Tobias Singhania, BSc

Reconstruction of Regulatory Networks from Epigenetic and Transcriptomic Data

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Information and Computer Engineering

submitted to

Graz University of Technology

Supervisor

Dr. Gerhard Thallinger

Institute of Neural Engineering
Institut für Computational Biotechnology

Graz, May 2018

AFFIDAVIT[1]

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Graz, _____ _____

Date Signature

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Deciphering gene regulation and its mode of function is one of the major challenges in systems biology. Many different factors impact gene regulation and many studies have been carried out to elucidate what triggers genetic transcription and how different genes impact each other. In the last decade it has been shown that epigenetic modifications have a major impact on gene regulation. New sequencing technologies have become available which allow analysis of complex relations in the living cell.
The innovations of RNA-seq and ChIP-seq made it feasible to perform whole genome sequencing for detection of novel splicing, quantification of transcriptional activity, detection of transcription factor binding sites and analysis of histone modifications. Integration of different *-omics* datasets is a promising approach to gain deeper insights in gene regulation.

In this thesis RNA-seq data and ChIP-seq data for *Saccharomyces cerevisiae* were analyzed. Clustering of the histone modification patterns was performed and a relation between histone modifications and their impact on expression was analyzed. The adaption of a genetic regulatory network reconstruction algorithm was evaluated and different approaches for integrative analysis were performed.
For a set of histone modifications a strong impact on gene expression could be established, as well as that rather combinations of modifications are responsible for regulation of gene expression than a single modification alone, i.e. that H3K9ac, H4K5ac, and H3K14ac act together in gene activation while H3K36me3 regulates the impact which H3K9ac, H4K5ac, and H3K14ac have on gene regulation.

Furthermore the use of the ARACNE algorithm on datasets other than gene expression data is discussed, and its mode of function is explained in detail. The different approaches for deduction of histone modifications impact on gene regulation which were used in this thesis are discussed in detail.

Contents

1 Introduction

Genetic regulatory networks play an important role in living organisms and their reconstruction is one of the major challenges of systems biology. In the past decade many studies have been carried out to elucidate the regulatory interactions of different genes, that is to say how RNA- and protein products from expressed DNA segments govern transcription.
To gain deeper insights into those complex processes, mathematical models and algorithms for inference of such regulatory networks from gene expression data have been developed. Integration of different types of 'omics' data (e.g. genomic, transcriptomic, epigenomic, proteomic data) and prior knowledge from literature could drastically improve the quality of network reconstruction [1].
Reconstruction of regulatory networks and the analysis how histone modifications control gene regulation by integrating epigenetic and transcriptomic data is the goal of this thesis, and different analysis approaches are evaluated.

1.1 Epigenomics

Epigenomics refers to the study of DNA and chromatin modifications which are, in contrast to mutations, not based on an alteration of the DNAs nucleotide sequence. The epigenome is associated with many regulatory processes in eukaryotic organisms. It plays an important role in cell differentiation [2], apoptosis [3] and DNA repair mechanisms [4]. Epigenetic modifications create a complex set of combinations of positive or negative regulatory signals, which provides a flexible means for a cell to react to environmental changes and adapt

quickly to them.
The modifications are accomplished by molecular markers on the DNA itself (e.g. DNA methylation) or on the proteins involved in chromatin packaging (histone modifications).
In contrast to mutations, epigenetic modifications are specific for a cell. Therefore each cell has its own epigenetic code which enables each cell to act differently than other cells of the same type. Thus, the analysis of the epigenome of a multicellular organism must consider the tissue from which the cell was taken.

With the development of next generation sequencing (NGS) technologies new possibilities for genome wide profiling of the epigenome have emerged, thus a vast variety of epigenetic datasets has become available in public databases.

1.1.1 DNA methylation

When methyl groups are covalently bound to specific parts of the DNA backbone, thus changing the chemical structure of the DNA, while keeping the sequence of nucleobases in place, is called DNA methylation. Those changes are heritable and are essential for normal development of an organism [5]. They are also tightly linked to gene regulation and histone modifications and it appears that a reciprocal influence between DNA-methylation and specific histone modifications exists [5].

1.1.2 Histone modifications

Structural organization of the chromatin

The chromatin of a eukaryotic cell consists of DNA, RNA, and proteins, and embodies the state in which DNA is packaged within the cell. The first level of the cascadic packaging is the assembly of the DNA in nucleosomes. The nucleosome, which represents a basic unit of chromatin, is composed of an octamer of the four core histones (H2A, H2B, H3, H4) and DNA coiled around it, which is about 146 base pairs long. Those proteins are composed of a high quantity of positively charged aminoacids which warrant the electrostatic attraction

to the negatively charged DNA. The nucleosome can be considered as a basic unit DNA packaging which repeats itself throughout the whole chromatin (figure 1.1).

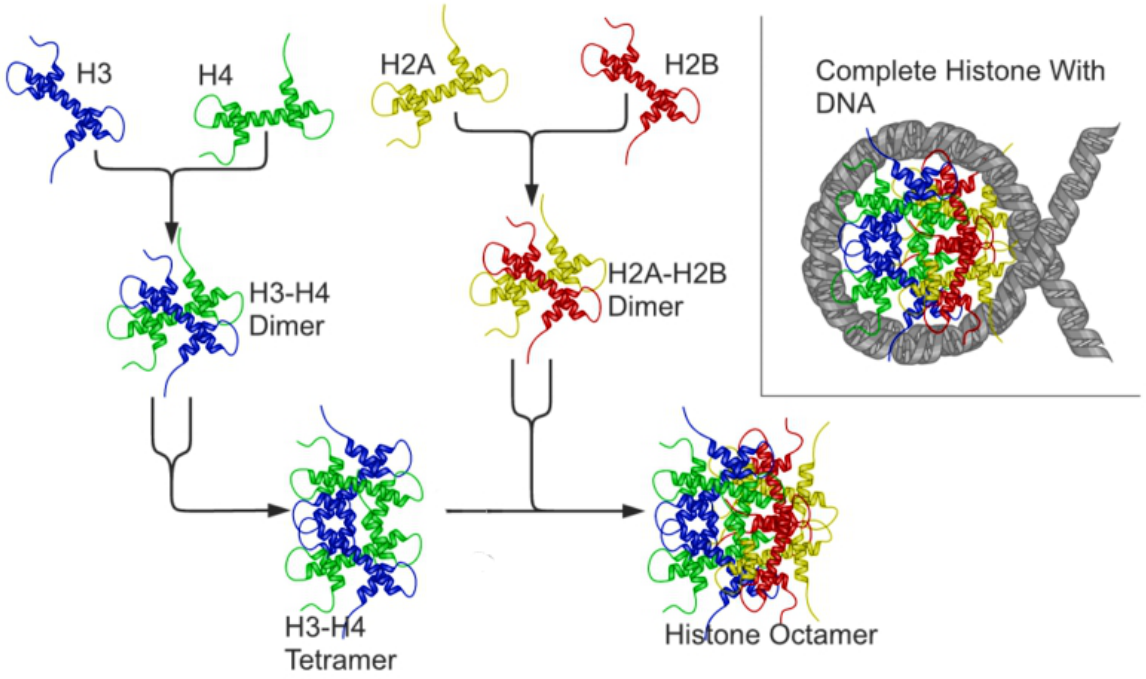

Figure 1.1: Schematic assembly of the nucleosome [6].

Nucleosomes are interconnected via so-called 'linker-DNA', which is associated with the linker-histone H1 (see figure 1.2).

The nucleosomal core histones have elongated tails which point outwards of the core and can be accessed by other proteins, thus providing mechanisms to control gene expression.
Those N-terminal tails of the core histones can be modified by mono-, di- and trimethylation, phosphorylation, ubiquitylation, acetylation and SUMOylation [8]. Those modifications change the strength of the bond between DNA and the histones, leading to a change of the three-dimensional structure of the packaged chromatin.
Tightly packed DNA is referred to as Heterochromatin whereas loosely assembled DNA is called Euchromatin. Due to this difference in condensation, genes are accessible or inaccessible for transcription, leading to activation or silencing of the gene.

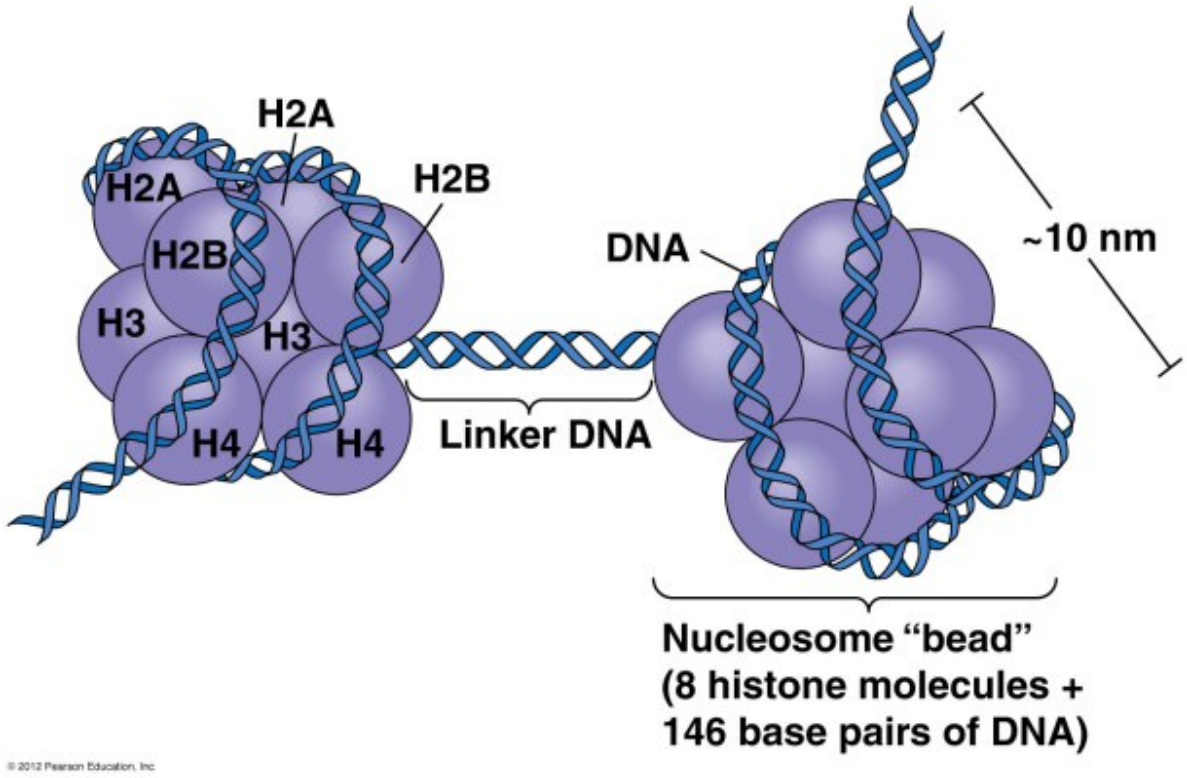

Figure 1.2: Schematic assembly of the the histones of two nucleosomes showing the linker DNA [7].

Acetylations and Methylations

A variety of studies has been carried out to characterize histone modifications for different organisms in detail.
Krogan et al. for example showed that in yeast, the methylation of Lys36 of the histone H3 (H3K36me) by enzyme Set2 is linked to active transcription [10], whereas Briggs et al. suggested that Set1-mediated H3 Lys4 methylation (H3K4me) is associated with transcriptional silencing [11].

In this thesis a set of lysine acetylations and trimethylations of the H3 and H4 histones for the organism *S. cerevisiae* is considered: H3K4me3, H3K36me3, H3K9ac, H3K14ac, H3K56ac, H4K16ac, and H4K5ac [12].

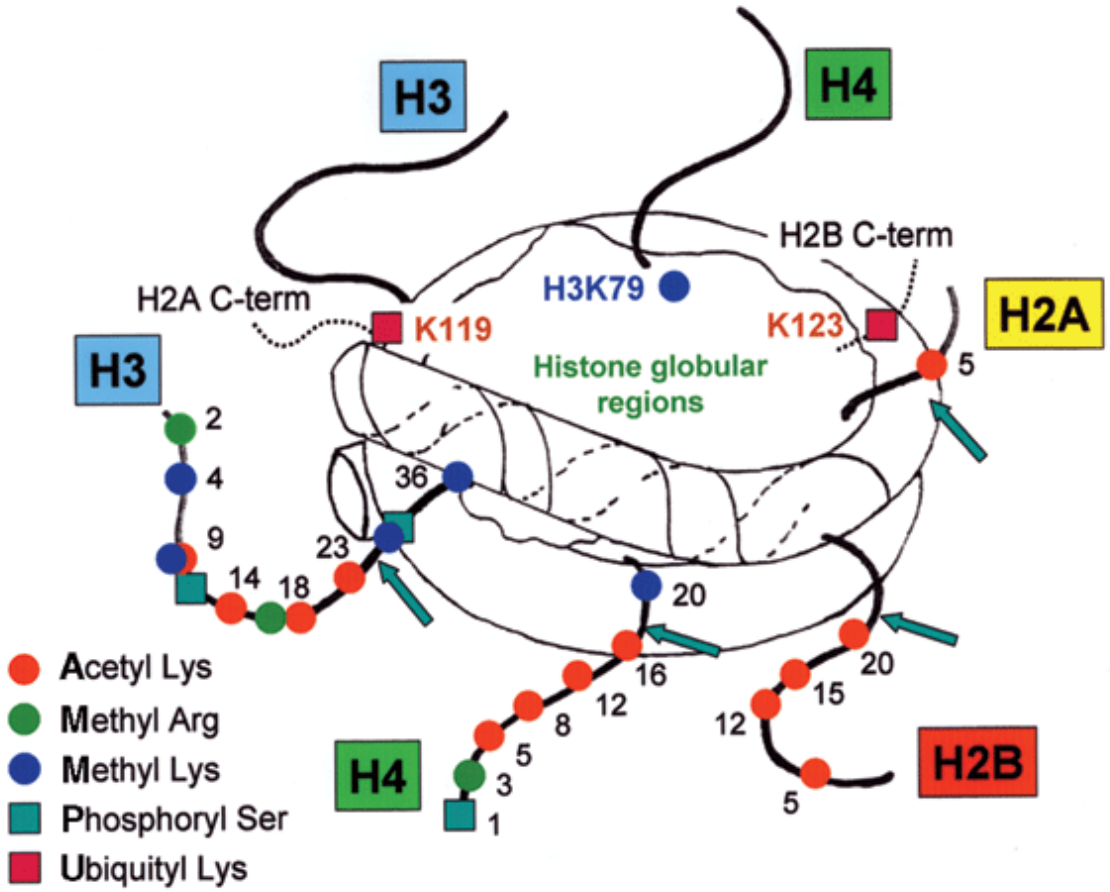

Figure 1.3: A nucleosome core particle, which shows six of the eight core histone N-terminal tail domains and two C-terminal tails. The colored dots show sites of modifications [9].

Histone acetylation

Acetylation of lysines is a common and highly dynamic process, carried out by two families of enzymes, the so-called histone acetyltransferase (HATs) and histone deacetylase (HDACs). HATs trigger a replacement of a hydrogen atom in a reactive amino group with an acetyl moiety ($CO\text{-}CH_3$) [13]. The N-terminal acetylation of the ϵ-amino group of the lateral chain of lysines is shown in figure 1.4. The positively charged residues of a histone make it a highly basic protein which leads to a high affinity for the negatively charged DNA. The acetylation of lysines and aginines neutralizes the positive charge and so weakens the bond between histone and DNA. It is thus often associated with euchromatin.

Table 1.1 shows a list of the yeast genes associated with the GO term histone

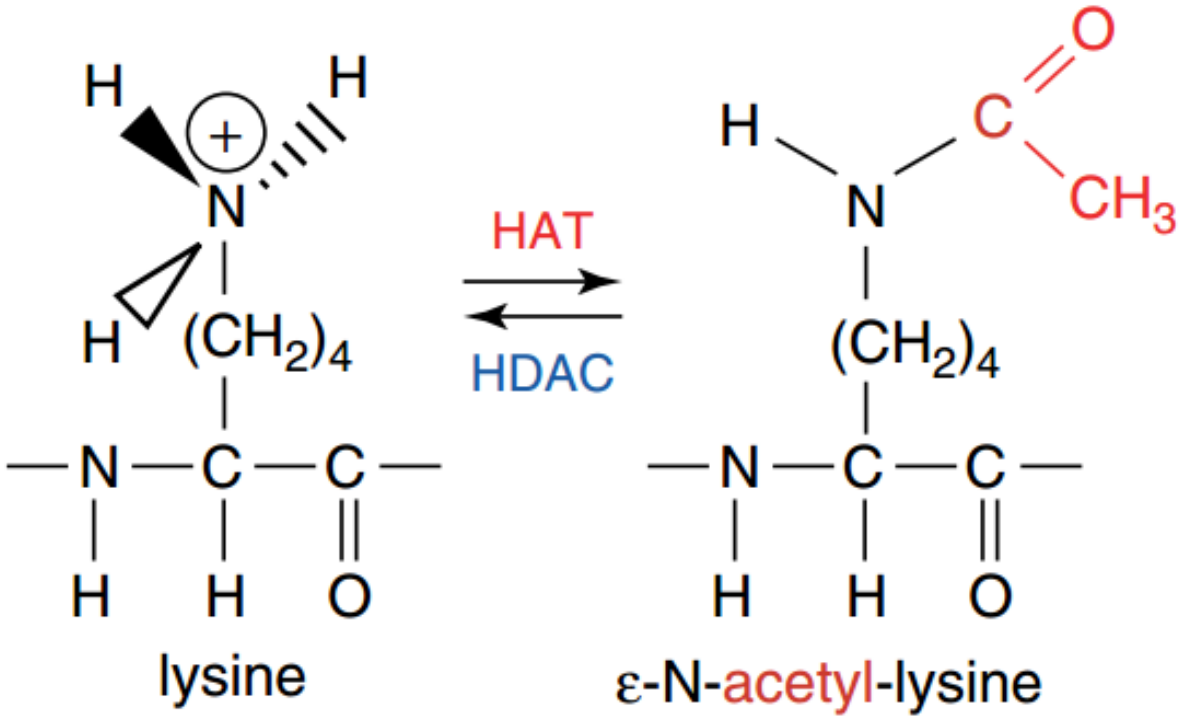

Figure 1.4: Structural change by N-terminal acetylation of the ϵ-amino group of the lateral chain of lysines.

acetyltransferase (GO:0004402)

Histone deacetylases (HDACs) revert the acetylation carried out by HATs and are associated with the GO term GO:0004407. An overview of genes associated with HDAC is shown in table 1.2.

Histone methylation

In constrast to acetylation and phosphorylation, methylation does not change the overall charge of a histone / a nucleosome. Therefore no clear linkage to Heterochromatin and Euchromatin can be established and histone methylation can be either repressive or activating, depending on the location of the methylated residue, which suggests that other proteins interact with the chromatin due to specific pattern of the histone marks. It has been shown that methylations occur on the side chains of lysine, arginine and histidine [13].

Table 1.1: *S. cerevisiae* genes which are related to histone acetyl transferase activity (GO:0004402) [14]

Name	Description
GCN5	Catalytic subunit of ADA and SAGA histone acetyltransferase complexes
HAT1	Catalytic subunit of the Hat1p-Hat2p histone acetyltransferase complex
ADA2	Transcription coactivator
YOR338W	Putative protein of unknown function
HPA2	Tetrameric histone acetyltransferase
SPT10	Histone H3 acetylase with a role in transcriptional regulation
ARP4	Nuclear actin-related protein involved in chromatin remodeling
TRA1	Subunit of SAGA and NuA4 histone acetyltransferase complexes
YNG2	Subunit of NuA4, an essential histone acetyltransferase complex
AHC1	Subunit of the Ada histone acetyltransferase complex

Histone methyltransferases (HMTs) are enzymes, which catalyse the mono-, di- and trimethylation of mainly lysine and arginine. This multiple stages of methylations add an additional level of complexity to the analysis of their modes of actions.
HMTs are highly specific enzymes when it comes to the location of the modification, which necessitates a more complex classification of the modifying proteins. Yet, most of the HMTs that methylate N-terminal lysines share the SET domain, with the exception of the Dot1 enzyme [16]. Nevertheless, abundance of proteins of the SET family is not soley linked to histone methylation, as they also modify some other proteins in addition to histones [17].
The N-terminal (de)methylation of lysine by histone lysine methyltransferase (KMT) and histone lysine demethylase (KDM) is shown in figure 1.5.

Table 1.2: *S. cerevisiae* genes which are related to histone deacetylation activity (GO:0004407) [15]

Name	Description
HST2	Cytoplasmic NAD(+)-dependent protein deacetylase
RXT2	Component of the histone deacetylase Rpd3L complex
HDA2	Subunit of the HDA1 histone deacetylase complex
HST3	Member of the Sir2 family of NAD(+)-dependent protein deacetylases
HDA1	Putative catalytic subunit of a class II histone deacetylase complex
RPD3	Histone deacetylase, component of both the Rpd3S and Rpd3L complexes
HOS3	Trichostatin A-insensitive homodimeric histone deacetylase (HDAC)
SIN3	Component of both the Rpd3S and Rpd3L histone deacetylase complexes
HST1	NAD(+)-dependent histone deacetylase
SAP30	Component of Rpd3L histone deacetylase complex
SIR2	Conserved NAD+ dependent histone deacetylase of the Sirtuin family
HOS2	Histone deacetylase and subunit of Set3 and Rpd3L complexes
SDS3	Component of the Rpd3L histone deacetylase complex
HDA3	Subunit of the HDA1 histone deacetylase complex
HOS1	Class I histone deacetylase (HDAC) family member
HST4	NAD(+)-dependent protein deacetylase

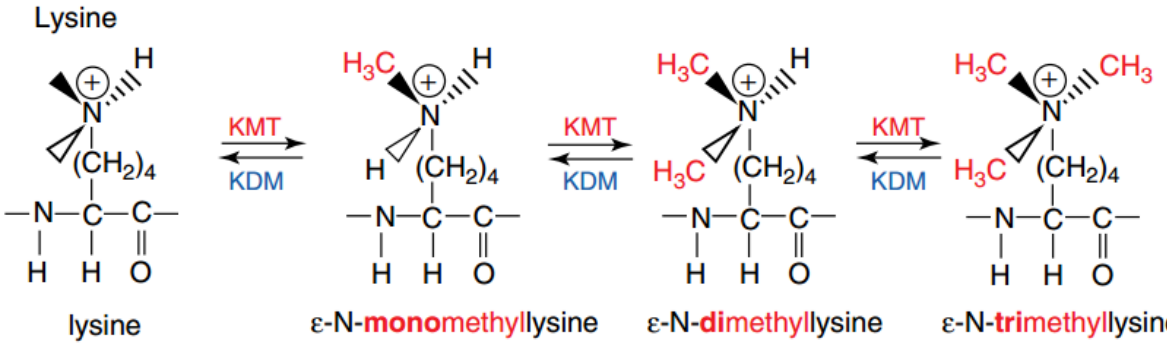

Figure 1.5: Structural change by N-terminal (de)methylation of lysine by histone lysine methyltransferase (KMT) and histone lysine demethylalse (KDM).

Histone code

The histone code hypothesis states that transcription of genes is not only regulated by the genetic information stored in the DNA, but also in the packaging of the chromatin [18]. A variety of proteins has been discovered, which can recognize specific patterns of histone modifications (Histone Readers) and interact with the chromatin accordingly [19, 20]. Therefore, it is standing to reason that many modifications serve as binding sites for proteins which translate the information stored in the histone modifications into certain processes. Thus, the combination of different histone modifications at different loci leads to an enormous amount of information, which could be possibly encoded in the chromatin, hence called the histone code.

Complexity of the histone code

To give an impression of the sheer complexity of histone code, the following example can be considered. Lysine methylation occurs on 5 sites on H3: H3K4, H3K36, H3K79, H3K9, and H3K27 [21]. If the modifications are considered independent from each other and each lysine can be un-, mono, di- or trimethylated, then 1024 different combinations times the number of H3 histones in a cell are possible, only for this small set of lysine methylations of a single histone.

For *S. cerevisiae*, the reference genome of the haploid strain S288c contains 12.3273 mega base pairs [22]. Assuming that the spacing of each nucleosome is about 200 base pairs [23] and

each nucleosome contains 2 H3 histones, for 1024 different lysine methylation patterns can be deducted that

$$\#combinations = \frac{12.3273 \cdot 10^6}{200} \cdot 2 \cdot 1024 = 1.2623155 \cdot 10^8 \tag{1.1}$$

Each other modification multiplies the number of patterns, thus the set of possible permutations grows exponentially with the number of different modifications.
Although this estimation is only a very rough estimate, it becomes clear how vast the number of different patters is and how much information can be encoded by such patterns.

1.1.3 NGS technologies for epigenetic profiling

Next generation sequencing technologies, also known as high throughput sequencing, aim at parallelizing the sequencing process to obtain millions of sequences concurrently [24].
The high demand for cheap whole genome sequencing has driven the development of high throughput methods. This progress facilitated the genome wide survey of epigenetic markers and the cost of DNA sequencing has been decreased more than a million fold in the last decade [25].
An overview of current NGS sequencing technologies [26] is given in table 1.3.

