BIOINFORMATIC METHODS FOR ANALYSES OF EPIGENOME AND RNA INTERACTOME

BY

PENGFEI YU

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biophysics and Computational Biology in the Graduate College of the University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Sheng Zhong, Chair
Professor Lisa Stubbs
Associate Professor Saurabh Sinha
Assistant Professor Jian Ma
ABSTRACT

The sequenced genetic codes for multiple species have provided great sources for understanding how a single genome can give rise to a complex organism. However, only a small percent of mammalian genome could be identified as protein-coding genes. Our understanding of the genome is far from complete, particularly on non-coding RNAs and regulatory sequences. Epigenomic modifications and variations were considered to contribute to the diversity of functions found across different cell types, play key roles in the establishment and maintenance of cellular identity during development. Here, we propose to annotate functional regulatory sequences based on combinatorial patterns of epigenomic marks. We will classify these patterns in two directions, through conservation information among mammalian species and through a mouse differentiation process. We are also trying to explore new functions of non-coding RNAs through analysis of genome-wide RNA-RNA interactions.

The first part of my thesis is focused on the cross-species direction. We provide a comparative approach to compare epigenomic patterns in pluripotent stem cells of three different mammalian species and find that certain combinations of epigenomic modifications tend to be co-localized. These co-localizations are also more likely to be present in conserved regions. Moreover, our results suggest that these conserved co-localization patterns could help to define strong regulatory elements within the genome. The functions of these locations are tested through a guided differentiation system.

The second part is about the other direction. With time-course epigenomic data from the guided differentiation system, genomic sequences are clustered based on the spatiotemporal epigenomic information. These analyses provide us better understanding how the single genome is dynamically regulated to ensure a diversity of cell types. A two-layer hierarchical model is presented and by applying this model, context-specific functions of some epigenomic
modifications are uncovered. Meanwhile, a genome-wide view of regulatory sequence locations as well as their activation time is provided to help clarify the transcription network.

The last part of my thesis is about a data analysis tool set and analytical results for a novel next generation sequencing based technology to identify RNA-RNA interactions. The technology allows simultaneous and unbiased identification of different types of RNA-RNA interactions within the cells. Our computational tools are able to evaluate the quality of sequencing data generated from this technology, identify strong interactions starting from raw sequencing reads and compare similarities of results from different samples. We provide two different types of visualizations for identified interactions. Novel regulatory functions of some small RNAs are also discovered in mouse ES cells. The interactions within the same RNA molecule can also provide useful information for in vivo RNA structures.
To my family, for their constant love and support.
It would have never been possible to finish my Ph.D dissertation without the guidance of my committee members, help from friends, and support from my family. To only some of them, it is possible to give particular mention here.

My first debt of gratitude must go to my advisor, Dr. Sheng Zhong. I want to thank Sheng for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. He was always patient to help preparing my future career by teaching me how to write scientific papers, by mentoring me how to present my works, and by providing me a pleasant environment to interact with our collaborators. Also, his positive attitude and determination encouraged me whenever I had a hard time for my research.

Special thanks to the members of my thesis defense committee: Dr. Saurabh Sinha, Dr. Jian Ma and Dr. Lisa Stubbs. All of them provided valuable comments and helpful suggestions during my Prelim exam and within this thesis. Besides that, Saurabh and Jian also generously offered me kind guidance in both area of computation and biology during biophysics tutorial studies. I would also like to thank Dr. Jongsook Kim Kemper and Dr. Alison Bell for the pleasant experiences to collaborate with experimental scientists in their groups.

This thesis also owes much to many colleagues and friends at UIUC. I had wonderful experience with all of my lab mates. Dr. Dan Xie, Dr. Xiaoyi Cao and Dr. Chieh-Chun Chen helped me a lot to be comfortable on Bioinformatics in the early days of my PhD studies, Dr. Wei Huang gave me valuable advices on statistical modeling, and Dr. Shu Xiao patiently mentored me on molecular and cellular biology skills. Besides research, I also appreciate their helps in many aspects of my personal life including career development. It was also my great pleasure to work with Marcelo Rivas-Astroza, Tri Nguyen,
Bharat Sridhar, and all other Zhong lab members. I would like to specially thank one of my best friends, Bo Liu. We've had a great time since the first day I came to Champaign. He was always supportive both for my researches and personal life. I also want to thank many of my other friends whom I met in Champaign and who shared a memorable life with me.

Finally, I wish to thank my parents, Yunhai Yu and Xingjuan Wang and my girlfriend Zhongyi (Lucy) Li for their unequivocal support. Their love provided me inspiration and was my driving force, for which my mere expression of thanks likewise does not suffice.
Chapter 1. Portions of the introductory text are modified from previously written introductory material from my Prelim proposal.

Chapter 2. A version of the materials has been published as S. Xiao, D. Xie, X. Cao, P. Yu, X. Xing, C. C. Chen, M. Musselman, M. Xie, F. D. West, H. A. Lewin, T. Wang, and S. Zhong (2012). The ChIP-seq, RNA-seq experiments were conducted by Dr. Shu Xiao. Xingyun Xing and Mingchao Xie in Dr. Wang Ting’s group (Washington University, Sr. Louis, USA) helped to generate and preprocess data for MRE-seq and MeDIP-seq. Pig iPS cell lines were provided by Dr. Franklin West’s group (University of Georgia, Athens, USA). Dr. Xiaoyi Cao generated Figure 2.3 and did corresponding analysis. I performed all the other analysis and generated other figures. Dr. Sheng Zhong wrote the manuscript.

Chapter 3. A version of these materials has been published as P. Yu, S. Xiao, X. Xin, C. X. Song, W. Huang, D. McDee, T. Tanaka, T. Wang, C. He, and S. Zhong (2013). The ChIP-seq, RNA-seq, smRNA-seq experiments were conducted by Dr. Shu Xiao. Xingyun Xing in Dr. Ting Wang’s group (Washington University, St. Louis, USA) helped to generate data for MRE-seq and MeDIP-seq. Chun-Xiao Song in Dr. Chuan He’s group (University of Chicago, Chicago, USA) helped to generate data for 5-hmC-seq. Dr. Wei Huang helped on the implementation of the GATE model. All the other analysis were performed by me. Dr. Sheng Zhong and I wrote the manuscript for the published paper.

Chapter 4. The experimental technology was designed and optimized by Cong Tri Nguyen. Figure 4.1, Table 4.1 and the experimental details were generated by Cong Tri Nguyen. The RNA Hi-C data for different samples were also generated by Cong Tri Nguyen. Dr. Xiaoyi Cao performed the network analysis and generated Figure 4.8B-C. Part of the conservation analysis was done by Jia Lu who generated Figure 4.12B. I performed all other com-
putational analysis and developed the computational tools for this new type of NGS data. I also helped to write the manuscript which was prepared for submission together with Dr. Sheng Zhong.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER 1  INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2  RELATIONSHIP BETWEEN EPIGENOME AND GENE REGULATION IN PLURIPOTENT STEM CELLS ACROSS SPECIES</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Epigenomic data in three species</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Methods</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Results for comparative epigenomic study</td>
<td>11</td>
</tr>
<tr>
<td>2.5 Discussion</td>
<td>20</td>
</tr>
<tr>
<td>CHAPTER 3  SPATIOTEMPORAL CLUSTERING OF EPIGENOMIC STATES</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.2 GATE model</td>
<td>25</td>
</tr>
<tr>
<td>3.3 Simulation analysis</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Application of GATE model on mESC guided differentiation</td>
<td>33</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>51</td>
</tr>
<tr>
<td>CHAPTER 4  RNA-HIC-TOOLS: A BIOINFORMATICS PIPELINE FOR ANALYSIS OF RNA HI-C DATA ON HIGH-THROUGHPUT MAPPING OF RNA-RNA INTERACTOME</td>
<td>54</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>54</td>
</tr>
<tr>
<td>4.2 Experimental design for RNA Hi-C technology</td>
<td>55</td>
</tr>
<tr>
<td>4.3 Computational methods of RNA-HiC-tools</td>
<td>57</td>
</tr>
<tr>
<td>4.4 Simulation study on synthetic paired-end RNA Hi-C data</td>
<td>71</td>
</tr>
<tr>
<td>4.5 Results on real data</td>
<td>74</td>
</tr>
<tr>
<td>4.6 Discussion</td>
<td>100</td>
</tr>
<tr>
<td>CHAPTER 5  CONCLUSION</td>
<td>105</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>109</td>
</tr>
</tbody>
</table>
LIST OF TABLES

2.1 Modeling fitting measured by $R^2$. .......................... 14

3.1 Parameters used in simulation for GATE model. ............... 32

4.1 Evaluation of fragment length prediction. ......................... 72
4.2 Evaluation of fragment class prediction. ......................... 72
4.3 Biological and technical differences of the four samples. .... 76
4.4 Evaluation of ligation efficiency. ............................... 77
4.5 Result summary of mapped pairs and strong interactions in four samples. ............................................. 79
4.6 Overlaps of strong inter-RNA interactions among different samples and replicates. ................................. 80
4.7 The distribution of interaction sites in various types of RNAs from mouse ES cells. ............................. 85
4.8 Other types of RNAs functioning as miRNAs ................... 91
4.9 The overlap between interacting RNA sites identified from RIP-seq and RNA Hi-C. .............................. 103
LIST OF FIGURES

2.1 The same epi-modifications are predicative of gene expression levels in all three species .................................................. 13
2.2 Correlations among evolutionary changes of epi-modification intensities, gene expression levels, TF binding intensities, and genomic sequences ........................................ 15
2.3 Interspecies conservation of co-occupancy of different epigenomic modifications. .................................................. 16
2.4 Epigenomic changes and gene expression changes during differentiation .................................................. 17
2.5 Poised promoters defined by H2A.Z and H3K4me2/3 in mouse ES cells .................................................. 19

3.1 Two types of correlations between the epigenome and gene expression .................................................. 23
3.2 The Genomic Annotation based on Time-course Epigenomic data (GATE) model .................................................. 26
3.3 An expectation-maximization algorithm for inferring model parameters .................................................. 29
3.4 Evaluation of model prediction on simulated dataset .................................................. 34
3.5 Differentiation of mouse ES cells into mesendoderm cells .................................................. 35
3.6 GATE predicted epigenomic clusters .................................................. 38
3.7 Dendrogram of epigenomic clusters .................................................. 39
3.8 Relative genomic locations of epigenomic clusters .................................................. 40
3.9 Predicting genomic features .................................................. 42
3.10 Distribution of miRNA promoters and piRNA genes in GATE clusters .................................................. 44
3.11 A model for combinatorial epigenomic changes and gene expression .................................................. 45
3.12 Temporal correlations between 5-hmC and other epigenomic marks .................................................. 48
3.13 Predicted mesendoderm enhancers harbor transcription factor binding sites .................................................. 50

4.1 RNA Hi-C technology to map RNA-RNA interactions .................................................. 56
4.2 Computational pipeline for analysis of RNA Hi-C data .................................................. 58
4.3 Workflow for reconstructing fragments. .......................... 61
4.4 A statistical test based Approach to identify strong RNA-RNA interactions. .................................................. 65
4.5 Evaluation of performance based on simulation studies. .... 74
4.6 Portions of different fragment types for four samples. ....... 76
4.7 Comparison of RNA Hi-C libraries. ............................... 78
4.8 An example for the interaction between snoRNA SNORA1 and 3’-UTR region of Trim25 gene. ......................... 80
4.9 Different types of inter-RNA interactions and Global visualization of RNA-RNA interactomes. .......................... 82
4.10 Interaction sites and their interaction partners on the transcribed region of Eef1a1 gene. ................................. 84
4.11 The hybridization energies between two RNA sequences that are interacting with each other supported by chimeric RNAs for different interaction types. .......................... 86
4.12 Examples of base-complement between two directly interacting RNAs for different interaction types. ............... 87
4.13 Increased conservation levels in interaction sites ............. 89
4.14 snoRNAs function as miRNAs to target UTR regions of mRNAs. ................................................................. 93
4.15 An example to show the interaction between processed SNORA14 and 3’ UTR of Mcl1. ................................. 94
4.16 Self ligations detected in RNA Hi-C. ............................. 96
4.17 Examples of self-ligations within same snoRNA molecules which provide information of RNA structures. ........... 98
4.18 A novel small RNA identified within the intronic region of Rpl23 gene. .......................................................... 100
The human genome project produced an almost complete order of the 3 billion pairs of chemical letters in the DNA that embodies the human genetic code, but little about the way how this blueprint works. Besides small portion of protein-coding sequences, the annotation for genomic sequences encoding non-coding RNAs (ncRNAs) or cis-regulatory sequences would be a more challenge task. The accomplishment in this field could definitely expand our knowledge on how human genome is precisely regulated to ensure a diversity of cell types created from a single genome.

In the eukaryotic genome, majority of DNA can mutate freely without deleterious effects, while certain sequence elements are more constrained which serves as evidence of structural or functional conservation [1]. Based on this theory, researchers have inferred functional genomic segments by examining genomic sequence conservation [2] and have identified human-specific regulatory DNA by looking for sequences with accelerated rates of evolutionary change [3]. Thus, Evolutionary comparisons on DNA sequences provide a powerful tool to interpret genome functions.

Since the genomic conservation has been proved a feasible approach to find functional elements within the genome, we start to ask whether the conservation of epigenome can help us determine its functions. To do so, the basic evolutionary properties of the epigenome must be established first, preferably in the contexts of both genomic and transcriptomic evolution. An epigenome consists of chemical modifications and protein variations to the DNA and histone, and some of these modifications and variations can be passed down to an organism’s offspring [4, 5], which supports the feasibility of this approach. The relationship between evolutionary changes to the genome and the transcriptome is weak. The degree of gene expression conservation cor-
relates poorly with the extent to which non-exonic sequences are conserved among vertebrates [6, 7]. Also, mammalian orthologous transcription factors (TF) often do not bind to orthologous DNA sequences [8], as only 5% of the OCT4 and NANOG binding sites occupy homologous sequences in human and mouse embryonic stem (ES) cells [9]. These discrepancies may be explained by the evolutionary changes to the epigenome.

In Chapter 2, we will extensively study the evolutionary changes of epigenome, genome and transcriptome. We will provide evidences to show how interspecies epigenomic changes influence the interspecies changes of transcriptome and transcription factor bindings. The conserved combinations of epigenomic modifications could also suggest a novel approach to identify functional elements within the genome.

It has been shown that epigenomic modification can influence transcription activities in different ways. Cytosine methylation generally marks packaged nucleosomes with limited TF accessibility or disrupt GC-rich transcription factor binding sites themselves [10]. Active promoters are usually marked by H3 lysine 4 di- or tri-methylation (H3K4me2/3), which may associate with transcriptional initiation. Enhancer regions are often marked with H3K4me1/2 [11, 12]. Repressive markers, H3K9me2/3 and H3K27me3, may block the access of TFs from DNA and are usually enriched in silenced regions [11]. While the exact mechanism is still not clear, histone variant H2A.Z is required for stem cell commitment [13]. Moreover, histone modifications were also been proved to be significantly different between pluripotent cells and committed ones [14], suggesting its vital role in the differentiation process.

Epigenomes are dynamic, and epigenetic modifications are associated with changes in gene expression [15, 16]. Thus, the epigenome adds an extra layer of information onto the genomic sequence and enables a genome to dynamically orchestrate gene expression in different cell types [17, 18]. Potentially, if we were presented with genome-wide distributions of epigenetic modifications at multiple time points during a developmental or differentiation process, we would be able to utilize them to find the genomic (cis-) regulatory sequences that regulate gene expression. Meanwhile, combinatorial functions of epigenetic modifications and regulatory sequences could also be identified.

The laboratory mouse is the most widely used mammalian model organ-
ism in biomedical research. The $2.6 \times 10^9$ bases of the mouse genome possess a high degree of conservation with the human genome, so a complete annotation of the mouse genome will be of significant value to understanding the function of the human genome. In chapter 2 and 3, we applied a differentiation assay in which mouse embryonic stem cells were differentiated into mesendoderm cells to investigate the functions of epigenetic modifications in a dynamic process. At 3 time points during this differentiation process, we mapped the genomic distributions of 9 epigenetic modifications, including DNA methylation (5-mC), hydroxymethylation (5-hmC), histone variant H2A.Z, and histone modifications H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3. At the same time points, we also assayed the expression of small non-coding RNAs (ncRNAs) and total RNAs.

With this invaluable data, in chapter 3, we will develop a statistical model to jointly combine the position effect and the temporal effect of the epigenome and capture combinatorial patterns of temporal epigenomic changes. This will help us achieve \textit{ab initio} identification and functional annotation of regulatory sequences.

The exploration of functional non-coding RNAs is as important as annotation of cis-regulatory elements for understanding of the huge mammalian genome. Non-coding RNAs are transcripts that are not translated into proteins but are also functionally or structurally important in the cellular processes. Some types of ncRNAs are well-known and fulfill central functions within the cells such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs). The functions of these ncRNAs have been studied for a long time. However, new discoveries over the last few years have shown that ncRNAs may have a much broader functional spectrum than we previously expected. They could play significant regulatory roles in complex organisms through interactions with other RNAs, chromatin, and proteins. For example, the discovery of miRNAs expanded our knowledge of how gene expressions are regulated [19]. Actually, it has been revealed that about 90% of the human genome is been transcribed at certain time within certain cell state [20]. The interpretation for the functional consequences of these abundant transcripts remains highly controversial [21].

Many evidences suggested that short and long ncRNAs served as versa-
tile regulators of cellular processes. For example, besides their well-known functions for the chemical modifications of rRNAs and tRNAs, some snoRNAs could be processed by Dicer into smaller RNAs, and these smaller RNAs could function as non-canonical miRNA by targeting 3'-UTR regions of other protein-coding genes [22, 23, 24]. Long ncRNAs (lincRNAs) also have diverse regulatory functions [25]. Some of them could mediate epigenetic modification by recruiting chromatin remodeling complex to specific chromatin loci and others could also be involved in post-transcriptional regulations of mRNAs. The variability for the functions of ncRNAs makes it more challenging for us to understand to full functional spectrum of ncRNAs.

Interactions between ncRNAs and other RNAs mediated by RNA binding proteins are the major means for ncRNAs to exhibit their regulatory functions. Recent years, with rapid development on high-throughput sequencing, many innovative technologies has been emerging for detection of RNA-RNA interactions such as PAR-CLIP [26], HITS-CLIP [27], and CLASH [28, 29].

Two major bottlenecks remain. First, one experiment can only analyzes the interactions mediated by one RNA-binding protein at a time. Second, each experiment requires either a protein-specific antibody (HITS-CLIP or PAR-CLIP) or stable expression of a tagged protein (CLASH). As a result, CLASH is only applicable to transformed cell lines. Furthermore, neither HITS-CLIP or PAR-CLIP provides physical evidence of RNA interactions, because any two RNAs co-appeared in HITS-CLIP or PAR-CLIP could have been attached to different protein copies of the same gene, such as different AGO copies, in the same cell or in different cells.

In chapter 4, we will introduce a novel method called “RNA Hi-C” to analyze all protein-assisted RNA-RNA interactions in vivo. The method offers several advantages and can tackle above bottlenecks for mapping RNA-RNA interactions. We will also introduce a bioinformatic toolkit (RNA-HiC-tools) to analyze and visualize RNA Hi-C data. RNA-HiC-tools is able to automate all the analysis steps. Through the analysis of RNA Hi-C data, we will study novel regulatory functions of some small RNAs in mouse ES cells. The interactions within the same RNA molecule can also provide information for in vivo RNA structures.

Finally, in chapter 5, we will give a brief summary and future direction of
this thesis work.
CHAPTER 2

RELATIONSHIP BETWEEN EPIGENOME AND GENE REGULATION IN PLURIPOTENT STEM CELLS ACROSS SPECIES

2.1 Introduction

Post-translational modifications on histones and chemical modifications on genomic DNA are prevalent in eukaryotic cells. They are called epigenomic (epi-) modifications. These epi-modifications enable eukaryotic genome to encode a more complex program of gene regulation by adding an extra layer of information onto the genomic sequence [18, 17]. Different epi-modifications affect how the DNA interacts with transcription factors, although many mechanisms remain unknown [30]. Due to these complexity, the genomes are far from being completely annotated on the functional level, making it necessary to first find regulatory genomic sequences before we can understand their complex regulatory roles.

Evolutionary comparisons provide a powerful tool to study genome functions. This became obvious when it was recognized that the majority of DNA can mutate freely without deleterious effects, while certain sequence elements are more constrained [1]. Leveraging this theory, researchers have inferred functional genomic segments by examining genomic sequence conservation [2] and have identified human-specific regulatory DNA by looking for sequences with accelerated rates of evolutionary change [3].

The successes in genomic comparisons beg the question: can we also use evolution to study the functions of the epigenome? To do so, the basic evolutionary properties of the epigenome must be established first, preferably in the contexts of both genomic and transcriptomic evolution.

Among many types of epi-modifications [31], a subset is known to correlate with gene transcription. For example, DNA cytosine methylation ($C^m$) [17], histone 3 lysine 27 trimethylation (H3K27me3), and histone 3 lysine 9 trimethylation (H3K9me3) may repress gene transcription, whereas histone 3
lysine 4 mono-, di-, and trimethylation (H3K4me1/2/3), lysine 27 acetylation (H3K27ac), and lysine 36 trimethylation (H3K36me3) are positively associated with transcription [18]. The functions of many epi-modifications have so far only been evaluated individually, primarily due to the difficulty of assessing the functional significances of colocalized epi-marks. Any two epi-marks can colocalize in some genomic regions, but such colocalizations do not necessarily serve any regulatory functions. The best documented epi-marks colocalization is probably the bivalent domain (H3K27me3+H3K4me3), which is hypothesized to be poised for activation during differentiation of embryonic stem cells (ESCs) [32]. We wish to develop a method to systematically examine the functions of epi-modifications and, more importantly, the functions of combinations of epi-marks. We propose to leverage the connection between evolutionary conservation and functional importance to achieve this goal.

