universität wien

DISSERTATION

Titel der Dissertation

„Bioinformatic tools for analyzing epigenomic profiling data“

Verfasser

Huy Q. Dinh

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 091 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreuer:	Univ.-Prof. Dr. Arndt von Haeseler

Work for this thesis was partially performed, supervised, and financed in the group of ***Ortrun Mittelsten Scheid*** *at the Gregor Mendel Institute of Molecular Plant Biology, Vienna, Austrian Academy of Science.*

I wonder why, I wonder why
I wonder why I wonder ...
I wonder why I wonder why,
I wonder why I wonder.

Richard Feynman

We cannot change the cards we are dealt, just how we play the hand.

Randy Pausch

Computers are to biology what mathematics is to physics.

Harold Morowitz

Abstract

Epigenetics, investigating the biological information of genomes not only encoded in the DNA sequence, has become a hot topic boosted by rapid development of high-throughput technologies. In the light of that, bioinformatics plays an important role in analyzing the massive datasets to further examine the data and to formulate biological hypotheses.

DNA methylation is one important epigenetic mark in developmental and disease biology. One widely-used technique to profile genome-wide DNA methylation is based on bisulfite conversion of unmethylated cytosines (C) to thymines (T), followed by deep sequencing technology, called BS-Seq data. The C-T conversion raises a number of challenges in mapping the bisulfite-converted short reads to the reference genome. Besides, the current technology cannot consider the heterogeneity of DNA methylation from mixtures of cells. This affects the accuracy of estimating the DNA methylation patterns in the genome. Hence, new bioinformatics methods are required to estimate the cell-type specific DNA methylation.

Integrating multiple datasets of profiling epigenetic/chromatin marks for many different samples, conditions and organisms is also an underdeveloped field in bioinformatics, given the rapid growth of biological data. It is essential for further studies to find epigenomic patterns like a chromatin-based epigenetic code. However, comparative bioinformatics procedure is difficult because of different distributions or different scales of the marks.

In this thesis, I have developed bioinformatics tools and applied them to the model organism, *Arabidopsis thaliana*. First, I have implemented a new and sensitive analysis tool for analyzing BS-Seq data based on Smith-Waterman local alignment mapping. Second, I have developed an efficient algorithm to deal with heterogeneity in DNA methylation data derived from BS-Seq. Finally, I have suggested a method to integrate epigenomic signals from multiple genome-wide profiling data for further data mining purpose, e.g. epigenetic signature discovery.

Zusammenfassung

Epigenetik, die Erforschung der biologischen Information in Genomen ausserhalb der DNA Sequenz, hat durch die rasche Entwicklung der Hochdurchsatz-Techniken besonders viele Impulse bekommen. Deshalb spielt die Bioinformatik eine wichtige Rolle bei der Analyse der ausserordentlich grossen Datenmengen und der Formulierung biologischer Hypothesen in der Epigenetik.

DNA Methylierung ist ein wichtiger epigenetischer Parameter in der normalen und pathologischen Entwicklungsbiologie. Genomweite DNA Methylierungsprofile werden hauptschlich durch Bisulfit-Konversion genomischer DNA erstellt, bei der unmethyliertes Cytosin (C) in Thymin (T) umgewandelt wird, gefolgt von Hochdurchsatz-Sequenzierung (BS-Seq). Die Umwandlung von C zu T erschwert die Zuordnung der Einzelsequenzen zum Referenzgenom in mehrerer Hinsicht. Ausserdem kann mit der herkmmlichen Technik die Heterogenitt der DNA Methylierung in Material aus mehreren Zellen oder Geweben nicht bercksichtigt werden. Das beeintrchtigt die Genauigkeit bei der Bestimmung der genomischen Methylierungsmuster. Deshalb sind neue bioinformatische Methoden erforderlich, um zellspezifische DNA Methylierung zu erkennen.

Aufgrund der schnell wachsenden Datenmengen ist die gleichzeitige Erfassung mehrerer epigenetischer Parameter in Form von Chromatineigenschaften in verschiedenen Proben, Bedingungen oder Organismen eine weitere Herausforderung und ein wenig bearbeitetes Gebiet der Bioinformatik, jedoch Voraussetzung zur Entdeckung eines chromatinbasierten epigenetischen Codes. Vergleichende bioinformatische Anstze werden hierbei durch unterschiedliche Verteilung und/oder Spannweite der Parameter erschwert.

In dieser Dissertation stelle ich von mir entwickelte bioinformatische Methoden zu diesen Themenbereichen vor und zeige deren Anwendung auf Daten aus dem Modellorganismus *Arabidopsis thaliana*. Als erstes habe ich ein neues und hochauflsendes Verfahren zur Analyse von BS-Seq Daten entwickelt, welches auf dem Smith-Waterman local alignment Prinzip beruht. Zweitens habe ich einen effizienten Algorithmus konzipiert, um den Grad der Heterogenitt in BS-Seq Daten zu bestimmen. Drittens habe

ich eine Methode entworfen, mit der man zahlreiche epigenetische Parameter und deren genomweite Profile zusammenfassen, vergleichen und optisch darstellen kann, um die weitere Analyse und Interpretation zu erleichtern.

Parts of this thesis have been published or submitted:

(i) **Dinh HQ, Dubin M, Sedlazeck FJ, Letter N, Mittelsten Scheid O, von Haeseler A.** (2012) Advanced methylome analysis after bisulfite conversion: An example in Arabidopsis. *PLoS ONE, 7, e41528.*

(ii) **Dinh HQ, Mittelsten Scheid O, von Haeseler A** (2011). MethColor: a computational approach to uncover DNA methylation heterogeneity. *In the Proceedings of the German Conference on Bioinformatics (GCB2011, Weihenstephan, Germany, September 2011).*

(iii) **Dinh HQ, Mittelsten Scheid O, von Haeseler A.** Epi-Speller: a tool to discover epigenetic signatures. *submitted to Epigenetics & Chromatin.*

Contents

List of Figures

List of Tables

Chapter 1.

Introduction & Background

Epigenomics is where genomics was 30 years ago, when everyone was working on part of the puzzle.

Peter Jones

1.1. Epigenetics

Epigenetics, a term coined by Conrad Waddington (Waddington, 1942) in the context of how cell fates are established during development, commonly refers nowadays to the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in the DNA sequence (Russo *et al.*, 1996; Allis *et al.*, 2007). Epigenetic changes are rather exerted by biochemical modifications of DNA and associated proteins, many common in all eukaryote species. They are conserved during cell division and development (Allis *et al.*, 2007). The most prominent and most studied epigenetic elements are DNA methylation (e.g. Klose and Bird, 2006) and histone modifications (e.g. Kouzarides, 2007), but other epigenetic features like nucleosome positioning (Jiang and Pugh, 2009) or chromatin remodeling (e.g. Alabert and Groth, 2012), histone variants (e.g. Talbert and Henikoff, 2010), etc. gain increasing attention.

Methodology in epigenetic research has quickly changed along with the rapid development of high-throughput genome technology. Nowadays, more and more epigenomes across different species, developmental stages, tissues or conditions have been successfully profiled by numerous research teams (e.g. Satterlee *et al.*, 2010). Not only did the genome-wide profiling technology help to gain high resolution (even at single bp level)

of epigenetic maps, but it also boosted the discovery of biological roles of epigenetic elements. Concomitantly, the challenges in data analysis have also risen in the context of more and more sophisticated high-throughput techniques and special properties of epigenetic data (Bock and Lengauer, 2008).

1.1.1. DNA methylation and bisulfite conversion

DNA methylation (Holliday and Pugh, 1975; Razin and Riggs, 1980) was the first heritable epigenetic mark discovered to correlate with the variability of gene expression across many eukaryotes. It occurs as an addition of methyl groups (CH3) to the DNA, usually at position 5 of cytosine residues (mC). DNA methylation patterns are established and maintained, depending on the nucleotide context of the cytosines, by the enzyme family of DNA methyltransferases (DNMTs) (Bird, 2002). While in differentiated mammalian cells mC is found nearly exclusively in the dinucleotide CG, all analyzed plants (e.g. Arabidopsis Lister *et al.*, 2008), some fungi (e.g *Neurospora crassa,* Selker *et al.*, 2003), insects (e.g honey bee Lyko *et al.*, 2010) and mammalian stem cells (Lister *et al.*, 2009) have in addition mC followed by H (= A, T, C) (Dyachenko *et al.*, 2010).

DNA methylation patterns vary according to different cells, cell types, developmental stages and external conditions and can correlate with different cellular phenotypes. In the course of practical methylation analysis, DNA samples are usually prepared from large cell populations that can have heterogeneous DNA methylation profiles (Laird, 2010). Given the important implication of DNA methylation in development and diseases (e.g. in cancer types), genome-wide profiling of this modification is essential to gain a comprehensive picture of epigenetic regulation.

The state-of-the-art technique to analyze DNA methylation applies bisulfite (BS) conversion that was discovered in 1980 (Wang *et al.*, 1980). With this biochemical technique and in combination with PCR amplification (Frommer *et al.*, 1992), the modification can be detected at single-bp resolution. The procedure chemically converts unmethylated cytosines to uracil, whereas the methylated cytosines remain unchanged. During subsequent PCR and sequencing, uracil residues become apparent as thymines (Fig. 1.1). Initially applied to single or a few genomic regions by amplification with specific primers and individual Sanger sequencing (Frommer *et al.*, 1992; Clark *et al.*, 1994), it is now also performed at a genome-wide scale by Next Generation Sequencing (NGS)

(Lister and Ecker, 2009; Laird, 2010). Complete conversion and sufficient coverage of the sequencing provided, this allows to generate a methylome, a single-bp map of DNA methylation along the whole genome.

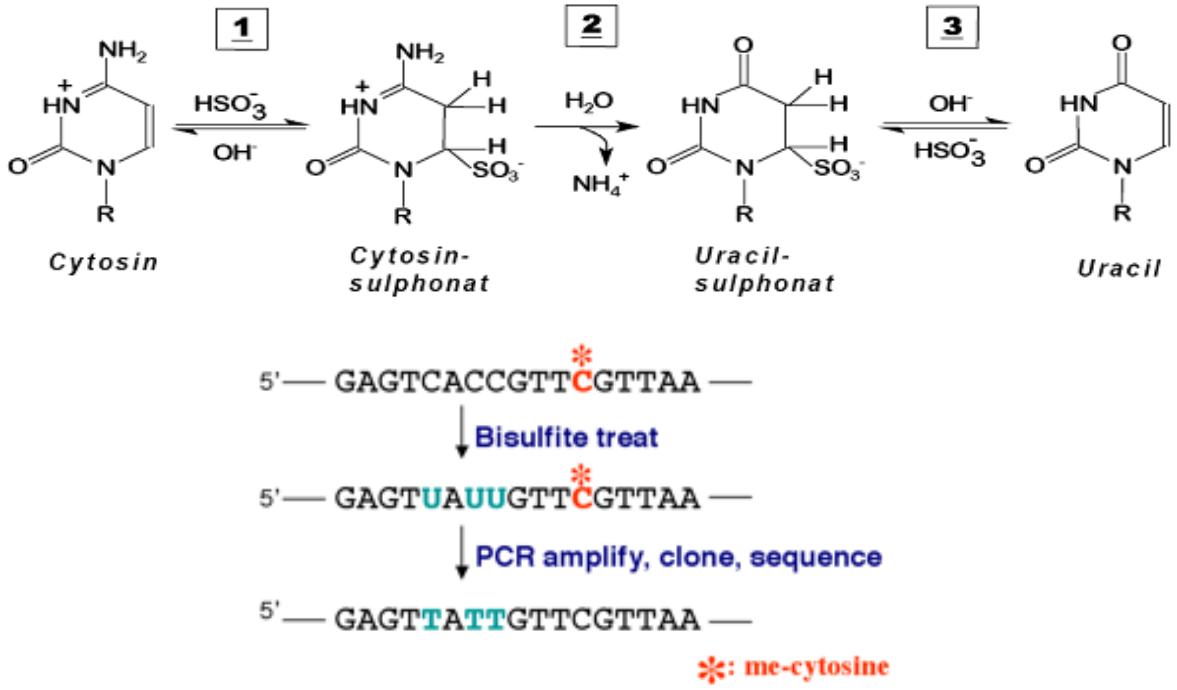


Figure 1.1.: **DNA methylation & BS conversion**: Figures from `http://www.methylogix.com` and `http://www.alphabiolab.com`

1.1.2. Histone modifications

Major elements of chromatin organization in eukaryotes are the nucleosomes, multimeric complexes formed by the assembly of different histone molecules that are in the center of 147 bp DNA wrapped around them to result in a bead-on-the-string structure (Fig. 1.2). Histones play an important role in gene regulation and are carriers of epigenetic information (Strahl and Allis, 2000) since histone tails can be biochemically modified by several different posttranslational processes, among others by acetylation, phosphorylation and methylation. The number of possible modifications at one genomic position is extremely high, since different residues at several amino acid positions of four canonical

histones (namely H2A, H2B, H3 and H4) and their variants (e.g. H3.3, H2A.Z, CENP-A) can be combined. These modifications can change the chromatin configuration and DNA accessibility and thereby regulate gene expression, recombination, or replication by cis- and trans- effects.

Chromatin ImmunoPrecipitation (ChIP) (Figure 1.2) is a biochemical technique, invented in 1988 by Solomon *et al.* (1988), used to determine whether a histone protein binds to or is localized at a specific DNA sequence. The genomic regions of interest are enriched by binding partially fragmented DNA chromatin complexes to specific antibodies corresponding to the profiled marks. The profiling of histone modifications is commonly studied by ChIP followed by microarray (ChIP-chip), (Ren *et al.*, 2000) or ChIP followed by next generation sequencing (ChIP-Seq), reviewed in Park (2009). The enriched genomic regions are quantified by relative abundance of microarray hybridization signal intensities or the number of mapped reads in deep sequencing data.

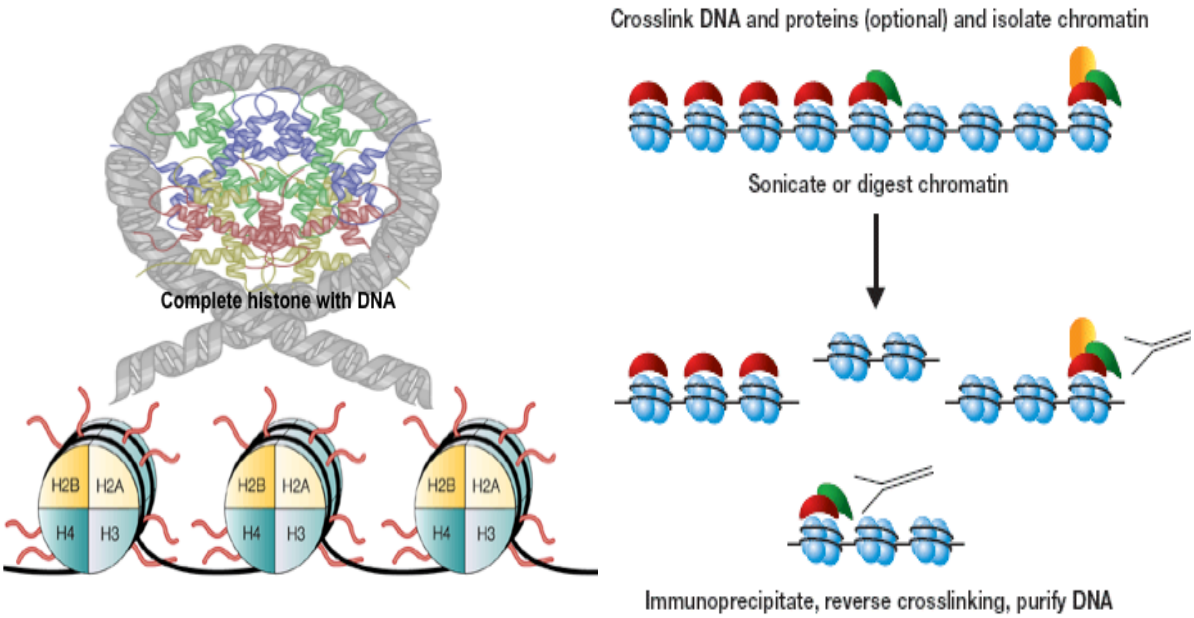


Figure 1.2.: **Histone structure and Chromatin Immunoprecipitation**, adapted from (Marks *et al.*, 2001) and English Wikipedia `http://en.wikipedia.org/wiki/Histone`.

It was suggested that the combination of histone modifications determines chromatin and gene expression activities in a way presenting a "histone code" or "chromatin-based

epigenetic code" (Jenuwein and Allis, 2001; Allis *et al.*, 2007). However, unlike the universal genetic code, any epigenetic code (if existing) must vary considerably across organisms, given the high diversity of histone modifications and their different "implementations". More than a decade after the idea of a code was proposed, there are many computational efforts to analyze the diversity of epigenomes for a large number of chromatins marks supporting the hypothesis across organisms (Hon *et al.*, 2008; Ernst and Kellis, 2010; Roudier *et al.*, 2011) (reviewed in (van Steensel, 2011)). The accumulating data about the genomic distribution of multiple chromatin marks in many organisms offer a great opportunity to test the hypothesis of an epigenetic code and to decipher it.

1.1.3. Plant/Arabidopsis and epigenetics

Plants have an impressive portfolio of epigenetic regulators and have contributed substantially to insight into epigenetic regulation. Especially the model organism *Arabidopsis thaliana* has proven very suitable for epigenetic research. In fact, the first single-bp resolution methylome published was that of Arabidopsis (Lister *et al.*, 2008). Beside of their contribution to pioneering discoveries and analyses of epigenetic phenomena from early on, plants provide some other benefits in epigenetic research. First, in terms of many epigenetic concepts, plants are more similar to mammals than other animals (Matzke and Mittelsten Scheid, 2006) and therefore offer approaches also for biomedically relevant research. In addition, plants offer experimental possibilities for epigenetic research that are difficult to apply in mammals, in terms of both methodology and concepts (more details reviewed in Matzke and Mittelsten Scheid (2006)). However, there are also some extra challenges in plants, like the significant amount of non-CG methylation in genomic DNA. Although Arabidopsis was such a widely applied model system due to its small genome size, the potential for C/mC at every C position made the analysis of BS-Seq data from Arabidopsis more difficult and potentially biased compared to other eukaryotic organisms. On the other hand, the rapid increase in data sets for multiple epigenetic parameters, gene expression data under numerous conditions, nucleosome position information and naturally existing variation in all these parameters in Arabidopsis make it a very interesting, though challenging organism for the study of the connection between genomic and epigenomic information on one hand and phenotype and adaptation of the organism on the other.

1.2. High-throughput technology for profiling epigenomic data at genome-wide scale

Low input, high throughput, no output science.

Sydney Brenner

Nowadays, biology research is significantly boosted by high-throughput technology. This allowed answering many long-standing questions quicker, better and more efficiently on a genome-wide scale. High throughput biological assays provide thousands of measurements per biological sample, and the sheer amount of high throughput data enables the biology community to tackle many important questions in a comprehensive and global way at very high resolution genome-wide scale.

1.2.1. Widely-used technologies used in epigenetic research

Microarrays, also called DNA chips, consist of thousands of microscopic DNA spots attached to a solid surface, each spot containing a short specific DNA sequence representing known genes or genomic regions called probes. A fluorescently labeled sample of DNA or RNA is hybridized against the microarray, and the relative abundance of a nucleotide sequence in the sample is detected and quantified by the probe-target hybridization signal (Schena *et al.*, 1995). Microarrays have been widely used for measuring gene expression or single nucleotide polymorphisms. Tiling microarrays are high-density oligonucleotide-based microarrays that refine genome-wide resolution (Mockler *et al.*, 2005) due to overlapping (tiled) probes. In contrast to the standard microarray, the signal at specific genomic regions can be normalized and summarized over all overlapping probes that contain those regions. Typical applications of tiling arrays are hybridizations with DNA from chromatin ChIP-chip to determine histone modifications, transcriptional factor binding or transcriptome profiles.

NGS, or deep sequencing, refers to the massive parallelization technology of DNA sequencing. Even though there are different NGS frameworks, they commonly share the same principle based on the so-called cyclic-array sequencing by iterative cycles of enzymatic manipulation and imaging-based data collection (Shendure and Ji, 2008). The most popular technologies are 454 pyrosequencing, Illumina (Solexa) sequencing, SOLiD sequencing, to name a few, in which the sequencing reactions proceed in parallel on

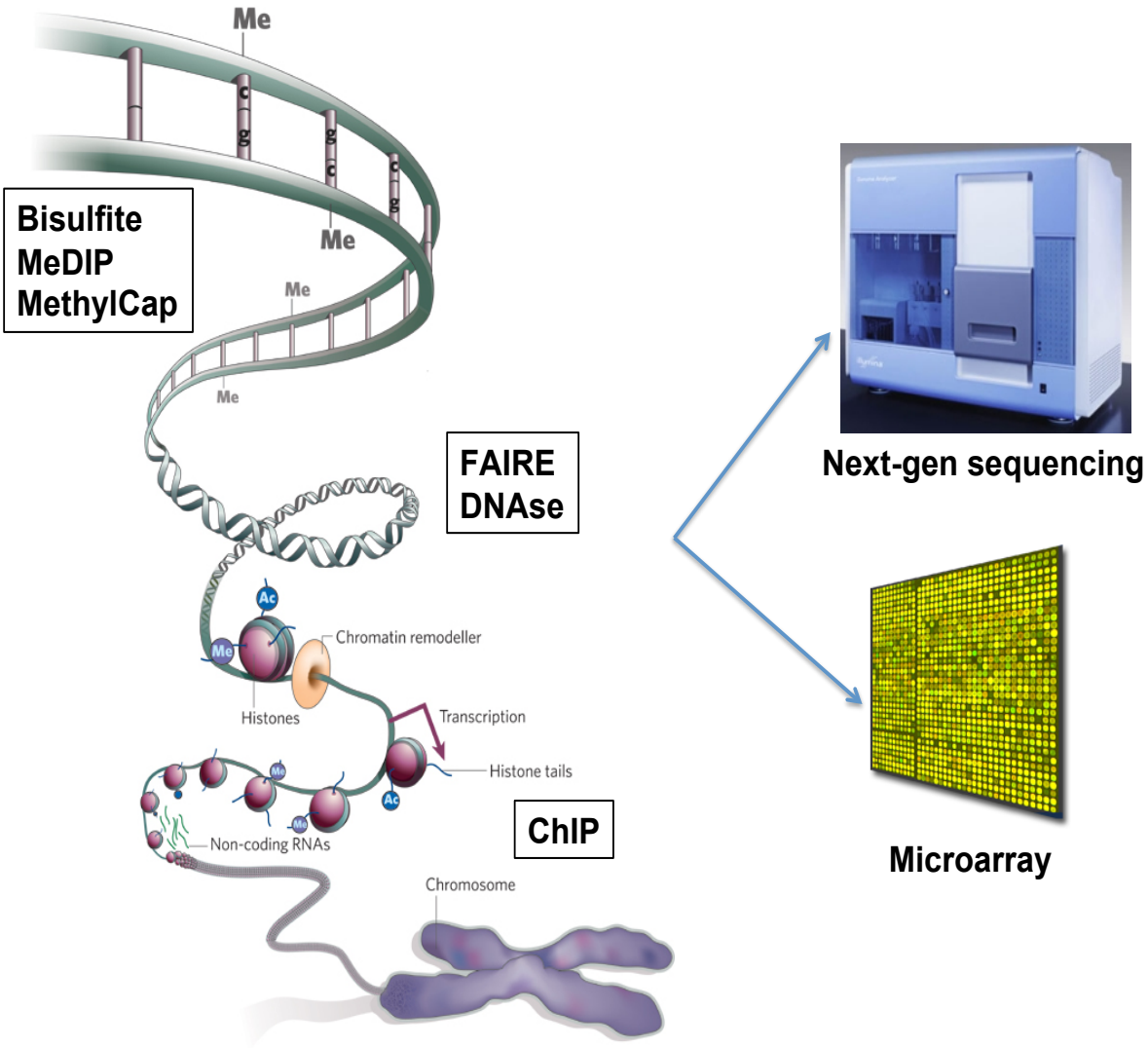


Figure 1.3.: **Epigenomics meet high-throughput data technology**, adapted from (Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, 2008). Many epigenetic components, e.g DNA methylation, histone modification, DNA sensitivity can be profiled by either microarray or NGS technology. For example, DNA methylation is profiled by BS-Seq, MeDIP(Methylated DNA immunoprecipitation) or MethylCap. ChIP is used to profile the histone modifications, and FAIRE (formaldehyde-assisted isolation of regulatory elements), and DNAse assays are used to profile the open chromatin regions and nucleosome positioning.

millions of template sequences. Applications of NGS are very diverse including genome (re-)sequencing, transcriptome/small-RNA sequencing or ChIP-Seq and sequencing of BS-treated DNA. Though microarrays are still useful and complementary for certain experimental purposes (e.g the Capture Sequencing technology Mercer *et al.*, 2012), NGS has remarkably boosted the epigenetic research. It offers the opportunities to discover many novel biological insights or to test the hypotheses based on small-scale observations over the past. However, it also raised a lot of computational challenges that need to be addressed.

1.2.2. Methylome profiling by bisulfite deep sequencing (BS-Seq)

Although there are many different technical frameworks to profile DNA methylomes (for a detailed review, see (Lister and Ecker, 2009)), all methods based on BS conversion share the same workflow: first, genomic DNA is fragmented and ligated to framework-specific methylation-annotated sequence adapters; in the second step, DNA is treated with BS under denaturing conditions, upon which unmethylated cytosines in adapter sequences and genomic DNA undergo deamination to uracil. Subsequent PCR amplification and enrichment with primers complementary to the adapters yield sequencing libraries, which are then sequenced in parallel by the appropriate sequencers.

The resulting fluorescence signals are detected and quantified to produce the library of short-length DNA sequences (called BS-Seq short reads) which are aligned with the reference genome for mapping and further downstream bioinformatics analysis. Depending on the choice of the adapters and the sequencing protocol, the short reads can represent the forward only or forward and reverse reads from the Watson ($W+$ or $W-$) or Crick strand ($C+$ or $C-$) of the reference genome (Fig. 1.4). In case of Arabidopsis BS-Seq, both possibilities have been applied: due to different sequence tags on the BS-converted sequences, Lister *et al.* (2008) have produced W+ and C+ reads, whereas Cokus *et al.* (2008) produced all four possibilities. This has to be considered for read mapping to the genome. According to a recent comparative study (Harris *et al.*, 2010b), BS-Seq is currently considered to be the most reliable and precise method to achieve a large-scale single-bp resolution DNA methylation profile.

