1_01	634	6026	10.52	81.95	2.104	4.31E-86
ARP1_01	400	3823	10.46	82.98	2.092	3.56E-54
CHOP_01	424	4053	10.46	82.98	2.092	2.71E-57
IRF2_01	631	6090	10.36	82.40	2.072	3.94E-82
ISRE_01	586	5721	10.24	82.70	2.048	5.59E-74
IK3_01	409	4048	10.1	82.28	2.02	3.33E-50
IRF7_01	518	5179	10	81.99	2	2.80E-61
E47_02	145	1455	9.97	82.59	1.994	3.60E-18
LYF1_01	229	2311	9.91	81.13	1.982	2.53E-27
SRF_C	252	2566	9.82	82.80	1.964	3.87E-29
CEBP_C	411	4189	9.81	81.09	1.962	2.59E-46
IRF1_01	682	6959	9.8	81.87	1.96	2.14E-75
LMO2COM_02	527	5556	9.49	80.66	1.898	4.14E-53
HSF2_01	234	2468	9.48	82.24	1.896	1.70E-24
CEBPB_02	447	4770	9.37	82.61	1.874	1.24E-43
GFI1_01	578	6204	9.32	81.85	1.864	7.27E-55

HOXA3_01	392	4231	9.26	81.11	1.852	4.09E-37
STAT_01	392	4266	9.19	82.92	1.838	3.80E-36
TST1_01	592	6488	9.12	81.26	1.824	1.82E-52
PPARG_01	710	7814	9.09	81.98	1.818	1.08E-61
PPARA_01	465	5164	9	82.03	1.8	8.30E-40
AP2REP_01	547	6091	8.98	80.63	1.796	4.24E-46
FAC1_01	493	5510	8.95	81.47	1.79	3.33E-41
FOXO4_01	269	3006	8.95	80.37	1.79	2.97E-23
HSF1_01	248	2782	8.91	80.46	1.782	2.71E-21
TCF11_01	72	815	8.83	83.08	1.766	5.10E-07
AP4_Q6	140	1595	8.78	81.66	1.756	4.45E-12
HLF_01	621	7240	8.58	81.75	1.716	2.50E-44
SOX9_B1	507	6001	8.45	82.07	1.69	1.53E-34
AREB6_04	114	1360	8.38	81.71	1.676	1.05E-08
GATA1_02	285	3463	8.23	79.69	1.646	2.76E-18
CDPCR3HD_01	255	3101	8.22	80.31	1.644	1.79E-16
CEBPB_01	340	4144	8.2	81.21	1.64	2.92E-21
BACH1_01	705	8622	8.18	80.01	1.636	9.79E-42
AP1_C	132	1616	8.17	78.33	1.634	5.10E-09
MSX1_01	475	5959	7.97	80.24	1.594	6.73E-26
HAND1E47_01	298	3774	7.9	79.14	1.58	3.26E-16
FREAC2_01	787	9980	7.89	81.27	1.578	6.08E-40
FOXO1_02	774	9846	7.86	80.67	1.572	8.71E-39
GR_Q6	247	3172	7.79	81.34	1.558	5.94E-13
CEBP_01	186	2407	7.73	80.47	1.546	8.27E-10
PPARG_02	312	4041	7.72	80.53	1.544	2.09E-15
E4BP4_01	777	10253	7.58	79.71	1.516	4.60E-33
FOXO4_02	687	9146	7.51	81.36	1.502	3.05E-28
CEBP_Q2	281	3774	7.45	79.74	1.49	5.44E-12
TCF11MAFG_01	459	6167	7.44	79.73	1.488	1.34E-18
TGIF_01	469	6312	7.43	80.20	1.486	8.06E-19
TAL1ALPHAE47_01	328	4430	7.4	80.38	1.48	2.11E-13
AP1_Q2	9	123	7.32	69.78	1.464	2.38E-01
SRY_02	453	6285	7.21	81.43	1.442	9.72E-16
TAL1BETAE47_01	189	2626	7.2	81.43	1.44	2.39E-07
BRACH_01	539	7551	7.14	79.71	1.428	1.53E-17
CDPCR3_01	411	5769	7.12	79.71	1.424	1.33E-13
GATA1_04	233	3283	7.1	80.67	1.42	3.52E-08
FOXO1_01	16	226	7.08	84.88	1.416	1.51E-01
GRE_C	223	3169	7.04	80.40	1.408	1.43E-07
P53_02	109	1555	7.01	78.67	1.402	2.77E-04
FREAC3_01	817	11874	6.88	79.38	1.376	5.33E-21
MRF2_01	355	5167	6.87	79.81	1.374	6.86E-10
PAX6_01	506	7404	6.83	79.59	1.366	4.44E-13
FOXO3_01	685	10077	6.8	80.14	1.36	1.23E-16
AP1_Q4	15	223	6.73	78.07	1.346	2.37E-01
SOX5_01	545	8141	6.69	80.59	1.338	2.30E-12
USF_02	5	75	6.67	74.80	1.334	5.08E-01
NFAT_Q6	182	2727	6.67	78.23	1.334	6.05E-05
GATA_C	492	7389	6.66	77.48	1.332	6.09E-11

RORA1_01	606	9119	6.65	79.76	1.33	5.61E-13
HFH3_01	444	6811	6.52	78.74	1.304	8.85E-09
CEBPA_01	242	3827	6.32	79.60	1.264	1.72E-04
FREAC4_01	677	10747	6.3	79.55	1.26	6.37E-10
MEF2_01	783	12444	6.29	78.68	1.258	3.74E-11
TAL1BETAITF2_01	426	6863	6.21	80.48	1.242	4.46E-06
SRF_01	362	5878	6.16	79.41	1.232	4.59E-05
RORA2_01	651	10690	6.09	79.52	1.218	2.34E-07
OCT1_03	268	4401	6.09	77.59	1.218	9.12E-04
STAT5A_01	175	2891	6.05	77.77	1.21	9.36E-03
MEF2_02	437	7256	6.02	77.79	1.204	6.42E-05
CHX10_01	895	14949	5.99	78.11	1.198	3.07E-08
OCT_C	708	11876	5.96	79.71	1.192	1.52E-06
RSRFC4_01	945	15901	5.94	77.56	1.188	4.87E-08
MEIS1AHOXA9_01	576	9767	5.9	77.33	1.18	4.71E-05
TATA_C	531	9043	5.87	77.86	1.174	1.42E-04
HNF1_01	481	8391	5.73	79.82	1.146	2.08E-03
NKX25_01	426	7440	5.73	79.31	1.146	4.07E-03
GATA1_05	416	7558	5.5	78.69	1.1	4.43E-02
PAX4_02	527	9584	5.5	77.74	1.1	2.51E-02
S8_01	711	13375	5.32	77.37	1.064	9.37E-02
FOXD3_01	441	8351	5.28	77.92	1.056	2.39E-01
OCT1_Q6	154	2938	5.24	78.15	1.048	5.48E-01
STAT5B_01	204	3900	5.23	77.87	1.046	5.08E-01
CDP_01	446	8602	5.18	79.51	1.036	4.32E-01
HNF3B_01	617	12096	5.1	77.92	1.02	6.11E-01
HNF1_C	509	10088	5.05	78.22	1.01	8.34E-01
FOXJ2_01	600	11956	5.02	77.86	1.004	9.26E-01
EVI1_02	229	4607	4.97	77.12	0.994	9.27E-01
MEF2_03	365	7360	4.96	75.89	0.992	8.73E-01
BRN2_01	476	9771	4.87	77.21	0.974	5.60E-01
HFH1_01	574	11803	4.86	77.93	0.972	4.95E-01
NKX25_02	559	11665	4.79	77.36	0.958	3.03E-01
OCT1_06	378	7906	4.78	75.55	0.956	3.72E-01
EVI1_06	360	7613	4.73	76.62	0.946	2.78E-01
NKX22_01	456	9733	4.69	76.82	0.938	1.54E-01
PBX1_01	194	4181	4.64	79.55	0.928	2.86E-01
EVI1_03	258	5602	4.61	79.38	0.922	1.75E-01
PBX1_02	532	11572	4.6	78.40	0.92	4.69E-02
OCT1_05	98	2155	4.55	78.93	0.91	3.35E-01
POU3F2_02	697	15759	4.42	77.23	0.884	8.87E-04
NKX61_01	548	12435	4.41	78.14	0.882	2.41E-03
OCT1_02	447	10155	4.4	76.69	0.88	5.67E-03
EVI1_04	348	7951	4.38	77.03	0.876	1.08E-02
CDC5_01	572	13069	4.38	76.90	0.876	1.08E-03
CART1_01	694	15877	4.37	78.01	0.874	2.77E-04
FREAC7_01	479	11075	4.33	77.26	0.866	1.12E-03
OCT1_01	252	5820	4.33	76.69	0.866	1.90E-02
MEF2_04	463	10751	4.31	76.46	0.862	9.70E-04
EVI1_01	426	10147	4.2	77.06	0.84	2.11E-04

POU6F1_01	560	13489	4.15	76.89	0.83	6.14E-06
OCT1_07	539	13158	4.1	77.80	0.82	1.97E-06
OCT1_04	345	8439	4.09	76.56	0.818	1.21E-04
LHX3_01	468	11982	3.91	76.31	0.782	3.90E-08
CDP_02	525	13440	3.91	76.55	0.782	5.96E-09
POU3F2_01	601	16211	3.71	76.38	0.742	4.30E-14
EVI1_05	181	4874	3.71	76.67	0.742	3.78E-05
NKX3A_01	372	10468	3.55	76.74	0.71	1.12E-11
FOXJ2_02	484	14006	3.46	76.44	0.692	5.03E-17

Appendix D - Overrepresentation of TFBS in unmethylated genome sequence

TFBS motifs are taken from the non redundant version of the JASPAR vertebrate core database. For the PSSM of each motif the odds ratio and the log odds ratio are reported. The list is sorted descending by the strength of overrepresentation in sequences that obey the unmethylated equilibrium distribution.

Rank	TFBS Name	Odds-Ratio	Log Odds
1	DAL81	67.59394	6.078822
2	RSC30	24.19466	4.596617
3	PDR3	17.12353	4.097909
4	RSC3	14.12888	3.820575
5	IME1	13.82657	3.789372
6	RDS1	10.41102	3.380039
7	SWI4	8.482207	3.08444
8	MBP1::SWI6	7.578759	2.921962
9	MIZF	7.335728	2.87494
10	UGA3	6.71728	2.747877
11	STP1	6.605268	2.723617
12	LEU3	6.503181	2.701146
13	YLL054C	6.424421	2.683566
14	STP2	5.754297	2.52464
15	PDR1	5.477768	2.453588
16	SUT1	5.361134	2.422538
17	E2F1	5.333781	2.415158
18	MBP1	5.22171	2.384522
19	NHP10	4.881334	2.287276
20	GAL4	4.745383	2.246525
21	UME6	4.665303	2.221971
22	CHA4	4.633145	2.211992
23	TEA1	4.578067	2.194738
24	PUT3	4.526733	2.17847
25	CAT8	4.398683	2.137072
26	SNT2	4.36613	2.126355
27	RDS2	4.319054	2.110715
28	YER184C	4.219052	2.076919
29	SIP4	4.215525	2.075712
30	YJL103C	4.199741	2.0703
31	HAL9	4.156235	2.055277
32	YBR239C	4.147621	2.052284
33	TBS1	4.071869	2.025691
34	RDR1	4.06493	2.023231
35	ASG1	4.056344	2.02018
36	CEP3	4.004818	2.001737

37	XBP1	3.959635	1.985368
38	STB4	3.94848	1.981298
39	YLR278C	3.910654	1.96741
40	PDR8	3.900859	1.963792
41	OAF1	3.8895	1.959585
42	ELK4	3.848583	1.944327
43	Deaf1	3.847955	1.944092
44	YNR063W	3.843349	1.942364
45	STB5	3.825168	1.935523
46	DAL82	3.823474	1.934884
47	STP3	3.820866	1.933899
48	STP4	3.809575	1.92963
49	FHL1	3.774212	1.916176
50	Hkb	3.748139	1.906175
51	bZIP911	3.741861	1.903756
52	ECM22	3.707702	1.890525
53	YKL222C	3.704265	1.889187
54	HAP1	3.69831	1.886866
55	Egr1	3.695222	1.885661
56	YDR520C	3.623047	1.857203
57	Brk	3.536054	1.82214
58	HIF1A::ARNT	3.53456	1.82153
59	GSM1	3.37006	1.752774
60	ARO80	3.346785	1.742776
61	UPC2	3.344767	1.741906
62	GABPA	3.327055	1.734246
63	HAC1	3.222956	1.688385
64	SUT2	3.21087	1.682964
65	ABF1	3.199971	1.678059
66	Btd	3.170028	1.664496
67	Arnt	3.13557	1.648728
68	CBF1	3.131628	1.646913
69	bZIP910	3.121059	1.642035
70	YAP3	3.109702	1.636776
71	IXR1	3.04807	1.607896
72	YDR026C	2.970045	1.570485
73	Abi4	2.943993	1.557774
74	YRM1	2.929688	1.550747
75	H	2.9095	1.540771
76	YRR1	2.855842	1.513916
77	Arnt::Ahr	2.764196	1.46686
78	Mycn	2.633639	1.397058
79	YPR196W	2.624127	1.391837
80	CST6	2.438098	1.285756
81	LYS14	2.416138	1.272703
82	TYE7	2.350596	1.233027
83	EmBP-1	2.331017	1.22096
84	RPN4	2.328342	1.219303
85	TOD6	2.306597	1.205766
86	Myc	2.277505	1.187454

87	MAX	2.231237	1.157844
88	DOT6	2.218969	1.14989
89	RGT1	2.202312	1.139019
90	RTG3	2.193796	1.133429
91	MYC::MAX	2.056461	1.040163
92	USF1	2.015267	1.010971
93	Zfx	2.008058	1.005801
94	PHO4	1.906954	0.93127
95	ELK1	1.901499	0.927137
96	OPI1	1.89748	0.924085
97	SKO1	1.857292	0.893201
98	Opa	1.801429	0.849142
99	TFAP2A	1.793237	0.842566
100	Eip74EF	1.792597	0.842051
101	CREB1	1.725464	0.786985
102	ASH1	1.678963	0.747571
103	Run::Bgb	1.666133	0.736504
104	TP53	1.663463	0.734189
105	Pax5	1.660415	0.731544
106	Che-1	1.636427	0.710549
107	MIG2	1.631432	0.706139
108	TGA1A	1.619272	0.695346
109	SKN7	1.601776	0.679672
110	YAP1	1.58366	0.663263
111	MIG3	1.570622	0.651336
112	dl_1	1.566872	0.647888
113	IRF2	1.544383	0.627031
114	SP1	1.542081	0.624879
115	MET31	1.541623	0.62445
116	Gamyb	1.527936	0.611584
117	MIG1	1.507739	0.592386
118	Usp	1.504089	0.58889
119	NFKB1	1.489428	0.574758
120	NHLH1	1.488007	0.573382
121	DAL80	1.487193	0.572592
122	PLAG1	1.461019	0.546975
123	REB1	1.453522	0.539553
124	ZMS1	1.448373	0.534433
125	Gt	1.444209	0.53028
126	RELA	1.38685	0.471812
127	Myb	1.364474	0.448345
128	REL	1.336639	0.41861
129	CRZ1	1.329676	0.411075
130	Pax6	1.318112	0.398473
131	NF-kappaB	1.31739	0.397682
132	YPR015C	1.305493	0.384594
133	EWSR1-FLI1	1.305427	0.384522
134	YGR067C	1.281075	0.357355
135	EDS1	1.277652	0.353495
136	lin-14	1.27562	0.351199

