The Dynamic Epigenome
Analysis of the Distribution of Histone Modifications

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)
im Fachgebiet
Informatik

vorgelegt
von Diplom-Informatikerin Lydia Steiner
egenom am 15. Mai 1986 in Leipzig

Die Annahme der Dissertation wurde empfohlen von:
1. Jun.-Prof. Dr. Sonja J. Prohaska, Institut für Informatik, Leipzig
2. Prof. Dr. Andrew Torda, Zentrum für Bioinformatik, Hamburg

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 27. Juni 2013 mit dem Gesamtprädikat 'magna cum laude'
Abstract

There is a genome in a cell, as everyone knows, but there is also an epigenome. The epigenome regulates the transcription of the underlying genome. In the last decade, it was discovered that the epigenome state and its regulation are important for differentiation and development. Correlation studies with aging samples had led to the hypothesis that misregulation of the epigenome causes aging and cancer. Furthermore, diseases were identified which are caused by errors in the epigenome state and its regulation.

Identification of erroneous epigenome states and misregulation requires the prior knowledge of the common state. Several studies aim at measuring epigenome states in different organisms and cell types and thus, provide a huge amount of data.

In this dissertation, a pipeline is developed to analyze and characterize histone modifications with respect to different cell types. Application of this pipeline is shown for a published data set of mouse consisting of data for H3K4me3, H3K27me3, and H3K9me3 measured in embryonic stem cells, embryonic fibroblasts and neuronal progenitors.

Furthermore, methods for the detection of the epigenetic patterns are presented in this dissertation. Therefore, a segmentation method is developed to segment the genome guided by the data sets. Based on this segmentation, the epigenome states as well as epigenetic variation can be studied. Different visualization methods are developed to highlight the epigenetic patterns in the segmentation data. Application of the segmentation AND visualization methods to the mouse data set had resulted in not only colorful squares but also in biological conclusions! It demonstrate the power of the developed methods.

Although the studied data set in this dissertation contains only ordinary tissue cells, the methods are not restricted to study the reference epigenome state. Comparison of normal and disease cells as well as comparison with aged cells are possible with all of the methods.

Finally, the methods are compared based on the obtained results. It shows that all methods highlight different aspects of the data. Thus, applying all methods to the same data sets, deep insights into the epigenome in murine embryonic stem cells, embryonic fibroblasts and neuronal progenitor cells are gained. For example, it had been found that several mechanisms exist setting H3K4me3 marks. Furthermore, not all mechanisms are found in all cell types. Strong evidence
had been found that catalysis of H3K4me3 and H3K27me3 is coupled.
Acknowledgment

Let me thank my family and friends for your great support in the last years. Particularly, thanks goes to Erik. You always understood long working days of thesis writing and you spoiled me during my thesis compilation. Thanks to my mom for having everlasting faith in me!

Thanks to Sonja: for the support, hands-on help, and faith. Thanks to Peter: for your supervision particularly in the last half a year.

Many thanks to Jens, Petra, and Corinna: for your assistance in bureaucracy and system administration. Particularly, many thanks for the chocolate.

Thanks to all who proofread my thesis.

I thank also Irma, Axel, Christian, Tobi, and the whole bioinformatics group.

Many thanks to Dirk: I honor you as a collaborator; working with you is fun. Thanks also to Sarah and Daniel for doing their bachelor theses at the junction of bioinformatics and visualization!

Thanks to Jörg, Jens, and Thimo.

Thanks to the BMBF and MAGE for funding my research.
Contents

Abbreviations ix

1 One Genotype but many Phenotypes 5
  1.1 The Modification Zoo ........................................ 7
  1.2 Origin of Data .................................................. 11

2 Hallmarks of Pluripotency 15
  2.1 Histone Marks and Modifying Complexes ....................... 18
  2.2 Where have all the CpGs gone? ............................... 21

3 Genome-wide Distribution of Epigenetic Marks 25
  3.1 Short Reads to Regions ......................................... 26
  3.2 Calculating Length- and CpG-Density-Distribution .......... 29
  3.3 Calculation of Overlaps between Epigenetic Marks .......... 30

4 Distribution of Epigenetic Marks in Mouse 33
  4.1 Epigenetic marks in ESCs ...................................... 34
     4.1.1 Length Distributions of the ESC data .................. 35
     4.1.2 Outliers of the Length Distributions .................. 37
     4.1.3 Correlation with CpG-Density ........................... 39
     4.1.4 Overlaps between Regions of Different Modifications .. 40
  4.2 Distribution of H3K4me3 in ESCs, MEFs, and NPCs .......... 42
     4.2.1 Length Distribution of H3K4me3 ......................... 42
     4.2.2 Changes in the CpG-Dependency .......................... 43
     4.2.3 Retention of H3K4me3 Modified Regions ................. 44
  4.3 Summary ......................................................... 48
5 Discovering Epigenetic Patterns 49
5.1 Finding a Reference 49
5.2 Visualization and Clustering by SOMs 52
5.3 Scatterplot Matrices 54
5.4 k-means++ Clustering and Star Plots 56

6 Pattern Discovery in Murine ES-Segmentation Data 59
6.1 SOMery 60
6.2 Scatterplot Matrix 65
6.3 k-means++ Clustering and Starplots 70

7 Patterns found in Murine EV-Segmentation Data 77
7.1 SOMery 77
7.2 Scatterplot Matrix 81
7.3 k-means++ Clustering and Starplots 86

8 Who is the Best? 91
8.1 Either SOMs or Scatterplot Matrices or Starplots! 92
8.2 Nobody! 96

9 Conclusions 99

 Appendices 107
A Long Modified Regions in Mouse ESCs 109

B Technical Details 119
B.1 SOM 119
B.2 Scatterplot Matrix 120
B.3 k-means, k-means++, and Starplots 121

C Additional SOM images for ES-Segmentation 123

D Scatterplot Matrices for ES-segmentation 127

E Starplots for the ES-segmentation 131

F Additional SOM images for EV-Segmentation 137

G Scatterplot Matrices for EV-segmentation 141

H Starplots for the EV-segmentation 145
Abbreviations

5mC  . . . . . . . . . . . . . . . . . . . . . . . . . . 5-methylcytosine
ADP-ribose  . . . . . . . . . . . . . . . . . . . . adenosine diphosphate ribose
bp  . . . . . . . . . . . . . . . . . . . . . . . . . . base pairs
C  . . . . . . . . . . . . . . . . . . . . . . . . . . Cystosine
CAF-1  . . . . . . . . . . . . . . . . . . . . . . . . . chromatin assembly factor 1
cDNA  . . . . . . . . . . . . . . . . . . . . . . . . complemented DNA
CEP  . . . . . . . . . . . . . . . . . . . . . . . . . combinatorial epigenetic profile
ChIP  . . . . . . . . . . . . . . . . . . . . . . . . . chromatin immunoprecipitation
ChIP-chip  . . . . . . . . . . . . . . . . . . . . . chromatin immunoprecipitation in combination with chip analysis
ChIP-seq  . . . . . . . . . . . . . . . . . . . . . chromatin immunoprecipitation in combination with sequencing
COMPASS  . . . . . . Complex of Proteins Associated with Set1
CpG  . . . . . . . . . . . . . . . . . . . . . . . . . cystosine followed by guanine
DNA  . . . . . . . . . . . . . . . . . . . . . . . . . desoxyribonucleic acid
Eed  . . . . . . . . . . . . . . . . . . . . . . . . . embryonic ectoderm development
EP  . . . . . . . . . . . . . . . . . . . . . . . . . epigenetic profile
ES  . . . . . . . . . . . . . . . . . . . . . . . . . epigenetic state
ESC  . . . . . . . . . . . . . . . . . . . . . . . . . embryonic stem cell
EV  . . . . . . . . . . . . . . . . . . . . . . . . . epigenetic variation
Ezh1  . . . . . . . . . . . . . . . . . . . . . . . . . enhancer of zeste 1
Ezh2  . . . . . . . . . . . . . . . . . . . . . . . . . enhancer of zeste 2
H19  . . . . . . . . . . . . . . . . . . . . . . . . . a imprinted maternally expressed transcript, long non-coding RNA
H2BK119  . . . . . . lysine at position 119 on the histone H2B
H2BK119ub  . . . . . . ubiquitination of the lysine at position 119 on the histone H2B
H3K27me3  . . . . . . trimethylation at histone H3 at lysine at position 27
H3K4me3 ................................ trimethylation at histone H3 at lysine at position 4
HEP ........................................ Human Epigenome Project
HP1 ........................................ heterochromatin protein 1
Igf2 ......................................... insulin-like growth factor 2
iPSC ......................................... induced pluripotent stem cell
MBD1 ....................................... Methyl-CpG binding domain 1 protein
MEF ........................................ murine embryonic fibroblast
MS ............................................. modification state
NPC ........................................ neuronal progenitor cell
nt ............................................. nucleotide(s) as length unit
PcG .......................................... Polycomb Group Proteins
PCR .......................................... polymerase chain reaction
PRC1 ........................................ Polycomb repressing complex 1
PRC2 ........................................ Polycomb repressive complex 2
RB1 .......................................... retinoblastoma 1
RBBP4 ...................................... retinoblast binding protein 4
SETdb1 ..................................... SET domain, bifurcated 1
SETdb2 ..................................... SET domain, bifurcated 2
SNP .......................................... single nucleotide polymorphism
SOM ......................................... self-organizing map
SUMO ....................................... small ubiquitin-like modifier
Suv39h1 ..................................... Suppressor of variegation 3-9 homolog 1
Suv39h2 ..................................... Suppressor of variegation 3-9 homolog 2
Suz12 ...................................... suppressor of zeste 12
TrxG .......................................... Trithorax Group Proteins
WCE .......................................... whole cell extract
WGS .......................................... whole genome segmentation
Different environments strongly influence the phenotype of every individual but also of every cell. One key role player in this process is the epigenome. It is a dynamic system that regulates gene expression and models changes in the environment. Furthermore, the epigenome functions as a memory for the cells state. The epigenome mainly acts on histone modifications which can be set and unset. Modification specific writers and erasers exist. Readers enable the interaction of the marks. Thus, it is a very dynamic system which is organized in cell type specific patterns.

Detection and analysis of the patterns shed light on the dynamics of the epigenome state. Knowing the epigenome state of the normal cells enables the comparison of the epigenome states of normal and disease tissues or normal and aged tissue. Therefore, it may reveal specific epigenetic misregulation causing epigenetic-related and age-related diseases.

Also for regenerative medicine, the epigenome state is important. Reprogramming of somatic cells often result in induced pluripotent stem cells which retain part of the epigenetic memory of the donor cells. Thus, analysis of the patterns highlights the differences between induced pluripotent and embryonic stem cells. Differences in the regulation may reveal target pathway which have to be activated or switched off to better fit the epigenome states of induced pluripotent to embryonic stem cells.

An introduction into epigenetic as historical review is given in chapter 1. At this, the basic concept and minimal units of the epigenome are described. The structure of a nucleosome is explained as well as the set of modifications which can be attached to a nucleosome. Correlations between the modifications and the transcriptional regulation and chromatin structure are described. The chapter ends with a description of ChIP-seq; a method to measure the genomic localization of histone marks. Here, not only the method but also the sources for errors are described.

Chapter 2 introduces into the concept of pluripotency and differentiation. Histone modifications which are prominent for regulating differentiation in early developmental stages and defining stem cell identity are named. The current knowledge about their modifying complexes and their dependencies and interactions are described. Finally, the connection to CpGs and DNA methylation is explained in more detail motivating the further analysis of the CpG-dependency of marks.

In the first part of the thesis, the analysis of a single modification and its characteristics is in the fore. It aims at analyzing the distribution of epigenetic marks in embryonic stem cells and multipotent cells. It is searched for qualitative and quantitative characteristics of different epigenetic marks in the same cell types. Description of the differences between the distribution
of the same modifications in different cell types will be generated referring to quantitative and qualitative descriptions as well. A model will be used to examine if changes can be simulated changing the parameters of the model. Thus, insights in the causes of the changes is gained.

For this reason, chapter 3 explains how ChIP-seq is analyzed to obtain the characteristics for a single modifications data sets. The first part describes the analysis of ChIP-seq read data to obtain the genomic regions which are modified with a specific mark. The second part explains how the data for the length distributions and the CpG-dependency-analysis is gathered. The analysis of the data obtained is described. The last part explains the calculation of the overlap of two data sets. The further analysis and visualization of this data is described.

In chapter 4, three modifications in murine embryonic stem cells, embryonic fibroblasts, and neuronal progenitor cells are analyzed as described in chapter 3. It is divided into two parts. The first part analyzes differences between the different modifications in ESCs. Characteristics in the length distributions are shown. Common features and differences of the length distributions are highlighted. Outlier of the length distribution are examined in detail. The CpG-dependency reported in literature is confirmed for all three marks and are furthermore described by the CpG-densities characteristic for the different modifications. Thus, the description of the CpG-dependency can be significantly improved. Overlaps are analyzed to describe the co-localization in detail and the connections between length characteristics, CpG-density, and combination of marks are shown. The second part of chapter 4, investigates the distribution of H3K4me3 in the three cell types. Changes in the length distribution highlight major changes in the recruitment of the marks. A model fitted to the data shows which changes in the parameter setting lead to best fits of the models simulated length distribution with the measured length distribution. Differences in the length distributions are complemented with changes observed in the CpG-dependencies. Using also the overlap data, changes in the distributions of H3K4me3 can be described in detail and allows for hypothesis about number of mechanisms and activity of mechanisms for H3K4 trimethylation.

This first part of the thesis, namely the analysis and characterization of a single modification, is mainly based on the following two publications:


In the second part of the thesis, analysis of epigenetic patterns consisting of the modifications and cell types studied in the first part is in the fore. It aims not only at the discovery of patterns in the distributions of the marks but also at the characterization and correlation with additional information from different sources. Again, length and CpG-density are used for characterization.
Furthermore, these data is complemented with gene annotations, chromosomal localization, and gene expression data. Changes in the patterns formed by the marks during differentiation and connections between the patterns formed by the same modifications in different cell types are analyzed and described.

For this purpose, chapter 5 introduces a segmentation method. Only this method enables efficient analysis of the patterns. Either the epigenome state or the variation of the distribution of one mark is highlighted. Visualization methods are applied to the segmentation data. While the methods itself are not new, the combinations and adaptations are a result of this thesis. Thus, technical details are described in appendix B.

Using the data produced with the methods from chapter 3, the data driven segmentation of the genome is performed. In this way, two segmentation data sets are produced. While the first data set highlights the epigenome state, the second enables analysis of the variation of the modifications. In chapter 6, the first segmentation data set is analyzed using the methods described in chapter 5. Not only the modification data itself but also additional data such as gene expression, chromosome position, CpG-density, or length are incorporated into the analysis. For each visualization, it is shown which results can be obtained from the application of the method. Biological conclusions are drawn. The second segmentation data set is analyzed in chapter 7. Also here, additional information is used during detection and analysis of the patterns. Valuable insights are gained into the co-localization of the marks and the changes during differentiation.

In chapter 8, the visualization methods are compared with each other. It focuses on answering the question whether an standard method can be named. Thus, comparison is based on the results obtained in chapter 6 and chapter 7. It is highlighted which results are obtained by which method. Patterns discovered with all methods are shown as well as pattern observable with only one method.

This second part of the thesis, namely the analysis and characterization of patterns, is based on the following two publications:


**Starplot Visualization for Analyzing Chromatin:** Dirk Zeckzer, Sarah Seifert, Lydia Steiner, Sonja J. Prohaska, *International Journal of the Eurographics Association*, submitted

Chapter 9 summarizes the biological results and draws conclusion for mechanisms leading to the observed patterns and their characteristics. Finally, the results of the thesis are summarized and a short outlook on further researched based on the thesis results is given.
One Genotype but many Phenotypes

The human being consists of trillions of cells of several hundreds of different cell types. Each of them carries the same genome with the same set of genes. However, there exist regulatory mechanisms which lead to different expression patterns and thus, to different cell types. The discovery of this mechanisms is illustrated in the following.

The DNA was already discovered in 1869 by F. Miescher. However, he did not identified the function of the substance which he called “nuclein”. Not until 1944, O. Avery characterized DNA as information carrier [8]. In 1952, A. Hershey and M. Chase proved that DNA contains the genetic information of an organism [61].

Independent of the discovery of the DNA, G. Mendel had shown the inheritance of traits from parents to their offspring in 1854. Responsible for the inheritance of the traits are biological entities. In 1909, W. Johannsen called this entities genes [69]. Accompanied with the discovery of the DNA, T. H. Morgan found that genes are located on the chromosomes [103]. W. Johannsen introduced also the terms genotype and phenotype [69]. While the genotype describes the heritable information, the phenotype denote the set of observable traits.

It was noticed that all cells of one human being have the same genotype. However, there are many cell types with different characteristics such as shape, size, and function. Thus, many
phenotypes are produced by one genotype by varying the expression patterns of the genotype. The phenotype which is produced depends on its environment. As long ago as 1942, C.H. Weddington introduced the term “epigenetics”. It designates the study of the processes by which the genotype get in contact with its environment to produce different phenotypes [162]. More than 50 years later, the definition was changed. R. Holliday proposed that epigenetics are the heritable changes of the gene expression that are not based on changes in the DNA sequence [64]. During cell division, not only the genotype but also the epigenotype is inherited. This is very important to retain the cell fate. Imaging differentiated cells such as fibroblasts or lymphocytes, they have specific epigenotypes. Genes are silenced which are expressed in other cell types, e.g. in embryonic stem cells, while fibroblast-specific or lymphocyte-specific genes are expressed. During proliferation, the epigenotype has to be inherited to retain the cell type of the cells.

Although the concept of epigenetics is already fairly old, only a little is known about the mechanisms responsible for the epigenetic phenomena. One mechanism is DNA-methylation. A cytosine (C) is methylated at the fifth carbon atom. The resulting nucleotid is called 5-methylcytosine (5mC). DNA-methylation was discovered in 1975 by Riggs [124] as well as Holliday and Pugh [65]. They found that DNA methylation leads to inactivation of one X chromosome. However, the phenomena of X inactivation itself was already known in 1959 [108]. Rings, Holliday, and Pugh proposed that DNA methylation is symmetric on both DNA strands. Once one strand gets methylated, the other strand will be methylated as well. Thus, the methylation mark can be transmitted to the next generation. Later, Bird found that CpG sites (cytosine followed by guanine) are either methylated or unmethylated but never semimethylated [21, 22]. DNA-methylation was not only found on the X chromosome. Repetitive and retroviral sequences are often target of DNA methylation. Thus, DNA-methylation also functions as protection against foreign DNA. Another phenomenon is imprinting. At it, either the DNA of the paternal or maternal allele is methylated and thus silenced. It was demonstrated in 1985 for the Igf2/H19 loci by Cattanach and Kirk [34]. In section 2.2, the mechanisms and inheritance of DNA methylation is explain in more detail.

Another epigenetic mechanism regards the histone complex. It consists of 8 proteins of 4 different types. These types are histone H2A, H2B, H3, and H4 [38]. The DNA is wrapped around this complex 1.6 times. This are approximately 146bp. Histones and the DNA together form the nucleosome. Two nucleosomes are separated by approximately 80bp called linker. The existence of histones is known for a long time. However, their function remains unraveled a long time. In 1950, it was proposed that histones work as gene repressors [143]. The repressive function of histones in vitro was shown in 1963 [23]. Furthermore, it was shown that removing the histones enables transcription. At that time, it was not known how the histones are removed or attached to DNA as well as the signals for both events. One year later, in 1964, it was discovered that histones can be posttranslationally methylated and acetylated [5]. While acetylation inhibit the repressive function of histones, methylation does not. Thus, transcription can take place without taking
of the histones as long as they are acetylated. Neither the mechanisms itself nor the acetylation or methylation sites were known. Before 1974, it was thought that the histone complex is a single protein. In 1974, the heterooctameric structure was discovered. Additionally, a fifth type was found which is optionally bound to the complex. More precisely, a nucleosome contains a heterotetramere consisting of two copies of each H3 and H4, respectively, and heterodimers of H2A and H2B [80]. The crystal structure of a nucleosome was determined in 1984 [123].

![Figure 1.1: 3D-Structure of a Nucleosome](image)

Figure 1.1: 3D-Structure of a Nucleosome: Colors indicate different molecules in the complex; blue: DNA, yellow: H2A, orange: H2B, green: H3, and red: H4. l.h.s.: X-ray structure of the nucleosome core particle at a resolution of 2.5 Å [57]. r.h.s.: schematic depiction of a nucleosome.

### 1.1 The Modification Zoo

In 1991, it was shown that removing the N-terminal sequence (also called tail) of histone H4 affects promoter activity in yeast dramatically. Especially, deletion of one arginine residue has effect on promoter activity. Normally, this residue can be reversibly acetylated and thus, may function as repressor and activator for gene expression [45]. The biochemical mechanism remained unclear. This changed five years later. Brownell et al. studied the histone acetylase A in *Tetrahymena*. They observed that it is a homolog of the yeast protein Gcn5p. A subset of transcriptional activators in yeast require Gcn5p for their full activity. Thus, Gcn5p was already linked to gene activation, while the histone acetylase A provides the link to acetylation of histone. Brownell et al. had shown that also Gcn5p has histone acetyltransferase activity. They concluded that complexes of Gcn5p link histone acetylation with gene activation [26].

In the more recent past, more and more types of modifications and residues that are modified were discovered [16, 81, 115, 150]. The different effects of different marks and combination of marks were studied highlighting the impact on differentiation, development and aging [17, 12,
163, 170, 101]. Modifying and demodifying complexes [100, 33, 47] as well as complexes reading
the marks on the histones [11, 173, 134] were discovered. Their interplay with regulators, such
as RNAs, was studied [155, 132]. Theories and models for chromatin mechanisms are proposed
[44, 118, 20]. Coupling of histone modifications and transcription is used to predict the gene
expression level based on the modification state of the promoter [74].

In Figure 1.2, the methylation, acetylation, phosphorylation, and ubiquitination sites at the C-
and N-terminal tails of the histones of the nucleosome core particle are shown. Further modifications
are, e.g., crotonylation, SUMOylation, and ADP ribosylation. Most of the modification can
be found at the tails, only a few are found in the globular domains of the histones. The location
and the type of modifications are used to reference to the different modifications. For example,
the acetylation at histone H4 at the lysine (one letter code K) at position 5 is named H4K5ac.
Similar, H3S28ph refers to the phosphorylation at the serine (one letter code S) at position 28 on
histone H3. Not every residue can be modified. Modifications are only attached to a small set of
amino acids. For example, lysine can be mono-, di-, trimethylated, acetylated, ubiquitinated, and
modified by other marks. Phosphorylation occur at, e.g., threonine.

Methyl-groups can be attached to lysine and arginine. Mono- and dimethylations are possible for
both, lysines and arginines. However, at arginine, dimethylation can be set either symmetrically or
asymmetrically (see Figure 1.3). A third methyl-group can be set at lysine leading to the formation
of trimethyl-lysine. The functions of the methylation marks depend on the histone, site, and

Figure 1.2: Histone Modifications: Methylation, acetylation, phosphorylation, and ubiquitination sites identified at
the C- and N-terminal tails of the histones of the nucleosome core particle. Only the modifications at histone H3
at position 56 and position 79 locate to the globular domain. Figure taken from Bhaumik et al. [18]
number of methyl-groups. Probably, asymmetric and symmetric dimethylation at the arginines do also impact the function of the marks. For example, H3K4me3 correlates with transcriptional activity, H3K27me3 correlates with formation of heterochromatin. Methylation at arginine have roles in, e.g., mRNA splicing, DNA damage, signal transduction [13]. In this thesis, three specific methylations are examined. The current knowledge about the function and modifying complexes for these modifications are reviewed in chapter 2.

![Figure 1.3: Arginine Methylation: unmethylated arginine, N\textsuperscript{G}-monomethyl L-arginine (L-NMMA), asymmetric dimethyl-arginine (ADMA), and symmetric dimethyl-arginine (SDMA). Figure taken from Mangoni [92].](image)

Acetylations can only be found at lysines. In contrast to methylation, only one acetyl-group can be attached to the residues. Doing so, the positively charged amino acid is neutralized. This reduces the affinity of the histone to the negatively charged DNA and increases the accessibility of the DNA [83]. Additionally, many complexes exist which read acetylations and recruit the RNA polymerase complex or enhance elongation of the mRNA [83]. Most often such readers were identified as part of the RNA polymerase complex before there ability to read acetylation marks was noticed. Moreover, the RNA polymerase complex contains several components able to acetylate histones [146]. Thus, acetylation is set and read by the RNA polymerase complex and thus, has a positive effect on transcription. Changes of the acetylation level are shown to be predictive for changes in expression level [95]. Here, no specific acetylation was measured but the overall level of acetylation marks in a specific region. This shows that acetylation has a strong direct effect on transcription. Nevertheless, many examples have been discovered showing complex patterns of acetylations without direct effect, for example H3K9ac and H3K14ac [51]. Here, the effect is carried out by reader proteins.

The amino acids serine, threonine, and tyrosine can be phosphorylated. It is known that phosphorylation can change the protein structure. However, it is still unknown if changes of the
structure count to the effects of histone phosphorylations. Moreover, many readers are identified for phosphorylation marks [10]. Phosphorylation is linked to many different functions including transcriptional control, chromatin structure, apoptosis, or DNA repair [10]. For example, H2AT119ph and H3S10ph are associated with chromatin structure formation during mitosis [42]. In the most cases, phosphorylation does not function alone but forms pattern with other marks [10].

Ubiquitin is a 76 amino acid large protein which can be attached to internal lysines of the histones. Since ubiquitin has internal lysines, ubiquitin can be attached to itself. In theory, any number of ubiquitin molecules could be found at the histones. However, in vivo mostly only monoubiquitination is observed [25]. The function of ubiquitination is diverse and not fully understood so far. Monoubiquitination at H2BK120 is found at the promoter of actively transcribed genes. Here, ubiquitination is directed by transcription activators [25]. It is proposed that ubiquitination may function as a check point for elongation of the transcripts [164]. Ubiquitination at H2AK119 is found to recruit the heterochromatin protein 1 leading to the formation of heterochromatin. This process is activated by trimethylation of H3K27 [25]. Not only here but in general ubiquitination shows strong interaction with other histone marks [25].

Likewise ubiquitin, SUMO (small ubiquitin-like modifier) is attached to lysines. SUMO is a protein 100 amino acid in size which exists in different isoforms. There is not much know about the function of SUMOylation at histones. However, it was shown that SUMOylations at the histone variant H2AX are important for the DNA repair mechanism [36]. Misregulation may lead to diseases like rheumatic arthritis [78].

Adenosine diphosphate ribose (ADP-ribose) is found to be attached to lysines in the histones tail of all three histones [66]. Histones can be mono- and poly-ADP ribosylated [98]. Little is known about the function of these marks and how their effect is carried out. Nevertheless, it was found that ADP ribosylation is activated due to environmental stress [31]. Furthermore, it is a component of the histone code, i.e. the interaction and combinatorial effect of histone marks [66].

Citrullination describes the conversion of arginine into citrulline within a protein. As most of the histone marks, it is correlated with transcription, replication, and DNA repair [140] and is associated with diseases [153]. It was found that it contributes to transcriptional repression mediated by the heterochromatin protein 1. However, the function of the modification is currently unknown [136].

Crotonylation is the attachment of crotonyl-groups to lysines and was only recently identified as a histone mark [89]. While it remains unknown how this mark is set, it was found that it marks genes important for post-meiotic processes of male germ cells [101].

The described modifications are not the full set of histone marks. Many more were discovered in the recent past [89]. However, it will take much more time to discover their functions and interactions with the other histone marks. Most of the modifications are not found frequently in the genome and usually occupy just a small fraction of the histones at all. More frequent
modifications such as acetylations and methylations are understood in more detail than infrequent modifications such as crotonylation and citrullination.

Effects of histone modifications are often not carried out by just one modification but by a combination of marks. An example was already mentioned where H3K27me3 recruits an ubiquitin transferase which lead to H2AK119ub. H2AK119ub serves as binding site for the heterochromatin protein 1 which is leads to the formation of heterochromatin and thus, to gene silencing. Thus, histone modification form complex pattern which lead to combinatorial effects. This observation results in formulation of the histone code hypothesis, i.e. that functions are assigned to patterns of marks rather than to a single mark.

It is observed that the epigenome state and thus, the histone modifications are passed on from the parent cell to the two daughter cells. The DNA is replicated and the histone with their marks are distributed randomly to the two daughter strands. Thus, on average half of the information on the epigenome state is passed on to each of the daughter cells. After replications, the full information has to be restored. This requires a mechanism to read the pattern and set marks corresponding to the read pattern. By this, an epigenetic memory is established and retained over several cell divisions. Epigenetic memory is subject to many models. Nevertheless, the system has to be dynamic to adapt to different environments and enable differentiation into different cell types.

Although much is known about the common modifications, enzymes and possible regulators, only little is known about the mechanisms behind. Already with the current knowledge, it is obvious that epigenetics play an important role in differentiation and development. Adaptation to different environments seems to be impossible without epigenetic regulation. Epigenetic patterns have a great influence on the regulation of the epigenome. Nevertheless, it is unknown which patterns exist and how they change over time. Thus, it is worthwhile to study the properties of epigenetic marks and epigenetic patterns.

### 1.2 Origin of Data

To study the epigenetic marks, several experiments can be performed to measure the epigenetic marks in vivo and in vitro. The set of marks on histones can be identified by mass spectrometry. This method provides information about the kind of the modification and its position within the histone. Several modifications at several sites were detected using this method, e.g. methylation, acetylation and recently crotonylation. The distribution of such marks can not be determined by mass spectrometry.

Different methods were developed to study the distribution of histone marks. All methods provide information of the genomic position of the measured mark. This means that one obtain the genomic position of the histone carrying the mark in different resolutions. Histone marks and also DNA-bound proteins, are usually measured using Chromatin Immunoprecipitation (ChIP)
followed by either sequencing (ChIP-seq) or chip analysis (ChIP-chip) of the obtain DNA (see Figure 1.4).

Chromatin immunoprecipitation is an antibody-based technique. It cannot be performed on a single cell but on a cell population to obtain a sufficient amount of chromatin. In a first step, the cells are treated with formaldehyde to fixate the DNA and all bound proteins, e.g. histones or modifying complexes. Thus, the bound proteins cannot be removed from the chromatin anymore. The cells are broken up and the chromatin is extracted [141].

The chromatin is purified and sonicated to produce small fragments. Depending on the sonication procedure the fragment size can vary. Usually, the size ranges from 200bp to 700bp. Thus, it contains approximately 1 to 4 nucleosomes. Affinity-purified antibodies are used to select fragments with a specific protein [112]. In the case of histone modifications, the antibodies are designed to detect a mark, e.g. trimethylation at histone H3 lysine 4. Antibodies detecting different histone variants exist as well. Such antibodies are also designed against different proteins. Using them, DNA-bound transcription factors or DNA-bound proteins of modifying complexes can be detected.

