Aalto University
School of Science
Degree Programme in Computational and Systems Biology

Gökcen Eraslan

A Dirichlet-Multinomial Mixture Model For Clustering Heterogeneous Epigenomics Data

Master's Thesis
Espoo, September 15, 2014

Supervisors:	Assistant Professor Harri Lähdesmäki, Aalto University
	Professor Jens Lagergren, KTH
Advisor:	Maria Osmala, M.Sc. (Tech)

Aalto University
School of Science
Degree Programme in Computational and Systems Biology

ABSTRACT OF MASTER'S THESIS

Author:	Gökcen Eraslan		
Title: A Dirichlet-Multinomial Mixture Model For Clustering Heterogeneous Epigenomics Data			
Date:	September 15, 2014	**Pages:**	vi + 72
Major:	Computational Systems Biology	**Code:**	T-61
Supervisors:	Assistant Professor Harri Lähdesmäki Professor Jens Lagergren		
Advisor:	Maria Osmala, M.Sc. (Tech)		

Epigenetic information sheds light on essential biological mechanisms including the regulation of gene expression. Among the major epigenetic mechanisms are histone tail modifications which can be utilized to identify cis-regulatory elements such as promoters and enhancers. Nucleosome positions and open chromatin regions are other key elements of the epigenomic landscape.

Thanks to the advances in high-throughput sequencing technologies, comprehensive genome-wide analyses of epigenetic signatures are possible at present. Despite the growing number of epigenetic datasets, the tools to discover novel patterns and combinatorial presence of epigenetic elements are still needed. In this thesis, we introduce a model-based clustering approach that uncovers epigenetic patterns by integrating multiple data tracks in a multi-view fashion where different views correspond to different epigenetic signals extracted from the same genomic location. Moreover, to address the inaccuracy of the positions of anchor points, such as TF ChIP-seq peak summits or TSS, a profile shifting feature is implemented. Finally, owing to the hyperprior regularization, our approach can also account for the correlation between the number of reads mapped to consecutive base pair positions.

We demonstrate that the genome-wide clustering of promoter and enhancer regions in human genome reveals distinct patterns in various histone modification and transcription factor ChIP-seq profiles. Furthermore, TFBS enrichment in different classes of enhancers and promoters that are identified by our method is investigated which shows that some transcription factors are significantly enriched in a subset of enhancer and promoter clusters.

Keywords:	chromatin, enhancers, promoters, multi-view clustering, histone modifications, epigenomics, generative models, Dirichlet-multinomial
Language:	English

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Harri Lähdesmäki. With his endless support and friendly guidance, he radically reshaped my understanding of mentoring.

I owe a great debt of gratitude to my advisor Maria Osmala. She has been very helpful and understanding throughout my studies and answered all my question patiently.

Besides my supervisor and advisor, my thanks also goes to all my friends in Aalto Computational Systems Biology research group. Discussions during the group meetings were truly influential and encouraging.

I wish to thank the program coordinators of Aalto University and KTH for organizing the euSYSBIO program and hence giving me the opportunity to have this wonderful experience on diverse applications of computational biology and machine learning.

Espoo, September 15, 2014

Gökcen Eraslan

Abbreviations and Acronyms

AUC	Area under curve
BFSG	Broyden-Fletcher-Goldfarb-Shanno
BIC	Bayesian Information Criterion
ChIP-seq	Chromatin immunoprecipitation sequencing
DNase-seq	DNase I hypersensitive sites sequencing
EM	Expectation-maximization
ENCODE	Encyclopedia of DNA Elements
eRNA	Enhancer RNA
MAP	Maximum a posteriori
MLE	Maximum likelihood estimation
MNase-seq	Micrococcal nuclease sequencing
ROC	Receiver operating characteristic
TFBS	Transcription factor binding site
TSS	Transcription start site
TTS	Transcription termination site

Contents

Chapter 1

Introduction

Each human cell has about 1.8 meters of DNA. In order for this long DNA to fit into the cell nucleus, utilization of an efficient method for compaction is inevitable. In the nucleus, compaction is achieved at different levels. At the simplest level, the protein complexes called *histones* act as key elements. These complexes are composed of eight proteins known as *core histones* and act as a spool around which the DNA strands are wound. Winding of DNA around histones causes it to be packaged into a much smaller volume. Along with approximately 147 base pair long segment of wound DNA, core histones are called *nucleosomes* which are the fundamental repeating structures of *chromatin*. It is known that each chromosome in humans contains millions of base pairs, therefore there are thousands of nucleosomes in every chromosome forming a *beads-on-a-string* structure [1]. Multiple nucleosomes can further arrange more compact forms by wrapping into structures called *30nm fibers*. Finally, at the highest level of compaction, these fibers lead to even denser structures known as *chromosomes* during the cell division. These levels are illustrated in Figure 1.1.

One major property of histones is their long *tails* on the N-terminal end of histone amino acid chain. These tails are the main factors involved in post-translational histone modifications caused by chromatin modifying enzymes. Histone modifications, which are discussed in Chapter 2, can cause chromatin structure to be loosened or tightened by altering the electrostatic attraction between positively charged histone and negatively charged DNA backbone or alternatively by recruiting other proteins modifying the chromatin structure. Loosened chromatin structure makes the DNA regions wrapped around histones accessible so that the DNA-binding proteins can attach to open DNA. This dynamic nature of DNA compaction is the essential mechanism behind the regulation of gene expression and thus of great importance.

Following the sequencing of the entire human genome, approximately 3.2 billion base pairs long DNA sequence is regarded as the essential source of information in genomics studies. This information made it possible for scientists to annotate motifs, transcription factor binding sites (TFBSs), protein-coding genes and regulatory elements. Nevertheless, the biological mechanisms underlying complex phenomena such as the multiplicity of cell types originating from the same genetic sequence cannot be explained solely by sequence-based analyses. In this respect, developing

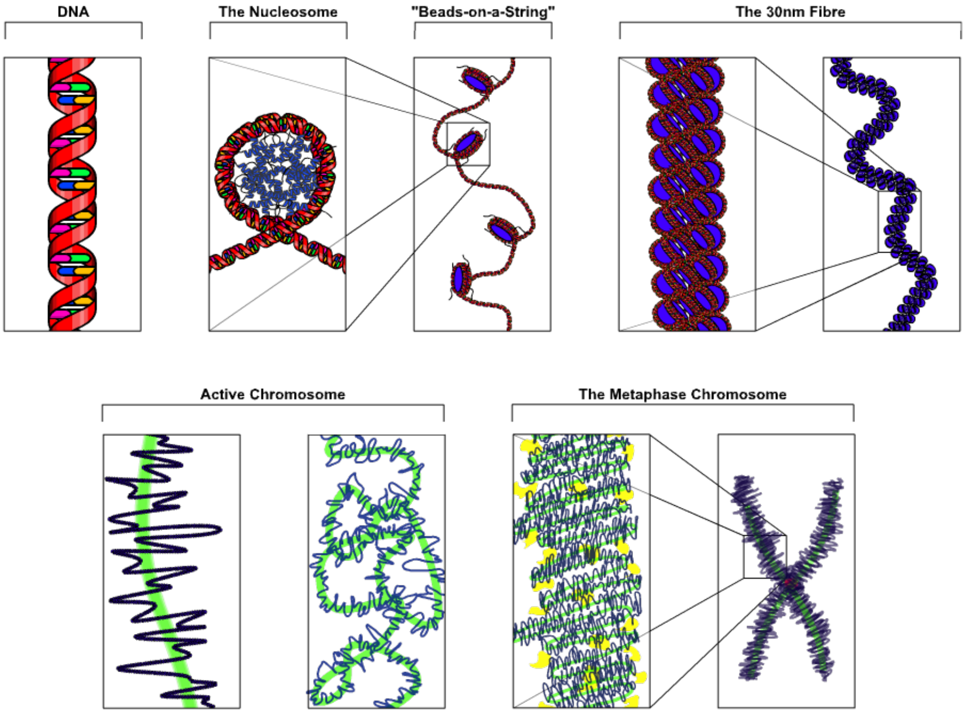


Figure 1.1: Levels of DNA compaction. Adapted from [43]

an understanding of the additional layer of information regarding the regulation of gene expression, called epigenetics, is of significant value.

Epigenetics describes the mechanisms which lead to changes in the regulation of gene expression and activity without altering the sequence of genome. Examples of such epigenetic mechanisms utilized in this study are histone modifications and nucleosome occupancies.

With the advent of next-generation techniques, the epigenetic enrichment data can be obtained from the genome through various strategies coupling high throughput sequencing with experimental techniques, most notably chromatin immunoprecipitation followed by sequencing (ChIP-seq) [25, 32], DNase-I hypersensitivity followed by sequencing (DNase-seq) [4] and micrococcal nuclease followed by sequencing (MNase-seq) [36]. These methods are discussed in Chapter 2.

Epigenetic patterns are utilized as indicators to identify cis-regulatory elements such as promoters and enhancers [12]. Discrimination of different classes of functional elements on the basis of epigenomics data is another challenge which was addressed in previous studies using various clustering approaches. The clustering of the enhancer and promoter regions on the basis of histone modification patterns was first performed by Heintzman et al. through a simple k-means approach [12]. ChromaSig [14] is a heuristic clustering method to discover frequently occurring histone modification patterns genome-wide without using any annotations. It has been shown that among the identified patterns are known signatures associated with promoters and enhancers, as well as the patterns yet to be linked to any functional element. Another clustering method for epigenomics analysis, CATCHprofiles, is proposed by Nielsen et al [30]. In this approach, the most similar profile pairs are merged and aligned to remaining profiles iteratively to determine the topology of the cluster hierarchy. Similar to ChromaSig, CATCHprofiles also does not use any annotation information thus the entire genome is examined to identify chromatin signatures. Kundaje et al. [21] proposed another hierarchical clustering-based approach called CAGT which can group chromatin profiles around functional elements into clusters using k-medians algorithm. This procedure is followed by merging redundant clusters through the hierarchical agglomerative clustering. Nair et al. [29] introduced a mixture model-based clustering method where the ChIP-seq data is binned and the number of reads fall into each bin are modeled using independent Poisson distributions. Similar to our approach, this method uses the EM algorithm to estimate the parameters of the distributions in the mixture and posterior membership probabilities.

Available methods have severe limitations in their ability to exploit the intrinsic structure of epigenetic signals. For instance, clustering results of distance-based approaches are highly sensitive to the choice of distance (or similarity) metric. Hierarchical clustering, a commonly used technique in available approaches, requires clusters to be subjectively determined and the interpretation of the hierarchy is problematic. Furthermore, an unbiased and principled method for determining the optimal number of clusters is lacking in available methods. Last but not least, current approaches do not provide rigorous methods for handling multiple data types meaning that the methods usually ignore the fact that different data types contribute to the clustering in different ways. Most commonly, ad-hoc approaches are preferred such as concatenating the data vectors which may lead to incorrect results, especially in cases where the signal magnitudes of different data types are varying.

To address given shortcomings of available approaches, we introduce a model-based clustering method which exploits the discrete, sparse and non-negative nature of epigenomic data and integrates multiple data tracks to account for the combinatorial presence of different epigenetic patterns.The probabilistic approach presented in this thesis is based on the hierarchical Bayesian model previously proposed by Holmes et al. [13] where the data is modeled using a mixture of Dirichlet-multinomial compound distributions. Modeling through Dirichlet-multinomial compound mixture yields a powerful means to capture the magnitude and shape of discrete data as well as the variation in clusters. Furthermore, our approach extends the original model by treating various epigenetic signals extracted from the same genomic locus as *multiple views* of the locus of interest, so that each cluster of the mixture exhibits a set of Dirichlet-multinomial compound distributions, whose elements correspond to the *views* of the data. This leads to $K \times M$ many independent Dirichlet-multinomial compounds for a K-cluster model fitted to the data with M different data tracks. A rigorous model selection-based method to determine the number of clusters is also presented in our study. Owing to the regularization of the hyperprior, our approach can also account for the correlation between the number of reads mapped to consecutive base pair positions. Moreover, we implemented a profile shifting technique to address the inaccuracy of anchor points positions such as TF ChIP-seq peak summits or TSS.

We demonstrate clustering of enhancer and promoter regions using multiple data tracks from the ENCODE dataset of cell type K562 including ChIP-seq profiles of various histone modifications and TFs as well as the DNase-seq profiles. Additionally, significant enrichment of TFBSs in identified clusters are reported in this study.

Chapter 2

Background

2.1 Enhancers and promoters

Transcription factors and other regulatory molecules bind to typically non-coding DNA regions called *cis-regulatory elements*. Among the mostly studied cis-regulatory elements, *promoters* are located on the DNA upstream to the transcription start sites of genes and act as a binding platform for the transcription machinery including the essential enzyme of transcription, RNA polymerase. However, the involvement of promoters alone results transcription at the basal level. Another cis-regulatory elements called *enhancers* are required to increase transcription rate by catalyzing the chemical reaction. Even though the enhancer sequences may be kilobases away from the gene they affect, they can physically interact with promoters. The widely accepted model elucidating enhancer-promoter interaction is DNA-looping model [33]. According to this model, promoter and enhancer sites are brought into direct contact through the establishment of a loop structure. This phenomenon is illustrated in Figure 2.1.

Given the critical roles of promoters and enhancers in the regulation of gene expression as well as in the cell development and differentiation, these elements are subject to deep interest in scientific literature. In this thesis, we focus on identifying distinct classes of regulatory elements based on the their epigenetic signatures.

2.2 Histone modifications

Histones are the principal protein components of chromatin. Two categories of histones exist: core histones and linker histones. Two copies of each core histone, namely H2A, H2B, H3 and H4, are found at the center of nucleosomes. An essential component of core histones is the histone tail protruding from the nucleosome. Histone tails provide sites for covalent modifications which in turn can alter the chromatin structure or lead to the recruitment of nuclear proteins. Furthermore, histone modifications, chromatin structure and gene expression levels are shown to be strongly correlated [18]. Linker histones, H1, are bound to the outside of nucleosome so that the DNA wrapped around nucleosomes is kept in place and also the

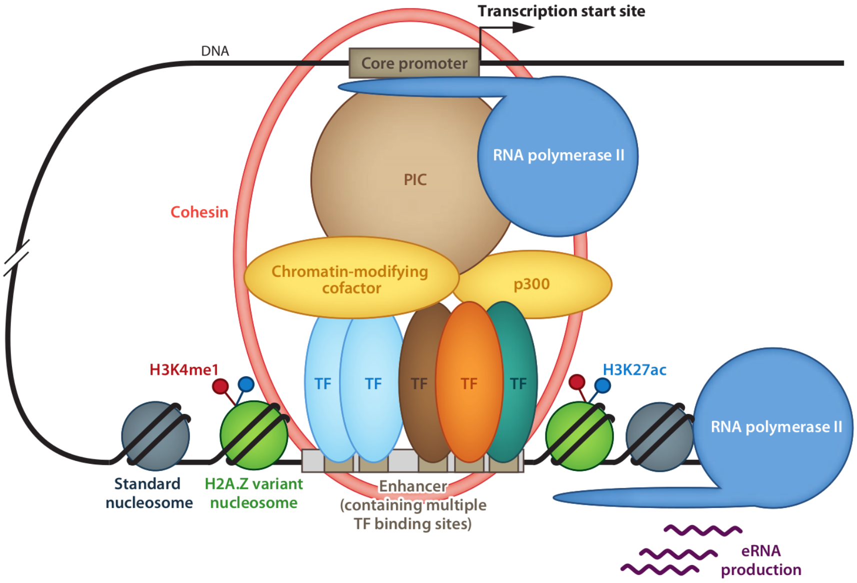


Figure 2.1: Interaction between enhancer and promoter. Adapted from [26]

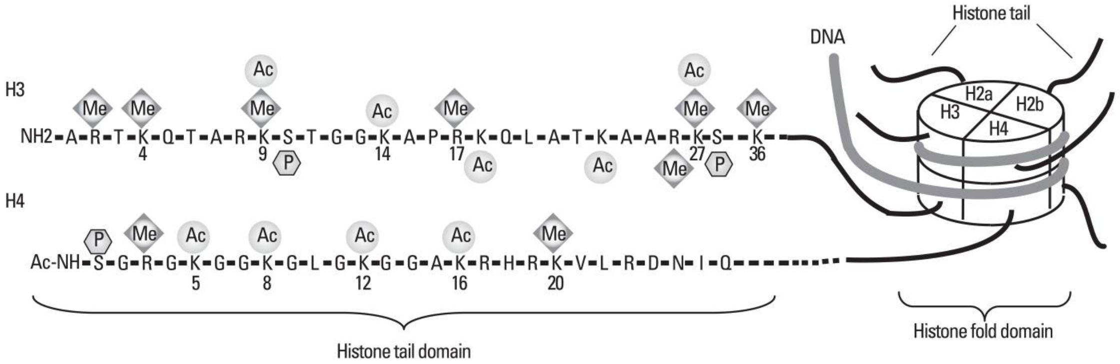

H3
NH2-A-R-T-K-Q-T-A-R-K-S-T-G-G-K-A-P-R-K-Q-L-A-T-K-A-A-R-K-S--K--
4
9
14
17
27
36
Me
Me
Ac
Me
Ac
Me
Ac
Me
Me
P
Ac
Ac
Me
P
H4
Ac-NH-S-G-R-G-K-G-G-K-G-L-G-K-G-G-A-K-R-H-R-K-V-L-R-D-N-I-Q--------
5
8
12
16
20
P
Me
Ac
Ac
Ac
Ac
Me
Histone tail domain
DNA
Histone tail
H2a
H2b
H3
H4
Histone fold domain

Figure 2.2: H3 and H4 core histone tails decorated with histone modifications [20].

structure of chromatin fibers are stabilized.

Although there are several different types of modifications including but not limited to methylation, acetylation, phosphorylation, ubiquitination and citrullination, we focus on acetylation and methylation modifications and utilize profiles of these modifications in our study. Acetylation of lysine residues reduces the attraction between histone and wrapped DNA by neutralizing their positive charge. On the other hand, methylation of lysine residues does not alter the charge of histone significantly but can be recognized by proteins that are quite sensitive to lysine methylation and are capable of changing chromatin structure. Some modifications occurring in H3 and H4 tails are given in Figure 2.2.

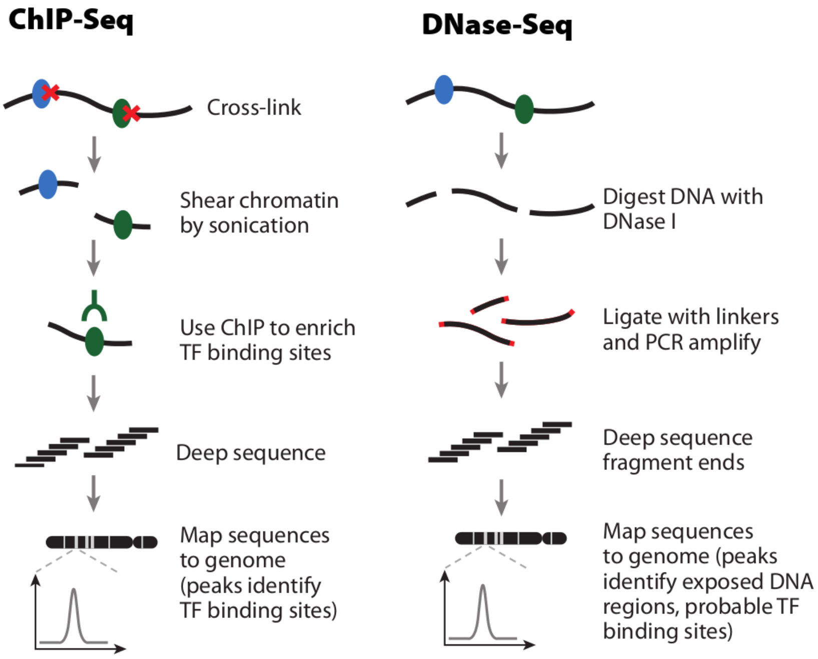
The common nomenclature of histone modifications is composed of four main parts, namely "histone name", "amino acid abbreviation and position", "type of modification" and "number of modifications", e.g. H3K79me2 corresponds to dimethyla-

Core Histone	Residue	Site/ Variant	Localization	Transcriptional Function
H3	K4	me1	Active/poised enhancers	Activation
		me2	Promoters/ enhancers	
		me3	Active/poised promoters	
	K9	ac	Promoters/ enhancers/ coding regions	Activation
		me1 me3	Poised enhancers	Repression
	K27	ac	Active enhancers	Initiation
		me3	Poised enhancers/ Poised promoters	Repression
	K36	me3	Gene bodies	Activation, elongation
	K79	me2	-	Activation
H4	K20	me1	Promoter/coding regions	Activation
H2A	-	Z	(Along with H3.3) Promoters and Enhancers	Regulation, DNA damage, chromosome stability

Table 2.1: Functions of various histone variant and modifications used in this study[3, 8, 15, 17, 31, 39].

tion of 79th residue (lysine) of H3 core histone.

With the advent of high-throughput sequencing technologies, association of histone tail modifications with regulatory genomic elements have become commonplace. For instance, it was previously reported that H3K4me1 and H3K27ac marks are linked to active enhancers [12]. Furthermore, according to the histone code hypothesis [16, 38], combinations of histone modifications dictate some biological function on the basis of DNA-chromatin interactions. Even though the knowledge of specific functions of these combinations are far from complete, this hypothesis leads to the idea that multiple histone modifications should be assessed together to carry out extensive studies. Histone modifications and variants used in this study as well as their associations with regulatory elements and transcriptional functions are presented in Table 2.1.


ChIP-Seq
Cross-link
Shear chromatin by sonication
Use ChIP to enrich TF binding sites
Deep sequence
Map sequences to genome (peaks identify TF binding sites)
DNase-Seq
Digest DNA with DNase I
Ligate with linkers and PCR amplify
Deep sequence fragment ends
Map sequences to genome (peaks identify exposed DNA regions, probable TF binding sites)

(a) (b)

Figure 2.3: ChIP-seq(a) and DNase-seq(b) protocols. Adapted from [26]

2.3 Experimental techniques

The data used in this study are generated through three different experimental strategies. In this section, these methods are discussed. First of these methods is the ChIP-seq protocol which allows for the detection of DNA-protein interactions genome-wide. In this method DNA-bound protein, e.g. a TF or a histone with a specific modification, is first crosslinked to DNA through formaldehyde, then DNA is fragmented by sonication and DNA-protein complexes are isolated. Next, the antibody specific to the protein of interest is utilized to capture the fragment of DNA crosslinked to the target protein. Proteins are released and DNA fragments are sequenced through massively parallel sequencing techniques and the resulting data are aligned to the reference genome [25, 32]. For each nucleotide position in the genome, the number of reads covering the position in question are computed. This procedure leads to genome-wide profiles which can then be assessed to identify regions where the reads are enriched. When applied to histone proteins with epigenetic marks (such as methylation), this method yields critical information regarding the epigenomic landscape.

In DNase-seq technique, an enzyme called Deoxyribonuclease I (DNase I), which digests nucleosome-depleted DNA, is employed. Akin to ChIP-seq, DNase-digested fragments are then sequenced and aligned to the reference genome. Enrichment in DNase-I hypersensitive sites provides a powerful means to identify open chromatin regions that are accessible to DNA-binding proteins e.g. TFs. The workflows of ChIP-seq and DNase-seq protocols are illustrated in Figure 2.3.