RNA-seq

RNA-seq is also referred to as whole transcriptome shotgun sequencing, as sequencing of all transcripts, including mRNAs, small RNAs, and non-coding RNAs [28] becomes feasible.
With RNA-seq it can be determined, which exons are expressed and the detection of novel splicing variants is possible.
For expression profiling, DNA microarrays have been, and still are a common technique. Due to background noise and cross-hybridization, microarrays perform poorly in detecting genes with low expression. Furthermore, the measured microarray probe intensity depends on the transcript affinity under the specific hybridization conditions. Due to this fact, the probe intensity is not actually proportional to transcript abundance. Those issues do not impact RNA-seq [29].

Table 1.3: Selection of current NGS technologies (adapted from [27]).

Sequencing devices	Chemistry	Read length (bp)	Run time	Throughput per run	Reads per run
High-end instruments					
Ion Proton	Proton detection	200	2.5h	15GB	60 - 80 million
Pacific BioSciences RS II	SMRT - P6-C4	50% $\geq$ 14kb 5 % $\geq$ 24kb max $\geq$ 40kb	0.5 - 6h	0.5–1 billion bases per SMRT cell	-
HiSeq 2000/2500 (Illumina)	Reversible terminator	2 x 150	High output: 11 days Rapid run: 27 h	High output: 600 Gb Rapid run: 120 Gb	High output: 3 billion x 2 Rapid run: 600 million x 2
5500xl W SOLiD (Life Technologies)	Ligation	1x75 Frag, 2x50 MP	8 days	320 Gb	1.4 billion x 2
Bench-top devices					
Ion PGM (Life Technologies)	Proton detection	100 or 200	3 h	100 Mb (314 chip) 1 Gb (316 chip) 2 Gb (318 chip)	400–550 thousand (314 chip) 2–3 million (316 chip) 4–5.5 million (318 chip)
MiSeq (Illumina)	Reversible terminator	2x250	27 h	8.35 Gb	6.8 million (LRGC routinely getting > 15 M)

In this thesis RNA-seq data from the study *High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast* [12] is used to profile the gene expression level of *S. cerevisiae* cells at different points in time.

ChIP-seq

Chromatin immunoprecipitation (ChIP) followed by sequencing is a direct way for the identification of histone modifications or DNA binding sites of proteins [30].
ChIP-seq is a methodology to study DNA-binding proteins or other factors that associate with DNA or chromatin. The cell nucleus is isolated and proteins bound to DNA are covalently cross-linked with DNA. The resulting chromatin-protein complexes are sheared into suitable fragments using sonication or nuclease digestion. Then an antibody against the specific protein or protein modification is used to precipitate the protein bound to DNA; cross-linking is reversed, and DNA is purified.

The purified DNA fragments are made suitable for sequencing by adding appropriate adapters in order to create libraries that can be loaded into deep sequencing instruments, such as the HiSeq 2500 from Illumina.
The sequenced reads are then mapped back to the genome, and enriched regions are identified (peaks) as areas where the protein was bound [31].
This technique is used to detect occupancy of transcription factor binding sites and histone modifications in the living cell.
In this thesis ChIP-seq data from the study *High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast* [12] is used to analyze 7 different histone modifications of *S. cerevisiae* cells at different points in time.

1.2 Goals of this thesis

In this thesis, publicly available RNA-seq and ChIP-seq should be used to investigate relations between histone modifications and gene expression, i.e. how certain histone modifications impact gene regulation and how well a given modification pattern can be used to predict a gene's expression level.
Specifically the following should be done/analyzed:

- Finding a suitable dataset
- Downloading and preprocessing the dataset
- Analysis of the relation between histone modifications and gene expression
- Using an algorithm for gene regulatory network reconstruction, i.e. ARACNE, for the epigenetic data
- Gene onthology enrichment analysis for epigenetic data

2 Methods

2.1 Data aquisition

In 2014, Kuang et al. carried out a study in which they analyzed transcriptome and chromation changes of *Saccharomyces cerevisiae* under glucose-limited conditions [12] for 16 different points in time. Under continuous, nutrient-limited conditions, a periodic respiratory cycle, called the yeast metabolic cycle (YMC), can be observed, where more than 3000 transcripts oscillate at the same pace. The metabolic cycle can be divided in the oxidative phase, the reductive building phase, and the reductive charging phase.
Kuang et al. exploited that robust oscillation to create a high-temporal-resolution dataset for transcription and chromatin states throughout this cycle.

ChIP-seq and RNA-seq samples were sequenced with Illumina HiSeq 2000, Illumina Genome Analyzer and AB SOLiD System and made publicly available in the NCBI GEO database [32] under accession GSE52339.
The raw RNA-seq reads were mapped and further processed in this thesis. Therefore, the 16 raw sequence read archives (SRA) were downloaded, one for each point in time (an overview over that data and it's dimensions are given in table 2.1).

The ChIP-seq reads are also available on the NCBI, but the mapping and peak calling was not performed in the thesis and the modification signals for H3K4me3, H3K36me3, H3K9ac, H3K56ac, H4K16ac, H4K5ac, and H3K14ac for 16 points in time were taken from the supplementary data of the publication.

Table 2.1: Overview over the raw RNA-seq data from the NCBI database [32].

Point in time	Sample	Run	# of Spots	# of Bases	Size
1	RNA-seq_t1	SRR1029143	11 339 219	567M	367.3Mb
2	RNA-seq_t2	SRR1029144	10 650 839	532.5M	344.8Mb
3	RNA-seq_t3	SRR1029145	9 251 840	462.6M	301.6Mb
4	RNA-seq_t4	SRR1029146	9 891 366	494.6M	319.8Mb
5	RNA-seq_t5	SRR1029147	10 120 711	506M	326.6Mb
6	RNA-seq_t6	SRR1029148	9 930 864	496.5M	321.6Mb
7	RNA-seq_t7	SRR1029149	10 099 686	505M	326.2Mb
8	RNA-seq_t8	SRR1029150	10 926 593	546.3M	352.6Mb
9	RNA-seq_t9	SRR1029151	9 017 221	450.9M	291.3Mb
10	RNA-seq_t10	SRR1029152	10 612 655	530.6M	343.2Mb
11	RNA-seq_t11	SRR1029153	8 732 109	436.6M	282.8Mb
12	RNA-seq_t12	SRR1029154	9 746 383	487.3M	315.4Mb
13	RNA-seq_t13	SRR1029155	10 524 434	526.2M	340.3Mb
14	RNA-seq_t14	SRR1029156	9 756 118	487.8M	315.9Mb
15	RNA-seq_t15	SRR1029157	9 753 159	487.7M	315.4Mb
16	RNA-seq_t16	SRR1029158	15 842 279	792.1M	514.8Mb

The download of the data from the NCBI was performed using the SRA Toolkit [33]. Using the NCBI GEO accession number GSE52339 a list of the entries was compiled in a file SSRindex (appendix A1). To download the data form the NCBI, the following command was issued:

```
1 cat SRRindex | sed -r 's/^(\w+).*/\1/g' | while read in; do (fastq-
    dump "$in") >> out ; done
```

Listing 2.1: Read archives download

This creates a set of fastq-files using the fastq-dump utility from the SRA Toolkit [33]. The files were then renamed to a human readable name with a Ruby [34] interpreter:

```
1 filenames = Dir.glob("*.fastq")
2 fileNamesMap = Hash[*File.read('../SRRindex').split(/[\t \n]+/)]
3 filenames.each do |filename|
4   File.rename(filename, fileNamesMap[filename.gsub(/.fastq/,'')].to_s
      + ".fastq")
5 end
```

Listing 2.2: Human readable file names

Although this script downloads all the RNA-seq and ChIP-seq files, only the RNA-seq files were used for further processing, as the preprocessed ChIP-seq data was taken from the original study.

2.1.1 Analysis

The operating system Ubuntu 16.04 was used as the platform for the analysis.
For the processing of the RNA-seq data Tophat 2.1.1 [35] and Cufflinks 2.2.1 [36] were used.

Used programming languages:

- Ruby 1.9 [34]
- Groovy 2.4 [37]
- Matlab 1205b [38]
- Linux Bash

2.1.2 Data preprocessing

RNA-seq

In the last few years many mapping programs have been developed, which use different concepts to achieve good mapping results [39–41].
Among those, a very efficient and popular one is Bowtie [42]. It uses a data structure called FM index for the storage of the reference genome which allows a fast search. Yet Bowtie is not the ideal choice for all alignment tasks as alignments between a read and the reference genome must not contain large gaps. This makes it unsuitable for RNA-seq data which often spans introns and is subjected to splicing.
Tophat [35] is a mapper which uses Bowtie internally, but overcomes those problems, which makes it suitable for RNA-seq data mapping.
The workflow for obtaining between-sample normalized gene expression levels is shown in figure 2.1.
To map the reads with Tophat, gene annotation information and a reference genome is required.

Those were downloaded from the Saccharomyces Genome Database.

```
1 url = "http://downloads.yeastgenome.org/sequence/S288C_reference/
     genome_releases/"
2 url = "${url}S288C_reference_genome_Current_Release.tgz"
3 curl $url | tar xvz
```

Listing 2.3: Annotation and reference download

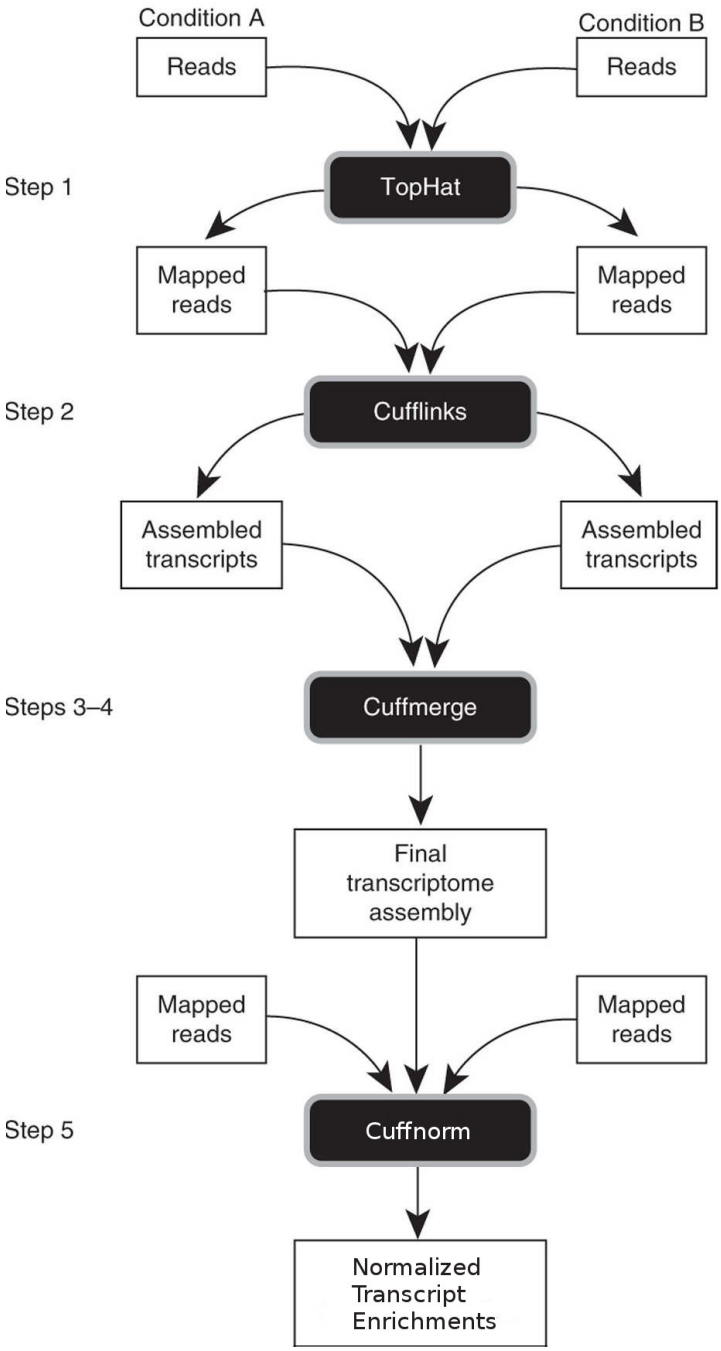

Figure 2.1: The Cufflinks toolchain shows the workflow used to obtain gene expression signals from the RNA-seq data (adapted from [36]).

The reference file in fasta-format must contain the same identifiers for the chromosomes as the annotation file. Therefore the identifiers were fixed with a 'regular expression':

```
1 cp S288C_reference_genome_R64-2-1_20150113/
     S288C_reference_sequence_R64-2-1_20150113.fsa ./reference.fa
2 sed -i 's/ref.*chromosome=\(.*\)\]/chr\1/g' reference.fa
3 sed -i 's/ref.*/chrmt/g' reference.fa
```

Listing 2.4: Rename chromosome identifiers in reference genome

Tophat was called with the following parameters:

```
1 while read p; do
2   tophat --bowtie1   -i 40 -N 4 -p 8  -o "tophatOut_$p"  \
3     --read-edit-dist 4  --segment-length 18 --no-coverage-search -
         G \
4       S288C_reference_genome_R64-2-1_20150113/
           saccharomyces_cerevisiae_R64-2-1_20150113.gff \
5       genome "$p.fastq" > "tophat_$p.out"
6 done < rnaseqfiles.txt
```

Listing 2.5: Mapping with Tophat

The manually created file *rnaseqfiles.txt* (appendix A2) contains a list of the 16 RNA-seq filenames.

The resulting alignment files were then passed to Cufflinks, which assembles the transcriptome for each time sample.

```
1 find ../ -type f -name 'accept*' -execdir cufflinks -p 8 {} +
2 find . -name transcripts.gtf > manifest
3 cuffmerge -p 8 -g S288C_reference_genome_R64-2-1_20150113/
     saccharomyces_cerevisiae_R64-2-1_20150113.gff manifest
```

Listing 2.6: Cufflinks toolchain

Then Cuffnorm was used to obtain the normalized gene expression fragments per kilobase million (FPKM) values for all samples:

```
1 cuffnorm -p 8 -o cuffdiff_out merged_asm/merged.gtf $samples
```

Listing 2.7: Cuffdiff

The resulting FPKM values for each gene represent the gene expression. Matlab was used to visualize the expression in a heatmap (figure A10).

ChIP-seq

The histone modification profiles were obtained from Kuang et al. [12]. They state that ChIP signals of histone modifications were evaluated by counting the reads which overlap defined regions of every gene.
For H3K9ac, H3K14ac, H3K56ac, H4K5ac, and H3K4me3 they used -100 to +400 base pair windows spanning the transcription start site (TSS) of each gene.
For H3K36me3, H4K16ac, and H4 signals were calculated from TSS to the transcription end site (TES). Furthermore, they used the MACS peak-calling package [43] with a p-value cutoff of 0.01 and annotated the peaks with the nearest TSS.
The ChIP-seq samples were normalized by Kuang et al. according to the total number of aligned reads, followed by standardization of the 16 observed values for each point in time, yielding 7 standardized modification levels for 5205 genes for each point in time.
The Pearson correlation was calculated with Matlab between each modification dimension and gene expression using the samples from all points in time at once.

2.2 Clustering

2.2.1 k-means

To find similar histone modification profiles, k-means clustering [44] was used. Therefore the modifications for all time samples were considered simultaneously. Each of the 5205 gene modification patterns at all 16 points in time was concatenated in one big matrix, which gave a total of 83280 profiles for 7 different modifications.

The k-means algorithm starts with a predefined number of random cluster centers and tries to position them in a way that the variance within a cluster is minimized. Different algorithms exist, which are commonly referred to as k-means. Here the Lloyd k-means algorithm variant [45] was used, which iteratively assigns each point to the closest

cluster center and then calculates the new cluster center as the center of all the points in a cluster.

Thus, it is obvious that the distance metric plays an important role. In this case the Euclidean distance measure was used, as the data for each column is normalized and therefore the distance between different modifications is comparable.

One drawback compared to other clustering algorithms, such as hierarchical clustering, is that the number of clusters k must be chosen in advance. Therefore, a criterion for quantifying the quality of the model is necessary.

In this thesis, several different approaches were tried to determine the optimum number of clusters.

A higher number of clusters results in a lower sum of squared distance within each cluster, but leads to overfitting and generalizes the data badly. Thus, the criterion should penalize a high number of parameters for the model which describes the data.

Several methods for rating the quality of the clustering results exist. One popular method is the silhouette method proposed by Peter J. Rousseeuw [46].

For each point in the dataset a silhouette value is calculated, which measures how similar that point is to other points in its cluster in comparison with points in other clusters.

For n clusters the silhouette value S_i for the ith point p_i is defined as

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{2.1}$$

where a_i is the average distance between the ith point and other points within the cluster c_k where $\{k \in \mathbb{N} | 1 \leq k \leq n\}$ to which the ith point was assigned and b_i is the average distance to points within another cluster c_l, where c_l is the cluster for which b_i is minimal with $\{l \in \mathbb{N} | \{1 \leq l \leq n\} \backslash k\}$. This results in a silhouette value for each point p_i which ranges between -1 to $+1$. The higher a silhouette S_I value for a point p_i the better the match to its own cluster, and the poorer the match to another cluster. If many points have a low or negative silhouette value, the number of clusters selected have been either too high or too low.

Using this measure for evaluation of a clustering result, different values k for the number of clusters can be compared.

In the given scenario this method proved to be computationally very intensive, as 83280 values in 7 dimensions were clustered, and was therefore not usable.
Another approach is Figure of Merit (FoM) [47], which quantifies the predictive power of the model. For each number of clusters the model is trained while omitting one sample. Then the variance within the omitted sample is calculated for each cluster.
The FoM was calculated, but promoted a very low number of clusters, which is why another criterion was used.

A well-known method is the Akaike Information Criterion (AIC) [48] which is based on an information theoretic approach, namely the Kullback-Leibler divergence which quantifies the information loss when a distribution $f(x)$ is represented by a model $g(x|\Theta)$.
Akaike was able to show the relation between the Kullback-Leibler information and the likelihood theory. He introduced his formula for an information criterion

$$AIC = -2\mathcal{L}(\Theta) + 2p \tag{2.2}$$

where $\mathcal{L}$ is the log-likelihood function, Θ are the model parameters and p is the number of parameters.
For k-means, the number of parameters is the number of cluster centers k times the dimensionality of each cluster center μ_i, because each dimension can be varied freely.
The AIC can be derived as

$$AIC = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\vec{x}_i - \vec{\mu})^2 + 2kd \tag{2.3}$$

The derivation of this formula can be found in theorem 1 in the appendix.
The AIC decreased very slowly which is why another scoring function, namely the Bayesian Information Criterion (BIC) was evaluated, which penalizes the number of parameters more strongly. The name of the BIC is misleading, as it is not based on an information theoretic approach, but on Bayesian inference [49].

Gideon Schwarz derived the Bayesian information criterion [50] as

$$BIC = -2\log(\mathcal{L}(\Theta)) + p\log(n) \tag{2.4}$$

Using the same assumptions about the likelihood function $\mathcal{L}$ for k-means as for the AIC (theorem 1), the BIC can be formulated in the following way

$$BIC = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\vec{x_i} - \vec{\mu})^2 + 2kd \log(n). \tag{2.5}$$

where σ denotes the standard deviation, μ is the mean of the data, k is the number of clusters, d is the dimensionality of the data, and n is the number of data points.

2.2.2 Cluster evaluation

Since the focus of clustering was not on detecting genes, which exhibit similar modification patterns, but on investigating the correlation between histone modification and transcription, the clusters were evaluated for their average expression. Each cluster contains a set of modification profiles where each profile can be mapped to a gene at a specific point in time. Thus, transcriptional activity for each profile is known from the RNA-seq dataset.

With that additional information, the expression values associated with the profiles within each cluster were obtained. In order to quantify transcriptional activity, the mean of all expression values within a cluster $cl_{modprof}$ was calculated and divided by the standard deviation (*z-score*) of the expression values within the cluster, as shown in equation 2.6, to make the expression values of the clusters comparable.

$$expr_{cl_{modprof}} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} (x_i - \mu_{cl})^2 \tag{2.6}$$

where μ is the mean expression in each cluster, σ is the expression standard deviation in each cluster, and n is the number of elements within the cluster.

Thresholds of -0.5 and 0.5 were used to assign each cluster a label of *silencing* for clusters with $expr_{cl_{modprof}}$ below -0.5, *no impact* for clusters between 0.5 to 0.5 and *activating* for clusters above 0.5.

For each of the groups (*silencing*, *no impact*, *activating*) all patterns belonging to the respective group were extracted and an estimate of the density function for each modification was plotted. The cluster centers of the groups *silencing* and *activating* were used to generate a color-encoded table, which gives an overview of the patterns and their impact on gene expression.

To visualize the clustering result and the clusters' impact on expression, a scatterplot matrix was created where each modification is plotted against all others and the clusters' segmentation into *silencing*, *no impact* and *activating* is encoded with different colors.

In order to elucidate linear relations between different modifications, the correlation of all modification patterns within clusters of the groups *silencing* and *activating* was calculated and compared with the correlation of clusters in the group *no impact*.
Tables with conditional formatting were created with OpenOffice Calc.
The give a better visualization, the correlation matrix was interpreted as an adjacency matrix and a graph was generated with Javascript [51] and D3 [52] (figures 3.7 and 3.6)

2.3 Mutual information

Detecting and analyzing interactions and dependencies in a multivariate setting is very useful in many different areas, like life sciences, finance and many machine learning applications.
Surprisingly, no formal definition of dependence in a multivariate setting exists, but a variety of dependence concepts has been formulated using standard statistical methods, which are mostly only applicable in a restricted context [53].
Correlation for example is very useful in finding linear dependencies between two random variables X and Y, but may result in a correlation coefficient $r = 0$ for non linear relations.

Mutual information [54] is a concept based on information theory, i.e. a comparison of the marginal entropies of the random variables and the entropy of the joint probability distribution (JPD) of the random variables.
The definition of mutual information $I(X;Y)$ is

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.7}$$

where X and Y are random variables, $p(x)$ and $p(y)$ are the marginal distributions and

$p(x,y)$ is the JPD.

This definition is very closely related to the Kullback-Leiber divergence $D(f||g)$, also known as relative entropy, which measures the distance between two models f and g [55]:

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{2.8}$$

$D(f||g)$ quantifies the loss of information, when g is used to approximate f. This leads to an intuitive understanding of mutual information which can be expressed as relative entropy:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.9}$$

This equation can be interpreted as "How much information is lost, when the product of the marginal distributions $p(x)p(y)$ is used to approximate the JPD $p(x,y)$?", which is why independence can be formulated as $I(X;Y) = 0$ *iff* $p(x,y) = p(x)p(y)$.

This statement can easily be visualized by using Gaussian mixtures of two dependent random variables and comparing the approximation of the JPD with the actual JPD (figure 2.2).

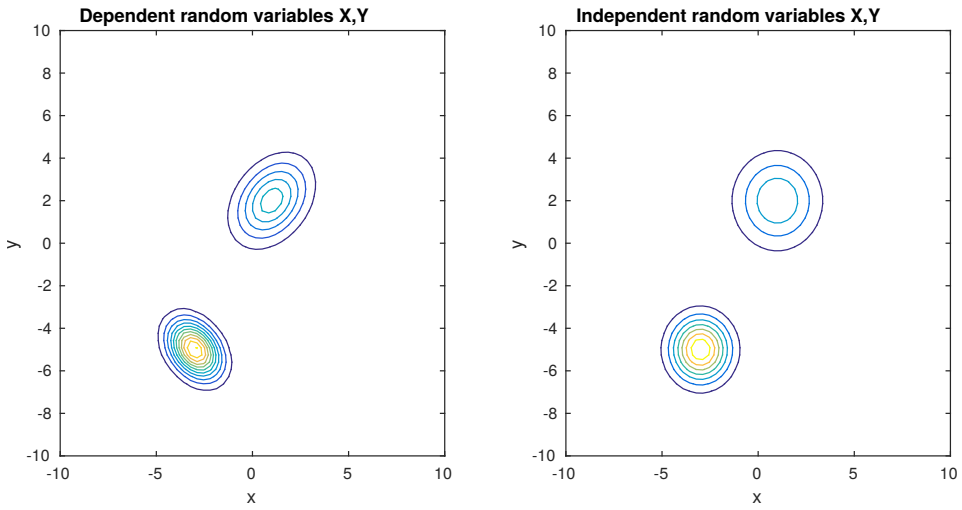

Figure 2.2: Example of Gaussian mixtures for dependent and independent random variables which share the same marginal distributions. The Gaussians on the right share no mutual information, as the joint distributions can be perfectly approximated by the product of the marginal distributions. The Gaussians on the left share mutual information, as the approximation of the joint distribution from the marginals is not possible - hence X and Y can be considered dependent on each other.

In the case of independent random variables X and Y the JPD is the product of the marginal distributions. In the case of dependent random variables in figure 2.2 the marginal distributions are exactly the same as in the independent case, but the distribution of the independent random variables would be a bad approximation of the JPD of the dependent case.