Here, we introduce “comparative epigenomics” - interspecies comparison of epigenomes - as an approach for annotation of the regulatory sequences of the genome. We created a multispecies epigenomic data set from pluripotent stem cells of humans, mice, and pigs, which is comprised of genomic distributions of DNA methylation and eight histone modifications, the binding intensities of four transcription regulators (NANOG, OCT4, and P300), and transcribed RNA sequences. These data suggest that epigenomic conservation is not completely dictated by genomic sequences. [10] On the other hand, interspecies epigenomic changes are linearly correlated with evolutionary changes of transcription factor binding and gene expression, suggesting that comparative epigenomics can directly reveal critical information on gene regulation. Based on these initial analyses, we set out to discover regulatory sequences by conserved colocalization of different epi-marks. To test the functions of these putative regulatory sequences, we developed a differentiation assay in which mouse embryonic stem cells were differentiated into mesendoderm cells. Our time course chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data in this differentiation process confirmed the regulatory functions of all seven pairs of epi-marks identified by conserved colocalization. Thus, conserved colocalization is an efficient approach to identify functional epi-mark combinations from a large (combinatorial) number of random combinations of epi-marks. More importantly, comparative epigenomics reveals regulatory features of the genome that cannot be discerned from sequence comparison alone.
2.2 Epigenomic data in three species

We generated and compiled from published work the genomic distributions of nine epigenetic modifications including $C^m$, H2A.Z, H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, H3K36me3, and the binding of four transcription regulators, P300, TAF1, OCT4, and NANOG, in pluripotent stem cells of humans, mice, and pigs ($Sus$ scrofa)[33]. $C^m$ was assayed by both methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) and DNA digestion by methyl-sensitive restriction enzymes followed by sequencing (MRE-seq)[17]. Histone modifications and binding of transcription regulators were assayed with chromatin immunoprecipitation followed by sequencing (ChIP-seq). Gene expression was measured by RNA sequencing (RNA-seq) technology. Taken together, a total of 48 sequencing datasets were compiled, among which 31 datasets (87 billion bases) were generated from this study and the 17 other datasets (13 billion bases) were compiled from 3 published works[34, 35, 36]

2.3 Methods

2.3.1 Conserved correlation of the epigenome and the transcriptome

Data preprocessing

RNA-seq data for hESCs and mESCs was obtained from NCBI Short Read Archive v0.6 under accession number SRX026668 and SRX003912. Pig iPS RNA-seq data was generated by us. The raw data were mapped using Tophat Software [37]. And followed by Cufflink software [38], the FPKM values for all the genes were calculated. The ChIP-seq data of 8 epigenetic markers (H3K36me3, H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K27me3, H3K9me3 and $C^m$) for human and mouse were obtained from the GEO data set public repository. We generated the ChIP-seq data of 8 epigenetic markers for pig iPS by ourselves. All ChIP-seq data are mapped by Bowtie software [39] allowing one mismatch for each read and only accept uniquely mapped reads. In case that there are alternative transcripts for same genes,
we only select one transcript to represent each gene for further analysis. The numbers of genes which both have expression and epigenetic data in three species are: 17645 (human), 32540 (mouse) and 20427 (pig).

**Linear regression model to predict gene expression**

Mapped tags from ChIP-seq experiments are extended to 300bp to mimic the length of fragments. Each epigenetic marker \( j \) and promoter for gene \( k \) was represented by the fragment number \( X_{k,j} \) in the region (TSS±2000bp). The expression level of gene \( k \) was represented by FPKM value \( Y_k \). The values of \( X_{k,j} \) were transformed to logarithmic scale to optimize the vectors \( X_j \) as normal distribution for linear regression model, and same for expression indicator \( Y_k \). To avoid undefined values of logarithm when \( X_{k,j} \) was zero, a pseudocount \( a_j \) was optimized for each epigenetic marker. These transformed data for epigenetic counts \( X_{k,j} \) were then used as predictor variables for training linear regression models [18], the full model was shown below:

\[
\log(Y_k + 1) = \sum_{j=1}^{8} (b_j \times \log (X_{k,j} + a_j)) + b_0 + \epsilon_k \tag{2.1}
\]

Here \( b_j (j = 1, \ldots, 8) \) are slope parameters, \( b_0 \) is intercept, and \( \epsilon_k \) are residuals.

**Model selection**

The minimum sufficient pool of covariates (total 8 for all the epigenetic markers) was selected to avoid overfitting based on Bayesian Information Criterion (BIC) and confirmed by both backward and forward stepwise selection. We used \( \Delta BIC > 300 \) as cutoff for forward stepwise selection. The process was done by “stats” package in R. The predicted logarithm of expressions were calculated based on final model for each species and the Spearmans rank correlation coefficient \( r \) between measured and predicted values was calculated to evaluate the prediction power.
Cross validation

Ten fold cross validation was applied to measure prediction accuracy for the finally selected pool of covariates. All the genes in one species were randomly divide into 10 groups with almost same number of genes in each group. Each time nine groups were used as training set to fit the model and the remaining one as test set for validation of the model. The Spearmans rank correlation coefficients between measured and predicted value of the 10 linear models were calculated for all the 10 test sets. The average of these correlation coefficients was generated as an indicator of prediction accuracy. Processes repeated for each species.

2.3.2 Interspecies differences of the epigenome explains evolutionary changes of transcriptome

Data normalization and ortholog genes

For each of the three between-species pairs (human vs mouse, human vs pig, mouse vs pig), ortholog genes were matched for each between-species analysis (14341 gene pairs for human/mouse, 11178 gene pairs for human/pig, 13367 gene pairs for pig/mouse). To make the data between species more comparable, $Y_{k,s}$ and $X_{k,j,s}$ for each species $s$ were normalized by 1/50 of mean values for all the ortholog genes as shown by following formulas.

\[
Y'_{k,s} = \frac{50 \times K \times Y_{k,s}}{\sum_{k=1}^{K} (Y_{k,s})} \\
X'_{k,j,s} = \frac{50 \times K \times X_{k,j,s}}{\sum_{k=1}^{K} (X_{k,j,s})}
\]  \hspace{1cm} (2.2)

Between-species linear model

The ratios of $Y'_k$ and $X'_{k,j}$ between two of the three species were used as a measure for interspecies differences of epigenome and transcriptome. Similar linear models were constructed as those for single species. A similar psudo-count $a_{j,s_1s_2}$ ($s_1, s_2$ represent two species) was optimized for each epigenetic
marker to avoid zero numbers of $Y'_{k}$ and $X'_{k,j}$. Below was the full model for between-species analysis:

$$\log \left( \frac{Y'_{k,s_1}}{Y'_{k,s_2}} + 1 \right) = \sum_{j=1}^{8} \left( b_{j,s_1,s_2} \times \log \left( \frac{X'_{k,j,s_1}}{X'_{k,j,s_2}} + a_{j,s_1,s_2} \right) \right) + b_{0,s_1,s_2} + \epsilon_{k,s_1,s_2}$$

(2.3)

Here $s_1$ and $s_2$ represent two species, $b_{j,s_1,s_2}$ ($j = 1, \ldots, 8$) are slope parameters, $b_{0,s_1,s_2}$ is intercept, and $\epsilon_{k,s_1,s_2}$ are residuals.

The selection for best between-species linear model was also conducted based on BIC change. The prediction power was measured by correlation coefficient between measured and predicted interspecies differences of gene expression. Epigenetic markers with most significant contribution for the prediction of evolutionary changes of transcriptome were reflected in the final model.

### 2.4 Results for comparative epigenomic study

#### 2.4.1 Interspecies epigenomic changes are predictive of interspecies changes of gene expression and transcription factor binding

Through cross-species comparison of epimodifications, we find that the relative difference in epimodification intensities on different genomic features is in general consistent in pluripotent stem cells of three different mammalian species, also the co-occupancy of different epimodifications are generally conserved across species. However, epigenomic conservation is not completely determined by interspecies sequence similarity. [10] From these observations, we hypothesize that to buffer sequence changes from negative selective pressure, the epigenome must buffer genomic changes from generating phenotypic outcomes through, for example, concomitant transcriptome changes.

To explore this possibility, we started by asking whether the same combination of epimodifications is predictive of gene expression in every species. In each species, we used a linear regression model to fit the expression value of every gene to the nine measured epi-mark intensities in its promoter, and
we used a model selection procedure to choose the epi-modifications that are predictive of gene expression. With only four epi-mark intensity values, the expression of every gene can be predicted in each species (largest p value $< 10^{-16}$) (Figure 2.1A). The models did not overfit (Figure 2.1D), and the epi-marks to expression predictive power matches $52.7\% \sim 81.3\%$ of using one RNA-seq data set to predict another (as measured by $R^2$; Figure 2.1C). The epi-modifications predictive of gene expression levels were almost identical among humans, mice, and pigs, including H3K4me3, H3K36me3, H3K27me3, and H3K27ac (Figure 2.1B). The only exception was that, in pigs, H3K9me3 replaced H3K27ac in the final model due to a large correlation (0.91) between H3K4me3 and H3K27ac data. These data show that gene expression can be predicted by a conserved set of epi-marks, reiterating the idea that epigenomic conservation can be used to study gene regulation.
Figure 2.1: The same epi-modifications are predicative of gene expression levels in all three species. (A) Scatter plots of the observed gene expression levels (y axis) and the values predicted by a linear model of the epi-intensities (x axis). In the case of the mouse, the predicative power from epi- to gene expression was larger than half of the predicative power of using one gene expression (RNA-seq) data set to predict another (C). (B) Model selection procedure for the linear model. The epi-modification with the largest explanatory power, measured by $\Delta BIC$, was first selected into the linear model as a covariate. This process continued until $\Delta BIC \leq 300$. The fitted linear coefficient for each epi-modification is shown in red (positive) or blue (negative). (D) R-squares for cross validation. Ten-fold cross validation for the linear regression models of gene expression and epi-modification data, in three species. Each cross validation reserved 1/10 of the data for testing, and the validations were performed 10 times.

We then asked whether interspecies epigenomic changes are correlated to transcriptomic changes. The interspecies differences of epi-modification intensities are predicative of interspecies gene expression differences ($p$-value $< 10^{-16}$) (Figure 2.2 left). In a control experiment, when interspecies epi-
Table 2.1: Modeling fitting measured by $R^2$. When interspecies epimodification changes are used to predict gene expression changes, the original epimodification intensities in each species provide little improvement to the prediction.

<table>
<thead>
<tr>
<th>Species pairs</th>
<th>Covariates used for prediction (number of covariates)</th>
<th>Prediction power ($R^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human/mouse</td>
<td>Interspecies epimodification changes (8)</td>
<td>0.2492</td>
</tr>
<tr>
<td></td>
<td>Adding original epimodification intensities (16, total 24)</td>
<td>0.2603</td>
</tr>
<tr>
<td>Mouse/pig</td>
<td>Interspecies epimodification changes (8)</td>
<td>0.2568</td>
</tr>
<tr>
<td></td>
<td>Adding original epimodification intensities (16, total 24)</td>
<td>0.2861</td>
</tr>
<tr>
<td>Human/pig</td>
<td>Interspecies epimodification changes (8)</td>
<td>0.2208</td>
</tr>
<tr>
<td></td>
<td>Adding original epimodification intensities (16, total 24)</td>
<td>0.2358</td>
</tr>
</tbody>
</table>

Mark intensity differences were considered, the original epimodification intensities in each species did not contribute to further explain gene expression difference (Table 2.1). This implies that the epigenomic information associated with changes of gene expression between species is distinct from the epigenomic information associated with gene expression variation within a single species. In contrast, published cross-species analysis of tissue expression data found no identifiable sequence-to-expression correlation in vertebrates [6]. Similarly, we could not find any apparent correlation between interspecies sequence difference and expression difference by using a simple model (Figure 2.2 right).

2.4.2 Evolutionarily conserved co-localization of different epigenomic marks defines several classes of cis-regulatory sequences

In previous section, we mentioned that co-occupancy of different epimodifications are generally conserved across species. A further step from that, we start to test whether the genomic regions with one or a pair of epigenomic modifications are correlated with conserved genomic sequences. Genomic regions with all assayed epigenomic modifications except for H3K9me3 are correlated with sequence-conserved regions. With the exception of H3K9me3-marked regions, the genomic regions with two epigenomic modifications are
Figure 2.2: Correlations among evolutionary changes of epi-modification intensities, gene expression levels, TF binding intensities, and genomic sequences. (Left) Evolutionary changes of epi-modification intensities are predictive of gene expression changes and TF binding intensity changes. x axis, predicted gene expression or TF binding intensity changes with a linear model of interspecies epi-intensity changes; y axis, observed interspecies changes. (Right) Scatter plots between interspecies gene expression difference (y axis) and promoter sequence difference (x axis). For every orthologous gene pair, sequence difference was measured by log(m) - log(n), where m is the maximum log blastn score of all orthologous promoters (4,000 bp centered at TSS), and n is the blastn score of the orthologous promoter pair under consideration. $R^2$, square of the sample correlation coefficient.

also correlated with conserved sequences, often with stronger correlations than single epigenomic modification regions (Figure 2.3).

Based on the log ratios between the number of conserved regions carrying two epigenomic modifications and expected number, Seven pairs of epigenomic marks were identified as conserved co-modifications in pluripotent
**Figure 2.3:** Interspecies conservation of co-occupancy of different epigenomic modifications. Log ratio between the number of conserved regions carrying one (diagonal boxes) or two (non-diagonal boxes) epigenomic modifications and the expected number, calculated from a null model in which conserved regions and epigenomic modified regions appear independently. Conserved genomic regions are determined by six pair-wise comparisons, shown in six small boxes outlined with a darker edge. For example, the left-most upper box refers to the human genomic regions conserved in a human vs. mouse comparison. All genomic regions with epigenomic modifications except H3K9me3 were positively associated with conserved regions (red). H3K9me3 selectively marks non-conserved regions (blue). Bivalent domains (co-marked by repression mark H3K27me3 and activation mark H3K4me2/3) exhibited the strongest association with conserved regions.

stem cells, namely H2A.Z + H3K4me2/3, H3K27ac + H3K4me1/2, H3K27ac + H3K4me2/3, H3K27me3 + H3K4me1/2, H3K27me3 + H3K4me2/3, H3K36me3 + H3K27ac, and H3K36me3 + H3K4me1.

Next, we set out to test the functions of evolutionarily conserved epigenomic mark combinations (Figure 2.3) using a cell differentiation assay. During ESC differentiation, the directions of epigenomic changes and expression changes of nearby genes were expected to reflect the function of an epigenomic mark combination [40]. We differentiated mESCs into mesendoderm cells, a lineage in which the dynamic changes of the epigenome has not been examined. On the 6th day of differentiation, almost all cells expressed the mesendoderm protein Goosecoid (GSC) and endoderm protein SOX17 (Figure 3.5A), and exhibited typical mesendoderm morphology.
Figure 2.4: Epigenomic changes and gene expression changes during differentiation. (AG) Each panel represents a set of genomic regions associated with a pair of epi-marks. Each set of regions is categorized into four subclasses, i.e., kept both marks during differentiation (1,1→1,1), lost the first mark (1,1→0,1), lost the second mark (1,1→1,0), and lost both marks (1,1→0,0). For example, the red line (1,1→0,1) in (A) (H3K27me3, H3K4me2/3) represents sequences with loss of H3K27me3 (the first sign changes from 1 to 0) and retention of H3K4me2/3 (the second sign stays at 1). Relative gene expression values of the nearest genes to the comarked regions are plotted on the y axis. Error bars show SD of the mean.
The bivalent domain (H3K27me3 + H3K4me2/3) was the most conserved epigenomic mark combination among all 36 pairs of modifications (Figure 2.3), lending credence to the approach of using epigenomic comparison for identifying gene regulatory regions. To illustrate how our time-course experiment can reveal the functions of an epigenomic mark combination, we examined how bivalent domains regulate the early stages of mESC differentiation in a lineage-specific manner. It turned out that not all bivalent domains behave the same. Four subclasses of bivalent domains with different dynamic behaviors were discovered. On Day 6, the majority of sequences either retained both marks (40.7% of the sequences) or lost H3K27me3 while retaining H3K4me2/3 (36.7%). These sequences were preferentially located near transcription start sites (TSS). As expected, the genes whose promoters lost H3K27me3 but retained H3K4me2/3 exhibited higher expression (red line, Figure 2.4A) than those that kept both marks (purple line), which in turn were higher than those that kept H3K27me3 but lost H3K4me2/3 (green line). These data indicate that there are subclasses of bivalent promoters, which may be activated, still-poised, repressed, or suffer a loss of both marks during mesendoderm formation. We further examined the functions of genes regulated by each subclass of bivalent promoters. The genes with activated promoters are enriched for Gene Ontology (GO) terms of “TGF and receptor binding” (p-value < 10^{-17}), “mesoderm formation” (p-value < 10^{-13}), and “positive regulation of BMP pathway” (p-value < 10^{-9}), consistent with the mesendoderm differentiation process. On the other hand, “Neuron fate commitment” was enriched in both the still-poised (p-value < 10^{-118}) and the repressed (p-value < 10^{-13}) subclasses. These data reveal an intricate coordination among distinct subclasses of bivalent promoters that facilitates lineage-specific differentiation.

The conserved co-modifications H3K27me3 + H3K4me1/2 were expected to mark poised enhancers. On Day 6 of differentiation, 31.8% of H3K27me3 + H3K4me1/2 marked regions removed repression mark H3K27me3 and kept activation mark H3K4me1/2. The genes next to this subset of hypothetically activated enhancers exhibited increased expression on Day 4 and Day 6 (red line, Figure fig:expressionChangeB), even greater than the expression of genes associated with other dynamic epigenomic patterns (purple, green, and blue lines). In addition, conserved H3K27ac + H3K4me1/2 and H3K27ac + H3K4me2/3 marked active enhancers and promoters as expected (Figure
Figure 2.5: Poised promoters defined by H2A.Z and H3K4me2/3 in mouse ES cells. (A) Expression levels of four classes of genes, with promoters marked by neither (0, 0), either (1, 0) and (0, 1), or both (1, 1) of H2A.Z and H3K4me2/3. Each box plot provides the mean (thick middle line), quantiles (box edges), minimum (overlapped with lower edge), and maximum (top thin line) expression (FPKM) of each class of genes. Expression levels (1, 0) < (0, 0), and (1, 1) < (0, 1) suggest that H2A.Z is repressive. (B) Scatter plot of promoter H2A.Z intensity and gene expression level for all genes. The correlation coefficient (R) is negative.

2.4C-D).

The most conserved co-mark of H2A.Z was H3K4me2/3. H2A.Z is a variant of H2A and is required for early mammalian development [41]. Despite the usual assumption that H2A.Z is associated with active gene expression in multicellular organisms [42, 43], we found H2A.Z not to be positively associated with gene expression levels in mESCs (Pearson correlation = 0.0066; Figure 2.5B). This is consistent with the lack of global anticorrelation of H2A.Z and C/or in all three mammals (Figure 2.3). Thus, H2A.Z could be a repressor mark in mammals, and H2A.Z+H3K4me2/3 could mark poised promoters rather than active promoters as is generally assumed. Indeed, H2A.Z+ and H3K4me2/3- promoters were less active than H2A.Z- and H3K4me2/3- promoters, and H2A.Z+ and H3K4me2/3+ promoters were less active than H2A.Z- and H3K4me2/3+ (Figure 2.5A). More importantly, during differentiation, the H2A.Z+ and H3K4me2/3+ promoters in ESCs that lost the H2A.Z mark became more active (Figure 2.4E, red line), and those that lost H3K4me2/3 but kept H2A.Z were downregulated (Figure 2.4E, green line). Thus, we propose that H2A.Z is a repressor mark in mammalian pluripotent stem cells and that H2A.Z+H3K4me3 marks a class of poised promoters.

H3K36me3, H3K27ac, and H3K4me1/2 exhibited pairwise conservation.
Whereas H3K4me1 and H3K27ac were previously associated with enhancers, H3K36me3 has been typically regarded as a mark for actively transcribed regions. The conserved colocalization of H3K36me3 with H3K27ac and H3K4me1/2 tempted us to explore H3K36me3 as an enhancer mark as well. Consistent with this thought, active enhancers make transcripts (eRNA) [44]; H3K36me3 could be associated with any transcribed regions, including active enhancers. If this hypothesis holds, we would predict that H3K36me3 should avoid overlapping with bivalent (poised) enhancers. Indeed, the epi-mark that has the least colocalization with H3K27ac is H3K36me3 (Figure 2.3). During differentiation, the genes near (not overlapping with) sequences that lost H3K36me3 and H3K27ac (or H3K4me1) exhibited lower expression than those with one mark lost, which in turn exhibited lower expression than those that retained two marks (Figure 2.4F, G). Thus, we propose that H3K36me3, when coappearing with H3K27ac or H3K4me1/2, is a mark of active enhancers. In summary, it is powerful to use epi-mark combinations to annotate regulatory sequences and to form hypotheses about their functions.

The difficulties of having too many epi-mark combinations and not knowing how to distinguish random versus functional colocalizations can be overcome by using evolutionary conservation.

### 2.5 Discussion

Comparative genomics has been proved as an efficient approach to identify functional regions in human genome based on evolutionary principles. Here, we provide an evolutionary view of the mammalian epigenome and illustrate coevolutionary relationships among genomes, transcriptomes, and epigenomes. These results show how comparative epigenomics, an emerging field that studies evolutionary patterns of epigenomes, can use epigenomic information to functionally annotate genomes.

Interspecies epigenomic changes have been compared to both genomic and transcriptomic change. Our data reveal that the degree of epigenomic conservation is not always correlated with the degree of genomic conservation but that epigenomic conservation can provide additional information to genomic conservation. More importantly, the conservation levels of epigenomes are more indicative of the conservation levels of gene expression, further il-
lustrating the idea that epigenomic comparison can shed light on regulatory functions of the genome.

Some epi-mark combinations colocalize in a conserved manner. The conservation of colocalized epi-marks is much stronger than the conservation level of each epi-mark, thus making the combinations computationally identifiable. This phenomenon provides evidence that evolution appears to have left traces on mammalian epigenomes in the combination of epi-marks. We used a ES cell differentiation system to test the regulatory functions of these co-localized and conserved epi-mark combinations. The results support the regulatory functions of all these epi-mark combinations, suggesting that interspecies comparison can efficiently distinguish functional colocalization of epi-marks from nonfunctional combinations. This confirms an efficient approach to identify functional epi-mark combinations from a large (combinatorial) number of candidate combinations.