137	BAS1	1.25974	0.333126
138	Klf4	1.257551	0.330617
139	YML081W	1.237606	0.307553
140	ARG80	1.236828	0.306645
141	Mafb	1.228359	0.296732
142	id1	1.2149	0.280837
143	ETS1	1.196565	0.258899
144	dl_2	1.189497	0.250351
145	MSN2	1.188369	0.248983
146	IRF1	1.187641	0.248098
147	AFT2	1.165851	0.221384
148	MZF1_1-4	1.157279	0.210737
149	Su(H)	1.141307	0.190687
150	YPR022C	1.141052	0.190365
151	Macho-1	1.14083	0.190084
152	Trl	1.13526	0.183022
153	YPR013C	1.132286	0.179238
154	SPIB	1.123779	0.168358
155	RIM101	1.11833	0.161346
156	CTCF	1.117496	0.16027
157	ADR1	1.107553	0.147376
158	MSN4	1.104662	0.143605
159	Kr	1.103068	0.141521
160	YAP6	1.098523	0.135565
161	HMG-1	1.082991	0.115022
162	GAT1	1.082448	0.114298
163	THI2	1.082207	0.113976
164	RGM1	1.081687	0.113284
165	RXRA::VDR	1.077706	0.107964
166	HLF	1.071744	0.09996
167	MZF1_5-13	1.065885	0.092052
168	HAP5	1.062988	0.088125
169	GZF3	1.060588	0.084865
170	mab-3	1.05847	0.08198
171	GCR2	1.044876	0.063332
172	MET32	1.043414	0.061312
173	Tcfcp2l	1.037517	0.053134
174	Ar	1.037327	0.052871
175	Oc	1.026213	0.037331
176	Pax2	1.025912	0.036907
177	ARR10	1.025577	0.036436
178	Spz1	1.024094	0.034348
179	Myb,Ph3	1.02281	0.032538
180	INSM1	1.019388	0.027704
181	EBF1	1.008873	0.012744
182	GATA2	1.007664	0.011014
183	FEV	1.007177	0.010317
184	GCR1	1.007019	0.010091
185	Zfp423	1.003631	0.005229
186	Stat3	1.00052	0.00075

187	Dof3	0.999948	-7.5E-05
188	BRCA1	0.99944	-0.00081
189	Ovo	0.997185	-0.00407
190	Prd	0.997185	-0.00407
191	MCM1	0.99706	-0.00425
192	Bcd	0.989191	-0.01568
193	ACE2	0.987584	-0.01802
194	ELF5	0.985656	-0.02084
195	STAT1	0.984537	-0.02248
196	Ptx1	0.982087	-0.02608
197	FOXC1	0.981429	-0.02704
198	NFYA	0.978348	-0.03158
199	RFX1	0.978094	-0.03196
200	Ct	0.977359	-0.03304
201	Gsc	0.977321	-0.0331
202	RPH1	0.977237	-0.03322
203	GLN3	0.970655	-0.04297
204	ZNF354C	0.967976	-0.04696
205	YER130C	0.967462	-0.04772
206	SRD1	0.967144	-0.0482
207	STE12	0.965965	-0.04996
208	MNB1A	0.960696	-0.05785
209	TBP	0.960309	-0.05843
210	ECM23	0.959158	-0.06016
211	REI1	0.958428	-0.06126
212	GIS1	0.957122	-0.06323
213	RME1	0.956455	-0.06423
214	SPI1	0.955369	-0.06587
215	PBF	0.951112	-0.07231
216	Dof2	0.943387	-0.08408
217	RXR::RAR_DR5	0.940224	-0.08892
218	GAT3	0.940046	-0.0892
219	NR3C1	0.935855	-0.09564
220	TBF1	0.935832	-0.09568
221	NRG1	0.933248	-0.09967
222	REST	0.927506	-0.10857
223	ESR1	0.926872	-0.10956
224	MAC1	0.926008	-0.1109
225	HNF4A	0.921543	-0.11788
226	NDT80	0.919688	-0.12078
227	GATA3	0.916015	-0.12656
228	ABF2	0.915085	-0.12802
229	ZEB1	0.906649	-0.14138
230	TEAD1	0.906557	-0.14153
231	Z	0.904816	-0.1443
232	CIN5	0.904459	-0.14487
233	NFIC	0.902247	-0.14841
234	GAT4	0.90206	-0.14871
235	FOXF2	0.901814	-0.1491
236	Hb	0.901181	-0.15011

237	Sd	0.895248	-0.15964
238	B-H1	0.893907	-0.1618
239	AFT1	0.883846	-0.17813
240	TEC1	0.881578	-0.18184
241	NR4A2	0.881162	-0.18252
242	Twi	0.878793	-0.18641
243	Esrrb	0.877395	-0.1887
244	AZF1	0.873484	-0.19515
245	AG	0.872391	-0.19695
246	SPT23	0.871138	-0.19903
247	HAP2	0.870815	-0.19956
248	PHD1	0.866314	-0.20704
249	En1	0.864821	-0.20953
250	HAP4	0.864524	-0.21002
251	NFATC2	0.863911	-0.21105
252	PPARG::RXRA	0.861203	-0.21558
253	NFE2L1::MafG	0.860913	-0.21606
254	YY1	0.858569	-0.22
255	HMG-I/Y	0.855898	-0.22449
256	Nr2e3	0.853114	-0.22919
257	Tal1::Gata1	0.852445	-0.23032
258	HOXA5	0.851923	-0.23121
259	Six4	0.85152	-0.23189
260	Ddit3::Cebpa	0.850783	-0.23314
261	Hand1::Tcfe2a	0.84668	-0.24011
262	HSF1	0.842337	-0.24753
263	MOT3	0.840792	-0.25018
264	ARG81	0.840732	-0.25028
265	Nkx3-2	0.840611	-0.25049
266	SWI5	0.839811	-0.25186
267	Optix	0.834396	-0.2612
268	USV1	0.834175	-0.26158
269	Gata1	0.833243	-0.26319
270	Odd	0.831845	-0.26561
271	CEBPA	0.829446	-0.26978
272	Lag1	0.828811	-0.27089
273	Gfi	0.820878	-0.28476
274	CUP2	0.820316	-0.28575
275	SPT2	0.819234	-0.28765
276	CAD1	0.818074	-0.2897
277	ttx-3::ceh-10	0.818043	-0.28975
278	FKH2	0.816373	-0.2927
279	ZAP1	0.816319	-0.2928
280	Ceh-22	0.815917	-0.29351
281	Exd	0.815754	-0.29379
282	Ttk	0.813183	-0.29835
283	Lbe	0.811773	-0.30085
284	Pdx1	0.81122	-0.30184
285	FOXD1	0.807919	-0.30772
286	Bsh	0.806283	-0.31064

287	Prrx2	0.805953	-0.31123
288	Pax4	0.804479	-0.31387
289	HAP3	0.803722	-0.31523
290	exex	0.803363	-0.31588
291	exex	0.803363	-0.31588
292	So	0.801989	-0.31835
293	NFIL3	0.801929	-0.31845
294	Pan	0.798816	-0.32406
295	Lbl	0.797571	-0.32632
296	TLX1::NFIC	0.795705	-0.3297
297	ESR2	0.795481	-0.3301
298	SRF	0.793959	-0.33286
299	C15	0.793783	-0.33318
300	MET28	0.793099	-0.33443
301	Hth	0.792099	-0.33625
302	Sna	0.791549	-0.33725
303	Vsx2	0.789858	-0.34034
304	Tll	0.788682	-0.34248
305	Dll	0.788399	-0.343
306	SOX10	0.787032	-0.34551
307	Dr	0.786887	-0.34577
308	znf143	0.78312	-0.35269
309	Hmx	0.781982	-0.35479
310	CG11085	0.781099	-0.35642
311	Nobox	0.779845	-0.35874
312	OdsH	0.779406	-0.35955
313	OdsH	0.779406	-0.35955
314	PPARG	0.779173	-0.35998
315	Nkx2-5	0.778513	-0.36121
316	SFL1	0.778004	-0.36215
317	MGA1	0.777967	-0.36222
318	Bap	0.777478	-0.36313
319	HMRA2	0.775859	-0.36613
320	RORA_1	0.775637	-0.36655
321	RAP1	0.775131	-0.36749
322	Ara	0.774531	-0.36861
323	Ro	0.774263	-0.3691
324	caup	0.774212	-0.3692
325	Slbo	0.774032	-0.36953
326	AP1	0.772817	-0.3718
327	Eve	0.772096	-0.37315
328	FZF1	0.771968	-0.37339
329	ARID3A	0.771288	-0.37466
330	Hltf	0.771272	-0.37469
331	PEND	0.769366	-0.37826
332	br_Z2	0.768769	-0.37938
333	NR2F1	0.768729	-0.37945
334	CG9876	0.767968	-0.38088
335	unpg	0.767372	-0.382
336	Zen2	0.76732	-0.3821

337	TOS8	0.767155	-0.38241
338	CG7056	0.767111	-0.38249
339	br_Z3	0.766466	-0.38371
340	Mirr	0.765766	-0.38502
341	Vvl	0.765375	-0.38576
342	Pph13	0.764999	-0.38647
343	Ind	0.764898	-0.38666
344	E5	0.763649	-0.38902
345	Tin	0.763419	-0.38945
346	Abd-B	0.763131	-0.39
347	Kni	0.762726	-0.39076
348	SOX9	0.762714	-0.39079
349	NKX3-1	0.762674	-0.39086
350	Otp	0.762663	-0.39088
351	RUNX1	0.761971	-0.39219
352	PHO2	0.761901	-0.39233
353	repo	0.761351	-0.39337
354	Vsx1	0.760108	-0.39572
355	Lim3	0.759088	-0.39766
356	YOX1	0.756856	-0.40191
357	Unc-4	0.75665	-0.4023
358	En	0.75656	-0.40247
359	CG11294	0.756403	-0.40277
360	Al	0.756349	-0.40288
361	Slou	0.755834	-0.40386
362	B-H2	0.755829	-0.40387
363	Lim1	0.755637	-0.40423
364	Ap	0.755395	-0.4047
365	PHDP	0.755173	-0.40512
366	Rx	0.754463	-0.40648
367	Abd-A	0.754254	-0.40688
368	CG32532	0.754188	-0.407
369	NK7,1	0.75418	-0.40702
370	CG13424	0.754033	-0.4073
371	ems	0.753489	-0.40834
372	Awh	0.753093	-0.4091
373	CG18599	0.752838	-0.40959
374	Hbn	0.75187	-0.41145
375	CG34031	0.751191	-0.41275
376	Btn	0.750557	-0.41397
377	CG32105	0.750016	-0.41501
378	HGTX	0.749937	-0.41516
379	Lab	0.749861	-0.41531
380	H2,0	0.748483	-0.41796
381	CG15696	0.748377	-0.41816
382	Oct	0.748362	-0.41819
383	Tup	0.748112	-0.41867
384	Inv	0.747983	-0.41892
385	Dfd	0.747457	-0.41994
386	HCM1	0.747239	-0.42036

387	Zen	0.747119	-0.42059
388	AGL3	0.746098	-0.42256
389	Vis	0.745854	-0.42304
390	Pb	0.745665	-0.4234
391	INO2	0.74548	-0.42376
392	YAP5	0.745354	-0.424
393	Cad	0.74513	-0.42444
394	Myf	0.744634	-0.4254
395	NR1H2::RXRA	0.74363	-0.42734
396	Ftz	0.743116	-0.42834
397	Ubx	0.741362	-0.43175
398	YHP1	0.740487	-0.43345
399	Antp	0.739659	-0.43507
400	FOXL1	0.739088	-0.43618
401	Scr	0.738591	-0.43715
402	SIG1	0.738442	-0.43744
403	CG42234	0.737426	-0.43943
404	SOK2	0.734336	-0.44549
405	Vnd	0.734184	-0.44579
406	FOXO3	0.733467	-0.4472
407	ROX1	0.732667	-0.44877
408	RREB1	0.73109	-0.45188
409	achi	0.730567	-0.45291
410	CG4328	0.730247	-0.45354
411	MET4	0.729513	-0.45499
412	SFP1	0.727421	-0.45914
413	HNF1A	0.727287	-0.4594
414	SUM1	0.726758	-0.46045
415	ARR1	0.724973	-0.464
416	slp1	0.724276	-0.46539
417	YAP7	0.723321	-0.46729
418	onecut	0.722886	-0.46816
419	T	0.722109	-0.46971
420	Sox17	0.721179	-0.47157
421	ATHB-5	0.719688	-0.47456
422	NFE2L2	0.719591	-0.47475
423	MEF2A	0.717025	-0.4799
424	Sox5	0.716026	-0.48192
425	SRY	0.712366	-0.48931
426	HNF1B	0.710516	-0.49306
427	br_Z4	0.710085	-0.49394
428	HAT5	0.703822	-0.50672
429	Evi1	0.70012	-0.51433
430	CUP9	0.690815	-0.53363
431	INO4	0.688644	-0.53817
432	Foxd3	0.688115	-0.53928
433	D	0.6856	-0.54456
434	RORA_2	0.685247	-0.5453
435	RLM1	0.68332	-0.54937
436	Lhx3	0.68327	-0.54947

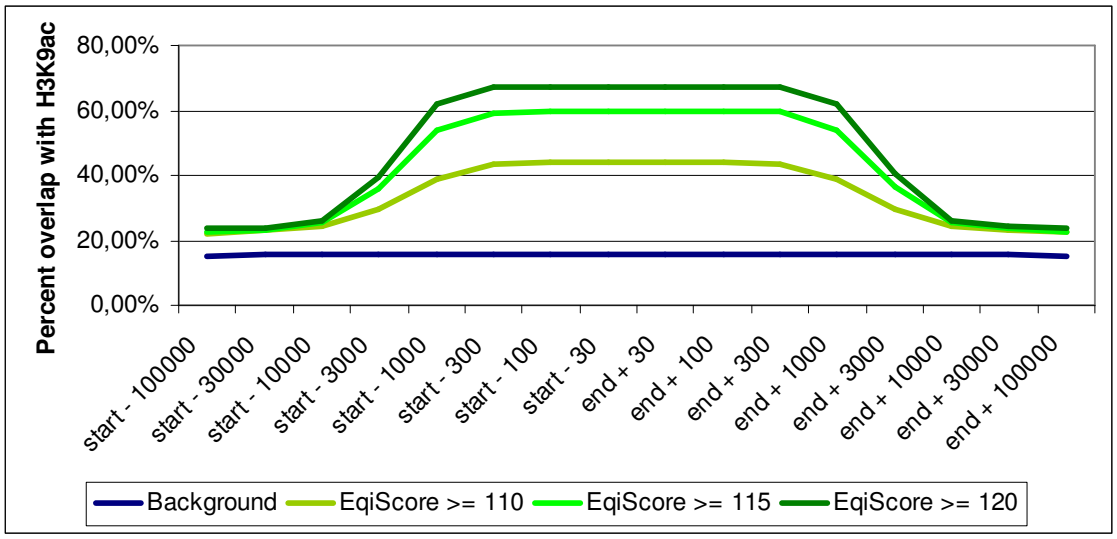
437	CG11617	0.683107	-0.54982
438	MATA1	0.683069	-0.5499
439	TAL1::TCF3	0.682308	-0.55151
440	STB3	0.680075	-0.55623
441	squamosa	0.676005	-0.56489
442	br_Z1	0.666923	-0.58441
443	Cf2_II	0.659881	-0.59972
444	GCN4	0.654738	-0.61101
445	Fkh	0.653439	-0.61388
446	SMP1	0.647859	-0.62625
447	MATALPHA2	0.646171	-0.63001
448	FOXI1	0.644423	-0.63392
449	FOXA1	0.63746	-0.64959
450	FKH1	0.63578	-0.6534
451	Foxa2	0.628975	-0.66893
452	Foxq1	0.627	-0.67346
453	Sox2	0.617805	-0.69478
454	PBX1	0.612739	-0.70666
455	Nub	0.610197	-0.71265
456	TBP	0.609175	-0.71507
457	NHP6B	0.569872	-0.81129
458	NHP6A	0.538871	-0.89199
459	Pou5f1	0.532767	-0.90842

