Title
Comparative functional genomics of mammalian developmental processes

Author
Jiang, Shan

Publication Date
2018

Peer reviewed|Thesis/dissertation

University of California,
Irvine

Comparative functional genomics of mammalian developmental processes

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Biological Sciences

by

Shan Jiang

Dissertation Committee:
Associate Professor Ali Mortazavi, Chair
Professor Ken W. Cho
Professor Tallie Z. Baram
Assistant Professor Sha Sun
Assistant Professor Zeba Wunderlich

2018

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without help and support from many people.

First of all, I deeply thank my academic advisor Dr Ali Mortazavi. Dr Mortazavi introduced me to the intersection of functional genomics and developmental biology. He has been intensively involved in all the work presented in this thesis. He helped in shaping my taste and style of research and taught me what is good science and what is worth pursuing. I appreciate all the opportunities that he has provided me to build my research career.

I would like to thank all my friends around for bringing funs and color into my life. The PhD journey is never boring with friends around.

I would like thank all my committee members, Professors Ken Cho, Tallie Z. Baram, Sha Sun and Zeba Wunderlich, who have helped guide my dissertation and demonstrated commitment in this regard.

I would also like to thank Dr Kyoko Yokomori, Dr Ricardo Ramirez, Dr Weihua Zeng, Dr Noriko Kamei and Dr Xiaojie Wang, for their tremendous advice and support in my research.

I would like to thank all the supports from my mom, dad, grandparents and my dear husband. They never stop encouraging me to pursue my dream bravely. Thank you mom for the hot food and thank you Zong for tolerating my struggling during studies.

CURRICULUM VITAE

SHAN JIANG

EDUCATION

Doctor of Philosophy in Developmental and Cell Biology 2018
University of California, Irvine Irvine, California, USA

Master of Biological Sciences 2015
University of California, Irvine Irvine, California, USA

Bachelor of Biotechnology 2013
Sun Yat-sen University Guangzhou, CHINA

RESEARCH EXPERIENCES

Department of Developmental and Cell Biology, UCI 2014-present
Graduate Research Assistant

Gateway program of Mathematical, Computational and Systems Biology 2013-2014
Graduate Research Assistant

Biotechnology Research Center, SYSU 2012-2013
Student researcher in RNA informatics core

Beijing Genomics Institute, BGI 2010
Bioinformatics Interns for undergraduate students

TEACHING EXPERIENCES

Teaching assistant Winter 2015
Introduction to personalized medicine
Instructor: Dr Ali Mortazavi

Teaching assistant Spring 2015&2016
Lab in developmental and Cell Biology
Instructor: Dr Debra Mauzy-Melitz

Teaching assistant Winter 2017
Organisms to ecological systems
Instructor: Dr Nancy Aguilar-Roca

PUBLICATIONS

Najafi AR.,Crapser J., **Jiang S.**, Ng W., Mortazavi A., West BL., Green KN. 2018. A limited capacity for microglial repopulation in the adult brain. (in press, ***GLIA***).

Jiang S., Mortazavi A., 2018. Integrating ChIP-seq with other functional genomics data. ***Briefings in Functional Genomics***, doi: 10.1093/bfgp/ely002.

Singh-Taylor, A., Molet, J#., **Jiang S#**., Korosi, A., Bolton, JL., Noam, Y., Simeone, K., Cope, J., Chen, YC., Mortazavi, A., and Baram, TZ., 2017. NRSF-dependent epigenetic mechanisms contribute to programming of stress-sensitive neurons by neonatal experience, promoting resilience. ***Molecular Psychiatry***, doi: 10.1038/mp.2016.240. (# Equal contributors)

Hernandez MX., **Jiang S**., Cole TA.,…, Mortazavi A., Tenner AJ., 2017. Prevention of C5aR1 Signaling Delays Microglial Inflammatory Polarization, Favors Clearance Pathways and Suppresses Cognitive Loss. ***Molecular Neurodegeneration,*** doi: 10.1186/s13024-017-0210-z.

Zeng, W., **Jiang, S**., Kong, X., El-Ali, N., Ball, A.R., Christopher, I., Ma, H., Hashimoto, N., Yokomori, K. and Mortazavi, A., 2016. Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. ***Nucleic Acids Research***, doi: 10.1093/nar/gkw739.

Yang, J.H., Li, J.H., **Jiang, S**., Zhou, H. and Qu, L.H., 2013. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. ***Nucleic Acids Research***,doi: 10.1093/nar/gks1060.

MANUSCRIPTS

Jiang, S.#, Williams, KE.#, Zeng, WH., Kong, XD., Tawil R., Mortazavi, A.* and Yokomori, K.*, Single-cell and single-nucleus RNA-seq of differentiating myoblasts from Facioscapulohumeral Muscular Dystrophy (FSHD) patients. (in preparation) (# Equal contributors)

Jiang, S#., Kamei, N#., Bolton, JL., Ma, XY., Stern HS., Baram, TZ .*, Mortazavi, A.*,. Rapid intra-individual methylation signatures of diverse early life experiences. (# Equal contributors) (in submission)

Jiang, S., Ramirez, R., El-Ali, N., Mortazavi, A. Dynamics of NRSF/REST motif evolution favor the canonical NRSE/RE1 form. (in resubmission)

Jiang, S., Wilcox, C., Ma, X Y., and Mortazavi, A., Comparative chromatin dynamics of definitive endoderm differentiation (in preparation)

Schulmann, A., Bolton, JL., Curran, MM., Regev, L., Kamei, N., Singh-Taylor, A., **Jiang, S.**, Molet, J., Mortazavi, A., Baram, TZ. Unexpected transcriptional programs underlie enduring memory deficits after early-life adversity. (under review)

Baglietto-Vargas, D., Cai, L., Martini, AC., Trujillo-Estrada, L., Forner, S., **Jiang, S**., Kramr, EA., Nuez-Diaz, C., Matheos, DP., Cunha, CD., Ager, RR., Gutierrez, A., McGregor, G., Green,

KN., Wood, MA., Mortazavi, A., Tenner, AJ., LaFerla, FM. A knock-in mouse model for sporadic Alzheimer's disease. (in preparation)

Hohsfield, LA., **Jiang, S**., Mortazavi, A., Green, KN. De novo myeloid cell generation from the adultsub-ventricular zone/rostral migratory stream. (in preparation)

Wang, XJ., Ramos, R., Oh, JW., Nguyen, TK.,...**Jiang, S**., ...Mortazavi, A., Kunisada, T., Li, J., Plikus, MV. Signaling by senescent cells hyper-activates skin stem cell niche. (in resubmission)

THESIS CHAPTERS

Comparative functional genomics of mammalian developmental processes

Chapter 1 Introduction - Integrating ChIP-seq with other functional genomics data

Chapter 2 Rapid intra-individual methylation signatures of diverse early life experiences

Chapter 3 Characterizing the heterogeneity of DUX4 and DUX4 targets expression during FSHD2 myoblast differentiation

Chapter 4 Comparative chromatin dynamics of definitive endoderm differentiation

Chapter 5 Dynamics of NRSF/REST motif evolution favor the canonical NRSE/RE1 form

Chapter 6 Future directions

ABSTRACT OF THE DISSERTATION

Comparative functional genomics of mammalian developmental processes

By

Shan Jiang

Doctor of Philosophy in Developmental and Cell Biology

University of California, Irvine, 2018

Associate Professor Ali Mortazavi, Chair

Individual development is a complex process with a myriad of developmental controls at multiple levels ranging from individual cells to organs and entire individuals. The development and specification of each cell ultimately encoded in the genome. But whereas the genome is the same for all cells of the same individual, cell differentiation, specialization and response to the environment is regulated at the epigenetic level by gene regulatory networks (GRNs). Functional genomics studies have revealed that protein- DNA interactions, DNA methylation and changes in chromatin accessibility are essential to maintain cell identity and that interruption of these GRNs causes defects in cell development that can lead to disease and abnormal behaviors in individuals. Given the importance of epigenetic regulation in cells, tissues, and individuals, it would be interesting to know how these GRNs are conserved and evolve during mammalian evolution and how they can go wrong in disease. In the thesis, I present functional genomics studies and expand the understanding of epigenetic control in development from four aspects: (1) changes in DNA methylation in the same individual can be used as signature of different life experiences; (2) mutations in a repressor can cause abnormal gene expression in a small group of cells that further induce the onset of muscle wasting disease FSHD; (3) comparative dynamics of

chromatin accessibility during definitive endoderm differentiation can identify conserved regulatory modules as well as species-specific enhancements; (4) The canonical form of the transcription factor NRSF is stabilized in genome through motifs conversion during mammalian evolution. These results show the versatility of epigenetic control during development and disease as well as highlight evolutionary forces shaping GRNs.

THEME OF THESIS

One of the most important questions in developmental biology is how development is precisely controlled temporally and spatially. Many studies have shown that transcription factors (TFs) selectively interacting with cis-regulatory elements in open chromatin regions is one of the most critical regulatory inputs in specifying cell types and in maintaining cell functions by tightly controlling gene expression. DNA methylation and histone modifications are also required in this process and they actively cooperate with sequence-specific TFs to regulate development precisely. The environment such as early life experiences, environmental chemicals, diet and aging, also influences developmental processes by triggering epigenetic changes at both cellular and whole-individual level. Disruptions in epigenetic regulation may cause serious diseases such as cancer, mental disorders, diabetes and immune diseases. The central theme of my thesis is to understand the plasticity of epigenetic regulation and also assess their conservation during mammalian evolution. I examined these regulatory changes from single cell to individual using a combination of both experimental and integrative computational methods.

Functional genomic assays have been widely used to observe regulatory changes in diverse developmental processes. **In Chapter 1**, I review the current methods being used to integrate ChIP-seq data of TF binding and histone modification with other data such as open chromatin accessibility and gene expression data. ChIP-seq has been used as a standard method to detect the binding of interest TFs and histone modifications *in vitro* and *in vivo*. It is hypothesized that the stage-specific binding profiles have the potential to predict the level of gene expression during cellular development. However, the binding of one transcription factor alone is rarely enough to directly infer functional effects on the expression levels of neighboring genes, which are typically under the combinatorial control of multiple transcription factors and other epigenetic regulation. Therefore, ChIP-seq data is often integrated with other functional genomic techniques to decipher the basic regulatory control of gene expression. In this chapter, I introduce the strategy and methods of integrative analysis of ChIP-seq with other functional genomic assays to understand the regulatory control of gene expression by incorporating the combinatorial control of gene expression levels (RNA-seq), open chromatin regions (ATAC-

/DNase-seq/FAIRE-seq), long-range chromatin interactions (ChIA-PET/Hi-C) and single-nucleotide polymorphism (SNP) variants. As the number of ChIP-seq datasets as well as datasets from other genome-wide assays grows, the power of integrative computational analyses continue to increase. Therefore, I discuss the application of probabilistic models and machine-learning methods to the analysis of TF and histone modification ChIP-seq data simultaneously in order to identify chromatin patterns across multiple genomes and cell types. Finally, I also discuss the challenges of analyzing ChIP-seq data using low amounts of input materials, and their further application in the emerging field of integrative analysis of single-cell sequencing functional genomics data.

In Chapter 2, I show that a DNA signature that can used to distinguish individual rats with different early life experiences. I found that the DNA methylation level of multiple genomic regions changes substantially in early rat postnatal development and that normal or fragmented maternal care during that window changes methylation patterns at the vicinity of multiple transcription factors. This study tries to understand how environment, i.e early life experiences, influence development in individuals via epigenetic changes. Although it is known that early life experiences drive gene expression changes and they further influence the maturation of brain and other organs in mammalian individuals, our knowledge about specific epigenetic regulations involved into these processes are limited. Among epigenetic regulations, DNA methylation is known to correlate with gene expression changes that can be used to predict aging and risk level of certain cancer types. However, it is not known if DNA methylation changes might provide a useful 'epigenetic signature' of early-life experiences in an individual child. Therefore, this study addresses two critical questions to understand the nature of DNA methylation changes in early life experiences: (1) does a short period of early postnatal life change methylation patterns in individuals? (2) can methylation changes be used to distinguish individuals with early-life adversity? In order to allow the future extension of our methods to human infants, we examine methylation by using peripheral cell population from buccal swabs (mixed epithelial and white blood cells) rather than brain cells directly. By comparing two samples from two time points, i.e. neonatal (Day 2) and infant (Day 10), of the same individual rat in groups exposed to distinct early-life experiences, we find changes in methylation patterns globally and these profiles can be used to distinguish as separate groups infant from neonatal rats. Consistent with previous studies,

these methylation changes cannot distinguish rats with different early life experiences as distinct groups. We develop a novel approach and demonstrate for the first time that intra-individual changes in methylation patterns can robustly distinguish individuals with adverse experiences and that they serve as a predictive signature in individuals. Given the predictive power of methylation signature in early life experiences and the accessibility of peripheral cell populations, we will apply in the future this technique to the study of human babies.

In Chapter 3, I show that epigenetic changes cause severe disease in human by only inducing the abnormal expression of one gene in a small population of cells. Fascioscapulohumeral muscular dystrophy (FSHD) is primarily caused by the expression of the normally repressed transcription factor DUX4 in skeletal muscle by turning on a set of target genes. FSHD has been classified into two subtypes based on the mechanism of DUX4 expression in skeletal muscle. FSHD1 is caused by a contraction of the D4Z4 macrosatellite repeat array containing the DUX4 gene. But FSHD2 is characterized by a normal D4Z4 repeat array size, and recurring mutations in genes such as the chromatin modifier *SMCHD1* (Structural maintenance of chromosomes flexible hinge domain-containing protein 1), which is important for maintenance of DNA methylation and epigenetic silencing of the D4Z4 repeat array. Mutations in SMCHD1 decrease repression of the D4Z4 repeat array and further result in the up-regulation of DUX4 target genes. Previous studies have shown that DUX4 is lowly expressed and rarely detected in patient samples and previous transcriptomic studies have been based on the overexpression of DUX4. However, overexpression methods cause higher DUX4 expression than in patients, which may not be appropriate to derive solid physiological and cellular conclusion on the disease progression. Furthermore, previous population-based studies have found that DUX4 target genes are not consistently expressed across all FSHD patient cells and given that DUX4 is presumably only expressed in a small subset of cells, it is important to investigate the cellular heterogeneity in FSHD patient samples and to understand how DUX4 regulates target genes directly, as well as how they are involved in the disease dysregulation. This study tries to understand the contribution of DUX4 expression and its target genes to the pathogenesis of FSHD2 by addressing two critical questions: (1) What are the targets of DUX4 in FSHD2? (2) Are DUX4 and its targets all expressed in the same nuclei? By using single nucleus RNA-seq methods in myoblast differentiation that our lab has developed, I present the

first direct detection of DUX4 expression in myotubes from FSHD2 patient-derived myoblasts (2.2% of FSHD2 myotube nuclei). Although DUX4 is only detected in a small number of nuclei, over 50% of FSHD2 myotube nuclei express multiple DUX4 targets. I show that DUX4 positive nuclei share similar profiles to a larger set of nuclei without DUX4 but with significantly higher expression of DUX4 targets compared with other DUX4 negative nuclei. This distinct population of DUX4-target positive nuclei clearly separated from other cells types using a variety of computational approaches, indicating these cells have entered a distinct biological program potentially driven by other transcription factors downstream of DUX4.

In Chapter 4, I study the role of epigenetic regulation during cell differentiation by changing chromatin accessibility and its effect on gene expression. One of the most important questions in developmental biology is how cell-fate commitment and differentiation are precisely controlled by the genome using gene regulatory network (GRNs). By observing the expression and chromatin-level interactions between TFs and other key genes, the regulation of cell type specification can be summarized into one GRNs with hierarchical regulatory structures. Embryonic stem cells (ESCs) is one of the most attractive model to study gene regulation in cellular development and specification as they are pluripotent and can produce all three germ layers. Previous studies have used ESCs differentiation into neurons, heart, liver and kidney in human and mouse and demonstrated that open chromatin regions coupled with TF binding and histone remodeling regulate stage-specific gene expression. However, it is unknown how conserved these GRNs are during mammalian evolution. Therefore, to address these questions, this study for the first time demonstrate the conservation and divergence of GRNs in three mammalian species (human, mouse and rat) by differentiating their ESCs into definitive endoderm *in vitro*. Although GRNs have been built during endoderm specification in vertebrates, such as Xenopus and zebrafish, our knowledge about endoderm GRNs in mammalian species is more limited. In this study, I describe the monolayer differentiation of embryonic stem cells into definitive endoderm in human and mouse, and for the first time in rat *in vitro*. Using RNA-seq and ATAC-seq during endoderm differentiation, I quantify the dynamics of gene expression and chromatin accessibility during differentiation for three species. I show that gene expression has higher conservation level (48%) than chromatin accessibility (25%) across three species. I then use chromatin accessibility footprinting to construct gene regulatory networks (GRNs) of key

TFs involved in endoderm formation and compared their conservation level across species. Known regulatory interactions are recovered in these GRNs and many novel interaction are also detected. I finally show more extensive rewiring of endoderm GRNs compared to ESC GRNs.