MNase-seq is a method to determine nucleosome occupancy using micrococcal nuclease enzyme. Nucleosomal DNA is protected from MNase enzyme, therefore nucleosomal DNA fragments having length of approximately 147bp can be isolated, sequenced and analyzed. Enriched regions identifies nucleosome-occupied regions.

Chapter 3

Materials

In this thesis, we utilize discrete genome-wide profiles of various histone modification, TF binding and nucleosome positioning data generated through high-throughput sequencing techniques, such as ChIP-seq and DNase-seq. Profiles flanking the genomic positions of functional genomic elements are clustered to identify different classes of elements. Therefore, histone modification profiles of regulatory elements are regarded as their *features* in the clustering algorithm so that elements with similar profiles are grouped into the same cluster.

In order to extract the data to be clustered, two types of information are required: *loci* and *signals*. A *locus* is the specific location on a chromosome, such as the range between 15,358,042nd and 15,358,142nd base pairs on chromosome 21. This location may refer to the nucleotide position of a cis-regulatory element such as an enhancer or an annotated transcription factor binding site. To extract the *signal*, also referred to as the *coverage signal*, first a 2000bp window is centered at the given locus and for every nucleotide position within the window, number of reads that cover the position in question are retained. This leads to a 2000-dimensional vector with integer elements. However, to satisfy the requirements of multinomial sampling, we used a slightly different definition of coverage signal where only the 5' ends of reads are taken into account. This topic is discussed in more detail in Section 4.3.

Considering many loci, these vectors form a matrix whose rows correspond to loci, columns to nucleotide positions and elements to number of reads starting from this position. This process is illustrated in Figure 3.1. Additionally, a form of binning can be applied to reduce the computational burden where the read counts of consecutive nucleotide positions are summed up and represented as a single value. This is also referred to as “50bp resolution” when the read counts of 50 nucleotide positions are represented as one integer.

3.1 Data

ENCODE is an international collaboration aiming to annotate the functional elements in human genome. The project provides a wealth of publicly available data. Among the datasets provided by the ENCODE project, we used ChIP-seq

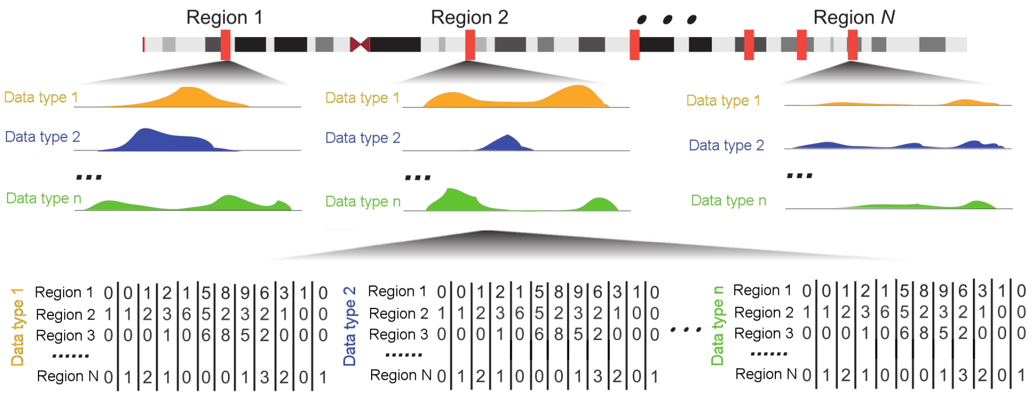

Region 1
Region 2
Region N
Data type 1
Data type 2
Data type n
Data type 1
Region 1 0 0 1 2 1 5 8 9 6 3 1 0
Region 2 1 1 2 3 6 5 2 3 2 1 0 0
Region 3 0 0 0 1 0 6 8 5 2 0 0 0
Region N 0 1 2 1 0 0 0 1 3 2 0 1
Data type 2
Region 1 0 0 1 2 1 5 8 9 6 3 1 0
Region 2 1 1 2 3 6 5 2 3 2 1 0 0
Region 3 0 0 0 1 0 6 8 5 2 0 0 0
Region N 0 1 2 1 0 0 0 1 3 2 0 1
Data type n
Region 1 0 0 1 2 1 5 8 9 6 3 1 0
Region 2 1 1 2 3 6 5 2 3 2 1 0 0
Region 3 0 0 0 1 0 6 8 5 2 0 0 0
Region N 0 1 2 1 0 0 0 1 3 2 0 1

Figure 3.1: An illustration showing how the coverage signals of n different data types are extracted from N genomic regions. Adapted from [40]

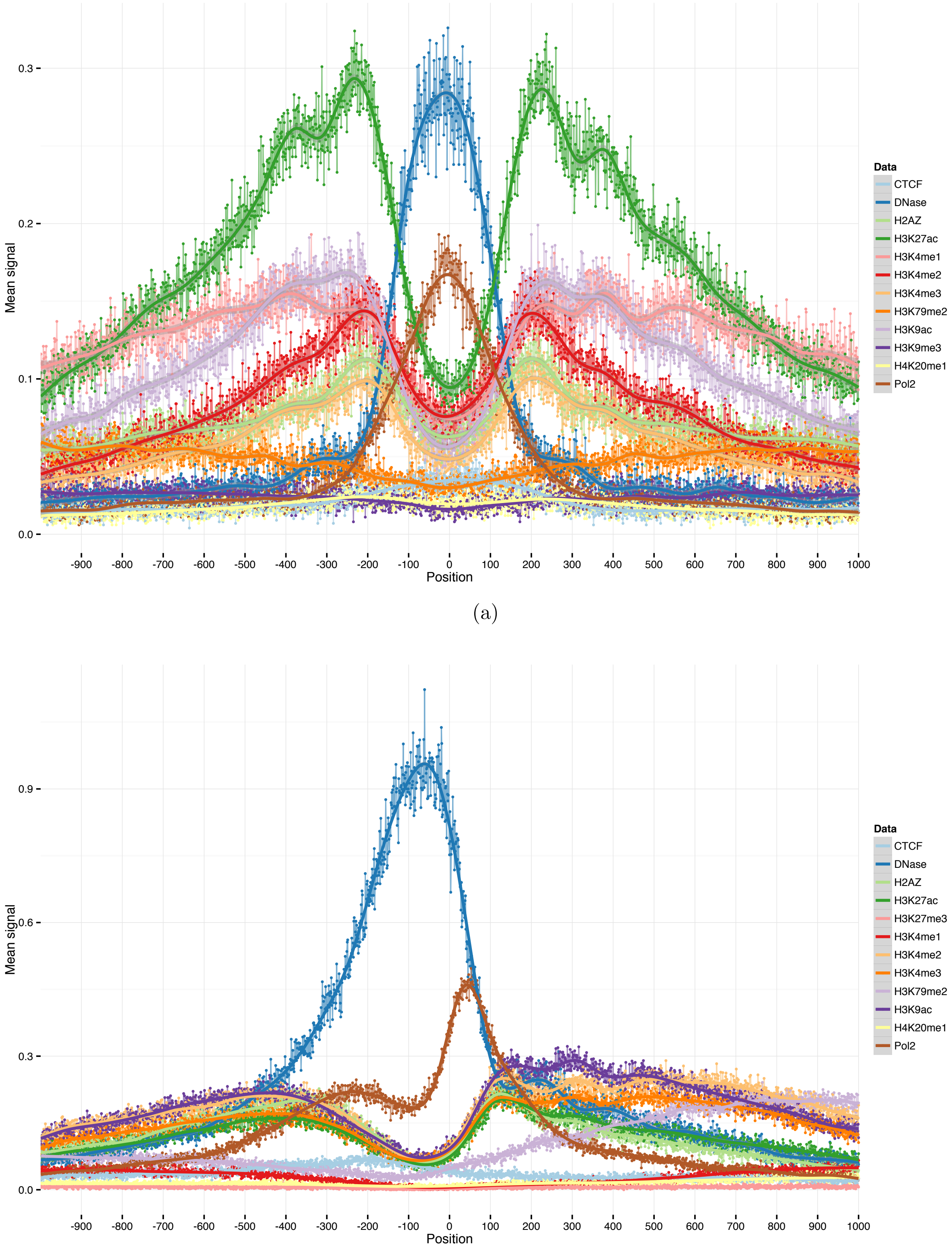
and DNase-seq signals. Details about the data are given in Table 3.1. The data has been downloaded in raw FASTQ format and preprocessed as described in Section 3.3.

3.2 Clustered loci

p300 binding sites generated in Stanford University are downloaded from ENCODE `wgEncodeAwgTfbsUniform` data track whereas DNase-I HS peaks generated in Duke University are downloaded from ENCODE data track `wgEncodeOpenChromDnase`. p300 binding sites which colocalize with DNase-I HS peaks are identified. Next, GENCODE v17 gene list are utilized to filter out identified p300 binding sites within 2000bp of a TSS. Furthermore, to eliminate lower quality p300 peaks, only the first thousand p300 peaks with the highest signal value[1] are included in the analysis.

For the clustering of promoter regions, first TSSs of protein-coding genes have been downloaded from the Ensembl database through `biomaRt` package [7] of the R programming language [35]. Then, the DNase hypersensitive sites are identified. The mean signals of the data extracted from promoter and enhancer regions are shown in Figure 3.2. Moreover, in Appendix D, mean signals at enhancer and promoter regions are plotted separately for different data types for clarity. Note that, throughout the thesis, in all plots where the mean signals are shown (i.e. aggregation plots) the mean signals are represented as small points, and the smoothers are plotted only to show the trend in the data. A combination of generalized linear models and additive models called *generalized additive models* (GAM) [47, 48], is used to plot smoothers through the implementation in `MGCV` [46] and `ggplot2` R packages [42].

[1] In the p300 peak list of ENCODE project, peaks with the highest signal value are also the most significant ones on the basis of q-values.



(a)

(b)

Figure 3.2: Mean profiles of reads within 2000bp windows centered at 1000 enhancer(a) and promoter(b) regions. Also note that the directionality is taken into account while extracting the promoter signal, meaning that the signal from the negative strand is reversed.

Experiment	Targeted Protein	Track	Input	Lab	Cell Type
ChIP-seq	H3K4me1 H3K4me2 H3K4me3 H2A.Z H3K27ac H3K27me3 H3K9ac H3K9me1 H3K9me3 H3K36me3 H4K20me1 H3K79me2 CTCF Pol2	`wgEncodeBroadHistone`	Yes	Broad Institute	K562
DNase-seq	-	`wgEncodeOpenChromDnase`	No	Duke	

Table 3.1: Details of ENCODE signal tracks used in this study

3.3 Data preprocessing

Preprocessing steps are as follows:

1. Downloaded raw FASTQ files are uncompressed and separate replicates are pooled through concatenation.

2. Reads are aligned to the reference genome (hg19 assembly) by Bowtie 0.12.7 read aligner [22]. "-m 1" option is used to retain only uniquely mappable reads.

3. SAM files produced by Bowtie are converted to sorted BAM files using samtools [23].

4. Polyclonal reads in BAM files are removed by "samtools rmdup" command to avoid potential PCR duplicates.

5. Resulting BAM files are converted to BED files using bedtools [34].

6. (Only for the ChIP-seq data) SPP is a ChIP-seq processing and peak calling tool [19] that can find the fragment length d (also referred to as peak separation distance) based on the peak of the cross-correlation between signals of the two strands. Using this feature of SPP (version 1.10), the correct amount of shifting is detected and reads from both strands are shifted towards the 3' end by the half of detected shifting distance, $\frac{d}{2}$, so that the signals from different strands colocalize. This process is shown in Figure 3.3.

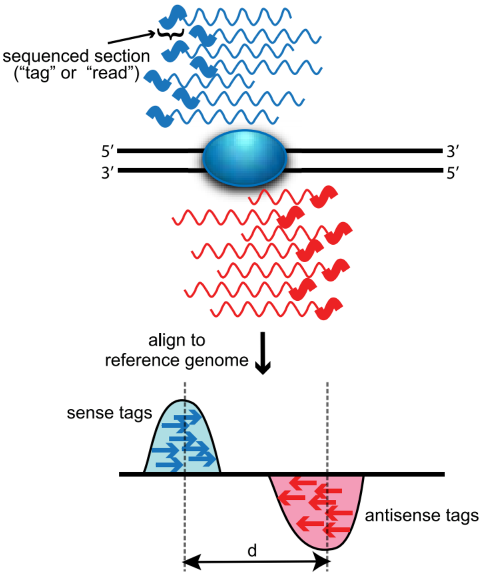


Figure 3.3: For the data from different strands to colocalize, reads from both strands must be shifted by the distance $\frac{d}{2}$ in 5' to 3' direction. Adapted from [45].

7. For the data types for which the control data is available, a normalization procedure is applied based on the following formula

$$\text{round}\left(S - C \cdot \frac{N_S}{N_C}\right)$$

where S and C represent the signal being normalized and the control data, whereas N_S and N_C denote total number of reads in the signal and in the control data, respectively. The control signal is first multiplied by the ratio of total number of reads in the signal to total number of reads in the control signal. Then normalized control signal is subtracted from the signal and the resulting values are rounded to integers.

8. p300 binding sites and DNase-I HS peaks provided by ENCODE in narrowPeak format are downloaded. Binding sites that are at least 2000bp away from GENCODE TSS list and overlapping with DNase-I HS peaks are recorded.

Chapter 4

Methods

4.1 Dirichlet-multinomial compound distribution

Binomial distribution is a discrete probability distribution where the number of successful outcomes of independent success/failure experiments (also called Bernoulli trials) are modeled. Following conditions must be met for modeling using a binomial distribution:

- Each trial must be independent,
- Each outcome must be either a success or a failure,
- The number of trials must be fixed,
- The probability of success must be equal in all trials.

Therefore, there are two parameters: probability of success p and number of trials J. Given these parameters, probability mass function can be given as

$$\text{Binomial}(k; J, p) = \binom{J}{k} p^k (1-p)^{N-k},$$

where k is the number of successful outcomes, p^k represents the probability of getting k independent successful outcomes, $(1-p)^{J-k}$ indicates the probability of getting $J-k$ unsuccessful outcomes whereas $\binom{J}{k}$ term, called the *binomial coefficient,* represents the number of permutations of k successful and $J-k$ unsuccessful outcomes.

Generalization of binomial distribution to more than two categories is called the multinomial distribution where the probabilities and the number outcomes of different categories are represented as vectors rather than scalars. The probability mass function is given below:

$$\begin{aligned}\text{Multinomial}(\vec{\mathbf{X}};\vec{\mathbf{p}},J) &= \binom{J}{X_1,X_2,\ldots,X_S}\prod_{i=1}^{S}p_i^{X_i}\\ &= \frac{J!}{X_1!\,X_2!\,\ldots\,X_S!}\prod_{i=1}^{S}p_i^{X_i}\,,\end{aligned}$$

where S is the number of rival categories, $\vec{\mathbf{X}}$ is an S-dimensional vector representing the number of outcomes for each category, N is the number of total trials, and $\vec{\mathbf{p}}$ is an S-dimensional vector denoting the probabilities for S categories. Akin to the binomial distribution, the first term represents the permutations of the outcome vector $\vec{\mathbf{X}}$ and is called the *multinomial coefficient*. For example, binomial and multinomial distributions can be used to model tossing a coin and rolling a die, respectively.

In Bayesian statistics, a common choice for the prior of multinomial distribution is the Dirichlet distribution which is a multivariate generalization of the beta distribution. For the outcome vector $\vec{\mathbf{X}}$ and parameter vector $\vec{\alpha}$, Dirichlet distribution returns the *probability vector* of events given that each event has occurred $\alpha_i - 1$ times[1]. Therefore, Dirichlet distribution can be defined as the probability distribution over probability vectors whose elements are real numbers in interval $(0,1)$ and sum up to 1.

The space of $S-$dimensional vectors whose elements sum up to a number, such as the support of the Dirichlet distribution with S parameters, defines a $(S-1)$-simplex. For instance, a 3-dimensional vector space can be represented as a $2-$simplex, which is a triangle. Figure 4.1 demonstrates 4 different Dirichlet distributions where x, y and z indicates three elements of $\vec{\alpha}$ parameter vector and the "height" of the bumps represents the density. In this example, it can be observed that the bumps are closer to the events (vertices) that are observed more often. For instance in the first Dirichlet distribution, x is observed $6-1=5$ times whereas the other events are observed $2-1=1$ times, and hence the bulk is closer to the x point.

Probability distribution function of Dirichlet distribution can be given as

$$\text{Dirichlet}(\vec{\mathbf{p}};\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{i=1}^{S}p_i^{\alpha_i-1}$$

where $\vec{\mathbf{p}}$ is a vector of probabilities, $\vec{\alpha}$ is a vector of Dirichlet parameters and the normalizing constant $B(\cdot)$ is the multinomial beta function[2].

Dirichlet-multinomial is a compound probability distribution. In compound probability distributions, the parameters of one distribution are assumed to be distributed according to another distribution. The compound distribution arises

[1] This is an intuitive but rough description because $\vec{\alpha}$ parameters can also take values less than 1.

[2] Multinomial beta function can also be represented in terms of the gamma function: $B(\vec{\alpha}) = \frac{\prod_{j=1}^{S}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{S}\alpha_j)}$

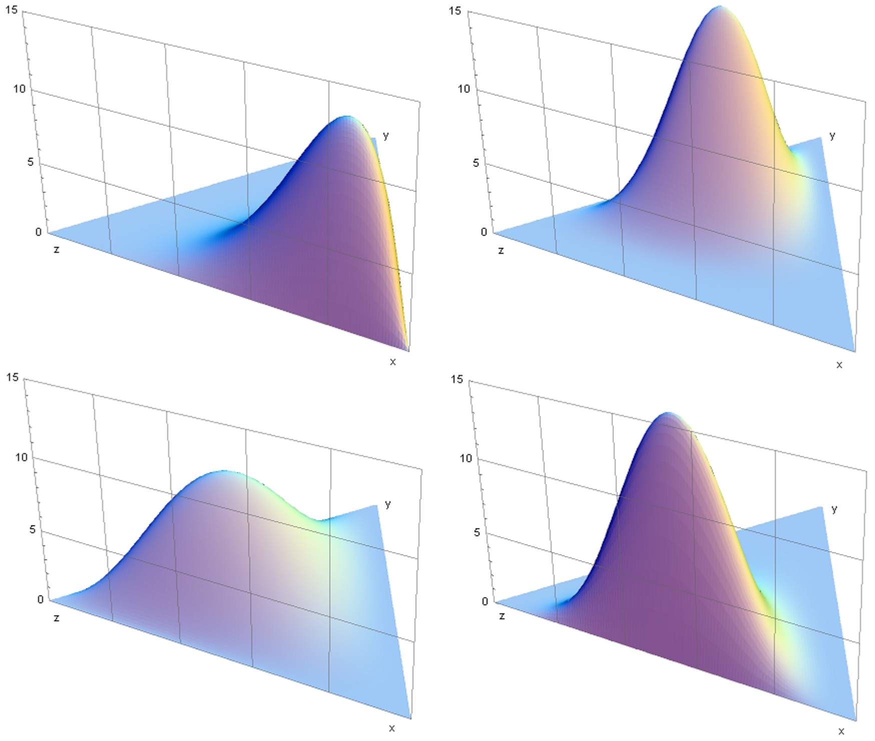


Figure 4.1: Densities of four Dirichlet distributions with parameters $\vec{\alpha} = (6, 2, 2)$, $(3, 7, 5)$, $(6, 2, 6)$, $(2, 3, 4)$ clockwise from top to left [44]. The order of the parameters is (x, y, z).

when the random variable parameters are marginalized out. Here we use Dirichlet-multinomial compound distribution (also called multivariate Pólya distribution) which results from compounding Dirichlet and multinomial distributions by marginalizing multinomial parameters $\vec{\mathbf{p}}$[3]

$$\begin{aligned}\mathrm{DirMult}(\vec{\mathbf{X}};\vec{\alpha}) &= \int_{\vec{\mathbf{p}}} P(\vec{\mathbf{X}}|\vec{\mathbf{p}})P(\vec{\mathbf{p}}|\vec{\alpha})d\vec{\mathbf{p}} \\ &= \frac{J!}{\prod_{i=1}^{S} X_i!}\frac{B(\vec{\mathbf{X}}+\vec{\alpha})}{B(\vec{\alpha})}\,.\end{aligned}$$

where $J = \sum_{i=1}^{S} X_i$. Although modeling integer data with this distribution provides a flexible and powerful means, this comes at a cost. Maximum likelihood estimate of Dirichlet-multinomial compound distribution does not have a closed-form solution; therefore, numerical methods are used to find out ML and MAP estimates. This may cause rather slow implementations if, especially, iterative algorithms such as the EM algorithm are involved.

One important reason to prefer Dirichlet-multinomial compound over multinomial distribution is that former can account for the variability in the data. This can be demonstrated easily by reparamerizing the compound distribution with $\vec{m}\alpha_0 \equiv \vec{\alpha}$ where $\alpha_0 = \sum_{j=1}^{S} \alpha_j$ and $\forall j \in 1, \dots S$, $m_j = \frac{\alpha_j}{\alpha_0}$ so that $\sum_{j=1}^{S} m_j = 1$. In this parameterization, $\vec{m}$ vector can be viewed as the proportion of successful trials for each category, similar to the multinomial parameters sum to one, whereas α_0 gives the overall precision (inverse variance) of the data.

4.2 Mixture models and the EM algorithm

Mixture models result from a weighted sum of multiple distributions where the mixing weights are non-negative and sum up to 1. This form of combination is also referred to as a *convex combination*. Convex combination assures that the mixture of distributions has a valid probability density function. Following is the general formula for mixture models [28]

$$\begin{aligned}p(x_i;\vec{\theta}) &= \sum_{k=1}^{K} \pi_k p(x_i|z_i = k, \vec{\theta}) \\ &= \sum_{k=1}^{K} \pi_k p_k(x|\vec{\theta})\ .\end{aligned}$$

It is assumed that each observed data point x_i has a corresponding unobserved data point z_i which is a discrete latent variable representing which distribution the data point belongs to, namely the *membership* of the data point. The prior of z_i variable is a categorical distribution with parameter vector $\vec{\pi}$, hence $p(z) =$

[3] Derivation of Dirichlet-multinomial density function can be found in Appendix A.3.

$\mathrm{Cat}(\vec{\pi})$. $\vec{\theta}$ represents the parameter vector of individual distributions. Intuitively, the generative process of mixture models can be described in two steps. First, one individual distribution out of K is chosen by sampling from $\mathrm{Cat}(\vec{\pi})$ which leads to z. Then, based on the value of z_i, a sample is drawn from the distribution with parameters matching z, namely $p(\cdot|\theta_{z_i})$.