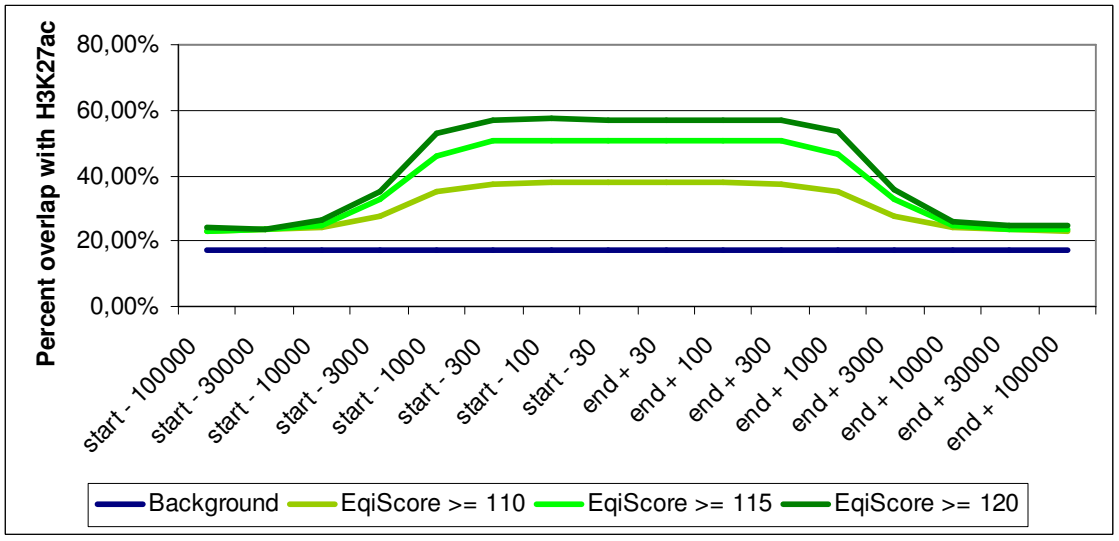
Appendix E – ALU consensus sequence

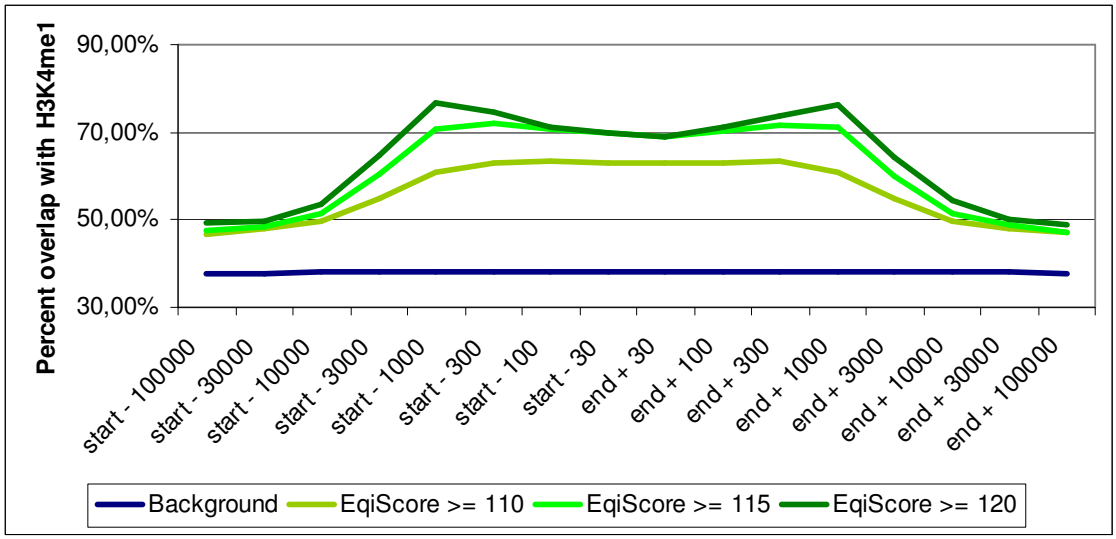
ALU consensus sequence (RepBase 14.05):

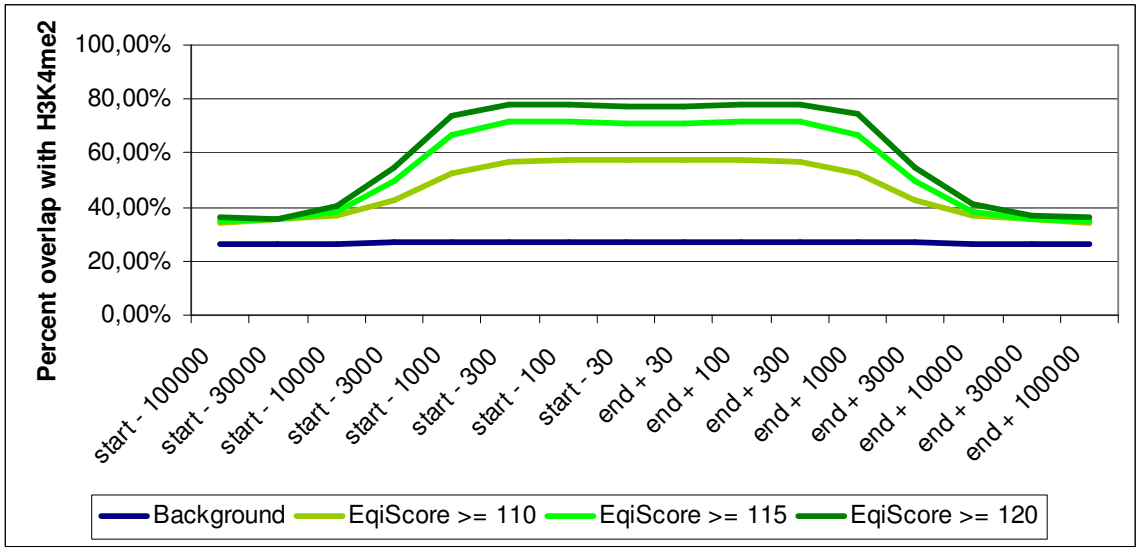
```
ggccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccgaggcgggagg
attgcttgagcccaggagttcgagaccagcctgggcaacatagcgagaccccgtctcta
caaaaaatacaaaaattagccgggcgtggtggcgcgcgcctgtagtcccagctactcgg
gaggctgaggcaggaggatcgcttgagcccaggagttcgaggctgcagtgagctatgat
cgcgccactgcactccagcctgggcgacagagcgagaccctgtctcaaaaaaaaaaaaa
aaaaaaaaaaaaaaaa
```

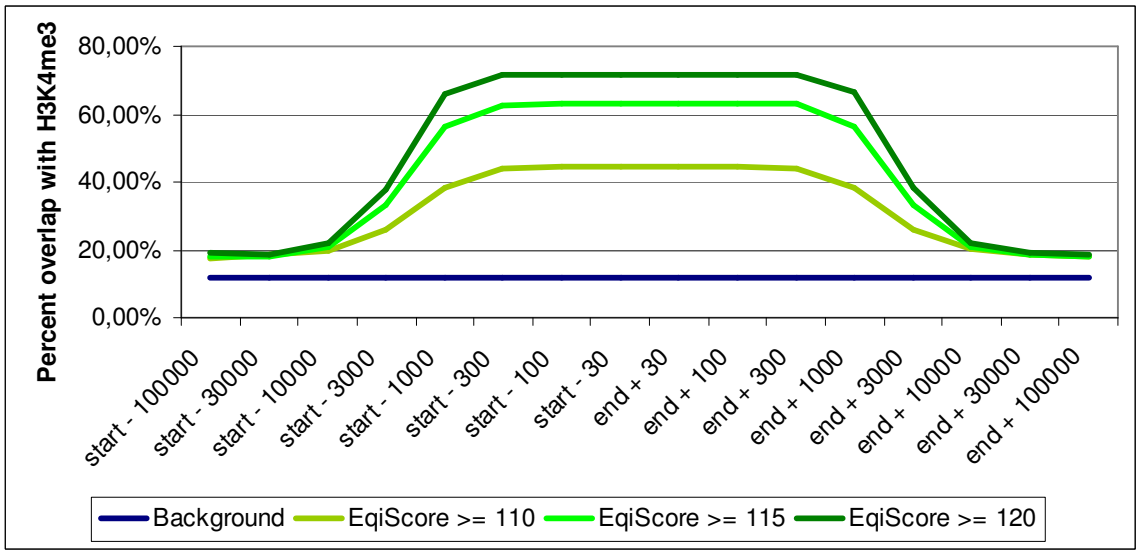
Appendix F – Epigenetic neighborhood of genome regions with high EqiScore

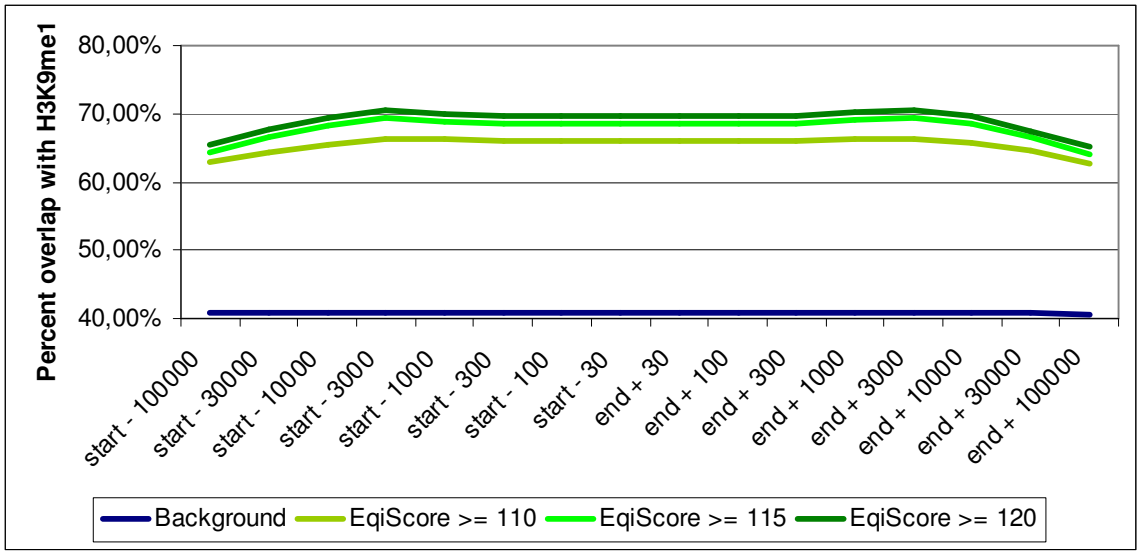


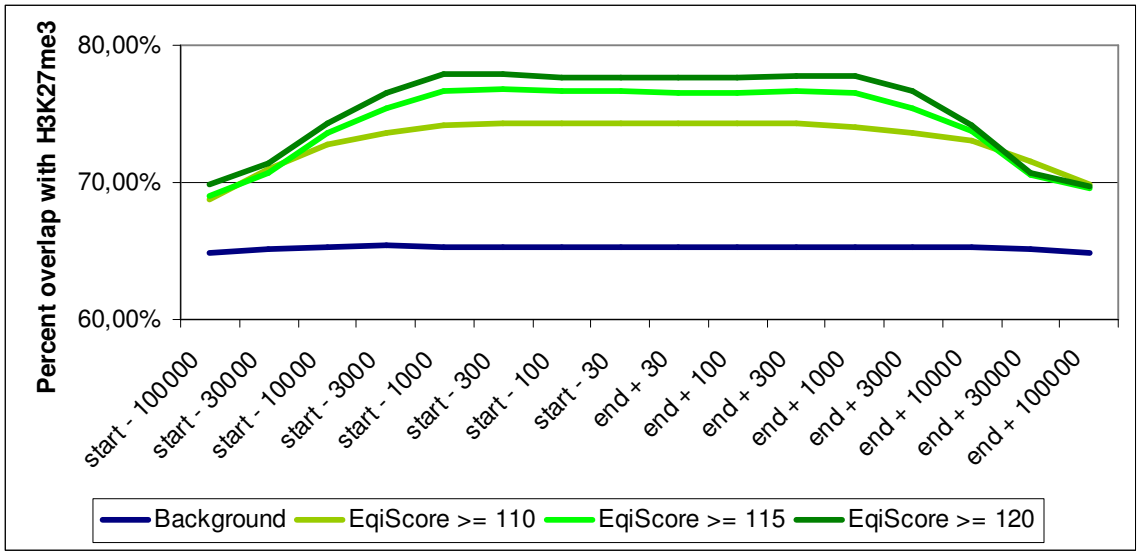


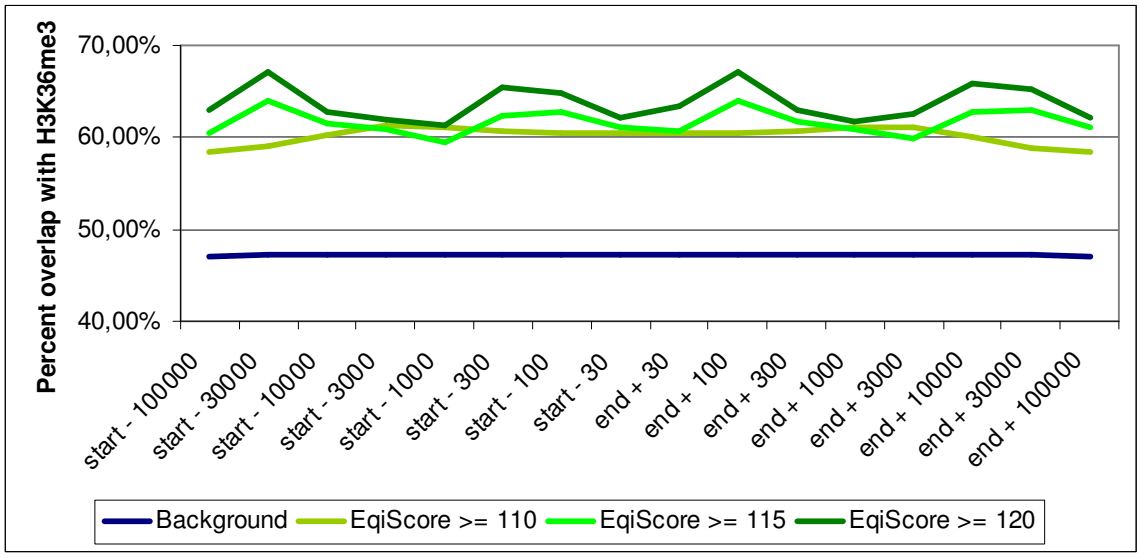

Percent overlap with H3K4me1
90,00%
70,00%
50,00%
30,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120