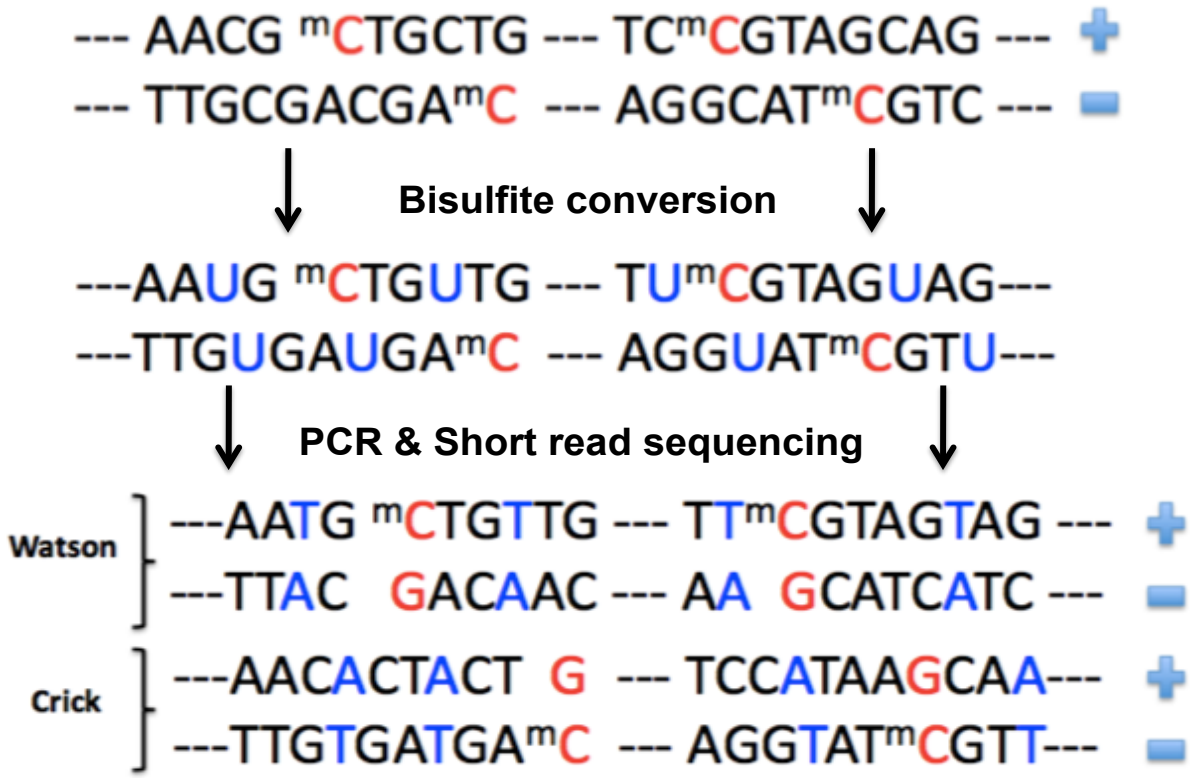


Figure 1.4.: **Bisulfite deep sequencing scheme:** first, the DNA fragments undergo BS conversion in which unmethylated cytosines are converted into uracil; this is followed by PCR amplification and sequencing to produce short reads which come from either of 4 directions.

1.3. Computational challenges

> *We are close to having a $1,000 genome sequence, but this may be accompanied by a $1,000,000 interpretation*
>
> Bruce Korf

Besides the computational challenges in high-throughput data analysis, like quality control, filtering, normalization, short-read mapping (reviewed in Bock and Lengauer (2008)), the analysis of epigenetic data poses some additional and specific difficulties. In this thesis, I discuss the challenges in computational analysis of DNA methylation and in data integration that come along with the rapid creation of profiles of chromatin marks based on genome-wide high-throughput technology.

1.3.1. BS-Seq mapping and downstream analysis

Important computational tasks in profiling genome-wide DNA methylation by BS-Seq arise from the C → T conversion of unmethylated cytosines and from DNA methylation heterogeneity within a mixture of cell or cell-types in the samples from which the DNA was prepared (reviewed in Zhang and Smith (2010)).

The loss of the perfect correspondence between read and reference genome in any sequence containing unmethylated Cs due to BS conversion (unmethylated C → T or G → A in the other strand), and the fact that the short reads are almost equally likely to be generated in 4 directions (forward and backward orientation from the two complementary strands) make BS-Seq computationally more difficult. First, the search space in finding the potential location of short reads in the reference genome is significantly increased with respect to the C/G-content, because T nucleotides in the short reads can represent either Cs or Ts in the genome. Second, the (di-)nucleotide complexity is reduced because the BS-reads are C/G-poor, especially in non-methylated genomic regions. Last, the asymmetry of the C-T matching process, i.e. T from short reads can be matched with either C or T in the reference, needs a special weighted function in the scoring-based mapping methods, like Smith-Waterman local alignment algorithm or other BLAST-like alignment methods.

Widely used procedures to map BS-Seq short read libraries apply the general-purpose mapping method with a reference genome restricted to a 3-letter nucleotide alphabet, with all Cs converted into Ts (Lister *et al.*, 2008; Harris *et al.*, 2010a; Chen *et al.*, 2010). Other, more efficient methods used a BS-converted wild-card (Xi and Li, 2009; Smith *et al.*, 2009) in searching read-reference correspondences after C → T conversion. However, one of the common major problems of current DNA methylome mapping strategies is the lack of mapping sensitivity, i.e. the number of mapped reads. This is a problem especially for experiments where a small amount of biological material limits the depth of sequencing. Hence, a highly sensitive mapping method is needed.

Apart from the mapping difficulties, there are additional hurdles in the downstream analysis of DNA methylation. One of the challenges is the choice of the appropriate statistical test for calling methylation status, e.g. how to define the systematic error rate for the widely-used binomial test in calling a cytosine methylated or unmethylated, based on the mapping coverage and profile at each individual position. The error rate,

which determines the confidence in calling a cytosine methylated or not, might cause inaccuracy due to the mixture of account sequencing errors, BS conversion errors as well as the mapping errors themselves. Hence, a comprehensive way of combining and considering these error types is also needed.

1.3.2. DNA methylation heterogeneity

Although developing nano-technology will probably allow single molecule analysis of DNA methylomes in future, current BS-Seq techniques require DNA amounts that can only be prepared from several thousands of cells. These can potentially differ in their methylation pattern, especially if the samples comprise different cell types or tissues. Current methylation profiles therefore might not reveal the extent of cell-specific differences. Understanding the heterogeneity of DNA methylation patterns, e.g. from a BS-Seq data set, is therefore an important issue (Mikeska *et al.*, 2010).

Methylation heterogeneity is expected in heterogeneous cell populations that may contain unmethylated, partially methylated and fully methylated alleles in varying proportions. The BS conversion-based methods, combined with deep sequencing, offer a great chance to investigate the degree of DNA methylation heterogeneity with computational approaches. Once the reads are unambiguously mapped and the methylation is determined for each cytosine, the information can be connected between all other cytosines combined in one read. Since each single read is expected to represent a single genomic template, i.e. one allele from one cell, we may ask if we can computationally infer the methylation characteristics of each cell/cell-types and the proportion of methylation alleles in a mixture of cells (Fig. 1.5).

This problem is analogous to that in metagenomic sequence assembly (Zhang and Smith, 2010; Peng and Smith, 2011) or to the problem of haplotype reconstruction in population genetics (Eriksson *et al.*, 2008). Although a reference sequence is available in the case of BS-Seq data, it is not trivial and classified as an Non-deterministic Polynomial-time hard (NP-HARD) problem in computer science.

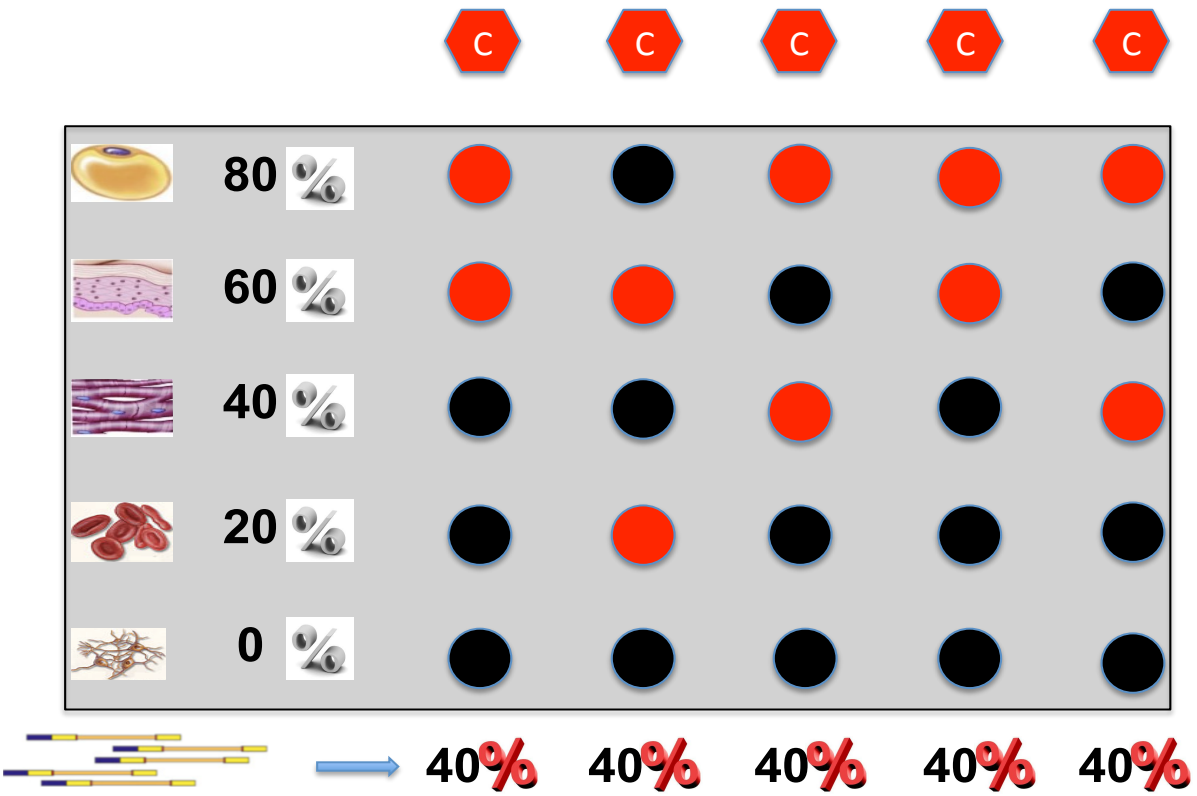


Figure 1.5.: **Uncovering DNA methylation heterogeneity from a mixture of different cell/cell-types via BS-Seq**. Red circles indicate methylated cytosine loci and black ones indicate unmethylated ones. Short-read sequencing allows to calculate the average methylation frequency (red %) for each locus. However, the actual frequency (black %) for tissues/cell(-types) can be different and needed to be inferred (in black-box).

1.3.3. Epigenomic data integration

As mentioned before, DNA methylation is just one of many different chromatin marks and should be evaluated in combination with other biologically relevant features, like profiles of histone modifications or nucleosome occupancy. Today, technologies allow to simultaneously profile multiple chromatin marks in the same or similar samples under different experimental or natural conditions (Fig. 1.6). Many biological questions can be addressed based on such synoptic data sets, but profiling multiple chromatin data presents the challenge to integrate the data, to obtain a unified view of the data and to

make them comparable. Hence, it is necessary to improve methods of data normalization or to transform and rescale different chromatin signals for comparative studies in an unbiased manner. It is further necessary to visualize multiple chromatin marks at the same genomic regions across different data scaling systems.

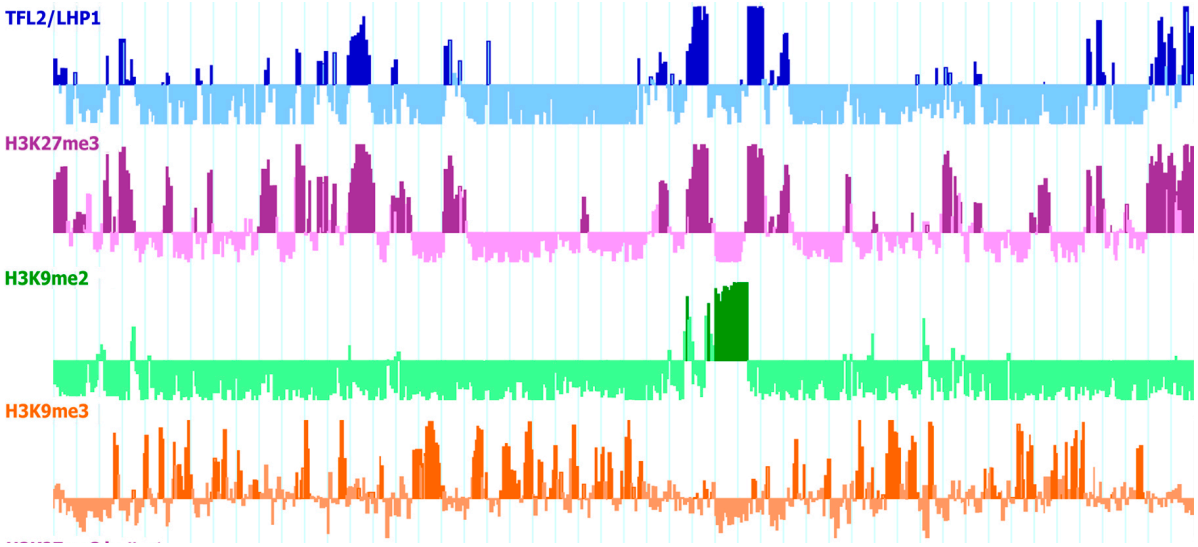


Figure 1.6.: **Multiple chromatin marks for the same genomic regions** (an example from *Arabidopsis* ChIP-chip data (Turck *et al.*, 2007))

Besides technical and biological issues to generate each individual profile for a comparative study, there is also the computational challenge to integrate the data sets and analyze them for epigenetic signatures. Scanning for common patterns across different chromatin mark data sets needs the same reference of a signal alphabet. The simplest could be the binary system of 0/1, as for DNA methylation at a single cytosine residue. However, this is not applicable for most other chromatin features that need an appropriate data transformation or representation, given the fuzziness of the epigenetic signals, often continuous scales for the signal strength, or limitations of experimental procedures. Hence, a better way to integrate and represent multiple profiling datasets is necessary.

1.4. Contributions of the thesis

In this thesis, we describe computational approaches to analyze two aspects of DNA methylation, namely how to improve the data analysis after BS-Seq, and how to infer

DNA methylation patterns in heterogeneous mixtures. In addition, we describe an approach how to integrate multiple epigenetic marks from high-throughput data sets. Accordingly, results of the thesis is organized in 3 chapters as follows.

(i) In Chapter 2, we introduce a new analysis pipeline incorporating the tolerance of C-T mismatches in the Smith-Waterman short-read local alignment for BS-Seq data, thereby increasing mapping efficiency. We also introduce a so-called adaptive error for binomial statistical test in presence of sequencing replicates in methylation calling, hence enhance the confidence in downstream analysis.

(ii) In Chapter 3, we formalize the computational problem of inferring DNA methylation patterns in mixtures of different methylomes. I developed a method called MethColor to estimate the number of distinct methylation profiles and to characterize their patterns by empirical mapping heuristics. we provided a proof-of-concept simulation study.

(iii) In Chapter 4, we propose a novel way to interpret multiple epigenetic parameters (i.e. DNA methylation level or chromatin modification distribution) in the search for biological patterns and correlations between chromatin features and gene expression data. The approach is based on the principle to convert the relative prevalence of epigenetic marks into a letter code that then can be further analyzed with existing data mining tools. we showed an application with epigenetic signatures in *Arabidopsis thaliana*.

Chapter 2.

Advanced methylome analysis after bisulfite deep sequencing: an example in Arabidopsis

In theory, there is no difference between theory and practice.
But, in practice, there is.

Jan L. A. van de Snepscheut/Yogi Berra

Summary

Deep sequencing after bisulfite conversion (BS-Seq) is the method of choice to generate whole genome maps of cytosine methylation at single base-pair resolution. Its application to genomic DNA of Arabidopsis flower bud tissue resulted in the first complete methylome, determining a methylation rate of 6.7% in this tissue. BS-Seq reads were mapped onto an in silico converted reference genome, applying the so-called 3-letter genome method. Here, we present BIsufite Sequencing Scorer (BiSS), a new method applying Smith-Waterman alignment to map bisulfite-converted reads to a reference genome. In addition, we introduce a comprehensive adaptive error estimate that accounts for sequencing errors, erroneous bisulfite conversion and also wrongly mapped reads. The re-analysis of the Arabidopsis methylome data with BiSS mapped substantially more reads to the genome. As a result, it determines the methylation status of an extra 10% of cytosines and estimates the methylation rate to be 7.7%. We validated

the results by individual traditional bisulfite sequencing for selected genomic regions. In addition to predicting the methylation status of each cytosine, BiSS also provides an estimate of the methylation degree at each genomic site. Thus, BiSS explores BS-Seq data more extensively and provides more information for downstream analysis.

2.1. Introduction

Whole genome sequencing of numerous species and individuals has considerably expanded our understanding of biological diversity and evolution, of normal and abnormal phenotypes. However, it also revealed that regulation of, and differences in, gene expression are not always connected with differences in DNA sequence information. The occurrence of different phenotypes or heritable changes of gene expression, in spite of identical genetic information, has driven the search for additional, epigenetic information transmitted from cell to cell or from parents to progeny. One major component of epigenetic inheritance and regulation is chemical DNA modification by methylation at the 5 position of cytosine residues (mC). This modification occurs in some fungi and insects, in all mammals and higher plants examined to date, and it is sometimes referred to as the fifth base. Research on the role of mC was stimulated by its potential to transmit epigenetic information during DNA replication. Its study was facilitated by the ground-breaking development of bisulfite sequencing, in which non-methylated cytosines get chemically converted into uracil and can be distinguished from methylated residues after PCR amplification and subsequent DNA sequencing (Frommer *et al.*, 1992). DNA methylation was the first epigenetic mark that could be analysed at high resolution, and its analysis profited substantially from the rapid development of sequencing technologies. It is now accepted as one of the most comprehensive and efficient methods (Harris *et al.*, 2010b).

Bisulfite conversion followed by deep sequencing (BS-Seq) has been successfully applied in many species and cell types to analyze the methylome. However, the mismatches after converting unmethylated cytosine residues make the mapping of short reads during BS-Seq more challenging than during genome sequencing. Not for the first time, pioneering epigenetic research was performed in plants, as the first whole methylome was established for Arabidopsis thaliana (Cokus *et al.*, 2008; Lister *et al.*, 2008). In addition, some plant genomes have lower levels of total mC compared to that of mammals,

therefore more mismatches after bisulfite conversion, and more mC in a non-CG context. BS-Seq of genomic DNA isolated from flower buds was fragmented, ligated with methylated adaptors, followed by bisulfite conversion prior to PCR amplification and deep sequencing. The total mC content was calculated as 6.7% of those C positions for which the methylation status could be determined (Lister *et al.*, 2008). This is in good agreement with experimentally measured values in the range from 4.6 to 8.6%, obtained by different methods and with different tissue (Kakutani *et al.*, 1999; Leutwiler *et al.*, 1986; Rozhon *et al.*, 2008). However, we noticed a discrepancy between the total mC content calculated after BS-Seq and the frequency estimated from counting cytosines occurring in the raw data from the short-read libraries (Lister *et al.*, 2008). Any C in the sequence between the methylated adaptors should directly correspond to methylated cytosines in the genome, complete conversion and low sequence error rates provided. We calculated three Illumina sequencing runs of the data set in (Lister *et al.*, 2008) to report roughly 10% mC, while the other two suggest $23 - 26\%$, probably due to incomplete bisulfite conversion, or sequencing errors. Pooling all five runs would correspond to 14.7% methylation. We suspected the discrepancy to originate from limited mapping of individual short reads to the reference genome since only 78.5% of genomic cytosines were included in at least 2 mapped reads (Lister *et al.*, 2008). This could have been due to the mapping procedure: the so-called three-letter genome method, in which all genomic cytosines are converted in silico to thymine, before the reads are mapped using ELAND software from the Illumina company, interpreting C-T mismatches as indicative for methylated cytosines during the downstream analysis (Lister *et al.*, 2008). We refer to it as Arabidopsis 3-letter Methylome (A3M) method in the following.

Aiming to improve mapping efficiency and accuracy for analysis of plant material, we have developed BiSS (BIsulfite Sequence Scorer), based on an efficient Smith-Waterman (SW) local alignment implementation for BS-Seq mapping with a customized alignment scoring function. SW has the potential to produce superior alignments due to the base-by-base resolution in sequence comparison. Previous (Ning *et al.*, 2001) and the recent work (Sedlazeck *et al.*, 2012, submitted) suggests that SW local alignment is in fact the most sensitive method for NGS mapping available to date. High specificity and confidence are obtained with this method, admittedly at the cost of increased computing time. SW local alignment was implemented in the MAQ program recently (Li *et al.*, 2008; Chen *et al.*, 2010) but not yet evaluated in comparison with other methods. Therefore,

we applied this SW approach using a special asymmetric score for BS-Seq data to re-analyze the Arabidopsis methylome data set with BiSS.

For the data analysis downstream of mapping the reads, we introduced a comprehensive adaptive error estimate that accounts for sequencing errors, erroneous bisulfite conversion and also wrongly mapped reads. With BiSS, we were able to map many more short reads unambiguously to the reference genome than other methods. The increased coverage gives increased power to call an individual cytosine methylated or un-methylated, thus allowing the determination of methylation status at significantly more sites. The re-analysis of the Arabidopsis methylome dataset using BiSS and the adaptive error estimate could identify the methylation status of an extra 10% of genomic cytosines and resulted in estimation for the global methylation to be 7.7% of all cytosines. We validated these results by traditional individual traditional bisulfite sequencing (ITBS) at several randomly chosen genomic regions. In all but one locus these results confirmed the prediction from our BiSS analysis. Moreover, these data show that the BiSS method provides an accurate estimation of the degree of methylation at individual partially methylated genomic sites.

2.2. Results

2.2.1. BiSS can map more reads unambiguously to the reference genome

BiSS calls a read uniquely mapped if it can identify only one SW-alignment with the highest score. To avoid mapping artefacts we excluded reads with an alignment identity (not considering bisulfite mismatches) below 85%. More than half (53.2%) of the raw reads were above this threshold and were used for the downstream analysis. In total, we were able to map approximately 77 million unique reads, 1.96 x times more than the A3Mapproach used in the original data analysis.

There are also several other published methods to analyze BS-Seq data, such as BSMAP (Xi and Li, 2009), RMAPBS (Smith *et al.*, 2009), BRAT (Harris *et al.*, 2010a), BS-Seeker (Chen *et al.*, 2010), PASH 3.0 (Coarfa *et al.*, 2010), BisMark (Krueger and Andrews, 2011), and MethylCoder (Pedersen *et al.*, 2011). We compared BiSS to a

Method[a]	Number of mapped reads	Number of analyzed reads[b]
A3M[c]	55,805,931(38.6%)	39,113,599(27.1%)
BSMAP[d]	**73,215,737 (50.7%)**	**47,922,346 (33.2%)**
RMAPBS[e]	64,061,732(44.4%)	47,859,115(33.1%)
BS-Seeker	51,657,927(35.8%)	37,939,172(26.3%)
BisMark	50,324,319(34.8%)	37,706,400(26.1%)
BiSS	**103,073,409 (71.4%)** [f]	**76,841,502 (53.2%)**[g]

[a]default parameters unless otherwise specified
[b]Uniquely mapping, except BiSS
[c]A3M results reported by (Lister et al. 2008)
[d]BSMAP parameters: -p 8 -s 12 -r 2 -w 100 -n 1 -v 5 -g 5, recommended by the authors, maximal 5 mismatches
[e]RMAP parameters: -m 5 v, default parameters
[f]uniquely highest SW scored alignments
[g]above 85% identity

Table 2.1.: **Comparison of mapping results for BiSS and other selected programs**

selection of these, including the most recently published aligners BisMark (Krueger and Andrews, 2011) and BSMAP (Xi and Li, 2009), the most sensitive mapping according to previous comparative studies (Harris *et al.*, 2010b; Chatterjee *et al.*, 2012). With the parameters recommended by the authors (Xi and Li, 2009), BSMAP mapped 1.60 times less reads than BiSS (Table 2.1). RMAP mapped 1.61 times less reads, and all other methods performed in a comparable range or less (Table 2.1). To compare the BiSS-generated Arabidopsis methylome with the previous interpretation of the same data, we chose A3M for a more detailed comparison, since this was applied in the pioneering approach to generate a single-bp methylation profile after mapping. Details on the comparison and mapping statistics can be found in Supplementary Tables 1-5 [1]. In summary, BiSS almost doubles the number of mapped reads that can be used for mC analysis. This translated to over 10% more cytosines in the genome that are covered by at least 2 mapped reads, the minimum required for methylation calling in the A3M approach.

[1]Supplementary tables 1-20 for this chapter are in a large EXCEL file deposited online at `http://www.cibiv.at/software/ngm/BiSS`

2.2.2. BiSS extends the methylome of Arabidopsis thaliana

Since the number of Cs for which the methylation status could be assigned differs between the two methods, we investigated the degree of overlap between them (Table 2.1).

		A3M		
		M (5.3%)	**U (73.6%)**	**X (21.1%)**
BiSS	**M (6.9%)**	1,839,780 M/M (4.3%)	371,418 M/U (0.9%)	749,156 M/X (1.7%)
	U (82.6%)	397,308 U/M(0.9%)	30,970,572 U/U(72.2%)	4,028,624 U/X(9.4%)
	X (10.5%)	30,359 X/M (0.07%)	230,988 X/U(0.54%)	4,257,806 X/X(9.9%)

Table 2.2.: **Congruency between methylation calling by A3M and BiSS:** M: methylated, U: unmethylated, X: not determined due to lack of sufficient sequencing coverage. Percentages refer to the total number of genomic cytosines.

There was good agreement (76.5%) between the two methods when classifying methylated (M/M) and unmethylated (U/U) cytosines. For 9.9% of cytosines neither A3M nor BiSS could make a call (X/X). However, BiSS was able to determine the methylation status of 89.5% of the genomic Cs, in contrast to 79% for A3M. In total, BiSS called 6.9% of all Cs methylated (Table 2.2), 30% more than A3M, which scored 5.3% methylated. The additional Cs called methylated by BiSS were mainly from the fraction where A3M was unable to make a call (X in Table 2.2). However, some Cs (0.9%) called unmethylated by A3M were assigned to the methylated category by BiSS (M/U). A substantial fraction (9.4%) of Cs, for which A3M could not call the methylation state, was assigned to the unmethylated category by BiSS (U/X). Small shifts also occur in the opposite directions: 0.9% A3M-called methylated Cs are considered unmethylated by BiSS (U/M), and only 0.6% of Cs not determined by BiSS were assigned by A3M (X/M and X/U, Table 2.2). Thus, the more efficient mapping procedure employed in BiSS was able to considerably reduce the uncharted portion of the methylome. In summary, the analysis of the data set by BiSS largely corroborates the previously published analysis on the amount and distribution of genomic mCs but was able to determine the

methylation status at 10.6% more sites in the reference genome. This reanalysis indicates higher levels of mC i (7.7%) in flower tissue than previously reported to (6.7%).

To gain a deeper insight into the different performance of both methylation assignment approaches, we computed the differences corresponding to the sequence context of the cytosines (CG, CHG, CHH; with H=A, C, or T). The results are illustrated in Figure 2.1. In the reference genome, CHH is naturally most frequent (73%), followed by CHG and CG, the latter occurring at almost equal frequency (Figure 2.1). The frequency distribution of mC with respect to the sequence context shows a strong preference for mCG as expected, and is nearly identical for A3M and BiSS (Figure 2.1). Thus, although BiSS assigns a methylation status to more genomic Cs, it does it without a bias for any sequence context.