As mutual information measures the Kullback-Leibler divergence between the product of the marginal distributions and the JPD function, it is also sensitive to dependencies which do not manifest themselves in the covariance and thus cannot be detected with a correlation coefficient [56].
It is also very common to formulate the mutual information between X and Y using the entropy:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.10}$$

where $H(X)$ is the entropy of X and $H(X,Y)$ is the entropy of the JPD of X and Y.

Another important feature of mutual information is its reparametrization invariance, i.e. if $X' = f(X)$ and $Y' = g(Y)$ are homeomorphisms then $I(X';Y') = I(X;Y)$ [56].

2.4 ARACNE

ARACNE [57] uses the concept of mutual information to calculate pairwise statistical dependencies between the samples of two genes.
As mutual information is reparametrization invariant, the expression values for g_1 and g_2 are first copula-transformed [57], which gives identically distributed marginals in the range of [0,1].
Therefore, a performant calculation of the mutual information, using a Gaussian Kernel estimator for the JPD, is possible, where the kernel width is not position dependent.
The resulting mutual information $I(g_i, g_j)$ is then filtered by an appropriate threshold I_0, which is computed for a specific p-value p_0. This approach suffers from the limitation that highly co-regulated genes may lead to false positives, as indirect relationships but no direct interactions exist.

To cope with that problem, ARACNE removes a majority of detected interactions by using the so called data processing inequality.
The data processing inequality states that if an interaction between g_1 and g_2 only exists over a third gene g_3, then

$$I(g_1, g_2) \leq min\,[I(g_1, g_3), I(g_2, g_3)] \tag{2.11}$$

i.e. the mutual information from the indirect interaction between g_1 and g_2 must be less or equal than the mutual information between g_1 and g_3 or g_2 and g_3.

The full algorithm can be described as calculating the mutual information for each gene pair after copula-transforming their expression values and building a graph from the mutual information values while rejecting adjacencies between genes g_i and g_j if $I(g_i, g_j) \leq I_0$.
All connected gene triplets in the graph are examined and the edge with the least mutual information is removed. It is worth noting that in order to account for variance of the mutual information a tolerance for the removal of edges can be specified for cases where the mutual information between the three genes exhibits little difference, thus leading to persistence of loops between three genes in some cases.
ARANCE was used for assessment of mutual information between the different modification profile clusters. A dataset containing the modification patterns from all clusters labeled with *no impact* was exported as **.tsv* file, as well as another file containing all patterns from clusters *activating* and *silencing*.

Then ARACNE was called from the commandline for both files, in a way that the algorithm searches mutual information between each modification, using the modification signals for each gene in that group as sample.

```
1 aracne2 -H ARACNE -i alltimeModsUnknExprAracne.tsv -o
     alltimeModsUnknExprAracne.adj -p 1e-20  -a variable_bandwidth
```

Listing 2.8: Aracne commandline call for patterns with no impact on expression

```
1 aracne2 -H ARACNE -i alltimeModsUnknExprAracne.tsv -o
    alltimeModsUnknExprAracne.adj -p 1e-20  -a variable_bandwidth
```

Listing 2.9: Aracne commandline call for patterns with impact on expression

The resulting mutual information was again conditionally-color-formatted with OpenOffice Calc [58] and graphs were generated with the same script which was used for generating the correlation graphs.

2.5 Decision Trees

To obtain a model which shows the relation between histone modifications and gene expression and is also well interpretable, decision trees [59] were trained with Matlab using the *trainTree* function [60].
In a first approach, all the patterns and their corresponding expression values were used as dataset. The modification profiles were segmented into two classes, based on the corresponding expression signal for the pattern, i.e. class -1 for expression level less than 0 and class 1 for expression level greater than 0. The data was split into a training set containing 30% of the data and a validation set containing the remaining 70%.

The tree was trained with the training set using 20 fold cross validation and pruned to a height giving the best cross validation result.
This was accomplished by training the tree to a much larger height than the optimum and then using cross-validated classification error for the cutting it back down to a hight where the performance on the cross validation data starts declining.

In an attempt to increase the classification accuracy, patterns with an expression level between -0.5 and 0.5 were removed.

In a third experiment, a tree was trained to distinguish between patterns which show an impact on expression and patterns which show no impact. Therefore, patterns which correspond to an expression value less than -0.5 and greater than 0.5 were

labeled as class 1 and all other patterns were assigned to class 0. Again, the data was split into 30% training and 70% validation. The size of the overall data set was 83280 samples and 20 fold cross validation was used for training.

Table 2.2: Overview of the trees which were trained on the modification patterns in order to obtain well interpretable results on how the modification patterns impact gene expression. Different data subsets were used, either depending on the associated standardized gene expression e of a modification pattern or on the association of the modification pattern with a k-means cluster of the groups *activating*, *silencing* or *no impact*.
The prediction targets of the different trees is a binary classification of either a binarized expression value or a cluster group.

Subset	#Samples	Prediction
Full dataset	83280	$e \leq 0$ or $e \geq 0$
$e \leq$ -0.5 or $e \geq 0.5$	52421	$e \leq$ -0.5 or $e \geq 0.5$
Full dataset	83280	$-0.5 \leq e \leq 0.5$
Patterns from cluster *activating* or *silencing*	19136	Cluster group
Patterns from cluster *activating* or *silencing*	25688	$e \leq 0$ or $e \geq 0$
Patterns from cluster *activating* or *silencing* where $e \leq$ -0.5 or $e \geq 0.5$	18574	$e \leq$ -0.5 or $e \geq 0.5$

Another experiment was to train a decision tree to predict whether a pattern belongs to one of the previously calculated k-means clusters, labeled with positive or negative impact on expression.

A similar task as the previous was executed by segmenting the modification patterns into three classes, based on the cluster group they were assigned to. The three class labels were -1 for patterns, assigned to clusters of the group *silencing*, 0 for *no impact* and 1 for *activating*.

Another data set was generated by selecting only patterns that belonged to a k-means cluster which was labeled with positive or negative impact on expression. In contrast to the previous experiments, the class labels were not the ones assigned to the cluster, but the actual expression values for the pattern itself. Those expression values were quantified to -1 and 1, thus dividing the dataset into two classes and resulting in an overall dataset of size 19136.
A last experiment was executed with the same setup as the previous one, but with the difference that all patterns which showed an expression between -0.5 and 0.5 were removed.

2.6 GO enrichment analysis

The set of genes belonging to two clusters, which showed the strongest relation to gene expression were exported and imported into GORILLA [61] to perform a GO enrichment analysis for *cellular component*, *molecular function* and *biological processes*.

3 Results

The Pearson correlation coefficient between each modification dimension and gene expression shows that the available modifications mostly have an activating impact on gene regulation (see table 3.1). H3K9ac and H4K5ac show the highest correlation with gene expression.

Table 3.1: Pearson correlation was calculated for each histone modification and gene expression using all time samples at once. H3K9ac and H4K5ac show the highest correlation, followed by H3K14ac and H3K4me3. The only negative correlation is seen for H3K56ac, but the coefficient is close to zero.

Modification	correlation coeff	p-Value
H3K9ac	0.37	0
H4K5ac	0.34	0
H3K14ac	0.25	0
H3K4me3	0.20	0
H3K36me3	0.15	0
H3K56ac	0.01	0
H4K16ac	-0.10	0

3.1 Clustering

The Bayesian Information Criterion was used for determination of the number of clusters for k-means (figure 3.1). Although the BIC exhibited a minimum at 400, only 200 cluster centers were chosen, as difference of the BIC at k=200 and k=400 is low.
The formula for the calculation of the AIC and BIC was derived for k-means (theorem 1).

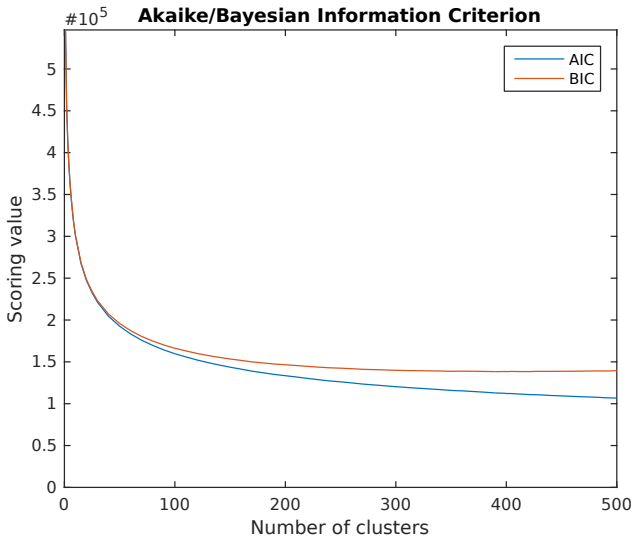

Figure 3.1: The cluster score for the Akaike Information Criterion and the Bayesian Information Criterion for k-mean clustering of the modification profiles. The selected value for k is 200, as the slope of the BIC curve is very flat.

Each of those clusters was assigned a class label based on the mean expression of the genes within a cluster. The threshold for assignment to class *silencing* was a mean expression level divided by the standard deviation (as described in section 2.2.2) within the cluster below -0.5, clusters with mean expression levels between -0.5 to 0.5 were labeled as *no impact* and clusters with mean expression above 0.5 were labeled as *activating*.

The majority of the clusters could not be related to having an activating or silencing effect on transcription; namely 135 clusters, containing 57588 modification profiles, which is 69.15% of all samples.

39 clusters were labeled as *activating*, containing 13284 modification profiles (15.95% of all samples) whereas 26 clusters were considered as *silencing*, which contained 12408 modification profiles (14.90% of all samples).

The density function for the mean expression was estimated and normalized by the clusters' standard deviation (figure 3.2) for the purpose of visualization.

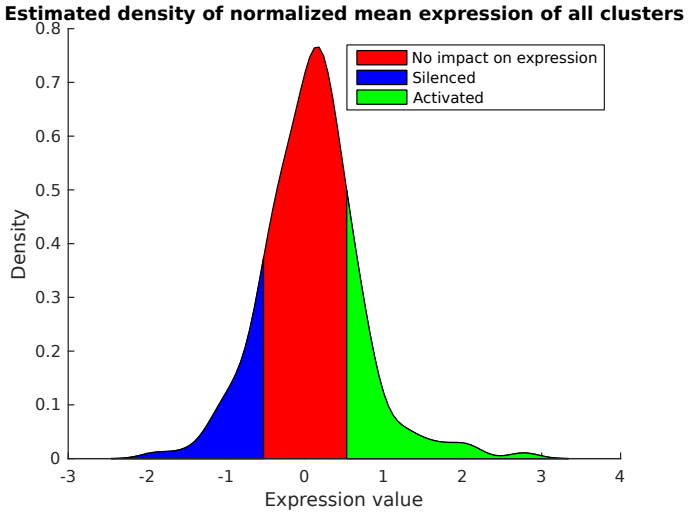

Figure 3.2: Estimated density function of mean expression values for each cluster, normalized by the standard deviation of each cluster. The blue area denotes the expression values which will be considered as *silenced*, the red area marks the expression values for which the modification patterns will be treated as *unknown impact* on expression and the green area will be considered as *activated*.

The density function for each modification dimension was estimated from all profiles for clusters for which *no impact* on expression was observed, as well as for upregulated clusters and downregulated ones (figure 3.4).
In the case of *no impact* on expression it can be seen that most modifications are well centered at zero, whereas in the other two cases almost all signal means are shifted away from zero.

It could be shown that certain modifications are related very closely to gene expression. H3K9ac and H4K5ac exhibit positive signals for almost all clusters associated with activation and negative signals for silenced ones. For H3K56ac and H4K16ac no obvious relation can be deducted visually from figure 3.3, which shows a matrix plot of all cluster centers for each modification dimension's cross section.
Centers of clusters whose average expression is not impacted are randomly distributed (figure 3.5).
To obtain another representation of the modifications' impact on expression, tables were

created for the cluster centers which were linked to *activation*, *silencing* and *no impact* and sorted by their mean expression within the cluster, normalized with its standard deviation (tables 3.2 and 3.3).

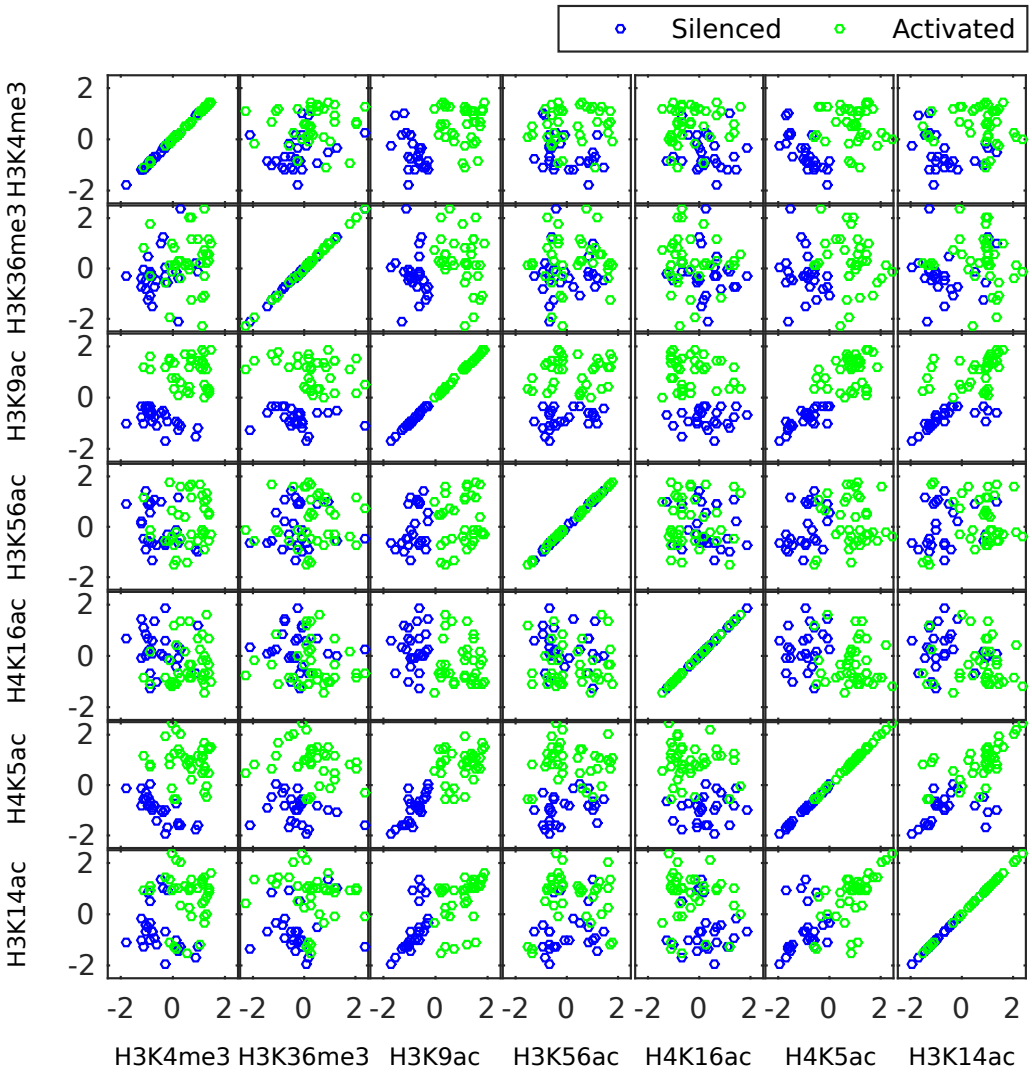

Figure 3.3: Cluster centers of the k-means clustering shown for each modification dimension tuple. The green circles show the cluster centers labeled as *activating* and the blue circles denote cluster centers related to *silencing*. In some dimensions the two groups are well linearly separable (e.g. H3K4me3 and H3K9ac), whereas in other dimensions no obvious relation on expression is visible.

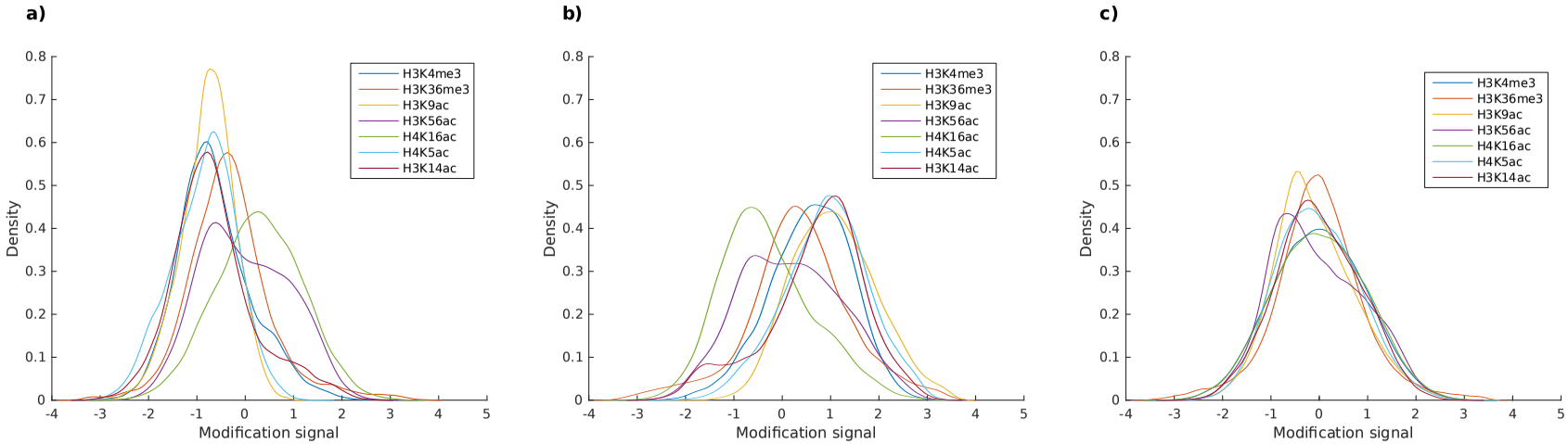

Figure 3.4: Estimated density functions of the histone modification signals for the *no impact*, *activating*, and *silencing* clusters. **a)** The density of modification signals for clusters for which a silencing impact on expression could be detected. Most of the means are below zero, only H4K16ac is centered at zero. Throughout all evaluations in this thesis, this modification proves to be the one having the least impact on expression. **b)** The density of all modifications for the clusters which are associated with an activating impact on expression. As well as in figure **a**, the means are not centered at zero and mostly positive. Interestingly, H4K16ac and H3K56ac are shifted towards the negative axis. **c)** The estimated density of modification signals for the clusters where no impact on expression could be detected. It can be observed that the mean value for most of the modifications is close to zero.

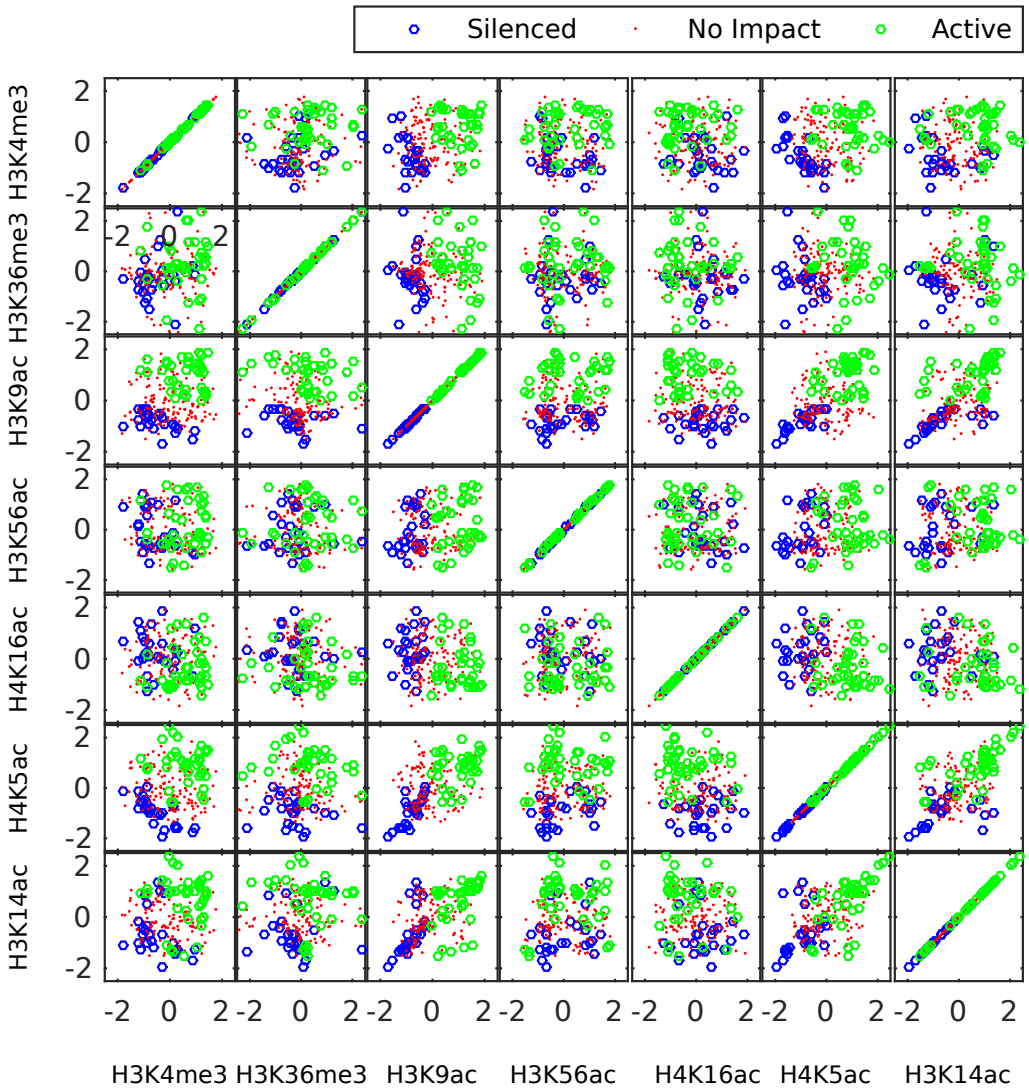

Figure 3.5:

Cluster centers of the k-means clustering shown for each modification dimension tuple. The green and blue circles show the same data as in figure 3.3, and the red dots show the cluster centers where no clear impact on expression could be deducted. In cases where the covariance for the *activating* and *silencing* is high, the *unknown impact* group shows no obvious covariance (e.g. H3K14ac and H4K4ac).

For some modifications, an obvious impact on expression becomes clearly visible, like for H3K9ac and H4K5ac, where a linear separation at a specific threshold is possible to distinguish between activated and silenced clusters. For others, like H4K16ac, no obvious correlation with the expression can be deducted from the visualization.

Table 3.2: The modification cluster centers for which low gene expression was observed. The size of each cluster as well as the normalized expression mean is shown.

H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac	Cluster Size	Normalized Mean
0.194	-2.117	-1.309	-0.65	0.309	-1.604	-1.334	287	-1.891
-0.272	0.049	-1.681	-0.659	0.59	-1.934	-1.939	316	-1.518
-1.2	-0.41	-0.785	0.1	-0.096	-0.805	-1.018	725	-1.214
0.265	2.385	-1.096	-0.446	0.254	-1.568	-1.232	309	-1.169
-0.202	-0.552	-0.914	-0.268	-0.111	-1.661	0.908	332	-1.153
0.913	0.183	-1.542	-0.889	-0.703	-1.733	-1.697	370	-1.063
-0.697	-0.765	-0.743	1.098	1.316	-0.804	-0.974	541	-1.036
-1.064	-0.488	-0.418	1.434	0.038	-0.585	-0.65	636	-1.002
-1.812	-0.264	-1.026	0.877	0.713	-0.12	-1.127	393	-0.967
0.981	-0.125	-1.228	-0.951	1.155	-1.569	-1.263	379	-0.904
-1.145	-0.043	-1.013	-0.475	1.134	-0.593	-1.239	504	-0.887
-0.837	-0.192	-0.648	1.029	-1.231	-0.628	-1.048	567	-0.829
-0.479	0.973	-0.631	1.034	0.107	-0.964	1.349	271	-0.812
0.185	-0.38	-0.976	1.205	-0.047	-1.514	-1.428	321	-0.804
-0.823	-1.481	-0.614	-0.689	0.092	-0.904	-0.657	513	-0.797
-0.663	-0.479	-1.06	-0.719	-1.043	-1.454	-1.175	482	-0.745
-1.076	-0.84	-0.38	-0.921	0.888	-0.26	-0.433	699	-0.654
-0.848	-1.078	-0.321	0.526	0.258	0.021	-0.332	614	-0.636
-0.89	-0.137	-0.59	0.862	-0.674	-0.845	0.515	389	-0.617
-0.316	-0.29	-0.665	-0.697	1.832	-0.952	-0.707	463	-0.615
0.143	-0.096	-0.963	-0.099	-0.462	-0.719	-1.023	759	-0.611
-1.189	-0.745	-0.33	0.233	1.407	-0.087	-0.2	594	-0.597
-0.977	-1.255	-0.366	-0.539	0.201	-0.467	0.869	404	-0.555
-0.38	1.244	-0.547	-0.566	-0.015	-1.12	1.044	388	-0.546
-0.788	-0.298	-0.924	-1.295	0.587	-1.024	-0.707	590	-0.54
-0.919	0.491	-0.587	0.867	0.656	-0.813	-0.525	562	-0.511

Table 3.3: The modification cluster centers for which high gene expression was observed. The size of each cluster as well as the normalized expression mean is shown.