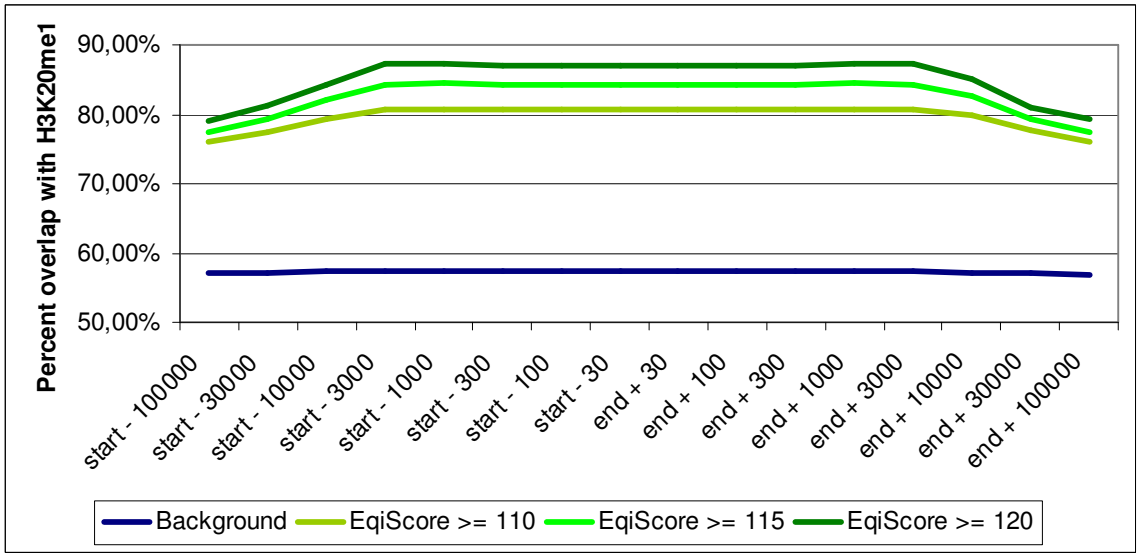

Percent overlap with H3K4me2
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120


Percent overlap with H3K4me3
80,00%
60,00%
40,00%
20,00%
0,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120


Percent overlap with H3K9me1
80,00%
70,00%
60,00%
50,00%
40,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120


Percent overlap with H3K27me3
80,00%
70,00%
60,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120


Percent overlap with H3K36me3
70,00%
60,00%
50,00%
40,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120


Percent overlap with H3K20me1
90,00%
80,00%
70,00%
60,00%
50,00%
start - 100000
start - 30000
start - 10000
start - 3000
start - 1000
start - 300
start - 100
start - 30
end + 30
end + 100
end + 300
end + 1000
end + 3000
end + 10000
end + 30000
end + 100000
Background
EqiScore >= 110
EqiScore >= 115
EqiScore >= 120

Appendix G - Finite State Continuous Time Markov Chains

Markov chains

A number of stochastic problems in genome research can be approached by modeling them as markov chains. This conceptual tool is very versatile and can be applied in a variety of subtypes, which have been discussed in depth before (Karlin&Taylor;1975).

The *Finite State Continuous Time Markov Chains* subtype is central for many models of DNA evolution, and therefore, the most important properties will be summarized below following p150-152 in (Karlin&Taylor;1975).

Considering a finite state space *S* with labels 0 to *N* the continuous time Markov chain X_t ($t > 0$) is described by the transition probability matrix *P* with *N+1* rows and columns. This transition probabilities are stationary. Thus, $P_{ij}(t)$ denotes the probability that $X_{t+s}=j$ if $X_s=i$. In matrix notation we have $P(t+s)=P(t)P(s)$ for $t,s >0$.
This implies that evolution is a stable stochastic process.

Under this assumption P(t) is continuous and can be described by its infinitesimal matrix Q that satisfies the differential equation $P'(t) = P(t)Q= QP(t)$. Under the condition that P(0) equals the identity matrix I, a system of ordinary differential equations can be applied to compute *P(t)* by e^{Qt}.

In other words, if we consider that state changes can not appear instantaneously, but at least take an infinitesimal small amount of time, the matrix Q fully describes *P(t)*.

Finally, from the properties of the model follows that $\sum_{j=0, j\neq i}^{N} q_{ij} = -q_{ii}$. Thus, from the rates that describe the state changes, we can compute the rate that describes the conservation of the state. It is worthwhile to note that the relative difference between the elements in *Q* characterize *P(t)* and that the scaling of their absolute values only impacts the unit in which *t* is measures.

Appendix H – Jonit probability of backmutation and conservation of CpG, TpG, CpA and TpA are independent from neighboring nucleotides

Substitution models: We call the substitution model described in the main text Q2-model, as its rate matrix pertains to dinucleotides. Its most important features are that (i) all transversion rates are expected to be equal and (ii) the only neighborhood-dependent process is the CpG decay effect (CpG -> TpG/CpA).

Central model property: The probability that a CpG is still (or again) a CpG after time $t > 0$ has passed is independent from genomic neighborhood of the dinucleotide.

Proof:

1) The CpG dinucleotide is not influenced by its neighborhood, i.e. for XCGY, with X and Y being any of the four nucleotides A,C,G,T, neither X nor Y can be selected in a way to form a CpG that overlaps with the central CpG.

2) By only applying transitions XCGY can be mutated into XTGY, XCAY and XTAY. These four patterns form a *transition group*. For none of these X or Y can be selected such a way that a novel CG is formed, which overlaps with the central dinucleotide. Hence, within a context-independent transition group, the probability of conserving or reestablishing the original state of the central dinucleotide is independent from the context (*context-independant transition group*).

3) There are three additional *transition groups*. The first group is reached by a transversion in the first position: XAGY, XGGY, XAAY and XGAY. The second group is reached by a transversion in the second position: XCCY, XTCY, XCTY and XTTY. The third group is reached by transversions in both positions: XACY, XGCY, XATY and XGTY.
Within these *transition groups* only those positions that underwent a transversion (first nucleotide in first group 2, second nucleotide in second group and both nucleotides in the third group) , can be influenced by a neighboring nucleotide. Therefore, these three groups are *context-dependant transition groups*. Furthermore, the two central positions cannot influence each other, as no group member contains a central CpG. To return, to the original transition group, the nucleotide that can be influence by a neighbor, has to undergo a transversion. The probability of this event is independent from the transition state of this nucleotide (C to A transversion has equal probability as T to A transversion etc.). Hence, the probability to leave and return to a *context-independant transition group* is independent of the neighboring nucleotides and always equals an even number of transversions.

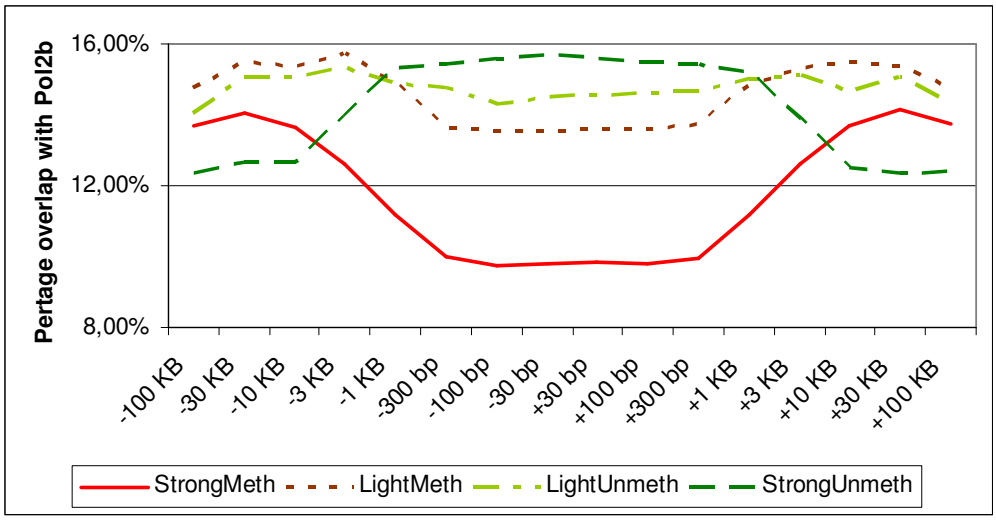
Neighter conservation 1) nor any chain of transitions 2) and transversions 3) enable a neighboring nucleotide to influence the probability that a CG is a CG after an arbitrary time interval *q.e.d.*

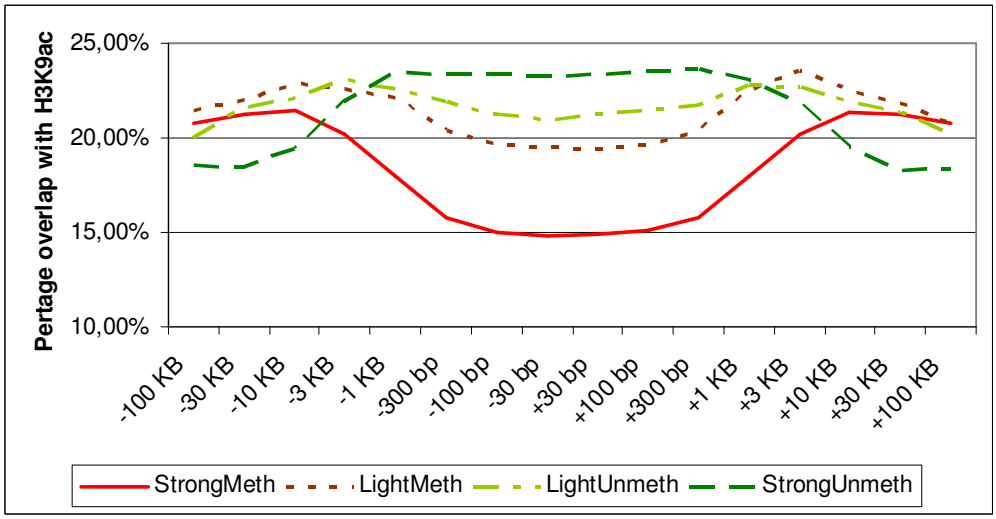
Corollary:

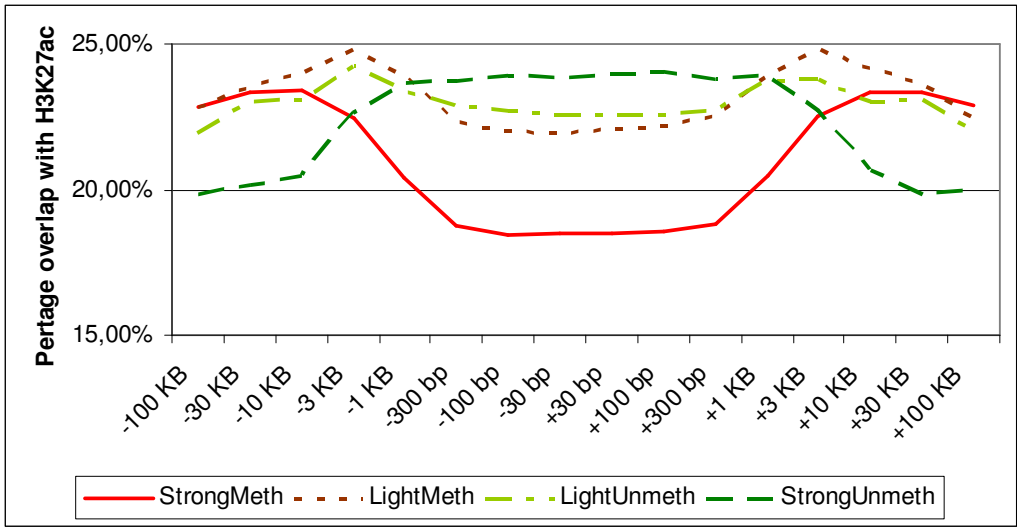
The probability that a CpA, TpG or TpA is still (or again) a CpA, TpG or TpA, respectively, after time $t > 0$ has passed is independent from genomic neighborhood of the dinucleotide.

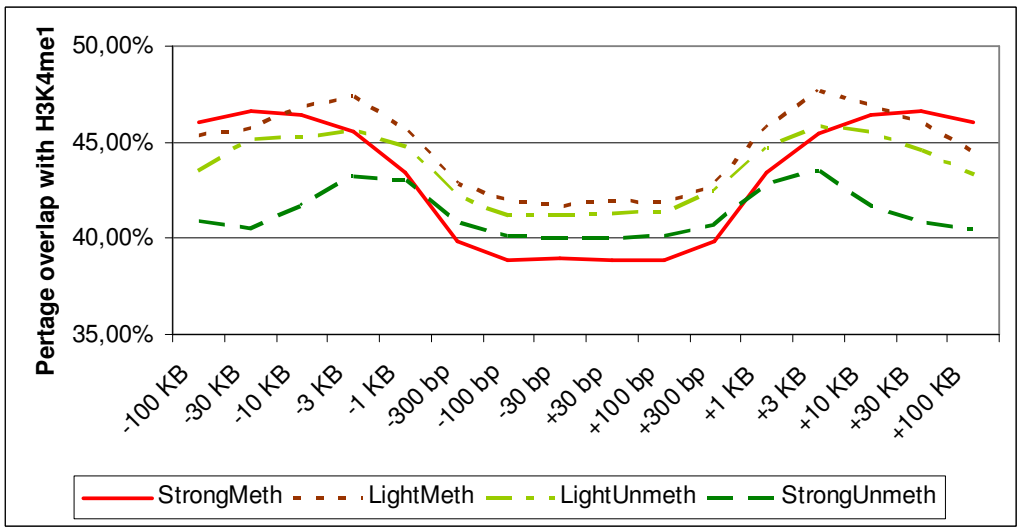
Appendix I – Epigenetic neighborhood of L-scored ALU repeats

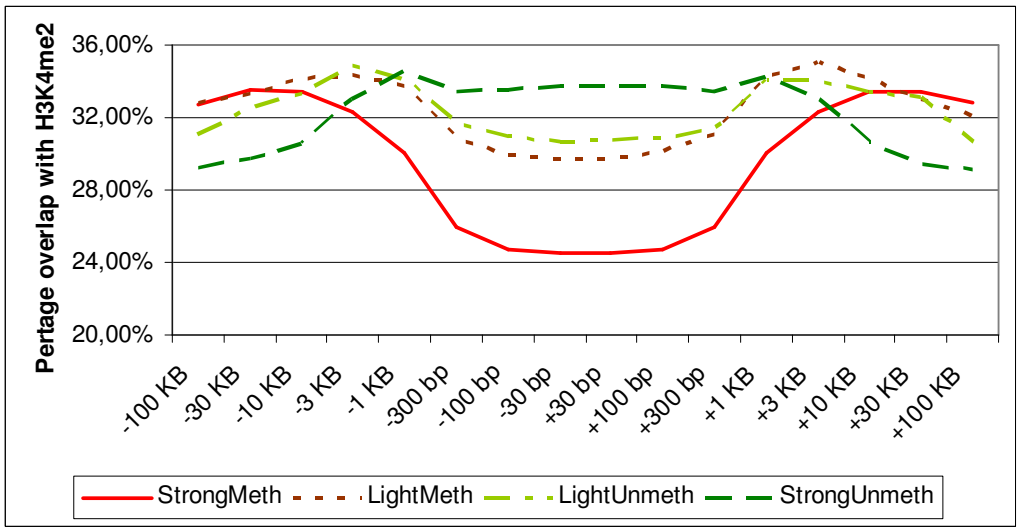
Overlap of RNA polymerase II binding sites and histone modifications with AluJudge predictions. Diagrams were generated from the output of the neighborhood analysis of the EpiExplorer tools (compare section 2.4.1).