After selecting the fragments by antibodies, the fixation is reversed and the DNA is purified [141]. After ChIP, one ends up with purified DNA which was bound to the protein of interest. The DNA can now be used to determine the genomic position of these fragments.

Two different methods are used to find the fragments position within the genome, chip analysis and sequencing. Chip analysis uses DNA microarrays with probes for thousands for genomic loci.
At this, one genomic loci is represented by several probes. The fragments are broken up into even smaller fragments and are fluorescently labeled. The smaller fragments are hybridized on the microarray. Unbound fragments are washed away. The strength of the fluorescent signal of a spot provides as measurement of amount of bound fragments to the probes. The signals are normalized and further analyzed to reduce the effect of, e.g., unspecific binding of fragments. For each genomic locus, a signal is generated from all probes for this locus. The resolution of this method is restricted to the resolution of the microarray used. The exact match of the fragment and the genomic loci measured cannot be determined. One microarray usually does not provide information of the whole genome but on a subset of genomic loci which can be discriminated by hybridization of sequences to small probes. One has to make compromise between signal quality and amount genome which can be identified.

The next generation sequencing techniques enables cheap and fast sequencing of huge amount of DNA. Thus, ChIP was also combined with sequencing techniques [12, 70, 99, 125]. The DNA obtained by ChIP is used as input for the library preparation. To the end of the fragments, sequencing adapters are ligated. Fragments with adapters are purified and afterwards enriched by PCR. The sequencing produces reads with length of $26 - 36\text{bp}$ covering the fragments. The reads can be longer for different protocols and different applications. Thus, one experiment outputs millions of sequencing reads. These reads have to mapped back onto the genome of origin.

ChIP-seq is a method which allows to measure histone marks and other signals genome-wide and on single base resolution. Nevertheless, there are limitations arising from experimental procedure as well as the subsequent analysis of the sequencing reads.

Formaldehyde fixates all DNA-bound proteins at the DNA as long as they are within a maximal distance. From this fact arise two problems. On the one hand, not only histones are fixated but also all other bound proteins such as polymerase or transcription factors or histone modifying complexes. Thus, histone bound to other proteins can not be accessed by the antibodies. As a consequence, these histone are not selected. On the other hand, nucleosomes which are bound very weakly or currently on exchange may be to far away from DNA to be fixated by formaldehyde. Thus, they are washed away and the DNA fragments can not be selected by antibodies because no histone which carries the modification is present. Such exchanges occur, for example, at highly transcribed genes.

The position of nucleosomes within the genome is not fixed. They slide a little bit along the DNA. Thus, within different cells of the same cell population, a specific genomic position is only occasional occupied by a histone. As a consequence, ChIP-seq signals may look fuzzy. The actual start and end of the signal has to be defined during the analysis of the signals.

The epigenetic system is very dynamic. New marks are set while other marks are removed. With ChIP-seq, a snapshot of the epigenome is made. Due to the fact that this is made on the base of whole cell populations, the signal represents a population wide average. Virtually, one obtain strong signals from sites which are modified in many cells and weak signals form sites which only
modified in some cells. While strong signals arise from more stably modified sites, weak signals arise from sites which are only occasional in a modified state. Thus, during the analysis of the ChIP-seq signals, the decision has to be made which ChIP-seq signal strength has to be reached to be regarded as modified.

The antibodies used in the experiment are affinity-purified to detect a specific signal. However, this might not work perfectly. It cannot be neglected that antibodies unspecifically bind to histones or other DNA-bound proteins. It is also the case that chromatin not bound to antibodies might not be perfectly washed away. Thus, there are some false positive reads. During analysis, reads or read counts have to be filtered to retain only genomic position which are associated with the measured mark.

During sequencing library preparation, adapters are ligated to the end of the DNA fragment. Depending on the fragments DNA sequence and the adapter used, the ligation occurs with different rates. It is also reported that PCR produces errors which are base-specific. For example, the most often occurring error for polymerase Taq is replacement of A and T with C and G [76]. Also the sequencing and PCR primer bind some sequences better than others. Thus, some sequences are more often copied than sequences with lower affinity to the primer. Combined with errors within the sequencing procedure itself, one has to deal with missing and duplicated reads and fragments as well as erroneous reads. Mapping back these set of reads to the genome, one may see enrichment of reads where no marks were enriched.

In addition, mapping produces additional errors. The major source are repeats. Here, the mapping is not unique and depending on the strategy, reads are discarded because of an ambiguous mapping. It also occurs that reads originate from genomic positions which cannot be positioned within the genome so far. Thus, they are ever mapped to the false position or they are unmapped.

An additional source of errors is the fact that the individual used in the experiment is not the same individual used to create the reference genome. Thus, the sequences of the two genomes, the reference genome and the genome of the individual used in the experiment, differ. Even if the experiment would not produce errors, not all reads can matched with 100% identity to the reference genome due to single nucleotide polymorphisms, inversions of parts of the genomic sequence, insertions, or deletions.

Summarizing the effects, one deals with multiple errors from different sources. In the further analysis, reads counts have to be corrected. Since there are several position dependent errors and sequence specific errors, correction has to take these effects into account and thus, correct with respect to the position. At this, sequencing of the whole cell extract (WCE) used as input for the selection by antibodies step, provides a position-wise estimation of such effects.
Embryonic stem cells are a very interesting cell type and thus, are subject to many research projects. The high interest arises from two properties of embryonic stem cells. The first property is **pluripotency**. They are able to differentiate into any cell type. Because of that, embryonic stem cells are especially interesting for regenerative medicine. The second property is the immortality. Embryonic stem cells are able to replicate indefinitely. Taken together, from one embryonic stem cell one can generate as much cells as needed and differentiate them into the cell type required. This opens up many possibilities for research and therapy. However, embryonic stem cells have one large drawback: the only single natural source for embryonic stem cells is the inner cell mass of the blastocyst. Blastocysts are formed in the early embryogenesis. After the oocyte is fertilized by the sperm, the zygote is formed (see Figure 2.1). The zygote is a single cell and totipotent, i.e. it can differentiate into any cell type, can divide indefinitely and build a complete individual. The zygote divides to build the blastocyst (also called blastula). In humans, this stage is reached after four to five days [154]. In vivo, the embryonic stem cells further divide and differentiate into any kind of fetal cell type and later into the adult cell types (see Figure 2.1). To obtain embryonic stem cells, the blastocyst has to be destroyed. This fact raises ethical questions.

Embryonic stem cells differentiated into terminal differentiated cells building several intermedi-
ate cell types (see Figure 2.2). In a first step, tissue-specific stem cells are formed. While they can still divide indefinitely, they lost the ability to form any cell type. Instead being pluripotent, tissue-specific stem cells are multipotent. Tissue-specific stem cells can be extracted from any organ in the fetal and adult organism [91]. They can be differentiated into all cell types in the organ from which they are extracted. Hence, tissue-specific stem cells are preprogrammed and build the natural source for regeneration of the tissues. Compared to embryonic stem cells, they are easier to obtain and thus, are popular cell types for research in regenerative medicine. Furthermore, in vivo and in vitro tissue-specific stem cells can be differentiate into precursor cells of the respective tissue and finally into differentiated cells (see Figure 2.2). Thus, mechanism of differentiation can be studied in vitro.

The low availability of embryonic stem cells induces many research projects aiming at transforming cells into pluripotent cells. The resulting cells are called induced pluripotent stem cells (iPSCs). Somatic cells are treated with different transcription factors to induce a specific expression pattern (see Figure 2.1). This leads to major changes in the regulation of the gene expression resulting in pluripotency [148]. Somatic cells cannot only be transformed into pluripotent cells but
also in multipotent cells [110] (see Figure 2.1) by the same method but different factors. Induced pluripotent as well as induced multipotent stem cells provide a second artificial source for stem cells which can be used for research and regenerative medicine. Generation of pluripotent stem cells using the patient’s own somatic cells provides the advantage that it is less likely to lead to rejection than using allogenic embryonic stem cells.

Fibroblasts derived from iPS cells and those derived from embryonic stem cells are very similar with respect to epigenetic and phenotypic properties [62]. Nevertheless, it was shown that iPS retain a memory on the source cell type at the epigenetic level [73]. Furthermore, it was discovered that they tend more often to build tumors then the cells derived from embryonic stem cells [55].

The problems observed leads to the investigation of differences between iPS cells, differentiated cells, tissue-specific stem cells, iPS-derived cells and embryonic stem cells. This includes projects regarding differences on the level of the epigenome as well as on transcription level and expression level. On the level of the epigenome, large scale projects try to establish reference epigenomes for many cell types of the human organism, for example the Roadmap Epigenomics Project [107], the Human Epigenome Project (HEP) [68] or the Blueprint epigenome [1]. Within these studied different histone marks, DNA methylations, transcription and chromatin accessibility are measured. Thus, a large amount of data sets is provided to study pluripotency, differentiation and differences between different cell types. Furthermore, differences in the epigenome during aging and in diseases can be studied once the epigenome state of normal cells is known.

Many histone modifications are discovered (see section 1.1). Only for a small part antibodies are designed to be used for genome-wide measurement of the respective mark. Nevertheless, all measurable modifications are still too many modifications to understand epigenetic regulation and the interplay of the marks due to the high complexity. Furthermore, differences in some of
the marks are less important for pluripotency and stem cell identity than other marks. Thus, examination of the most promising modifications for pluripotency enable discovery of the major and most important differences in the cell types studied. Some modifications are described as hallmark of pluripotency playing a crucial role in development, among them trimethylation at histone H3 lysines 4, 27, and 9. Thus, in this thesis, the focus lies only on this small set of histone modifications. Additionally, they can be found at the same loci and hence, it is worthwhile to study the patterns of these marks. The trimethylations at H3 lysine 4, 27, and 9 as well as their function and modifying complexes are described in the following. Knowledge on the interplay of these marks is described.

2.1 Histone Marks and Modifying Complexes

It is found that two histone modifications, namely the trimethylations at histone H3 at the lysines at position 4 and 27 (H3K4me3 and H3K27me3, respectively), play a key role in embryonic stem cells [99]. Genes required for cell differentiation often carry both marks, H3K4me3 and H3K27me3, at their promoters in embryonic stem cells [17]. One of these marks is lost, if they differentiate. The patterns of these marks seem to be important for staying in a pluripotent state. Thus, they are often called hallmarks of pluripotency.

In general, H3K4me3 and H3K27me3 are counteracting modifications. Trimethylation at H3K4 is often found in euchromatin and can be positively correlated with transcription [133, 15, 77, 59]. Furthermore, model predicting gene expression levels based on the chromatin state are most predictive if they also regard H3K4me3 [74]. H3K27 trimethylation is associated with the formation of heterochromatin and gene silencing [33, 116]. Both marks are catalyzed by enzymes containing a SET domain. This enzymes belong to the Trithorax Group Proteins (TrxG) and Polycomb Group Proteins (PcG) catalyzing H3K4me3 and H3K27me3, respectively [88, 2].

Several complexes containing TrxG or PcG are discovered. COMPASS and COMPASS-like complexes catalyze the methylations at H3K4 [149] (see Figure 2.3). The catalytic SET domain as well as most of the other proteins in the complexes are evolutionary conserved [138, 48] from yeast to human. The COMPASS-like complex consists of MLL, WDR5, RbBP5, Ash2L, and two DPY-30 proteins. MLL contains a SET domain which catalyzes the reaction from unmodified H3K4 to monomethylated H3K4. Ash2L is capable of catalyzing the reaction from H3K4me1 to dimethylated H3K4 [114]. It is unknown which domain in Ash2L has the catalytic activity.

Some of the proteins contained in the core of COMPASS-like complexes are also contained in COMPASS core, namely WDR5, RbBP5, Ash2L, and DPY-3 [149]. Furthermore, Set1, WDR82, and Cxxc1 are part of the COMPASS core [85, 86]. Set1 contains a SET domain able to methylate H3K4me2 and thus, setting a H3K4 trimethylation mark. Since it also contains Ash2L, COMPASS might be also able to dimethylate H3K4me1. Cxxc1 contains a Cxxc domain which binds unmethylated CpGs [40]. Thus, H3K4me3 is preferentially set at regions with high CpG-density.
Figure 2.3: Methylation of H3K4: COMPASS-like complexes can monomethylate H3K4 mediated by the MLL. Dimethylation of H3K4me is catalyzed by Ash2L which can be found in COMPASS and COMPASS-like complexes. Trimethylation of H3K4me2 is carried out by Set1 as part of the COMPASS complex.

and anti-correlate with DNA methylation [30].

Polycomb repressive complex 2 consists of four core proteins, namely enhancer of zeste 2 (Ezh2), embryonic ectoderm development (Eed), suppressor of zeste 12 (Suz12), and the retinoblast binding protein 4 (RBBP4) [94]. It is able to catalyze the trimethylation of H3K27 (see Figure 2.4). The catalytic SET domain is part of Ezh2. Furthermore, it was observed that it can also bind to H3K27me3 without losing the ability to set the trimethylation at H3K27 [56, 93]. Responsible for binding to H3K27me3, is the Suz12 protein and thus, inducing cooperativity in the methylation process.

Figure 2.4: Repression mediated by the Polycomb Repression Complexes: PRC2 binds the nucleosome and trimethylates H3K27. The catalytic activity of PRC2 arises from the SET domain in the Ezh2 protein. H3K27me3 serves as binding site for PRC1 which ubiquitinates H2AK119. The catalytic unit is RING1A or RING1B. H2AK119 to transcriptional repression by formation of heterochromatin. Figure taken from Spivakov et al. [142].

The PRC2 core complex can be alternatively composed of enhancer of zeste 1 (Ezh1), Eed, and Suz12. Ezh1 is a homolog of Ezh2 and set the H3K27me3 mark. Both complexes are required
to set the H3K27me3 in ESCs. However, depletion of Ezh2 shows stronger effects on H3K27me3 distribution in ESCs than Ezh1. Thus, Ezh2 is the major H3K27 trimethylases in ESCs [137].

H3K27me3 is associated with gene silencing and formation of heterochromatin but the effect is not carried out directly by H3K27me3 but indirectly (see Figure 2.4). A second complex called Polycomb repressing complex 1 binds to H3K27me3 marks and catalyze the ubiquitination of H2BK119 [38]. The heterochromatin protein 1 (HP1) binds H2BK119ub and links two neighboring nucleosomes. Doing so, the chromatin is tightened and forms heterochromatin [38].

There are 4 major H3K9 trimethyltransferases known, namely SETdb1, SETdb2, Suv39h1 and Suv39h2 [53, 120, 50, 151]. All of them are able to set trimethylation at H3K9me3. Not much is known about the complexes which are formed by these proteins. Thus, no core complex can be described.

SETdb1 is known to interact with MBD1; a protein which binds to methylated DNA. Thus, H3K9me3 is coupled with DNA-methylation and has a repressive function. This is even more strengthened by the fact that H3K9me3 is known to recruit HP1 [105] to form heterochromatin. It was also found that SETdb1 is recruited by CAF-1. CAF-1 brings newly synthesized H3 and H4 histones to the replicated DNA. In this way, H3K9me3 can be efficiently set on newly synthesized nucleosomes after DNA replication [130]. Suv39h1 is recruited by RB1 [156]. Literature reports that Suv39h1 is required for DNA methylation at pericentric repeats [87]. The recruitment of DNA methylation to repeats and the interaction with cell cycle and DNA replication processes shows the importance of H3K9me3 not only for pluripotent cells but for all cells.

The interaction with DNA methylation is reported for H3K9me3 in general. H3K9me3 is not only found where DNA methylation is already set (e.g. recruitment by MBD1) but activates DNA methylation pathways. The latter mechanism is not known in detail. Nevertheless, H3K9me3 marks are often found at regions which also carry H3K4me3 and H3K27me3 [19]. Since H3K4me3 and DNA methylation cannot co-occur [9], H3K9me3 do not necessarily lead to DNA methylation.

Co-occurrence of different histone modifications are observed genome-wide. However, the mechanism behind is usually unknown. The last years, different long non-coding RNAs are discovered which can bind usually one or two modifying complexes and guide them to specific loci in the genome. Thus, co-occurrence can be explained by guidance of both complexes to the same loci. Examples are Kcnq1ot1 binding G9a and PRC2 and thus, leading to co-localization of H3K9me3 and H3K27me3 marks [113] and HOTAIR binding PRC2 and LSD1 (H3K27me3 trimethylases and H3K4me2/1 demethylases) [132]. The conservation of the function of HOTAIR is low and varies in vertebrates [132]. Although RNAs might explain some of the co-occurrence and play key role in regulation of epigenetic marks, they can not explain the very abundant observations of co-occurrences of specific marks.

Summarizing, the three introduced marks are found at the same genomic locus. Certain combinations are found to co-localize but the mechanism of co-localization is only identified for some examples. Studying the occurrence and abundance of the different combinations can reveal the
Table 2.1: Combinations of Marks: Special names and functions are highlighted for combinations of H3K4me3, H3K27me3, and H3K9me3.

<table>
<thead>
<tr>
<th>Present Marks</th>
<th>Function/Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me3</td>
<td>euchromatin marker, activates transcription</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>heterochromatin marker, represses transcription</td>
</tr>
<tr>
<td>H3K4me3 &amp; H3K27me3</td>
<td>poised chromatin, bivalently marked chromatin, represses transcription but can be transformed by one change to euchromatin</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>heterochromatin marker, indicator for DNA methylation, represses transcription</td>
</tr>
<tr>
<td>H3K4me3, H3K27me3, H3K9me3</td>
<td>heterochromatin marker, represses transcription, frequently found in ESCs</td>
</tr>
</tbody>
</table>

commonness of the combination. Changes during differentiation can be investigated by including several cell types in the study. Changes in the distributions of the combinations provides insights into the function and cell type specificity of the combination and can indicate switches in the mechanism leading to the observed epigenome state. Moreover, the function of most of the combinations observed is still unclear. Specific functions are only assigned to some of the combinations (see Table 2.1). Analyses, thus, may reveal whether a combination has a specific function (is expected to occur frequent and show specific transition depending on the cell type) or just act at a transition state (specific transition state no matter in which cell type, less abundant).

### 2.2 Where have all the CpGs gone?

Looking at sequence pattern is interesting no matter which pattern is regarded. However, looking at CpG-density, as frequently done for the three marks studied, is often of higher interest than any other sequence pattern. The reason for it is the evolutionary conservation of CpGs. They are very rare in the whole genome and tend to occur in clusters. Clusters with a CpG-density higher than expected are called CpG-islands. The distribution of CpGs is surprising since single nucleotide frequencies do not differ much for the different bases [54]. On dinucleotide level, frequencies are not distributed uniformly anymore. Thus, there is an evolutionary drift leading to a genome-wide depletion of CpGs which is often used to estimate the phylogenetic originality of a genome [41].

Long time, the mechanism leading to a higher mutation rate for CpGs remained unknown. With the detection of DNA methylation, it was discovered that methylated CpGs are more often mutated into TpGs and CpAs than unmethylated CpGs. The process converting C’s into T’s is called deamination [41] (see Figure 2.5). At this, the amino group is removed from the cytosine which becomes than an uracil. Uracil is replaced by thymine due to the DNA-replication mechanisms [161].

CpGs are associated with high transcription rates and thus, are found in the promoters of many
Figure 2.5: DNA Methylation: DNA methyltransferases attach methyl-groups to cytosine. Demethylation can remove this group. Cytosine is turned into uracil by deamination. DNA-repair mechanism can reverse this reaction. During DNA replication uracil is replaced by thymidine. Similarly, 5-methyl cytosine is converted into thymidine by deamination. Figure taken from Singal et al. [139].

housekeeping genes [174, 135]. They are organized in CpG-islands which are genomic regions with a GC-content of more than 50% and a observed to expected ration of 0.6 for the CpGs [54]. While this is a rather computational definition, it was found that such regions provide a high amount of binding sites for transcription factors and thus, increase the affinity of the RNA polymerase complex for the respective promoter [29].

Since housekeeping genes are important for the function of all cells, evolutionary pressure is put on the conservation of the CpGs. For this reason, in some part of the genome CpGs are conserved in some part of the genome, e.g. promoter regions, and are depleted elsewhere. Nevertheless, deamination also found in promoters and contribute to emergence of transcription factor binding sites [172]. Thus, deamination is considered when modeling evolution of promoter sequences [14]. Conservation of CpGs are easier if the CpGs are unmethylated. Thus, preventing DNA methylation leads to a better evolutionary conservation of the CpGs.

DNA methylation is also functional. Methylating the CpGs in promoter regions leads to stable silencing of the corresponding gene [71]. Imprinting describes the methylation of one of the allele which lead to the expression of the gene from only one allele. An example for it is described in chapter 1. Furthermore, foreign DNA integrated into the genome and repetitive sequences get methylated to suppress transcription of such elements. Hence, DNA methylation retains the stability of the genome. Stable inheritance, thus, is important. Indeed DNA methylation is inherited with high efficiency encoded in a specific DNA methyltransferase. It is based on the fact that CpGs are found symmetrically on the DNA double strand, i.e. the reverse complement of a
CpG is again a CpG. In mammals, two types of DNA methyltransferases, de novo and maintenance DNA methyltransferases exist. The latter one binds to hemimethylated DNA, namely DNA regions where one strand carries DNA methylation marks and the other strand does not. Doing so, it methylates the unmethylated strand leading to a DNA double strand which is fully methylated. After DNA replication, genomic loci, which were fully methylated before the replication, remain hemimethylated. The maintenance DNA methyltransferase restore the information and turn the hemimethylated loci into fully methylated loci. This makes inheritance of DNA methylation more stable than inheritance of histone marks.

CpGs are targeted by the Cxxc domain in the set complex as explained in the previous section. Thus, CpG-rich regions are know to be H3K4me3 marked, among them the CpG-rich promoters. H3K4me3 marks at such promoters enhances transcription even more than CpGs alone by opening the chromatin and recruiting the RNA polymerase (see Figure 2.6). However, the Cxxc domain recognizes CpGs only if they are not methylated. Thus, H3K4me3 can only be found in regions which are unmethylated [40]. Furthermore, DNA-methyltransferases can not bind H3K4me3 loci [30]. Thus, H3K4me3 marks and DNA methylation marks exclude each other on the genome. This correlates with gene expression at such loci. CpG-rich promoters enhance transcription of the corresponding gene. Due to the CpG-richness, they are targeted by H3K4 trimethyltransferases. This results in higher transcription rates of the corresponding gene compared to the unmodified state of the promoter. Such genes can be efficiently silenced by DNA methylation of the promoter. In this case, H3K4me3 cannot be set anymore and transcription is silenced. However, DNA methylation marks can only be set once the H3K4me3 marks are removed (see Figure 2.6).

The other way around is the connection between H3K9me3 and DNA methylation. MBD1 was found to recruit H3K9 trimethylases [59]. It contains a CpG-binding domain (MBD) which only binds to methylated CpGs. Thus, H3K9me3 is localized at CpG-rich loci if they are methylated.
Also for the opposite direction, evidence was found. H3K9me3 seems to be crucial for the recruitment of DNA methyltransferases to pericentric repeats. However, the connections between H3K9me3 and DNA methylation is not fully understood [59]. Especially, H3K9me3 does not always correlate with DNA methylation, e.g. in ESCs this is probably not the case [9].

Considering H3K4me3 and H3K9me3, there are two counter acting modification not only in the meaning of their effect on transcription but also in conservation of CpGs. While H3K9me3 is known to recruit DNA methylation and thus, leads to depletion of CpGs, H3K4me3 prevent DNA methylation [109] and doing so, conserves CpGs. Examination of the association of these marks with CpGs provides insights in which amount and at which position and time CpGs are protected from DNA methylation or may be specifically marked as target for DNA methylation.
Chapter 3

Genome-wide Distribution of Epigenetic Marks

Epigenetic marks target specific locations in the genome. Often, not just one but several consecutive nucleosomes are modified by the same modification. The length distributions of these regions, namely consecutive nucleosomes with the same epigenetic mark, are characteristic for each modification and cell type. Doing this requires the analysis of the genome-wide distribution of the marks. Since histone marks can be found on the whole genome, a large amount of data has to be analyzed to obtain genome-wide distribution. For example, H3K9me3 is very abundant on chromosome 19 in embryonic stem cells of the mouse (see Figure 3.1). Thus, computational methods are necessary to handle the data and characterize epigenetic marks.

Figure 3.1: Distribution of H3K9me3 on chr19 in embryonic stem cells: Blue blocks indicate position with H3K9me3 marks. The red line shows the genes annotated by ensembl and in orange the genscan gene predictions for the chromosome 19 in mm9 mouse genome is depict.
For several modifying complexes (e.g. Polycomb), it is known that they not only bind to DNA motifs and unmodified histones but also to histones which already carry the modification they catalyze [111]. Thus, they act cooperatively. Nucleosomes are preferentially modified if nucleosome in close proximity already have a certain modification. This has a high impact on the distribution of the marks. Modified regions are typically more stable the longer they are [119, 126, 20]. Long regions can be more easily maintained over several proliferation steps by recovering the parental state from the partial information passed to the daughter cells. During differentiation, these region can be lost or specifically demodified leading to switch in the regulation of the underlying genes [84, 32]. Analyses of the distribution of an epigenetic mark in one cell type then provide insight in the mechanisms leading to the modification and changes during differentiation.

### 3.1 Short Reads to Regions

In this thesis, measurement of epigenetic marks by chromatin-immunoprecipitation followed by sequencing of the obtained DNA as described in section 1.2 is regarded but not measurement using ChIP-chip. To identify the modified regions, the short reads obtained in the experiments have to be mapped against the corresponding reference genome. For this purpose, several short read aligner have been designed to efficiently find the position of origin of each single read [131]. Here, Segemehl [63] is used as it allows mapping with mismatches, insertions, and deletions. It uses an enhanced suffix array for searching for nearly exact hits with a limit number of mismatches on a reference sequence. These hits are used as seed alignments of the full reads with corresponding positions in the reference sequence.

Figure 3.2: Filtering of Mapped Hits: Reads are mapped onto the reference sequence (black). Reads which fulfill all parameters are depict in green. Reads which do not map at all to the reference sequence are shown in gray. Orange reads do not pass the accuracy filter. Seeds for the red reads have more than two differences. Blue reads represent hits which for which better hits are found. Only green reads pass all filters and thus, are returned by the mapping.

Segemehl was used with the parameters shown in Table 3.1. For some parameters, the default values were kept. Three parameters were changed to better fit the purpose of analysis of epigenetic data. Only those regions position should be obtained which are most likely to be modified in the while cell population sequenced. By default, minimal accuracy of the reported alignments is 85%. The differences can result from sequencing errors and SNPs (namely differences in the genomic
sequences of the sequenced genome and the reference genome) but can also be artifacts of mapping the read to the wrong genomic position (e.g. because the position of origin is not sequenced so far). To improve the specificity of the mapping, the accuracy threshold was accordingly set to 90%. This means that for a reported alignment of a read with a length of 36nt up to 3nt ($\leq 10\%$ of 36nt $\leq 3.6$nt $\leq 3$nt) can differ from the reference sequence. Instead of taking into account all alignments satisfying the accuracy threshold, only those alignments were reported which had the best e-value (*best hit strategy*). It is likely that alignments with e-values worse than the best one are falsely mapped reads and better match another genomic position. The changes in both parameters, accuracy and hit strategy, improve the specificity of the mapping but also increase the amount of false negatives (reads which cannot be mapped onto the reference sequence). The maximal number of mismatches in the initial hits was increased from one difference (default) to two mismatches. Thus, the set of initial hits was increased. The resulting alignments are then filtered by the parameters explained above (see Figure 3.2).

Table 3.1: Segemehl Parameters: Parameter setting of the most important parameters for mapping ChIP-seq data against reference genome. For three parameters the values were set differently from the default value, while for the last two parameters the default values were kept.

<table>
<thead>
<tr>
<th>parameter</th>
<th>value</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>minimal accuracy of reported hits</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>best hit strategy - only the best hits with the best e-value are reported</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>at most 2 mismatches are allowed for the initial hit on the reference sequence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>parameter</th>
<th>value</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>5</td>
<td>maximum e-value</td>
</tr>
<tr>
<td>M</td>
<td>100</td>
<td>maximum number of initial hits</td>
</tr>
</tbody>
</table>

For each position in the reference genome, the number of reads which were mapped to this position was counted. This can be easily done with `samtools`. Since most of the positions will be zero, positions with zero counts are excluded. Thus, position which do not occur in the list can be assumed to have no reads assigned to it. The position-wise counts, and thus the mapping as well, is not only calculated for the ChIP-seq data for the epigenetic marks but also for the whole cell extract and histone H3 data if available.

All histone modifications studied in this work are modifications at histone H3. Thus, they can only be found at genomic positions wrapped around histones complexes containing H3. For this reason, the counts for positions in the genome where the number of reads for the marks is greater than zero are set to zero if no reads of the H3 where mapped to the positions within the distance of one read length. Since H3 data is not available for all cell types studied, it was checked whether H3 corrected and uncorrected data is still comparable. This is the case, since less than 1% of the positions are set to zero by this procedure. It highlights also the good affinity of the antibodies
used in the experiments.

ChIP extracted the DNA bound to the protein of interest (namely histone with a specific epigenetic mark). However, this selection process is not perfect. Thus, the sequenced DNA probe always contain DNA fragments which were not bound by the protein studied (see section 1.2). Hence, it is necessary to normalize the counts with respect to the background model. In the case of ChIP-seq, the typical background model is the whole cell extract. For the whole cell extract, the chromatin is extracted from the cell without selecting for a DNA bound protein. Thus, the whole genomic DNA is extracted and sequenced. While in the whole cell extract each position is theoretically sequenced equally often, in the ChIP library those positions are enriched which are associated with the studied mark. For this reason, the enrichment of reads mapped to a position with respect to the read counts in the whole cell extract is used as a filter. A cut-off value of three is chosen, meaning that only those position are classified as modified to which at least three times more reads where mapped in the ChIP experiment than in the whole cell extract (see Figure 3.3). By this procedure, positions which are hard to obtain by sequencing and thus, have low coverage in WCE, are taken even if they have only low coverage in the ChIP experiment. Positions which are highly covered by the WCE show that sequences are easy to sequence and to map. Thus, proportionally more reads have to be found at this position in the data of the ChIP experiment.

The position-wise counts are used to define the DNA regions associated with a specific epigenetic mark. Consecutive positions which are all classified as modified are joined to the initial set of regions (see Figure 3.3). The strict parameter setting during mapping and strict cut-off value for the normalization, may introduce gaps in the regions. Also fragments which were not sequenced as well as the fact that the fragment’s ends have a lower coverage by sequencing reads than the fragments mid contribute to the gaps which are artifact of the experiment and mapping. Thus, neighboring regions are joined if there distance is less than 100bp (approximately half length of a nucleosome, see Figure 3.3). Short regions with a length of 100bp or smaller are discarded if they can not be joined with neighboring regions. It is likely that these short regions are mapping artifacts since the experiments produces fragments of length 200-700bp. Regions with length smaller than 200bp are not expected but can occur if the region’s ends are not classified as modified.

