Unsupervised Continuous Feature Annotation of the Human Genome

by

Habib Daneshpajouh

B.Tech, University Technology PETRONAS, 2017

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in the School of Computing Science Faculty of Applied Sciences

© Habib Daneshpajouh 2020
SIMON FRASER UNIVERSITY
Spring 2020

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Approval

Name: Habib Daneshpajouh
Degree: Master of Science (Computing Science)
Title: Unsupervised Continuous Feature Annotation of the Human Genome

Examining Committee: Chair: Manolis Savva
                        Assistant Professor

                       Kay C Wiese
                       Co-Supervisor
                       Associate Professor

                       Maxwell W Libbrecht
                       Co-Supervisor
                       Assistant Professor

                       Martin Ester
                       Examiner
                       Professor
                       School of Computing Science

Date Defended: Apr 14, 2020
Abstract

Genome annotation methods are widely used to understand the function of the genome. For example, they can be used to identify the activity of a genomic position that is associated with a disease. Existing genome annotation methods produce discrete annotations that assign a single label to each genomic position. However, these discrete annotation methods have several limitations. For example, these methods cannot easily represent varying strengths of genomic elements, and they cannot easily represent combinatorial elements that simultaneously exhibit multiple types of activity. To remedy these limitations, an annotation strategy is proposed that instead outputs a vector of chromatin state features at each position. Also a method, epigenome-ssm is proposed to annotate the genome with chromatin state features. It is shown that chromatin state features from epigenome-ssm are more useful for several downstream applications than both continuous and discrete alternatives, including their ability to identify expressed genes and enhancers.

Keywords: unsupervised learning; epigenomics; sequencing-based assays; genome annotation; human genome; continuous modelling.
Dedication

To my parents, my wife, and my brother, for being a true family.
Acknowledgements

First of all, I would like to thank my co-supervisors Dr. Kay C Wiese and Dr. Maxwell W Libbrecht. I was lucky to have two wonderful mentors in my graduate school. Kay guided me since the first day that I joined SFU and made sure that I stayed on the right track to finish my master’s fruitfully and on time. Max introduced me to the problem of genome annotation and patiently helped me in grasping the subtleties of this area. Both Max and Kay provided me with financial support for my study and for attending relevant conferences and workshops.

I would like to thank Dr. Martin Ester for serving as my examiner and Dr. Manolis Savva for serving as the chair for my thesis defence. Also, a special thanks to Dr. Leonid Chindelevitch for helping me in realizing my true potential for graduate studies right at the beginning of my master’s degree.

Needless to say, a group project will not succeed without the contribution and dedication of each and every member; so I would like to thank my peers who worked with me in this project. Particularly, I want to thank Bowen Chen for his contributions to design and implementation of the models; and Shohre Masoumi and Neda Shokraneh for their contributions to visualizations.
"...the value of the genome is only as good as its annotation."

–Lincoln Stein
## Table of Contents

Approval ................................................. ii  
Abstract ................................................... iii  
Dedication ................................................... iv  
Acknowledgements ........................................ v  
Quotation ..................................................... vi  

Table of Contents .......................................... vii  
List of Figures ................................................ ix  

1 Introduction .............................................. 1  
1.1 Overview ............................................... 1  
1.2 Summary ............................................... 3  
1.3 Collaborator contributions ............................. 3  

2 Background and Literature Review ...................... 5  
2.1 A high-level introduction to basics of molecular biology and genomics for genome annotation .......................... 5  
2.2 Genomics and epigenomics datasets .................... 6  
2.3 Genome annotation ...................................... 8  
2.3.1 Overview ........................................... 8  
2.3.2 Processing pipeline .................................. 8  
2.3.3 Unsupervised techniques for annotation .......... 9  
2.3.4 Discrete vs. continuous annotation ................. 9  
2.3.5 Single cell-type vs. multiple cell-type annotation .. 10  
2.3.6 Cell-type-specific vs. cell-type-agnostic annotation .. 10  
2.4 A survey of previous work on genome annotation .... 10  
2.5 Deeper insights into main works on cell-type-agnostic genome annotation .......................... 12  
2.5.1 Combined Annotation-Dependent Depletion (CADD) ........................................... 12
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.2</td>
<td>Deep Learning-based Sequence Analyzer (DeepSea)</td>
<td>12</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Segway Encyclopedia</td>
<td>12</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Avocado</td>
<td>13</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary</td>
<td>14</td>
</tr>
<tr>
<td>2.7</td>
<td>Collaborator contributions</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Methods</td>
<td>15</td>
</tr>
<tr>
<td>3.1</td>
<td>Reference genome</td>
<td>15</td>
</tr>
<tr>
<td>3.2</td>
<td>ChIP-seq histone modification data sets</td>
<td>15</td>
</tr>
<tr>
<td>3.3</td>
<td>State space model</td>
<td>16</td>
</tr>
<tr>
<td>3.4</td>
<td>Non-negativity constraint</td>
<td>16</td>
</tr>
<tr>
<td>3.5</td>
<td>Hidden Markov model (HMM)</td>
<td>17</td>
</tr>
<tr>
<td>3.6</td>
<td>Alternative models</td>
<td>18</td>
</tr>
<tr>
<td>3.7</td>
<td>Gene expression evaluation</td>
<td>19</td>
</tr>
<tr>
<td>3.8</td>
<td>Enhancer evaluation</td>
<td>19</td>
</tr>
<tr>
<td>3.9</td>
<td>Source code</td>
<td>19</td>
</tr>
<tr>
<td>3.10</td>
<td>Summary</td>
<td>20</td>
</tr>
<tr>
<td>3.11</td>
<td>Collaborator contributions</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>Results</td>
<td>21</td>
</tr>
<tr>
<td>4.1</td>
<td>Chromatin state features are predictive of gene expression</td>
<td>21</td>
</tr>
<tr>
<td>4.2</td>
<td>Chromatin state features at enhancer elements are predictive of enhancer activity</td>
<td>23</td>
</tr>
<tr>
<td>4.3</td>
<td>Chromatin state features recapitulate known genome biology</td>
<td>24</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>26</td>
</tr>
<tr>
<td>4.5</td>
<td>Collaborator contributions</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>Discussion and Conclusion</td>
<td>33</td>
</tr>
</tbody>
</table>

Bibliography | 36
List of Figures

Figure 1.1  Chromatin state feature annotation ........................................ 4
Figure 3.1  Gene expression evaluation ...................................................... 20

Figure 4.4  **(a-c) Distribution of promoter features.** Each point corresponds to a gene. Teal points correspond to genes in the top 33% of gene expression; orange points correspond to genes in the lower 66%. Horizontal and vertical axes indicate average values of two features respectively at the gene’s promoter (average value in the 10 kb upstream of the TSS). For the sake of visibility, the plot shows a random 10% subset of all genes. **(d-f) Distribution of enhancer features.** Same as (a-c), but expression corresponds to CAGE enhancer activity. All plots are generated using data from human lymphoblastoid cell line. ........................................ 28
Figure 4.5  **Distribution of promoter and enhancer features.** Each point corresponds to either an enhancer or promoter depending on its color. The plots show the distribution of these elements with respect to different pairs of chromatin state features. All plots are generated using data from human lymphoblastoid cell line. ................................. 29
Figure 4.6  **(a) Distribution of feature values. (c) Autocorrelation of features. (e) Empirical cumulative distribution function (ECDF) of features.** All plots are generated with $k = 3$ and using data from eight cell types. ........................................ 30
Figure 4.7  **(a) Relationship of features to the input genome data. (b) Relationship between features in neighboring positions. (c-e) Averaged enrichment of chromatin state features with relative position to TSS, TTS and middle of enhancer respectively.** All plots are generated with $k = 5$ using data from eight cell types. ................................. 31
Figure 4.8  (a) Distribution of feature values. (c) Autocorrelation of features. (e) Empirical cumulative distribution function (ECDF) of features. All plots are generated with $k = 5$ and using data from eight cell types.
Chapter 1

Introduction

1.1 Overview

Sequencing-based genomic assays can measure many types of genomic biochemical activity, including transcription factor binding, chromatin accessibility, transcription, and histone modifications. Thanks to The Encyclopedia of DNA Elements (ENCODE), Roadmap Epigenomics and other international projects working on massively parallel sequencing of the human DNA, data from sequencing-based genomic assays is now available from hundreds of human cellular conditions, including varying tissues, individuals, disease states, and drug perturbations.