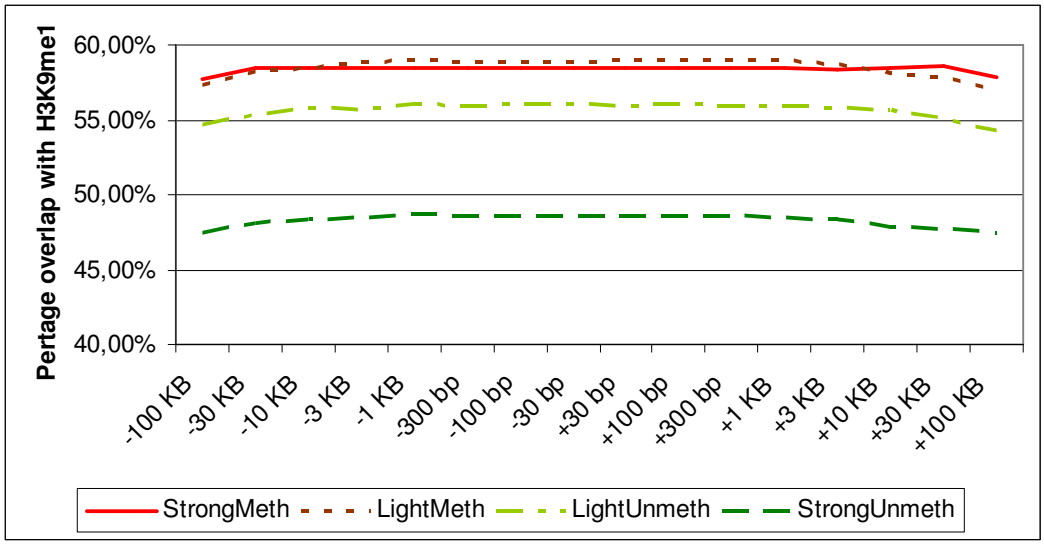


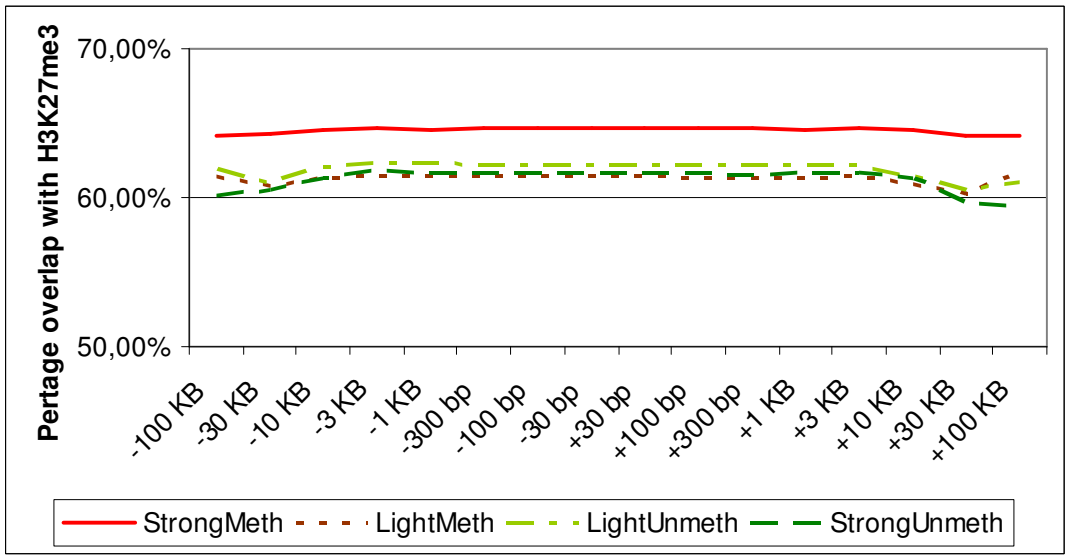


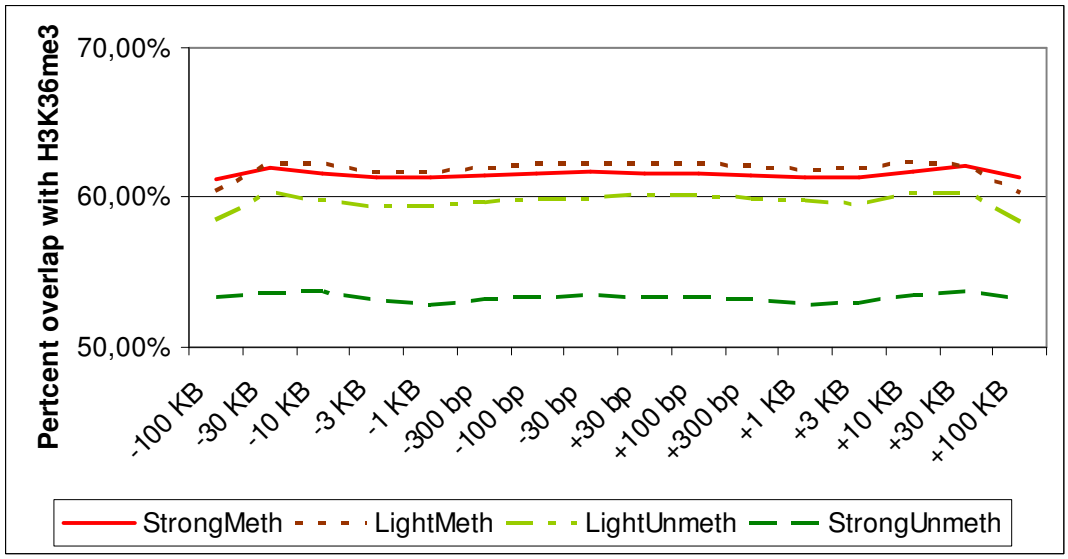

Pertage overlap with H3K27ac
25,00%
20,00%
15,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth

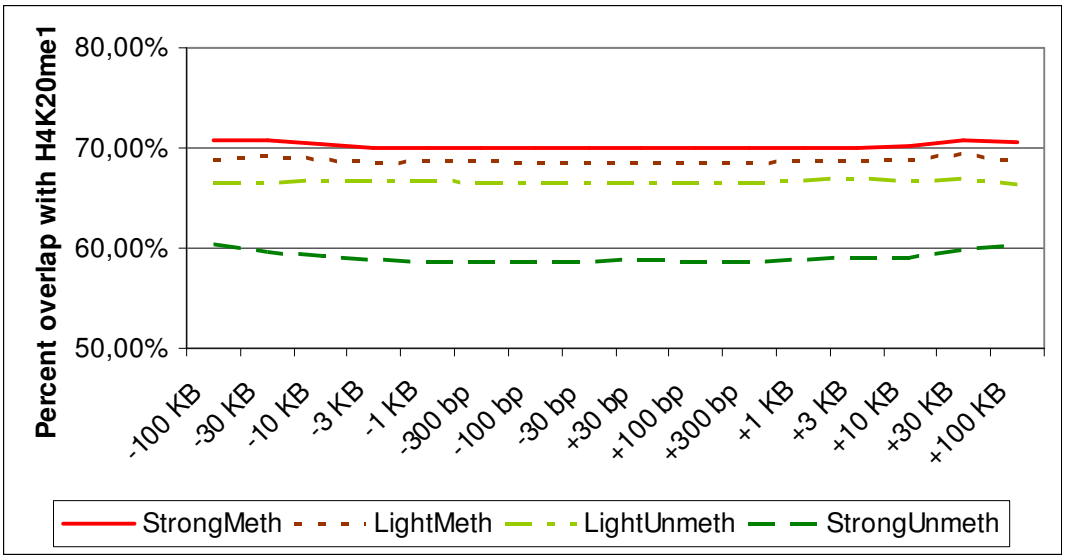

Pertage overlap with H3K4me1
50,00%
45,00%
40,00%
35,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth


Pertage overlap with H3K4me2
36,00%
32,00%
28,00%
24,00%
20,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth


Pertage overlap with H3K9me1
60,00%
55,00%
50,00%
45,00%
40,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth


Pertage overlap with H3K27me3
70,00%
60,00%
50,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth


Pertcent overlap with H3K36me3
70,00%
60,00%
50,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth


Percent overlap with H4K20me1
80,00%
70,00%
60,00%
50,00%
-100 KB
-30 KB
-10 KB
-3 KB
-1 KB
-300 bp
-100 bp
-30 bp
+30 bp
+100 bp
+300 bp
+1 KB
+3 KB
+10 KB
+30 KB
+100 KB
StrongMeth
LightMeth
LightUnmeth
StrongUnmeth

Appendix J – Predicted and measured methylation of human promoters

This table contains the database of Figure 5.11. Additionally the coordinates and the names of the associated genes of the promoters are given.

Gene	Chrom	From	To	Strand	Predicted	Measured	Difference
A2M	chr12	9,159,092	9,162,092	-	50.00%	61.10%	-11.10%
ABCC4	chr13	94,750,688	94,753,688	-	10.00%	2.79%	7.21%
ABHD3	chr18	17,537,764	17,540,764	-	20.00%	4.16%	15.84%
ACACB	chr12	108,036,783	108,039,783	+	100.00%	81.98%	18.02%
ACTR10	chr14	57,734,551	57,737,551	+	0.00%	4.04%	-4.04%
ACVR1B	chr12	50,629,718	50,632,718	+	0.00%	5.30%	-5.30%
AICDA	chr12	8,655,734	8,658,734	-	100.00%	88.19%	11.81%
AKAP3	chr12	4,627,474	4,630,474	-	0.00%	3.69%	-3.69%
ALAS2	chrX	55,073,222	55,076,222	-	100.00%	86.96%	13.04%
ALS2	chr2	202,353,157	202,356,157	-	20.00%	3.44%	16.56%
AMMECR1	chrX	109,569,117	109,572,117	-	0.00%	87.66%	-87.66%
AMOT	chrX	111,969,699	111,972,699	-	0.00%	1.29%	-1.29%
ANAPC5	chr12	120,321,082	120,324,082	-	10.00%	3.01%	6.99%
ANKRD5	chr20	9,961,689	9,964,689	+	50.00%	8.42%	41.58%
ANKRD9	chr14	102,044,889	102,047,889	-	10.00%	6.08%	3.92%
AP4S1	chr14	30,562,063	30,565,063	+	20.00%	3.14%	16.86%
APOBEC2	chr6	41,127,021	41,130,021	+	90.00%	89.15%	0.85%
APOD	chr3	196,791,365	196,794,365	-	100.00%	81.97%	18.03%
AQP2	chr12	48,628,791	48,631,791	+	30.00%	89.85%	-59.85%
ARF6	chr14	49,427,560	49,430,560	+	0.00%	2.21%	-2.21%
ARHGAP5	chr14	31,613,071	31,616,071	+	0.00%	6.36%	-6.36%
ARPC2	chr2	218,788,062	218,791,062	+	10.00%	4.67%	5.33%
ARVCF	chr22	18,383,331	18,386,331	-	10.00%	5.87%	4.13%
ASB1	chr2	238,998,122	239,001,122	+	10.00%	3.34%	6.66%
ASB13	chr10	5,747,774	5,750,774	-	0.00%	6.04%	-6.04%
ASB8	chr12	46,860,263	46,863,263	-	60.00%	32.39%	27.61%
ASB9	chrX	15,197,510	15,200,510	-	50.00%	50.65%	-0.65%
ASPH	chr8	62,788,753	62,791,753	-	0.00%	3.48%	-3.48%
ATP2A2	chr12	109,200,944	109,203,944	+	0.00%	6.89%	-6.89%
ATP5G2	chr12	52,356,459	52,359,459	-	0.00%	3.36%	-3.36%
ATP6V0D2	chr8	87,066,668	87,069,668	+	30.00%	76.48%	-46.48%
ATP6V1H	chr8	54,917,671	54,920,671	-	10.00%	6.20%	3.80%
ATP7A	chrX	77,050,850	77,053,850	+	0.00%	6.77%	-6.77%
ATP8B4	chr15	48,261,306	48,264,306	-	0.00%	11.48%	-11.48%
B3GAT2	chr6	71,722,462	71,725,462	-	10.00%	2.33%	7.67%
BACH2	chr6	91,062,348	91,065,348	-	0.00%	2.24%	-2.24%
BAI3	chr6	69,399,980	69,402,980	+	0.00%	2.08%	-2.08%
BARD1	chr2	215,381,673	215,384,673	-	10.00%	1.72%	8.28%
BATF	chr14	75,056,521	75,059,521	+	0.00%	86.98%	-86.98%
BCOR	chrX	39,920,526	39,923,526	-	0.00%	9.77%	-9.77%
BGN	chrX	152,411,591	152,414,591	+	40.00%	82.10%	-42.10%

BICD2	chr9	94,565,915	94,568,915	-	0.00%	3.49%	-3.49%
BMP2	chr20	6,694,311	6,697,311	+	0.00%	1.93%	-1.93%
BMPR2	chr2	202,947,904	202,950,904	+	0.00%	1.16%	-1.16%
BMX	chrX	15,390,290	15,393,290	+	10.00%	86.18%	-76.18%
BRCA2	chr13	31,785,611	31,788,611	+	10.00%	5.02%	4.98%
BRF2	chr8	37,825,617	37,828,617	-	30.00%	6.72%	23.28%
BRP44L	chr6	166,715,476	166,718,476	-	0.00%	2.05%	-2.05%
BRS3	chrX	135,395,712	135,398,712	+	100.00%	87.75%	12.25%
BTBD3	chr20	11,817,371	11,820,371	+	0.00%	3.82%	-3.82%
BTK	chrX	100,526,839	100,529,839	-	30.00%	80.18%	-50.18%
CABP1	chr12	119,560,738	119,563,738	+	0.00%	9.75%	-9.75%
CACNA1C	chr12	1,948,213	1,951,213	+	80.00%	88.99%	-8.99%
CACNG3	chr16	24,172,375	24,175,375	+	10.00%	14.23%	-4.23%
CAPN6	chrX	110,399,407	110,402,407	-	30.00%	56.82%	-26.82%
CAPZA2	chr7	116,236,360	116,239,360	+	100.00%	86.20%	13.80%
CARD10	chr22	36,244,495	36,247,495	-	10.00%	11.71%	-1.71%
CASP8	chr2	201,804,411	201,807,411	+	70.00%	14.39%	55.61%
CBLL1	chr7	107,169,378	107,172,378	+	10.00%	2.29%	7.71%
CBLN2	chr18	68,361,754	68,364,754	-	0.00%	5.57%	-5.57%
CCR6	chr6	167,443,285	167,446,285	+	90.00%	90.48%	-0.48%
CD28	chr2	204,277,443	204,280,443	+	90.00%	84.28%	5.72%
CD2AP	chr6	47,551,484	47,554,484	+	10.00%	2.19%	7.81%
CD33	chr19	56,418,132	56,421,132	+	70.00%	85.01%	-15.01%
CD9	chr12	6,177,142	6,180,142	+	10.00%	12.50%	-2.50%
CDC5L	chr6	44,461,240	44,464,240	+	20.00%	5.12%	14.88%
CDH17	chr8	95,297,707	95,300,707	-	40.00%	35.31%	4.69%
CDH20	chr18	57,149,968	57,152,968	+	20.00%	4.31%	15.69%
CDH7	chr18	61,566,468	61,569,468	+	10.00%	3.80%	6.20%
CDK8	chr13	25,724,276	25,727,276	+	30.00%	2.40%	27.60%
CDKL1	chr14	49,951,929	49,954,929	-	100.00%	87.12%	12.88%
CDKN3	chr14	53,931,317	53,934,317	+	0.00%	4.14%	-4.14%
CDX1	chr5	149,524,552	149,527,552	+	0.00%	11.41%	-11.41%
CECR6	chr22	15,981,257	15,984,257	-	0.00%	4.33%	-4.33%
CFL2	chr14	34,252,780	34,255,780	-	0.00%	4.36%	-4.36%
CFTR	chr7	116,891,074	116,894,074	+	50.00%	84.21%	-34.21%
CGA	chr6	87,860,543	87,863,543	-	70.00%	78.42%	-8.42%
CHGA	chr14	92,457,178	92,460,178	+	20.00%	5.33%	14.67%
CHGB	chr20	5,838,076	5,841,076	+	0.00%	5.38%	-5.38%
CHM	chrX	85,188,222	85,191,222	-	0.00%	3.47%	-3.47%
CHRDL1	chrX	109,924,942	109,927,942	-	10.00%	7.45%	2.55%
CHST8	chr19	38,802,701	38,805,701	+	10.00%	4.77%	5.23%
CIT	chr12	118,798,478	118,801,478	-	10.00%	12.11%	-2.11%
CITED1	chrX	71,442,762	71,445,762	-	0.00%	1.35%	-1.35%
CKS2	chr9	91,113,933	91,116,933	+	10.00%	3.31%	6.69%
CLCN5	chrX	49,571,965	49,574,965	+	0.00%	7.26%	-7.26%
CLDN1	chr3	191,521,958	191,524,958	-	0.00%	14.68%	-14.68%
CLDN2	chrX	106,028,050	106,031,050	+	100.00%	86.55%	13.45%
CLDN5	chr22	17,894,068	17,897,068	-	40.00%	82.07%	-42.07%
CLIC5	chr6	46,155,091	46,158,091	-	100.00%	91.64%	8.36%
CLTB	chr5	175,775,176	175,778,176	-	0.00%	8.78%	-8.78%