In Chapter 5, I focus on the evolution of the binding repertoire of TFs in mammalian species. TFs selectively bind to non-coding elements regulate specific gene expression. Changes in these TFs binding sites are known to be involved in the regulation of gain, loss or modification of traits. TFs binding repertoires have expanded during vertebrate evolution, which provide more evolutionary raw material for *cis*-regulatory elements in rewiring GRNs. 50 years of biochemical experiments have shown that specific DNA sequences, known as motifs, are located within TFs binding sites and they are believed to be the direct target of TFs binding. In this study, I answer one important question in gene regulation: how do these TFs binding motifs evolve in the mammalian genome? Previous studies have shown rapid turn-over of TFs motifs during mammalian evolution. However, these studies only focus on activator and leave repressors behind. In addition, many TFs are known to have more than one type of motif and the selection between these alternative forms during genome evolution is poorly understood. In order to address these questions, I observe the binding profile of the well-known repressor NRSF/REST (repressor element 1 silencing transcription factor), which is a neuronal repressive transcription factor primarily repressing neuronal gene expression in non-neuronal cells and neuronal stem cells, using ChIP-seq in human, mouse, dog and horse. NRSF can regulate gene expression by binding to three form of motifs, including the most common 21bp canonical NRSE/RE1 motif; a smaller class of non-canonical NRSEs, which consist of two half-motifs separated by 10, 16-19 base pairs; left- and right- half motifs. I show the frequent turnover of NRSF binding sites across four species, which are mediated by genomic feature changes, including repetitive elements, and insertion or deletion in motif sequences. I also show that canonical motifs are not only the major form mediating NRSF binding in four species but also show dominant role in conserved NRSF binding across four species. I further propose a model to explain this prevalence of canonical form of NRSF motifs in which the biased conversion from non-canonical or half-motif contribute to the accumulation of canonical motifs during mammalian evolution.

In Chapter 6, I focus on extensive discussion and future direction for each of the chapters above. I suggest possible experimental designs to validate some of the hypotheses suggested by our analyses that could be carried out in future studies. As each chapter heavily relies on the usage of different functional genomics assays, I focus on discussing the limitation of specific techniques and improving the methods both experimentally and computationally to optimize the results. I also discuss the application of machine learning methods in some of the chapters to overcome the challenges of large-scale computation on genomics data.

CHAPTER 1

Introduction

Integrating ChIP-seq with other functional genomics data

Note: This chapter is an adaption of materials that appears in publication

Jiang, S., & Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. *Briefings in functional genomics*, *17*(2), 104-115.

Chapter 1
Introduction - Integrating ChIP-seq with other functional genomics data

1.1 Abstract

Transcription is regulated by transcription factor (TF) binding at promoters and distal regulatory elements and histone modifications that control the accessibility of these elements. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become the standard assay for identifying genome-wide protein-DNA interactions *in vitro* and *in vivo*. As large-scale ChIP-seq datasets have been collected for different TFs and histone modifications, their potential to predict gene expression can be used to test hypotheses about the mechanisms of gene regulation. In addition, complementary functional genomics assays provide a global view of chromatin accessibility and long-range cis-regulatory interactions that are being combined with TF binding and histone remodeling to study the regulation of gene expression. Thus, ChIP-seq analysis is now widely integrated with other functional genomics assays to better understand gene regulatory mechanisms. In this review, we discuss advances and challenges in integrating ChIP-seq data to identify context-specific chromatin states associated with gene activity. We describe the overall computational design of integrating ChIP-seq data with other functional genomics assays. We also discuss the challenges of extending these methods to low-input ChIP-seq assays and related single-cell assays.

1.2 Introduction

DNA-protein interactions and epigenetic modifications are crucial for transcriptional regulation. Genome-wide profiling of transcription factor (TF) binding sites, regions with covalently modified histones, and other DNA-binding proteins reveal cell/tissue-, species-, and disease-specific cis-regulatory repertoires, which are vital for understanding gene regulation. Chromatin immunoprecipitation (ChIP) methodologies [1-3] use an antibody that recognizes a transcription factor or histone modification to pull down attached DNA for identifying binding locations. With the rapid development of sequencing technology, chromatin immunoprecipitation followed by sequencing (ChIP-seq) [2-5] has become the most common and effective assay to identify bound loci genome-wide *in vitro* and *in vivo*. The basic computational pipeline and software for analyzing ChIP-seq data has been established and

optimized alongside advances in sequencing library preparation and ChIP-seq techniques [6-8], including read quality control, alignment, peak calling, and evaluation of reproducibility. ChIP peaks can be visualized using genome browsers as a simple quality check of signal over known true positives. Confirmed peaks can be further analyzed with differential density analysis for different treatments, gene-associated annotation, motif discovery, and other downstream analyses. Limitations and advances in these steps are reviewed in detail elsewhere [9].

However, the binding of one transcription factor alone is rarely enough to directly infer functional effects on the gene expression levels of neighboring genes, which are typically under the combinatorial control of multiple transcription factors. Therefore, ChIP-seq data is often actively integrated with other functional genomic techniques to decipher the basic regulatory control of gene expression by incorporating open chromatin regions, long-range chromatin interactions and SNP (single-nucleotide polymorphism) variants. With the increasing availability of multiple ChIP-seq datasets [10, 11], as well as datasets from other genome-wide assays, the power of integrative computational analysis is of ever-increasing interest. In this review, we discuss the application of probabilistic models and machine-learning methods to the analysis of TF and histone modification ChIP data simultaneously in order to identify chromatin patterns across multiple genomes and cell types. We also focus on the computational integration of ChIP-seq with other functional genomic assays such as RNA-seq for gene expression levels, ATAC-/DNase-seq/FAIRE-seq for chromatin accessibility, and ChIA-PET/Hi-C for chromatin interactions that affect regulation of gene expression. Finally, we discuss the development of ChIP-seq assays that use low amounts of input materials, and their further application in the emerging field of integrative analysis of single-cell sequencing functional genomics data.

1.3 Identifying distinct chromatin states using histone modifications and TF occupancy

Histone modifications are often found in recurring combinations at promoters, enhancers, and repressed regions. These combinations are referred to as “chromatin states” and can be used to annotate regulatory regions in genomes [12, 13]. For example, H3K4me1 alone marks primed enhancers, while H3K4me1 combined with H3K27ac mark active enhancers. Promoters are characterized by a detectable level of H3K4me3 coupled with a high ratio of H3K4me3 to H3K4me1. Furthermore, H3K36me3 histone modifications and RNA Pol II ChIP signal are

associated with transcribed regions, while the presence of H3K27me3 or H3K9me3 is associated with repressive chromatin states (Figure 1.1) [14, 15]. The goal of software packages analyzing chromatin states is to first discover these relationships in the data,
and to then check for changes in states assigned to a particular region in different cell types. Large-scale datasets produced by ENCODE [10] and Roadmap Epigenomics [11] have been used to train and to test with statistical or machine-learning methods that assign chromatin states to genomic segments (typically 100 bp or longer). These state assignments can then be interpreted through comparisons with known annotations and gene expression.

Hidden Markov Models (HMMs) were originally developed for speech recognition, but have since been used extensively in other fields to identify hidden states from observed signal data [16]. In genomics studies, it has been successfully applied to gene annotation [17] and protein domain characterization [18]. HMMseg [19] was the earliest software package to partition and annotate a genome by training HMMs on functional genomics data. However, this tool can only identify two states ("active" or "inactive"), which limits its application in annotating chromatin states in greater detail, e.g. active/poised promoters and enhancers. ChromHMM [13] and Segway [20] were developed with the goal of capturing more comprehensive combinatorial patterns of multiple histone modifications, RNA Polymerase II binding, and insulator CTCF binding genome-wide (Figure 1.2). ChromHMM segments the genome into minimum 200bp intervals (default) and converts raw reads counts into binary code using a product of independent Bernoulli random variables for each interval, which are then used to train a hidden Markov model. Similarly, Segway was developed based on dynamic Bayesian networks (DBNs). It transforms raw read counts to coverage signal and can segment the genome down to 1-bp resolution, although 100bp segments are more practical. Additional tools have been developed to extend and speed up the identification of chromatin states. For example, TreeHMM [21] also uses binary vectors, but is position-dependent when inferring chromatin patterns during cell differentiation and across different cell types. hiHMM [22] uses a hierarchically linked infinite HMM model to not only identify chromatin states across multiple ChIP-seq data sets, but also address species variance for cross-species inference. diHMM [23] inherits from ChromHMM but uses a hierarchical hidden Markov model to identify combinatorial patterns at variable length scale that range from nucleosome-level to higher-order

domain-level states. Another joint analysis platform, IDEAS [24, 25], can infer chromatin states using both position-dependency and cell-type specific cases at multiple range scales, and can run faster than both ChromHMM and Segway using single core mode. Additional tools have been developed for comparing chromatin patterns between different experimental treatments [26] and expanding the comprehensiveness of epigenomic maps [27]. The combinatorial patterns generated by these methods have been correlated with gene expression profiles to find context specific signatures across cell types using linear regression model [24, 25]. However, the difficulty of interpreting large numbers of states has led to a practical preference for models with lower numbers of states. Typically, the focus is on the discovered states rather than their transition probabilities, unlike more traditional applications of HMM to gene annotation. The assumption is that a limited number of chromatin states and a small number of histone markers combinations covering significant fractions of the genome will capture most of the biologically relevant features.

While useful for predicting chromatin states, HMM-based methods have been relatively less successful when applied to a large number of transcription factors with very restricted, presumably combinatorial binding patterns, which cover small fractions of the genome. Self-organizing maps (SOMs) are an alternative, unsupervised machine-learning method for integratively analyzing such high-dimensional, comparatively sparse data. SOMs consist of individual units (which can be thought of as either neurons or mini-clusters) arranged on a scaffold that is trained with data to capture the high-density parts of high-dimensional datasets while preserving similarity relationships, i.e. data that is close in the input will also be close on the SOM. Chromatin SOMs identify TF-TF localization and co-binding pairs of TFs across cell types and tissues [28]. SOMs have been trained on the same data as chromHMM and Segway in ENCODE, namely histone modification markers, RNA Polymerase II, and CTCF. These are then overlaid post-training with additional data such as EP300 ChIP-seq signals to confirm cell-type specific and commonly shared enhancer activity of groups of DNA segments [29]. For example, a trained SOM would distinguish open chromatin regions from promoters and enhancers based on their difference in H3K4me3 and H3K4me1 signal density (Figure 1.3). The individual units in SOM maps can be grouped into map regions called metaclusters [29, 30], which can then be analyzed for their ChIP-seq signal enrichments and used to automatically identify sets of

potentially co-regulated regions [29]. Once a unit or metacluster of interest has been identified, proximal genes can be associated with bound DNA elements by using tools like GREAT [31] and Homer [32], and their gene expression profiles can be correlated [24] and visualized together with DNA element activity. Co-associated genes can then be analyzed for gene ontology enrichment using GREAT and Homer but other tools such as DAVID [33] and Metascape [34], can also be applied to identify potential functional enrichments. While SOM does not impose a state transition model like HMMs, it recovers similar high-level states at the level of metaclusters, but allows for further granular mining of "microstates" corresponding to very specific chromatin profiles in individual such as distinct combination of transcription factors that are present in small sections of the genome [29]. SOM can therefore be used to deeply data-mine for complicated relationships in highly-dimensional ChIP-seq datasets.

1.4 Incorporating chromatin accessibility with ChIP-seq

Eukaryotic chromatin is tightly packaged into nucleosomes and the positioning of nucleosomes regulated by TFs and histone modifications show dynamic patterns during cell differentiation and development [35]. Specific proteins, often called pioneer factors, can control nucleosome repositioning via recruitment of chromatin remodelers, thus exposing cis-regulatory elements to lineage- or cell type-specific TFs that activate or repress gene expression [15, 36]. Additionally, nucleosomes with H3.3/H2A.Z histone variants show hypermobility, which make them less stable and the DNA more easily accessible for TFs binding [37, 38]. Histone-depleted regions are referred to as open chromatin (Figure 1.1), and several sequencing assays have been developed to capture chromatin accessibility directly at high resolution such as DNase-seq [39-41], FAIRE-seq [42, 43] and ATAC-seq [44]. MNase-seq [35, 45, 46] is a related assay for identifying DNA regions occupied by nucleosomes instead of detecting open chromatin regions directly. DNase- and ATAC-seq depend on enzymatic digestion and Tn5 transposase insertion, respectively, to detect open chromatin regions *in vivo*. Both of them have a higher signal-to-noise ratio than the other methods, and ATAC-seq has become increasingly popular because of its ease of use. All of these methods need deep sequencing (about 50-100 million reads per sample) to get accurate, high-resolution profiles. The basic computational pipeline for open chromatin assays includes reads alignment, visualization for QC, peak calling, and footprint analysis for DNase- and ATAC-seq or nucleosome profiling for MNase- and ATAC-seq (each step has been

reviewed in detail elsewhere) [35, 47]. Specific software packages have been developed to detect signal-enriched regions for each assay. For example, Hotspot [48] detects DNase I hypersensitive (DHS) regions for DNase-seq; GeneTrack [49] and DANPOS [50] do nucleosome calling for MNase-seq; NucleoATAC [51] calls nucleosome positions and occupancy for ATAC-seq. In addition, tools developed for ChIP-seq and DNase-seq peak calling also work effectively for ATAC-seq, such as MACS [52], Hotspot [48] and Homer [32]. DNAse-seq open chromatin data have been used alongside histone modification ChIP-seq data to define chromatin states using HMMs and SOMs in the ENCODE project [10, 29].

Deeper sequencing of open chromatin data to 200-500 million reads per sample can also be used to detect TF binding occupancy "footprints" at nucleotide resolution [35]. The ability of DNase- and ATAC-seq to perform footprint calling is the consequence of TF occupancy protecting DNA from nuclease cleavage and Tn5 transposition, which results in small stretches of fewer cuts within otherwise open regions. The sequences within these footprints can be compared to known motifs for identification [53-55]. The power of footprinting is that a single experiment can identify the binding sites for hundreds of transcription factors, a task that would be still gargantuan with hundreds of TF-specific ChIP-seq experiments. However, many TF motifs are very similar to each other and can be difficult to distinguish based on sequence alone. For these cases, ChIP-seq of selected TFs can be used to validate the footprints when they are critical to the inferred gene regulatory networks [56]. Additionally, histone modification ChIP-seq data can be mapped to open chromatin peaks to confirm the chromatin state of regulatory elements [44, 57-59]. The profiling of chromatin accessibility and TFs/histone occupancy have revealed that cis-regulatory elements show both transitory and stable activity during development and differentiation process for different lineages [60, 61]. Integrative analysis of chromatin accessibility and TFs occupancy from ChIP-seq has revealed that the two processes are not necessarily synchronous. Some TFs commonly referred to as pioneer factors can induce and remodel chromatin accessibility [62-65]. On the other hand, chromatin can be opened and activated before TFs binding [48] or closed well-after the TF has ceased to be bound. Since open chromatin assays such as ATAC-seq are relatively easier to do and require less starting material than ChIP-seq, we expect that an increasing number of studies will start with open chromatin

data followed with selected ChIP-seq for TFs and/or histone modifications. These data will be analyzed integratively with additional packages developed to facilitate their joint analysis.

1.5 Integrative analysis of gene expression with ChIP-seq

Most users of ChIP-seq data are interested in understanding the impact of transcription factor binding or histone modifications on the expression of nearby genes, and therefore ChIP-seq and RNA-seq are analyzed jointly to estimate this effect [6, 7, 14, 66, 67]. In the ideal case, a high ChIP-seq signal of a transcriptional activator would be found near highly expressed genes, while a high ChIP-seq signal of a repressor would be found near silenced genes. In another case, differentially expressed genes are first identified and classified into up- or down- regulated genes between different experimental treatments. Then, differential TF and epigenetic occupancy are correlated with differential gene expression levels. TF binding peaks and histone modification-enriched regions are associated with genes based on which gene is nearest, or using a particular distance radius. However, TF and epigenetic occupancy alone are seldom effective in predicting nearby target gene expression level accurately, because (a) they cannot account for post-transcriptional turnover of the transcript, (b) it is difficult to accuratly associate ChIP-seq peaks with their target genes, and (c) we may not have the ChIP-seq data for all of the TFs controlling the expression of the target genes. One study has reported that the binding signal of twelve embryonic stem cell (ESC) TFs can explain 65% of the variance in mES gene expression and the correlation coefficient between predicted and observed gene expression is 0.8 [68]. However, the predictive power of the same set of TFs in differentiated mES decreased dramatically (r=0.2) [68], and they can only explain 30% of gene expression variance in GM12878 [69]. In addition, while histone modifications alone can explain high gene expression variance in human CD4 T+ cells (r=0.7), combinatorial histone modification combinations show different predictive power [70]. Inferring the effect of TFs on expression is complicated by the fact that TFs may activate a subset of target genes but repress others. Furthermore, TFs and histone marks have different power in predicting gene expression levels [71-73]. Thus, this approach is only practical for predicting gene expression in well-studied systems, where there are plenty of TFs and histone modifications datasets available that can be selected based on biological significance.