Segmentation and genome annotation (SAGA) methods are widely used to understand genome activity and gene regulation. These algorithms take as input a collection of sequencing-based genomic data sets from a particular tissue. They output an annotation of the genome that assigns a label to each genomic position. These annotations are widely used to understand the function of the genome. For instance, they can be used to identify the activity of a genomic position that is associated with a disease. SAGA methods are unsupervised; they discover categories of activity (such as promoters, enhancers or genes) without any prior knowledge of known genomic elements and a human interprets these categories, similar to a clustering algorithm.

Many SAGA methods have been proposed for discrete chromatin state label annotation [5, 12, 22, 31, 53, 37, 60, 4]. The primary model used by these methods is the hidden Markov model (HMM). An HMM is a probabilistic model that assumes that there is a latent (unknown) chromatin state label at each position, that the observed genomic data sets are generated as a function of this label and that the label at position \( i \) depends on the label at position \( i - 1 \). Later work extended this basic approach in a number of ways. First, there are three methods for modeling the input genomic data: one can binarize the data and model 0/1 values with a Bernoulli distribution [12], use a continuous measure of signal strength such as fold enrichment over control modelled with a Gaussian distribution [22, 53], or model raw read counts with a negative Binomial distribution [37]. Second, some methods
use statistical marginalization to handle unmappable regions. Third, several strategies
exist for modelling segment lengths or for producing annotations on multiple length scales
[22, 31]. Finally, several strategies have been proposed to guide the choice of the number of
labels [53, 60, 57, 4].

A related class of joint annotation methods aim to improve epigenome annotations
by simultaneously annotating many cell types and sharing position-specific information
between the annotations [57, 58, 1, 34, 33]. Such joint annotations can be more accurate,
but have the drawback that they may mask differences among cell types. We do not consider
the joint annotation task in this work, but adapting the continuous annotation approach to
this task is a promising direction for future work.

Another related task aims to take data from all available cell types as input to produce
a single cell-type-agnostic (as opposed to cell-type-specific) annotation, a task sometimes
known as “stacked” annotation. Several methods have been proposed to produce a discrete
[23, 36] and continuous [8, 52] cell-type-agnostic annotations.

As mentioned above, majority of existing SAGA methods output a discrete annotation
that assigns a single label to each position. This discrete annotation strategy has several
limitations. First, discrete annotations cannot represent the strength of genomic elements.
Variation among genomic elements in intensity or frequency of activity of cells in the sample
is captured in variation in the intensity of the associated marks. Such variation is lost if
all such elements are assigned the same label. In practice, SAGA methods often output
several labels corresponding to the same type of activity with different strengths, such as
“Promoter” and “WeakPromoter” [23, 12]. Second, a discrete annotation cannot represent
combinatorial elements that simultaneously exhibit multiple types of activity. To model
combinatorial activity, a discrete annotation must use a separate label to represent each
pair (or triplet etc.) of activity types. For example, intronic enhancers usually exhibit marks
of both transcription and regulation [23]. However, representing all possible combinations of
activity types with discrete labels would require a number of labels that grows exponentially
in the number of activity types.

In this work, a continuous genome annotation strategy is proposed (Fig. 1.1). That
is, the proposed method outputs a vector of real-valued chromatin state features for each
genomic position, where each chromatin state feature putatively represents a different type
of activity. Continuous chromatin state features have a number of benefits over discrete
labels. First, chromatin state features preserve the underlying continuous nature of the input
signal tracks, so they preserve more of the information present in the raw data. Second, in
contrast to discrete labels, continuous features can easily capture the strength of a given
element. Third, chromatin state features can easily handle positions with combinatorial
activity by assigning a high weight to multiple features. Fourth, chromatin state features
lend themselves to expressive visualizations because they project complex data sets onto a
small number of dimensions that can be shown in a plot. For these reasons, in other fields,
continuous modeling is often preferred over discrete. For example, the widely-used method of topic modeling for text documents assigns a continuous weight to each of a number of categories (such as “sports” or “politics”) for each document [28].

In this work the utility of chromatin state feature annotation is explored. Several measures of the quality of a chromatin state feature annotation are proposed and the performance of several alternative methods are compared according to these quality measures. A non-negative Kalman filter state space model, epigenome-ssm, is proposed for this problem that produces the highest-quality continuous annotations of the methods that are compared.

1.2 Summary

In summary, this chapter introduced the task of genome annotation and its applications in analyzing the functions and activities of the human genome. Moreover, limitations of the existing annotation methods were explained and a novel method to address those issues was proposed. In the following, chapter 2 provides a background for genome annotation and surveys the major works in this area. Chapter 3 presents the proposed method for continuous annotation of the genome and several methods for evaluating the resulting annotations. The evaluation results and interpretations of the continuous annotations are presented in Chapter 4. Finally, chapter 5 concludes the thesis and provides some directions for future work.

1.3 Collaborator contributions

This thesis is based on a group project that includes four students and two supervisors. The students are Habib Daneshpajouh (HD), Bowen Chen (BC), Neda Shokraneh (NS), and Shohre Masoumi (SM). The supervisors are Maxwell W Libbrecht (MWL) and Kay C Wiese (KCW). The contributions of each member is mentioned at the end of each chapter.
Figure 1.1: **Chromatin state feature annotation.** epigenome-SSM takes as input a set of genomics assays, each represented as a real-valued track over the genome. It outputs a set of real-valued chromatin state features for each position in the genome, using a Kalman filter state space model.
Chapter 2

Background and Literature Review

2.1 A high-level introduction to basics of molecular biology and genomics for genome annotation

The instructions needed for developing and directing the activities of nearly all living organisms are encoded in their deoxyribonucleic acid (DNA), a chemical compound that is made of four nucleotide bases: adenine (A), thymine (T), guanine (G) and cytosine (C). These nucleotide bases are often called base pairs (bp), because they form a paired structure by bonding with each other. The order of the base pairs determine the meaning of the information that is encoded in each segment of DNA. A complete set of an organism’s DNA is called genome. Each and every single cell contains a complete copy of the genome. The human genome is comprised of 23 pairs of thread-like structures of DNA called chromosomes.

One of the main functions of each cell is the production of proteins. Units of DNA called genes contain the instructions for making proteins. Each human chromosome may contain tens to thousands of protein-coding genes. Besides genes, enzymes and a family of molecules called ribonucleic acids (RNAs) play important roles in protein production. The gene’s information is transcribed into a RNA molecule which will further be translated into a protein. The process in which a gene’s information is used in producing a gene product is called gene expression.

For various reasons, genes should only be expressed whenever necessary. The regulation of gene expression is handled by some structural regions of DNA called transcriptional regulators. The main transcriptional regulators are promoters, enhancers and silencers. Promoters are short regions of genes that contain transcription start sites (TSS). Enhancers are longer than promoters and can be located up to 1,000,000 bp away from the gene, though the chromatin structure of DNA is folded in a way that enhancers are spatially located near the promoter of the gene. By binding to transcription factors (TFs) (i.e. the proteins that control transcription rate by binding to transcriptional regulators), enhancers increase the likelihood of a gene being transcribed. Silencers on the
other hand act opposite to the enhancers by binding to those TFs that prevent a gene from transcription. Such a TF is called **repressor**.

**Eukariotic** cells (human cells for instance) unlike **prokaryotic** cells (bacterial cells for instance) contain a complex of DNA and protein called **chromatin** that packs very long sequences of DNA into a more compact and stable form by wrapping them around some proteins called **histones**. The chromatin components (i.e. DNA and histones) can be modified by some **epigenetic modifications** that result in **phenotype** (i.e. observable characteristics or traits) changes without changing the **genotype** (i.e. the actual DNA sequence). Two of the most well-known such modifications are **DNA methylation** and **histone modification**. In DNA methylation, methyl groups are added to DNA and may result in repressing the gene transcription if it is located at the gene’s promoter. Likewise, histone modification is the process of modifying histone proteins that can impact gene expression.