Figure 3.3: Definition of Regions: For each position in the genome, it is checked whether the the experiments contains three times more reads than the WCE. Subsequent position passing the enrichment filtered are defined as modified regions. Two neighboring regions which are closer than 100nt are joint into one region.
due to insufficient enrichment. Thus, a minimal region length of 100bp corresponding to the half length of the DNA wrapped around a nucleosome was chosen.

Since the definition of the regions crucially depends on the cut-off value for normalization, the stability of the results was tested for different cut-off values. Thus, regions were defined for varying enrichments, namely cut-off values of one, three, and five. Most changes were observed in the short regions from which some were lost with higher cut-off values (see Figure 3.4). Regions consisting of one nucleosome show strongest changes in their abundance when varying the cut-off value. For long regions, one can observe a shortening of the length. However, the overall result, namely characteristic CpG-density and slope of the length distribution, vary only marginally. For this reason, a medium enrichment cut-off of three was chosen.

3.2 Calculating Length- and CpG-Density-Distribution

Characteristic properties of one modification are its length distribution and in the case of the studied marks also its CpG-density (see section 2.2). The length distribution provides inside about the abundance of epigenetic marks and the cooperativity of the modifying complexes for the studied marks. CpG-density analyses reveal the existence of CpG-dependent and CpG-independent recruitment of the modifying complexes. Combined with the length distribution, differences in the abundance and rate in the recruitment mechanisms can be described.
The length of a region is determined as the number of base pairs between start and stop position. This is the length of the DNA wrapped around the modified nucleosome. As an rough estimate of the number of consecutively modified nucleosomes, the length is divided by 200bp and rounded. The region lengths in nucleosomes is used to calculate the distribution of the lengths. For each region length in nucleosomes, it is counted how often it occurred in the data.

As a measurement of the CpG-density, the number of CG-dinucleotides (CpGs) in regions are counted. Only the plus strand as defined in the genomic reference is used. The actual CpG-density is than obtained by dividing the number of CpG by the number of dinucleotides in the region. Investigation of CpG-dependency simultaneously with the region length can be done by plotting the length on the x-axis and the number of CpGs on the y-axis (e.g. see Figure 4.2ii in the next chapter).

CpG-dinucleotides are not uniformly distributed in the genomic sequence. Therefore, the background CpG-density has to be determined. Genomic sequences are chosen randomly from the genome. The lengths are chosen to be between the smallest and largest length of the modified regions and reflect the length distribution of the experimental data. Given a random genomic sequence, the number of CpGs is determined in the same way as for the modified regions. Taking the artificial data, the average number of CpGs as well as the highest number of CpGs obtainable for certain region lengths can be computed. Since the number of data points is high, the density plot was created using the smoothScatter function in R [121]. At it, nearby points are reduced to one point and each point color intensity is choose according to the number of data points assigned to it.

The background distribution of the CpG-density is used as reference to assess the CpG-density distribution in the studied data sets. Regions with a CpG-density near to the average CpG-density measured in the genome are produced by the CpG-independent mechanism. If many regions of a mark have a CpG-density similar to the maximal measured density, a CpG-dependent mechanism can be inferred. The modifying complex is recruited only to those parts in the genome with the highest CpG-density. A characteristic CpG-density between the background and maximal CpG-density refers to a mechanism using a specific CpG-motif as binding site. Mechanisms avoiding CpGs would produce data sets with regions having a CpG-density below the average density.

### 3.3 Calculation of Overlaps between Epigenetic Marks

Epigenetic marks are known to co-occur with other marks which influences the function of the marks. Thus, one mark can have several different functions depending on which marks are present in the nearby environment. For this reason, it is interesting to analyze the overlap of different epigenetic marks. Not only for different marks, the overlap can be computed but also for the same mark in different cell types. It provides inside into the changes during differentiation. Overlap calculations are a common method to analyze co-occurrence of two features [168, 96, 152].
Given two sets of regions, the overlap can be calculated genome-wide or region-wise. Genome-wide overlap is the number of nucleotides in the set of reference regions which are also covered by an alternative set of regions. This number is normalized by the number of nucleotides in the regions of the reference. Since this is not symmetric, the overlap has to be calculated in both directions. In this way, the fraction of the reference modifications covered by the alternative modification is calculated and vice versa.

In the case of the region-wise measurement of the overlap, for each region in the reference data set overlap with alternative data set is calculated (see Figure 3.5). The distribution of these values provides insights into the characteristic values for the coverage values. Also here, reference and alternative data set have to be switched and the coverage values have to be calculated again to gain full insight into the overlap of two data sets (see Figure 3.5).

The distribution of the coverage values is used to identify a threshold to decide whether a region is also observed in another data set or not. For a reference data set, the regions can be divided into subsets which are covered with a set of different modifications. For each subset, length distributions and CpG-density can be analyzed. Differences in the distribution coming along with certain combinations of epigenetic marks can be detected as well as changes in both properties during differentiation.

Although in theory, this can be done with several sets of modifications or several set of cell types, the results are difficult to interpret with many different data sets. Thus, important connection may be missed. Another drawback is that this method is not designed to do genome-wide analyses with respect to both modifications used. Only the part modified by the reference modification data set is taken into account, while the part modified with the alternative modification data set but not with the reference is not regarded. In chapter 5, more sophisticated methods for data analysis and visualization are presented especially designed to analyze epigenetic patterns genome-wide. They also allow for correlating patterns with e.g. CpG-density, length, expression data, and gene annotations.
Chapter 4

Distribution of Epigenetic Marks in Mouse

Analysis of epigenetic marks in mouse is performed on data for three modifications, namely H3K4me3, H3K27me3, and H3K9me3, in three cell types, namely embryonic stem cells (ESCs), murine embryonic fibroblasts (MEFs), and neuronal progenitor cells (NPCs). The data was published by Mikkelsen et al [99] and is available under the GEO accession number GSE12241. It was analyzed as described in section 3.1. Calculations of length and CpG-distributions was performed according to section 3.2. Pair-wise overlaps are calculated for all modifications in ESCs and for the H3K4me3 modifications in all three cell types. The methods explained in section 3.3 are used for the calculations of the overlaps.

The results obtained by the analysis of the data described above are shown in the following sections. Firstly, all length and CpG-distributions of the modifications measured in ESCs are analyzed to obtain insight in the mechanism leading to the modification states observed. Pair-wise overlaps are calculated to estimate to which extent the marks overlap in general. Using the data, an threshold can be defined to classify regions from a given modification into modified and unmodified by a second modification. Overlaps between all three marks are calculated to
encounter the co-occurrence of the marks in ESCs. Analogously to the analysis of the marks in ESCs, the H3K4me3 data in all three cell types is analyzed. The changes in the distribution of the modification are considered in this study.

For comparison reasons, the CpG-distribution of the mouse genome has to be measured which is explained in section 3.2. The mm9 version [104] of the mouse genome downloaded from the UCSC browser [75] was used. Only the primary assembly (no haplotype data) without the unplaced contig data and the mitochondrial sequence was used. This is the same genomic sequence also used for mapping the read data. Sequences were drawn randomly from the genome and the CpG-content was determined. Average and maximal CpG-density was determined (see Figure 4.1). Therefore, the maximal and average CpG-density for all lengths was calculated. Linear regression was used to calculate the average and maximal CpG-density.

### 4.1 Epigenetic marks in ESCs

Three marks are studied in ESCs, namely H3K4me3, H3K27me3, and H3K9me3. It is reported that all of them are important for embryonic stem cell identity and are needed for pluripotency. Correlation with CpG-density of the underlying DNA was shown [99]. For this reason, length and CpG-distributions are calculated for the three marks in ESCs and the correlation and characteristics of these marks are evaluated. In Figure 4.2, length and CpG-distributions for the marks in ESCs are shown.
4.1.1 Length Distributions of the ESC data

Figure 4.2i shows the length distribution for the three marks in ESCs. Regions length in nucleosomes is multiplied by 200nt for a better comparison with CpG-distribution data. All three distributions show exponentially decreasing frequency with higher lengths. Regions with length greater 10000nt occur very rare and thus, are excluded from the plot. Exceptional long regions (lengths greater than 20000nt) are described in subsection 4.1.2.

All three distributions show a small enrichment of regions with lengths near to 8000-9000nt (40-45 nucleosomes). The strongest enrichment is found in the H3K4me3 data (compare Figure 4.2i,a with Figure 4.2i,b and 4.2i,c). This might occur due to joining of neighboring modified regions with a distance less than 100nt (see section 3.1). These neighboring regions are highly modified and much shorter than 10000nt. Very near to them another modified regions is found as well with length significantly shorter than 10000nt. Due to the fuzziness of the data, the distance between them is shorter than 100nt. Since this length cannot be occupied by the unmodified nucleosome, the regions are joined into one region.

Dotted lines in Figure 4.2i show the lengths where the length distributions change their slopes and thus, more strongly decreases the frequency with higher lengths. A change of the slope occur in all three distribution at the length of 400nt which corresponds to 2 nucleosomes. Thus, regions consisting of only one or two modified nucleosomes are distributed in a different way than regions with at least 3 nucleosomes. More precisely, if all lengths would originate from the distribution of the longer regions, more regions with one and two nucleosomes than observed would be expected. Once can formulate the hypothesis that regions with less than 3 nucleosomes are more unstable than regions with more than 2 nucleosomes. Thus, one and two continuously modified nucleosomes do more rarely exists at the same loci in all cells of a cell population. For this
reason, the measurement of these regions might be more difficult than the measurement of longer regions. This effect is stronger in the H3K27me3 data and only slightly observable in H3K4me3 which may indicate lower stability of H3K27me3 marks compared to the H3K4me3 marks.

The length distribution can be fitted using a differential equation model as in Binder et al [20]. An interaction complex binds to the chromatin either to binding sites or already modified histones. Once bound, it modifies a histone. Demodification is not modeled by another complex but occurs constantly, e.g. caused by cell division. It is assumed that a certain set of nucleosomes (response element or cooperative unit) acts cooperatively regarding the binding of the interaction complex. The cooperative unit’s lengths are distributed according to a hypergeometric distribution simulating binding sites for insulator proteins, e.g. CTCF [60], at the boundaries of the cooperative units. The model (see Figure 4.3) consists of three parts:

i) A repulsion term representing the energy which is needed to e.g. making the chromatin accessible or building the interaction complex.

ii) A term for binding to DNA binding sites. Each cooperative unit contains a small number of binding sites. The number of binding sites is proportional to the length of the cooperative unit with a upper cut-off value of 5. The binding energy of the interaction complex to the binding site is chosen to fit to the data.

iii) A cooperativity term models the binding of the complex to another already modified histone in the cooperative unit. The binding energy is fitted to the data. In each time step, it is multiplied by the number of modified histones in the cooperative unit.

Although the model is formulated for modifications which lead to gene silencing (e.g. H3K27me3), it can also be used for modifications which activate transcription. For example, the model can be used to simulate H3K4me3 marks. For that purpose, one has to change the outcome of the modified and unmodified state of the histones. Thus, modified CUs lead to activation of transcription, while transcription in unmodified CUs is silenced. However, changing the outcome of the states, does not change the model itself but the interpretation of the states.

The model shows that the length distribution does not depend on the modification rate itself but on the ratio of modification rate to demodification rate [126, 20]. Depending on the parameters and the length of the cooperative unit, 3 different states can be distinguished: monostable off, monostable on, and bistable. The latter one can switch between on and off with the same parameter setting induced by, e.g., proliferation. It is a consequence of the cooperativity [126]. Cooperative units with bistable behavior have a certain length. However, the range of these length is very small but increases with the noise ratio assumed for the simulation. Noise models for example, replacement or depletion of nucleosomes.

Highest stability is predicted for the longest regions [126, 20]. This can be seen in the data of MEFs, the length distribution show lower numbers of short regions while the long regions are retained corresponding to a higher stability of long regions compared to short regions. In later
differentiation level such as NPCs, the longer regions can not be retained anymore and thus, get lost [20]. Stability of long regions is a consequence of the cooperativity implemented in the model. This can be assumed since it is observed that the modifying complexes are found more often at genomic locations which are already modified than at loci which are unmodified. At least of PRC2, the complex member responsible for binding to H3K27me3 is known; it is Suz12 [56, 93].

4.1.2 Outliers of the Length Distributions

The three length distribution have a cut-off of approximately 10000bp. Only a few regions exist being longer which fit nevertheless to the fitted distribution. Most of them are shorter than 20000bp. However, some of them are extremely long. These are listed in Table 4.1. The modification states in all three cell types are shown. The genomic location was examined to discover the genes which are annotated at these loci. Additionally, expression data for different cell types available as UCSC Genome Browser [75] track was used to detect whether the regions are expressed in ESCs and differentiated cells. The expression data was published as part of the ENCODE project [127]. Depiction of the regions in the UCSC Genome Browser can be found in the appendix in Figures A.1 to Figure A.9. An example is shown in Figure 4.4. Only the modifications in ESCs and the gene annotations are shown. This region contains the SNORD116. Further details of this and the other outliers are reported below.

The regions on chromosome X and Y contain neither gene annotations nor gene predictions.
Figure 4.4: Modification State of chr7:67.000.000-67.050.000. Marks of H3K27me3, H3K4me3, and H3K9me3 in ESCs are depicted in red, green, and blue, respectively. Ensembl annotation and genscan prediction are shown below the modification state. In black, repeats are indicated.

Table 4.1: Long Modified Regions in ESCs: Exceptional long modified regions found in any of the three modification data sets in ESCs are shown. For each region, genomic coordinates and modification state in all three cell types (ESC, MEF, NPC) is shown. Green squares indicate present modification, blue lozenges stand for a slightly diminished modification state, a magenta triangle is taken for only partly modified regions, and red stars show absent modifications. Annotated/predicted genes within the regions are listed. The genomic location is shown in the figures referenced in the table.

<table>
<thead>
<tr>
<th>Location</th>
<th>Figure</th>
<th>H3K4me3</th>
<th>H3K27me3</th>
<th>H3K9me3</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr8:126061166-126148424</td>
<td>Figure A.1</td>
<td></td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>5S rRNA predicted miRNA, one genscan prediction</td>
</tr>
<tr>
<td>chr6:47601570-47723339</td>
<td>Figure A.2</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>ensembl pipeline annotated genes</td>
</tr>
<tr>
<td>chr9:3000001-3038591</td>
<td>Figure A.3</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>genscan prediction SNORD116</td>
</tr>
<tr>
<td>chr6:114444142-114587871</td>
<td>Figure A.4</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>SNORD115</td>
</tr>
<tr>
<td>chr7:67002206-67047765</td>
<td>Figure A.5</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td></td>
</tr>
<tr>
<td>chr7:66763853-66855402</td>
<td>Figure A.6</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td></td>
</tr>
<tr>
<td>chrX:166601240-166633574</td>
<td>Figure A.7</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td></td>
</tr>
<tr>
<td>chrY:2688290-2765959</td>
<td>Figure A.8</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td></td>
</tr>
<tr>
<td>chr3:3009790-3033629</td>
<td>Figure A.9</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td>♦♦♦</td>
<td></td>
</tr>
</tbody>
</table>

Also no significant expression could be detected in any of the ENCODE data sets currently available at the UCSC browser. Thus, the regions might not function in gene regulation but may play a role in chromatin structuring. One region on chromosome 3 overlaps with the position of two genes predicted by genscan [27]. However, gene annotation have not annotated these genes. Thus, no function can be inferred from the genes.

In the region on chromosome 9, 13 genes are annotated by ensembl [117]. Except one gene, all of them are located on the positive strand and encode for proteins. Nearly all of them (10 of 12) contain frame-shift introns which are small introns (1,2, 4, or 5 bases long) which are inserted to make a given cDNA fit the genomic positions. Such frame-shift introns occur rarely. Discovering them in 10 of 12 genes is very unlikely. On that account, the annotation might not be correct. One the reversed strand, one gene is annotated having 3 transcript. The transcripts do not encode for proteins but are long intergenic non-coding RNAs. Their sequences match cDNA which was extracted from the spleen. The boundaries of the region found modified matches
very good the start and end of the genes (see Figure A.3 transcripts ENSMUST00000143083, ENSMUST00000135951, ENSMUST00000155807). This is not surprising since chromatin data was incorporated in the annotation of the lincRNA.

Both regions on chromosome 7 overlap with the IC-SNURF-SNRPN transcript. It is known that this transcript is maternally imprinted (DNA methylated on only one allele) and thus, only paternally expressed [128]. The regions cover intronic regions of a non-coding transcript which is associated with the Prader-Willi syndrome [128]. Exactly at the regions position, two snoRNAs are found, SNORD116 (see Figure A.5 and Figure 4.4) and SNORD115 (see Figure A.6). The genomic organization is very unusual since these snoRNAs occur not only once but as multiple copies in the genome. Both snoRNAs show brain-specific expression [128]. This correlates with the observation that the marks get lost during differentiation into NPCs. However, H3K27me3 seems to remain longer on this regions suggesting a late activation of this regions (see Figure A.5 and A.6). At least for SNORD115, it was shown that it is important for regulation of alternative splicing and regulation of gene expression in brain tissues [24]. While over-expression of SNORD115 leads to autistic phenotypes in mouse [106], microdeletions cause the Prader-Willi syndrome [129].

The region on chromosome 8 matches the locus of the 5S rRNA gene cluster. It contains 151 copies of the 5S rRNA. It is known that the clusters arose from head-to-tail tandem repeats [147]. The 5S rRNA is part of the large 50S subunit of the ribosome which assemble the proteins given an mRNA template [102]. In ESCs, all three marks cover the region. In the other two cell types they diminish or vanish completely except H3K27me3 (see Figure A.1).

On chromosome 6, two region were found with exceptional large lengths. Genscan predictions span those regions. They consist of many short exons separated by introns which are much larger than the exons. Such structure can be also observed for non-coding RNA gene such as SNORD115 and SONRD116. However, in this case it is predicted as protein-coding. The region more closely located to the annotated beginning of the chromosome also contains a miRNA predicted by ensembl.

### 4.1.3 Correlation with CpG-Density

In Figure 4.2ii the number of CpGs for each regions as a function of the length is shown. Two lines are plotted indicating the maximum number CpGs observed in long regions genome-wide (red regression line) and the average CpG count in the genome (green line). For all three modifications, different distributions can be observed. All distributions have in common that regions accumulate at high CpG-density as extensively reported in literature [99, 67, 40]. It can be seen in Figure 4.2ii where a lot of points are densely located around the red line. The red line corresponds to a CpG-density of about 1/32 which is half of the CpG-density expected for uniformly distributed CpGs. Certainly, it is known that during evolution CpGs are more often converted into other dinucleotides than any other dinucleotide [97].

The distribution of H3K4me3 exhibit a set of short regions with even higher CpG-density than
the red line. Such high CpG-densities can only be observed in regions smaller than 2000nt of H3K4me3 and co-locate with CpG-islands [20]. Except the small regions with very high CpG-density, the points do solely spread around the highest CpG-density for long genomic regions. Thus, any mechanism setting H3K4me3 marks in mouse ESCs is CpG-dependent. This is congruent with literature. A possible function of H3K4me3 is also to protect CpGs from being DNA methylation. Since methylated CpGs are more easily mutated into TpGs and CpAs, H3K4me3 modification helps conserving CpGs in the genome [97].

Likewise H3K4me3, H3K27me3 is found at a CpG-density of about 1/32. However, not all H3K27me3 modified regions are CpG-rich, there is a set of regions which accumulate at a CpG-density of 1/211 (see green line in Figure 4.2ii,b), the average genome-wide CpG-density. Thus, one can assume two mechanisms, a CpG-dependent and a CpG-independent mechanism. So far, it is just know that H3K27me3 is often found at CpG-islands and regions with high CpG-density. Which part of PRC2 result in this distribution or whether another complex is involved is not known. Since it is known that there are many different proteins occasionally joining PRC2, one can hypothesize that the PRC2 complex exists in at least two different configuration. In the one configuration, PRC2 is joined by a CpG-binding protein and thus, set H3K27me3 at regions with high CpG-density. In the other configuration, PRC2 acts independently of a CpG-binding protein. The data suggest that at least in ESCs the configuration with CpG-binding protein occur more often than the configuration without CpG-binding protein.

Similarly to the distribution of H3K27me3, also in the distribution of H3K9me3 two branches at different CpG-densities can be observed (see Figure 4.2ii,c). Certainly, less regions are located at low CpG-density (green line). Also the variance in the CpG-density of the upper branch (red line) is higher than for H3K4me3 and H3K27me3. Thus, the modifying complex for H3K9me3 does not bind strongly to CpG-density. Up to now, the modifying complex for H3K9me3 is unknown. Furthermore, the only protein which binds CpGs and cooperate with writers for H3K9me3 is MBD1. However, it binds methylated CpGs [130]. Since H3K4me3 can not be set at DNA methylated regions, H3K9me3 and H3K4me3 should not co-occur. Nevertheless, in ESCs, it is observed that most of the H3K9me3 modified regions are also modified by H3K4me3 and H3K27me3 (see subsection 4.1.4 and [19]). The mechanisms which couples these marks is so far unknown.

### 4.1.4 Overlaps between Regions of Different Modifications

Literature reports that all three marks can occur together [99, 19]. Thus, the overlap of the marks in ESCs were calculated. Since the overlap calculations are not symmetric, the overlap between each pair of modifications was calculated with either mark as basis.

Surprisingly, all coverage distribution have a similar shape. Many regions are covered either almost fully (coverage value of almost 1) or to nearly 0% (coverage value of nearly 0) by the other modifications. Thus, sharp peaks can be observed at 1 and 0 in all distributions (see Figure 4.5). The distribution differ in the high of the peaks which means that the amount of overlap between
the modification is different.

With H3K4me3 modified regions and H3K27me3 as second modification, the peak at 1 is slightly higher than the peak at 0 (see Figure 4.5a,ii). Thus, more than half of the H3K4me3 regions are also covered with H3K27me3. The same is observed for H3K4me3 regions covered with H3K9me3 (see Figure 4.5a,iii). Contrary, the opposite is not true. In Figures 4.5b,i and 4.5c,i, the distribution have only a very small peak at 1 but a high peak at 0. This means that most of the regions of H3K27me3 and H3K9me3, respectively, are not covered with H3K4me3. There are nearly as many regions which are modified by both, H3K27me3 and H3K9me3, as regions modified with H3K9me3 only (see Figure 4.5c,ii). Regions modified by H3K27me3 and H3K9me3 are slightly less abundant than regions modified by H3K27me3 only (see Figure 4.5b,iii). This can be explained by the fact that there are more regions in the H3K27me3 data set than in the H3K9me3 data set.

Given the facts above, it is of high interest to investigate the difference between regions shared by modifications and the regions just modified by one modification. Furthermore, regions modified by all three marks may show different characteristics than regions with just two or one mark. To understand this phenomena, H3K4me3 is taken as reference. It is the data set with the lowest number of regions and most of these regions are also covered by other modifications. All regions are sorted by the combination of marks present on them (see section 3.3).

In Figure 4.6 all combinations of marks are shown. Most of the long regions are modified by all three marks. Only a few of the long regions are bivalently marked and do not carry H3K9me3.
These strong co-occurrence of all three marks is reported in literature as well [19]. However, it is not reported that the regions with all three marks or at least H3K4me3 and H3K27me3 made up the whole set on long, CpG-rich regions.

The H3K4me3 and H3K9me3 modified regions as well as the H3K4me3 regions are short. In the case of H3K4me3, there are also regions which are not CpG-rich. Thus, the mechanism for catalyzing trimethylation on H3K4 can not only act on CpG-rich regions but also on region with low CpG-density. The latter one is only possible on short regions and thus not very stable.

### 4.2 Distribution of H3K4me3 in ESCs, MEFs, and NPCs

During differentiation, many changes in the epigenome state have to be made. Genes required for the next cell type are activated and cell type specific genes of the previous stage are repressed. Thus, analyzing the changes in the distribution of H3K4me3 in the three cell types reveals mechanism for differentiation. Length distributions as well as CpG-distributions are shown for H3K4me3 in ESCs, MEFs, and NPCs in Figure 4.7.

#### 4.2.1 Length Distribution of H3K4me3

Figure 4.7i shows the length distributions for H3K4me3 in all three cell types. Region lengths were prepared as for ESC data (see subsection 4.1.1). Progressive shortening of the regions can be identified as a general trend for differentiation into NPCs. While ESCs regions reach lengths until 10000nt, regions in MEFs are mostly shorter than 8000nt (compare Figure 4.7i,a and Figure 4.7i,b). This corresponds to shortening of about 10 nucleosomes from ESCs to MEFs. Even more dramatically is the change to NPCs. Lengths above 3000nt are only observed sporadically (see Figure 4.7i,c). Thus, regions in NPCs are usually not longer than 15 nucleosomes.

Dotted lines in Figure 4.7i represent the lengths where the slopes of the distributions change. In all three distributions, such a changes can be observed at the length of 400nt (red dotted lines). Compared to MEFs and NPCs, the change in ESCs is less strong. In NPCs, a second change can
Figure 4.7: Length and CpG-Distribution of H3K4me3 in ESCs, MEFs, and NPCs: i) Length distributions are shown for all cell types as black lines, namely a) ESCs, b) MEFs, and c) NPCs. Changes in the slope of the distributions are indicated as dotted lines. ii) Length and number of CpG are depicted for H3K4me3 modification in each cell type. Red lines indicate the highest CpG-count observable in the mouse genome on long regions, while green lines show the average CpG-count in the mouse genome.

be observed at a length of 1000nt (green dotted line). This change is even stronger than the change at length of 400nt. Furthermore, the shape of the length distribution in NPCs strongly differs from the shape of the length distributions in ESCs and MEFs. Frequencies decreasing even faster with length in NPCs than in MEFs and ESCs. Thus, the underlying mechanisms changes dramatically in NPCs leading to the measurement of a very different length distribution.

A biological interpretation of the phenomenon above might be the loss of cooperativity in the underlying mechanism. Thus, only short regions are stably modified and long regions can not be established. The step-wise reduction of the regions size from ESCs over MEFs to NPCs vote for a step-wise reduction of the influence of cooperativity in the underlying mechanism. Since the length distribution in NPCs has completely different shape than the distributions for ESCs and MEFs, the mechanism in NPCs may not include cooperativity at all.

4.2.2 Changes in the CpG-Dependency

Figure 4.7ii shows the distribution of the CpGs for H3K4me3 marks in ESC, MEFs, and NPCs. The genome-wide maximal number of CpGs in long regions is indicated by a red line and the average CpG-density by a green line. Most H3K4me3 regions accumulate around the red line in ESCs (see Figure 4.7ii,a). Thus, catalyzing H3K4me3 marks is strictly CpG-dependent and specifically select regions with a CpG-density of approximately $\frac{1}{32}$ (see also subsection 4.1.3).

In MEFs, the same kind of mechanism can be observed. Still, many regions accumulate around the red line. Nevertheless in Figure 4.7ii,b, once can see several region which are short and have higher CpG-density then indicated by the red line. This very high CpG-density can only be reached on region shorter than 2000nt. Mostly, this region are CpG-islands. Given changes in the CpG-distribution, changes in the underlying mechanism can be conducted. While regions established
in ESCs become demodified (regions located at the red line in Figure 4.7ii,b), new regions are established with different mechanism depending on a higher CpG-dependency than in ESCs.

CpG-dependency is completely lost in NPCs (see Figure 4.7ii,c). Regions in NPCs are short and there CpG-density varies greatly. In contrast to the MEFs, regions with very high CpG-density can not be found. The maximal number of CpGs is around 100 CpGs (see Figure 4.7ii,c), while MEFs can have nearly 200 CpGs on regions of the same length (see Figure 4.7ii,b). Thus, a CpG-independent mechanism for H3K4me3 trimethylation can be assumed for NPCs. Seeing this picture, it is getting even more clear that there have to be different mechanisms for setting H3K4me3.

While differentiation, CpG-dependency of H3K4me3 mechanisms changes from strictly CpG-dependent in ESCs into CpG-independent in NPCs. This can not be explained by having just one complex. Thus, at least two mechanisms exists may using the same core complex but employing a number of different proteins defining different binding sites for the complex. There is evidence for a third mechanism existing only in MEFs but not in ESCs and NPCs. It selectively bind the CpG-richest regions of the genome. Taken together, H3K4me3 modifying complexes in ESCs are CpG-dependent and modify regions with a CpG-density of approximately $1/32$. While differentiation into MEFs, the mechanism still remains but may reduce the catalytic activity leading to a reduction of the maximal length from 10000nt to around 8000nt. Additionally, another process modifies regions which are extremely CpG-rich but short. Doing the next step in differentiation into NPCs, the both mechanisms are switched off. At the latest in NPCs, a new mechanism is activated resulting in short regions without defined CpG-density. It might be the case that this mechanism already exists in MEFs but can not be recognized in the current analysis of the data sets.

4.2.3 Retention of H3K4me3 Modified Regions

In ESCs, many H3K4me3 regions are established. The amount of retained regions in MEFs or NPCs provides insight into the stability of the underlying mechanism. To study the retention of the regions, the overlaps of the ESCs H3K4me3 regions with the H3K4me3 regions in MEFs and NPCs are calculated. As in subsection 4.1.4, a threshold for the coverage has to be defined to distinguish between retained regions and lost regions. This threshold is defined with the help of Figure 4.8.

Most of the plots in Figure 4.8 have similar shape; a huge peak at zero and a small peak at one. Only two plot have high peaks at a coverage of approximately one, namely ESC regions covered by MEF regions and NPC regions covered by MEF regions (Figure 4.8a,ii and Figure 4.8c,ii, respectively). Thus, H3K4me3 set in ESCs are retained in MEFs to a large amount. This is a evidence that the mechanism for guiding the modifier to this regions are active in both, ESC and MEFs. An additional mechanism for H3K4me3 modifications have to exist in MEFs since most H3K4me3 regions MEFs do not exit in ESCs (see Figure 4.8b,i). This is already suggested by the number of regions which is in MEFs about twice as large as in ESCs (see Figure 4.8a,ii and
A large number of H3K4me3 regions in NPCs did already exists in MEFs (see Figure 4.8c,ii peak of coverage 1). Roughly the same number of H3K4me3 regions are newly modified regions in NPCs (see Figure 4.8c,ii peak of coverage 1). Again, this is evidence for two mechanism, one which exist in both MEFs and NPCs and one which exist in NPCs but not in MEFs. However, the mechanism mentioned first is not dominant in MEFs. Only a small part of the H3K4me3 regions in MEF are still modified by H3K4me3 in NPCs (see Figure 4.8b,iii).

