Leveraging high-throughput datasets for studies of gene regulation

by

Angela Yen

S.B., Massachusetts Institute of Technology (2010)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....
Department of Electrical Engineering and Computer Science
May 10, 2011

Certified by .....
Manolis Kellis
Associate Professor
Thesis Supervisor

Accepted by .....
Christopher J. Terman
Chairman, Master of Engineering Thesis Committee

Leveraging high-throughput datasets for studies of gene regulation

by

Angela Yen

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 2011, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, I leveraged computational methods on biological data to better understand gene regulation and development of the human body, as well as of the model organisms mouse and yeast. Firstly, I tackled biological questions with machine learning techniques by studying pre-transcriptional gene regulation through nucleosome positioning, which resulted in the identification of function-specific factors and improved predictive performance. Next, computational analysis enabled the discovery of genome-wide epigenetic modifications that play a foundational role in silencing for the monoallelic and monogenic expression of olfactory receptor genes in mice. Lastly, signatures of functional, bound RNA regions provide insight into a potential protocol-specific bias and produce a new avenue for de novo discovery of functional regions.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor

0.1 Acknowledgments

To my supervisor, Manolis Kellis, I am indebted for providing immeasurable opportunities, valuable direction, creative ideas, and constant support. To my collaborators at UCSF, Stavros Lomvardas, Angeliki Magklara, and Josie Clowney, I am so grateful for your tireless dedication, high-quality work, and focused guidance that made our project so successful. To Nathan Haseley, my collaborator on the nucleosome positioning project, I appreciate all your patience in helping me grow my biological understanding. To my fellow Kellis lab members, especially Pouya Kheradpour, Luke Ward, and Jason Ernst, I would like to thank for all the insightful advice, fun conversations, and helpful mentoring.

I would also like to acknowledge the Siebel Foundation for financial support through the Siebel Scholarship.

To Mom, Dad, and Philip, I owe you everything for making me the person I am today, and I love you all very much. To my friends, especially Gabriel, Jess, Dustin, David, Kent, Lindsay, Simone, Cathy, Steph S., Nina, May, Elizabeth K., Michael, Kate, and Larissa, thank you for always believing in me and putting up with me, but most of all, for making the past 5 years some of the best years of my life.

Contents

Chapter 1

Introduction

1.1 Problem Statement

The sequencing of genomes has revolutionized the ways in which scientists can investigate biological processes and disease pathways; new genome-wide, high-throughput experiments require computer scientists with a biological understanding to analyze and interpret the data to improve our understanding about life science.

The question of how the complexity of a human body arises is fundamental to both biological knowledge and medical treatment. Nevertheless, development has largely remained an open question due to its complexity. Unraveling the tangled interactions between the genome and other factors will pave the way for improvements in both diagnosis and treatment of diseases.

While it is clear that each person's unique DNA, or genetic code, plays a central role in development, knowing the genome is not enough. This can be evidenced by the fact that different types of cells, from skin cells to heart cells, can encode different phenotypes, even though they all have identical DNA. In this thesis, I study various factors that influence and control development through pre-transcriptional and post-transcriptional regulatory control.

1.2 Background

1.2.1 Regulation of gene activity

One of the central tenets of biology is that DNA is transcribed into messenger RNA (mRNA), and mRNA is translated into proteins. The overall process of producing proteins from the genetic code is generally considered gene expression, and controlling the time period or quantity in which genes are expressed is often referred to as gene regulation.

The underlying mechanisms of gene regulation are complex and vary widely in different contexts. While the genes provide the genetic 'code' necessary for biological processes, gene regulation acts as the 'control' level. Just as computer programs must decide which sub-functions to run in which contexts, gene regulation ensures that specific genes are expressed in specific cell types during specific time points: this enables the same initial stem cells to differentiate into the hundreds of distinct cell types in an adult human.

Gene regulation can imply that expression of a gene is increased or decreased, and different types of gene regulation can occur at different points along the path of gene expression. Since gene expression is the act of transcription (DNA to RNA), followed by translation (RNA to proteins), some mechanisms of gene regulation occurr at the transcriptional level, while some occurr at the post-transcriptional level. For example, regulatory regions at the transcriptional level can be promoter regions, regions at the the beginning of genes at which proteins bind to for initiation of transcription, or enhancer regions, which can distantly regulate the transcription of genes. At the post-transcriptional level, many regulatory regions fall in the 3' and 5' UTRs, at binding sites for microRNA and RNA-binding-proteins.

Epigenetic modifications can both control and record gene regulation, as they are heritable changes made either to DNA (DNA methylation) or to its associated histone proteins (histone modifications). Epigenomics, which specifically studies epigenetic modifications on a genome-wide scale, makes discoveries of large-scale patterns of gene regulation, such as regulation of entire gene families.

Another aspect to study is the genome-wide transcription of genes, as this can give insight into both pre-transcriptional and post-transcriptional regulation. With new technologies such as RNA-Seq, I can identify what regions of DNA are transcribed into RNA, the intermediate step before the RNA is translated into proteins. In this thesis, I study protein-bound RNA, as these regions are often functional regions closely tied to gene regulation; for example, protein-bound regions might be transcription factor binding sites or areas of post-transcriptional modifications. By identifying signatures in RNA-Seq data for protein-bound regions of RNA across different tissues, I gain insight into tissue-specificities and reveal a potential bias of the RNA-Seq protocol. Furthermore, this project provides the fundamental groundwork for de novo annotation of protein-bound RNA regions of the genome.

1.2.2 Epigenomics and pre-transcriptional regulation

Though the sequencing of genomes was a landmark event in biology, sequence information is not enough; epigenetic modifications also play a crucial rule in gene regulation. On a cellular level, epigenetic modifications can play a causal role in the regulation of genes - for example, a modification might serve as a 'sign' that the surrounding genes should be expressed. On the other hand, the epigenome might show the history of how the genome has been used through different developmental stages; just like hunters can find clues about nearby animals through tracks in the dirt, scientists can see the history of a cell by observing the locations and types of epigenetic modifications.

Specifically, the two main types of epigenetic modifications are DNA methylation and histone modifications. DNA is tightly packed due to being wound around protein sets called nucleosomes. The combination of nucleosomes and the DNA wrapped around it is called chromatin, and these nucleosomes are octamers of histone proteins. Therefore, DNA methylation is the addition of a methyl (-CH3) group to DNA, the genetic code. Additionally, histone modifications are molecular post-translational changes made to either the core or the long tail of certain histone proteins in the histone octamer.

These epigenetic modifications have been shown to be associated with pre-transcriptional gene regulation, as certain marks often have 'repressive' or 'activating' effects on the surrounding genes. The mechanism through which this occurs is still unclear, but there is evidence that it is related to nucleosome positioning.

As mentioned above, DNA is tightly packed in our cells by being wound around nucleosomes. This means that nucleosome positioning can play an epigenetic role in pre-transcriptional gene regulation. Specifically, regions of the DNA that wrap around nucleosomes are less accessible and more closed to transcription factors. On the other hand, the regions of DNA that link the nucleosomes are more accessible and open to transcription factors. The state of the DNA being more or less accessible due to nucleosome positioning is often referred to as an 'open chromatin state' or 'closed chromatin state,' respectively. In general, it has been shown that chromatin states are often correlated with the transcription state of the corresponding genes; they can act as instructions for the genes present in the surrounding DNA, or they can record the 'history' of the transcriptional state.

The naming mechanism of histone modifications provides an implicit description about the modification. There are five major classes of histones, and the name of the histone modification starts with the class of histone (e.g. H3). This is followed by the single-letter amino acid abbreviation, such as K for Lysine, and the number of the position of the amino acid in the protein. The final part of the naming procedure is the type of modification that was applied to the amino acid, such as Me3 for trimethylation.

1.2.3 Post-transcriptional regulatory mechanisms

The general model for gene expression is transcription (DNA to mRNA) followed by translation (mRNA to protein). Transcription occurs through the production of complementary RNA to the DNA of the gene, as each DNA nucleotide has a complementary RNA nucleotide: Adenine DNA nucleotides are paired with Uracil RNA nucleotides, Thymine DNA nucleotides are paired with Adenine RNA nucleotides, Cytosine DNA nucleotides are paired with Guanine RNA nucleotides, and Guanine

DNA nucleotides are paired with Cytosine RNA nucleotides.

While the classic model is that translation from RNA into proteins follows transcription, post-transcriptional regulation sometimes prevent this from happening. Post-transcriptional regulation, as the name implies, is control of gene expression at the RNA level, in between the transcription and translation of the gene; this can be done through modifying the stability of the transcript or regulating the act of translation.

There are various methods of post-transcriptional regulation. One common mechanism is the binding of RNA-Binding Proteins or regulatory RNA to the 5' or 3' untranslated regions (UTRs) of the RNA transcripts. For example, AU-rich elements (regions rich in Adenine and Uracil nucleotides) in the 3' UTR often serve as binding sites for proteins that can either stabilize or destabilize the transcript. On the other hand, regulatory sequences in the 5' UTR can more directly affect translation through prevention or initiation, as it is an important area for initiation of translation. Additionally, a common example of regulatory RNA is microRNA (miRNA) binding sites in 3' UTR, as miRNAs and their respective RNA-induced Silencing Complexes (RISCs) can be responsble for post-transcriptional silencing through either degradation or translation prevention.

RNA-Seq data aims to measure the amount of RNA present by isolating the RNA in cells, fragmenting and isolating it, amplifying it, and then sequencing it. Aligning the sequences back to the reference genome gives the numbers of RNA 'reads' that were found for each position in the genome. However, since this is an experimental process, there is a possibility for signatures or biases in the resulting data that can provide scientific insights about post-transcriptional regulation, as well as factors should be accounted for in other applications of this data.

1.2.4 Model organisms for understanding human biology

Studying model organisms, such as yeast and mice, in addition to studying humans, has proven to be an incredibly powerful technique. As there are obvious ethical limitations on human experimental techniques, studying these model species with a

larger toolbox of techniques can reveal biological findings that can then be confirmed in humans. Similarly, if the interactions of factors in human are too complex to immediately unravel, some model organisms, such as yeast, provide similar but simpler systems that are a crucial stepping stone for understanding humans.

1.3 Summary of research Contributions

For this thesis, I leveraged high-throughput datasets for three studies of gene regulation. The initial project used the yeast genome to predict positioning of nucleosomes, the complexes of histones attached to DNA. The second project was a study of how epigenetic modifications play a role in the monoallelic and monogenic olfactory receptor gene regulation in mice. Lastly, I characterized a signature of functional, protein-bound RNA in transcriptome data, which I could use in the future for de novo discovery of functional regions.

1.3.1 Nucleosome positioning

As mentioned above, nucleosome positioning has been shown to play a critical role in gene regulation, DNA repair and replication, and recombination. Large-scale analysis of nucleosome positions have been carried out in multiple organisms, but the underlying factors contributing to these placements remain poorly understood. Many factors such as the frequencies of short k-mers and the periodic repeats of GC and AT rich dinucleotides have been associated with nucleosome positioning, but their significance has often been questioned, and models developed based on these features give only modest performance.

I evaluated the hypothesis that nucleosomes are regulated in different ways across varying functional classes of DNA. As yeast is the organism with the most thorough and cleanest annnotations and sequencing, I used pre-existing yeast data for our methods; I can immediately apply our findings in yeast to other similar organisms, such as mice and human, both in terms of feature importance and prediction methods. I divided nucleosome-bound and nucleosome-free regions sequences from

Saccharomyces cerevisiae into 4 functional subclasses - coding, noncoding, promoter, and centromere/telomere groups.

Based on the existing literature, the features I chose to use were the frequency of all k-mers of length less than or equal to 6, as well as scores to measure the occurrence of periodic GC and AT rich dinucleotides. For feature selection, we used F-scores to approximate the 20 most important features for discriminating nucleosome positioning in each class and trained SVMs on data with only those 20 features.

We found that SVMs trained on specific subclasses gave, on average, at least 1.89% better performance over an SVM trained on a more general set of sequences. Further analysis of our models suggests that their discriminatory power mostly lies in the periodicity features, which are the three features that measure the repeating signatures of AT and GC rich dinucleotide repeats. These are by far the most discriminating features across all of our subsets, according to the F-scores, and models composed of only these features perform nearly as well as our initial models.

We have also shown, however, that different k-mer frequencies appear to be selected more frequently in some subclasses than in others. Additionally, though our periodicity features had high f-scores across all classes relative to the other features, the actual values of the scores varied greatly between the classes. Combined, these facts suggest that a few very strong, general characteristics whose effects are relatively universal may dominate nucleosome positioning; these few characteristics also likely mask weaker function-specific signatures of positioning.

1.3.2 Epigenetic regulation of olfactory neuron specification

As multicellular organisms develop from an initial single zygote into a complex system, cellular differentiation turns less specialized cells into more specialized cells. For example, pluripotent cells are unspecialized, and therefore, have the potential to differentiate into any cell type in the organism. Differentiation changes a cell's size, shape, activity, and other physical characteristics, largely through the strict regulation of gene activity.

Olfactory receptor neurons, the neurons responsible for our sense of cell, are one

type of specialized cell that has a strict 'one neuron - one receptor' rule: specifically, each olfactory neuron expresses exactly one olfactory receptor (OR) gene, while all the other OR genes are silenced. This means that each olfactory neuron has the genetic capacity to detect any odor molecules, but the receptors are regulated so every neuron actually detects exactly one smell. The chosen olfactory receptor gene that is expressed in the neuron largely defines the functional essence of that neuron. The combined power of all the olfactory neurons is what enables the brain to detect a wide variety of smells. In this project, we idemtified the regulatory role of epigenetic modifications for the monogenic expression of olfactory receptor genes in mice.

Olfactory receptor gene regulation is especially crucial in mice, as their sense of smell is even more discriminating than humans; mice have over 1300 olfactory receptor genes (approximately 5% of their genes), while humans have only about 900 OR genes. Furthermore, mice are biologically very similar to humans, so findings in mice can often be generalized to holding in humans as well. Clearly, however, mice provide advantages over humans due to limits on data collection for humans. The lifespan of mice, as well as the increased experimental power provided by such a model organism, made it a clear choice to use mice for this study.

We found that in the mouse olfactory epithelium, OR genes are specifically and sensitively correlated with the histone modifications H3K9me3 and H4K20me3; these marks were much less present in our control tissue, liver. We also found that other familes of chemoreceptors, such as vomeronasal receptors and formyl peptide receptors were also marked than the same histone modifications, although at a lesser degree. As a result, the cell-type and developmentally dependent deposition of these marks along the OR clusters is, most likely, reversed at a single OR allele during OR choice, to allow for monogenic and monoallelic OR expression. In contrast to the current view of OR choice, our data suggest that OR silencing takes place developmentally before OR expression, indicating that it is not the product of an OR-elicited feedback signal; this can be considered a conservative starting state for this strict regulatory mechanism. Overall, this suggests a new role for chromatin-mediated silencing as the molecular foundation upon which singular and stochastic selection can be applied.

1.3.3 Post-transcriptional regulation of RNA

Studies have increasingly found that post-transcriptional regulation plays a crucial role in many scenarios of gene regulation. Furthermore, there is increasing availability of high-throughput datasets of various human cell lines. In this project, we combine these two factors to use deep human RNA-Seq data to study post-transcriptional regulation across. Specifically, we hypothesized that protein-bound RNA regions may be less accessible in the RNA-Seq protocol, resulting in an artificially reduced signal for protein-bound regions.

To approximate protein-bound regions of RNA, we investigate conserved regions of 3' and 5' UTRs. This is based on the fact that the majority of protein-bound post-transcriptional regulation takes place in the 3' and 5' UTRs, as well as the fact that conservation often implies functional importance, which is present in protein-binding regions of the genome. The technique we use to study these specific types of regions are aggregating RNA-Seq signal across the distinct instances of these conserved regions.

The study of aggregate plots in different conditions gives insights to transcription in different environments. For example, we can compare conserved regions in 3' UTRs and 5' UTRs. We can also aggregate the data in different ways - either by looking at the arithmetic sum or the geometric sum of the RNA-Seq counts. Additionally, since we have data for 20 different tissue types, we can search for any tissue-specific differences in signal. Furthermore, we can require a minimum window size for each conserved region, and varying this window size allows us to see how this affects the signature in the plot. Lastly, since conserved regions across the genome will vary in size, we must somehow account for these differences: options are to align based on the start or the end of the region, as well as to align in the center but scale the different regions so they can be aligned end-to-end.

Our preliminary findings show a promising signature of a dip at the alignment point, especially in the 3' UTR when aligned to the ends of the conserved regions. We also see a significant correlation between the general slope of the plots and the

alignment point. We also generally see a distinct signal between the 3' UTR and 5' UTR regions, which makes some sense, since they often regulate with distinct mechanisms.

These findings are mainly applicable in two ways. First of all, they lay the foundation for de novo prediction of genetic regions that are transcribed into functional, protein-bound RNA regions. Secondly, these experimental artifacts must be taken into account and corrected for when RNA-Seq data is used for other studies.

Chapter 2

Nucleosome positioning

In a joint project with graduate student Nathan Haseley, I used a supervised machine learning technique, Support Vector Machines (SVMs), to produce classifiers that predict whether a DNA sequence is in a nucleosome-binding or linker region and investigate the factors contributing to nucleosome positioning in various functional genomic regions. Specifically, I divided all nucleosome-bound and nucleosome-free regions of DNA from the yeast genome into 4 subclasses: centromeric and telomeric regions, promoter binding sites, protein-coding genes, and non-coding regions. I chose to use SVMs because they are a straightforward technique of supervised learning to produce a classifier from training data. Additionally, they have been shown to be moderately successfully in previous papers concerning nucleosome positioning[52].

I measured 5462 features for each sequence and used subsets of data from each class to select the 20 most significant features. I used both a linear and radial basis kernel to build classifiers based on our training data. The accuracy of our models was evaluated using cross-validation, as well as testing on labeled test sets, when sufficient data was available. I compared these subclass-specific (centromeric/telomeric, promoter, coding, non-coding) models to a general model, constructed in the same manner, using a training set composed of data from all four classes of genomic sequences. By calculating the accuracy of the subclass-specific model and general model on class-specific test sets, I showed that our subclass-specific modes perform slightly better in all cases.

I then further investigated the features chosen in our subclass-specific SVMs to better understand why they were more accurate than the general model. Finally, by investigating the features selected for each subset of data, as well as the accuracy of the models, I gained some insight into how different features can play different functional roles for nucleosome positioning.

2.1 Background

2.1.1 Problem Statement

DNA in eukaryotic cells is organized into a highly compact and structured form known as chromatin. This process is mediated by interactions with histone octamers, which bind with DNA to form nucleosomes. DNA segments of approximately 147 base pairs in length are wrapped around each histone octamer, and the 'free', unattached DNA in between nucleosomes are called linker regions, and are generally 10-50 bp in length[30]. This system of DNA packaging permits cellular DNA, which can be meters long, to fit into the nucleus, which is usually only a few micrometers in diameter. Significantly, nucleosomes have been shown to play critical roles in gene regulation both by sequestering specific DNA sequences, and by interacting with protein complexes[37, 82].

Nucleosome placement in vivo is far from random. Many nucleosomes have extremely stable positions that seem to hold across a variety of cellular conditions, while other nucleosomes seem to migrate in response to specific signals[60]. There are also more general trends, such as how promoter regions have been shown to be highly enriched on linker sequences, permitting optimal access of transcription factors[60].

The importance of nucleosome placement is recognized for many reasons. Nucleosome placement seems to influence DNA transcription, repair, recombination, and replication[84]. Furthermore, the incorporation of nucleosome binding site information has been shown to improve the identification of transcription factor binding site and other regulatory motifs[63]. It has also been suggested that understanding nu-

cleosome positioning may shed light on novel selective forces operating on DNA[78]. Finally, nucleosomes have been reported to be involved in the regulation of tissue-specific transcripts in humans[40]. An understanding of the mechanisms used to control these placements would provide key insight into mechanisms of cellular genetic regulation[61] and genomic processes.