Efforts have been made to integrate chromatin accessibility data and ChIP-seq together to predict gene expression, and this combination is more accurate than using ChIP-seq alone [69]. However, the asynchrony between binding and chromatin accessibility also accounts for the less than perfect correlation between changes in these metrics and changes in gene expression. This is because transcription is the sum total of the multitude of effects of chromatin remodelers, TFs co-occupancy, different combination of histone marks and even DNA methylation, which are laborious to capture and profile simultaneously. Using regression models of RNA-seq, ChIP-seq and chromatin accessibility data, gene expression can be predicted from TFs/histone binding [69] and ChIP-seq-identified TF binding motifs in open chromatin regions [74]. Mixed linear models of gene expression correlated with chromatin accessibility corrected with ChIP-seq TF binding can predict TFs triggering or binding prior to chromatin remodeling [75]. Furthermore, TF-TFs co-occupancy can be predicted using support vector machines trained on open chromatin, histone markers, and TFs ChIP-seq data [76]. The predictive power of integrated chromatin feature data can also be extended to the inference of gene regulatory networks. In one recent study [77], chromatin feature data was not only used to predict gene expression, but also to predict the activation status of regulatory elements and further infer a context-specific gene regulatory network. The expression of TFs, target genes, and chromatin remodelers as well as the accessibility of cis-regulatory elements and TFs motifs in regulatory elements are integrated together and fed into a statistical Paired Expression and Chromatin Accessibility (PECA) model. This model predicts active cis-regulatory elements, TF expression, and expression of related target genes within the same context-specific gene regulatory network, which are confirmed by knocking down key TFs in the network [78]. Although combining TFs/histone modifications ChIP-seq and chromatin accessibility data is an effective strategy for predicting gene expression and inferring gene regulatory networks, more software packages and platforms are still needed to be developed for integrating data from different functional assays. We expect that the next generation of packages will improve the predictive power of ChIP-seq for gene expression prediction using ever-more sophisticated and robust statistical methods.

1.6 Incorporating long-range chromatin interactions with ChIP-seq

Most gene regulatory analyses only consider the effects of histone modifications and TFs on the nearest gene, thus not taking into account long-range interactions of cis-regulatory elements

with more distal genes. Promoters and enhancers are physically coupled with target genes by chromatin loops mediated by transcription factors, cohesin, mediator, and some non-coding RNAs to control gene expression [79-83]. A single promoter or enhancer can interact with multiple enhancers or promoters within the same chromatin loops [10, 84]. Recruitment of cofactors such as EP300 by TFs ultimately mediate these complex promoter-enhancer interactions. Chromosome conformation capture (3C)-based sequencing assays such as Hi-C [85, 86] and ChIA-PET [87] can be used to detect these long range interactions. In particular, ChIA-PET (Chromatin interaction analysis by paired-end tag sequencing) combines ChIP and 3C-based methods to detect chromatin interactions between sites bound by specific proteins such as RNA Polymerase II or CTCF on a genome-wide scale [79, 88], but requires hundreds of millions of cells as starting materials. Compared with ChIA-PET, Hi-C can capture all sites interactions in the genome but at the expense of deep sequencing, as it needs at least a billion reads to achieve 1 kb resolution in mammalian genomes [85, 86, 89]. ChIA-PET can capture promoter-enhancer, promoter-promoter, and enhancer-enhancer interactions that involve RNA Polymerase II directly, while Hi-C identifies TADs (topologically associated domains) in chromatin structure. Newer methods such as HiChIP [90] and PLAC-seq [91], combine the advantages of ChIA-PET and Hi-C to capture long-range interactions more efficiently and accurately. 3C-based methods and the basic computational analysis pipelines for each of the techniques have been reviewed previously [92, 93].

Although the mechanisms of long-range interactions are not completely understood, it is known that TFs and histone modifications are actively involved in the interactions and may help alter the chromatin structures [94]. By coupling ChIP-seq with long-range interaction data, studies find that TFs such as CTCF and YY1, are highly enriched in interacting loci or the boundaries of TADs in long-range interactions [86, 88, 89, 95- 101]. Multiple studies have reported that CTCF can also co-bind with other TFs to form lineage - or cell type - specific long-range interactions and activate context-specific gene expression [101- 104]. It has also been shown that disruptions to TF binding at TADs boundaries or cis-regulatory elements, whether caused by mutations, methylation of TF binding sites, or deletion of a transcription factor, can cause remodeling of chromatin interactions and abnormal expression of target genes, which may lead to disease [105, 106]. To integrate ChIP-seq data with ChIP-based long-range interaction

data (i.e. ChIA-PET), peak callers are used to find TF co-binding and histone modifications in anchor sites of PETs [79, 107, 108]. For example, RNA Pol II ChIA-PET detects promoter-promoter and enhancer-promoter interactions directly. Enhancers or promoters can be further confirmed by comparing ChIP signal between H3K4me3 and H3K4me1 modifications [79]. In addition, distal enhancers have been thought to interact with promoters via cohesin-associated CTCF-CTCF loops that also insulate enhancers from genes that they are not supposed to target. The insulators are identified by overlapping anchor sites of cohesin ChIA-PET with cohesin and CTCF ChIP signal, while active enhancers are marked with H3K27ac ChIP signal [108]. Specific TFs co-binding patterns involved in the cis-interactions can be detected with ChIP-seq peak calling in the anchor regions [107]. Furthermore, differential promoter-promoter, enhancer-promoter, and enhancer-enhancer interactions can be identified using ChIP-seq of histone modifications and comparing ChIP signal between conditions [79, 108]. For example, CTCF ChIP-seq signal at the anchor sites of cohesin PETs was used to confirm CTCF-CTCF loops in hESC. Although CTCF-CTCF loops are highly conserved between naïve and primed ES cells, the loop structures are different in terms of enhancer-promoter and enhancer-enhancer interactions, as can be seen by comparing H3K27ac ChIP-seq signal between the two states [108]. A popular strategy is to segment the genome into TADs using HiC when available, or predicting TADs using CTCF and/or Cohesin component ChIP-seq in order to constrain interactions between TFs and *cis*-regulatory elements within these ~100-1000kb regions [109]. By matching ChIP-seq peaks of CTCF and cohesin complex proteins to non- ChIP-based long-range interaction data, like Hi-C, TADs boundaries can be defined and TADs can be segmented into sub transcription units more accurately [108]. Although TADs have relatively conserved segmentation structure during cell development and differentiation [105, 110], the intra-TADs interactions and epigenetic states of TADs are less stable in terms of outside stimulus and differentiation conditions [110, 111]. By comparing normalized ChIP signal of histone modifications within TADs before and after treatment, it is possible to define activated or repressed TAD states that are then correlated with differentially expressed genes within the same TADs. As ChIP-seq has been performed routinely in many labs and large consortiums such as the ENCODE [10] and modENCODE [112] projects, many ChIP-seq datasets are available for public use. Frequent chromatin interaction loci (“hubs”) and TAD boundaries can be predicted accurately from published histone ChIP-seq data integrated with customized Hi-C [113].

Interestingly, a recent study shows that cohesin loss causes loop domains to disappear based on Hi-C data, but CTCF and histone modification ChIP-seq data shows that their patterns are unaffected. The disappearance of loop domains only affects the expression levels of a small percentage of genes, which suggests that cohesin-mediated loops only have modest effects on transcription for most genes and those super-enhancers of genes seem to keep their activity intact without cohesin looping [114]. Thus, given the complex relationship between long-range interactions and gene expression, more studies applying Hi-C/ChIA-PET coupled with ChIP-seq are needed to understand the exact role of chromatin loops in gene expression and to further categorize genes based on their response to the disruption of loop formation.

1.7 Predicting regulatory sequence variants by integrative analysis with ChIP-seq

Sequence variants or single nucleotide polymorphisms (SNPs) are known to be associated with genetic traits and diseases [115, 116]. Most SNPs identified by genome-wide association studies (GWAS) as associated with traits or diseases are found outside of protein-coding regions, with the majority of these non-coding SNPs located in open chromatin regions [117, 118]. As open chromatin regions map to enhancers and promoters, non-coding SNPs in the accessible regions may interrupt or strengthen protein-DNA interactions by introducing sequence variants into binding motifs, and thus causing gene expression and traits to vary between individuals. Indeed, multiple studies have reported that many disease-causing nucleotide changes are in TFs binding sites and affect TFs-DNA binding events [10, 119- 131]. The interruption in TFs binding can not only influence proximal gene expression, but also that of distal genes [122, 125, 129, 132]. However, only a minority of differential TF-DNA binding causes can be explained by sequence variation in binding motifs [133]. Besides, allelic occupancy profiling of more than 20 TFs using ChIP-seq data revealed that only a small proportion of these events have sequencing variants in binding motifs for specific TFs [134]. Although local variants in motifs are not necessarily affecting specific TFs binding, sequence context is still an important source of differential TFs-DNA binding. For example, proximal sequence changes may influence cooperative TFs-TFs binding [133, 135- 138], and distal variants can affect TF-DNA and TF-TF interactions by changing chromatin state and conformation [133, 139- 141].

Many efforts have been made to integrate ChIP-seq and other experimental data to predict regulatory sequence variants. One of the most straightforward methods is to match SNPs to known TFs binding motifs from database such as JASPAR [142] and TRANSFAC [143], or to look for putative TFs bindings sites using hidden Markov models (HMM). The binding affinity score can be calculated based on a position weight matrix (PWM) representation of the motif. When comparing the motif affinity score between two alleles, a greater motif score difference indicates that the variant is more likely to be regulatory [144- 146]. However, these methods rely on known TFs binding sites and do not leverage the predictive power of chromatin signatures to filter out a large set of false positive predictions. Recent studies have successfully integrated ChIP-seq and DNase-seq data into predictive analyses without relying on TFs binding motifs databases [147, 148]. In these studies, peak calls from ChIP-seq and DNase-seq are scanned for k-mers of a given length and the putative regulatory sequences are used to train a support vector machine (SVM) to predict the regulatory power of any k-mer sequence. The weighted sequences can then be used to predict the impacts of single nucleotide changes on regulatory activity in the variant sequences [147]. Another version of this method is to weigh the predictive power of k-mer sequences and compute DNase-seq covariates from ChIP-seq data using regression methods. The trained k-mers and DNase-seq signals are then used to predict ChIP-seq binding signals at two alleles. By comparing the predicted ChIP-seq signal between the reference and variant alleles, the variant can be predicted to be regulatory or not [148]. Other studies have applied deep learning methods such as convolutional neural nets in order to more comprehensively integrate sequence variants, chromatin states, chromatin accessibility and even RNA-binding protein data to predict which regulatory variants will be functional [149, 150]. Some regulatory variants are disease-associated and we can predict the effect of those variants on the binding affinity of transcription factors by evaluating the change in score for the motif [149]. We expect additional work on the development algorithms that can predict potential causal disease variants from the integration of functional genomics data, which will require experimental validation. The validation data in turn will be of great value for training the next set of methods to analyze variants from ChIP-seq data.

1.8 ChIP-seq integrative analysis in the era of low cell count and single-cell genomics

ChIP-seq has been the standard method for identifying genome-wide protein-DNA interactions when a specific antibody is available [151]. However, the traditional ChIP-seq technique requires a large amount of starting material (preferably more than ten million of cells) to get high resolution profiles, which limits its applicability for small organisms, rare cell types, and single cells. Efforts have been made to optimize the ChIP-seq protocol for a low amount of starting materials, which successfully detect TFs binding signals with as few as 5,000 cells [152] and H3K4me3 binding signals with only 500 cells [153]. Although these methods generate binding profiles at a good resolution with a small number of cells, the experimental procedures are still time consuming and costly. Due to the need for high PCR amplification in the low-input ChIP protocols, the number of identical aligned reads need to be carefully corrected for during data analysis. The low-input ChIP-seq peaks can also be compared with open chromatin regions from ATAC-seq to show high correlation between enhancer histone modifications and open chromatin regions. By doing motif discovery analysis, people also identify lineage specific TFs binding to lineage specific open chromatin regions. TFs expression levels have been observed to correlate with differential open chromatin regions accessibility across cell types [153]. Another advancement in low-input ChIP-seq technique is to couple ChIP and Tn5 transposase tagmentation to add sequencing adapters to the beads-bound chromatin in a single step [154]. This protocol is both fast as well as cost-effective and it successfully identifies TFs binding with 100,000 cells and histone markers with 10,000 cells. The ChIP signal needs to be normalized to genomic tagmentated DNA to remove tagmentation bias. However, the protocol also benefits from Tn5 insertions in open chromatin regions to detect transcription factor footprints and nucleosome positioning [154].

Single-cell epigenetics is a rapidly emerging area because of the development of new techniques [155, 156]. While we know that TF binding, histone modifications, chromatin accessibility, DNA methylation, and long-range interactions work together to generate context-specific patterns, these results are primarily based on experiments with bulk samples. Individual cells may have different epigenetic patterns that influence their random behaviors [157]. Therefore many single-cell epigenetics assays [156] have been developed to study this, including scATAC-seq [158, 159], scHi-C [160], scBS-seq [161- 163]. In addition, several techniques have been developed to couple multiple functional assays together to get transcriptomic and epigenetic

data from the same cell simultaneously [164- 166]. Compared with these methods, single-cell ChIP-seq seems more limited, due to the technical difficulties of working from so little material. Only one protocol has successfully performed ChIP-seq at single cell level [167], identifying hundreds of histone modification peaks per cell. The authors successfully distinguished three cell types by doing unsupervised hierarchical clustering and identifying subpopulations with different chromatin signatures. However, the low input and antibody sensitivity cause single cell ChIP-seq to suffer from high technical variance and low sensitivity across individual cells. Similarly, recent advances in single-cell ATAC-seq [158, 159] successfully identified individual open chromatin regions in single cells, with the downside of low signal-to-noise compared with bulk ATAC-seq. However, scACTAC-seq reads are aggregated to be validated when comparing with bulk ATAC-seq data, which shows less technical variance and higher sensitivity compared with single cell ChIP-seq. The high background IP noise probably limits scChIP-seq to histone modifications, and extensive computational analysis needs to be carried out to remove the noise in peak calling. The strategy used for now is to segment ChIP-ed DNA for peak calling for individual cells and then cluster cells based on fractions of reads in known ChIP peaks from bulk samples. Thus, the analysis is still performed at low-input level rather than the true single-cell level [9]. Future studies need to develop methods to remove IP noise and improve solid peak calling in individual cells, since bulk analysis methods cannot be applied directly in single-cell assays. We can further expect that methods will appear combining single-cell ChIP-seq and single-cell RNA-seq from the same cells, which will open up new possibilities when working from mixed cell types and difficult-to-obtain samples.

1.9 Future direction and conclusion

ChIP-seq has become the standard method for profiling protein-DNA binding over the last decade, and it has been actively integrated with newer functional genomics assays such as RNA-seq, DNase/ATAC-seq and Hi-C/ChIA-PET in order to generate models of gene regulation. In the best studies, the integrative analysis is validated with a series of validation experiments to show that the binding of particular TFs is critical for target genes expression. As ATAC-seq and RNA-seq protocols continue to become easier, we expect that ChIP-seq will be routinely integrated with these functional genomics assays. While most current studies compare different "static" cell types, transcription changes temporally in response to stimuli that involve changes in

transcription factor binding, and will become more often the subject of study using ChIP-seq during development and/or stimulation. ChIP-seq following perturbations will also become more routine, and will need to be integrated when building predictive models to identify potentially active cis-regulatory elements and key TFs, which would guide experimental validation and will feed back into further model building.