Most H3K4me3 regions in ESCs can not be retained modified until NPCs (see Figure 4.8a,iii). Likewise, most of the H3K4me3 region in NPCs do not exist in ESCs (see Figure 4.8c,i). This fact is additional evidence for different mechanism for H3K4me3 mark catalysis in ESCs and NPCs.
they can be found again. For each set of regions, length and CpG-density are depict. Most of the long CpG-rich branch found in the ESC data can only be retained in MEFs (see Figure 4.9c). Some regions are already lost in MEFs and thus, can only be found in Figure 4.9a. A minor part of the CpG-rich regions, especially the shorter regions, can be retained in all cell types. Regions which get lost and re-modified in NPCs, are extremely short and do not exhibit a defined CpG-density. Conceivably, the signal and the mechanism leading to the regions in ESCs is completely different from the signal and mechanism in NPCs.

![Figure 4.10: Length Distributions of H3K4me3 and H3K27me3](image)

Figure 4.10: Length Distributions of H3K4me3 and H3K27me3: Length distribution obtained from simulations with the model from [20] (black, red, and green lines). Dots indicate experimental data from mouse ESCs, MEFs, and NPCs. Distributions of ESCs are shown in black, red indicates distributions of MEFs and NPCs distributions are shown in green. The blue solid line shows the double exponential fit of the ESCs data.

Taking the data from Figure 4.9, three length distributions can be generated for ESC H3K4me3 regions. For each of the cell types, all ESC H3K4me3 regions are regarded which can be re-discovered in the respective cell type. All regions from Figure 4.9 are used for the ESC length distribution. Regions in Figure 4.9b and 4.9c are included in the length distribution for MEFs and 4.9b and 4.9d contribute to the length distribution of NPCs. This length distributions are used to fit the parameters for the model described in subsection 4.1.1. Afterwards, the model is used to simulate the dynamics of H3K4me3 regions for the different cell types. Data from this simulations were analyzed in the same way as the experimental data itself and length distributions for each cell type are calculated. Additionally, the analytic solution of the length distribution were calculated for the ESCs parameters. Analytic solution, simulated distributions, and length distribution ESCs H3K4me3 regions in the three cell types are shown in Figure 4.10.

The predicted length distribution fit very well with the experimentally determined distributions. It was found that differences in the length distributions in ESCs and MEFs can be reached by shifting the modification-demodification-ratio to higher demodification from ESCs to MEFs. This can either be implemented by increasing the demodification rate or by decreasing the modification rate. The transition from ESCs to NPCs can be modeled in two different ways. (1) The modification-demodification-ratio is shifted even further to higher demodification. (2) The distribution of insulator elements is changed such that the cooperative units are much shorter.
Comparing the length distribution in the different cell types, one can see that long regions are retained in MEFs while many short regions are already lost. This shows that cooperativity which is explicitly assumed in the model leads to higher stability of long regions versus short regions. Not until NPCs, the long H3K4me3 created in ESCs get lost. Similar results can be obtained for H3K27me3 length distribution in the three cell types which are shown in Figure 4.10.

With the described model and parameter settings, it can be shown that stably modified regions (e.g. the long modified regions in ESCs) can switch into bistability with increasing feedback-to-noise ratio [126], e.g. cell divisions. Thus, DNA replication is source for perturbations of the system. During each cell division, the epigenetic state has to be inherited to both daughter cells. Thus, approximately only half of the histone modifications are transmitted to each of the daughter cells, respectively. After replication, the lost information has to be re-established. In the range of bistable parameters, only a few generations are necessary to pass into a stable differentiated state [126]. Thus, bistability and proliferation may be driving forces of differentiation.

The data presented in this section shed light on the mechanisms for regulation of H3K4me3 marks in early developmental stages. As extensively reported in literature, H3K4me3 modifications are found at a CpG-rich regions. However, that is only half the story. The marks are not found at all CpG-rich regions but only those exhibiting a maximal genome-wide CpG-density on long regions. Short regions which reach even higher CpG-densities are not modified in ESCs or NPCs but only in MEFs. In NPCs, CpG-density of the regions is not an issue anymore. It is shown that bistability and cooperativity play a major role for regulation of H3K4me3 and are required for differentiation into different cell types.

Evidence is found that not only one mechanism exists which is activated in different differentiational stages to regulate the epigenome state. ESCs regions are mainly regulated by a CpG-dependent and cooperative mechanisms while NPCs H3K4me3 regions are generated by a CpG-independent mechanism. There is no hard boundary between presence of two mechanisms. Moreover, in MEFs both mechanisms seems to be present. While the CpG-dependent mechanism becomes switched off, the CpG-independent mechanism is activated.

In conclusion, H3K4me3 is not retained over large time scale such as many differentiation stages. This is required for regulation of cell type specific genes. Not only H3K4me3 but the other marks in this study take part in the regulation of genes. All of them are important for early developmental stages such as ESCs, MEFs, and NPCs. Thus, it is worthwhile to analyze the pattern established in the different cell types by not only one modification but a combination of the marks. Changes in the patterns give insights into regulation of the epigenome. However, this can not be done with the methods used in this chapter. Thus, more sophisticated methods are presented in chapter 5. Results for this methods on the same data set are presented in chapter 6 and chapter 7.
4.3 Summary

In this chapter, new knowledge had been acquired about the distribution of the H3K4me3, H3K27me3, and H3K9me3 in ESCs and H3K4me3 in ESCs, MEFs, and NPCs. It had been found that length and CpG-distributions are specific for each cell type and modification.

Length distributions are pair-wise distinct and thus, can be used to characterize the different modifications in the different cell types. Especially, the length distribution of H3K4me3 in NPCs had been found to differ in slope and shape from the length distribution of the other cell types. With a model assuming cooperative binding of the modifying complexes, the length distribution of H3K4me3 and H3K27me3 had been simulated. Simulations had shown that a shift to higher demodification allows to change from the length distributions in ESCs to the length distributions in MEFs. The switch to NPCs can be performed by re-defining the cooperative units of the epigenome by changing the arrangement of the insulators between the cooperative units. This had indicated an extensive re-organization of the chromatin during differentiation which has a high impact on the distribution of histone modifications.

Examination of the CpG-dependency had confirmed the CpG-dependency reported for the marks studied. The analysis, however, had shown that CpG-dependency exist in ESCs and MEFs only but not in NPCs. Additionally, the CpG-dependency had been described by a set of CpG-densities. Thus, reported characteristics of the marks are not static but changes over time. Furthermore, each of this marks can be characterized by CpG-density. Also each cell type can be described with a set of characteristic CpG-densities which changes during differentiation.

It has been found that modified regions either overlap completely or do not overlap at all while analyzing the overlaps of the different marks. Thus, defined regions of the genome are modified and are the same for all studied modifications. Regions covered by all three modifications had been shown to be significantly different in length and CpG-density than the regions covered by a subset of the modifications. The same had been discovered for regions covered by H3K4me3 in different combinations of cell type. Thus, the high dynamics of the epigenome had been shown and had revealed that modification states change in a specific manner rather than randomly. Different recruitment modes, thus, had been assumed for the different modifications in the different cell types.
Discovering Epigenetic Patterns

After the formulation of the histone code hypothesis, a lot of attention was paid to epigenetic patterns. The histone code hypothesis describes the interplay of epigenetics marks. The transcriptional state of a gene is not only determined by a single epigenetic marks but by a pattern of epigenetic marks. However, it is not clear which combination of epigenetic marks are possible and which information content is assigned to which pattern. Thus, comparison of multiple data sets covering several modifications and cell types provides new insights into epigenetic patterns and their regulation.

5.1 Finding a Reference

The analysis of ChIP-seq data as described in section 3.1 provides for each experiment a list of regions in form of genomic coordinates. Analyses as done in chapter 4 provides information of typical length distribution and CpG-density characteristics. However, this information can only be calculated for one experiment at once. Discovery of epigenetic patterns enforces the comparison of multiple ChIP-seq experiments. By default, the analyzed ChIP-seq data is not comparable. Several reasons, including technical and biological aspects, lead to slightly different border for the
same regions in different experiments. This calls for a method creating a reference for comparing ChIP-seq data of different experiments for cells of the same species.

A reference can be created by segmenting the genome. For each segment, the modification state can be extracted from each single experiment. The resulting vector of modification states provides an intuitive base for comparison and combines information about epigenetic patterns and changes during differentiation or aging. In general, any meaningful segmentation of the genome can be used. Depending on the information of interest, a specialized segmentation highlighting aspects of interest have to be designed.

In the case of discovering epigenetic patterns and understanding the underlying mechanisms, two main aspects play a key role, (1) combinations of marks and (2) their behavior during differentiation. Obviously, a proper segmentation highlights these aspects. Highlighting both aspects at the same time make it hard to identify the source of the patterns. The most fruitful approach will be able to highlight both aspects but not at the same time. Thus, two segmentations are created in the manner to keep them comparable but emphasizing either combinations of marks or behavior of marks during differentiation.

In the following, a whole genome segmentation (WGS) is introduced using ChIP-seq data itself as basis. Once the segments are defined the same ChIP-seq data sets are used to calculate the modification state of the segments. The WGS can be applied to segment the genome by a set of modifications from the same cell type. Doing so, the epigenetic state of the segments and its changes during differentiation are highlighted. Thus, this segmentation is referred to as epigenetic state segmentation (ES-segmentation). Segments for the ES-segmentation are determined by the following procedure:

1. For a cell type of reference, all \( n \) modification states (MS) are selected. Notice that the discretized representation of the chosen data holds a list of modified regions specified by the genomic position of the boundaries.

2. All boundaries of modified regions from the \( n \) modification states are projected onto the genome (ES-segmentation). The superposition of all boundaries subdivides the genome into segments. Short segments (here, length <200bp) are omitted in the applications since they are below the discretization limit (characteristic length of ∼200bp DNA per nucleosome).

3. As a consequence of segmentation, each segment is covered either 0% or 100% by each of the \( n \) modification data sets used for segmentation (see Figure 5.1c). It is therefore possible to describe a segment \( x \) by an \( n \)-dimensional binary vector which we call combinatorial epigenetic profile (CEP). Evidently, we can observe at most \( 2^n \) different combinatorial epigenetic profiles.

4. Information from the remaining \( n \times (m - 1) \), so far disregarded, modification states can be integrated where \( m \) is the number of cell types. We therefore intersect the segment \( x \)
Figure 5.1: Segmentation Procedure: WGS applied two 3 modification in 3 cell types to highlight both, changes during differentiation and combinations of different marks. a) Sets of modification states choose for ES- and EV-segmentation, respectively. b) Segments and CEPs obtained by EV-segmentation. c) Segments and CEPs obtained by ES-segmentation.

Combinations of epigenetic marks in one cell type can be studied using the CEPs. It is easy to distinguish frequent epigenetic patterns from very rare pattern or correlate them with genomic features such as CpG-density, gene expression, chromosomal distribution and overlap with genes and promoters. Thus, the ES-segmentation provides insights into the function (e.g. activation of housekeeping genes) and regulation (e.g. CpG-dependency) of epigenetic marks. Using the full EPs, the epigenetic patterns can be traced through the studied cell types. It provides information of retention of the epigenetic marks and patterns established in the cell type of reference. Furthermore, the modification of previously unmodified parts can be detected and changes in the fraction of the genome modified by one specific or all studied modification is easy to identify.

Analogously to the ES-segmentation, we can apply the WGS to one modification in several cell types. Instead of using the modification states from one cell type, all modification states of one mark are used to determine the segments. The EPs complemented with the coverage value of the modifications states from so far unused marks. It emphasizes the changes of the
chosen modification during differentiation while preserving the combinatorial features given by the other modifications in the same cell types. This variant of the WGS is called epigenetic variation segmentation (EV-segmentation). Using this segmentation, regions can be identified which are cell type specific or common to all studied cell types. Changes in the composition of the underlying DNA sequences are revealed by correlating the epigenetic profiles with the genomic feature and thus, indicating switches in the epigenetic regulation, e.g. from CpG-dependent to CpG-independent mechanisms or vice versa.

The WGS method presented above enables a comparison of epigenetic marks in different cell types but also allows for correlation with additional data. The CpG-density is already mentioned as example. Analogously to the CpG-density, any kind of DNA motif can be used for analyze. Genomic localization relative to specific motifs or genes can be analyzed by assigning presence and absence states to the segments, while the fraction of expressed genes can be used to analyze the correlation with gene expression. Any other kind of ChIP-seq data, e.g. bound proteins, can be handle in the same way as the modification states which were not used for the segmentation. Thus, the WGS provides a flexible tool to analyze epigenetic marks in the context of multiple cell types and different types of additional information.

Data obtained by WGS is visualized in three different ways highlighting different aspects of the data. The first visualization method utilizes SOMs to cluster and visualize the data and thus enables detection of epigenetic patterns. The second method uses scatterplots of pairs of modifications and filtering to specifically analyze the coverage values of the modifications not used for the segmentation. Lastly, k-means++ clustering is used to cluster the data for each cell type and visualize switching of the segments between the different clusters during differentiation with the help of starplots. While the methods by themselves are not new, the combination and adaptation to data is a result of this thesis. Thus, technical details can be found in appendix B.

5.2 Visualization and Clustering by Self-Organizing Maps

Segmentation of the genome as described in section 5.1 results in a data sets consisting of several thousands of segments. Each of them is characterized by a epigenetic profile and may be combined with additional information. The high number of segments desires an efficient clustering of similar data points which creates a representative data point for each cluster. Furthermore, formulating hypotheses requires a proper representation of this clusters for a visual examination. Both, visualization and clustering is combined in self-organizing maps (SOMs) [79]. It was already used to analyze high-dimensional transcriptomics data originating from several different tissues [167].

Details of the method and parameter settings, such as size and learning iterations, for the calculation of the SOMs are described in section B.1. The SOMs are calculated using the epigenetic profiles only. However, the segments are not only characterized by the epigenetic profile but also by a feature vector containing additional information such as segment length, CpG-density, or
expression state. The information from the feature vector is so far disregarded. Nevertheless, SOM-images can be calculated for the additional information. Therefore, representative values are calculated for each node and feature in the feature vector. Analogously to the meta-EP, a meta feature vector can be assigned to each node in the SOM. Using the meta feature vectors, additional SOMs are calculated.

Depending on the type of the feature and the information which should be represented, the values for meta feature vectors are calculated in different way. The population map is an indispensable map for the interpretation of the SOM-images and the formulation of hypotheses. It depict the number of segments assigned to the nodes. As a effect of the construction of the epigenetic values, the SOM is structured into island with many segments per node and borders with just a few segments per node. The borders and islands are easily detectable in the population map. Furthermore, it provides information about the frequency of epigenetic patterns. Those which are depict by nodes with many segments are more frequent than the ones depict by nodes with smaller numbers of assigned segments. To identify the islands and borders also on the other SOM-images, the border can be projected to them and nodes of the borders are colored in white.

To correlate epigenetic patterns with genes or gene groups, a gene map can be created. In the simplest way, analogously to the population map the number of genes assigned to segments in a node is counted and the counts are used to color the map. This is sufficient for showing the localization of all genes in the map. For groups of genes, this may be insufficient. Thus, the fraction of genes assigned to a node’s segments which belong to a certain gene group is shown in the map. This is important because the genes are not uniformly distributed over the SOM.

Gene groups can be derived from the Gene Ontology [7]. For example, one can choose the genes for specific biological process to highlight its epigenetic regulation. Other sources for gene groups can be experiments or specific motifs near to the genes.

To detect epigenetic patterns which are specific for certain chromosomes, chromosome enrichment maps can be calculated. Hereby, for each node the expected number of segments from a chromosome is calculated using the hyper geometrical distribution. Enrichment of the segments from a chromosome towards the expectation value is depict in the SOM-image. Thus, a SOM-image for each chromosome can be created.

For features for which numbers (either floating point number or integers) can be assigned to each segment, the representative value is defined as mean over all values of this feature of the segments assigned to the node. CpG-density and segment length are examples for this kind of features.

Expression data can be incorporated to explore the influence of epigenetic marks on transcription. By analyzing expression data, genes can be marked as present (expressed) or absent (not expressed). An expression value of 1 is assigned to expressed genes and 0 otherwise. For each segment, the average over all assigned expression values is calculated. Only segments with an average expression value of 1 are considered as active. In the present call map, SOM-nodes are
colored according to the fraction of segments considered as active.

By default the color scale ranges from blue for small values over green, yellow, and orange to red for large values. In the case of the segment length, the color scale is inverted. Thus, blue values mean long segments and red values mean short segments.

### 5.3 Scatterplot Matrices

Although the WGS is designed for one modification or one cell type, it also provides the coverage values for the remaining modification states in the data set. Using SOMs, these modification states are hard to analyze. Thus, another visualization method is required, enabling direct comparison of these coverage values. Scatterplot Matrices are powerful tool for this purpose.

Taking the data from WGS, the CEPs are converted into codes by concatenating their entries and interpreting these 0-1 strings as binary numbers. For example, in the case of 3 modifications which are all present, the segments with CEP = (1, 1, 1) has code 111 = 7. A segment with CEP = (1, 0, 1) is encoded in 101 = 5. Thus, in case of 3 modification, there are $2^3 = 8$ different codes.

Theoretically, plotting the coverage values of two modifications for each segment against each other using the code for coloring of the points would give a detailed view on the correlation of two modification states. However, this is only true for small data sets. In the case of WGS data, plotting several thousand points does not provide much insight into the data sets since it will be impossible to differentiate the individual points. Thus, more sophisticated methods have to be used which do not plot each individual data point but condense several close data point to one data point plotted.

![Figure 5.2: Scatterplot Matrix](image)

Figure 5.2: Scatterplot Matrix: Schematic depiction of a Scatterplot Matrix for three modifications and one cell type. a) Overall layout of a scatterplot matrix. H indicates histograms and SP stands for scatterplot. b) Histogram in diagonals. Each color represent the code defined from the reference cell type. Each pile represents the number of segments in a bin with a specific code. All piles of one bin are grouped together. c) Each non-diagonal element is a scatterplot of two modifications whereby the range of the coverage values are splitted into equally distributed bins. Each bin-pair (BP) is further subdivides as seen in d. d) Each bin-pair in a scatterplot consist of tiles colored by code. Saturation depict the amount of segments with a specific code with respect to the total amount of segments in the bin (local distribution) or the total amount of segments with the respective code (global distribution).
Taking the coverage values of the modifications, one can create histograms for each modification as well as scatterplots for each pair of modifications (see Figure 5.2). Arranging scatterplots and histograms in a matrix results in a scatterplot matrix. More details on scatterplot matrices can be found in section B.2. For each bin-pair, it is counted how much segments have a coverage value of the first bin for the first modification and a coverage value of the second bin for the second modification. In each of these pairs of bins, the segments can be grouped together by the code. In the scatterplot, either the relative frequency of segments falling in a bin-pair from the different codes is highlighted (global distribution) or relative frequency of the code in the bin-pair is depicted (local distribution). Each code has a different color (see Figure 5.2d and Table 5.1). Colors are chosen in a way that neighboring colors in the plots can be visually differentiated. Due to the fact that most of the segments coverage values near to 1 or near to 0, the relative frequencies are logarithmically scaled to enable the analysis of less frequent cases of values between 0 and 1. Each cell which does not belong to the diagonal contains the scatterplot for the respective pair of modifications (see Figure 5.2c). The diagonal contains the histograms for the single modifications (see Figure 5.2b). The scatterplot matrices can be extended by additional rows and columns for additional features with continuous values, e.g. CpG-density or length.

Table 5.1: Color Code for the Scatterplot Matrix: Assignment of colors to the codes are shown in the table below. Color name as well as hue is given for each value. For low and maximal coverage values, the color is shown.

<table>
<thead>
<tr>
<th>CEP</th>
<th>Code</th>
<th>Color</th>
<th>Hue</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>red</td>
<td>0°</td>
<td>-</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>green</td>
<td>120°</td>
<td>-</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
<td>blue</td>
<td>240°</td>
<td>-</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>yellow</td>
<td>60°</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>cyan</td>
<td>180°</td>
<td>-</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
<td>amber</td>
<td>30°</td>
<td>-</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
<td>guppy green</td>
<td>150°</td>
<td>-</td>
</tr>
<tr>
<td>111</td>
<td>7</td>
<td>purple</td>
<td>300°</td>
<td>-</td>
</tr>
</tbody>
</table>

For each segmentation, two scatterplot matrices can be created, one for the local distribution and one for the global distribution. The local distribution highlights the fraction of segments in a bin-pair having a certain CEP. With this information, one can answer the question whether segments with certain coverages by two modifications predominately have a certain state in ESCs. Consequently, it provides new insights into the dependency on the previous modifications state. The fraction of segments of given CEP which are located in a certain bin-pair is emphasized by the global distribution. Thus it is easily recognizable whether a certain CEP is preferentially turned into a certain modification state. Doing so, hypotheses about the existence of demodification/modification complexes can be formulated, recognizing certain epigenetic patterns.

Filters can be applied to select an interval of the coverage values from a certain modification. They are available for all modification which are not included in the code and additional feature values. Thus, a detailed look can be taken of segments, e.g., with a certain CpG-density or a
certain length. The distribution of the coverage values can be analyzed, e.g., for segments with high H3K4me3 coverage in MEFs only. Filtering allows to detect epigenetic patterns which are induced by presence or absence of a certain modification or characteristics that may have been missed in the whole data set.

The scatterplot matrix is implemented as a Java application. It loads the segmentation data as semicolon separated file. Each line represents a segment. The first line does not contain data but is used by the program to identify the file as data file. The second line contains the columns name and is used, for example, to name the filters in the GUI. The tool does not require specific hardware but runs on any computer with Java installed. Thus, it is independent of the operation system. However, to load big data sets sufficient memory is required.

5.4 k-means++ Clustering and Star Plots

The WGS is clustered using a k-means [90] variant called k-means++ [6]. The data is clustered according to the set of modifications from one cell type or according to one modification using the data from all cell types. Since the clustering for the binary data is trivial, only the parts of the epigenetic profile with continuous values are used for clustering. The distribution of the clusters are depicted using starplots [35, 39].

The k-means clustering algorithm requests the data to cluster, a distance function between the data points and the number of clusters $k$ as input. The data points are the EPs. As distance function, the euclidean distance is used but restricted to the part of the EPs used for clustering (subprofile). In the case of the ES-segmentation, for each cell type one clustering run is done. At this, the subprofiles are composed of the modifications measured in the chosen cell type. Analogously, the clustering is done in the case of EV-segmentation; a subprofile is composed of the measurements of one modification in the different cell types available in the data sets. Given a subprofile of length $l$, the number of clusters $k$ is set to $2^l$ by default. Here, it is assumed that segments are either modified or unmodified by the modification although the coverage values are continuous. Thus, for each modification, there are two states leading to $2^l$ different combinations of states for $l$ modifications in the used subprofile. However, if necessary, $k$ can be set to any values greater than 1. Details on the clustering can be found in section B.3.

For each cluster, it is counted how many segments clustered into this cluster correspond to which CEP. Thus, the clustering results are compressed to a matrix $C$ such that $C_{i,j}$ contain the number of segments clustered in cluster $i$ with CEP code $j$. Based on this counts, the global distribution matrix $G$ and the local distribution matrix $L$ can be retrieved whereby $G_{ij} = C_{ij} / \sum_{k \in \text{clusters}} C_{kj}$ and $L_{ij} = C_{ij} / \sum_{k \in \text{codes}} C_{ik}$, respectively.

To visualize the clustering result, starplots are used as shown in Figure 5.3. All axes range from 0 to 1 (see Figure 5.3b). The base line representing 0 is put some pixel away from the middle of the star plot to make CEPs with a local or global distribution value of 0 easy to observe. For each
Figure 5.3: Starplot: Representation of distribution of a specific code to the different clusters. a) Each axis represents a cluster. The clusters number is shown at the end of each axis. b) Thin gray lines placed in 0.1 steps starting from 0 as innermost line (not starplot center) until 1 as outermost line. c+d) Polygons indicate the distribution of the CEPs in the clusters. c) Vertexes position represents the fraction of segments in the cluster with a specific CEP (local distribution). d) Vertexes position represents the number of segments with a specific CEP normalized by the number of segments in the axis’ cluster (global distribution).

CEP, a polygon is drawn (see Figure 5.3c and Figure 5.3d). To distinguish between the different codes, each polygon is drawn in a different color (see Table 5.2). In total, \( c + 1 \) (\( c \) - number of codes) star plots are created for each distribution. One star plot containing all polygons and for each of the \( c \) codes one star plot containing the polygon of the code only. Starplots with all codes provide an overview on the distributions of the codes and can be used to compare the codes with each other. Single code starplots enable analyses of the distribution of one code without interference with the polygons of the other codes [171].

Table 5.2: Color Code for the Starplots: Assignment of colors to the codes are shown in the table below. Color name as well as colors code in RGB system are given.

<table>
<thead>
<tr>
<th>CEP</th>
<th>Code</th>
<th>Color</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>red</td>
<td>(255,0,0)</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>blue</td>
<td>(0,0,255)</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
<td>maroon</td>
<td>(128,0,0)</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>rose</td>
<td>(255,0,128)</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>cyan</td>
<td>(0,255,255)</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
<td>dark green</td>
<td>(0,96,64)</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
<td>green</td>
<td>(0,255,0)</td>
</tr>
<tr>
<td>111</td>
<td>7</td>
<td>chrome yellow</td>
<td>(255,160,0)</td>
</tr>
</tbody>
</table>

The local distribution allows the inspection of the composition of one cluster (see Figure 5.3c). The main source CEP of each cluster can be found as well as CEPs which do not contribute to the clusters segments. Regarding the ES-segmentation, dependencies of previous states of the epigenome can be detected and compared among different cell types. Impossible or very rare transitions from one CEP in one cell type between into another CEP in the next cell type combined with the most frequent transitions allow for conclusions about the main mechanisms for differentiation into the different cell types. In the case of the EV-segmentation, combinations of
different epigenetic marks and their changes during development can be analyzed. Mechanisms which are dependent on the existence of a certain mark in specific developmental stages can be detected by means of this method by comparing the two stages which influences each other. Here, focus is put onto a reference modification while the presence of a second modification is examined carefully.

To analyze in which cluster a CEP mainly appears, the global distribution star plots are designed (see Figure 5.3d). Thus, the prevalent successive states of a certain CEP can be observed. Since the fraction is calculated according to the number of segments with a certain CEP, also assumptions about very rare cases can be made. Hypotheses about readers detecting certain states and mechanistically turning them into other states can be made using the ES-segmentation data. Also the ability to retain a specific CEP can be analyzed easily. Co-regulation of modifications could be examined by applying clustering and visualization to the WGS for the EV-segmentation.

This method is implemented as Java application and uses the same framework as the scatterplot matrix. For this reason, the input format for both tools is the same. Requirements are similar to the requirements of the scatterplot matrix tool; Java has to be installed and sufficient amount of memory has to be available to load big data sets. However, even with the biggest data set used in this thesis, the program uses not more than 3 GB [171].
The last chapter presents several methods to analyze and visualize epigenetic data. They are designed to highlight epigenetic patterns. All of them can be combined with the segmentation method presented.

Segmentation was applied to the mouse data set already described in chapter 4. Biological results generated with the different methods are described in the following section for the ES-segmentation.

ES-segmentation is designed to examine the epigenetic state of a reference cell type. Thus, each segment’s CEP represents the combination of marks in the reference cell type. The modification state of the segment can be tracked in the other cell types. Epigenetic patterns in the combination of marks can be investigated with the help of SOMs, Scatterplot Matrices, and Starplots. ES-segmentation was applied to segment the whole genome with respect to H3K4me3, H3K27me3 and H3K9me3 modified regions in ESCs.
6.1 SOMery

The set of EPs from the ES-segmentation with respect to the modification in ESCs was subjected to SOM-learning as described in chapter 5. The population map in Figure 6.1a shows the number of EPs assigned to each of the 1600 SOM-nodes after SOM-training. One can easily identify a certain number of green islands with a density of at least 10 EPs per node that are separated from each other by sparsely populated nodes with less than 10 EPs per node (dark blue).

![Population map and CEP cluster map](image)

**Figure 6.1:** SOM-properties after ES-segmentation: a) The population map shows the number of EPs per node. It is logarithmically scaled and ranges from 0 (blue) to 150,000 (red). b) CEP cluster map. Islands with more than 10 EPs per node are shown in white and are labeled with their CEP. Borderlines with less than 10 EPs per node are shown in black. EPs with the same CEP cluster together in one island, except for ES110, ES010 and ES011 which split into two separated islands each (postfixes a,b). Colored are islands of active chromatin (ES100, green) and poised chromatin (ES110a,b, yellow).

It can be found that all EPs of each particular island refer to the same CEP. Figure 6.1b assigns the CEPs to the respective subregions of the map. All eight possible CEPs are present and allocated to eleven subregions of the SOM. Three CEPs, namely ES110, ES010, and ES011 locate to two different subregions of the SOM each, indicated by the postfixes “a” and “b”.

The largest island of the SOM refers to the CEP ES000 representing the unmodified state in ESCs. It corresponds to 97% of the total chromatin. The CEP ES100 marks active chromatin in ESCs, i.e. only H3K4me3 modified (see green island in Figure 6.1b), ES110 indicates poised chromatin, which is H3K4me3 and H3K27me3 modified but not H3K9me3 modified (yellow islands in Figure 6.1b). All other CEPs mark different repressive states of chromatin.

Figure 6.2 shows the complete atlas of SOM-images after ES-segmentation, three images for each modification and each cell type, respectively. Here, the border lines of the CEP clusters are shown in white. The three SOM-images for ESCs provide the distribution of the CEPs ES1**, ES*1* and ES**1 in dark red, i.e. the distribution of segments modified at least by either H3K4me3, H3K27me3 or H3K9me3, respectively (compare with Figure 6.1b). For MEFs and NPCs, the EPs adopt continuous values due to the particular coverage of predefined segments with the respective modification. This causes the color gradient in Figure 6.2d-i.

Figure 6.2 demonstrates a genome-wide reorganization of the modification patterns during
Figure 6.2: SOM-atlas after ES-segmentation: a)-i) Each tile in the mosaic images is colored by the average coverage of the segments assigned to it by the respective modification. High and low coverage is depicted in red and blue, respectively (linear scale). White tiles indicate the borderlines between islands with different CEPs (see also Figure 6.1). De novo formation of H3K4me3 and H3K27me3 in MEFs is a frequent event (black arrows in d and e). Many of these marks remain stable in NPCs (arrows in g and h). The arrow in panel f indicates poised chromatin de novo formed in MEFs. It partially turns into repressed chromatin in NPCs by H3K9me3 modification (arrow in i). b'), e'), and h') Subregions corresponding to the same CEP, as ES011a and ES011b, differ by their H3K27me3 status in MEF and NPC.

specification of ESCs into MEFs and NPCs. This reorganization includes de-modification of segments (red areas turn to blue ones) as well as de novo modification (blue areas turn to red ones). These changes appear to follow different rules for the different marks.