H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac	Cluster Size	Normalized Mean
-0.039	-0.122	2.768	-0.414	-1.18	2.484	2.347	316	2.774
0.257	-0.393	2.605	1.598	-0.826	2.009	2.056	242	2.093
0.101	0.1	2.607	-0.181	0.391	2.21	2.119	307	2.048
1.469	0.975	1.857	-0.315	-0.988	1.54	1.57	318	1.843
1.119	2.689	1.83	-0.331	-0.194	1.45	1.362	227	1.633
1.203	-1.036	1.763	-0.221	-1.136	1.668	1.476	269	1.546
0.692	2.024	1.543	1.197	-0.711	0.663	1.024	263	1.306
1.14	0.521	1.35	1.274	-0.906	1.087	1.252	388	1.286
1.066	0.214	1.2	1.705	0.24	0.887	-1.195	233	1.259
0.54	0.244	0.216	-1.397	-0.672	0.842	-1.526	298	1.177
0.075	0.786	1.394	-0.757	-1.074	1.208	1.037	403	0.991
0.622	0.091	1.168	1.61	1.321	1.023	1.011	323	0.955
-0.097	0.129	1.501	1.792	-1.13	1.068	-1.084	250	0.877
1.364	1.174	0.169	0.505	-0.327	0.444	0.84	431	0.859
0.479	0.145	1.694	0.43	0.209	1.051	1.075	483	0.844
0.969	1.17	1.419	-0.596	0.638	1.123	1	325	0.842
0.676	-1.277	1.706	1.67	-0.305	0.68	1.358	258	0.801
1.341	0.532	0.352	-0.601	-0.655	1.483	0.998	471	0.787
0.6	2.042	0.208	-0.728	-0.89	0.89	0.903	329	0.779
1.404	0.282	0.267	0.583	0.664	1.422	0.943	389	0.755
1.291	0.598	0.344	1.332	1.578	-0.058	0.015	296	0.725
-0.251	-0.56	1.148	-0.59	-0.806	1.906	1.412	360	0.721
0.162	0.145	0.744	-1.318	-1.068	-0.533	-1.369	387	0.675
0.673	1.158	1.475	-0.341	-0.813	0.245	-0.072	397	0.652
1.132	-2.287	1.09	-0.132	-0.801	0.442	0.391	296	0.652
1.142	-1.139	1.481	-0.135	0.861	1.185	1.218	339	0.642
1.191	0.732	-0.03	0.966	-1.457	0.555	-0.299	388	0.634
1.302	0.231	1.121	1.051	-0.192	-0.456	0.687	397	0.634
-0.916	1.784	1.074	-0.357	0.169	0.838	0.917	282	0.632
0.629	-0.0	0.397	1.617	-0.378	1.008	0.23	416	0.599
0.252	0.259	0.629	-0.495	1.339	1.266	0.367	394	0.592
-0.136	-1.964	1.482	-0.593	-0.191	0.778	1.366	269	0.591
0.21	0.328	0.107	0.489	-0.665	1.431	1.041	532	0.585
1.241	2.366	0.542	0.774	-0.706	-0.342	-0.064	244	0.579
-0.864	-0.29	1.651	-0.151	-1.135	0.735	1.116	331	0.547
1.168	-0.259	1.235	-0.883	0.012	0.747	0.536	406	0.544

Furthermore, it can be observed that H3K14ac, H3K9ac and H4K5ac exhibit a strong covariance and therefore a linear dependency for the *activating* and *silencing* clusters (green, blue), but apparently none for the clusters where *no impact* on expression could be detected (red).

The distribution of clusters with *no impact* on expression shows that a modification level above average of H3K9ac or H4K5ac alone does not automatically result in high transcriptional activity.

This leads to the conclusion that specific patterns of different modifications can be associated with activation or silencing, but no single modification is responsible for activating or silencing a gene.

As mentioned before, clusters which have an impact on expression show a higher corr for H3K14ac-H3K9ac, H3k14ac-H4K5ac and H4K5ac-H3K9ac as the ones which do not have an impact (figure 3.5). To validate this hypothesis, the correlation was calculated for genes in the clusters of the *no impact* group (table 3.4 and figure 3.6) as well as for the clusters where activation or silencing was observed (table 3.5 and figure 3.7).

Only weak correlation between the modifications exist for the clusters from the *no impact* group, whereas the correlation of some modifications for clusters which were linked with activation or silencing is much higher; H4K5ac, H3K9ac and H3K14ac show very strong correlation, whereas H4K16ac exhibits a negative correlation with most of the other modifications, especially with H4K5ac, H3K9ac and H3K14ac.

Table 3.4: Correlation of modification signals for the genes assigned to a cluster which is not associated with an impact on transcription. Red signifies a negative correlation, yellow a correlation close to zero and green marks a highly positive correlation (a color gradient is used). Only weak correlation exists; its range is between -0.14462 to 0.26143.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3		0.06	-0.021	-0.088	0.033	-0.145	-0.023
H3K36me3	0.06		-0.092	-0.039	-0.028	-0.078	-0.092
H3K9ac	-0.021	-0.092		0.052	-0.013	-0.006	0.261
H3K56ac	-0.088	-0.039	0.052		0.093	0.052	-0.04
H4K16ac	0.033	-0.028	-0.013	0.093		-0.006	0.092
H4K5ac	-0.145	-0.078	-0.006	0.052	-0.006		0.19
H3K14ac	-0.023	-0.092	0.261	-0.04	0.092	0.19	

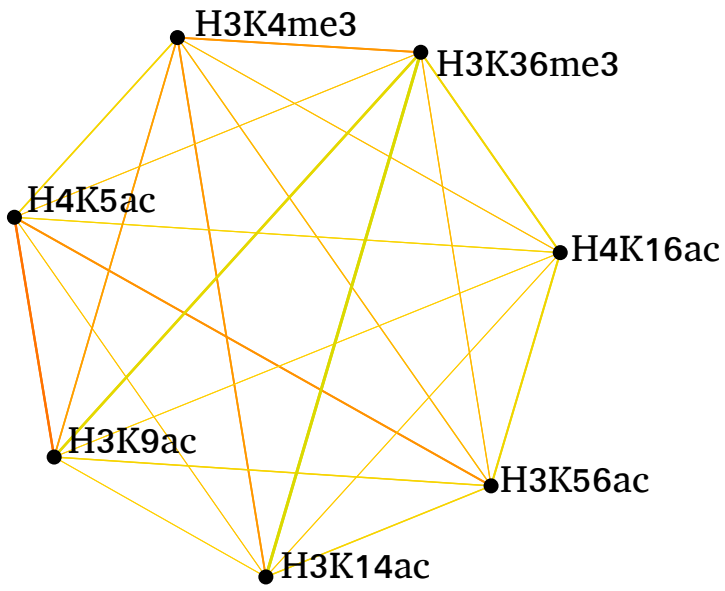

Figure 3.6: Correlation between modifications for genes assigned to clusters which are not associated with an impact on transcription. It uses the values from table 3.4 as adjacency matrix. The thicker the edge, the higher the absolute value of the correlation. Red signifies a negative correlation, yellow a correlation close to zero and green marks a highly positive correlation.

Table 3.5: Correlation of modification signals for the genes assigned to a cluster which is associated with activation or silencing. Again, red signifies a negative correlation, yellow a correlation close to zero and green marks a highly positive correlation. Compared to the correlation in figure 3.4, much stronger correlation between the modifications exist, ranging from -0.24555 to 0.7095. This suggests a high codependency between the different modifications in relation to their impact on gene expression.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3		0.266	0.372	0.025	-0.182	0.325	0.278
H3K36me3	0.266		0.159	0.055	-0.155	0.145	0.16
H3K9ac	0.372	0.159		0.118	-0.246	0.71	0.635
H3K56ac	0.025	0.055	0.118		-0.073	0.125	0.12
H4K16ac	-0.182	-0.155	-0.246	-0.073		-0.213	-0.198
H4K5ac	0.325	0.145	0.71	0.125	-0.213		0.62
H3K14ac	0.278	0.16	0.635	0.12	-0.198	0.62	

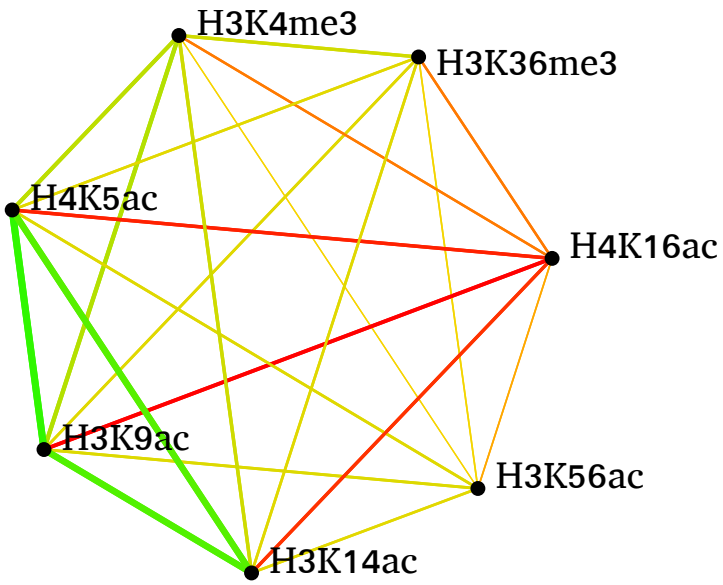

Figure 3.7: Graphical visualization of the correlation between modifications for genes assigned to a cluster which is not associated with activation or silencing It uses the values from table 3.5 as adjacency matrix. The thicker the edge, the higher the absolute value of the correlation. The color scheme is the same as in table 3.5.

3.2 Quantification of the impact of histone modifications on expression

3.2.1 ARACNE

To further elaborate the codependency between different histone modifications and their impact on expression, ARACNE was used to calculate the mutual information between the different modifications.

It was previously shown in this thesis that a higher linear relationship between the different histone modifications exists for clusters of histone profiles (table 3.5), which exhibit an impact on transcription, compared to those clusters which show no impact (see section 2.2.2). This relationship was further investigated with ARACNE.

ARACNE uses the data processing inequality to remove the edge with the least mutual information between each gene triplet loop. This is useful when dealing with gene expression, where one gene might indirectly be influencing another gene, but is not meaningful when the algorithm is presented with different modification signals instead of genes and genes instead of different time samples.

Therefore, the data processing inequality postprocessing was turned off, thus giving a graph where each node is connected with every other node. Nonlinear relationships between some of the modifications, which were not detected by calculating the correlation, would have been revealed by ARACNE, but the modifications with high mutual information also had a high correlation and vice versa.

The strong linear dependency for modifications, which have an impact on gene expression (table 3.5), can also be seen in the mutual information results of ARACNE (tables 3.6, 3.7 and figures 3.8, 3.9).

Table 3.6: Mutual information as calculated with ARACNE for the genes assigned to clusters which are not associated with an impact on transcription. Red signifies a mutual information close to 0 and green represents a high mutual information. It can be seen that no strong mutual information exists.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3		0.013	0.01	0.011	0.008	0.019	0.008
H3K36me3	0.013		0.024	0.014	0.01	0.023	0.021
H3K9ac	0.01	0.024		0.014	0.007	0.011	0.054
H3K56ac	0.011	0.014	0.014		0.013	0.012	0.016
H4K16ac	0.008	0.01	0.007	0.013		0.007	0.017
H4K5ac	0.019	0.023	0.011	0.012	0.007		0.035
H3K14ac	0.008	0.021	0.054	0.016	0.017	0.035	

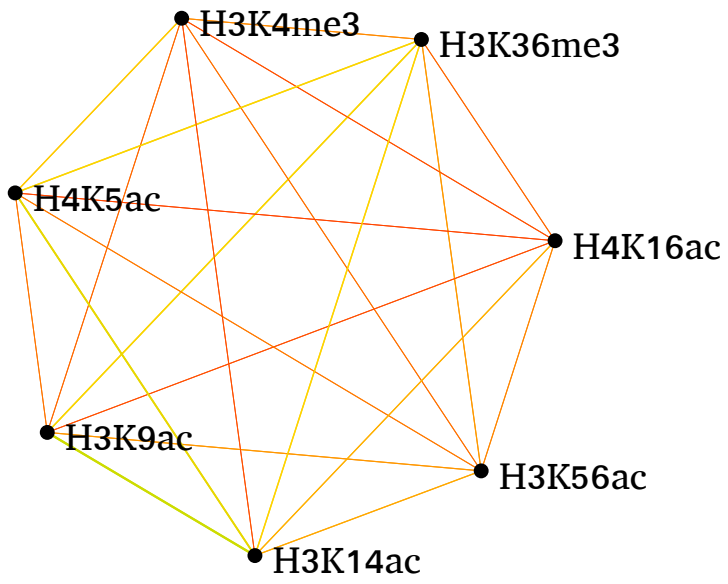

Figure 3.8: Visualization of the mutual information between modifications for genes assigned to clusters which are not associated with an impact on transcription. It uses the values from table 3.6 as adjacency matrix. The thicker the edge, the higher the mutual information shared between the two modifications. The color scheme is the same as in table 3.6.

Table 3.7: Mutual information as calculated with ARACNE for the genes assigned to clusters which are associated with an impact on transcription. Red signifies a mutual information close to 0 and green represents a high mutual information. It can be seen that much more mutual information exists than in table 3.6. The modification pairs with the strongest mutual information are also the same as the ones with the strongest correlation.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3		0.081	0.145	0.019	0.036	0.139	0.093
H3K36me3	0.081		0.072	0.027	0.038	0.061	0.075
H3K9ac	0.145	0.072		0.036	0.048	0.384	0.317
H3K56ac	0.019	0.027	0.036		0.014	0.04	0.038
H4K16ac	0.036	0.038	0.048	0.014		0.039	0.045
H4K5ac	0.139	0.061	0.384	0.04	0.039		0.301
H3K14ac	0.093	0.075	0.317	0.038	0.045	0.301	

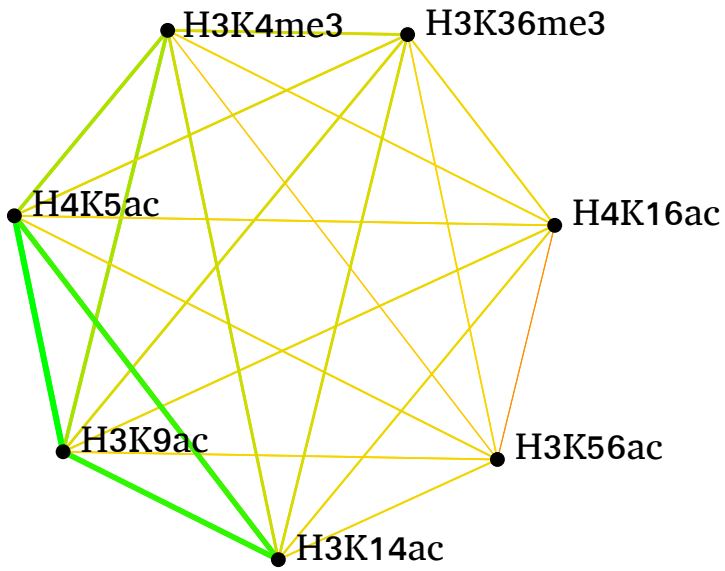

Figure 3.9: This graph visualizes the mutual information between modifications for genes assigned to clusters which are associated with activation or silencing. It uses the values from table 3.7 as adjacency matrix. The thicker the edge, the higher the absolute value of the correlation. The color scheme is the same as in table 3.7.

3.2.2 Decision trees

Interpretation of how the different histone modifications impact gene expression becomes feasible, when decision trees are used.

Table 3.8: Overview of the trees which were trained on the modification patterns in order to obtain well interpretable results on how the modification patterns impact gene expression. Different data subsets were used, either depending on the associated standardized gene expression e of a modification pattern or on the association of the modification pattern with a k-means cluster of the groups *activating*, *silencing* or *no impact*.
The prediction targets of the different trees is a binary classification of either a binarized expression value or a cluster group.

Subset	#Samples	Prediction	Accuracy	Figure
Full dataset	83280	$e \leq 0$ or $e \geq 0$	66.52%	3.10
$e \leq$ -0.5 or $e \geq 0.5$	52421	$e \leq$ -0.5 or $e \geq 0.5$	72.35%	3.11, 3.12
Full dataset	83280	$-0.5 \leq e \leq 0.5$	63.48%	3.13
Patterns from cluster *activating* or *silencing*	19136	Cluster group	98.24%	3.14
Patterns from cluster *activating* or *silencing*	25688	$e \leq 0$ or $e \geq 0$	78.72%	3.15
Patterns from cluster *activating* or *silencing* where $e \leq$ -0.5 or $e \geq 0.5$	18574	$e \leq$ -0.5 or $e \geq 0.5$	85.35%	3.16

A decision tree was trained with all 83280 samples, using class label 1 for modification patterns associated with a standardized expression greater than zero and class label -1 for the rest (figure 3.10). The tree was pruned to the optimal height using the cross validation data, as shown in figure A1. For training and pruning 30% of all available data was used and the tree was pruned to a height of 6.
Its classification accuracy is low, as it only classifies 66.52% of the validation data

correctly.

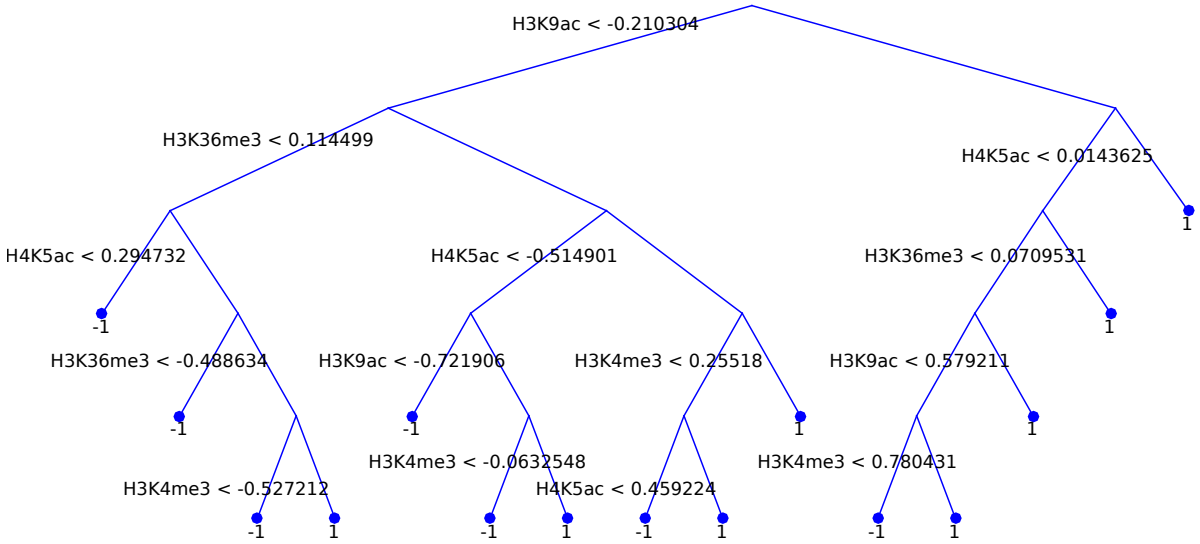

Figure 3.10: A tree which predicts whether gene expression is greater than zero or less, based on the histone modification profile. All modification patterns were used; patterns which were associated with standardized expression greater than one are labeled with 1 and the rest with -1. For the training of this tree 30% of available data was used. Classification accuracy on the validation data is low at 66.52%

To avoid misclassification of modification patterns with expression close to zero, an additional tree was trained, using only patterns for which expression less than -0.5 (class label -1) and greater than 0.5 (class label 1) was observed, which gave a total of 52421 samples (figures 3.11 and A2). The accuracy on the validation set (70% of the data) was 72.35%.

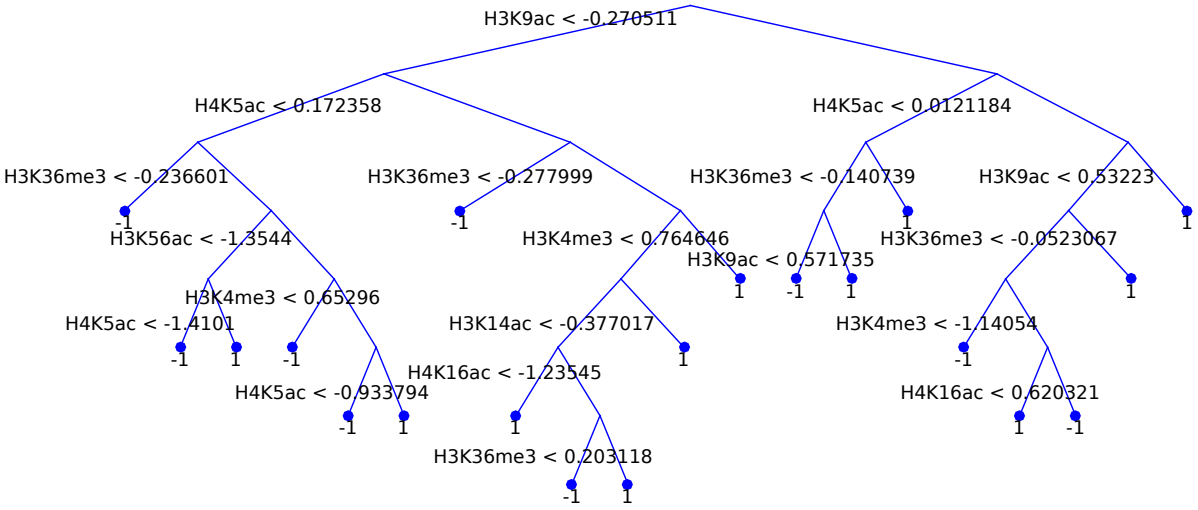

Figure 3.11: Tree predicting whether gene expression is greater or less than zero based on the histone modification profile. It was only trained and validated with histone profiles for which the expression value was greater than 0.5 or less than -0.5. For the training of this tree 30% of available data was used. Classification accuracy on the validation data is 72.35%.

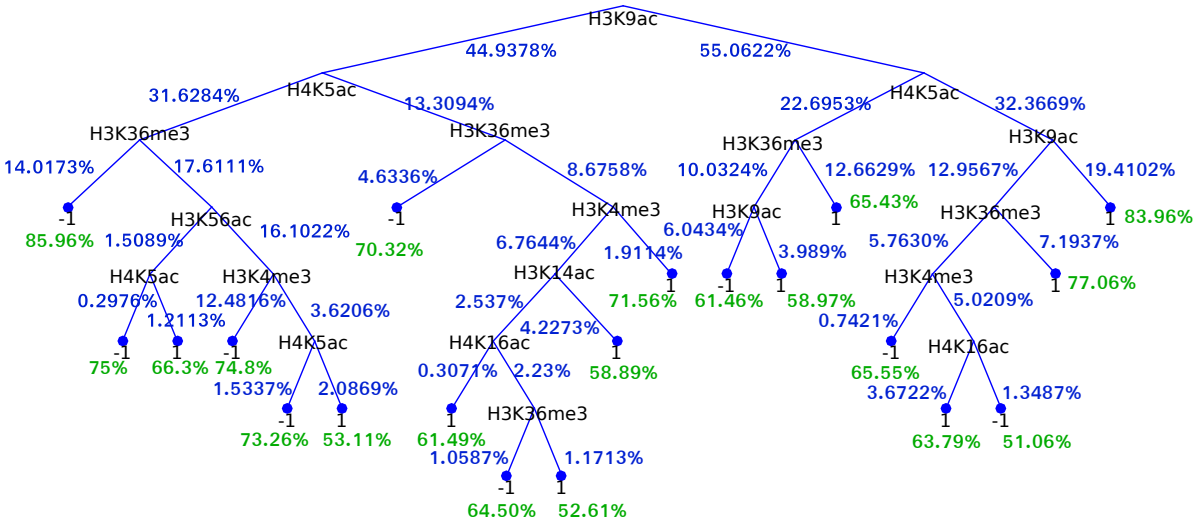

Figure 3.12: The same tree as in figure 3.11 is shown, but instead of the conditions, the data flow (blue percentage values) and the classification accuracy is shown (green).

Both trees (figure 3.10 and 3.11) have H3K9ac as root node with a decision value between -0.2 and -0.1, but differ already on the second level. The tree trained with the full dataset tends to classify a pattern as activating based on very few decisions, if H3K9ac is greater than -0.21, but is only able to achieve 66.52% classification accuracy

on its validation data.