Mixture models are widely used for clustering where the mixed distributions represent the distributions of individual clusters. In this case, the probability of an observation belonging to a particular cluster is estimated. This probability is called the posterior probability of latent membership variable z, namely $p(z_i = k|x_i, \vec{\theta})$. The posterior, which is also known as the *responsibility*, can be written using Bayes rule

$$\begin{aligned} p(z_i = k|x_i, \vec{\theta}) &= \frac{p(z_i = k)p(x_i|z_i = k, \vec{\theta})}{\sum_{k'=1}^{K} p(z_i = k')p(x_i|z_i = k', \vec{\theta})} \\ &= \frac{\pi_k p(x_i|z_i = k, \vec{\theta})}{\sum_{k'=1}^{K} \pi_{k'} p(x_i|z_i = k', \vec{\theta})} \, . \end{aligned} \tag{4.1}$$

$\vec{\theta}$ and $\vec{\pi}$ parameters of the mixture model can be easily estimated through maximum likelihood estimation (MLE) or maximum a posteriori estimation (MAP), if the values of z are observed. However, z is hidden, z and thus it must be marginalized to achieve the log likelihood to be maximized. For the entire dataset this likelihood is

$$\log p(\vec{\mathbf{X}}; \vec{\theta}) = \log \sum_{\vec{Z}} p(\vec{\mathbf{X}}, \vec{Z}; \vec{\theta})$$

which is often intractable. Expectation-maximization (EM) algorithm [6] provides an iterative algorithm to estimate the parameters of models involving latent variables. The trick that is used in the EM algorithm is to maximize a lower bound on the log-likelihood given above, which is more tractable [27]. Using Jensen's inequality, a lower bound can be obtained by

$$\begin{aligned} \log p(\vec{\mathbf{X}}; \vec{\theta}) &= \log \sum_{\vec{Z}} p(\vec{\mathbf{X}}, \vec{Z}; \vec{\theta}) \\ &= \log \sum_{\vec{Z}} \frac{q(\vec{Z})}{q(\vec{Z})} p(\vec{\mathbf{X}}, \vec{Z}; \vec{\theta}) \\ &= \log \sum_{\vec{Z}} q(\vec{Z}) \frac{p(\vec{\mathbf{X}}, \vec{Z}; \vec{\theta})}{q(\vec{Z})} \\ &\geq \sum_{\vec{Z}} q(\vec{Z}) \log \frac{p(\vec{\mathbf{X}}, \vec{Z}; \vec{\theta})}{q(\vec{Z})} \, , \end{aligned} \tag{4.2}$$

where $q(\vec{Z})$ represents and arbitrary distribution over the hidden variable $\vec{Z}$. In the EM algorithm the posterior of $\vec{Z}$ variable, namely $p(\vec{Z}|\vec{\mathbf{X}}, \vec{\theta})$, is used as $q(\vec{Z})$ to obtain a tight lower bound on complete data likelihood.

EM algorithm consists of two steps. First, $q(\vec{Z})$, which is $p(\vec{Z}|\vec{\mathbf{X}},\vec{\theta})$, is estimated using Equation 4.1 by assuming that the model parameters $\vec{\theta}$ and $\vec{\pi}$ are known. This is known as the *expectation* step. Then, in the second step, know as the *maximization step*, based on the posterior $\vec{Z}$ values computed previously, which can now be denoted as $p(\vec{Z}|\vec{\mathbf{X}},\vec{\theta}^{old})$, the lower bound is maximized and the new model parameters are obtained

$$\begin{aligned}\log p(\vec{\mathbf{X}};\vec{\theta}) &\geq \sum_{\vec{Z}} q(\vec{Z})\log\frac{p(\vec{\mathbf{X}},\vec{Z};\vec{\theta})}{q(\vec{Z})}\\ &= \sum_{\vec{Z}} q(\vec{Z})\log p(\vec{\mathbf{X}},\vec{Z};\vec{\theta}) - q(\vec{Z})\log q(\vec{Z})\\ &= \sum_{\vec{Z}}\left[p(\vec{Z}|\vec{\mathbf{X}},\vec{\theta}^{old})\log p(\vec{\mathbf{X}},\vec{Z};\vec{\theta})\right] + \mathcal{H}(q)\\ &= E_{\vec{Z}}[\log p(\vec{\mathbf{X}},\vec{Z};\vec{\theta})] + \mathcal{H}(q)\,,\end{aligned}$$

where $\mathcal{H}(q) = -\sum_{\vec{Z}} p(\vec{Z}|\vec{\mathbf{X}},\vec{\theta}^{old})\log p(\vec{Z}|\vec{\mathbf{X}},\vec{\theta}^{old})$ is a non-negative entropy term which can be ignored in maximization since it does not depend on $\vec{\theta}$. Therefore

$$\vec{\theta}^{new} = \arg\max_{\vec{\theta}} E_{\vec{Z}}[\log p(\vec{\mathbf{X}},\vec{Z};\vec{\theta})]\,. \tag{4.3}$$

This can also be applied to the MAP estimation

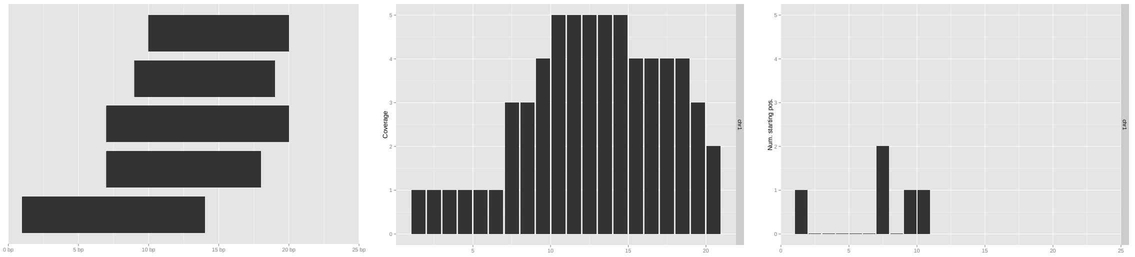
$$\vec{\theta}^{new} = \arg\max_{\vec{\theta}} E_{\vec{Z}}[\log p(\vec{\theta},\vec{Z};\vec{\mathbf{X}})]\,. \tag{4.4}$$

The iterations of these two steps continue until the increase in the likelihood of the data (or the posterior in MAP case) is negligible. When the iterations are over, the posterior probability of memberships, which are also called *soft labels*, can be converted to *hard labels* by using a MAP estimate

$$z_i^* = \arg\max_k p(z_i = k|x_i,\hat{\vec{\theta}})$$

4.3 Model description

In our model, the process of aligning the reads that are generated by a high-throughput sequencing technique to the reference genome is modeled using a multinomial distribution. Let us assume that a 2kb window is centered at a genomic region of interest and for every 2000 positions the number of reads covering these positions are counted. Resulting 2000-dimensional vector can be interpreted as a sample from a multinomial distribution where the number of categories is 2000 and the number of trials is the sum of the counts of 2000-dimensional vector. An analogy



(a) Reads

(b) Coverage signal

(c) Coverage signal with only 5' ends

Figure 4.2: Difference between two forms of coverage signals.

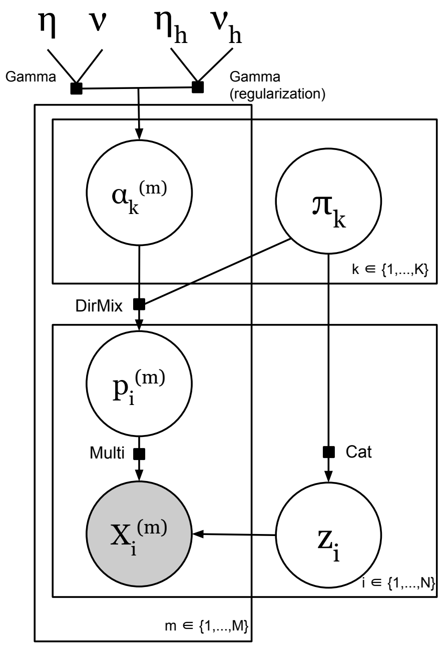
is that aligning a read to the genome is similar to rolling a 2000-sided die. However, as described in Section 4.1, one important requirement for binomial (and thus multinomial) modeling is that every trial must lead to the success of exactly one category. However, if the alignment of a read to the genome is regarded as a multinomial trial and the coverage of whole reads is used in the model, one read affects more than one position which corresponds to one trial leading to the success of multiple categories. Therefore the requirement of multinomial distribution is violated. To fulfill this requirement, only 5' ends of reads are counted in the generation of signals which leads very sparse data matrices. The difference between considering whole reads and 5' ends only are illustrated in Figure 4.2.

Given a window size of 2kb, a single data type such as H3K4me1 and 1000 genomic loci to cluster, entire data can be represented as one 2000 × 1000 matrix where the rows represent loci and columns represent nucleotide positions. However, as discussed in Chapter 3, a lower nucleotide resolution, such as 50bp, is preferred to reduce the computational cost of clustering. In this case, our coverage matrix becomes a 40×1000 matrix since the number of bins is $2000/50 = 40$.

In this thesis, we propose a Dirichlet-multinomial mixture model to cluster genomic loci by exploiting various epigenetic data extracted from the loci being clustered. Dirichlet-multinomial mixture model which is previously applied to the microbial metagenomics data by Holmes et al. [13] is extended in the following ways:

- Multiple data types can be incorporated in a multi-view fashion where different views correspond to different epigenetic signals extracted from the same genomic locus,
- Hyperprior can be regularized to reflect the dependence of consecutive base pair positions in the real data and,
- Profile shifting can account for the uncertainty in the genomic loci being clustered.

The data matrix is represented as $\mathbf{X}$ with elements $\mathbf{X}_{ij}$ denoting the number of input-subtracted reads whose starting position (5' end) is mapped to bin j of locus i. Total



Notation	Description
η	Fixed Gamma shape parameter
ν	Fixed Gamma rate parameter
η_h	Fixed Gamma shape parameter (regularization)
ν_h	Fixed Gamma rate parameter (regularization)
$\boldsymbol{\alpha}$	Dirichlet parameter vector (of length S)
π	Dirichlet mixture weights
$\mathbf{p}$	Multinomial parameter vector
$\mathbf{X}$	Data generated through multinomial sampling
K	Number of mixture components
N	Number of loci
M	Number of data types

Figure 4.3: Model diagram in directed factor graph notation

number of genomic loci are N and the number of bins in a 2kb window is S, e.g. for 2kb window length and 50bp nucleotide resolution, $S = 40$. Therefore $\mathbf{X}$ is an $N \times S$ matrix. Different data types of N regions are represented as $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}$ where M represents the number of data types such as H3K4me1, H3K27ac, etc. Entire dataset consisting of all loci and all data types is denoted as $\mathbf{X}^{(*)}$ whereas all data types of a single locus i is represented as $\vec{X}_i^{(*)}$.

Every row of $\mathbf{X}$ is assumed to be generated from a multinomial distribution with parameter vector $\vec{p_i}$ where the elements p_{ij} represents the probability that 5' end of a read mapped to bin j of locus i. Multinomial parameters $\vec{p_i}$ are sampled from Dirichlet mixtures of size K with parameters $\vec{\alpha}_k$ and mixture weights π_k where K indicates the number of clusters. $\mathbf{Z}$ matrix represents binary memberships. Directed factor graph notation is given in Figure 4.3.

4.3.1 Likelihood

Likelihood for observing one data type (e.g. H3K4me1) of a genomic locus is

$$L_i(\vec{X}_i|\vec{p}_i) = \binom{J_i}{X_{i1}\, X_{i2} \ldots X_{iS}} \prod_{j=1}^{S} p_{ij}^{X_{ij}} \tag{4.5}$$

$$= \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!} \prod_{j=1}^{S} p_{ij}^{X_{ij}} \tag{4.6}$$

where $J_i = \sum_{j=1}^{S} X_{ij}$. Here, we assume that different data types of a locus are independent, hence the likelihood for all data types of a locus i can be written as

$$L_i(\vec{X}_i^{(*)}|\vec{p}_i^{(*)}) = \prod_{m=1}^{M} L_i(\vec{X}_i^{(m)}|\vec{p}_i^{(m)}) \tag{4.7}$$

thus the multinomial likelihood for all loci is

$$L(\mathbf{X}^{(*)}|\mathbf{p}^{(*)}) = \prod_{i=1}^{N}\prod_{m=1}^{M} L_i(\vec{X}_i^{(m)}|\vec{p}_i^{(m)}) \tag{4.8}$$

4.3.2 Prior, posterior and marginal likelihood

Dirichlet prior for multinomial parameters $\vec{p_i}$ is

$$\mathrm{Dir}(\vec{p_i}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{j=1}^{S} p_{ij}^{\alpha_j - 1}\,, \tag{4.9}$$

where $\vec{\alpha}$ vector represents Dirichlet parameters. Assuming that the multinomial parameters, $\vec{p_i}$, are generated by a mixture of Dirichlet distributions, the prior can be expressed as

$$\mathrm{Dir}(\vec{p_i}|\vec{\alpha},\vec{\pi}) = \sum_{k=1}^{K} \mathrm{Dir}(\vec{p_i}|\vec{\alpha}_k)\pi_k\,.$$

For multiple data types (views), priors are multiplied since priors of different data types are assumed to be independent

$$\begin{aligned}\mathrm{Dir}(\vec{p}_i^{(*)}|\vec{\alpha}^{(*)},\vec{\pi}) &= \sum_{k=1}^{K}\pi_k \mathrm{Dir}(\vec{p_i}^{(*)}|\vec{\alpha}_k^{(*)}) \\ &= \sum_{k=1}^{K}\pi_k \prod_{m=1}^{M}\mathrm{Dir}(\vec{p_i}^{(m)}|\vec{\alpha}_k^{(m)})\,,\end{aligned}$$

which leads to the following posterior

$$P(\vec{p}_i^{(*)}|\vec{X}_i^{(*)},\vec{\alpha}^{(*)},\vec{\pi}) = \frac{\sum_{k=1}^{K}\pi_k L(X_i^{(*)}|\vec{p_i}^{(*)})\mathrm{Dir}(\vec{p_i}^{(*)}|\vec{\alpha}_k^{(*)})}{\sum_{k=1}^{K}\pi_k P(X_i^{(*)}|\vec{\alpha}_k^{(*)})}\,. \tag{4.10}$$

The "evidence" or "marginal likelihood" term in the denominator of Equation 4.10 has a key role in the mixture model and can be obtained by integrating out multinomial parameters $\vec{p_i}$ which in turn yields the Dirichlet-multinomial compound distribution

$$
\begin{aligned}
P(\vec{X}_i^{(*)}|\vec{\alpha}_k^{(*)}) &= \prod_{m=1}^{M} \int P(\vec{X}_i^{(m)}|\vec{p}_i^{(m)}) \mathrm{Dir}(\vec{p}_i^{(m)}|\vec{\alpha}_k^{(m)}) d\vec{p}_i^{(m)} && (4.11) \\
&= \prod_{m=1}^{M} \frac{J_i^{(m)}!}{\prod_{j=1}^{S} X_{ij}^{(m)}!} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})} && (4.12)
\end{aligned}
$$

Since the actual denominator of Equation 4.10 is denoted as a mixture of Dirichlet-multinomial compounds, the following gives the marginal likelihood for one locus, i

$$
\sum_{k=1}^{K} P(\vec{X}_i^{(*)}|\vec{\alpha}_k^{(*)})\pi_k = \sum_{k=1}^{K} \left(\pi_k \prod_{m=1}^{M} \frac{J_i^{(m)}!}{\prod_{j=1}^{S} X_{ij}^{(m)}!} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})} \right) \tag{4.13}
$$

which can also be written for all genomic loci as

$$
P(\mathbf{X}^{(*)}|\vec{\alpha}^{(*)}, \vec{\pi}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \left(\pi_k \prod_{m=1}^{M} \frac{J_i^{(m)}!}{\prod_{j=1}^{S} X_{ij}^{(m)}!} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})} \right). \tag{4.14}
$$

Without any hyperpriors, $\vec{\alpha}$ parameters maximizing this expression (Type II MLE) could have been estimated through the expectation-maximization algorithm, however, to provide more control over the mean and the variance of parameters, Gamma hyperpriors are utilized and therefore the MAP estimate is computed.

4.3.3 Hyperprior

Gamma hyperpriors are defined on the Dirichlet parameters i.e. $\alpha_{jk} \sim \Gamma(\eta, \nu)$. Here η and ν are fixed shape and rate parameters of the Gamma distribution. Thus independent and identically distributed Gamma hyperpriors can be written as

$$
\begin{aligned}
p(\vec{\alpha}_1^{(1)}, ..., \vec{\alpha}_K^{(1)}, \vec{\alpha}_1^{(M)} ..., \vec{\alpha}_K^{(M)}) &= \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{j=1}^{S} \mathrm{Gamma}(\alpha_{jk}^{(m)}; \eta, \nu) && (4.15) \\
&= \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{j=1}^{S} \frac{\nu^{\eta} \alpha_{jk}^{(m)\,\eta-1} e^{-\nu\alpha_{jk}^{(m)}}}{\Gamma(\eta)}.
\end{aligned}
$$

However, considering the real data (e.g. ChIP-seq, DNase etc.), it would be more accurate to reflect the correlation between the number of reads mapped to consecutive base pair positions to the model rather than assuming all Dirichlet parameters are independently drawn from the same Gamma distribution. 2000 Dirichlet parameters sampled independently from Gamma distribution with rate and shape parameters 0.1 and 0.1 are plotted in Figure 4.4. It can be clearly observed that there is no

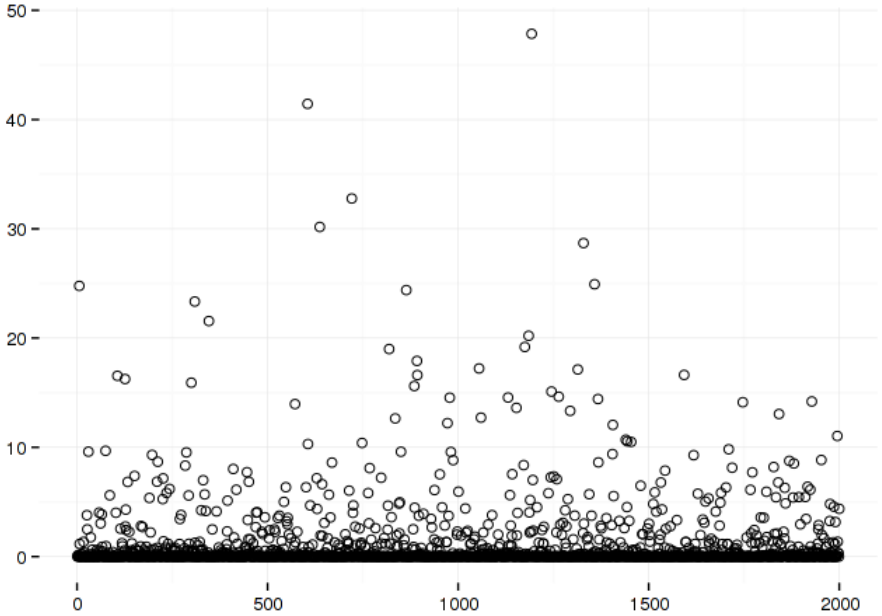


Figure 4.4: 2000 Dirichlet parameters sampled from $Gamma(0.1, 0.1)$.

resemblance between these samples and the mean profiles of the data given in Figure **??** as samples are independently drawn from the same Gamma distribution and hence have no bimodal or unimodal shape commonly observed in the signal of enriched regions.

To reflect this fact to our model, a regularization term is incorporated into Gamma hyperprior to favor smoother Dirichlet parameters over the ones which have higher read count differences between consecutive base pair positions[4]:

$$\begin{aligned} p(\vec{\alpha}_1^{(1)}, ..., \vec{\alpha}_K^{(1)}, \ldots, \vec{\alpha}_1^{(M)} ..., \vec{\alpha}_K^{(M)}) \;\propto\; & \Gamma(\eta_h)^{-MK} \nu_h^{\eta_h MK} \Gamma(\eta)^{-MKS} \nu^{\eta MKS} \\ & \exp\left\{ -\sum_{m=1}^{M} \sum_{k=1}^{K} \left(\nu_h h_k^{(m)} + \sum_{j=1}^{S} \nu \alpha_{jk}^{(m)} \right) \right\} \\ & \prod_{m=1}^{M} \prod_{k=1}^{K} h_k^{(m)\,\eta_h - 1} \prod_{j=1}^{S} \alpha_{jk}^{(m)\,\eta-1}, \end{aligned} \tag{4.16}$$

where $h_k^{(m)} = \sum_{j=2}^{S} (\alpha_{jk}^{(m)} - \alpha_{j-1\,k}^{(m)})^2$ and η_h and ν_h are the fixed parameters of the smoothing Gamma term.

To keep $\alpha_{jk}^{(m)}$ values positive during the optimization step of EM, we use $\vec{\lambda} = \log \vec{\alpha}$ transformation through the multivariate change of variables method where $\vec{\lambda}$ and $\vec{\alpha}$

[4] The derivation is given in Appendix A.1.

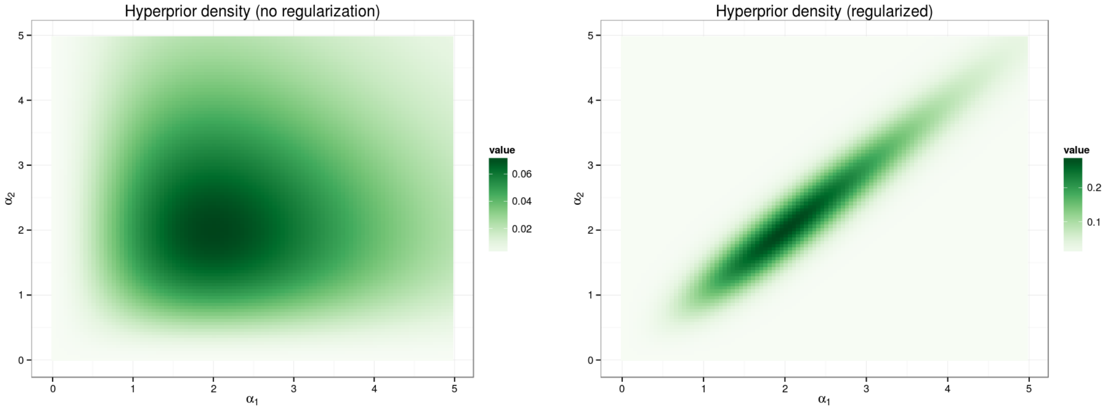


Figure 4.5: The prior densities of two Dirichlet parameters (α_1 and α_2) with and without regularization

vectors represent all Dirichlet parameters across different data types and clusters[5]

$$
\begin{aligned}
p(\vec{\lambda}_1^{(1)}, ..., \vec{\lambda}_K^{(1)}, \dots, \vec{\lambda}_1^{(M)} ..., \vec{\lambda}_K^{(M)}) \propto\; & \Gamma(\eta_h)^{-MK} \nu_h^{\eta_h MK} \Gamma(\eta)^{-MKS} \nu^{\eta MKS} \\
& \exp\left\{ - \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\nu_h h_k^{(m)} + \sum_{j=1}^{S} \nu \alpha_{jk}^{(m)} \right) \right\} \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} h_k^{(m)\,\eta_h - 1} \prod_{j=1}^{S} \alpha_{jk}^{(m)\,\eta},
\end{aligned}
\tag{4.17}
$$

where $\alpha_{jk}^{(m)} = e^{\lambda_{jk}^{(m)}}$.

The effect of the regularization is demonstrated in Figure 4.5. On the left-hand side, prior joint density of two independent Dirichlet parameters is plotted where

$$p(\alpha_1, \alpha_2) = \text{Gamma}(\alpha_1; \eta = 3, \nu = 1) \cdot \text{Gamma}(\alpha_2; \eta = 3, \nu = 1).$$

Second plot on the right-hand side shows the regularized prior density based on Equation 4.16 where parameters are $\eta = 3, \nu = 1, \eta_h = 1, \nu_h = 4$. Since Gamma distribution reduces to an exponential distribution with parameter ν_h, when the shape parameter (η_h) is one, these regularization parameters favor α values with small differences. Therefore, this reduces the differences between consecutive Dirichlet parameters and yields smoother Dirichlet parameters.