Another challenge in ChIP-seq integrative analysis will be how to incorporate long-range interaction and gene expression data into the chromatin state analyses that are being done with HMMs and SOMs. Currently, all of these analysis include multiple ChIP-seq datasets and can incorporate chromatin accessibility, but are not designed to incorporate connectivity between distant regions or gene expression data as part of their training as opposed to post-training analysis and annotation. A challenge is that while at least ChIA-PET and HiC are working in a similar "feature space" of chromatin as ChIP-seq, regular RNA-seq is measuring the steady state of transcripts, which is affected by several post-transcriptional processes such as mRNA turnover mediated by microRNAs. Since chromatin will always be more predictive of transcriptional initiation, it may be more fruitful to compare the predicted models of expression to GRO-seq and other measurements of transcriptional activity than regular RNA-seq.

In recent years, ChIP-seq techniques for low input materials have been developed to expand its applications to rare tissues or cell types, and even single-cell studies. Other functional genomics assays have also been developed at single-cell level to answer new biological questions. However, the integrative analysis of single-cell ChIP-seq with these functional genomics assays in single cells is a difficult challenge. One reason is that the experimental protocols to capture protein binding, transcriptomes, and DNA methylation data from the same cell are still not available. However, it may still be worthwhile to integrate data from scChIP-seq and other functional genomics assays in different individual cells from the same pool based on the assumption that protein binding profiles would match to the gene expression profiles from the assay because these cells are from the same pool. Once protocols are available to do scRNA-seq and scChIP-seq from the same single cell, algorithms will need to be developed to integrate these single-cell data types together in order to understand the connection between binding and gene expression heterogeneity in subsets of a cell population. As single-cell data is sparser than

bulk data, new statistical methods and tools are required for integration. In a hopefully not-so-distant future where robust single-cell ChIP-seq and RNA-seq are practical, they could become the method of choice for studying samples where the amount of material or the heterogeneity of the population make the bulk version of these experiments less attractive.

1.10 Figures

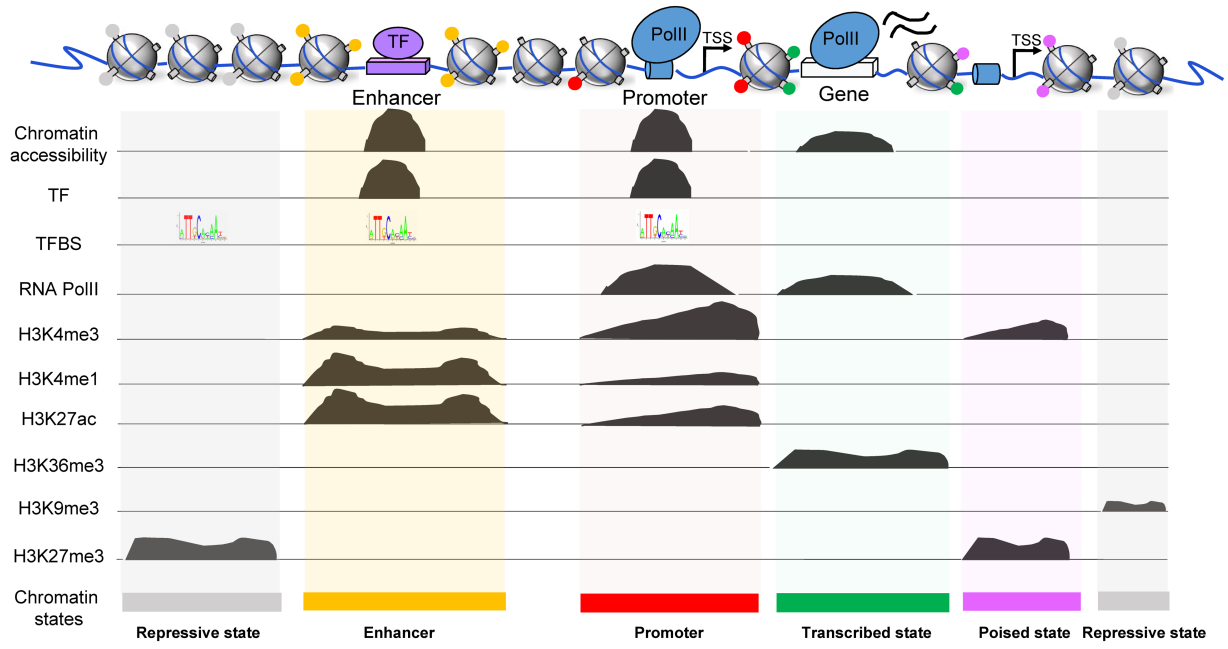


Figure 1.1. Chromatin states are defined by different combinations of histone modifications, transcription factors and RNA Pol II binding. In this example, a typical repressive state (gray) is characterized by high H3K27me3 signal or H3K9me3 signal, an enhancer state (yellow) would show a high occupancy ratio of H3K4me1 to H3K4me3 as well as high H3K27ac, and the promoter state (red) would show a high occupancy ratio of H3K4me3 to H3K4me1 as well as RNA Pol II binding at the promoter, whereas poised promoter state (magenta) would show the occupancy of H3K4me3 and H3K27me3 bivalent modifications. Actively transcribed region (green) is characterized by a high occupancy of H3K36me3 with some RNA Pol II binding along the gene body.

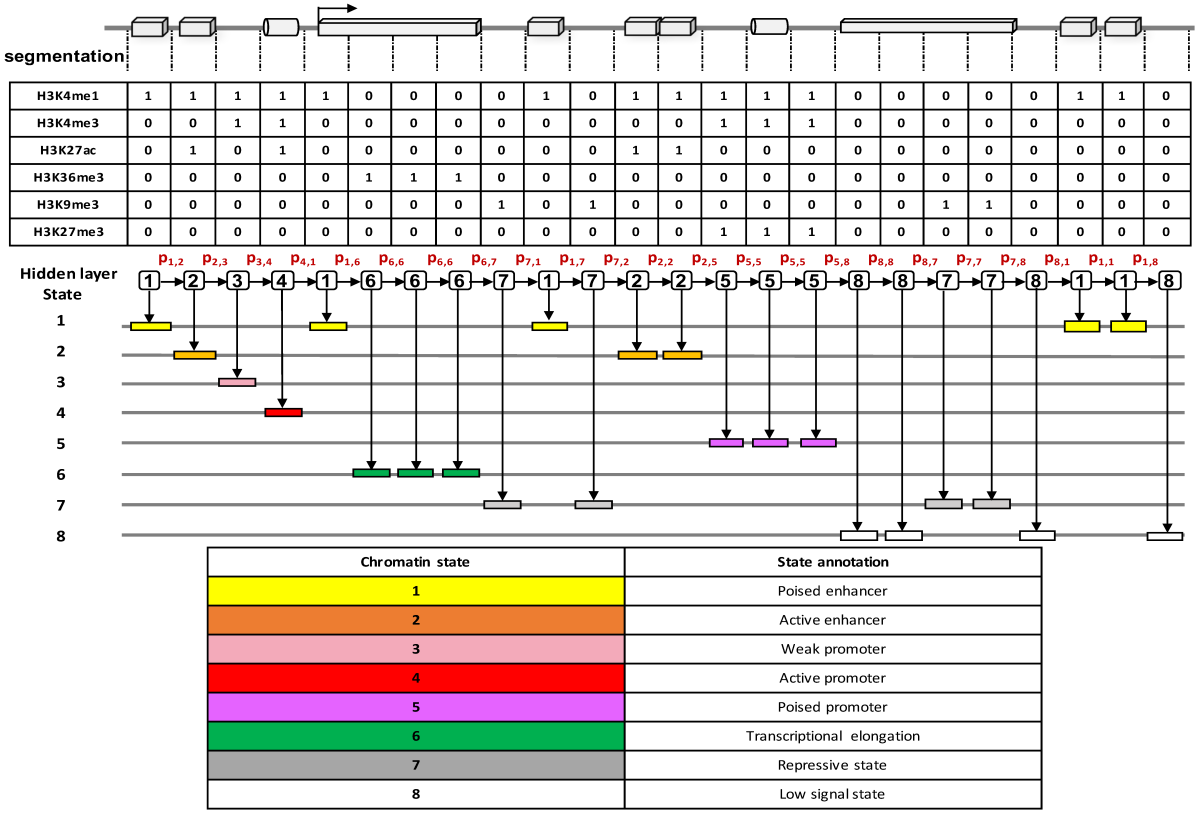
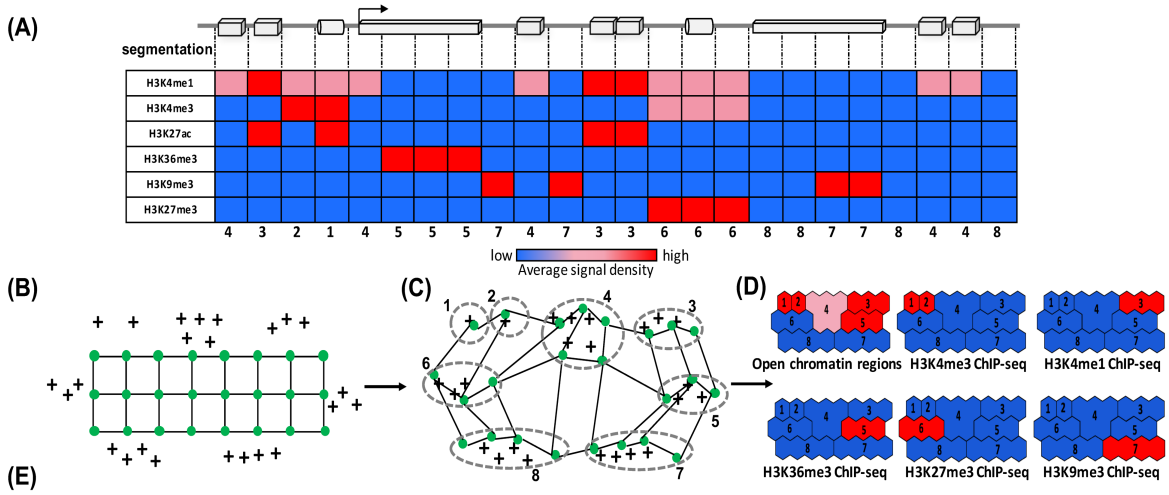


Figure 1.2. Graphical structure of annotating chromatin states using a Hidden Markov Model (HMM) method such as ChromHMM. The genome is split into non-overlapping segments and ChIP-seq signal for histone modifications are binarized (0 or 1) and collected for each segment, which are further built into input matrix for HMM training. The hidden state of the current segment is dependent on the state of the previous one and the transition probabilities (in red) of changing from one state to another are learnt from training on the input matrix. ChromHMM outputs trained hidden states for each segmentation, which are then interpreted as chromatin states based on the chromatin profile and gene annotations, such as active promoter/enhancer, transcriptional elongation, or repressive states.



Metacluster	H3K4me1	H3K4me3	H3K27ac	H3K36me3	H3K9me3	H3K27me3	Annotation
1							Active promoter
2							Weak promoter
3							Active enhancer
4							Poised enhancer
5							Transcriptional elongation
6							Poised promoter
7							Repressed region
8							Low signal region

Figure 1.3. Graphical structure of annotating chromatin states using Self-Organizing Maps (SOMs)**. (A)** The genome is split into non-overlapping segments and ChIP-seq signal for histone modifications are collected for each segment to build an input matrix for SOM training, where each segment represents a vector of signal. **(B)** The map is initialized with units (green dots) and the input signal vectors (black plus signs) are spread around the map in high dimensional space. **(C)** For each step of training, a signal vector is selected and the closest unit is found. Then the unit is pulled as well as other units around it towards to the selected signal vector, which causes units cluster together to represent signal vectors sharing same features. **(D)** The trained SOM map can be divided into metacluster regions (metacluster 1-8) based on combinations of signal enrichments to recover regions that are high in H3K4me3 and open chromatin (promoters), high in H3K4me1 and open chromatin (potential enhancers), high in H3K36m3 (transcribed regions) or high in H3K27me3 and H3K9me3 (repressed). **(E)** Metaclusters are further manually assigned chromatin state labels based on annotations and the combinations of signal enrichments just as in the HMM case.

1.11 References

[1] Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. *Science* 2000; **290**: 2306-2309.

[2] Johnson DS, Mortazavi A, Myers RM, et al. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 2007; **316**: 1497–1502.

[3] Barski A, Cuddapah S, Cui K, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 2007; **129**: 823–837.

[4] Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007; **4**: 651–657.

[5] Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature 2007;* **448**: 553–560.

[6] Park PJ. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009; **10**: 669–680.

[7] Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* 2012; **13**: 840–852.

[8] Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012; **22**: 1813–1831.

[9] Nakato R, Shirahige RNK. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform* 2017; **18**: 279–290.

[10] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.

[11] Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317–330.

[12] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010; **28**: 817–825.

[13] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012; **9**: 215–216.

[14] Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011; **12**: 7–18.

[15] Calo E, Wysocka J. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* 2013; **49**: 825–837.

[16] Eddy SR. What is a hidden Markov model? *Nature biotechnology* 2004; **22**: 1315-1316.

[17] Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 2012; **13**:329-342.

[18] Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* 2016; **44**: D279-D285.

[19] Day N, Hemmaplardh A, Thurman RE, et al. Unsupervised segmentation of continuous genomic data. *Bioinformatics* 2007; **23**: 1424–1426.

[20] Hoffman MM, Buske OJ, Wang J, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012; **9**: 473–476.

[21] Biesinger J, Wang Y, Xie X. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics* 2013; **14 Suppl 5**: S4.

[22] Sohn KA, Ho JWK, Djordjevic D, et al. HiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics* 2015; **31**: 2066–2074.

[23] Marco E, Meuleman W, Huang J, et al. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nat Commun* 2017; **8**: 15011.

[24] Zhang Y, An L, Yue F, et al. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* 2016; **44**: 6721–6731.

[25] Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res* 2017; **45**: 9823–9836.

[26] Yen A, Kellis M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat Commun* 2015; **6**: 7973.

[27] Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 2015; **33**: 364–376.

[28] Xie D, Boyle AP, Wu L, et al. Dynamic trans-acting factor colocalization in human cells. *Cell* 2013; **155**: 713-724.

[29] Mortazavi A, Pepke S, Jansen C, et al. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. *Genome Res* 2013; **23**: 2136–2148.

[30] Longabaugh WJR, Zeng W, Zhang JA, et al. Bcl11b and combinatorial resolution of cell fate in the T-cell gene regulatory network. *Proc Natl Acad Sci* 2017; **114**: 5800-5807.

[31] McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010; **28**: 495–501.

[32] Heinz S, Benner C, Spann N, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010; **38**: 576–589.

[33] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: 1–13.

[34] Tripathi S, Pohl MO, Zhou Y, et al. Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe* 2015; **18**: 723–735.

[35] Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014; **7**: 33.

[36] Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. *Genes Dev* 2014; **28**: 2679–2692.

[37] Creyghton MP, Markoulaki S, Levine SS, et al. H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is Necessary for Lineage Commitment. *Cell* 2008; **135**: 649–661.

[38] Jin C, Felsenfeld G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes Dev* 2007; **21**: 1519–1529.

[39] Boyle AP, Song L, Lee BK, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* 2011; **21**: 456–464.

[40] Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein– DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 2009; **6**: 283-289.

[41] Neph S, Vierstra J, Stergachis AB, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012; **489**: 83–90.

[42] Simon JM, Giresi PG, Davis IJ, et al. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc* 2012; **7**: 256–267.

[43] Bianco S, Rodrigue S, Murphy BD, et al. Global mapping of open chromatin regulatory elements by formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq). *Methods in Molecular Biology 2015;* **1334**: 261–272.

[44] Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013; **10**: 1213–1218.

[45] Ponts N, Harris EY, Prudhomme J, et al. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res* 2010; **20**: 228–238.

[46] Schones DE, Cui KR, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008; **132**: 887–898.

[47] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 2014; **15**: 709-721.

[48] John S, Sabo PJ, Thurman RE, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 2011; **43**: 264–268.

[49] Albert I, Wachi S, Jiang C, et al. GeneTrack--a genomic data processing and visualization framework. *Bioinformatics* 2008; **24**: 1305–1306.

[50] Chen K, Xi Y, Pan X, et al. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* 2013; **23**: 341–351.

[51] Schep AN, Buenrostro JD, Denny SK, et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* 2015; **25**: 1757–1770.

[52] Zhang Y, Liu T, Meyer CA, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**: R137.

[53] Piper J, Elze MC, Cauchy P, et al. Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res* 2013; **41**: e201.

[54] Pique-Regi R, Degner JF, Pai AA, et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 2011; **21**: 447–455.

[55] Jankowski A, Tiuryn J, Prabhakar S. Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* 2016; **32**: 2419–2426.

[56] Ramirez RN, El-Ali NC. Mager MA, et al. Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. *Cell Systems* 2016; **4**: 416-429.