Although DNA sequence is safe from being altered by epigenetic modifications, it can indeed be modified by **genetic variations (mutations)**. The most common types of genetic variation are **single-nucleotide polymorphism (SNP)**, **single-nucleotide variation (SNV)** and **copy-number variation (CNV)**. SNPs are variations of a single nucleotide in a specific position of the genome that occur with a tangible frequency (e.g. more than one percent) within a population. On the other hand, SNVs are the SNPs that can occur with any level of frequency. Lastly, CNVs refer to the case where some segments of the genome are repeated. The genome-wide study of association between genetic variations and particular phenotype traits (like major human diseases) in different individuals is called **genome-wide association study (GWAS)**. Genome annotation is a useful tool in GWAS.

Further information on fundamentals of molecular biology and genomics can be found in [24] and [40].

### 2.2 Genomics and epigenomics datasets

The largest collaborative biological project in the world, the **Human Genome Project**\(^1\) (HGP) (1984-2003) [29] was an effort to determine the complete set of the DNA sequence in the human genome. A HGP’s follow-up project was the **Encyclopedia of DNA Elements**\(^2\) (ENCODE) (2003-) [39, 9, 6], aiming to identify functional elements in the human genome. Another project in this area was the **Roadmap Epigenomics Project**\(^3\) which aimed to complement the data in the ENCODE project. These projects produced a vast amount of data that enable scientists to gain deeper insights into the human genome.

---

\(^1\)https://www.genome.gov/human-genome-project

\(^2\)https://www.encodeproject.org

\(^3\)http://www.roadmapepigenomics.org
A vital component for the success of the above-mentioned projects was the high-throughput DNA sequencing techniques [46]. Using these sequencing techniques, several methods were developed to study the human genome, including chromatin immunoprecipitation sequencing (ChIP-seq) for transcription factor binding, DNase I hypersensitive sites sequencing (DNase-seq) for chromatin structure and RNA sequencing (RNA-seq) for RNA transcribed regions, among others.

Using ChIP-seq method for instance, one can assay histone modifications of a particular protein in the cells of a particular epigenome[^4]. The assay process in this method starts by using formaldehyde (or a different alternative) to glue all the proteins that are bound to the DNA together with the DNA itself. The DNA sequence is then cut into small fragments (approximately 300 bp). Next, an antibody is used to isolate the protein of interest while everything else is washed away. The formaldehyde used in the initial step is reversed by warming up the fragments. At this stage, the sole DNA fragments can be isolated by washing away all the proteins (including the histones).

All the above-mentioned steps are separately applied to all the chromosomes in the cell. A typical ChIP-seq assay is performed on a pool of about six million cells, that ends up with a lot of DNA fragments (or DNA reads) from a lot of cells. The next step is to add sequencing adapters to both ends of the DNA fragments to build a sequencing library. After some cloning and cleaning processes, the high quality reads (usually between 50 and 100 million reads) are gathered and aligned to a reference genome, so that the genomic position for each DNA fragment is identified. Eventually, a long list of genomic coordinates along with the number of reads for each position will be produced. Obviously, a higher number of reads for a position indicates that this particular position is more involved with bonding the protein of interest.

The tracks of ChIP-seq assays can be visualized using tools like the UCSC Genome Browser[^5] for further downstream analysis. This genome browser is compatible with a majority of file formats for genomic data including bedGraph, GTF, PSL, BED, bigBed, WIG and bigWig among others, while providing sophisticated tools for many types of genomic and genetic analysis such as mapping and sequencing, genes and gene prediction, phenotype and literature, mRNA and EST, expression, regulation, comparative genomics, variation, and repeats.

The Roadmap Epigenomics currently possesses data of several assays for 127 human epigenomes. The data files are genome-wide and contain a signal value for each genomics position which indicates the extend to which that position participates in the particular

[^4]: In the Roadmap Epigenomics Project terminology, an epigenome refers to a particular tissue, cell type or cell line.

[^5]: [https://genome.ucsc.edu](https://genome.ucsc.edu)
activity that the assay is looking for. The data files are usually in bigwig format and their
names start with an epigenome identifier followed by the name of the assay.

A data file named "E116-H3K4me3.fc.signal.bigwig" for instance, belongs to the assay
H3K4me3 which measures histone modification in epigenome 116 (human lymphoblastoid
cells). The "H" in the name of the assay stands for histone, "3" is the type of histone, "K" is
lysine (the amino acid that is modified) followed by a number which is the residue of that
amino acid, and finally me3 is the type of modification (tri-methylation in this case). The
assay names are followed by either "fc" or "pval" abbreviations which refer to fold change
and p-value pre-processing transformation respectively.

2.3 Genome annotation

2.3.1 Overview

Genomes, and in particular, the human genome, contains many functional and non-functional
elements that lead the biochemical activities in the cell. Whether it is a particular disease,
certain super abilities or any other phenotypic phenomenon, these may always be caused by
a genomic activity. Activities such as protein production and genomic variations may have
serious phenotypic impacts. However, the huge size of the genome makes it a non-trivial
task to get a comprehensive picture of genomic elements and their activities. Genome an-
notation could be a first but extremely useful step in generating a more compact view of
the genome. This compact representation removes much of the redundancy in the genome
data while being information-rich.

2.3.2 Processing pipeline

The complex nature and large size of genomics data makes it necessary to go through some
pre-processing steps. These steps could vary from some typical processes on data such as
cleaning, normalization and filtering to more context specific techniques like representing
the data in various spatial resolutions. Looking into the genome in different granularity
levels turns out to be of great importance. For instance, one observes different genomic
behavior at domain scale (about one to few thousands bp) compared to nucleosome scale
(about one to few hundreds bp). Some annotation methods are able to operate at very fine
scales, for example, at 1 bp.

The annotation process includes three main steps: model training, annotation, and eval-
uation. In the training step, the training data, that is usually a small subset of the actual
dataset is passed to the model to learn its parameters. The learned model will then be used
to annotate the whole data by assigning a label to each position of the genome. The aim
here is to have labels that represent an actual biological phenomenon. Hence, a challenge
here is to choose the best number of labels that leads to a meaningful annotation. Although
having more labels will increase the model accuracy, most of the works choose a number of
labels that makes more biological sense and is easier to interpret, i.e. a number between 3 and 20.

The next step is to evaluate the annotations. The evaluation step can be tricky as one should consider the type of input data and the main target of annotation before choosing appropriate evaluation techniques. Generally, one wants the resulting annotation to be a good predictor of regulatory phenomena such as gene expression and enhancer activity while recapitulating known genome biology. One may require the annotations to not only be a bird’s eye view of the genome and its activities, but a set of features for further downstream analysis of such activities and their impacts.

The unsupervised annotation methods mentioned above require a post-processing step to interpret the generated labels, i.e. mapping each label to a particular biological phenomenon. The interpretation step could be one of the most tricky parts of the annotation process and it requires expert knowledge in molecular biology. Sometimes, having a hierarchical architecture for the labels helps in reducing the complexity and to ease the interpretation. In this type of architecture, there may be several main categories of labels, each with its own sub-categories. The interpretation can be done either manually by the expert or automatically, for example by a machine learning method that learns from previous manual interpretations.

### 2.3.3 Unsupervised techniques for annotation

Unsupervised algorithms based on Hidden Markov Model (HMM) and Bayesian Network (which is a generalization of HMM) are widely used for genome annotation. Basically, these are sequential methods that look at each position of the genome as a component that its characteristics depend on the characteristics of its neighboring positions, and its activities may be positively or negatively correlated with activities of other positions depending on how far apart they are from each other. This strategy of looking at the genome as a collection of dependent components turns out to be able to better capture the genome coordinates. An introduction to HMMs and Bayesian Networks can be found in [43, 17].

### 2.3.4 Discrete vs. continuous annotation

Genome annotation methods can be divided into two groups based on the type of labels that they produce. Discrete annotation methods assign a single-element label (an integer label for example) to each genomic position. This type of annotation is similar to clustering or segmentation. Although being simple to understand and easier to interpret, discrete annotation has several limitations. For instance, there may be two elements with enhancer activity, but one with a weak level and the other with a strong level of such activity. Discrete methods will either ignore such a difference in strength of activity or generate separate labels for each. Likewise, there may be some elements that exhibit multiple types of activity. For a discrete method to label such an element, it needs a separate label for representing each
group of activities, which results in an exponential relationship between the number of labels required and number of activities.