4.3.4 Expectation-maximization

MAP estimates of Dirichlet parameters $\vec{\alpha}$ and $\vec{\pi}$ are computed through the EM algorithm

[5] The derivation is given in Appendix A.2.

$$
\begin{aligned}
\hat{Q} &= \arg\max_{Q} P(Q|\mathbf{X}) && (4.18)\\
&= \arg\max_{Q} P(\vec{\alpha}^{(*)}, \vec{\pi}|\mathbf{X}^{(*)})\\
&= \arg\max_{Q} P(\mathbf{X}^{(*)}|\vec{\alpha}^{(*)}, \vec{\pi})P(\vec{\alpha}^{(*)})\\
&= \arg\max_{Q} P(\mathbf{X}^{(*)}|\vec{\alpha}^{(*)}, \vec{\pi}) \prod_{m=1}^{M} P(\vec{\alpha}^{(m)})\\
&= \arg\max_{Q} \prod_{m=1}^{M} P(\mathbf{X}^{(m)}|\vec{\alpha}^{(m)}, \vec{\pi})P(\vec{\alpha}^{(m)})\,, && (4.19)
\end{aligned}
$$

where hyperparameter vector Q is defined as $Q = (\vec{\alpha}^{(1)}, \ldots, \vec{\alpha}^{(m)}, \vec{\pi})$ and it is assumed that $\vec{\pi}$ has a uniform prior. The logarithm of the posterior $P(Q|X)$ can be written as follows

$$
\begin{aligned}
\log P(Q|X) \;\propto\;& \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K}\left(\pi_k \prod_{m=1}^{M} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})}\right)\right] && (4.20)\\
&+\sum_{m=1}^{M}\sum_{k=1}^{K}\left((\eta_h - 1)\log h_k^{(m)} - \nu_h h_k^{(m)} + \sum_{j=1}^{S} \eta \log \alpha_{jk}^{(m)} - \nu \alpha_{jk}^{(m)}\right)\\
&+\text{terms independent of } Q\,. && (4.21)
\end{aligned}
$$

Using Equation 4.14, we can write the marginal likelihood augmented with $\mathbf{Z}$ latent variable as

$$
P(\mathbf{X}, \mathbf{Z}|Q) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left(\pi_k \prod_{m=1}^{M} \frac{J_i^{(m)}!}{\prod_{j=1}^{S} X_{ij}^{(m)}!} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})}\right)^{Z_{ik}} \qquad (4.22)
$$

and the log posterior is

$$
\log P(Q, \mathbf{Z}|\mathbf{X}) \;\propto\; \log P(\mathbf{X}, \mathbf{Z}|Q) + \log P(Q)\,. \qquad (4.23)
$$

Then $\log P(\mathbf{X}, \mathbf{Z}|Q)$ and $\log P(Q)$ can be written using Equations 4.22 and 4.17 respectively

$$\begin{aligned}
\log P(\mathbf{X}, \mathbf{Z}|Q) &\propto \log\left(\prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k \prod_{m=1}^{M} \frac{B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})})\right]^{Z_{ik}}\right) \qquad (4.24)\\
&\propto \sum_{i=1}^{N}\sum_{k=1}^{K}\left\{z_{ik}\left[\log\pi_k + \sum_{m=1}^{M}\left(\log B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)}) - \log B(\vec{\alpha}_k^{(m)})\right)\right]\right\}\\
\log\, P(Q) &= MK(\eta S\log\nu - S\log\Gamma(\eta) + \eta_h\log\nu_h - \log\Gamma(\eta_h)) \qquad (4.25)\\
&+\sum_{m=1}^{M}\sum_{k=1}^{K}\left((\eta_h - 1)\log h_k^{(m)} - \nu_h h_k^{(m)} + \sum_{j=1}^{S}\eta\log\alpha_{jk}^{(m)} - \nu\alpha_{jk}^{(m)}\right)
\end{aligned}$$

As shown in Equation 4.4, the expected log posterior, which is the lower bound to $\log P(Q|\mathbf{X})$, is maximized

$$\begin{aligned}
\log P(Q|\mathbf{X}) &\geq E_{\mathbf{Z}}[\log P(Q, \mathbf{Z}|\mathbf{X})] + \mathcal{H} \qquad (4.26)\\
E_{\mathbf{Z}}[\log P(Q, \mathbf{Z}|\mathbf{X})] &= \sum_{i=1}^{N}\sum_{k=1}^{K}\left\{E[z_{ik}]\left[\log\pi_k + \sum_{m=1}^{M}\left(\log B(\vec{X}_i^{(m)} + \vec{\alpha}_k^{(m)}) - \log B(\vec{\alpha}_k^{(m)})\right)\right]\right\}\\
&+\sum_{m=1}^{M}\sum_{k=1}^{K}\left((\eta_h - 1)\log h_k^{(m)} - \nu_h h_k^{(m)} + \sum_{j=1}^{S}\eta\log\alpha_{jk}^{(m)} - \nu\alpha_{jk}^{(m)}\right)\\
&+\text{terms independent of } Q. \qquad (4.27)
\end{aligned}$$

Here, $E[z_{ik}]$ denotes the membership probabilities (responsibilities) which is given below

$$\begin{aligned}
E[z_{ik}] &= P(z_{ik} = 1|\vec{X}_i^{(*)}) \qquad (4.28)\\
&= \frac{P(z_{ik} = 1)\prod_{m=1}^{M} P(\vec{X}_i^{(m)}|z_{ik} = 1)}{\sum_{k'} P(z_{ik'} = 1)\prod_{m=1}^{M} P(\vec{X}_i^{(m)}|z_{ik'} = 1)}\\
&= \frac{P(z_{ik} = 1)\prod_{m=1}^{M} P(\vec{X}_i^{(m)}|\vec{\alpha}_k^{(m)})}{\sum_{k'} P(z_{ik'} = 1)\prod_{m=1}^{M} P(\vec{X}_i^{(m)}|\vec{\alpha}_{k'}^{(m)})} \qquad (4.29)\\
&= \frac{\pi_k\prod_{m=1}^{M}\frac{B(\vec{X}_i^{(m)}+\vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})}}{\sum_{k'}\pi_{k'}\prod_{m=1}^{M}\frac{B(\vec{X}_i^{(m)}+\vec{\alpha}_{k'}^{(m)})}{B(\vec{\alpha}_{k'}^{(m)})}}, \qquad (4.30)
\end{aligned}$$

where $P(\vec{X}_i^{(m)}|z_{ik} = 1) = P(\vec{X}_i^{(m)}|\vec{\alpha}_k^{(m)})$. Note that, this has the same form as the membership probability equation given in Equation 4.1 in Section 4.2.

In the maximization step of the EM algorithm $E_{\mathbf{Z}}[\log P(Q, \mathbf{Z}|\mathbf{X})]$ is maximized w.r.t. π_k and $\alpha_{jk}^{(m)}$. Maximization w.r.t. π_k leads to

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} E[z_{ik}] \tag{4.31}$$

and maximization w.r.t $\alpha_{jk}^{(m)}$ gives us[6]

$$\begin{aligned}
\frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \alpha_{jk}^{(m)}} &= \sum_{i=1}^{N}\Bigg\{E[z_{ik}]\Bigg[\Bigg(\psi(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg) \\
&\quad + \Bigg(\psi(\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\} \qquad (4.32)\\
&\quad +(\eta_h-1)\frac{g_{jk}^{(m)}}{h_k^{(m)}}-\nu_h g_{jk}^{(m)}+\frac{\eta}{\alpha_{jk}^{(m)}}-\nu\,, \qquad (4.33)
\end{aligned}$$

where

$$g_{jk}^{(m)} = \frac{\partial h_k^{(m)}}{\partial \alpha_{jk}^{(m)}} = \frac{\partial \sum_{j=2}^{S}(\alpha_{jk}^{(m)}-\alpha_{j-1\,k}^{(m)})^2}{\partial \alpha_{jk}^{(m)}} = \begin{cases} 2(\alpha_{jk}^{(m)}-\alpha_{j+1\,k}^{(m)}) & \text{for } j=1 \\ 2(\alpha_{jk}^{(m)}-\alpha_{j-1\,k}^{(m)}) & \text{for } j=S \\ 2(2\alpha_{jk}^{(m)}-\alpha_{j+1\,k}^{(m)}-\alpha_{j-1\,k}^{(m)}) & \text{otherwise} \end{cases} \tag{4.34}$$

As shown in Equation 4.17, we optimize the function with respect to $\lambda_{jk}^{(m)} = \log \alpha_{jk}^{(m)}$ to keep $\alpha_{jk}^{(m)}$ positive. The derivative w.r.t. $\lambda_{jk}^{(m)}$ can be given as[7]

$$\begin{aligned}
\frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \lambda_{jk}^{(m)}} &= \alpha_{jk}^{(m)}\sum_{i=1}^{N}\Bigg\{E[z_{ik}]\Bigg[\Bigg(\psi(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg) \\
&\quad - \Bigg(\psi(\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\} \qquad (4.35)\\
&\quad +\alpha_{jk}^{(m)}\Bigg((\eta_h-1)\frac{g_{jk}^{(m)}}{h_k^{(m)}}-\nu_h g_{jk}^{(m)}+\frac{\eta}{\alpha_{jk}^{(m)}}-\nu\Bigg)\,. \qquad (4.36)
\end{aligned}$$

Steps of the EM algorithm are summarized in Algorithm 4.1.

[6] The derivation is given in Appendix A.4.

[7] The derivation is given in Appendix A.5.

Algorithm 4.1 Steps of the EM algorithm

```
1. Initialize membership probabilities using soft k-means algorithm given in Appendix B.
2. Initialize $\lambda_{jk}^{(m)}$ parameters by minimizing the negative of Equation 4.27 w.r.t. $\lambda_{jk}^{(m)}$. Broyden-Fletcher-Goldfarb-Shanno (BFGS) [5, 9, 10, 37] method provided by $R$ is used for numerical optimization.
3. Calculate membership probabilities, $E[z_{ik}]$ using Equation 4.30.
4. Update $\lambda_{jk}^{(m)}$ parameters using $E[z_{ik}]$ values from previous step.
5. Calculate mixing weights, $\vec{\pi}$ using Equation 4.31.
6. Go to step 3, until convergence of expected log posterior given in Equation 4.26.
```

4.4 Model comparison

Choosing the right number of components, K, is one of the most challenging steps of data clustering. Here we take two Bayesian model selection approaches, namely Bayesian information criterion (BIC) and Laplace approximation which provide methods to determine the most suitable model for the given problem.

In Bayesian model selection approaches, the *posterior probability of the model* is utilized to select a model from a set of models so that the model having the highest probability of generating the given data is favored

$$p(\mathcal{M}_K|\mathbf{X}) \propto p(\mathbf{X}|\mathcal{M}_K)p(\mathcal{M}_K)\,,$$

where $p(\mathcal{M}_K)$ represents the prior probability of K component model.

BIC approach provides an asymptotic approximation to the model posterior probability

$$\text{BIC} = -2\log\hat{L} + k\log(n)\,,$$

where $\hat{L}$, k and n represent the likelihood[8], number of parameters in the model and number of observations, respectively. Here, $k\log(n)$ acts a penalty term depending on the number of model parameters so that more complex models are penalized more strongly. For example, in our model, clustering of a data with S many bins, M different data types and K components

$$k = S \times K \times M + (K-1)$$

where $S \times K \times M$ denotes the number of Dirichlet parameters estimated and $K-1$

[8] Since the MAP estimation is used in this study, $\hat{L}$ parameter represents the posterior rather than the likelihood.

represents K many mixture weights ($\vec{\pi}$) sum to one. Subsequent to the calculation of BIC values for clusterings with different number of components, the model with the lowest BIC value is preferred. This provides a reliable means to find the optimal number of clusters in the data.

Alternatively, $p(\mathbf{X}|\mathcal{M}_K)$ term, which is also called *"the model evidence"*, can be computed by marginalizing out the Dirichlet parameters assuming that $p(\mathcal{M}_K)$ is uniform

$$p(\mathbf{X}|\mathcal{M}_K) = \int p(\mathbf{X}|Q, \mathcal{M}_K)p(Q|\mathcal{M}_K)dQ\,.$$

Although this integral cannot be calculated analytically, it can be estimated using Laplace approximation

$$\log p(\mathbf{X}|\mathcal{M}_K) \approx \log p(\mathbf{X}|\hat{Q}, \mathcal{M}_K) + \log p(\hat{Q}|\mathcal{M}_K) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|H| \quad (4.37)$$

where $\hat{Q}$ is the parameters maximizing the posterior which are estimated through EM, H is the Hessian matrix[9] of the second derivatives of negative log posterior evaluated at $\hat{Q}$ and M is the number of parameters in Q.

4.5 Profile shifting

Genomic regions being clustered, such as enhancers, usually involve a peak calling process which identifies the regions where the data is enriched. However, this process is not 100% accurate and hence peak summits, where we center a large window of length 2kb, might not be actually at the center of region of interest.

To account for this phenomena, we propose a profile shifting process where an additional integer parameter of d denoting the amount of distance between the given location of the peak summit and the actual peak summit is estimated during the EM algorithm.

Profile shifting adds one more step to the EM algorithm which is described in Algorithm 4.1. After Step 5, for each genomic region separately, windows are shifted by the amount of d^* which ranges between $-\frac{w}{8}$ and $\frac{w}{8}$ where w is the window length, and the values of d^* parameters which yield the highest likelihood of the data are retained. Since we keep separate d^* values for each genomic region, in the end a d^*vector of length N is obtained. Next EM iteration uses the data shifted by the amount of d^* and this continues until the convergence of the EM algorithm.

4.6 Artificial data generation

Histone modification-enriched genomic regions typically have a unimodal or bimodal shape due to the occupancy of nucleosomes with histone modification flanking a func-

[9] Calculation of the elements of Hessian matrix is given in Appendix C.

tional element. To reflect this aspect of the real data, we utilized sum of Gaussian functions to generate Dirichlet-Multinomial parameters from which the data can be sampled. The following equation provides a means to generate a multimodal signal conveniently:

$$f(\mu_1, \sigma_1, scale_1, \ldots, \mu_m, \sigma_m, scale_m, w) = \sum_{i=1}^{m} \exp\left(-\frac{(x - (\mu_i \cdot w))^2}{2(\sigma_i \cdot w)^2}\right) \cdot scale_i$$

where

- μ_i is the relative position of the peak summit within the window in $[0, 1]$ interval,
- σ_i represents the "width" of the peak similar to the standard deviation of a Normal distribution,
- $scale_i$ is the y-component of the peak summit,
- m is the number of peaks,
- w is the length of the window in base pairs.

For instance, the following function describes a signal with two peaks on both sides of the window center:

$$f(\mu_1 = 0.3, \sigma_1 = 0.2, scale_1 = 100, \mu_2 = 0.7, \sigma_2 = 0.2, scale_2 = 100, w = 2000)$$

The resulting template profile is shown in Figure 4.6.

Through sampling from a Dirichlet-multinomial compound with generated parameters, it is easy to create artificial data similar to those observed in enriched genomic loci. Results of clustering on the data generated by this scheme is given in the next chapter. Additionally, a method for generating artificial data randomly is given in Appendix G.

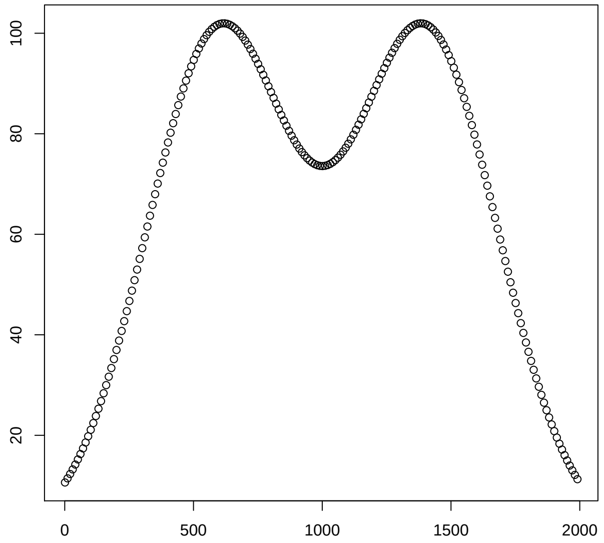


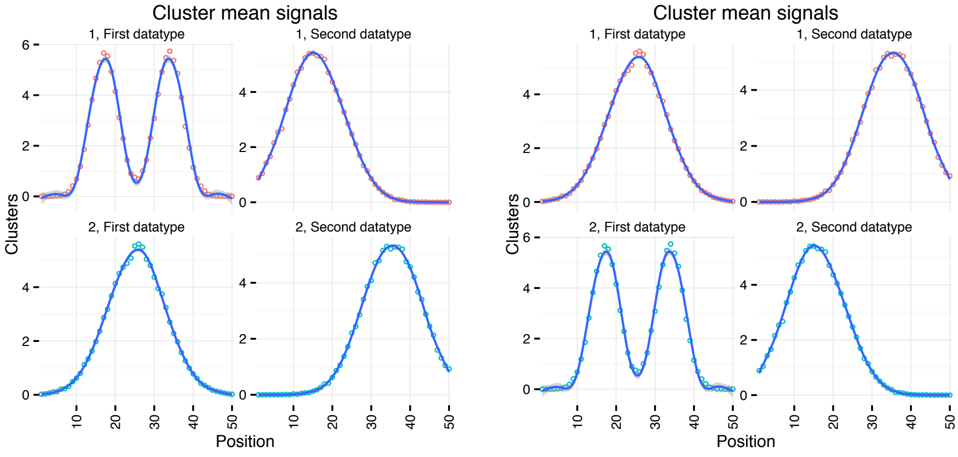
Figure 4.6: A bimodal shaped data which can be used as parameters of Dirichlet-multinomial compound to generate artificial data.

Chapter 5

Results

5.1 Artificial data clustering

Using the method described in Section 4.6, artificial data consisting of two data types and two clusters are generated. Generated data vectors has the length of 2000 and clusterings were performed in 40bp resolution. Number of samples per cluster is 500 with 1000 being the total sample size. The total number of reads per sample is 100 which leads to the expected bin count (coverage) of $\frac{100}{2000} \cdot 40 = 2$. Mean signals of the true clusters and the ones identified by the model are given in Figure 5.1.



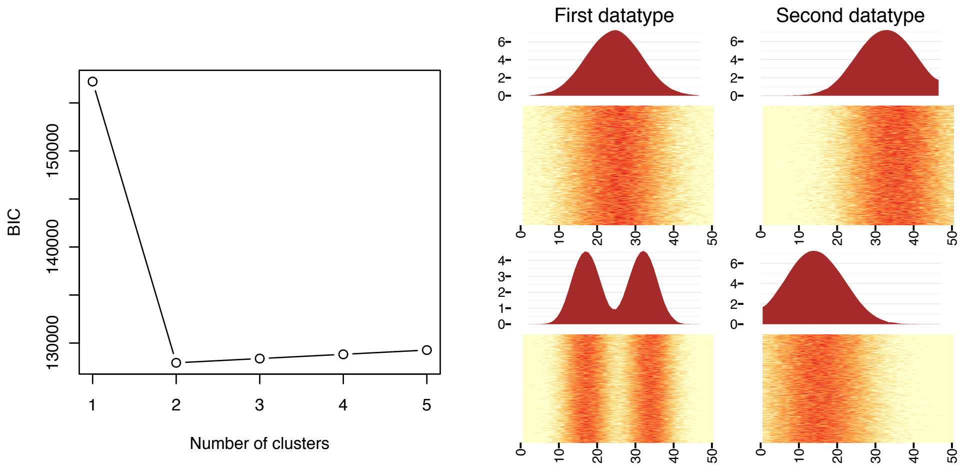
(a)

(b)

Figure 5.1: Mean signals of the true clusters(a) and the clusters identified by the model(b)

Using this proof-of-priciple example, it can be said that the model perfectly clustered the artificial data with the only difference being the switch of cluster labels.[1] BIC curve of clustering results with varying number of components is given

[1] The area under curve (AUC) value for the clustering is computed as 1.



(a) BIC values for clustering results with varying number of components

(b) The estimated Dirichlet parameters along with the data

Figure 5.2

in Figure 5.2a, where number two can be identified as the *knee* of the curve. In Figure 5.2b, estimated Dirichlet parameters are shown along with the clustered data.

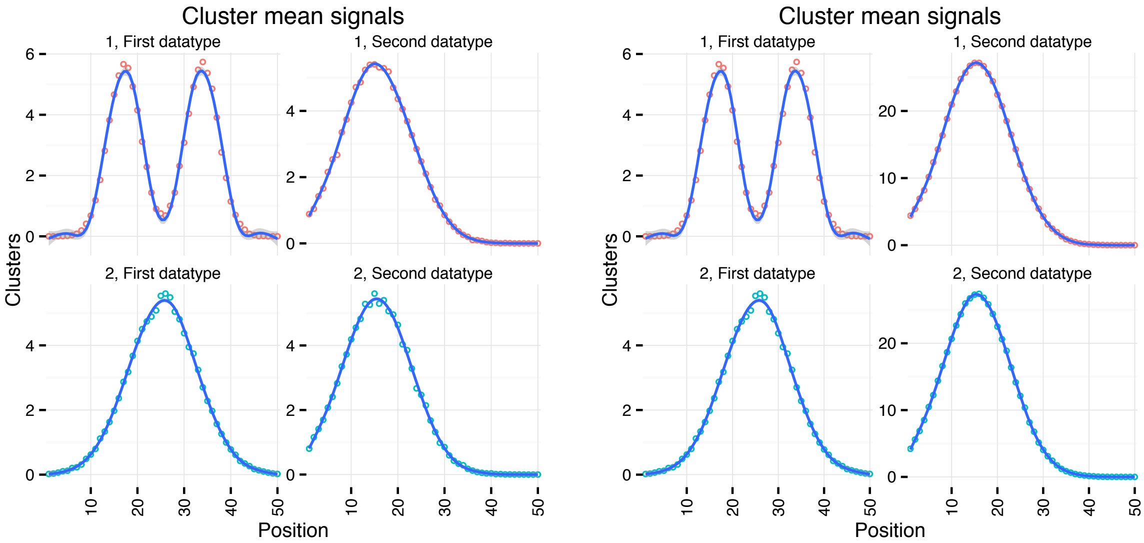
5.1.1 Comparison with data concatenation

When multiple data types are involved in clustering, most commonly used approach is to concatenate data vectors. However, this approach has some drawbacks. For instance, if the signal of one of the data types has a higher magnitude than the others, it can dominate the clustering and lead to misleading results, especially if the cluster structure of the data type in question is different than the others.

Here, we compare the mixture model extended with multi-view approach to the same model without the multi-view extension to demonstrate the advantages of the multi-view approach. Two data sets with one thousand samples are generated for this purpose. In both of the data sets, the first data type has a distinct two-cluster structure, whereas the second data type has only one. While the signal magnitudes of the data types are similar in the first data set, second one has differing magnitudes such that the data type with a single component structure has a higher signal magnitude[2]. To cluster the generated data using the model without the multi-view extension, data vectors are concatenated. Mean signals of the generated data is given in Figure 5.3.