Finally, the correlated evolutionary changes of the epigenome, the transcriptome, and TF binding suggest the functional importance of the epigenome in mammalian transcription networks (TNs). This may explain the limited successes in human TN reconstruction using only the information of DNA sequence motifs and gene expression, which were sufficient for reconstruction of yeast TNs [45].

In the next chapter, we will discuss how the dynamic changes of epi-modifications in a combinatorial manner during a differentiation process could correlated with temporal transcriptomic changes and how we could use this information to infer transcriptional networks. We will further strengthen the connection between epigenome and regulatory networks by adding an extra dimension.
3.1 Introduction

An epigenome consists of chemical modifications and protein variations to the DNA and histones, and some of these modifications and variations can be passed down to an organism’s offspring [4]. Epigenomes are dynamic, and epigenetic modifications are associated with changes in gene expression [15, 16]. Thus, the epigenome adds an extra layer of information onto the genomic sequence and enables a genome to dynamically orchestrate gene expression in different cell types [18, 17]. It is argued that organismal development can be viewed as a progression of epigenomic states [4, 14]. To gain mechanistic support for this view, a number of challenges have to be addressed. First, when presented with genome-wide distributions of epigenetic modifications at multiple time points during a developmental or differentiation process, how can we find the genomic (cis-) regulatory sequences that regulate gene expression? What are the combinatorial functions of epigenetic modifications and regulatory sequences? Here, we present experimental data and a probabilistic model that utilizes the temporal changes of the epigenome to annotate the regulatory sequences. This approach classifies regulatory sequences by their temporal epigenomic patterns, and thus it can identify subclasses of cis-regulatory sequences with different regulatory functions.

Two types of associations were observed between the epigenome and gene expression. First, in a given cell type, the transcription levels of different genes are associated with the epigenomic modifications in the genomic neighborhoods of these genes. In other words, without changing cell types, epigenomic modifications at different chromosomal locations are indicative of the relative abundance of RNAs transcribed from these locations [18] (spatial correlation (S), Figure 3.1). Second, for a given gene, the temporal change
Figure 3.1: Two types of correlations between the epigenome and gene expression. Spatial correlation (S) examines different genes in a fixed cell type, and temporal correlation (T) examines different differentiation stages or cell types for a fixed gene. Spatial correlation is often much more pronounced than temporal correlation.

in its expression during a developmental or differentiation process is associated with temporal epigenomic changes [40] (temporal correlation (T), Figure 3.1). The first type of association facilitated the use of invariant epigenomic signatures in a static cellular condition to annotate genomic features [46]. However, genes are dynamically regulated in nearly all biological processes. It is important to incorporate the dynamic aspect of gene regulation into the annotation of the regulatory sequences. Here, we jointly model the position effect and the temporal effect of the epigenome, thus achieving ab initio identification and functional annotation of regulatory sequences [47].

The regulatory functions of a number of epigenetic modifications remain elusive. A case in point is DNA hydroxymethylation [48]. Methylated cytosine (5-mC) can be converted to an oxidized form 5-hydroxymethylcytosine (5-hmC) by a family of Ten-eleven translocation (TET) proteins [49, 50]. In embryonic stem (ES) cells, 5-hmC is enriched in gene bodies of actively transcribed genes [51], promoters of inactive genes [52, 53, 51], and active enhancers [54]. These seemingly conflicting data are thought-provoking for analyzing epigenetic modifications in a combinatorial manner, such that the function of each modification is investigated in the context of other modifications as well as the underlying genomic sequence. By modeling the co-appearance of different epigenetic modifications in each cell type, a pioneering method
demonstrated the power in predicting different genomic features, including enhancers and genes [46]. However, epigenomic co-appearance in static cell types does not reveal all epigenetic mechanisms of gene regulation. Two major questions remain unsolved. First, what are the upstream signals that guide specific epigenomic modifications, such as 5-hmC, to appear in specific genomic regions? Second, the regulatory functions for several epigenomic marks including 5-hmC and H2A.Z remain elusive. New ideas for combinatorial epigenomic analyses beyond the co-appearance in static cell types are needed. A natural extension in this direction is to utilize a dynamic process in which both the epigenome and the transcriptome have changes. Ideally, we need some methods that can capture combinatorial patterns of temporal epigenomic changes and correlate them with gene expression changes.

A major difficulty in analyzing epigenomic dynamics lies in the asynchronous nature of epigenomic changes in different genomic regions. Suppose a type of epigenomic change, for example the induction of H3K4me1 and 5-hmC, is a recurring pattern shared by many genomic regions. Such a pattern can be difficult to find because different genomic regions can accumulate either modification at different times. Furthermore, the corresponding changes in gene expression are not synchronized either, making it difficult to associate epigenomic dynamics with gene expression changes. To reveal the hidden rules of epigenomic dynamics and gene expression, we developed a spatiotemporal clustering model. This model clusters genomic regions by shared epigenomic changes but does not require the changes to be synchronized among a cluster of genomic regions. This was achieved by allowing each region to have its own time-specific epigenomic states and then integrating out the time of transition between the epigenomic states in the clustering model.

To investigate the functions of epigenetic modifications in a dynamic process, we differentiated mouse embryonic stem (ES) cells into mesendoderm cells [55], the common precursor of mesoderm and endoderm. At 3 time points during this differentiation process, we mapped the genomic distributions of 9 epigenetic modifications, including DNA methylation (5-mC), hydroxymethylation (5-hmC), histone variant H2A.Z, and histone modifications H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3. At the same time points, we also assayed the expression of small non-coding RNAs (ncRNAs) and total RNAs. Our model-based analysis of these temporal data
revealed several fundamental properties of epigenome dynamics, characterizing regulatory roles for functionally elusive epigenomic modifications. As an analogy to the sequence “rules” of gene regulation [56, 57], these discoveries may provide epigenomic “rules” of gene regulation.

3.2 GATE model

We developed a probabilistic model to annotate the genome using temporal epigenomic data. Two main features of this model include explicit treatment of combinatorial epigenomic changes and detecting similar but asynchronous epigenomic changes in different genomic segments.

As input data to the model, the genome is represented as consecutive genomic segments, with a typical segment size of 200 nucleotides (nt). Each segment is associated with the time-specific intensities of a set of epigenetic modifications. The model clusters the genomic segments, such that each cluster shares a similar combination of epigenetic modifications as well as their temporal changes. We call the combination of epigenetic modifications shared by a cluster of genomic segments at a given time an epigenomic state. Each cluster represents a time-series of related epigenomic states. Essentially this model assigns epigenomic states based on time-series epigenomic data.

We call this model Genomic Annotation using Temporal Epigenomic data (GATE). GATE is a hierarchical model (Pearl 1985) with two layers (Figure 3.2). The top layer is a Finite Mixture Model (FMM) [58], in which each component of the mixture represents a cluster of genomic segments that share temporal epigenomic patterns. Without considering the time factor, each component (cluster) degenerates into a set of genomic segments sharing an epigenomic state. The bottom layer models the epigenomic data in each cluster. Each cluster is modeled as a Hidden Markov Model (HMM) [59] that represents the temporal changes of epigenetic modification intensities. The hidden states are binary activity states (inactive and active), which are allowed to change with respect to time. For example, if an enhancer changes from an active enhancer into an inactive enhancer during differentiation, the hidden states for this enhancer would change from 0 (inactive) to 1 (active). In a differentiation process, State 0 can be interpreted as the initial state before differentiation (Undifferentiated, U), whereas State 1 can be regarded as
Figure 3.2: The Genomic Annotation based on Time-course Epigenomic data (GATE) model. (A-B) GATE models the genome as equal sized genomic segments, and each segment is associated with temporal epigenomic data. The model assumes that there are shared temporal epigenomic patterns among different genomic segments. GATE is a hierarchical model. The top layer is a Finite Mixture Model for clustering genomic segments (A). The bottom layer models the temporal changes within each cluster as a Hidden Markov Model (B). The hidden variables (circles) are a binary variables indicating the time of a change of regulatory activities. Emitted (vertical arrows) from the hidden variable are the intensities of each epigenomic mark.

the other state in differentiated cells (Differentiated, D). The observed data are the epigenetic modification intensities for each genomic segment at every time point (Figure 3.2). The sequencing reads from ChIP-seq experiments for each epigenetic modification on a genomic segment are modeled with a Poisson distribution, with the Poisson parameter reflecting the cluster and time-dependent epigenomic state. Thus, GATE has been completely specified as a generative probabilistic model (see Methods). In short, GATE is a mixture (FMM) of HMMs.
3.2.1 Symbols

**Indices,** $w$: genomic segments; $t$: time points; $m$: epigenomic marks; $k$: epigenomic clusters.

**Observed data,** $W$: the number of genomic segments; $M$: the number of epigenomic marks; $T$: the number of time points; $\nu$: normalized sequence counts. Note that $\nu_{w,t,m}$ is the normalized sequence count for epigenomic mark $m$ in genomic segment $w$ at time $t$. $O$: all the observed data.

**Hidden variables,** $C_w$: cluster membership of the genomic segment $w$; $H$: activity states, taking values 0 or 1.

**Pre-computed parameters,** $K$: the number of epigenomic clusters.

**Model parameters,** $\pi_k$: the proportion of genomic segments in cluster $k$; $b_{i,j}$: transition probability from state $i$ to state $j$; $\lambda$: the Poisson parameter for emission probabilities; $\Lambda$: all the model parameters.

3.2.2 The model

**The top layer.** The top layer FMM models the cluster memberships of every genomic segment. The cluster membership of genomic segment $w$ is modeled as a categorical distribution with probability $\pi=(\pi_1, \ldots, \pi_K)$:

$$C_w \sim \text{Categorical}(\pi)$$

$$P(C_w) = \sum_{k=1}^{K} P(C_w = k)\pi_k$$  \hspace{1cm} (3.1)

**the HMM at the bottom layer.** Given the cluster membership $C_w$, the potential changes of regulatory activities for genomic segment $w$ are modeled as a Markov chain. As a hidden variable, $H_{w,t} \in (0, 1)$ represents the activity state of genomic segment $w$ at time $t$.

The transition probability matrix ($b_{C_w}$) is written as:

$$b_{i,j}^{C_w} = P(H_{w,t+1} = j|H_{w,t} = i, C_w)$$  \hspace{1cm} (3.2)

where $i, j \in (0, 1)$. 

27
The conditional probability of $H_w$ given $C_w$ is:

$$P\left(H_w | b^C_w, C_w\right) = P\left(H_w, 1 | C_w\right) \prod_{t=1}^{T-1} P\left(H_w, t+1 | H_w, t, C_w\right)$$

$$= P\left(H_w, 1 | C_w\right) \prod_{t=1}^{T-1} b^C_{H_w, t, H_w, t+1} \quad (3.3)$$

Given the hidden variable, the observed sequence count for each epigenomic mark is modeled to follow a Poisson distribution (emission distribution). The Poisson parameter depends on the cluster membership and the hidden regulatory state.

$$\nu_{w,t,m} \sim Poisson \left(\lambda^C_{H_w, t, m}\right), \quad (3.4)$$

where $1 \leq m \leq M$, $1 \leq w \leq W$, $1 \leq t \leq T$.

Conditional on the cluster membership and the hidden variables, the different epigenomic marks are modeled as independent, and thus:

$$P\left(\nu_w | C_w, H_w; \lambda^C_w\right) = \prod_{t=1}^{T} P\left(\nu_{w,t} | C_w, H_w, t; \lambda^C_w\right)$$

$$= \prod_{t=1}^{T} \prod_{m=1}^{M} P\left(\nu_{w,t,m} | C_w, H_w, t; \lambda^C_w\right) \quad (3.5)$$

where $\nu_w = (\nu_{w,1}, \nu_{w,2}, \ldots, \nu_{w,T})$ and $\nu_{w,t} = (\nu_{w,t, 1}, \nu_{w,t, 2}, \ldots, \nu_{w,t, M})$.

Thus, a generative probabilistic model for all data has been fully specified.

**Likelihood function.** Under model assumptions, the likelihood function of observed data ($O$) is

$$P\left(O | b, \lambda, \pi\right) = \prod_{w=1}^{W} P\left(\nu_w | b, \lambda, \pi\right) = \prod_{w=1}^{W} \left(\sum_{C_w=1}^{K} P\left(\nu_w | C_w; b^C_w, \lambda^C_w\right) P\left(C_w\right)\right)$$

$$= \prod_{w=1}^{W} \left(\sum_{C_w=1}^{K} \sum_{H_w} \left[P\left(\nu_w | C_w, H_w; \lambda^C_w\right) P\left(H_w | b^C_w, C_w\right)\right]\right)$$

$$\quad (3.6)$$

where $b = (b^1, b^2, \ldots, b^K)$, $\lambda = (\lambda^1, \lambda^2, \ldots, \lambda^K)$. 

28
Parameter inference. The hidden variables of interest \((C, H)\) were estimated by maximum likelihood estimation. We implemented a nested expectation-maximization (EM) algorithm in which the transition and emission parameters of the HMM were estimated by a Baum-Welch algorithm. (Figure 3.3; Section 3.2.3)

3.2.3 The EM implementation for parameter estimation of GATE model

We implemented a nested expectation-maximization (EM) algorithm to estimate the parameters within the model. Baum-Welch algorithm was applied for transition and emission parameters of each HMM within the Maximization (M) step. The Expectation and Maximization steps are described as follows:
E-step
Let \( z_{w,k} \) be the (0,1) cluster membership indicator of genomic segment \( w \). \( z_{w,k} = 1 \) if genomic segment \( w \) is in cluster \( k \); otherwise, \( z_{w,k} = 0 \). \( z_{w,k} \) is the missing data. The Q-function is

\[
Q \left( \Lambda, \Lambda^{(old)} \right) = \sum_{w=1}^{W} \sum_{C_w=1}^{K} T_{w,C_w}^{(old)} \left[ \log \pi_{C_w} + \log f(V_w | \lambda, b, C_w) \right] \tag{3.7}
\]

where \( \Lambda = \{ \pi, b, \lambda \} \) and

\[
T_{w,k}^{(old)} = P \left( Z_{w,k} = 1 | V_w = \nu_w, \Lambda^{(old)} \right) = \frac{\pi_k^{(old)} f(V_w | \lambda_k^{(old)}, b_k^{(old)}, k)}{\sum_{s=1}^{K} \pi_s^{(old)} f(V_w | \lambda_s^{(old)}, b_s^{(old)}, s)}
\]

\[
f(V_w | \lambda, b, k) = \sum_{H_w} \left[ \prod_{t=1}^{T-1} b_{H_{w,t}, H_{w,t+1}}^{k} \left( \prod_{t=1}^{T} \prod_{m=1}^{M} \frac{\lambda_{H_{w,t},m}^{k} \nu_{w, t,m} e^{-\lambda_{H_{w,t},m}^{k}}}{\nu_{w, t,m}} \right) \right]
\]

where \( \Lambda^{(old)} \) is the value of \( \Lambda \) obtained from the previous step.

M-step
The parameter estimate \( \Lambda^{(new)} \) is obtained by maximizing the Q-function.

\[
\Lambda^{(new)} = \left( \pi^{(new)}, \lambda^{(new)}, b^{(new)} \right) = \arg \max_{\Lambda} \left( Q \left( \Lambda, \Lambda^{(old)} \right) \right) \tag{3.8}
\]

In particular,

\[
\pi_k^{(new)} = \frac{\sum_{w=1}^{W} T_{w,k}^{(old)}}{\sum_{w=1}^{W} \sum_{s=1}^{K} T_{w,s}^{(old)}} = \frac{1}{W} \sum_{w=1}^{W} T_{w,k}^{(old)}
\]

where \( z_{w,k}^{(new)} = 1 \), if \( k = \arg \max_s \left( E \left( z_{w,s} | O, \Lambda \right) \right) \), otherwise \( z_{w,k}^{(new)} = 0 \). \( O \) is the whole dataset.

Furthermore,

\[
\left( \lambda_k^{(new)}, b_k^{(new)} \right) = \arg \max_{\lambda_k, b_k} \sum_{w=1}^{W} z_{w,k}^{(new)} \{ \log f(V_w | \lambda_k, b_k, k) \} \tag{3.9}
\]
The Baum-Welch algorithm [60] is used to maximize (Equation 3.9) and obtain parameter estimate $\hat{\Lambda}$.

### 3.2.4 Fitting data to the GATE model

For simulation data (Section 3.3), the cluster number ($i$) was estimated by the Bayesian Information Criteria (BIC):

$$BIC = -2 \times Q(\Lambda|\Lambda^{(\text{final})}) + (2 \times K - 1 + K \times M \times 2) \times \ln(W)$$

where $\Lambda = \{b, \lambda, \pi\}$ is the collection of all parameters.

For real data, we initially run the model with a relatively large cluster number (55), which was estimated from a previous study [46]. The model-generated clusters were then merged into larger groups based on hierarchical clustering (Section 3.2.5). We set the hidden state at the first time point as 0. We set $b_{i,0}^c = 1 - b_{i,1}^c = 0$, because it is unlikely to make two switches of regulatory states within this short differentiation time course.

### 3.2.5 Hierarchical clustering

The hierarchical clustering was done using the $\lambda$ parameters for each cluster with R function "hclust". Euclidean distance was chosen as metric and average linkage method was used. The Hierarchical tree was cut to ensure largest height range within that cut (we consider the cuts with same group numbers as same cut), together with the requirement that group number larger than 10 after cut.

### 3.3 Simulation analysis

We simulated 4 epigenomic marks on 8,000 genomic segments from 4 clusters. Epigenomic data were simulated in three time points. Each data point in cluster $k$ was sampled from a HMM with the transition probability matrix $b^k$ and the emission distribution of Poisson ($\lambda_{i,m}^k$), where $i$ is the hidden state and $m$ is the epigenomic mark (Table 3.1).
Table 3.1: Parameters used in simulation for GATE model.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of segments</th>
<th>Markers</th>
<th>Transition probability (b)</th>
<th>$\lambda_{0,m}^k$</th>
<th>$\lambda_{1,m}^k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster1</td>
<td>2000</td>
<td>M1</td>
<td>0</td>
<td>1</td>
<td>0.2 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>10</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M4</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Cluster2</td>
<td>2000</td>
<td>M1</td>
<td>20</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>5</td>
<td>20</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>5</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M4</td>
<td>5</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Cluster3</td>
<td>2000</td>
<td>M1</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>20</td>
<td>10</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M4</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Cluster4</td>
<td>2000</td>
<td>M1</td>
<td>15</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2</td>
<td>25</td>
<td>50</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M3</td>
<td>40</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M4</td>
<td>45</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>
The simulated data mimicked real data in three aspects. First, the real data formed 4 large clusters (groups) corresponding to enhancers, promoters, gene bodies, and repeats. Second, the simulated data had different temporal patterns for different clusters. Furthermore, different hidden states of the same genomic segment emitted data with different emitting distributions, mimicking the change of regulatory functions. Finally, the number of simulated time points is the same as the time points of the ES cell differentiation experiment.

The simulated data were provided to the GATE model as input data. The parameter sets we used for the program are: maxIteration = 1000, nstep = 20, ndistance = 0.001, initial = 2. We ran the program for cluster numbers 2 to 8 and used BIC to choose the best cluster number. Four was found to be the best cluster number, which was consistent with our simulation.

Then, we compared the clustering accuracy of GATE with K-means algorithm. The average misclassification rate (the proportion of genomic segments that are incorrectly clustered) of K-means was 23.91%, which was 133 times larger than that of GATE (0.18%). The optimized cluster number (4) was used for k-means clustering with the algorithm of Hartigan and Wong.

The GATE estimated parameters ($\lambda$, $b$) were close to real parameters (Figure 3.4A). More importantly, 99.10% of the hidden states were correctly predicted (Figure 3.4B). This is a useful feature because the hidden states reflect when the regulatory function changes.

### 3.4 Application of GATE model on mESC guided differentiation

#### 3.4.1 Differentiation of mouse ES cells to mesendoderm

Mesendoderm is the diverging point of definitive endoderm and mesoderm [55], which represents an important cell lineage besides the neural lineage (ectoderm) during the early stages of ES cell differentiation. We differentiated mouse ES cells to mesendoderm using Activin and a previously described culture medium [55]. On the sixth day of differentiation, almost all cells exhibited typical mesendoderm morphology and expressed mesendoderm pro-
Figure 3.4: Evaluation of model prediction on simulated dataset. (A) Comparison of model-learned parameters with real parameters in simulation. Here M1 - M4 are simulation for four different histone marks. $b_{0 \rightarrow 0}$ is the transition probability from state 0 to state 0. (B) Comparison of model-learned hidden states (orange) with real hidden states in simulation (blue). 0 and 1 correspond to the two hidden activity states. "0 $\rightarrow$ 0 $\rightarrow$ 0" denotes State 0, 0, 0 in the three time points.

tein Goosecoid (GSC) [61] and endoderm protein SOX17 [62] (Figure 3.5A). Pluripotency genes Pou5f1 (a.k.a. Oct4), Sox2, and Nanog were down-regulated, whereas endoderm and mesoderm genes Gsc, Chordin, Foxa2, Sox17, Lim1, and Hnf4 were up-regulated (Figure 3.5B).

We measured a total of 9 epigenomic marks at 3 time points (Day 0, 4, and 6) during the differentiation process. These marks included 7 histone modifications or variants (H3K4me1/2/3, H3K27ac, H3K27me3, H3K36me3, and H2A.Z), which were assayed by chromatin immunoprecipitation followed by sequencing (ChIP-seq) [10]. We supplemented the histone data with 2 types of DNA modifications, including 5-hmC by chemical labeling and pull-down followed by sequencing (5-hmC-seq) [63] and 5-mC by both methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) and DNA digestion by methyl-sensitive restriction enzymes followed by sequencing (MRE-seq) [17]. The 5-hmC pull-down specifically used the chemical property of the hydroxyl-group and thus was efficient to distinguish 5-hmC from 5-mC [63].
Figure 3.5: Differentiation of mouse ES cells into mesendoderm cells. (A) Immunofluorescence staining of cells at 0, 4, and 6 days of differentiation. DNA was stained by Hoechst in blue. GSC and SOX17 proteins were stained in red and green. At the 6th day of differentiation, almost all cells express GSC and SOX17. (B) Comparison of mRNA expression of pluripotency and lineage-specific marker genes. Real time quantitative PCR analyses were carried out in three biological replicates at each time point. Fold change: the ratio of expression levels between Day 4 and Day 0 (yellow), and the ratio between Day 6 and Day 0 (green).