It can also be noticed how H3K36me3 plays a role in gene regulation (figure 3.12). This modification only shows a slight change in mutual information between clusters from the group *no impact* and the rest (see figures 3.9 and 3.8), but can be found on a high level of the tree. This indicates that H3K36me3 and the modifications with high mutual information (H3K9ac, H3K14ac and H4K5ac) together impact gene regulation, but are independent from each other.

19.4% of the whole data were classified as upregulated, only based on the conditions that H3K9ac $>$ 0.5322 and H4K5ac $>$ 0.0121184 with an accuracy of 83.96%, whereas 14% where classified as downreguated by the condition that H3K9ac $<$ 0.5322 and H4K5ac $<$ 0.172358 and H3K36me3 $<$ -0.2366 with an accuracy of 85.96%.

It is also worth noting that 67.95% of the data were only classified based on H3K9ac, H4K5ac and H3K36me3 with an average accuracy of 75.79%.

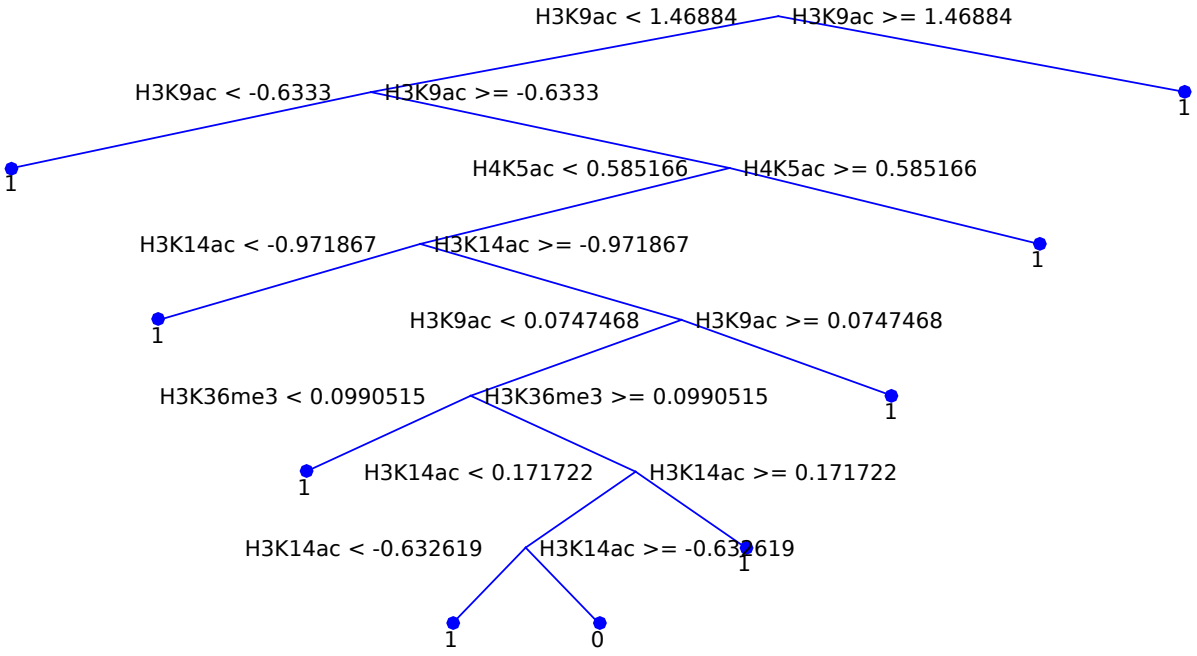

Figure 3.13: This tree was trained to predict whether a profile has an impact on gene expression or not. The classification performance was low at 66.52%, but the structure shows an interesting pattern of looking only for extremes of 4 different modifications, but at especially at H3K9ac, H4K5ac and H3K14ac, which are the ones that exhibit the greatest covarinace and mutual information (see figures 3.5 and 3.7).

The classification accuracy of a tree to predict whether a cluster was labeled as activating or silencing achieved 98.08% correct classifications (figure 3.14). The tree exhibits a somewhat more complex structure than the previous ones, but it can be seen as well that H3K9ac $>$ 0.5743 alone is sufficient to classify a pattern as belonging to an activating cluster. The shortest path to classify a pattern as silencing consists of H3K9ac $<$ -0.0275, H4K5ac $<$ 0.557, H3K4me3 $<$ -0.546 and interestingly H4K16ac $>$ -0.1687. The first three modifications have been shown to have the most impact on expression and also show the highest codependency but for H4K16ac only a slight negative impact has been observed so far.

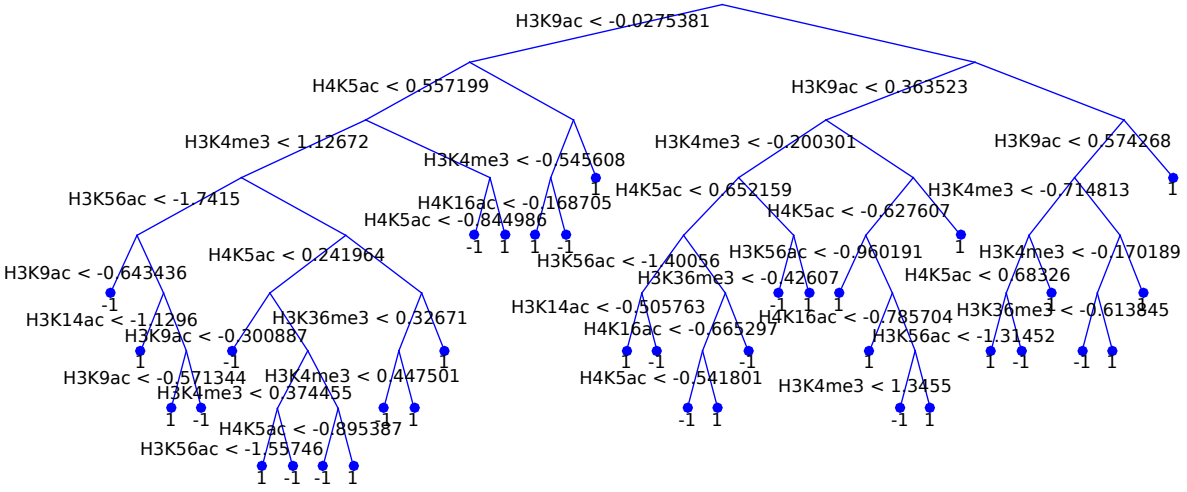

Figure 3.14: Tree trained to distinguish between patterns belonging to an activating or silencing cluster from the k-means clustering. It yields a classification performance of 98.08% on the validation data.

The interpretation of the tree reveals some interesting relations between the modifications, but has little value for classification of the patterns, as it does not predict if a gene is activated or silenced by a modification pattern, but only if that gene was assigned to a cluster where most of the genes where activated or silenced.

Nevertheless, the approach of only using patterns from clusters in those groups was promising, as they are apparently much better classifiable.

Therefore, only the data from the clusters *activating* and *silencing* was taken for the next experiment. The actual expression levels of the profiles were passed through a signum function, so that two expression classes (*-1* and *1*) were obtained. The decision tree was

then trained to predict those class labels based on the modification pattern.

The resulting tree (figure 3.15) shows a very simple structure and achieves a classification performance of 78.72% on the validation data, which was 70% of the dataset with a size of 25688. Again, the modifications with the highest mutual information, i.e. H3K9ac and H4K5ac, but not H3K14ac, are mainly used for the classification. The only other modification in the tree is H3K56ac, which is only present on the last level and shows a negative impact on expression (H3K56ac < -0.757434 signifies higher expression).

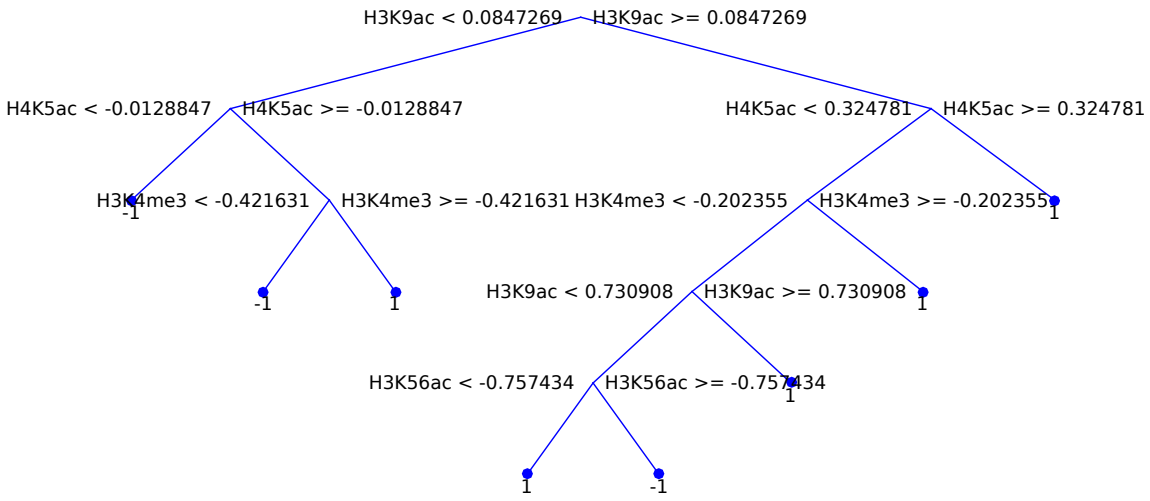

Figure 3.15: For training and evaluating this tree, only the patterns from clusters labeled with *activating* and *silencing* were used. In contrast to the tree in figure 3.14 the actual expression values were binarized (label -1 and 1) at a threshold of zero and used as target values.

The classification accuracy can be further improved by filtering the dataset which was used for the tree in figure 3.15, so that only patterns which show an expression level below -0.5 or above 0.5, are used for the classification. This dataset contained 18574 values and resulted in a tree with a very simple structure (figure 3.16). A classification performance on the validation data of 85.35% was achieved.

This tree only contains three different modifications, namely H3K9ac, H4K5ac and H3K36me3, and like the tree in figure 3.15, classification is mainly based on H3K9ac

and H4K5ac and does not include H3K14ac, which also shares high mutual information with H3K9ac and H4K5ac.

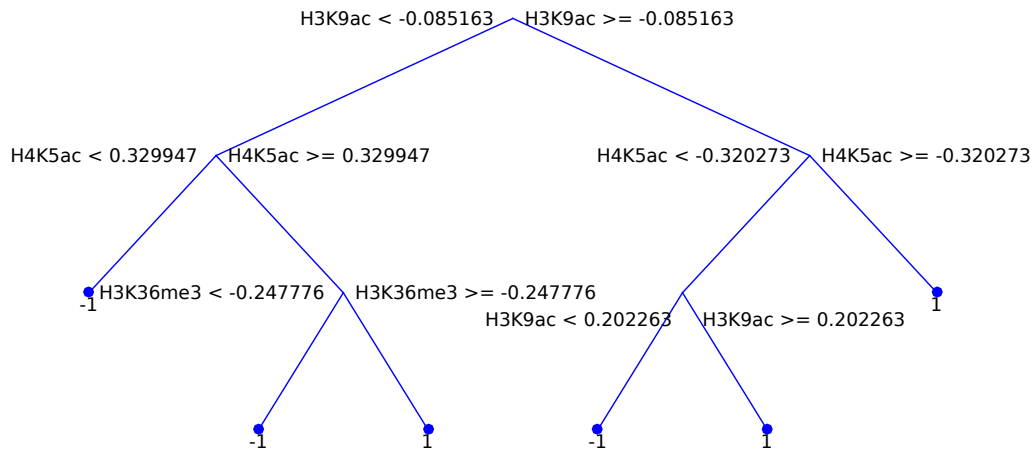

Figure 3.16: The tree was trained to predict the associated expression for a pattern. Only patterns contained in the cluster groups *silencing* and *activating* which showed an expression level below -0.5 or above 0.5 were used to train and evaluate the tree.

The comparison of the decision trees shows that H3K9ac has the strongest impact on expression, as this modification is in the root of every trained tree.
Also, H4K5ac, H3K14ac and H3K36me3, exhibit a significant positive impact on expression, whereas H4K16ac shows a slight negative impact.

It could be shown that by selecting only the patterns for which H4K5ac, H3K14ac and H3K9ac have the same sign as H3K36me3 (19396 samples) and removing those for which the corresponding transcription level is between -0.15 and 0.15 to prevent noise effects, only the signs of either H4K5ac, H3K14ac or H3K9ac allow prediction with an accuracy of 80.25%. If patterns with expression between -0.5 and 0.5 are also removed, the accuracy increases to 84.59%.
These results implicate that H4K5ac, H3K14ac and H3K9ac have the same sign.
Therefore a check was made by only selecting patterns where H4K5ac, H3K14ac and H3K9ac have the same sign. A tree was trained to quantify the classifiability of the data.

A performance of 74.50% was achieved.
As another control experiment, all those patterns where H4K5ac, H3K14ac and H3K9ac have the opposite signs as H3K36me3 (for which samples with expression between -0.15 and 0.15 were removed again - giving 15839 patterns), a tree was trained to predict whether expression is below or above zero. This tree performed significantly worse, classifying only 68.81% correctly.

3.2.3 Analysis of false positives

In the previous section, prediction of gene expression based on histone modification profiles was evaluated. The analysis of patterns, where the prediction contradicts the actual gene expression, reveals further insights.
The full dataset was queried for patterns where H3K9ac gives a modification signal greater than 1 but with an expression level less then -1. Using this criterion, only 744 patterns were found. For those patterns, the mean of each modification dimension was calculated and divided by the standard deviation of the modification values. The result which was obtained that way for H3K14ac was positive with a value of 0.61; H4K5ac was close to zero with 0.064, and H3K36me3 was the only modification which showed a strong impact on expression in this thesis that gave a negative mean modification of -0.21.
The same test was repeated for the inverted scenario, i.e. a modification signal less than -1 and an expression value greater 1. Here 1037 patterns were found. H3K14ac again moved in the same direction as H3K9ac, giving a mean of -0.24 and H3K36me3 again moved in the opposite direction, resulting in a mean of 0.25. H4K5ac has a mean value of 0.22.

These two tests were rerun for H4K5ac (yielding 936 for the first test and 1002 for the second), H3K14ac (yielding 1511 for the first and 1556 for the second) and H3K36me3 (yielding 1159 for the first and 1099 for the second).
In each of those tests it could be observed that H3K36me3 moves in the opposite direction as the modification under test (tables 3.9 and 3.10)

Table 3.9: For each modification all patterns were selected, where the modification under test has a signal strength less than -1. The resulting patterns were then filtered for corresponding expression values greater than 1. For modifications associated with silencing, this gives a set of patterns, in which prediction of expression based on that modification is incorrect.
For each modification dimension in the resulting dataset the *z-score* was calculated as described in section 2.2.2. Each row in the table contains the mean values of the patterns for the specific modification under test.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3	-4.244	0.048	0.209	0.061	-0.136	0.352	0.216
H3K36me3	0.332	-2.909	1.063	0.007	-0.218	0.559	0.676
H3K9ac	0.001	0.25	-4.056	-0.275	0.054	0.217	-0.241
H3K56ac	0.28	0.208	0.182	-4.226	-0.251	0.265	-0.165
H4K16ac	0.346	0.2	0.684	-0.063	-3.947	0.641	0.339
H4K5ac	0.286	0.172	0.043	-0.166	-0.034	-4.451	-0.44
H3K14ac	0.271	0.336	0.015	-0.24	-0.248	-0.019	-4.296

Table 3.10: In analogy to table 3.9 all the patterns for each modification were selected, where the modification under test has a signal strength greater than 1. The resulting patterns were then filtered for corresponding expression values less than -1. For modifications associated with activation, this also gives a set of patterns, in which prediction of expression based on that modification is incorrect.
For each modification dimension in the resulting dataset the *z-score* was calculated. Each row in the table contains the mean values of the patterns for the specific modification under test.

	H3K4me3	H3K36me3	H3K9ac	H3K56ac	H4K16ac	H4K5ac	H3K14ac
H3K4me3	3.79	-0.095	-0.479	-0.355	0.134	-0.582	-0.225
H3K36me3	-0.024	2.702	-0.664	-0.209	0.071	-0.828	-0.297
H3K9ac	-0.057	-0.21	3.603	0.07	-0.19	0.064	0.615
H3K56ac	-0.631	-0.254	-0.521	4.413	0.127	-0.486	-0.375
H4K16ac	-0.202	-0.291	-0.79	-0.061	3.835	-0.534	-0.288
H4K5ac	-0.281	-0.265	-0.146	-0.02	0.084	3.915	0.383
H3K14ac	-0.089	-0.189	-0.013	-0.011	0.168	-0.098	4.43

3.3 GO enrichment analysis

It could be shown that the histone modification clusters which exhibit an impact on gene expression can also be linked to biological processes and molecular functions, i.e. that histone modifications do not only regulate specific genes but also specific functions and processes within the cell.

For the two k-means histone profile clusters with the greatest average gene expression a GO term enrichment analysis was performed with GORILLA, i.e. for *biological process* (figures 3.17, A7), *molecular function* (figures 3.18, A8) and *cellular component* (figures 3.19, A9). A biological process is a recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end.

For the cluster with the most activating impact on gene expression a relation to RNA, rRNA, and ncRNA processing was observed (figures 3.17, 3.18, 3.19).

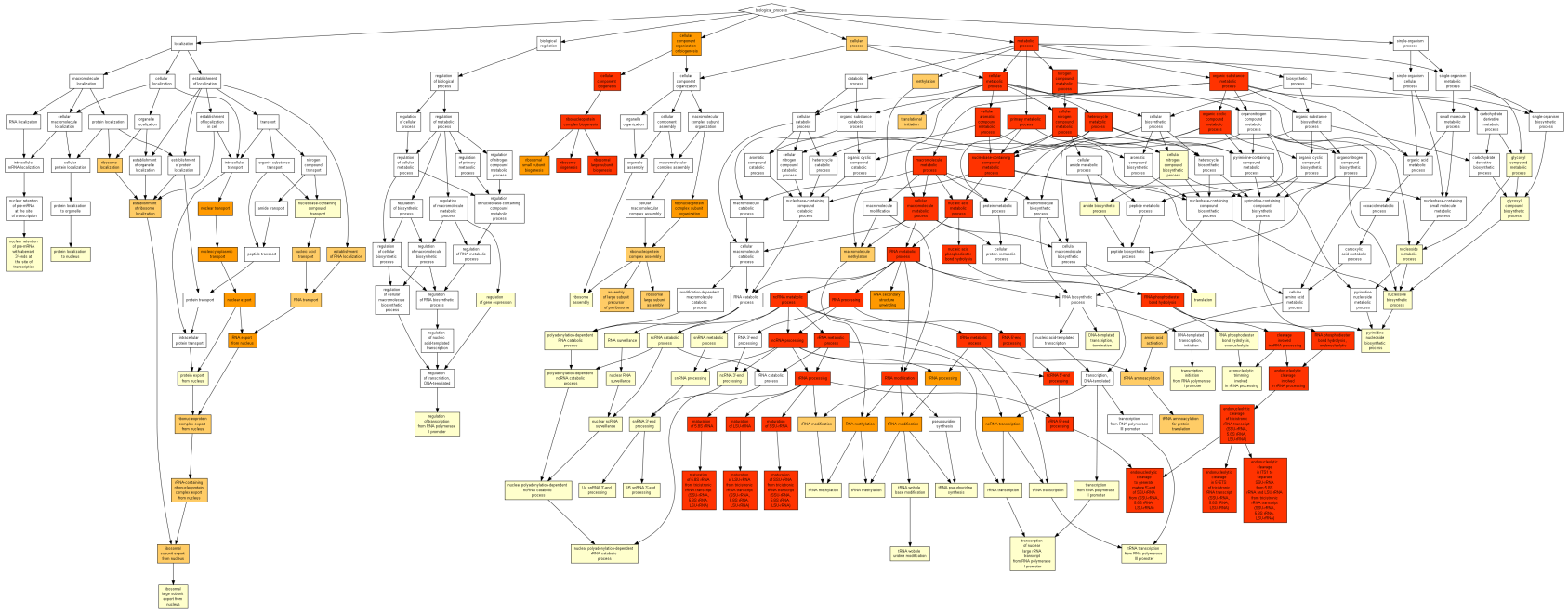

Figure 3.17: GO term analysis for *biological process* for the cluster with the greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A1.

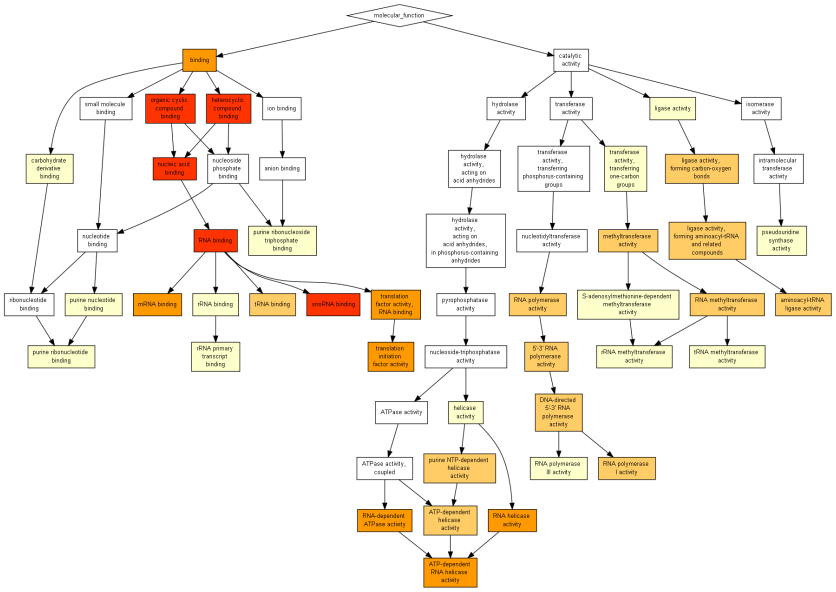

Figure 3.18: GO term analysis for *molecular functions* for the cluster with the greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A2.

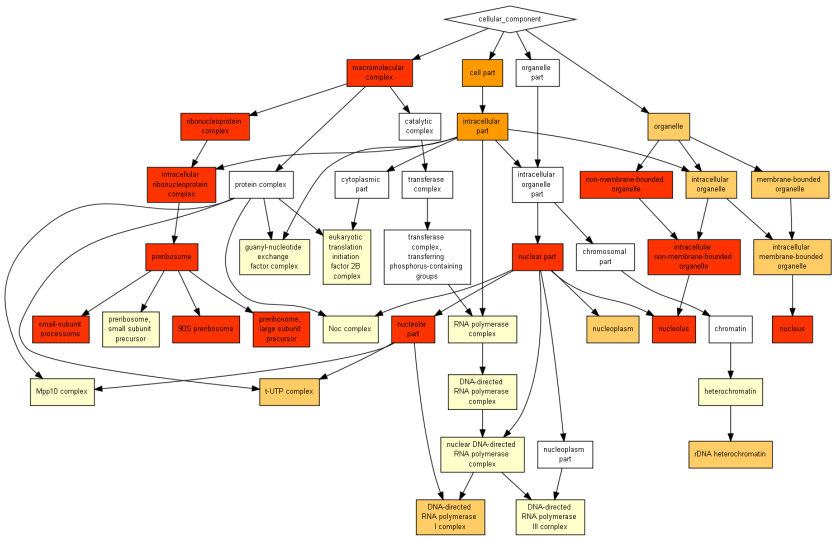

Figure 3.19: GO term analysis for *cellular component* for the cluster with the greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A3.

4 Discussion

In this thesis RNA-seq data and ChIP-seq data for *Saccharomyces cerevisiae* were analyzed. The RNA-seq data spanned 16 points in time, and provided transcript enrichment signals for 5205 genes after mapping. The ChIP-seq data was used to obtain modification signals for 7 different histone modifications for the same 16 points in time as the RNA-seq data and the same 5205 genes. Both datasets are publicly available on the NCBI GEO database [32].
Clustering of the histone modification patterns was performed, and a relation between histone modifications and their impact on gene expression was analyzed.
The adaption of a genetic regulatory network reconstruction algorithm was evaluated, and different approaches for integrative analysis were performed.
For a set of histone modifications a strong impact on gene expression could be shown as well as that rather combinations of modifications are responsible for regulation of gene expression than a single modification, i.e. that H3K9ac, H4K5ac and H3K14ac act together in gene activation while H3K36me3 regulates the impact which H3K9ac, H4K5ac and H3K14ac have on gene regulation.

4.1 Clustering

For clustering of the histone modifications the k-means algorithm was used because of its efficiency for calculating 200 cluster centers for the dataset which contained 83280 entries with 7 dimensions.
In this thesis different numbers of cluster centers were evaluated for most of the calculations and it was observed that, as long as the number of clusters was not very low (≤ 20), all further calculations showed very similar results.
This can be explained by the fact that each cluster was labeled as *activating*, *silencing*

or *no impact* and that those labels were then used to categorize the 83280 modification profiles into three groups. So even for a very small number of clusters most of the profiles still got the same labels.