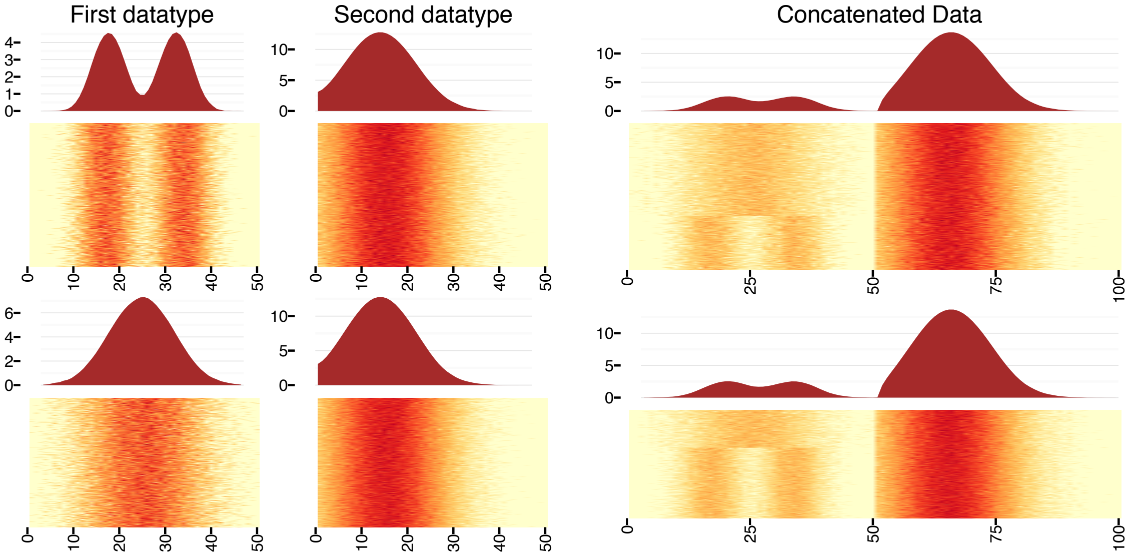
While both approaches can cluster the first data set perfectly (AUC $= 1$), the second data set is clustered accurately only by the multi-view approach. Results showing the estimated Dirichlet parameters and the data is given in Figure 5.4.

[2] Data types with lower signal magnitudes has 100 reads per sample which leads to the expected bin count of $\frac{100}{2000} \cdot 40 = 2$. Second data type of the second data set has 500 reads per sample which leads to the expected bin count of $\frac{500}{2000} \cdot 40 = 10$.



(a) First data set with similar signal magnitudes (b) Second data set with differing signal magnitudes

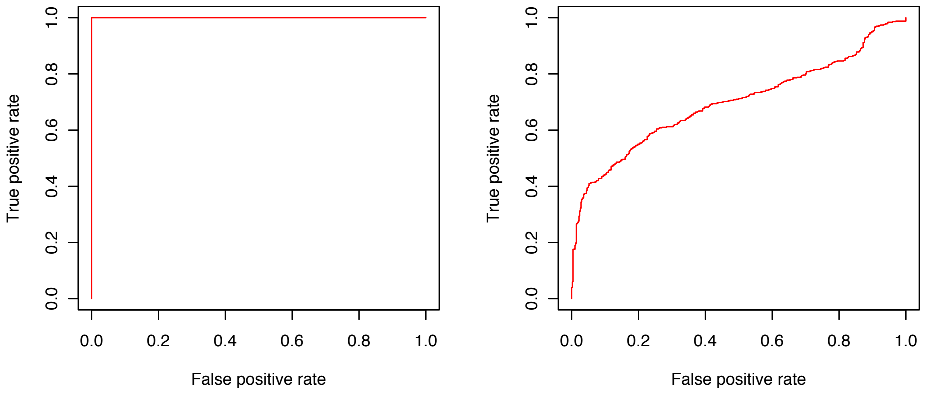
Figure 5.3: Mean signals of the true clusters in the artificial data.



(a)

(b)

Figure 5.4: Clustering results demonstrating multi-view(a) and data concatenation(b) approaches on the data set with varying signal magnitudes. Dirichlet parameters along with the heatmaps showing the clustered artificial data. The rows represent clusters whereas the columns represent different data types.



(a) Multi-view approach (b) Data concatenation

Figure 5.5: The comparison of clustering performance through ROC curves. AUC values are 1(a) and 0.6976(b).

It can be clearly seen that, in the concatenation approach, high signal magnitude dominated the clustering so that the method was not able to distinguish between two distinct clusters that exist in the first data type. Additionally, receiver operating characteristic (ROC) curves for both results are given in Figure 5.5.

5.1.2 Profile shifting

The epigenomics data that we aim to analyze are extracted from the genomic regions of interest. However, the methods that determine these genomic loci, such as peak callers, may not perform with 100% accuracy. So, the positions where the data are extracted might need to be corrected by some means. As described in Section 4.5, our model can account for this uncertainty of the anchor points such as TF ChIP-seq peaks or TSSs, in other words, during the clustering, a discrete parameter vector of length N, denoted as d in the previous chapter, representing the amount of shifts is estimated in the EM iterations.

Here, to demonstrate the effect of profile shifting on clustering, artificial data composed of 100 samples and two equally sized clusters are generated. Data vectors are of length 1600 and the resolution is 40bp. To define shifting amounts for each sample, 100 integers are sampled from a Skellam distribution[3] with mean 0 and standard deviation 120. The initial artificial data, randomly shifted data and the clustering result of our model are given in Figure 5.6. It can be seen that the shifting distances are correctly recovered by our model.

[3] Skellam distribution can be defined as the difference between two independent random variables each having Poisson distributions.

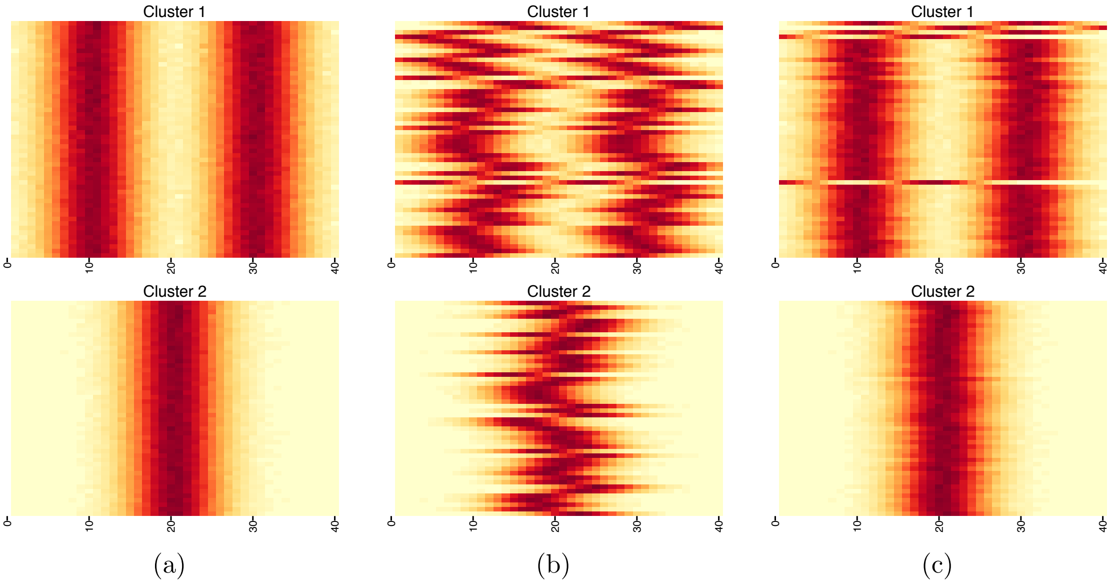


Figure 5.6: Artificial data(a), randomly shifted data(b) and the clustering result of the randomly shifted data(c).

5.2 Enhancer clustering

We applied our method to heterogeneous epigenomics data extracted from enhancer regions. As described in Chapter 3, enhancer loci consist of DNase-I sensitive and TSS-distal p300 binding sites. For every binding site, 5' ends of reads for the data types that are relevant to enhancer presence are counted in bins of 40bp over a region of -1kb to +1kb relative to p300 peak summit. Data types chosen for the clustering are CTCF, H2A.Z, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac and Pol2.

First step of the analysis is to perform clustering using varying number of components and to determine optimum number of clusters present in the data using model selection methods, e.g. BIC

$$\hat{K}_{\text{BIC}} \quad = \quad \underset{K \in \{1,2,\ldots,12\}}{\arg\min} \; \text{BIC}_{\mathcal{M}_K} \,.$$

Figure 5.7 shows BIC goodness-of-fit values[4] for clusterings with different number of components ($\text{BIC}_{\mathcal{M}_K}$) which indicates that the optimal number of clusters present in the data is five, i.e. $\hat{K}_{\text{BIC}} = 5$. Therefore, rest of the figures in this section are based on the clustering with five components. In Figure 5.8, mean profiles of clusters are shown for each data type. The data type with highest signal magnitude is H3K27ac which is the indicative of active enhance regions. It is noteworthy that clusters 2, 3 and 4, 5 are mirror images of each other, while cluster 1 seems to be

[4] Since these values are quite similar to Laplace approximation values, only BIC figures are given. Laplace approximation curves can also be found in Appendix E.

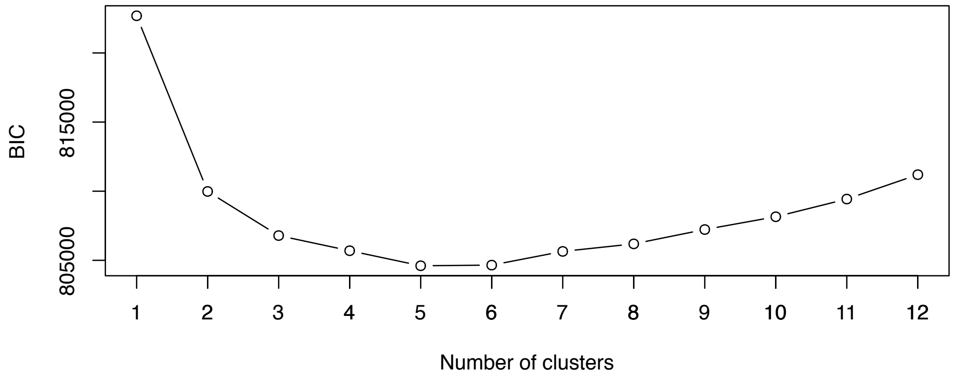


Figure 5.7: BIC values for different clusterings of enhancer loci

slightly more symmetrical than the rest.

In addition to the cluster mean signals, parameters of the Dirichlet distributions give useful information. While the shape of the Dirichlet parameters represents the relative read counts fall into the bins, the magnitude of the parameters indicates the inverse variance (precision) of the read counts in the corresponding bin. In Figure 5.9, the clustered data are shown along with the Dirichlet parameters where the data are indicated as heatmaps and the parameters are given above the data. Note that, the height of the heatmaps are proportional to the number of genomic loci within the cluster hence it can be said that smallest cluster is number 5. When the Dirichlet profile magnitudes of data types are compared, it can be said that CTCF clusters has the highest variation, which is also related to the fact that no distinct patterns are observed in the heatmaps showing the clustered CTCF data. Finally, the asymmetrical Dirichlet profiles of cluster 4 and 5 for the RNA polymerase II data type are worth mentioning. Such asymmetrical Pol2 profiles, which are also present in Figure 5.8, may indicate the direction of the enhancer RNA (eRNA) transcription.

In Appendix F, another visualization shows how the cluster memberships of each enhancer change as the number of components increases. These figures give us an idea about how cluster memberships would differ, if we partitioned the data into smaller or larger number of clusters. Therefore, observing a group of genomic loci falling into the same cluster in most clusterings suggests that this cluster is more distinct and separable than the others. It can be said that enhancer clusters 3 and 4 are examples of such cases.

5.2.1 Transcription factor binding site enrichment

The TFBS enrichment in different classes of enhancers that are identified by our method is also investigated. For this analysis, ChIP-seq peaks of 60 transcription factors that are generated by ENCODE Uniform pipeline (track `wgEncodeAwgTfb-`

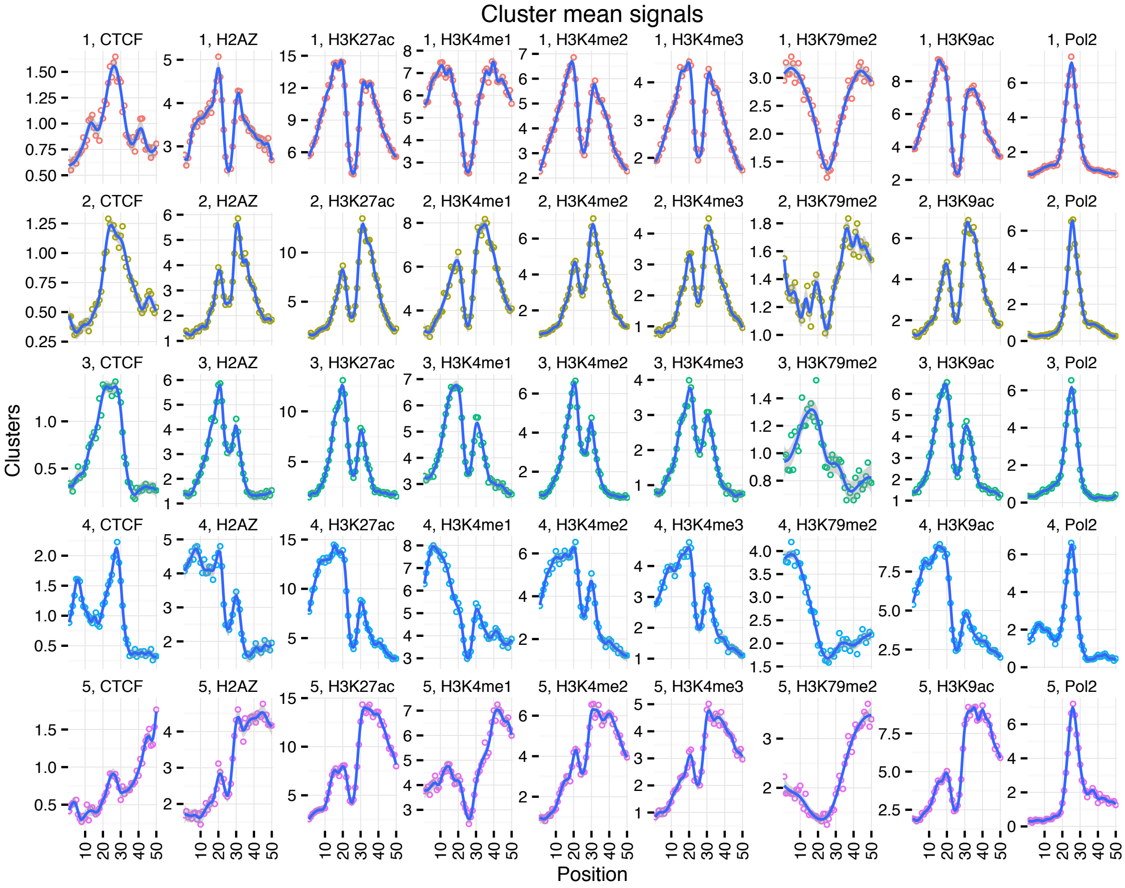


Figure 5.8: Mean profiles of five enhancer clusters. The rows represent clusters whereas the columns represent different data types. Similar to other plots showing the mean signal, mean data are indicated as small points and the smoothers are plotted only to display the trend in the data.

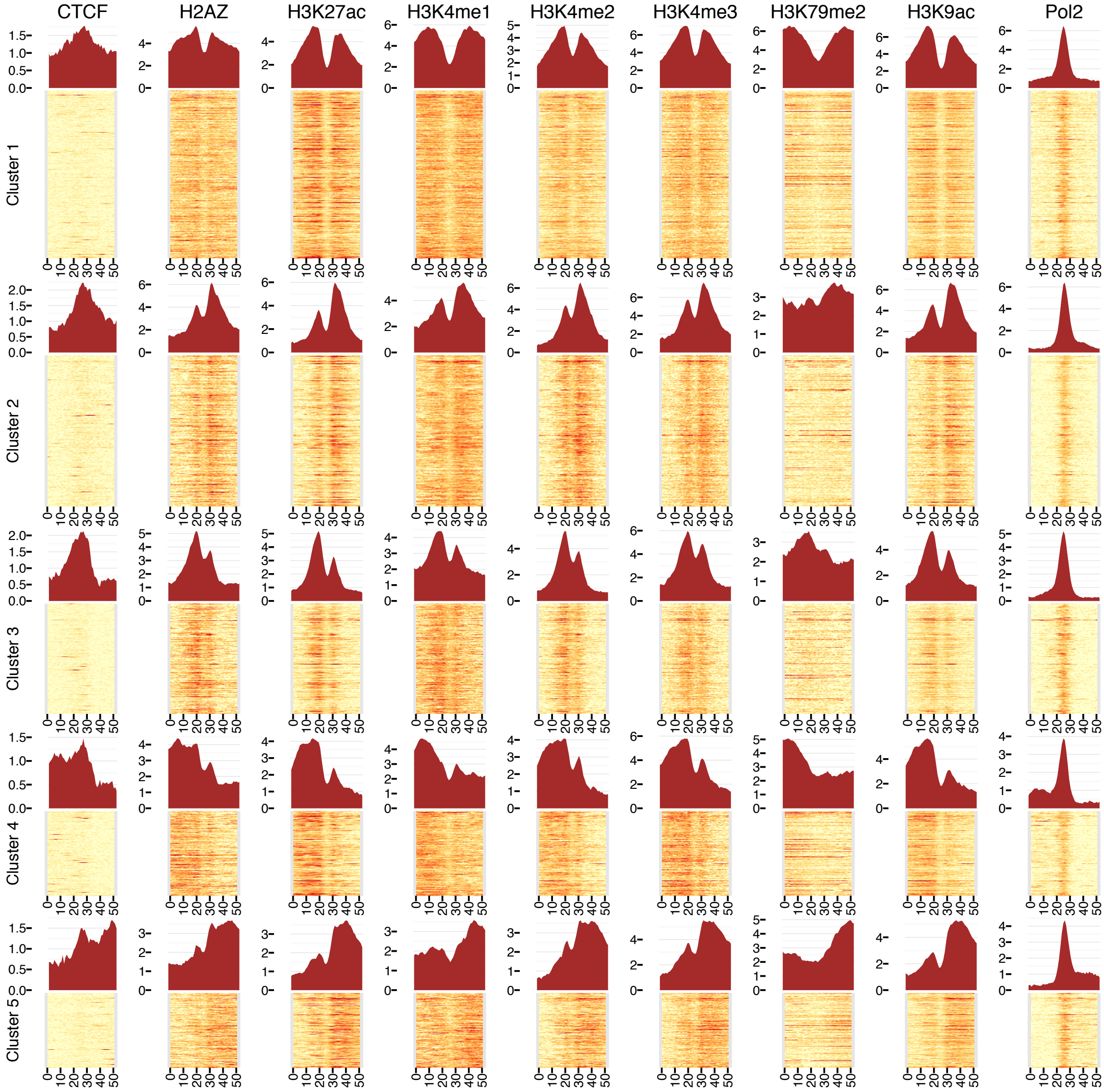


Figure 5.9: Dirichlet parameters along with the heatmap showing the clustered enhancer data. Again, the rows represent clusters whereas the columns represent different data types.

TF	Enrichment	Cluster(s)	Function of the TF
BDP1	Not enriched in	4	TFIIIB subunit and Pol III recruiter involved in transcriptional activation
SETDB1		3	H3K9-selective histone methyltransferase, transcriptional repression
RAD21		5	DNA repair
BRF2		4, 5	TFIIIB subunit
RPC155	Enriched in	1, 3	Largest component of Pol III which transcribes housekeeping genes
ATF3		2, 3	Transcriptional activator and repressor
TR4		5	Transcriptional activator and repressor
NELFe		3	A complex which binds to Pol II to suppress elongation

Table 5.1: Summary of TFBS enrichment in enhancer clusters for selected TF.

sSydhK562) are downloaded. Then, for each TF-cluster pair a significance test was performed by the GAT tool [11] to determine whether the TF enrichment in the cluster of interest is more than expected by chance. To achieve this, first, random locations with same size distribution of TF peaks are created within the mappable regions[5] of the genome, and then the number of randomly created regions overlapping with the enhancer cluster is compared to the number of observed overlaps between the cluster and TFBSs in question. This comparison yields to the empirical p-value of the enrichment which can then be adjusted through Benjamini-Hochberg correction to obtain the q-value. Rows and columns of the heatmap consisting of $-\log$ transformed q-values are clustered using hierarchical clustering by `gplots` `R` package. The results are shown in Figure 5.10

Results of the significance test enabled us to discover TFs that exhibit interesting enrichment patterns e.g. TFs significantly enriched only in some clusters but not in the others. One such interesting group of TFs are BDP1 (B double-prime 1), SETDB1 (SET domain, bifurcated 1), RAD21 and BRF2 (B-related factor 2) which are not enriched in a subset of enhancer clusters. On the other hand, RPC155, ATF3 (activating transcription factor 3), TR4 (testicular receptor 4) and NELFe (negative elongation factor E) are only enriched in clusters (1, 3), (2, 3), (5) and (3), respectively. Summary of these results are given in Table 5.1. The function of TFs are provided by Factorbook [41].

[5] ENCODE 36mer mappability data generated in Guigó lab at Centre de Regualció Genòmica were used. During the simulations conducted to compute q-values, only the positions with mappability score of one in the genome were taken into account.