In contrast, continuous annotation methods assign a vector of features to each genomic position. This strategy is more similar to dimensionality reduction where one wants to reduce the number of features for each observation in the dataset to come up with a lower-dimensional representation of data, though in the case of continuous annotation, it is technically possible to have more annotation features than the original input features. By having more than one feature for each position, continuous methods provide a great level of flexibility in the annotations that enables them to represent various strengths and types of activity for genomic elements. However, the interpretation of continuous annotations are more difficult than discrete ones.

2.3.5 Single cell-type vs. multiple cell-type annotation

On the input side, one can perform genome annotation on data from only a single cell type. For instance, multiple data tracks from several assays can be placed in an input matrix, and an annotation method can be used to assign a label (using discrete methods) or a vector of features (using continuous methods) to each genomic position. However, with having more and more genomics data available, it is natural to put data tracks from multiple cell types together for annotation.

2.3.6 Cell-type-specific vs. cell-type-agnostic annotation

On the output side, one can design the annotation method in such a way that it produces separate annotations for each cell type presented in the input data. This approach is called cell type-specific annotation and enables the user to analyze each cell type in isolation. However, from a biological point of view, it makes more sense to produce a single annotation of the genome regardless of which cell type one is referring to. Such a cell type-agnostic annotation provides a collective view of the regulatory elements in genome.

2.4 A survey of previous work on genome annotation

Early works on genome annotation focused on cell type-specific annotation of well-studied cell lines. [25] presents a genome-wide annotation of Drosophila melanogaster based on data from 18 histone modifications, and an integrative analysis with other data from non-histone chromatin proteins, DNase I hypersensitivity, GRO-Seq reads produced by engaged polymerase, and short/long RNA products. ChromHMM [12, 14] is a widely-used tool for annotation and visualization of chromatin states based on their biological functions and correlations with large-scale functional datasets which is based on a multivariate hidden Markov model. In [22], the authors used Segway, an annotation tool based on Dynamic
Bayesian Networks (DBNs) to annotate chromatin data from different experiments on human chronic myeloid leukemia cell line, such as histone modifications, transcription-factor binding and open chromatin. [3] proposes a new version of Segway that incorporates Gaussian mixture models and mini-batch training.

In a large-scale study [23], an integrative analysis of chromatin states is performed on several human cell lines available from the ENCODE datasets. A method based on tiered hidden Markov model is proposed in [30] for modelling epigenomic patterns over multiscales applied to a ChIP-seq dataset of human embryonic stem cells. A more recent work on multi-scale chromatin state annotation is proposed in [38] that applies diHMM, a hierarchical hidden Markov model to ChIP-seq data. EpiCSeq [37] is another algorithm that was designed to annotate histone modification data by applying a probabilistic model to ChIP-seq read counts. Yet another work on learning chromatin states is presented in [21] that uses HMMs combined with a model selection method called factorized information criteria. A computational tool called chromstaR is proposed in [55] for mapping genomics regions to discrete combinatorial chromatin states based on the presence or absence of various histone modifications. [60] proposed a data-driven probabilistic modelling of chromatin factors dependencies that results in a systematic map of Drosophila chromatin states.

One of the early works on cell type-agnostic annotation is presented in [1] where a Bayesian network is used to extract chromatin states from histone modifications data on nine human cell types. A general annotation framework called Combined Annotation-Dependent Depletion (CADD) is proposed in [26] that for each possible human genetic variant, it aims to integrate different annotations into a single score. DeepSEA is a deep learning-based sequence model proposed in [59] to predict the effects of non-coding functional variants. The NIH Roadmap Epigenomics Consortium published one of the most comprehensive annotations on 111 reference human epigenomes [27] analyzing histone modification patterns, DNA accessibility, DNA methylation and RNA expression. An integrative and discriminative epigenome annotation system (IDEAS) was proposed in [57] that jointly characterizes epigenetic landscapes and detect differential regulatory regions in many cell types. In [58], the IDEAS method is applied to the same data used in [27] to have a reproducible functional maps of 127 human cell types.

An annotation strategy is proposed in [35] that includes a fully automated pipeline starting from taking the epigenomics datasets and ending at providing annotations for each position of the genome. A deep neural network tensor factorization methods called Avocado was proposed in [52] for jointly annotating data from several cell types while having the ability to impute the missing data. Other methods for genome annotation with the ability to impute missing data are ChromImpute [13] and PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition (PREDICTD) [8]. A method called enhancer HMM (eHMM) [56] was specifically designed to predict enhancers from ChIP-seq data using a supervised hidden Markov model.
2.5 Deeper insights into main works on cell-type-agnostic genome annotation

2.5.1 Combined Annotation-Dependent Depletion (CADD)

A general annotation framework named Combined Annotation-Dependent Depletion (CADD) is proposed in [26] that for each possible human genetic variant (for example SNV or indel events), it aims to integrate different annotations into a single score called "C score". Unlike traditional annotation metrics that focus on molecular functionality and pathogenicity, the C score is considered as a metric for deleterious variants (i.e. those variants that cause damage to the organism and will eventually be depleted by natural selection). The experiments in this work were done on annotations of different data types such as conservation metrics (GERP, phastCons and phyloP), regulatory information from genomics regions of DNase hypersensitivity and transcription factor binding, transcript information, and finally protein-level scores generated by Grantham, SIFT and PolyPhen. They examined the best performing individual annotations and also the correlations between different annotations.

Despite its usefulness from many aspects, CADD has some limitations. For instance, although the C score measurement correlates with deleteriousness, it is affected by some other factors such as local mutation rate and background selection that can potentially reduce its accuracy. The latest updates to CADD including its most recent version (v1.4) are reviewed in [45].

2.5.2 Deep Learning-based Sequence Analyzer (DeepSea)

DeepSEA is a deep learning-based sequence model proposed in [59] to predict the effects of non-coding functional variants. DeepSEA aims to provide a flexible quantitative platform for modelling complex dependencies of transcription factor binding and histone modifications from DNA sequence which may then be used to estimate the functional effects of non-coding variants. This method learns regulatory sequence code by learning to simultaneously predict histone modification, DNase I sensitivity and transcription factor binding using ENCODE and Roadmap epigenomics datasets.

DeepSEA framework surpassed the performance of the best methods (of its time) for predicting chromatin prediction. However, this method focused on single nucleotide variants in isolation while it is now more evident that most phenotypic effects are results of a group of single nucleotide variations.

2.5.3 Segway Encyclopedia

An annotation strategy is proposed in [35] that includes a fully automated pipeline starting from taking the epigenomics datasets and ending at providing annotations (with biological labels assigned to them) for each position of the genome. In this work, they used ChIP-seq, DNase-seq and Repli-seq from the ENCODE and Roadmap Epigenomics consortia.
for annotation, while RNA-seq expression data was used for evaluation. This work trained separate unsupervised annotation models on each cell type as opposed to previous works that used a 'concatenated' approach where a single model is shared among all the cell types. One advantage of the proposed approach over the concatenated approach is that it does not require each cell type to have data from the same set of assays.

However, one downside of this so-called 'independent' approach is that it requires independent interpretation of states for each cell type, which could be cumbersome in the case of a large number of cell types. To overcome this burden, at the end of the pipeline, a machine learning classifier is used to interpret the annotations by assigning a known biological label (such as 'promoter', 'enhancer', 'inactive region' etc.) to each annotation. This machine learning approach turns out to be able to recapitulate manual interpretation of annotation states that was previously done. However, a minor drawback of this approach is that it is bounded to a fixed set of known biological labels and not able to discover new ones.

As part of the proposed strategy in this work, a functionality score is evaluated for each locus to differentiate functional activities from non-functional biochemical ones. This functionality score aims to depict how important a given type of activity is to the organism’s phenotype.