4.2 Quantification of the impact of histone modifications on expression

4.2.1 ARACNE

When a dataset with linear, monotonic relationships is to be analyzed, classical correlation functions like Pearson correlation are a suitable choice. In the case of non-linear relationships, methods which are based on mutual information are better suited to detect interrelations.

ARACNE [57] is an algorithm which was developed for reconstructing gene regulatory networks from microarray expression profiles. The algorithm is based on mutual information and uses an information theoretic approach for eliminating most of the indirect interactions inferred by co-expression methods. ARACNE has proven to be a powerful tool for reverse engineering of regulatory networks from microarray experiments for the human organism [57].

Due to the fact that ARACNE's way of detecting related genes is based on calculating the pairwise mutual information, it can be used to infer relations in other datasets as well. Hereby, an important caveat has to be considered, namely that ARACNE is pruning the graph using the data processing inequality and therefore no loops between gene triplets are allowed unless a tolerance for pruning is set. This may lead to false negatives and it depends on the underlying data whether such triplet loops are valid or not. Furthermore, ARACNE only investigates pairwise interactions and is therefore not able to find any interactions that are only expressed indirectly between two genes.

An interaction between two genes g_1 and g_2 as detected by ARACNE from expression data can be interpreted as "If g_1 exhibits a high expression then it is probable that the expression of g_2 is also impacted".

When histone modification signals for a specific modification at the TSS are used

instead of expression values, high mutual information between two genes implies that a high modification of one gene impacts the modification of another gene.
Unfortunately, this kind of analysis could not be performed, as the number of samples per modification was limited to 16 samples per gene and ARACNE does not perform well with less than 100 samples [57].
Nevertheless, ARACNE proved to be a useful tool for epigenetic research. Instead of calculating the mutual information between tuples of genes, the algorithm was utilized to obtain information of the relationship between combinations of modifications. It was previously shown in this thesis that a higher linear relationship between the different histone modifications exists for clusters of histone profiles (table 3.5), which exhibit an impact on transcription, compared to those clusters which show no impact (see section 2.2.2). This relationship was further investigated with ARACNE.
As already mentioned, ARACNE uses the data processing inequality to remove the edge with the least mutual information between each gene triplet loop. This is useful when dealing with gene expression, where one gene might indirectly be influencing another gene, but is not meaningful when the algorithm is presented with different modification signals instead of genes and genes instead of different time samples.
Therefore, the data processing inequality postprocessing was turned off, thus giving a graph where each node is connected with every other node. Nonlinear relationships between some of the modifications, which were not detected by calculating the correlation, would have been revealed by ARACNE, but the modifications with high mutual information also had a high correlation and vice versa.
When the data is split into two groups, and one group consists of all histone modification patterns from the clusters labeled with *no impact*, and the other group contains all the remaining patters, an interesting fact can be observed. The patterns from the group *no impact* show very little mutual information, whereas the patterns which exhibit an impact on gene expression are highly correlated and therefore also share high mutual information.
This shows that not a single modification alone is responsible for gene regulation. A liaison of different modifications is involved in the regulatory process. In this thesis we could show that H3K9ac, H4K5ac and H3K14ac have a high covariance for modification

profiles with a non-average expression level and furthermore correlate with gene expression.

For H3K36me3 only a low mutual information/covariance with other modifications was observed, but it was shown that H3K36me3 also plays an important role in gene regulation, but is modified independently of H3K9ac, H4K5ac and H3K14ac.

Kuang et al. also studied the correlation of different modifications, but in contrast to this thesis, where the data was split into groups based on the modification profiles impact on gene expression, they ran their tests on subsets of the data based on the yeast metabolic cycles phases. They observed a high correlation for the distribution of H3K4me3, H3K36me3, H3K9ac, H3K14ac, and H4K5ac in the oxidative phase, whereas little correlation was observed in the reductive building phase.

According to their study, H3K9ac is a very active modification which primarily regulates oxidative growth genes. As the distinction between different sets of genes, which are related to different phases of the YMC, was not made in this thesis, it was not investigated how the modification of H3K9ac relates to the genes on which it acts as a regulatory mechanism.

4.2.2 Decision trees

Decision trees are predictive models which can be used not only for classification of observations but also for regression of continuous targets. They are often used in data mining for classification in multivariate systems for obtaining easily interpretable models. Common usages are variable selection and assessment of the relative importance of variables, prediction or how to subdivide heavily skewed variables into ranges [62].

In this thesis decision trees were used to quantify the impact of histone modifications on gene regulation.

The training of decision trees proved to be a useful tool to obtain models that can be interpreted. Here they were not used for predicton − the classification results on the validation data rather gave an indication of how trustworthy the model is.

The trees were mainly used for binary classification. Prefiltering the dataset and defining the target classes, to find out which configuration results in a model that

abstracts the data well, was useful for quantifying the importance of specific histone modifications for gene expression.

The decision tree derived from the full dataset achieved only a classification performance of 66.52 % (figure 3.10). It is not surprising that such a model has such a rather low performance, as many other factors influencing gene expression, like other epigenetic modifications or completely different regulatory mechanisms, are not covered in the training data. Therefore, a strategy for selection of the patterns, where the given set of histone modifications plays a major role in gene regulation, was necessary for increasing the model performance.
The prediction whether a pattern belongs to a cluster for which an impact on expression was observed worked almost perfectly; a fact that is not further surprising as the data space had already been segmented by the clustering algorithm. Nevertheless, this shows that this space could be further segmented based on expression signals – information which was not used during clustering. Furthermore, is was valuable to observe that a tree with the shortest path to a leave of 3 intermediate nodes and a maximum depth of 8 was sufficient to achieve more than 98% classification performance on the validation data.
More significance can be attributed to the analysis, where only data belonging to clusters of the groups *activating* and *silencing* was used to create the decision tree, but the target classes were determined by the actual expression values.
This model was still further improved by removing genes with an average expression value (between -0.5 and 0.5). This resulted in good prediction of gene expression based on the modification signals (85.35% validation performance).
This tree was also the one with the least height and decided only based on H3K9ac, H4K5ac, and H3K36me3, but not based on H3K14ac. This observation shows that, although correlation of H3K14ac with expression is high, H3K14ac does not serve as a good predictor for gene expression.
Although the tree which distinguishes between profiles that have no impact on expression and the rest (figure 3.13) yields a low classification accuracy of 63.48%, it shows an interesting structure. The modification signal of H3K9ac $\geq$ 1.46884 alone is

considered sufficient to classify the profile as having an impact. Based on what was observed in figure 3.3, it can be assumed that primarily activating profiles are evaluated with that branch. On the other hand H3K9ac $\leq$ -0.6333 is also a sufficient signal to classify the pattern as having an impact. This can be assumed to account for silencing profiles.

If the pattern falls within the H3K9ac range of -0.6333 $\leq$ H3K9ac $\geq$ 1.46884, H4K5ac is checked if it shows a greater signal than 0.585166. If that is the case, the pattern is again classified as having *an impact*. On the next level the modification signal of H3K14ac is reviewed for having a value less than 0.971867. On the following levels, each of those modifications is checked once more and H3K36me3 is the only other one used in that classification tree.

Interestingly, H3K9ac, H4K5ac and H3K14ac are exactly those modifications, which correlate the most with each other and share the highest mutual information and therefore contain interdependent information. When the structure of the tree is analyzed, it can be observed that each of the mentioned modification signals is only checked for extremes to decide whether the pattern has an impact on gene expression or not.

Throughout all decision tree setups, H3K9ac was always in the root of the trees. In the model, distinguishing patterns that are associated with an impact on expression from the rest, the first two decisions are solely based on H3K9ac. This also shows how closely this modification is linked to gene expression.

4.2.3 Analysis of false positives

The prediction of gene expression based on H3K9ac, H4K5ac and H3K14ac is viable to some extent, but obviously many other factors influence gene expression, thus reducing the accuracy of the prediction.

Investigating those factors was beyond the scope of this thesis. Nevertheless, it was important to analyze the modification patterns where the interpretation of the modification signals leads to false assumptions about expression.

To this end, the modification patterns where H3K9ac, H4K5ac and H3K14ac exhibit strong signals, but gene expression is low and vice versa, were extracted.

It was observed that H3K36me3 always moved in the opposite direction to the modification signal under test. Although no direct impact of H3K36me3 on expression could be shown, this behavior suggests that H3K36me3 plays a role in how H3K9ac, H4K5ac and H3K14ac impact gene expression.

4.3 GO enchrichment analysis

GO term enrichment analysis showed that the two most activating clusters can be significantly related to specific biological processes, molecular functions and cellular components. This reveals that sets of genes which work together to accomplish processes within a cell are regulated by histone modifications. As the GO enrichment analysis was not the main focus of this thesis, it was only performed for the two clusters for which the strongest impact on gene expression was observed. In a more exhaustive analysis, a broader set of modification pattern clusters could be tested.

4.4 Conclusion

In this thesis RNA-seq data and ChIP-seq was analyzed to investigate how histone modifications affect gene expression.

The modification data was clustered using k-means and 39 clusters were identified which can be associated with increased transcription and 26 clusters were related to reduced transcription.

The mode of function of ARACNE was studied in detail and its use for modification profiling was discussed. A promising application of the algorithm, i.e. creating a network of how the modification level for a specific modification for each gene affects other genes, could not be executed, as an insufficient number of time samples was available to produce meaningful results.

Nevertheless, ARACNE was used to calculate the mutual information between the different modifications for genes assigned to the clusters of the *no impact* group and the groups *activating* and *silencing*.

This revealed that H3K9ac, H4K5ac and H3K14ac exhibit a much higher mutual

information for these clusters, compared to those which were not associated with an impact on gene expression. This can also be interpreted as that H3K9ac, H4K5ac and H3K14ac together activate genes, thus showing that a combination of modifications is responsible for gene expression and not a single one of them is sufficient to activate a gene. Furthermore, it was shown that H3K36me3 which shows no comodification with H3K9ac, H4K5ac and H3K14ac also impacts how those three modifications regulate gene expression.

For H4K16ac and H3K56ac no significant impact on expression could be detected, whereas a small impact for H3K4me3 could be observed, although H3K4me3 shows a higher correlation value than the one of H3K36me3 (H3K36me3).

GO term analysis was performed on genes that exhibit similar modification patterns which were linked with transcriptional activation. It was possible to show that sets of genes which work together to accomplish processes within a cell are regulated by histone modifications.

4.5 Outlook

Several aspects of the influence of histone modifications on gene expression were analyzed in this thesis, and interesting results were found, which show the complexity of the topic. Nevertheless, additional aspects can be investigated using similar methods and data.

A very promising approach would be using ARACNE to detect how a modification of a gene impacts the same modification for other genes. However, in order to apply ARACNE a much larger number of samples is necessary. The authors of ARACNE state that at least 100 different samples for each gene must be available for the algorithm to produce reliable results.

This thesis evaluated a dataset of 7 different histone modifications for 16 points in time, but only modifications with no impact or an activating impact on gene expression were available. A far more detailed study would be possible if a more diverse set of histone modifications was available. The analysis of which modifications share mutual information could reveal how they work together to regulate gene expression.

Gene onthology enrichment analysis was only performed on those two clusters which showed the greatest activating impact on gene expression. In a further study, more clusters could be examined to obtain a model of how histone modifications govern functional processes.

Bibliography

[1] Hecker M, Lambeck S *et al.*: **Gene regulatory network inference: Data integration in dynamic models—A review**. *Biosystems* 2009. **96**(1):86–103.

[2] Mohn F and Schübeler D: **Genetics and epigenetics: stability and plasticity during cellular differentiation**. *Trends in Genetics* 2009. **25**(3):129–136.

[3] Carmona-Gutierrez D, Eisenberg T *et al.*: **Apoptosis in yeast: triggers, pathways, subroutines**. *Cell Death & Differentiation* 2010. **17**(5):763–773.

[4] Hassa PO and Hottiger MO: **An epigenetic code for DNA damage repair pathways?** *Biochemistry and Cell Biology* 2005. **83**(3):270–285.

[5] Jin B, Li Y *et al.*: **DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?** *Genes & Cancer* 2011. **2**(6):607–617.

[6] Wheeler R: **Nucleosome_structure.png, CC BY-SA 3.0**. 2005. URL `https://commons.wikimedia.org/wiki/File:Nucleosome_structure.png`. Accessed 01 May 2018.

[7] Bettkober M: **Charakterisierung der Histondemethylase JMJD1A als neues Zielgen von HIF-1** 2014. URL `http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000096950`. Accessed 01 May 2018.

[8] Flotho A and Melchior F: **Sumoylation: A Regulatory Protein Modification in Health and Disease**. *Annual Review of Biochemistry* 2013. **82**(1):357–385.

[9] Turner BM: **Reading signals on the nucleosome with a new nomenclature for modified histones**. *Nature Structural & Molecular Biology* 2005. **12**(2):110–112. Accessed 01 May 2018.

[10] Krogan NJ, Kim M *et al.*: **Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II**. *Molecular and Cellular Biology* 2003. **23**(12):4207–4218.

[11] Briggs SD: **Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae**. *Genes & Development* 2001. **15**(24):3286–3295.

[12] Kuang Z, Cai L *et al.*: **High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast**. *Nature Structural & Molecular Biology* 2014. **21**(10):854–863.

[13] Xhemalce B, Dawson MA *et al.*: **Histone Modifications**. *Reviews in Cell Biology and Molecular Medicine* 2011.

[14] Carbon S, Ireland A *et al.*: **AmiGO: online access to ontology and annotation data**. 2008. URL `http://amigo.geneontology.org/amigo/search/bioentity?q=*:*&fq=regulates_closure:"GO:0004402"&fq=source:"SGD"&sfq=document_category:"bioentity"`. Accessed 01 May 2018.

[15] Carbon S, Ireland A *et al.*: **AmiGO: online access to ontology and annotation data**. 2008. URL `http://amigo.geneontology.org/amigo/search/bioentity?q=*:*&fq=regulates_closure:"GO:0004407"&fq=source:"SGD"&sfq=document_category:"bioentity"`. Accessed 01 May 2018.

[16] Bannister AJ and Kouzarides T: **Regulation of chromatin by histone modifications**. *Cell Research* 2011. **21**(3):381–395.

[17] Dillon SC, Zhang X *et al.*: **The SET-domain protein superfamily: protein lysine methyltransferases**. *Genome Biology* 2005. **6**(8):227.

[18] Jenuwein T: **Translating the Histone Code**. *Science* 2001. **293**(5532):1074–1080.

[19] Yun M, Wu J *et al.*: **Readers of histone modifications**. *Cell Research* 2011. **21**(4):564–578.

[20] Taverna SD, Li H *et al.*: **How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers**. *Nature Structural & Molecular Biology* 2007. **14**(11):1025–1040.

[21] Brock T: **Histone Methylation: SET versus Jumonji** 2010. URL `https://www.caymanchem.com/news/histone-methylation`. Accessed 01 May 2018.

[22] Engel SR, Dietrich FS *et al.*: **The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now** . *G3: Genes—Genomes—Genetics* 2013. **4**(3):389–398.

[23] Szerlong HJ and Hansen JC: **Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure.** *Biochemistry and Cell Biology* 2011. **89**(1):24–34.

[24] Hall N: **Advanced sequencing technologies and their wider impact in microbiology**. *Journal of Experimental Biology* 2007. **210**(9):1518–1525.

[25] Liang S, Lu Y *et al.*: **Analysis of epigenetic modifications by next generation sequencing**. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society: 03-06 Sept 2009; Minneapolis, Minnesota.* 2009.

[26] Di Bella JM, Bao Y *et al.*: **High throughput sequencing methods and analysis for microbiome research**. *Journal of Microbiological Methods* 2013. **95**(3):401–414.

[27] Ghorbanpour M, Manika K *et al.*: **Nanoscience and Plant–Soil Systems**. Springer International Publishing, 2017. ISBN 97833194683 58.

[28] Wang Z, Gerstein M *et al.*: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009. **10**(1):57–63.

[29] Zhao S, Fung-Leung WP *et al.*: **Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells**. *PLoS ONE* 2014. **9**(1):e78644.

[30] Furey TS: **ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions**. *Nature Reviews Genetics* 2012. **13**(12):840–852.

[31] Landt SG, Marinov GK *et al.*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome Research* 2012. **22**(9):1813–1831.

[32] Kuang Z, Cai L *et al.*: **GEO Series GSE52339**. 2014. URL `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52339`. Accessed 01 May 2018.

[33] Leinonen R, Sugawara H *et al.*: **The Sequence Read Archive**. *Nucleic Acids Research* 2010. **39**(Database):D19–D21. Accessed 01 May 2018.

[34] **The Ruby programming language**. 2017. URL `https://www.ruby-lang.org/`. Accessed 01 May 2018.

[35] Trapnell C, Pachter L *et al.*: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009. **25**(9):1105–1111.

[36] Trapnell C, Roberts A *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature Protocols* 2012. **7**(3):562–578. Accessed 01 May 2018.

[37] **Apache Groovy**. 2017. URL `http://groovy-lang.org/`. Accessed 01 May 2018.

[38] The MathWorks, Inc.: **Mathworks Matlab**. 2017. URL `https://www.mathworks.com/products/new_products/release2015b.html`. Accessed 01 May 2018.

[39] Li H and Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009. **25**(14):1754–1760.

[40] Lunter G and Goodson M: **Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads**. *Genome Research* 2010. **21**(6):936–939.

[41] Santana-Quintero L, Dingerdissen H *et al.*: **HIVE-Hexagon: High-Performance, Parallelized Sequence Alignment for Next-Generation Sequencing Data Analysis**. *PLoS ONE* 2014. **9**(6):e99033.

[42] Langmead B, Trapnell C *et al.*: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009. **10**(3):R25.

[43] Zhang Y, Liu T *et al.*: **Model-based Analysis of ChIP-Seq (MACS)**. *Genome Biology* 2008. **9**(9):R137.

[44] Hartigan JA and Wong MA: **Algorithm AS 136: A K-Means Clustering Algorithm**. *Applied Statistics* 1979. **28**(1):100.

[45] Lloyd S: **Least squares quantization in PCM**. *IEEE Transactions on Information Theory* 1982. **28**(2):129–137.

[46] Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. *Journal of Computational and Applied Mathematics* 1987. **20**:53–65.

[47] Yeung KY, Haynor DR *et al.*: **Validating clustering for gene expression data**. *Bioinformatics* 2001. **17**(4):309–318.

[48] Akaike H: **Akaike's Information Criterion**. In M Lovric, editor, **International Encyclopedia of Statistical Science**, pages 25–25. Springer Berlin Heidelberg. ISBN 9783642048982, 2011.

[49] Burnham KP and Anderson DR: **Multimodel Inference - Understanding AIC and BIC in Model Selection**. *Sociological Methods & Research* 2004. **33**(2):261–304.

[50] Schwarz G *et al.*: **Estimating the dimension of a model**. *The Annals of Statistics* 1978. **6**(2):461–464.

[51] **Javascript** 2017. URL `https://www.javascript.com/`. Accessed 01 May 2018.

[52] Bostock M, Ogievetsky V *et al.*: **D^3 data-driven documents**. *IEEE Transactions on Visualization and Computer Graphics* 2011. **17**(12):2301–2309. URL `https://d3js.org/`. Accessed 01 May 2018.

[53] Nemenman I: **Information theory, multivariate dependence, and genetic network inference**. *eprint arXiv:q-bio/0406015* 2004. URL `https://arxiv.org/abs/q-bio/0406015`.

[54] De Mántaras RL: **A distance-based attribute selection measure for decision tree induction**. *Machine Learning* 1991. **6**(1):81–92.

[55] Burnham KP and Anderson D: **Model selection and multi-model inference**, volume **1229**. Taylor & Francis, 2003. ISBN 9780387953649.

[56] Kraskov A, Stögbauer H *et al.*: **Estimating mutual information**. *Physical Review E* 2004. **69**(6).

[57] Margolin AA, Nemenman I *et al.*: **ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context**. *BMC Bioinformatics* 2006. **7**(Suppl 1):S7.

[58] **OpenOffice**. 2017. URL `https://www.openoffice.org/de/`. Accessed 14 May 2018.

[59] Breiman L, Friedman J *et al.*: **Classification and regression trees**, volume **1**. Chapman and Hall/CRC, 1984. ISBN 9780412048418.

[60] The MathWorks, Inc.: **Matlab decisiontree function**. 2017. URL `https://de.mathworks.com/help/stats/decision-trees.html?requestedDomain=true`. Accessed 01 May 2018.

[61] Eden E, Navon R *et al.*: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009. **10**(1):48.

[62] Song YY and Ying L: **Decision tree methods: applications for classification and prediction**. *Shanghai Archives of Psychiatry* 2015. **27**(2):130.

5 Appendix

Theorem 1 *If a normal distribution is assumed for each cluster*

$$f_{cl_i}(\vec{x}, \vec{\mu}_i, \Sigma) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma)}} \exp(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})) \tag{A1}$$

where Σ is the covariance matrix, d is the dimensionality of the space and μ is the cluster center. As the k-means algorithm minimizes the sum of squares within each cluster and assigns each point to the cluster with minimum distance to the center, each cluster forms a spheroid. Thus the covariance matrix can be written as

$$\Sigma = \sigma^2 I \tag{A2}$$

where σ^2 is the variance and I is the identity matrix. Therefore equation A1 can be simplified to

$$f_{cl_i}(\vec{x}, \vec{\mu}_i, \Sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp(-\frac{(\vec{x} - \vec{\mu})^2}{2\sigma^2}) \tag{A3}$$

As a consequence the log-likelihood can be written as

$$\begin{aligned} \mathcal{L}(\Theta; \vec{x}) &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp(-\frac{(\vec{x}_i - \vec{\mu})^2}{2\sigma^2})) \\ &= -n \log(\sqrt{(2\pi\sigma^2)^d}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})^2 \end{aligned} \tag{A4}$$

When using the log-likelihood function in the context of the AIC, we only want to compare differences between different AICs. The term $-n \log(\sqrt{(2\pi\sigma^2)^d})$ is constant over experiments with different values for k and can therefore be omitted.
This results in a formula for the AIC as shown in A5

$$AIC = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})^2 + 2kd \tag{A5}$$

The derivation of the likelihood function for k-means considers the algorithm as a version of an Expectation Maximization algorithm with a Gaussian mixture model with a constant value for

σ for all Gaussians. Using the Lloyd k-means algorithm, the value for σ is not a free parameter but can be chosen arbitrarily. Unfortunately, the value for σ does have an impact on the AIC in a way that a greater value for σ favors a model with less parameters. In this thesis, the value for σ was set to 1*, as the data is standardized in each dimension and this would be the correct value, if a single cluster was calculated.*

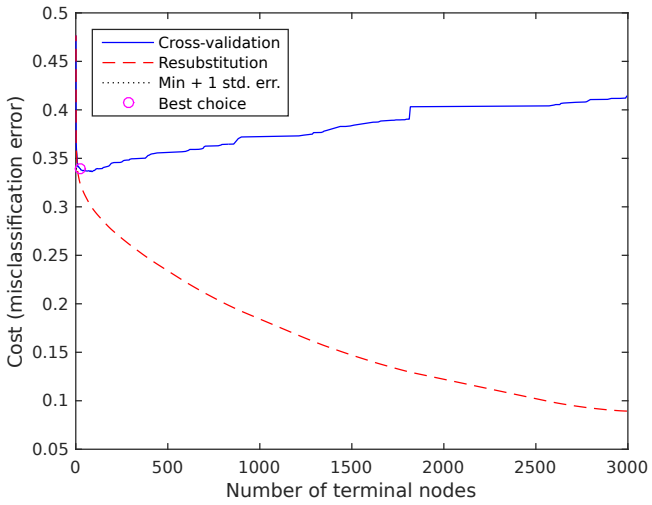

Figure A1: Training result for the tree shown in figure 3.10 where the full dataset was used. The tree generalizes very badly, as training took around 3000 terminal nodes to approximate the training data, but at a height of 8 the cross validation error started to increase.

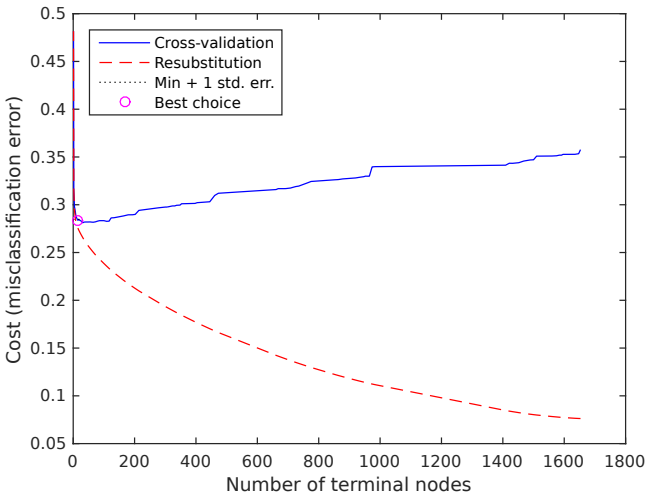

Figure A2: Training result of the tree shown in figure 3.11 where only patterns with expression values less than 0.5 or greater than -0.5 were used. The tree took less height to approximate the training data compared to 3.10, but cross validation still diverges very fast.