CMKLR1	chr12	107,256,248	107,259,248	-	20.00%	41.78%	-21.78%
CNGA3	chr2	98,327,050	98,330,050	+	10.00%	8.09%	1.91%
CNGB3	chr8	87,824,019	87,827,019	-	100.00%	86.15%	13.85%
CNR1	chr6	88,931,797	88,934,797	-	0.00%	1.46%	-1.46%
COL19A1	chr6	70,631,184	70,634,184	+	0.00%	2.90%	-2.90%
COL2A1	chr12	46,683,536	46,686,536	-	0.00%	2.93%	-2.93%
COL3A1	chr2	189,545,291	189,548,291	+	0.00%	26.00%	-26.00%
COL4A4	chr2	227,736,073	227,739,073	-	0.00%	2.61%	-2.61%
COL5A2	chr2	189,751,850	189,754,850	-	50.00%	26.11%	23.89%
COL9A1	chr6	71,068,507	71,071,507	-	50.00%	81.59%	-31.59%
CPNE3	chr8	87,564,175	87,567,175	+	100.00%	89.84%	10.16%
CPNE5	chr6	36,914,756	36,917,756	-	0.00%	3.69%	-3.69%
CREM	chr10	35,453,725	35,456,725	+	30.00%	5.49%	24.51%
CRISP2	chr6	49,788,258	49,791,258	-	0.00%	15.88%	-15.88%
CRYBA2	chr2	219,565,387	219,568,387	-	0.00%	7.69%	-7.69%
CSMD1	chr8	4,838,902	4,841,902	-	0.00%	2.54%	-2.54%
CSNK1A1	chr5	148,910,200	148,913,200	-	10.00%	4.67%	5.33%
CSTF2	chrX	99,960,040	99,963,040	+	10.00%	3.96%	6.04%
CSTL1	chr20	23,366,322	23,369,322	+	100.00%	88.96%	11.04%
CTSG	chr14	24,114,306	24,117,306	-	50.00%	61.93%	-11.93%
CUBN	chr10	17,210,836	17,213,836	-	70.00%	30.58%	39.42%
CUL2	chr10	35,418,576	35,421,576	-	0.00%	15.33%	-15.33%
CXCL14	chr5	134,941,868	134,944,868	-	10.00%	9.28%	0.72%
CXCR3	chrX	70,754,092	70,757,092	-	80.00%	58.18%	21.82%
CYP20A1	chr2	203,809,908	203,812,908	+	0.00%	2.87%	-2.87%
CYP27A1	chr2	219,352,716	219,355,716	+	20.00%	9.95%	10.05%
CYP2S1	chr19	46,388,955	46,391,955	+	30.00%	12.52%	17.48%
DAAM1	chr14	58,723,117	58,726,117	+	0.00%	1.58%	-1.58%
DACH2	chrX	85,288,111	85,291,111	+	10.00%	5.71%	4.29%
DACT1	chr14	58,168,438	58,171,438	+	100.00%	62.08%	37.92%
DAPK1	chr9	89,299,963	89,302,963	+	0.00%	8.71%	-8.71%
DCT	chr13	93,928,937	93,931,937	-	60.00%	82.35%	-22.35%
DCX	chrX	110,541,259	110,544,259	-	100.00%	66.08%	33.92%
DDX3X	chrX	41,075,595	41,078,595	+	0.00%	1.74%	-1.74%
DEFB1	chr8	6,721,954	6,724,954	-	90.00%	85.01%	4.99%
DES	chr2	219,989,343	219,992,343	+	10.00%	10.39%	-0.39%
DHRS4	chr14	23,490,635	23,493,635	+	30.00%	10.71%	19.29%
DICER1	chr14	94,693,100	94,696,100	-	0.00%	7.37%	-7.37%
DIO2	chr14	79,922,853	79,925,853	-	100.00%	84.25%	15.75%
DLL1	chr6	170,440,486	170,443,486	-	0.00%	1.15%	-1.15%
DMBT1	chr10	124,308,171	124,311,171	+	80.00%	80.41%	-0.41%
DNAJB7	chr22	39,587,076	39,590,076	-	30.00%	84.16%	-54.16%
DNAJC1	chr10	22,331,704	22,334,704	-	30.00%	6.20%	23.80%
DNPEP	chr2	220,293,637	220,296,637	-	80.00%	59.76%	20.24%
DPF3	chr14	72,429,562	72,432,562	-	0.00%	3.75%	-3.75%
DPYSL3	chr5	146,868,812	146,871,812	-	20.00%	12.26%	7.74%
DSC1	chr18	26,995,817	26,998,817	-	70.00%	76.42%	-6.42%
DSC2	chr18	26,935,376	26,938,376	-	0.00%	6.57%	-6.57%
DSTN	chr20	17,496,508	17,499,508	+	10.00%	2.62%	7.38%
DTNA	chr18	30,325,252	30,328,252	+	10.00%	2.61%	7.39%

DUSP4	chr8	29,263,104	29,266,104	-	0.00%	1.17%	-1.17%
DUSP9	chrX	152,559,140	152,562,140	+	10.00%	10.34%	-0.34%
EGFL6	chrX	13,495,645	13,498,645	+	0.00%	9.49%	-9.49%
EGR1	chr5	137,827,068	137,830,068	+	0.00%	1.17%	-1.17%
EML1	chr14	99,271,783	99,274,783	+	40.00%	82.86%	-42.86%
EMP2	chr16	10,581,056	10,584,056	-	20.00%	4.93%	15.07%
ENPP5	chr6	46,245,667	46,248,667	-	0.00%	7.44%	-7.44%
EPC1	chr10	32,706,732	32,709,732	-	10.00%	20.71%	-10.71%
ERBB4	chr2	213,110,810	213,113,810	-	0.00%	2.17%	-2.17%
ERN2	chr16	23,631,322	23,634,322	-	20.00%	23.57%	-3.57%
ESRRB	chr14	75,844,710	75,847,710	+	20.00%	49.76%	-29.76%
EYA1	chr8	72,436,021	72,439,021	-	20.00%	5.55%	14.45%
FABP4	chr8	82,557,053	82,560,053	-	100.00%	85.09%	14.91%
FABP5	chr8	82,353,153	82,356,153	+	0.00%	3.23%	-3.23%
FBLN5	chr14	91,483,084	91,486,084	-	10.00%	13.67%	-3.67%
FBN1	chr15	46,724,338	46,727,338	-	0.00%	2.59%	-2.59%
FBXO25	chr8	344,428	347,428	+	20.00%	4.68%	15.32%
FGD2	chr6	37,079,400	37,082,400	+	40.00%	87.18%	-47.18%
FGF1	chr5	142,056,801	142,059,801	-	50.00%	75.68%	-25.68%
FGF13	chrX	138,131,605	138,134,605	-	20.00%	87.88%	-67.88%
FGF23	chr12	4,358,155	4,361,155	-	40.00%	83.08%	-43.08%
FGF7	chr15	47,500,585	47,503,585	+	100.00%	63.68%	36.32%
FGF9	chr13	21,141,522	21,144,522	+	0.00%	1.41%	-1.41%
FHL1	chrX	135,055,225	135,058,225	+	10.00%	2.91%	7.09%
FHL5	chr6	97,115,145	97,118,145	+	40.00%	85.41%	-45.41%
FKBP11	chr12	47,605,524	47,608,524	-	50.00%	49.00%	1.00%
FKBP5	chr6	35,803,338	35,806,338	-	10.00%	15.38%	-5.38%
FLT1	chr13	27,966,265	27,969,265	-	0.00%	2.10%	-2.10%
FN1	chr2	216,008,140	216,011,140	-	0.00%	2.62%	-2.62%
FNTB	chr14	64,521,191	64,524,191	+	10.00%	8.60%	1.40%
FOS	chr14	74,813,230	74,816,230	+	0.00%	1.13%	-1.13%
FOXA1	chr14	37,137,996	37,140,996	-	10.00%	4.75%	5.25%
FOXA2	chr20	22,513,093	22,516,093	-	10.00%	1.74%	8.26%
FOXN4	chr12	108,230,408	108,233,408	-	0.00%	1.96%	-1.96%
FOXP2	chr7	113,511,618	113,514,618	+	0.00%	1.26%	-1.26%
FOXP3	chrX	49,007,232	49,010,232	-	50.00%	89.27%	-39.27%
FOXP4	chr6	41,620,142	41,623,142	+	10.00%	2.66%	7.34%
FZD5	chr2	208,341,532	208,344,532	-	0.00%	1.83%	-1.83%
FZD7	chr2	202,605,555	202,608,555	+	0.00%	1.43%	-1.43%
GABRA3	chrX	151,369,993	151,372,993	-	10.00%	29.46%	-19.46%
GABRQ	chrX	151,555,293	151,558,293	+	0.00%	5.83%	-5.83%
GADD45G	chr9	91,407,748	91,410,748	+	0.00%	1.13%	-1.13%
GALC	chr14	87,528,762	87,531,762	-	0.00%	3.51%	-3.51%
GALR1	chr18	73,089,493	73,092,493	+	0.00%	1.72%	-1.72%
GAS1	chr9	88,750,924	88,753,924	-	0.00%	1.29%	-1.29%
GATM	chr15	43,480,708	43,483,708	-	0.00%	1.07%	-1.07%
GBX2	chr2	236,740,751	236,743,751	-	0.00%	1.20%	-1.20%
GEM	chr8	95,342,754	95,345,754	-	30.00%	14.39%	15.61%
GEMIN7	chr19	50,272,370	50,275,370	+	0.00%	21.05%	-21.05%
GJA10	chr6	90,658,909	90,661,909	+	40.00%	83.44%	-43.44%

GLO1	chr6	38,777,895	38,780,895	-	0.00%	11.07%	-11.07%
GLP1R	chr6	39,122,552	39,125,552	+	20.00%	11.88%	8.12%
GLRA2	chrX	14,455,341	14,458,341	+	20.00%	21.15%	-1.15%
GLS	chr2	191,451,798	191,454,798	+	10.00%	1.67%	8.33%
GNPDA1	chr5	141,371,790	141,374,790	-	10.00%	12.27%	-2.27%
GNPNAT1	chr14	52,327,136	52,330,136	-	10.00%	5.21%	4.79%
GOLGA1	chr9	126,749,113	126,752,113	-	100.00%	87.63%	12.37%
GOLGA5	chr14	92,328,329	92,331,329	+	10.00%	8.49%	1.51%
GPHN	chr14	66,041,878	66,044,878	+	20.00%	8.83%	11.17%
GPM6B	chrX	13,865,678	13,868,678	-	0.00%	2.93%	-2.93%
GPR1	chr2	206,790,016	206,793,016	-	0.00%	75.66%	-75.66%
GPR119	chrX	129,346,192	129,349,192	-	30.00%	86.89%	-56.89%
GPR12	chr13	26,231,922	26,234,922	-	20.00%	6.56%	13.44%
GPR18	chr13	98,710,999	98,713,999	-	100.00%	87.56%	12.44%
GPR26	chr10	125,413,861	125,416,861	+	0.00%	5.78%	-5.78%
GPR63	chr6	97,391,074	97,394,074	-	10.00%	4.04%	5.96%
GPR64	chrX	19,049,676	19,052,676	-	0.00%	4.11%	-4.11%
GPR85	chr7	112,514,069	112,517,069	-	40.00%	17.47%	22.53%
GRIA3	chrX	122,143,687	122,146,687	+	20.00%	4.69%	15.31%
GRPR	chrX	16,049,600	16,052,600	+	0.00%	65.68%	-65.68%
GSC	chr14	94,305,315	94,308,315	-	0.00%	1.46%	-1.46%
GTF2A1	chr14	80,756,474	80,759,474	-	0.00%	2.53%	-2.53%
GTF3A	chr13	26,894,681	26,897,681	+	20.00%	8.50%	11.50%
GTF3C1	chr16	27,467,752	27,470,752	-	20.00%	7.51%	12.49%
GTPBP4	chr10	1,022,338	1,025,338	+	10.00%	4.28%	5.72%
GUCY2F	chrX	108,610,957	108,613,957	-	40.00%	83.16%	-43.16%
H2AFY	chr5	134,762,503	134,765,503	-	0.00%	3.11%	-3.11%
HAO1	chr20	7,868,121	7,871,121	-	80.00%	90.77%	-10.77%
HBP1	chr7	106,594,642	106,597,642	+	10.00%	1.32%	8.68%
HCCS	chrX	11,037,342	11,040,342	+	20.00%	8.07%	11.93%
HDC	chr15	48,344,551	48,347,551	-	60.00%	81.46%	-21.46%
HDLBP	chr2	241,904,149	241,907,149	-	30.00%	6.34%	23.66%
HEY1	chr8	80,841,653	80,844,653	-	0.00%	1.79%	-1.79%
HIF1A	chr14	61,229,984	61,232,984	+	0.00%	2.59%	-2.59%
HMGB1	chr13	30,088,734	30,091,734	-	0.00%	1.35%	-1.35%
HMGB3	chrX	149,897,640	149,900,640	+	100.00%	82.61%	17.39%
HNF4G	chr8	76,480,704	76,483,704	+	10.00%	1.76%	8.24%
HOXC13	chr12	52,616,802	52,619,802	+	0.00%	1.56%	-1.56%
HRASLS	chr3	194,439,608	194,442,608	+	10.00%	3.77%	6.23%
HS6ST2	chrX	131,922,093	131,925,093	-	100.00%	88.44%	11.56%
HTR2B	chr2	231,697,076	231,700,076	-	80.00%	87.58%	-7.58%
HTR2C	chrX	113,722,807	113,725,807	+	0.00%	1.14%	-1.14%
HTR4	chr5	148,035,991	148,038,991	-	100.00%	49.33%	50.67%
ICOS	chr2	204,507,716	204,510,716	+	70.00%	87.11%	-17.11%
IDH1	chr2	208,838,043	208,841,043	-	10.00%	3.14%	6.86%
IFRD1	chr7	111,848,259	111,851,259	+	100.00%	78.98%	21.02%
IGBP1	chrX	69,268,024	69,271,024	+	10.00%	7.64%	2.36%
IGSF1	chrX	130,360,358	130,363,358	-	100.00%	64.70%	35.30%
IL1R1	chr2	102,045,436	102,048,436	+	60.00%	85.36%	-25.36%
IL1RAP	chr3	191,712,534	191,715,534	+	10.00%	4.63%	5.37%