De novo trimethylation of H3K4 and/or H3K27 is observed during differentiation of ESCs into MEFs (see arrows in Figure 6.2d and 6.2e indicating regions which turn from blue in ESCs into red in MEFs). Many of these marks found in MEFs but not in ESCs are also observed in NPCs (see arrows in Figure 6.2g and 6.2h). Interestingly, most of the associated segments are in a poised chromatin state in MEFs (H3K4me3 and H3K27me3 modified but not H3K9me3 modified) but in an inactive state in NPCs (additional modified with H3K9me3, arrows in Figure 6.2f and 6.2i). In contrast to H3K4me3 and H3K27me3, for which the amount of modified segments is roughly constant in all three cell types, a large-scale H3K9 de-methylation is observed in MEFs and NPCs compared to ESCs (all ES**1 islands turn, at least partly, into blue). De novo
H3K9 tri-methylation of segments without H3K9me3 in ESCs (ES**0) is only rarely observed. Therefore, it turns out that H3K9me3 remodeling is almost exclusively restricted to segments carrying this mark in ESCs. Interestingly, the unmodified chromatin in ESCs (ES000) largely remains unmodified (colored in blue) also in MEFs and NPCs. Thus, while about 97% of the total chromatin is located in this region, changes in the chromatin state with respect to H3K4me3, H3K27me3 and H3K9me3 is mainly observed in the remaining 3% of the chromatin.

It was found that SOM-training splits part of the islands referring to a particular CEP into two subregions distinguished with the postfixes “a” and “b” (see Figure 6.1b). Remarkably, all these paired subregions (ES110a/b, ES010a/b, and ES011a/b) strongly differ in their H3K27me3 modification status in MEFs and NPCs (see Figure 6.2e' and 6.2h'). We labeled subregions with a high H3K27me3 coverage in MEFs and NPCs with a and those with a low coverage with b. The finding that segments with the same CEP in ESCs can acquire different modification states in MEFs compared to NPCs opens up interesting questions about the underlying mechanisms. Examples are possibly different modes of recruitment of the marks in ESCs or targeted (de-)modification events during differentiation into MEFs and NPCs. Further, these reorganization processes may depend on other histone marks not included in the current study.

Recruitment of modifications and its regulatory consequences can be further analyzed using complementary information summarized in supporting maps. Figure 6.3 shows supporting maps displaying the average segment length (a), the average CpG-density of a segment (b), the expression status of genes overlapping with a segment (c) and the distribution of segments overlapping developmental genes (d). Note that none of these additional data were used for SOM-training. Instead, they were projected onto the SOM-topology which is governed solely by the EPs.

In the segment length map (Figure 6.3a) one observes that long chromatin segments mostly accumulate in the island referring to segments unmodified in ESC (ES000 in Figure 6.3a). However, a small region of long segments is also observed in the island referring to triple-modified segments in ESCs (ES111). It has been demonstrated that the appearance of long modified segments may refer to cooperative mechanisms in recruitment of modifications [44, 126]. Accordingly, the appearance of long modified segments in triple-modified ESCs suggests cooperative recruitment of H3K4me3, H3K27me3 and H3K9me3 to chromatin.

Inspection of the CpG-density map (Figure 6.3b) reveals that segments carrying H3K4me3 in ESCs (ES1**) can be CpG-enriched. This has already been observed by Mikkelsen et al. [99] and is consistent with the finding that H3K4me3-modifying complexes include a binding motif for unmethylated CpGs [4]. However, beside in (ES1**) segments, CpG-enrichment is found for a subset of the unmodified segments (ES000) only. In particular, it is not found for segments that become de novo modified with H3K4me3 in MEFs and NPCs as e.g. part of the segments of (ES011). Hence, H3K4me3 recruitment in MEFs and NPCs is obviously not associated with high CpG-densities which suggests a different mechanism compared to the CpG-dependent recruitment in ESCs discussed above. Note that this alternative mechanism may exist also in ESCs, since
Figure 6.3: Supporting maps after ES-segmentation: a) Average segment length per node ranging from 200 (red) to 135,000bp (blue, linear color scale). b) Average CpG-density of the segments per node ranging from 0 (blue) to 0.08 CpGs/bp (red, linear color scale). c) Fraction of segments that exclusively overlap with genes, which are significantly expressed (present) in all three cell types (low-to-high refers to blue-to-red). d) Fraction of segments overlapping with genes associated with the GO-term cellular developmental process. A logarithmically scale is used ranging from 0.004 (dark blue) to 0.3 (red). Tiles of nodes without any overlap are colored in white.

ES111 and ES101 in part also contain segments with no CpG-enrichment. Correlations between H3K4me3 and CpG-density will be discussed later in the context of EV-segmentation.

H3K27me3 recruitment has been associated with high CpG-density as well [99]. Moreover, a H3K4me3 dependent mechanism has been suggested [82]. Indeed, in all cell types a large fraction of segments with a high H3K27me3 coverage lies in regions of the SOM that overlap regions of high H3K4me3 coverage. However, this overlap is not exhaustive. This is most obvious for segments of ES010 and ES011 in ESCs (see Figure 6.2a and 6.2b). Again, this observation opens up interesting research questions: it may either indicate that H3K4me3 is not necessarily required for H3K27me3 recruitment, or that previously existing H3K4me3 marks were removed at an earlier stage of development. Moreover, all ES01* islands lack CpG-enrichment. Consequently, H3K27me3 recruitment is also not necessarily associated with local CpG-enrichment. As the ES01* segments are predominately short segments (<1kb), and thus cooperative binding of modifying complexes is limited, we suggest H3K27me3 recruitment to ES01* to be sequence specific but
not CpG-dependent. Different binding motifs for segments of ES01*a and ES01*b, would explain the observed differences in their modification status in MEFs and NPCs. However, in the case of ES110, the difference in the CpG-density (ES110a low, ES110b high) may contribute to the different modification status in MEFs and NPCs.

Note that the occurrence of e.g. short ES01* segments could result from insufficient saturation of H3K27me3 ChIP-seq libraries. However, if this was the dominant effect, we would not expect to measure a large number of long H3K27me3 segments which are also H3K4me3 modified (ES11*). Instead we would expect that H3K4me3 modified segments are randomly associated with H3K27me3 modified segments. It is one of the advantages of our combinatorial method that it is quite robust against this kind of measurement biases in the data for single modifications.

The chromatin-associated information considered so far can be easily correlated with the expression status of genes. In Figure 6.3c and 6.3d, this is demonstrated for two important classes, namely housekeeping and developmental genes. We find that ESC chromatin associated with particularly high CpG-density, is either active (ES100) or can be at least viewed as not actively silenced (spots in ES000). Strikingly, as seen in Figure 6.3c, it associates with “active genes”, namely genes that are expressed in all three cell types.

According to Mikkelsen et al., genes associated with high CpG-density and monovalent H3K4me3 in ESCs may be considered as “housekeeping genes”. Actually, compared to all other islands in the SOM, we found that ES100 shows a more than 10-fold enrichment in segments that have a housekeeping probability >0.75 according to De Ferrari et al. [43]. By contrast genes associated with high CpG-density and bivalent chromatin (H3K4me3 and H3K27me3 modified) in ESCs have been linked to genes with more complex expression patterns among them key developmental genes [99]. Indeed, we see that segments of ES110b are associated with developmental genes (see Figure 6.3d). Strong enrichment, however, is mostly found in the CpG-rich spots of ES000.

Figure 6.4: Chromosomal enrichment maps for ES-segmentation: Enrichment of segments from chromosome 1 and chromosome X in the nodes of the SOM are shown in a and b, respectively. The logarithmic scale ranges from 0.1 to 4.8.
Chromosomal distributions are calculated for all chromosomes of the genome. Most of the chromosomes are not significantly enriched in any of the spots, e.g. chromosome 1 (see Figure 6.4a). Appendix C contains chromosomal-specific population maps and chromosomal enrichment maps for all chromosomes of the mouse genome. In this section, only chromosome X showing a specific enrichment pattern and chromosome 1 for comparison is shown. Chromosome X shows enrichment pattern which correlate well with the location of spots (see 6.4b). Enrichment is predominantly found in ES001, the spot associated with H3K9me3 mark only. Depletion is found in spots associated with active or poised chromatin, namely ES110 “a” / “b” and ES100. So far, the origin of the phenomena is unknown and thus, can not be explained by literature.

6.2 Scatterplot Matrix

Scatterplot matrices are created with 50 bins for each histogram and 50×50 bin-pairs for each scatterplot. The image for the global distribution of the ES-segmentation data can be found in the supplemental at the web¹ or on the enclosed CD. Scatterplots are not shown in the thesis itself, since details can only be observed on large images in high-resolution. However, an overview can be gained from the Figures D.1, Figure D.2, and Figure D.3 in the appendix of the dissertation. The scatterplot matrices for the global and local distribution for all segments and segments with a length of at most 10.000nt are shown as down scaled figures.

The diagonal contains the histograms for the coverage by the three modifications in MEFs and NPCs for the segments defined by H3K4me3, H3K27me3, and H3K9me3 in ESCs. Non-diagonal elements contain scatterplots for each pair of modification data sets. In Table 6.1, the mapping of color to code and CEPs in ESCs is shown.

Table 6.1: Color of ES-CEPs for the Scatterplot Matrix: For each CEP of the ES-segmentation the presence of marks as well as the decimal code is shown. Color used in the scatterplot matrix for respective code is named and shown.

<table>
<thead>
<tr>
<th>ES-CEP</th>
<th>H3K4me3</th>
<th>H3K27me3</th>
<th>H3K9me3</th>
<th>Code</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>ES001</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>ES010</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>blue</td>
</tr>
<tr>
<td>ES011</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>yellow</td>
</tr>
<tr>
<td>ES100</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>cyan</td>
</tr>
<tr>
<td>ES101</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>amber</td>
</tr>
<tr>
<td>ES110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>guppy green</td>
</tr>
<tr>
<td>ES111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>purple</td>
</tr>
</tbody>
</table>

A significant difference can be observed between scatterplots of modifications of the same cell type and scatterplots of modifications of different cell types. While in the latter one, highly populated nodes are located at the borders of the scatterplots only, scatterplot of modification of

¹http://evodevo.bioinf.uni-leipzig.de/~lydia/thesis_supplemental/supplemental.html
Figure 6.5: Coverage of Highly Populated Nodes: For any position of a node in the scatterplot matrix which is highly populated, the structure of the segments is depicted: a) top and right border of the scatterplot, b) bottom and left border of the scatterplot, c) diagonal of the scatterplot if pattern is produced by two independent processes for the two modifications, d) diagonal if it is result of two dependent processes or the same process.

The same cell type also have highly populated nodes in the diagonal between the bin-pairs (0,0) and (1,1). Nodes at the border correspond to segments which are covered by one modification totally and only partly by the second modification ((x,1) or (1,x) nodes) or segments which are not covered at all by the one modification and are partly covered by the other modification ((x,0) or (0,x) nodes, see Figure 6.5a and Figure 6.5b). This is not surprising since two modification overlap either completely or not at all. Thus, enrichment of the segments at the borders and especially in the edges as seen in the data has to be expected. Nodes at the diagonal ((x,x) nodes) are segments which are covered by both modification to the same extent. In theory, 2 cases can be assumed: (1) two independent processes which lead to the same coverage by either modification (see Figure 6.5c) or (2) two processes dependent on each other or the same process leading to this effect (see Figure 6.5d). Most likely is the second case since a high number of such segments are observed. Only a small amount of segments are located apart from the diagonals and borders. Thus, the pattern observed in the scatterplot matrices suggests a reorganization of the segments between MEFs and NPCs. Such a reorganization may occur due to structural changes in the organization of the chromatin within the nucleus. Referring to the model explained in chapter 4, the cooperative units may be changed by switching the boundary elements. For example, different boundary elements are used in MEFs and NPCs and thus, different loops defining the cooperative units are formed. This hypothesis is depict in Figure 6.6.

The distribution of the coverage values of the modifications MEFs can be seen in the histograms for MEFs. Most of the segments which were already modified in ESCs, especially 111, 110, and 011, are still highly modified in MEFs. In the histogram for H3K9me3 in MEF also 101 segments are enriched at high coverage values. Segments which were unmodified in ESCs are less often modified by H3K9me3 than by H3K4me3 and H3K27me3 in MEFs. Thus, segments modified in ESCs mostly stay modified by H3K4me3 and H3K27me3 in MEFs. Several segments which where unmodified in ESCs (code 0, color red), are found at high coverage values in the histograms for MEF H3K4me3 and MEF H3K27me3. Thus, many segments are newly modified in MEFs. The
Figure 6.6: Loop Reorganization Hypothesis: The usage of different boundary elements (here depicted as green or pink) in different cell types leads to the different chromatin structures due to formation of different loops. Thus, cooperative units (depicted as blue, red, and black lines) of one cell type are destroyed and new cooperative units are defined by the new boundary elements.

overall modification level is therefore increased in MEFs.

In MEFs, the histogram of H3K9me3 differs from the histograms of H3K4me3 and H3K27me3. Segments with code 0 (unmodified in ESCs) are mostly found at low coverage values and only rarely at high coverage values. Thus, newly H3K9me3 marked segments are rare in MEFs. Segments which were modified by H3K9me3 in ESCs are found at high coverage values but also frequently at low coverage values. This shows that the H3K9me3 marks are massively lost in MEFs.

In NPCs, the histograms look more diverse than in MEFs. The histogram for H3K4me3 in NPCs shows that segments unmodified in ESCs are located at the low coverage values and thus, indicates that H3K4me3 is only set on segments which were already modified in ESCs. The unmodified part of the genome in ESCs stays mostly unmodified in NPCs as well. Many segments which carry H3K4me3 in ESCs are now unmodified (located at low coverage values). Especially bivalent segments (guppy green), can be found at low coverage values. Thus, bivalently marked domains in ESCs lose the H3K4me3 and are switched into repressed state. The opposite case, namely bivalently marked domains in ESC switch to active state can be found only rarely.

The histogram for H3K27me3 in NPCs shows that regions which were unmodified in ESCs are not modified in NPCs. However, high amount of segments locate at high coverage values which already carry H3K27me3 in ESCs except segments with code 010. Thus, H3K27me3 is retained in NPCs when one of the other marks was set at this segments. Surprisingly, segments which solely carry H3K27me3 in ESCs are enriched in the middle of the histograms and thus, are an exception from the overlapping principle of totally or nothing. In other words, segments with code 010 in ESCs, loose the mark on approximately half of the segment length.

Coverage of H3K9me3 in NPCs shows that many segments which were modified by H3K9me3 in ESCs are only slightly covered by H3K9me3 in NPCs. This emphasizes the loss of H3K9me3 which is even stronger in NPCs than in MEFs. While H3K9me3 seems to play a great role in ESCs, in MEFs and NPCs H3K9m3 has only a minor role.

The histograms of the CpG-density can be parted into three ranges, namely high CpG-density (> 0.77), middle CpG-density (> 0.015 and < 0.077) and low CpG-density (< 0.015). This is best seen in the local distribution of the CEPs. High CpG-density is dominated by red (000) and
cyan (100). This is according to the literature reporting that CpG-islands are often modified with H3K4me3 to activate transcription [99]. Since many CpG-islands are located near to the transcription start sites of housekeeping genes, these segments may be associated with housekeeping genes. At middle CpG-density, not only cyan and red segments can be found but also pink segments (code 7, ES111). Segments modified with all three marks in ESCs have mostly a high CpG-density but do not have the highest CpG-density possible in the genome. Low CpG-density is not dominated by a specific color. Thus, any combination of marks can be found at low CpG-density segments.

Since many of the unmodified segments and some outliers of the modified segments show extreme lengths, the scatterplots and histograms for the length do not reveal many detail. For this reason, segments were filtered by length to obtain a detailed look at segments with length smaller 10000.

In the length histogram filtered for lengths smaller 10000, large segments are mostly unmodified (red) or modified with all three marks (pink) in ESCs. Segments longer than 4400nt (approx. 22 nucleosomes) do not originate from another state in ESCs. Other CEPs in ESCs are mainly observed for segments shorter than 2000nt. Thus, setting all three marks on a segment may stabilize the modifications on these segments and so, enable the formation of long modified regions.

The scatterplot of CpG-density and length shows that high CpG-density correlates with short lengths. High CpG-density is mainly found for segments shorter than 2000nt but only very short segments with high CpG-density were modified in ESCs. Segments with a length smaller than 2000nt and a CpG-density smaller than 0.023 can have any state in ESCs. The same can be observed for segments shorter than 400nt and a CpG-density smaller than 0.06. Long modified segments (pink) have a mid-level CpG-density. Surprisingly, long segments with a CpG-density lower than 0.015 where not modified in ESCs. Taken together, segments in ESCs are modified with any modification if they are short but do not have the highest CpG-density possible in the genome. Modified segments are only extended if they have the highest CpG-density possible on this length (no long segments at all with CpG-density larger than 0.05).

In the scatterplots for length and the modifications in MEFs, it can be seen that the long regions modified by all three marks are retained to a high amount (pink segments of high length are located above a coverage of 0.8 for all of the three modifications in MEFs). Short segments are mostly unmodified or completely modified. Only a small fraction has intermediate coverage values. Thus, they are either retained (coverage near to 1) or lost (coverage near to 0). The retention of H3K27me3 and H3K4me3 is stronger on segments shorter than 800nt than on segments with a length between 800nt and 1400nt. For H3K9me3, this interval is extended to 400nt to 1400nt. Thus, modifications on segments with intermediate length can be harder obtained than short and long segments.

Modifications on long segments can not be retained in NPCs at all. Also short segments are mostly unmodified (low coverage values for all three modifications). Some specific patterns
can be found for the short segments which are modified by H3K4me3 and H3K9me3 in NPCs. In both scatterplots, namely length and NPC H3K4me3 as well as length and NPC H3K9me3, high coverage is dominated by green and yellow. Thus, segment which carry H3K9me3 only (green, ES001) or H3K9me3 and H3K27me3 (yellow, ES011) in ESCs are in NPCs modified by H3K9me3 and additionally acquire H3K4me3 in NPCs. While these segments were able to get DNA methylated in ESCs (recruited by H3K9me3 marks), they are protected against DNA methylation in NPCs (H3K4me3 mark).

The scatterplots and histograms for the local distribution can also be found in the supplemental. Histograms as well as scatterplots for the modifications within MEFs have a similar structure. Most of the segments with high coverage values were modified with all three modifications in ESCs and vice versa most of the segments with low coverage values are unmodified in ESCs as well. In all three histograms, a clear border can be seen such that segments with lower coverage values were mostly unmodified in ESCs and segments with higher coverage values were often modified by all three modifications. While this border is located at a coverage value of 0.68 for H3K4me3 and H3K27me3, a coverage value of 0.4 marks the border in the H3K9me3 histogram. Segment carrying H3K9me3 only or H3K9me3 and H3K27me3 in ESCs are the most abundant segments around the borders in all three histograms. In the scatterplots, in the (1,1) corner are almost only pink segments (ES111) and in the (0,0) corner are almost only red segments (ES000). Between this two corners, namely the diagonal between (1,0) and (0,1) is not dominated by a single color but by green (ES001) and yellow (ES011) segments. However, also the other color can be easily observed. Taken together, the ESC modification states of this segments with all three modifications or without modifications are retained in MEFs. Changes in the modification state occur mostly for the segments carrying H3K9me only or H3K9me3 and H3K27me3 in ESCs. They are partly modified by all three marks in MEFs.

Scatterplots and histograms within NPCs have the a similar structure as the one within MEFs. However, high coverage values are not only dominated by pink segments. Segment which have code 7 (ES111, pink), 3 (ES011, yellow), or 1 (ES001, green) are the most abundant codes for modified segments in NPC. Thus, also the (1,1) corners of the scatterplots are colored in pink, yellow, and green. The composition of the border in the histograms do not change. Thus, a great part of the modifications in ESCs is lost in NPCs, especially the modification of segments which were modified by all three marks in ESCs. Greatest loss can be observed for H3K9me3 which seems to be an ESC-specific mark.

In the scatterplots of modifications of different cell types, no specific pattern can be observed. In most of the nodes any color can be observed. Dominating colors are mainly explainable with the distribution of the codes in the histogram.

In the histogram of the CpG-density, three levels can be observed as in the global distribution. Low CpG-density is mostly found on segments unmodified in ESCs. High CpG-density is mostly found on segments unmodified in ESCs as well. Nevertheless, segments which carry H3K4me3 only
in ESCs can be observed at this level. The mid-level is mainly composed of segments unmodified in ESC or segments with the CEP 111 in ESCs (pink). Segments which are in an active state in ESCs (ES100, cyan) are less frequent than segments without or with all modifications but still significantly more abundant than all the other CEPs in ESCs. Since most CEPs in ESCs do not occur very often in the whole data set, they do not dominate a specific CpG-density but locate to low and mid-level CpG-density. In the scatterplots for CpG-density and the modifications, it can be seen that the combination of low CpG-density and high coverage values is dominated by green and yellow which are segments with CEP ES001 and ES011, respectively. High coverage and mid-level CpG-density, is mostly composed of segments which are modified by all three modifications in ESCs. Thus, CpG-density plays a key role in the regulation of the three marks.

Again, the data sets were filtered by length to obtain a detailed look at segments with length smaller 10000nt, namely the length range in which most of the modified segments are observed (see chapter 4). Scatterplots of length and modifications of MEFs clearly show that long segments are either unmodified (coverage near 0) or completely modified (coverage near 1). Only for short segments (<2000nt), the whole range of coverage values can be observed. In NPCs, long segments locate to coverage near 0 only and thus, long segments are unmodified in NPC. Modifications on long segments are lost during differentiation into NPCs. Short segments modified in ESCs stay modified in MEFs and NPCs but may switch the set of modification. The important and surprising fact is that in contrast to the long segments, short segments are not only demodified but also modified and thus do not switch into the CEP 000. They are still regulated by the studied marks. Since most of the segments are unmodified in any cell type little information can be drawn from local distribution of the length because most of the nodes are dominated by red, i.e. unmodified segments.

6.3 k-means++ Clustering and Starplots

ES-segmentation data is analyzed by k-means++ and starplots. Results for the clustering of the coverage values for H3K4me3, H3K27me3, and H3K9me3 in murine embryonic fibroblasts and for H3K4me3, H3K27me3, and H3K9me3 in neuronal progenitor cells are shown below. Codes for the modifications in ESCs are assigned to colors used to colorize the polygons in the starplot. Codes to color assignments are shown in Table 6.2.

Murine Embryonic Fibroblasts The starplots resulting from clustering are depicted in Figure 6.7 and the centroid coordinates are shown in Table 6.3. Given the coordinates of the centroids, the dominant combination of modifications can be assigned to each cluster and thus, codes can be abstracted from them for each cluster (Table 6.3). All eight possible combination are present.

Figures E.1a and E.2a clearly show that most of the segments unmodified in ESCs stay unmodified in MEFs (87%) and the main source of unmodified MEF segments are segments with
Table 6.2: Color of ES-CEPs for the Starplots: For each CEP of the ES-segmentation, the presence of marks as well as the decimal code is shown. Color used in the starplots for respective codes are named and shown.

<table>
<thead>
<tr>
<th>ES-CEP</th>
<th>H3K4me3</th>
<th>H3K27me3</th>
<th>H3K9me3</th>
<th>Code</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>ES001</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>blue</td>
</tr>
<tr>
<td>ES010</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>maroon</td>
</tr>
<tr>
<td>ES011</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>rose</td>
</tr>
<tr>
<td>ES100</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>cyan</td>
</tr>
<tr>
<td>ES101</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>dark green</td>
</tr>
<tr>
<td>ES110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>green</td>
</tr>
<tr>
<td>ES111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>chrome yellow</td>
</tr>
</tbody>
</table>

Figure 6.7: Starplots for Clustering of the Modifications in MEFs: a) Local distribution of the CEPs observed in ESCs in the clusters. b) Global distribution of the CEPs observed in ESCs in the clusters.

ESC code 000 (92%). This was already shown in [145], however the exact amount could not be determined using the SOMs. Further, it was not detectable in the SOMs that a large amount of segments carrying a solitary mark in MEFs (100: 50%, 010: 45%, 001: 35%) were not modified at all in ESCs.

The combination of all three marks at the same segment can be stably retained and thus amounts to 66% of the segments modified by all three marks in MEFs. Fully modified segments in ESCs are also the main sources of segments with H3K4me3 and H3K9me3 (code 101: 41%) and for segments with H3K27me3 and H3K9me3 (code 011: 39%). Thus, it could be concluded that segments marked with these combination of marks in MEFs are a result of histone demodification during differentiation.

Surprisingly, 34% of the segments in MEF which are in a bivalent state had switched their modification state completely. They originate from segments with code 001 in ESCs (Figure E.1b).
Table 6.3: Centroids in MEF: For each cluster, the coordinates in MEF are given as well as the abstracted code.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroids in MEFs</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H3K4me3</td>
<td>H3K27me3</td>
</tr>
<tr>
<td>1</td>
<td>0.019</td>
<td>0.025</td>
</tr>
<tr>
<td>2</td>
<td>0.789</td>
<td>0.340</td>
</tr>
<tr>
<td>3</td>
<td>0.514</td>
<td>0.219</td>
</tr>
<tr>
<td>4</td>
<td>0.180</td>
<td>0.191</td>
</tr>
<tr>
<td>5</td>
<td>0.802</td>
<td>0.849</td>
</tr>
<tr>
<td>6</td>
<td>0.138</td>
<td>0.614</td>
</tr>
<tr>
<td>7</td>
<td>0.386</td>
<td>0.814</td>
</tr>
<tr>
<td>8</td>
<td>0.903</td>
<td>0.913</td>
</tr>
</tbody>
</table>

Since H3K9me3 is correlated with DNA methylation and H3K4me3 can not be set on histones associated with methylated DNA, these results indicate that H3K9me3 modified segments are not necessarily DNA methylated, but stay unmodified to switch into a bivalent state for later activation or repression. A second source for bivalent segments in MEFs are segments with code 011 in ESCs (25% of the bivalent segments (Figure E.1d). This might indicate that the switching is performed in steps setting a H3K27me3 for repression of the segments, followed by removing the H3K9me3 mark and adding the trimethylation at H3K4. Doing so, the repressive state of the segment is retained all the time, but a complete switch of the modification can be performed.

Except state 000, all other states have a clear secondary source. Surprisingly, these secondary sources are only two distinct states in ESCs. ESC code 011 is secondary source for 101, 110, 011, and 111, while ESC code 001 is secondary source for 010, 001, and 100 (see Figure 6.7 local). It seems that segments carrying already multiple marks in ESCs are segments, which carry at least two marks in MEFs. This means that at least one of the marks in ESCs is retained and at most one mark is lost. Thus, all segments retain the repressive state set in ESCs also in MEFs although the combination of marks might change. Segments with just one mark are able to switch the modification and thus, changing from active to repressed and vice versa.

There are two ways how ESC codes are distributed over the clusters of MEF: (1) Some codes, e.g. code 100, prefer certain clusters, while the other clusters are avoided. (2) Segments are distributed uniformly of the all clusters. Examples for the latter one are 101 and 110 (Figures E.2f, E.2g). Since the latter two combinations do not show specific transitions, the combinations might not have a specific function. In the following, examples of specific distributions are described.

The most unstable state in ESCs is 001, i.e., segments with H3K9me3 only. They lose the H3K9me3 mark completely and often acquire new modifications. ESCs 001 is about uniformly distributed to 000, 100, 110, and 010 in MEF (Figure E.2b). H3K9me3 marks alone cannot be transmitted into MEFs, but are retained over multiple cell divisions in ESCs. Thus, they might play a crucial role in ESCs identity, e.g., suppressing differentiation-specific genes that are otherwise easily activated.
Segments solely marked with H3K27me3 can be retained at a level of 21%. More often (35%), they become unmodified. 11% of the segments change into a bivalent state, which means that they gather a new mark, but do not lose the old one. With the approximately same rate (12%), the H3K27me3 mark is removed and the H3K4me3 mark is set. Only in a few cases (8%), H3K27me3 is exchanged with H3K9me3 (Figure E.2c).

The active segments in ESCs (100, H3K4me3 mark only) loose the H3K4me3 in 58% of the cases. Nevertheless, most of the remaining segments retain the H3K4me3 and stay in an active chromatin state (32%). These segments might be associated with house keeping genes, which have to be active in all cell types. The remaining segments turn into fully modified segments (Figure E.2e).

Bivalently marked segments in ESCs are either retained or completely demodified (110 → 110: 20% and 110 → 000: 18%). As reported in the literature, one of these marks can easily be removed to put the segment in an active (100) or stable repressed (010) state. During differentiation into MEFs, these switches occur with rates of 10% and 15%, respectively. However, 10% turn into double repressed segments (011).

Segments carrying all three marks in ESCs retain their marks in 40% of the segments. Loosing exactly one of the three marks happens with a rate of around 15% each. The cases that more than one mark is lost can be neglected due to very rare occurrence (≤ 5%). For this reason, carrying all three marks appears to be very stable during differentiation.

Predominant transitions from the ESC codes into the MEF codes are highlighted by clustering. The results show that some states do not change in a specific manner discriminating them from functional combination of modifications. Main sources of MEF states could be detected and show new insights into the origin of combinations of modifications.

**Neuronal Progenitor Cells** Analysis of the overlap between ESC and NPC data showed again that segments are either covered completely or not at all by the modifications. Thus, as for the MEFs 8 clusters are expected. However, clustering the NPC data with $k = 8$ results in unreliable centroid coordinates (data not shown). For example, cluster 2 has the following coordinates in NPC: (H3K4me3, H3K27me3, H3K9me) = (0.879, 0.934, 0.829). All three modification are still set on segments in this cluster. The coordinates for cluster 5 are (H3K4me3, H3K27me3, H3K9me) = (0.555, 0.480, 0.509). This would correspond to the segments where all three marks are retained on the half segment length. However, there are two reason making this cluster unreliable: (1) there is no cluster with code 101, (2) analysis shows that modifications are either retained on the whole segment or lost completely. For this reasons, we conclude that cluster 5 is an artifact arising from the clustering with too many cluster centroids. Further analysis of the centroid coordinates shows that 3 codes are missing, namely 101, 001, and 100, and 3 code occur twice, namely 111, 011, and 110. This means that there are only 5 patterns clustering in the NPC data. Thus, clustering is repeated with $k \in \{7, 6, 5, 4\}$.

Results show that there are indeed only 5 cluster but also one additional cluster collecting
segments which do not change the state completely. Thus, result for $k = 6$ are used, representing the 5 codes still observed in NPCs and one cluster for miscellaneous segments. Coordinates and classification of the clusters are shown in Table 6.4. The starplots with polygons for all CEPs are shown in Figure 6.8. Starplots with one CEP only are shown in the appendix in Figure E.3 and Figure E.4.