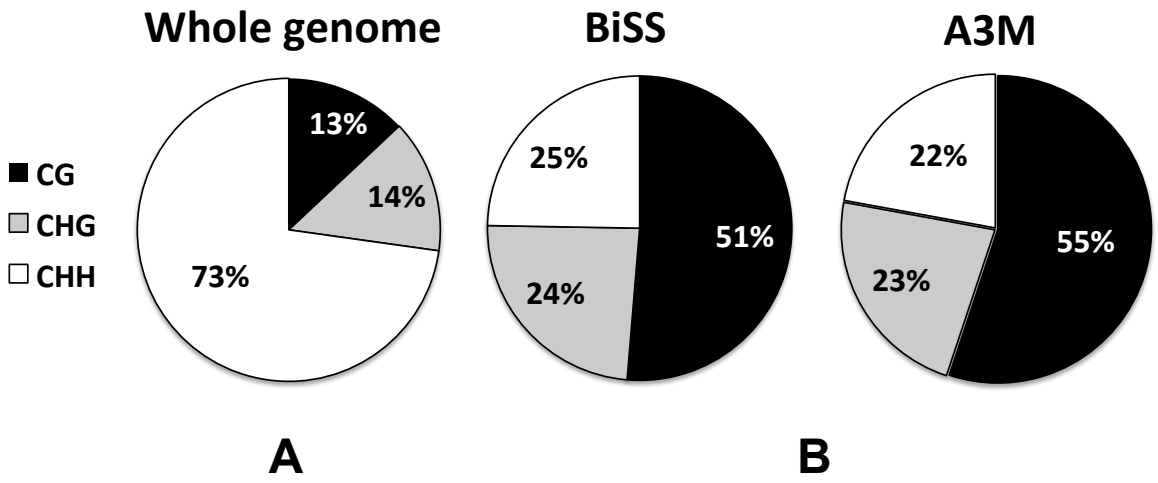


Figure 2.1.: **Distribution of cytosine sequence context.** (A) Frequency of sequence context in the reference genome. (B) Frequency of sequence context of methylated C according to BiSS and A3M.

We have further split the congruency assignment (Table 2.2) into the C-sequence context (Figure 2.2). Sixty five per cent of methylated Cs identified by both methods (M/M) occur in a CG dinucleotide context. The frequencies of C-contexts for unmethylated Cs (U/U) are almost identical to their genomic frequencies. The methylated Cs only called by BiSS (M/X) were in all sequence contexts, while unmethylated Cs only called by BiSS (U/X) occurred largely in the CHH context (Figure 2.2). The few Cs only called

by A3M have a similar distribution. Taken together with the absolute numbers in Table 2.2, it can be concluded that the SW scoring method used by BiSS is able to assign a methylation status to a significant number of CHH sites that could not be called by A3M. This suggests that the CHH context is more challenging to map, as seen from the large fraction of CHH sites where the two methods either disagree or both fail to make a call.

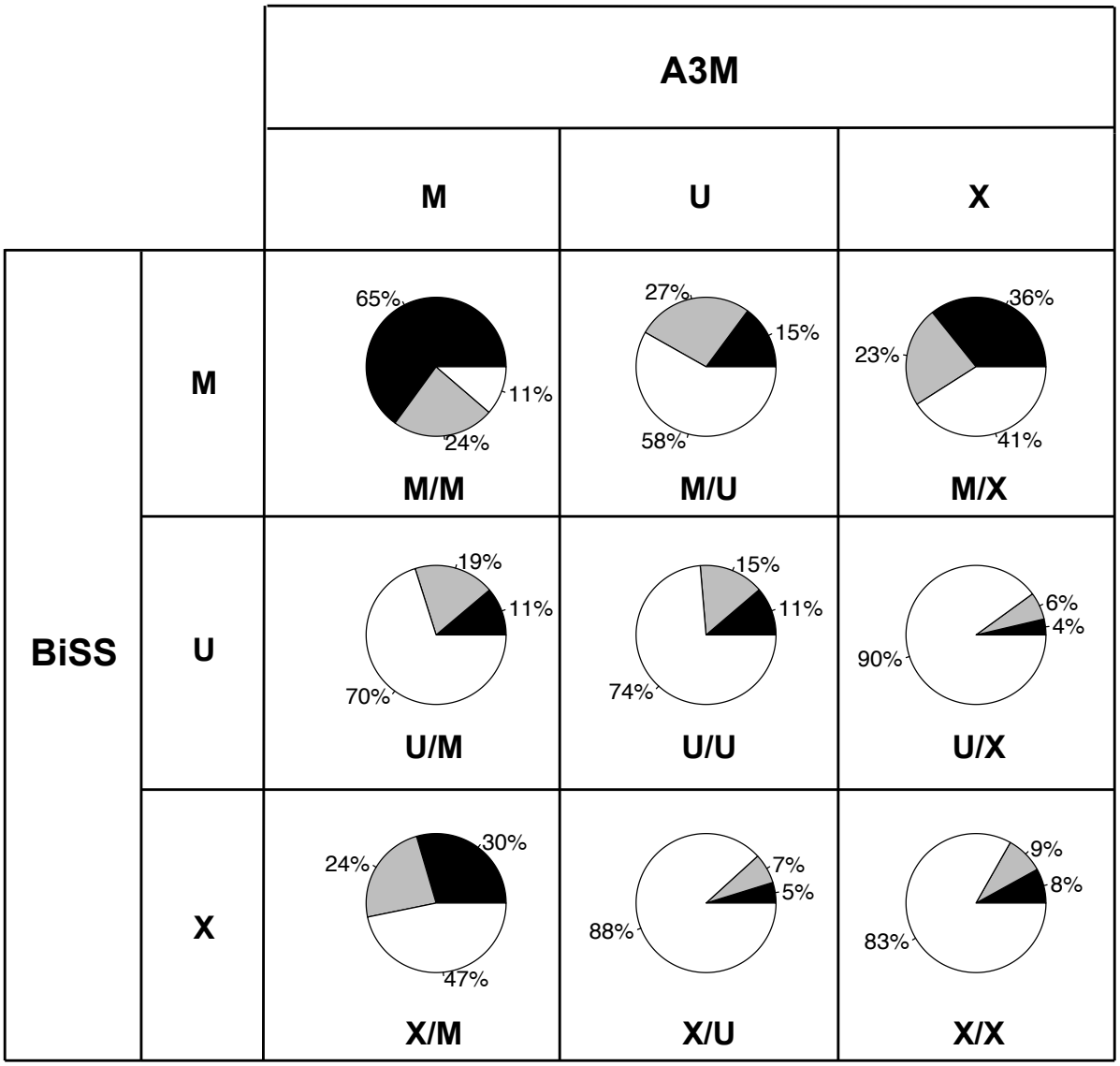


Figure 2.2.: **Methylation status according to BiSS and A3M split into distribution of cytosine sequence context.** M, methylated; U, unmethylated; X, not determined. Percentages refer to the total numbers in Table 2.2.

2.2.3. Evaluating the methylation level

The decision to call a genomic C as methylated is based on a statistical test that considers the number of reads mapped to a genomic C, the C/T counts at the site and the estimated adaptive error (see Material and Methods). Thus, a genomic C can be called methylated with confidence even if not every mapped read contains a C at that site. The coexistence of Cs and Ts at individual positions reflects the biologically well-known heterogeneity of methylation between alleles in the same genome or in different cell types, tissues or individuals. A plot of the degree of mC, calculated as the C/(C+T) ratio of mapped reads for each C in the reference genome shows that this ratio varies across the entire range from 0-1 (Figure 2.3).

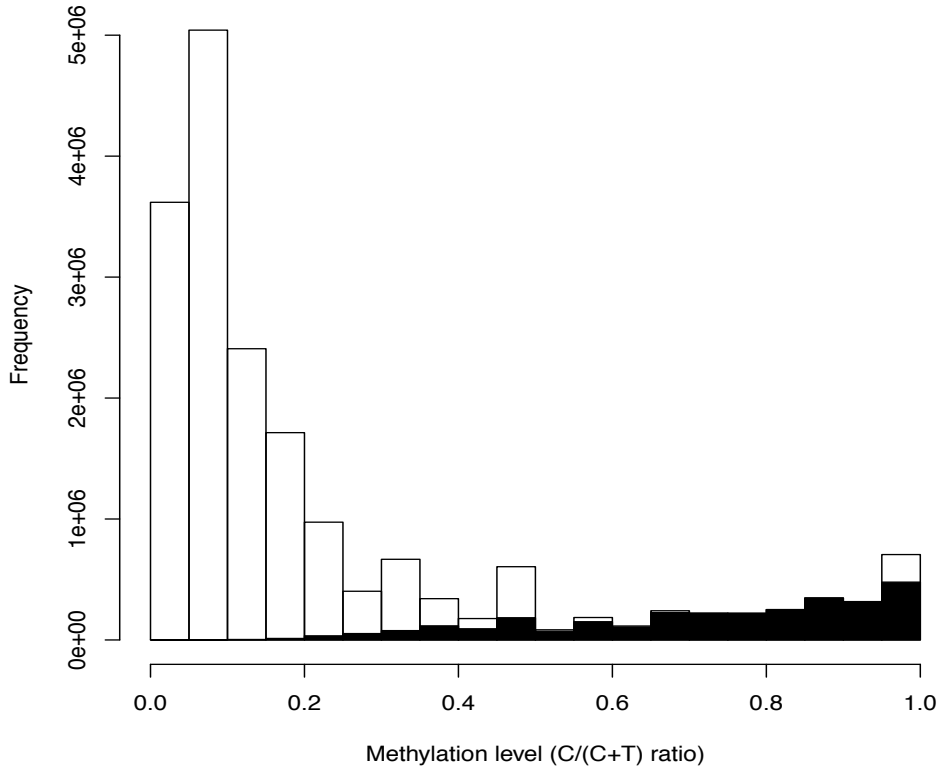


Figure 2.3.: **Global methylation level.** Ratio of the number of mapped Cs divided by the number of mapped C plus T for all classified Cs. Black: Cs that are called methylated, White: C that are not called methylated.

For the majority of the Cs that BiSS calls methylated, the ratio is typically above or equal to 0.5. However, 1.1% of the mCs in the genome display a C/(C+T) less than 0.5. This suggests that genomic Cs with a C/(C+T) ratio larger than 0.4, say, are probably methylated but were not called as such due to the very conservative nature of the test. Thus our estimate of 7.7% methylated Cs is likely to be an underestimate, and the true methylation level may be higher (for example, if all Cs with a C/(C+T) above 0.4 are called methylated we get an estimate of 9.1%.

2.2.4. BiSS methylation calling is validated by independent bisulfite sequencing

To confirm the improved accuracy of the BiSS method compared to A3M, methylation levels at selected regions of the genome were independently determined by individual traditional bisulfite sequencing (ITBS) and compared to the results from BiSS and A3M. We selected 2 regions where both methods reported high methylation (M/M) and two regions were both reported no methylation (U/U). We further identified 4 regions where the two methods disagreed in methylation calling (2 x M/U and 2 x U/M), and 6 regions that were mapped by BiSS but lacked sufficient sequencing coverage in A3M results (M/X and U/X). The regions represent genic, intergenic and repetitive sequences (Table A.1). For each region, we compute the Pearson correlation coefficient between calculated methylation levels from A3M/BiSS and ITBS. Representative correlations for each comparison category are shown in Figure 2.4.

The BiSS/A3M methylation levels were confirmed for the U/U and M/M regions, Figure 2.4, (Supplementary Figure A.1, A.2 and Supplementary Tables 6-9). However the BiSS prediction had a higher correlation with the ITBS data, even if both methods call high methylation. For one of the two regions of the M/U category (Figure 2.4, Supplementary Table 10), BiSS, but not A3M, results were in good agreement with the ITBS data. However, at the second M/U region the BiSS results did not agree with the ITBS data (Supplementary Figure A.1, Supplementary Table 11). This appears to be because there were only 1-2 sequencing runs on which BiSS based the methylation call. The low coverage in this region is likely due to having only a few cytosines, exclusively in CHH context, which are more difficult to map. However, even at this locus the correlation is not much different for both methods ($r = 0.37$ for A3M and $r = 0.33$ for

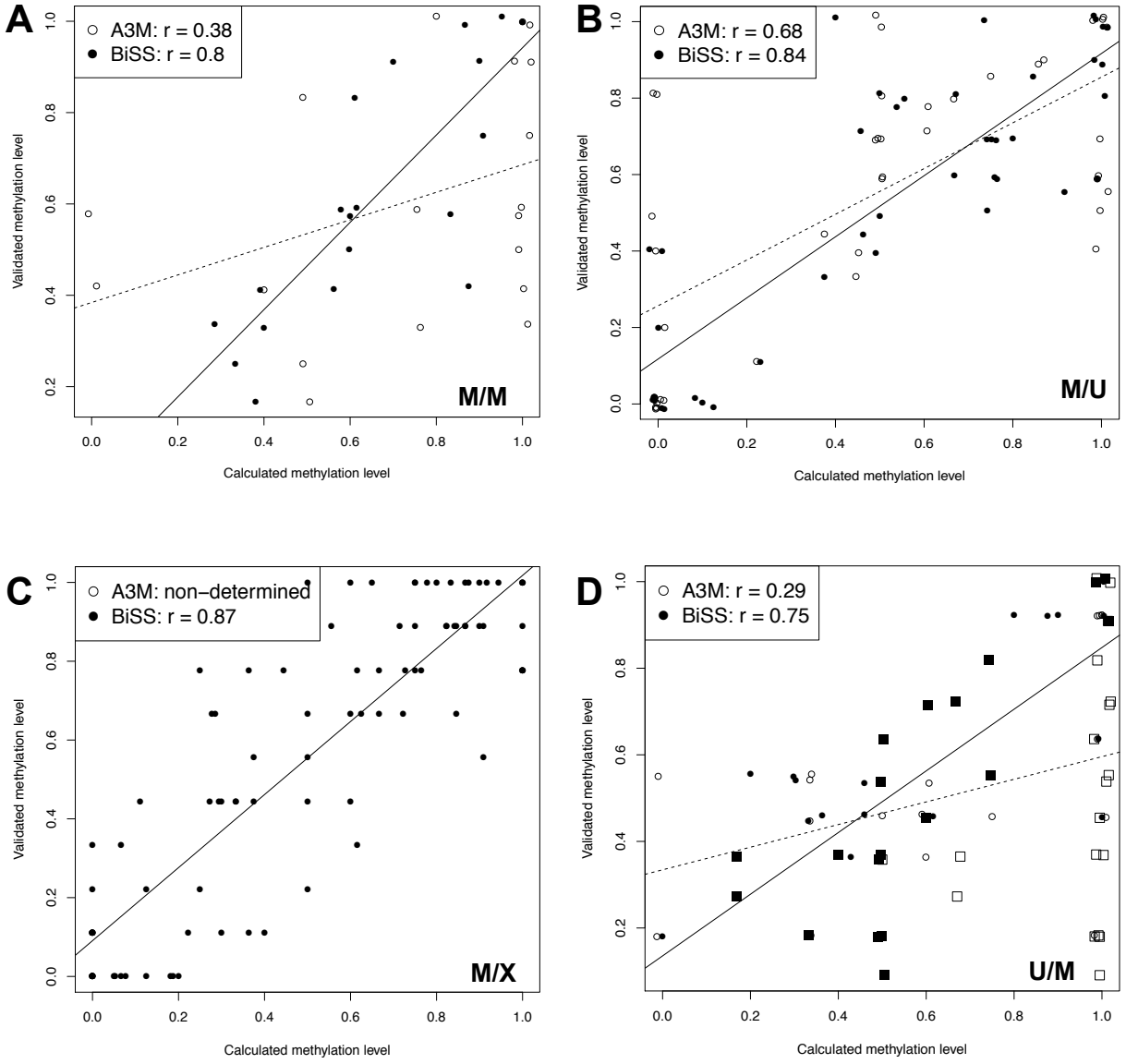


Figure 2.4.: **Examples for validation by individual bisulfite sequencing.** The plots show the correlation between calculated and validated methylation levels (C/(C+T)) from regions selected for congruency (A) or disagreement (B-D) between BiSS and A3M. Each point represents one cytosine position. The x-axis corresponds to the methylation levels calculated from either BiSS (filled circles and black regression lines) or A3M (open circles and dotted regression lines); the y-axis shows the result of individual bisulfite sequencing. The legends show the Pearson correlation coefficients. (A) Methylated region according to both methods (M/M). (B) A region called methylated by BiSS but not by A3M (M/U); the rectangles indicate experimentally validated Cs congruent to BiSS (filled) and discrepant to A3M (open). (C) A region called methylated by BiSS but not by A3M due to insufficient sequencing coverage (M/X). (D) A region called methylated by A3M but not by BiSS (U/M), the rectangle symbols are the same as in (B).

BiSS). Apart from this exception, there was good agreement between BiSS results and Sanger data for regions where Cs could not be classified by A3M. One example is shown in Figure 2.4, five others in the Supplementary Figure A.3 and Supplementary Tables 12-17. The U/M regions classified by A3M as highly methylated but categorized by BiSS as unmethylated were indeed unmethylated according to ITBS. The reason for better performance of BiSS appears to be due to either run-specific information (Figure 2.4, Supplementary Table 18) or simply by gaining higher coverage (Supplementary Figure A.1, Supplementary Table 19). Thus, the BiSS predictions had a higher correlation with Sanger data compared to A3M not only for the methylation calling but also for predicting the level of methylation.

An interesting case was a region called methylated by A3M (with fully methylated for many individual sites), which was in strong disagreement with the ITBS data, with a negative correlation coefficient. The BiSS results had a high correlation with the ITBS data but did not call any methylation there (Supplementary Figure A.4 and Supplementary Table 20). A closer look revealed that the A3M-determined methylation level was based on a read coverage of only 2. Notably, this indicates that we still underestimate the methylation rate due to the very conservative test mentioned above (Figure 2.3). Thus, BiSS could obtain a higher coverage and indicated rather low methylation here. This supports the notion that BiSS can provide a more realistic interpretation of the BS-Seq data, especially in regions where A3M suffers from low coverage, and at many CHH sites. In summary, BiSS can help to improve mapping of BS-Seq reads to the reference genome, to provide higher coverage, and to provide a refined and more accurate methylome map.

2.3. Discussion

BiSS, a scoring method for whole genome bisulfite deep sequencing data, takes advantages of SW alignment to evaluate the bisulfite conversion as an add-on for the general SW-based mapping package (NextGenMap Sedlazeck *et al.*, 2012, submitted). In addition, BiSS incorporates an adaptive error into the binomial test to correct for the mismatch ratio including sequencing or mapping errors in the downstream analysis. Moreover, BiSS also exploits the potential of considering run-specific information, which can reduce the effect of errors introduced by sequencing bias. It also allows the separate

analysis of individual sequencing runs representing biological replicates.

The re-analysis of previously published BS-Seq data from Arabidopsis by BiSS increased the number of cytosines for which the methylation status could be reliably determined by 10% largely due to higher mapping efficiency. In particular, BiSS successfully identified the methylation status at a significant number of CHH-context cytosines, where the A3M method performs poorly. Independent bisulfite sequencing confirmed the BiSS predictions at regions where it disagreed with the A3M method. It also confirmed that BiSS more accurately predicted the level of methylation at partially methylated cytosines. Thus, BiSS provides a new and more accurate reference for the floral Arabidopsis methylome.

We note that some researchers prefer to trim or filter reads prior to the alignment step to remove bases with low quality scores. To test if filtering prior to alignment affected the performance of BiSS we repeated the alignment after filtering the raw reads using the FASTX toolkit `http://hannonlab.cshl.edu/fastx_toolkit/index.html`. When mapping this filtered data set of 107 Mio (71%) reads, BiSS was still able to map 10% more reads than BSMAP, the best performing of the published aligners. BiSS could also be applied to other Arabidopsis methylome datasets obtained from different material (Cokus *et al.*, 2008; Zemach *et al.*, 2010), and to data from human DNA (Table A) in the same order of magnitude in term of running times compared to existing methods.

The improved performance of BiSS with respect to the number of mapped reads for methylation analysis is mainly due to the SW-based mapping method, applied here to bisulfite deep sequencing data in open-source software. The algorithm compares subsequence of different lengths and thereby optimizes the similarity detection, compared to other BS-Seq mapping methods, which either encode the reference genome in a three letters alphabet or use special bisulfite conversion masks for mapping the reads with general-purpose software. SW alignment has also the ability to stop aligning, if the reads get too different, due to increasing sequencing errors towards the end of longer NGS reads.

In attempting to maximize the number of aligned reads, one runs the risk of generating a data set containing a significant portion of incorrect alignments. Therefore, we independently validated the mC frequency at selected genomic regions (Figure 2.3), with results that excluded this possibility (Table A). Naturally, the SW-based method requires extra computing time (Table A) compared to other methods. However, in many

cases the gain of extra mapping information will outweigh this disadvantage. As long as the costs of NGS remain an issue, at least for researchers outside of large genome centers, it is reasonable to apply optimized evaluation methods. BiSS can also align both single & paired-end as is described in the manual. Our results suggest that the improved performance of BiSS compared to competing methods is in large part due to its superior ability to assign the methylation status to cytosines outside of CG context. This is sure to be appreciated by some researchers, given the growing evidence for mCHG and mCHH in specialized mammalian cells (Lister *et al.*, 2009).

2.4. Experimental and computational procedures

2.4.1. Mapping deep sequencing reads after bisulfite conversion

BiSS uses the SW local alignment to map the sequencing reads after bisulfite conversion (BS-reads) to the reference genome with a special scoring function. To speed up computation, a hash-table stores the positions of all k letter words (k-mers) in the genome. The k-mers are encoded as numbers (keys) as follows: Nucleotide A is converted into 00, C to 01, G to 10, and T to 11. Thus an 8-mer results in a string of 16 zeros or ones; this string can be converted into an integer number that serves as key to point to the genomic positions. Keys were also computed for all k-mers in a BS-read. Together with the hash-table, the BS-read keys allow a quick retrieval of the genomic positions in the reference genome, where read and genome share the same k-mers.

Because bisulfite conversion turns unmethylated Cs into Ts, the 1-to-1 correspondence between k-mer in a read and in the reference genome is lost. To account for this, a precomputed look-up table was generated that for each k-mer stores the alternative keys that can be computed by switching a C into a T. Figure 2.5 exemplifies this for the 8-mer A1C2G3T4C5G6C7T8 (key 7015), switching C7 into a T provides key 7023 and so on. Thus, the k-mer A1T2G3T4C5G6T7T8 with key 15215 from a BS-read will automatically be associated with the potential genomic regions from the additional keys in the look-up table. The look-up table needs to be computed only once, thus saving computing time. In this study k was set to 12.

The hashing table is searched for k-mer co-occurring in a BS-read and the reference genome to determine the potential locations of SW alignment. To reduce the number

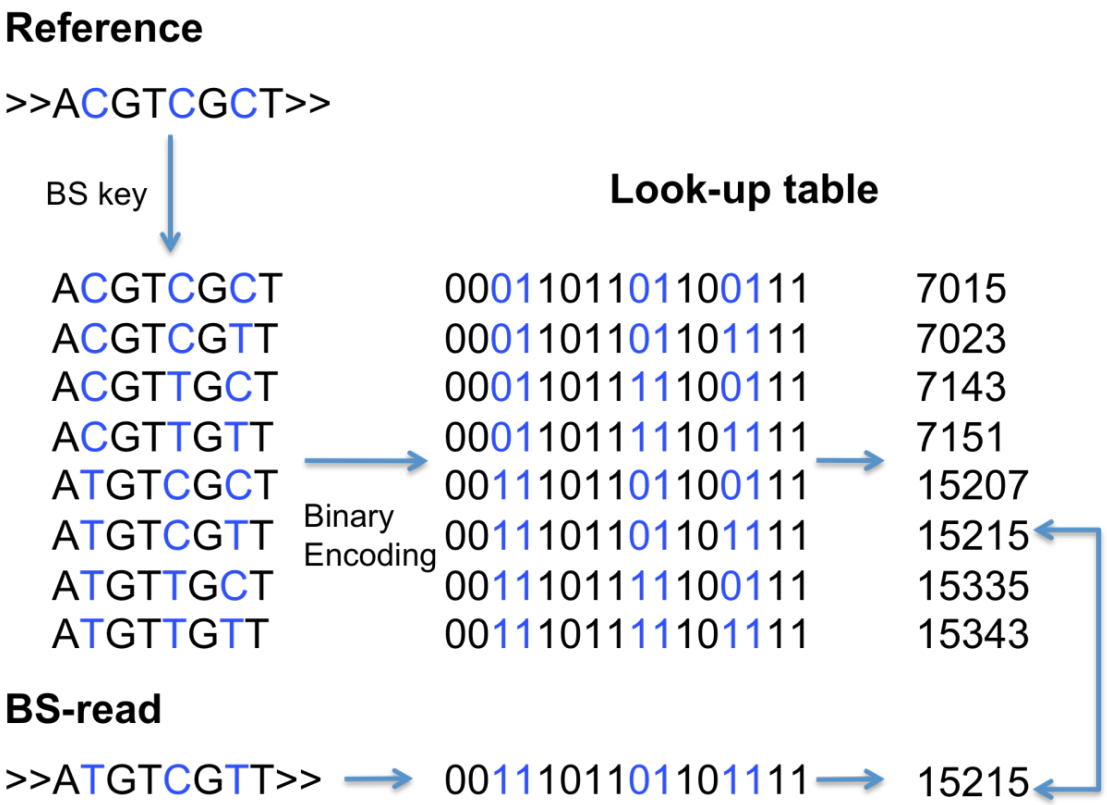


Figure 2.5.: **Example for an asymmetric look-up table for 8 k-mers.** The 8-mer ACGTCGCT (corresponds to key 7015) generates 7 other keys. The 8-mer ATGTCGTT (key 15215) from the BS-read can be looked up to find its referenced key as ACGTCGCT (key 7015), ACGTCGTT (key 7023), ATGTCGCT (key 15207) and ATGTCGTT (key 15215) but no others.

of potentially matching genomic regions that need to be scored, at least two k-mers in the BS-read must occur in close proximity in the reference genome. Moreover, we allow that the distance between two neighboring k-mers in the read and their distance in the corresponding genomic sequence can differ by 3 nucleotides. Finally, to reduce the number of unspecific matches, we excluded all 12-mers from hashing that contain 8 or more Ts. The parameter 8 is the default parameter that can be specified by the user. For these reads, the SW algorithm is applied with a special scoring function, specifically a score 4 for match (including C-T mismatch, where C occurs in the reference genome and G-A in other strand), -2 for mismatch, -10 as gap penalty. The genomic region

providing the highest alignment score is considered as the genomic origin of the read. To cope with the huge number of SW alignment computation the package NextGenMap (Sedlazeck *et al.*, 2012, submitted) which implements a banded SW algorithm speeded up by Graphical Processing Unit (GPU) computing was used.