### 2.5.4 Avocado

A deep neural network tensor factorization technique called Avocado was proposed in [52] to compress epigenomics data into a more compact, information-rich representation of the human epigenome. The aim of using such a method was to reduce redundancy, bias and noise in data that will result in a representation with most of its variance covering meaningful biological differences. The input for this method is a three-dimensional tensor of epigenomics data with the orthogonal axes being cell type, assay type, and genomic position. Tensor factorization and a deep neural network is used to decompose this input tensor and learn a lower-dimensional latent representation of it.

Avocado was applied to the Roadmap DNase-seq and ChIP-seq datasets containing only those assays that had data available for at least five cell types. The performance of this method was compared with two other methods for epigenomics data imputation: ChromImpute and PREDICTD. In most of the evaluations, Avocado performed either better than or equal to the other methods. The latent representation learned showed to be a good predictor of gene expression activities and promoter-enhancer interactions. However, a downside of non-sequential methods like Avocado is that these methods look at each position of the genome independently as opposed to sequential methods (like HMMs or recurrent neural networks) which look at the genome as a sequence of dependent elements. The latter strategy is biologically more plausible.
2.6 Summary

In this chapter, some of the most frequently-used terms in molecular biology were concisely reviewed. The concept, processing pipeline and types of genome annotation were explained. Moreover, this chapter reviewed main approaches to the problem of genome annotation and went through well-known methods that incorporate each approach. Particularly, the main strategies for modelling input and output data were reviewed. Also, previous work on genome annotation were surveyed, starting from the early works on a limited set of assays and cell types, all the way forward to more recent works on joint annotations of multiple cell types. Finally, the main methods for cell-type-agnostic annotation were reviewed in detail.

2.7 Collaborator contributions

HD did the literature review. KCW helped in designing the literature review.
Chapter 3

Methods

3.1 Reference genome

All analysis are performed on the human reference genome hg19. To improve computational efficiency, following previous work [12, 36], the genome is divided into 200 bp bins and all analysis are performed at the bin level. Note that the 200 bp resolution was chosen to facilitate comparison to previous work; however a finer resolution (e.g. 50 or 100 bp) is recommended in practice. Following previous work [22], to train the model, the ENCODE Pilot regions [11] were used, which cover about 1% of the human genome. Also, the ENCODE blacklist regions [10] were removed from all analysis.

3.2 ChIP-seq histone modification data sets

The ChIP-seq histone modification data sets are downloaded from the Roadmap Epigenomics data portal [27]. See [27] for a full description of the data processing pipeline. Briefly, reads are mapped to the reference genome, shifted and extended according to the fragment length and compared to an input control. The signal at a given position is represented as the negative log P-value of the observed read count compared to input, according to a Poisson model [27]. To reduce the influence of large outliers, following previous work [22], ChIP-seq values were transformed using the transformation arc sinh(x) = log(x + \sqrt{x^2 + 1}). Data from five histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) in the human lymphoblastoid cell line GM12878 [48] were used. For further analysis and interpretation, more data from eight cell types were included in the analysis: H1-hESC cells, IMR90 fetal lung fibroblasts cell line, A549 EtOH 0.02pct lung carcinoma cell line, human lymphoblastoid cell line, HUVEC umbilical vein endothelial primary cells, K562 leukemia cells, NHEK-Epidermal keratinocyte primary cells, and NHLF lung fibroblast primary cells using a concatenated approach.
3.3 State space model

A Kalman filter state space model [7] is developed for annotating the genome with chromatin state features. This model takes as input a vector of $E$ observed genomic data sets for each position, $y_g \in \mathbb{R}^E$, for $g \in 1 \ldots G$. This model assumes that at position $g$ there is a latent vector $\alpha_g \in \mathbb{R}^M$ that encodes the chromatin state features of that position. It assumes that the observed data vector at that position $y_g$ is generated as a linear function of $\alpha_g$ plus Gaussian noise,

$$y_g = Z\alpha_g + \epsilon_g \quad \epsilon_g \sim N(0, I). \quad (3.1)$$

It further assumes that the latent vector $\alpha_{g+1}$ is generated as a linear function of $\alpha_g$ plus Gaussian noise,

$$\alpha_{g+1} = T\alpha_g + v_g \quad v_g \sim N(0, I). \quad (3.2)$$

To learn the SSM model, the Expectation-maximization (EM) algorithm is used to maximize the log likelihood of the model as a function of its parameters, $Z \in \mathbb{R}^{E \times M}$ and $T \in \mathbb{R}^{M \times M}$. Briefly, this algorithm alternates two steps, the E step and the M step. In the E step, $Z$ and $T$ are held fixed and a message-passing algorithm is used to efficiently estimate $\alpha_{1:g}$ and compute sufficient statistics for updates to $Z$ and $T$. In the M step, these sufficient statistics are used to update $Z$ and $T$. The $Z$ matrix is initialized to Uniform$(0,1)^{E \times M}$ while the $T$ matrix is initialized to $I_M$.

To limit the model’s capacity to overfit and its sensitivity to local optima, several $L_2$ regularization terms were added to the optimization’s objective function $J(Z, T)$, which encourage $Z$ and $T$ to have small values:

$$J(Z, T) = \log P(\alpha, Z, T|Y) + \lambda_1 \|Z\|_F + \lambda_2 \|T\|_F. \quad (3.3)$$

The E step was parallelized over a cluster by dividing the genome into windows, computing sufficient statistics from each window on worker nodes, and combining the sufficient statistics to compute updates to $Z$ and $T$ on a master node.

3.4 Non-negativity constraint

A version of the proposed model is developed, epigenome-ssm-nonneg, in which the chromatin state features $\alpha_g$ and the emission parameters $Z$ are both constrained to be non-negative.

To optimize $Z$, an active set method of Lagrange multipliers was used to enforce the non-negativity constraint [20]. Specifically, a Lagrange multiplier term is added to the objective function:
\[
J^\Lambda(Z, T, \Lambda, \Lambda_\alpha) = J(Z, T) + \text{tr}(\Lambda_Z^T \Lambda Z).
\] (3.4)

The active set method takes advantage of the property of complementary slackness: \(\Lambda_{Ze, m} Z_{e, m} = 0\). At each iteration, a list of parameters with non-negative Lagrange multipliers is maintained. Parameters are added or removed from the active set when the optimization assigns them a zero or non-zero value respectively. At each iteration, the Lagrange multipliers associated with each parameter in the active set are updated in order to stop the parameters from becoming negative.

For \(\alpha\), non-negativity constraint is added by projecting the optimized state to lie in the constraint space, where \(\hat{\alpha}^P\) is the constrained estimate:

\[
\hat{\alpha}^P_{g|g} = \arg\min_{\alpha} \left\{ (\alpha - \hat{\alpha}_{g|g})^T (\alpha - \hat{\alpha}_{g|g}) : A\alpha = 0 \right\}
\]

\[A_{ij} = 1 \text{ iff } i = j \text{ and } \hat{\alpha}_{g|g} < 0 \text{ otherwise } A_{ij} = 0\] (3.5)

\(\hat{\alpha}_{g|g}^P\) is the constrained value, and \(\hat{\alpha}_{g|g}\) is the unconstrained Kalman Filter update estimate at time \(g\).