[57] Prescott SL, Srinivasan R, Marchetto MC, et al. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* 2015; **163**: 68–84.

[58] Wu J, Huang B, Chen H, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* 2016; **534**: 652–657.

[59] Minoux M, Holwerda S, Vitobello A, et al. Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science* 2017; **355**: pii: eaal2913.

[60] Stavreva DA, Coulon A, Baek S, et al. Dynamics of chromatin accessibility and long-range interactions in response to glucocorticoid pulsing. *Genome Res* 2015; **25**: 845–857.

[61] Su Y, Shin J, Zhong C, et al. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat Neurosci* 2017; **20**: 476-483.

[62] Biddie SC, John S, Sabo PJ, et al. Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. *Mol Cell* 2011; **43**: 145–155.

[63] Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature* 2012; **489**: 75–82.

[64] Sherwood RI, Hashimoto T, O'Donnell CW, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014; **32**: 171–178.

[65] Takaku M, Grimm SA, Shimbo T, et al. GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biol* 2016; **17**: 36.

[66] Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009; **6**: S22–S32.

[67] Angelini C, Costa V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front Cell Dev Biol* 2014; **2**: 51.

[68] Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* 2009; **106**: 21521–6.

[69] McLeay RC, Lesluyes T, Cuellar Partida G, et al. Genome-wide in silico prediction of gene expression. *Bioinformatics* 2012; **28**: 2789–2796.

[70] Karlic R, Chung H-R, Lasserre J, et al. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci* U S A 2010; **107**: 2926–2931.

[71] Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 2012; **40**: 553–68.

[72] Cheng C, Yan KK, Yip KY, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 2011; **12**: R15.

[73] Dong X, Greven MC, Kundaje A, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* 2012; **13**: R53.

[74] Natarajan A, Yardimci GG, Sheffield NC, et al. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* 2012; **22**: 1711–1722.

[75] Lamparter D, Marbach D, Rueedi R, et al. Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility. *PLoS Comput Biol* 2017; **13**: e1005311.

[76] Liu L, Zhao W, Zhou X. Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res* 2016; **44**: e49.

[77] Duren Z, Chen X, Jiang R, et al. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci* U S A 2017; **114**: E4914-4923.

[78] Wang J, Jiang W, Yan Y, et al. Knockdown of EWSR1/FLI1 expression alters the transcriptome of Ewing sarcoma cells in vitro. *J Bone Oncol* 2016; **5**: 153–158.

[79] Li G, Ruan X, Auerbach RK, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012; **148**: 84–98.

[80] Harmston N, Lenhard B. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res* 2013; **41**: 7185–7199.

[81] Cavalli G, Misteli T. Functional implications of genome topology. *Nat Struct Mol Biol* 2013; **20**: 290–299.

[82] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014; **15**: 272–286.

[83] Böhmdorfer G, Wierzbicki AT. Control of Chromatin Structure by Long Noncoding RNA. *Trends Cell Biol* 2015; **25**: 623–632.

[84] Sanyal A, Lajoie BR, Jain G, et al. The long-range interaction landscape of gene promoters. *Nature* 2012; **489**: 109-113.

[85] Lieberman-aiden E, Berkum NL Van, Williams L, et al. Comprehensive Mapping of Long-Range Interactions Revelas Folding Principles of the Human Genome. *Science* 2009; **326**: 289–294.

[86] Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; **485**: 376–380.

[87] Fullwood MJ, Liu MH, Pan YF, et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 2009; **462**: 58–64.

[88] Handoko L, Xu H, Li G, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011; **43**: 630–638.

[89] Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014; **159**: 1665–1680.

[90] Mumbach MR, Rubin AJ, Flynn RA, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016; **13**: 919–922.

[91] Fang R, Yu M, Li G, et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* 2016; **26**: 1345–1348.

[92] Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* 2016; **17**: 743–755.

[93] Davies JO, Oudelaar AM, Higgs DR, et al. How best to identify chromosomal interactions: a comparison of approaches. *Nat Methods* 2017; **14**: 125–134.

[94] Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev* 2012; **22**: 79–85.

[95] Donohoe ME, Zhang LF, Xu N, et al. Identification of a Ctcf cofactor, Yy1, for the X Chromosome Binary Switch. *Mol Cell* 2007; **25**: 43–56.

[96] Seitan VC, Faure AJ, Zhan Y, et al. Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res* 2013; **23**: 2066–2077.

[97] Giorgetti L, Galupa R, Nora EP, et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* 2014; **157**: 950–963.

[98] Zuin J, Dixon JR, van der Reijden MI, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci* U S A 2014; **111**: 996–1001.

[99] Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014; **15**: 234–246.

[100] Gómez-Marín C, Tena JJ, Acemel RD, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci* U S A 2015; **112**: 7542–7547.

[101] Beagan JA, Duong MT, Titus KR, et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* 2017; **27**: 1139-1152.

[102] Donohoe ME, Silva SS, Pinter SF, et al. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature* 2009; **460**: 128–132.

[103] Lee J, Krivega I, Dale RK, et al. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. *Cell Rep* 2017; **19**: 2490-2502.

[104] Jerković I, Ibrahim DM, Andrey G, et al. Genome-Wide Binding of Posterior HOXA/D Transcription Factors Reveals Subgrouping and Association with CTCF. *PLoS Genet* 2017; **13**: e1006567.

[105] Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012; **485**: 381–385.

[106] Krijger PH, de Laat W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* 2016; **17**: 771–782.

[107] Zhang Y, Wong CH, Birnbaum RY, et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* 2013; **504**: 306–310.

[108] Ji X, Dadon DB, Powell BE, et al. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* 2016; **18**: 262–275.

[109] Jost D, Vaillant C, Meister P. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr Opin Cell Biol* 2017; **44**: 20–27.

[110] Le Dily F, Baù D, Pohl A, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* 2014; **28**: 2151–2162.

[111] Neems DS, Garza-Gongora AG, Smith ED, et al. Topologically associated domains enriched for lineage-specific genes reveal expression-dependent nuclear topologies during myogenesis. *Proc Natl Acad Sci* U S A 2016; **113**: E1691–E1700.

[112] Celniker SE, Dillon LA, Gerstein MB, et al. Unlocking the secrets of the genome. *Nature* 2009; **459**: 927-930.

[113] Huang J, Marco E, Pinello L, et al. Predicting chromatin organization using histone marks. *Genome Biol* 2015; **16**: 162.

[114] Rao, SS, Huang SC, St Hilaire BG, et al. Cohesin Loss Eliminates All Loop Domains. *Cell*, 2017; **171**: 305-320.

[115] Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* U S A 2009; **106**: 9362–9367.

[116] The 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

[117] Maurano MT, Humbert R, Rynes E, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 2012; **337**: 1190–1195.

[118] Schaub MA, Boyle AP, Kundaje A, et al. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012; **22**: 1748-1759.

[119] Martin DI, Tsai SF, Orkin SH. Increased gamma-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature* 1989; **338**: 435–8.

[120] Matsuda M, Sakamoto N, Fukumaki Y. Delta-thalassemia caused by disruption of the site for an erythroid- specific transcription factor, GATA-1, in the delta-globin gene promoter. *Blood* 1992; **80**: 1347–51.

[121] De Gobbi M, Viprakasit V, Hughes JR, et al. A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter. *Science* 2006; **312**: 1215–1217.

[122] Jeong Y, Leskow FC, El-Jaick K, et al. Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat Genet* 2008; **40**: 1348–1353.

[123] Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010; **466**: 714–719.

[124] Al Zadjali S, Wali Y, Al Lawatiya F, et al. The β-globin promoter −71 C>T mutation is a β+ thalassemic allele. *Eur J Haematol* 2011; **87**: 457–460.

[125] Claussnitzer M, Dankel SN, Klocke B, et al. Leveraging cross-species transcription factor binding site patterns: From diabetes risk loci to disease mechanisms. *Cell* 2014; **156**: 343–358.

[126] Kulzer JR, Stitzel ML, Morken MA, et al. A Common Functional Regulatory Variant at a Type 2 Diabetes Locus Upregulates ARAP1 Expression in the Pancreatic Beta Cell. *Am J Hum Genet* 2014; **94**: 186–197.

[127] Weinhold N, Jacobsen A, Schultz N, et al. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 2014; **46**: 1160–5.

[128] Weedon MN, Cebola I, Patch AM, et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* 2014; **46**: 61–64.

[129] Claussnitzer M, Dankel SN, Kim KH, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* 2015; **373**: 895–907.

[130] Wienert B, Funnell APW, Norton LJ, et al. Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nat Commun* 2015; **6**: 7085.

[131] Wang S, Wu S, Meng Q, et al. FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin. *Sci Rep* 2016; **6**: 19229.

[132] Smemo S, Tena JJ, Kim KH, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 2014; **507**: 371–375.

[133] Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* 2016; **166**: 538-554.

[134] Reddy TE, Gertz J, Pauli F, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* 2012; **22**: 860–869.

[135] Kilpinen H, Waszak SM, Gschwind AR, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 2013; **342**: 744–7.

[136] Siersbaek R, Rabiee A, Nielsen R, et al. Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Rep* 2014; **7**: 1443–1455.

[137] Tijssen MR, Cvejic A, Joshi A, et al. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* 2011; **20**: 597–609.

[138] Domcke S, Bardet AF, Adrian Ginno P, et al. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* 2015; **528**: 575–579.

[139] Ding Z, Ni Y, Timmer SW, et al. Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. *PLoS Genet* 2014; **10**: e1004798.

[140] Waszak SM, Delaneau O, Gschwind AR, et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* 2015; **162**: 1039–1050.

[141] Grubert F, Zaugg JB, Kasowski M, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* 2015; **162**: 1051–1065.

[142] Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016; **44**: D110–D115.

[143] Wingender E, Dietze P, Karas H, et al. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996; **24**: 238–241.

[144] Andersen MC, Engström PG, Lithwick S, et al. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* 2008; **4**: e5.

[145] Macintyre G, Bailey J, Haviv I, et al. Is-rSNP: A novel technique for in silico regulatory SNP detection. *Bioinformatics* 2010; **26**: i524–i530.

[146] Riva A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics* 2012; **13 Suppl 4**: S7.

[147] Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 2015; **47**: 955–961.

[148] Zeng H, Hashimoto T, Kang DD, et al. GERV: A statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* 2015; **32**: 490–496.

[149] Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015; **33**: 831–838.

[150] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015; **12**: 931–934.

[151] Adli M, Bernstein BE. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc* 2011; **6**: 1656–1668.

[152] Shankaranarayanan P, Mendoza-Parra MA, Walia M, et al. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat Methods* 2011; **8**: 565–567.

[153] Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, et al. Chromatin state dynamics during blood formation. *Science* 2014; **345**: 943–949.

[154] Schmidl C, Rendeiro AF, Sheffield NC, et al. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* 2015; **12**: 963–965.

[155] Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015; **16**: 716–726.

[156] Clark SJ, Lee HJ, Smallwood SA, et al. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* 2016; **17**: 1–10.

[157] Sekelja M, Paulsen J, Collas P. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? *Genome Biol* 2016; **17**: 54.

[158] Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015; **523**: 486–490.

[159] Cusanovich DA, Daza R, Adey A, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015; **348**: 910–4.

[160] Nagano T, Lubling Y, Stevens TJ, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013; **502**: 59–64.

[161] Guo H, Zhu P, Wu X, et al. Single-Cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 2013; **23**: 2126–2135.

[162] Smallwood SA, Lee HJ, Angermueller C, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014; **11**: 817–820.

[163] Farlik M, Sheffield NC, Nuzzo A, et al. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep* 2015; **10**: 1386–1397.

[164] Macaulay IC, Haerty W, Kumar P, et al. G & T-seq : parallel sequencing of single- cell genomes and transcriptomes. *Nat Methods* 2015; **12**: 519–522.

[165] Angermueller C, Clark SJ, Lee HJ, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 2016; **13**: 229–232.

[166] Hu Y, Huang K, An Q, et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 2016; **17**: 88.

[167] Rotem A, Ram O, Shoresh N, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* 2015; **33**: 1165–1172.

CHAPTER 2

Rapid intra-individual methylation signatures of diverse early life experiences

Notes: (1) Dr. Noriko Kamei and I equally contributed to the material in this chapter. She designed and performed experiments, as well as providing import suggestions on interpreting the results.

(2) Xinyi Ma performed quality checks and sequenced all experimental samples.

(2) Dr. Hal S. Stern provided valuable suggestions on statistical analysis in this chapter.

(3) Dr. Tallie Z. Baram and Dr. Ali Mortazavi conceived the idea and provided continued support and guidance throughout the project.

Chapter 2

Rapid intra-individual methylation signatures of diverse early life experiences

2.1 Abstract

Genetic and environmental factors interact during sensitive periods early in life to influence mental health and disease *(1-4)*. These influences involve modulating the function of neurons and neuronal networks via epigenetic processes such as DNA methylation *(2-7)*. However, it is not known if DNA methylation changes outside the brain provide an 'epigenetic signature' of early-life experiences in an individual child that may serve as a marker for vulnerability or resilience to mental illness. Here, to obviate the massive variance among individuals, we employed a novel intra-individual approach by testing DNA methylation from buccal cells of individual rats before and immediately after exposure to one week of typical or adverse life experience. We show that whereas inter-individual changes in DNA methylation reflect the effect of age, DNA methylation changes within paired DNA samples from the same individual reflect the impact of diverse neonatal experiences on the individual. The methylome signature of early-life experience is enriched in genes encoding transcription factors and key molecular cellular pathways. Specifically, genes involved in cell morphogenesis and differentiation were more methylated in pups exposed to the adverse environment whereas pathways of response to injury and stress were less methylated. Thus, intra-individual methylome signatures indicate large-scale transcription-driven alterations of cellular fate, growth and function. Our observations in rats--that distinct early-life experiences generate specific individual methylome signatures in accessible peripheral cells--should be readily testable in humans.

2.2 Introduction

Experience, particularly during sensitive periods early in life, leaves indelible marks on an individual's ability to cope with life's challenges, influencing resilience or vulnerability to emotional disorders (*2–4*, *7*, *8*). There is evidence that the mechanisms by which early-life experiences influence the function of neurons and neuronal networks involve modifying the repertoire and levels of gene expression via epigenetic processes (*2–7*, *9–12*). Among epigenetic processes, changes in DNA methylation of individual genes and at the genomic scale have been reported, and these generally correlate with gene expression (*3*, *5*, *13–15*). However, it is not

known if DNA methylation changes might provide a useful 'epigenetic signature' of early-life experiences in an individual child. Such a readily-accessible measure might serve as a biomarker for vulnerability or resilience to mental illness. Obviously, it is not possible to repeatedly sample DNA from brain cells in humans in order to assess DNA methylation changes for predicting and preventing disease. Therefore, current approaches employ peripheral cells including white blood cells (WBC) or buccal swabs (mixed epithelial/WBC), which are available repeatedly and noninvasively. Here we tested the feasibility of using peripheral DNA samples to assess the impact of diverse neonatal experiences on an individual by directly comparing two samples collected at different time points from the same individual rat in groups exposed to distinct early-life experiences with defined onset and duration. We have previously established that these diverse experiences provoke specific phenotypic outcomes later in life (*7*, *16*, *17*). Specifically, we imposed 'simulated poverty' by raising pups for a week (from postnatal day P2 to P10) in cages with limited bedding and nesting materials (LBN). This manipulation disrupts the care provided by the rat dam to her pups and results in profound yet transient stress in the pups, devoid of major weight-loss or physical changes. This transient experience provokes significant and life-long deficits in memory and generates increases in emotional measures of anhedonia and depression (*16–18*).

Here we tested if adversity during a defined sensitive developmental period in rats leads to a detectable epigenomic signature in DNA from buccal-swab cells. We obtained intra-individual epigenomic signatures of early-life adversity using reduced representation bisulfite sequencing (RRBS) (*19*) to identify changes in DNA methylation profiles. Comparisons were made both between two samples from an individual rat (P2 vs P10) and between samples from rats subjected to the two neonatal experiences. We found that assessing the methylation profile of samples enabled detection of age and development effects (*18*, *20*), distinguishing P2 samples from those obtained on P10, but did not separate the two groups of pups based on their experience. In contrast, the changes in DNA methylation in two samples obtained from the same rat enabled clear differentiation of the control vs the adverse experience, likely by obviating large inter-individual variance. Thus, our findings establish the feasibility of identifying markers of adverse experiences that portend risk or resilience to mental illness, with major potential translational impact.