Table 6.4: Centroids in NPC: For each cluster, the coordinates in NPC are given as well as the abstracted code.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroids in NPCs</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H3K4me3</td>
<td>H3K27me3</td>
</tr>
<tr>
<td>1</td>
<td>0.009</td>
<td>0.023</td>
</tr>
<tr>
<td>2</td>
<td>0.860</td>
<td>0.897</td>
</tr>
<tr>
<td>3</td>
<td>0.642</td>
<td>0.879</td>
</tr>
<tr>
<td>4</td>
<td>0.045</td>
<td>0.630</td>
</tr>
<tr>
<td>5</td>
<td>0.420</td>
<td>0.239</td>
</tr>
<tr>
<td>6</td>
<td>0.225</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Segments, which do not exhibit any of the codes, are found in cluster 5 in NPCs. Coverage values of the modifications on these segments may result from reorganization of these segments leading to new units for regulation. Thus, it is not surprising that most of the segments (55%) were unmodified in ESCs (code 000). Such segments represent the largest part of the genome and are usually large in length. Thus, subsegments unmodified in ESCs are now modified in NPCs. The different modifications however, need not necessarily occupy the same position on the segment. Segments modified in ESCs and found in cluster 5 in NPCs are now either reorganized in new segments or they just incompletely switched the state, which is not uncommon in such data sets due to several sources of noise.
Most of the unmodified segments in NPCs are segments also unmodified in ESCs (90%). 5% carried the H3K4me3 mark only in ESCs. Thus, a large amount of the unmodified chromatin in ESCs stays unmodified. Regarding also the local distribution, most of the codes in ESCs are located in cluster 0 in NPCs representing the unmodified state. Taken together, many segments become unmodified and unmodified segments stay in this state resulting in an overall reduction of epigenetic marks in NPCs!

In ESCs, code 111, i.e., segments with all three modifications, was often observed. In NPCs, this is not the case. The local distribution of code 111 shows that this code can not be retained. Only 12% of the ESCs code 111 segments are still in this state. Most frequently, ESCs 111 segments loose H3K4me3 and H3K9me3 marks and turn into H3K27me3 repressed segments (30%). The global distribution of code 111 shows that only 22% of the segments with all three modifications in NPCs carry these modifications also in ESCs. This highlights again that code 111 is not retained. Additional to the code 111, code 001 and code 011 are the main sources of segments with all three modifications in NPCs. The segments retain the repressive character, but acquire H3K4me3 and thus, can be activated by removing the repressive marks. Furthermore, the H3K4me3 mark prevents DNA methylation on these segments and thus, keeps the segments in an epigenetic state able to switch into transcriptional active state.

Three codes are rare in ESCs, namely, 100, 101, and 110. Thus, the local distribution does not reveal many insights. The codes 101 and 110 mostly change into unmodified (101: 35%, 110: 38%) and H3K27me3 only modified (101: 20%, 110: 30%). Code 100 also becomes unmodified in NPCs (70%) and is also found in cluster 5 (miscellaneous code: 25%). Since code 100 and code 101 do not exist in NPCs, segments with these combination of marks have to change there modification state. Therefor, it is not surprising to find most these segments in the two most frequently observed clusters in NPCs, namely, 000 and 010.

The third code lost in NPCs is 001. Most segments carrying H3K9me3 only in ESCs loose this mark in NPCs and stay unmodified (30%). However, this does not necessarily mean that these segments are unregulated. It is known that H3K9me3 recruits DNA methylases. Thus, the segments may be DNA methylated in NPCs and do not need to be repressed by histone methylations. ESCs code 001 also highly contribute to 110 and 011 in NPCs. The latter one represents segments, which are double repressed. Both marks have to be removed and at least one mark has to be set to activate such segments leading to a more stable repression. The segments in code 110 change the state of all three marks and turn from repressed by H3K9me3 into poised chromatin state. Removing just one mark can lead to activation of transcription on these segments.

The local distribution of code 010 indicates that this code is rare in ESCs. 15% of segments carrying H3K27me3 only in NPCs retained this mark during differentiation from ESCs into NPCs. The contribution of ESCs 010 to the other clusters is on a marginal level, namely less than 10% for each cluster. Half of the segments with H3K27me3 only in ESCs loose this mark and are
unmodified in NPCs. Only 30% retained the state from ESCs.

Code 011 in ESCs shows pattern different from those of the other codes. It contributes to a high amount to segments with code 111 (35%) and code 110 (25%) in NPCs. Only 20% of the segments with H3K27me3 and H3K9me3 marks in ESCs retained this mark in NPCs. In each of the clusters in NPCs with codes 000, 111, 110, and 010, approximately 20% of the segments with code 011 in ESCs are found. The other two clusters contains 10% each of the 011 segments in ESCs.

The global and the local distributions of the codes reveal that the amount of H3K4me3 marks and H3K9me3 is dramatically reduced in NPCs. Many segments carry only H3K27me3 or are unmodified w.r.t. the studied modification. Thus, the level of the H3K27me3 marks is high. A high amount of the unmodified segments in ESCs acquire any of these marks. The main source for the segments in NPCs that do not belong to cluster 1 (code 000) or cluster 4 (code 010) are segments with code 001, 111, and 011. Since 101 is rare in ESCs, this suggests that H3K9me3 modified segments mark segments that are regulated by H3K9me3 and H3K4me3 in NPCs.

Summarizing, the results obtained from clustering the NPCs indicate a major reorganization of the segments in NPCs and strong changes in the epigenetic regulation.
Patterns found in Murine EV-Segmentation Data

For investigation of fate of modifications during differentiation, the EV-segmentation can be applied to data sets of one modifications in different cell types. CEPs therefore represent the existence of a specific mark in the different stages. All visualization methods presented in chapter 5, namely SOMs, Scatterplot Matrices, and Starplots, can be combined with EV-segmentation to examine the epigenetic patterns for variation of modifications. EV-segmentation was applied to segment the whole genome with respect to the variation among ESCs, MEFs and NPCs in the distribution of H3K4me3 modified regions.

7.1 SOMery

The EP data set obtained in the EV-segmentation for H3K4me3 was subjected to SOM-training. The structures of the obtained population map and CEP cluster map are analogous to the corresponding maps after ES-segmentation (compare Figure 7.1 to Figure 6.1). The map shows highly populated islands that are separated by sparsely populated border zones. Again we found that
all EPs of a particular island refer to the same CEP (see Figure 7.1b for assignments) The eight possible CEPs form nine islands where EV100 splits into two subregions labeled with “a” and “b”.

Figure 7.1: SOM-images after EV-segmentation: a) Population map. The number of EPs per node is logarithmically scaled and ranges from 1 (blue) to 24.000 (red). b) CEP cluster map. Islands with more than 10 EPs per node are shown in white and are labeled with their CEP. Borderlines with less than 10 EPs per node are shown in black. Chromatin stable marked by H3K4me3 refers to EV111 (green island).

The largest part of the SOM is again occupied by unmodified segments, i.e. segments not carrying H3K4me3 modifications in any of the three studied cell types (EV000). EV000 corresponds to 98% of the total chromatin. Regarding the modified part of the chromatin, the CEP EV111 marks stable H3K4me3 modifications, while all other CEPs mark dynamic H3K4me3 modifications, namely segments that either lose or acquire this epigenetic mark in the course of differentiation.

Figure 7.2 shows the complete SOM-atlas obtained after EV-segmentation. It enables direct insights into the global reorganization of H3K4me3 modified chromatin during differentiation of ESCs into MEFs and NPCs. Inspection of the atlas shows that the paired subregions (EV100a/b) carrying H3K4me3 marks exclusively in ESCs (but not in MEFs and NPCs) can be distinguished by their combined H3K27me3 and H3K9me3 modification status in ESCs and thus by their chromatin activity status in these cells (see Figure 7.2a’-7.2c’). Quantification of the qualitative information demonstrate that 11065 out of 21759 (51%) segments carry repressive marks of which 6501 (48%) are in a repressed and only 685 (3%) in a poised chromatin state. An active chromatin state (H3K4me3 only) is observed in 4652 (21%) segments. While the ”b”-subregion (EV100b) contains segments with activating marks only, EV100a is associated with segments of poised and repressive chromatin states, namely carrying also H3K27me3 and/or H3K9m3 marks. Having a closer look at the sub-regions in the island EV100a that show loss of H3K27me3 or H3K9me3 in either MEFs or NPCs, it can be concluded that H3K4me3 de-methylation is independent of repressive marks.

The supporting maps of the EV-segmentation (Figure 7.3) project additional information about the segment lengths, their CpG-density and the expression status of genes onto the SOM-topology. Long segments mostly accumulate in the EV000 island (see Figure 7.3a), i.e. they remain
H3K4me3 unmodified under all conditions studied. A small amount of long segments associates with the EV110-island (see Figure 7.3a). These segments refer to H3K4me3 modified chromatin in ESCs that loose H3K4me3 in NPCs but not in MEFs. Segments of comparable length and modification status locate to island EV111 in the ES-segmentation maps. Consistently, the long segments of EV110 are modified with H3K27me3 and H3K9me3 in ESCs. Inspection of Figure 7.3b clearly shows that the associated DNA is enriched in CpGs. This suggest that the recruitment of the respective histone marks may thus be governed by the binding of modifying complexes.

Figure 7.2: SOM-atlas after EV-segmentation: Color codes as in Figure 6.2. a')-c')Segments with CEP (EV100) locate to two islands EV100a and EV100b differing in their coverage by H3K27me3 (and H3K9me3) in ESCs.
to CpG-enriched chromatin as previously proposed [4, 99]. The association between the length of the segments, their CpG-density and histone modification status was already established using ES-segmentation, however, the orthogonal view via EV-segmentation more clearly shows that long and CpG-rich subregions lose H3K27me3 in NPCs when H3K4me3 is lost while the shorter and CpG-poor subregions more often stay H3K27me3 trimethylated.

Figure 7.3: Supporting maps after EV-segmentation: a) Average segment length per node ranging from 200 (red) to 31,000bp (blue). b) Average CpG-density of the segments per node ranging from 0 (blue) to 0.06 CpGs/bp (red). c) Fraction of segments that exclusively overlap with genes, which are significantly expressed (present) in all three cell types.

Applying ES-segmentation we observed that many segments carrying the H3K4me3 mark in ESCs are enriched in CpGs. EV-segmentation clearly shows that such enrichment is present for EV100, EV100a and EV100b but only for a few spots of islands EV101 and EV111. The length (Figure 7.3a) and CpG-density maps (Figure 7.3b) clearly show that the EV111 segments are mostly short and CpG-poor. Interestingly, EV111 is the only large island with H3K4me3 marked segments in ESCs that are CpG-poor. In addition, EV111 is stably H3K4 methylated in all cell types. This surprising observation leads us to the conclusion that stable maintenance of H3K4me3 is CpG-independent. Quantitative analysis of the epigenetic state of the segments reveals that 10719 segments carry H3K4me3 in all cell types of which 6589 are either H3K9me3 or H3K27me3 marked in all cell types. Only 644 segments (about 9%) have no repressive marks either in ESCs, MEFs, or NPCs. Consistently, the present-call map (Figure 7.3c) indicates that EV111 segments do not associate with a higher fraction of active genes then other segments although being stably H3K4me3 modified in all three cell types.

The combination of supporting maps on segment length and CpG-density also shed light on the potential dependencies of H3K4me3 de-modification (EV100a,b). While the two islands for the CEP EV100 show no commonalities with respect to H3K27me3 and H3K9me3 marks, the do have short segment length and high CpG-densities in common. As a consequence, we conclude that only short, CpG-rich segments are H3K4me3 de-modified in MEFs (EV100a,b) while in NPCs also long, CpG-rich segments lose their H3K4me3 mark.

Figure 7.4 shows the chromosomal enrichment maps for chromosome 1 and X of the EV-
segmentation. As for the ES-segmentation, for most of the chromosomes no specific enrichment pattern could be detected. Thus, chromosome 1 is used as reference. The whole set of chromosome-specific population maps and chromosome enrichment maps is shown in Appendix F. Enrichment is found in spots which are not associated with H3K4me3 in ESCs (EV010, EV011, and EV001). However, in the case of the spot associated with H3K4me3 in MEFs only, no enrichment is found in the part which is associated with H3K27me3 in ESCs (compare Figure 7.4b and Figure 7.2). While a feature like the latter may be hard to explain with state of the art knowledge about epigenetic processes, it opens new questions on the genome-wide organization of chromatin states.

### 7.2 Scatterplot Matrix

Scatterplot matrices are created with 50 bins per histogram and 50×50 bin-pairs for each scatterplot. The images for the global and local distribution are generated for the EV-segmentation data. Since details can only be seen on large images with high resolution, the images are provided in supplement on the web\(^1\) or on the enclosed CD. In the appendix in Figure G.1, Figure G.2, and Figure G.3, down scaled figures for the global and local distribution of all segments and segments with a length of at most 2,000nt are shown. In Table 7.1, the mapping from the presence of the H3K4me3 in the three modifications to color is shown. Results, derived from the EV-segmentation using the scatterplot matrix, are described in the following.

Again, diagonal elements are histograms while non-diagonal elements are scatterplots. Only modifications with continuous coverage value, CpG-density, and length are used. As described in Section 5.3, frequencies are counted for each code separately.

As already discussed in section 6.2, scatterplots of modifications in the same cell type have

---

\(^1\)http://evodevo.bioinf.uni-leipzig.de/~lydia/thesis_supplemental/supplemental.html
Table 7.1: Color of EV-CEPs for Scatterplot Matrix: For each CEP of the EV-segmentation the presence of the H3K4me3 mark in the three cell types as well as the decimal code is shown. Color used in the scatterplot matrix for respective code is named and shown.

<table>
<thead>
<tr>
<th>EV-CEP</th>
<th>ESCs</th>
<th>MEFs</th>
<th>NPCs</th>
<th>Code</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>EV001</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>green</td>
</tr>
<tr>
<td>EV010</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>blue</td>
</tr>
<tr>
<td>EV011</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>yellow</td>
</tr>
<tr>
<td>EV100</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>cyan</td>
</tr>
<tr>
<td>EV101</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>amber</td>
</tr>
<tr>
<td>EV110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>guppy green</td>
</tr>
<tr>
<td>EV111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>purple</td>
</tr>
</tbody>
</table>

highly populated nodes on the borders and on the diagonal between node (0,0) and node (1,1). In scatterplots between modifications of different cell types, highly populated nodes can only be observed on the borders but not on the diagonals. This shows that the diagonals are not an artifact of the ES-segmentation.

The histogram for H3K27me3 reveals that high coverage values, namely H3K27me3 marked segments, often also carry H3K4me3 (see Table 7.2). In the histogram for H3K27me3 in ESCs, high coverage segments make up a large part of the segment with codes EV110 (guppy green), EV111 (pink), EV101 (amber), and EV100 (cyan). All three codes representing segments with H3K4me3 in ESCs are observed. Nearly half the segments with H3K4me3 in ESCs carry also H3K27me3 in ESCs on the whole segment (49.4%). EV110 and EV111 as well as EV011 (yellow) and EV010 (blue) dominate the high coverage values in the histogram for H3K27me3 in MEFs. Again, all four codes representing segments with H3K4me3 in MEFs are found. The fraction of segments with H3K4me3 in MEFs, carrying on the full length also H3K27me3, is even higher than in ESCs, namely 85.8%. Also for H3K27me3 in NPCs, it is observed that all four codes, representing segments with H3K4me3 in NPCs, namely EV001 (green), EV011 (yellow), EV111 (pink), and EV101 (amber), are enriched at high coverage values. The fraction of H3K4me3 modified segments in NPCs with H3K27me3 in NPCs is approximately as high as in ESCs (57.3%). The strict co-occurrence can also be seen in the scatterplots of H3K27me3 modifications in ESCs, MEFs, and NPCs. The (0,0)-corners in the scatterplots are dominated by red (EV000) while the (1,1)-corners mainly consist of pink (EV111), guppy green (EV110), amber (EV101), and yellow (EV011). Thus, H3K27me3 and H3K4me3 are likely to be found at the same segments in all three cell types.

In the local distribution, it can be seen that both marks often are set at the same time and removed at the same time. Thus, high amounts of CEP EV110 (guppy green) are found at high coverage values in the histograms for H3K27me3 in ESCs and MEFs. Similar, CEP EV010 is predominant at high coverage values of H3K27me3 in MEFs and CEP EV011 at high coverage values of H3K27me3 in MEFs and NPCs.
Table 7.2: Co-occurrence of H3K4me3 and H3K27me3: Fraction of H3K4me3 modified segments which do not carry H3K27me3, the fraction of H3K4me3 unmodified segment which carries H3K27me3 and the the fraction of H3K4me3 modified segments which also carry H3K27me3 is listed for the three cell types. Only segments with full segment coverage (coverage of 1) and 0 coverage are regarded. The fraction of H3K4me3 unmodified segments which carry H3K27me3 is calculated including the data of the EV000 code and without including this data.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>H3K4me3 only</th>
<th>H3K27me3 only</th>
<th>Both marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCs</td>
<td>13.6%</td>
<td>2.6%</td>
<td>49.4%</td>
</tr>
<tr>
<td>MEFs</td>
<td>12.8%</td>
<td>1.5%</td>
<td>85.8%</td>
</tr>
<tr>
<td>NPCs</td>
<td>7.5%</td>
<td>3.1%</td>
<td>57.3%</td>
</tr>
</tbody>
</table>

Although H3K27me3 and H3K4me3 prefer to co-occur, many segments found carrying just one of those marks, i.e. either H3K4me3 or H3K27me3 (see Table 7.2). In ESCs and NPCs, the fractions of H3K4me3 unmodified segments that carry H3K27me3 on the full segment length are 2.6% and 3.1%, respectively. Disregarding the segments which belong to EV000 (mostly unmodified completely), increases these numbers to 10.2% and 13.3%, respectively. In MEFs, only 1.5% with EV000 and 4.2% without EV000 of the not H3K4me3 modified segments are fully H3K27me3 modified. The fraction of H3K4me3 modified segments which do not carry any H3K27me3 mark, is highest in ESCs (13.6%), slightly decreases in MEFs (12.8%) and is reduced to 7.5% in NPCs.

The pattern observed for H3K4me3 and H3K27me3 co-occurrence suggests that a strong coupling of these modifications and a mechanisms catalyzing them. However, there is no coupling of the mechanisms itself (e.g. binding of the co-occurring mark or interaction of the modification writers) on a genome-wide scale reported. Thus, the coupling might be indirectly, e.g. writers using the same or very similar binding motifs. Furthermore, strongest coupling is found in MEFs leading to reduced amount of H3K27me3 marks only. Reduced abundance of H3K4me3 marks in NPCs indicates a reduced activity of the H3K4 trimethylases and thus, results in lower levels of bivalently marked and H3K4me3 only marked segments and a higher level of segments with H3K27me3. Summarizing, in ESCs H3K4me3 and H3K27me3 likely co-occur. Nevertheless, there are mechanisms leading to segments with only one of those marks. In MEFs, less segments, carrying only H3K27me3 marks, are observed while fraction of segments only marked by H3K4me3 is kept on the same level as in ESCs. Thus, in MEFs the H3K4 trimethylases prefer much strong segments carrying already H3K27me3 marks than they prefer in ESCs. Also unmodified segments are more often bivalently marked than marked with H3K27me3 only. For this reason, a high abundance of bivalently marked segments is found in MEFs. The activity of H3K4 trimethylating complexes seems to be reduced in NPCs. Thus, the level of H3K4me3 only marked segments and bivalently marked segments is reduced. Since the fraction of H3K27me3 only marked segment is higher in MEFs, the activity of the trimethylases for H3K27me3 is not or only weakly reduced. Overall, the abundance of H3K4me3 marks changes strongly during differentiation into MEFs and NPCs, while the frequency of H3K27me3 marks stays at the same level.
Analysis of co-occurrence and avoidance of H3K4me3 and H3K9me3 provides insight to which amount H3K9me3 can recruit DNA methylation since this is only possible in absence of H3K4me3. However, this is just the upper bound and does not provide information about the actual existence of DNA methylation marks.

Comparing the amount of H3K4me3 modified segments carrying also H3K9me3 between the three cell types, it can be seen that in ESCs cells both marks occur together very frequently (62%). This amount decreases to 15% and 22.9% in MEFs and NPCs, respectively. Regarding also the facts that H3K4me3 modified segments without H3K9me3 marks show the opposite trend and segments without H3K4me3 marks but modified by H3K9me3 as well decreases dramatically (see Table 7.3), indicates that H3K9me3 marks are frequent in ESCs but very rare in MEFs and NPCs.

Table 7.3: Co-occurrence of H3K4me3 and H3K9me3: Fraction of H3K4me3 modified segments which do not carry H3K9me3, the fraction of H3K4me3 unmodified segment which carries H3K9me3 and the the fraction of H3K4me3 modified segments which also carry H3K9me3 is listed for the three cell types. Only segments with full segment coverage (coverage of 1) and 0 coverage are regarded. The fraction of H3K4me3 unmodified segments which carry H3K9me3 is calculated including the data of the EV000 code and without including this data.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>H3K4me3 Only with EV000</th>
<th>H3K9me3 Only without EV000</th>
<th>Both Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCs</td>
<td>14.6%</td>
<td>6.1%</td>
<td>24.1%</td>
</tr>
<tr>
<td>MEFs</td>
<td>48.0%</td>
<td>0.4%</td>
<td>3.4%</td>
</tr>
<tr>
<td>NPCs</td>
<td>21.9%</td>
<td>0.4%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

In the histograms of the local and global distribution of the coverage values of H3K9me3 stand out that H3K9me3 specifically locate on segments with H3K4me3 in MEFs. Most strongly this effect can be seen in ESCs; CEPs EV110 (guppy green), EV010 (blue), EV011 (yellow), and EV111 (pink) predominantly can be found in bins with high coverage values for H3K9me3. The same CEPs frequently occur at high coverage bins of H3K9me3 in MEFs, but EV010 and EV011 are less frequent than EV110 and EV111. In NPCs, only EV011 and EV111 are found frequently. Additionally, CEP EV001 (green) is frequent. Regarding only the most frequent CEP in bins for high coverage values, the list of CEPs is reduced to EV110 in MEFs and EV011 in NPCs. H3K9me3 seems to be set to segments which carry H3K4me3 to repress them in the next differentiation level.

In the scatterplots for H3K9me3 in ESCs and MEFs as well as in the scatterplot for H3K9me3 in ESCs and NPCs, high coverage values of MEFs and NPCs, respectively, are rarely populated. This indicates a strong reduction of H3K9 trimethylation level in MEFs and NPCs compared to ESCs. Furthermore, the borders for low coverage of H3K9me3 in ESCs and higher segment coverage by H3K9me3 in MEFs and NPCs, are also rarely populated. In MEFs and NPCs, H3K9me3 is mostly found on those segments which carry this mark in ESCs as well.

The subdivision of the CpG-density into three levels by the CEPs can be seen in the histogram of the CpG density in the global distribution as well as in the local distribution. High CpG-density
is dominated by red (EV000) and blue (EV010). A high fraction of amber (EV101) is also found at high CpG-density. Guppy green (EV110), cyan (EV100), and red (EV000) are found frequently at mid-level CpG-density. Low CpG-density is found on segments with any CEP. Mid-level CpG-density characterizes the segments with H3K4me3 modification in ESCs only or ESCs and MEFs. Thus, setting H3K4me3 marks in ESCs requires CpG-density around 0.05 (0.017 - 0.071 with maximum at 0.05). This correlates also with high coverage by H3K27me3 and H3K9me3. However, the mid-level CpG-density spread into the low coverage value for H3K27me3 and H3K9me3 and is often found at low coverage values of H3K9me3 in NPCs. Trimethylating H3K4 in MEFs requires high CpG-density, indicating that the modifying complex for the modifications MEFs and ESCs are different and recognize different loci in the genome. There is an additional mechanism active in ESCs and NPCs which recognizes high CpG-density and set H3K4me3. Neither the segments with CEP EV010 nor the segments with CEP EV101 are associated with high coverage by H3K27me3 and H3K9me3 in any cell type. In NPCs, H3K4me3 does not correlate with high or mid-level CpG-density. Segments with low CpG-density can be modified and unmodified in any cell type with respect to all three modifications.

Taken together, this indicates that there are several mechanisms to set H3K4me3 marks. At least two of them are CpG-dependent but distinguished in the CpG-density targeted. Furthermore, there is also a CpG-independent mechanism. Not all mechanisms are used in every cell type. Dependency of the H3K4me3 mechanism on mid-level CpG-density is found in ESCs and MEFs. It correlates well with high coverage by H3K27me3 and H3K9me3 in the same cell types. This suggests a coupling of the mechanisms for the three modifications. The mechanism, dependent on high CpG-density, is only found in MEFs. It is independent of the other modifications studied. In all cell types, CpG-dependent mechanisms exit and are uncorrelated with the occurrence of H3K27me3 and H3K9me3 marks. In NPC, no other mechanism is found.

Since the length of the unmodified segments largely exceeds the length of the modified segments, the segments are filtered by a length of 10.000bp as it was done for the ES-segmentation. Even with this length filter, not much can be seen. Most of the segments are smaller than 2000bp. Larger segments are H3K4me3 marked in ESCs and MEFs (guppy green, EV110). They have high coverage values for H3K27me3 and H3K9me3 marks in ESCs and MEFs. In NPCs, these segments have mostly low coverage values for both modifications. A second filter was applied to examine only those segments which are at most 2.000bp long. Segments which are modified in ESCs or MEFs tend to be longer than segments which are modified in NPCs only or NPCs and ESCs.

Summarizing, the results form CpG-density and length, it can be seen that H3K4me3 regions are long and exhibit a mid-level CpG-density. On such segments, H3K27me3 and H3K9me3 marks can be found as well. Most of them retain the modification state in MEFs but lose the marks in NPCs. Additionally, short segments can be found which carry H3K4me3 in ESCs. Typically, the CpG-density is low and H3K4me3 modifications do not correlate with the occurrence of the other
modifications. Thus, they are set by a different mechanism than the long segments in ESCs. A second mechanism seems to be active in MEFs. It leads to H3K4me3 modification of segments with high CpG-density. The segments are much shorter and do not exceed 2,000bp. This length restriction may arose from the genome architecture since one can see that larger regions do not show high CpG-densities in the mouse genome. Correlations with the other marks can not be found.

7.3 k-means++ Clustering and Starplots

Starplots are generated for the EV-segmentation data. At first, clustering is applied to the H3K27me3 modifications in the three cell types, namely ESCs, MEFs, and NPCs. Afterwards, H3K9me3 modification data in ESCs, MEFs, and NPCs are clustered. Starplots are shown and described in this subsection. In Table 7.4, the colors used to represent the presence of H3K4me3 in ESCs, MEFs, and NPCs are shown and mapped to the CEPs of the EV-segmentation.

Table 7.4: Color of EV-CEPs for the Starplots: For each CEP of the EV-segmentation, the presence of H3K4me3 in the three cell types as well as the decimal code is shown. Color used in the starplots for respective code is named and shown.

<table>
<thead>
<tr>
<th>EV-CEP</th>
<th>ESCs</th>
<th>MEFs</th>
<th>NPCs</th>
<th>Code</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>red</td>
</tr>
<tr>
<td>EV001</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>blue</td>
</tr>
<tr>
<td>EV010</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>maroon</td>
</tr>
<tr>
<td>EV011</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>rose</td>
</tr>
<tr>
<td>EV100</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>cyan</td>
</tr>
<tr>
<td>EV101</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>dark green</td>
</tr>
<tr>
<td>EV110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>green</td>
</tr>
<tr>
<td>EV111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>chrome yellow</td>
</tr>
</tbody>
</table>

**Trimethylation at histone H3 lysine K27** Starplots for the clustering of H3K27me3 coverage values from all three cell types are shown in Figure 7.5. Additional starplots can be found in the Appendix in Figure H.1 and Figure H.2. Each centroid can be assigned to the combination of cell types in which H3K27me3 is present. The centroid coordinates and assigned CEPs are shown in Table 7.5.

The main source for segments with H3K27me3 in no or one cell type are segments which are unmodified with respect to H3K4me3 all the time. Genes in these segments are not specifically activated by H3K4me3 marks but shut down in one cell type. Thus, this is an efficient mechanism to prevent expression of certain genes in specific cell types. It is an mechanism independent of the H3K4me3 modifications.

Main source of segments carrying H3K27me3 in ESCs and MEFs carry H3K4me3 in these cell types. This is also partly valid for segments with CEP EV010, EV111, and EV001. This is
Figure 7.5: Starplots for Clustering of H3K27me3 in all Three Cell Types: a) Local distribution of the CEPs observed for H3K4me3 in the clusters. b) Global distribution of the CEPs observed for H3K4me3 over the clusters.

Table 7.5: Centroids for H3K27me3: For each cluster, the coordinates for H3K27me3 are given as well as the abstracted code.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroids for H3K27me3</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ESCs</td>
<td>MEFs</td>
</tr>
<tr>
<td>1</td>
<td>0.907</td>
<td>0.939</td>
</tr>
<tr>
<td>2</td>
<td>0.024</td>
<td>0.028</td>
</tr>
<tr>
<td>3</td>
<td>0.092</td>
<td>0.664</td>
</tr>
<tr>
<td>4</td>
<td>0.796</td>
<td>0.301</td>
</tr>
<tr>
<td>5</td>
<td>0.862</td>
<td>0.857</td>
</tr>
<tr>
<td>6</td>
<td>0.534</td>
<td>0.169</td>
</tr>
<tr>
<td>7</td>
<td>0.186</td>
<td>0.869</td>
</tr>
<tr>
<td>8</td>
<td>0.089</td>
<td>0.198</td>
</tr>
</tbody>
</table>

an evidence for co-regulation of H3K4me3 and H3K27me3. Taken the global distribution into account, many more evidence can be found for co-regulation of H3K4me3 and H3K27me3. For each cluster in the H3K27me3 data, the main source is always the CEP representing the same combination of cell types. Both marks are set in the same cell types and are removed in the same cell type.

There are two special cases, namely local distribution of cluster 101 (w.r.t. H3K27me3) and global distribution of 111. For the first one, one can find that 25% of segments carrying H3K27me3 in ESCs and NPCs, carry H3K4me3 in MEF only. Thus, they are transcriptionally silenced in ESCs, become active in MEFs and are shut down in NPCs. It would be interesting to discover whether there is a mechanism specifically targeting the H3K27me3 marked genes to switch them on for just one cell type and switch them off in the next differentiation stage. There are several segments
carrying H3K4me3 marks in all three cell types. Analyzing the H3K27me3 distribution in the three cell types reveals that most segments miss H3K27me3 in just one cell type. This might be segments which are kept bivalent and switch on in a cell type specific manner. In these segments, genes might be located which act as marker for the three cell types and are only needed in one cell type but not in the other two.

Summarizing the clustering of H3K27me3, the existence of at least two mechanisms for setting H3K4me3 and H3K27me3 could be hypothesized. One mechanism sets both marks independently of each other and one mechanism which ensures that both marks are set at the same time. Additionally, cell type specific demodification of H3K27me3 is seen which might lead to expression of cell type specific genes.