2.1.2 Previous Work

Nucleosome positioning remains a puzzle, though it has been thoroughly studied from many perspectives. Previous work has shown that some contributing factors are nucleosome-protein interactions, sequence specificity, and steric interactions with other histones[60, 63, 64, 84]. The majority of nucleosome prediction methods currently available rely on signature sequence characteristics, including the enrichment of particular k-mers that influence the flexibility of DNA[63, 84] and the periodicity of nucleosome-bound DNA fragments[81]. Recurrent patterns of AT and GC rich dinucleotides have been reported every 10-11 nucleotides in many nucleosome-bound sequences, corresponding to a single turn in the DNA helix. It has been suggested that these primarily sequence-based methods account for the majority of factors influencing nucleosome positioning based on the high level of agreement between in vitro nucleosome maps (created without the influence of proteins) and known in vivo maps[63]. Despite this, and the amount of analysis that has gone into analyzing nucleosome-bound sequences, these classification methods show only modest improvement over a null model which identifies every nucleotide as a nucleosome bound site.

2.1.3 Approach: function-specific SVMs

I hypothesized that one failing in previous models that they have been binary in nature; that is, all nucleosome-bound DNA sequences were treated as if they were functionally equivalent and regulated by the same mechanisms. I did not believe that this assumption is warranted. Intuitively, it seems that nucleosomes near functional elements, such as transcription factor binding sites, are more likely to be regula-

tory in nature. Therefore, their position may be governed by different interactionsor mechanisms than nucleosomes that play a more structural role. Furthermore, it has recently been shown that histone-modifying proteins tend to target specific subsets of histones and produce different sets of epigenetic modifications[81]. These tags may lead to different protein interactions and affect the sterics of DNA binding.

I therefore felt that it was necessary for nucleosome positioning to be evaluated in a functional, as opposed to structural, context. I decided to evaluate coding regions, noncoding regions, transcription factor binding sites, and structural regions (centromeric and telomeric regions) separately, using a robust supervised learning SVM approach to improve prediction performance and identify factors important in nucleosome positioning in each of these contexts. These specific regions were chosen because they represent a variety of functional contexts and histone modification patterns.

Model organism choice: yeast

Budding yeast, also known as *Saccharomyces cerevisiae*, was a clear choice for this study. Yeast provides two crucial advantages: it is one of the most cleanly annotated and deeply sequenced organisms, and due to the ease with which experimental techniques can be used on yeast, there was already publicly available experimental data for nucleosome positions.

2.1.4 Datasets used: annotation and nucleosome positions

Nucleosome positions

Nucleosome bound and free regions of DNA from across the Saccharomyces cerevisiae genome were taken from the "reference set" of nucleosome positions described by Jiang et al[33]. This data set was compiled based on agreement between six different experimental datasets measuring nucleosome positioning using different technologies, and it is the most comprehensive set of nucleosome locations available for the yeast genome. I filtered these DNA sequences, eliminating regions with nucleosome occupancy less

than 50% (as described by Jiang et al), to remove ambiguities. Additionally, I ignored all hypothetical nucleosomes and all linker regions shorter than 10bp, reasoning that very short linker regions could bias our feature selection (see below) and that these regions may exist primarily because of steric interactions between nucleosomes instead of any sequence specific signals that I could detect with our SVM. All remaining nucleosome-bound and nucleosome-free sequences were sorted into coding, noncoding, promoter, and structural subsets as described below and used as input for our various SVMs. The resulting number of nucleosome-bound and nucleosome-free sequences in each subset is given in Table 2.1.

	Nucleosome-bound	Nucleosome-free
Coding	39756	38598
Promoter	10680	11087
Noncoding	6396	6690
Structural	669	619

Table 2.1: Number of Nucleosome-bound and nucleosome free regions of DNA per subset. All subsets were derived from annotations in the SGD. An approximately equal number of nucleosome-bound and nucleosome free regions are present in each subset. As is expected, based on the contents of the yeast genome, most DNA regions fall within coding sequences, with very few being present in structural elements (centromeres and telomeres).

Annotation data

I used information from the Saccharomyces Genome Database[11] to parse the entire Saccharomyces cerevisiae genome into coding, noncoding, promoter, and structural regions. Any sequences contained in verified, protein-coding ORFs were considered to be coding. All sequences located within 1000 bases upstream of a verified, protein-coding ORF were classified as promoters. Structural regions were composed of both annotated centromere and telomere sites. Noncoding sequence included all regions not classified as coding or promoter regions that also did not contain putative protein coding or RNA coding genes. All other regions, such as those coding tRNA genes were ignored for this study.

2.2 Methods

2.2.1 Feature types and selection

K-mer and periodicity features

There is a surprising amount of disagreement in literature concerning the factors that affect nucleosome positioning, and even how critical genetic sequence is in this process. Most research supporting the prevalence direct sequence signatures concur that the frequencies of relatively short k-mers and the periodic occurrence of GC and AT rich regions are important for controlling the ability of a given sequence to bend around and interact with histone proteins. Despite this agreement, the relative importance of these features, including which k-mers should be used, is less certain. Because of these ambiguities and our suspicion that specific k-mers may especially serve as sites of protein interaction, and therefore, might have function-specific importance, I did not want to rely exclusively on previously published literature to select our specific features.

First, I systematically considered all the possible k-mers from k=1 to 6, and created a feature to represent the frequency of each k-mer. The value six was a somewhat arbitrary limit, chosen to include most of the specific k-mer instances that have been used in previous studies while still maintaining a reasonable limit for the number features I had to consider. For each k-mer feature, I generated a representative value x using the following formula:

$$x = \frac{n * 4k}{l}$$

where n represents the number of occurrences of that k-mer in the sequence, and l is the length of the sequence. This was to give our scores a rough probabilistic interpretation; with four possible nucleotides (Adenine, Cytosine, Guanine, and Thymine), the probability of a given k-mer given a random sequence of length k is $frac14k$. The division by l serves as a normalization for length, which was necessary because nucleosome free sequences tend to be much shorter than nucleosome bound sequences.

Each of these k-mer features was mapped to an feature 'index': for example, the 1-mers were mapped to the values 0-3 while the 6-mers were mapped to the values 1364 through 5459.

Additionally, I considered three features that measure in the periodicity of the sequence. It has been suggested that nucleosome bound sequences often contain both patterns of AT-rich dinucleotides repeating with a frequency of approximately 10.3 nucleotides (the length of a single turn of a DNA helix) and GC-rich dinucleotides with the same frequency but at an off-set of five base pairs from the AT-rich repeats. Therefore, I performed a pair-wise comparison of AT-rich and GC-rich dinucleotides found in each given sequence. One feature, 5460, represented the number of pairs of AT-rich dinucleotides that were approximately a multiple of 10.3 bps away from one another. Another feature, 5461, represented the same frequency for GC-rich dinucleotides, while a third feature, 5462, represented the number of pairs of GC-rich and AT-rich dinucleotides that were approximately a multiple of 5 base pairs away from each other. The actual formula used was:

$$x = \frac{n}{16 * l}$$

where x is the value calculated for the feature, l is the length of the sequence, and n is the number of times the respective dinucleotides were found at the expected frequency. Again, I normalized the counts by dividing by sequence length and multiplying by 16, since the probability of finding pairs of AT-rich or GC-rich dinucleotides at expected frequencies is $\frac{1}{2^2} * \frac{1}{2^2} = \frac{1}{16}$.

Feature selection: Fisher information metric

Although I initially considered a huge number of potential features ($\tilde{5}500$) for maximum coverage, I had to choose the most discriminating features for our SVM, as complexity quickly increases with the number of features. Due to the large amount of data, I chose to use a relatively rudimentary information analysis technique to rank the features, rather than a variant of feature subset selection. Specifically, I calcu-

lated a Fisher information metric for each of our features, with the implementation provided with the software package LIB SVM[8]. Fisher scores represent how discriminative each feature is for classifying between positive and negative examples. That is for a set of data points, let μ and σ^2 be the mean and variance for the data set, while μ_+, μ_-, σ_+^2, and σ_-^2 are the means and variances for the specified feature values of the set of nucleosome-bound sequences and nucleosome-free sequences, respectively. Then, the formula for the Fisher statistic for a specific feature is shown below:

$$F = \frac{n_+(\mu_+ - \mu)^2 + n_-(\mu_- - \mu)^2}{n_+{\mu_+}^2 + n_-{\mu_-}^2}$$

While this metric did not allow us to capture discriminatory power that lied in combinations of our features, it made the feature selection computationally feasible and straightforward. Even using this greedy approach, I was limited to running feature selection on subsets of 1000 sequences of our data due to the cost of computing F-scores. Separate feature selection was performed for each of the five models - one for each of the four subclass-specific models, and one for the general model. The 20 features with the top-ranking F-scores were chosen to be used for training and testing each model.

2.2.2 Kernel types and parameter selection

Kernel types: radial basis and linear kernel

I chose to initially use both a radial basis and a linear kernel. In theory, the radial basis should always result in lower training error; however, that is only true assuming one is able to find the optimal cost parameter (penalty for an incorrect classification). Since the programs that chose parameters and trained the SVM utilized simplifications that resulted in not searching the entire space, the radial basis kernel was not guaranteed to perform better. Furthermore, since the RBF kernel is much more complex, it is more prone to overfitting. Lastly, using both kernels provides an opportunity to consider the trade-off between shorter processing times and decreased accuracy.

Parameter selection

For the radial basis kernel, I chose the cost parameter with the help of a program (grid.py) from LIB SVM. The program took the training sets for each class, each with its twenty most important features as input. The program outputted the ideal parameter values for cost found from its limited search space; specifically, it considered powers of 2 as the possible values.

Scaling feature values

The authors of LIB SVM also recommended that the feature values for training data be scaled so that they fall between -1 and 1. Therefore, all training data sets were scaled, and the parameters used for the scaling was stored in a range file so that the test data could be scaled the same way. The scaling increased the speed of the SVM model generation.

2.2.3 Training, Cross Validation, and Accuracy

Training with cross validation

The training sets of 3000 sequences with selected and scaled features were used to train a SVM model with both radial basis and linear kernels with the provided programs (svm-train from LIB SVM and LIB LINEAR, respectively). I obtained 5-fold cross validation results for each of the models; this means that I subdivided the training dataset into 5 separate subsets, and each subset took turns being the 'hold-out' set, where training was performed on the other 4 subsets, and the hold out set was used for testing validation.

Testing and performance measures

Additionally, I used test sets of 3000 sequences that were non-overlapping with the training sets, when enough data was available, to calculate testing accuracy. I quantified performance by overall accuracy (what percent of sequences were classified correctly), sensitivity (what percent of nucleosome-bound sequences were classified as

nucleosome-bound) and specificity (what percent of nucleosome-free sequences classified as nucleosome-free).

2.3 Results and Discussion

2.3.1 Feature selection

I selected features based on the Fischer scores, and the scores for the features of each subclass are shown in Figure 2-1. Features 5460, 5461, and 5462, the features measuring the periodicity of AT and GC rich dinucleotides, were by far the most significant features in each class, though the score magnitude and relative rankings change across sub-classes. In comparison to the periodicity features, the F-scores for our k-mer features are all very small, but a comparison of the k-mer features is shown in Figure 2-2. The results show that there does seem to be a slight bias towards shorter k-mers, which would be expected from the literature. The significance of these values will be addressed further shortly.

Feature selection was one aspect of our study that could have easily been improved. In the future, I would like to run feature selection again on larger subsets of our data. Additionally, I could also use F-scores to eliminate the weakest features, and then use subset selection methods to choose the best subset of features together. While I did not expect all the k-mer feature f-scores to perform so poorly in comparison to the periodicity features, it was more difficult than expected to identify features that have sub-class specific importance. However, as shown later, I still believe there is a function-specific signal in these f-scores, but that it is overshadowed by the universal dominance of the periodicity features.

After feature selection, I w able to use larger subsets for the remainder of the study, since the decrease in number of features reduced the complexity so drastically. Therefore, for the rest of the project (parameter selection, training, and cross validation), I used a subset of 3000 sequences for each of the noncoding, coding, and promoter regions as training sets. For the fourth class, telomeric and centromeric

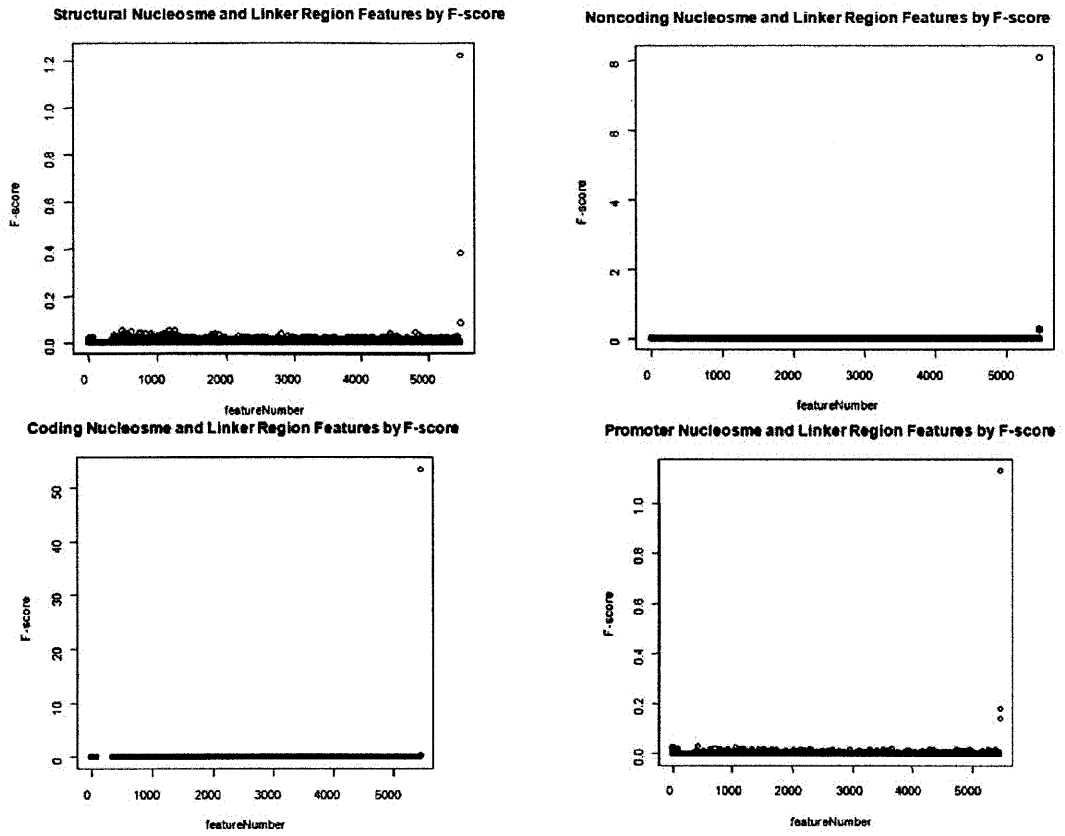


Figure 2-1: F values for all features. Note the difference in scale. Features 5460, 5461, and 5462, which represent AT rich dinucleotides with a period of 10.3 nucleotides, GC rich dinucleotides with a period of 10.3 nucleotides, and alternating AT, GC rich dinucleotides with a period of 5 nucleotides respectively, dominate the F scores in all subsets.

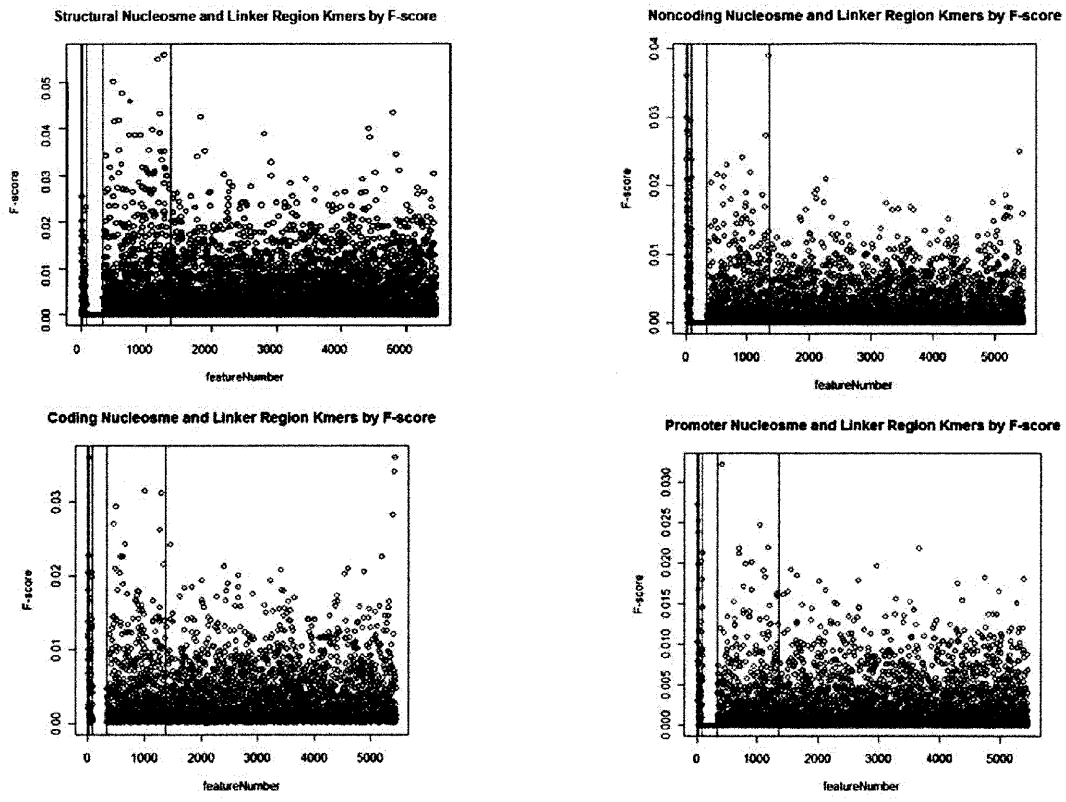


Figure 2-2: F-values for kmer features only. Lines denote the demarcation between differing values of k. 1-mers are shown on the left, while 6-mers are displayed on the right. While these values do not score nearly as well as our repeat scores, they show significant differences between our subsets.

regions, I had less than 3000 sequences to begin with, so I used the entire data set, leaving no sequences for a test set.

2.3.2 Training, Cross Validation, and Testing

Training with 5-fold cross validation

I trained our SVM models on datasets of 3000 sequences with 5-fold cross validation. I obtained the resulting cross-validation accuracy, shown in Table 2.2.

Subset-specific model	RBF kernel CV accuracy	Linear kernel CV accuracy
Promoter	85.47%	81.67%
Coding	90.83%	90.03%
Noncoding	90.57%	85.37%
Structural	92.33%	82.12%

Table 2.2: Cross-validation accuracy of subset specific SVMs on training data. Separate SVMs were trained to distinguish between nucleosome-bound and nucleosome-free regions of DNA in promoter, coding, noncoding, and structural (telomere and centromere) regions of the genome. These were tested on subsets of 3000 pre-categorized DNA sequences. The 5-fold cross validation accuracies are shown above.

Testing performance

For the three data sets with sufficient data (noncoding, coding, and promoter regions), I also created test datasets of 3000 sequences. After selecting and scaling the feature values in the exact same way as was done for the training set, I calculated the predictive performance of the model on the test set: the results are shown below in Tables 2.3 and 2.4. Due to the small number of structural sequences, a test set of 3000 could not be withheld for the structural sub-class. The results show strong performance from all the models, for both sensitivity and specificity, though the models are better at correctly identifying nucleosome-bound sequences. This performance is much better than the 50% accuracy expected by chance.

Subset-specific model	RBF kernel testing accuracy	Linear kernel testing accuracy
Promoter	85.47%	81.67%
Coding	90.90%	90.23%
Noncoding	90.17%	86.53%
Structural*	N/A	N/A

Table 2.3: Percent accuracy of subset specific SVMs on testing data. Separate SVMs were trained to distinguish nucleosome-bound and nucleosome-free regions of DNA in promoter, coding, noncoding, and structural (telomere and centromere) regions of the genome. These were tested on subsets of 3000 pre-categorized DNA sequences, none of which were in the original training data. The percent of correct classifications on the training set is shown above. *Structural set was not tested on a full set of size 3000 because of limited data availability

Subset-specific model	Sensitivity	Specificity
Promoter	97.7%	82.18%
Coding	98.7%	81.7%
Noncoding	97.2%	72.97%
Structural*	N/A	N/A

Table 2.4: Sensitivity and Specificity of subset-specific RBF SVMs. Sensitivity and specificity were calculated for promoter, coding, noncoding, and structural specific SVMs. Sensitivity is a measure of the models ability to detect nucleosome-bound regions. Specificity is a measure of the number of nucleosome-free sites identified. As can be seen, our models are much more effective at correctly identifying nucleosome-bound sequences, though both are detected far better than would be expected by chance. *Structural set was not analyzed because of limited data availability

2.3.3 Kernel and parameter selection

Kernel selection

Due to the fact that every subset-specific model with the RBF kernel performed better on the test set than every linear model, I can conclude that the radial basis models generalize better. That is, even though the complexity penalty based on the VC-dimension would give a lower bound for the linear kernel, it is evident that the overall generalization error of the radial basis model is still smaller. Therefore, the linear models bound must be much tighter than the radial basis models bound, so the generalization guarantee is not a good metric, in this case, for model selection.