Given the optimized state \(\hat{\alpha}^P_{g|g}\), the constraint matrix \(A\) is constructed as equation 3.5 and \(b\) is made as a zero vector. Then the best constrained estimate is given by:

\[
\hat{\alpha}^P_{g|g} = \hat{\alpha}_{g|g}^* - A^T (AA^T)^{-1} (A \hat{\alpha}_{g|g} - b)
\] (3.6)

### 3.5 Hidden Markov model (HMM)

The most common model used in previously-proposed genome annotation methods is the hidden Markov model (HMM). The HMM assumes that at position \(g\) there is a latent state \(x_g \in \{1 \ldots k\}\) that represents the chromatin state label of that position. It assumes that the observed data is generated as a function of the latent state \(x_g\) (see next paragraph). It further assumes that the state at position \(g\) depends just on the state at position \(g - 1\), and that state \(l\) transitions to state \(l'\) with probability \(\phi_{l,l'}\)

\[
P(x_g = l'|x_{g-1} = l) = \phi_{l,l'} \sum_{l'=1}^k \phi_{l,l'} = 1
\] (3.7)

In this work, two versions of the HMM model are considered, which take as input continuous or discrete data respectively. As described in Chapter 2, some existing methods take continuous input [5, 22] while others take discrete input [16, 12]. In the continuous input case, as with the state space model, the continuous-input HMM takes as input a vector of \(m\) observed genomic data sets for each position, \(y_g \in \mathbb{R}^m\), for \(g \in 1 \ldots T\). It assumes that there is a mean vector associated with each state \(\mu_k\), and that the observed
data vector equals $\mu_k$ plus Gaussian noise

$$y_g = \mu_x + \epsilon_g \quad \epsilon_g \sim N(0, \Sigma_{x_g}^2).$$  \hspace{1cm} (3.8)

In the discrete input case, input data is thresholded into binary values such that the input data at position $g$ is represented by a binary vector $\bar{y}_g \in \{0,1\}^m$. The discrete-input HMM assumes that the observed data is generated as a multivariate Bernoulli distribution. That is, for track $i$ at position $g$,

$$P(\bar{y}_{i,g} = 1) = \theta_{x_k,i}.$$  \hspace{1cm} (3.9)

To learn the HMM model, the EM algorithm is used to maximize the log likelihood of the model as a function of the model’s parameters: $\mu_{1:k}, \Sigma_{1:k}$ and $\phi_{1:k,1:k}$ for the continuous-input model, and $\theta_{1:m,1:k}$ and $\phi_{1:k,1:k}$ for the discrete-input model.

Moreover, two possible ways of representing the output of the HMM model are considered. Existing methods generally output a discrete value for the inferred state $x_g$ as the chromatin state label at position $g$. However, an HMM can be re-purposed to output continuous chromatin state features by defining the chromatin state feature $\ell$ at position $g$ as $\alpha_{\ell,t} = P(x_g = \ell)$.

In summary, two choices for input (continuous input with a Gaussian distribution or discrete input with a Bernoulli distribution) and two choices for output (output continuous chromatin state features or discrete chromatin state labels) are considered, resulting in four HMM variants: HMMgaus-con, HMMgaus-dis, HMMber-con and HMMber-dis.

The Python package pomegranate [51] was used for all training and inference of HMM models.

### 3.6 Alternative models

For the purpose of evaluation, epigenome-ssm is compared with well-known annotation methods including chromHMM and Segway. Additionally, three other alternative models are employed. First, the raw input tracks were simply used as the chromatin state features. To produce a chromatin feature annotation with $k$ features, a subset of $k$ of the input tracks were randomly selected. Second, principle component analysis (PCA) [41] and non-negative matrix factorization (NMF) [54] were investigated. Both methods take as input a vector $y_g \in \mathbb{R}^m$ and output a vector of features $\alpha_g \in \mathbb{R}^k$ to optimize a likelihood function. They differ from the SSM and HMM methods in that PCA and NMF treat each position independently, without considering the genome coordinate. NMF differs from PCA in that it outputs non-negative values $\alpha_{i,t} \geq 0$. The implementations of PCA and NMF from scikit-learn [42] were used in the evaluations.
3.7 Gene expression evaluation

Following previous work [58], an annotation is evaluated according to the strength of correlation between the labels at a promoter of a given gene and that gene’s expression. The RNA-seq data is downloaded from the Roadmap Epigenomics data portal [47].

A linear regression model is used to evaluate the degree to which annotations at a promoter region are predictive of gene expression (Fig. 3.1). Two types of evaluation models are trained: region-specific models and a whole-gene model. For the region-specific models, each gene is divided into 10 evenly-spaced bins. Also, a bin is defined for each 1 kb interval out to 5 kb upstream of the gene’s transcription start site (TSS) and 5 kb downstream of the gene’s transcription termination site (TTS), for a total of 20 bins.

For each bin, a linear regression model is trained. As the regression feature vector, the average feature vector in the respective bin is used. For discrete annotations, a one-hot encoding is used—that is, the feature vector has a 1 in the position corresponding to the label and 0’s elsewhere. As the regression response value, the RNA-seq RPKM (reads per kb per million mapped reads) values is used. As with ChIP-seq data, RPKM values are transformed with an arcsinh transformation. To measure the predictive power of a regressor, the fraction of variance explained ($r^2$, also known as the coefficient of determination) is used. To control for the complexity of the regressor, the standard adjustment $ar{r}^2 = 1 - (1 - r^2)(n-1)/(n-p-1)$ is used, where $n$ and $p$ are the number of examples and regressor parameters respectively.

For the whole-gene model, another linear regression model is applied on the entire region [TSS, TTS]. This model is identical to the region-specific models, except that the average feature vector is computed over the entire region.

3.8 Enhancer evaluation

Following previous work [58], an annotation is evaluated according to the strength of correlation between the labels at an enhancer and the strength of activity of that enhancer. The enhancer data based on Cap Analysis of Gene Expression (CAGE) is downloaded from FANTOM5 website [15]. This analysis is similar to the gene expression evaluation. As the regression response value, the arcsinh transformation of average TPM (tags per million mapped reads) values is taken from three replicates of GM12878 (FANTOM library ids CNhs12331, CNhs12332, and CNhs12333) as the response values. As with the gene expression evaluation, the average feature vector within the enhancer region is used as the feature vector.

3.9 Source code

Source code for the epigenome-ssm model, other alternative models, and all the experiments and evaluations is available on Github [18].
Figure 3.1: **Gene expression evaluation.** Chromatin state features are passed as input features to a linear regression model to predict RNA-seq data.

### 3.10 Summary

This chapter presented all the methods used in this work. Particularly, the SSM and HMM models were explained in detail. Also, the non-negativity constraint on state features and emission matrix of the SSM model using Lagrange multipliers and active set method was presented. Finally, different techniques were proposed to evaluate the quality of chromatin state features in predicting important genomic phenomena. In other words, vectors of chromatin state features can be used as input features to a linear regression model in order to predict gene expression and enhancer activity.

### 3.11 Collaborator contributions

BC initially implemented all the methods. HD and BC tested and debugged the methods. HD extended the methods to work with multiple cell types (in concatenated mode), performed model refinement, developed new evaluation methods and tested the whole processing pipeline on large datasets.
Chapter 4

Results

4.1 Chromatin state features are predictive of gene expression

To compare annotation methods, RNA-seq gene expression data is used to evaluate the annotations, following previous work [57, 36] (Section Gene expression evaluation). Briefly, in a high-quality annotation, high-expression and low-expression genes should be annotated in a distinct way. In other words, there should be a strong correlation between the annotation at the promoter and the gene’s expression level, as measured by RNA-seq. Specifically, the quality of an annotation is evaluated according to the variance explained by a predictive model that takes as input the annotation state at a gene’s promoter and outputs predicted RNA-seq expression. Four variants of a HMM-based method are compared, with the intention to represent existing methods like ChromHMM (Bernoulli) [12] and Segway (Gaussian) [22], as well as two dimensionality reduction methods and, for completeness, Segway and ChromHMM themselves. For chromHMM, both the segmentation (chromHMM-dis) and posterior (chromHMM-con) results are considered.

It was found that all models are predictive of gene expression, but the SSM model outperforms alternatives by this measure (Fig 4.1a). An epigenome-ssm annotation explains noticeably more variance (Adj $r^2 = 0.66$ for $k = 5$) than a discrete HMM, whether or not the HMM models continuous input with a Gaussian distribution or models discrete input with a Bernoulli distribution (0.52 and 0.23 respectively). This is also true for Segway, ChromHMM-dis and ChromHMM-con (0.62, 0.58 and 0.59 respectively). Part of this improvement is a result of the greater richness of a continuous model— if a continuous annotation is produced from the HMM model by using the probability of each state, the performance improves in most cases of $k$. Even though adding non-negativity constraints to the SSM reduces the model’s restricted optimization space, it was found that doing so would not significantly worsen the performance on this task; the performance of epigenome-ssm(nonneg) is quite close to that of epigenome-ssm, particularly for $k \geq 8$. 