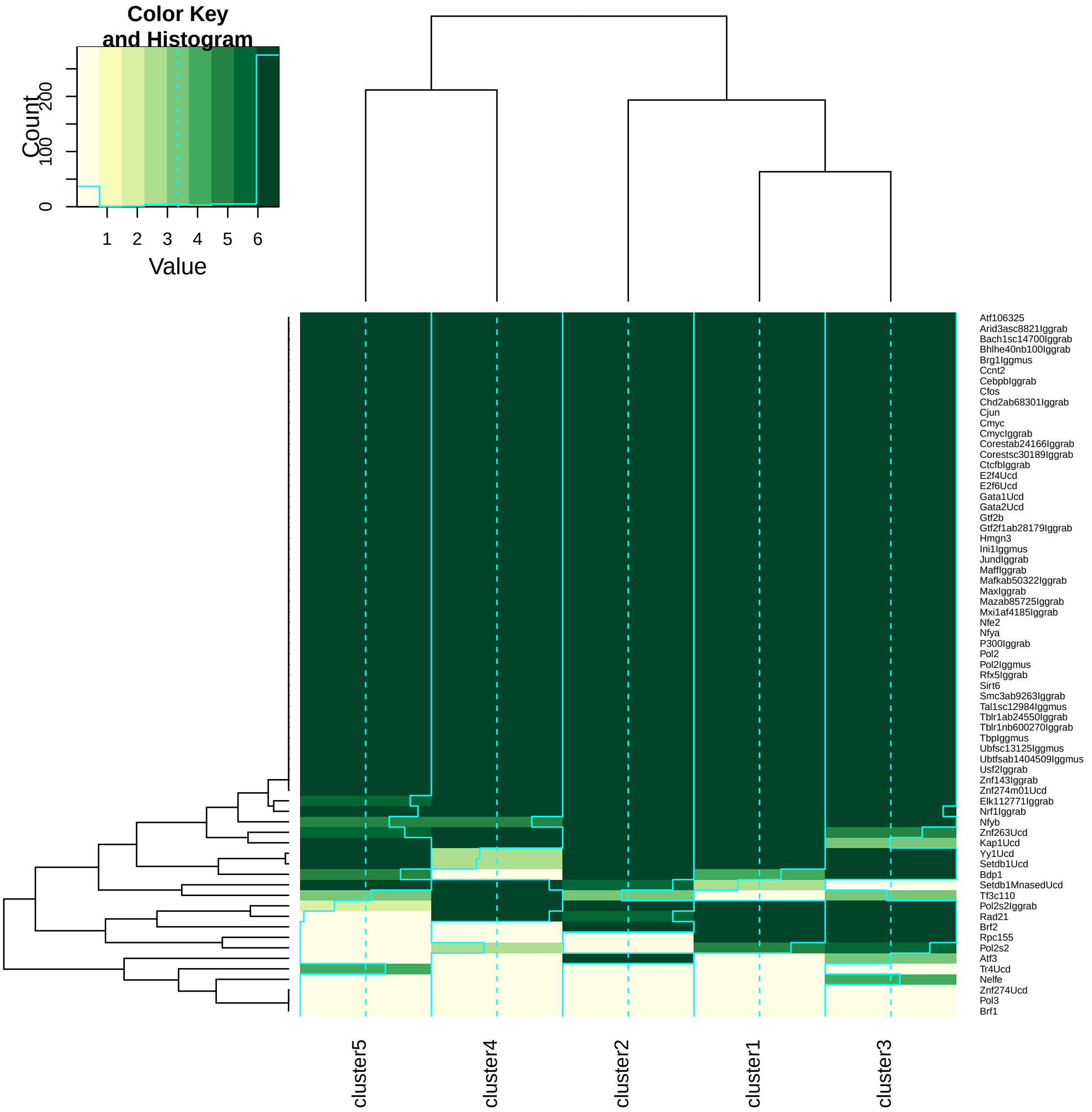
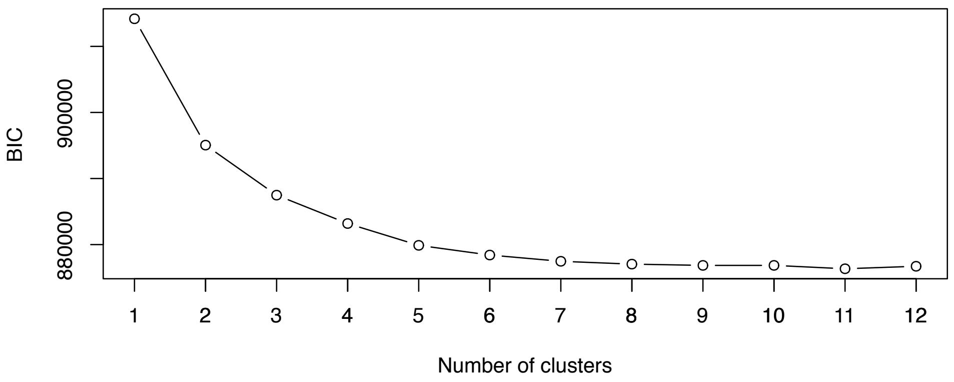


Figure 5.10: Significance test results of TFBS enrichment analysis. For each TF-enhancer cluster pair, empirical p-values with Benjamini-Hochberg correction are calculated using GAT tool [11]. Negative log-transformed q-values are presented.



(a)

Figure 5.11: BIC values for different clusterings of promoter loci

5.3 Promoter clustering

In addition to discover epigenomic patterns in enhancer regions, we aimed to cluster promoters by taking multiple data types which are relevant to promoter presence into account. In this section, results of this analysis are presented. As described in Chapter 3, first, TSSs overlapping with DNase-I hypersensitivity peaks are determined. Subsequently, 5' ends of reads falling into 40bp bins over a region of -1kb to +1kb relative to TSS centers are counted. The data types used in this analysis consist of H2A.Z, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac and Pol2.

Similar to the enhancer clustering, we used BIC goodness-of-fit values to determine a reasonable number of clusters. BIC curve for the clustering results with varying number of components is given in Figure 5.11. Unlike the one for the enhancer clustering results, here we observe a BIC curve with monotonically decreasing values. For that reason, seven was visually chosen as a reasonable number of components, since the decrease in BIC for the numbers larger than seven is negligible.

Mean signals of identified promoter clusters are given in Figure 5.12. First of all, in all clusters the signal of H3K4me3, a characteristic histone mark for promoter regions, has high magnitude and mostly bimodal-like shapes representing two histones flanking the TSS. Except for cluster 6, all histone modifications exhibit an asymmetrical profile which may indicate the direction of transcription [3].

Figure 5.13 shows the estimated Dirichlet parameters along with the heatmaps representing the clustered data. Similar to Figure 5.9 given in the previous section, in addition to the epigenetic signatures of various promoter clusters, this figure also shows the variance of the data within clusters through the magnitude of Dirichlet parameters.

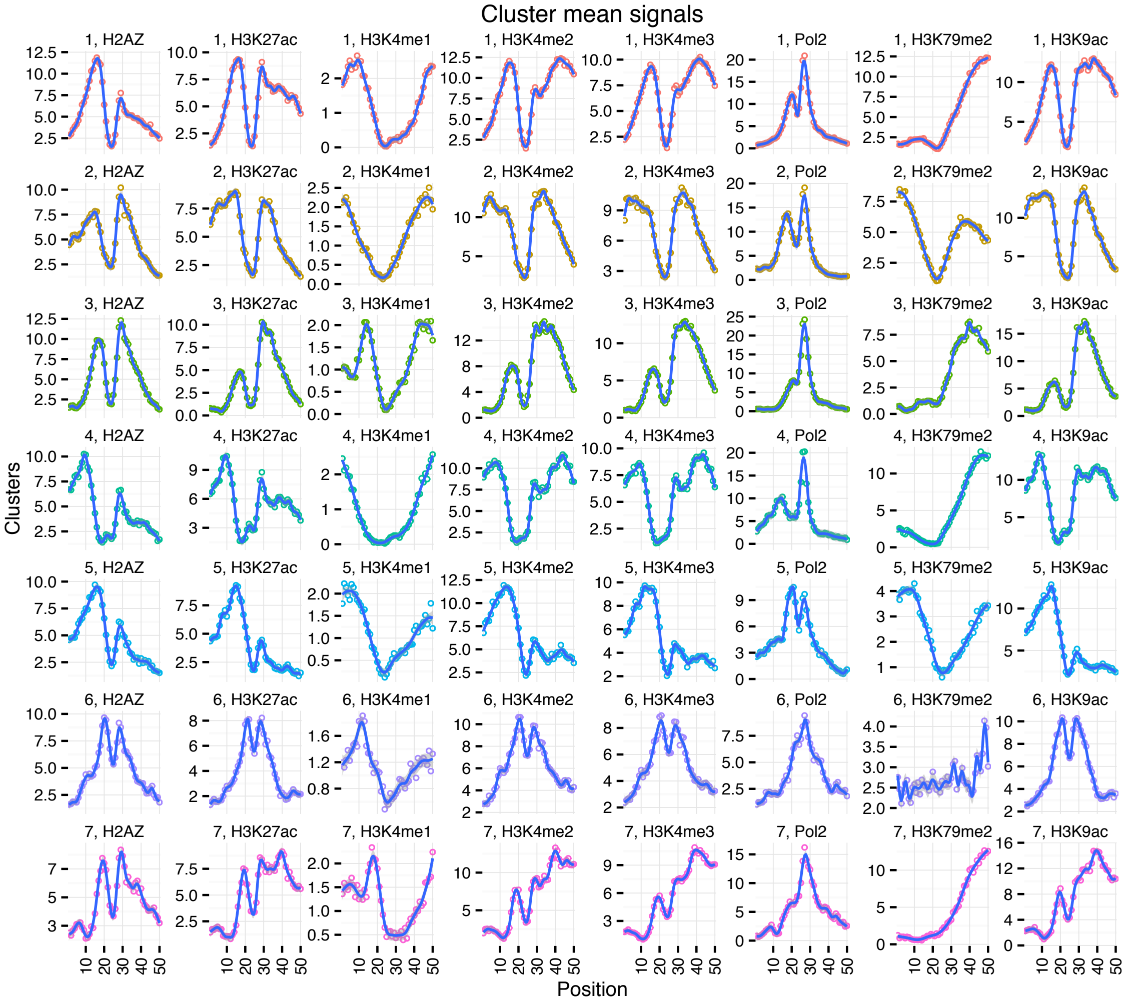


Figure 5.12: Mean profiles of seven promoter clusters. The rows represent clusters whereas the columns represent different data types. Similar to other plots showing the mean signal, mean data are indicated as small points and the smoothers are plotted only to display the trend in the data.

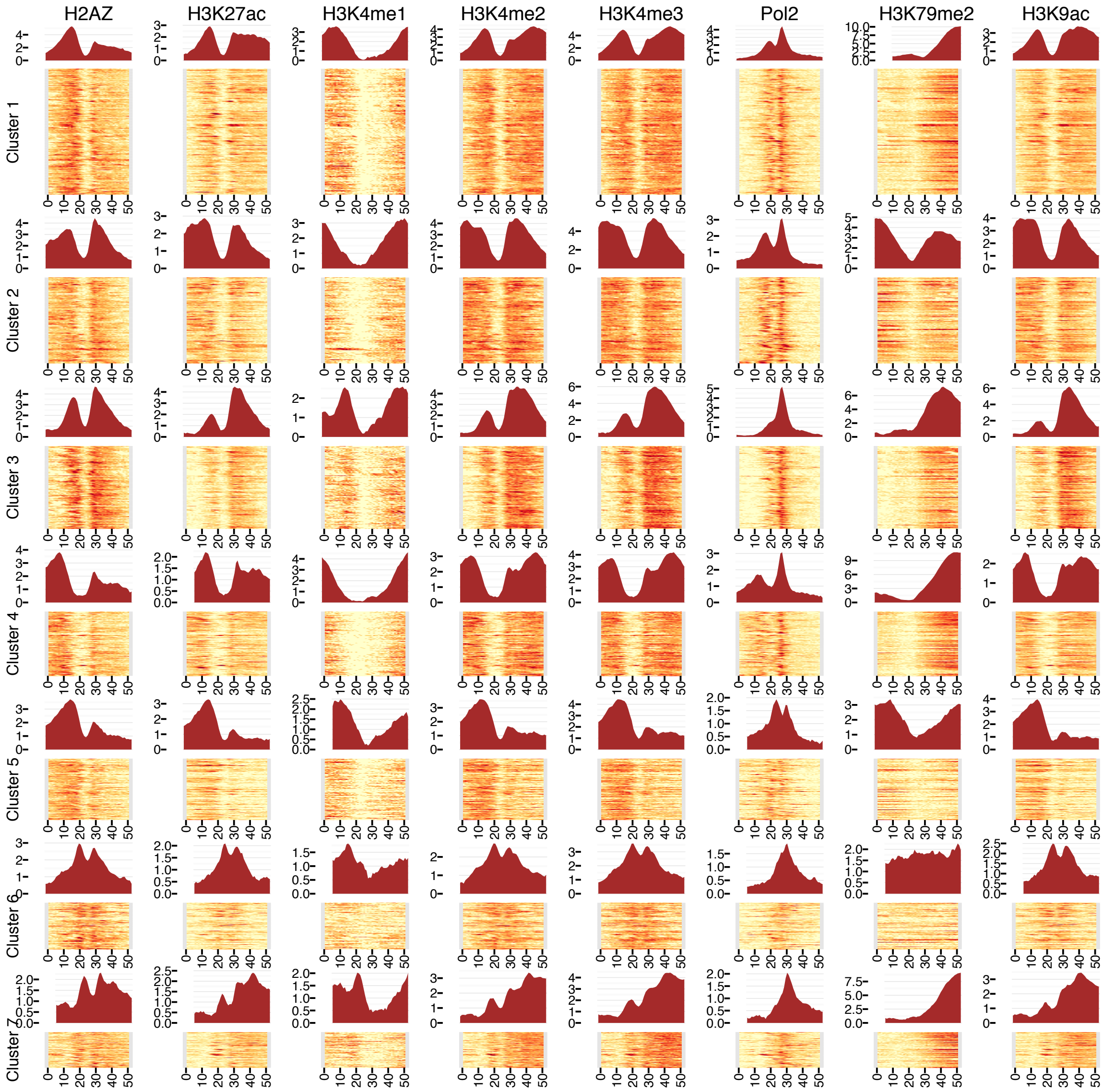


Figure 5.13: Dirichlet parameters along with the heatmap showing the clustered promoter data. Again, the rows represent clusters whereas the columns represent different data types.

TF	Enrichment	Cluster(s)	Function of The TF
GATA1	Not enriched in	6	TF that binds to promoter regions and regulates the transcription
SMC3		3	Involved in DNA repair and chromosome maintenance
GATA2		2	TF that binds to promoter regions and regulates the transcription
CoREST/RCOR1		1	Protein that binds to REST transcriptional repressor
SIRT6	Enriched in	7	Regulates epigenetic gene silencing
TFIIIC-110		5,7	Subunit of Pol III transcription factor TFIIIC required for transcription

Table 5.2: Summary of TFBS enrichment in promoter clusters for selected TF.

5.3.1 Transcription factor binding site enrichment

Similar to the analysis given in Section 5.2.1, we also assessed the enrichment of various TFs in promoter clusters. The GAT tool was used similarly to compute q-values for every cluster-TF pair where the TFs that are enriched in promoter regions more than expected by chance are revealed. The heatmap showing $-\log$ transformed q-values is presented in Figure 5.14.

While quite a few transcription factors are enriched in all promoter clusters, the lower half of the heatmap shows patterns that require further investigation. Table 5.2 summarizes TFs that are only enriched in (or only not enriched in) a few clusters along with the function of TF in question.

5.4 Hyperprior regularization

In Figure 5.15, clustering result of H3K4me1 and H3K27ac profiles of width 2kb centered at thousand enhancer regions are shown. Every vertical panel consists of three clusters of a single data track. Within a panel, clusters are represented as heatmaps below the Dirichlet parameters fitted to the cluster in question. The effect of regularization can be observed by comparing the Dirichlet parameters of two clusterings performed with(5.15a) and without(5.15b) regularization. Although the profile of Dirichlet parameters is closer to the real data profiles when regularization is applied, no improvement on the clustering results are observed.

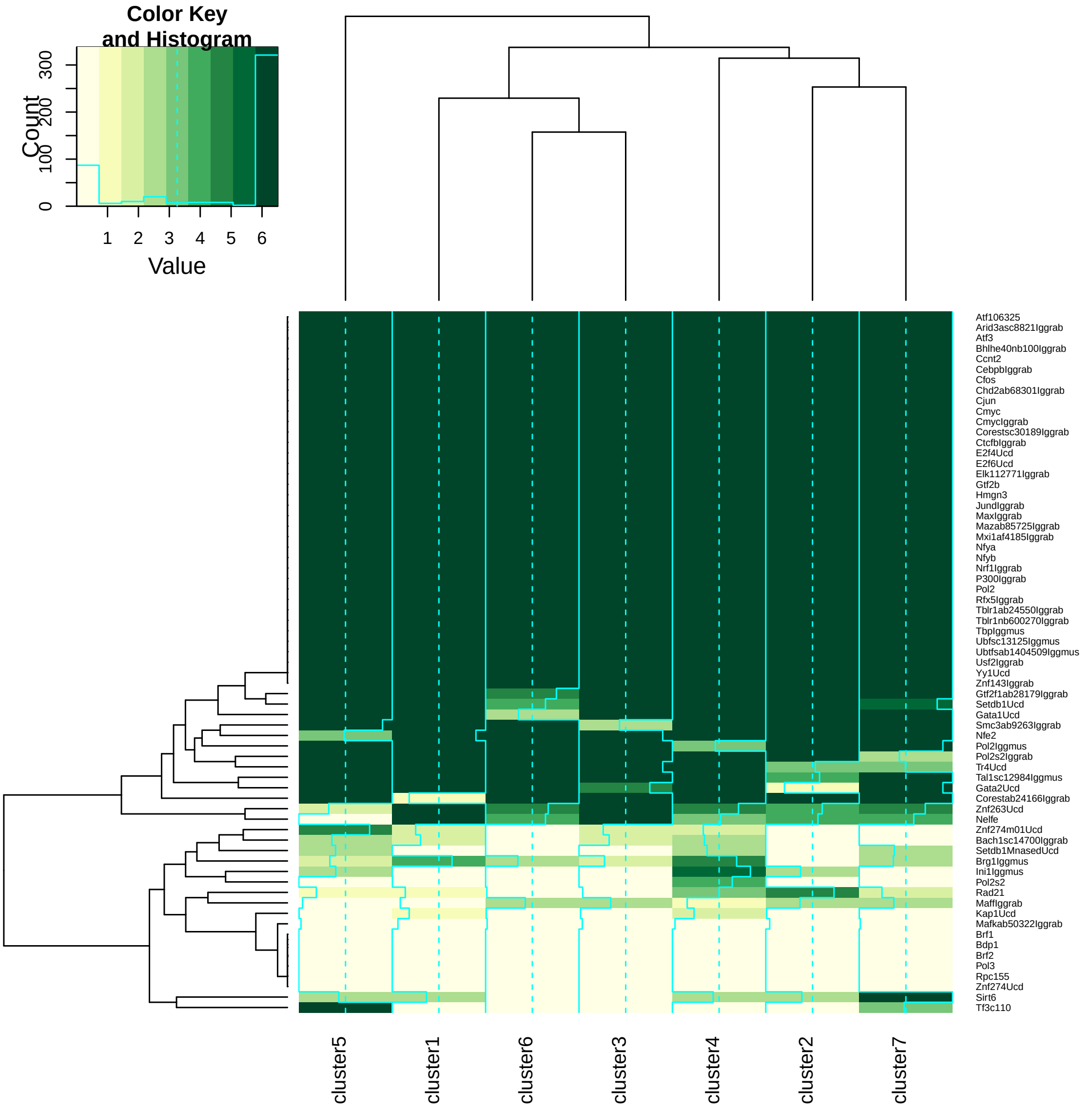
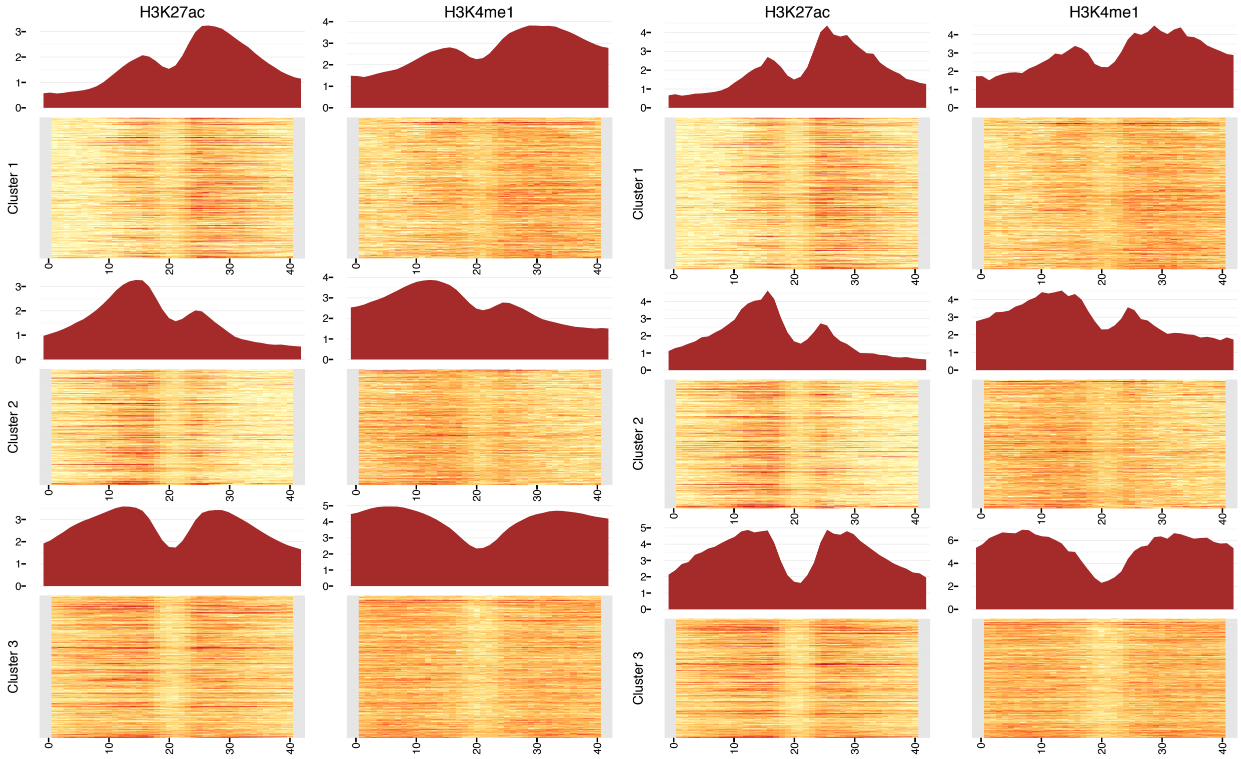


Figure 5.14: In this figure, significance test results of TFBS enrichment analysis are given. For each TF-promoter cluster pair, empirical p-values with Benjamini-Hochberg correction are calculated using GAT tool [11]. Negative log-transformed q-values are presented.



(a) Regularization

(b) No regularization

Figure 5.15: The Dirichlet parameters of the model-based clustering method fitted to the data.

Chapter 6

Discussion

The study of epigenomics provides crucial information concerning the key biological mechanisms such as the gene regulation. The analysis of vast amount of publicly available epigenomic datasets is essential as it may shed light on many critical questions of biology. However, the datasets are growing faster than the computational means required for the analysis and therefore novel methods are still needed. One of the most commonly used analysis techniques is clustering methods aiming to uncover commonly occurring epigenetic signatures. A vital aspect lacking in currently available clustering methods is the handling of multiple data types in a principled manner to account for the combinatorial presence of epigenetic marks. The determination of a reasonable number of clusters emerges as another challenging task.

In this study, we provide a probabilistic clustering technique tailored for the heterogeneous epigenomics data that are intrinsically sparse and discrete. The epigenomics data of various data types that are extracted from the genomic locus of interest are handled rigorously by using a well-known technique called the multi-view clustering approach. This technique can accurately model the epigenomics data with various number of marks each having differing signal magnitudes and shapes. We demonstrated this aspect of the model by comparing the technique with the same mixture model lacking the multi-view approach which can only model concatenated epigenetic data vectors. Results show that concatenation may fail especially in cases where the data types have varying signal magnitudes.

Another challenge that needs to be addressed is the uncertainty of the locations of anchor points such as TSSs. In this work, we implemented a profile shifting technique by exploiting the iterative essence of the expectation-maximization algorithm, so that the profiles are aligned by means of shifting during the process of clustering. The effect of this aspect is also demonstrated. Considering the fact that there is a correlation between the count data of consecutive nucleotide positions which results from the unimodal or bimodal shape of histone modification signals, we introduced an hyperprior regulation to account for this phenomena.