It is clear from the testing accuracies that the RBF kernel model will result in more accurate predictions. However, since the linear kernel is less complex than the RBF, it should hypothetically result in faster running times by the programs. In this case, run-time was not an issue because I used both LIB-SVM and its sister library, LIB-LINEAR; LIB-SVM was optimized for the RBF kernel and LIB-LINEAR was optimized for a linear kernel. Interestingly enough, this meant that LIB-SVMs programs ran slower for the linear kernel than the RBF kernel (while the RBF kernel was not an option for LIB-LINEAR). Runtimes for training of LIB-SVM with a RBF kernel and LIB-LINEAR with a linear kernel were comparable for practical reasons; though LIB-LINEAR was consistently faster, averaging less than a second, LIB-SVM only took 2-3 seconds on average.

Since the generalization error of the RBF kernel was smaller and runtimes were comparable, the RBF kernel was a better choice for our purposes; when SVMs are mentioned in the rest of the paper, they are implied to be the models associated with the RBF kernels.

Cost parameter selection

The cost parameter for the radial basis kernel was calculated by searching for the optimal value over a limited search space, as discussed in the previous methods section. The resulting values found are shown below in Table 2.3.3

Subset-specific data	Cost parameter
Promoter	8192
Coding	8192
Noncoding	32768
Structural	2048

Table 2.5: Cost parameters found for subset-specific training sets.

2.3.4 Classification performance comparisons

Comparison with previous work

Our SVM methods of predictions far exceed the accuracy that would be achieved by random chance (note the near equal composition of nucleosome-bound and nucleosome-free sequences of DNA in our subsets from table 1). This, of course, brings up the question as to how I did compared with other classifiers and predictors that have been developed. Unfortunately, this is not an easy question to answer. Our classifier does much better than the majority of predictors available[52, 63, 84] , but a significant contributor to this performance is likely that I simplified the problem by using DNA sequences that are already divided into nucleosome-bound and nucleosome-free segments. I only had to assign a label to each sequence, not the positions where state changes occurred. The majority of computational methods used to predict nucleosome positioning score the probability that a given stretch of DNA occupied by a nucleosome and then use some type of HMM or other process to trace an optimal path of nucleosomes through a large segment of DNA. Thus, their methods are far more relevant for predictions in unknown sequences. I am mostly interested in using SVMs to probe how factors that affect nucleosome positioning vary among DNA sequences and thus felt that trying to predict state transition locations was an unnecessary complication, especially considering the time frame of our project. I, therefore, decided to construct a general model, using the same procedures described above, to determine the effectiveness of our class-specific classifiers.

Comparison between subclass-specific and general models

By comparing the general model to our function-specific models, I can quantify the improvement gained from dividing the genome into our classes before building the models. I therefore constructed a subset of 1000 DNA regions (nucleosome-bound and nucleosome-free) composed of equal numbers of promoter sequences, coding sequences, noncoding sequences, and structural sequences. Feature selection was performed, as described above. A general model was trained, using an RBF kernel on a subset of 3000 sequences, created in a similar manner. The cross-validation used during training estimates that the accuracy of this general model is 88.3%, with a cost parameter of 2. I then tested this general model on the subsets used to test our other classes. The accuracies are shown in Table 2.6 below. Again, the structural set was not included in this table because no independent test set was available to validate this sample.

As Table 2.6 shows, sub-setting DNA segments into promoter, coding, noncoding, and presumably structural classes provided a small but clear benefit to predictions. This improvement cannot be the result of over fitting to the subset-specific models, since the test sets and training sets were independent for those models. Therefore, I can conclude that the subset-specific models performed better because the nucleosome positioning within each class must be more similar than the positioning of other classes. The improvement derived from sorting our genomic information could occur in 2 ways: either the features that are influential in nucleosome positioning vary between our genomic classes, or the weight of each feature varies between the classes. To try to better understand the differences between our general and linear models, I proceeded to look deeper into our feature selection results below.

2.3.5 Function-specific features

Our feature selection scores can be used to better illuminate the performance our predictive models and determine the likely overlap between significant features across the classes. In performing this analysis, however, potential problems with our scores

Subset-specific model	Accuracy on Test Set	Benefit of testing on subset
Promoter	83.7%	1.77%
Coding	89.7%	1.43%
Noncoding	87.7%	2.47%

Table 2.6: Test set accuracy of subset specific SVMs on a general test set: I developed class-specific SVM models that were trained to distinguish nucleosome-bound and nucleosome-free DNA sequences in promoter, coding, and noncoding regions. I tested the accuracy of these models on a general set of nucleosome-bound and nucleosome-free sequences and measured the decrease in accuracy from class-specific subsets (see table 4). Note that all models performed better on class specific test sets. This suggests that some form of class-specific regulation.

must be considered. In retrospect, given the small amount of variability between the F-scores of the various k-mers that I used, the sample size of 1000 is a likely source of error. To quantify the significant features between our classes accurately, I must first verify that our feature selection within a class is reproducible. To this end, I selected our promoter subset as the class most likely to contain biologically relevant k-mers and repeated the F-score analysis on 5 independent subsets, each time taking the 20 most significant features; the second column of table 7 counts the number of times each feature was selected as significant. A similar tabulation was done across our different classes (column 3). The reproducibility of our feature selection is quite low. Based on these results, I cannot confidently assert set of significant k-mers within each class. Given more time, I would repeat our feature selection with different methods and larger sample sizes. Even better methods and a larger sample size, however, may not clarify our analysis. Biologically, the lack of reproducibility could also indicate that many of the kmers contribute relatively equally to nucleosome positioning. If this is the case, vary large sample sizes, even beyond the scope of what I have for our project, may be required to adequately and reproducibly select the most significant set of k-mers.

Dominance of periodicity features

Despite possible shortcomings, however, some results are still apparent from our data. Features 5460, 5461, and 5462 were the top three features in all promoter subsets,

strongly suggesting that this is a biologically relevant signal. Furthermore, despite the relatively low F-values, 3 features were present in at least 4 out of 5 promoter subsets (frequency of C, CTG, and AT). All of these sequences were relatively short (k_i=3) and contain definite GC content biases. Thus they are similar to kmers that have been shown to be significant in previous studies8. Additionally it appears as though I may be observing some region-specific feature differences. Although I do not have the statistical power to point to specific features, there appears to be far more overlap between our various promoter subsets then there are between different classes, as shown in Table 2.7. Many of the features that repeatedly occurred among promoter subsets were not repeated across different functional categories (AT frequency across 4 promoter subtypes but in none of the other classes, CC was significant in 3 promoter subsets but in none of the other classes, and 5 other features were significant in two of the promoter subsets but in none of the other classes). It is also worth noting that our overlaps between classes may be artificially inflated. Between the various classes that I compared, all overlapping features occurred between the promoter class and the noncoding class (with the exception of the periodicity features). Biologically, this is interesting as the actual regions responsible for gene regulation within promoters are small and likely dispersed throughout the sequence. Thus, our promoter regions likely contain many regions that could be more accurately classified as noncoding. With the exception of feature 16 (frequency of TA) which occurred in the coding, noncoding, and promoter classes, none of the other classes overlapped at all. This cannot be merely the result of sequence bias in our different classes, as the F-score specifically measures the ability of a feature to distinguish between nucleosome-bound and nucleosome-free sequences. Given these results, despite our inability to select specific, significant k-mers, it appears quite clear that I am seeing the signatures of class-specific feature selection.

Function-specific features

The very slight improvement in performance for our SVMs given by subsetting our data made us suspect that functional differences controlling nucleosome localization

Number of subsets	Number of features within different promoter subsets	Number of features across different functional categories
1	59	61
2	5	2
3	1	1
4	2	3
5	4	N/A

Table 2.7: Significant Features Across Subsets within a Class and Between Different Classes. 5 independent subsets of size 1000 were taken from our promoter nucleosome-bound and nucleosome-free genetic sequences and the F-statistic for each feature was measured. Features were sorted by the number of subsets in which they ranked among the top 20 features (column two). In column three features were sorted by how often they ranked among the top 20 features between subsets taken from coding, noncoding, promoter, and structural genomic regions. The large number of features occurring in one subset among promoter subsets is likely due to the relatively small number of samples used to do feature selection. Note, however, that the variability between different genomic regions is larger than that within the promoter subsets.

in promoter, coding, noncoding, and structural regions may be disguised by the strong role of features 5460, 5461, and 5462 in all of our subsets. I therefore wished to see if these features were dominating our models as much as the F-score may suggest (figure 1). To this end I retrained our SVMs using only the periodicity features to indicate if our k-mer features played a significant role in our predictions. The accuracy of these models, based on cross validation are shown in table 8.

As can be seen in Table 2.8, models composed of only of features 5460, 5461, and 5462 (which measure dinucleotide periodicities) perform almost as well as our models including 17 additional features. This adds credence to the explanation described above that nucleosome positioning may be dominated by a few features common to most, if not all, genomic regions. This also explains how general predictors can work as well as they do even if region specific kmers play a role. Furthermore it is worth noting that the improvements of our initial models over our general model correlate strongly with the benefit of kmer sequences in each class (Pearson's R =.9867). While this is extremely tentative with only 3 points, it may suggest that the general model performed worse, in part, because it did not pick up on region specific k-mer frequencies that are important for nucleosome localization.

Subset-specific model	Cross-Validation Accuracy	Benefit from k-mer sequences
Promoter	84.4%	1.07%
Coding	90.2%	.63%
Noncoding	87.4%	3.17%
Structural	89.6%	2.73%

Table 2.8: CV Accuracy of SVMs without kmers. SVMs were trained and tested on the same subsets as before, but only with the use of periodicity features. It can be seen that adding k-mer based parameters only slightly enhanced the accuracy of predictions. Values are all based on cross-validation.

2.4 Contributions

In this project, I have learned that I can slightly improve predictive performance by leveraging annotation data on DNA fragments to allow for function-specific nucleosome positioning prediction. However, the dominant sequence features are features that capture the periodicity of alternating AT and GC-rich dinucleotides that corresponds with a single turn in the DNA helix. This suggests that using only sequence information for nucleosome positioning prediction may miss other important factors. Finally, I provided evidence that there is likely function-specific kmers that influence nucleosome positioning, though this signal is overshadowed by dominance of universally-important periodicity features.

Chapter 3

Epigenetic regulation of olfactory neuron specification

In olfactory neurons, there is a strict rule that each neuron must express exactly one allele of one of the 1300 olfactory receptor genes. However, the mechanism behind this monogenic expression is not yet fully understood. In this project, I found that in the olfactory epithelium of mice, olfactory receptor genes are marked in a highly dynamic fashion with the molecular landmarks of constitutive heterochromatin. The cell-type-dependent deposition of H3K9Me3 and H4K20Me3 along the clusters of OR genes is differentiation-dependent, and it is most likely reversed during the process of OR choice for monogenic and monoallelic expression. In contrast to the current view of olfactory receptor choice, which suggests that the silencing of the OR genes results from a feedback signal initiated by OR gene expression, our data suggests that OR silencing takes place before OR expression. This implies a new molecular role of chromatin-mediated silencing as the foundation upon which singular and stochastic selection can be applied, shown here in OR genes, but generally applicable.

3.1 Introduction

3.1.1 Problem Statement

Olfactory receptor (OR) genes are the genes that code for the receptors that detect smells. OR gene regulation is a topic of general interest, as OR genes are regulated in an unusual way: specifically, in each olfactory neuron, exactly one OR gene is expressed, while all the other OR genes must be silenced. This means that while each neuron has the *genetic* capacity to detect any smell, the receptor genes are regulated so every neuron actually detects exactly one smell. The combined power of all the neurons enable detection of a variety of smells.

The sense of smell is especially important to mice, as they are scavengers by nature, and they must take advantage of their powerful sense of smell to find food. Furthermore, mice are a well-studied model organism for humans, with many genetic similarities that allow findings in mice to often be applied to humans. Of course, with more experimental options for mouse than for human, it made mice an obvious choice for our study.

This project was a partnership with Prof. Stavros Lomvardass group of UCSF's Neuroscience Department, and we worked to discover and understand what the mechanism is behind monoallelic and monogenic olfactory receptor gene regulation in mice. Specifically, I performed computational analysis of genome-wide experimental data of epigenetic modifications (ChIP-chip data).

3.1.2 Background and previous work

Olfaction

Olfactory perception, or the sense of smell, takes place through the detection of volatile chemicals in the olfactory epithelium; the detection of these chemicals is then transmitted to the brain, which processes the information. In contrast to other sensory systems, olfaction requires a large family of $\tilde{1}000$ OR genes olfactory receptor (OR) genes, and these genes undergo a strict "one neuron-one receptor" rule.

That is, olfactory sensory neuron (OSN) are responsible for the detection of odors through olfactory receptors, and in each OSN, exactly one allele of one OR gene is expressed[6, 12]. This means that each olfactory neuron can detect exactly one kind of odor, dependent on which of the $\tilde{1}000$ olfactory receptors is expressed. Once OSNs detect the chemicals, they transmit signals through their axons to the olfactory bulb, the region of the brain involved in olfaction. The axons of olfactory neurons that express the same receptor meet up in the same glomerulus, a spherical structure in the olfactory bulb[48, 56, 74]; this is possible because the ORs play a role both in odor detection, as well as guiding the axons to the proper glomeruli, effectively determining the OSN's identity in this way[1, 20, 66, 76]. As ORs are important in both the wiring and physiology of olfaction, their proper expression is especially crucial.

The monoallelic and monogenic expression of OR genes is a difficult task: exactly one allele must be expressed at high levels, while the other $\tilde{1}000$ genes must be kept silent. The repression of the non-chosen OR genes must be extremely effective, since even a low level of transcription would result in thousands of inappropriately expressed OR molecules, due to the high number of OR genes; each individual receptor type would have low representation, the total OR activity of non-chosen alleles could be comparable to the activity of the chosen allele, possibly resulting in sensory confusion.

Previous work on olfactory regulation

In the mouse, about 1400 olfactory receptors are expressed in total in the main olfactory epithelium (MOE); they appear to be organized in a spatial and temporal fashion determined by positional clues[55, 58, 73]. Within each zone of expression, however, there are still several hundred alleles that could be expressed; only one of these alleles is actually transcribed in a seemingly stochastic fashion[67]. Previous experiments have implied that the production of OR protein elicits a feedback signal that prevents the expression of any other OR alleles, while stabilizing the expression of the chosen OR[41, 65, 68]. Additionally, the OR coding sequence seems to play an important role in the OR regulation, as there has been evidence to show that the coding sequence represses heterologous promoters[49]. Furthermore, both enhancers

and promoters contain regulatory information[59, 65]. In the past, the Lomvardas lab had shown that a specific enhancer, the H enhancer, interacts with active OR alleles, suggesting that this enhancer might be instructive for OR expression[46]. However, genetic ablation of the H enhancer only disrupted the expression of three proximal ORs, which makes it unlikely that it is singularly responsible for orchestrating OR choice[22, 51]. Therefore, the overall molecular mechanisms responsible for monoallelic and monogenic gene regulation are still unknown.

Chromatin-mediated silencing

Chromatin-mediated silencing is an effective form of transcriptional repression, and transcriptionally inactive chromatin is known as heterochromatin. Facultative heterochromatin is chromatin of silenced genes, and it is generally represented by hypoacetylation and di-methyl or tri-methyl groups on lysine 27 and/or dimethyl groups on lysine 9 of histone H3[72].Since facultative heterochromatin often silences genes in some environments and not in others, it is dynamic and appears to be developmentally regulated[3, 19]. On the other hand, constitutive heterochromatin is usually found in structural regions, such as pericentromeric and telomeric repeats, and it remains tightly packed during the cell cycle and stable during differentiation[21, 62].

3.1.3 Approach

In our project, we tested the hypothesis that chromatin-mediated silencing prevents the expression of OR genes in the sensory neurons. The Lomvardas lab generated Chromatin ImmunoPrecipitation on chip (ChIP-chip) data, which provides genome-wide data for presence of epigenetic modifications, as well quantitative PCR (qPCR) validation at specific locations. I computationally analyzed ChIP-chip data for quality control, normalization, identification of regions with histone marks, and statistical quantification of significance. Furthermore, the Lomvardas lab performed additional experiments to explain and validate our findings.

3.2 Methods

3.2.1 ChIP-chip experiments

As aforementioned, Chromatin ImmunoPrecipitation on chip, abbreviated as ChIP-chip, is a technique that can be used to find what regions of the genome have a certain histone modification, among other uses. The ChIP portion of the protocol isolates out DNA that is bound to specific types of histones. In ChIP-chip experiments, the isolated DNA sequences are then washed over a microarray chip that contains a matrix of probes, which are complementary DNA fragments; this allows for the identification of the isolated DNA sequences that match the probes. Based on the coloring of the cells on the chip, one can identify the intensity of the signal for each probe. By mapping each probe sequence back to the genome, one can estimate how likely it is that the chosen histone modification was present across the genome.

3.2.2 Data processing, normalization, and quality control

Quality control

Since ChIP-chip is an experimental method, the possibility of experimental problems is always a threat. Therefore, I generated graphs to assess the quality of each set of data in a number of ways, through a standard set of techniques[71]. Since ChIP-chip data gives both an amount for ImmunoPrecipitation (IP) - which is the type of DNA the protocol specifically pulled down - and 'input' - which is our control, I can compare the data distributions between the two. Ideally, one hopes for a normal distribution for both the IP and input data, so I plotted the distributions of the two types of data and observed if they look approximately normal. To better quantify how 'close' to normal the distributions are, I used quantile-quantile plots (also known as Q-Q plots), which compares the actual distribution with the normal distribution based on their respective quantiles. Additionally, I generated a 'MA' plot of the ratio of IP/input (on the y-axis) against the average of IP and input signal. This checks if our data has the common problem where ratios of IP/input tend to increase with

increasingly strong average signals. Therefore, the ideal average of the cloud of points would be horizontal line, where the average ratio does not change with the average signal. Finally, I calculated the standard deviation for various signal intensities, again checking for major skews.

Dataset normalization

Since ChIP-chip data is an experimental method, noise will inevitably be present in the data; this requires normalization within one set of data, as well as across multiple sets of data, as they are being compared to one another. I chose to try a number of different number of normalization techniques, trying to determine which best corrected for our experimental noise and bias. I used pre-existing normalization methods such as quantile normalization, which is a conservative normalization that fits the experimental data to a standard distribution, variance stabilization and normalization[31], which normalizes for the varying intensities of microarrays, and global normalization[80], which uses the median and standard deviation of log intensity ratios to correct the data for comparison across datasets.