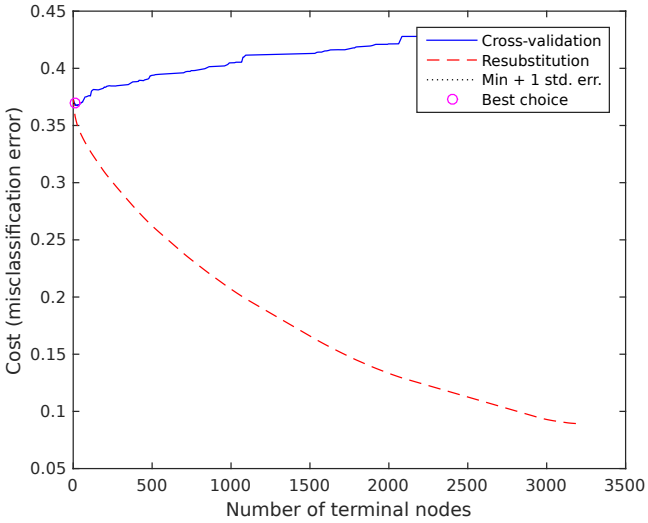

Figure A3: The training performance for the tree shown in figure 3.13 where the full dataset was used. The tree resulted in bad generalization with classification performance of 66.52% on the validation data.

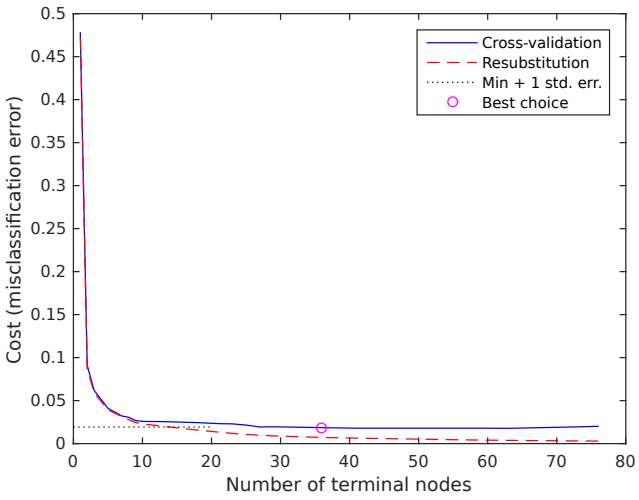

Figure A4: Training result for the tree in figure 3.14 where only patterns from the clusters *activating* and *silencing* were used. The tree converges fast and also approximates the cross validation data very well.

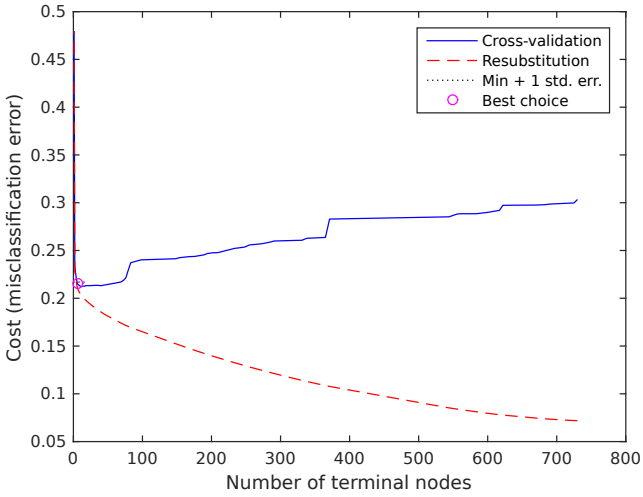

Figure A5: Training result for the tree in figure 3.15 where only patterns from the clusters *activating* and *silencing* were used. It can be observed that the cross-valdation still diverges fast, but the overall performance on the validation data was 78.72% accuracy.

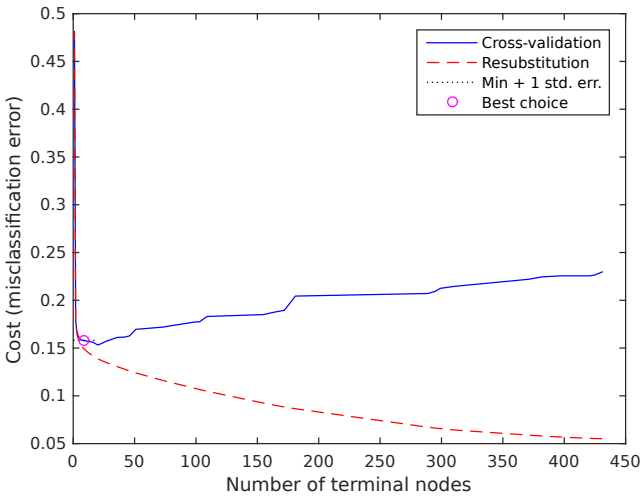

Figure A6: Training performance for the tree shown in figure 3.16 where only patterns from the clusters *activating* and *silencing* with mean expression between -0.5 and 0.5 were used.

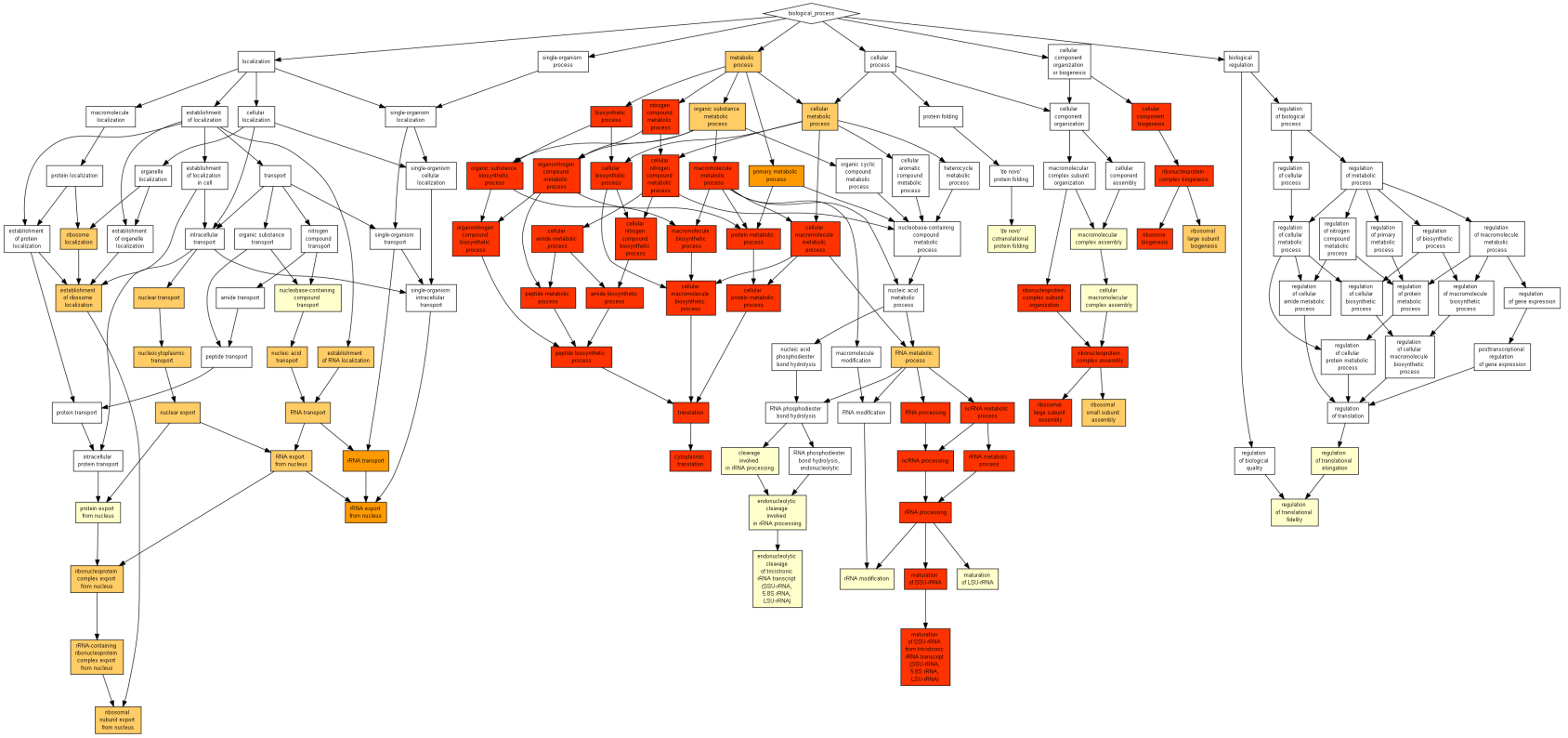

Figure A7: GO term analysis for *biological process* for the cluster with the second greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A4.

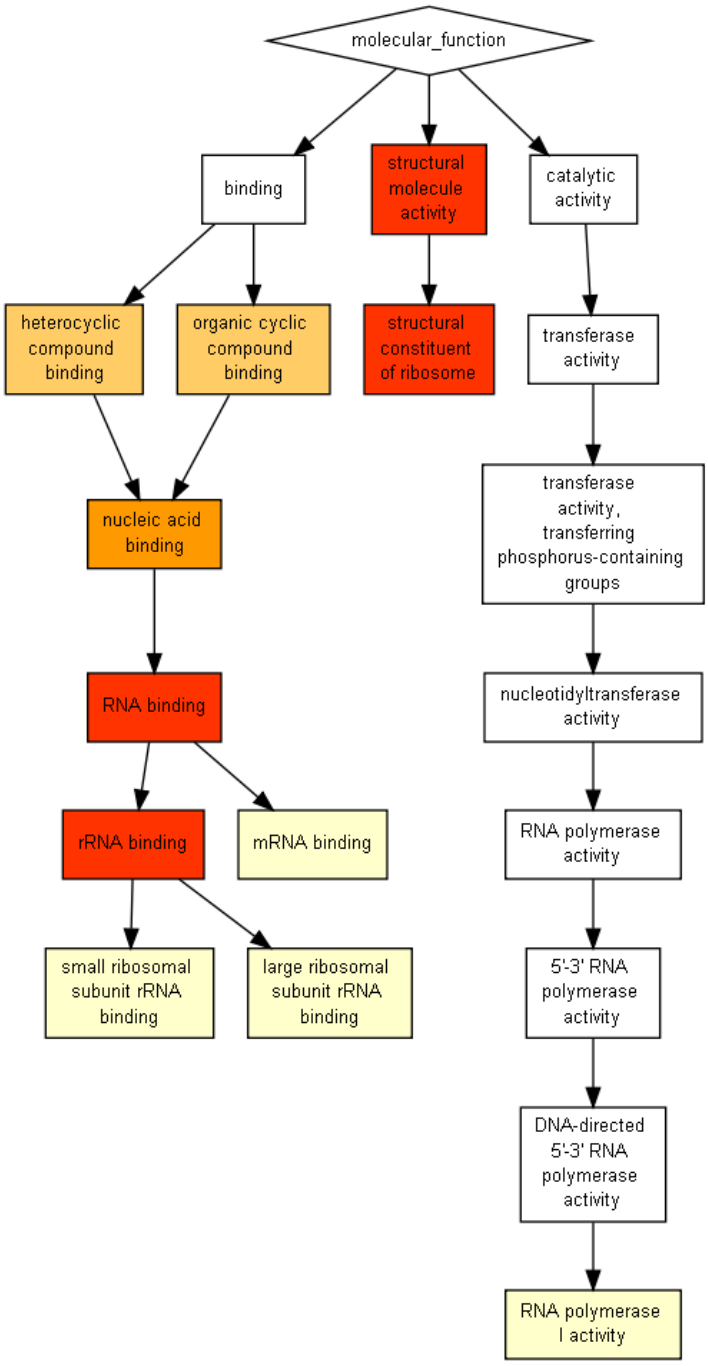

Figure A8: GO term analysis for *molecular functions* for the cluster with the second greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A5.

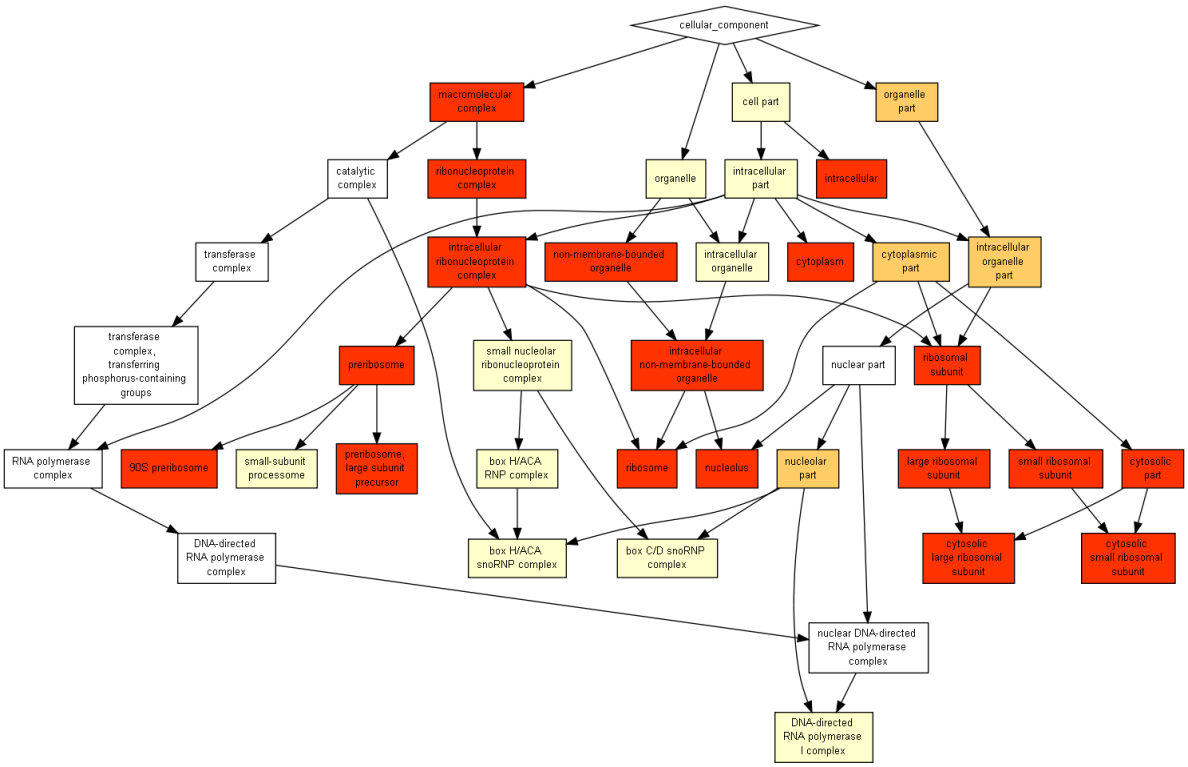

Figure A9: GO term analysis for *cellular component* for the cluster with the second greatest observed impact on gene expression. A list of the GO terms with the according p-value is provided in table A6.

Table A1: Enriched GO terms for biological process of the cluster with the highest expression mean.

GO term	Description	P-value
GO:0034660	ncRNA metabolic process	3.76E-77
GO:0022613	ribonucleoprotein complex biogenesis	9.99E-65
GO:0034470	ncRNA processing	1.51E-63
GO:0016072	rRNA metabolic process	8.91E-60
GO:0042254	ribosome biogenesis	1.04E-59
GO:0006364	rRNA processing	5.55E-57
GO:0044085	cellular component biogenesis	3.66E-54
GO:0006396	RNA processing	1.03E-47
GO:0016070	RNA metabolic process	5.61E-39

GO term	Description	P-value
GO:0090304	nucleic acid metabolic process	1.06E-32
GO:0006139	nucleobase-containing compound metabolic process	9.01E-32
GO:0006725	cellular aromatic compound metabolic process	9.33E-32
GO:0046483	heterocycle metabolic process	3.02E-30
GO:0034641	cellular nitrogen compound metabolic process	3.62E-30
GO:1901360	organic cyclic compound metabolic process	2.32E-29
GO:0044260	cellular macromolecule metabolic process	3.42E-23
GO:0006807	nitrogen compound metabolic process	3.95E-23
GO:0043170	macromolecule metabolic process	9.15E-23
GO:0000469	cleavage involved in rRNA processing	2.78E-21
GO:0000447	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	2.2E-20
GO:0000478	endonucleolytic cleavage involved in rRNA processing	1.78E-19
GO:0000479	endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	1.78E-19
GO:0000463	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	2.33E-19
GO:0000480	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	1.15E-18
GO:0034471	ncRNA 5'-end processing	7.07E-18
GO:0000967	rRNA 5'-end processing	7.07E-18
GO:0044238	primary metabolic process	1.09E-17
GO:0000966	RNA 5'-end processing	1.63E-17
GO:0006399	tRNA metabolic process	3.25E-17
GO:0000470	maturation of LSU-rRNA	4.25E-17
GO:0030490	maturation of SSU-rRNA	4.39E-17
GO:0000472	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	8.43E-17

GO term	Description	P-value
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	2.84E-16
GO:0044237	cellular metabolic process	3.96E-16
GO:0071704	organic substance metabolic process	6.34E-16
GO:0008152	metabolic process	3.67E-14
GO:0042273	ribosomal large subunit biogenesis	7.76E-14
GO:0009451	RNA modification	5.69E-13
GO:0090501	RNA phosphodiester bond hydrolysis	1.87E-12
GO:0000466	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	1.05E-11
GO:0000460	maturation of 5.8S rRNA	8.61E-11
GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic	3.49E-10
GO:0090305	nucleic acid phosphodiester bond hydrolysis	3.62E-10
GO:0042274	ribosomal small subunit biogenesis	1.31E-09
GO:0006400	tRNA modification	3.36E-09
GO:0071840	cellular component organization or biogenesis	8.78E-09
GO:0006405	RNA export from nucleus	1.65E-08
GO:0001510	RNA methylation	1.74E-08
GO:0051169	nuclear transport	2.24E-08
GO:0051168	nuclear export	2.27E-08
GO:0010501	RNA secondary structure unwinding	3.47E-08
GO:0008033	tRNA processing	4.59E-08
GO:0071826	ribonucleoprotein complex subunit organization	5.34E-08
GO:0098781	ncRNA transcription	6.45E-08
GO:0006913	nucleocytoplasmic transport	7.61E-08
GO:0022618	ribonucleoprotein complex assembly	0.000000163
GO:0032259	methylation	0.000000322
GO:0043414	macromolecule methylation	0.000000329
GO:0009987	cellular process	0.000000357

GO term	Description	P-value
GO:0050658	RNA transport	0.000000436
GO:0051236	establishment of RNA localization	0.000000436
GO:0050657	nucleic acid transport	0.000000558
GO:0006413	translational initiation	0.00000236
GO:0006418	tRNA aminoacylation for protein translation	0.00000301
GO:0000027	ribosomal large subunit assembly	0.00000301
GO:0071426	ribonucleoprotein complex export from nucleus	0.00000358
GO:0071428	rRNA-containing ribonucleoprotein complex export from nucleus	0.00000358
GO:0033753	establishment of ribosome localization	0.00000358
GO:0000054	ribosomal subunit export from nucleus	0.00000358
GO:1902626	assembly of large subunit precursor of preribosome	0.00000406
GO:0033750	ribosome localization	0.00000458
GO:0000154	rRNA modification	0.0000047
GO:0043038	amino acid activation	0.00000683
GO:0043039	tRNA aminoacylation	0.00000683
GO:0002097	tRNA wobble base modification	0.000013
GO:0006360	transcription from RNA polymerase I promoter	0.000013
GO:0009303	rRNA transcription	0.000013
GO:0042790	transcription of nuclear large rRNA transcript from RNA polymerase I promoter	0.000015
GO:0015931	nucleobase-containing compound transport	0.0000152
GO:0016073	snRNA metabolic process	0.0000178
GO:0006611	protein export from nucleus	0.0000214
GO:0071035	nuclear polyadenylation-dependent rRNA catabolic process	0.0000353
GO:0000055	ribosomal large subunit export from nucleus	0.000043
GO:0031167	rRNA methylation	0.0000483
GO:0043634	polyadenylation-dependent ncRNA catabolic process	0.0000518
GO:0071029	nuclear ncRNA surveillance	0.0000518
GO:0071046	nuclear polyadenylation-dependent ncRNA catabolic process	0.0000518

GO term	Description	P-value
GO:0009116	nucleoside metabolic process	0.0000693
GO:0044271	cellular nitrogen compound biosynthetic process	0.0000908
GO:0043633	polyadenylation-dependent RNA catabolic process	0.000104
GO:0030488	tRNA methylation	0.000142
GO:0034476	U5 snRNA 3′-end processing	0.000155
GO:1901657	glycosyl compound metabolic process	0.000186
GO:0042797	tRNA transcription from RNA polymerase III promoter	0.000248
GO:0071027	nuclear RNA surveillance	0.000254
GO:0009163	nucleoside biosynthetic process	0.000254
GO:0006356	regulation of transcription from RNA polymerase I promoter	0.000257
GO:0043604	amide biosynthetic process	0.000292
GO:0000459	exonucleolytic trimming involved in rRNA processing	0.000347
GO:0009304	tRNA transcription	0.000347
GO:0043628	ncRNA 3′-end processing	0.000376
GO:0002098	tRNA wobble uridine modification	0.000428
GO:0071025	RNA surveillance	0.000428
GO:0034504	protein localization to nucleus	0.000459
GO:0034472	snRNA 3′-end processing	0.000473
GO:0071049	nuclear retention of pre-mRNA with aberrant 3′-ends at the site of transcription	0.000594
GO:0006361	transcription initiation from RNA polymerase I promoter	0.000594
GO:0016180	snRNA processing	0.000633
GO:1901659	glycosyl compound biosynthetic process	0.000685
GO:0034661	ncRNA catabolic process	0.000764
GO:0010468	regulation of gene expression	0.000783
GO:0031119	tRNA pseudouridine synthesis	0.000837
GO:0042255	ribosome assembly	0.000837
GO:0046134	pyrimidine nucleoside biosynthetic process	0.000837
GO:0034475	U4 snRNA 3′-end processing	0.000841

GO term	Description	P-value
GO:0090503	RNA phosphodiester bond hydrolysis, exonucleolytic	0.000853
GO:0006353	DNA-templated transcription, termination	0.000917
GO:0006412	translation	0.000969

Table A2: Enriched GO terms for molecular function of the cluster with the highest expression mean.

GO term	Description	P-value
GO:0003723	RNA binding	6.34E-28
GO:0003676	nucleic acid binding	3.87E-19
GO:0097159	organic cyclic compound binding	1.44E-15
GO:1901363	heterocyclic compound binding	2E-15
GO:0030515	snoRNA binding	1.38E-11
GO:0004004	ATP-dependent RNA helicase activity	1.75E-09
GO:0008186	RNA-dependent ATPase activity	1.75E-09
GO:0003729	mRNA binding	3.64E-09
GO:0003724	RNA helicase activity	3.7E-09
GO:0005488	binding	1.33E-08
GO:0003743	translation initiation factor activity	1.88E-08
GO:0008135	translation factor activity, RNA binding	2.64E-08
GO:0008173	RNA methyltransferase activity	0.000000131
GO:0000049	tRNA binding	0.000000188
GO:0008026	ATP-dependent helicase activity	0.000000596
GO:0070035	purine NTP-dependent helicase activity	0.000000596
GO:0034062	5′-3′ RNA polymerase activity	0.000000869
GO:0003899	DNA-directed 5′-3′ RNA polymerase activity	0.000000869
GO:0097747	RNA polymerase activity	0.000000869
GO:0004812	aminoacyl-tRNA ligase activity	0.00000225
GO:0016875	ligase activity, forming carbon-oxygen bonds	0.00000225
GO:0016876	ligase activity, forming aminoacyl-tRNA and related compounds	0.00000225

GO term	Description	P-value
GO:0001054	RNA polymerase I activity	0.00000304
GO:0008168	methyltransferase activity	0.00000357
GO:0016741	transferase activity, transferring one-carbon groups	0.0000129
GO:0004386	helicase activity	0.0000191
GO:0008649	rRNA methyltransferase activity	0.0000483
GO:0016874	ligase activity	0.000136
GO:0001056	RNA polymerase III activity	0.000173
GO:0008757	S-adenosylmethionine-dependent methyltransferase activity	0.000262
GO:0008175	tRNA methyltransferase activity	0.000347
GO:0009982	pseudouridine synthase activity	0.000394
GO:0097367	carbohydrate derivative binding	0.000496
GO:0019843	rRNA binding	0.000528
GO:0035639	purine ribonucleoside triphosphate binding	0.000576
GO:0032555	purine ribonucleotide binding	0.000597
GO:0017076	purine nucleotide binding	0.000709
GO:0042134	rRNA primary transcript binding	0.000837

Table A3: Enriched GO terms for cellular component of the cluster with the highest expression mean.