We applied our model to various histone modification and transcription factor ChIP-seq data extracted from the enhancer and promoter regions to identify distinct epigenetic patterns. Moreover, a TF enrichment analysis is conducted to determine the transcription factors that are significantly enriched in a subset of enhancer or promoter clusters.

6.1 Future work

Our approach can be further used to discover novel patterns present in other types of genomic loci such as transcription factor binding sites, transcription termination sites (TTS) or intron-exon boundaries. Alternatively, the clustering can also be applied to the data generated through other experimental protocols.

Biological meaning of the relation between enhancer/promoter clusters and TFs that are significantly enriched in specific clusters requires further investigation. The association between identified clusters and TFs can be supported by motif enrichment analysis where the motifs encountered in clusters are compared to those of TFs.

Another potential research direction is to extend the model to a classifier similar to the approach taken by Holmes et al.[13] where a mixture model is trained for each class separately and the probabilities of the new data being generated by the trained models are compared.

We introduced a finite mixture model where the data are clustered for varying number of components iteratively and then, a reasonable number of clusters is identified by means of a model selection approach. An alternative approach for overcoming the identification of correct number of clusters would be to extend the model into an infinite mixture model using a Dirichlet process prior.

Bibliography

[1] Annunziato, A. Dna packaging: Nucleosomes and chromatin. *Nature Education 1*, 1 (2008), 26.

[2] Arthur, D., and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 1027–1035.

[3] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell 129*, 4 (2007), 823–837.

[4] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell 132*, 2 (2008), 311–322.

[5] Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics 6*, 1 (1970), 76–90.

[6] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), 1–38.

[7] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols 4*, 8 (2009), 1184–1191.

[8] Falvo, J. V., Jasenosky, L. D., Kruidenier, L., and Goldfeld, A. E. Epigenetic control of cytokine gene expression: regulation of the tnf/lt locus and t helper cell differentiation. *Advances in immunology 118* (2012), 37–128.

[9] Fletcher, R. A new approach to variable metric algorithms. *The computer journal 13*, 3 (1970), 317–322.

[10] Goldfarb, D. A family of variable-metric methods derived by variational means. *Mathematics of computation 24*, 109 (1970), 23–26.

[11] Heger, A., Webber, C., Goodson, M., Ponting, C. P., and Lunter, G. Gat: a simulation framework for testing the association of genomic intervals. *Bioinformatics 29*, 16 (2013), 2046–2048.

[12] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics 39*, 3 (2007), 311–318.

[13] Holmes, I., Harris, K., and Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One 7*, 2 (2012), e30126.

[14] Hon, G., Ren, B., and Wang, W. Chromasig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS computational biology 4*, 10 (2008), e1000201.

[15] Hon, G. C., Hawkins, R. D., and Ren, B. Predictive chromatin signatures in the mammalian genome. *Human molecular genetics 18*, R2 (2009), R195–R201.

[16] Jenuwein, T., and Allis, C. D. Translating the histone code. *Science 293*, 5532 (2001), 1074–1080.

[17] Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. H3.3/h2a.z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics 41*, 8 (2009), 941–945.

[18] Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences 107*, 7 (2010), 2926–2931.

[19] Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. Design and analysis of chip-seq experiments for dna-binding proteins. *Nature biotechnology 26*, 12 (2008), 1351–1359.

[20] Kondo, Y. Epigenetic cross-talk between dna methylation and histone modifications in human cancers. *Yonsei medical journal 50*, 4 (2009), 455–463.

[21] Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C. L., Raha, D., Winters, E. E., Johnson, S. M., Snyder, M., Batzoglou, S., and Sidow, A. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research 22*, 9 (2012), 1735–1747.

[22] Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol 10*, 3 (2009), R25.

[23] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. The sequence alignment/map format and samtools. *Bioinformatics 25*, 16 (2009), 2078–2079.

[24] MacKay, D. J. *Information theory, inference, and learning algorithms*, vol. 7. Citeseer, 2003.

[25] Mardis, E. R., et al. Chip-seq: welcome to the new frontier. *Nature methods 4*, 8 (2007), 613–613.

[26] Maston, G. A., Landt, S. G., Snyder, M., and Green, M. R. Characterization of enhancer function from genome-wide analyses. *Annual review of genomics and human genetics 13* (2012), 29–57.

[27] Minka, T. Expectation-maximization as lower bound maximization. *Tutorial published on the web at http://www-white. media. mit. edu/tpminka/papers/em. html* (1998).

[28] Murphy, K. P. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[29] Nair, N. U., Kumar, S., Moret, B. M., and Bucher, P. Probabilistic partitioning methods to find significant patterns in chip-seq data. *Bioinformatics* (2014), btu318.

[30] Nielsen, F. G., Markus, K. G., Friborg, R. M., Favrholdt, L. M., Stunnenberg, H. G., and Huynen, M. Catchprofiles: clustering and alignment tool for chip profiles. *PloS one 7*, 1 (2012), e28272.

[31] Ong, C.-T., and Corces, V. G. Enhancers: emerging roles in cell fate specification. *EMBO reports 13*, 5 (2012), 423–430.

[32] Park, P. J. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics 10*, 10 (2009), 669–680.

[33] Ptashne, M. Gene regulation by proteins acting nearby and at a distance. *Nature 322*, 6081 (1985), 697–701.

[34] Quinlan, A. R., and Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics 26*, 6 (2010), 841–842.

[35] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[36] Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. Dynamic regulation of nucleosome positioning in the human genome. *Cell 132*, 5 (2008), 887–898.

[37] Shanno, D. F. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation 24*, 111 (1970), 647–656.

[38] Strahl, B. D., and Allis, C. D. The language of covalent histone modifications. *Nature 403*, 6765 (2000), 41–45.

[39] Tian, X., and Fang, J. Current perspectives on histone demethylases. *Acta biochimica et biophysica Sinica 39*, 2 (2007), 81–88.

[40] Wang, J., Lunyak, V. V., and Jordan, I. K. Chromatin signature discovery via histone modification profile alignments. *Nucleic acids research 40*, 21 (2012), 10642–10656.

[41] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research 22*, 9 (2012), 1798–1812.

[42] Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[43] Wikipedia. Chromatin. `http://en.wikipedia.org/wiki/Chromatin`. Accessed: 2014-08-21.

[44] Wikipedia. Dirichlet distribution. `http://en.wikipedia.org/wiki/Dirichlet_distribution`. Accessed: 2014-08-08.

[45] Wilbanks, E. G., and Facciotti, M. T. Evaluation of algorithm performance in chip-seq peak detection. *PloS one 5*, 7 (2010), e11471.

[46] Wood, S. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2006.

[47] Wood, S. N. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society (B) 62*, 2 (2000), 413–428.

[48] Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73*, 1 (2011), 3–36.

Appendix A

Derivations

A.1 Derivation of Equation 4.16

$$
\begin{aligned}
p(\vec{\alpha}_1^{(1)}, ..., \vec{\alpha}_K^{(1)}, \vec{\alpha}_1^{(M)} ..., \vec{\alpha}_K^{(M)}) &\propto \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{j=1}^{S} \Gamma(\alpha_{jk}^{(m)}; \eta, \nu) \\
&\quad \prod_{m=1}^{M} \prod_{k=1}^{K} \Gamma(h_k^{(m)}; \eta_h, \nu_h) && \text{(A.1)} \\
&\propto \prod_{m=1}^{M} \prod_{k=1}^{K} \Gamma(h_k^{(m)}; \eta_h, \nu_h) && \text{(A.2)} \\
&\quad \prod_{j=1}^{S} \Gamma(\alpha_{jk}^{(m)}; \eta, \nu) && \text{(A.3)} \\
&\propto \prod_{m=1}^{M} \prod_{k=1}^{K} \frac{\nu_h^{\eta_h} h_k^{(m)\,\eta_h - 1} e^{-\nu_h h_k^{(m)}}}{\Gamma(\eta_h)} \\
&\quad \prod_{j=1}^{S} \frac{\nu^{\eta} \alpha_{jk}^{(m)\,\eta-1} e^{-\nu \alpha_{jk}^{(m)}}}{\Gamma(\eta)} && \text{(A.4)} \\
&\propto \Gamma(\eta_h)^{-MK} \nu_h^{\eta_h MK} \Gamma(\eta)^{-MKS} \nu^{\eta MKS} && \text{(A.5)} \\
&\quad \exp\left\{ - \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\nu_h h_k^{(m)} + \sum_{j=1}^{S} \nu \alpha_{jk}^{(m)} \right) \right\} \\
&\quad \prod_{m=1}^{M} \prod_{k=1}^{K} h_k^{(m)\,\eta_h - 1} \prod_{j=1}^{S} \alpha_{jk}^{(m)\,\eta - 1}
\end{aligned}
$$

A.2 Derivation of Equation 4.17

Multiplying $p(\vec{\alpha})$ by the determinant of Jacobian matrix of the inverse mapping $\vec{\lambda} \to \vec{\alpha}$, which is a diagonal matrix, yields $p(\vec{\lambda})$:

$$\begin{aligned}
p(\vec{\lambda}_1^{(1)},...,\vec{\lambda}_K^{(1)},\vec{\lambda}_1^{(M)}...,\vec{\lambda}_K^{(M)}) &\propto \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(h_k^{(m)};\eta_h,\nu_h)\prod_{j=1}^{S}\Gamma(\alpha_{jk}^{(m)};\eta,\nu)\,|\det\mathbf{J}_{\lambda\to\alpha}| \\
&\propto \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(h_k^{(m)};\eta_h,\nu_h)\prod_{j=1}^{S}\Gamma(\alpha_{jk}^{(m)};\eta,\nu) \\
&\quad \left|\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{j=1}^{S}\frac{\partial\, e^{\lambda_{jk}^{(m)}}}{\partial\lambda_{jk}^{(m)}}\right| \\
&\propto \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(h_k^{(m)};\eta_h,\nu_h)\prod_{j=1}^{S}\Gamma(\alpha_{jk}^{(m)};\eta,\nu) \\
&\quad \left|\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{j=1}^{S}e^{\lambda_{jk}^{(m)}}\right| \\
&\propto \prod_{m=1}^{M}\prod_{k=1}^{K}\Gamma(h_k^{(m)};\eta_h,\nu_h)\prod_{j=1}^{S}\Gamma(\alpha_{jk}^{(m)};\eta,\nu) \\
&\quad \left|\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{j=1}^{S}\alpha_{jk}^{(m)}\right| \\
&\propto \Gamma(\eta_h)^{-MK}\nu_h^{\eta_h MK}\Gamma(\eta)^{-MKS}\nu^{\eta MKS} \\
&\quad \exp\left\{-\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\nu_h h_k^{(m)}+\sum_{j=1}^{S}\nu\alpha_{jk}^{(m)}\right)\right\} \\
&\quad \prod_{m=1}^{M}\prod_{k=1}^{K}h_k^{(m)\,\eta_h-1}\prod_{j=1}^{S}\alpha_{jk}^{(m)\,\eta-1}\alpha_{jk}^{(m)} \\
&\propto \Gamma(\eta_h)^{-MK}\nu_h^{\eta_h MK}\Gamma(\eta)^{-MKS}\nu^{\eta MKS} \\
&\quad \exp\left\{-\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\nu_h h_k^{(m)}+\sum_{j=1}^{S}\nu\alpha_{jk}^{(m)}\right)\right\} \\
&\quad \prod_{m=1}^{M}\prod_{k=1}^{K}h_k^{(m)\,\eta_h-1}\prod_{j=1}^{S}\alpha_{jk}^{(m)\,\eta}
\end{aligned}$$

A.3 Derivation of Dirichlet-multinomial distribution density

Since Dirichlet is conjugate prior of multinomial, it is mathematically convenient such that the following integration has a closed form. The integration can be calculated using equations multinomial(4.6) and Dirichlet(4.9) mass and density functions

$$
\begin{aligned}
P(\vec{X}_i|\vec{\alpha}_k) &= \int P(\vec{X}_i|\vec{p}_i)\cdot \mathrm{Dir}(\vec{p}_i|\vec{\alpha}_k)d\vec{p}_i \\
&= \int \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!}\prod_{j=1}^{S} p_{ij}^{X_{ij}} \cdot \frac{1}{B(\vec{\alpha}_k)}\prod_{j=1}^{S} p_{ij}^{\alpha_{jk}-1}d\vec{p}_i \\
&= \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!}\frac{1}{B(\vec{\alpha}_k)}\cdot\int\prod_{j=1}^{S} p_{ij}^{X_{ij}+\alpha_{jk}-1}d\vec{p}_i \\
&= \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!}\frac{1}{B(\vec{\alpha}_k)}\cdot B(\vec{X}_i+\vec{\alpha}_k)\int\frac{1}{B(\vec{X}_i+\vec{\alpha}_k)}\prod_{j=1}^{S} p_{ij}^{X_{ij}+\alpha_{jk}-1}d\vec{p}_i \\
&= \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!}\frac{1}{B(\vec{\alpha}_k)}\cdot B(\vec{X}_i+\vec{\alpha}_k)\int \mathrm{Dir}(\vec{p};\vec{X}_i+\vec{\alpha}_k)\,d\vec{p}_i \\
&= \frac{J_i!}{\prod_{j=1}^{S} X_{ij}!}\frac{B(\vec{X}_i+\vec{\alpha}_k)}{B(\vec{\alpha}_k)}
\end{aligned}
\tag{A.6}
$$

In the fourth line, we multiply and divide the integrand by $B(\vec{X}_i+\vec{\alpha}_k)$ to achieve the posterior term $\mathrm{Dir}(\vec{p};\vec{X}_i+\vec{\alpha}_k)$ which integrates to one.

Since we assume that the probabilities of multiple data types of the same genomic loci are independent, the likelihood of a locus i with M data types can be calculated by multiplying likelihoods of individual data types

$$
\begin{aligned}
P(\vec{X}_i^{(*)}|\vec{\alpha}_k^{(*)}) &= \prod_{m=1}^{M}\int P(\vec{X}_i^{(m)}|\vec{p}_i^{(m)})\mathrm{Dir}(\vec{p}_i^{(m)}|\vec{\alpha}_k^{(m)})d\vec{p}_i^{(m)} \\
&= \prod_{m=1}^{M}\frac{J_i^{(m)}!}{\prod_{j=1}^{S} X_{ij}^{(m)}!}\frac{B(\vec{X}_i^{(m)}+\vec{\alpha}_k^{(m)})}{B(\vec{\alpha}_k^{(m)})}.
\end{aligned}
\tag{A.7}
$$

A.4 Derivation of Equation 4.33

Logarithm of multinomial beta function can be written using its representation in terms of gamma function

$$\begin{aligned} B(\vec{\alpha}) &= \frac{\prod_{j=1}^{S}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{S}\alpha_j)} & \text{(A.8)} \\ \log B(\vec{\alpha}) &= \log\left(\frac{\prod_{j=1}^{S}\Gamma(\alpha_{jk})}{\Gamma(\sum_{j=1}^{S}\alpha_{jk})}\right) & \text{(A.9)} \\ &= \sum_{j=1}^{S}\log\Gamma(\alpha_{jk}) - \log\Gamma(\sum_{j=1}^{S}\alpha_{jk}) \end{aligned}$$

To take the derivative of $E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]$ w.r.t $\alpha_{jk}^{(m)}$, first let's write expected log posterior (Equation 4.26) using only terms depend on $\alpha_{jk}^{(m)}$ and then rearrange the equation:

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{k=1}^{K}\Bigg\{E[z_{ik}]\Bigg\{\sum_{m=1}^{M}\Bigg[\Bigg(\sum_{j=1}^{S}\log\Gamma(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg)\\
&\qquad-\Bigg(\sum_{j=1}^{S}\log\Gamma(\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\}\Bigg\}\\
&\qquad+\sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg((\eta_h-1)\log h_k^{(m)}-\nu_h h_k^{(m)}+\sum_{j=1}^{S}\eta\log\alpha_{jk}^{(m)}-\nu\alpha_{jk}^{(m)}\Bigg)\\
&=\sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg\{\sum_{i=1}^{N}\Bigg\{E[z_{ik}]\Bigg[\Bigg(\sum_{j=1}^{S}\log\Gamma(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg)\\
&\qquad-\Bigg(\sum_{j=1}^{S}\log\Gamma(\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\}\Bigg\}\\
&\qquad+\sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg((\eta_h-1)\log h_k^{(m)}-\nu_h h_k^{(m)}+\sum_{j=1}^{S}\eta\log\alpha_{jk}^{(m)}-\nu\alpha_{jk}^{(m)}\Bigg)\\
&=\sum_{m=1}^{M}\sum_{k=1}^{K}\Bigg\{\sum_{i=1}^{N}\Bigg\{E[z_{ik}]\Bigg[\Bigg(\sum_{j=1}^{S}\log\Gamma(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg)\\
&\qquad-\Bigg(\sum_{j=1}^{S}\log\Gamma(\alpha_{jk}^{(m)})-\log\Gamma(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\}\\
&\qquad+\Bigg((\eta_h-1)\log h_k^{(m)}-\nu_h h_k^{(m)}+\sum_{j=1}^{S}\eta\log\alpha_{jk}^{(m)}-\nu\alpha_{jk}^{(m)}\Bigg)\Bigg\}
\end{aligned}\tag{A.10}$$

Using Psi (digamma) function $\psi(x)=\frac{d(\log\Gamma(x))}{dx}=\frac{\Gamma'(x)}{\Gamma(x)}$, we can write the derivative of Equation A.10 w.r.t $\alpha_{jk}^{(m)}$ as follows:

$$\begin{aligned}
\frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial\alpha_{jk}^{(m)}}&=\sum_{i=1}^{N}\Bigg\{E[z_{ik}\Bigg[\Bigg(\psi(X_{ij}^{(m)}+\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\Bigg)\\
&-\Bigg(\psi(\alpha_{jk}^{(m)})-\psi(\sum_{j=1}^{S}\alpha_{jk}^{(m)})\Bigg)\Bigg]\Bigg\}+(\eta_h-1)\frac{g_{jk}^{(m)}}{h_k^{(m)}}-\nu_h g_{jk}^{(m)}+\frac{\eta}{\alpha_{jk}^{(m)}}-\nu
\end{aligned}\tag{A.11}$$

A.5 Derivation of Equation 4.36

$$
\begin{aligned}
\frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \lambda_{jk}^{(m)}} &= \frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \alpha_{jk}^{(m)}} \frac{d\alpha_{jk}^{(m)}}{d\lambda_{jk}^{(m)}} \\
&= \frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \alpha_{jk}^{(m)}} \frac{1}{\frac{d\lambda_{jk}^{(m)}}{d\alpha_{jk}^{(m)}}} \\
&= \frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \alpha_{jk}^{(m)}} \frac{1}{\frac{1}{\alpha_{jk}^{(m)}}} \\
&= \frac{\partial E_{\mathbf{Z}}[\log P(Q,\mathbf{Z}|\mathbf{X})]}{\partial \alpha_{jk}^{(m)}} \alpha_{jk}^{(m)} \\
&= \alpha_{jk}^{(m)} \sum_{i=1}^{N} \left\{ E[z_{ik}] \left[\left(\psi(X_{ij}^{(m)} + \alpha_{jk}^{(m)}) - \psi(\sum_{j=1}^{S} X_{ij}^{(m)} + \alpha_{jk}^{(m)}) \right) \right.\right. \\
&\quad \left.\left. - \left(\psi(\alpha_{jk}^{(m)}) - \psi(\sum_{j=1}^{S} \alpha_{jk}^{(m)}) \right) \right] \right\} \qquad (A.12) \\
&\quad + \alpha_{jk}^{(m)} \left((\eta_h - 1) \frac{g_{jk}^{(m)}}{h_k^{(m)}} - \nu_h g_{jk}^{(m)} + \frac{\eta}{\alpha_{jk}^{(m)}} - \nu \right) \qquad (A.13)
\end{aligned}
$$

Appendix B

Soft k-means clustering algorithm

Algorithm B.1 Soft k-means clustering algorithm [24].

1. Cluster centers, denoted as $m^{(k)}$, are initialized using kmeans++ algorithm [2].

2. Membership probabilities of the data points are calculated using the following formula:
$$r_k^{(n)} = \frac{\exp(-\beta\, d(m^{(k)}, x^{(n)}))}{\sum_{k'} \exp(-\beta\, d(m^{(k')}, x^{(n)}))}$$
where $r_k^{(n)}$ shows the probability that data point $x^{(n)}$ is a member of cluster k. The only parameter of the method,β, defines the *stiffness* which makes the algorithm identical to hard k-means as it goes to infinity. $d(\cdot)$ function defines the distance which is the Euclidean distance in our case.

3. Cluster centers are computed based on the following equation:
$$m^{(k)} = \frac{\sum_n r_k^{(n)} x^{(n)}}{\sum_n r_k^{(n)}}$$

4. If the sum of differences between the previous cluster centers and last cluster centers are negligible stop, otherwise go to step 2.