Additionally, I developed a tailored form of normalization to suit our use of replicates and data states, which I call 'weighted global normalization;' this method was similar to the standard global normalization, except that it weighted the data for each of the states (H3K9me3 in OE, for example) equally, in spite of how many replicates there are for a given state. Specifically, each sample of data is subtracted by its median and divided by its mean absolute deviation (MAD), as in usual global normalization. Then, the weighted global median and weighted global MAD is calculated by first finding the average median and average MAD within each state, and then averaging these values across all four states. Then, analogous to global normalization, these 'weighted global' statistics are used to scale all the dataset values back through multiplication of the data by the weighted global MAD and addition of the weighted global median. Since I expect each state to have a similar distribution, this allows each state to weight the global distribution similarly, even if certain states have more replicates than others. The formula for weighted global normalization is

shown below, with X_i representing the post-normalized log ratio for dataset i; IP_i is the original immunopreciptation signal and $input_i$ is the original input signal, while m_i and d_i are the median and mean absolute deviation, respectively, for dataset i.

$$X_i = (\log \frac{IP_i}{input_i} - m_i)(\frac{1}{d_i})d_g + m_g$$

m_g and d_g are the weighted global median and MAD, as described above. That is, let m_s and d_s represent the average mean and MAD for a single state, where n_s is the number of replicates for state s. Then, in the formula, one sums over all i's for $i \in S$ where S is the set of indices of the datasets in the state s.

$$m_s = \frac{\sum m_i}{n_s}$$

$$d_s = \frac{\sum d_i}{n}$$

Then, the weighted global median m_g and d_g are calculated as follows, given that n_g is the total number of states:

$$m_g = \frac{\sum m_s}{n_g}$$

$$d_g = \frac{\sum d_s}{n_g}$$

3.2.3 Detection of heterochromatin domains

I chose to detect heterochromatin domains through two general approaches: sliding window and hidden markov models (HMMs).

Sliding window approach

Since ChIP-chip data gives us an analog signal rather than a digital one, the data must be interpreted into regions that have the presence of the histone modification and regions that do not. Many techniques can be used to turn the probe data into

finite binary peaks of enrichment. One powerful method for this is the sliding window approach[34], which slides a window of fixed size across the genome, averaging over the probes present in that window; if the resulting average meets the enrichment threshold, that window is considered a peak.. Variations on this general approach have been developed for specific experimental data: for example, the Model-based analysis of 2-color arrays[70] (MA2C) specifically corrected for sequence-specific biases based on GC probe content. Another consideration was the recent suggestion of large regions of chromatin k9 modifications, or LOCKs[79]; we hypothesized we might find large regions of modifications, or heterochromatin domains, as OR genes are often already clustered together in the genome. Therefore, while most research usually focuses on finding peaks via peak-calling, I specifically checked for large regions of enrichment, or what I call blocks. This was accomplished by using both the MA2C[70] and LOCKs[79] protocol, but adjusting the parameters for our data and goals. In the LOCKs methods, averaging was performed across 500 base pair windows, while the minimum block size was 10,000. In the MA2C pipeline, we used 2 sets of parameters: one to find smaller 'peaks,' and the other to find broader 'blocks.' For the peaks, we used a window to be 500 bp, with a FDR $\leq$ 5%, while the 'blocks,' were found by using a window of 10kb, with the minimum number of probes in a window being 20, and the maximum gap of being 1,000 base pairs.

Hidden Markov Model approach

I also used Hidden Markov Models (HMMs) to detect domains of histone modifications. A HMM is a statistical model where there are various states and probabilities of transitions and emissions. In this context, the emission was the intensity of the ChIP-chip signal, and there were either two or three states: enriched and repressed was used for the two-state HMM, or enriched, neutral, or repressed was used for the three-state HMM. However, since I did not know which areas were enriched or not, I had to use unsupervised learning with random initializations to train the model, and then find the assignment of states to the signal that maximizes the probability of it being produced by the model.

3.2.4 Clustering and ranking

Gene representation

To represent each OR gene, I chose to follow a protocol previous used to identify histone modifications at human enhancers[29]. Specifically, for each gene and modification, I centered a 10k basepair window at the translation start site. Each 2kb window consisted of 20 buckets of 100 basepairs each, where every probe's $\log \frac{IP}{input}$ was added to the appropriate bucket, and all values in a bucket were averaged, including data from replicate experiments. Since there were many modifications, the values for each modification H4K20me3 in OE tissue, H3K9me3 in OE tissue, H3K9me3 in liver tissue, and H4K20me3 in liver tissue - were concatenated.

Clustering

Once I had generated the representation for each gene, I chose to use a standard k-means clustering algorithm[15] to group genes based on their signal; this allowed us to identify potential patterns in signal across the four states. Cluster 3.0[15] was used to group the genes into four clusters. By tracking which genes were OR genes, I was able to calculate how many OR genes and non-OR genes were in each cluster, and whether there were patterns in histone enrichment for different subclasses of OR genes.

Ranking

I also ranked the genes by average enrichment for the histone modifications in olfactory epithelial tissue, to see which genes had the most enrichment for these heterochromatic marks. This was done by taking the 20 buckets for each OE state (H3K9me3 in OE and H4K20me3 in OE) and averaging across all 40 buckets, and then simply ranking the genes from highest average value to lowest average value.

3.3 Results

Our data show that, in the olfactory epithelial tissue, an unusual form of heterochromatic silencing is present at OR genes. Our ChIP-on-chip experiments show a very strong signal for H3K9me3 and H4K20me3 both specifically and sensitively at OR genes in MOE tissue. The cell-type and differentiation-dependent presence of these trimethyl histone modifications at clusters of OR genes results in compacted and inaccessible heterochromatic macrodomains. Surprisingly, these heterochromatic marks are found developmentally before OR transcription, implying that it is not the product of a feedback signal from OR expression. At an active OR allele, I see a significant reduction for the H3K9me3 and H4K20me3 modifications, with a strong signal instead for the H3K4me3 histone modification, often associated with active gene expression. Lastly, I found that insertion of a reporter transgene within a heterochromatic macrodomain results in OR-like expression of this transgene instead of ubiquitous expression, as the transgene is silenced in most of the olfactory neurons. With this evidence, we believe that stochastic escape from heterochromatic silencing might be the basis of monogenic and monoallelic OR gene expression.

3.3.1 Quality control

While most of the data sets were, unsurprisingly, not ideal distributions, they generally could be shown to have no major problems. An example of the quality plots that were generated for a single set of data are shown below in Figure 3-1.

3.3.2 Whole-genome analysis of H3K9me3 and H4K20me3 in MOE tissue

Using the gene representation described in the methods section, I was able to observe the presence of histone marks at genes all across the genome. Using heatmaps to represent enrichment with red and absence with green, I could organize the genes by chromosomal positions. For example, in Figures 3-2, 3-3, and 3-4, I show the genes

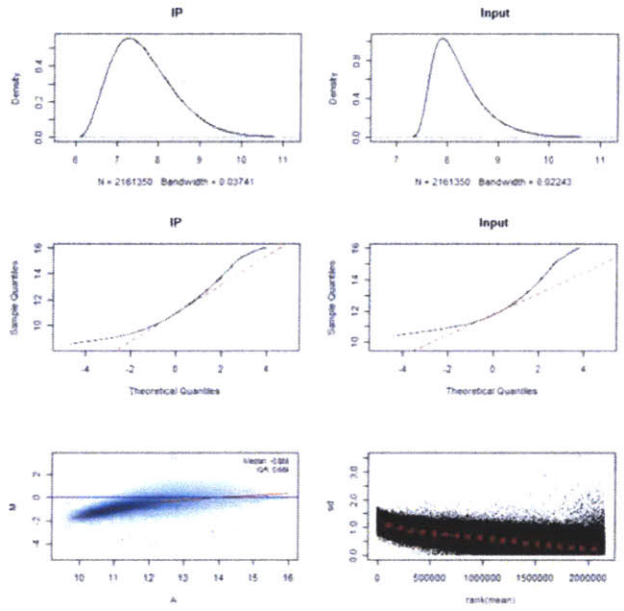


Figure 3-1: Figure 1: Quality control plots are shown for a sample array, 39849902, which represented liver tissue with the trimethyl k9 modification. IP indicated the immunoprecipitated DNA; Input shows the control. First row shows distributions (a normal distribution is expected); the second row demonstrates how close to normal the distribution is (the red line is perfectly normal). The bottom left plot shows the log ratio (M) of IP over input against average signal (A) of IP and input, where each dot is a probe; the ideal trend is the horizontal blue line. The bottom right plots the standard deviation (sd) against rank of signal intensities; the ideal trend is again a horizontal line.

in chromosomes 2, 7, and 9, in chromosomal order, with the rows that correspond to OR genes represented in blue, while other chemoreceptor genes are indicated in orange. This is an effective way to qualitatively study the correlation between the heterochromatic marks and OR genes, as OR genes are positionally clustered together in a few chromosomes, especially in the presented chromosomes.

It is immediately obvious that the histone modification enrichment is specifically and sensitively correlated with OR genes in MOE tissue, as can be especially seen in the OR clusters in Figures 3-2 and 3-3. Most genes, independently of their transcription status, appear to be devoid of both modifications in both tissues. However, in the MOE, there is significant enrichment for H3K9me3 and H4K20me3 on ORs. Additionally, it should be noted that the presence of these marks is present in a tissue-specific manner; that is, the correlation is very strong in OE tissue (left columns) and much less strong in our control liver tissue (right columns). Vomeronasal receptor (VR) genes, which encode receptors that detect pheromones, are also enriched for H3K9me3 and H4K20me3 in the MOE, as can be seen in Figure 3-3 by the genes marked with orange. ORs and VRs are hypomethylated in the liver in agreement with published observations that report the complete absence or the low abundance of these marks on OR genes in numerous cell types[7, 27, 38].

Additionally, as can be seen in Figure 3-4, there is also some enrichment for H3K9me3 and H4K20me3 non-OR chemoreceptor genes, although it is not strong as the enrichment at OR genes. Specifically, clusters of Vomeronasal Receptor (VR) and Formyl-Peptide Receptor (FPR) genes shown in Figure 3-4 reveal presence of heterochromatic markers similar to that of OR genes, but at a slightly lower level.

3.3.3 Heterochromatic signature for chemoreceptors

Using the aforementioned method for clustering, I performed an unsupervised 4-means clustering on the genes in chromosome 2 to identify potential epigenetic signatures of OR genes. The results of the signals in the 4 clusters are shown below in Figure 3-5. The clusters roughly correspond to tiers of strength of enrichment for the histone marks.

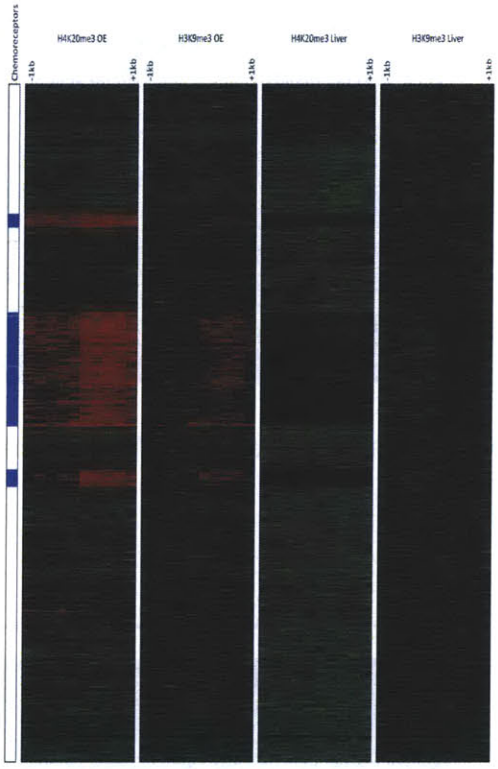


Figure 3-2: Genome-wide mapping of H3K9me3 and H4K20me3 reveal a tissue-dependent heterochromatinization of the ORs in the MOE. ChIP-on-chip experiments with antibodies against H3K9me3 and H4K20me3 using native chromatin preparations from the MOE and liver. The log2 ratio of IP/input was calculated and used for the construction of the heatmaps presented here. Positional heatmaps of chromosomes 2 is shown here. Each row represents one gene in 100 bp windows from -1kb to +1kb of the TSS. Four states are shown as adjacent columns: OE-H4K20me3, OE-H3K9me3, liver-H4K20me3, and liver-H3K9me3. OR genes are indicated in blue, while other chemoreceptor genes are indicated in orange.

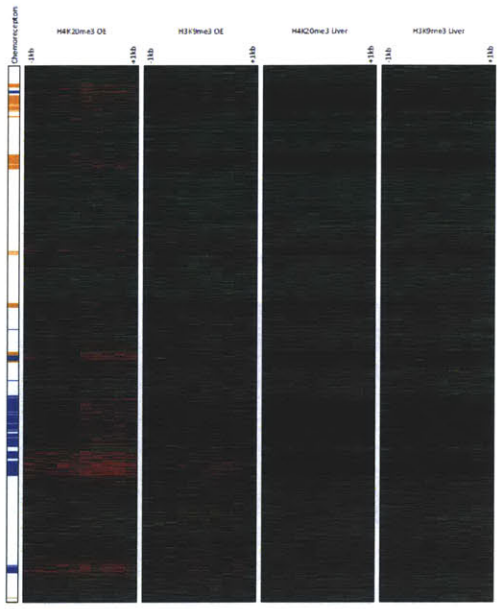


Figure 3-3: Positional heatmap of chromosome 7, as described above for chromosome 2.

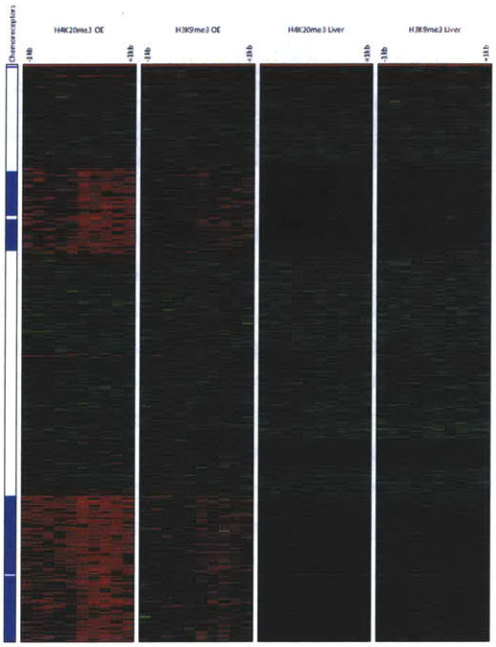


Figure 3-4: Positional heatmap of chromosome 9, as described above for chromosome 2.

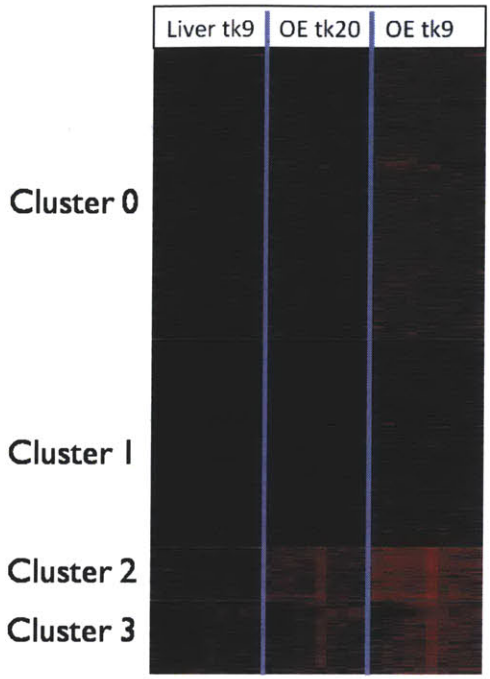


Figure 3-5: Result of unsupervised 4-means clustering on chromosome 2

By tracking which genes were OR genes, I was able to identify that OR genes were strongly clustered together, as shown in Table 3.1. Almost all the OR genes are present in the 2 clusters that correlate with significant enrichment for histone modifications; furthermore, the cluster with the strongest signal is nearly solely composed of OR genes. All these findings show that the H3K9me3 and H4K20me3 modifications are strongly associated with OR genes. The similarity between the pattern of epigenetic marks on OR genes indicates that these histone modifications are likely involved in OR gene regulation.

Type of genes	Cluster 0	Cluster 1	Cluster 2	Cluster 3
OR genes	33	15	161	163
non-OR genes	895	649	9	71

Table 3.1: Distribution of OR genes and non-OR genes in clusters. OR genes are almost universally grouped into the clusters representing high enrichment for histone modifications.

After studying the clustering of the OR genes based on histone modification enrichment, it quantitatively confirmed that the 'pattern' for OR genes was simply a strong presence for the heterochromatic marks, as suggested qualitatively by Figures 3-2, 3-3, and 3-4. Therefore, I ranked the all genes in mouse based on the average signal intensity of H3K9me3 and H4K20me3 as previously described in the methods to see how strongly genes enriched for the histone modifications correlated with OR genes. To present the data in a visually comprehensive manner I included only every 15th mouse gene in the presentation, although the analysis was performed for all the genes. In Figure 3-6, on the left, 1,000 randomly selected genes are ranked in descending order based on their average enrichment values for the two modifications. OR genes, depicted by blue lines at the side of the heatmap, are clustered on the very top, showing that they are the most enriched genes for H3K9me3 and H4K20me3 in the MOE. In a zoomed-in view of the top 1,000 genes in Figure 3-6 on the right, OR genes constitute the majority of genes with significant enrichment for both trimethyl-marks. Using the rank-sum test, I calculated a p-value of less than $10^{-}7$ for the OR genes having such a high level of enrichment. Notably, as shown also in Figure 3-3, type I OR genes that are organized in a unique cluster on chromosome 7 have the lowest enrichment values among ORs.

Most of the non-OR genes that are enriched for H3K9me3 and H4K20me3, represented by orange lines in Figure 3-6 are also chemoreceptors, namely VRs and Formyl-Peptide receptors (FPRs), which matches our previous findings from Figure 3-4. These VR and FPR genes are generally clustered in extremely AT-rich isochores and likely follow the same regulatory logic as ORs, which explains their similar, but lower-level, heterochromatinization[18, 42, 57]

3.3.4 Heterochromatic macrodomains cover OR clusters in MOE tissue

I identified regions across the genome with a strong signal for the histone modifications H3K9me3 and H4K20me3 with both the sliding window and hidden markov models,

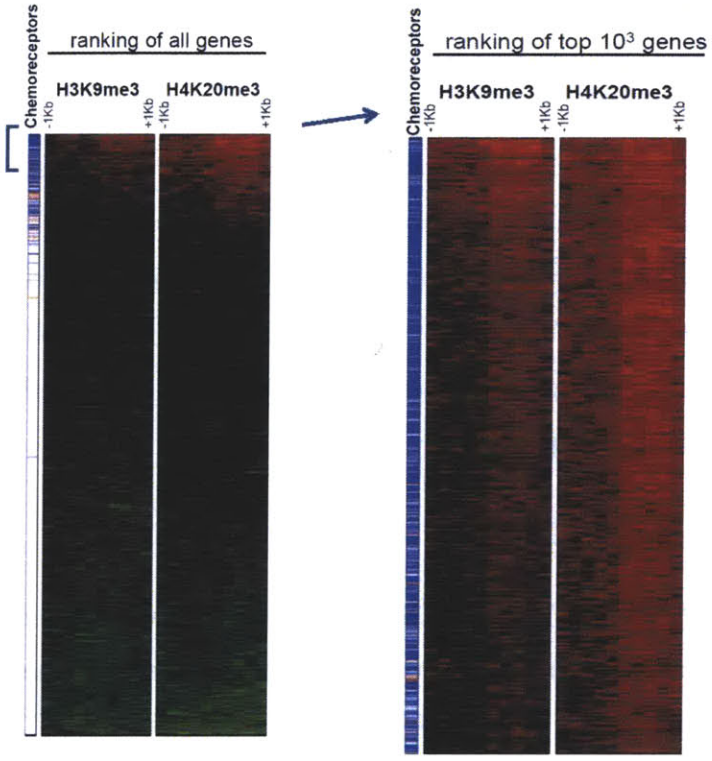


Figure 3-6: Ranking of genes based on enrichment for heterochromatic marks from strongest to weakest, using the previously described gene representation. The p-value for OR genes ranking so highly is less than $10^{-}7$ with the rank-sum test.

but I found the sliding window approach to be much more appropriate for our data.

Hidden Markov Models

I analyzed the data with Hidden Markov Models (HMMs) as described in the methods. However, with our data, I did not have success with unsupervised HMMs, as the maximum probability assignments resulted in each state having approximately the same proportion of the genome assigned to it (50% for 2-state HMMS or 33% for 3-state HMMs). Though this increased the sensitivity for classifying OR genes as enriched, specificity was very important to us, due to the size of the genome, so these results did not correspond with biological significance. Furthermore, when I adjusted the initialization parameters to make the enriched state have a lower probability, as this would increase specificity, the unsupervised learning struggled with a lack of data for the enriched state. Therefore, since I found much more biologically meaningful results with both sliding window techniques, discussed below, I chose not to use the HMM analysis for this study.