IL1RN	chr2	113,579,262	113,582,262	+	100.00%	86.27%	13.73%
IL21R	chr16	27,319,187	27,322,187	+	50.00%	85.72%	-35.72%
IL7	chr8	79,879,313	79,882,313	-	20.00%	5.27%	14.73%
IL9	chr5	135,258,415	135,261,415	-	100.00%	77.07%	22.93%
INPP1	chr2	190,914,441	190,917,441	+	10.00%	8.36%	1.64%
IRS1	chr2	227,371,719	227,374,719	-	0.00%	1.12%	-1.12%
ITGAV	chr2	187,161,035	187,164,035	+	0.00%	4.80%	-4.80%
ITM2A	chrX	78,508,820	78,511,820	-	0.00%	14.67%	-14.67%
ITPR3	chr6	33,694,500	33,697,500	+	10.00%	4.98%	5.02%
JAG1	chr20	10,601,608	10,604,608	-	0.00%	1.52%	-1.52%
KCNA1	chr12	4,887,334	4,890,334	+	0.00%	1.94%	-1.94%
KCNA6	chr12	4,786,603	4,789,603	+	0.00%	6.14%	-6.14%
KCNE1L	chrX	108,754,049	108,757,049	-	10.00%	10.63%	-0.63%
KCNE4	chr2	223,622,776	223,625,776	+	40.00%	31.62%	8.38%
KLF7	chr2	207,739,236	207,742,236	-	0.00%	2.26%	-2.26%
KLHDC2	chr14	49,302,076	49,305,076	+	10.00%	5.19%	4.81%
KLK4	chr19	56,104,806	56,107,806	-	40.00%	34.35%	5.65%
KTN1	chr14	55,093,543	55,096,543	+	100.00%	85.82%	14.18%
LAMA3	chr18	19,521,560	19,524,560	+	0.00%	4.83%	-4.83%
LANCL1	chr2	211,049,621	211,052,621	-	50.00%	26.43%	23.57%
LECT2	chr5	135,317,622	135,320,622	-	100.00%	58.20%	41.80%
LIPG	chr18	45,340,425	45,343,425	+	0.00%	3.81%	-3.81%
LMAN1	chr18	55,176,483	55,179,483	-	20.00%	3.93%	16.07%
LRRC15	chr3	195,570,761	195,573,761	-	70.00%	84.57%	-14.57%
LTA	chr6	31,645,810	31,648,810	+	50.00%	88.51%	-38.51%
LTBP2	chr14	74,148,059	74,151,059	-	20.00%	5.60%	14.40%
LYN	chr8	56,952,926	56,955,926	+	10.00%	2.39%	7.61%
M6PR	chr12	8,992,818	8,995,818	-	10.00%	2.55%	7.45%
MAB21L1	chr13	34,947,832	34,950,832	-	0.00%	2.68%	-2.68%
MAG	chr19	40,472,868	40,475,868	+	90.00%	86.95%	3.05%
MAGEB2	chrX	30,141,598	30,144,598	+	40.00%	19.82%	20.18%
MAL	chr2	95,053,149	95,056,149	+	10.00%	6.41%	3.59%
MALT1	chr18	54,487,598	54,490,598	+	0.00%	3.07%	-3.07%
MAP3K10	chr19	45,387,491	45,390,491	+	0.00%	3.93%	-3.93%
MAP3K12	chr12	52,179,114	52,182,114	-	0.00%	1.11%	-1.11%
MAP3K7	chr6	91,352,507	91,355,507	-	0.00%	1.27%	-1.27%
MAP3K9	chr14	70,344,976	70,347,976	-	0.00%	2.06%	-2.06%
MAPK4	chr18	46,338,482	46,341,482	+	10.00%	3.08%	6.92%
MAPRE2	chr18	30,808,890	30,811,890	+	0.00%	5.34%	-5.34%
MAS1	chr6	160,245,964	160,248,964	+	100.00%	86.14%	13.86%
MASP1	chr3	188,491,504	188,494,504	-	100.00%	85.51%	14.49%
MATR3	chr5	138,635,340	138,638,340	+	0.00%	6.92%	-6.92%
MAX	chr14	64,638,166	64,641,166	-	0.00%	3.28%	-3.28%
MBP	chr18	72,973,627	72,976,627	-	30.00%	16.72%	13.28%
MBTPS2	chrX	21,765,675	21,768,675	+	40.00%	8.85%	31.15%
MC4R	chr18	56,189,981	56,192,981	-	0.00%	38.77%	-38.77%
ME2	chr18	46,657,417	46,660,417	+	0.00%	1.48%	-1.48%
MED6	chr14	70,136,137	70,139,137	-	30.00%	5.57%	24.43%
MEP1B	chr18	28,021,985	28,024,985	+	100.00%	83.34%	16.66%
MET	chr7	116,097,484	116,100,484	+	10.00%	2.53%	7.47%

MIPOL1	chr14	36,734,869	36,737,869	+	10.00%	2.62%	7.38%
MKI67	chr10	129,813,639	129,816,639	-	0.00%	5.39%	-5.39%
MKKS	chr20	10,361,870	10,364,870	-	0.00%	1.12%	-1.12%
MKRN3	chr15	21,359,547	21,362,547	+	10.00%	15.26%	-5.26%
MLPH	chr2	238,056,810	238,059,810	+	100.00%	75.79%	24.21%
MMP16	chr8	89,408,370	89,411,370	-	0.00%	2.28%	-2.28%
MNAT1	chr14	60,269,213	60,272,213	+	10.00%	12.28%	-2.28%
MOCS1	chr6	40,009,268	40,012,268	-	0.00%	5.91%	-5.91%
MORF4L2	chrX	102,828,742	102,831,742	-	20.00%	10.89%	9.11%
MOS	chr8	57,188,095	57,191,095	-	10.00%	12.33%	-2.33%
MRPL44	chr2	224,528,365	224,531,365	+	0.00%	5.46%	-5.46%
MRPS10	chr6	42,292,581	42,295,581	-	30.00%	24.76%	5.24%
MRPS9	chr2	105,018,873	105,021,873	+	20.00%	12.73%	7.27%
MSC	chr8	72,918,257	72,921,257	-	10.00%	2.92%	7.08%
MTCH1	chr6	37,061,052	37,064,052	-	0.00%	3.62%	-3.62%
MTM1	chrX	149,485,727	149,488,727	+	0.00%	7.03%	-7.03%
MTMR6	chr13	24,759,147	24,762,147	-	20.00%	11.64%	8.36%
MYADM	chr19	59,059,289	59,062,289	+	10.00%	25.38%	-15.38%
MYH7	chr14	22,973,767	22,976,767	-	50.00%	88.74%	-38.74%
MYL2	chr12	109,841,909	109,844,909	-	50.00%	81.62%	-31.62%
MYO1B	chr2	191,816,156	191,819,156	+	0.00%	4.73%	-4.73%
MYO5B	chr18	45,974,449	45,977,449	-	10.00%	3.26%	6.74%
NAB1	chr2	191,217,717	191,220,717	+	0.00%	32.38%	-32.38%
NCF4	chr22	35,584,976	35,587,976	+	100.00%	82.68%	17.32%
NCOA4	chr10	51,233,114	51,236,114	+	0.00%	9.74%	-9.74%
NDFIP1	chr5	141,466,254	141,469,254	+	0.00%	7.56%	-7.56%
NDN	chr15	21,482,543	21,485,543	-	0.00%	11.64%	-11.64%
NEU2	chr2	233,603,626	233,606,626	+	100.00%	83.11%	16.89%
NEUROG1	chr5	134,898,538	134,901,538	-	0.00%	7.03%	-7.03%
NFKBIA	chr14	34,942,706	34,945,706	-	0.00%	1.60%	-1.60%
NID2	chr14	51,605,295	51,608,295	-	10.00%	6.63%	3.37%
NIN	chr14	50,366,589	50,369,589	-	0.00%	2.57%	-2.57%
NOL4	chr18	30,056,513	30,059,513	-	0.00%	1.33%	-1.33%
NOS1	chr12	116,373,358	116,376,358	-	100.00%	86.01%	13.99%
NOTCH4	chr6	32,298,822	32,301,822	-	100.00%	85.67%	14.33%
NOVA1	chr14	26,135,800	26,138,800	-	0.00%	2.17%	-2.17%
NPC1	chr18	19,419,449	19,422,449	-	0.00%	2.66%	-2.66%
NPPC	chr2	232,498,357	232,501,357	-	0.00%	3.06%	-3.06%
NR2C2	chr3	14,962,095	14,965,095	+	0.00%	5.03%	-5.03%
NRG1	chr8	31,614,444	31,617,444	+	0.00%	6.76%	-6.76%
NRL	chr14	23,653,063	23,656,063	-	10.00%	5.11%	4.89%
NRXN3	chr14	77,776,487	77,779,487	+	40.00%	87.42%	-47.42%
NTF3	chr12	5,409,539	5,412,539	+	0.00%	1.24%	-1.24%
NUDT14	chr14	104,717,705	104,720,705	-	0.00%	22.11%	-22.11%
NUPL1	chr13	24,771,662	24,774,662	+	20.00%	6.75%	13.25%
NXT1	chr20	23,277,373	23,280,373	+	0.00%	2.83%	-2.83%
NYX	chrX	41,189,631	41,192,631	+	50.00%	83.07%	-33.07%
OAT	chr10	126,096,535	126,099,535	-	20.00%	4.70%	15.30%
OGN	chr9	94,205,799	94,208,799	-	10.00%	89.92%	-79.92%
OGT	chrX	70,667,658	70,670,658	+	20.00%	2.30%	17.70%

OMD	chr9	94,225,564	94,228,564	-	100.00%	73.54%	26.46%
ONECUT2	chr18	53,251,915	53,254,915	+	10.00%	2.11%	7.89%
OPRK1	chr8	54,325,810	54,328,810	-	0.00%	2.70%	-2.70%
OTX2	chr14	56,345,950	56,348,950	-	0.00%	1.36%	-1.36%
P2RX7	chr12	120,053,005	120,056,005	+	90.00%	83.32%	6.68%
P2RY10	chrX	78,085,485	78,088,485	+	100.00%	68.79%	31.21%
PABPC5	chrX	90,574,250	90,577,250	+	0.00%	8.33%	-8.33%
PAK3	chrX	110,072,169	110,075,169	+	10.00%	8.24%	1.76%
PAK4	chr19	44,306,260	44,309,260	+	20.00%	9.32%	10.68%
PARVB	chr22	42,724,424	42,727,424	+	90.00%	63.74%	26.26%
PAX8	chr2	113,751,997	113,754,997	-	0.00%	13.84%	-13.84%
PAX9	chr14	36,194,524	36,197,524	+	0.00%	2.12%	-2.12%
PCCA	chr13	99,537,270	99,540,270	+	10.00%	6.33%	3.67%
PCDH12	chr5	141,328,488	141,331,488	-	20.00%	6.40%	13.60%
PCDHB1	chr5	140,409,163	140,412,163	+	10.00%	5.39%	4.61%
PCDHB2	chr5	140,452,411	140,455,411	+	20.00%	16.09%	3.91%
PCDHB5	chr5	140,492,984	140,495,984	+	0.00%	10.70%	-10.70%
PCDHB6	chr5	140,507,867	140,510,867	+	10.00%	14.67%	-4.67%
PDE6A	chr5	149,303,549	149,306,549	-	100.00%	88.87%	11.13%
PDE7A	chr8	66,916,111	66,919,111	-	0.00%	2.29%	-2.29%
PDGFB	chr22	37,969,702	37,972,702	-	20.00%	10.68%	9.32%
PDGFRB	chr5	149,514,616	149,517,616	-	50.00%	84.36%	-34.36%
PGF	chr14	74,491,240	74,494,240	-	0.00%	20.11%	-20.11%
PGRMC1	chrX	118,252,244	118,255,244	+	0.00%	4.28%	-4.28%
PI15	chr8	75,897,327	75,900,327	+	20.00%	73.99%	-53.99%
PIGA	chrX	15,262,597	15,265,597	-	0.00%	3.69%	-3.69%
PIK3CG	chr7	106,290,959	106,293,959	+	100.00%	69.74%	30.26%
PITPNB	chr22	26,645,122	26,648,122	-	10.00%	9.41%	0.59%
PKHD1	chr6	52,059,382	52,062,382	-	90.00%	86.63%	3.37%
PKIA	chr8	79,588,929	79,591,929	+	30.00%	3.17%	26.83%
PLA2G7	chr6	46,810,389	46,813,389	-	20.00%	7.37%	12.63%
PLAG1	chr8	57,285,437	57,288,437	-	0.00%	1.41%	-1.41%
PLAT	chr8	42,183,399	42,186,399	-	80.00%	86.01%	-6.01%
PLCB1	chr20	8,058,824	8,061,824	+	10.00%	4.85%	5.15%
PLCB4	chr20	8,995,410	8,998,410	+	0.00%	2.93%	-2.93%
PLCD4	chr2	219,178,732	219,181,732	+	70.00%	75.38%	-5.38%
PLDN	chr15	43,664,709	43,667,709	+	0.00%	4.76%	-4.76%
PLEK2	chr14	66,947,670	66,950,670	-	10.00%	10.52%	-0.52%
PLG	chr6	161,041,260	161,044,260	+	100.00%	57.61%	42.39%
PLS3	chrX	114,699,757	114,702,757	+	10.00%	7.24%	2.76%
PLXDC2	chr10	20,143,174	20,146,174	+	0.00%	1.55%	-1.55%
PMAIP1	chr18	55,716,160	55,719,160	+	0.00%	2.67%	-2.67%
POU4F3	chr5	145,696,780	145,699,780	+	0.00%	1.23%	-1.23%
PPP1CC	chr12	109,664,127	109,667,127	-	0.00%	3.47%	-3.47%
PPP1R2	chr3	196,750,498	196,753,498	-	10.00%	4.08%	5.92%
PPP2R5E	chr14	63,078,845	63,081,845	-	10.00%	2.34%	7.66%
PRDM13	chr6	100,159,327	100,162,327	+	20.00%	2.92%	17.08%
PRKAB1	chr12	118,587,941	118,590,941	+	40.00%	4.24%	35.76%
PROM2	chr2	95,301,928	95,304,928	+	70.00%	83.77%	-13.77%
PRPF39	chr14	44,621,052	44,624,052	+	50.00%	16.60%	33.40%