2.3 Results

2.3.1 Methylation changes reflect postnatal ages rather than maternal experiences

We obtained a mix of epithelial & white blood cell DNA from rat pups, on P2 and on P10 from the same pup using buccal swabs (Methods). We obtained buccal swabs rather than peripheral white blood cells for three reasons. First, the swab, lasting seconds, is much less stressful than a painful needle prick to obtain peripheral blood, and this stress might influence methylation in itself. Second, this approach provides a more direct comparison with human studies where ethical reasons preclude needle pain whereas buccal swabs are routinely implemented (*21*, *22*). Finally, several studies found that DNA methylation profiles in buccal swab cells are more similar to patterns from several brain regions than methylation profiles in white blood cells (*22–25*). Following the initial sample collected on P2, rats were exposed to either simulated poverty or to a typical environment for one week, followed by a second sample on P10. We examined for intra-individual epigenomic signatures of early-life adversity and compared both P10 samples from groups with two divergent neonatal experiences as well as the changes in methylation levels between matched samples from the same individual rat (P2 vs P10; Figure 2.1A).

DNA methylation status was assessed using RRBS, with libraries sequenced to an average of 20 million mapped reads, and we reliably detected an average of 482 thousand CpGs in both time points of the same individual (Figure 2.2; Methods). We performed differential methylation analysis between P2 and P10 for each individual and identified 3417 significantly differential methylation regions (DMRs) after coalescing CpGs within 100 basepairs that were shared in at least two individuals from each experience group (Figure 2.1B). These were further analyzed. Specifically, we analyzed the DNA methylation levels of these DMR in P2 and in P10 for both the control and adversity-experiencing (LBN) groups across individuals using k-means clustering and observed substantial changes in DNA methylation level during the one-week interval in both control and LBN (Figure 2.3A). The DNA methylation levels within a given individual clearly distinguished rats at different ages (Figure 2.3A). We further performed principal component analysis (PCA) on the percentage of DNA methylation of these DMR and found that individual samples were clearly separated by age using the first three principal components (up to 62.1% variances explained), indicating the large change in DNA methylation associated with age (Figure 2.3B). Note that this result held when cohort effects were considered (Figure 2.4B,F & 2.5A-C).

These data demonstrate that development and age modify the buccal swab methylome (*20*, *24*, *26*, *27*) in conjunction with experience. We then examined the DMRs with the top weights in PC2, which explained 20.7% of the variance, and was the dominant component distinguishing individuals of different ages (Figure 2.3B). We found that DMRs with reduced methylation level in P10 were associated with genes involved in cellular response to hormones, negative regulation of growth and regulation of kinase activity whereas DMRs with increased methylation level in P10 were enriched with genes in pattern specification processes such as nervous system and mesodermal development (Figure 2.6). Notably, the PCA analyses of the P2 and P10 methylome profiles did not separate the control group from the adversity-experiencing group (Figure 2.3C). Thus, whereas the level of DNA methylation in buccal swabs may denote an epigenetic signature of age, it provides little information about antecedent life experiences.

2.3.2 Intra-individual changes in methylation can distinguish early-life experience

To probe the impact of the early-life adversity experienced by an individual on DNA methylation patterns of the same individual, we explored intra-individual fold changes in methylation (referred to as "delta methylation") rather than the absolute value of methylation levels for each pup by taking advantage of the two samples collected immediately before and after a week of imposed adversity. We clustered and aligned these delta methylation profiles (differential methylation between P2 and P10 defined as $\log_2$(P10/P2)) in both early-life experiences (Figure 2.7A). We then examined the intra-individual methylation changes in detail and found that the patterns of changes in methylation within an individual were distinct depending on group assignment (Figure 2.7A). We performed a PCA analysis on the individual delta methylation samples and the resulting principal components revealed that delta methylation within an individual clearly distinguished the control and LBN groups (Figure 2.7B). Specifically, the first three principal components accounted for 65.0% of the variance and the third principal component (PC3) (4.9% of the variances) distinguished most LBNs from controls. Importantly, the adverse and control experiences differentially reduced or increased levels of methylation in an experience-specific manner. These results indicate that intra-individual changes in methylation-level profiles before and after a defined experience provide a novel epigenetic signature that identifies the nature of the experience.

2.3.3 Differential methylation regions are at the vicinity of multiple transcription factors

The paragraph above demonstrates that levels of DNA methylation in mixed epithelial / white blood cell samples from buccal swab can separate pups by age, whereas the nature of methylation changes in the same individual (delta methylation) distinguishes different early-life experiences. We examined the relative contribution of individual DMRs to the overall difference in PC3, and, guided by the slope of the weight distribution selected a cutoff threshold at +/- $2x10^{-2}$ (Figure 2.8A) to identify the 346 largest positive weights that account 254 DMRs, which are associated with 246 genes (Figure 2.8A & 2.9). The methylation regions with maximal positive weights are thus major contributors to the differential methylation profiles of early-life adversity (LBN) compared to typical development, and genes that have generally increased DNA methylation after the LBN experience typically denote reduced expression. Gene Ontology (GO) analysis of these "positive weight" genes identified enrichment in terms associated with cell and organ development and cell-morphogenesis (Figure 2.8B). Inspection of "positive weight" genes contributing to the adversity-provoked methylome signature uncovered a strong enrichment of genes involved in growth and response to growth factors (26/246 annotated genes; 10.6%), pathways of injury, inflammation, and death (25 of 246 annotated genes; 10.2%), as well as transcription factors (15/246; 6.1%).

A strongly regulated program of gene expression was also suggested when the same approach was applied to the top 311 negative weights (241 DMRs) associated with 225 genes that contributed most to the methylome signature associated with a typical developmental epoch. First, GO analysis indicated enrichment in genes involved in cell morphogenesis and differentiation. Notably, inspection of the individual "negative weight" genes uncovered likely mechanisms for a regulated gene expression program: 17.8% of the genes in this group (40/225: 17.8%) were transcription factors. Indeed, transcription factors accounted for 30% (7/23) of the top-contributing genes (genes associated with DMRs having weights more significant than $-5x10^{-2}$ in Figure 2.8A) to the typical methylation phenotype. Furthermore, 53 of the 225 negative weight genes (23.6%) were involved in cell morphogenesis, cytoskeleton stability and growth, and cell-cell communication.

Taken together, these findings suggest that early-life experiences set in motion genome-wide changes in methylation patterns of crucial gene-sets, including transcription factors. Early-

life adversity associates with altered methylation of gene-sets involved in growth and differentiation as well as response to injury and death. These changes are likely driven by molecular signals, including hormones and nutrients that modulate the complex enzymatic processes that govern DNA methylation status (*28–30*).

2.4 Conclusion

In summary, we find here that comparing cohort-wide DNA samples obtained at different developmental ages reveals the signature of age and development on the peripheral methylome, as widely reported. However, these inter-individual analyses do not distinguish the divergent impacts of diverse experiences that take place during the intervening developmental epoch. By contrast, here we demonstrate that paired samples from the same individual before and after an adverse or typical developmental experience enable clear distinction of each of these experiences: we identify epigenetic ‘scars’ and ‘kisses’ that, at least in the rodent model, precede and predict later-life emotional functions.

2.5 Figures

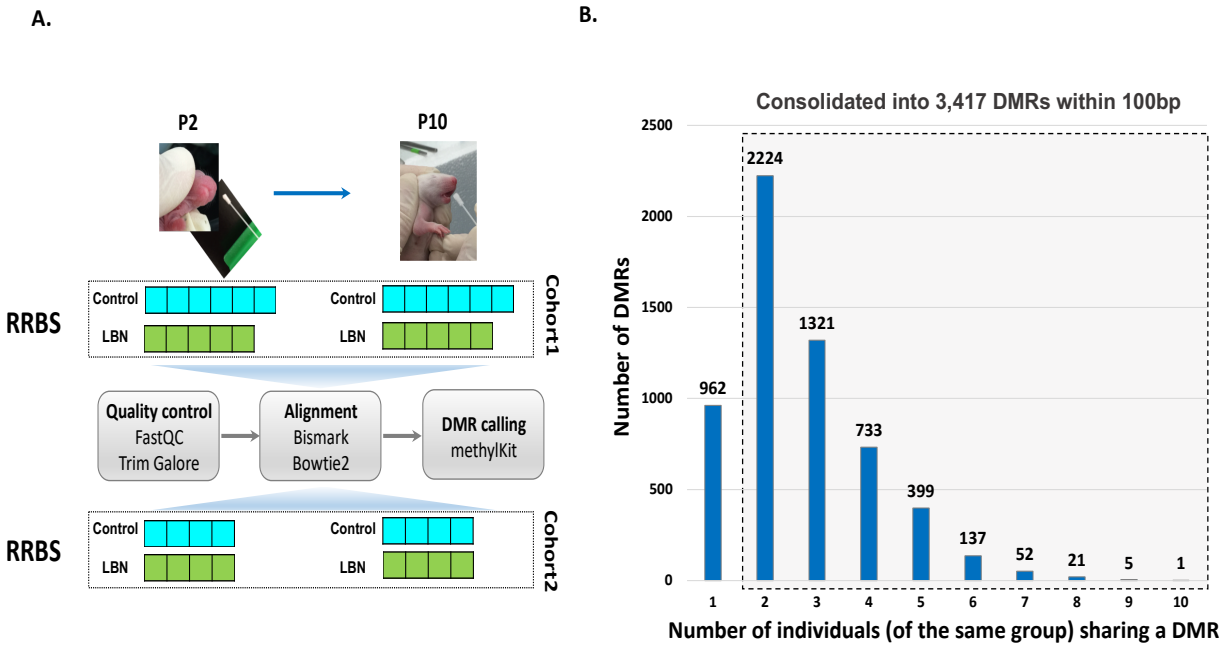


Figure 2.1. (A) Experimental design and analysis pipeline. **(B)** Histogram of the number of significant differentially methylated regions (DMRs) based on the number of individuals sharing the same experience. RRBS- reduced representation bisulfite sequencing; LBN- limited nesting and bedding cages, a paradigm of adversity. P2,P10 = postnatal days 2,10.

A.

Cohort1

P2	Mapped reads (million)	Efficiency	P10	Mapped reads (million)	Efficiency
P2C3	13	43.30%	P10C3	23	46.00%
P2C4	9	36.1%	P10C4	19	55.1%
P2C6	12	51.30%	P10C6	17	44.40%
P2C7	22	59.50%	P10C7	26	52.00%
P2C9	18	41.10%	P10C9	26	53.70%
P2C12	20	49.20%	P10C12	24	52.20%
P2LBN2	15	39.90%	P10LBN2	11	45.90%
P2LBN3	18	43.20%	P10LBN3	30	58.90%
P2LBN4	23	48.80%	P10LBN4	28	51.00%
P2LBN9	12	46.40%	P10LBN9	24	50.40%
P2LBN12	15	45.40%	P10LBN12	26	56.40%

Individual	Detected CpGs in both P2 and P10	Significant DMRs
C3	418860	870
C4	358726	4664
C6	404260	4384
C7	521418	1959
C9	514738	2679
C12	496953	2318
LBN2	415216	5540
LBN3	482076	6345
LBN4	539139	629
LBN9	418466	1851
LBN12	443196	1963

B.

Cohort2

P2	Mapped reads (million)	Efficiency	P10	Mapped reads (million)	Efficiency
P2C1	21	46.0%	P10C1	22	42.5%
P2C5	20	47.4%	P10C5	20	45.6%
P2C8	17	60.4%	P10C8	22	49.2%
P2C11	19	55.7%	P10C11	27	58.1%
P2LBN2	28	57.8%	P10LBN2	30	61.9%
P2LBN4	21	49.2%	P10LBN4	20	42.0%
P2LBN9	21	45.8%	P10LBN9	22	54.2%
P2LBN11	23	47.2%	P10LBN11	20	50.8%

Individual	Detected CpGs in both P2 and P10	Significant DMRs
C1	510525	3073
C5	505100	1879
C8	508787	1278
C11	525733	7040
LBN2	584480	1064
LBN4	498882	2566
LBN9	521111	1190
LBN11	504507	2745

Figure 2.2. RRBS quality control (QC) matrix for cohort1 **(A)** and cohort2 **(B),** including the number of uniquely mapped reads, mapping efficiency and significant DMRs calling for each individual.

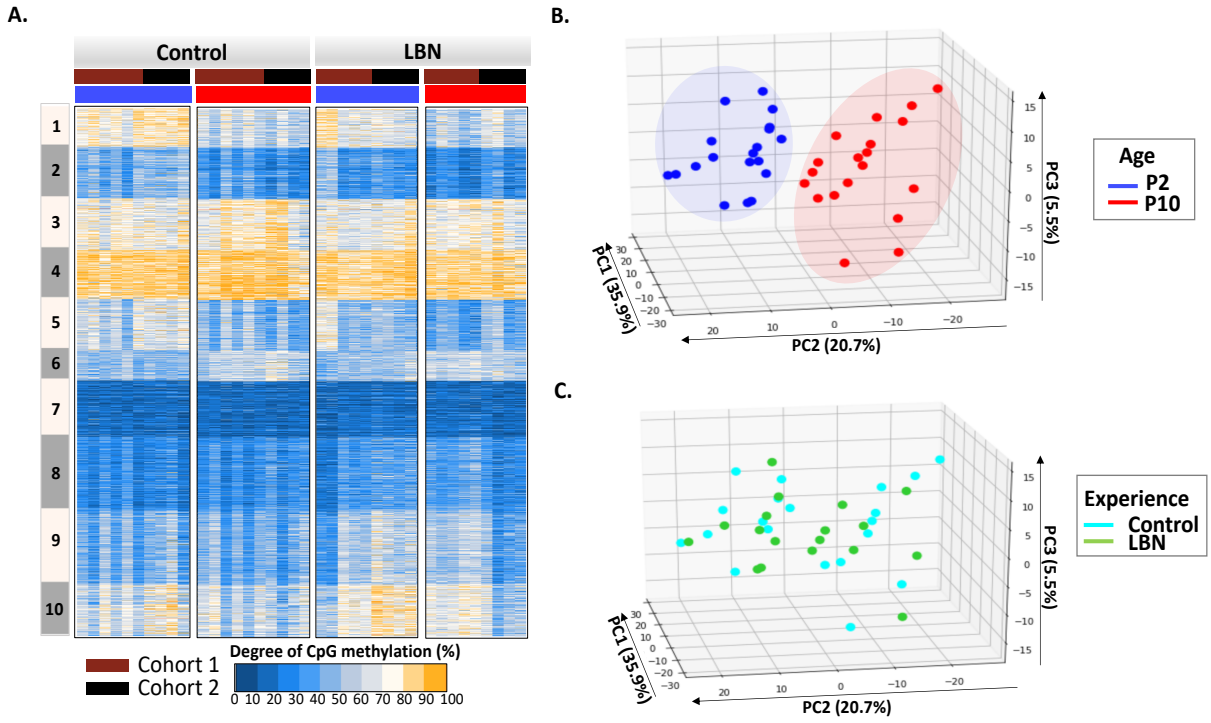


Figure 2.3. **(A)** Heatmap of CpG methylation percentage on 3,417 DMRs across individuals. The profile is presented into 10 clusters that are clustered using K-means clustering. Blue, low methylation percentage; orange, high methylation percentage. **(B)** Principal component analysis (PCA) of individuals on 3,417 DMRs. Individuals are labeled by age, P2, blue; P10, red. **(C)** Principal component analysis (PCA) of individuals on 3,417 DMRs. Individuals are labeled by experience, Control, cyan; LBN, green.