Sliding Window

Using both the MA2C[70] and LOCKs[79] protocol, we were able to identify large domains of histone modifications. Furthermore, it was clear that these heterochromatic macrodomains covered clusters of OR genes in MOE tissue.

Since the two protocols had different benefits and drawbacks, as described in the methods, we decided to use both of them on our data and compare the results, with both algorithms set up to find broader range 'blocks.' From our results, visualized in Figure 3-7 with the Integrated Genome Browser[50], it was clear that both methods found very similar domains of histone modification; this was very promising as it told us that the biological signal from the data was so strong that the small technical differences in protocols did not strongly affect our findings.

Since both protocols produced such similar results, we chose to continue the analysis using only the results from the MA2C algorithm. As hinted at in Figure 3-7, we found that in the MOE, the 'peaks' for the two histone modifications were strongly

Figure 3-7: Blocks (in blue) were found from ChIP-chip for trimethyl K20 modifications in OE tissue with the LOCKs protocol (first row) and the 'blocks' parameters for the MA2C protocol (second row). Mouse genes are shown in green (from mm8 reference), with positive strand on top, and negative strand on bottom. The range of the image is on chromosome 2, spanning from about 84 Mb - 94 Mb, while the blocks range from 85 Mb 90.2 Mb, which matches with a cluster of 300 OR genes.

clustered together in broadly enriched genomic regions throughout the OR clusters in an almost continuous arrangement (Figure 3-8). Therefore, we modified the parameters to find broad 'blocks' of enrichment, as described in the methods, and we confirmed that H3K9me3 and H4K20me3 form heterochromatic macrodomains (blocks) that cover megabases of clustered OR genes in the MOE (Figure 3-8). Quantitatively, we found that 1376 ORs fall in H4K20me3 blocks and 1109 ORs fall in H3K9me3 blocks, out of a total of about 1441 annotated OR genes, which corresponds to a p-value $< 10^{-7}$.

Again, as expected, the low signal for H3K9me3 and H4K20me3 in the liver tissue resulted in very few peaks and blocks on the OR genes in the liver (Figure 3-8); furthermore, the few peaks or blocks that were found were not close together, and ChIP-qPCR confirmed that their enrichment for the histone marks was, in fact, very low. It is unsurprising that there were a few spurious peaks or blocks found, since these sliding window algorithms somewhat base their enrichment threshold relative to the signal in the entire dataset; therefore, if there was a low signal all across the genome in liver, then peaks and blocks would be called for regions that showed stronger enrichment than the rest of the genome in liver, but that still corresponded

to low enrichment when compared to the strong enrichment at OR genes in OE tissue.

We further validated our ChIP-chip results by quantitative PCR (qPCR) for multiple OR gene clusters in both tissues. Whereas ChIP-chip can give a noisy signal across the entire genome, ChIP-qPCR can give a more precise signal for a very specific location. qPCR for representative genes, as boxed in 3-8 are shown in Figure 3-9.

We also noted that the borders of the heterochromatic marks strongly coincided with the borders of OR loci, as shown in Figures 3-8 and 3-10. The reported binding of CTCF outside of OR clusters [35] or other insulating elements [16], may play a role in the borders of OR heterochromatin aligning with OR clusters. Additionally, the data shows that the presence of transcriptionally active non-OR genes in an OR cluster interrupts the heterochromatin blocks, until the next OR gene reconstitutes the heterochromatin (Figure 3-10). On the other hand, transcriptionally inactive non-OR genes in OR clusters are partially covered by the histone modifications, which implies that in the absence of a competing need for transcription or insulating activity, the heterochromatin can extend over non-OR genes within an OR cluster.

3.3.5 Further experimental validation

To further study and validate the findings of the ChIP-chip and ChIP-qPCR data, the Lomvardas lab performed more experiments to investigate the relation between the heterochromatic histone modifications and the OR genes in OE tissue.

Characterization of OR heterochromatin

To determine if the histone modifications present at OR genes in MOE tissue resulted in functional differences of the chromatin, the Lomvardas lab analyzed the accessibility of the DNA at different loci. This was accomplished through the treatment of nuclei from MOE and liver tissue with DNaseI to cleave the DNA, and then measuring the amount of DNA at specific loci with qPCR. As demonstrated by Figure 3-11, we found that OR genes in MOE tissue were much less digested, and therefore, most likely less accessible, than transcriptionally active genes, while silent non-OR genes

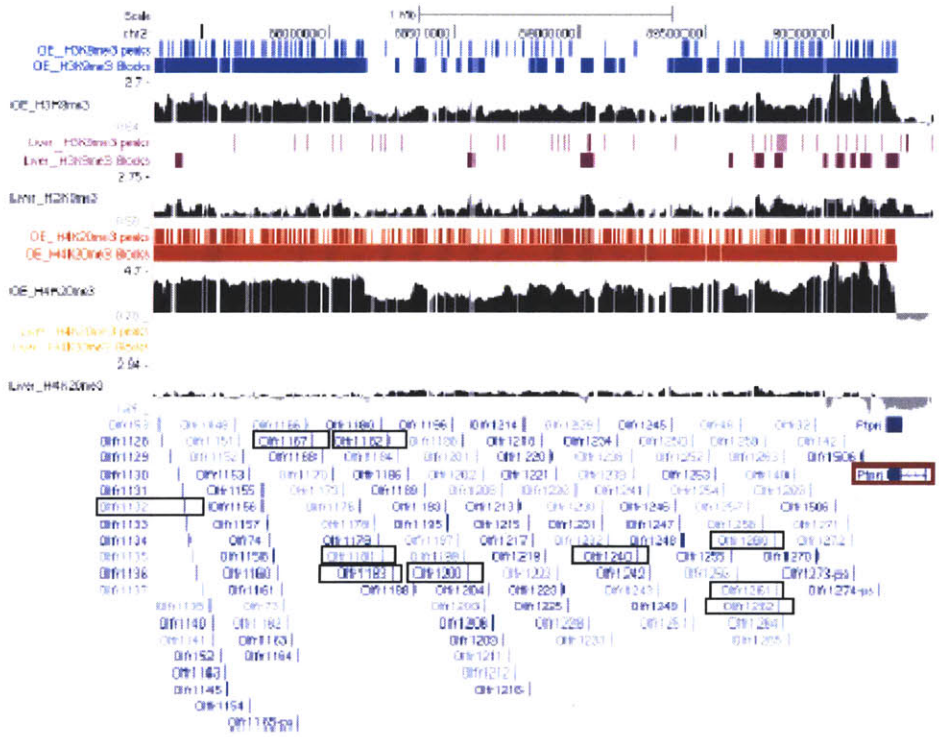


Figure 3-8: OR clusters in the MOE are surrounded by tissue-specific heterochromatic blocks of H3K9me3 and H4K20me3. Ma2C analysis of our ChIP-on-chip data viewed on the UCSC genome browser. This figure shows part of the biggest OR cluster located on chromosome 2, which contains 240 genes and spans a 5MB region. The thin blue (H3K9me3) or red (H4K20me3) bars represent significant peaks (FDR $\leq$ 5%) identified in the MOE by MA2C using standard parameters (window=0.5 kb, min number of probes= 5, max gap=0.25 kb); the thick blue or red bars represent the blocks identified with parameters for the identification of large-scale enrichment (window=10 kb, min number of probes= 20, max gap= 1kb). In the liver, there are only a few, sporadic H3K9me3 peaks and blocks (purple). Boxed genes have qPCR data in the following figures.

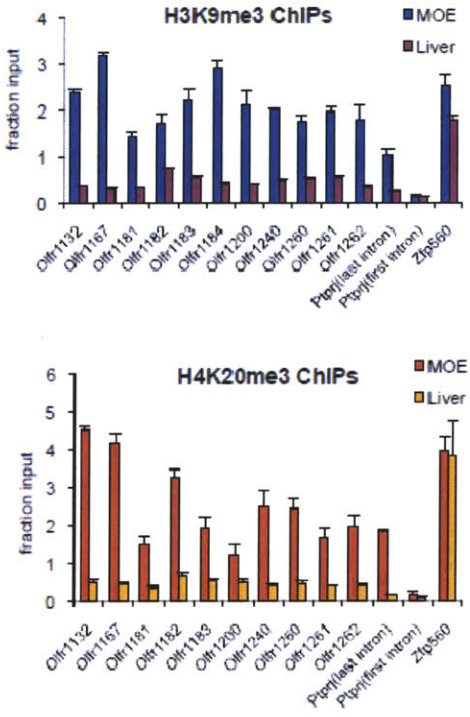


Figure 3-9: Results from H3K9me3 and H4K20me3 ChIP-qPCR analysis using native chromatin preparations from MOE and liver. The Ptprj gene stands at the border of the OR cluster which coincides with the border of the heterochromatic block. Its intron that is most proximal to the OR cluster is enriched for H3K9me3 and H4K20me3k, while its most distal intron is free of these modifications. Zfp560 serves as positive control.

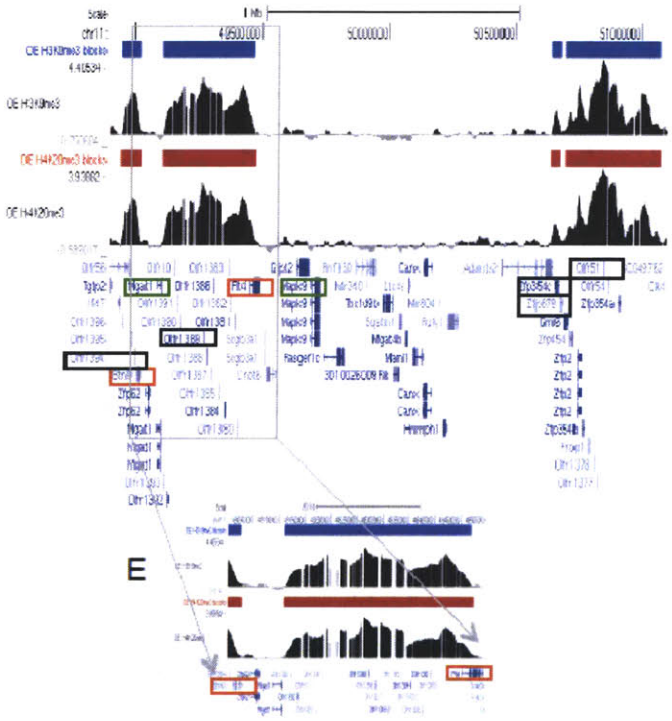


Figure 3-10: Part of an OR cluster on chromosome 11 is interrupted by a small group of transcriptionally active non-OR genes, marked in green. Genes marked by red rectangles do not have detectable transcripts, and they heterochromatic blocks extend over these genes.

had intermediate accessibility. In liver, OR loci were similar to other genes in terms of DNase I accessibility. These findings were also supported by other methods, such as southern blot analysis with a degenerate OR probe (not shown here).

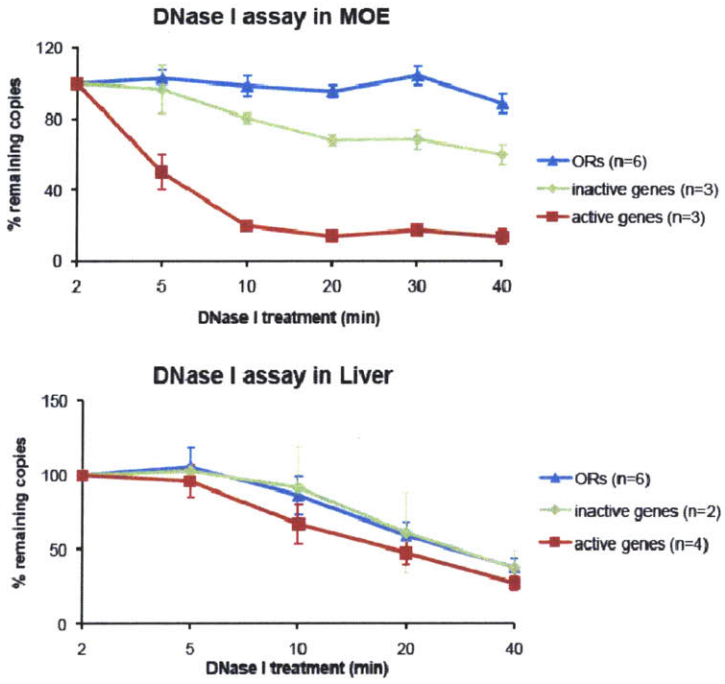


Figure 3-11: The ORs acquire a highly compacted chromatin structure in the MOE. DNase I accessibility assay with nuclei from both MOE and liver is presented here. Nuclei were treated with DNase I, DNA was isolated at various time points (2 to 40 min) and equal amounts were used for qPCR. The amount of DNA measured at each interval was expressed as a fraction of the DNA present at 2 min of enzyme treatment and was plotted over time. We assayed several ORs as well as genes that are active or inactive in the MOE or liver, and their mean is shown here, with representative data from one experiment. In MOE, the ORs appear to be more resistant, suggesting they are less accessible.

OR silencing independent of OR expression

Since the MOE tissue is composed of multiple cell types[17], we performed experiments to confirm that our results in OE tissue actually reflected the state of the OSNs specifically. For this purpose, the Lomvardas lab performed fluorescence-activated

cell sorting (FACS) experiments followed by ChIP-qPCR. That is, we isolated mature OSNs from OMP-IRES-GFP mice and, as seen in Figure 3-12, the OR genes tested have high levels of enrichment for both H3K9me3 and H4K20me3 in OSNs. Each OR gene was expressed in 0.1% of the OSNs, which supports the idea that the majority of OR genes would need to be silenced.

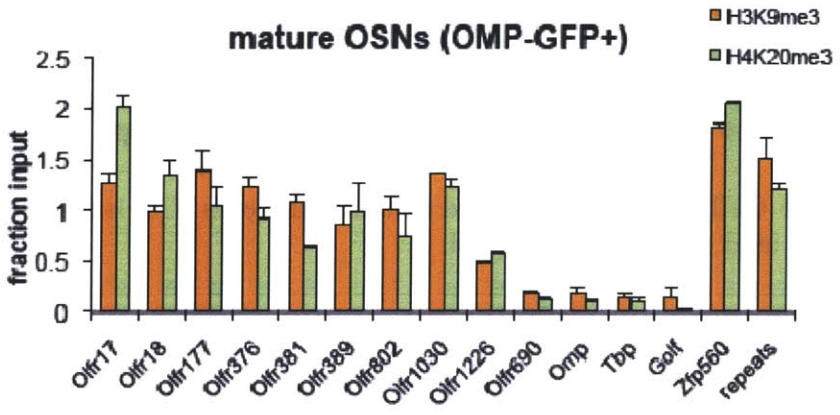


Figure 3-12: ChIP-qPCR assays for H3K9me3 and H4K20me3 in sorted cell populations from the MOE. GFP+ cells (mature OSNs) were isolated with FACS from OMP-IRES-GFP mice and were used for ChIP-qPCR experiments. Golf, Tbp and Omp are active genes in these cells that are used as negative controls, while Zfp560 and major satellite repeats are used as positive controls. Olfr690 is a type I OR.

Additionally, to determine whether the heterochromatic silencing was independent of or a result of OR expression, we sorted sustentacular cells from the MOE[9]; sustentacular cells are present in OE tissue and have common developmental ancestors with the OSNs, but they do not express ORs. As shown in Figure 3-13, we found similar levels of H3K9me3 and H4K20me3 in the sustentacular cells as in the OSNs, suggesting that marking of OR genes with H3K9me3 and H4K20me3 occurs in the absence of OR expression. This raises the possibility that trimethylation of lysines 9 and 20 takes place before OR activation.

To further investigate the possibility of heterochromatic silencing before OR expression, we performed ChIP-qPCR analysis in progenitor cells, starting with the most multipotent cells of the MOE, the HBCs[39]. Our results, as shown in Figure 3-14, indicate that there is no enrichment for H3K9me3 and H4K20me3 on OR

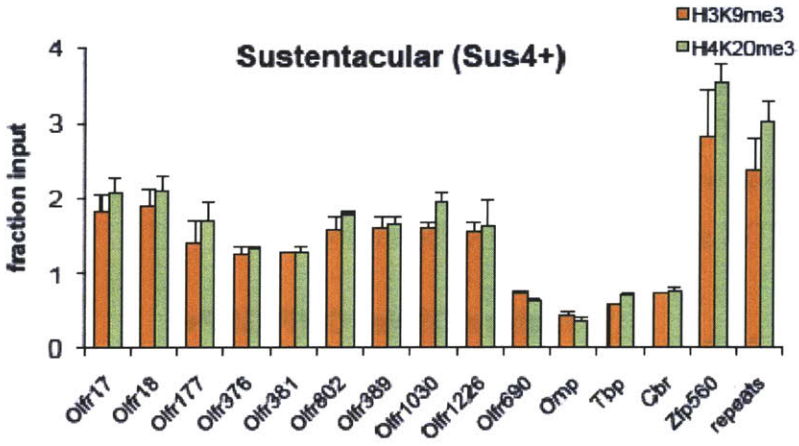


Figure 3-13: ChIP-qPCR with isolated sustentacular cells. Cbr is transcribed in these cells and is used as a negative control.

genes, although there is a strong signal for H3K9me2 (not shown), suggesting that in this multipotent cell, ORs are repressed via mechanisms that differ from repression in OSNs. Additionally, we checked the chromatin state of OR genes in other progenitor cells from the MOE that are negative for OMP, ICAM-1, iLR, and SUS4; the result was that the enrichment for H3K9me3 and H4K20me3 appeared to be as high as in the OMP+ cells in Figure 3-15, even though, according to RT-PCR, this population does not express ORs (Figure 3-16). Again, this suggests that the trimethylation of OR genes occur developmentally before OR expression.

To study a cell population that is more well-defined, we studied a Neurogenin1-GFP (Ngn1-GFP) BAC transgenic reporter mouse from GENSAT[28]. RT-PCR analysis showed that these cells represent a mixed population of progenitors and immature neurons (not shown). We found that Ngn1+ cells had 8-fold lower mRNA levels than the mature OSNs for 1185 OR genes (not shown), and, importantly, in the Ngn1+ cells, 95% of OR genes have transcript levels similar to the transcript levels of silent genes (data not shown). Therefore, the low levels of OR mRNA in these cells likely reflects a small percentage of contaminating mature OSNs. When we performed FACs and ChIP-qPCR on the Ngn1+ cell population, we found high levels of enrichment for H3K9me3 and H4K20me3 on OR genes, demonstrating similar heterochromatic

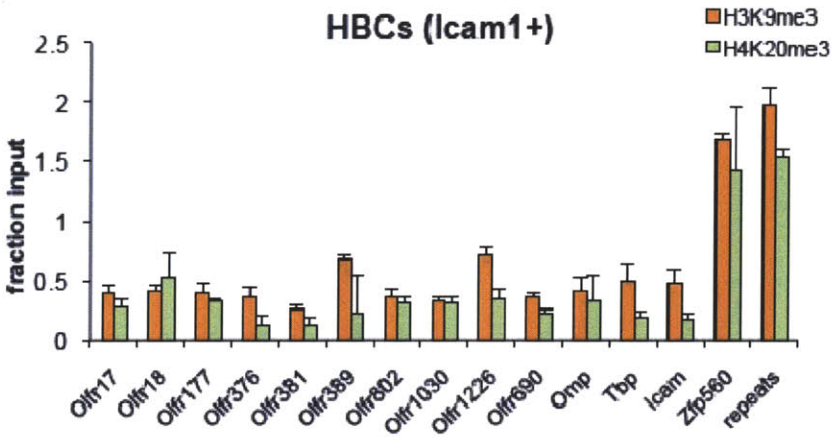


Figure 3-14: ChIP-qPCR experiments with isolated HBCs.

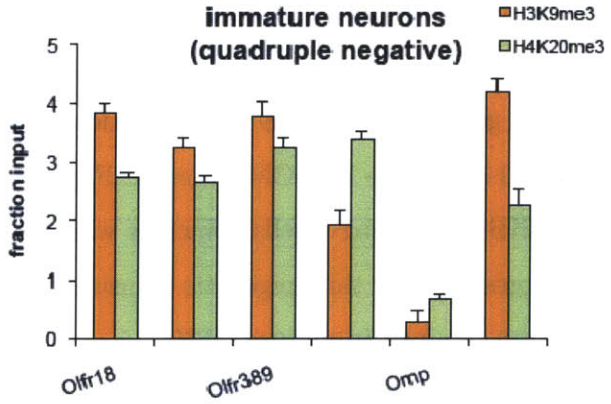


Figure 3-15: ChIP-qPCR with immature neurons and progenitors from the MOE isolated by collecting cells that are quadruple negative for OMP-, ICAM-, iLR- and Sus4-.