MeDIP-seq was representative of 5-mC, and MRE-seq was representative of unmethylated CpGs (uCpG) [17].

To analyze the transcriptome, we sequenced ncRNAs using the Illumina Small RNA Sample Preparation procedure followed by sequencing [64] and mRNAs using RNA-seq [65] at the same 3 time points. Taken together, 36 sequencing datasets comprised of 1.94 billion 75 nt or 100 nt uniquely
alignable sequencing reads were generated.

These data allowed us to estimate that 11.5% of the mouse genome is associated with at least one type of epigenetic modification in undifferentiated ES cells. Nearly half of these regions (5.60% of the genome) exhibited significant changes in at least one epigenetic modification during differentiation. About 1.92% of the genome was transcribed into mRNAs in ES cells, and 0.43% of the genome exhibited change of mRNA expression levels during differentiation.

3.4.2 Processing of ChIP-seq data to be fed into GATE model

The ChIP-seq reads of 8 epigenomic marks (H3K36me3, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H2A.Z, and 5-hmC) in three time points were mapped onto the mouse genome (mm9) with Bowtie software [39] allowing 1 mismatch. The number of sequence reads for each genomic segment (200 nt) was counted and then normalized by the total number of mappable reads. The standardized sequence counts were log-transformed, multiplied by 10 and rounded to the nearest integer:

$$\nu_{w,t,m} = \left[10 \times \ln (n_{w,t,m} + 1)\right]$$  \hspace{1cm} (3.10)

where \([n]\) is the largest integer no larger than \(n\), and \(n_{w,t,m}\) is the normalized ChIP-seq read count for epigenomic mark \(m\) on genomic segment \(w\) at time \(t\).

The preprocessed variables \(\nu_{w,t,m}\) were calculated here as visible observation input of GATE model.

3.4.3 Spatiotemporal clustering of epigenomic states

GATE clusters genomic segments based on both spatial distributions of epigenomic modifications and temporal changes of these modifications. Applying GATE to the ES cell to mesendoderm differentiation dataset, we initially obtained 55 clusters, consistent with the previously estimated number of chromatin states [46]. These clusters formed 14 larger groups (Figure 3.7). Twelve of the 14 groups showed epigenomic characteristics that are typical to gene bodies, promoters, and enhancers. For example, Groups 2,
3, 6, and 9 shared enhancer characteristics including low H3K36me3, high H3K4me1 in either undifferentiated (U, Figure 3.6A) or differentiated states (D), and either high H3K27me3 or high H3K27ac (vertical green bar, Figure 3.6). Groups 1b, 5, 8, and 12 shared promoter characteristics including high H3K4me3 and low H3K36me3 (red bar). Groups 1a, 4, 10, and 11 shared high levels of H3K36me3, which was associated with gene bodies [66, 67] (blue bar, Figure 3.6).

By assigning clusters as promoter, enhancer, gene body, and repeat clusters (red, green, blue, and gray bars, Figure 3.6), we turned the unsupervised spatiotemporal epigenomic clusters into predictions of different genomic features. To check these predictions, we compared the locations of the genomic segments in every cluster to their nearest genes. Indeed, the relative locations of genomic segments in each cluster corroborated the unsupervised predictions (Figure 3.8). Chromosome 11 was randomly chosen for quantifying the prediction accuracies. The sensitivities for detecting promoters increased from 0 to 60% when the false positive rate (1-specificity) increased from 0 to 0.25% (Figure 3.9B). Similar tradeoffs between sensitivity and specificity were found for gene body predictions (Figure 3.9C). These quantities reinforced the visual impression (Figure 3.8) that spatiotemporal clusters correlate with different genomic features. Changing the input size of genomic segments from 200 nt to 100 nt did not change any qualitative characteristics of the clustering results (data not shown).
Figure 3.6: GATE predicted epigenomic clusters. (A) Average intensities of each epigenomic mark (column) in each cluster (row). The model allows each sequence segment to have two activity states, denoted as undifferentiated (U) and differentiated (D). The model assumes that each cluster has a shared mean intensity for each epigenomic mark at either activity state. These mean intensities are plotted in this matrix and are color coded. Clusters with similar intensities were merged into larger clusters (Groups. See Figure 3.7 for the merging procedure). Based on the epigenomic patterns, the groups were assigned with representative names, including promoters, enhancers, genes, and repeats (vertical color bars). Consistent intensity changes across multiple clusters in a group were highlighted (red circles). (B) Fold enrichment of genomic and epigenomic features (column) in each cluster (row). Fold enrichment was calculated as the ratio between the average signal of a cluster to the average signal of all clusters. mRNA: transcription levels of nearest genes, derived from RNA-seq data. ncRNA: transcription levels of the genomic segment in each cluster, derived from ncRNA-seq data. CpG: CpG density. P300: P300 binding intensity. PolII: PolII binding intensity. Repeats: repeat density. Significant temporal changes were highlighted (orange circles).
Cluster dendrogram
Height
Groups: 14 12 5 8 10 1a 2 6 3 9 4 13 11 7 1b

Figure 3.7: Dendrogram of epigenomic clusters. The hierarchical tree represents the relative distances between the clusters. Each leaf is an epigenomic cluster (cluster number shown below each leaf). Cutting the dendrogram at the height of 30, 14 groups were formed. Each group contained several clusters with similar spatiotemporal epigenomic patterns. Group 1a and Group 1b are the two main branches of Group 1. The input size of genomic segments for GATE clustering was 200 nt.

3.4.4 Spatiotemporal epigenomic clusters are predictive of transposons, bidirectional promoters, miRNA promoters, and piRNAs

The spatiotemporal epigenomic clusters did not only predict usual genomic features such as enhancers, promoters, and gene bodies, unexpectedly, these clusters were also capable of predicting genomic features including repeats, bidirectional promoters [68, 69], microRNA (miRNA) promoters [70], and PIWI RNAs (piRNA) [71].

Two groups (Groups 13 and 14) exhibited unfamiliar epigenomic characteristics, including high H3K36me3 in parallel with high H3K27me3. These clusters did not locate in gene bodies or promoters (Figure 3.8). They contained the highest proportions of repeats among all 55 clusters (p-value = 5.45 × 10^{-37}, Fishers exact test, “Repeats” column, Figure 3.6B). ncRNA expression showed that Groups 13 and 14 corresponded to transcribed and silenced repeat sequences (ncRNA, Figure 3.6B). These data suggest a distinct
Figure 3.8: Genomic locations of epigenomic clusters. The relative location of every genomic segment with respect to the nearest gene was categorized (columns). These relative locations were summarized for each cluster (row). The relative abundance of each cluster (row) in each location category (column) is quantified by fold enrichment. The fold enrichments are shown in a green-red heatmap. The order of groups and clusters (rows) is the same as in Figure 3.6. Vertical color bars on the left indicate an unsupervised assignment of genomic features to the clusters.
spatiotemporal epigenomic signature for repeats and transposons.

Bidirectional promoters were strongly enriched in Group 8 (Clusters 19, 12, 48) (p-value = $3.9 \times 10^{-290}$, Figure 3.9A). A simple but powerful classifier for identifying bidirectional promoters can be built based on the clusters. Based on whether a genomic segment belongs to Cluster 19, one can reach 80% sensitivity with a specificity of 99.6% (Figure 3.9D). Besides bidirectional promoters, miRNA promoters were also enriched in specific clusters, including Cluster 28 (p-value = $6.96 \times 10^{-28}$, Fishers exact test) and Cluster 52 (p-value = $1.51 \times 10^{-11}$, Fishers exact test), (Figure 3.10A). Thus, through unsupervised clustering, GATE revealed distinct spatiotemporal epigenomic patterns in several specific types of promoters.

PiRNAs and PIWI proteins were discovered in germ cells [71] and were thought to be silenced in ES cells. Unexpectedly, piRNAs were specifically enriched in Cluster 11 (p-value = $3.08 \times 10^{-16}$, Fishers exact test) and Cluster 28 (p-value = $2.05 \times 10^{-6}$, Fishers exact test) (Figure 3.10B). As a control, in Cluster 11 where piRNA was strongly enriched (9.53-fold more enriched than expected), miRNAs were depleted (0.89-fold less than expected). Thus piRNAs and miRNAs had different temporal epigenomic characteristics. These data suggest that even though piRNAs were produced and functional in germ cells, specific epigenomic patterns were formed on piRNA genes much earlier than germ cell development. PiRNA genes may be epigenetically prepared for activation in ES cells.

The distinct epigenomic characteristics in ES cells opened the possibility that a subset of piRNAs is produced in ES cells. Indeed, a cluster of piRNA genes (piRNA cluster) on Chromosome 5 was clearly expressed (Figure 3.10C). Moreover, the expression of this piRNA cluster was specifically induced in undifferentiated ES cells (Days 0, 4, 6, Figure 3.10C). As a control, Sgsm1, the neighboring gene to this piRNA cluster, showed a slight increase in expression during differentiation (Figure 3.10C). Even more strikingly, Piwil2 (a.k.a. Mili), a mouse orthologue of the drosophila PIWI gene, is expressed in ES cells, and its expression decreased below a detectable level 4 days after differentiation (Figure 3.10D). The consistent inductions of piRNAs and the Piwil2 gene in undifferentiated ES cells further entertained the hypothesis that some piRNAs were produced not only in germ cells.
Figure 3.9: Predicting genomic features. (A) Distribution of bidirectional promoters in the 55 epigenomic clusters. Fold enrichment: the ratio between the percentage of bidirectional promoters in a cluster and the average percentage of all clusters. *: p-value < $10^{-60}$. **: p-value < $10^{-90}$. (B-D) Accuracies of predicting genomic features as measured by receiver operating characteristic (ROC) curves. AUC: area under the curve. Promoters were predicted by Groups 1b, 5, 8, and 12 (p-value < $2.2 \times 10^{-16}$, Wilcoxon test). Gene bodies were predicted by Groups 1a, 4, 10, and 11 (p-value < $2.2 \times 10^{-16}$). Bidirectional promoters were predicted by Cluster 19 (p-value < $2.2 \times 10^{-16}$). Inserts: details of the high specificity regions. (E) A predicted bidirectional promoter and a regular promoter. Along a fraction of Chromosome 11, each genomic segment is colored by the cluster it belonged to. The stretch of DNA belonging to Cluster 19 (yellow) corresponded to a bidirectional promoter.

3.4.5 Recurrent themes of epigenomic and transcriptome changes

Recurrent themes appeared in the majority of the spatiotemporal clusters. These recurring patterns may represent basic properties of temporal gene
regulation by the epigenome.

First, combinatorial epigenomic changes are prevalent. In every case except one, we observed combinatorial changes of 3 to 7 epigenomic modifications. Furthermore, different genomic features (promoters, enhancers, genes, and repeats) have different combinations of temporal changes.

Second, combinatorial patterns of epigenomic changes are predictive of gene expression changes (temporal axis, Figure 3.1). While gene expression data were not used in clustering epigenomic data, the epigenomic clusters clearly distinguish gene groups with different temporal expression changes (Figure 3.6). Third, almost all combinatorial epigenomic changes correspond with changes in ncRNA expression levels. The direction of ncRNA expression changes was not associated with the changes in any single epigenomic modification but was strongly predictable by combinatorial changes (Figure 3.11).

Fourth, epigenomic changes in enhancers instead of promoters are indicative of mRNA expression changes. This reconciles previous observations that temporal epigenomic changes were poorly correlated with gene expression changes during cell differentiation [72] by reproducing such results in promoter regions with more epigenomic modifications; but also it points out the importance of epigenomic changes in enhancers. Consistently, human ChIP-chip analysis showed enhancer associated modifications including H3K27ac and H3K4me1 had greater dynamic changes than other modifications during ES cell differentiation [16].

Fifth, all assayed modifications except H3K36me3 have robust temporal changes in multiple genomic features (promoters, enhancers, gene bodies, and repeats). The robust and recurrent temporal changes appeared not only in genomic regions where the epigenomic modifications were abundant, but also in genomic regions where the modification levels were low. Previously, H3K4me1 was associated with enhancers and H3K4me3 and H2A.Z were associated with promoters due to their abundance in these regions. However, H3K4me1 showed reliable changes not only in enhancers, but also in promoters and gene bodies where its modification level was low. Similarly, H3K4me3 (Group 7) and H2A.Z (Group 9) showed reliable changes in enhancers. These data may suggest regulatory functions of epigenomic modifications in previously ignored genomic regions.
Figure 3.10: Distribution of miRNA promoters and piRNA genes in GATE clusters. Fold enrichment: the ratio between the percentage of miRNA promoters (A) or piRNA genes (B) in a cluster and the average percentage of all clusters. *: p-value < 10\(^{-5}\). **: p-value < 10\(^{-15}\). (C) An piRNA gene cluster on Chromosome 5. Cluster 11 genomic segments are enriched in this piRNA gene cluster. This piRNA gene cluster is expressed in ES cells (RNA-seq Day 0), but is not expressed after differentiation (Day 4, Day 6). The nearby gene, *Sgsm1*, shows a relatively constant expression. (D) Expression of PIWI protein genes *Piwil2* (Lane a), *Piwil1* (Lane b), *Piwil4* (Lanes c-f) in ES cells and during differentiation. FPKM: fragments per kilobase of exon per million fragments mapped.
3.4.6 Temporal changes of H3K4me2, non-CpG methylation, and H2A.Z are predictive of DNA hydroxymethylation

It remains unknown what guides TET enzymes to specific parts of the genome to convert 5-mC to 5-hmC. To explore the upstream signals that might specify where in the genome 5-mC should be converted to 5-hmC, we asked if there were any epigenomic modifications that correlate with 5-hmC in terms of temporal changes. Across all 55 epigenomic clusters, the temporal changes of 5-hmC were on average most correlated with H3K4me2, unmethylated CpG (\(^5\)CpG, measured by MRE-seq), H3K4me1, and H2A.Z (Figure 3.6). Next, we checked the temporal correlations between 5-hmC and every other assayed epigenomic mark on every genomic segment (200 nt window).
78.1% of the genomic segments the temporal correlations between 5-hmC and H3K4me2 were larger than 0.8 (p-value $< 10^{-300}$, Figure 3.12B). Strong temporal correlations between 5-hmC and $^u$CpG, 5-hmC and H4K4me1, and 5-hmC and H2A.Z were also observed (Figure 3.12E,F). In contrast, H3K4me3 and H3K36me3 did not show large temporal correlations with 5-hmC. Categorizing genomic segments by their clusters, promoter segments showed the strongest temporal correlations between 5-hmC and H3K4me2, $^u$CpG, and H2A.Z (Figure 3.12). These data indicate strong associations between the di- and mono- methylation of H3K4 and hydroxymethylation of nearby cytosines. The exchange of histone variants H2A and H2A.Z may also associate with 5-hmC synthesis.

The temporal changes of unmethylated CpG ($^u$CpG) and 5-hmC were strongly correlated (Figure 3.12K). Two scenarios can fit this data. First, the $^u$CpG was generated by a cytosine demethylation preprocess that involves the conversion of 5-mC to 5-hmC [73]. In other words, the same cytosine is converted from 5-mC into 5-hmC and then into C (same cytosine hypothesis). Alternatively, $^u$CpG signals TET enzymes to the genomic neighborhood to convert neighboring 5-mC into 5-hmC (neighbor hypothesis). If the same cytosine hypothesis is true, we would predict that 5-mC and 5-hmC are anti-correlated during ES cell differentiation. However, temporal changes of 5-mC did not anti-correlate with 5-hmC changes (Figure 3.12M). More genome segments had the same direction of 5-mC changes and 5-hmC changes than expected at random (Dashed line, Figure 3.12M, p-value $< 10^{-200}$). These data are inconsistent with the same cytosine hypothesis. Conversely, TET1 contains a Znf_CXXC domain that interacts with $^u$CpG, which is in line with the neighbor hypothesis.

We then explored the roles of non-CpG methylation (mCpH). mCpH was reported in oocytes without known functions [74]. These mCpHs were presumably due to high levels of de novo methylation enzymes DNMT3a/b in oocytes [75]. The simultaneous increase of 5-mC and decrease of CpG methylation in the same promoters (Figure 3.11) suggest de novo non-CpG methylation. This is consistent with the increased expression levels of DMNT3b during guided differentiation of ES cells towards mesendoderm cells (Figure 3.12N). Temporal changes of mCpH were strongly correlated with 5-hmC changes (Figure 3.12D), suggesting the genomic segments undergoing non-CpG methylation were also experiencing hydroxymethylation. Complement-
tary to these data, 5-hmC level was most enriched in low CpG regions in ES cells [76]. In summary, temporal changes of H3K4me1/2, mCpH, and H2A.Z are predictive of 5-hmC changes throughout the mouse genome. These temporal correlations do not provide any causal information, but they may help to prioritize some hypotheses for future biochemical analyses of 5-hmC pathways.
Figure 3.12: Temporal correlations between 5-hmC and other epigenomic marks. (A) Distribution of Pearson correlation coefficients between 5-hmC and the expression of nearby genes (the nearest gene within 100,000 bp for each segment). (B-L) Distribution of Pearson correlation coefficients between 5-hmC and other epigenomic marks. The distributions were separately calculated for genomic segments in promoter (red), enhancer (green), and gene body clusters (blue). A background distribution was derived by permuting the data from the three time points (dotted line). (M) The calculation procedure for the distributions in Panels A-L. First, on every genomic segment, a correlation was calculated between two marks using their intensities on three time points. Second, the correlations on all genomic segments were summarized into a histogram. (N) Expression levels of DNA methylation enzyme genes Dmnt3a (different transcript isoforms in Lanes a-b) and Dmnt3b (Lanes c-n). Several Dmnt3b transcripts showed increased expression during differentiation.
3.4.7 Epigenomic states correlate with transcription networks

The GATE model infers epigenomic states of every genomic segment as undifferentiated (U) or differentiated (D), indicating when a genomic segment may change its regulatory functions. Thus, the GATE model provides a genome-wide view of the locations of regulatory sequences as well as their time of activation. Such information may help to clarify the transcription network [16]. To explore this potential, we did a case study for three mesendoderm genes: Fgf8, Sox17, and Foxa2. A set of Group 3 enhancers were found in the introns and 3 intergenic regions of Fgf8 (yellow boxes, Figure 3.13A). These enhancers were predicted to shift from inactive to active epigenomic states (U → D). Transcription factor binding sites (TFBS) of GSC and IRF-1 appeared in these enhancers (Figure 3.13B,C). Both GSC and IRF-1 are key regulators of mesendoderm differentiation [61, 77]. An isolated enhancer was identified around 50,000 nt upstream to the Sox17 gene (Figure 3.13D). Another enhancer was found around 7,000 nt upstream to the Foxa2 gene (Figure 3.13E). Using epigenomic data, GATE suggested that both the Sox17 enhancer and the Foxa2 enhancers shifted from inactive to active states (Figure 3.6, Group 3), which was in line with the increased expression of these genes (FPKM, Figure 3.13D, E). A strong FOXA2 TFBS appeared in the Sox17 enhancer, and a strong SOX17 TFBS appeared in a Foxa2 enhancer. These data suggest a positive feedback loop between Sox17 and Foxa2. Coincidently, a peak in a FOXA2 ChIP-seq experiment in mouse liver (GEO accession number: GSM427089 [78]) co-localized with the GATE predicted Sox17 enhancer (data not shown). This is the strongest peak (p-value < 10^{-6}) in the 70,000 nt sequence neighborhood of the Sox17 gene. Moreover, the predicted FOXA2 TFBS appeared at the center of this peak (Figure 3.13D). Reversely, when Sox17 expression was induced in mouse ES cells, a strong peak (fold change = 8.1, p-value < 2.5 × 10^{-7}) of SOX17 ChIP-chip (GEO accession number: GSM470844 [79]) appeared 7,000 nt upstream to the Foxa2 gene, co-localizing with the GATE predicted Foxa2 enhancer (data not shown). Moreover, the predicted SOX17 TFBS located precisely at the center of this peak (Figure 3.13E). These ChIP-seq/chip data reinforced the GATE predicted feedback loop. This feedback loop can stabilize the activation of two master transcription factors, and thus may be essential for mesendoderm differentiation.
Figure 3.13: Predicted mesendoderm enhancers harbor transcription factor binding sites. Epigenomic clusters near the *Fgf8* (A), *Sox17* (D), and *Foxa2* (E) genes. Genomic segments (colored bars) were marked by their cluster numbers on the left. Their variable widths indicate their activity states. A left-thin-right-fat bar indicates a change of the activity states. A strong GSC motif (B) and a strong IRF-1 motif (C) appeared in predicted enhancer segments in the 3 and the intron of *Fgf8* gene. Both predicted enhancers showed changes of activities during the differentiation (left-thin-right-fat). A strong FOXA2 motif appeared in a predicted enhancer 50,000bp upstream to the *Sox17* gene (D). In turn, a strong SOX17 motif appeared in a predicted *Foxa2* enhancer (E). FPKM: fragments per kilobase of exon per million fragments mapped.
3.5 Discussion

Previous computational methods primarily utilized spatial information of epigenomic marks to predict chromatin states. For example, an HMM model was developed to annotate genomic sequences by co-localization of multiple epigenomic marks [46]. GATE connects with the Ernst-Kellis model in that with only one time point, GATE degenerates into a zero-order HMM. Unlike the Ernst-Kellis model, though, GATE added an extra dimension of temporal epigenomic changes while maintaining the prediction accuracy. The unsupervised nature of GATE also makes it capable of predicting the genomic features that were not included in a training process. In the ES cell differentiation process, GATE predicted bidirectional promoters, miRNA promoters, and piRNA genes with high accuracies.