21
Figure 4.1: Evaluation of annotations relative to gene expression and enhancer activity (using data from the human lymphoblastoid cell line). (a) Vertical axis is the fraction of variance in gene expression explained ($r^2$, Gene expression evaluation section). Horizontal axis is the number of features or states in a given continuous or discrete model respectively ($k$). Each line corresponds to a model. (b) Same as (a), but using enhancer activity from FANTOM5 (Enhancer evaluation section). (c) Same as (a), but using the annotation at a specific position relative to the gene rather than the entire TSS as the predictor (Gene expression evaluation section). The horizontal axis indicates which region relative to the gene was used as a predictor. All methods use $k = 3$.

Moreover, the superior performance of the SSM model is maintained even when the HMM model uses more labels than the number of SSM features (Fig 4.1a). Even for the largest number of labels that was tried ($k = 15$), the performance of any HMM model was no higher than $r^2 = 0.57$, which is lower than the performance of the SSM model with 3 features (0.58). This comparison offsets the potential disadvantage that continuous features are more complex than discrete labels. According to this analysis, if one is interested in obtaining a very simple annotation, it is preferable to use a continuous model with a small number of features rather than a discrete model with many labels.

It was also found that different types of annotations differed in which positions relative to the transcription start site were predictive of expression (Fig 4.1c). As expected, positions near the transcription start site (TSS) are most predictive of expression, whereas annotations at positions over 3 kp from the TSS have less predictive power. Notably, however,
discrete annotations at the TSS itself are no more predictive of expression than annotations at its flanks. This is likely because the discrete methods label the TSS as such regardless of whether it is expressed, whereas the continuous features can express both whether or not the position is a TSS and whether or not it is expressed. This behaviour shows how a continuous model can incorporate longer-range dependencies between positions. Continuous PCA and NMF models perform worse than epigenome-ssm(nonneg) in the whole-gene model, indicating the value of using a model that incorporates the genomic axis, although PCA and NMF perform slightly better in the region-based regression (Figure 4.1c).

4.2 Chromatin state features at enhancer elements are predictive of enhancer activity

These annotation methods are further evaluated by measuring how predictive each annotation is of experimentally-validated enhancer elements, again following previous work [57]. FANTOM5 enhancer RNA data was used as a measurement of the activity of each enhancer element. Briefly, in a high-quality annotation, highly active and inactive enhancers should be annotated in a distinct way. In other words, there should be a strong correlation between the annotation at an enhancer and that enhancer’s activity level, as measured by CAGE. Specifically, the quality of an annotation is evaluated according to the variance explained by a predictive model that takes as input the annotation state at an enhancer and outputs the predicted CAGE activity.

Similar to prediction of gene expression, it was found that all models are predictive of enhancer activity but epigenome-ssm outperforms them on this evaluation metric too (Figure 4.1b). In overall, for all the values of $k$, the variance explained by epigenome-ssm annotations is the highest among all the methods. Specifically, an epigenome-ssm annotation with $k = 5$ ($r^2 = 0.18$) is superior in predicting enhancer activity than all the alternative methods even with much larger values of $k$ except for Segway with $k = 15$ ($r^2 = 0.19$). Also, despite the results of the epigenome-ssm(non-neg) being slightly lower than the unconstrained version, it is still among the highest ones, particularly for larger values of $k$. On the other hand, from the perspective of output type, continuous versions of each method performed better than discrete versions (see the results of chromHMM, HMMber and HM-Mgaus). Also, increasing the value of $k$ did not necessarily help the annotations to be a better predictor of enhancer activity; most of the methods have their best or near the best performance with $k = 8$. 

23
4.3 Chromatin state features recapitulate known genome biology

While the results above show quantitatively that chromatin state features are predictive of many genomic phenomena, it was additionally found that these features qualitatively recapitulate known genome biology (Figures 4.2-4.6). The focus here is on an annotation using three features ($k=3$).

Feature 1 is a mark of repression, characterized by the repressive histone modifications H3K27me3 and H3K9me3 (Figure 4.2a) [23]. Thus, this feature is found at the promoters of repressed genes and is absent from promoters of active genes (Figure 4.4b). This feature shows a distinctive pattern at transcription termination sites (TTSs) that shows enrichment just before the TTS and 2.5k bp downstream and depletion immediately following the TTS (Figure 4.3b).

Feature 2 is a mark of transcribed genes, characterized by the histone modification H3K36me3 (Figure 4.2a) [23]. Thus, this feature is found downstream of transcription start sites, within the gene body (Figure 4.3a).

Feature 3 is a mark of regulatory activity, characterized by open chromatin (DNase-seq) and the regulation-associated histone modifications H3K4me3 and H3K27ac (Figure 4.2a) [23]. This feature occurs at transcription start sites and enhancers (Figure 4.3a, 4.3c). Opposite to feature 1, feature 3 is found at the promoters of active genes and is absent from promoters of repressed genes (Figure 4.4b). Also, feature 3 is found at active enhancers and is absent from inactive enhancers (Figure 4.4f).

As discussed in the first chapter, continuous chromatin state features are very useful for visualization purposes. Figure 4.5 shows the distribution of promoter and enhancer features with respect to different pairs of chromatin state features where each point corresponds to either an enhancer or promoter depending on its color. Although a bit noisy, one can see the separation of promoters and enhancers in those plots.

Moreover, Figure 4.6 provides some additional insights into distribution of the state features. In particular, Figure 4.6a shows that all the three features have a value of zero or close to zero in majority of the genomic positions. The same view is provided by the empirical cumulative distribution function (ECDF) plot in Figure 4.6c (note that an ECDF value at each point of the plot indicates the fraction of the feature values that are less than or equal to the value on the horizontal axis). This indicates that majority of genomic positions are not involved in any functionality.

One may be interested in analyzing the value of each feature at different genomic positions. For this purpose, autocorrelation plots of the three features for up to 10k positions are given by Figure 4.6b. In this figure, feature 1 and feature 2 have a high autocorrelation in longer distances, which means that each of those features at position $a$ is highly correlated with itself at position $b$ where $a$ and $b$ are far apart from each other. This indicates that
Figure 4.2: (a) **Relationship of features to the input genome data.** Color corresponds to the mean signal value of a given assay at positions annotated with a given feature. (b) **Relationship between features in neighboring positions.** Color in cell \(i, k\) represents the correlation of feature \(i\) at one position with feature \(k\) in the following position.

Feature 1 and feature 2 represent large genomic elements, for instance genes or quiescent DNA. In contrast, feature 3 has a low autocorrelation in longer distances, meaning that it represents small genomic elements, for instance promoters and enhancers. The autocorrelation result is in line with the interpretations in the previous paragraphs that feature 1 is a mark of repression, feature 2 is a mark of transcribed genes, while feature 3 is a mark of general regulatory activity.

A chromatin state annotation with five features \((k=5)\) shows similar patterns, but can express more granularity in activity types. In particular, the five-feature model includes features that represent enhancer- and promoter-specific regulatory activity (Figures 4.7-4.8). As in the case of the three-feature model, a combination of different plots can be used to interpret the associated elements and activities of each feature.

In that regard, feature 2 is a mark of promoter activity, characterized by transcription-associated histone modification H3K4me2 (Figure 4.7a). Thus, this feature is found at transcription start sites (Figure 4.7c) and has very low autocorrelation in long distances (Figure 4.8b).

Feature 3 is a mark of quiescent regions, given the fact that there is no strong association between this feature and any of the histone modifications in Figure 4.7a. Thus, in Figure 4.8a the plot for feature 3 around the value zero has the highest peak among all the features. This can be seen by the ECDF plot in Figure 4.8c as well, where the value of feature 3 in more than 80% of the positions is shown to be zero.

Feature 4 is a mark of transcribed genes, characterized by the histone modification H3K79me2 (Figure 4.7a). Thus, this feature is found downstream of the transcription start sites, within the gene body (Figure 4.7c). This can also be seen in Figure 4.8b where feature 4 has high autocorrelation in long distances.
4.4 Summary

This chapter presented evaluation results of the chromatin state features produced by epigenome-ssm. Those evaluations showed that continuous features are more powerful than discrete labels in analyzing human genome. First of all, chromatin state features proved to be more predictive of gene expression and enhancer activity compared to alternative methods. Moreover, continuous features were able to recapitulate known genome biology phenomena such as gene expression, repression, transcription start sites, enhancer activity and so on.