2.4.2. Re-analyzing the Arabidopsis thaliana methylome

BiSS was used to re-analyze the BS-read data of Arabidopsis thaliana Col-0 wild type generated by (Lister et al. 2008). This read library consists of 5 Illumina runs with approximately 150 million 56-bp BS-reads, thus providing a theoretical coverage of roughly 56. After mapping the reads, only the uniquely mapped reads with at least 85% similarity (calculated after excluding C-T mismatches on the Watson strand and G-A mismatches on the Crick strand) were further analyzed. Following (Lister *et al.*, 2008), only genomic cytosines with at least 2 mapped reads are further used for statistical calling of the methylation status (see below). To compare BiSS results with those of A3M, the same assembly version (TAIR7 Arabidopsis thaliana) was used as a reference. The alignment profile of the A3M method was downloaded from NCBI Gene Expression Omnibus (accession number GSE10877), the list of methylcytosines was provided by the authors of (Lister *et al.*, 2008).

2.4.3. Methylcytosine calling

To determine if a specific cytosine was methylated a binomial test was performed. The parameters of the binomial distribution are n as the coverage at a genomic cytosine position, p=0.04 as the assumed sequencing or conversion error, and m as the number of BS-reads that carry a C at that position, indicating methylation. In addition, a so-called adaptive error was introduced. Whenever the frequency of non C-T mismatches at a given genomic cytosine position is bigger than the default error, we used the local mismatch frequency as parameter for the binomial distribution. Thus, the adaptive error will reduce a potential bias due to reads that are aligned to the wrong genomic region. To account for multiple testing, the False Discovery Rate p-value adjustment based on Benjamini and Hochberg (Benjamini and Hochberg, 1995) from the R statistical computing package (`www.R-project.org`).

The actual methylcytosine calling was done in several steps: First the pool of all mapped reads from the 5 sequencing runs was considered to identify genomic Cs with a read coverage of at least 2. Then a binomial test was applied as described above, together with the FDR correction, to test the methylated cytosines with significance cut-off of 5%. The resulting list of methylated cytosines was then analyzed to account for differences between sequencing runs. The binomial test was then applied for all mapped reads from individual runs, again requiring coverage of at least two. If the majority of runs where the test could be performed suggested methylcytosine, the genomic C was called as methylated, otherwise not. This approach considers the experiments with varying bisulfite conversion rates in different runs. In case of no sufficient coverage in any individual run, the methylation decision was based on the global test in the pooled set.

2.4.4. Experimental validation

A window-scanning strategy was used to identify genomic regions of 250-500 bp for which the BiSS and A3M methods were in disagreement as to the extent of calculated methylation. These sequences were analysed for their methylation level by conventional bisulfite sequencing of individual regions. For this, plants of the Arabidopsis thaliana accession Col-0 (Columbia) were grown under long day (16h/d) light condition at 21C and DNA was extracted from 100 mg of unopened flower buds using the Phytopure DNA extraction kit (GE Healthcare; Little Chalfont, UK). After an additional RNase A treatment, 1 g of DNA was digested with either EcoRI or KpnI (excluding a restriction recognition site between the primers for the region) and purified with a PCR purification kit (Qiagen; Hilden, Germany) according to the manufacturers protocol. Five hundred ng of DNA were bisulfite-converted using the EpiTect Bisulfite Kit (Qiagen; Hilden, Germany) according to the manufacturers alternative protocol for dilute solutions. The sequences of interest were amplified using the polymerase PfuTurbo Cx (Agilent Technologies; Santa Clara, CA) and methylation-neutral primers. Amplicons were cloned into the pJET1.2/blunt vector (Fermentas; Vilnius, Lithuania) and transformed into E. coli. For each amplicon at least 8 independent clones were sequenced, aligned and analysed as described in (Foerster and Mittelsten Scheid, 2010).

Data access

The BiSS analysis pipeline is based on Graphic Processing Unit computation on CUDA (Computer Unified Device Architecture) framework, and details of the Arabidopsis methylome generated by BiSS are available at `http://www.cibiv.at/software/ngm/ \gls{BiSS}`.

System Requirements: CPU: SSE enabled dual-core (quad-core recommended), RAM: 4 GB (16 GB recommended), GPU (optional): CUDA (Nvidia) or ATI Stream Technology (ATI) enabled, OS: Linux (OpenSUSE with gcc 4.3.4 recommended), Software: CUDA 3.2 (or higher), AMD Accelerated Parallel Processing SDK 2.5.

Chapter 3.

MethColor - a computational approach for uncovering DNA methylation heterogeneity in deep sequencing data

It's all about heterogeneity.

Bill Gates

Summary

DNA methylation is an important epigenetic biomarker, for instance in cancer medicine, but the degree of modification at a specific genomic cytosine may vary according to individual cells, tissues, or developmental stages. To capture and elucidate the heterogenous nature of DNA methylation is one of the challenges in both, biological and computational aspects. Deep sequencing data after bisulfite conversion of non-methylated cytosines which provide the methylation state at single-bp resolution allow to tackle this problem since individual reads represent individual genomic copies, even if the DNA was prepared from multiple cell types.

We introduce a computational approach to identify the minimal number of distinct methylation profiles in such pooled data sets. We use a graph coloring method, called MethColor, together with empirical heuristics to characterize different, cell- or tissue-type-specific methylation profiles. A simulation study demonstrates the potential of this method.

3.1. Introduction

DNA methylation is an important epigenetic mark that plays a crucial role in the cellular development of many eukaryote organisms (Laird, 2010). It is also known as a crucial biomarker in nearly all types of cancers (Jaenisch and Bird, 2003). DNA methylation profiling as a diagnostic tool has become more and more attractive along with the development of simplified, accelerated and cost-efficient high-throughput data technology. DNA methylation is widely-known as the addition of a methyl group to the position 5 of genomic cytosines. Hence, the modification leaves the DNA sequence unaltered but changes chromatin features and gene expression. DNA methylation is detected by bisulfite conversion (Frommer *et al.*, 1992) which converts unmethylated cytosines (C) to uracil, after PCR to thymines (T), while the methylated cytosines are not affected. With the advent of the next-generation sequencing technology, the so-called bisulfite deep sequencing (BS-Seq) technology has recently been developed to obtain high resolution DNA methylation maps (Lister *et al.*, 2008; Cokus *et al.*, 2008). To obtain the map, the BS-Seq short reads (thereafter referred to as reads) are mapped to the reference genome. Then, the methylation profile for every genomic cytosine is generated: C-T mismatches indicate genomic unmethylated Cs whereas C-C matches indicate methylated genomic Cs in absence of sequencing error. Thus, one can characterize the methylation state for every cytosine at single-bp resolution (so-called whole-genome methylation profiles or methylomes). This is referred to as methylation profiles thereafter.

Current sequencing technologies use DNA samples prepared from a mixture of cells with potentially heterogenous DNA methylation profiles(Laird, 2010; Pelizzola and Ecker, 2011). This leads to average measurements across DNA molecules but inaccurate estimation of DNA methylation levels for single sites as the frequency of cell types in the mixture is typically not determined (Laird, 2010). This has raised a challenge in both computational and biological aspects, although each read in the BS-Seq approach provides a discrete DNA methylation pattern for a single genomic DNA molecule. Thereby, inferring the distinct cell type-specific profiles in DNA mixtures is important for uncovering the heterogeneity of DNA methylation.

Here we suggest a computational approach, called MethColor, to differentiate the methylation profiles across cells in DNA mixtures based on mapped read libraries. First, a mapped read library is transformed into a graph in which each node represents one

read, and an edge is formed if the reads overlap on the reference genome but display a different methylation patterns (i.e. incompatible reads). Then, we propose an optimization problem of finding the minimal number of distinct methylation profiles as a well-known graph coloring problem in computer science (Cormen *et al.*, 1990). The read nodes that have the same color constitute a methylation profile. Different colors imply distinct profiles in the cell population, and the minimal number of profiles provides the most parsimonious explanation of the observed methylation state from the mapped reads. In addition, we use a simple heuristic based on the empirical methylation frequency at a single locus estimated from the mapped reads to obtain the final methylation profiles. To demonstrate the efficiency of our approach, we use simulated data and evaluate both the minimal number of inferred profiles and the similarity between the inferred profiles and the original. The results offer a promising perspective for the proposed approach in understanding DNA methylation heterogeneity.

3.2. Computational problem formulation

3.2.1. Input data and optimization problem

The input data are a reference genome consisting of n genomic cytosines and a library of mapped BS-Seq reads, where we only consider reads that map uniquely to the reference genome without sequencing errors (the sequencing errors can be corrected, Eriksson *et al.*, 2008). Each mapped read r is represented by a so-called location vector $p(r) = (p_1(r), \ldots, p_k(r))$, $p_1(r) < \ldots < p_k(r)$ which indicates the positions of the cytosines in the reference genome where k is the number of genomic cytosines mapped by read r. We use p_i instead of $p_i(r)$ for short. Each mapped read has an associated methylation string of length k, $S(r) = s_1 \ldots s_k$ where

$$s_i = \begin{cases} 1 & \text{the cytosine at the genomic position } p_i \text{ is methylated} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The goal is to find the minimal set of methylation profiles $\mathcal{G} = \{G_1, \ldots, G_\gamma\}$, where $G_i = g_{i1}, \ldots g_{in}$ with $g_{ij} \in \{0, 1\}, j = 1 \ldots n$ indicates the methylation state of the cytosine at position j, such that the methylation string of every read is the substring of at least one profile in $\mathcal{G}$. In other words, one needs to find a minimal number of distinct

DNA methylation profiles of the whole genome, such that the methylation pattern of every read is contained in at least one profile. This problem is classified as a NP-HARD problem (Peng and Smith, 2011, and references therein).

Figure 3.1 (a) shows an input of 9 reads across 5 genomic cytosines. The methylation strings derived from the corresponding reads are displayed in Figure 3.1 (b). For this example, three profiles $G_1 = 11000, G_2 = 00011, G_3 = 10101$ are the unique optimal solution to explain the data. Specifically, the reads $\{R1, R4, R7\}$ are generated by G_1, $\{R2, R5, R8\}$ by $G2$ and the rest by G_3.

3.2.2. Graph coloring problem

We reformulate the above optimization problem as a graph coloring problem. Each read is represented as a node, two nodes are connected by an edge if the mapped regions in the reference genome corresponding to the two reads r and r' overlap. It means $\exists(p_i, \ldots, p_j)$ such that $min(p(r), p(r')) \leq p_i \leq p_j \leq max(p(r), p(r'))$ and if the two methylation patterns in the overlap are different. That is, an edge indicates that the methylation patterns of the two reads are different although both reads map at least partially to the same genomic region. We solve the minimal number of methylation profile problem by assigning colors to the nodes such that any two nodes connected by an edge must have different colors. Figure 3.1 (b) illustrates the graph for the data in Figure 3.1 (a). The nodes are colored by 3 colors corresponding to the 3 distinct profiles G_1, G_2, G_3 mentioned above.

3.3. MethColor method

3.3.1. Estimate the minimal number of methylation profiles

We suggest a greedy algorithm as a quick approximation to find the minimal number of colors to color the nodes of the read graph. The greedy idea is similar to the McColor algorithm presented in Lvque and Maffray (2005). However, we apply it to arbitrary graphs and use a heap data structure for efficiency in choosing the coloring nodes.

Given M nodes, $\mathcal{C} = \{1, \ldots, M\}$ is the set containing the maximal number of colors. For each node x, $nbCol(x)$ denotes the set containing the colors of its colored neighbors

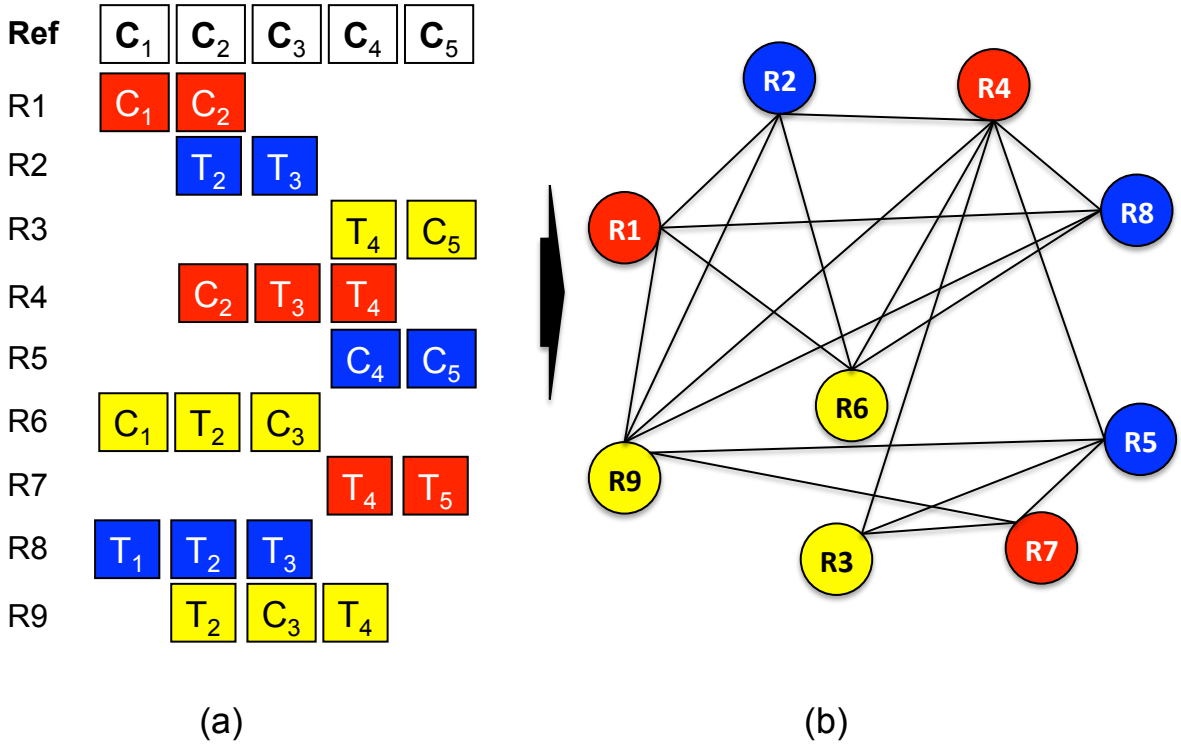


Figure 3.1.: **An illustrative example where the reference genome has 5 cytosines ($n = 5$):** (a) 9 reads derived from 3 cells with different DNA methylation patterns indicated by 3 colors: red(G_1), blue (G_2), and yellow (G_3); the subscripts indicate the mapped positions at cytosines enumerated by the order of their relative positions in the reference genome; (b) the graph with 9 nodes built from the input data, three colors are assigned to the reads according to their cell of origins.

and $color(x)$ denotes the color of a node x, $color(x) = 0$ if x is not yet colored. Our algorithm (see pseudo code: Algorithm 3.1) works as follows: we start with color 1 assigned to a random node. At each step, we select from the set of yet uncolored nodes ($color(x) = 0$) the node x for which $|nbCol(x)|$ is maximal; and the color assigned to x is $color(x) = min\{\mathcal{C} \setminus nbCol(x)\}$. For every uncolored y connected by an edge with x, $nbCol(y)$ is then updated. The process is repeated until all nodes are colored. For efficient implementation, we use the heap data structure (Cormen *et al.*, 1990) as a binary tree where each node x of the heap H has the key value $key(x) = |nbCol(x)|$. Thus, the root of the heap is always the uncolored node x with the largest $|nbCol(x)|$.

Algorithm 3.1: Pseudo code of graph coloring algorithm

Data: A mapped read library.

Output: An assignment of colors for every node.

begin

Building graph $G = (V, E)$ corresponding to the mapped read library;

Initialization: $\forall x : nbCol(x) = \emptyset$; $color(x) = 0$; $push(x, H)$;

foreach $i = 1..|V|$ **do**

$x = pop(H)$;

$color(x) = min\{\mathcal{C} \setminus nbCol(x)\}$;

forall $y : (x, y) \in E$ *and* $color(y) = 0$ **do**

$nbCol(y) = nbCol(y) \cup color(x)$;

update H: $key(y) = |nbCol(y)|$;

end

The graph coloring algorithm has the worst complexity of $O(|E| \log |V|)$ where $|V|$ is the number of nodes, i.e the number of reads and $|E|$ is the number of edges. Generating the graph has $O(|E| * k_{max})$ complexity where k_{max} is length of the longest location vector. Fig. B.1 show a more detailed illustration via an example.

3.3.2. Heuristics for determining the DNA methylation profiles

After coloring the read graph, the reads that belong to one color are used to construct the intermediate profile as the consensus string of those reads' methylation strings. $\hat{\mathcal{G}} = \{\hat{G}_1, \ldots, \hat{G}_\gamma\}$ is the set of inferred methylation profiles, in which $\hat{G}_i = \hat{g}_{i1} \ldots \hat{g}_{iN}, \hat{g}_{ij} \in \{0, 1, 2\}$, where 0 indicates unmethylated, 1 methylated and 2, unresolved, i.e there is no read assigned at genomic cytosine j and profile i. We note that due to the greedy strategy, some methylation profiles are only comprising very few reads, i.e the profile consists of many 2 (unresolved states).

To determine unresolved methylation states, we use a heuristic based on the empirical methylation level $f(C_i)$ at cytosine position i that can be computed from mapped reads as the ratio of reads with cytosine at that position over the total number of mapped reads at position i. From the intermediate profiles we compute $\hat{f}(C_i)$ as the preliminary methylation level computed from the profiles ignoring unresolved states (Figure 3.2

shows an example). To minimize the difference between $f(C_i)$ and $\hat{f}(C_i)$ we apply the following heuristic. A read r is called movable from one profile $\hat{G}$ to another $\hat{G}'$ if its methylation pattern is compatible with the methylation profile of $\hat{G}'$. If we move a read from $\hat{G}$ to $\hat{G}'$, we will recalculate the $\hat{f}(C_i)$ for all cytosines affected by this move. Then, the goal is now to replace most of the unresolved states in the intermediate profiles and at the same time to minimize the difference i for which $|f(C_i) - \hat{f}(C_i)|$. To do this we perform a greedy strategy by starting with genomic position i if the $|f(C_i) - \hat{f}(C_i)|$ are maximal. This process is repeated until we cannot find any suitable movement. Fig (2c) displays the final result (see Fig. B.2 for a more detailed illustration).

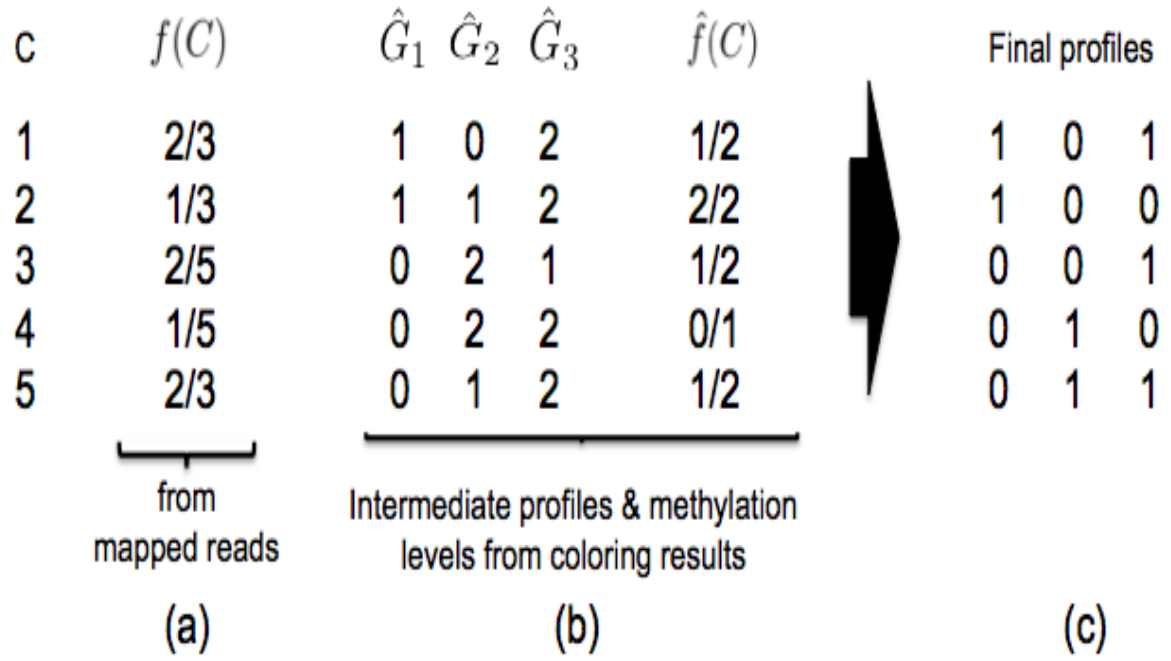


Figure 3.2.: **Illustration of building the final methylation profiles:**
(a) the 5 cytosines in the example from Figure 3.1 and their empirical methylation level $f(C)$;
(b) the columns represent 3 intermediate patterns with 1: methylated, 0: unmethylated, 2: unresolved, and their methylation level $\hat{f}(C)$;
(c) The heuristics move reads from the $\hat{G}_1$ to $\hat{G}_2$ or $\hat{G}_3$ to resolve unresolved positions such that $\hat{f}(C_i)$ is as close to $f(C_i)$ as possilbe. This leads to the final profile.

3.4. Simulation study

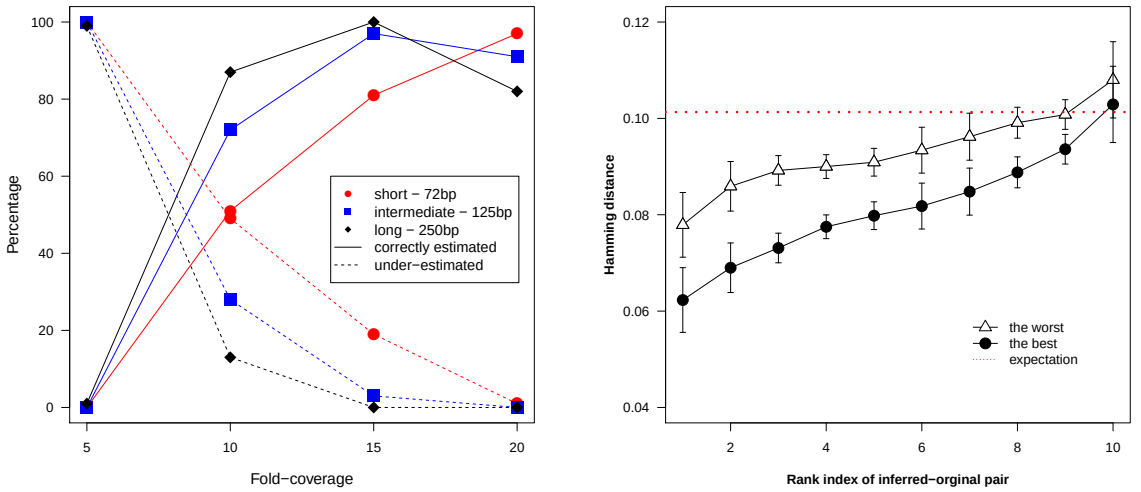
3.4.1. Datasets

We simulated the BS-Seq short reads from 10 kbp reference assuming no sequencing error, a 6% methylation rate (motivated from an empirical study by Lister *et al.* (2008)) and 10 distinct DNA methylation profiles. Mapped short (72-bp), intermediate (125-bp), and long (250-bp) reads were randomly generated with sequencing coverage of $5, 10, 15$, and 20 after mapping. A coverage of 20 is high compared to the available BS-Seq datasets (Lister *et al.*, 2008, 2011). Here we simulated equal number of reads for each profile. Hundred datasets were generated for each combination of read length and sequencing fold-coverage. Thus, we had 1200 datasets in total.

For evaluation, we first computed the empirical distribution of the number of inferred methylation profiles for all simulations, i.e how often MethColor estimated exactly the original number of methylation profiles or under-/overestimated it. Then, we used Hamming distance to evaluate the similarity between inferred and originally simulated methylation profiles. As the assignment of the inferred profile and the original profile is not predictable, we applied a greedy strategy. First, the pair of predicted and simulated profiles with smallest Hamming distance was chosen, deleted from the set of unassigned ones, and then the process was repeated until no pair can be selected. Then we tested if the Hamming distance for each selected pair is significantly smaller in comparison with the distribution of Hamming distances between random patterns at the same methylation rate. The cutoff 5% is used as the lower-bound distance.

3.4.2. Simulation results

Figure 3.3(a) shows that the ability to predict the correct number of profiles depends on the read length and the coverage. In case of low coverage (of coverage 5), all the test underestimate the number of distinct profiles regardless the read length. Increasing the number of mapped reads leads to more cases in which the algorithm estimates exactly the number of profiles. In addition, longer reads also improve the coloring results for the case of high coverage (sequencing fold of 10 and 15). However, when the coverage exceeds 15, the precision of the algorithm decreases, it tends to overestimate the number of profiles especially for long read lengths. We speculate that the graph gets too complex due to



(a) Number of inferred methylation profiles (b) Similarity between inferred and original profiles

Figure 3.3.: **Performance of MethColor with diverse sequencing coverages and read lengths:** (a) solid/dashed lines indicate the number of exactly/under-estimated number of profiles, (b) the average Hamming distance of the inferred patterns and original ones for the worst and the best cases that corresponds to the short and long read length at 20-fold coverage. The horizontal dashed line indicates the 5% lower bound of Hamming distance between two random profiles.

the number of read nodes and graph edges are much risen. This simulation also provides a hint of how much sequencing is needed to have an exact estimation of methylation profiles.

Finally, not only the precise number of different profiles is estimated but also the accuracy of individual methylation state of each profile. Figure 3.3(b) presents the average Hamming distance of 10 inferred-original pairs indexing from the closest pair to the most distant one. Here we only show two examples from simulated datasets (the short and the long read case with the sequencing coverage of 20), that provide the best fit to the simulated profiles (small Hamming distance, lower curve) and the worst fit (large Hamming distance, upper curve). The results demonstrate that MethColor can estimate well the methylation patterns of up to 8 out of 10 DNA methylation profiles as the Hamming distance is ranging from 0.06 to 0.1. The average Hamming distance

is then significantly smaller than the analytical lower bound from the distribution of distance between random patterns given the same methylation rate (dotted horizontal line in figure 3.3(b)), assuming a p-value 0.05.

We also performed the experiments with non-uniform distribution of methylation profile frequency, and the results are similar in terms of inferring the number of methylation profiles.

Taking all together, the simulation results prove the potential MethColor to be applied for experimental biological data.

3.5. Discussion

We formulated a computational problem aiming to understand the heterogeneity of DNA methylation based on the mapping profiles from deep sequencing data after bisulfite conversion. Our approach, MethColor, efficiently estimates the number of distinct methylation profiles as well as their specific profiles. Despite a rough estimation, this can provide valuable information for research and diagnosis and help to understand the diversity of DNA methylation patterns.

In addition, the DNA methylation profile frequency can be estimated based on the inferred patterns by the Expectation Maximization approach used in haplotype construction problem (Eriksson *et al.*, 2008). Our approach can also be generalized to the haplotype reconstruction, for instance, in the viral population according to (Eriksson *et al.*, 2008). Our approach can also work in a context of pooled sequencing of different individuals or ecotypes (Docherty *et al.*, 2010), in presence of one available reference genome, e.g. help to assemble shortreads mapped to common reference genome (Peng and Smith, 2011).

Future works will address: (i) incorporating sequencing errors, (ii) applying the MethColor to analyze available biological datasets, for example the recent diverse set of methylome data from human embryonic stem cells (Lister *et al.*, 2011). In dealing with sequencing error, we can either do error correction before or incorporate the non C-T mismatches as the weights of the edges in the read graph.