Temporal information is as important as spatial information in studying epigenomic functions. A case in point is that although TET was known to interact with trithorax homolog MLL [49], the MLL targets H3K4me1/2 were not pursued as a major clue for guiding TETs to specific genomic regions. This was probably due to the lack of a very strong spatial correlation between H3K4me1/2 and 5-hmC in any studied cell types. Indeed, 5-hmC was enriched not only in enhancers where H3K4me1/2 levels were high, but also in promoters [76], CTCF binding sites [76], and gene bodies [80] where H3K4me1/2 levels were not necessarily high. However, temporal correlations between H3K4me1/2 and 5-hmC were particularly strong, in that more than 85%

People questioned "CpGs capability to attract TET1 [81], despite that TET1 contains a zinc finger CXXC domain that can bind "CpG [49]. Indeed, no genome-wide mapping has showed strong overlaps of 5-hmC and "CpG. Furthermore, the information content of CpG is small, making it hard to believe that such a weak sequence signal can confer specificity to guide TET1. In this study, we reported striking temporal correlations of 5-hmC and "CpG throughout the genome, highlighting the necessity of analyzing epigenome dynamics and providing genome-wide data to support the role of "CpG in guiding TET1. TET1s interacting partner MLL contains a zinc finger CXXC domain as well. Theoretically, the MLL-TET1-"CpGs three way interaction can be a lot more stable than a two way interaction of a protein and its DNA recognition site [82]. This MLL-TET1-"CpGs interaction is reinforced by
MLLs roles to methylate H3K4 and interact with methylated H3K4. These analyses provide a model that \(^4\)CpGs guide TET1 to specific genomic locations by initiating self-reinforcing \(^4\)CpGs-MLL-H3K4me1/2-TET1 interactions.

It remains controversial whether 5-hmC predominantly exists in the CpG context. Stand bias analysis suggested presence of 5-hmC on CpH in ES cells [52]. However, this result was not supported by single-base resolution mapping of 5-hmC [76]. We observed positive temporal correlations of \(^4\)mCpH and 5-hmC in multiple genomic regions. Two scenarios fit this observation. First, 5-hmC existed on CpH; alternatively, the temporal changes of mCpH were associated with 5-hmC changes in nearby CpGs. In promoters, \(^4\)CpG levels measured by MRE-seq often increased as 5-hmC levels increased, which suggests that at least a subset of the newly converted 5-hmC in promoters were on CpHs. These results suggest perhaps examining differentiated ES cells may resolve the controversy of the presence of 5-hmC on CpHs.

Most of the epigenomic marks (all assayed except H3K36me3) showed robust temporal changes in multiple genomic features, including promoters, enhancers, gene bodies, and repeats. This recurring theme can have large implications for studying gene regulation. In the canonical view, certain modifications are indicative of certain genomic features; for example, H3K27ac and H3K4me1 are enhancer marks and H3K4me1 is a promoter mark. This canonical view was built on the observation that these modifications were a lot more abundant in certain genomic features than others. This view made it tempting to ignore the regulatory roles of modification in places where it is not abundant. However, the robust temporal changes of many modifications in their non-canonical (low-abundance) regions, such as H3K27ac in gene bodies, H3K4me1 in gene bodies and promoters, and H3K4me3 in gene bodies and enhancers, indicate that they can play regulatory roles in more genomic features than in the canonical view. Future experiments are needed to test this hypothesis. Consistent with this new view, the temporal epigenomic changes in low-abundance regions were sometimes correlated with mRNA or ncRNA expression changes (Figure 3.6 and Figure 3.11).

Another dilemma in the epigenomic field is as follows. On the one hand, epigenomic changes are essential to organismal development, supported by the fact that different cell types exhibit clearly different epigenomic patterns [14, 83]. Thus, the epigenome is expected to regulate gene expression during
development and differentiation [16]. However, on the other hand, temporal epigenomic changes during cell differentiation were reported to not correlate with gene expression changes [72]. Our method and data allow us to investigate this dilemma from a new perspective.

From the methodological perspective, GATE provides two advantages. First, the unsupervised clustering summarizes the combinatorial changes of multiple epigenomic marks. Previously, one had to compare gene expression changes with every epigenomic mark one-by-one, resulting in inconclusive or even conflicting results. This was because the association between gene expression and an epigenomic mark may be confounded by other epigenomic marks. GATE enables us to correlate gene expression changes with the combinatorial changes of multiple epigenomic marks. Second, GATE makes unsynchronized changes in different parts of the genome comparable. This enables us to effectively pull information together from different genomic segments with similar but unsynchronized temporal changes.

Consistent with the previous erythroid differentiation study [72], temporal epigenomic changes did not correlate with gene expression changes in several GATE clusters. These clusters were all in promoters (Groups 5, 8, 12, Figure 3.6). However, temporal epigenomic changes in enhancer-associated clusters were clearly correlated with gene expression changes. In particular, changes in DNA methylation alone were associated with ncRNA expression change (Group 6); changes that involve different combinations of DNA and histone modifications were associated with both mRNA and ncRNA changes in different directions (Groups 3, 7, 9, Figure 3.6); changes in modifications on repeats were predictive of repeat expression (Groups 13, 14). Including non-uniquely-mapped reads into the analysis may impact the results on repeat regions. These data suggest that the epigenome mediated gene regulation during cell differentiation, although clear in enhancers, may not be discernible in promoters, thus helping to resolve the hitherto mentioned dilemma.
CHAPTER 4

RNA-HIC-TOOLS: A BIOINFORMATICS PIPELINE FOR ANALYSIS OF RNA HI-C DATA ON HIGH-THROUGHPUT MAPPING OF RNA-RNA INTERACTOME

4.1 Introduction

A large number of RNAs exhibited regulatory roles by interacting with other RNAs. Such interactions were often mediated by RNA binding proteins [84], including Argonaute proteins [85], PUM2, QKI [26], snoRNP proteins [86] and others. Despite powerful technological innovations, including the PAR-CLIP [26], HITS-CLIP [27], and CLASH [28, 29], it remains a formidable challenge to map nearly the entire protein-assisted RNA-RNA interactome.

PAR-CLIP and HITS-CLIP analyzes one RNA-binding protein at a time and requires protein-specific antibodies. Even though two RNAs showed up in HITS-CLIP or PAR-CLIP, they could have attached to different copies of the same protein, for example Argonaute protein, in the same cell or even in different cells. CLASH utilizes ectopic expression of Argonaute protein and is therefore restricted to analyze miRNA interactions in transformed cell lines.

The colleagues in our lab developed the RNA Hi-C experimental procedure to analyze all protein-assisted RNA-RNA interactions in vivo. This procedure crosslinks RNAs with their bound proteins, and ligates the RNAs co-bound by the same protein into a chimeric RNA. The chimeric RNA is interspersed by a predesigned biotinylated RNA linker, in the form of RNA1-linker-RNA2. These linker-containing chimeric RNAs are selected by streptavidin and then subjected to pair-end sequencing (Figure 4.1). Thus, each non-redundant read-pair sequence provides a piece of evidence to a physical molecular interaction.

The RNA Hi-C method offers several advantages for mapping RNA-RNA interactions. First, the one-to-one pairing information of interacting RNAs is experimentally captured. Second, by using the biotinylated linker as a selec-
tion marker, it avoids antibody based selections of specific proteins, allowing for an as unbiased mapping of the entire RNA interactome as possible. Third, the RNA ligation is done in a dilute condition which minimizes the chances of ligating random RNAs that happened to be proximal in space. Fourth, the predesigned RNA linker provides a clear boundary to split any sequencing read that spans across the ligation spot, thus avoids ambiguities in mapping the sequencing reads. Fifth, RNA Hi-C directly analyzes the endogenous cellular condition without introducing any exogenous nucleotides [26, 87] or proteins [28] before crosslinking. Sixth, potential PCR amplification biases were removed by attaching a random 6nt barcode to every chimeric RNA before PCR amplification; the overlapping sequencing reads with identical barcodes are counted only once [27, 88, 89].

We created a suite of bioinformatic tools, dubbed RNA-HiC-tools, to analyze and visualize RNA Hi-C data (http://systemsbio.ucsd.edu/RNA-Hi-C). This software package automated all the analysis steps, starting from handling raw sequencing reads and ending at several visualization capabilities. Its analysis modules included removing PCR duplicates, splitting multiplexed samples, identifying the linker sequence, splitting junction reads, calling interacting RNAs, statistical assessments, categorizing RNA interaction types, and RNA structure analysis. The visualization modules enabled global and local views of the identified interactions between RNAs or proximal sites within an RNA (ToolManual: http://systemsbio.ucsd.edu/RNA-Hi-C/sources/RNA-Hi-C-tool-ToolManual.pdf).

4.2 Experimental design for RNA Hi-C technology

The RNA Hi-C technology to identify RNA-RNA interactome contains the following main steps (Figure 4.1). Cells are cross-linked, causing the interacting RNAs mediated by a RNA binding protein to be covalently linked to this protein. RNAs are then fragmentized with RNase I and the cysteine residues on proteins are biotinylated. The proteins including protein-RNA complexes are immobilized on streptavidin beads. The 5’ end of the RNA is then ligated with a biotin-tagged RNA linker (24nt) to facilitate subsequent selective purification of chimeric RNAs. Next, proximity-based ligation is carried out on beads under dilute conditions that favor ligations between cross-linked RNA
fragments. Protein-RNA complex is then eluted from streptavidin beads and RNAs are recovered by digesting the bound protein. Eluted RNAs are subjected to rigorous DNase treatment to eliminate DNA contamination. Purified RNAs are then hybridized with a DNA probe that is complementary to the 24nt RNA linker, and treated with T7 exonuclease to remove the non-ligated biotinylated RNA linkers. As a result, mainly the successfully ligated chimeric RNAs contain a biotin-tagged linker at the junction. This chimeric RNA library is fragmented again to an average of 150 nucleotides, and the ligation junctions will be pulled-down with streptavidin-coated magnetic beads. The end product is a library of 150nt chimeric RNAs. This library is enriched with chimeras of in the form of R1-linker-R2, where R1 and R2 are fragments of interacting RNAs. This library is converted into cDNAs and sequenced with paired-end next-generation sequencing.

**Figure 4.1:** RNA Hi-C technology to map genome-wide RNA-RNA interactions. The major experimental steps: 1. crosslinking RNAs to proteins, 2. RNA fragmentation and protein biotinylation, 3. immobilization, 4. ligation of a biotinylated RNA linker, 5. proximity ligation under an extremely dilute condition, 6. RNA purification and reverse transcription, 7. biotin pull-down. 8. construction of sequencing library.
4.3 Computational methods of RNA-HiC-tools

4.3.1 Computational pipeline for RNA Hi-C data

RNA-HiC-tools is a package of command-line tools for analyses of RNA Hi-C data. It is written in Python and R and is version controlled by GitHub. The full documentation is at \url{http://systemsbio.ucsd.edu/RNA-Hi-C}. The pipeline takes raw pair-end sequencing reads from a mixture of different samples as input. Each input read pair contains a 4-nt sample barcode together with a 6-nt random barcode at the 5’ end of read1 (Figure 4.2A). The sequences of 4-nt sample barcodes for multiplexed sequencing and the sequence of RNA linker (purple lines in Figure 4.2C) are also necessary for the computational pipeline. The main outputs include: 1. a parsed cDNA library, including the list of chimeric cDNAs in the form of “RNA1-Linker-RNA2” (Figure 4.2C), 2. the genomic locations of RNA1 and RNA2 of every chimeric cDNA (Figure 4.2D), 3. interacting RNA pairs inferred from statistical enrichment of chimeric cDNAs (Figure 4.2E).

The pipeline contains following major computational steps.
Figure 4.2: Computational pipeline for analysis of RNA Hi-C data. From the raw input paired-end sequencing reads, PCR duplicates were first removed (Step 1, panel A. ‘N’: random barcode; ‘B’: sample barcode), and then multiplexed samples were separated based on the 4-nt sample barcodes (Step 2, panel A). Later, the cDNA sequences between two illumina sequencing adapters (P5 and P7) were reconstructed based on the sequences of read1 and read2 (Step 3, panel B). Then, with the reconstructed cDNA fragments and the pre-designed linker sequence, different fragment forms could be categorized (“BarcodeOnly”, “RNA1-RNA2”, “single RNA”, “LinkerOnly”, “RNA1-Linker”, “linker-RNA2” and “RNA1-Linker-RNA2”) and the desired “RNA1-linker-RNA2” form could be selected (Step 4, panel C). After that, we split RNA1/RNA2 from chimeric fragments and mapped them back to the genome. Only pairs with both RNA1 and RNA2 mappable were selected (Step 5, panel D). Finally, with all mappable chimeric fragments, strong RNA-RNA interactions could be identified based on a hypergeometric test together with their interaction sites (Step 6, panel E).
Removing PCR duplicates

High-throughput sequencing of cDNA libraries prepared in this study was performed on an Illumina HiSeq 2500 (run length 100 nt). The cDNA libraries contain a 4-nt sample barcode plus a 6-nt random barcode at 5’ end of read1, which allows multiplexing and the removal of PCR amplicons, respectively. A read pair is classified as a PCR duplicate of another read pair and is therefore discarded if the two read pairs had identical sequences at both read ends and contained identical barcodes (10nt). The tool ‘remove_dup_Pe.py’ provides this function, and generates a fastq/fasta file containing the non-duplicated reads, and reports the number of duplicates been removed.

Assigning multiplexed sequencing reads into corresponding experimental samples

The tool ‘split_library_pairend.py’ assigns each pair-end read into a sample by matching the sample barcode in each read with those in the list of sample barcodes (a user input text file), generates a fastq/fasta file for the reads assigned to each sample, as well as a fastq/fasta file for the unassigned reads. We local-aligned the 4-nt sample barcodes against the forward-end of each sequenced pairs while allowing one either mismatches or indels. In doing so, we allowed skipping 3 or 4 nucleotides in the forward end up until we found the matching.

Recovering the cDNAs in the sequencing library

This step identifies the overlapping regions of the two ends of every read pair, if any. It also recovers the entire sequences of the cDNAs in the sequencing library, whenever possible. The cDNA fragments are defined as the cDNA sequences between two illumina sequencing adapters P5 and P7 within the library (Figure 4.2B).

1. If an overlap existed, this read pair was sequenced from a cDNA between 100bp and 200bp (not counting the lengths of P5 and P7) (Type 2, Figure 4.3). In this case the entire sequence of the cDNA was completely covered by concatenating the forward read (Read1) with the non-overlapping region of the reverse read (Read2).
(a) If the cDNA was shorter than 100bp, we verified the presence of the P5 and the P7 primers at the two ends of the cDNA (Type 1 Figure 4.3). The ones did not contain P5 or P7 were discarded (Type 4).

2. Without an overlap, the read pair was sequenced from a cDNA longer than 200bp, whose sequence only be partially recovered (Type 3, Figure 4.3).

This function is achieved by ‘recoverFragment.py’, which uses local alignment to identify the overlapping regions. When the overlap was small (15bp or less) compared to read length (100bp on each end), local alignment could be insensitive. To overcome this insensitivity, recoverFragment.py collects the read pairs without identifiable overlaps after the first alignment (ALIGN1, Figure S-RECOV), truncates each read into one third of its length (retaining 33bp at the 3’ end of each read), and repeats local alignment (ALIGN4). The presence of P5 and P7 sequences at the end of two reads for Type 1 fragments are determined by (ALIGN2-3).
Figure 4.3: Workflow for reconstructing fragments. Local alignments could help us to identify the overlap between two read ends within a pair and reconstruct the whole chimeric fragments if the lengths of the fragments are less than twice the length of each end. We could also determine whether the fragments were longer than twice the sizes of each end. We used results of four different local alignments (ALIGN1-4) to reconstruct the fragments and assign them into different types. Each alignment was expressed as “local-align (seq1,seq2) {M,m,o,e}”. Here ‘seq1’ and ‘seq2’ were two input sequences and ‘M’, ‘m’, ‘o’, ‘e’ were score parameters for match, mismatch, open-gap and extend-gap penalties. The output of each alignment ‘X’ were the alignment score (ScoreX), the beginning and end points of alignment in first sequence (BeginPos1_X, EndPos1_X) and second sequence (BeginPos2_X, EndPos2_X). Type1 fragments were the original cDNA fragments with length less than 100bp. Type2 were the original fragments with lengths from 100bp to 200bp. Type3 were the fragments with lengths larger than 200bp. Type4 contained complete overlapping read-ends, which indicated that the original cDNA fragments should be shorter than 100bp. However, the P5 and P7 sequences were not observed in Type4 fragments.

Parsing the chimeric cDNAs

The next step in our analysis pipeline is to locate the position of the linker in the reconstructed fragments and for any true chimeric transcripts with the linker inserted in the middle of the fragments, split them into 2 interacting parts. In order to achieve that, we local-aligned the linker sequence against the whole fragments. Based on whether we could find the linker and where
the linker was in the fragment, we divided the fragments into six categories (Figure 4.2C):

1. Fragments without RNA linker
   
   (a) BarcodeOnly: These fragments contained only the 6-nt random barcode and the sample barcode. These fragments were contaminating RT primers. Since the RT primers contained every features that could allow them to be converted into cDNA library, ineffective removal of them during the size selection by gel purification step would lead to these fragments presenting in the sequencing reads.

   (b) Single RNA and RNA1-RNA2: Although there was cDNA insert, the linker was not present. These fragments were either from linker-absent fragments binding non-specifically to streptavidin beads at the streptavidin-biotin pulldown step or truncated fragments that were produced due to the stopping of Reverse Transcriptase at the crosslinking sites before it could reach the linker in the middle.

2. Fragments with RNA linker
   
   (a) RNA1-Linker-RNA2: These were from desirable chimeric RNAs that had the linker in the middle of the fragments.

   (b) linker-RNA2: These were fragments from the RNAs that had the linker ligated to the 5’-end, the step before proximity ligation. However, these linker-containing RNAs failed to be proximity-ligated to generate chimeric RNA transcripts. Although these fragments were undesirable and we tried to remove them by T7 Exonuclease. However, the efficiency of the purification step was still low. Therefore, a lot of these fragments were observed in the sequencing reads.

   (c) RNA1-linker: These were fragments from the RNAs that had the linker ligated to the 3’-end. The RNA linker didn’t have the 5’ Phosphate group, so theoretically, it shouldn’t be ligated to the 3’ end of RNAs. However, a trace amount of contaminating T4 PNK, which phosphorylated the 5’-end of all nucleic acids, added
a 5’-Phosphate to the linker and allowed them to be ligated to the 3’-end of RNAs as well. Another side-effect of these 5’-phosphate linkers was that they could concatenate and form sequences with many copies of the linkers.

(d) LinkerOnly: Since the linker contained the biotin group, they were also pulled downed by streptavidin. So if they were not effectively removed by size selection or they formed sequences of many copies concatenated together as explained above, they would be present in the sequencing reads as well.

We focused our analysis on the “RNA1-Linker-RNA2” class and split the fragments into two interacting parts (RNA1 and RNA2) that flank the linker. For fragments that were longer than 200bp, we were not able to recover them so most of them would be considered as “RNA1-Linker-RNA2” configuration if both of the two read-ends were still longer than 15bp (otherwise could be difficult for mapping) after trimming off the potential linker sequences at each end (Figure 4.2C, right panel).

Mapping to the genome

Hereafter, all analyses were based on the “RNA1-Linker-RNA2” type of read pairs. First, any cDNA containing less than 15bp on either the RNA1 or RNA2 side of linker was discarded, because it is unlikely to uniquely map a 15bp or less sequence to the genome in the mapping step. The two RNA fragments on each side of the linker (RNA1 and RNA2) were separately mapped to the mouse genome mm9/NCBI37 using Bowtie version 0.12.7. This step was implemented in ‘Stitch-seq.Aligner.py’. We allowed one mismatch in the first 15 base pairs of the read and the maximum permitted total of quality values at all mismatched read positions throughout the entire alignment to be 200 (Bowtie parameters -best -n 1 -l 15 -e 200 -p 9 -S). We also provided options to map using Bowtie2 and only select unique aligned reads. For Bowtie2 option, we utilized “-sensitive-local” mode with parameter set “-D 15 -R 2 -N 0 -L 20 -i S,1,0.75”. Bowtie2 follows “Multiseed alignment” strategy, and this mode uses seed lengths of 20bp, with no mismatch permitted per seed and interval between seeds as $1 + 0.75 \times \sqrt{100} = 8.5 \approx 9bp$. It also allows up to 15 consecutive seed extension attempt and up to 2 times of
“re-seeding”. Since Bowtie2 didn’t give us significant gains on mapping rate, we decided to keep on Bowtie for our current four samples. Only “RNA1-Linker-RNA2” chimeric fragments with both RNA1 and RNA2 mapped into genome were selected for following analyses.

Identifying interacting RNA pairs

The locations of RNA genes were retrieved from Ensembl (release 67, mouse NCBIM37), including the genes of mRNAs, lincRNAs, rRNAs, snRNAs, snoRNAs, miRNAs, misc RNAs, tRNAs, and transposons. The different genomic copies of the same transposon were considered as different genes in this analysis. The number of uniquely aligned reads (from either RNA1 or RNA2 of the RNA1-Linker-RNA2 type) was counted on every gene. Any gene with a read count smaller than 5 was filtered out. Next, the association between any two genes were tested with Fisher’s exact test. The null hypothesis was that gene A and gene B independently contributed to the sequencing reads. The alternative hypothesis was that their contributions to read counts were associated. We denote $c_A$, $c_B$ as the read counts for gene A and gene B, respectively, and $I_{A,B}$ as the read counts of co-appearance, where the two genes co-appeared on the same read pair. A Fisher’s exact test was carried out on each gene pair, with $I_{A,B}$, $c_A$, $c_B$, $c_A^{cA},c_B^{cB}$ as the test statistics, where $c_A^{cA}$ ($c_B^{cB}$) was the read counts on other genes besides gene A (gene B) on each end. Both p-values and FDRs (Benjamini-Hochberg procedure) were calculated for every gene pair. This step outputs gene pairs with FDR < 0.05 and fold-change (FC) ≥ 3. The FC was calculated as $(I_{A,B} + 0.5)/(I'_{A,B} + 0.5)$, where $I'_{A,B}$ was the co-appearing read counts in the control sample (ES-indirect). This step was implemented in ‘Select_strongInteraction_RNA.py’ which outputs strong interacting RNA pairs with information of their interaction regions, number of supporting pairs, p-value of significance, FDR and fold changes.