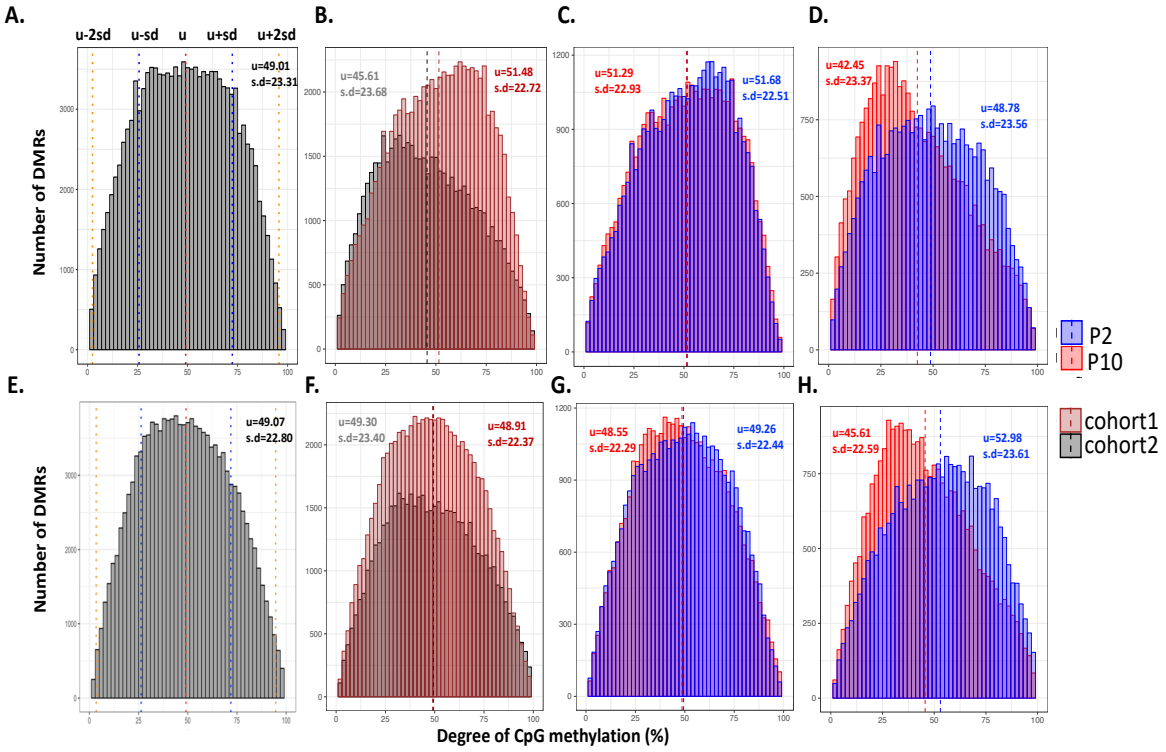


Figure 2.4. (A) Histogram of DNA methylation level on 3,417 DMRs across 19 individuals from two cohorts before batch correction by cohorts. For each DMR, each individual has one methylation level for P2 and one for P10. 11 individuals are in cohort1 and 8 individuals are in cohort2.
(B) Histogram of DNA methylation level on 3,417 DMRs for two cohorts separately (before correction by cohorts).
(C) Histogram of DNA methylation level on 3,417 DMRs on cohort1 for P2 and P10 separately (before correction by cohorts).
(D) Histogram of DNA methylation level on 3,417 DMRs on cohort2 for P2 and P10 separately (before correction by cohorts).
(E) Histogram of DNA methylation level on 3,417 DMRs across 19 individuals from two cohorts after correction by cohorts. For each DMR, each individual has one methylation level for P2 and one for P10. 11 individuals are in cohort1 and 8 individuals are in cohort2.
(F) Histogram of DNA methylation level on 3,417 DMRs for two cohorts separately (after correction by cohorts).
(G) Histogram of DNA methylation level on 3,417 DMRs on cohort1 for P2 and P10 separately (after correction by cohorts).
(H) Histogram of DNA methylation level on 3,417 DMRs on cohort2 for P2 and P10 separately (after correction by cohorts).
u represents mean value; s.d represent standard deviation.

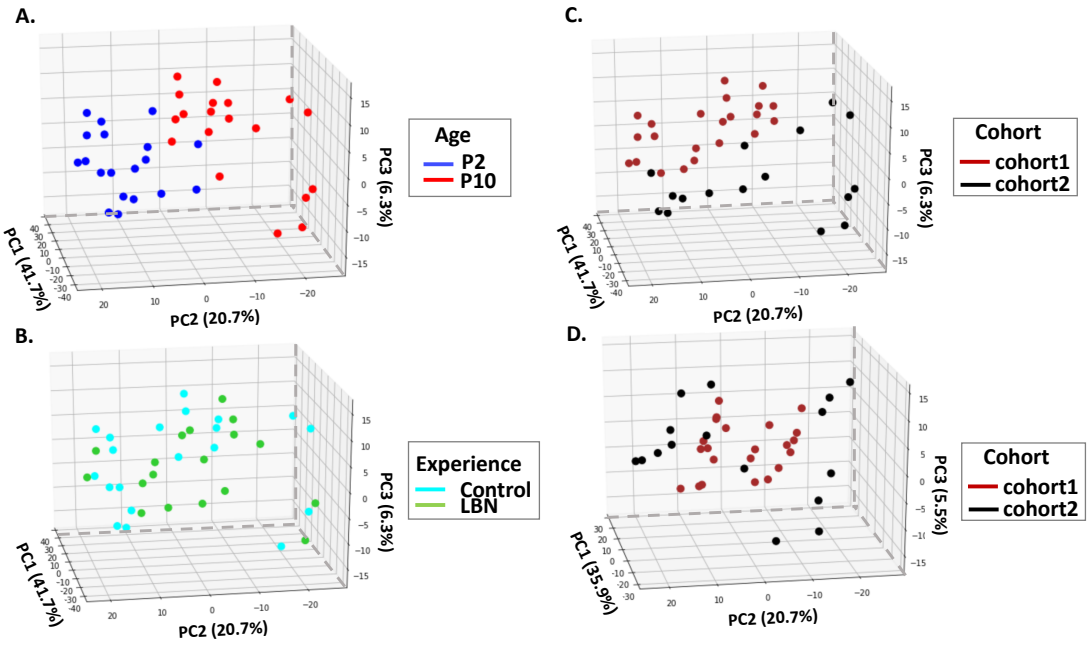


Figure 2.5. PCA analysis on 3,417 DMRs before batch correction by cohort and labeled by **(A)** age, **(B)** experience and **(C)** cohort. **(D)** PCA on 3,417 DMRs after batch correction by cohort.

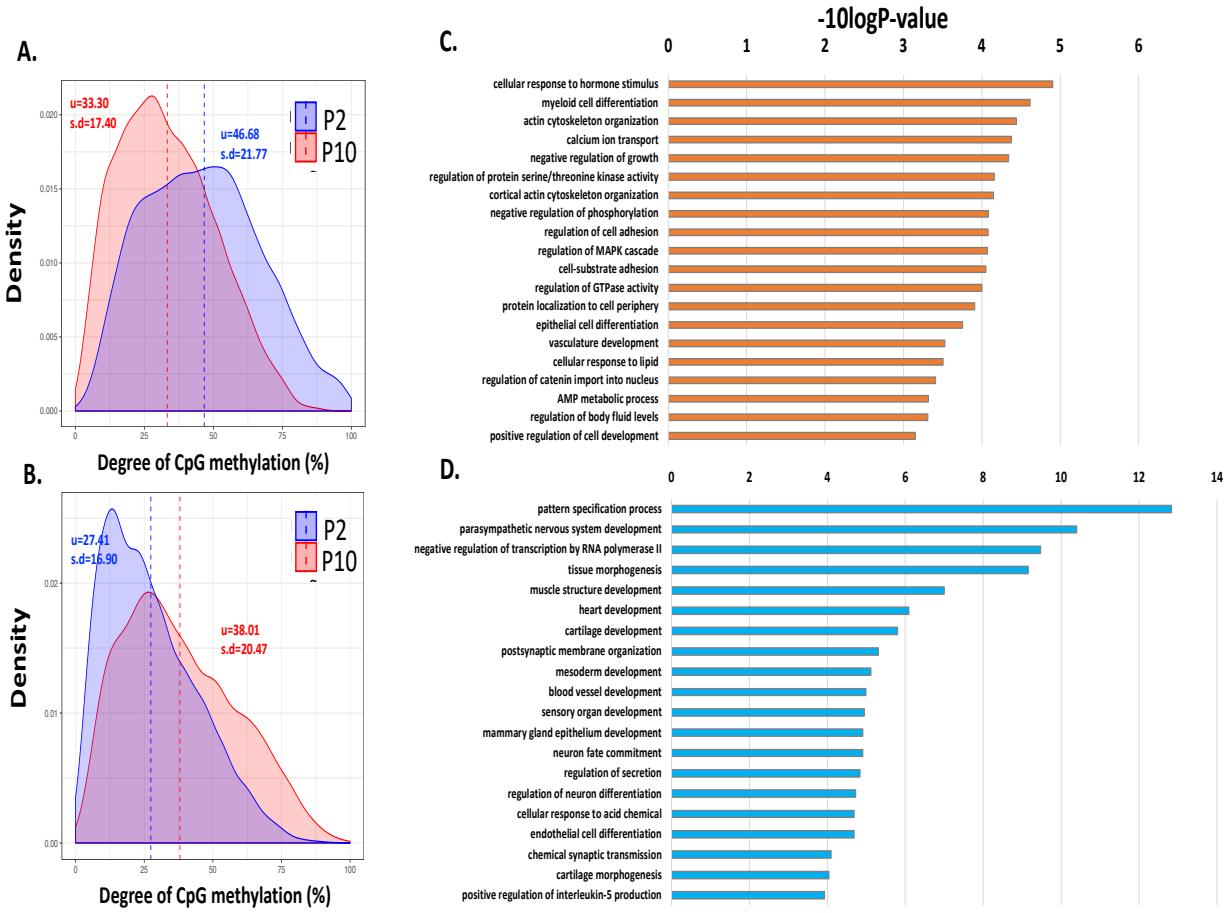


Figure 2.6. (A) DNA methylation level of top-positive DMRs (from PC2 in Figure 2.3B) and density distributions are plotted by age separately.

(B) DNA methylation level of bottom-negative DMRs (from PC2 in Figure 2.3B) and density distributions are plotted by age separately.

(C) Gene ontology terms enriched in genes associated with top-positive DMRs.

(D) Gene ontology terms enriched in genes associated with bottom-negative DMRs.

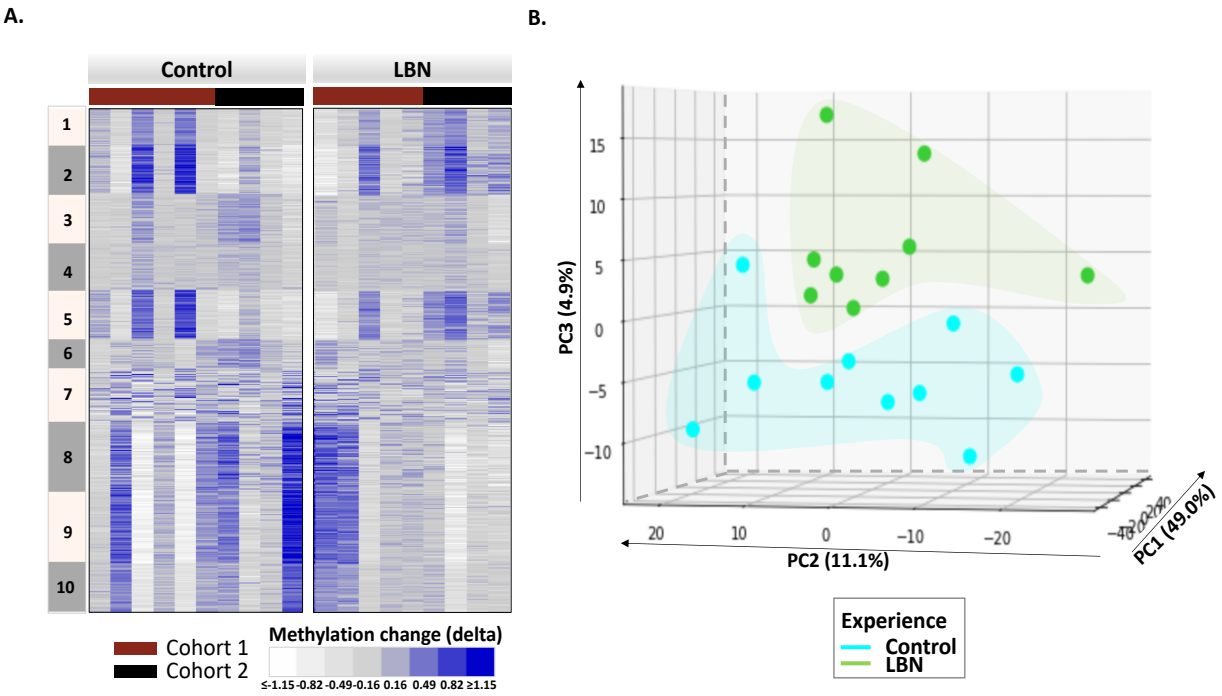


Figure 2.7. (A) Heatmap of delta methylation between P10 and P2 (log2(P10/P2)) of 3,417 DMRs. DMRs are in the same order as in Figure 2.3A. **(B)** Principal component analysis (PCA) of the DMRs of individuals labeled by experience, Control, cyan; LBN, green.

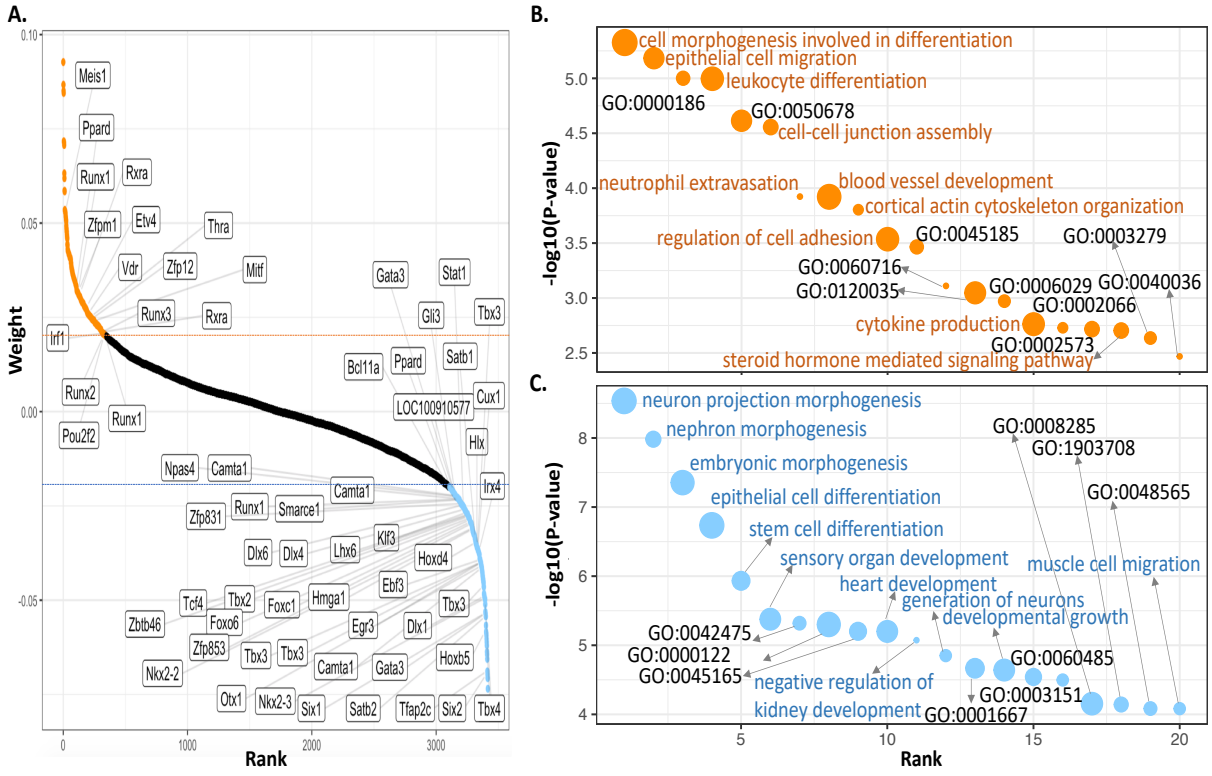


Figure 2.8. Analysis of PC3 weights **(A)** Most significant positive (orange) and negative (blue) weights are enriched in transcription factors. **(B)** GO terms of genes associated with 346 most positively weighted DMRs (orange) in panel A. **(C)** GO terms of genes associated with 311 most negatively weighted DMRs (blue) in panel A.

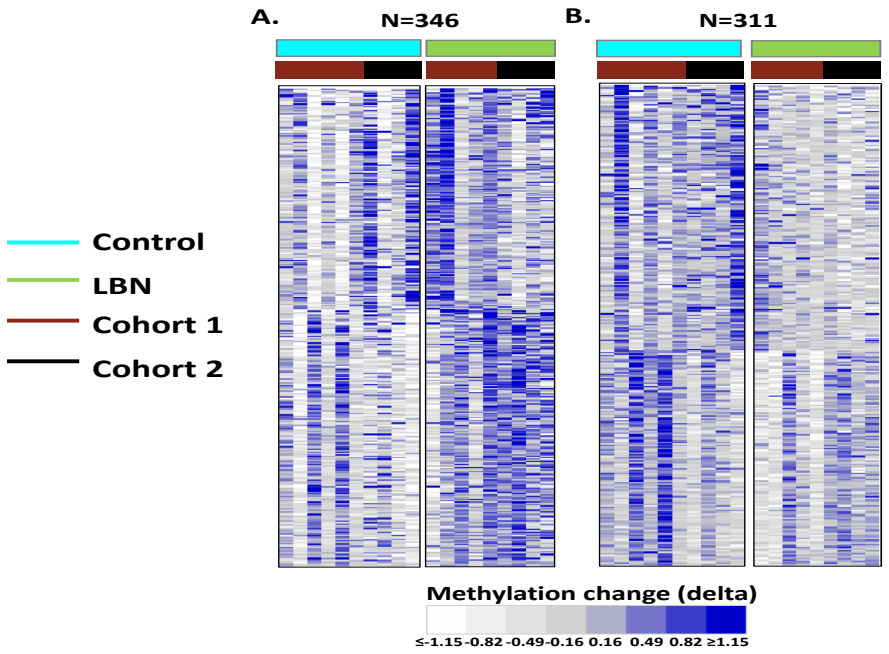


Figure 2.9. Heatmap of delta methylation changing profiles between P10 and P2 on **657** DMRs (top weights were collected from PC3 in Figure 2.7B, which can separate control and LBN). **(A)** 346 "top-positive" weights predict LBN, showing increased methylation in P10 LBN while **(B)** 311 "bottom-negative" weights predict control, showing increased methylation in P10 Control.