Appendix C

Calculation of the Hessian matrix elements in Laplace approximation

Diagonal and off-diagonal elements of the Hessian matrix H in Equation 4.37 can be calculated using Equation 4.36 as follows:

$$
\begin{aligned}
-\frac{\partial^2 E_Z[F(Q,\mathbf{Z})]}{\partial \lambda_{jk}^{2\,(m)}} &= -\frac{\partial^2 E_Z[F(Q,\mathbf{Z})]}{\partial \lambda_{jk}^{(m)} \partial \alpha_{jk}^{(m)}} \alpha_{jk}^{(m)} \\
&= -\alpha_{jk}^{(m)} \sum_{i=1}^{N} \Bigg\{ E[z_{ik}] \Bigg[\Bigg(\psi(X_{ij}^{(m)} + \alpha_{jk}^{(m)}) - \psi(\sum_{j=1}^{S} X_{ij}^{(m)} + \alpha_{jk}^{(m)}) \Bigg) \\
&\quad - \Bigg(\psi(\alpha_{jk}^{(m)}) - \psi(\sum_{j=1}^{S} \alpha_{jk}^{(m)}) \Bigg) \Bigg] \Bigg\} \\
&\quad - \alpha_{jk}^{2\,(m)} \sum_{i=1}^{N} \Bigg\{ E[z_{ik}] \Bigg[\Bigg(\psi_1(X_{ij}^{(m)} + \alpha_{jk}^{(m)}) - \psi_1(\sum_{j=1}^{S} X_{ij}^{(m)} + \alpha_{jk}^{(m)}) \Bigg) \\
&\quad - \Bigg(\psi_1(\alpha_{jk}^{(m)}) - \psi_1(\sum_{j=1}^{S} \alpha_{jk}^{(m)}) \Bigg) \Bigg] \Bigg\} \\
&\quad - \alpha_{jk}^{(m)} \left((\eta_h - 1) \frac{g_{jk}^{(m)}}{h_k^{(m)}} - \nu_h g_{jk}^{(m)} + \frac{\eta}{\alpha_{jk}^{(m)}} - \nu \right) \\
&\quad - \alpha_{jk}^{2\,(m)} \left[(\eta_h - 1) \left(\frac{r_{jjk}^{(m)} h_k^{(m)} - g_{jk}^{2\,(m)}}{h_k^{2\,(m)}} \right) - \nu_h r_{jjk}^{(m)} - \frac{\eta}{\alpha_{jk}^{2\,(m)}} \right]
\end{aligned}
$$

$$\begin{aligned}
-\frac{\partial^2 E_Z[F(Q,\mathbf{Z})]}{\partial\lambda_{j'k}^{(m)}\partial\lambda_{jk}^{(m)}} &= -\frac{\partial^2 E_Z[F(Q,\mathbf{Z})]}{\partial\lambda_{j'k}^{(m)}\partial\alpha_{jk}^{(m)}}\alpha_{jk}^{(m)} \\
&\quad -\alpha_{jk}^{(m)}\alpha_{j'k}^{(m)}\sum_{i=1}^{N}\left\{E[z_{ik}]\left[\psi_1(\sum_{j=1}^{S}\alpha_{jk}^{(m)})-\psi_1(\sum_{j=1}^{S}X_{ij}^{(m)}+\alpha_{jk}^{(m)})\right]\right\} \\
&\quad -\alpha_{jk}^{(m)}\alpha_{j'k}^{(m)}\left((\eta_h-1)\frac{r_{jj'k}^{(m)}h_k^{(m)}-g_{jk}^{(m)}g_{j'k}^{(m)}}{h_k^{2\,(m)}}-\nu_h r_{jj'k}^{(m)}\right)
\end{aligned}$$

where $r_{abk}^{(m)}$ represents an element of the Hessian matrix of regulation term $h_k^{(m)}$ with indices a and b. Therefore

$$r_{abk}^{(m)} = \frac{\partial g_{ak}^{(m)}}{\partial\alpha_{bk}^{(m)}} = \frac{\partial^2 h_k^{(m)}}{\partial\alpha_{ak}^{(m)}\partial\alpha_{bk}^{(m)}}$$

Hessian matrix of $h_k^{(m)}$, which is denoted by $r_{abk}^{(m)}$, is given below:

$$\begin{bmatrix}
2 & -2 & 0 & \cdots & 0 & 0 \\
-2 & 4 & -2 & 0 & \cdots & 0 \\
0 & -2 & \ddots & \ddots & \ddots & \vdots \\
\vdots & 0 & \ddots & \ddots & -2 & 0 \\
0 & \vdots & \ddots & -2 & 4 & -2 \\
0 & 0 & \cdots & 0 & -2 & 2
\end{bmatrix}$$

Super- and sub-diagonal elements are -2 since the first partial derivative of Dirichlet parameters (last case in Equation 4.34) between the first and last ones also depend on the previous and next Dirichlet parameters.

Appendix D

Promoter and enhancer mean signals

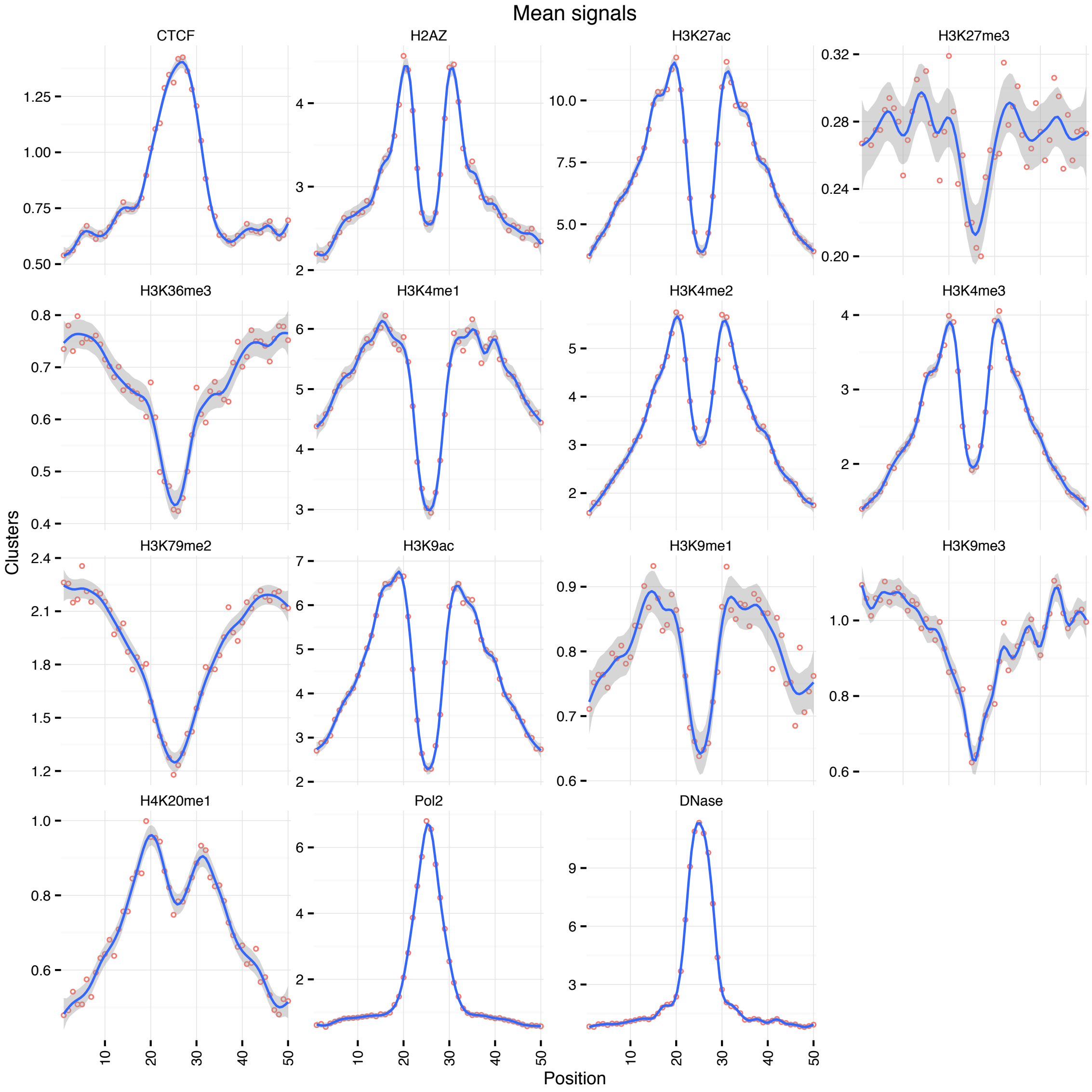


Figure D.1: Enhancer mean signals plotted separately for different data types

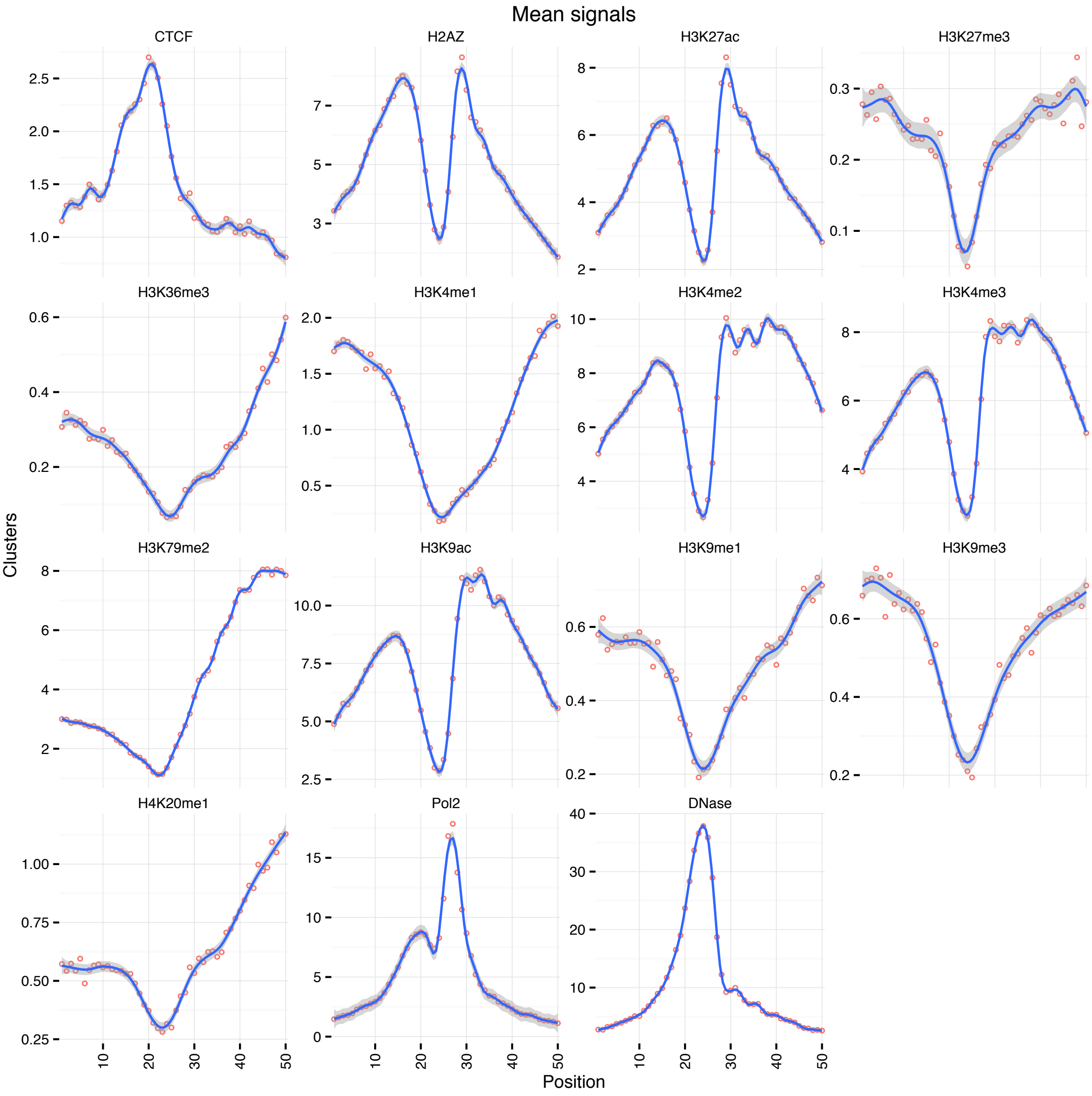
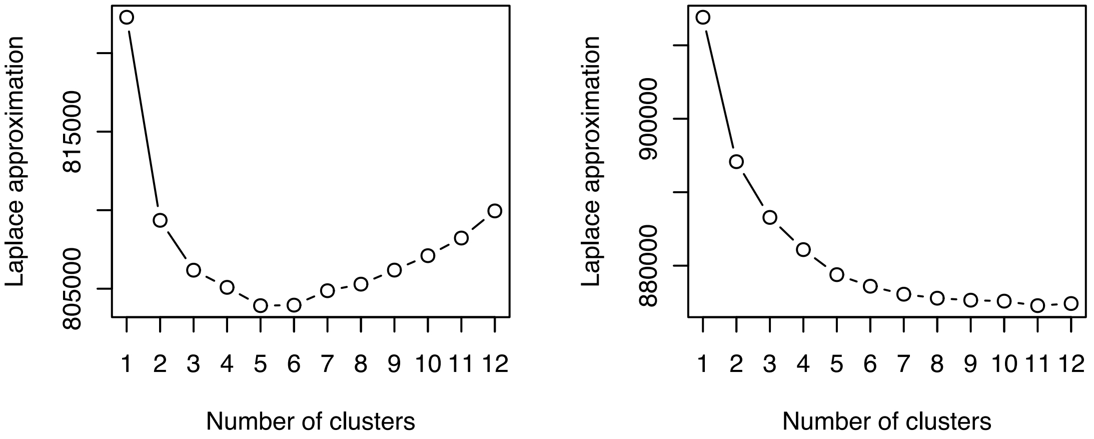


Figure D.2: Promoter mean signals plotted separately for different data types

Appendix E

Laplace approximation results



(a)

(b)

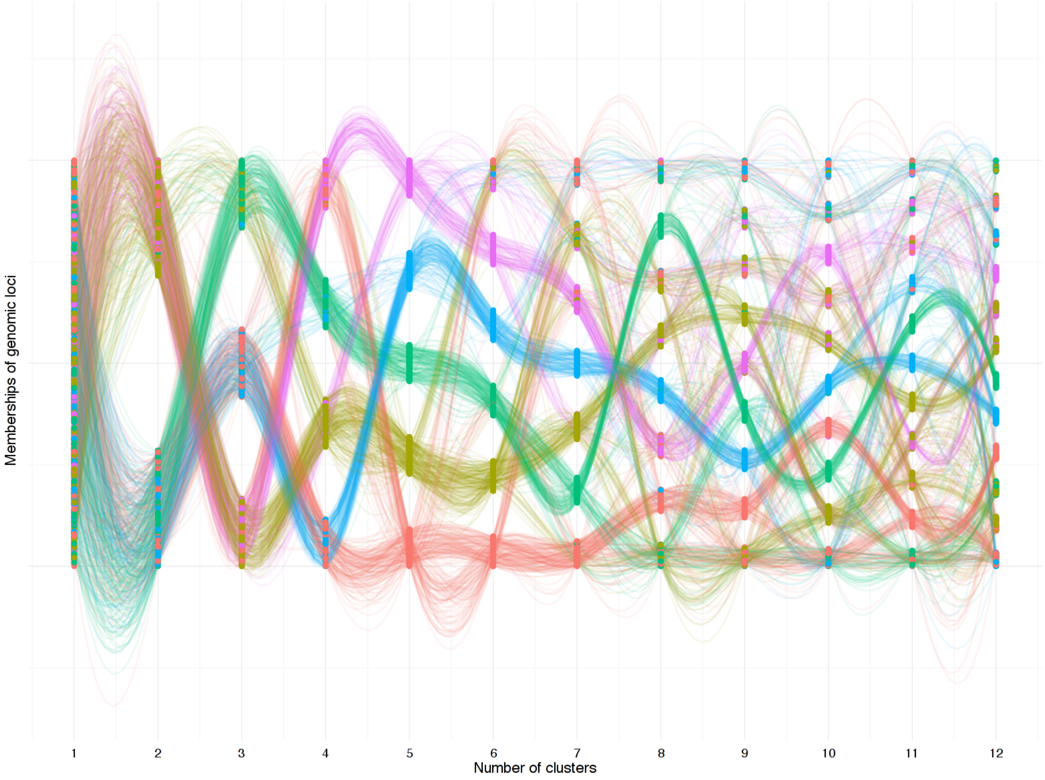
Figure E.1: Laplace approximation goodnes-of-fit values for the mixture models modeling the enhancer(a) and promoter(b) data.

Appendix F

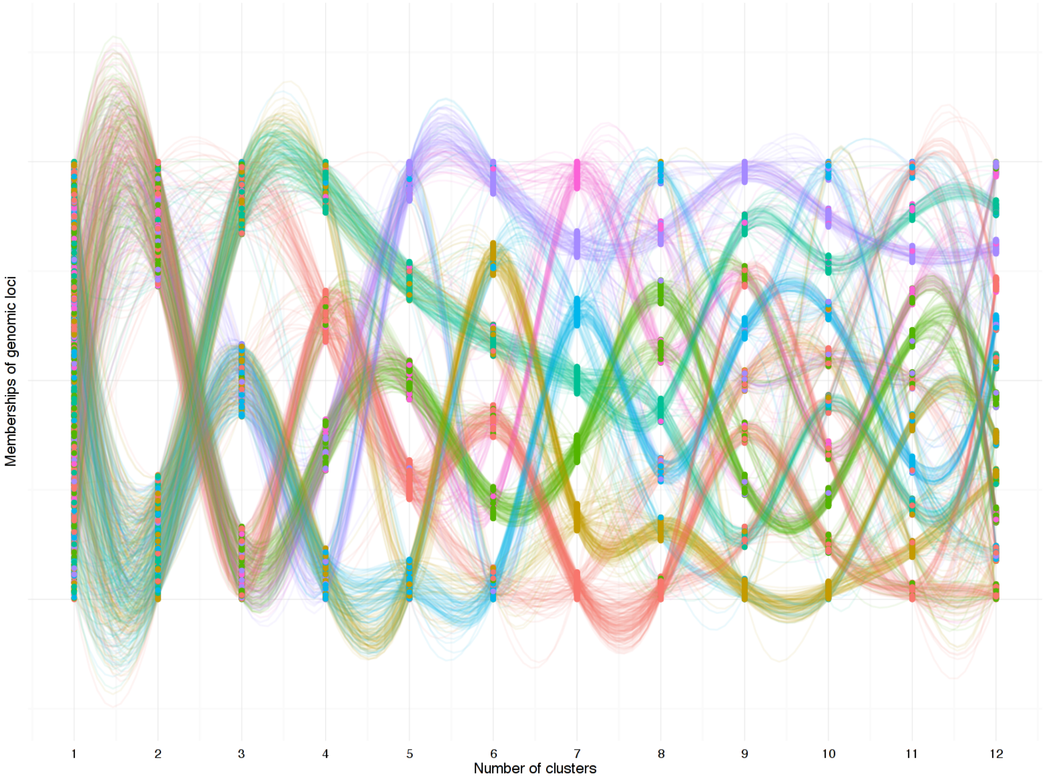
Change of labels across different clusterings

Visualizing which cluster a genomic locus falls into when the clustering is performed with various number of components may provide useful information. To create such a visualization, after adding an amount of jitter to integer cluster labels, we plotted the points that represent the labels of genomic loci of interest on the parallel coordinate system such that each vertical line (axes of parallel coordinate system) shows a different clustering result. Next, a spline interpolation was performed to connect the labels of a locus across different clustering results. Colors of the splines are chosen based on the result of clustering with optimum number of clusters determined by BIC, which are 5 and 7 for enhancer and promoter clusterings, respectively. The cluster labels in each vertical line are ordered in a way that the cluster label with the highest number is at the top.

Observing that some loci consistently fall into the same cluster in different clustering results suggests that this group of loci is a major cluster which is more distinct and separable than the others. For instance, in Figure F.1a, 4 out of 5 clusters, with the exception being the fifth cluster represented in purple color, are consistent throughout the clustering results with more than 5 components. In other words, no major bifurcations are observed for these clusters. Fewer such cases are observed in promoter clustering results i.e. clusters 3 and 6 given in green and purple colors.



(a) Enhancer labels



(b) Promoter labels

Figure F.1: Cluster labels of each regulatory element resulting from clusterings with different number of components

Appendix G

Random generation of artificial data

Artificial data can be generated through Gaussian functions described in Section 4.6. However, three parameters characterizing a bell-shaped curve, namely μ, σ and $scale$ are required to be specified manually for each data type and each cluster. Specifying these parameters by hand can be cumbersome. Therefore, we defined probability distributions appropriate for each parameter to make the process truly random. The description of the generative process showing how to create these parameters randomly is given in Algorithm G.1. Generated Dirichlet parameters can then be used for generating the artificial data by sampling from the mixture of Dirichlet-multinomial compound with generated parameters. Mean profiles of the clusters of the data generated through this scheme is shown in Figure G.1.

Algorithm G.1 Generative process for creating Dirichlet-multinomial parameters in a random fashion

```
1. Number of clusters is drawn from a zero-truncated Poisson: $K \sim \text{ztPois}(\lambda_K)$
2. For each data type:
    (a) For each cluster:
        i. Sample number of peaks in a window from a zero-truncated Poisson:
           $nPeaks \sim z\text{tPois}(\lambda_{peak})$
        ii. For each peak $i$ in $nPeaks$:
            A. Sample $\mu_i$ parameter which is in $[0,1]$ interval: $\mu_i \sim \text{Beta}(\alpha_\mu, \beta_\mu)$
            B. Sample $\sigma_i$ parameter: $\sigma_i \sim \text{Beta}(\alpha_\sigma, \beta_\sigma)$
            C. Sample $scale_i$ parameter: $scale_i \sim \text{Gamma}(\alpha_{scale}, \beta_{scale})$
3. Sample Dirichlet-multinomial mixture weights from a symmetric Dirichlet:
   $\pi \sim \text{Dirichlet}(\alpha_\pi \mathbf{1})$
```

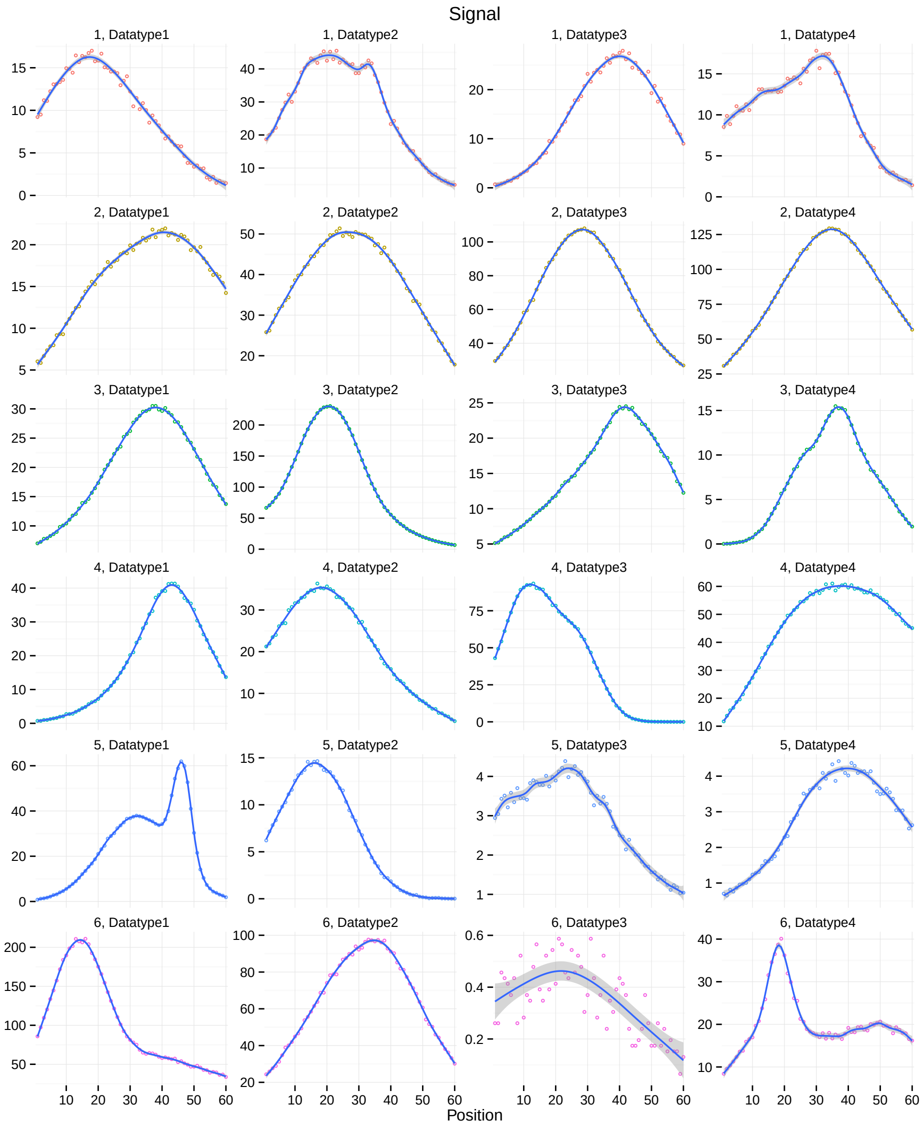


Figure G.1: Cluster mean signals of artificially generated data in 50bp resolution. Rows represent different clusters whereas columns correspond to different artificial data types.