Identifying RNA interaction sites

We defined the RNA interaction site as a continuous RNA segment that frequently participated RNA-RNA interactions. RNA interaction sites were inferred from RNA Hi-C data as continuous RNA segments with multiple
overlapping reads and frequent co-appearance (proximity ligation) with other RNAs (Figure 4.4). First, any continuous RNA segment covered by 5 or more uniquely aligned reads was identified as a candidate interaction site. Second, the association between any two candidate sites were tested with Fisher’s exact test. The null hypothesis was that candidate sites A and gene B independently contributed to the sequencing reads. The alternative hypothesis was that their contributions to read counts were associated. We denote $c_A, c_B$ as the read counts for candidate sites A and B, respectively, and $I_{A,B}$ as the read counts of co-appearance, where the two sites co-appeared on the same read pair. A Fisher’s exact test was carried out on each site pair, with $I_{A,B}, c_A, c_B, c_Ac_B$ as the test statistics, where $c_A (c_B)$ was the read counts on other candidate sites besides A (B). Both p-values and FDRs (Benjamini-Hochberg procedure) were calculated for every pair of candidate sites. The candidate sites exhibiting significant associations (FDR < 0.05) were regarded as RNA interaction sites. This step was automated in Select_strongInteraction_pp.py which outputs the identified RNA interaction sites.

![Figure 4.4: A statistical test based Approach to identify strong RNA-RNA interactions.](image)

**Visualization of RNA-RNA interactions**

We developed two different types of visualizations for the RNA-RNA interactions that have been discovered.

The tool ‘Plot_interaction.py’ was developed for visualizing RNA inter-
action sites and the ligation events of these sites. Given any two genomic regions as input, for example the locations of two genes, this tool displays all the supporting read pairs in the form of RNA1-Linker-RNA2, where RNA1 and RNA2 were aligned to each of the two genomic locations. The linker of each RNA pair was plotted as well. This tool also plots RNA interaction sites in the input regions, if any, as well as the identified interactions between these sites.

The tool ‘Plot_Circos.R’ provides a global view of the RNA-RNA interactome. It plots the entire genome as a circle, and any RNA-RNA interaction as a curve between the two genes. The interactions involving different types of RNAs are coded with different colors. The densities of RNA1 and RNA2 read fragments are displayed along with every chromosome as inner circles.

Both two types of visualizations could be generated automatically using intermediate output from the computational pipeline as the input. The detailed description of how to use them are presented here: http://systemsbio.ucsd.edu/RNA-Hi-C/Visualization.html.

4.3.2 Simulation study on synthetic paired-end RNA Hi-C data

We first simulated 1,000,000 pairs of synthetic paired-end RNA Hi-C data with the parameters to mimic the real data we obtained. The procedure to generate each synthetic read pairs is as following:

1. Assign a sample barcode and generate random barcode

2. Randomly choose a fragment class from: [“linkerOnly”, “NoLinker” (“singleRNA” and “RNA1-RNA2”), “RNA1-linker”, “linker-RNA2”, “RNA1-linker-RNA2”] with a given discrete probability distribution. (probability: [0.1,0.3,0.1,0.3,0.2], similar as the portions in real data)

3. If the class contains linker, randomly choose linker numbers from [1, 2].

4. Generate sequences for RNA1 and RNA2. The steps for either of these two are the same as following:

   (a) Choose the length of RNA sequence \( l \sim Unif(15, 150) \)
Choose a RNA type from [“miRNA”, “mRNA”, “lincRNA”, “snoRNA”, “snRNA”, “tRNA”] based on a discrete probability distribution (similar as portions in real data):

i. If length \( l < 50 \), distribution: \([0.2, 0.2, 0.1, 0.2, 0.2, 0.1]\)

ii. Else, distribution: \([0.05, 0.4, 0.2, 0.2, 0.1, 0.05]\)

Randomly pick a RNA with chosen RNA type from the RNA annotation database

(d) Randomly pick a sequence with length \( l \) within that RNA.

5. Concatenate the whole fragment based on sequences from step 1, 3, 4.

6. Generate two ends of read pair based on cDNA fragment from step 5.

7. Add P5 and P7 sequences if fragment length is smaller than read length.

8. Add mutations for each read based on sequencing error rate 0.01 [90].

The information of fragment length, class, and genomic locations of two RNAs for each read pair was stored in file for comparison purpose. After generating the synthetic paired-end sequencing data, we ran our whole computational pipeline to create intermediate and final results of fragment lengths, classes, genomic locations for the chimeric RNAs and interactions. These results were compared with the actual information of synthetic data to evaluate the sensitivity and specificity of RNA-HiC-tools.

4.3.3 Other computational methods.

Evaluation of overlap between strong interactions in different samples

The Overlap between different sets of strong interactions were determined through two methods: (I) by genomic locations of the pairs; (II) by annotated gene pairs. In the first method, two interactions were considered as overlapped with each other if and only if both of two partners between interactions were close to each other within the genome with distance less than 100bp. In the second method, if the chromosome names, and annotated RNA
names were the same for both partners between two interactions, those two interactions were considered to be overlapped.

For both of two methods, we did 100 permutations to determine the statistical significance of the number of overlaps. In each permutation, we fixed one interaction set and shuffle the partners of the other interaction set. Then we determined the number of overlaps between the fixed interaction set and the other shuffled interaction set as the result for one permutation. The p-value can be calculated by comparing the real number of overlaps and a set of permuted numbers of overlaps.

**Generation of RNA-RNA interaction network**

RNA-RNA interaction pairs generated from step 6 of computational pipeline were converted to network format and imported into Cytoscape 3.1.0 for analysis and visualization. Nodes in network were colored according to their annotation and arranged for better display. Degrees of all nodes were obtained in Cytoscape and log-log plot for degree v.s. node proportion were generated in R. Filtered network were generated by removing all snRNAs, snoRNAs and tRNAs.

**Binding energies between RNA interaction sites**

The hybridization energies and folding between two RNA interaction sites were achieved by “DuplexFold” program from RNAstructure software version 5.6 [91]. The base-parings between two interaction sites were determined by MiRanda software [92].

**Conservation levels of RNA interaction sites**

We used PhyloP conservation scores[93] to evaluate conservation level across RNAs.

Figure 4.13A: For each end (RNA1 or RNA2) of the chimeric RNA pairs, we identified a 1000 bp region on the genome from 500 bp upstream of the ligation site to 500 bp downstream. We also randomly selected a region with identical length on the same chromosome as control. Then the averaged PhyloP scores across all the RNAs and controls were calculated respectively.
on each nucleotide in the region. We used the merged ES-1 and ES-2 chimeric RNA pairs to generate the general conservation level distribution around interaction sites.

Figure 4.13B: To explore conservation levels of RNA interaction sites of different types, we calculated average PhyloP score per nucleotide for each interaction site and generated boxplots for different annotation categories. To compare with genomic average PhyloP scores for different annotation categories, we randomly generated 200,000 genomic sites with same size mean and standard deviation as real interaction sites, and annotated them with our annotation module. The PhyloP score distributions were also plotted for different categories the same as real interaction sites. We then used two-sample t-test to evaluate the PhyloP score differences between real interaction sites and random sites.

Detection of other types of RNAs functioning as miRNAs

We selected candidates of interaction sites (clusters) that might function as miRNAs by incorporating the small RNA-seq (GSM945907) [47] and AGO-CLIP-seq (GSM622570) [94] data. In doing so, we calculated the RPKMs of small RNA-seq and CLIP-seq data for interaction sites of each interaction. Here RPKM was the normalized read intensity of small RNA-seq and CLIP-seq data scaled by interaction site length (per kilobase) and total number of mapped reads (per millions). We first selected interaction sites with both small RNA-seq signals (RPKM > 0) and CLIP-seq signals (RPKM > 0). Then from these candidates interaction sites, we further selected those that had interactions with exon or 3'-UTR regions of mRNAs which also had CLIP-seq signals. For example, for interaction sites on snoRNAs, we selected all snoRNA-mRNA interactions with small RNA-seq signals on the snoRNA side and CLIP-seq signals on both sides.

After selecting these candidate interaction sites, we further checked their base-pairing with their mRNA targets supported by “RNA-linker-RNA2” fragments. We collected these corresponding fragments (suppose N fragments for each interaction) and calculated the hybridization energies for all RNA1/RNA2 pairs and also the random hybridization energies by shuffling sequences. From that, we determined whether the hybridization energies for each of the selected interactions were significantly lower than random based
on these N pairs of hybridization energies from real interaction and random shuffling with a Wilcoxon signed-rank test. With p-value < 0.05, we refined the list of candidate interaction sites that might indicate a specific type of RNAs which could function as miRNA to target mRNA (See table 4.8).

Detection of self-ligation within the same RNA molecule

From the mapped pairs of fragments, we first removed those that were indicative of non-chimeric fragments if they could be considered as concordant pairs considering the aligned locations, strands and RNA annotations. The criterion to call a mapped pair as intact (non-chimeric) RNA fragment was as follows:

1. The two reads within the pair were mapped to the same chromosome;
2. Whole reads were mapped to the genome for both reads (90bp for read1, since index was removed at the beginning, and 100bp for read2), no signal of linker sequences;
3. The strands of mapped read1 and read2 were different and the mapped genomic locations for read1 and read2 were no longer than 2000bp away from each other.
4. the read mapped to plus strand has smaller coordinates than the read mapped to minus strand in the genome within the pair.

After removing these intact (non-chimeric) fragments, from the mapped chimeric pairs of fragments, we selected those that had the same RNA annotation and mapped to same chromosome for both sides as candidates for self-ligation within the same RNA molecules.

RNA folding and secondary structure prediction

Structure information of some snoRNA was downloaded from fRNAdb [95] in dot format and converted into graphs using command line version of VARNA applet version 3.9 [96]. For the RNAs without known structure information, we predicted their secondary structures using “Fold” program in RNAstructure software version 5.6 [91] based on RNA sequences. The information of
RNase I digested site distributions were utilized as single strand offset with “-sso” option to refine the prediction of the secondary structure.

4.4 Simulation study on synthetic paired-end RNA Hi-C data

The way to simulate 1,000,000 pairs of paired synthetic paired-end RNA Hi-C data was described in Method section. By comparing the actual cDNA fragment information from synthetic data and the results generated from RNA-HiC-tools, we could assess the performances of different modules in RNA-HiC-tools.

4.4.1 Evaluation of fragment length prediction

The “recoverFragment” function recovers the cDNA fragments from paired-end reads and classified them into four types based on fragment lengths: Type1 (< 100bp); Type2 (100 ∼ 200bp); Type3 (> 200bp); Type4 (Weird) (Figure 4.3). The comparison between recovered fragment types and actual fragment types showed that the types of most fragments (99.10%) could be accurately recovered with high sensitivity and specificity (Table 4.1). There were about 0.58% of all cDNA fragments which were actually less than 200bp but recovered as larger than 200bp. The reason was that for the fragments with lengths slightly less than 200bp (overlap of two read ends smaller than 5bp but larger than 0bp), our local alignment based recovery method could not detect the overlap, and the fragments were assigned as Type3.

If two read ends had detectable overlaps, we were able to recover the whole fragment and predict the whole fragment length. The predicted fragment lengths for recovered Type1&2 fragments were identical with the actual fragment lengths for most of the fragment (Figure 4.5A).

4.4.2 Evaluation of fragment class prediction

The “split_partner.py” function divides all the Type1-3 fragments into different classes (Figure 4.2C). Here “NoLinker” class includes both “SingleRNA”
Table 4.1: Evaluation of fragment length prediction. (Top) The number of pairs for each category comparing the synthetic data and the prediction. The overall accuracy is 99.10%. (Bottom) The sensitivity and specificity for the prediction of each fragment length types.

<table>
<thead>
<tr>
<th></th>
<th>Type1_predict</th>
<th>Type2_predict</th>
<th>Type3_predict</th>
<th>Type4_predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type1_synthetic</td>
<td>312,411</td>
<td>24</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Type2_synthetic</td>
<td>65</td>
<td>480,835</td>
<td>5,750</td>
<td>898</td>
</tr>
<tr>
<td>Type3_synthetic</td>
<td>126</td>
<td>1,322</td>
<td>197,716</td>
<td>853</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type1</td>
<td>99.99%</td>
<td>99.97%</td>
</tr>
<tr>
<td>Type2</td>
<td>98.62%</td>
<td>99.73%</td>
</tr>
<tr>
<td>Type3</td>
<td>98.84%</td>
<td>99.28%</td>
</tr>
</tbody>
</table>

and “RNA1-RNA2” classes. The overall accuracy is 96.62% (Table 4.2). Both the sensitivity (99.89%) and specificity (95.82%) for the identification of desired “RNA1-linker-RNA2” class were high. The major source of false positives (3.34% of all pairs) were from actual “SingleRNA” and “RNA1-RNA2” fragments which were longer than 200bp. If these fragments were actually “RNA1-RNA2”, then they could also be contributed as chimeric RNA pairs. Otherwise, if they were actually “SingleRNA”, we had a step to remove these false positives after alignment and annotation steps (see section “Detection of self-ligation within the same RNA molecule”).

Table 4.2: Evaluation of fragment class prediction. The number of pairs for each category comparing the actual fragment classes and the predict fragment classes. All the rows represent the predicted classes (P) and the columns represent the actual classes from synthetic data (S). The overall accuracy is 96.62%. The “R1-link-R2” here represents the “RNA1-linker-RNA2” class.

<table>
<thead>
<tr>
<th></th>
<th>NoLinker</th>
<th>LinkerOnly</th>
<th>R1-linker</th>
<th>Linker-R2</th>
<th>R1-link-R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoLinker</td>
<td>266,554</td>
<td>10</td>
<td>—</td>
<td>—</td>
<td>33,402</td>
</tr>
<tr>
<td>LinkerOnly</td>
<td>—</td>
<td>100,230</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>R1-linker</td>
<td>24</td>
<td>25</td>
<td>100,267</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Linker-R2</td>
<td>50</td>
<td>58</td>
<td>—</td>
<td>299,180</td>
<td>—</td>
</tr>
<tr>
<td>R1-link-R2</td>
<td>57</td>
<td>116</td>
<td>24</td>
<td>22</td>
<td>199,981</td>
</tr>
</tbody>
</table>
4.4.3 Evaluation of interaction prediction

After mapping step, the predicted “RNA1-Linker-RNA2” chimeric fragments with both RNA1 and RNA2 uniquely mapped into genome were collected and these interaction pairs were compared with actual interaction pairs from synthetic data. 131571 out of the 200200 actual chimeric RNA pairs with “RNA1-linker-RNA2” configuration could be identified after mapping the annotation step (65.72% sensitivity and 92.57% specificity). For each interaction type, we also calculated the identification sensitivity and specificity. The sensitivities were lower for transposons and snRNAs associated interactions (Figure 4.5C). This was because many of them have different gene copies for one gene and could not be uniquely aligned to the genome. Interactions involving mRNAs, lincRNAs and snoRNAs could be detected with higher sensitivities (Figure 4.5C). So the identified RNA interactome could be slightly biased to these interaction types. Overall, the biases on different types of interactions were not very significant.
4.5 Results on real data

We applied RNA Hi-C to generated four libraries for mouse embryonic stem (ES) cells and mouse embryonic fibroblast (MEF) cells. Three libraries were prepared from direct RNA-RNA interactions (assisted by a single protein) of primarily cytoplasmic contents of mouse ES cells (ES-1, ES-2, Table 4.3) and embryonic fibroblasts (MEF, Table 4.3). As a control, we generated the
fourth library using dual crosslink agents (formaldehyde and EGS) that form covalent bonds between both nucleotides and proteins and between proteins (ES-indirect, Table 4.3). Then each library was sequenced with an average of 63 million pair-end reads (Table 4.4). The quality of sequencing data were evaluated using some bioinformatic assessments.

4.5.1 Bioinformatic assessments of data quality

Presence of chimeric products.

We asked whether the expected chimeric products existed in the sequencing library. To do so, we put the reads into five categories: single RNA, RNA1-RNA2 (RNA ligation without linker), RNA1-linker, linker-RNA2, RNA1-linker-RNA2 based on the existence of linker sequences and their relative locations within the fragments (See Methods, based on input chimeric cD-NAs shorter than 200nt; barcode only and linker only reads are removed). And then we analyzed their proportions (Figure 4.6). Of the four experimental samples, 12% to 31% of the RNA sequence reads belonged to the RNA1-linker-RNA2 configuration, which was much larger than the proportions of chimeric products from CLASH-seq experiment [29]. This result confirmed the formation and enrichment of the expected chimeric products.

We then evaluated the efficiencies of the two ligation steps: Step 4, ligation of the RNA linkers to the 5’ end of RNAs, and Step 5, ligation of the linkers to the 3’ end of other RNAs (Figure 4.1). Of the four samples, 42.0% to 82.6% of sequence reads belonged to link-RNA2 and RNA1-linker-RNA2, indicating a reasonable efficiency of linker ligation to the 5’ end. Second, 20.0% to 52.6% of 5’ ligated linkers belonged to RNA1-linker-RNA2 (RNA1-linker-RNA2/(linker-RNA2+ RNA1-linker-RNA2)), indicating a significant proportion of the linker-attached RNAs participated proximity ligation (Table 4.4). We recognize that these sequence based estimates were not precisely estimating the efficiency of any ligation reaction, but rather the “observed” outcomes of these reactions in conjunction with other biochemical reactions, such as biotin selection. Finally, each library gave on average about 15.1 million read pairs of the desired chimeric form (RNA1-linker-RNA2). After mapping them into mouse genome (mm9), we obtained an average of 2.2 mil-
lion read pairs with both the RNA1 end and the RNA2 end being mapped (Table 4.5). These read pairs were used for following analyses.

**Table 4.3:** Biological and technical differences of the four samples. 254nm UV specifically crosslinks nucleotides with proteins without generating protein-protein crosslinks, thus only allowing for the RNAs bound to the same protein to be ligated.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>ES-1</th>
<th>ES-2</th>
<th>ES-indirect</th>
<th>MEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell type</td>
<td>ES cells</td>
<td>ES cells</td>
<td>ES cells</td>
<td>MEF</td>
</tr>
<tr>
<td>Crosslinking</td>
<td>254nm UV</td>
<td>254nm UV</td>
<td>Dual crosslink</td>
<td>254nm UV</td>
</tr>
<tr>
<td>RNA-protein interactions</td>
<td>Direct</td>
<td>Direct</td>
<td>Indirect</td>
<td>Direct</td>
</tr>
<tr>
<td>Protein solubilization</td>
<td>Detergents</td>
<td>Detergents</td>
<td>Sonication</td>
<td>Detergents</td>
</tr>
<tr>
<td>1st fragmentation</td>
<td>1000-2000 nt</td>
<td>1000 nt</td>
<td>1000 nt</td>
<td>300 nt</td>
</tr>
<tr>
<td>rRNA removal</td>
<td>Duplex-specific nuclease</td>
<td>Antibody based</td>
<td>Duplex-specific nuclease</td>
<td>Antibody based</td>
</tr>
<tr>
<td>Cellular compartment</td>
<td>Primarily cytoplasmic</td>
<td>Primarily cytoplasmic</td>
<td>Entire cell</td>
<td>Primarily cytoplasmic</td>
</tr>
</tbody>
</table>

**Figure 4.6:** Portions of different fragment types for four samples. The portions of different fragment types are inferred from original cDNA fragments with lengths less than 200nt.

**Comparison of RNA Hi-C libraries**

We then asked how consistent were the interaction-participating RNAs across the samples. To do so, we split each sample into two RNA-seq like datasets, one containing the read fragments at the 5’ side of the linker (RNA1) and the
Table 4.4: Evaluation of ligation efficiency. Total pairs are total numbers of read pairs from Paired-end sequencing data. Fragments are defined as the original cDNA sequences between two illumina sequencing adapters P5 and P7 within the constructed library. All the index after that are estimated from cDNA fragments with lengths less than 200bp (Type1&2 in Figure 4.3). The efficiency of first ligation is defined by the percentage of \( \frac{\text{linker-RNA2} + \text{RNA1-linker-RNA2}}{\text{five categories}} \). The efficiency of the second ligation is defined as the percentage of \( \frac{\text{RNA1-linker-RNA2}}{\text{linker-RNA2} + \text{RNA1-linker-RNA2}} \).

<table>
<thead>
<tr>
<th>Name</th>
<th>ES-1</th>
<th>ES-2</th>
<th>ES-indirect</th>
<th>MEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample barcode</td>
<td>ACCT</td>
<td>GGCG</td>
<td>AATG</td>
<td>GGCG</td>
</tr>
<tr>
<td># of total pairs</td>
<td>45,702,794</td>
<td>49,316,127</td>
<td>74,009,386</td>
<td>83,083,324</td>
</tr>
<tr>
<td># of total fragments &lt; 200nt</td>
<td>26,428,442</td>
<td>41,054,803</td>
<td>47,824,454</td>
<td>56,000,615</td>
</tr>
<tr>
<td>% of linker sequence</td>
<td>63.40%</td>
<td>73.35%</td>
<td>61.06%</td>
<td>93.39%</td>
</tr>
<tr>
<td>First ligation efficiency</td>
<td>58.30%</td>
<td>55.67%</td>
<td>41.99%</td>
<td>82.60%</td>
</tr>
<tr>
<td>Second ligation efficiency</td>
<td>52.61%</td>
<td>39.11%</td>
<td>29.53%</td>
<td>20.04%</td>
</tr>
</tbody>
</table>

other containing all fragments at the 3’ end of the linker (RNA2). We borrowed the FPKM metric to estimate the amount of interaction-participating RNAs of each dataset (ES-1_RNA1, etc.), and then clustered the datasets (Figure 4.7). Despite their technical differences on fragmentation size and rRNA removal (Table 4.3), the four datasets from UV crosslinked ES cells clustered together, next to the dual crosslinked ES cell datasets (ES-indirect_RNA1 and ES-indirect_RNA2). Furthermore, the dual crosslinked ES cell sample was prepared such that whole cell proteins were effectively included while the nature of the lysis procedure used in the UV-crosslinked samples rendered nuclear RNA-binding proteins less accessible. The MEF datasets formed their own cluster, despite that its fragmentation size and rRNA removal procedure was identical to the second ES cell sample (ES-2, Table 4.3).
Figure 4.7: Comparison of RNA Hi-C libraries. (A-B) The read fragment at the 5’ end (RNA1) and the 3’ end (RNA2) of the linker were separately analyzed as two RNA-seq experiments. Scatter plots of read count distribution (FPKM) of all known RNAs between ES-1 and ES-2. R: Pearson correlation. S: Spearman correlation. (C) Hierarchical clustering of RPKMs of each sample.