Chapter 4.

Epi-Speller: a bioinformatic tool to discover epigenetic signatures

Deciphering the epigenetic code will illuminate some of the most profound questions in biology

Stephan Beck

Summary

The concept of a chromatin-based epigenetic code, proposed more than a decade ago, associates specific combinations of chromatin marks with different gene expression states and their maintenance. High-throughput technologies like microarray profiling or next generation sequencing enable us to examine the validity of the concept, by profiling transcriptomes and multiple chromatin marks for many different samples, conditions and organisms. The large amounts of generated data require efficient and instructive computational methods to identify and interpret biologically relevant correlations and to challenge the hypothesis of an epigenetic code.

Here, we introduce a generally applicable bioinformatic method to group epigenetic information across genome-wide chromatin data sets. It automatically classifies the abundance of chromatin-based signals into discrete categories and transforms the categories into so-called epi-letters. Each genomic region can then be represented as a combined string of epi-letters referring to different chromatin marks. This synoptic compilation can be used for further clustering to determine common epigenetic signatures and can

be represented applying the concept of the DNA motif sequence logo. We present the results of applying the epi-letter principle using published data from 12 chromatin marks in the model organism *Arabidopsis thaliana*.

We propose a new and simple tool for finding and representing epigenetic patterns across genome-wide profiling data of different chromatin marks. We provide a proof-of-concept application with published data, resulting in a classification of epigenetic signatures in Arabidopsis thaliana. The method has also other potentials for de novo discovery and visualization of general genome-wide profiling patterns.

4.1. Background

The concept of a chromatin-based epigenetic code was proposed more than ten years ago (Strahl and Allis, 2000; Jenuwein and Allis, 2001). It describes the principle that a certain combination of chromatin marks and their distribution across the genome would influence the accessibility of DNA for the transcription machinery and thereby determine gene activity. In the last decade, various epigenetic marks, including DNA methylation, multiple histone modifications and nucleosome occupancy, were quantified in many different eukaryotic organisms, and genome-wide profiles are available after hybridization to chip arrays or deep sequencing. This has produced a massive amount of data that can now be mined to associate the different combinations of chromatin patterns with biological features (van Steensel, 2011). This task can only be accomplished with the power of computational methods. Chromatin signature analysis (for review see van Steensel (2011)) has applied different methods including clustering, considering all marks (Roudier *et al.*, 2011) or a data set reduced by Principal Component Analysis (PCA) (Filion *et al.*, 2010), genome-scanning methods (e.g. with a Hidden Markov Model (HMM) (Ernst and Kellis, 2010)), or genome segmentation method with a Bayesian network (Hoffman *et al.*, 2012). All approaches share the principle to compare regions defined by common chromatin features with gene expression signatures across the genome. Extensive analyses were performed for three organisms: Drosophila (Filion *et al.*, 2010), Arabidopsis (Roudier *et al.*, 2011) and Homo sapiens (Ernst and Kellis, 2010), applying chromatin immunoprecipitation with modification-specific antibodies (ChIP) or DNA adenine methyltransferase identification (DamID), combined with either tiling arrays or deep sequencing technology. The common finding was that

the combinatorial complexity of multiple chromatin marks is surprisingly low (Rando, 2012), with a few chromatin types characterized by presence (or absence) of a specific combination of chromatin marks. An example is a compilation of 53 different chromatin features in Drosophila that resulted in only five different types (Filion *et al.*, 2010). A data set for twelve chromatin marks in Arabidopsis identified only four different combinations (Roudier *et al.*, 2011). The recently completed work with human CD4 T-cells and 38 distinct marks (Ernst and Kellis, 2010) discovered 51 chromatin states summarized in 5 big groups. However, the low complexity might be misleading. ChIP and DamID require a substantial amount of biological material where epigenetic heterogeneity might go undetected. Further, the number of different chromatin modifications discovered to be relevant for DNA accessibility and transcriptional regulation is still increasing, making a synoptic view more demanding while more difficult. Hence, for already existing as well as for future data, there is a growing need of more sophisticated analysis methods, including bioinformatics tools, to discover chromatin signatures in complex data sets and their functional relevance.

Considering the different ranges of chromatin parameters, one important task in signature discovery is providing an unbiased comparative framework to determine the level of occupancy at all genomic regions. Chromatin marks and DNA methylation are usually profiled and summarized as enrichment levels of each mark at a given genomic region. For a synoptic comparison, it is essential to normalize absolute values and to consider the genomic resolution that is determined by the experimental procedure. Ideally, a common scale with similar resolution along the reference genome would be optimal to visualize multiple chromatin features within the same genomic regions, e.g. in a genome browser.

To uncover biologically relevant correlations and to shape hypotheses about their causal relationship, the various chromatin features should be seen in connection with gene activity, gene structure, and the nature of the genetic information. Gene expression, clearly correlated with specific chromatin states (reviewed in van Steensel (2011)), is quantified in all organisms similarly as amount of RNA homologous to the reference genome, and results can be displayed as for the different chromatin marks. However, a connection of chromatin states with gene structures and sequence categories cannot be documented in a linear fashion. A rough breakdown separates sequence categories into three basic types, i.e. genic and intergenic regions and repetitive DNA like transposable

elements or satellites (Roudier *et al.*, 2011). A finer resolution of genic regions considers gene components like enhancers, promoters, introns and exons, and untranslated regions (Ernst and Kellis, 2010). Further, known or assumed functions of genes, summarized in GO terms (Ashburner *et al.*, 2000) are frequently tested for a connection with specific chromatin states. For all studies asking for a significant correlation of any parameter with epigenetic marks, a quantitative yet simple synoptic summary of chromatin organization at a defined site of the genome would be supportive.

Another challenge for comparative studies comes from the different dynamic ranges of several chromatin parameters and non-Gaussian distributions. While diagrams indicating the measured values as continuous parameter allow informative graphic documentation, a comparison of more than three parameters in such presentation is difficult. This problem of complex vertical comparison is known from multiple DNA or protein sequence alignments, where it has found an elegant solution by motif discovery algorithms (Bailey, 2008). Similar principles have also been applied to real-valued histone motif scanning (Hon *et al.*, 2008, 2009). The unsupervised method, ChromaSig, first exploits the progressive alignment method, and then discovers the histone motifs from the pre-computed seeds. However, this method is limited to handling only a small amount of chromatin signatures across the whole genome. More recent work (Ernst and Kellis, 2010) applies alternatively the multivariate HMM for scanning common states across the human genome in CD4+ T-cells, based on binary representation (presence/absence). Further, clustering methods have been applied to find chromatin patterns, like PCA (Filion *et al.*, 2010) and k-mean-based methods (Roudier *et al.*, 2011). However, the conventional heatmap representation is complex, and a more condensed display could be helpful.

In this chapter, we introduce a straightforward and generally applicable method to analyze large sets of epigenetic data along a reference genome. Epi-Speller transforms chromatin-based information into discrete categories, called epi-letters. Each genomic region is then represented as a combined string of epi-letters, referring to different chromatin marks. This synoptic compilation resembles the DNA or single amino acid letter code and can be used accordingly for further clustering, alignments, or *de novo* discovery of common chromatin signatures. The signatures consisting of similar patterns can be represented applying the sequence logo concept (Schneider and Stephens, 1990; Crooks *et al.*, 2004). We show the applicability of Epi-Speller with data from twelve chromatin

marks (including DNA methylation and histone modifications) in the model organism *Arabidopsis thaliana* (Roudier *et al.*, 2011) and compare them with previously published results. While there is good general agreement, we discuss some interesting differences obtained with Epi-Speller.

4.2. Results

4.2.1. Epi-Speller - a bioinformatic tool for grouping and summarizing chromatin data

We introduce a dynamic programming approach to group the intensities of epigenetic signatures. The algorithm, called Epi-Speller, minimizes the heterogeneity of value distribution within each group. Three letters represent different signature intensities between the groups. This so-called epi-letter is then mapped to the genomic region, where the signature was found (Figure 4.1). Contrary to conventional approaches, Epi-Speller categorizes values automatically independent of the intensity distribution of the respective epigenetic mark. The dynamic programming approach is computationally efficient and can be applied to large genomic data sets. The approach allows the bioinformatic analysis of many different epigenetic marks across the genome. For a proof of concept, we analyzed 17.365 genomic regions as defined by the Arabidopsis tiling array and the intensity values for twelve chromatin marks as described by Roudier *et al.* (2011). Details are presented in Figure C.1 and the Material and Methods (section 4.4).

Figure 4.2 displays the empirical signal strength distributions for the chromatin marks. Most of the distributions are uni-modal and some of them resemble a normal distribution (e.g. H3K27me2, H3K56Ac, H4K20me1). However, we also note that H3K27me1, H3K4me2, and H3K9me2 are clearly bi-modal. To capture the different types of distributions, we introduced a three letter alphabet of epi-letters, representing signal intensities H(igh), M(edium), L(ow). The results of the epigenetic signal group algorithms are displayed in Figure 4.1.

A categorization in only two letters (high or low) would split the normally-distributed intensities along the mode or mean of the distribution, a classification that is not plausible (figure C.2(a)). For normally distributed signal intensities the split in three groups

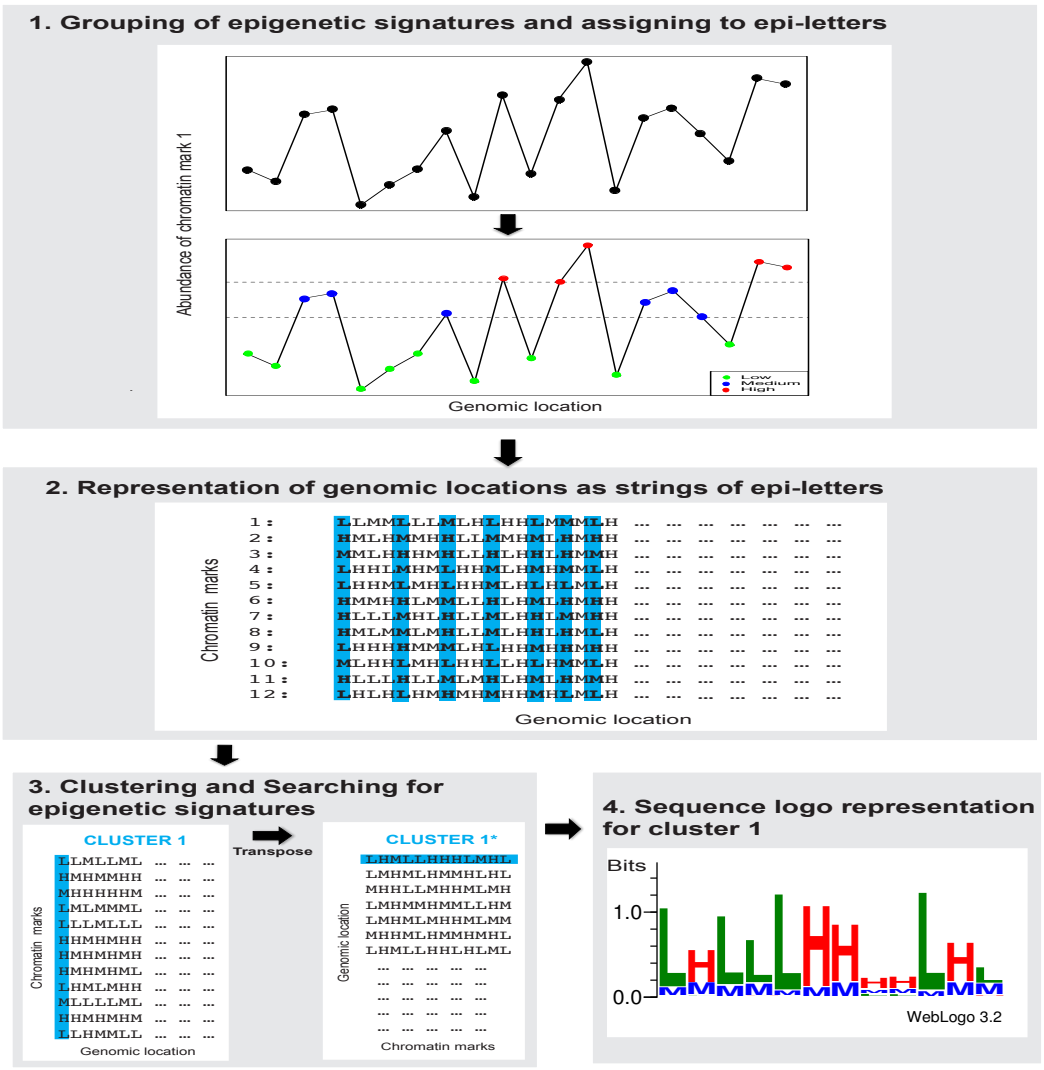


Figure 4.1.: **Workflow of the Epi-Speller method.** (1) The abundance of the chromatin marks along the genome is converted into three categories (epi-letters Low, Medium, High) according to their individual distribution (see Figure 4.2). (2) The abundance of all analyzed marks at each genomic location is represented as strings of epi-letters. (3) Clusters representing different chromatin marks at specific genomic locations by strings of epi-letters are transposed into clusters of genomic locations that share similar epigenetic signatures. (4) Letter-based information is condensed and represented by sequence logo.

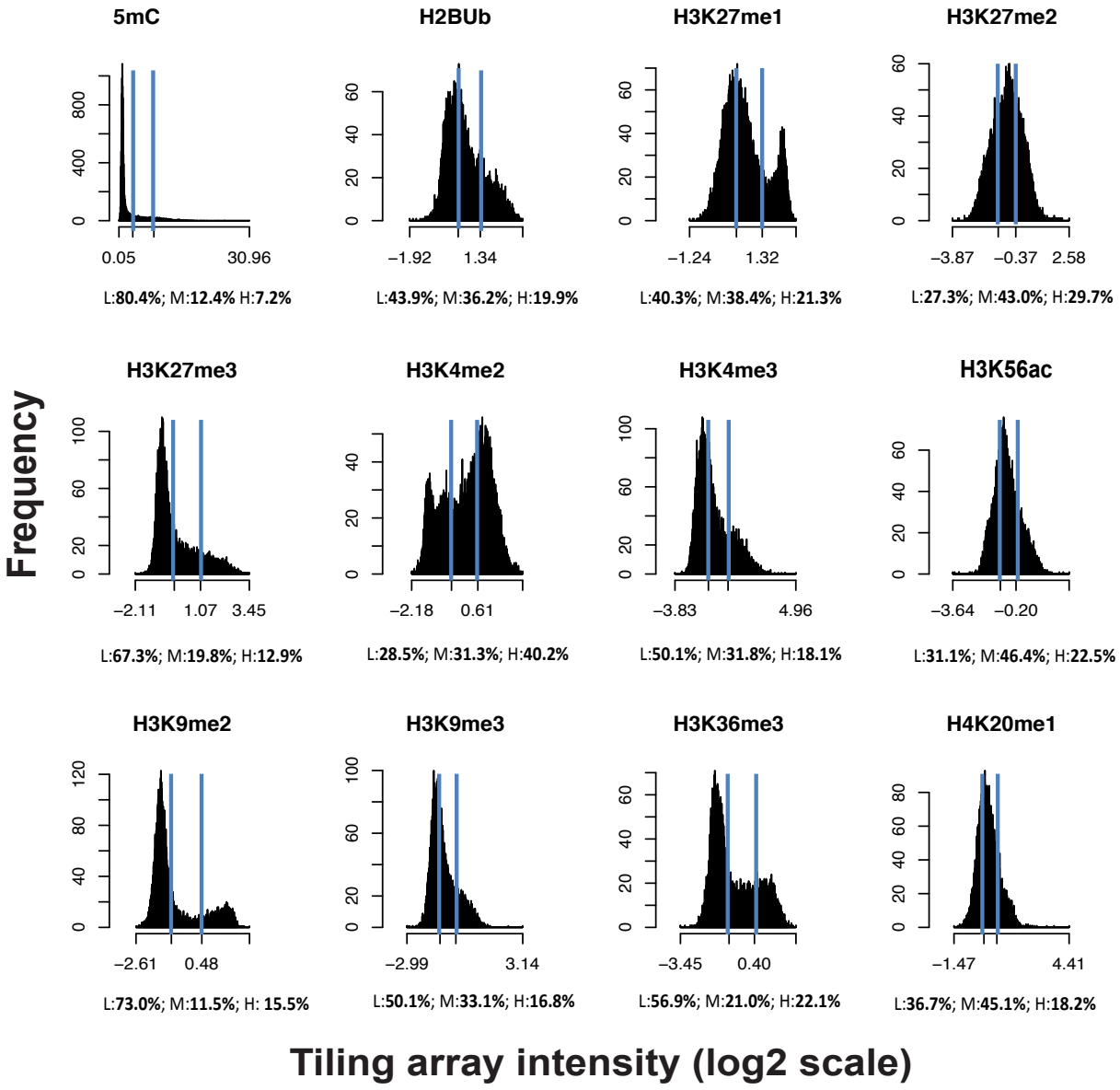


Figure 4.2.: **Frequency distribution of different chromatin marks.** The x-axis displays the range of abundance, here the log2 value of hybridization intensity from the tiling arrays. The y-axis shows the frequency of tiles with a given signal intensity. The vertical blue lines indicate the borders of the categories H(igh), M(edium), and L(ow) using the epigenetic signal group algorithms.

is sensible, because the upper- and lower tails of the distribution constitute their own groups and the bulk of the distribution belongs to the M group (figure C.2(b)). The three

letters categorization also nicely separates the bimodal distributions e.g. H3K27me1 or H3K9me2. Note that a refined categorization in 4 or 5 groups does not help for this data set (as we showed with simulated standard normal distribution, figure C.2(c) and C.2(d) respectively); except for H3K4me2 all other distributions are at best bimodal. Thus, for this data the three-way classification appears optimal.

We assigned one of the three epi-letters H, M, and L to each chromatin mark and each tile. We noted that the frequency of each letter varies across different chromatin marks. High frequencies of L were found for 5mC, H3K27me3, H3K36me3 and H3K9me2, in accordance with the relatively low amount of heterochromatin in the small Arabidopsis genome. M appears mainly in the marks with normal distribution like H3K27me2, H3K56ac, and H4K20me1, whereas H is rare in almost all marks, with the exception of H3K4me2 (Figure 4.2). The assignment of the letters, based on the individual distribution of the marks, converted the continuous values and the different dynamic ranges for each chromatin mark to discrete letters along the Arabidopsis chromosome. Finally, the Epi-state, a string of maximal 12 letters representing each mark, characterizes each tile.

To find out if some genomic regions are similar to each other with respect to the chromatin signatures, it is now possible to use standard clustering routines, e.g. k-means (Roudier *et al.*, 2011) or principal component analysis (Filion *et al.*, 2010). The resulting clusters of genomic regions can then be further characterized. Here, we employ the sequence logo concept (Schneider and Stephens, 1990; Crooks *et al.*, 2004) to the letter representation of tiles occurring in the same cluster. The sequence logo provides a variety of information (Schneider and Stephens, 1990): (a) the amount of information present at each position (measured in bits), (b) the order of predominance (the most frequent on top) of the letters for every chromatin mark, (c) a consensus word constructed from the letters on the top of the graph, and (d) the relative frequency of the letters at every mark.

4.2.2. Epi-Speller representation of Arabidopsis chromatin states

To provide a proof-of-concept for the application of Epi-Speller to experimental chromatin data, we computed the logos of the four chromatin-states (clusters) CS1, CS2, CS3, and CS4, as described by Roudier *et al.* (2011). The height of the stack for each chromatin mark (Figure 4.3) is proportional to the information of the mark in the clus-

ters. A high stack is indicative of a large amount of total information, indicating that large letters are very reliable and therefore characteristic for the corresponding cluster. Thus, the chromatin states are in general characterized by the lack of chromatin marks, as reflected by the prevalence of letter L. In fact, an L for 5mC, H3K27me3, and H3K9me2 is characteristic for CS1; the scarcity of 5mC, H3K36me3, H3K9me2, and H3K9me3 defines CS2, while low amounts of H2Bub, H3K36me3, H3K4me2, H3K4me3, and H3K9me3 are indicative for CS3. Finally, CS4 is characterized by only little 5mC, H3K27me3, H3K36me3, and H3K9me2. Chromatin states are further distinguishable by characteristic H letters: H3K36me3 and H3K4me2 for CS1, H3K27me2 and H3K27me3 for CS2, H3K27me1, H3K9me2, and H4K20me1 in CS3. CS4 is not enriched for any of the marks.

Despite the low information content, the assignment of the letters H or M correlates with the characterization of marks in the cluster according to Roudier *et al.* (2011): the color code for the boxes below each panel in Figure 4.3 displays the percentage of tiles within a cluster that is enriched with the corresponding mark (dark purple = 100% of the tiles marked), whereas the numbers in the boxes represent the percentage of marked tiles that occur in the cluster. For example, only 56.0% of the tiles with the H3K4me2 mark occur in CS1, but each of these tiles is marked. Thus the predominance of letter H correlates well with a large fraction of marked tiles in that cluster, whereas letter M is more often associated with a light purple color, a smaller fraction of marked tiles in a cluster. Therefore, the sequence logo is in good agreement with the information from the previous analysis and provides additionally the information content of individual marks in the cluster.

This becomes even clearer if we compute the consensus sequence. For instance in CS1, where the tiles are mainly associated with genes (Roudier *et al.*, 2011), the letter H in the consensus string LHLLLHHHHLHL neatly coincides with the proportion of marked chromatin markers, i.e. the marks H2Bub, H3K4me2, H3K4me3, H3K56ac, H3K9me3, and H3K36me3, consistent with high amounts of associated tiles as reflected by the purple color. Similarly, L matches to tiles with little marks. The same holds true for the remaining clusters. In some instances the order of M, H or M, L is swapped although the tiles were classified as associated with the mark. This is certainly due to the more refined partitioning of the signature intensities (Figure 4.1). Whenever a mark was quantified by dark purple as having the majority of tiles in the cluster, the automatic cutoff-defined

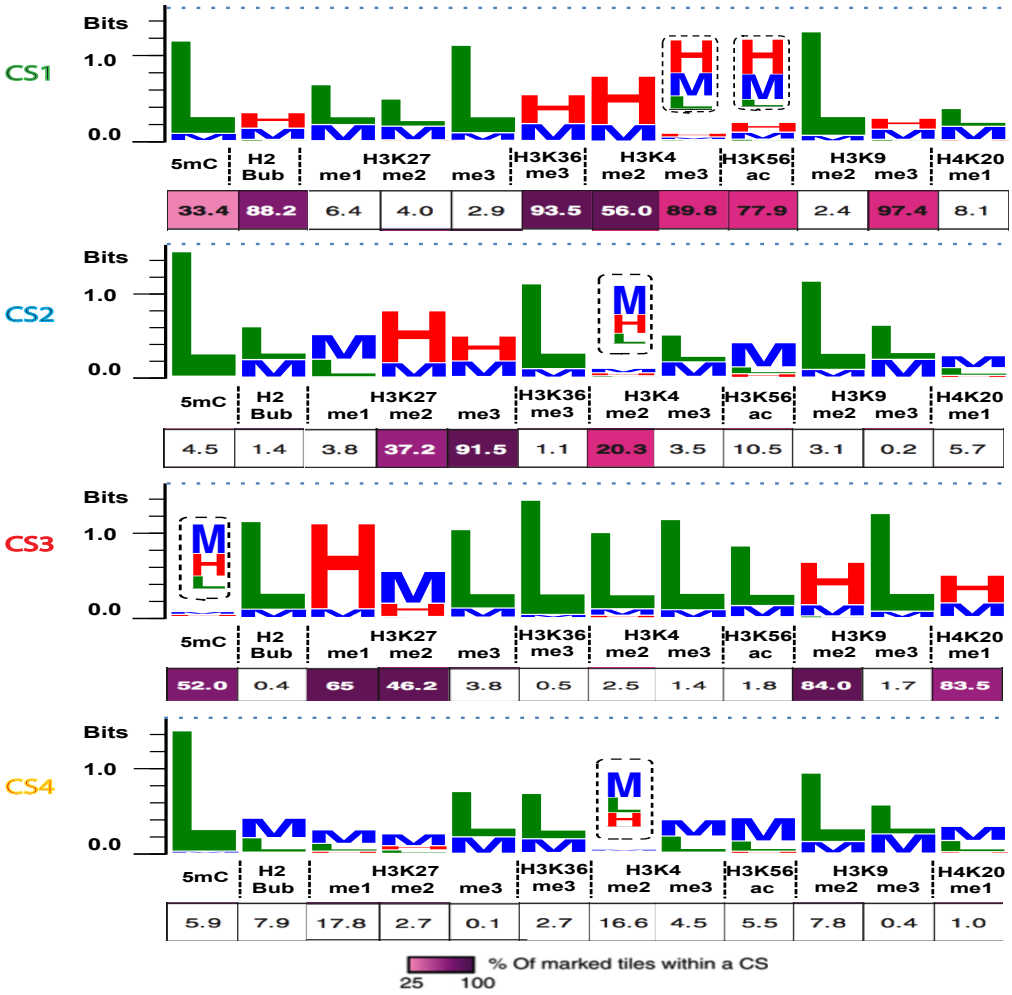


Figure 4.3.: **The epi-letter logos of four main chromatin states.** Epi-speller was applied to generate sequence logos for clusters CS1-4 in (Roudier *et al.*, 2011). Each stack represents the distribution and information content (in bit scores, blue dash line indicates the maximum bit score) of chromatin signatures with Low, Medium or High intensities within a cluster of regions with similar signatures. The bar below is adapted from (Roudier *et al.*, 2011): colors indicate the distribution of chromatin marks from 25% (light-) to 100% (dark purple); numbers inside cells indicate the percentage of tiles that are associated with each chromatin mark. The logos are generated using Weblogo 3.2 program (Crooks *et al.*, 2004). The epi-letters for chromatin marks for which information contents are low are enlarged in inset.

approach assigned an H.

To summarize, Epi-Speller provides a comprehensive representation of the chromatin states based on clustering the signature intensities from tiling arrays. The logo is easy to interpret and provides additional information compared to other summary statistics.

4.2.3. Letter-based clustering for finding chromatin states

Epi-Speller provides an innovative approach to integrate and represent multiple chromatin-profiling datasets. We now show that it can be used to infer clusters of different genomic regions based on epi-letter representation. To this end, we performed a standard k-mean clustering of the tiles simply by computing the Hamming distance for pairs of epi-letter tiles (see Method section).

We determined four clusters (called ES1 to ES4) as optimal number of chromatin states. Cluster ES2 comprises 5.817 tiles from a total of 17.329 tiles whereas three other clusters represent around 3.800 tiles each (Table 4.1).

Table 4.1.: **Congruency between two clustering results**. The entries in the table are the number of tiles that are categorized in CS1-4 as previously published and ES1-4 according to Epi-speller. Ambiguous tiles indicate those that are not assigned to any cluster in (Roudier *et al.*, 2011)

	CS1	CS2	CS3	CS4	Ambiguous	Total
ES1	**3805**	10	6	28	14	3863
ES2	53	**2968**	136	**2251**	409	5817
ES3	2	235	**3242**	278	164	3921
ES4	**2144**	297	59	**1000**	228	3728
Total	6004	3510	3443	3557	815	17329

Table 4.1 also displays the agreement between the epi-letter-based and the previously defined clusters (Roudier, Ahmed et al. 2011). Clearly, cluster ES3 and CS3 comprise the same tiles, with the exception of an insignificant number belonging to the other clusters. Clusters ES3 or CS3 comprises 19% of all tiles, and the resulting sequence logos are almost identical (Figure 4.4). Moreover, the sequence logo for ES3 or CS3

is quite distinct from all other logos, because Ls dominate for H3K27me3, H3K36me3, H3K4me2, H3K4me3, and H3K56ac. The analysis therefore confirms and corroborates the existence of a unique chromatin state CS3/ES3.