Figure 7.6: Starplots for Clustering of H3K9me3 in all three cell types: a) Local distribution of the CEPs observed for H3K4me3 in the clusters. b) Global distribution of the CEPs observed for H3K4me3 over the clusters.

**Trimethylation at histone H3 lysine K9** Clustering was applied to the trimethylation of H3K9 in all three cell types. Thus, centroids represent the combination of cell types in which H3K9me3 was found on the segments. Clustering was done with \( k = 8 \) clusters. However, the centroids coordinates revealed that one cluster is missing, namely 011. For this reason, the k-means++ is started again with \( k = 7 \) clusters. The centroids and assigned CEPs are shown in Table 7.6. The corresponding starplots can be found in Figure 7.6 and in the appendix in Figure H.3 and Figure H.4.

The starplots for the unmodified segments of both distributions show that most of the segments which do not carry H3K4me3 marks also do not carry H3K9me3 marks. 90% of the unmodified segment w.r.t. H3K4me3 do not carry H3K9me3 marks and 92% of the unmodified segments w.r.t. H3K9me3 do not carry H3K4me3 marks. This shows that modifications are restricted to a
Table 7.6: Centroids for H3K9me3: For each cluster, the coordinates for H3K9me3 are given as well as the abstracted code.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Centroids for H3K9me3</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ESCs</td>
<td>MEFs</td>
</tr>
<tr>
<td>1</td>
<td>0.032</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>0.922</td>
<td>0.095</td>
</tr>
<tr>
<td>3</td>
<td>0.899</td>
<td>0.838</td>
</tr>
<tr>
<td>4</td>
<td>0.271</td>
<td>0.067</td>
</tr>
<tr>
<td>5</td>
<td>0.960</td>
<td>0.815</td>
</tr>
<tr>
<td>6</td>
<td>0.242</td>
<td>0.521</td>
</tr>
<tr>
<td>7</td>
<td>0.760</td>
<td>0.058</td>
</tr>
</tbody>
</table>

small part of the whole genome.

Trimethylation of H3K9 which are found in only one cell type are often located on segments which do not carry H3K4me3 modifications (ESCs: 40%, MEFs: 38%, NPCs: 28%). Thus, they are set independent of the H3K4 trimethylations. Surprisingly, such segments frequently carry H3K4me3 in MEFs (ESCs: 28%, MEFs: 28%, NPCs: 20%). Additionally, segments with H3K9me3 in NPCs carry H3K4me3 also in NPCs only (20%) or in MEFs and NPCs (28%). Hence, writers for H3K9me3 in NPCs target several different types of segments.

Whenever H3K9me3 is found in two or three cell types, many segments carry H3K4me3 in the same combination of cell types (ESCs and MEFs: 55%, ESC and NPCs: 50%, all cell types: 32%). If the combination is not exactly the same combination, than there is only one cell type in which H3K4me3 is not found on the segments (H3K9me3 in ESCs and MEFs, H3K4me3 in MEFs: 18%; H3K9me3 in all cell types: H3K4me3 in ESCs and MEFs; 22%, H3K4me3 in MEFs and NPCs 28%). The combination H3K9me3 in all cell types and H3K4me3 in ESCs and NPCs do not occur. Taken together, there seems to be a mechanism to couple both modifications. It is mostly restricted to segments which retain their marks over more than one cell type. This is even more supported by the global distributions of the codes to the clusters in NPCs. 60% of the segments carrying H3K4me3 in ESCs and MEFs are also H3K9me3 modified in both cell types. Also 38% of segments with H3K4me3 marks in MEFs and NPCs and 30% of the segments H3K4 trimethylated in ESCs and NPCs are covered by H3K9me3 as well in the corresponding cell types. Half of the segments with H3K4me3 marks in all cell types do also carry H3K9me3 marks.

In contrast to the local distribution strong co-location of the marks is also found for the segments with H3K4me3 in only one cell type. So, 32% of the segments with H3K9me3 in ESCs only, 15% of the segments with the mark in MEFs only, and 25% of the segments with H3K4me3 in NPCs only are covered as well with H3K9me3 in the respective cell types. Furthermore, partial co-location can be observed. For example, 20% of the segments with H3K4me3 in all three cell types have H3K9me3 marks in ESCs and MEFs. Likewise, 15% of the EV111 segments loose the H3K9me3 mark only in MEFs.

While for most CEPs a high amount of segments do not carry H3K9me3 in any cell type (EV100:
32%, EV010: 39%, EV001: 25%, EV101: 25%), segment with CEPs EV111, EV110, and EV011 do almost always carry H3K9me3 in at least one of the three cell types. Both marks mostly collocate on these segments. Thus, modifying complexes for H3K9me3 seems to be coupled on H3K4me3 modifications. Nevertheless, the mechanism remains unclear.

Interestingly, 42% of the segments with H3K4me3 in NPCs carry H3K9me3 in ESCs (22% in ESCs and NPCs, 20% in ESCs only). It is unclear if a mechanism can exist which can explain such a pattern incorporating several developmental stages of a cell. However, if such mechanism exists, it cannot only be explained with the data studied. Likely, other signals such as other epigenetic marks are used to establish the pattern observed.
In the last two chapters, the results for both segmentations with the different visualizations are shown. So far, it is hard to decide which method is the best and thus, which method is most suitable for analysis of further data sets. Performance test for time and memory could be performed. While SOM training needs the longest time, calculation of the scatterplot matrix is the fastest method. Scatterplots and k-means++ clustering with starplots can be performed on a normal computer and require no specialized hardware settings. For fast training of the SOMs, a higher amount of memory is needed but it can be performed on nowadays computers. More important than memory consumption and time consumption are stability and amount of the results.

Results which cannot be reproduced, cannot be used for the most purposes. Thus, stability of the results is required for all of the methods. Since scatterplots do not include any random aspect, results are stable by construction of the method. SOMs and k-means++ clustering using random numbers for initialization. Using different initialization the qualitative results stay the same but may look different. In the case of the SOMs, different initializations usually lead to re-arrangement of the islands and spots observed as well as slightly different assignments of the segments to the nodes. Different initializations for the k-means++ lead to different orders of the obtained clusters and slightly changed centroids coordinates without changing their interpretation.
Thus, the starplots strongly differ for different initialization. To prevent such effects, SOMs are always initialized in the same way using linear initialization [158]. The same seed is used for the random number generator for the k-means++ clustering. Summarizing, all three methods provide stable results in the current implementations. In the following, the results will be summarized and will be used to compare the methods performance.

8.1 Either SOMs or Scatterplot Matrices or Starplots!

A high amount of the genome is not regulated by the studied histone modifications in the studied cell types. This can be seen in all three visualizations. It does not mean that these parts are not regulated by the epigenome but either carry other marks or are not modified in a high number of cells in the measured cell population.

The epigenetic patterns of the marks strongly change during differentiation and thus, every cell type exhibit its one distinct epigenome state. This can be seen in all visualizations. Different regulatory mechanism are necessary to switch between the epigenome states. Modifying complexes reported by literature may play a key role but cannot form the full set of complexes for the studied modifications. This can be easily seen in the following two examples: (1) in NPCs, H3K4me3 marks are found at low CpG-density only although the complex binds to CpG (SOM and scatterplot matrix). (2) In MEFs, segments with significantly higher CpG-density than in ESCs are found (scatterplots matrix only) although literature do not report different complexes for ESCs and MEFs.

The strong reduction of H3K9me3 modifications from ESCs to MEFs and NPCs (all visualizations) indicates again that not all modifying complexes exiting in the cell are so far described in literature. Furthermore, H3K9 is only rarely de novo trimethylated in MEFs and NPCs which can be seen in all three visualizations. Similarly, H3K4me3 show reduced abundance in NPCs. However, this is not seen in the SOMs. Both changes in the frequency of the studied marks are not reported in the literature to date.

Comparing the amount of changes from ESCs to MEF and ESCs to NPCs, it is obvious that more changes have to be made to turn into NPCs (all visualizations). It is most emphasized in the starplots where in MEFs all eight CEPs are discovered while in NPCs only 5 CEPs remain. Many segments retain their modification state while switching from ESCs to MEFs. Nevertheless, many segments change the combination of marks and some segments get newly modified (all visualizations). Retention of epigenetic patterns from ESCs to NPCs is rare. Although this can be seen in all visualizations, the combination of clustering and starplots emphasizes this aspect most.

The scatterplots matrix shows that the histone marks locate on segments with fixed boundaries in one cell type. However, the boundaries of the segments changes during differentiation. Starplots show at least that for the transition from ESCs to NPCs. A cluster with miscellaneous segments
is found in the clustering of the NPCs modifications. Segments in this cluster are covered only partly with modifications and thus, suggest re-organizations of the segments in NPCs.

The existence of several mechanisms for H3K4 trimethylation are proposed by all three visualizations. A CpG-dependent mechanism for ESCs can be seen in the scatterplots and in the SOMs. Most of these segments are retained in MEFs and carry in both cell types, ESCs and MEFs, all three marks. The latter fact can also be observed in the starplots. Another characteristic is that the described segments are the longest modified segments found. This can be seen in the scatterplot matrix and in the SOMs. However, it is more obvious in the scatterplot matrices than in the SOMs. To retrieve this fact in the scatterplot matrix only one plot has to be analyzed. In the case of SOMs, the simultaneous analysis of several SOM images is required to find this observation.

A second mechanism recruiting H3K4me3 exists which can be found in all three cell types. Segments are short and have a low CpG-density. Any of the other marks are found on these segments. Thus, coupling of the three modification can not be proposed for this mechanism. Furthermore, no other mechanism for the recruitment of H3K4me3 is found in NPCs. While the existence of these segments is observed in all visualizations, length and CpG-density characteristics can not be seen in the starplots.

A third mechanism recruiting H3K4me3 marks is observed in MEFs only. Segments are shorter than 2000bp but have highest CpG-density in the mouse genome. This can be seen in the scatterplot matrices. The SOMs reveal at least that there are many segments which are H3K4me3 modified in MEFs only. Since the described segments are not as abundant as the short and CpG-poor segments, the segments locate in only two nodes in the SOM for the EV-segmentation. Thus, they are easily overseen. The length characteristics can not be caught at all. Starplots only show that there are segments which are exclusively H3K4 trimethylated in MEFs. All three methods reveal that H3K4me3 modification is not coupled with the other modifications studied on these segments.

Summarizing, segments with three different combination of marks can be described which are modified by H3K4me3. This suggest that there are three different ways to recruit H3K4 trimethylases. All of them can be discovered in the scatterplots. Starplots suggest that there are at least two; one leads to strong co-localization of the three studied marks in ESCs and MEFs, the other one leads to any combination of H3K4me3 with the other marks. Since neither length nor CpG-density are incorporated in the clustering, the latter two described mechanisms can not be distinguished by starplots. SOMs show only two of the mechanism. The mechanism in MEFs is missed.

CpG-independent recruitment of H3K27me3 and H3K9me3 marks to short segments is also observed in all three methods. It was seen that these recruitment mechanisms may be not coupled with the recruitment of any of the other marks studied.

All three visualization methods highlight that H3K9me3 is almost only found on segments
which are already H3K9me3 modified in ESCs. Newly modified segments are only found rarely. Furthermore, in MEFs and NPCs, H3K9me3 marks are likely to be found on segments carrying also H3K4me3. In the contrary, segments with H3K4me3 do not necessarily carry H3K9me3. It can be seen in all three visualizations. This result suggests that H3K9me3 are set based on the distribution of the H3K4me3. At the same time, the mechanism trimethylating H3K9 does not target all H3K4me3 modified segments. It is unclear which signal is used to determine which segment gets modified by H3K9 trimethylases and which segments are not bound.

Starplots revealed that marks co-occur with other marks if they persist during differentiation into other cell types while marks found in one cell type only do not co-occur. This may indicate that marks occurring in combination with other marks are less frequently removed and thus, are more stable if at least one other mark is set on the histones.

In general, H3K4me3 and H3K27me3 are often found on the same segment and thus, seem to be coupled or prefer at least the same target. So far, it can not be explained by which mechanism this pattern is produced. However, it can be seen in all visualization methods presented but is emphasized the strongest by starplots.

Although literature reports that bivalent regions turn into H3K4me3 only or H3K27me3 only segments, this does not seem to be the main transitions for bivalently marked regions. This is observed in all three methods. However, in starplots it can be seen that at least a small fraction of the bivalent regions indeed shows this behavior.

Additionally, starplots strongly emphasizes that in NPCs H3K27me3 is more frequent than H3K4me3 and H3K9 trimethylation. Thus, a high number of segments can be found marked my H3K27me3 solely. This is a surprising results, since it is not reported in literature so far.

Furthermore, starplots enables the discovery of complex pattern such as H3K4me3 marks in ESCs may lead to H3K9me3 marks in NPCs. Even rare patterns can be observed and transitions which are prevented can be discovered. For example, segments which are modified with H3K4me3 in ESCs and MEFs do only contribute small amounts to the segments never modified by H3K9me3.

For some modification states, specific transitions can be observed. This is easy to observe in starplots where segments with a specific code are non-uniformly distributed to the other clusters. Also modification states with unspecific transitions and thus, likely non-functional CEPs are detectable in starplots.

SOMs enable correlation of the modification data with gene annotation and gene expression. Thus, SOMs show which CEPs are associated with expression in all cell types studied and which with developmental genes. Enrichment of CEPs on chromosomes could be detected.

In the current implementation of the starplots, comparison can only be done between the reference used for segmentation and the clusters generated in one run of the k-means++ algorithm. For example, comparison of the clusters in the MEFs and the cluster in NPCs are not possible. To the contrary, scatterplot matrices and SOMs enable the comparison of each combination of modification and cell type.
Scatterplots allow only for pair-wise comparison of the modifications. Patterns regarding more than two modifications thus are hard to discover.

While scatterplot matrices and starplots provide the amount of segments found in the bins, bin-pairs, and clusters, SOMs do not provide numbers. Thus, statistics can be done with the results from scatterplot matrices and starplots but not with the results obtained by using SOMs.

Furthermore, export of the information from the starplots is simple. For each segment, the cluster number has to be exported. In the case of scatterplot matrix, the bin and bin-pairs have to be exported. For SOMs, this is more complicated. Native export exist for segment to node assignment. However, the information required are the segment to island assignment. Thus, the position and CEP of an island has to be determined to assign it to all segments in the island.

The results are summarized in Table 8.1. All results are listed. A “+” indicates that a result is found by the corresponding visualization method. If a result could not be found with the help of a specific visualization method, a “-” is shown.

<table>
<thead>
<tr>
<th>Result</th>
<th>SOMs</th>
<th>Scatterplot Matrix</th>
<th>Starplots</th>
</tr>
</thead>
<tbody>
<tr>
<td>high fraction of unmodified segments</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>strong changes in the epigenetic patterns during differentiation</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CpG-independent H3K4me3 marks in NPCs</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>different CpG-densities of H3K4me3 segments in ESCs and MEFs</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>strong demethylation of H3K9me3 during differentiation into MEFs and NPCs</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>only rare de novo trimethylation</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>reduced abundance of H3K4me3 marks in NPCs</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>more changes between ESC and NPCs than between ESCs and MEFs</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>reduction of CEPs in NPCs</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>retention of modification state of segments in MEFs</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>no retention of epigenetic patterns in NPCs</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>reorganization of the segments</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CpG-dependent mechanism in ESCs which lead to long segments</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>many segments in ESCs are modified with all three marks and are retained in MEFs</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
many segments with any mark combination including H3K4me3 in any cell type

short segments with low CpG-density recruit H3K4me3 in all cell types

segments with H3K4me3 in MEFs only

segments H3K4me3 in MEFs only exhibits highest CpG-density

segments H3K4me3 in MEFs only are up to 2000 bp long

H3K4me3 recruitment is independent of H3K27me3 and H3K9me3

CpG- and mark-independent recruitment mechanisms for H3K27me3 and H3K9me3

in MEFs and NPCs, H3K9me3 is found on segments carrying H3K4me3

segments with more than one modification stay longer modified

general trend of co-occurrence of H3K4me3 and H3K27me3

bivalent regions do not mainly turn into H3K4 trimethylated or H3K27 trimethylated regions

behavior of bivalent regions reported in literature

high abundance of H3K27me3 marks in NPCs

detection of complex pattern and rare pattern

prevention of pattern

specific and unspecific transitions

correlation with gene annotation and expression data

chromosomal specific distribution of CEPs

comparison of any cell type and modification regardless of chosen reference

detection of patterns with more than two modifications

visual representation is complemented with numbers

easy export of the results

| 8.2 Nobody! |

Table 8.1 clearly shows that all methods provide many new insights into epigenetic patterns and their regulation. However, with no method all results could be detected. Depending on the goal of the analysis, the best suited method has to be chosen.

Analysis of properties and patterns of one and two modifications is best suited for scatterplot matrices. Furthermore, scatterplot matrices present the coverage values and additional data. Complex patterns and rare occurring patterns are best discovered by starplots. SOMs provide
an overview and additional data, not available for the whole data set but only for a part of the segments, can be easily integrated.

Best results are obtained by combining the methods. Starplots and scatterplot matrices are already implemented using the same framework and thus, can be easily combined. This enables the inspection of the coverage values and distribution of length CpG-densities within the cluster. Doing so, the cluster can be better analyzed and characterized.

Gene annotations and expression data can be analyzed by SOMs. This is not possible with scatterplots and starplots. However, SOMs can not be easily combined with the other methods. Furthermore, identification of the segments with a specific modification state is possible for the segmentation reference but hard for the remaining modifications and cell types. Additionally, the assignment of segments to islands is not possible so far. Thus, the other methods can not be used to analyze single islands or specific modification states observed in the SOM. Aiming at detection of all results, requires the application of all methods to the data set.
Epigenetic research come to the fore in the last years. Not much is known about the modifications and demodification processes as well as interplay of the marks. Due to the development of high throughput sequencing methods, extensive measurement of the distribution of epigenetic marks is possible. Several projects aiming at building reference epigenomes and thus, a large amount of data is available.

Analysis of the data, however, refers often to gene centered approaches. While gene annotations may correlate with epigenome state, the process leading to modification do not necessarily rely on gene organization on the genome. Thus, correlation with gene annotation and expression data is worthwhile but should not base the analysis. In chapter 3, analyses had been described to analyze ChIP-seq data sets without focusing on gene annotation. Instead of analyzing the enrichment of modification around genes, characteristics properties had been determined. The calculation of the overlaps of different modifications and different cell types had been proposed as method for comparison of the properties of the different data sets.

Chapter 4 had shown the application of the described method using data from mouse. Three modifications (H3K4me3, H3K27me3, H3K9me3) had been studied in 3 cell types (ESC, MEFs, NPCs) and had been characterized without using gene annotation and expression data. Charac-
teristic length and CpG-density distributions had been shown as well as co-occurrence of the three studied trimethylation. Changes during differentiation had been analyzed for at least trimethylation on H3K4. It had been shown that trimethylation at H3K4 can be modeled by differential equations. The parameters had been fitted using the length distribution calculated for the H3K4me3 data sets.

Comparison of the different modifications in ESCs and the comparison of H3K4me3 in all three cell types had revealed the existence of patterns in the epigenome. The three marks had been found frequently on the same loci and it had been shown that retention of H3K4me3 is specific to long regions. However, the method had not allowed for detection of epigenetic patterns.

A segmentation methods had been developed in chapter 5 which had been based on an observation which had been made in chapter 4. Detected regions of the data sets overlap to nearly 100% or to 0% with each other no matter whether different modifications or different cell types had been compared. The segmentation methods had been designed to infer the breakpoints for the segments form the data itself. Thus, a set of reference modification data sets had been used for segment definition while information for the remaining modifications is calculated for each of the segments. Due to construction of the segments, resulting data shows a strict structure. Values for the reference modification data sets are either 0 or 1. The remaining values lie in the interval $[0, 1]$.

The segmentation procedure proposed had not incorporated gene annotation data leading to several advantages. It segments the whole genome and not only the 2%-5% of the genome on which genes are annotated. Nevertheless, correlation with gene annotations and also gene expression is possible. As described in chapter 5, the segmentation highlights different aspects depending on the chosen reference data sets. Segmentation according to different modifications in one cell type emphasizes the combination of mark and their changes. Using the data sets for the same modification in different cell types shows the variation of the chosen mark and the correlation with the remaining marks. Additional data collected in further experiments and databases or calculated from the genomic sequence can be combined with the modification data.

The result of the segmentation is a list of segments. Each segment is described by a epigenetic profile. Optionally, a vector with additional data is assigned to each segments. Due to the segmentation, the data typically consists of some hundred thousand segments. Thus, visualization methods are required which are capable the huge amount of data. Making use of the structure of the data 3 different visualization methods had been introduced in chapter 5. A main aspect of the all three methods is dimension reduction to enable examination of the data and detection of epigenetic pattern.

Self-organizing maps had been trained on the whole modification data available. Due to the combinatorial epigenetic profiles, training had led to the establishment of islands. All segments in an island had had the same combinatorial epigenetic profile. Any kind of additional data can be integrated in a straight forward way. Thus, SOMs enable not only inspection of the epigenome
state and epigenetic variation but also correlation with CpG-density, length, expression data and gene annotations classes. Furthermore, the distribution of the chromosomes in the SOMs can be analyzed. In this visualization, each modification or additional information had been represented by a map; an image of the SOM where each node is colored by the modifications value. Furthermore, the segments had been sorted into the nodes of the SOM. To each node the average of all assigned epigenetic profiles had been used. Thus, the number of observable points had been reduced from several hundred thousand to the number of nodes in the SOM.

Scatterplot matrices are the second the methods which had been introduced. The combinatorial epigenetic profile had been used to sort all segments into classes. Each class had been described by its CEPs which is encode by a color. Modifications values with the continuous range are depict as histograms. Each pair of them had been represented as scatterplot whereby the density of the of the data points in a bin had been depicted as saturation of the color. In the same way, CpG-density and length had been handled by this method. Switching between local and global distribution of the CEPs, the distribution of the CEPs in bin or the distribution of a CEP in the scatterplot matrix had been analyzed. So far, expression data and gene annotation data cannot be integrated. This method is best suited for analysis of the behavior and patterns of single modifications and pairs of modifications. Here, the origin from a combinatorial epigenetic profile is highlighted. The numbers of bins can be varied. Thus, also strength of dimension reduction varies.

Starplots had been used to represent the results of the k-means++ clustering. It makes use of the fact that regions in two modification data sets either overlap to almost 100% or nearly 0%. Thus, the number of expected clustered can be estimated. Depending on the definition of the reference, either the set of modification in one cell type or one modification in the set of cell types had been clustered. Contribution of a CEP to a cluster had been analyzed in the local distribution. The global distribution had enabled analysis of the distribution of the CEPs to the clusters. Complex and rare epigenetic pattern had easily been seen in the starplots. Furthermore, transitions from the reference CEPs to the clusters had enabled detection of predominant rewriting of combinations or co-regulation of marks.

Summarizing, chapter 5 had described a segmentation method and three specialized visualization methods for exactly the data obtained by the segmentation method. All methods are suited for the discovery of different aspects of the data. Combined with the two segmentation variants, they provide a powerful tool for analysis of epigenetic patterns. While the SOM method is currently implemented in R, scatterplot matrices and starplots are implemented in the same Java framework and can be easily combined. Furthermore, user interactions with the framework are realized. Thus, the framework is a tool ready to be use by bioinformations as well as biologist. To the contrary, the R package requires expert knowledge to generate not only the default SOM images but also any kind of map for additional information.

The power of the two segmentation method and the three visualization methods had been
shown in chapter 6 and chapter 7. While some of the results are consistent with the literature reports, several so far undetected patterns had been shown. More precisely, it had been found that reported observations are mostly not universal.

In chapter 6, the ES-segmentation data had been analyzed. It had shown that the methods scale well with the current size of the data sets. ES-segmentation had provided more than 811,000 segments. Each of them described by an 9-elementic epigenetic profile and at least two value in the additional data vector. Thus, the methods had dealt with almost 10 million numbers. In the case of SOMs, the additional data vector had contained two more data point for each segments. Thus, in the case of SOMs more than 10 million numbers had been handled.

Analysis of the epigenome states of the three cell types had revealed new insights into the co-occurrence of marks. Different dependencies had been highlighted and had shown that the current knowledge of modifying complexes do not cover the full set of writer complexes or complex members. Characteristics described for the modifications are valid and had been detected in the visualization. However, segments with different characteristic had been observed abundantly. Furthermore, it had been found that the described mechanism for H3K4me3 is not active in NPCs at all. Most of the literature described mechanisms and relationships between the studied marked had been validated in ESCs but only partly in the MEFs and NPCs. The data sets used, however, do not provide information about the modifying complexes itself. Thus, it cannot be revealed how the pattern are established and which chromatin modifier are responsible for the pattern.

Results from SOMs had included also correlation with expression data and annotation of developmental genes. Genes expressed in all cell types studied had been found to be, consistently with literature, marked with H3K4me3 in ESCs. However, most of these genes had not been covered by H3K4me3 in the other two cell types. Thus, H3K4me3 marks housekeeping genes in ESCs but their active state is not retained by H3K4me3. Potentially, other marks such as acetylations mark these genes. However, acetylations had not been studied in this thesis. Similarly, developmental genes had been found modified as reported in literature in ESCs only. The data for the MEFs and NPCs, however, had suggested that trends reported in literature do not represent the main trend for these genes. It had been found that chromosome X segments more frequently H3K9me3 marked than expected. However, literature review had not given rise to a hypothesis for this pattern.

Scatterplots had provided insights into the distribution of the coverage values and the CEPs. Comparison of the modification state in ESCs with the coverage by the modifications in the other cell types had revealed that segment definition correlate within the cell types but not within the modifications. Dependencies of the modification state on previous epigenetic states had be seen. Distribution of CpG-density and length had revealed the existence several mechanisms. Different correlations of the modifications with the CpG-densities had been shown.

Starplot had shown that new marks in MEFs are often found on segments which were unmodified in ESCs. While turning a cell from ESCs into MEFs, the epigenetic state of a segment usually do
not change completely but often in only one mark. Especially, demodification of segments with all three marks leads only to segments with two of the three marks. The same had been observed for modification: in most of the cases only one mark is added to the modification state. On the contrary, strong changes occur during differentiation into NPCs. H3K27me3 is dominating mark and almost nothing of the epigenetic state in the ESCs is retained in NPCs. Thus, the epigenome state changes gradually. A lot information from the last stage of differentiation is kept and new one is added. However, over longer time scales the epigenome state do not retain the information established in early stage of development.

In chapter 7, the epigenetic variation based on H3K4me3 in ESCs, MEFs, and NPCs had been analyzed and had highlighted the existence of in total three different recruitment modes for H3K4me3. All three modes had been characterized by the dependency on CpG-density, length of the segments, and co-occurrence with H3K27me3 and H3K9me3.

Distribution of the chromosomes in the SOMs had been analyzed. It had been found that chromosome X segments more frequently do not carry H3K4me3 in ESCs than expected. However, so far this can not be explained. Literature review had not given rise to a hypothesis for this pattern.

Scatterplots again had shown that coverage values of modifications from on cell type correlate while coverage values of the same modification measured in different cell types do not correlate. Additionally, the different co-occurrence level of H3K27me3 and H3K9me3 with H3K4me3 had been observed. This had led to hypothesis that two kinds of mechanisms indeed exist, namely coupled modification setting all three marks and uncoupled setting one mark. Furthermore, coupled modification had been found to be characteristic for ESCs.

Starplots had given rise to the hypothesis that modification are more stable and thus, longer retained, if they co-occur with other marks. Hence, pattern comprising two and more marks had been found to be retained from ESCs into MEFs or NPCs. However, literature had not reported the ability of the writer complexes for the studied mark to recognize the other marks. The exact mechanisms for the increased stability and retention of co-occurring marks had remained unraveled.

**Final Remarks**

In the thesis' first part, three histone modifications in three cell types had been characterized. Differences in the length distributions and CpG-density had been highlighted. This had shown that the epigenome state and thus, the distribution of the marks gradual changes during differentiation. Quantitative description of the marks and their changes had enabled fitting of a model for the dynamics of the epigenome. Simulations with model fitted to the data had shown the impact of the cooperativity on the stability of the epigenome state.

The second part of this thesis had shown that the combination of segmentation and visualization reveals insights into the epigenetic patterns and helps to formulate hypothesis about the interplay of histone marks. It had been learned that current knowledge of histone modifications and the
catalyzing complexes is far away from being complete. However, analysis of the ChIP-seq data combined with genomic features, expression data, and functional annotations had revealed many pattern and had complement the current knowledge. Already today, many epigenetic data sets are public available and can be used for gathering knowledge about the epigenome. Although the introduced methods had been designed for the purpose of the analysis of such data sets, they can be further improved to handle even higher dimensional data and more kinds of additional data. Doing so, they will provide deeper insight into the dynamics of the epigenome.

This thesis had presented a template for the analysis of epigenetic data sets. It had been shown how characteristic properties for epigenetic marks can be extracted from ChIP-seq data sets. Applying the analysis template to a data set of pluripotent and multipotent cells had led to the characterization of the hallmarks of pluripotency, namely H3K4me3, H3K27me3, and H3K9me3. Doing so, the characteristic properties had been determined in a qualitative and quantitative way. Based on this analysis, a model for the dynamics of the epigenome state had been fitted and had provided insights on the stability the epigenome state and the influence of cooperativity. Furthermore, differences in the epigenome state of pluripotent and multipotent cells had been shown. Analysis of these differences had enforced the formulation of the hypothesis that so far unknown mechanisms exist leading to the establishment of very different distributions of H3K4me3 in the different cell types.

A data-driven segmentation method had been developed within the scope of the thesis. It had used ChIP-seq data sets to segment the genome. The segmentation method had been designed to allow for the discovery of patterns in the data sets. Specialized on the segmentation data generated, three visualization methods had been designed. Here, established methods in the field of information visualization had been combined and adapted to enable the analysis of epigenetic pattern. Application of the segmentation and visualization methods had shown the power of the designed methods. A variety of patterns had been discovered in the data set used already in the first part of the thesis. Using the visualization methods, the patterns had not only been characterized by the modifications present but also by additional information incorporated into the analysis. Thus, each modification and cell type had been described by the detected pattern. Furthermore, patterns in different cell types had been compared revealing the differences of the cell types studied and had provided insights into the regulation of the epigenome state. Additionally, regulation of the different modifications had been compared. Thus, it had been shown that marks are co-regulated but also independently of each regulated at the same time.

Several projects will provide huge data set on epigenetic marks in several cell types. Studying a greater set of cell types including also re-programmed, aged, and terminal differentiated cells will provide deep insights into regulation of epigenome and differentiation. Inclusion of additional modification, e.g. H3K4me2 or H3K27ac, DNA methylation, and transcriptomic data will expand the insights into the histone code and connections to transcriptions. Furthermore, measurement of the binding sites for readers, writers, and erasers of the studied marks and chromatin structure
data will enable deep understanding of the regulation of the epigenome state. Understanding the epigenetic may enables improved re-programming of somatic cells and successful differentiation in any cell type. This can enable new therapies for, e.g., cancer or degenerative diseases.