Next, we examined the overlaps of identified interacting RNA pairs between the libraries. The lists of strong interactions between different RNAs were generated for all four samples as described in Methods (Table 4.5). Then, we calculated the overlaps between different pairs of samples. Two interactions in two different samples were considered as the same interaction when both sides of the interactions had the same RNA annotations. From a permutation test (by shuffling the interaction partners), we found that ES-1, ES-2, and ES-indirect exhibited strongest overlaps (permutation p-values < $2.2 \times 10^{-16}$), whereas MEF exhibited fewer overlaps with ES-indirect, and even fewer overlaps with ES-1 and ES-2 (Table 4.6). For example, an interaction between the 3’ UTR of Trim25 and SNORA1 was supported by 24 and 22 pair-end reads in ES-1 and ES-2 libraries, respectively, but not supported by any reads in ES-indirect or MEF libraries (Figure 4.8). The predicted “interacting site” on Trim25 by the overlapping RNA Hi-C reads was associated with the Argonaute protein in ES cells, so was the SNORA1 RNA (AGO CLIP-seq, Figure 4.8). All the above results suggested that data from the first two ES cell samples (ES-1 and ES-2) were quite similar to each other. Thus, we decided to merge ES-1 and ES-2 libraries for analysis of RNA interactome in ES cells.
Table 4.5: Result summary of mapped pairs and strong interactions in four samples. Total pairs are total numbers of read pairs from Paired-end sequencing data. The mapped pairs are the pairs within the “RNA1-Linker-RNA2” group with both RNA1 and RNA2 mapped into genome. The clusters are generated based on genomic locations for all the mapped reads in RNA1 and RNA2 separately. Interactions are identified between clusters in RNA1 and RNA2. Interactions removing rRNA are interactions after removing all interactions involving rRNAs in either part. The intra-RNA interactions are defined if both sides of the interactions have the same RNA annotation. The inter-RNA interactions are interactions removing rRNA but other than intra-RNA interactions.

<table>
<thead>
<tr>
<th>Name</th>
<th>ES-1</th>
<th>ES-2</th>
<th>ES-indirect</th>
<th>MEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample barcode</td>
<td>ACCT</td>
<td>GGCG</td>
<td>AATG</td>
<td>GGCG</td>
</tr>
<tr>
<td># of total pairs</td>
<td>45,702,794</td>
<td>49,316,127</td>
<td>74,009,386</td>
<td>83,083,324</td>
</tr>
<tr>
<td># of RNA1-linker-RNA2 pairs</td>
<td>13,848,413</td>
<td>9,553,722</td>
<td>19,554,316</td>
<td>17,616,980</td>
</tr>
<tr>
<td># of mapped pairs</td>
<td>3,230,403</td>
<td>1,312,314</td>
<td>3,302,146</td>
<td>1,110,098</td>
</tr>
<tr>
<td># of clusters in RNA1</td>
<td>17,130</td>
<td>14,005</td>
<td>16,674</td>
<td>8,064</td>
</tr>
<tr>
<td># of clusters in RNA2</td>
<td>13,419</td>
<td>5,001</td>
<td>14,694</td>
<td>5,953</td>
</tr>
<tr>
<td># of interactions</td>
<td>51,635</td>
<td>20,248</td>
<td>20,013</td>
<td>16,100</td>
</tr>
<tr>
<td># of interactions removing rRNA</td>
<td>21,787</td>
<td>3,562</td>
<td>10,930</td>
<td>7,254</td>
</tr>
<tr>
<td># of inter-RNA inter-</td>
<td>21,249</td>
<td>3,328</td>
<td>8,844</td>
<td>7,168</td>
</tr>
<tr>
<td>actions</td>
<td>538</td>
<td>234</td>
<td>2,086</td>
<td>86</td>
</tr>
</tbody>
</table>


Table 4.6: Overlaps of strong inter-RNA interactions among different samples and replicates. Two interactions are considered to be the same in different samples or replicates if the chromosome names, annotated RNA names are the same for both RNA1 and RNA2. Permutations are done in 100 times by shuffling the identified interacting clusters in both samples (see method). The total number of inter-RNA interactions in these four samples are: 21,249, 3,328, 8,844, 7,168, respectively for ES-1, ES-2, ES-indirect and MEF.

<table>
<thead>
<tr>
<th>Sample A</th>
<th>Sample B</th>
<th>num of A overlapped with B</th>
<th>num of B overlapped with A</th>
<th>Permutation p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES-1</td>
<td>ES-2</td>
<td>1,531</td>
<td>992</td>
<td>&lt; 2.2 × 10^{−16}</td>
</tr>
<tr>
<td>ES-1</td>
<td>ES-indirect</td>
<td>796</td>
<td>1,122</td>
<td>&lt; 2.2 × 10^{−16}</td>
</tr>
<tr>
<td>ES-indirect</td>
<td>ES-2</td>
<td>287</td>
<td>163</td>
<td>&lt; 2.2 × 10^{−16}</td>
</tr>
<tr>
<td>ES-1</td>
<td>MEF</td>
<td>1,014</td>
<td>838</td>
<td>0.41037</td>
</tr>
<tr>
<td>ES-2</td>
<td>MEF</td>
<td>466</td>
<td>595</td>
<td>0.00054</td>
</tr>
<tr>
<td>ES-indirect</td>
<td>MEF</td>
<td>733</td>
<td>968</td>
<td>4.05 × 10^{−9}</td>
</tr>
</tbody>
</table>

Figure 4.8: An example for the interaction between snoRNA SNORA1 and 3’-UTR region of Trim25 gene. This interaction was detected in both of first two ES samples, but not in the third ES sample and the MEF sample. There are AGO CLIP-seq signals on both sides of interaction based on published data in same cell type (mouse ES cells).
4.5.2 Overview of ES cell RNA interactome

We analyzed the ES cell RNA interactome based on the merged data of ES-1 and ES-2, which included 4.54 million unambiguous and non-redundant chimeric RNAs. Since the ES-indirect sample used dual crosslinking method, which also included indirect RNA-RNA interactions for RNAs bound to different interacting proteins, we decided to use this sample as control to identify only direct RNA-RNA interactions where only RNAs bound to same protein were ligated. We identified 46,780 inter-RNA interactions under the criteria of FDR $<$ 0.05 (p-values from hypergeometric test and FDR from Benjamini-Hochberg procedure) and each interaction being supported by fold change between test and control no less than 3 (See Methods). The most abundant types of interactions were between an mRNA and a snoRNA (27,375 unique interactions), two mRNAs (7,076 interactions), an mRNA and a tRNA (2,894 interactions) and two snoRNAs (2,781 interactions). Among other types of detected interactions were mRNA-snRNA, lincRNA-mRNA, and miRNA-mRNA interactions (Figure 4.9A).
To see if the detected interactions were biologically meaningful, we constructed an RNA-RNA interaction network based on the identified inter-RNA interactions. Each RNA was linked with another RNA if they showed interactions from our merged ES data. The ES cell RNA-RNA interactome was a scale-free network, with a degree distribution conformed to power law ($P(k) \sim k^{-\gamma}$, $\gamma = 3$, Figure 4.9B) [97]. To see whether the scale-free property was driven by a small number of highly connected snoRNAs, snRNAs,
and tRNAs, we removed them from the network. The interactions composed only of mRNAs, lincRNAs, miRNAs, pseudogeneRNAs, and antisenseRNAs remained scale-free. A number of mRNAs, pseudogene RNAs, and lincRNAs emerged as hubs (Figure 4.9C). The largest mRNA hub was Suv420h2, which interacted with 21 mRNAs and 2 lincRNAs. The largest lincRNA hub was Malat1, which interacted 4 mRNAs, including an mRNA hub of Slc2a3. The scale-free property of RNA-RNA interactome network confirmed that the identified inter-RNA interactions were not random.

4.5.3 Enriched RNA interaction sites involved in RNA-RNA interactions

We asked whether the identified interactions were enriched in certain parts of the transcriptome. The read fragments of the hub RNAs appeared to concentrate at specific locations on the transcripts (Figure 4.10), and majorities (83.05%) of the interacting RNAs exhibited overlapping RNA Hi-C reads. We called “clusters” of overlapping read fragments from each end of the chimeric RNAs (Figure 4.4, see Methods, clusters were called when supported by at least 5 overlapping reads), similar to calling peaks from ChIP-seq data. These clusters represented the sites on transcripts that spatially close to other RNAs. In analogy of protein interaction domains, we termed these clusters RNA interaction sites. These interaction sites appeared not only on mRNAs, lincRNAs, miRNAs, but also on pseudogene RNAs and transposon RNAs, especially the L1 LINEs and the ERVK and MaLR LTRs (Table 4.7), indicative of their interactions of other RNAs [98, 99]. Hereafter, we identified interactions between different “clusters” based on mapped read pairs.
Figure 4.10: Interaction sites and their interaction partners on the transcribed region of *Eef1a1* gene. The interaction sites within *Eef1a1* gene (black blocks) are aligned to some clusters within different exon regions of this gene. Two clusters on top of exon 2 and 3 are plotted in detail with their interaction partners represented as grey blocks. These grey blocks are clusters within transcribed regions of other genes and connected with black blocks by the linker sequences.

Base pairing in RNA-RNA interactions

In RNA metabolism, many RNAs interacted with other RNAs through base-pairing to find their specific targets, such as snRNAs’ function for pre-mRNA splicing [100], snoRNAs’ function for ribosome synthesis [101] and miRNAs’ function for regulation of mRNA stability [19]. We asked how many types of RNA-RNA interactions preferentially used complementary RNA bases. To do that, we quantified the hybridization energy of the two fragments (RNA1, RNA2) of each chimeric RNA, and averaged the hybridization energies of the chimeric RNAs representing the same pair of interacting RNAs. We then compared the hybridization energies of the interacting RNAs with control RNAs generated by random shuffling of the bases. Base pairing was preferred in most types of RNA-RNA interactions, and was most pronounced in transposonRNA-mRNA, mRNA-mRNA, pseudogeneRNA-mRNA, lincRNA-mRNA, miRNA-mRNA interactions (p-values < 2.4 × 10^{-18}), but was not observed in LTR-pseudogene interactions (Figure 4.11). Examples for different types of RNA-RNA interactions with preferred base-pairing were also presented (Figure 4.12). These data suggest another layer of posttranscrip-
<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Involved genes</th>
<th>Total genes</th>
<th>Involved genomic locations</th>
<th>Total genomic locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>12,439</td>
<td>6,600</td>
<td>22,562</td>
<td>6,600</td>
<td>22,562</td>
</tr>
<tr>
<td>snoRNA</td>
<td>553</td>
<td>511</td>
<td>1,561</td>
<td>511</td>
<td>1,561</td>
</tr>
<tr>
<td>tRNA</td>
<td>365</td>
<td>57</td>
<td>60</td>
<td>337</td>
<td>4,760</td>
</tr>
<tr>
<td>lincRNA</td>
<td>363</td>
<td>243</td>
<td>2,054</td>
<td>243</td>
<td>2,054</td>
</tr>
<tr>
<td>snRNA</td>
<td>226</td>
<td>13</td>
<td>32</td>
<td>214</td>
<td>1,429</td>
</tr>
<tr>
<td>miRNA</td>
<td>27</td>
<td>25</td>
<td>1,630</td>
<td>25</td>
<td>1,630</td>
</tr>
<tr>
<td>misc_RNA</td>
<td>33</td>
<td>17</td>
<td>114</td>
<td>32</td>
<td>487</td>
</tr>
<tr>
<td>pseudogene</td>
<td>234</td>
<td>131</td>
<td>5,306</td>
<td>131</td>
<td>5,306</td>
</tr>
<tr>
<td>antisense</td>
<td>34</td>
<td>31</td>
<td>1,351</td>
<td>31</td>
<td>1,351</td>
</tr>
<tr>
<td>LINE (L1)</td>
<td>726</td>
<td>76</td>
<td>112</td>
<td>725</td>
<td>884,320</td>
</tr>
<tr>
<td>LINE (L2)</td>
<td>26</td>
<td>4</td>
<td>4</td>
<td>26</td>
<td>65,481</td>
</tr>
<tr>
<td>LTR (ERVK)</td>
<td>346</td>
<td>96</td>
<td>150</td>
<td>346</td>
<td>245,391</td>
</tr>
<tr>
<td>LTR (MaLR)</td>
<td>274</td>
<td>60</td>
<td>102</td>
<td>274</td>
<td>430,745</td>
</tr>
<tr>
<td>LTR (ERV1)</td>
<td>235</td>
<td>39</td>
<td>113</td>
<td>235</td>
<td>61,660</td>
</tr>
<tr>
<td>LTR (ERVL)</td>
<td>78</td>
<td>31</td>
<td>88</td>
<td>78</td>
<td>111,531</td>
</tr>
<tr>
<td>SINE</td>
<td>458</td>
<td>32</td>
<td>40</td>
<td>458</td>
<td>1,521,108</td>
</tr>
<tr>
<td>Novel</td>
<td>4,426</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Figure 4.11: The hybridization energies between two RNA sequences that are interacting with each other supported by chimeric RNAs for different interaction types. The chimeric RNAs which were defined as mapped pairs are used for this analysis. They were assigned into different interaction types based on RNA annotations. Interactions involving intron regions of mRNAs/lincRNAs are not included. The lowest binding energies between two paired RNA fragments were determined by “DuplexFold” program in RNAstructure package. Random controls are generated by shuffling the sequences of each participating RNAs and controlling the length of the sequence the same as data. The numbers on the bottom right are p-values based on wilcoxon signed-rank test. The groups with most significant differences between chimeric RNAs and random controls are marked with red color (p-value $< 2.2 \times 10^{-16}$). The unit of $\Delta G$ is kcal/mol.
Figure 4.12: Examples of base-complement between two directly interacting RNAs for different interaction types. (A) mRNA-mRNA; (B) lincRNA-mRNA; (C) pseudogene-mRNA; (D) mRNA-LTR; (E) LINE-mRNA; (F) mRNA-miRNA. The purple lines indicate the linker sequences and the purple numbers within the bracket give the numbers of supported chimeric RNAs. The lowest hybridization energies between two paired RNA fragments were determined by “DuplexFold” program in RNAstructure package. Random controls are generated by shuffling the sequences of each participating RNAs and controlling the length of the sequence the same as data.

Increased interspecies conservation of RNA interaction sites

Sequence conservation and evolutionary constraints were indicative of functional sequences within the genome [102, 103]. For the identified interactions in mouse ES cells, we were trying to see if they show some evidences of regulatory functions. To achieve that, we located the linkage points of each mapped “RNA1-linker-RNA2” chimeric fragment within the genome and plotted the average PhyloP scores [93] flanking the two linkage points (Figure 4.13A). When we plotted the average PhyloP scores for all interactions, the interspecies conservation levels strongly increased at the interaction sites, and the peak of conservation precisely pinpointed the junction of the two RNA fragments (Figure 4.13A). This result suggested that the regions that were involved in the interactions underwent stronger negative selections compared with their neighboring regions. Since our experimental technology selected the transcribed regions, which were more conserved, the overall PhyloP scores near linkage points were higher than genome-wide controls. But we
could still see significant differences between the interaction regions and other transcribed regions nearby. To prove that the differences of conservation levels between interaction sites and flanking regions were not just contributed by the surrounding intronic regions, we compared the average PhyloP scores per nucleotide in interaction regions with their random genomic counterpart within the same RNA categories. The interaction sites within the transcribed regions of mRNAs, lincRNAs, pseudogene RNAs and snoRNAs exhibited increased conservation levels compared to other transcribed regions of the same RNA types (Figure 4.13B). Taken together from last two sub-sections, base complementation is widespread in various types of RNA interactions, and is evolutionarily selected. Both of these two phenomena are consistent with the idea of posttranscriptional regulation by RNA interaction sites.
Figure 4.13: Increased conservation levels in interaction sites. (A) Average PhyloP scores [93] are plotted, centered at the 5’ (blue) and the 3’ (red) junction points (arrows). Conservation levels of randomly selected genomic regions are shown in light blue and pink. The merged data of ES cell samples 1 and 2 were used for this figure. (B) The average PhyloP score per nucleotide was calculated for each interaction site and boxplots were generated for different annotation categories (red). 200,000 random genomic sites were also sampled with same size mean and standard deviation as real interaction sites. They were assigned to different RNA types utilizing our annotation module. The average PhyloP scores were plotted (blue) the same as real interaction sites. P-values were calculated based on one-side two-sample t-test (**: p-value < 10^{-12}; *: p-value < 10^{-6}).
4.5.4 New classes of Argonaute associated small RNAs

We asked whether other RNAs could experience a similar biogenesis process to miRNAs and hybridize to their target RNAs. To do so, we intersected the RNA Hi-C identified interacting RNAs with those found by small RNA sequencing (small RNA-seq) and those bond to the Argonaute protein (CLIP-seq [94]) in ES cells. The small RNA-seq selectively sequenced “miRNAs and other small RNAs that have a 3’ hydroxyl group resulting from enzymatic cleavage by Dicer or other RNA processing enzymes” [64]. Besides miRNA, other RNA types like snoRNA and snRNA also contributed to the small RNA pool, and were attached to Argonaute (Figure 4.14A). Then, we selected RNA Hi-C identified interacting RNA pairs co-appeared in CLIP-seq data, in the meanwhile one RNA in the pair was cleaved into a small RNA and the other RNA in the pair was the exon/utr3 of mRNA. From that, we further selected those interacting RNA pairs with strong base-paring (compared with control by random shuffling of the bases, see Methods). After this filtering, most of processed small RNAs interacting with mRNAs supported by Argonaute binding and base-paring were from snoRNAs (Table 4.8). These data led to the hypothesis that short interfering RNAs could be produced by other genes like snoRNAs than the miRNA genes.

There were some evidences showing that some snoRNAs could be processed by Dicer into smaller RNAs, and these smaller RNAs could function as non-canonical miRNA by targeting 3’-UTR regions of other genes [22, 23, 24]. We asked whether it is a general mechanism for snoRNA genes to produce short interfering RNAs. This hypothesis was first supported by staggering amount of 919 physical snoRNA-mRNA interactions where both the mRNA and the snoRNA were bound by Argonaute. Second, Argonaute bound snoRNAs and their interacting mRNAs exhibited anti-correlated expression changes during guided differentiation of ES cells toward mesendoderm [47] (Figure 4.14B). Third, the average hybridization energy was smaller between Argonaute bound snoRNAs and their target transcripts than that without Argonaute binding (Figure 4.14C). Finally, the siRNA originated from snoRNA genes preferentially interacted with the UTR regions. Out of the 497 snoRNAs involved in RNA-RNA interactions, 243 interacted with UTR regions, among which 223 (92%) were detected in small RNA-seq, suggesting the experience of an enzymatic cut (Figure 4.14D). In comparison, the other 254
Table 4.8: Other types of RNAs functioning as miRNAs. RNA Hi-C identified interacting RNA pairs between a specific type of RNA and exon/utr3 regions of mRNAs which have Argonaute binding supported by CLIP-seq signals on both sides and small RNA-seq signals on the interacting locations of this type of RNA were first selected. Then based on further hybridization energy selection (See Methods), a refined list of interactions was obtained for each specific type of RNAs that might function as miRNA to target mRNA. The table listed number of miRNA-like interacting sites (and corresponding interactions) for each specific type of RNAs.