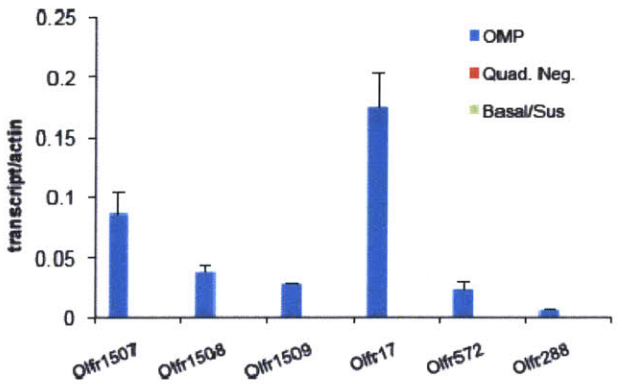


Figure 3-16: RNA isolated from combined OMP-GFP+, sustentacular and basal cells, or quadruple negative cells was used in qRT-PCR reactions with primers for different ORs. Actin was used as endogenous control.

signature with the mature OSNs (Figure 3-17). This confirms our belief in the contamination of the population; if only the few cells that exhibited expression of OR genes had contributed signal for the histone modifications, then the trimethylation signal would have also been 8-fold lower in the Ngn1+ cells. Therefore, the ChIP-qPCR data from the quadruple negative cells and Ngn1+ cells are consistent with H3K9me3 and H4K20me3 having been deposited on OR genes before OR expression.

We wanted to test the significance of the transition from di-methylation to tri-methylation at the OR genes during MOE differentiation, so we performed southern blot analysis on ICAM1+, Ngn1+ and OMP+ cells. Figure 3-18 demonstrates that the differentiation of HBCs to Ngn1+ cells coincides with increased protection from DNase I digestion, suggesting that this epigenetic transition results to a less accessible OR chromatin structure retained in mature OSNs.

Epigenetic switch accompanies OR choice

To investigate the state of the single active OR allele in OSNs, we used FACS to select neurons expressing the olfactory receptor P2 from P2-IRES-GFP knocked-in mice. We isolated 40,000 GFP+ and GFP- neurons, which, respectively, do and do not express the P2 allele, from P2-IRES-GFP heterozygote mice, and we performed

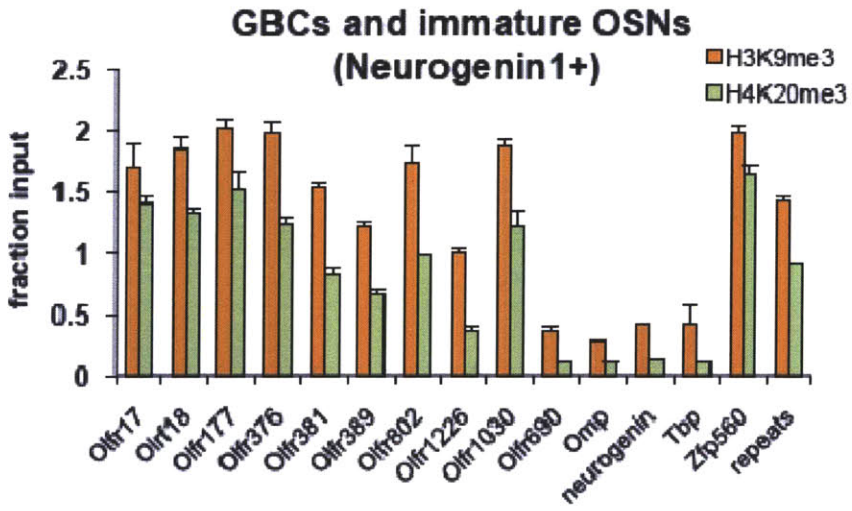


Figure 3-17: ChIP-qPCR analysis of the GFP+ cells that were isolated by FACS from Ngn1-GFP mice and were used for ChIP experiments for H3K9me3 and H4K20me3.

ChIP-qPCR for H3K9me3 and H4K20me3 on them. As seen in Figures 3-19 and 3-20, the enrichment for H3K9me3 and H4K20me3 is significantly reduced on the active OR allele, in comparison to the strong presence of the marks on P2 where it is not the active allele. Though the presence of these marks was reduced on the active allele, they were not completely removed; control experiments indicate that this is due to 30% contamination of the population, which is unsurprising since we were selecting for an extremely rare population (0.05% of total cells in the MOE).

Next, we performed a double FACS experiment to obtain a purer population; the GFP+ cells were sorted again, resulting in a $> 95\%$ GFP+ population, using MOR28-IRES-GFP heterozygote knock-in mice ,as they provide more GFP+ cells. As seen in Figure 3-21, ChIP-qPCRs from this extremely pure population provides strong evidence that H3K9me3 is absent from the transcriptionally active allele, MOR28.

To further probe the epigenetic state of the single active allele, we performed ChIP-qPCR on P2 for H3K4me3, a histone mark commonly associated with active promoters[25] that has a mutually exclusive distribution with H3K9me3 and

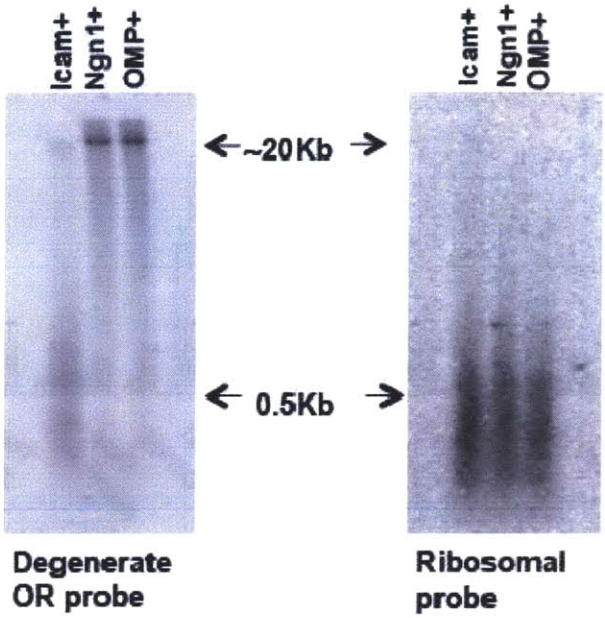


Figure 3-18: ICAM+ cells, Ngn1+ cells and OMP+ cells were sorted from the MOE tissue of adult mice. Their nuclei were extracted, digested with DNase I, and analyzed by agarose gel electrophoresis and Southern blot with a degenerate OR probe or a ribosomal probe.

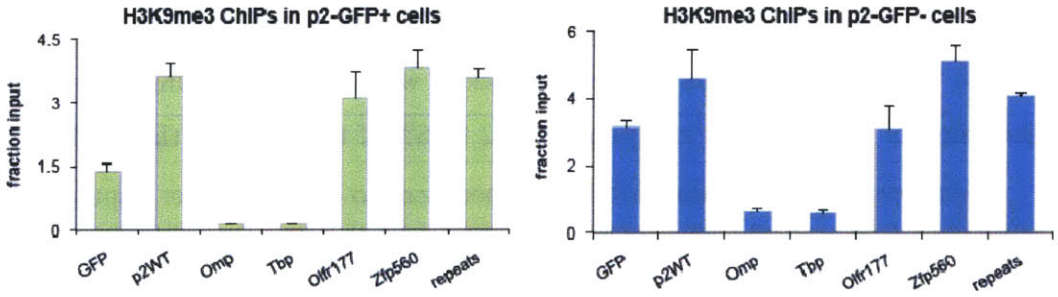


Figure 3-19: GFP is hypomethylated on H3K9 (left) in the GFP+ cells, where it is transcribed, but not in the GFP- cells (right), where this P2 allele is inactive. The inactive allele, amplified specifically by the p2WT primers, shows high enrichment for H3K9me3 in both GFP+ and GFP- populations. Omp and Tbp are used as negative and Zfp560 and repeats (major satellite) as positive controls.

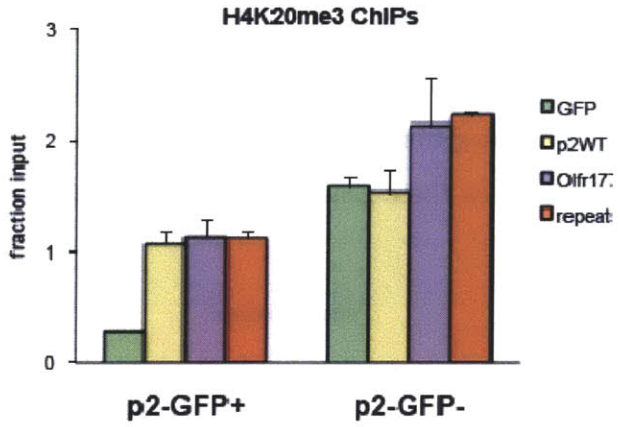


Figure 3-20: H4K20me3 ChIPs with P2-GFP sorted cells.

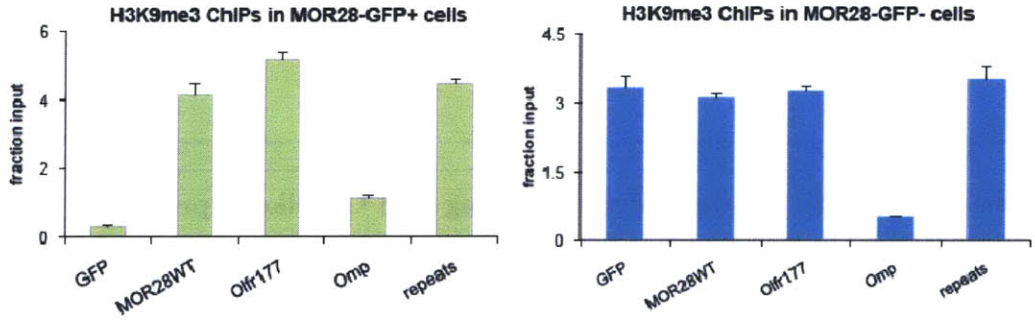


Figure 3-21: As above, but the GFP+ cells from the MOR28-IRES-GFP heterozygous mice were subject to a second round of FACS to yield a $> 95\%$ pure population, which were then used for H3K9me3 ChIPs.

H4K20me3[54]. As expected, H3K4me3 cannot be detected on OR promoters using chromatin preparations from the whole MOE (data not shown), but in Figure 3-22 there is enrichment for H3K4me3 only on the P2 promoter and CDS in the GFP+ population. This supports the idea that selection of the P2 allele is associated with the removal of H3K9me3 and H4K20me3. Although H3K4me3 is strongly present the active P2 allele, it is missing from the neighboring P3 and P4 genes 3-22, despite the sequence similarity between these genes and their expression in the same zone.

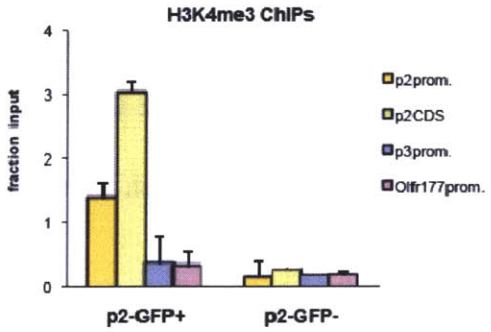


Figure 3-22: We repeated the same ChIP-qPCR experiment with an antibody against H3K4me3. There is significant enrichment for H3K4me3 throughout the P2 gene, but not on the neighboring P3 gene or a distant OR gene (Olfr177) in the GFP+ cells. As expected, there was no H3K4me3 on the P2 gene, or any other OR gene, in the GFP- cells. Values are the mean of triplicate qPCR, while error bars represent the SEM.

Heterochromatic marks induce silencing and OR-like expression

Our data suggested that heterochromatinization of OR loci was universally repressing the OR genes, so to test this hypothesis, we examined a transgenic mouse, where a OMP-LacZ transgene had been inserted proximal to a singular OR gene. Unlike numerous other OMP-LacZ or OMP-GFP independent transgenes that are expressed in the majority of olfactory neurons[49, 75], this transgene was silent in 99.9% of the neurons and has a sporadic and mostly zonal expression reminiscent of that of the neighboring OR[53].

Mapping the exact insertion site of this transgene revealed that it resides approximately 55kbs from Olfr459, as shown in Figure 3-23. ChIP-qPCR experiments showed that the insertion site is heterochromatinized in both the wild type and transgenic mice, as shown in Figures 3-24 and 3-25; ChIP-qPCR also indicates that the reporter is itself marked by H3K9me3/H4K20me3 in an tissue-specific fashion, in contrast to the endogenous OMP promoter, which is unmethylated (Figure 3-25).

Figure 3-23: Graphic representation of the Olfr459 locus and the OMP-LacZ insertion site located 55 kb away. Positions marked A, B, and C depict assayed regions in the qPCR analysis below.

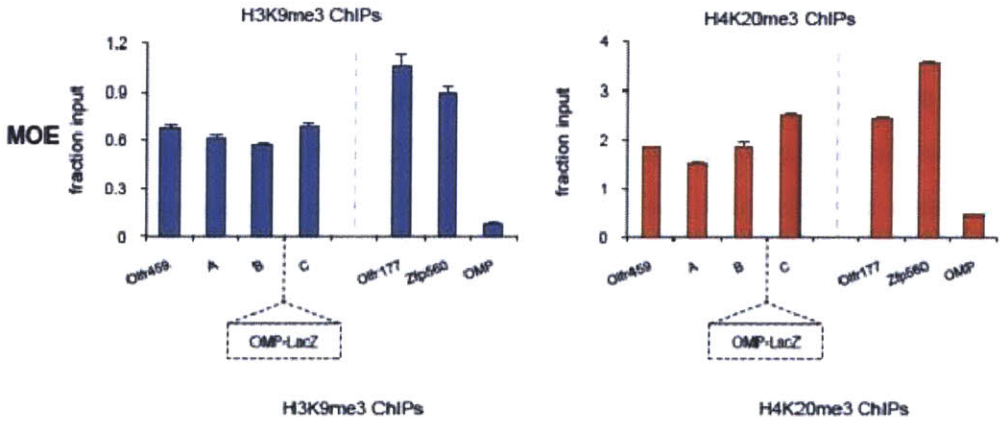


Figure 3-24: ChIP-qPCRs with chromatin from the MOE of wild type mouse show that the Olfr459 is enriched for H3K9me3 and H4K20me3. Both modifications appear to extend to the insertion site.

To examine whether the insertion of the OMP transgene resulted in monoallelic expression, we compared the number of β-gal+ cells between homo- and heterozygous mice. As seen in Figure 3-26, OMP-LacZ homozygotes have approximately 1.8 fold more β-gal+ cells than heterozygotes, consistent with a monoallelic expression pattern. Finally, to test whether the transgene is under the transcriptional control

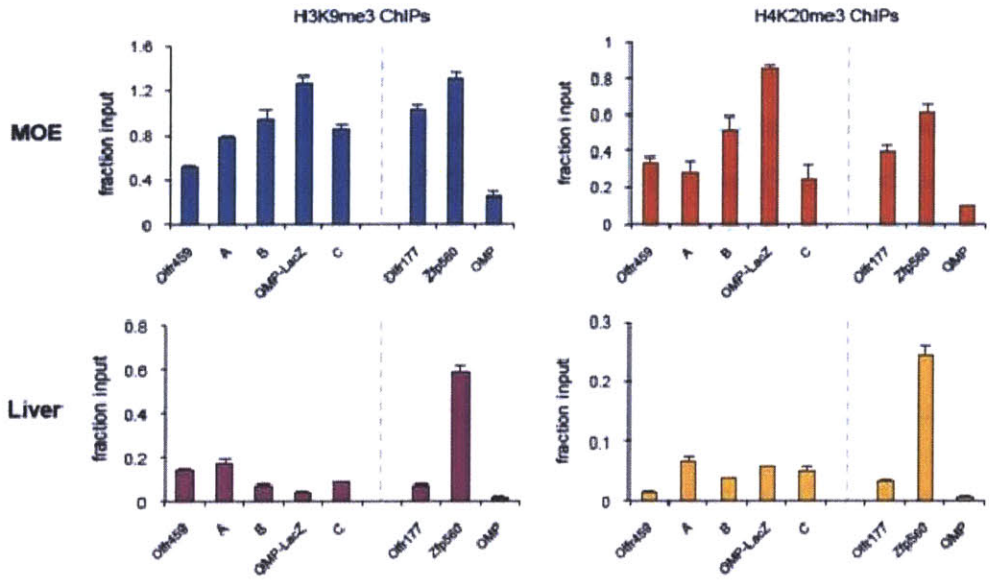
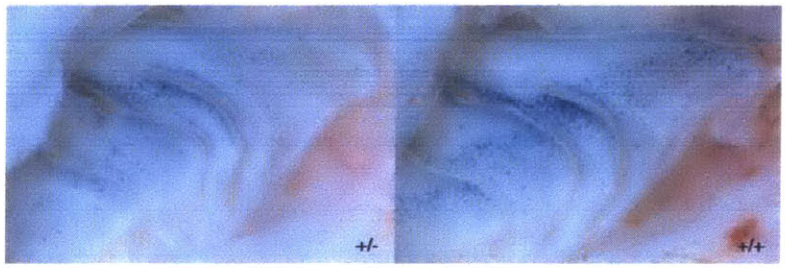


Figure 3-25: ChIP-qPCR analysis of the MOE and liver from OMP-LacZ positive animals. Both H3K9me3 and H4K20me3 show MOE-specific deposition on Olfr459, the OMP-LacZ transgene, and the regions proximal to these loci.

of the proximal OR locus, we crossed this transgenic mouse to the Emx2 knockout mice, as Emx2 is required for the expression of Olfr459[47]. We found that reporter expression is abolished in the transgenic - Emx2 knockout offspring, suggesting that this transgene conforms to the regulatory logic of the neighboring OR (not shown).

3.4 Contributions

In sum, our data strongly suggests that the presence of histone modifications H3K9me3 and H4K20me3 result in chromatin-mediated silencing of Olfactory Receptor (OR) genes developmentally before OR expression. The transcriptional activity of a single OR allele in an olfactory neuron is likely then made possible through the de-repression of that allele, with the repressive marks replaced with the active histone modification H3K9me2. This transcriptional activity then most likely triggers the previously-supported feedback signal that prevents the de-repression of any other OR alleles in that neuron.



OMP-LacZ	LacZ+ cells / area		N
+/-	51±3	} 1.8 *	3
+/+	93±8		4

Figure 3-26: X-gal stains of lateral whole mounts of the nasal cavities from hemizygote and homozygote OMP-LacZ animals. N, number of biological replicates. The calculated p-value was less than $10^{-}4$, as calculated by the Students t-test.

Chapter 4

Post-transcriptional regulation of RNA

Though mRNA is the classic example of RNA, other types of functional RNA have been shown to play an increasingly important role in gene regulation. For example, the discovery of hundreds of long intergenic non coding RNAs (lincRNAs)[26] and enhancer RNAs (eRNAs)[36] has given increasing evidence of the important role of post-transcriptional regulation. Post-transcriptional regulation of RNA often occurs through binding to the RNA, either with protein-RNA interactions or RNA-RNA interactions. Past attempts at identification of binding sites for post-transcriptional have had moderate success through use of sequence information[36].

Here, I search for signatures of functional RNA regions through RNA-Seq data, rather than sequence information. RNA-Seq measures the amount of transcribed RNA present across the genome, and I believe that this protein-RNA or RNA-RNA binding for post-transcriptional regulation can leave a signature in the RNA-Seq data. In this project, I computationally identified signatures of human RNA-Seq data for RNA regions of post-transcriptional regulation based on aggregate plots aligned for conserved regions of 3 UTRs and 5 UTRs. By comparing across different tissues, average types, alignment points, and UTR regions I found specificities for different conditions.