PRPS2	chrX	12,717,395	12,720,395	+	0.00%	3.73%	-3.73%
PRX	chr19	45,610,113	45,613,113	-	60.00%	77.96%	-17.96%
PSMA3	chr14	57,779,302	57,782,302	+	10.00%	1.82%	8.18%
PSMC6	chr14	52,241,640	52,244,640	+	10.00%	11.98%	-1.98%
PTER	chr10	16,516,970	16,519,970	+	10.00%	5.78%	4.22%
PTGDR	chr14	51,802,181	51,805,181	+	0.00%	3.23%	-3.23%
PTGER2	chr14	51,848,773	51,851,773	+	0.00%	2.60%	-2.60%
PTGIR	chr19	51,819,194	51,822,194	-	50.00%	71.93%	-21.93%
PTOV1	chr19	55,043,950	55,046,950	+	0.00%	3.73%	-3.73%
PTPN21	chr14	88,089,830	88,092,830	-	0.00%	6.03%	-6.03%
PURA	chr5	139,465,546	139,468,546	+	10.00%	2.69%	7.31%
PYGB	chr20	25,174,705	25,177,705	+	0.00%	2.34%	-2.34%
RAB27B	chr18	50,644,706	50,647,706	+	40.00%	13.58%	26.42%
RAD52	chr12	969,617	972,617	-	0.00%	1.19%	-1.19%
RASAL1	chr12	112,057,427	112,060,427	-	50.00%	20.83%	29.17%
RAX	chr18	55,091,298	55,094,298	-	10.00%	8.08%	1.92%
RBX1	chr22	39,675,297	39,678,297	+	10.00%	4.54%	5.46%
RET	chr10	42,890,481	42,893,481	+	0.00%	6.61%	-6.61%
RFC4	chr3	188,006,541	188,009,541	-	40.00%	7.68%	32.32%
RFC5	chr12	116,933,776	116,936,776	+	100.00%	87.19%	12.81%
RHAG	chr6	49,711,511	49,714,511	-	90.00%	86.75%	3.25%
RIMS1	chr6	72,651,127	72,654,127	+	0.00%	3.00%	-3.00%
RIPK3	chr14	23,878,091	23,881,091	-	0.00%	10.95%	-10.95%
RNF14	chr5	141,316,077	141,319,077	+	20.00%	86.74%	-66.74%
RNF26	chr11	118,708,447	118,711,447	+	0.00%	17.39%	-17.39%
RNF34	chr12	120,320,227	120,323,227	+	10.00%	3.51%	6.49%
ROR2	chr9	93,751,265	93,754,265	-	20.00%	2.14%	17.86%
RPL21	chr13	26,721,446	26,724,446	+	0.00%	10.30%	-10.30%
RPS14	chr5	149,808,512	149,811,512	-	10.00%	6.73%	3.27%
RPS16	chr19	44,617,458	44,620,458	-	10.00%	13.32%	-3.32%
RPS5	chr19	63,588,448	63,591,448	+	20.00%	3.81%	16.19%
RPS6KA3	chrX	20,194,444	20,197,444	-	0.00%	1.86%	-1.86%
RRS1	chr8	67,501,817	67,504,817	+	0.00%	3.39%	-3.39%
RTN1	chr14	59,406,437	59,409,437	-	0.00%	1.81%	-1.81%
RTN4R	chr22	18,649,769	18,652,769	-	100.00%	89.22%	10.78%
SAG	chr2	233,879,048	233,882,048	+	90.00%	84.98%	5.02%
SAV1	chr14	50,203,806	50,206,806	-	0.00%	2.41%	-2.41%
SCG2	chr2	224,174,465	224,177,465	-	20.00%	25.41%	-5.41%
SCGN	chr6	25,758,443	25,761,443	+	50.00%	12.73%	37.27%
SCNN1G	chr16	23,099,537	23,102,537	+	10.00%	4.72%	5.28%
SDC2	chr8	97,572,755	97,575,755	+	0.00%	5.65%	-5.65%
SDCBP	chr8	59,626,037	59,629,037	+	20.00%	3.60%	16.40%
SDPR	chr2	192,419,226	192,422,226	-	30.00%	23.20%	6.80%
SEMA6B	chr19	4,508,507	4,511,507	-	30.00%	33.75%	-3.75%
SERPINA10	chr14	93,828,361	93,831,361	-	100.00%	82.42%	17.58%
SERPINA5	chr14	94,095,532	94,098,532	+	80.00%	87.51%	-7.51%
SERPINA6	chr14	93,858,484	93,861,484	-	100.00%	65.07%	34.93%
SERPINB2	chr18	59,687,906	59,690,906	+	100.00%	80.76%	19.24%
SERPINB7	chr18	59,569,149	59,572,149	+	100.00%	90.03%	9.97%
SERPINB8	chr18	59,786,139	59,789,139	+	100.00%	15.23%	84.77%

SETBP1	chr18	40,512,136	40,515,136	+	0.00%	1.50%	-1.50%
SFRP1	chr8	41,285,173	41,288,173	-	10.00%	4.16%	5.84%
SGCG	chr13	22,651,091	22,654,091	+	0.00%	91.29%	-91.29%
SH2D1A	chrX	123,305,875	123,308,875	+	40.00%	29.83%	10.17%
SHC3	chr9	90,982,502	90,985,502	-	0.00%	1.47%	-1.47%
SIX1	chr14	60,193,730	60,196,730	-	10.00%	7.16%	2.84%
SIX6	chr14	60,043,422	60,046,422	+	0.00%	1.33%	-1.33%
SLC10A2	chr13	102,516,197	102,519,197	-	80.00%	85.73%	-5.73%
SLC11A2	chr12	49,707,616	49,710,616	-	50.00%	48.87%	1.13%
SLC12A6	chr15	32,416,553	32,419,553	-	0.00%	4.87%	-4.87%
SLC20A2	chr8	42,515,226	42,518,226	-	10.00%	2.20%	7.80%
SLC22A3	chr6	160,687,290	160,690,290	+	0.00%	2.76%	-2.76%
SLC24A4	chr14	91,856,678	91,859,678	+	0.00%	9.18%	-9.18%
SLC25A14	chrX	129,299,555	129,302,555	+	0.00%	3.55%	-3.55%
SLC25A5	chrX	118,484,391	118,487,391	+	10.00%	4.42%	5.58%
SLC26A2	chr5	149,318,493	149,321,493	+	10.00%	5.53%	4.47%
SLC26A3	chr7	107,229,906	107,232,906	-	100.00%	85.61%	14.39%
SLC26A4	chr7	107,086,316	107,089,316	+	10.00%	5.96%	4.04%
SLC27A2	chr15	48,259,685	48,262,685	+	0.00%	5.68%	-5.68%
SLC28A2	chr15	43,329,720	43,332,720	+	20.00%	25.06%	-5.06%
SLC29A1	chr6	44,293,220	44,296,220	+	20.00%	18.24%	1.76%
SLC5A7	chr2	107,967,411	107,970,411	+	20.00%	8.16%	11.84%
SLC6A14	chrX	115,479,818	115,482,818	+	10.00%	55.91%	-45.91%
SLC6A6	chr3	14,417,080	14,420,080	+	10.00%	13.24%	-3.24%
SLC7A13	chr8	87,401,491	87,404,491	-	100.00%	83.85%	16.15%
SLC7A8	chr14	22,721,723	22,724,723	-	0.00%	10.95%	-10.95%
SLC8A3	chr14	69,724,540	69,727,540	-	10.00%	2.67%	7.33%
SLITRK4	chrX	142,550,262	142,553,262	-	10.00%	5.70%	4.30%
SMAP1	chr6	71,432,200	71,435,200	+	20.00%	6.42%	13.58%
SMARCAL1	chr2	216,983,382	216,986,382	+	40.00%	13.90%	26.10%
SMOC1	chr14	69,388,601	69,391,601	+	90.00%	85.48%	4.52%
SMOC2	chr6	168,582,680	168,585,680	+	20.00%	5.83%	14.17%
SMPX	chrX	21,685,202	21,688,202	-	0.00%	81.82%	-81.82%
SNAI2	chr8	49,995,852	49,998,852	-	10.00%	2.86%	7.14%
SNAP25	chr20	10,145,478	10,148,478	+	0.00%	1.16%	-1.16%
SNRPB2	chr20	16,656,606	16,659,606	+	20.00%	6.67%	13.33%
SNTG1	chr8	50,982,902	50,985,902	+	10.00%	6.10%	3.90%
SOX17	chr8	55,531,048	55,534,048	+	0.00%	3.81%	-3.81%
SPAG6	chr10	22,672,405	22,675,405	+	0.00%	2.36%	-2.36%
SPATS1	chr6	44,416,375	44,419,375	+	40.00%	17.72%	22.28%
SPATS2	chr12	48,044,634	48,047,634	+	0.00%	4.20%	-4.20%
SPDEF	chr6	34,631,088	34,634,088	-	70.00%	84.99%	-14.99%
SPG20	chr13	35,841,317	35,844,317	-	40.00%	91.93%	-51.93%
SPRY4	chr5	141,685,204	141,688,204	-	10.00%	9.56%	0.44%
SPTLC2	chr14	77,151,869	77,154,869	-	10.00%	6.64%	3.36%
SQRDL	chr15	43,712,293	43,715,293	+	20.00%	10.98%	9.02%
SRPK1	chr6	35,996,097	35,999,097	-	20.00%	9.73%	10.27%
SST	chr3	188,869,881	188,872,881	-	20.00%	15.12%	4.88%
SSTR1	chr14	37,744,955	37,747,955	+	10.00%	10.07%	-0.07%
SSTR3	chr22	35,937,308	35,940,308	-	50.00%	87.80%	-37.80%

SSTR4	chr20	22,962,057	22,965,057	+	30.00%	8.88%	21.12%
ST18	chr8	53,535,072	53,538,072	-	50.00%	85.90%	-35.90%
STAG2	chrX	122,919,743	122,922,743	+	0.00%	1.90%	-1.90%
STAM	chr10	17,724,130	17,727,130	+	10.00%	4.30%	5.70%
STARD7	chr2	96,237,290	96,240,290	-	10.00%	6.65%	3.35%
STK17B	chr2	196,748,472	196,751,472	-	70.00%	83.32%	-13.32%
STK38	chr6	36,622,225	36,625,225	-	0.00%	13.99%	-13.99%
STMN2	chr8	80,683,604	80,686,604	+	30.00%	9.29%	20.71%
STXBP6	chr14	24,588,343	24,591,343	-	0.00%	2.97%	-2.97%
SULF1	chr8	70,539,413	70,542,413	+	40.00%	28.47%	11.53%
SUV39H2	chr10	14,958,825	14,961,825	+	0.00%	1.49%	-1.49%
SYNPO	chr5	149,958,835	149,961,835	+	30.00%	88.04%	-58.04%
SYT4	chr18	39,110,613	39,113,613	-	0.00%	11.20%	-11.20%
SYTL4	chrX	99,872,766	99,875,766	-	0.00%	13.65%	-13.65%
TAF7L	chrX	100,433,715	100,436,715	-	30.00%	14.66%	15.34%
TBC1D8	chr2	101,234,760	101,237,760	-	10.00%	2.43%	7.57%
TBX1	chr22	18,122,226	18,125,226	+	10.00%	5.75%	4.25%
TBX22	chrX	79,154,911	79,157,911	+	20.00%	55.02%	-35.02%
TBX3	chr12	113,605,352	113,608,352	-	0.00%	1.13%	-1.13%
TCERG1	chr5	145,805,067	145,808,067	+	0.00%	7.51%	-7.51%
TCOF1	chr5	149,715,395	149,718,395	+	0.00%	9.02%	-9.02%
TCP11	chr6	35,223,365	35,226,365	-	10.00%	45.71%	-35.71%
TEP1	chr14	19,950,428	19,953,428	-	0.00%	18.74%	-18.74%
TERF1	chr8	74,081,653	74,084,653	+	10.00%	2.83%	7.17%
TFIP11	chr22	25,237,471	25,240,471	-	10.00%	9.67%	0.33%
TGFBI	chr5	135,390,483	135,393,483	+	30.00%	13.55%	16.45%
THAP1	chr8	42,816,625	42,819,625	-	0.00%	6.75%	-6.75%
TJP1	chr15	28,047,360	28,050,360	-	20.00%	4.69%	15.31%
TLR8	chrX	12,832,660	12,835,660	+	80.00%	86.56%	-6.56%
TMEFF2	chr2	192,767,680	192,770,680	-	0.00%	1.22%	-1.22%
TMSB4X	chrX	12,901,148	12,904,148	+	0.00%	4.88%	-4.88%
TNFAIP2	chr14	102,657,532	102,660,532	+	0.00%	10.02%	-10.02%
TNFRSF11A	chr18	58,141,500	58,144,500	+	0.00%	1.86%	-1.86%
TNFRSF19	chr13	23,040,509	23,043,509	+	100.00%	83.90%	16.10%
TNFRSF21	chr6	47,384,600	47,387,600	-	0.00%	2.78%	-2.78%
TNP1	chr2	217,432,032	217,435,032	-	100.00%	88.25%	11.75%
TPST2	chr22	25,321,681	25,324,681	-	100.00%	86.00%	14.00%
TREM1	chr6	41,361,435	41,364,435	-	30.00%	86.15%	-56.15%
TRIM39	chr6	30,400,235	30,403,235	+	0.00%	2.34%	-2.34%
TRPC4	chr13	37,341,562	37,344,562	-	20.00%	3.66%	16.34%
TRPC5	chrX	111,211,660	111,214,660	-	10.00%	3.47%	6.53%
TRPM8	chr2	234,488,782	234,491,782	+	100.00%	88.83%	11.17%
TTPA	chr8	64,160,166	64,163,166	-	0.00%	9.95%	-9.95%
TUBGCP3	chr13	112,289,482	112,292,482	-	10.00%	6.61%	3.39%
UBE2A	chrX	118,590,529	118,593,529	+	0.00%	4.29%	-4.29%
UBE2J1	chr6	90,118,286	90,121,286	-	20.00%	4.44%	15.56%
UBE3A	chr15	23,234,221	23,237,221	-	0.00%	4.87%	-4.87%
UBL3	chr13	29,321,821	29,324,821	-	0.00%	5.23%	-5.23%
UBQLN2	chrX	56,604,751	56,607,751	+	0.00%	4.09%	-4.09%
UNC5D	chr8	35,210,517	35,213,517	+	0.00%	1.65%	-1.65%

USP26	chrX	132,057,803	132,060,803	-	40.00%	25.98%	14.02%
UXT	chrX	47,402,504	47,405,504	-	30.00%	14.22%	15.78%
VDR	chr12	46,622,098	46,625,098	-	40.00%	29.19%	10.81%
VPS4B	chr18	59,239,673	59,242,673	-	0.00%	5.81%	-5.81%
VRK1	chr14	96,331,394	96,334,394	+	0.00%	9.91%	-9.91%
VSX1	chr20	25,009,996	25,012,996	-	20.00%	1.35%	18.65%
WASF3	chr13	26,027,840	26,030,840	+	0.00%	3.34%	-3.34%
WDFY1	chr2	224,517,348	224,520,348	-	0.00%	9.96%	-9.96%
WHSC1L1	chr8	38,357,947	38,360,947	-	0.00%	2.41%	-2.41%
WNT2	chr7	116,749,579	116,752,579	-	0.00%	1.15%	-1.15%
XPNPEP2	chrX	128,698,631	128,701,631	+	90.00%	76.27%	13.73%
XRN2	chr20	21,229,942	21,232,942	+	30.00%	4.53%	25.47%
ZDHHC4	chr7	6,581,590	6,584,590	+	0.00%	23.65%	-23.65%
ZFP36L1	chr14	68,331,943	68,334,943	-	10.00%	13.59%	-3.59%
ZP2	chr16	21,129,369	21,132,369	-	100.00%	88.29%	11.71%