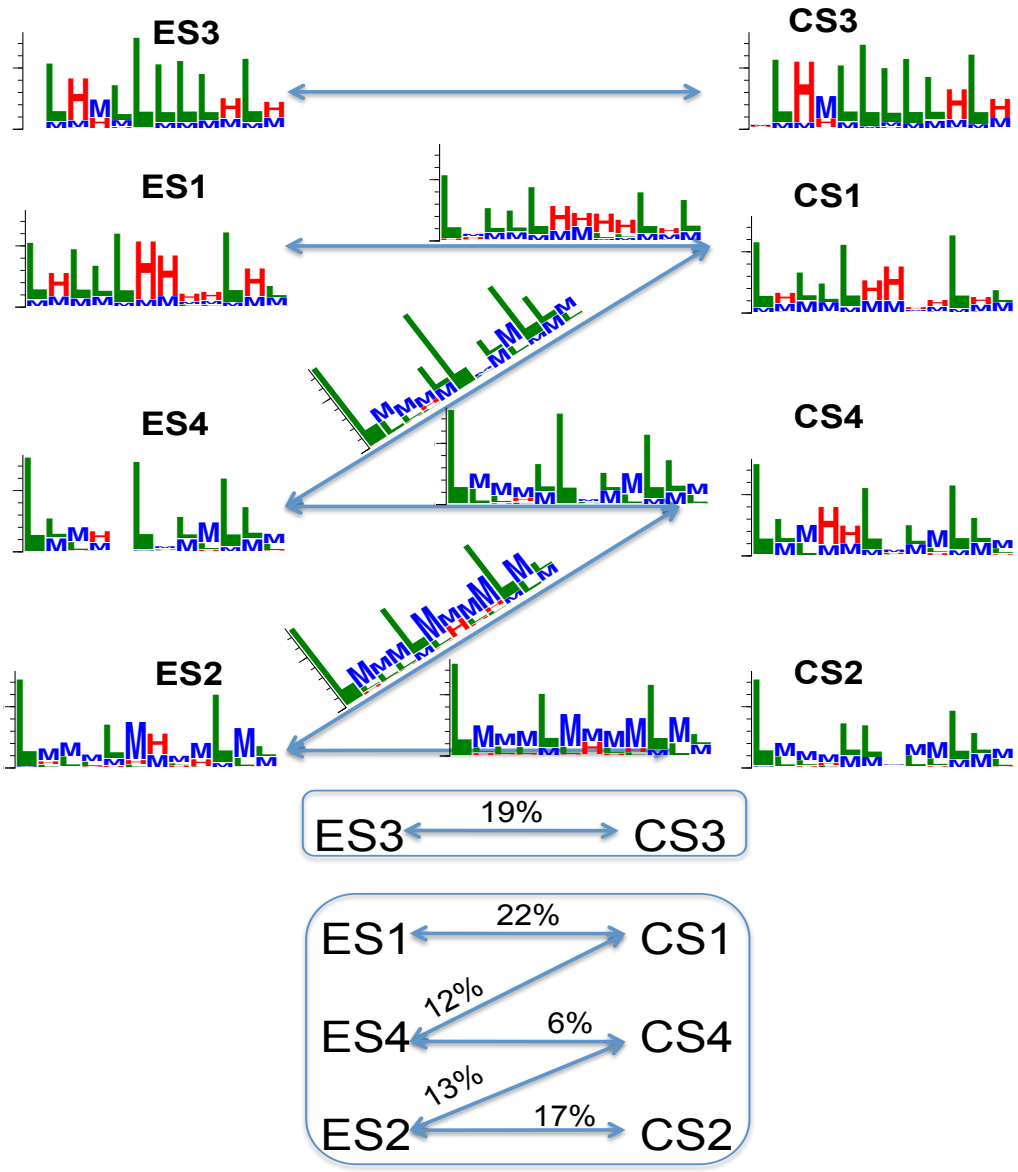


Figure 4.4.: **Comparison of two clustering methods with logo representation.** Sequence logos were built for two clustering results, according to the Epi-speller clustering (ES, left) or the previous clusters (CS, right), as well as the significant overlap between them (arrows with logos in the upper part, from Table 4.1, or quantified as percentage, lower part).

To simplify the following discussion, we will ignore the entries in Table 4.1 with less than 866 tiles (5% of all tiles). This lets the following picture emerge for the other clusters: ES1 is a subset of CS1 and comprises 22% of all tiles. CS1 contains ES1 and parts from ES4, CS4 is composed of tiles from ES4 and ES2, ES2 contains tiles from CS4 and CS2. CS2 is a subset of ES2 and comprises 17% of all tiles. The clusters ES2/ES4 and CS1/CS4 are therefore not totally congruent. Figure 4 summarizes the relationships between the two classification schemes together with the sequence logos. The logos for CS1 and ES1 are very similar, characterized by H signatures at markers H3K36me3, H3K4me2/3, and H3K56ac. However, the Hs are less pronounced in CS1 because CS1 contains some tiles from ES4 where tiles do not show high levels of the mark at those sites. The logo of CS2 as a subset of ES2 is identical to the ES2 logo (Figure 4.4). Four subset clusters that belong to different CS and ES main clusters are characterized by the prevalence of M signatures. While the clustering results are partially different, the sequence logos allow a quick comparison of the main characteristics. The method allows also monitoring the changes in signature strengths and information amount when moving from the top left cluster ES1 to the bottom right cluster CS2. According to the epi-letter-based clustering, CS1 is a mixture of quite distinct tile elements of ES1 and ES4, and the same holds for CS4 that comprises tiles from ES2 and ES4.

4.2.4. Biological annotation analysis of chromatin states

The concept of a chromatin-based epigenetic code proposed that specific combinations of chromatin states associate with different gene expression states, or with other biological annotations. In that context, we re-analyzed the biological annotations for chromatin states inferred by the clustering methods based on continuous values (Roudier *et al.*, 2011), called CS-state, or on letter representation generated by Epi-Speller, called ES-state.

First, we examined which functional groups of the genome would be found in the respective clusters (Figure 4.5). ES1 and CS1 compose of about 94% of tiles that represent gene-rich regions. Similarly, the chromatin states ES3 and CS3 are largely located in regions with many transposable elements (TEs) (79% of tiles in ES3, and 82% of tiles in CS3). ES2 and CS2 show a similar composition of tiles containing genes, TEs, and intergenic regions. The main difference between the two methods is found for the last

cluster where 86% of ES4 tiles are gene-rich, in contrast to only 51% in CS4. Hence, the association of three major genomic features with the four clusters shows both similarity and differences between two methods.

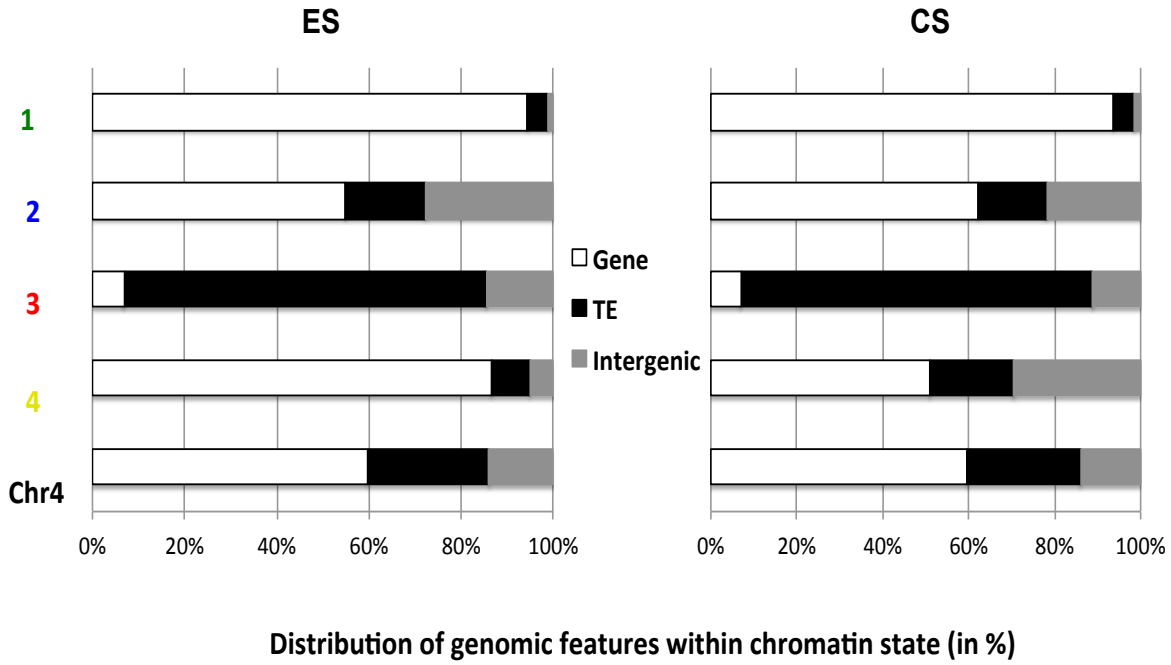


Figure 4.5.: **Annotation of genomic features with chromatin states.** The annotation as genes or transposon elements is taken from the TAIR8 version of the Arabidopsis genome. Tiles that overlap with an annotated feature at least 50 bp are called associated with the corresponding group. The remaining tiles are considered as intergenic regions. The axis shows the relative proportion (in percentage) for each cluster, according to the Epi-speller clustering (ES, left) or the previous clusters (CS, right).

To investigate the relevance of the differences among the E and C clusters for gene expression, transcript profiling data (Schmid *et al.*, 2005) from the same Arabidopsis tissue as used to generate the chromatin profiles were analyzed for a correlation with different clusters in both methods (Figure 4.6). We applied the same method as described before (Roudier *et al.*, 2011) to assign each tile with a corresponding expression level of its genes. Tiles below 8 (log2 scale) were called lowly expressed (Figure 4.6), in accordance with Roudier *et al.* (2011). There is good agreement between both clustering

methods. CS1 and ES1 consist of tiles with highly expressed genes whereas CS3 and ES3 are enriched for lowly or non-expressed sequences. The relative distribution of expression levels within these clusters is well separated from the average over the whole chromosome (Figure 4.6, upper panel). The distribution for the other clusters is closer to that along the non-clustered whole chromosome. The only, but interesting difference was observed for ES4, containing more expressed sequences than the chromosome on average, in contrast to nearly no difference to the average expression distribution in CS4 (Figure 4.6, lower panel). This could indicate a more refined clustering by ES in distinguishing gene expression states associated with chromatin signatures.

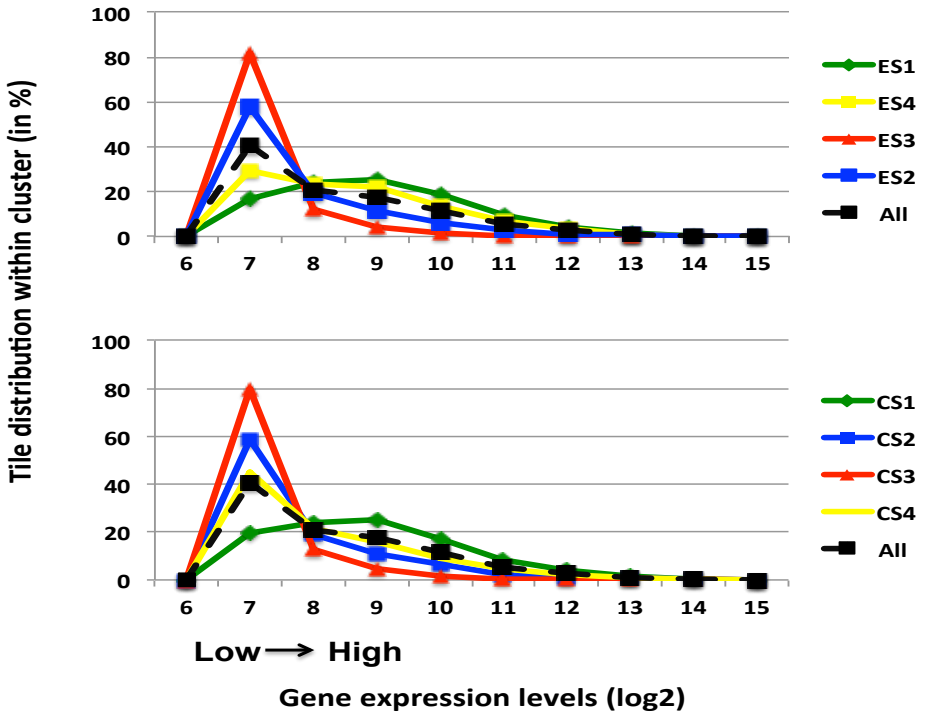
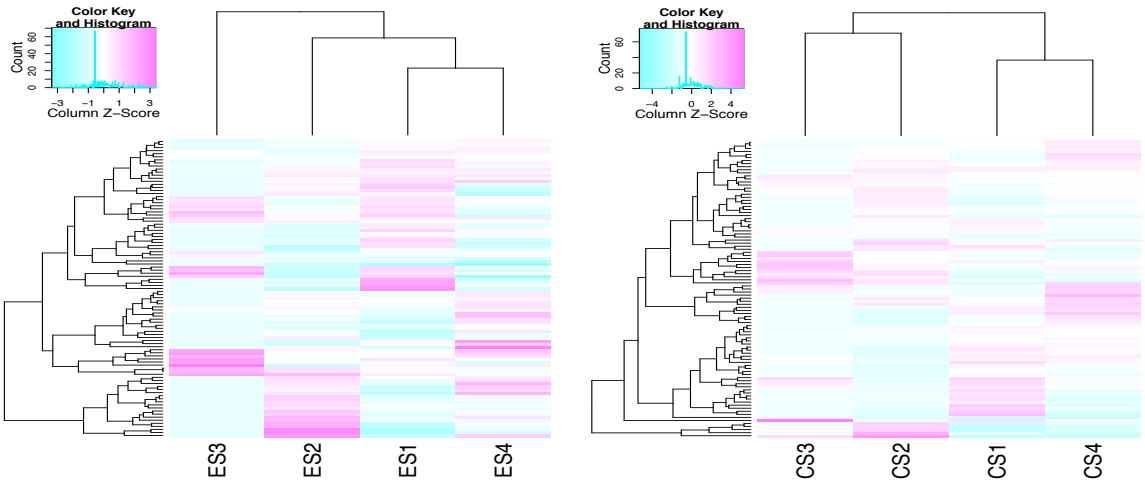


Figure 4.6.: **Distribution of gene expression connected with chromatin states.** The expression data are taken from Schmid *et al.* (2005). They are quantile-normalized and re-centered in order to consider the difference of signal dynamics between Affymetrix ATH1 array and the tiling arrays from (Roudier *et al.*, 2011). The distribution of gene expression is plotted against the association with the 4 different clusters according to the Epi-speller clustering (ES1-4, top) or the previous clusters (CS1-4, bottom). The black dashed lines show the distribution of expression levels among all tiles.

We further analyzed the results of the clustering by Epi-Speller with regard to functional annotations, applying the widely used Gene Ontology categories (Ashburner *et al.*, 2000). While shuffling cluster labels randomly 1000 times, with the same cluster size, did not result in any GO category enrichment by chance (Figure C.3), several groups of GO terms are enriched with genes associated with specific ES clusters, especially with the gene-rich cluster ES2 (Figure 4.7(a)). Importantly, this specification comprises more terms, and more pronounced enrichments, than the same analysis for the CS clusters (Figure 4.7(b)).



(a) Epi-Speller clustering

(b) Previous analysis

Figure 4.7.: **Gene Ontology analysis.** Enrichment of certain functional categories within the clusters was tested by comparing Gene Ontology (GO Slim Biological Process annotation from TAIR `www.arabidopsis.org`) distributions between clusters according to the (A) Epi-speller clustering (ES1-4) or (B) the previous clusters (CS1-4). The enrichment level is represented from high (magenta) to low (green). GO categories with less than 10 associated genes were excluded.

All together, the downstream analysis of the chromatin states indicates that the reduction of continuous values to a three-letter scheme for twelve chromatin marks and their analysis by Epi-Speller does not discard information, but rather allows detecting more refined differences between chromatin states in correlation with distinct biological annotations. This is due to the concept of direct comparisons between epigenomic

signatures, using the same alphabet size but considering differences in the distribution and prevalence of individual marks. Thus, Epi-Speller can help summarizing complex chromatin analyses, interpreting chromatin states and finding epigenomic signatures.

4.3. Discussion

Epi-Speller is a bioinformatic tool that finds common patterns among genome-wide data of abundance of different chromatin marks. This is done by automatically assigning one of three given categories for each mark at any location, transforming this information into strings of alphabet letters, grouping similar patterns, and representing them as a sequence logo. This combinatorial representation of chromatin states compiles large data amounts, helps visualizing them and allows the application of existing tools for summary statistics, comparative analysis, or motif search.

Although the three-letter alphabet was found optimal for the example provided here, the Epi-speller principle is not limited to this alphabet size. In fact, the most suitable number of categories depends on the shape of the distributions and can be easily adapted to different datasets by changing one parameter in the algorithm. Multi-modal distributions may be better represented with an increased alphabet size, whereas these may introduce unnecessary fine classification for bell-shaped distributions.

Another flexible parameter that users can adapt to their need is the choice of the clustering procedure, based on either profiling levels or epi-letter-based levels (e.g. Hamming distance). The congruency between the epi-letter-based clustering method of Epi-speller and a previously applied method, as shown for the Arabidopsis data here, is evidence for the similarity between the two principles. Nevertheless, the comparison revealed also some differences, especially in connection with annotation of genomic regions, gene expression within the clusters, and functional annotation. These differences can potentially refine the analysis of biologically relevant parameters and thereby the interpretation of chromatin data.

The application of Epi-speller presented here has so far considered only the vertical synopsis of several epigenetic parameters at individual genomic regions. The conversion of this compiled information into the letter code will also allow applying the principle of motif search in horizontal direction, along the genome. Comparing different epigenomes,

or associated with functional parameters like the location of replication origins, recombination hot- or cold spots, matrix attachment sites or three-dimensional conformation information, the analysis might reveal epi-motifs, i.e. a defined order of combinations of chromatin marks. Horizontal and vertical analysis can further be applied to decipher the chromatin signatures in connection with different organs, tissues, or cell types (Ernst *et al.*, 2011). Furthermore, in addition to Epi-Speller as a discovery tool for epigenetic signatures, it could be useful to visualize complex data sets. Integrating the Epi-speller principle into genome browsers will allow to document complex synoptic datasets in an informative and largely self-explanatory way.

4.4. Materials & Methods

4.4.1. Epigenetic signal grouping algorithm

We want to automatically group n (epigenetic) signals with intensities into g disjoint groups. The indices $i = 1, \ldots, n$ indicate the genomic regions where the signals were derived from. We want to find a clustering such that the sum of squares

$$\mathcal{D} = \sum_{i=1}^{n} (r_i - \bar{r}_i)^2 \tag{4.1}$$

is minimal, where $\bar{r}_i$ is the average of the r's that are belonging to the same group as r_i. This so-called grouping problem has been studied before (Fisher, 1958). Here we describe a dynamic programming algorithm to find the minimum value of $\mathcal{D}$ and also an assignment of the intensities to the g groups. To simplify the problem, we furthermore assume that the intensities are ordered, i.e $r_1 \leq r_2 \leq \ldots \leq r_n$. Then it was shown, that the only interesting groupings are contiguous (Fisher, 1958), that is, if $r_i \leq r_j \leq r_k$ and r_i, r_k are in one group, then r_j the intensity between r_i, r_k belongs to same group. To formulate the dynamic programming algorithm, some notation is necessary. With

$$\mathcal{D}[a, b] = \sum_{i=a}^{b} (r_i - \bar{r}_{[a,b]})^2 \tag{4.2}$$

we denote the least square value for an arbitrary contiguous group $[a, b]$, $1 \leq a \leq \ldots \leq b \leq n$, where

$$\bar{r}_{[a,b]} = \frac{1}{b - a + 1} \sum_{i=a}^{b} r_i \tag{4.3}$$

is the average intensity in $[a, b]$. Then, a dynamic programming algorithm solves the partitioning problem:

Let $\mathcal{P} = (P(t, i))$ be a $g \times n$ matrix where each element $P(t, i)$ represents the minimal sum of squared differences for partitioning $[1, i]$ into t groups. The first row of the matrix is initialized as follows: $P(1, i) = D[1, i]$ for $i = 1, \ldots, n$. The following rows $t = 2, \ldots, g$ are computed iteratively according to

$$P(t, i) = \min_{j<i}\{P(t-1, j) + D[j+1, i]\} \text{ for } i = t, \ldots, n \tag{4.4}$$

This is done until the full matrix is computed:

$$\mathcal{P} = \begin{pmatrix} \mathrm{P(1,1)} & \mathrm{P(1,2)} & \mathrm{P(1,3)} & \ldots & \ldots & \mathrm{P(1,n)} \\ 0 & \mathrm{P(2,2)} & \mathrm{P(2,3)} & & & \mathrm{P(2,n)} \\ & 0 & \mathrm{P(3,3)} & & & \mathrm{P(3,n)} \\ \vdots & & 0 & & & \\ & & \vdots & & & \\ 0 & \ldots & 0 & \mathrm{P(g,g)} & \ldots & \mathrm{P(g,n)} \end{pmatrix}$$

The entry $P(g, n)$ minimizes equation 4.1. Finally an optimal partition is constructed by a trace back procedure: It starts at $P(g, n)$ and determines in the $(g-1)^{th}$ row the matrix entry that minimizes equation 4.4; this is repeated until the first row is hit. Thus, we have determined the minimum value for equation 4.1 and a corresponding optimal partition of the intensities. The computing complexity of the dynamic programming approach is $O(gn^2)$. The algorithm was implemented in C^{++}.

In all analyses the number of groups was set equal to three. One group collects the genomic regions with a low signal intensity (L), one groups collects genomic regions with medium intensity (M), and the third group contains the genomic regions with a high intensity (H). The partitioning of the data was done as described above. Subsequently, we replaced the continuous signal r_i in genomic region i by exactly one epi-letter L, M, H depending on the group the signal intensity was assigned to.

4.4.2. Published data from 12 chromatin mark tiling arrays from Arabidopsis thaliana

Roudier *et al.* (2011) published signal intensities measured on ChIP tiling arrays (ChIP-chip) for 12 chromatin marks on *Arabidopsis thaliana* chromosome 4. They measured

the signal strength for two methylation forms (di- and tri-) of H3K4 and H3K9, three forms (mono-, di- and tri-) methylation of H3K27, H3K36me3, H3K56ac, H4K20me1, H2Bub, and DNA methylation. The tiles cover approx 90% of the chromosome. Our analysis is based on $n = 17.356$ tiles, where data for all 12 chromatin marks are available. For each epigenetic mark, we computed the optimal classification of the signals into three groups as described. Each tile represents one genomic region. Thus the twelve epigenetic signals per tile are replaced by their corresponding epi-letters L, M, or H. Details on the distribution of tiles for each letter across different chromatin marks were shown in the Supp. Figure C.1.

4.4.3. Clustering the genomic tiles by Epi-Speller

Each tile is represented by an epi-letter tile composed of 12 epi-letters where each epi-letter represents the relative strength of the corresponding chromatin mark with respect to its chromosome-wide distribution. To detect similarities between different tiles/genomic regions, we computed the Hamming distance between all pairs of epi-letter tiles, asking how many pairs of letters are different between each other for the same mark. For example, the distance between LLMHMLHMLHML and LLMHMLHMLHHM is 2 because of the epi-letter pairs (M,H) and (L,M) are different at mark 11 and 12 in the epi-letter-based representation. The resulting distance matrix was used to group the tiles into $k = 3, , 12$ mutually, disjoint clusters such that the tiles in the same cluster have smaller distance compared to the ones in other clusters by classical "k-mean" clustering (Lloyd, 1982), using the kmeans function in the R core package (`www.r-project.org`) with default parameters.

4.4.4. Summarizing and representing the tile cluster using sequence logo

To visualize the similarity between epi-strings in the same cluster, we compute the corresponding sequence logo. The sequence logo is a visualization of the degree of information content (in bits with maximal $log2(3)$ in 3-letter alphabet) at each position in the epi-strings. Here, we used the WebLogo tool (`http://weblogo.berkeley.edu` (Crooks *et al.*, 2004)) to illustrate the epi-string logo for the tile cluster. Each tile

cluster is also called a chromatin state and will be validated by the following biological annotation analysis in supporting the chromatin-based epigenetic code. The logo is also used to check the similarity between different chromatin states to then choose the final number of states.

4.4.5. Biological annotation analysis

Finally, we investigated if the clusters can be further validated by a biological annotation analysis. Three aspects were investigated: genomic feature (i.e gene/transposon) distribution, gene expression analysis, GO functional category analysis.

The gene and transposon element annotation is from TAIR8 version of Arabidopsis genome to compare our results with previously published results (Roudier *et al.*, 2011). The gene expression analysis was based on developmental microarray data from (Schmid *et al.*, 2005). We discarded tiles with expression level less than 7 (log2 scale) suggested as noise in Roudier *et al.* (2011). The functional category analysis is based on GO Slim specially categorized for the Arabidopsis genome by TAIR website (`http://www.arabidopsis.org/help/helppages/go_slim_help.jsp`). The enrichment level for a GO category from a given chromatin state is computed by the number of genes in that state which belongs to the given category normalized by total number of genes in the category and the total number of genes clustered in the state.

Appendix A.

Supplementary Figures & Tables to chapter 2

Table A.1.: **Annotated of validated regions by Sanger sequencing**

Category	Coordinate	Annotation	AGI Name
M-M	Chr2W_5782325:5782536	Repeat	
	Chr2W_5782325:5782536	Repeat	
U-U	Chr4W_13980472:13980941	Genic	AT4G28160
	Chr5C_5380040:5380291	Genic	AT5G11770
M-U	Chr1W_22875577:22875875	TE	AT1E75404
	Chr1C_21132234:21132534	Inter-TE	AT1TE69815 & AT1TE69865
M-X	Chr1W_7353537:7354023	TE	AT1G21020
	Chr1C_10803163:10803412	Intergenic	AT1TE34895&AT1G30500
	Chr2W_5087102:5087413	TE	AT2TE20895
	Chr3C_14331158:14331560	Intergenic	AT3TE58820&AT3G42180
	Chr4C_8421678:8422043	Intergenic	AT4G14690&AT4G14700
	Chr5W_7494919:7495438	TE	
U-M	Chr2W_7306350:7306698	Intergenic	AT1TE23555&AT1G20967
	Chr3W_6728921:6729226	5 upstream	AT3G19410

Table A.2.: **Comparison of mapping results for BiSS and other selected aligning programs[a] with simulated data[b]**

	Running time[c]	#of uniquely [d]	# of correctly [e]	False Positives(%)
Arabidopsis chromosome 1[f]				
BSMAP[g]	**1 min**	896,665	894,121	0.28
RMAP[h]	1.89 mins	892,905	890,732	0.24
BS-Seeker	1.93 mins	875,949	860,737	1.74
BisMark[i]	6.58 mins[j]	875,190	873,697	0.17
Human chromosome 21[k]				
BSMAP	**1.68 mins**	974,176	972,792	0.14
RMAP	2.04 mins	972,599	970,001	0.27
BS-Seeker	2.14 mins	961,751	958,205	0.37
BisMark	6.57 mins	976,237	974,746	0.15
BiSS	6.02 mins	**977,578**	**974,914**	0.27

[a] default parameters unless otherwise specified.