4.1 Introduction

4.1.1 Problem Statement

Animal genomes encode a variety of RNAs, including both protein-coding and non-coding RNA (ncRNA). Recently, thousands of instances of one class of ncRNAs, long intergenic non-coding RNAs (lincRNAs), have been identified[26]. Experimental results implicate lincRNAs as major contributors in the regulation of gene expression[32], but their mechanism of action is not known. Additionally, post-transcriptional regulation is a prevalent mechanism for gene regulation, as large amounts of mRNA are never translated into proteins. Lastly, the technique of transcriptome sequencing, or RNA-Seq, has become popular since its inception, due to the importance of measuring levels of RNA at high coverage with a reasonable cost. In this project, I combined these findings by leveraging biases in RNA-Seq data to identify signatures for RNA regions involved in post-transcriptional binding. This will allow us to eventually pursue *de novo* detection of function RNA regions. Furthermore, this provides information on what biases to expect to see in RNA-Seq regions, so future studies can more accurately use RNA-Seq data.

4.1.2 Related work

Dr. Loyal Goff, who works with both Kellis and Rinn laboratories, is currently working to develop an RNase assay. This assay will identify DNA that codes for protein-bound RNA, as these RNA are candidate functional regions involved in post-transcriptional modifications. The protocol for this assay will be to cross-link the RNA with proteins, digest the RNA that is not protected by proteins, uncross-link the RNA-protein complexes, reverse transcribe the remaining RNA, and sequence and align the resulting cDNA.

4.1.3 Approach

However, as this novel protocol presents many biological obstacles, I developed a novel computational method for Transcriptome Sequencing (RNA-Seq) data analysis to study protein-bound RNA. The Broad Institute had previously generated RNA-Seq data across 16 human tissues, I was able to directly use this data. RNA-Seq experiments measure levels of RNA transcription by using reverse transcription, but RNA bound to proteins will be 'protected' from reverse transcription, which should result in reduced levels of RNA-Seq reads in these regions. Since conservation generally signals functional importance[2], I identified an RNA-Seq 'signature' of functional RNA by examining aggregate plots of RNA-Seq signal at conserved 3' and 5' UTR regions of the genome.

4.2 Methods

4.2.1 RNA-Seq

RNA-Seq is a technique that takes advantage of next-generation sequencing technologies to profile the transcriptome[77]. In the past, genomic tiling microarrays were common to approximate the transcriptome, as they were high throughput and relatively inexpensive, and could reach a high resolution with specialized chips[14, 83, 4, 10]. However, drawbacks of genomic tiling microarrays include assumptions about the genomic sequence, problems of cross-hybridization and complicated normalization[77]. Other sequence-based approaches before RNA-Seq included Sanger sequencing of cDNA, but this approach was low throughput, expensive, and not quantitative[5, 24].

However, RNA-Seq has quickly become the dominant method of transcriptome profiling. The technique is to convert a population of RNA into a library of cDNA fragments; by sequencing the resulting cDNA in a high-throughput manner with either single-end or paired-end sequencing, one can obtain reads between 30 and 400 bp. The alignment of these sequenced cDNA fragments to the genome results in a genome-wide quantitative measure at the single-nucleotide level for the amount of

transcript present.

Advantages of RNA-Seq include single-base precision, no need for previous knowledge about the genomic sequence, low background signal, less RNA sample, a larger possible range of expression, high reproducibility, and lower cost[77].

In the available RNA-Seq data, the 16 human tissues profiled were adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cell.

4.2.2 Conservation

Many different conservation scores have been generated with varying methods and are publicly available. Existing examples are phastCons[69], GERP[13], PhyloCSF[43], and SiPhy[23]. Because I was looking for conserved elements in 3' and 5' UTR regions, rather than coding regions or single-nucleotide resolution scores, I chose to use SiPhy[23] with analysis of 12-mers.

SiPhy[23], or SIte-specific PHYlogenetic analysis, uses a probabilistic model for aligned sequences to describe molecular evolution by taking advantage of deeply sequenced clades and biased substitution patterns. For our purposes, I used human (hg18) elements found with SiPhy using 12mers across 30 mammals as generated for the 29 mammals paper[44] currently in review. Furthermore, since I am looking specifically in 3' UTRs and 5' UTRs regions, I only kept conserved elements that overlapped with 3' UTRs and 5' UTRs, and we distinguished between elements for each UTR, as they have functional differences.

4.2.3 Aggregate plots

An aggregate plot takes a number of regions across the genomes and aggregates, or stacks, the information at each region on top of one another. This can be accomplished by taking the arithmetic or geometric mean across the regions, dependent upon the type of signal you are looking for and the expected distribution of signal across the regions.

One of the main decisions that must be made for aggregate plots is how to align the regions so that they are comparable. They can either be centered at comparable relative locations, or they can be scaled to match each other in size. However, care must be taken so that artifacts are not created in the plots through selection of the regions or alignment.

Aligning the conserved elements by their end points or start points shows a clear distinction between the relative positions that are conserved or unconserved, correcting for strand orientation. However, the use of all the found elements could result in an artificial signal due to a large discrepancy between the number of elements available at varying distances from the alignment point. Therefore, I chose to use varying window sizes for our plots, filtering out any elements that were shorter than the half of the window size. Therefore, the end result is aggregate plots from a range of relative positions of $\frac{-window}{2}$ to $\frac{window}{2}$, where the relative position of 0 represents the alignment point, and negative positions are upstream of the alignment point, while positive positions are downstream. The filtering of conserved elements ensures that across the entire window, the same number of conserved elements should be aggregated, except for bases where RNA-Seq reads are not available, due to factors such as repetitive sequences.

For the aggregation of the data over many regions, the use of an geometric or arithmetic mean could be appropriate. While the arithmetic mean is a more straightforward average that assumes every read is weighted equally, regardless of which conserved element it is associated with, the geometric mean will minimize the weight of reads that come from elements with many reads associated with; this could help avoid dominance of the aggregate plot by a few elements that were sequenced at abnormally high levels.

In summary, the parameters that must be determined for each aggregate plot are 1) the tissue type the RNA-Seq data came from; 2) the selection of conserved regions overlapping either the 3' UTR or the 5' UTR; 3) the window size; 4) the use of a geometric or arithmetic mean; and 5) the alignment point of the start or end of the conserved region.

4.3 Results

By comparing across the different parameters, patterns and signatures appear in the aggregate plots. A small subset of plots is presented below in Figure 4-1 for brief comparison. On the whole, varying window sizes and cell types tended to result in consistent plots, given that the other parameters were held constant. However, the alignment point and chosen UTR made a significant difference in the shape of the plot.

4.3.1 Window sizes

Investigating plots with varying window sizes seem to indicate that the appropriate window size varies dependent on the other parameters. For example, in the 5' UTR conserved regions aligned by the end points, the signal is relatively consistent across window sizes, but it is certainly strongest and clearest with small window sizes, as shown below in Figure 4-2. The zoomed-out window and noise makes it difficult to interpret the plot with a window of 800bp, while the window of 50bp makes the dip at the alignment point obvious.

However, in another representative example, with 3' UTR conserved regions aligned by end points, as shown in Figure 4-3, the signals seem to significantly vary across different window sizes. The small window sizes seem to merely indicate a downward trend, while the larger window sizes seem to show a dip. One possible explanation for that is that the type of binding happening in the 3' UTRs is occurring with larger proteins or complexes that result in broader dips, while 5' UTRs, as shown in 4-2, are being affected by smaller regions of binding. However, this is probably not the case, as 5' UTRs have a median size of about 150 nt[45]. Therefore, due to this fact and the variation across window size, it is unclear whether the signal is biological and simply only shows up with certain computational measurements, or if it is merely an artifact of the methods. For these scenarios, it would be very useful to integrate our findings with experimental validation to make that important distinction.

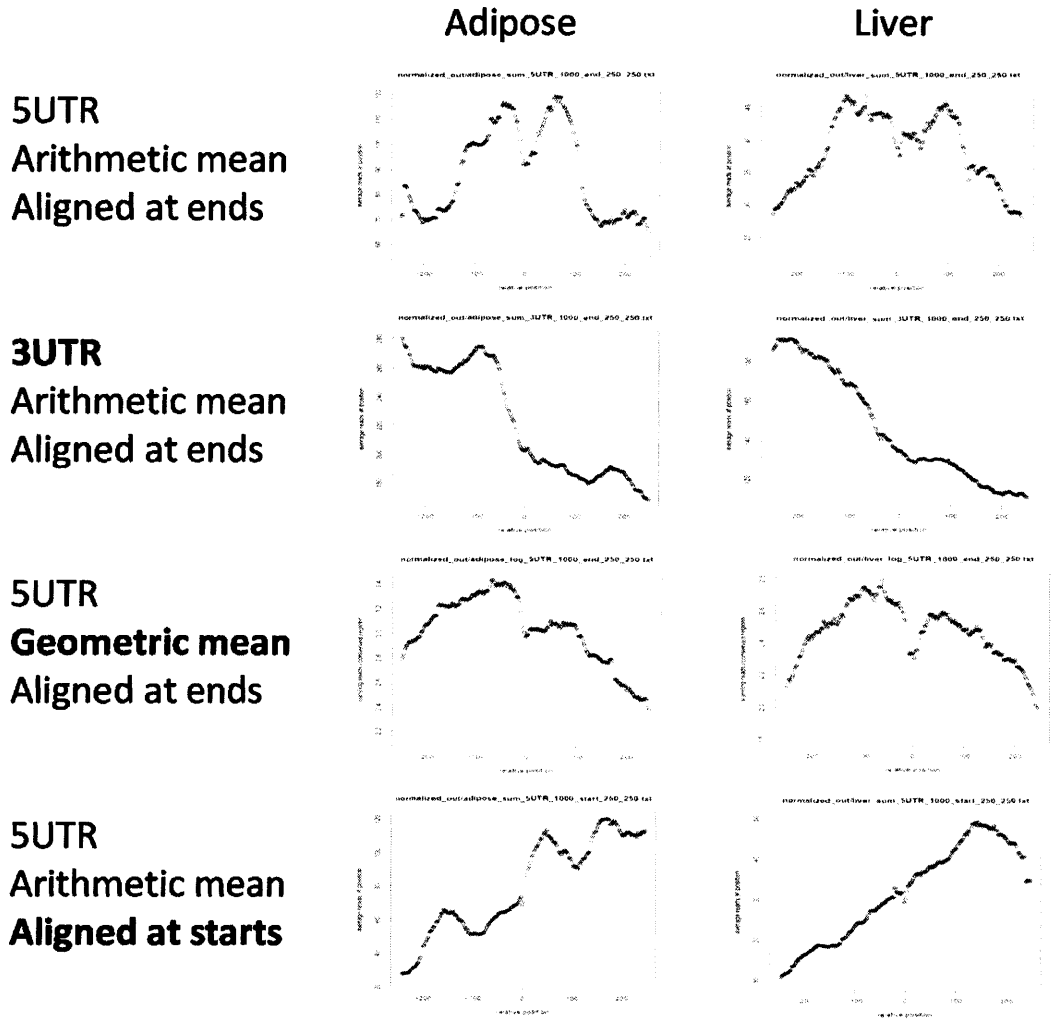


Figure 4-1: RNA-Seq aggregate plots at a 500 bp window with varying parameters. RNA-Seq data measures the amount of transcribed RNA present across the genome. These plots allow for the identification of signatures of human RNA-Seq data for functional, protein-bound RNA regions based on aggregate plots aligned for conserved regions of 3 UTRs and 5 UTRs. By comparing across different tissues, average types, alignment points, and UTR regions, I found specificities for different conditions.

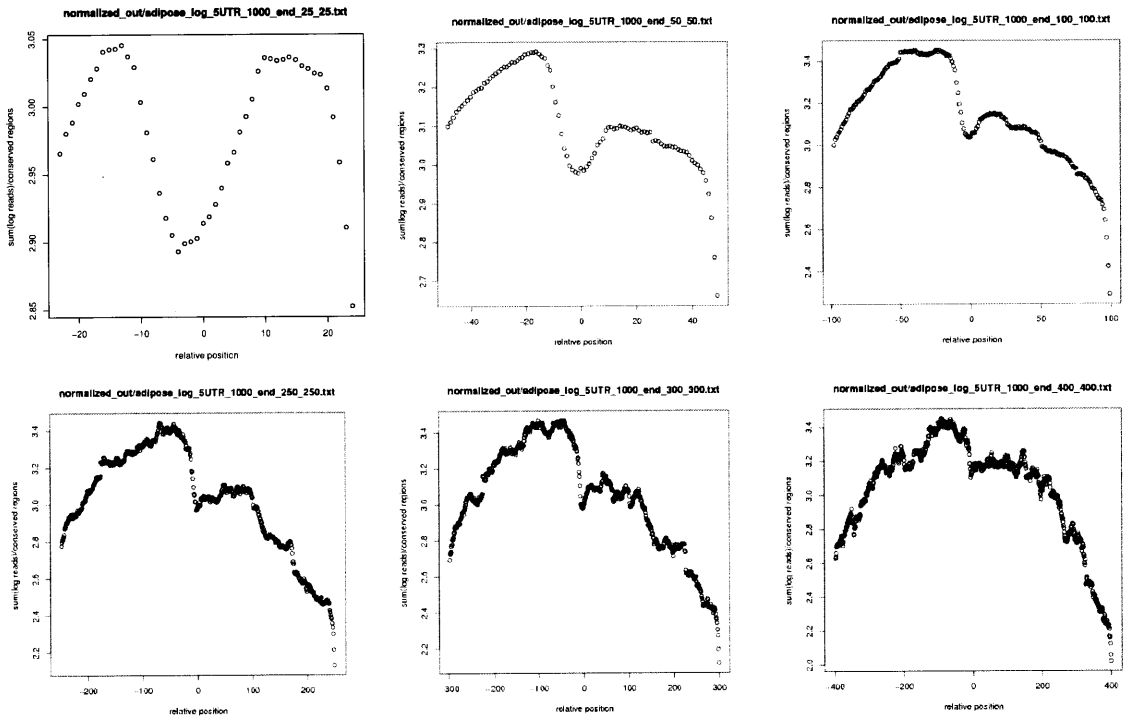


Figure 4-2: RNA-Seq aggregate plots with varying window sizes, for adipose tissue, using a geometric mean with conserved elements overlapping 5' UTRs. Window sizes represented are 50bp, 100bp, 500bp, 600bp, and 800bp, from left to right, top to bottom.

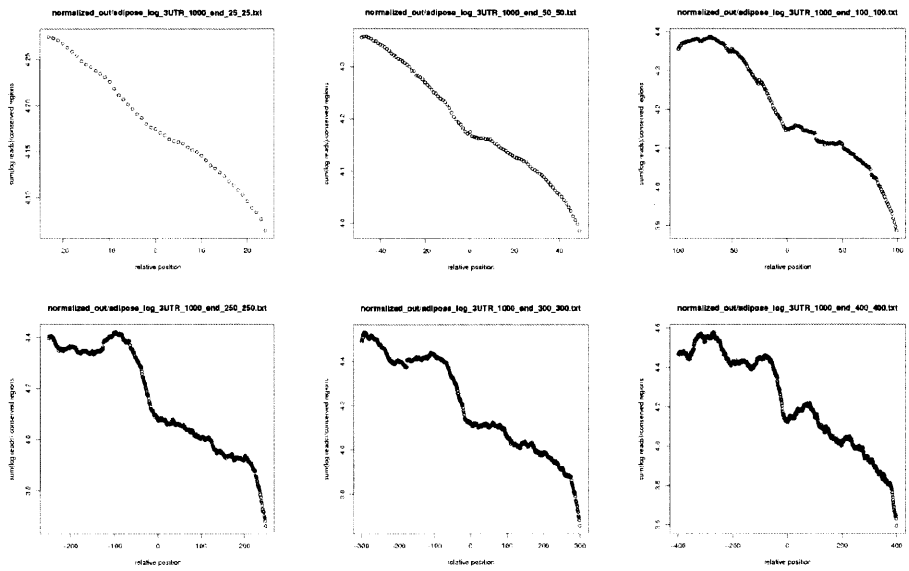


Figure 4-3: RNA-Seq aggregate plots with varying window sizes, for adipose tissue, using a geometric mean with conserved elements overlapping 3' UTRs. Window sizes represented are 50bp, 100bp, 500bp, 600bp, and 800bp, from left to right, top to bottom.

4.3.2 Cell types

Generally, the signatures between different tissue types seem quite similar, as can be seen below in Figure 4-4 this is not surprising, as the conserved regions identified were not tissue-specific, so functional regions for specific tissues most likely either would not be included or would be overshadowed by the general conservation signal. Interestingly, however, it can be noted that, though the shape of the plot is consistent, the absolute values generated for the mean fall in differing ranges, suggesting that tissue-specific variances in level of transcription or sensitivity for RNA-Seq data may exist.

4.3.3 Alignment point

Somewhat surprisingly, the point of alignment made a big difference in the shapes of the aggregate plots. One obvious trend is that aligning at start points results in a generally positive slope in the aggregate plot, while aligning at end points results in

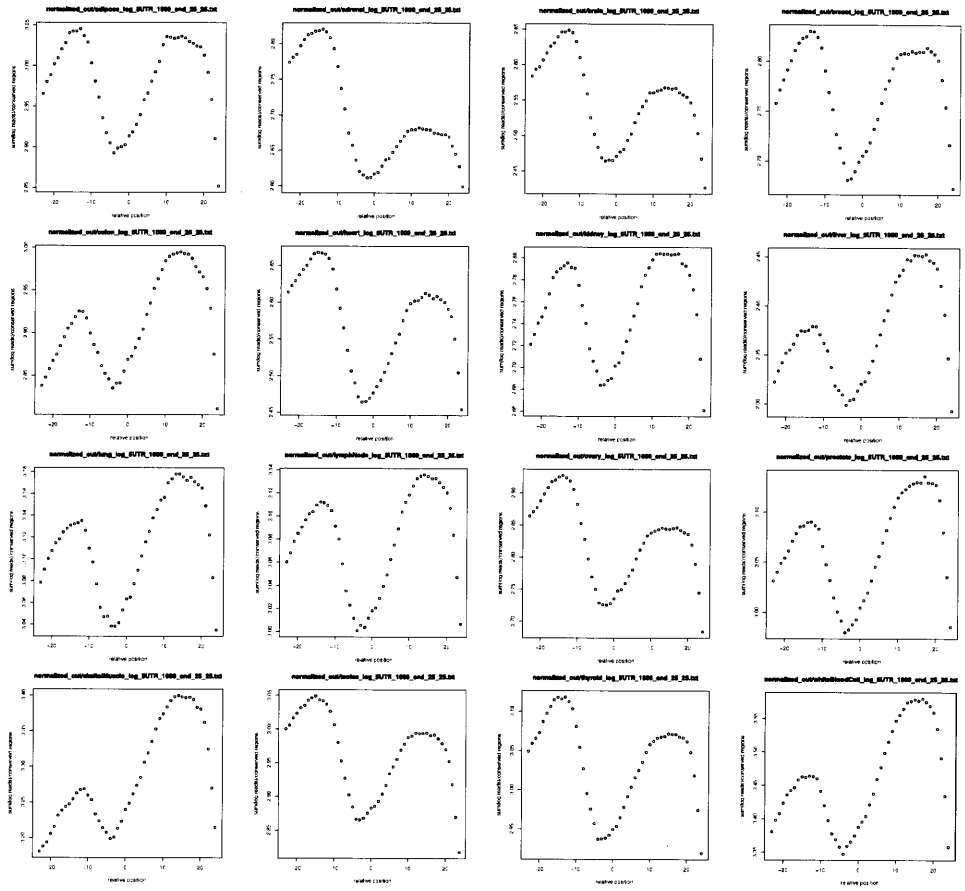


Figure 4-4: RNA-Seq aggregate plots with varying cell types. Cell types represented are adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cell, from left to right, top to bottom. The plots were generated with a geometric mean for conserved elements overlapping 5' UTRs.

a generally negative slope, as shown in Figure 4-5; this indicates that in general, the conserved parts of the plot have a higher amount of sequencing, since the alignment at start points result in right half of the plot falling in the conserved region, while alignment at end points results in the left half of the plot falling in the conserved region. However, it is notable that this trend seems to be much stronger with the 3' UTR aggregate plots aligned at the end point and 5' UTR aggregate plots aligned at the start points than with the 5' UTR-end and 3' UTR-start plots.