<table>
<thead>
<tr>
<th>specific type of RNAs</th>
<th># of miRNA-like interaction sites (interactions)</th>
<th>locations of interaction sites (mm9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>snoRNA</td>
<td>83 (226)</td>
<td>See table note for details</td>
</tr>
<tr>
<td>snRNA</td>
<td>8 (16)</td>
<td>See table note for details</td>
</tr>
<tr>
<td>LINE</td>
<td>1 (8)</td>
<td>chr2:90235277-90235370</td>
</tr>
<tr>
<td>misc_RNA</td>
<td>2 (4)</td>
<td>chr2:6997218-6997460, chr4:43505643-43505934</td>
</tr>
<tr>
<td>SINE</td>
<td>1 (2)</td>
<td>chr6:128748868-128748976, chr13: 107911768-107911832</td>
</tr>
<tr>
<td>pseudogene</td>
<td>1 (1)</td>
<td>chr11:86444105-86444271</td>
</tr>
<tr>
<td>LTR</td>
<td>1 (1)</td>
<td>chr18:10052120-10052158</td>
</tr>
</tbody>
</table>

miRNA-like interaction sites were determined by CLIP-seq, small RNA-seq and hybridization energies, see Methods. Detailed information for all miRNA-like interaction sites and interactions were listed here: http://systemsbio.ucsd.edu/RNA-Hi-C/Data/OtherRNAs_as_miRNA.htm
snoRNAs interacting with non-UTR regions contained fewer (55%) small RNAs (Figure 4.14D). Furthermore, two times more UTR-interacting sno-siRNAs were Argonaute bound than the non-UTR interacting snoRNAs (p-value $< 2.2 \times 10^{-16}$, Chi-square test). For example, **SNORA14** RNA targeted the 3' UTR of *Mcl1* mRNA (Figure 4.15A). The protein encoded by *Mcl1* gene was known to be important for the maintenance of undifferentiated state of mouse ES cells. The deficiency of *Mcl1* could result in peri-implantation embryonic lethality [104], and *Mcl1* expression was robust in mouse ES cells [105]. The interacting site on **SNORA14** RNA (110 - 135nt) precisely overlapped with the enzymatically cut form (small RNA-seq) as well as the Argonaute bound region (CLIP-seq). The enzymatically processed portion located precisely on one side of a hairpin loop (orange line, Figure 4.15B), and exhibited a strong binding affinity ($-60$ kCal/mol) to the target UTR site of *Mcl1*. Temporal expression change analysis also suggested the repressive role of this sno-siRNA targeting on the 3'-UTR region of *Mcl1* during guide differentiation (Figure 4.15C). These data suggested a large number (223) of small interfering RNAs originated from snoRNA genes, dubbed sno-siRNAs, which interact with more than 900 mRNAs in ES cells. Thus, the transcripts of these non-miRNA genes could go through a miRNA-like biogenesis and interaction pathway.
Figure 4.14: snoRNAs function as miRNAs to target UTR regions of mRNAs. (A) Average FPKM of each type of RNA Hi-C identified interaction participating RNAs in small RNA-seq (orange) and CLIP-seq (blue). (B) snoRNAs and mRNAs that are interacting with these snoRNAs in UTR/coding regions show anti-correlated temporal patterns of expression changes during a guide differentiation process when their interactions are supported by AGO-binding data. The purple distributions are correlations of temporal changes for snoRNA-mRNA(UTR/exon) pairs supported by AGO-CLIP-seq data; green distributions are for other snoRNA-mRNA pairs (not supported by AGO-CLIP-seq). The difference between two distributions are significant (p-value= 4.18 × 10⁻⁵, Kolmogorov-Smirnov Test) (C) For the interactions between snoRNAs and mRNAs, we separate them into two groups as panel A (AGO-bound and other). The hybridization energy for the AGO-bound snoRNA-mRNA interactions are significantly lower than other snoRNA-mRNA interactions (p-value< 2.2 × 10⁻¹⁶, t-test). For both two groups, significantly lower hybridization energies are found compared to random shuffling (p-value< 2.2 × 10⁻¹⁶, wilcoxon signed-rank test and paired-t-test). (D) snoRNAs involved in the RNA-RNA interactions are separated into two groups based on whether they interact with UTR regions of mRNA or not. The snoRNAs interacting with UTRs are more likely to show signals on small RNA-seq data and AGO-CLIP-seq data from the same cell type.
Figure 4.15: An example to show the interaction between processed $\text{SNORA}_{14}$ and 3' UTR of $\text{Mcl}_1$. (A) Strong interaction between 3' end of $\text{SNORA}_{14}$ and 3' UTR of $\text{Mcl}_1$ has been identified by our technology on mouse ES cells and also supported by our small RNA-seq data and public AGO-CLIP-seq data (GSM622570) [94]. (B) Dominate (orange) and minority (green) strands predicted from small RNA-seq data were base-paired in predicted structure of $\text{SNORA}_{14}$ (by RNAstructure software [91]). The dominate strand can also fold with the interaction region on 3' UTR of $\text{Mcl}_1$ with low folding energy ($\Delta G = -59.96 \text{kCal/Mol}$, from MIRANDA software [92]). (C) $\text{SNORA}_{14}$ and $\text{Mcl}_1$ show anti-correlated expression changes during a guide differentiation process.

4.5.5 RNA Hi-C reveals unique information on RNA structure

From our mapped chimeric fragment pairs and strong interactions, we found that a significant portion of interactions were actually within the same RNA molecule. A further investigation of these self-ligations within the same RNA molecule showed that the ligation points were not always close to each other in terms of the RNA sequences. Thus, the pairing of ligation points could reflect the information of spatial proximity in complex RNA structures like secondary or tertiary structures. Also, single stranded parts of RNAs were preferentially cut with RNase I in Step 3 (Figure 4.1), and therefore the sequencing library partially represented RNA footprinting [106]. A total of
67,221 “chimeric” RNAs mapped to the same gene (Figure 4.16A). These self-ligated RNAs went through the “enzymic cut and ligate” process, and could result from a total of two structural classes (Figure 4.16B).

Each self-ligated RNA could be unambiguously assigned to one of the two structural classes by comparing the orientations of RNA1 and RNA2 in the self-ligated RNA with their orientations in the genome (Figure 4.16B). Compared to other RNA types, snoRNAs possessed the largest fraction (86.8%) of the self-ligated RNAs (Figure 4.16A,C)), likely reflecting the abundance or the structural stability of this class of RNAs. The rest of self-ligated RNAs were primarily contributed by mRNAs and novel transcripts (Figure 4.16A). Longer mRNAs and lincRNAs exhibited a moderate tendency to produce more self-ligations, perhaps reflecting their larger chances to fold (Figure 4.16D).
Figure 4.16: Self ligations detected in RNA Hi-C. (A) Self-ligation mapped pairs has been selected from all mapped pairs after removing single non-chimeric fragments, which is based on the locations, the annotations and the orientations. Most of the self-ligation mapped pairs located in snoRNA regions, and some in mRNA, snRNA and lincRNA regions. (B) An illustration for all possible cases of structure information that can be supported by the self ligation events. Two major groups of different configurations based on the relative locations of ligation points are listed. Each can help us to infer spatial proximity on secondary or tertiary structures. The orange triangles mark RNase I digested sites and the purple lines are the linker sequence. (C) The relationship between the average lengths of RNAs and the average numbers of self-ligation fragments per RNA for different types of RNAs. Here is error bars are the standard errors for average numbers of fragments and average lengths. (D) The relationship between the numbers of self-ligation fragments within RNAs and the lengths of RNAs. It is plotted only for mRNAs and lincRNAs, because the length of snoRNAs and snRNAs are small and very similar.

We examined a set of H/ACA box snoRNAs with generally accepted
secondary structures. A total of 277 chimeric RNAs were produced from SNORA73, representing 1008 RNase I cut positions (Figure 4.17A-B), which concentrated on single stranded regions (heatmap, Figure 4.17D). Six pairs of proximal site pairs were detected, corresponding to 6 clusters of self-ligated RNAs (each cluster was indicated by a colored vertical bar, Figure 4.17A). Each cluster was supported by 4 or more self-ligated RNAs with overlapping ligation positions (black spots, Figure 4.17A). We thought the frequencies of the ligated positions in RNA Hi-C reads (Figure 4.17C) might be reflective of RNA looping. Indeed, 5 of the 6 proximal site pairs (circles, Figure 4.17C) were physically close in the generally accepted secondary structure (arrows of the same color, Figure 4.17D). The remaining pair (arrows 3 and 4) could indicate proximity in the 3 dimensional space. The authenticity of the RNA Hi-C reported proximal site pairs that was not supported by sequence-determined secondary structures could also be supported by other known higher level structures. For example, the SNORA14 RNA folded into a parallel-fingers structure in vivo enforced by its associated proteins. The two pseudouridylation pockets were spatially close to each other on one side (Figure 4.17E, green arrows), which explained one cluster of proximal site pairs from RNA Hi-C data.
Figure 4.17: Examples of self-ligations within same snoRNA molecules which provide information of RNA structures. (A) The relative locations of RNA fragment pairs from self-ligated RNAs belonging to SNORA73. The ligation ends has been marked as black. 6 clusters of self-ligated RNAs are marked by different colors with different numbers. (B) The distributions of RNase I digested sites across the RNA molecule. (C) Heatmap for the frequencies of linkage location pairs. The clusters in panel (A) are circled respectively. These are the region pairs that show spatial proximity within the RNA structures. (D) The distributions of RNase I digested sites across the RNA molecule are marked using a color scale on top of accepted snoRNA structures. The redder the nucleotide, the more frequently it could be digested by RNase I as suggested by our data. Non-base-paired regions tend to be more likely digested. The spatial proximity clusters were pointed on the structure. (E) another example for one pair of proximal sites within H/ACA snoRNA SNORA14 to show that the spatial proximity information are not restricted to secondary structures.

The self-ligation events could also help to interpret structures for novel transcripts. From the 67,221 self-ligated RNAs, we generated 937 clusters using the same methods as before. Since small non-coding RNA transcripts
were usually located within the intronic regions of other genes [107], we searched for novel transcripts from those self-ligation clusters which were located in the intronic regions and supported by more than 10 chimeric fragments. This procedure resulted in 9 clusters located within the intronic regions and also with sufficient self-ligation information. After removing three of them which were overlapped with transposon regions or other annotated small RNAs, 5 out of 6 clusters representing novel transcripts showed increased conservation levels compared with their flanking regions. Based on the relative locations of chimeric RNAs belonging to these novel transcripts, the structure information of these novel transcripts could be derived. For example, a self-ligation cluster was discovered within the third intronic region of \( Rpl23 \) gene (Figure 4.18A), which was highly conserved across mammalian species. Two different types of proximal site pairs were detected from 74 chimeric RNAs (Figure 4.18B, orange and green). The proximal site pair supported by larger number of chimeric RNAs (green) was consistent with the secondary structure predicted solely by sequences (RNAstructure software [91]). The footprinting information derived from frequently digested positions were also in line with the single stranded regions from the predicted secondary structure (heatmap in Figure 4.18C). The other type of proximal site pair could support a shorter version of transcript produced from this region (Figure 4.18C).

In summary, the spatial proximity information together with footprinting information in RNA Hi-C expanded our capacity to examine RNA structures.
Figure 4.18: A novel small RNA identified within the intronic region of Rpl23 gene. (A) the location of this novel small RNA within mm9 genome (pink region). (B) The intra-RNA ligation pairs within this novel small RNA. Two clusters of self-ligated RNAs are marked by different colors with different numbers. (C) Bottom: The secondary structure of this novel small RNA (predicted by RNAstructure software) with the distribution of RNase I digested sites marked with a color scale. Top: A shorter version of transcript can also be folded with the start and end proximate to each other to reflect the orange cluster in panel A.

4.6 Discussion

RNA-RNA interactions are fundamental to many processes including splicing, translation, and gene regulation. There have been several high-throughput technologies that allow mapping of RNA interactomes for a variety of proteins, such as PAR-CLIP [26], HITS-CLIP [27], and CLASH [28, 29]. But still, these technologies are not perfect for mapping of the entire protein assisted RNA interactomes. Several major bottlenecks remain. First, one experiment can only analyzes the interactions mediated by one RNA-binding
protein at a time. Second, each experiment requires either a protein-specific antibody (HITS-CLIP or PAR-CLIP) or stable expression of a tagged protein (CLASH). As a result, CLASH is only applicable to transformed cell lines. Another problem of HITS-CLIP and PAR-CLIP is that the identities of discovered interacting partners by them are not observed directly, which needs to be inferred bioinformatically and is not always reliable.

Our RNA Hi-C method offers several advantages to tackle these bottlenecks. First, the physical interactions among RNAs are experimentally captured. Second, using a biotinylated linker as a selection marker circumvents the requirement for either a protein-specific antibody or the need to express a tagged protein. This allows for the unbiased mapping of the RNA interactome. Third, false positive interactions which are produced by ligation of RNAs randomly proximal in space are minimized by performing RNA ligation on streptavidin beads in dilute conditions. Fourth, the predesigned RNA linker provides a clear boundary delineating sequencing reads that span across the ligation site, thus avoiding ambiguities in mapping the sequencing reads. Fifth, RNA Hi-C directly analyzes the endogenous cellular condition without introducing any exogenous nucleotides [26, 87] or protein-coding genes [28] prior to crosslinking. Sixth, potential PCR amplification biases are removed by attaching a random 6nt barcode to each chimeric RNA before PCR amplification and subsequently only counting completely overlapping sequencing reads with identical barcodes once [27, 88, 89].

We created a novel bioinformatics pipeline to analyze and visualize RNA Hi-C data. It is designed to recover confident RNA-RNA interactions starting from raw reads of linker-ligated chimeric fragments. The paired-end sequencing data from this new experimental technology (RNA Hi-C) is very unique and was built from non-traditional cDNA fragment structures (See figure 4.2, barcode positions and linker positions) in the sequencing library. Because of this feature, how to assemble and organize different modules in the analysis steps is a challenging task. In the RNA-HiC-tools, we assemble different analysis modules together as a complete computational workflow with different functions, i.e. calling interacting RNAs, statistical assessments, categorizing RNA interaction types, and RNA structure analysis. Some modules are adopted from existing tools, such as BOWTIE and BLAST. However, since this is a completely new type of data without existing analysis tools, most of the modules or steps were designed and implemented specifically for this
project. Here are some examples:

1. Existing barcode splitters only works for single-end sequencing data, we wrote a script to split libraries based on given barcodes for paired-end sequencing data.

2. Since our pipeline has several necessary steps before mapping, we decided to remove PCR duplicates before that to reduce the computational burden. Samtools’s “rmdup” and Picards “MarkDuplicates” functions are able to remove PCR duplicates after alignment. “fastx_collapser” is able to remove duplicates before alignment for single-end sequencing data only. Thus, removing PCR duplicates before alignment for paired-end sequencing data is still new. We created a function to solve this problem with similar idea as “fastx_collapser”. After we finished this function, we found “FastUniq” was just published with same purpose [108]. Our test showed that “FastUniq” is faster than our function but consumes more memories (about 4 times more memory consumption than ours), which makes it not suitable for our large dataset.

3. The function to recover fragments is also novel. There were existing tools to merge paired-end reads when the fragments to be recovered were equally long. However, when the fragment lengths are not the same and especially when either fragment length is shorter than one read end (Type1 in Figure 4.3) or longer than twice the size of one read end (Type3 in Figure 4.3), their results were not reliable [109]. PEAR works for all these situations [109] as well as our function. Studies on synthetic data proved that our function could recover the fragments with high accuracy. The idea of our function was described in the Method section and Figure 4.3.

4. There is no existing method to infer strong RNA interactions from such type of data. Therefore, we first applied Union-Find algorithm with weighted quick-union to find clusters based on mapped read pairs and listed all interactions between clusters. We then use a Fisher’s exact test to evaluate the significance of each interaction based on numbers of mapped reads. This method is novel. Results from HITS-CLIP and PAR-CLIP could not identify one-to-one correspondence of interaction partners. The bioinformatics pipeline for CLASH-seq doesn’t have a
statistical evaluation for the chimeric products it found, and it doesn’t support paired-end sequencing [110].

5. We also created computational tools to compare the interactions between different samples and visualize inter-RNA and intra-RNA interactions in different scales.

The sensitivities of identified RNA interaction sites could be partially assessed by comparing with the results from other technologies which were designed to detect RNAs interacting with one specific type of protein. Polycomb-associated RNAs were identified genome-wide using RIP-Seq also in mouse ES cells [111]. In theory, these RNAs should also be detected from our RNA Hi-C data in the same cell type. The comparison of Polycomb-associated RNAs they identified (Table S1 of reference [111]) and the interaction sites from our RNA Hi-C data could help us to assess the sensitivity of our method. More than half (50.98%) of the Polycomb-associated RNA locations were presented in RNA Hi-C interaction sites (Table 4.9), which implied good sensitivity of RNA Hi-C on Polycomb associated RNA interactions. However, since there is no reliable technologies to unbiasedly identify the whole RNA interactome existing yet, we are not able to evaluate the sensitivity of our method in real data completely and comprehensively.

**Table 4.9:** The overlap between interacting RNA sites identified from RIP-seq and RNA Hi-C. Random interaction sites within the RNA regions were sampled 50 times with the same number as RNA Hi-C interaction sites each time. The mean number and standard deviation of overlaps between random interaction sites and Polycomb-associated RNAs are calculated. P-value for the actual overlaps between RNA Hi-C interaction sites and Polycomb-associated RNAs is calculated assuming normal distribution on random overlaps.

<table>
<thead>
<tr>
<th></th>
<th>Polycomb-associated RNAs (8,670)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA Hi-C interaction sites</td>
<td>4,420</td>
</tr>
<tr>
<td></td>
<td>(21,278)</td>
</tr>
<tr>
<td>Random interaction sites</td>
<td>1,472.18 ± 39.48</td>
</tr>
<tr>
<td>(21,278), 50 times</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>&lt; 2.26 × 10⁻³⁰⁸</td>
</tr>
</tbody>
</table>

Detailed information for all overlapping regions (4420) and numbers of overlaps categorized by RNA types were listed here: http://systemsbio.ucsd.edu/RNA-HiC/Data/Overlap-RIP-seq_RNA-HiC_types.htm

Besides those inter-RNA interactions, the clear boundaries of intra-RNA
interactions determined by precise localization of linker sequence within the fragments could provide information on *in vivo* RNA structure in two aspects: 1, the footprint of single stranded parts of an RNA based on RNase I cutting sites; 2, the spatially proximal sites of each RNA based on the ligation locations. Other high-throughput methods can also map RNA secondary structures by examining whether each nucleotide is base-paired or not. In parallel analysis of RNA structure (PARS) [106] and fragmentation sequencing (FragSeq) [112], RNA is partially digested with ribonuclease and analyzed by deep sequencing. In SHAPE-Seq [113], it combines selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry with a multiplexed hierarchical bar coding and deep sequencing strategy. These methods could provide higher resolution for the base-pairing preference of each nucleotide compared to the information from RNA Hi-C data. That is because they were specifically designed for probing RNA structures with high coverage. But none of these methods could provide information of spatial proximity of two base-pairing partners in secondary structure or two proximal locations in tertiary structure without base-pairing. The intra-RNA interactions recovered from RNA Hi-C data could provide this type of information and add extra values for *in vivo* RNA structure prediction.
With the rapid advance of next generation sequencing (NGS) technologies, the expense of sequencing a genome has been greatly decreased. The scope of genetic analysis has been expanded from inspection of the coding regions of dozens of genes to the current methods of examination of the entire 3 billion base pair genome. Although the cost and time of sequencing data generation will no longer be a problem, the major bottlenecks on the data analysis side are still there for the genomic research. The main objective of this thesis is to explore different computational approaches to handle these massive amounts of data, and to uncover the mysteries of regulation on gene expression through studies of epigenome and RNA interactome.

Genome-wide epigenetic and RNA interaction data could be obtained from different types of NGS based technologies such as ChIP-seq, MeDIP-seq, MRE-seq, MNase-seq, RNA-seq and RNA Hi-C. These data provide invaluable information about how different parts of the genome execute their regulatory functions to ensure essential cell activities and cell-type specific behaviors. All three parts of my thesis have focused on building and implementing bioinformatic methodologies to extract these invaluable information and infer different regulatory roles of epigenome and RNA interactome. Meanwhile, each of these three parts has specific focal point in different dimensions.

In the first part, we have evaluated the conservation of epigenome in mammalian species, as well as its relationship with conservation of genomic sequences and transcriptome. We first find that the relative difference in epimodification intensities on different genomic features is in general consistent across species. This pattern of conservation provides evolutionary support for the idea of using epigenomic data to predict functional noncoding genomic features. Then from our analysis, it also shows that conservation of
epigenome cannot be fully explained by genomic sequence conservation. Instead, interspecies epigenomic changes are predictive of interspecies changes of gene expression. These results provide further support to use conserved epigenomic patterns to identify regulatory elements. We select several evolutionarily conserved epigenomic mark combinations and test their functions through a cell differentiation assay. The tests confirmed the functions of all conserved epigenomic mark combinations. Thus, by comparing interspecies epigenomic combinatorial patterns, we are able to distinguish functional colocalization of epigenomic marks from nonfunctional combinations.

From the guided differentiation assay, it is obvious that genome-wide distributions of epigenomic modifications are dynamic with the changes of cell states. The associations between epigenome and gene expression have two orthogonal dimensions. First, within a specific cell type, the relative differences of epigenomic modifications among different genomic locations are indicative of the relative expression differences of genes in the neighborhoods of these locations. Second, when cell type changes during a development process, the temporal epigenomic changes within a specific genomic region are correlated with the transcription level changes of genes near these regions. Previous works have taken advantages of the first dimension to annotate genomic features based on invariant epigenomic signatures in a static cellular condition. Here in the second part of this thesis, we developed a method to jointly model the effect of these two dimensions together for the first time. We call it GATE. This model can achieve more precise identification and functional annotation of regulatory sequences. The output of GATE model are first confirmed by known cis-regulatory annotations. By careful interpretation from these output, we get some interesting and novel findings. First, specific epigenomic patterns are found for piRNA genes which confirms the activation of piRNA genes in mouse ES cells. Second, some recurrent themes representing temporal gene regulation by the epigenome are uncovered. Third, our model reveals that temporal changes of H3K4me1/2, mCpH and H2A.Z are highly correlated with temporal 5-hmC changes. This may help for future biochemical analysis of 5-hmC regulation pathways. Finally, since the epigenomic states uncovered by the model also contain information of the time of their activation, this information could help us to infer the transcription network.
A case study has been presented within the genomic regions belonging to the enhancer cluster which are only activated during the differentiation process. A positive feedback loop between SOX17 and FOXA2 has been deciphered, which is also supported by other independent data.

Non-coding RNAs have been shown as versatile regulators of cellular processes. Many of their regulatory functions have been executed through interacting with other RNAs mediated by RNA binding proteins. To overcome the challenges on mapping the entire protein assisted RNA-RNA interactome, our lab developed a novel experimental technology called “RNA Hi-C” to allow simultaneous and unbiased identification of different types of RNA-RNA interactions within the cells. This was achieved by coupling high-throughput sequencing with proximity-ligation-based methods. We also created a bioinformatics toolbox (RNA-HiC-tools) to analyze and visualize RNA Hi-C data. The package included modules like removing PCR duplicates, splitting multiplexed samples, identifying linker sequences, splitting chimeric fragments, calling interacting RNAs, statistical assessments, reproducibility analysis, categorizing RNA interaction types, and RNA structure analysis. Following the computational pipeline on this toolbox, we were able to evaluate the quality of the RNA Hi-C data and obtain a comprehensive view of genome-wide RNA interactomes. Some novel ARGONAUTE associated small RNAs were discovered in mouse ES cells. Also, the interactions within the same RNA molecules provided useful information on RNA structures.

Some aspects within this thesis could be further improved by future work, most of them are in the modeling part. First, for the spatiotemporal GATE model, the cluster numbers need to be manually assigned by the user. To allow these numbers to be automatically determined by input data could further increase the performance powers of these two models. Second, at the end of the second part, we mentioned that transcription network could be inferred by the output of GATE model. There, only a case study has been presented. A more systematical approach for inferring transcription networks from temporal epigenomic data could greatly elevate the significance of our model. In terms of the bioinformatic pipeline for identification of RNA inter-
actome, there are several features could be improved. First, more mapping options could be added with the option to map the reads to transcriptome instead of whole genome. Second, the two types of RNA structure information interpreted from RNA Hi-C data could be incorporated with current secondary and tertiary structure prediction software to give a more accurate prediction of *in vivo* RNA structures.
REFERENCES


[121]