Analysis of the epigenome state of equivalent cell types in different species will shed light in the evolution of the regulation of the epigenome. It had been shown that the genome organization, e.g. distribution of CpGs, strongly influences the distribution of marks. Thus, conservation of the epigenetic regulation might be displaced by genomic rearrangement and mutation signals. This leads to the problem that epigenome states cannot be compared directly between species when it is based on genomic coordinates. Hence, comparison of the patterns discovered during the analysis of epigenome state in different species enables detection of conserved mechanisms and invention of new interactions between marks.

Better analysis enables the discovery of more patterns and better description of patterns. Thus, the visualization methods can be improved and extended to handle more kinds of data and even higher dimensional data. Not only data analysis but also modeling and simulation of the epigenome state provide new insights into epigenome state and its regulation. Thus, extension of the model and fitting to new data available allows deeper understanding of epigenetic processes.

Not only the designed methods had led to new research question but also the results obtained by the analyses shown in this thesis. For example, very extreme differences are observed between ESCs and NPCs leading to the finding that the regulation mechanisms have to be different. However, with the current known set on complex and complex configurations, this differences cannot be explained. The experimental examination of the modifying complexes and their composition within different cell types would provide insights into the size and complexity of the configuration landscape for the modifying complexes. Questions on the number of complex configurations, activity and regulation of the them can be answered in this way.

The patterns shown in this thesis had highlighted the complexity of the histone code. It had been shown that patterns can be vanish and thus, are not longer available in certain cell types. The reason for it remains unclear but would be worthwhile to be analyzed. Furthermore, the question whether the same pattern carry out the same function in all cell types remains unanswered. Analysis of the available readers and interplay of readers and effectors as well as analysis of the patterns correlation with structure, transcription, and further data would help to touch this question.

The discovery that combinations of marks are longer retained and thus, more stable than single marks open up the question on the mechanisms behind. It remains unclear whether the phenomenon is restricted to certain combinations or affect any combination. Furthermore, the question rises whether the stability is proportional to the number of marks in the pattern.

The method design in the scope of this thesis allow the detection of different sets of patterns. Thus, no standard method can be named. However, results questions for further testing whether combination of the methods enhances the analysis of the data sets.
Appendices
The genomic location of long modified regions in mouse ESCs will be presented in the following. Selection of these regions is explained in section 4.1. For depiction of the genomic location, the UCSC browser [75] is utilized. Gene annotation from ensemble genebuild [117] and the gene prediction track of genscan [27] enable identification of function in terms of regulation of expression. Expression data from the ENCODE project [127] are used to obtain a brief glimpse into the expression of the genes in different cell types. A description of the genes and reference expression is shown in the figures captions in this appendix.
Figure A.1: chr8:126061166-126148424: This long modified regions overlap with the 5s rRNA gene. It is a tandem repeated rRNA. Ensembl annotation contains this gene and it is predicted by genscan. Embryonic stem cells (ES-E14) show expression of small parts of this region, i.e. only single copies are expressed. Transcription is also found in erythrocytes. H3K27me3 is found in all three cell types, suggesting repression in all three studied cell types. H3K9me3 is lost in all cell types. H3K4me3 is diminished in MEFs and NPCs.
Figure A.2: chr6:47601570-47723339: This region composed of two subregions which correlate with the gene predicted by genscan. Only the part of the gene containing exons are covered with the modifications. Expression of this gene can be found in erythrocytes and embryonic stem cells. All three modifications are found in ESCs. H3K9me3 is almost lost in MEFs and only small leftovers can be detected in NPCs. H3K4me3 and H3K27me3 are diminished but not completely lost in MEFs and NPCs.
Figure A.3: chr9:3000001-3038591: On the plus strand 12 protein-coding genes with frame-shift introns are annotated. The high abundance of the frame-shift suggest that their may a problem with the annotation in this region. On the minus strand, 3 lincRNA are predicted. However, this is not surprising since histone modification data was used for annotation of lincRNAs in the enssembl pipeline. Expression data from erythrocytes and embryonic stem cells indeed show expression. Nevertheless, it is questionable whether the annotation is correct. Most strongly, H3K4me3 and H3K9me3 modifications are retained. H3K27me3 is only found in ESCs.
Figure A.4: chr6:114444142-114587871: Only a gene prediction by genscan can be found in this region. Expression signals are detectable in erythrocytes and embryonic stem cells. All three modifications are found in ESCs whereby H3K9me3 modifications consists of many subregions. In MEFs and NPCs, also the other marks are broken into smaller subregions or are almost completely lost.
Figure A.5: chr7:67002206-67047765: In this region, the SNORD116, a brain-specific snoRNA is annotated. Expression can be found, e.g., in the cerebellum but only marginal in embryonic stem cells. While marked with all three marks in ESCs and thus, repressed, marks get lost in NPCs. However, H3K27me3 remains almost over the full length. Only locally regions are not H3K27me3 modified. In these gaps, H3K4me3 can found in some cases leading to the hypothesis that some of the copies of SNORD116 are marked for expression while other copies are repressed. In MEFs, all three modifications a found on large parts of the region.
Figure A.6: chr7:66763853-66855402: The SNORD115 is annotated in this region. It is also a brain-specific snoRNA. Modifications are found on almost the whole region in all cell types suggesting a repression of transcription at this locus. The expression in cerebellum shows only very rare expression.
Figure A.7: chrX:166601240-166633574: Neither annotated genes nor gene predictions nor expression is found in this region. Modifications get lost in this region. Strongest losses are observed for H3K4me3. H3K9me3 is only slightly lost. A function of the modification at this region can not be derived from the data.
Figure A.8: chrY:2688290-2765959: Similar to the region in Figure A.7 no gene is annotated or predicted. Also no expression signal is found. While all three modifications are found in NPCs, in ESCs and MEFs the modifications exists only as many small subregions. H3K4me3 is not found ESCs but the regions completely modified by H3K9me3 in ESCs.
Figure A.9: chr3:3009790-3033629: In this region, marks get more or less lost during differentiation. Most strongest, H3K9me3 is retained. Gene predictions can be found but no corresponding gene prediction. Thus, no function for this region can be derived from the data.
The visualization methods presented in chapter 5 and used in chapter 6 and chapter 7 are not developed within the scope of this thesis but are known long time before. However, the combination of the methods and adjustments to the epigenetic patterns are new and thus, a result of this thesis. In this appendix the technical details for the used visualization methods are described.

B.1 SOM

In 1990, Kohonen developed the self-organizing maps and thus, they are also called Kohonen-Maps [79]. Such a map combines clustering and visualization. Thus, SOMs are capable to handle high-dimensional data and enable examination of the data. SOMs are used in a variety of biological application among them transcriptome analysis [167], mass spectronomy image data analysis [52], or drug design [72].

Self-organizing maps are a unsupervised learning method for neuronal networks. Hereby, high-dimensional data points are mapped onto the low-dimensional space. To cluster the segments, a two-dimensional rectangular grid is used. Each node in the grid can incorporate several data points which are sufficiently similar with respect to a given distance function. At the beginning,
an initialization method is used to assign a specific feature vector to each node in the grid which represents the feature vectors assigned to the node. This feature vector is called meta-feature. To initialize the nodes, a principle component analysis is performed on the input data. The first two principal components are mapped onto the width and the height of the SOM, respectively. This ensures that the mean is located in the center of the SOM, while the strong effects are broken down along the sides. This initialization method is called linear initialization [158] and initializes the nodes in a manner such that neighboring nodes have similar meta-features. After initialization, the meta-features are updated during learning in about 200,000 iterations over all feature vectors. In each step, one feature vector from the input data is taken. It is assigned to the node with the most similar meta-feature with respect to the euclidean distance between feature vector and meta-feature. Afterwards, the meta-features are updated to better fit the assigned feature vectors. Hereby, not only the meta-feature is updated which is most similar to the added feature vector but also all other nodes using a weight which is inverse proportional to the distance between meta-feature most similar to the feature vector and the updated meta-feature. In this way, the feature vectors are placed on the map such that similar meta-features are located in close proximity while dissimilar meta-features are located more distantly. The resulting SOM is finally visualized by mosaic images which we further refer to as SOM-images. Each of the tiles represents one node and its color corresponds to the respective meta-feature-value [157]. In this thesis, SOMs are trained with the epigenetic profiles resulting from the segmentation. Thus, the feature vectors correspond to the epigenetic profiles [145].

B.2 Scatterplot Matrix

Scatterplots Matrices were developed by Hartigan in 1975 [58]. For an $n$-dimensional data set, an $n \times n$ matrix is created, were each tile contains a scatterplot for the respective two dimensions. In the case of very large data sets referring to the number of points for each scatterplot, simple scatterplots are not appropriated. Thus, the data range for each dimension is binned. For each bin-pair, it is counted how many data points fall into the first bin for the first dimension and into the second bin for the second dimension. Each bin-pair is colorized such that the color saturation represents the number of data points in the bin-pair. Since the diagonal elements contain the scatterplots for the same dimension on the x- and y-axis, they can be used to depict the histograms for each dimension.

Scatterplot matrices are widely used as visualization technique for multi-dimensional data [49]. For example, for network exploration [159], analysis of the distribution of health influencing factors such as body weight and fat [28], and analysis of the storms [144], scatterplot matrices are used.
B.3 k-means, k-means++, and Starplots

The k-means algorithm was developed in 1967 by MacQueen. Additionally to the data points, the number of clusters \( k \) and a distance function for calculating the distance between two points are required as input. Initially, \( k \) random cluster centroids are chosen. Clustering is performed in two phases. They are alternately repeated until convergence of the cluster composition or the maximum number of iterations is reached. In the first phase, all data points are assigned to the nearest centroid given the distance function. In the second phase, for each cluster, a new centroid is calculated as mean of all data points in the cluster [90]. Many applications of k-means clustering are reported in literature such as for gene expression analysis [3] and similarity search on covariance matrices [37].

The algorithm may require a long time to reach convergence of the clusters, especially for large data sets. Thus, several improvements were developed to speed up convergence. One of them is the k-means++ algorithm developed in 2007 by Arthur et al. The k-means++ algorithm changes the way the initial cluster centroids are chosen to faster reach the convergence of the clusters. It also ensures high accuracy of the clustering. Only the first centroid is chosen randomly while the remaining \( k - 1 \) centroids are calculated based on the first centroid and the data. The k-means++ algorithm works as follows: As first centroid, a data point is chosen with a uniform probability distribution. The next centroid is chosen from the data points with a probability proportional to its distance to the closest centroid [6]. This is repeated until all \( k \) centroids are chosen. Afterwards, the k-means algorithm is continued [90].

Starplots can visualize multivariate data in two dimensional plots [35, 39, 171]. Each variable is depicted on one axis and all axes start at the same point (see Figure 5.3). Position and angel of the axes are chosen uniformly. For each of the clusters calculated in the k-means++ algorithm, an axis is put into the starplot and labeled with the cluster number. For each CEP, a polygone is drawn. The corners are defined by the values from the distribution which are drawn as points on the axes of the respective cluster. Starplots are widely used in research and economy. For example, for NMR spectroscopy [160] or characterization of petroleum [165].
Additional SOM images for ES-Segmentation

Chromosomal-specific population maps (Figure C.1) showing the fraction of segments in a node originating from a specific chromosome where generated for all chromosomes in the mm9 mouse genome. Most chromosomes show no specific pattern but an almost uniform distribution of the segments. Maps for chromosome 12 and 14 are slightly speckled. The sex chromosomes show specific pattern. Chromosome X is located to specific spots which is already explained in section 6.1. Chromosome Y do not has high contributions to the node because it is short in length. Thus, only a small number of segments originate from chromosome Y in total.

Similar results can be obtained from the chromosome enrichment maps in Figure C.2. Enrichment over the expected number of segments originating from a specific chromosome is computed node-wise. The color-scale has the same range for all chromosomes. Most chromosome show a uniform distributions. Again, chromosome 12 and chromosome 14 are slightly speckled and chromosome X show a specific pattern.
Figure C.1: Chromosome-specific population maps of ES-segmentation for each chromosome: The logarithmic color scale ranges from 0 to 1.
Figure C.2: Chromosome enrichment maps of ES-segmentation for each chromosome: The logarithmic color scale ranges from 0 to 46. Light blue corresponds to the expected number of segments within one node (enrichment of 1). Nodes with a 10-fold enrichment are colored in yellow.
Appendix D

Scatterplot Matrices for ES-segmentation

The scatterplot matrices generated for the ES-segmentation are shown in this appendix. High-resolution images can be found in the web-supplemental at http://evodevo.bioinf.uni-leipzig.de/~lydia/thesis_supplemental/supplemental.html.

Global and local distribution of all segments are shown in Figure D.1 and Figure D.2, respectively. For the segments with length smaller than 10,000nt, the global and local distribution is shown in Figure D.3. Thereby, the rows for the modifications are omitted, since the characteristics are extracted for the modifications from filtered and unfiltered data.
Figure D.1: Scatterplot Matrix (global distribution) of the ES-segmentation with all segments
Figure D.2: Scatterplot Matrix (local distribution) of the ES-segmentation with all segments
Figure D.3: Scatterplot Matrix for ES-segmentation filtered by length: Only segments with a length $\leq 10,000$nt are shown. Rows corresponding to the modification data are deleted, since they do not show different characteristics than the same rows with all segments. In a) the global distributions is shown and in b) the local distribution is shown.
Starplots for the ES-segmentation

Additional to the starplots presented in chapter 6, two starplots for each CEP in the ESCs are computed showing the local and global distribution of the CEP. In the following, the starplots for the local and global distribution of the CEPs in MEFs and NPCs are shown.
Figure E.1: Starplots for local distributions in MEF; CEP from ESC
Figure E.2: Starplots for global distributions in MEF; CEP from ESC
Figure E.3: Starplots for local distributions in NPC; CEP from ESC
Figure E.4: Starplots for global distributions in NPC; CEP from ESC
Chromosomal-specific population maps (Figure F.1) showing the fraction of segments in a node originating from a specific chromosome where generated for all chromosomes in the mm9 mouse genome. A uniform distribution of the segments can be observed for most of the chromosomes. Maps for chromosome 12 and 14 are slightly speckled as in Appendix C. Specific patterns can be seen in the maps for the sex chromosomes. Chromosome X is located to specific spots which is already explained in subsection 7.1. Chromosome Y do not has high contributions to the node because it is short in length. Thus, only a small number of segments originate from chromosome Y in total.

Similar results can be obtained from the chromosome enrichment maps in Figure F.2. Enrichment over the expected number of segments originating from a specific chromosome is computed node-wise. The color-scale has the same range for all chromosomes. Most chromosome show a uniform distributions. Again, chromosome 12 and chromosome 14 are slightly speckled and chromosome X show a specific pattern.
Figure F.1: Chromosomal-specific population map of EV-segmentation for each chromosome: The logarithmic color scale ranges from 0 to 1.
Figure F.2: Chromosome enrichment maps of EV-segmentation for each chromosome: The logarithmic color scale ranges from 0 to 24. Light blue corresponds to the expected number of segments within one node (enrichment of 1). Nodes with a 10-fold enrichment are colored in orange.
The scatterplot matrices generated for the EV-segmentation are shown in this appendix. High-resolution images can be found in the web-supplemental at http://evodevo.bioinf.uni-leipzig.de/~lydia/thesis_supplemental/supplemental.html.

Global and local distribution of all segments are shown in Figure G.1 and Figure G.2, respectively. For the segments with length smaller than 2.000nt, the global and local distribution is shown in Figure G.3. Thereby, the rows for the modifications are omitted, since the characteristics are extracted for the modifications from filtered and unfiltered data.
Figure G.1: Scatterplot Matrix (global distribution) of the EV-segmentation with all segments
Figure G.2: Scatterplot Matrix (local distribution) of the EV-segmentation with all segments
Figure G.3: Scatterplot Matrix for EV-segmentation filtered by length: Only segments with a length ≤ 2.000nt are shown. Rows corresponding to the modification data are deleted, since they do not show different characteristics than the same rows with all segments. In a) the global distributions is shown and in b) the local distribution is shown.
Starplots for the EV-segmentation

Additional to the starplots presented in chapter 7, two starplot for each CEP of H3K4me3 are computed showing the local and global distribution of the CEP. In the following, the starplots for the local and global distribution of the CEPs for H3K27me3 and H3K9me3 are shown.
Figure H.1: Starplots for local distributions in H3K27me3; CEP from ESC
Figure H.2: Starplots for global distributions in H3K27me3; CEP from H3K4me3
Figure H.3: Starplots for local distributions in H3K9me3; CEP from ESC.
Figure H.4: Starplots for global distributions in H3K9me3; CEP from H3K4me3
List of Tables

2.1 Combinations of Marks ................................................. 21
3.1 Segemehl Parameters ................................................. 27
4.1 Long Modified Regions in ESCs ..................................... 38
5.1 Color Code for the Scatterplot Matrix ............................. 55
5.2 Color Code for the Starplots ....................................... 57
6.1 Color of ES-CEPs for the Scatterplot Matrix ....................... 65
6.2 Color of ES-CEPs for the Starplots ................................. 71
6.3 Centroids in MEF .................................................. 72
6.4 Centroids in NPC .................................................. 74
7.1 Color of EV-CEPs for Scatterplot Matrix ......................... 82
7.2 Co-occurrence of H3K4me3 and H3K27me3 ....................... 83
7.3 Co-occurrence of H3K4me3 and H3K9me3 ......................... 84
7.4 Color of EV-CEPs for the Starplots ............................... 86
7.5 Centroids for H3K27me3 ........................................... 87
7.6 Centroids for H3K9me3 ........................................... 89
8.1 Comparison of the results obtained in the visualizations ....... 95
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>3D-Structure of a Nucleosome</td>
<td>7</td>
</tr>
<tr>
<td>1.2</td>
<td>Histone Modifications</td>
<td>8</td>
</tr>
<tr>
<td>1.3</td>
<td>Arginine Methylation</td>
<td>9</td>
</tr>
<tr>
<td>1.4</td>
<td>ChIP-chip and ChIP-seq</td>
<td>12</td>
</tr>
<tr>
<td>2.1</td>
<td>Differentiation into Neuronal Cells</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>From the Blastula to Differentiated Cells</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Methylation of H3K4</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Repression mediated by the Polycomb Repression Complexes</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>DNA Methylation</td>
<td>22</td>
</tr>
<tr>
<td>2.6</td>
<td>CpG-rich Promoters</td>
<td>23</td>
</tr>
<tr>
<td>3.1</td>
<td>Distribution of H3K9me3 on chr19 in ESCs</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>Filtering of Mapped Hits</td>
<td>26</td>
</tr>
<tr>
<td>3.3</td>
<td>Definition of Regions</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>Length Distribution for Various Cut-off Values</td>
<td>29</td>
</tr>
<tr>
<td>3.5</td>
<td>Calculation of the Region-wise Coverage</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Background CpG-Distribution of mm9</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Length and CpG-Distribution of Epigenetic Marks in ESCs</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Model for Histone Modifications</td>
<td>37</td>
</tr>
<tr>
<td>4.4</td>
<td>Modification State of chr7:67.000.000-67.050.000</td>
<td>38</td>
</tr>
<tr>
<td>4.5</td>
<td>Overlaps between Modifications in ESCs</td>
<td>41</td>
</tr>
<tr>
<td>4.6</td>
<td>ESC H3K4me3 Region Coverage</td>
<td>42</td>
</tr>
<tr>
<td>4.7</td>
<td>Length and CpG-Distribution of H3K4me3 in ESCs, MEFs, and NPCs</td>
<td>43</td>
</tr>
<tr>
<td>4.8</td>
<td>Overlaps between H3K4me3 in ESCs, MEFs, and NPCs</td>
<td>45</td>
</tr>
<tr>
<td>4.9</td>
<td>H3K4me3 Region Coverage</td>
<td>45</td>
</tr>
<tr>
<td>4.10</td>
<td>Length Distributions of H3K4me3 and H3K27me3</td>
<td>46</td>
</tr>
<tr>
<td>5.1</td>
<td>Segmentation Procedure</td>
<td>51</td>
</tr>
<tr>
<td>5.2</td>
<td>Scatterplot Matrix</td>
<td>54</td>
</tr>
<tr>
<td>5.3</td>
<td>Starplot</td>
<td>57</td>
</tr>
<tr>
<td>6.1</td>
<td>SOM-properties after ES-segmentation</td>
<td>60</td>
</tr>
<tr>
<td>6.2</td>
<td>SOM-atlas after ES-segmentation</td>
<td>61</td>
</tr>
<tr>
<td>6.3</td>
<td>Supporting maps after ES-segmentation</td>
<td>63</td>
</tr>
<tr>
<td>6.4</td>
<td>Chromosomal enrichment maps for ES-segmentation</td>
<td>64</td>
</tr>
<tr>
<td>6.5</td>
<td>Coverage of Highly Populated Nodes</td>
<td>66</td>
</tr>
<tr>
<td>6.6</td>
<td>Loop Reorganization Hypothesis</td>
<td>67</td>
</tr>
<tr>
<td>6.7</td>
<td>Starplots for Clustering of the Modifications in MEF</td>
<td>71</td>
</tr>
<tr>
<td>6.8</td>
<td>Starplots for combined local (a) and global (b) distributions in NPCs</td>
<td>74</td>
</tr>
<tr>
<td>7.1</td>
<td>SOM-images after EV-segmentation</td>
<td>78</td>
</tr>
<tr>
<td>7.2</td>
<td>SOM-atlas after EV-segmentation</td>
<td>79</td>
</tr>
<tr>
<td>7.3</td>
<td>Supporting maps after EV-segmentation</td>
<td>80</td>
</tr>
<tr>
<td>7.4</td>
<td>Chromosomal enrichment maps for EV-segmentation</td>
<td>81</td>
</tr>
<tr>
<td>7.5</td>
<td>Starplots for Clustering of H3K27me3 in all Three Cell Types</td>
<td>87</td>
</tr>
<tr>
<td>7.6</td>
<td>Starplots for Clustering of H3K9me3 in all three cell types</td>
<td>88</td>
</tr>
<tr>
<td>A.1</td>
<td>chr8:126061166-126148424</td>
<td>110</td>
</tr>
<tr>
<td>A.2</td>
<td>chr6:47601570-47723339</td>
<td>111</td>
</tr>
<tr>
<td>A.3</td>
<td>chr9:3000001-3038591</td>
<td>112</td>
</tr>
<tr>
<td>A.4</td>
<td>chr6:114444142-114587871</td>
<td>113</td>
</tr>
<tr>
<td>A.5</td>
<td>chr7:67002206-67047765</td>
<td>114</td>
</tr>
<tr>
<td>A.6</td>
<td>chr7:66763853-66855402</td>
<td>115</td>
</tr>
<tr>
<td>A.7</td>
<td>chrX:166601240-166633574</td>
<td>116</td>
</tr>
<tr>
<td>A.8</td>
<td>chrY:2688290-2765959</td>
<td>117</td>
</tr>
<tr>
<td>A.9</td>
<td>chr3:3009790-3033629</td>
<td>118</td>
</tr>
<tr>
<td>C.1</td>
<td>Chromosome-specific population maps of ES-segmentation for each chromosome</td>
<td>124</td>
</tr>
<tr>
<td>C.2</td>
<td>Chromosome enrichment maps of ES-segmentation for each chromosome</td>
<td>125</td>
</tr>
<tr>
<td>D.1</td>
<td>Scatterplot Matrix (global distribution) of the ES-segmentation with all segments</td>
<td>128</td>
</tr>
<tr>
<td>D.2</td>
<td>Scatterplot Matrix (local distribution) of the ES-segmentation with all segments</td>
<td>129</td>
</tr>
<tr>
<td>D.3</td>
<td>Scatterplot Matrix for ES-segmentation filtered by length</td>
<td>130</td>
</tr>
<tr>
<td>E.1</td>
<td>Starplots for local distributions in MEF; CEP from ESC</td>
<td>132</td>
</tr>
<tr>
<td>E.2</td>
<td>Starplots for global distributions in MEF; CEP from ESC</td>
<td>133</td>
</tr>
<tr>
<td>E.3</td>
<td>Starplots for local distributions in NPC; CEP from ESC</td>
<td>134</td>
</tr>
<tr>
<td>E.4</td>
<td>Starplots for global distributions in NPC; CEP from ESC</td>
<td>135</td>
</tr>
</tbody>
</table>
F.1 Chromosomal-specific population map of EV-segmentation for each chromosome 138
F.2 Chromosome enrichment maps of EV-segmentation for each chromosome: 139
G.1 Scatterplot Matrix (global distribution) of the EV-segmentation with all segments 142
G.2 Scatterplot Matrix (local distribution) of the EV-segmentation with all segments 143
G.3 Scatterplot Matrix for EV-segmentation filtered by length .......................... 144
H.1 Starplots for local distributions in H3K27me3; CEP from ESC ................. 146
H.2 Starplots for global distributions in H3K27me3; CEP from H3K4me3 .... 147
H.3 Starplots for local distributions in H3K9me3; CEP from ESC ............... 148
H.4 Starplots for global distributions in H3K9me3; CEP from H3K4me3 .... 149


Curriculum Vitae

Lydia Steiner
Date of Birth: 15. May 1986
Place of Birth: Leipzig, Germany

Academic Appointments

Ph.D. candidate September 2009 to present
Computational EvoDevo Group, Dept. of Computer Science, Universität Leipzig
Supervisor: Sonja J. Prohaska
Topic: The Dynamic Epigenome - Analysis of the Distribution of Histone Modifications

Trainee January 2008 to September 2008
Max-Planck-Institute for Mathematics in the Science, Leipzig

Research Assistant October 2006 to March 2008
Natural Language Processing Group, Universität Leipzig, Leipzig

Education

Universität Leipzig, Leipzig, Germany

Diplom (equivalent to M.S., grade 1.4 equivalent to A), Computer Science, August 2009
• Main Focus: Bioinformatics
• Thesis Topic: *www - world wide words*
• Advisers: Peter F. Stadler, Michael Cysouw

Publications for the Dissertation

• **Starplot Visualization for Analyzing Chromatin**: Dirk Zeckzer, Sarah Seifert, Lydia Steiner, Sonja J. Prohaska, *International Journal of the Eurographics Association, submitted*

• **Modeling the dynamic epigenome: from histone modifications towards self-organizing chromatin:** Thimo Rohlf, Lydia Steiner, Jens Przybilla, Sonja Prohaska, Hans Binder, Jörg Galle. *Epigenomics*, 2012. doi:10.2217/epi.11.117


### Publications

• **Pitfalls of Ascertainment Biases in Genome Annotations — Computing Comparable Protein Domain Distributions in Eukarya:** Arli A. Parikesit, Lydia Steiner, Peter F. stadler, Sonja J. Prohaska. *Computation*, submitted

• **A Pipeline for Computational Historical Linguistics:** Lydia Steiner, Peter F. Stadler, Michael Cysouw. *Language Dynamics and Change*, 2011. doi: 10.1163/221058211X570358

• **Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis:** Marcus Lechner, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F. Stadler, Sonja J. Prohaska. *BMC Bioinformatics*, 2011. doi:10.1186/1471-2105-12-124

### Conferences and Workshops

• **Talk**
  Title: Somthing about Chromatin - Modelling Phenotypes by Epigenetic Regulation
  10. Herbstseminar 2012, Doubice, Czech Republic

• **Talk**
  Title: A Scientific Picture Book - The Epigenome of the a Mouse
  9. Herbstseminar 2011, Vysoka Lipa, Decin, Czech Republic

• **Poster**
  Title: Visualizing the Dynamic Epigenome
  Epigenetics Europe 2011, Munich, Germany

• **Conference Attendance**
  Fourth Weißenburg Symposium ”Epigenetics and the Control of Gene Expression”, Weißenburg (Bayern), Germany
• **Talk**
  Title: A Pipeline for Computational Historical Linguistics  
  CITEC Workshop on Evolution of Human Language, Bielefeld, Germany

• **Talk**
  Title: Assisted reconstruction: The cases of Panoan and Mataco-Guaiacuruan  
  14th Annual Workshop on American Indigenous Languages, Santa Barbara, CA, USA

• **Talk**
  Title: Identify Homologous Words  
  26th TBI Winterseminar in Bled, 5th Annual Meeting of the Bompfnewerer Consortium, Bled, Slovenia

• **Talk**
  Title: Tracing Histone Modifications  
  8. Herbstseminar 2010, Vysoka Lipa, Decin, Czech Republic

• **Talk**
  title: Models of Epigenetic Regulation: Histone Modifications - part I  
  INRIA-IZBI-Workshop 2010, Leipzig, Germany

• **Poster presentation**
  title: Novel findings on the genome-wide correlation of chromatin marks and CpGs  
  Transcription, chromatin structure and DNA repair in development and differentiation, Essen, Germany

• **Talk** title: An Example for Chromatin Regulation  
  25th TBI Winterseminar in Bled, 4th Annual Meeting of the Bompfnewerer Consortium, Bled, Slovenia

• **Talk**
  title: www - world wide words  
  7. Herbstseminar 2009, Vysoka Lipa, Decin, Czech Republic

• **Talk**
  title: something about languages  
  6. Herbstseminar, 2008, Ceska Kamenice, Czech Republic

**Teaching Experience**

*Universität Leipzig*, Leipzig, Germany
Teaching Assistant  
**July 2012**
- Teaching Assistant for practical course in “Structural Biology”
- part of master module “Structural Biology”
- students predict 3D-structure of protein given the protein sequence

Teaching Assistant  
**March 2012**
- Teaching Assistant for “Programming for Evolutionary Biology course”
- designed to teach biologist programming skills in Perl and R for the analysis of huge data sets, e.g. arising from high throughput sequencing

Teaching Assistant  
**January 2011**
- Teaching Assistant for practical course in “Sequence Analysis and Genomics”
  - annotation of the let-7 micro RNA family in Deuterostomia

Teaching Assistant  
**October 2010 to February 2011**
- Teaching Assistant for “Algorithms and Data Structures I”
- I hold some tutorials in place of Sonja J. Prohaska.
- tutorials topics: different algorithms for sorting and searching lists and their complexity

Teaching Assistant  
**July 2010**
- Teaching Assistant for practical course in “Sequence Analysis and Genomics”
  - prediction of interactions of zinc-finger domains with RNA, DNA, or proteins based on the sequence and structure of the given zinc-finger domains

**Language Skills and Qualifications**

Languages: native in German, fluent in English, basic knowledge in Spanish

Programming:
- advanced programming skills in C, C++, Perl, Java, Javascript, HTML, PHP, R
- advanced knowledge of LaTeX and BibTeX
Eigenständigkeitserklärung


Lydia Steiner