[b] Arabidopsis chromosome 1 and human chromosome 21 one million single-end reads

[c] using multi-processor support (up to 8 cores, for the program that support like BSMAP) in the same computer. Other programs used 1 core.

[d] Number of uniquely mapped reads, except BiSS - we use only the #1 top hit among the highest SW-scored read alignments

[e] Number of correctly mapped reads. A read is called correctly mapped, if the coordinates of the first aligned position and the last aligned position agree with the origin of the read in the genome

[f] The simulation is generated based on Arabidopsis methylation data (0.55:0.23:0.22 for CG:CHG:CHH methylation at 6% frequency) using our adaption from **wgsim** program (from **SAMtools**) with 2% sequence error and 2% bisulfite conversion error.

[g] BSMAP version 1 parameters: -p 8 -s 12 -r 2 -w 100 -n 1 -v 5 -g 5, suggested by the authors, maximal 5 mismatches

[h] RMAP parameters: -m 5 v, default parameters

[i] BisMark parameters: -n 3 -l 56 -e 150

[j] * exluding the processing/indexing reference genome by **bowtie**

[k] The simulation is generated for human chromosome 21 (6% CpG methylation frequency) using RMAP package (well-designed for CpG methylation in mammalian) with 98% bisulfite conversion and maximal 5 mismatches . Command line: **./simreadsbs chr21.fa -o read_chr21_rmap.fa -n 1000000 -w 72 -e 5 -m 6 -b 98**

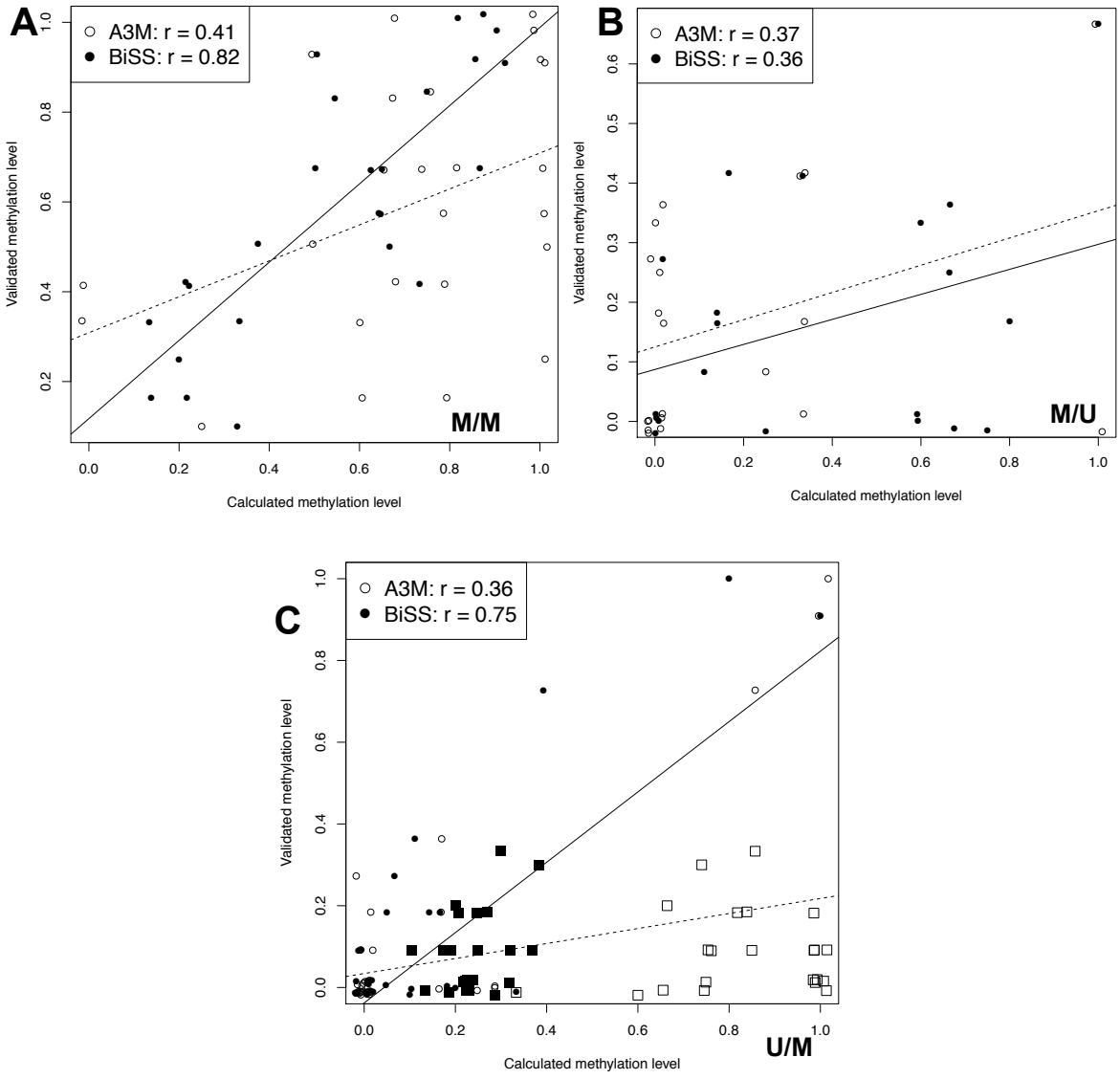


Figure A.1.: **Examples for validation by individual bisulfite sequencing.** The plots show the correlation between calculated and validated methylation levels (C/(C+T)) from regions selected for congruency (A) or disagreement (B-C) between BiSS and A3M. Each point represents one cytosine position. The x-axis corresponds to the methylation levels calculated from either BiSS (filled circles and black regression lines) or A3M (open circles and dotted regression lines); the y-axis shows the result of individual bisulfite sequencing. The legends show the Pearson correlation coefficients. (A) Methylated region according to both methods (M/M). (B) A region called methylated by BiSS but not by A3M (M/U); the rectangles indicate experimentally validated Cs congruent to BiSS (filled) and discrepant to A3M (open). (C) A region called unmethylated by BiSS but methylated by A3M (U/M); the rectangle symbols are the same as in (B).

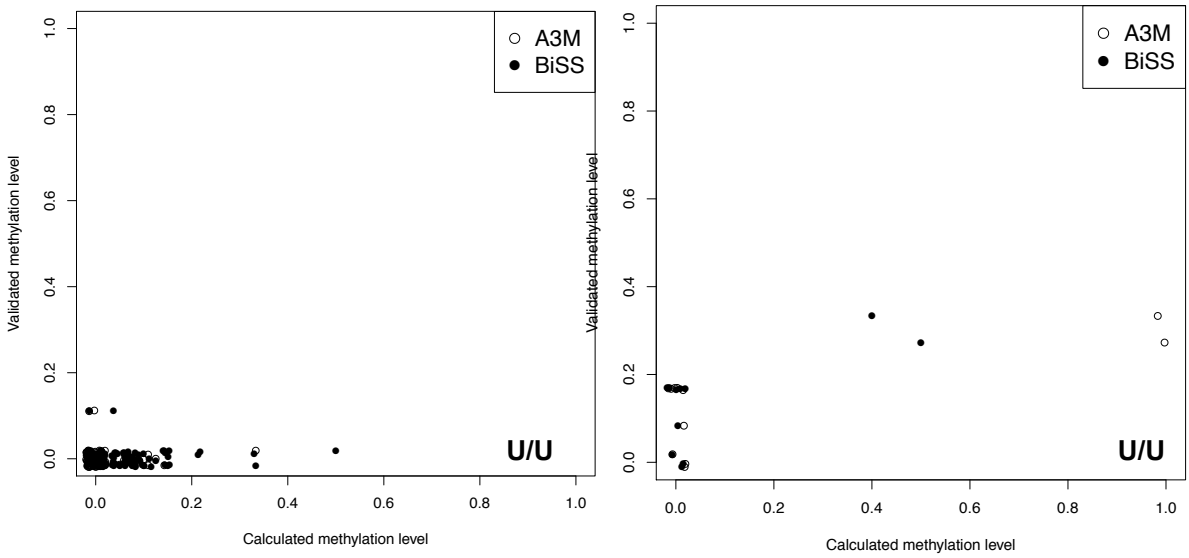


Figure A.2.: **Examples for validation by individual bisulfite sequencing.** The plots show the correlation between calculated and validated methylation levels (C/(C+T)) from regions selected for congruency (U/U) between BiSS and A3M. Each point represents one cytosine position. The x-axis corresponds to the methylation levels calculated from either BiSS (filled circles) or A3M (open circles); the y-axis shows the result of individual bisulfite sequencing.

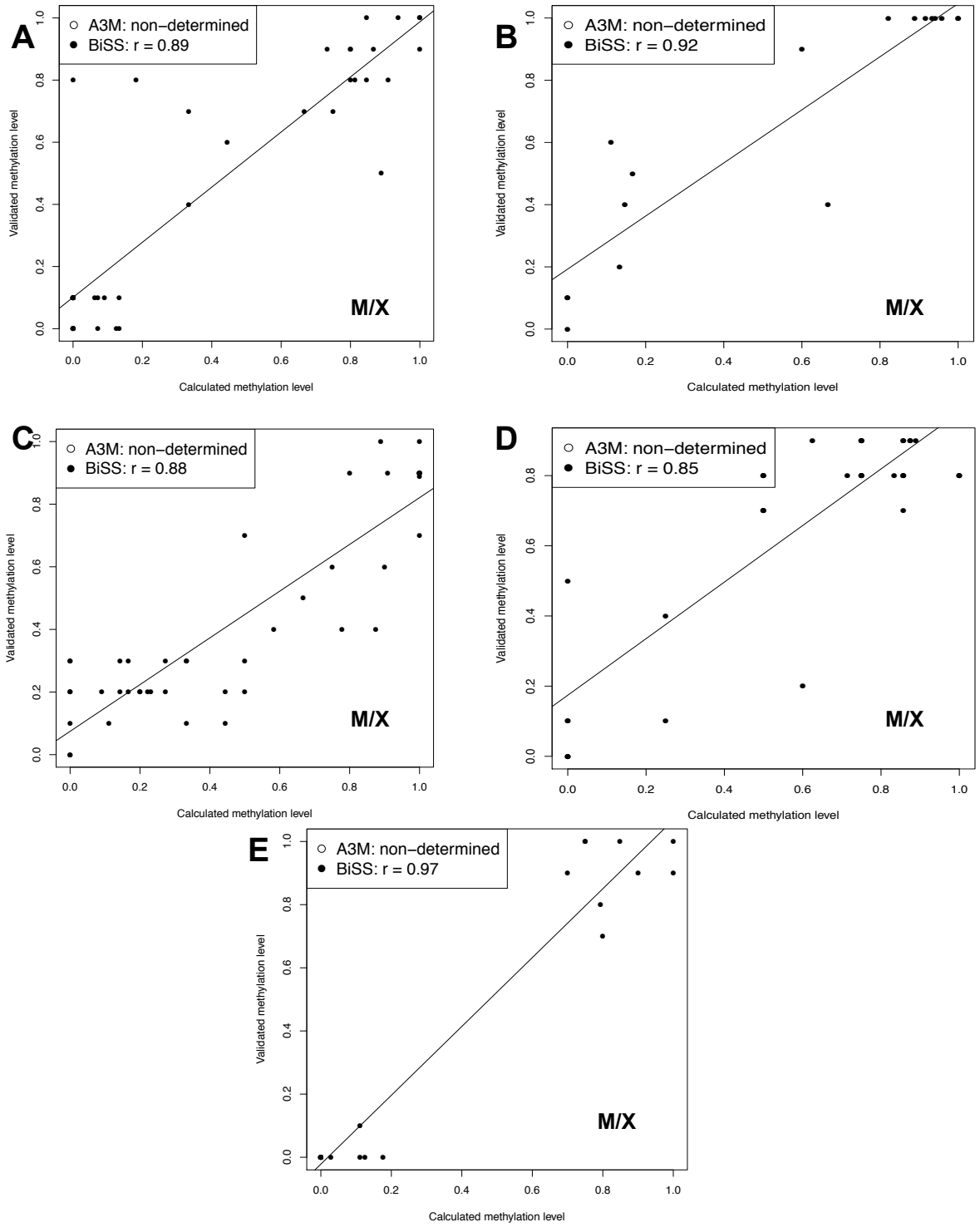


Figure A.3.: **Examples for validation by individual bisulfite sequencing.** The plots show the correlation between calculated and validated methylation levels (C/(C+T)) from 5 different regions (A-E) selected for methylation calling by BiSS versus undetermined state by A3M. Each point represents one cytosine position. The x-axis corresponds to the methylation levels calculated from BiSS; the y-axis shows the result of individual bisulfite sequencing. The legends show the Pearson correlation coefficients.

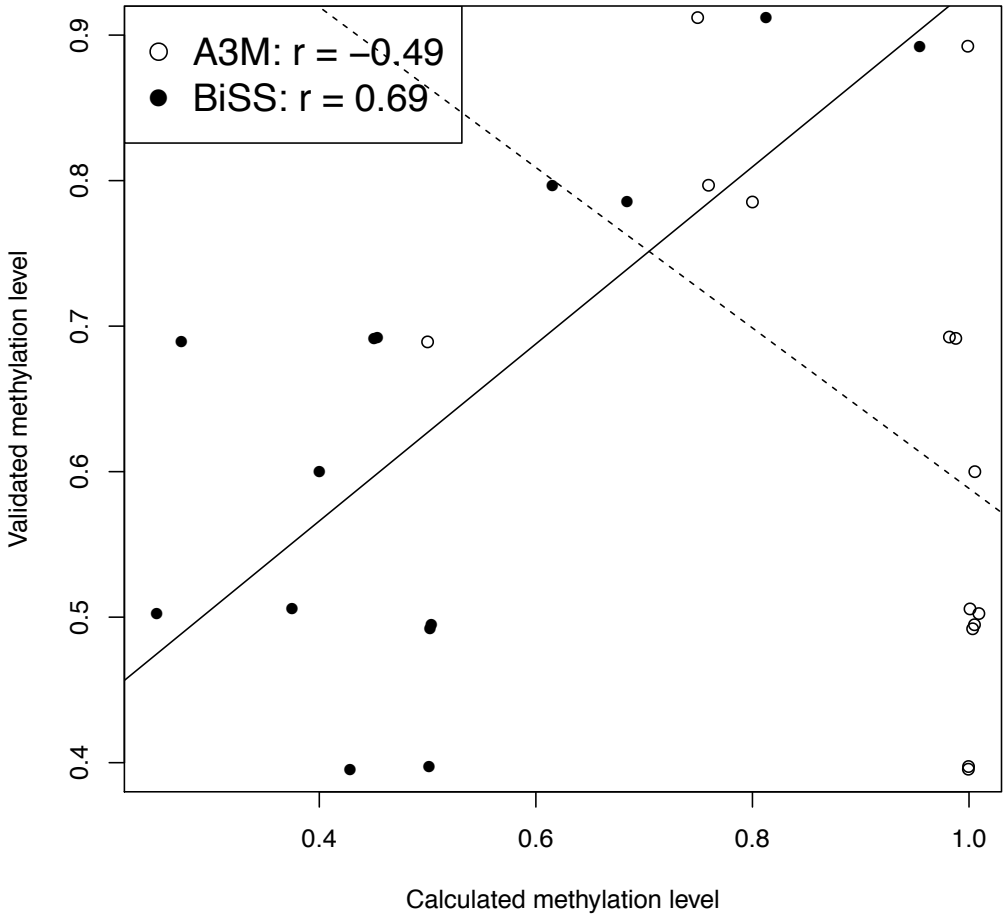


Figure A.4.: **Examples for validation by individual bisulfite sequencing.** The plot shows the correlation between calculated and validated methylation levels (C/(C+T)) from a region selected for disagreement between BiSS (calling it unmethylated) and A3M (calling it methylated). Each point represents one cytosine position. The x-axis corresponds to the methylation levels calculated from either BiSS (filled circles and black regression lines) or A3M (open circles and dotted regression lines); the y-axis shows the result of individual bisulfite sequencing. The legends show the Pearson correlation coefficients.

Appendix B.

Supplementary Figures to chapter 3

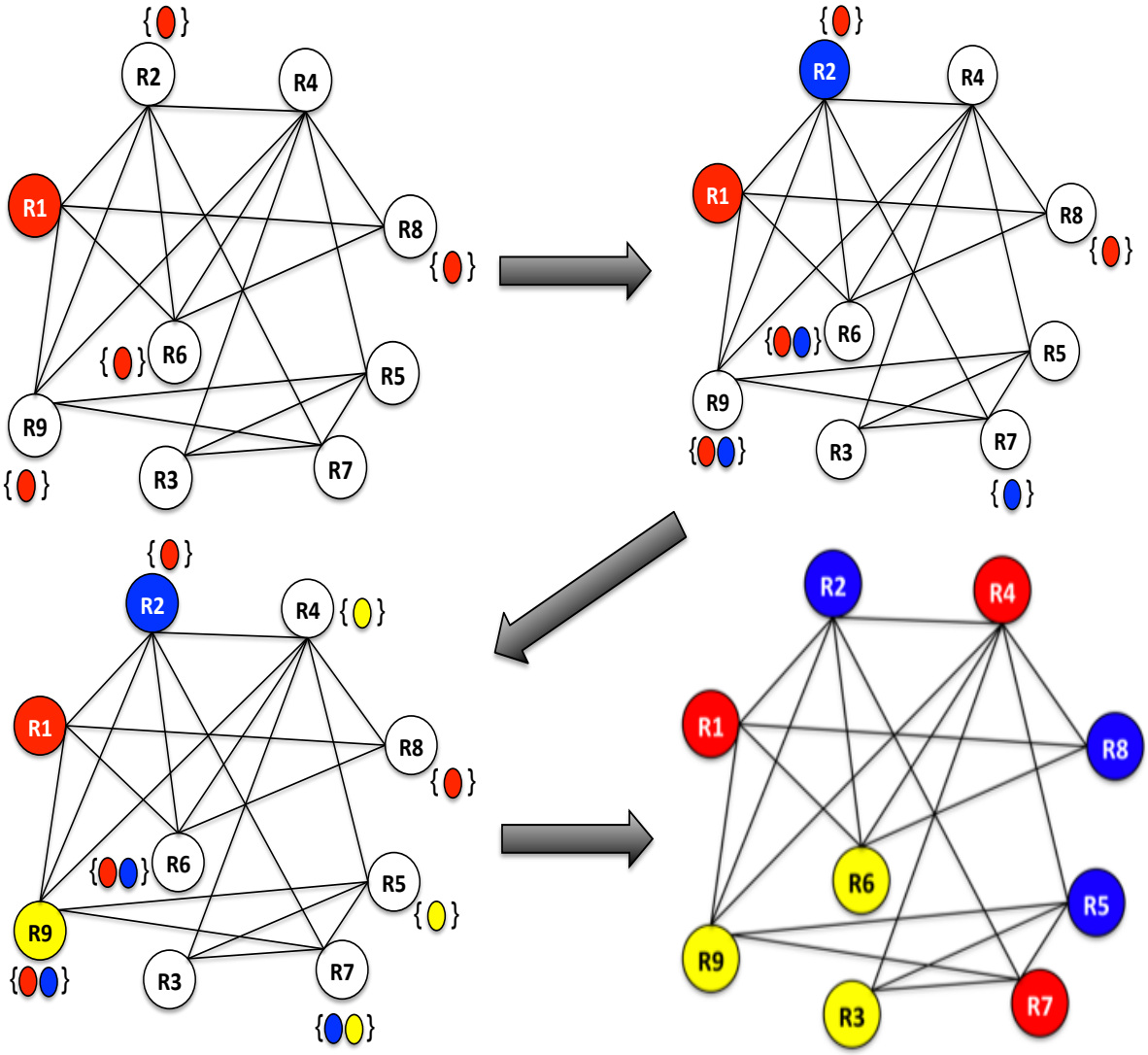
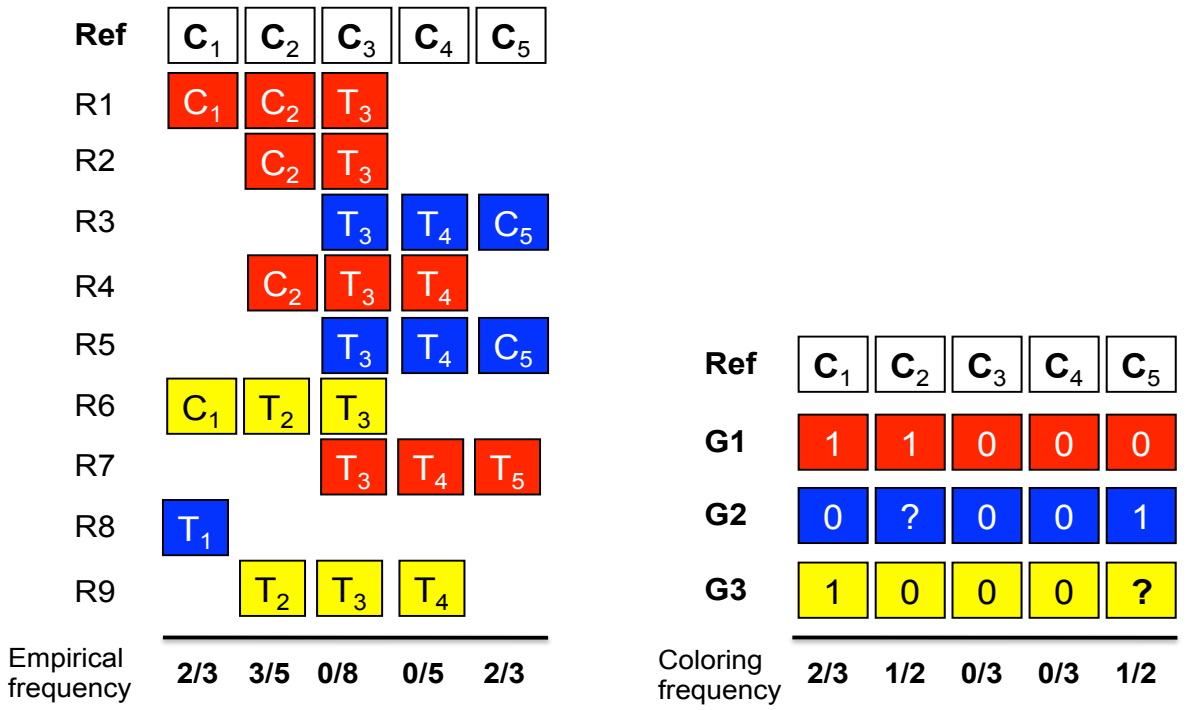
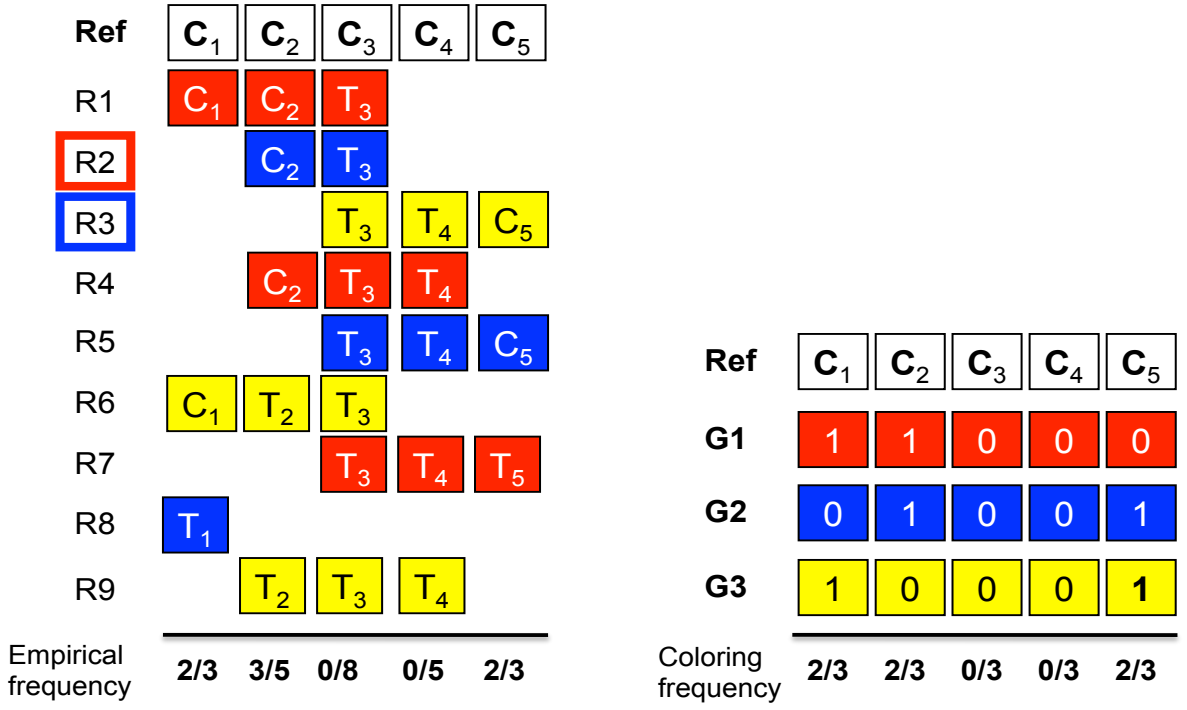


Figure B.1.: **A working example of greedy algorithm**:
The red color is randomly assigned to the read node $R1$, hence red cannot be assigned any more to the nodes $R2, R6, R8, R9$ (denoted in the neighboring color list {}).
The node $R2$ is chosen for next color (blue) as it has the maximal cardinality of neighboring color list.
Similarly, blue is excluded for the nodes $R6, R7, R9$. The node $R9$ is chosen for the next color (yellow, $R6$ could have been chosen).
After several steps, all the nodes are colored and it is clear that the input graph needs only 3 colors to fulfill the requirement that there should be no connected nodes with the same color.



(a) An example of read mapping profiles for 9 reads with 3 assigned colors from coloring algorithm: R - read, G - methylation profile with corresponding color: 1 - methylated, 0 - unmethylalated and ? indicates cytosines whose no assigned read. Color of the read nodes are heuristically re-assgined by directly comparing the methylation frequency from the empirical mapping and from the coloring inference.



(b) Mapping profile after reassigning the colors for read R2 (red → blue) and read R3 (blue → yellow). The ?-state are uncovered, and the empirical and coloring frequencies of methylation are closer.

Figure B.2.: **Color reassignment based on the empirical heuristics**

Appendix C.