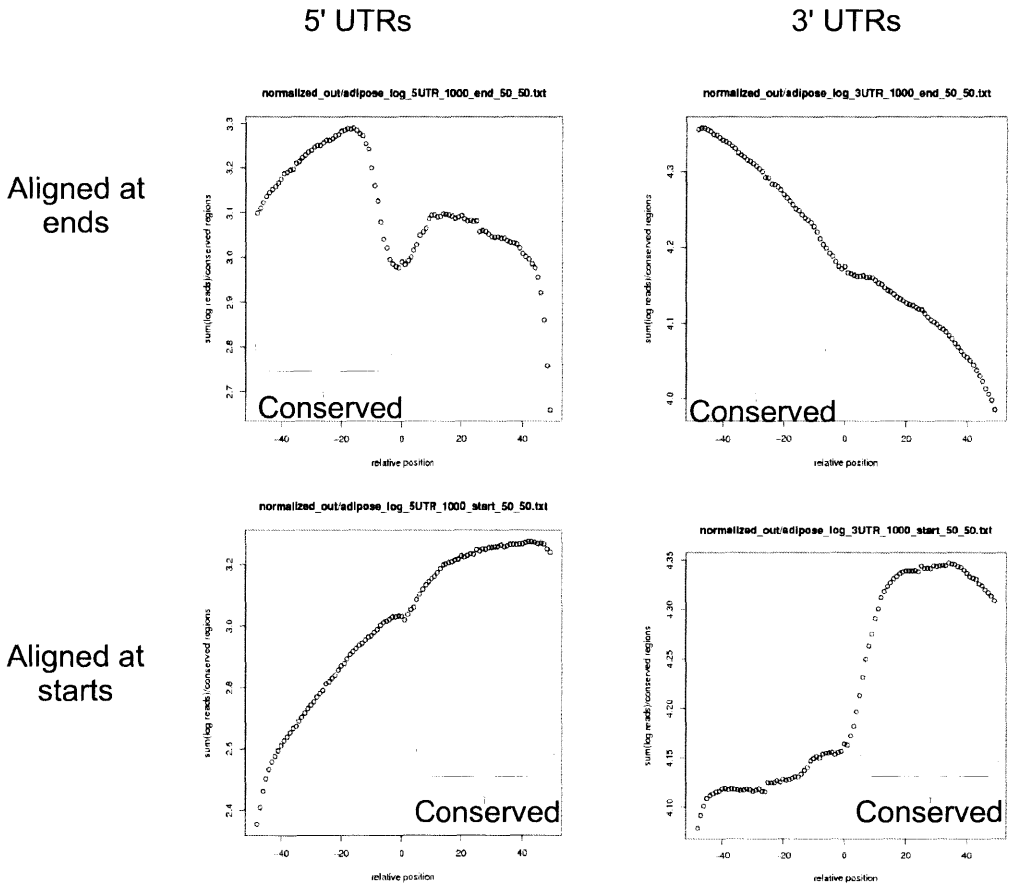


Figure 4-5: RNA-Seq aggregate plots with both start and end alignment points, for both 3' and 5' UTRs. The data shown is for adipose tissue, using a geometric mean with conserved elements.

4.4 Future work

4.4.1 *De novo* discovery of functional regions

In the future, I plan to leverage the signatures found here in end alignments of 3' UTRs and start alignments of 5' UTRs to computationally detect similar signatures in the RNA-Seq across the genome as potential regions for functional binding. I will develop statistics to quantify how closely a window of RNA-Seq data matches the signature, and use these statistics to identify novel functional regions.

4.4.2 Validation with RIP-Seq data

Integration with experimental RIP-Seq data will improve the power of the signature identification, as well as provide validation of our findings. The RIP-Seq protocol has been successfully developed and tested, so the resulting data should be available soon. Hopefully, the RIP-Seq data will immediately validate our current signatures for protein-bound regions, as I can simply substitute our conserved regions for enriched regions from the RIP-Seq data. If the signatures for regions in the RIP-Seq data diverge from the signatures of the conserved regions, we can further investigate what varying signals each might target. Furthermore, we can use the experimentally identified regions to validate the accuracy of regions identified through *de novo* discovery of functional regions.

4.4.3 Identification of bias in RNA-Seq method

Preliminary findings suggest that RNA-Seq has a bias for lower levels of transcription for functional RNA regions, most likely due to inaccessibility as a result of protein binding. In the future, I can validate this both with RIP-Seq, as described above, and with transcriptome data from micro-array experiments. The comparison of transcriptome data from RNA-Seq with targeted microarray data will also determine if the signal is a result of biological changes in transcription levels (if the levels of transcription are similar) or if it is a protocol-specific bias (if we only see the signature in

RNA-Seq data).

4.5 Contributions

In this project, I compared the transcription of different types of regions and identified a signature of conserved functional regions. Using this signature, I will be able to perform *de novo* discovery of functional regions and identify whether there is a bias in the RNA-Seq method. These insights increase our understanding of genome-wide transcription, and the identification of specific novel functional regions improve the annotation of the genome, influencing the development and conclusions of future studies. The identification of any protocol-specific bias will also influence future studies that use RNA-Seq data, as they can leverage the conclusions to correct or target data.

Chapter 5

Conclusion

5.1 Contributions

The research presented here has contributed to our understanding of gene regulation and development in the following ways:

- the unveiling of the mechanism of silencing that serves as a foundation for a crucial process of olfactory neuron specification,
- the discovery of an unusual form of tissue-specific heterochromatic silencing associated with histone marks H3K9me3 and H4K20me3,
- the identification of universal periodicity features and function-specific k-mers important in nucleosome positioning,
- the improvement of supervised prediction of nucleosome positioning through the subdivision of functional regions,
- a comparison of signatures for RNA regions of post-transcriptional modifications,
- the platform for *de novo* prediction of functional RNA regions,
- and a potential RNA-Seq specific experimental bias.

5.2 Further work

Further research that builds on the work presented here will be very valuable. For the nucleosome positioning project, the quantification of the presence of function-specific k-mers would give insight into known and novel motifs for pre-transcriptional modifications. Current work regarding the olfactory receptor gene regulation project is already being performed to determine the controlling factors that give rise to the abrupt and strategically located borders of the heterochromatin, such as motifs, nucleosome positioning, binding sites for the transcriptional repressor CTCF, counteracting chromatin modifications, or other insulator elements. Furthermore, the Lomvardas lab is also looking for motifs in the coding regions of the olfactory receptor genes, as that may be the nucleation site for the heterochromatin. Lastly, for the signature of functional RNA regions, the integration of RIP-Seq data will play a crucial role in validating our findings; these signatures will then be used for the discovery of novel functional RNA regions.

5.3 Conclusion

The work presented in this thesis studies the regulation of gene expression in different ways. We studied three organisms with varying degrees of complexity, with each project pinpointing different stages and mechanisms of gene regulation and utilizing different experimental and computational techniques. Through these studies, we made both general and specific discoveries that apply experimentally and computationally. While there is still much to be learned about epigenomics and gene regulation, we are hopeful that, with the progress we have made in these studies, as well as ever-improving experimental protocols and computational approaches, our understanding of these complex systems will continue to grow in both depth and quantity, ultimately improving medical techniques and the quality of human life.

Bibliography

[1] G. Barnea, S. O'Donnell, F. Mancia, X. Sun, A. Nemes, M. Mendelsohn, and R. Axel. Odorant receptors on axon termini in the brain. *Science*, 304:1468, Jun 2004.

[2] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233, Jan 2009.

[3] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125:315–326, Apr 2006.

[4] P. Bertone, M. Gerstein, and M. Snyder. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res.*, 13:259–274, 2005.

[5] M. S. Boguski, C. M. Tolstoshev, and D. E. Bassett. Gene discovery in dbEST. *Science*, 265:1993–1994, Sep 1994.

[6] L. Buck and R. Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65:175–187, Apr 1991.

[7] S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston. Unlocking the secrets of the genome. *Nature*, 459:927–930, Jun 2009.

[8] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[9] X. Chen, H. Fang, and J. E. Schwob. Multipotency of purified, transplanted globose basal cells in olfactory epithelium. *J. Comp. Neurol.*, 469:457–474, Feb 2004.

[10] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308:1149–1154, May 2005.

[11] J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer, and D. Botstein. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387:67–73, May 1997.

[12] A. Chess, I. Simon, H. Cedar, and R. Axel. Allelic inactivation regulates olfactory receptor gene expression. *Cell*, 78:823–834, Sep 1994.

[13] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, 15:901–913, Jul 2005.

[14] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5320–5325, Apr 2006.

[15] M.J.L. de Hoon, S. Imoto, J. Nolan, and S.Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[16] J. Dickson, H. Gowher, R. Strogantsev, M. Gaszner, A. Hair, G. Felsenfeld, and A. G. West. VEZF1 elements mediate protection from DNA methylation. *PLoS Genet.*, 6:e1000804, Jan 2010.

[17] C. D. Duggan and J. Ngai. Scent of a stem cell. *Nat. Neurosci.*, 10:673–674, Jun 2007.

[18] C. Dulac and R. Axel. A novel family of genes encoding putative pheromone receptors in mammals. *Cell*, 83:195–206, Oct 1995.

[19] E. Ezhkova, H. A. Pasolli, J. S. Parker, N. Stokes, I. H. Su, G. Hannon, A. Tarakhovsky, and E. Fuchs. Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell*, 136:1122–1135, Mar 2009.

[20] P. Feinstein, T. Bozza, I. Rodriguez, A. Vassalli, and P. Mombaerts. Axon guidance of mouse olfactory sensory neurons by odorant receptors and the beta2 adrenergic receptor. *Cell*, 117:833–846, Jun 2004.

[21] B. D. Fodor, N. Shukeir, G. Reuter, and T. Jenuwein. Mammalian Su(var) genes in chromatin control. *Annu. Rev. Cell Dev. Biol.*, 26:471–501, Nov 2010.

[22] S. H. Fuss, M. Omura, and P. Mombaerts. Local and cis effects of the H element on expression of odorant receptor genes in mouse. *Cell*, 130:373–384, Jul 2007.

[23] M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25:54–62, Jun 2009.

[24] D. S. Gerhard, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brown-stein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei, S. Mathavan, P. H. Gunaratne, J. Wu, A. M. Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G. Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morin, R. Morrin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak, and J. Malek. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, 14:2121–2127, Oct 2004.

[25] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130:77–88, Jul 2007.

[26] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227, Mar 2009.

[27] R. D. Hawkins, G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker, and B. Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, 6:479–491, May 2010.

[28] N. Heintz. Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.*, 7:483, May 2004.

[29] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifica-

tions at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108–112, May 2009.

[30] K. E. Van Holde. *Chromatin*. Springer-Verlag, 1988.

[31] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104, 2002.

[32] T. Hung and H. Y. Chang. Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol*, 7:582–585, 2010.

[33] C. Jiang and B. F. Pugh. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol.*, 10:R109, 2009.

[34] W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. U.S.A.*, 103:12457–12462, Aug 2006.

[35] T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128:1231–1245, Mar 2007.

[36] T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465:182–187, May 2010.

[37] R. D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98:285–294, Aug 1999.

[38] J. L. Larson and G. C. Yuan. Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. *BMC Bioinformatics*, 11:557, 2010.

[39] C. T. Leung, P. A. Coulombe, and R. R. Reed. Contribution of olfactory neural stem cells to tissue maintenance and regeneration. *Nat. Neurosci.*, 10:720–726, Jun 2007.

[40] V. G. Levitsky, O. A. Podkolodnaya, N. A. Kolchanov, and N. L. Podkolodny. Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics*, 17:998–1010, Nov 2001.

[41] J. W. Lewcock and R. R. Reed. A feedback mechanism regulates monoallelic odorant receptor expression. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1069–1074, Jan 2004.

[42] S. D. Liberles, L. F. Horowitz, D. Kuang, J. J. Contos, K. L. Wilson, J. Siltberg-Liberles, D. A. Liberles, and L. B. Buck. Formyl peptide receptors are candidate chemosensory receptors in the vomeronasal organ. *Proc. Natl. Acad. Sci. U.S.A.*, 106:9842–9847, Jun 2009.

[43] M. Lin, I. Jungreis, and Manolis. Kellis. Phylocsf: a comparative genomics method to distinguish protein-coding and non-coding regions., 2010. Available from Nature Precedings.

[44] K. Lindblad-Toh, M. Garber, O. Zuk, M.F. Lin, B.J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L.D. Ward, C.B. Lowe, A.K. Holloway, M. Clamp, S. Gnerre, J. Alfoldi, K. Beal, J. Chang, H. Clawson, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M.J. Hubisz, D.B. Jaffe, I. Jungreis, D. Kostka, M. Lara, A.L. Martins, T. Massingham, I. Moltke, B.J. Raney, M.D. Rasmussen, A. Stark, A.J. Vilella, J. Wen, X. Xie, M.C. Zody, Broad Institute Sequencing Platform, K.C. Whole Genome Assembly Team, Worley, C.L. Kovar, D.M. Muzny, R.A. Gibbs, W.C. Baylor College of Medicine Human Genome Sequencing Center, Warren, E.R. Mardis, Weinstock G.M., R.K. Wilson, E. Washington University Genome Center, Birney, E.H. Margulies, J. Herrero, E.D. Green, D. Haussler, A. Siepel, N. Goldman, K.S. Pollard, J.S. Pedersen, E.S. Lander, and M. Kellis. A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals. In review.

[45] H. Lodish, A. Berk, C. Kaiser, M. Krieger, M. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology.* W.H.Freeman, 6th edition.

[46] S. Lomvardas, G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland, and R. Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126:403–413, Jul 2006.

[47] J. C. McIntyre, S. C. Bose, A. J. Stromberg, and T. S. McClintock. Emx2 stimulates odorant receptor gene expression. *Chem. Senses*, 33:825–837, Nov 2008.

[48] P. Mombaerts, F. Wang, C. Dulac, S. K. Chao, A. Nemes, M. Mendelsohn, J. Edmondson, and R. Axel. Visualizing an olfactory sensory map. *Cell*, 87:675–686, Nov 1996.

[49] M. Q. Nguyen, Z. Zhou, C. A. Marks, N. J. Ryba, and L. Belluscio. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell*, 131:1009–1017, Nov 2007.

[50] John W. Nicol, Gregg A. Helt, Steven G. Blanchard, Archana Raja, and Ann E. Loraine. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, 2009.

[51] H. Nishizumi, K. Kumasaka, N. Inoue, A. Nakashima, and H. Sakano. Deletion of the core-H region in mice abolishes the expression of three proximal odorant receptor genes in cis. *Proc. Natl. Acad. Sci. U.S.A.*, 104:20067–20072, Dec 2007.

[52] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Res.*, 17:1170–1177, Aug 2007.

[53] M. Pyrski, Z. Xu, E. Walters, D. J. Gilbert, N. A. Jenkins, N. G. Copeland, and F. L. Margolis. The OMP-lacZ transgene mimics the unusual expression pattern of OR-Z6, a new odorant receptor gene on mouse chromosome 6: implication for locus-dependent gene expression. *J. Neurosci.*, 21:4637–4648, Jul 2001.

[54] K. Regha, M. A. Sloane, R. Huang, F. M. Pauler, K. E. Warczok, B. Melikant, M. Radolf, J. H. Martens, G. Schotta, T. Jenuwein, and D. P. Barlow. Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. *Mol. Cell*, 27:353–366, Aug 2007.

[55] K. J. Ressler, S. L. Sullivan, and L. B. Buck. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell*, 73:597–609, May 1993.

[56] K. J. Ressler, S. L. Sullivan, and L. B. Buck. Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. *Cell*, 79:1245–1255, Dec 1994.

[57] S. Riviere, L. Challet, D. Fluegge, M. Spehr, and I. Rodriguez. Formyl peptide receptor-like proteins are a novel family of vomeronasal chemosensors. *Nature*, 459:574–577, May 2009.

[58] D. J. Rodriguez-Gil, H. B. Treloar, X. Zhang, A. M. Miller, A. Two, C. Iwema, S. J. Firestein, and C. A. Greer. Chromosomal location-dependent nonstochastic onset of odor receptor expression. *J. Neurosci.*, 30:10067–10075, Jul 2010.

[59] A. Rothman, P. Feinstein, J. Hirota, and P. Mombaerts. The promoter of the mouse odorant receptor gene M71. *Mol. Cell. Neurosci.*, 28:535–546, Mar 2005.

[60] G. R. Schnitzler. Control of nucleosome positions by DNA sequence and remodeling machines. *Cell Biochem. Biophys.*, 51:67–80, 2008.

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

[62] G. Schotta, M. Lachner, K. Sarma, A. Ebert, R. Sengupta, G. Reuter, D. Reinberg, and T. Jenuwein. A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.*, 18:1251–1262, Jun 2004.

[63] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, Aug 2006.

[64] M. R. Segal. Re-cracking the nucleosome positioning code. *Stat Appl Genet Mol Biol*, 7:Article14, 2008.

[65] S. Serizawa, K. Miyamichi, H. Nakatani, M. Suzuki, M. Saito, Y. Yoshihara, and H. Sakano. Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. *Science*, 302:2088–2094, Dec 2003.

[66] S. Serizawa, K. Miyamichi, H. Takeuchi, Y. Yamagishi, M. Suzuki, and H. Sakano. A neuronal identity code for the odorant receptor-specific and activity-dependent axon sorting. *Cell*, 127:1057–1069, Dec 2006.

[67] B. M. Shykind. Regulation of odorant receptors: one allele at a time. *Hum. Mol. Genet.*, 14 Spec No 1:R33–39, Apr 2005.

[68] B. M. Shykind, S. C. Rohani, S. O'Donnell, A. Nemes, M. Mendelsohn, Y. Sun, R. Axel, and G. Barnea. Gene switching and the stability of odorant receptor gene choice. *Cell*, 117:801–815, Jun 2004.

[69] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15:1034–1050, Aug 2005.

[70] Jun Song, W. Evan Johnson, Xiaopeng Zhu, Xinmin Zhang, Wei Li, Arjun Manrai, Jun Liu, Runsheng Chen, and X. Shirley Liu. Model-based analysis of two-color arrays (MA2C). *Genome Biology*, 8(8):R178+, August 2007.

[71] Tobias Straub. Basic analysis of nimblegen chip-on-chip data using bioconductor/r, Apr 2009.

[72] P. Trojer and D. Reinberg. Facultative heterochromatin: is there a distinctive molecular signature? *Mol. Cell*, 28:1–13, Oct 2007.

[73] R. Vassar, S. K. Chao, R. Sitcheran, J. M. Nunez, L. B. Vosshall, and R. Axel. Topographic organization of sensory projections to the olfactory bulb. *Cell*, 79:981–991, Dec 1994.

[74] R. Vassar, J. Ngai, and R. Axel. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell*, 74:309–318, Jul 1993.

[75] E. Walters, M. Grillo, A. B. Oestreicher, and F. L. Margolis. LacZ and OMP are co-expressed during ontogeny and regeneration in olfactory receptor neurons of OMP promoter-lacZ transgenic mice. *Int. J. Dev. Neurosci.*, 14:813–822, Nov 1996.

[76] F. Wang, A. Nemes, M. Mendelsohn, and R. Axel. Odorant receptors govern the formation of a precise topographic map. *Cell*, 93:47–60, Apr 1998.

[77] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63, Jan 2009.

[78] T. Warnecke, N. N. Batada, and L. D. Hurst. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.*, 4:e1000250, Nov 2008.

[79] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A. Irizarry, and Andrew P. Feinberg. Large histone h3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet*, 41(2):246–250, Feb 2009.

[80] Ernst Wit and John D. McClure. *Statistics for microarrays: design, analysis and inference.* John Wiley & Sons.

[81] Q. Wu, J. Wang, and H. Yan. Prediction of nucleosome positions in the yeast genome based on matched mirror position filtering. *Bioinformation*, 3:454–459, 2009.

[82] J. J. Wyrick, F. C. Holstege, E. G. Jennings, H. C. Causton, D. Shore, M. Grunstein, E. S. Lander, and R. A. Young. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*, 402:418–421, Nov 1999.

[83] K. Yamada, J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, and J. R. Ecker. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, 302:842–846, Oct 2003.

[84] G. C. Yuan and J. S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, 4:e13, Jan 2008.