GO term	Description	P-value
GO:0005730	nucleolus	2.52E-75
GO:0030684	preribosome	5.5E-50
GO:0030686	90S preribosome	1.43E-28
GO:0043228	non-membrane-bounded organelle	8.32E-27
GO:0043232	intracellular non-membrane-bounded organelle	8.32E-27
GO:0005634	nucleus	3.25E-25
GO:1990904	ribonucleoprotein complex	2.56E-24
GO:0030529	intracellular ribonucleoprotein complex	2.56E-24
GO:0044428	nuclear part	4.04E-23

GO term	Description	P-value
GO:0032040	small-subunit processome	5.89E-23
GO:0030687	preribosome, large subunit precursor	1.75E-18
GO:0044452	nucleolar part	1.92E-17
GO:0032991	macromolecular complex	1.9E-10
GO:0044424	intracellular part	3.49E-08
GO:0044464	cell part	4.05E-08
GO:0043229	intracellular organelle	0.000000109
GO:0043226	organelle	0.000000113
GO:0034455	t-UTP complex	0.000000156
GO:0005654	nucleoplasm	0.000000273
GO:0043231	intracellular membrane-bounded organelle	0.00000117
GO:0043227	membrane-bounded organelle	0.00000259
GO:0005736	DNA-directed RNA polymerase I complex	0.00000304
GO:0033553	rDNA heterochromatin	0.00000854
GO:0055029	nuclear DNA-directed RNA polymerase complex	0.0000113
GO:0000428	DNA-directed RNA polymerase complex	0.0000181
GO:0030880	RNA polymerase complex	0.0000181
GO:0030689	Noc complex	0.0000395
GO:0005851	eukaryotic translation initiation factor 2B complex	0.000114
GO:0034457	Mpp10 complex	0.000155
GO:0005666	DNA-directed RNA polymerase III complex	0.000173
GO:0000792	heterochromatin	0.000248
GO:0032045	guanyl-nucleotide exchange factor complex	0.000254
GO:0030688	preribosome, small subunit precursor	0.000586

Table A4: Enriched GO terms for biological process of the cluster with the second greatest expression mean.

GO term	Description	P-value
GO:0002181	cytoplasmic translation	1.53E-82

GO term	Description	P-value
GO:0006412	translation	2.51E-56
GO:0043043	peptide biosynthetic process	8.22E-56
GO:0043604	amide biosynthetic process	6.09E-53
GO:0006518	peptide metabolic process	5.93E-51
GO:0043603	cellular amide metabolic process	4.26E-47
GO:1901566	organonitrogen compound biosynthetic process	5.6E-35
GO:0044271	cellular nitrogen compound biosynthetic process	3.24E-25
GO:0044267	cellular protein metabolic process	8.45E-24
GO:0034645	cellular macromolecule biosynthetic process	1.74E-23
GO:0009059	macromolecule biosynthetic process	2.09E-23
GO:0006364	rRNA processing	8E-22
GO:0022613	ribonucleoprotein complex biogenesis	5.44E-21
GO:0042254	ribosome biogenesis	8.1E-20
GO:0016072	rRNA metabolic process	1.36E-19
GO:0034641	cellular nitrogen compound metabolic process	1.75E-19
GO:0034470	ncRNA processing	3.55E-18
GO:0019538	protein metabolic process	2.23E-17
GO:0044085	cellular component biogenesis	1.44E-16
GO:0034660	ncRNA metabolic process	1.46E-16
GO:1901576	organic substance biosynthetic process	4.46E-16
GO:0009058	biosynthetic process	8.4E-16
GO:0044249	cellular biosynthetic process	9.66E-16
GO:0044260	cellular macromolecule metabolic process	1.03E-14
GO:0022618	ribonucleoprotein complex assembly	7.8E-14
GO:0043170	macromolecule metabolic process	1.07E-13
GO:0030490	maturation of SSU-rRNA	3.82E-13
GO:0071826	ribonucleoprotein complex subunit organization	5.27E-13
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	9.29E-13

GO term	Description	P-value
GO:0006396	RNA processing	1.06E-11
GO:1901564	organonitrogen compound metabolic process	1.6E-11
GO:0006807	nitrogen compound metabolic process	7.51E-11
GO:0000027	ribosomal large subunit assembly	7.07E-10
GO:0006407	rRNA export from nucleus	3.07E-09
GO:0051029	rRNA transport	3.07E-09
GO:0044238	primary metabolic process	6.58E-09
GO:0000028	ribosomal small subunit assembly	0.000000197
GO:0071704	organic substance metabolic process	0.000000376
GO:0006405	RNA export from nucleus	0.000000452
GO:0006913	nucleocytoplasmic transport	0.000000479
GO:0051169	nuclear transport	0.000000531
GO:0051168	nuclear export	0.000000576
GO:0044237	cellular metabolic process	0.00000152
GO:0008152	metabolic process	0.00000293
GO:0050658	RNA transport	0.00000379
GO:0051236	establishment of RNA localization	0.00000379
GO:0042273	ribosomal large subunit biogenesis	0.00000413
GO:0050657	nucleic acid transport	0.00000459
GO:0071426	ribonucleoprotein complex export from nucleus	0.00000736
GO:0071428	rRNA-containing ribonucleoprotein complex export from nucleus	0.00000736
GO:0033753	establishment of ribosome localization	0.00000736
GO:0000054	ribosomal subunit export from nucleus	0.00000736
GO:0016070	RNA metabolic process	0.00000887
GO:0033750	ribosome localization	0.00000904
GO:0006611	protein export from nucleus	0.0000329
GO:0000469	cleavage involved in rRNA processing	0.0000331
GO:0000470	maturation of LSU-rRNA	0.0000359
GO:0034622	cellular macromolecular complex assembly	0.0000438

GO term	Description	P-value
GO:0065003	macromolecular complex assembly	0.0000959
GO:0015931	nucleobase-containing compound transport	0.000137
GO:0000478	endonucleolytic cleavage involved in rRNA processing	0.000192
GO:0000479	endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	0.000192
GO:0006448	regulation of translational elongation	0.000314
GO:0006450	regulation of translational fidelity	0.000376
GO:0000154	rRNA modification	0.000376
GO:0051083	'de novo' cotranslational protein folding	0.000539

Table A5: Enriched GO terms for molecular function of the cluster with the second greatest expression mean.

GO term	Description	P-value
GO:0003735	structural constituent of ribosome	1.46E-67
GO:0005198	structural molecule activity	2.63E-44
GO:0003723	RNA binding	6.63E-16
GO:0019843	rRNA binding	1.42E-11
GO:0003676	nucleic acid binding	2.77E-08
GO:1901363	heterocyclic compound binding	0.0000051
GO:0097159	organic cyclic compound binding	0.00000681
GO:0001054	RNA polymerase I activity	0.000125
GO:0070181	small ribosomal subunit rRNA binding	0.000135
GO:0070180	large ribosomal subunit rRNA binding	0.000381
GO:0003729	mRNA binding	0.000892

Table A6: Enriched GO terms for cellular component of the cluster with the second greatest expression mean.

GO term	Description	P-value
GO:0044391	ribosomal subunit	1.3E-65
GO:0044445	cytosolic part	2.81E-64
GO:0005840	ribosome	3.54E-59
GO:0022625	cytosolic large ribosomal subunit	2.66E-55
GO:1990904	ribonucleoprotein complex	1.1E-54
GO:0030529	intracellular ribonucleoprotein complex	1.1E-54
GO:0005622	intracellular	2.67E-45
GO:0043228	non-membrane-bounded organelle	3.97E-42
GO:0043232	intracellular non-membrane-bounded organelle	3.97E-42
GO:0015934	large ribosomal subunit	8.47E-41
GO:0022627	cytosolic small ribosomal subunit	4.5E-29
GO:0030684	preribosome	5.94E-25
GO:0015935	small ribosomal subunit	3.64E-23
GO:0032991	macromolecular complex	2.12E-16
GO:0030687	preribosome, large subunit precursor	5.62E-15
GO:0005737	cytoplasm	1.09E-13
GO:0005730	nucleolus	2.67E-11
GO:0030686	90S preribosome	4.75E-11
GO:0044446	intracellular organelle part	0.000000153
GO:0044422	organelle part	0.000000175
GO:0044444	cytoplasmic part	0.00000235
GO:0044452	nucleolar part	0.00000395
GO:0032040	small-subunit processome	0.0000195
GO:0031428	box C/D snoRNP complex	0.0000308
GO:0043229	intracellular organelle	0.0000441
GO:0043226	organelle	0.0000451
GO:0044424	intracellular part	0.0000912

GO term	Description	P-value
GO:0005732	small nucleolar ribonucleoprotein complex	0.000112
GO:0005736	DNA-directed RNA polymerase I complex	0.000125
GO:0044464	cell part	0.000449
GO:0072588	box H/ACA RNP complex	0.000539
GO:0031429	box H/ACA snoRNP complex	0.000539

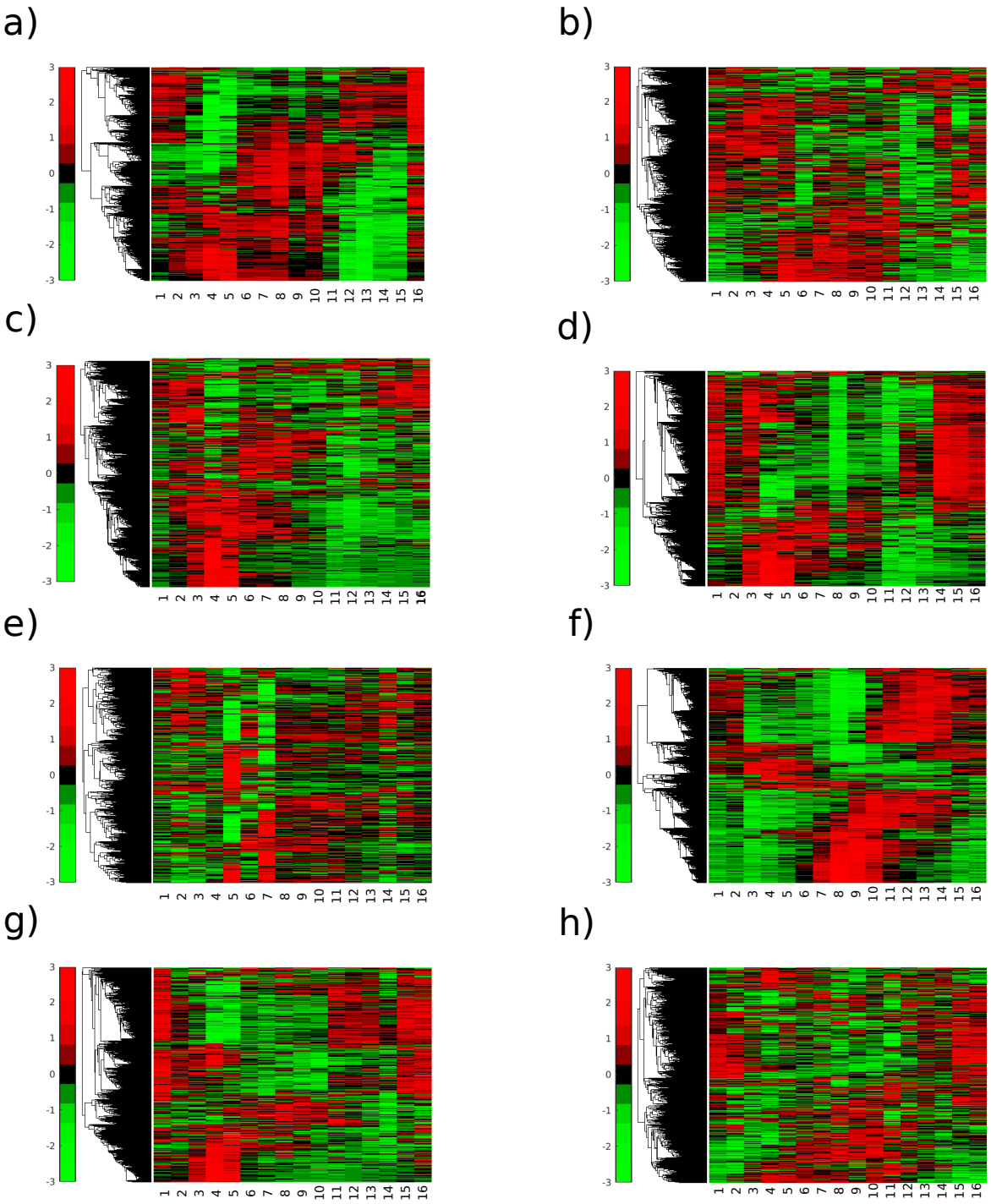

Figure A10: a) Heatmap visualizing the expression for the full set of genes for 16 points in time **b)** H3K4me3 modification signals for the 16 points in time **c)** H3K9ac modification signals **d)** H3K14ac modification signals **e)** H3K36me3 modification signals **f)** H3K56ac modification signals **g)** H4K5ac modification signals **h)** H4K16ac modification signals

```
SRR1029143  RNA-seq_t1
SRR1029144  RNA-seq_t2
SRR1029145  RNA-seq_t3
SRR1029146  RNA-seq_t4
SRR1029147  RNA-seq_t5
SRR1029148  RNA-seq_t6
SRR1029149  RNA-seq_t7
SRR1029150  RNA-seq_t8
SRR1029151  RNA-seq_t9
SRR1029152  RNA-seq_t10
SRR1029153  RNA-seq_t11
SRR1029154  RNA-seq_t12
SRR1029155  RNA-seq_t13
SRR1029156  RNA-seq_t14
SRR1029157  RNA-seq_t15
SRR1029158  RNA-seq_t16
SRR1029159  H3K9ac_time alignment_ChIP-seq_t1
SRR1029160  H3K9ac_time alignment_ChIP-seq_t2
SRR1029161  H3K9ac_time alignment_ChIP-seq_t3
SRR1029162  H3K9ac_time alignment_ChIP-seq_t4
SRR1029163  H3K9ac_time alignment_ChIP-seq_t5
SRR1029164  H3K9ac_time alignment_ChIP-seq_t6
SRR1029165  H3K9ac_time alignment_ChIP-seq_t7
SRR1029166  H3K9ac_time alignment_ChIP-seq_t8
SRR1029167  H3K9ac_time alignment_ChIP-seq_t9
SRR1029168  H3K9ac_time alignment_ChIP-seq_t10
SRR1029169  H3K9ac_time alignment_ChIP-seq_t11
SRR1029170  H3K9ac_time alignment_ChIP-seq_t12
SRR1029171  H3K9ac_time alignment_ChIP-seq_t13
SRR1029172  H3K9ac_time alignment_ChIP-seq_t14
SRR1029173  H3K9ac_time alignment_ChIP-seq_t15
SRR1029174  H3K9ac_time alignment_ChIP-seq_t16
SRR1029175  H3K9ac_ChIP-seq_t1
SRR1029176  H3K9ac_ChIP-seq_t2
```

```
SRR1029177  H3K9ac_ChIP-seq_t3
SRR1029178  H3K9ac_ChIP-seq_t4
SRR1029179  H3K9ac_ChIP-seq_t5
SRR1029180  H3K9ac_ChIP-seq_t6
SRR1029181  H3K9ac_ChIP-seq_t7
SRR1029182  H3K9ac_ChIP-seq_t8
SRR1029183  H3K9ac_ChIP-seq_t9
SRR1029184  H3K9ac_ChIP-seq_t10
SRR1029185  H3K9ac_ChIP-seq_t11
SRR1029186  H3K9ac_ChIP-seq_t12
SRR1029187  H3K9ac_ChIP-seq_t13
SRR1029188  H3K9ac_ChIP-seq_t14
SRR1029189  H3K9ac_ChIP-seq_t15
SRR1029190  H3K9ac_ChIP-seq_t16
SRR1029191  H3K14ac_ChIP-seq_t1
SRR1029192  H3K14ac_ChIP-seq_t2
SRR1029193  H3K14ac_ChIP-seq_t3
SRR1029194  H3K14ac_ChIP-seq_t4
SRR1029195  H3K14ac_ChIP-seq_t5
SRR1029196  H3K14ac_ChIP-seq_t6
SRR1029197  H3K14ac_ChIP-seq_t7
SRR1029198  H3K14ac_ChIP-seq_t8
SRR1029199  H3K14ac_ChIP-seq_t9
SRR1029200  H3K14ac_ChIP-seq_t10
SRR1029201  H3K14ac_ChIP-seq_t11
SRR1029202  H3K14ac_ChIP-seq_t12
SRR1029203  H3K14ac_ChIP-seq_t13
SRR1029204  H3K14ac_ChIP-seq_t14
SRR1029205  H3K14ac_ChIP-seq_t15
SRR1029206  H3K14ac_ChIP-seq_t16
SRR1029207  H3K56ac_ChIP-seq_t1
SRR1029208  H3K56ac_ChIP-seq_t2
SRR1029209  H3K56ac_ChIP-seq_t3
SRR1029210  H3K56ac_ChIP-seq_t4
SRR1029211  H3K56ac_ChIP-seq_t5
```

```
SRR1029212  H3K56ac_ChIP-seq_t6
SRR1029213  H3K56ac_ChIP-seq_t7
SRR1029214  H3K56ac_ChIP-seq_t8
SRR1029215  H3K56ac_ChIP-seq_t9
SRR1029216  H3K56ac_ChIP-seq_t10
SRR1029217  H3K56ac_ChIP-seq_t11
SRR1029218  H3K56ac_ChIP-seq_t12
SRR1029219  H3K56ac_ChIP-seq_t13
SRR1029220  H3K56ac_ChIP-seq_t14
SRR1029221  H3K56ac_ChIP-seq_t15
SRR1029222  H3K56ac_ChIP-seq_t16
SRR1029223  H4K5ac_ChIP-seq_t1
SRR1029224  H4K5ac_ChIP-seq_t2
SRR1029225  H4K5ac_ChIP-seq_t3
SRR1029226  H4K5ac_ChIP-seq_t4
SRR1029227  H4K5ac_ChIP-seq_t5
SRR1029228  H4K5ac_ChIP-seq_t6
SRR1029229  H4K5ac_ChIP-seq_t7
SRR1029230  H4K5ac_ChIP-seq_t8
SRR1029231  H4K5ac_ChIP-seq_t9
SRR1029232  H4K5ac_ChIP-seq_t10
SRR1029233  H4K5ac_ChIP-seq_t11
SRR1029234  H4K5ac_ChIP-seq_t12
SRR1029235  H4K5ac_ChIP-seq_t13
SRR1029236  H4K5ac_ChIP-seq_t14
SRR1029237  H4K5ac_ChIP-seq_t15
SRR1029238  H4K5ac_ChIP-seq_t16
SRR1029239  H3_ChIP-seq_t1
SRR1029240  H3_ChIP-seq_t2
SRR1029241  H3_ChIP-seq_t3
SRR1029242  H3_ChIP-seq_t4
SRR1029243  H3_ChIP-seq_t5
SRR1029244  H3_ChIP-seq_t6
SRR1029245  H3_ChIP-seq_t7
SRR1029246  H3_ChIP-seq_t8
```

```
SRR1029247  H3_ChIP-seq_t9
SRR1029248  H3_ChIP-seq_t10
SRR1029249  H3_ChIP-seq_t11
SRR1029250  H3_ChIP-seq_t12
SRR1029251  H3_ChIP-seq_t13
SRR1029252  H3_ChIP-seq_t14
SRR1029253  H3_ChIP-seq_t15
SRR1029254  H3_ChIP-seq_t16
SRR1029255  H3K4me3_ChIP-seq_t1
SRR1029256  H3K4me3_ChIP-seq_t2
SRR1029257  H3K4me3_ChIP-seq_t3
SRR1029258  H3K4me3_ChIP-seq_t4
SRR1029259  H3K4me3_ChIP-seq_t5
SRR1029260  H3K4me3_ChIP-seq_t6
SRR1029261  H3K4me3_ChIP-seq_t7
SRR1029262  H3K4me3_ChIP-seq_t8
SRR1029263  H3K4me3_ChIP-seq_t9
SRR1029264  H3K4me3_ChIP-seq_t10
SRR1029265  H3K4me3_ChIP-seq_t11
SRR1029266  H3K4me3_ChIP-seq_t12
SRR1029267  H3K4me3_ChIP-seq_t13
SRR1029268  H3K4me3_ChIP-seq_t14
SRR1029269  H3K4me3_ChIP-seq_t15
SRR1029270  H3K4me3_ChIP-seq_t16
SRR1029271  H3K36me3_ChIP-seq_t1
SRR1029272  H3K36me3_ChIP-seq_t2
SRR1029273  H3K36me3_ChIP-seq_t3
SRR1029274  H3K36me3_ChIP-seq_t4
SRR1029275  H3K36me3_ChIP-seq_t5
SRR1029276  H3K36me3_ChIP-seq_t6
SRR1029277  H3K36me3_ChIP-seq_t7
SRR1029278  H3K36me3_ChIP-seq_t8
SRR1029279  H3K36me3_ChIP-seq_t9
SRR1029280  H3K36me3_ChIP-seq_t10
SRR1029281  H3K36me3_ChIP-seq_t11
```

```
SRR1029282  H3K36me3_ChIP-seq_t12
SRR1029283  H3K36me3_ChIP-seq_t13
SRR1029284  H3K36me3_ChIP-seq_t14
SRR1029285  H3K36me3_ChIP-seq_t15
SRR1029286  H3K36me3_ChIP-seq_t16
SRR1029287  H4K16ac_ChIP-seq_t1
SRR1029288  H4K16ac_ChIP-seq_t2
SRR1029289  H4K16ac_ChIP-seq_t3
SRR1029290  H4K16ac_ChIP-seq_t4
SRR1029291  H4K16ac_ChIP-seq_t5
SRR1029292  H4K16ac_ChIP-seq_t6
SRR1029293  H4K16ac_ChIP-seq_t7
SRR1029294  H4K16ac_ChIP-seq_t8
SRR1029295  H4K16ac_ChIP-seq_t9
SRR1029296  H4K16ac_ChIP-seq_t10
SRR1029297  H4K16ac_ChIP-seq_t11
SRR1029298  H4K16ac_ChIP-seq_t12
SRR1029299  H4K16ac_ChIP-seq_t13
SRR1029300  H4K16ac_ChIP-seq_t14
SRR1029301  H4K16ac_ChIP-seq_t15
SRR1029302  H4K16ac_ChIP-seq_t16
SRR1029303  Gcn5_ChIP-seq_t1
SRR1029304  Gcn5_ChIP-seq_t2
SRR1029305  Gcn5_ChIP-seq_t3
SRR1029306  Gcn5_ChIP-seq_t4
SRR1029307  Gcn5_ChIP-seq_t5
SRR1029308  Gcn5_ChIP-seq_t6
SRR1029309  Gcn5_ChIP-seq_t7
SRR1029310  Gcn5_ChIP-seq_t8
SRR1029311  Gcn5_ChIP-seq_t9
SRR1029312  Gcn5_ChIP-seq_t10
SRR1029313  Gcn5_ChIP-seq_t11
SRR1029314  Gcn5_ChIP-seq_t12
SRR1029315  Gcn5_ChIP-seq_t13
SRR1029316  Gcn5_ChIP-seq_t14
```

```
SRR1029317  Set1_ChIP-seq_t1
SRR1029318  Set1_ChIP-seq_t2
SRR1029319  Set1_ChIP-seq_t3
SRR1029320  Set1_ChIP-seq_t4
SRR1029321  Set1_ChIP-seq_t5
SRR1029322  Set1_ChIP-seq_t6
SRR1029323  Set1_ChIP-seq_t7
SRR1029324  Set1_ChIP-seq_t8
SRR1029325  Set1_ChIP-seq_t9
SRR1029326  Set1_ChIP-seq_t10
SRR1029327  Set1_ChIP-seq_t11
SRR1029328  Set1_ChIP-seq_t12
SRR1029329  Set1_ChIP-seq_t13
SRR1029330  Set1_ChIP-seq_t14
SRR1029331  Esa1_ChIP-seq_t1
SRR1029332  Esa1_ChIP-seq_t2
SRR1029333  Esa1_ChIP-seq_t3
SRR1029334  Esa1_ChIP-seq_t4
SRR1029335  Esa1_ChIP-seq_t5
SRR1029336  Esa1_ChIP-seq_t6
SRR1029337  Esa1_ChIP-seq_t7
SRR1029338  Esa1_ChIP-seq_t8
SRR1029339  Esa1_ChIP-seq_t9
SRR1029340  Esa1_ChIP-seq_t10
SRR1029341  Esa1_ChIP-seq_t11
SRR1029342  Esa1_ChIP-seq_t12
SRR1029343  Esa1_ChIP-seq_t13
SRR1029344  Esa1_ChIP-seq_t14
```

Listing A1: SRRindex file for downloading the data from the NCBI.

```
RNA-seq_t1.fastq
RNA-seq_t2.fastq
RNA-seq_t3.fastq
RNA-seq_t4.fastq
RNA-seq_t5.fastq
```

```
RNA-seq_t6.fastq
RNA-seq_t7.fastq
RNA-seq_t8.fastq
RNA-seq_t9.fastq
RNA-seq_t10.fastq
RNA-seq_t11.fastq
RNA-seq_t12.fastq
RNA-seq_t13.fastq
RNA-seq_t14.fastq
RNA-seq_t15.fastq
RNA-seq_t16.fastq
```

Listing A2: rnaseqfiless.txt file used for mapping of the RNA-seq data.

List of Figures

List of Tables