Supplementary Figures to chapter 4

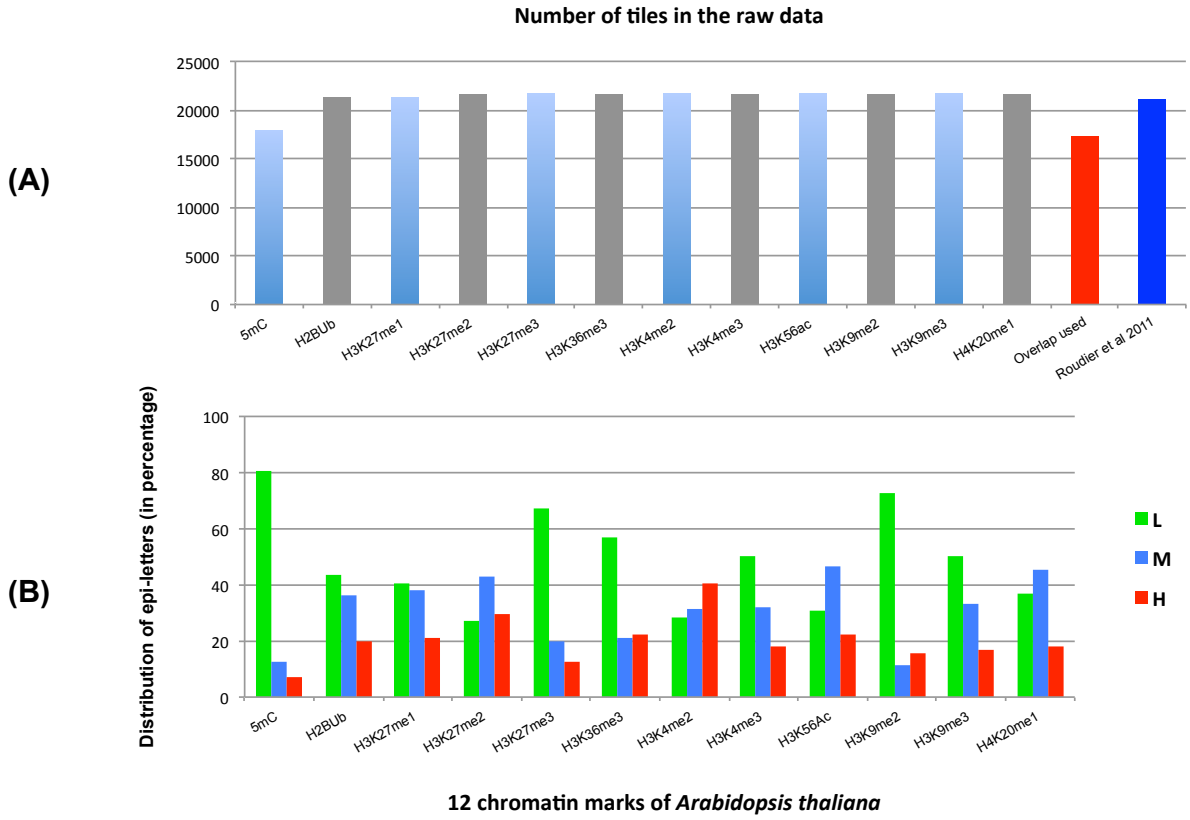
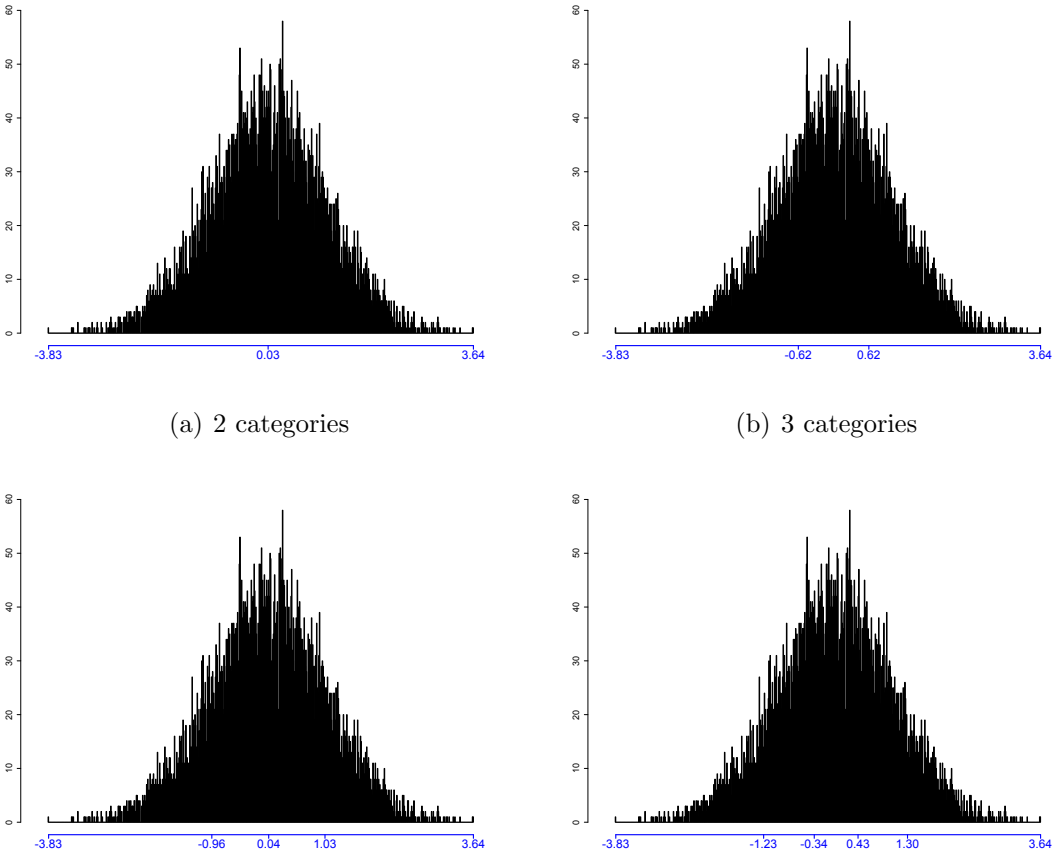


Figure C.1.: **Data summary.**
(A) Absolute number of tiles corresponding to the 12 respective tiling arrays. (B) 3-letter distribution for 12 chromatin marks.



(a) 2 categories

(b) 3 categories

(c) 4 categories

(d) 5 categories

Figure C.2.: **Grouping algorithm for simulated standard normal distribution.** Cutoffs (as tick in the x-axis) indicate the results of the grouping algorithm in case of (A) 2, (B) 3, (C) 4, or (D) 5 categories.

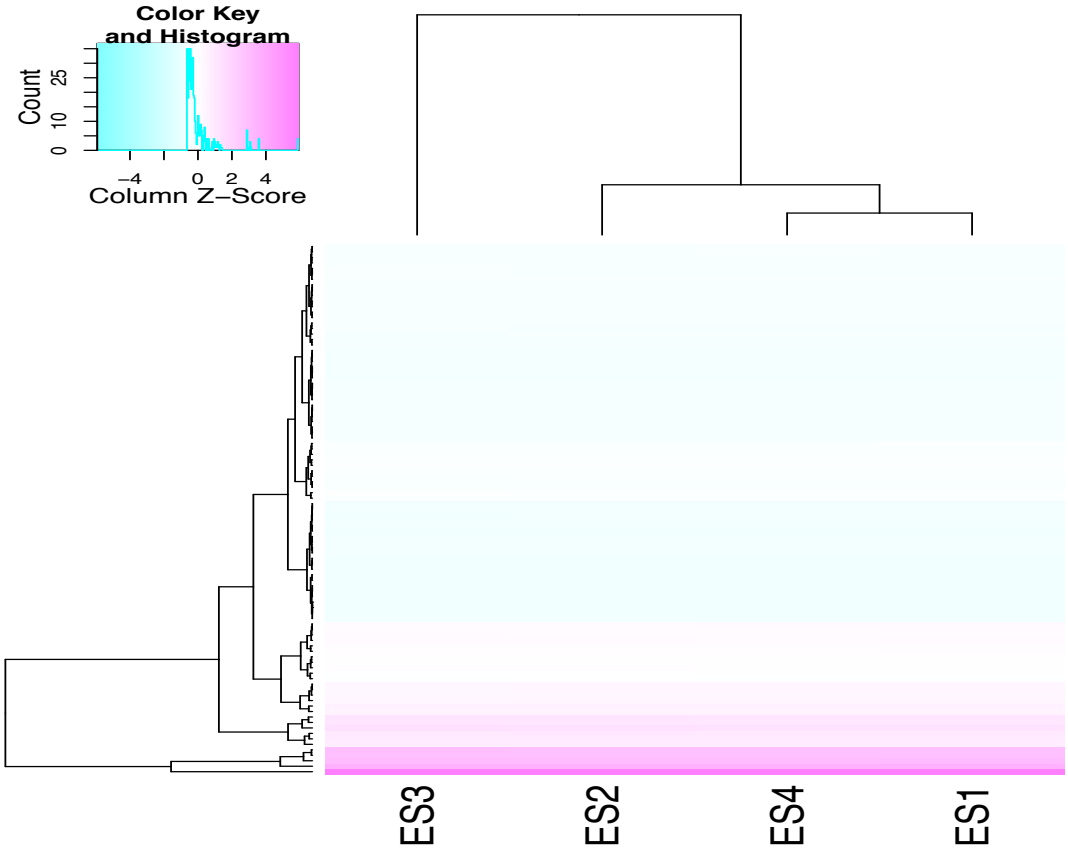


Figure C.3.: **GO analysis for random cluster.** Random clusters were produced by 1000-time bootstrap given the same cluster size from ES clustering results. GO enrichment levels (with the same condition, see text in chapter 4 and Figure 4.7) were then computed as average of 1000 times.

Acknowledgments

The writing of this dissertation is one of the most adventurous challenges in my academic career up to now. Without the support and collaboration of the following people, it would never been completed. It is to them I owe my deepest gratitude.

- Prof. Ortrun Mittelsten Scheid (GMI), my biology supervisor, thanks to whom I am engaged in epigenetic research. Her patiences (especially to address/encourage all of my never-ending ignorant biological questions and ideas) and constant cares in both non- and science stuffs have made my life here much easier.
- Prof. Arndt von Haeseler (CIBIV/MFPL), my bioinformatic supervisor, who introduced me to Ortrun and always encourages my new ideas while insisting on "getting things done first". His saying "Biologists read papers by figures" and "Do not read much, think first" are among my favorites.
- Prof. Martin Vingron (MPI, Berlin) and Prof. Joern Walter (University of Saarlandes), who accepted to review my dissertation despite their busy schedules.
- The (co-)authors of all the projects I was proud of being involved during my PhD time here, in particular Fritz Sedlazeck (CIBIV) for his very kind help in incorporating bisulfite mapping in his NextGenMap tool.
- Colleagues and administrative personnel from GMI, MFPL, IMP and the whole campus whom I have opportunities to collaborate/interact with in such a world-class research environment. I would like to thank all group members and alumni from the OMS lab (GMI) and from the CIBIV (MFPL) where I did learn a lot of complementary things from each side. What an unique experience!
- My family in Vietnam, especially my father, who always have faith in the career I am pursuing and never ask why I did not come back home for such a long time.
- Hany, my wife, without whom this effort would be worth nothing. This dissertation is dedicated to her never-ending love and patience. Now, the next adventure!

Curriculum Vitae

Huy Q. Dinh

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Dr. Bohr Gasse 9, 1030 Vienna, Austria
Phone: +43 1 / 42772-4030
Email: huy.dinh@univie.ac.at or huy.dinh@gmi.oeaw.ac.at
Homepage: www.cibiv.univie.ac.at/~huy

Personal Information

Full Name: Huy Quang Dinh (*in native* Đinh Quang Huy).
Date of Birth: 3, Oct, 1982.
Place of Birth: Phu Tho, Vietnam.
Nationality: Vietnamese.
Family Status: married.

Research Interests

- Computational methods for analyzing genome-wide (epigenomic) profiling data.
- Machine learning approaches towards understanding epigenetic/gene regulation.
- Combinatorial optimization algorithms and applications in computational biology.

Education

- 2008 - 2012: PhD in Bioinformatics at the Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories (University of Vienna and Medical University of Vienna) & Gregor Mendel Institute of Molecular Plant Biology, Vienna (Austrian Academy of Sciences).
- 2005 - 2007: Master in Computer Science, Vietnam National University, Hanoi.
- 2000 - 2004: Bachelor in Computer Science, Vietnam National University, Hanoi.

Degree

- 'M.Sc.' 2007, Vietnam National Univeristy, Hanoi, Vietnam.
  Thesis: An Ant Colony Optimization approach for phylogenetic reconstruction problem.
- 'B.Sc.' 2004, Vietnam National University, Hanoi, Vietnam.
  Thesis: Ant Colony Optimization and Scheduling problem.

Selected Awards

- 2012: Traveling fellowship by Swiss Institute of Bioinformatics for the most promising abstract at the 20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2012), Long Beach, CA, USA, July 2012.
- 2010: Traveling fellowship by National Institute of Health to attend the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2010), Boston, MA, USA, July 2010.
- 2006: Traveling fellowship by Institute of Electrical and Electronics Engineers (IEEE) to attend and contribute an oral presentation at the 4th IEEE International Conference in Computer Sciences, Research, Innovation, and Vision for Future (RIVF2006), Ho Chi Minh city, Vietnam, February, 2006.
- 2004-2006: Research and teaching assistant stipends in Vietnam National University, Hanoi.

- 2004: Two first prizes at student scientific conference for the undergraduate works at Vietnam National University, Hanoi.
- 2000-2004: Annual stipends for excellent students at Vietnam National University, Hanoi.
- 2003: Young Representative Prize for best students at Vietnam National University, Hanoi.
- 2000: Third Prize in National Informatics Olympiad, Vietnam.

Publications

1. **Dinh, HQ, Mittelsten Scheid O, von Haeseler A (2012)**: Epi-Speller - a tool to discover epigenetic signatures. (*submitted*).
2. **Dinh HQ, Dubin M, Sedlazeck FJ, Letter N, Mittelsten Scheid O, von Haeseler A (2012)**: Advanced methylome analysis after bisulfite deep sequencing: an example in Arabidopsis. *PLoS ONE 7, e41528.*
3. **Yanez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A (2012):** Uncovering cis-regulatory sequence requirements for context specific transcriptional factor binding. *Genome Research. [Epub ahead of print].*
4. **Do DD, Dinh HQ, Dang TH, Laukens K, Hoang XH (2012)**: ACOSEED: an Ant Colony Optimization for finding optimalspaced seeds in biological sequence search. *Swarm Intelligence - 8th International Conference (ANTS2012), Belgium: 360-367, LNCS in press.*
5. **Popova OV, Dinh HQ, Aufsatz W, Jonak C (2012)**: RNA-directed DNA methylation pathway is involved in heat response. (*in preparation*).
6. **Dinh, HQ, Mittelsten Scheid O, von Haeseler(2011):** MethColor: a computational approach to uncover DNA methylation heterogeneity. *In the Proceeding of the German Conference on Bioinformatics 2011 (GCB2011), Germany.*
7. **Kalyna M, Simpson C, Syed N, Lewandowska D, Marquez Y, Kusenda B, Marshall JZ, Fuller J, Cardle L, McNicol J, Dinh HQ, Barta A, Brown J (2011):** Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res. 40:2454-2469.*

8. **Foerster A, Dinh HQ, Sedman L, Wohlrab B, Mittelsten Scheid O (2011):** Genetic rearrangements can modify chromatin features at epialleles. *PLoS Genetics 7, e1002331*

9. **Brini F, Yamamoto A, Jlaiel L, Takeda S, Hobo T, Dinh HQ, Hattori T, Masmoudi K, Hanin M (2011):** Pleiotropic effects of the wheat dehydrin DHN-5 on stress responses in Arabidopsis. *Plant Cell Physiol. 52:676-688.*

10. **Pecinka A, Dinh HQ, Baubec T, Rosa M, Lettner N, Mittelsten Scheid O (2010):** Epigenetic control of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell 22:3118-3129.*

11. **Dinh HQ, Minh BQ, Hoang XH, von Haeseler A (2010):** ACOPHY: A simple and general ant colony optimization approach for phylogenetic tree reconstruction. *Swarm Intelligence - 7th International Conference (ANTS2010), Belgium: 360-367, LNCS.*

12. **Baubec T, Dinh HQ, Pecinka A, Rakic B, Rozhon W, Wohlrab B, von Haeseler A, Mittelsten Scheid O (2010):** Cooperation of multiple chromatin modifications can generate unanticipated stability of epigenetic states in Arabidopsis. *Plant Cell 22:34-47.*

13. **Dong Do Duc, Tri-Thanh Le, Trung-Nghia Vu, Dinh HQ, Hoang Xuan Huan (2012):** GA_SVM: A Genetic Algorithm for Improving Gene Regulatory Activity Prediction. *IEEE RIVF 2012 conference: 1-4, Vietnam.*

14. **Dong Do Duc, Dinh HQ, and Huan Hoang Xuan (2008)**: On the Pheromone Update Rules of Ant Colony Optimization Approaches for the Job Shop Scheduling Problem. *PRIMA 2008 conference: 153-160, LNCS.*

15. **Dinh Quang Huy, Dong Do Duc, Hoang Xuan Huan (2006)**: Multi-level ant system - a new approach through the new pheromone update for ant colony optimization. *IEEE RIVF 2006 conference: 55-58.*

16. **Dinh Quang Huy, Dinh Manh Tuong (2005)**: Link-Connected: A novel approach of clustering algorithm for categorical attributes. *In Proceedings of the 9th national conference in Information Technology, pp 354-363*

Vienna, August 2012

List of Abbreviations

A3M Arabidopsis 3-letter Methylome. 17–22, 24–27, 30, 31, 67–70

BiSS BIsufite Sequencing Scorer. 15–22, 24–28, 30–32, 66–70

BS bisulfite. 2, 3, 8–11

BS-Seq BiSulfite deep Sequencing. 7, 8, 10, 12, 13, 15–18, 26, 27

ChIP Chromatin ImmunoPrecipitation. 4, 6–8, 13, 44, 45, 61

ChIP-chip ChIP followed by microarray. 4, 61

ChIP-Seq ChIP followed by next generation sequencing. 4

DamID DNA adenine methyltransferase identification. 44, 45

HMM Hidden Markov Model. 44, 46

ITBS traditional individual traditional bisulfite sequencing. 18, 24, 26

NGS Next Generation Sequencing. 2, 6–8, 17, 28

NP-hard Non-deterministic Polynomial-time hard. 11, 36

PCA Principal Component Analysis. 44, 46

Bibliography

Alabert, C. and Groth, A. (2012) Chromatin replication and epigenome maintenance. *Nat Rev Mol Cell Biol*, **13**, 153–167.

Allis, C., Jenuwein, T. and Reinberg, D. (2007) Overview and concepts. In *Epigenetics*, Cold Spring Harbor Laboratory Press.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**, 25–29.

Bailey, T. L. (2008) Discovering sequence motifs. *Methods Mol Biol*, **452**, 231–251.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, pp. 289–300.

Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**, 6–21.

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

Cancer Research Human Epigenome Task Force and European Union, Network of Excellence (2008) Moving ahead with an international human epigenome project. *Nature*, **454**, 711–715.

Chatterjee, A., Stockwell, P. A., Rodger, E. J. and Morison, I. M. (2012) Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Res*, **40**, e79.

Chen, P.-Y., Cokus, S. J. and Pellegrini, M. (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.

Clark, S. J., Harrison, J., Paul, C. L. and Frommer, M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*, **22**, 2990–2997.

Coarfa, C., Yu, F., Miller, C. A., Chen, Z., Harris, R. A. and Milosavljevic, A. (2010) Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel dna sequencing. *BMC Bioinformatics*, **11**, 572.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Cormen, T. H., Leiserson, C. E. and Rivest, R. L. (1990) *Introduction to algorithms*. The MIT Press/McGraw-Hill, Cambridge, Masachusetts.

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

Docherty, S. J., Davis, O. S. P., Haworth, C. M. A., Plomin, R. and Mill, J. (2010) DNA methylation profiling using bisulfite-based epityping of pooled genomic dna. *Methods*, **52**, 255–258.

Dyachenko, O. V., Schevchuk, T. V., Kretzner, L., Buryanov, Y. I. and Smith, S. S. (2010) Human non-cg methylation: are human stem cells plant-like? *Epigenetics*, **5**, 569–572.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W. and Beerenwinkel, N. (2008) Viral population estimation using pyrosequencing. *PLoS Comput Biol*, **4**, e1000074.

Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, **28**, 817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and

Bernstein, B. E. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. and van Steensel, B. (2010) Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, **143**, 212–224.

Fisher, W. D. (1958) On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53**, 789–798.

Foerster, A. M. and Mittelsten Scheid, O. (2010) Analysis of dna methylation in plants by bisulfite sequencing. *Methods Mol Biol*, **631**, 1–11.

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

Harris, E. Y., Ponts, N., Levchuk, A., Roch, K. L. and Lonardi, S. (2010a) BRAT: bisulfite-treated reads analysis tool. *Bioinformatics*, **26**, 572–573.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., Olshen, A., Ballinger, T., Zhou, X., Forsberg, K. J., Gu, J., Echipare, L., O'Geen, H., Lister, R., Pelizzola, M., Xi, Y., Epstein, C. B., Bernstein, B. E., Hawkins, R. D., Ren, B., Chung, W.-Y., Gu, H., Bock, C., Gnirke, A., Zhang, M. Q., Haussler, D., Ecker, J. R., Li, W., Farnham, P. J., Waterland, R. A., Meissner, A., Marra, M. A., Hirst, M., Milosavljevic, A. and Costello, J. F. (2010b) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*, **28**, 1097–1105.

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

Holliday, R. and Pugh, J. E. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.

Hon, G., Ren, B. and Wang, W. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*, **4**, e1000201.

Hon, G., Wang, W. and Ren, B. (2009) Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, **5**, e1000566.

Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, **33 Suppl**, 245–254.

Jenuwein, T. and Allis, C. D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Jiang, C. and Pugh, B. F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, **10**, 161–172.

Kakutani, T., Munakata, K., Richards, E. J. and Hirochika, H. (1999) Meiotically and mitotically stable inheritance of dna hypomethylation induced by ddm1 mutation of arabidopsis thaliana. *Genetics*, **151**, 831–838.

Klose, R. J. and Bird, A. P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci*, **31**, 89–97.

Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.

Krueger, F. and Andrews, S. R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, **27**, 1571–1572.

Laird, P. W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*, **11**, 191–203.

Leutwiler, L. S., Meyerowitz, E. M. and Tobin, E. M. (1986) Structure and expression of three light-harvesting chlorophyll a/b-binding protein genes in Arabidopsis thaliana. *Nucleic Acids Res*, **14**, 4051–4064.

Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**, 1851–1858.

Lister, R. and Ecker, J. R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res*, **19**, 959–966.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. and Ecker, J. R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J. A., Evans, R. M. and Ecker, J. R. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.

Lloyd, S. (1982) Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**, 129–137.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C. and Maleszka, R. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol*, **8**, e1000506.

Lvque, B. and Maffray, F. (2005) Coloring meyniel graphs in linear time. *Electronic Notes in Discrete Mathematics*, **22**, 25 – 28, 7th International Colloquium on Graph Theory.

Marks, P., Rifkind, R. A., Richon, V. M., Breslow, R., Miller, T. and Kelly, W. K. (2001) Histone deacetylases and cancer: causes and therapies. *Nat Rev Cancer*, **1**, 194–202.

Matzke, M. and Mittelsten Scheid, O. (2006) Epigenetic regulation in plants. In David Allis, C., Jenuwein, T. and Reinberg, D. (eds.), *Epigenetics*, Cold Spring Harbor Laboratory Press.

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S. and Rinn, J. L. (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol*, **30**, 99–104.

Mikeska, T., Candiloro, I. L. and Dobrovic, A. (2010) The implications of heterogeneous DNA methylation for the accurate quantification of methylation. *Epigenomics*, **2:4**, 561–573.

Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E. and Ecker, J. R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.

Ning, Z., Cox, A. J. and Mullikin, J. C. (2001) SSAHA: a fast search method for large dna databases. *Genome Res*, **11**, 1725–1729.

Park, P. J. (2009) Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**, 669–680.

Pedersen, B., Hsieh, T.-F., Ibarra, C. and Fischer, R. L. (2011) MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, **27**, 2435–2436.

Pelizzola, M. and Ecker, J. R. (2011) The DNA methylome. *FEBS Lett*, **585**, 1994–2000.

Peng, Q. and Smith, A. D. (2011) Multiple sequence assembly from reads alignable to a common reference genome. *IEEE/ACM Trans Comput Biol Bioinform*, **8**, 1283–1295.

Rando, O. J. (2012) Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev*, **22**, 148–155.

Razin, A. and Riggs, A. D. (1980) DNA methylation and gene function. *Science*, **210**, 604–610.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Roudier, F., Ahmed, I., Brard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Desprs, B., Drevensek, S., Barneche, F., Drozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M. and Colot, V. (2011) Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J*, **30**, 1928–1938.

Rozhon, W., Baubec, T., Mayerhofer, J., Mittelsten Scheid, O. and Jonak, C. (2008) Rapid quantification of global DNA methylation by isocratic cation exchange high-performance liquid chromatography. *Anal Biochem*, **375**, 354–360.

Russo, V., Martienssen, R. and Riggs, A. (1996) Epigenetic mechanisms of gene regulation. *Cold Spring Harbor Monograph Archive*, **32**.

Satterlee, J. S., Schuebeler, D. and Ng, H.-H. (2010) Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*, **28**, 1039–1044.

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

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schlkopf, B., Weigel, D. and Lohmann, J. U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet*, **37**, 501–506.

Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097–6100.

Sedlazeck, F. J., Rescheneder, P., Tauber, S. and von Heaseler, A. (2012) NextGenMap: High throughput mapping for high throughput sequencing. *submitted*.

Selker, E. U., Tountas, N. A., Cross, S. H., Margolin, B. S., Murphy, J. G., Bird, A. P. and Freitag, M. (2003) The methylated component of the neurospora crassa genome. *Nature*, **422**, 893–897.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol*, **26**, 1135–1145.

Smith, A. D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z. and Zhang, M. Q. (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**, 2841–2842.

Solomon, M. J., Larsen, P. L. and Varshavsky, A. (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.

van Steensel, B. (2011) Chromatin: constructing the big picture. *EMBO J*, **30**, 1885–1895.

Strahl, B. D. and Allis, C. D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

Talbert, P. B. and Henikoff, S. (2010) Histone variants–ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol*, **11**, 264–275.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.-L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R. A., Coupland, G. and Colot, V. (2007) Arabidopsis tfl2/lhp1 specifically associates with genes marked by trimethylation of histone h3 lysine 27. *PLoS Genet*, **3**, e86.

Waddington, C. (1942) The epigenotype. *Endeavour*, **1**, 18–20.

Wang, R. Y., Gehrke, C. W. and Ehrlich, M. (1980) Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res*, **8**, 4777–4790.

Xi, Y. and Li, W. (2009) BSMAP: whole genome bisulfite sequence mapping program. *BMC Bioinformatics*, **10**, 232.

Zemach, A., McDaniel, I. E., Silva, P. and Zilberman, D. (2010) Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, **328**, 916–919.

Zhang, M. Q. and Smith, A. D. (2010) Challenges in understanding genome-wide DNA methylation. *Journal of Computer Science and Technology*, **25(1)**, 2634.