4.5 Collaborator contributions

HD performed data pre-processing and evaluation experiments. MWL and HD performed the interpretation of the three-feature model and HD performed the interpretation of the five-feature model. NS performed the visualizations for Figure 4.3 and 4.7c-e. SM performed the visualizations for Figure 4.6 and 4.8. All the other visualizations are done by HD and BC.
Figure 4.3: (a-c) Averaged enrichment of chromatin state features with relative position to TSS, TTS and middle of enhancer respectively. Vertical axis indicates average signal of feature at position relative to locus. All features have a genome-wide mean of 1. All plots are generated using data from eight cell types.
Figure 4.4: (a-c) **Distribution of promoter features.** Each point corresponds to a gene. Teal points correspond to genes in the top 33% of gene expression; orange points correspond to genes in the lower 66%. Horizontal and vertical axes indicate average values of two features respectively at the gene's promoter (average value in the 10 kb upstream of the TSS). For the sake of visibility, the plot shows a random 10% subset of all genes. (d-f) **Distribution of enhancer features.** Same as (a-c), but expression corresponds to CAGE enhancer activity. All plots are generated using data from human lymphoblastoid cell line.
Figure 4.5: Distribution of promoter and enhancer features. Each point corresponds to either an enhancer or promoter depending on its color. The plots show the distribution of these elements with respect to different pairs of chromatin state features. All plots are generated using data from human lymphoblastoid cell line.
Figure 4.6: (a) Distribution of feature values. (c) Autocorrelation of features. (e) Empirical cumulative distribution function (ECDF) of features. All plots are generated with $k = 3$ and using data from eight cell types.
Figure 4.7: (a) Relationship of features to the input genome data. (b) Relationship between features in neighboring positions. (c-e) Averaged enrichment of chromatin state features with relative position to TSS, TTS and middle of enhancer respectively. All plots are generated with $k = 5$ using data from eight cell types.
Figure 4.8: (a) Distribution of feature values. (c) Autocorrelation of features. (e) Empirical cumulative distribution function (ECDF) of features. All plots are generated with $k = 5$ and using data from eight cell types.
In this work, continuous chromatin state features for genome annotation are introduced. These chromatin state features are analogous to existing discrete chromatin state labels, but continuous features have several benefits: they preserve the underlying continuous nature of the input signal tracks; they can represent varying strength among elements; and they can easily represent combinatorial patterns of activity. Due to these benefits, it was shown that chromatin state features outperform existing discrete annotations at predicting gene expression and enhancer activity.

Continuous chromatin state features present an alternative representation of genomic activity to existing SAGA labels. It is expected that both types of annotations will be used in practice. Discrete labels are most effective when a fixed set of elements is needed. However, it is expected that continuous features will be used in applications where the limitations of discrete labels make such labels ineffective.

In particular, chromatin state features are useful for producing expressive visualizations. It was shown that visualizing chromatin state features from epigenome-ssm correctly depicts the continuum of active to inactive promoters and promoters to enhancers. Such a continuum is impossible in a discrete framework, which must use hard thresholds. Moreover, although each continuous feature is more complicated to interpret than a discrete label, it was shown that a small number of continuous features outperform even a large number of discrete labels in all of the evaluations. Therefore, a small number of chromatin state features can replace a much larger number of discrete labels, decreasing the overall complexity of the annotation.

Because continuous annotations maintain much more of the information in the input data than discrete annotations do, they are more useful for complex downstream applications. For example, a variant effect predictor might take chromatin state features as input in order to predict the functional impact of a given mutation. This is preferable to using raw tracks for two reasons. First, a small number of chromatin state features concisely summarize a large number of input tracks and therefore a predictive model based on these features will be less prone to overfitting. Second, chromatin state features can be used for variant
effect interpretation; that is, a model could report that its prediction of high variant effect is
due to the fact that a specific feature is present at that position. Such interpretation is more
difficult with raw tracks because most types of activity are associated with a combination
of many marks.

In this work, two versions of the SSM method are proposed: epigenome-ssm and epigenome-
ssm(non-neg). The main difference between the two versions is that in epigenome-ssm(non-
eg), there is a non-negativity constraint on emission matrix and state features. It was
shown in the gene expression and enhancer activity evaluations that despite the results of
the epigenome-ssm(non-neg) being slightly lower than epigenome-ssm, it is still among the
highest performing methods. Therefore, it is expected that both versions would be useful
in practice. In particular, if one requires a representation of the genome only, epigenome-
ssm should be used. However, if one requires easier interpretations of the state features in
addition to the representation, then epigenome-ssm(non-neg) is a better choice.

In the future, the continuous annotation approach can be used to create reference chro-
matin state feature annotations for all tissues with sufficient available data. These reference
annotations could be beneficial to several groups of people. Biologists for instance, can
use these annotations to better understand the underlying activities of the genome; this is
particularly important in the case of human genome. The reason is that unlike prokaryotic
genomes (a typical bacterial genome for example) that have 90% of them covered by protein-
coding DNA, only about 1.3% of the human genome comprises protein-coding regions [49].
Hence, annotation of the human genome is a much harder task compared to a bacterial
genome simply because there are huge gaps between genomic elements. The reference anno-
tations can also be useful for development of future annotation techniques where previous
annotations are required as a baseline for evaluating the newly produced annotations.

Also, despite having some visualization tools like the UCSC Genome Browser [32], Seg-
tools [2], and DeepTools [44] for analyzing genome annotations, these existing tools have
several limitations. For instance, the UCSC Genome Browser is a useful tool for visualizing
small regions of the genome; however, it is not very useful when one needs to get a holistic
view of the genome, e.g. some aggregate information from a whole chromosome. Even if a
visualization tool is very effective in a particular view of genomic data, it is often initially
designed for discrete annotations. Therefore, since continuous annotation is an emerging
approach, new tools are needed to be developed to better exploit chromatin state features.

Moreover, with ever-increasing generation of new genomic datasets from various tissues,
development of cell-type-agnostic annotation methods is an area with great potential for
future work. Early works on genome annotation were focused on small groups of assays
coming from a single or few cell types, and producing different annotations for each cell
type. However, there has been no doubt that from a biological point of view, having a
single annotation on assays from multiple cell-types is much more useful. So there is a need
for developing an epigenome annotation method that will be able to handle datasets from several cell types and identify regulatory elements such as enhancers and promoters.

The resulting annotation from a cell-type-agnostic method provides a global picture of where the regulatory elements are located in the human genome regardless of what specific cell type one is focusing on. Also, this type of annotation will be biologically more plausible in the sense that biologists consider the genomics position of the regulatory elements to be the same across all the cell types, and it is only the activity level of these elements that may be different from cell type to cell type. Therefore, having such methods will help biologists to gain a more realistic view of the human epigenome that will ease further downstream and upstream analysis.

The state space model for chromatin state feature annotation proposed in this work can be used as the base method for a cell-type-agnostic platform. However, there are some changes required for the base method to produce cell-type-agnostic annotation. First, the design of the input matrix needs to be changed, so that assay data from several cell types are stacked together. Second, a typical state space model may not work as expected for this kind of annotation; hence, different variations of SSM could be used for this purpose. Third, a linear method alone like the state space model may not be capable enough to capture the complex structure of the genomic data; hence, it could be used besides (or as part of) a non-linear method such as recurrent neural networks [19] or autoencoders [50].

Another direction for future work is to make the whole annotation pipeline automated, starting from data pre-processing to the final stages which include several validations and evaluations of the produced annotations. Supervised machine learning methods have a great potential to replace manual interpretations of the state features. In addition to the standard validation techniques, new evaluation techniques should be developed if necessary to be applied to the new types of annotations, for example, continuous and cell-type-agnostic annotations.

Last but not least, given the abundance of genomic data in very large sizes, it is crucial to have robust annotation platforms that can handle very large amounts of data in a reasonable time—incorporating parallel computing techniques such as multi-core and many-core computing might help in this regard. Those annotation tools that are able to run on computing clusters or supercomputers can of course perform better in large-scale annotation tasks. The annotation tools can also be provided as a web service in order to bypass many issues that can arise from installing and running the software on user’s personal computer.
Bibliography


[40] NHGRI. The National Human Genome Research Institute (NHGRI) Fact Sheets, 2015.