2.6 Methods

2.6.1 Animals

Subjects were born to primiparous time-pregnant Sprague-Dawley rat dams (around P75) that were maintained in the quiet animal facility room on a 12 h light/dark cycle with *ad libitum* access to lab chow and water. Parturition was checked daily, and the day of birth was considered postnatal day 0 (P0). Litter size was adjusted 12 per dam on P1, if needed. On P2, pups from several litters were gathered, and 12 pups (6 males and 6 females) were assigned at random to each dam, to obviate the potential confounding effects of genetic variables and of litter size. Each pup was identified by a rapid (<2 minute) foot pad tattooing using animal tattoo ink (Ketchum).

2.6.2 Early-life adversity paradigm

The experimental paradigm involved rearing pups and dams in "impoverished" cages for a week (P2-P9) as described elsewhere (31-33). Briefly, routine rat cages were fitted with a plastic-coated aluminum mesh platform sitting ~2.5 cm above the cage floor (allowing collection of droppings). Bedding was reduced to only cover cage floor sparsely, and one-half of a single paper towel was provided for nesting material, creating a limited bedding and nesting (LBN) cage. Control dams and their litters resided in standard bedded cages, containing 0.33 cubic feet of cob bedding, which was also used for nest building. Control and experimental cages were undisturbed during P2–P9, housed in a quiet room with constant temperature and a strong laminar airflow, preventing ammonia accumulation.

2.6.3 Collection of buccal swab from each pup

The first buccal swab was collected from both cheeks of each pup prior to randomization on P2, using Hydraflock swab (Puritan diagnostics, LLC). After an hour's rest with their mother, a second buccal swab was collected, enabling sufficient DNA from each pup. Pups were then randomized to controls or LBN cages. During P3-P9, behaviors of dams in both control and adversity/LBN cages was observed daily, to ascertain the generation of fragmented unpredictable caring patterns by the adverse environment (34,35). On P10, buccal swabs were collected as described for P2, then all litters were transferred to normal bedded cages.

2.6.4 Isolation and quantification of DNA for making RRBS libraries from Rat Buccal swab

The Buccal swab was placed into the DNA shields™ (Zymo research) immediately after swabbing. DNA was prepared from the DNA shields solution using the Quick gDNA MiniPrep kit following the manufacturer's protocol. The quantity of double stranded DNA was analyzed using Qubit, and RRBS Libraries were prepared from 40 ng of genomic DNA digested with Msp I and then extracted with ZR-DNA Clean & Concentrator™-5 kit (Zymo Research). Fragments were ligated to pre-annealed adapters containing 5'-methyl-cytosine instead of cytosine according to Illumina's specified guidelines (www.illumina.com). Adaptor-ligated fragments were then bisulfite-treated using the EZ DNA Methylation-Lightning™ Kit (Zymo Research). Preparative-scale PCR was performed and the resulting products were purified with DNA Clean & Concentrator for sequencing. Amplified RRBS libraries were quantified and qualified by Qubit, Bioanalyzer (Agilent), and Kapa library quant (Kapa systems), and then sequenced on the Illumina NextSeq 500 platform.

2.6.5 RRBS data processing and detection of differentially methylated regions (DMRs)

Adaptor and low quality reads were trimmed and filtered using Trim Galore! 0.4.3 (36) with parameter '--fastqc --stringency 5 --rrbs --length 30 --non_directional'. Reads were aligned to the rat genome (RGSC 6.0/rn6) by using Bismark 0.16.3 (37) with '--non_directional' mode. CpG sites were called by "bismark_methylation_extractor" function from Bismark.

Single CpG sites with more than ten reads coverage were kept for DMR calling. Differential methylation sites (DMSs) were first called using MethyKit (R 3.3.2) (38) with a false discovery rate (FDR) lower than 0.05; DMRs falling within 100 base pairs were then merged.

2.6.6 Calculation of DNA methylation level/percentage and Delta methylation

The methylation percentage/level was calculated as the ratio of the methylated read counts over the sum of both methylated and unmethylated read counts for a single CpG site or across CpGs for a region.

The delta methylation was calculated using the log2 transformation of the ratio of methylation level in the P10 sample and the methylation level in the P2 sample. Increased methylation in P10 is shown as a positive value while decreased methylation in P10 is shown as a negative value.

2.6.7 Principal component analysis (PCA) and K-Means clustering

PCA analysis was performed by using *IncrementalPCA* function from scikit-learn (39) using python 2 for both Figure 2 and 3. The value of k was set to 10 for the k-means clustering based on a preliminary hierarchical clustering analysis. A DNA methylation heatmap was generated with heatmap.2 function in R 3.5.0 and a delta methylation heatmap was generated using Java TreeView (40).

2.6.8 Gene association analysis

Genes associated with DMRs were identified using Homer 4.7 (41). For subsequent analyses, genes were kept if (1) CpGs were located within 20kb of TSS in intergenic, promoter-TSS and TTS positions; (2) CpGs were located within gene exons or introns. Gene ontology analysis was performed using Metascape (42) using the hypergeometric test with corrected P-value lower than 0.05.

2.7 References

1. T. N. and P. A. S. of the P. G. Consortium, Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* 18, 199 (2015).
2. T. L. Bale *et al.*, Early Life Programming and Neurodevelopmental Disorders. *Biol. Psychiatry*. 68, 314–319 (2010).
3. T. Klengel, E. B. Binder, Epigenetics of Stress-Related Psychiatric Disorders and Gene × Environment Interactions. *Neuron*. 86, 1343–1357 (2015).
4. E. J. Nestler, C. J. Peña, M. Kundakovic, A. Mitchell, S. Akbarian, Epigenetic Basis of Mental Illness. *Neurosci.* 22, 447–463 (2015).
5. M. Szyf, Nongenetic inheritance and transgenerational epigenetics. *Trends Mol. Med.* 21, 134–144 (2015).
6. T. A. Bedrosian, C. Quayle, N. Novaresi, F. H. Gage, *Science (80-.).*, in press (available at http://science.sciencemag.org/content/359/6382/1395.abstract).
7. Y. Chen, T. Z. Baram, Toward Understanding How Early-Life Stress Reprograms Cognitive and Emotional Brain Networks. *Neuropsychopharmacology*. 41, 197–206 (2016).
8. C. A. Nelson *et al.*, *Science (80-.).*, in press (available at http://science.sciencemag.org/content/318/5858/1937.abstract).
9. A. Singh-Taylor *et al.*, NRSF-dependent epigenetic mechanisms contribute to programming of stress-sensitive neurons by neonatal experience, promoting resilience. *Mol. Psychiatry*. 23, 648 (2017).
10. T. L. Bale, Epigenetic and transgenerational reprogramming of brain development. *Nat.*

Rev. Neurosci. 16, 332 (2015).
11. J. Bohacek, I. M. Mansuy, Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nat. Rev. Genet.* 16, 641 (2015).
12. B. G. Dias, K. J. Ressler, Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nat. Neurosci.* 17, 89 (2013).
13. C. J. Peter *et al.*, DNA Methylation Signatures of Early Childhood Malnutrition Associated With Impairments in Attention and Cognition. *Biol. Psychiatry*. 80, 765–774 (2016).
14. I. C. G. Weaver *et al.*, Epigenetic programming by maternal behavior. *Nat. Neurosci.* 7, 847 (2004).
15. Z. Nemoda *et al.*, Maternal depression is associated with DNA methylation changes in cord blood T lymphocytes and adult hippocampi. *Transl. Psychiatry*. 5, e545 (2015).
16. A. S. Ivy *et al.*, Hippocampal Dysfunction and Cognitive Impairments Provoked by Chronic Early-Life Stress Involve Excessive Activation of CRH Receptors. *J. Neurosci.* 30, 13005–13015 (2010).
17. J. L. Bolton *et al.*, Anhedonia Following Early-Life Adversity Involves Aberrant Interaction of Reward and Anxiety Circuits and Is Reversed by Partial Silencing of Amygdala Corticotropin-Releasing Hormone Gene. *Biol. Psychiatry*. 83, 137–147 (2018).
18. R. Lister *et al.*, Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science (80-.).* 341 (2013) (available at http://science.sciencemag.org/content/341/6146/1237905.abstract).
19. A. Meissner *et al.*, Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877 (2005).
20. Y. Reizel *et al.*, Postnatal DNA demethylation and its role in tissue maturation. *Nat. Commun.* 9, 2040 (2018).
21. M. Said *et al.*, Genomics In Premature Infants: A Non-Invasive Strategy To Obtain High-Quality DNA. *Sci. Rep.* 4, 4286 (2014).
22. R. Lowe *et al.*, Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics*. 8, 445–454 (2013).
23. P. Braun *et al.*, Genome-Wide Dna Methylation Comparison Between Live Human Brain and Peripheral Tissues Within Individuals. *Eur. Neuropsychopharmacol.* 27, S506 (2017).
24. A. K. Smith *et al.*, DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 168, 36–44 (2015).
25. M. N. Davies *et al.*, Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 13, R43 (2012).
26. M. Eipel *et al.*, Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. *Aging (Albany NY)*. 8, 1034–1044 (2016).
27. S. Horvath, K. Raj, DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* 19, 371–384 (2018).
28. E. Borrelli, E. J. Nestler, C. D. Allis, P. Sassone-Corsi, Decoding the Epigenetic Language of Neuronal Plasticity. *Neuron*. 60, 961–974 (2008).
29. T. S. Doherty, T. L. Roth, in *Epigenetics and Psychiatric Disease*, D. R. B. T.-P. in M. B. and T. S. Grayson, Ed. (Academic Press, 2018; http://www.sciencedirect.com/science/article/pii/S187711731730203X), vol. 157, pp. 1–

19.
30. L. D. Moore, T. Le, G. Fan, DNA Methylation and Its Basic Function. *Neuropsychopharmacology*. 38, 23–38 (2013).

31. Molet J, Maras PM, Avishai-Eliner S, Baram TZ. Naturalistic rodent models of chronic early-life stress. *Dev Psychobiol*. 2014;56(8):1675-1688. doi:doi:10.1002/dev.21230
32. Ivy AS, Brunson KL, Sandman C, Baram TZ. Dysfunctional nurturing behavior in rat dams with limited access to nesting material: a clinically relevant model for early-life stress. Neuroscience. 2008 Jun 26;154(3):1132-42. doi:
33. Walker CD, Bath KG, Joels M, Korosi A, Larauche M, Lucassen PJ, Morris MJ, Raineki C, Roth TL, Sullivan RM, Taché Y, Baram TZ. Chronic early life stress induced by limited bedding and nesting (LBN) material in rodents: critical considerations of methodology, outcomes and translational potential. Stress. 2017 Sep;20(5):421-448.
34. Molet J, Heins K, Zhuo X, Mei YT, Regev L, Baram TZ, Stern H. Fragmentation and high entropy of neonatal experience predict adolescent emotional outcome. Transl Psychiatry. 2016 Jan 5;6:e702. doi: 10.1038/tp.2015.200.
35. Davis EP, Stout SA, Molet J, Vegetabile B, Glynn LM, Sandman CA, Heins K, Stern H, Baram TZ. Exposure to unpredictable maternal sensory signals influences cognitive development across species. Proc Natl Acad Sci U S A. 2017 Sep26;114(39):10390-10395
36. Krueger F: Trim Galore!. [http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/]
37. Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*, *27*(11), 1571-1572.
38. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, *13*(10), R87.
39. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.
40. Saldanha, A. J. (2004). Java Treeview—extensible visualization of microarray data. *Bioinformatics*, *20*(17), 3246-3248.
41. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, *38*(4), 576-589.
42. Tripathi, S., Pohl, M. O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D. A., ... & Yángüez, E. (2015). Meta-and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. *Cell host & microbe*, *18*(6), 723-735.

CHAPTER 3

Characterizing the heterogeneity of DUX4 and DUX4 targets expression during FSHD2 myoblast differentiation

Notes: (1) Katherine Elizabeth Williams performed some of single cell and nucleus RNA-seq experiments. She also did the regular RNA-seq analysis and made Figure 3.3-3.5.
(2) Dr. Xiangduo Kong performed cell culture experiments and provided images in Figure 3.1.
(3) Dr. Weihua Zeng performed some of the single cell and nucleus RNA-seq experiments.
(4) Xinyi Ma sequenced all samples.
(5) Dr. Rabi Tawil provided FSHD2 cells in this research.
(6) Dr. Kyoko Yokomori and Dr. Ali Mortazavi conceived the idea and provided continued support and guidance throughout the project.

Chapter 3

Characterizing the heterogeneity of DUX4 and DUX4 targets expression during FSHD2 myoblast differentiation

3.1 Abstract

Fascioscapulohumeral muscular dystrophy (FSHD) is primarily caused by the expression of the normally repressed transcription factor DUX4 in skeletal muscle turning on a set of target genes. DUX4 expression in skeletal muscle is either caused by a contraction of the D4Z4 macrosatellite repeat array containing the DUX4 gene (FSHD1) or by mutation in other genes such as SMCHD1 involved in the repression of the D4Z4 repeat array (FSHD2). However, DUX4 is lowly expressed in patient samples and previous studies have focused on overexpression of DUX4. To better understand the expression profile of DUX4 and its targets in FSHD2, we first performed pooled RNA-seq on a 6-day differentiation time-course in FSHD2 patient-derived myoblasts and found upregulation DUX4 target genes starting from day 3. Using single-cell/nucleus RNA-seq on FSHD2 myoblasts and day 3 myotubes, we successfully detected a set of DUX4 positive nuclei in FSHD2 myotubes. We found that substantially more FSHD myotube nuclei expressed with DUX4 targets than the nuclei that expressed DUX4. We also found that both FSHD2 myotube nuclei expressed more DUX4 target genes compared to FSHD myoblasts and control myotubes. A pseudo time-course of single cells and nuclei shows a distinct bifurcation between DUX4 target-positive and DUX target-negative FSHD2 myotubes compared to control myotubes. DUX4 target-positive myotubes formed a distinct population that were separated from other cells types that imply that these cells have entered a distinct biological program potentially driven by other transcription factors downstream of DUX4.

3.2 Introduction

Fascioscapulohumeral muscular dystrophy (FSHD) is one of the most common forms of inherited muscular dystrophy [1], which is characterized by progressive wasting of facial, shoulder and upper arm musculature [1]. The most common form of FSHD, FSHD1 (>95% of cases), is linked to the mono-allelic contraction of the D4Z4 macrosatellite repeat array on chromosome 4q, shrinking from 11-100 units to 1-10 units, with each 3.3 kb containing the open reading frame for the double-homeobox transcription factor *DUX4* [2-4]. By contrast, there is no

contraction of the chromosome 4q repeat in FSHD2 (<5% of FSHD cases). Instead many FSHD2 cases are characterized by recurring mutations in the chromatin modifier *SMCHD1* (Structural maintenance of chromosomes flexible hinge domain-containing protein 1) on chromosome 18 [5]. *SMCHD1* is important for maintenance of DNA methylation and epigenetic silencing of multiple genomic loci —including the D4Z4 repeat array. Studies found that SMCHD1 mutation shown in a significant number of FSHD2 cases [5] as well as in severe cases of FSHD1 [6,7].

FSHD is associated with the expression of the full-length DUX4 transcript (DUX4fl), which is stabilized by a specific single-nucleotide polymorphism in chromosomal region distal to the last D4Z4 repeat that creates a canonical polyadenylation signal [8-10]. DUX4fl encodes a transcriptional activator that binds to a double-homeobox sequence motif in genome [3,4] and overexpression of DUX4fl causes differentiation defects in human myoblasts and mouse C2C12 cells [11,12]. However, DUX4fl is expressed at extremely low level in FSHD and DUX4 protein is only detected in 0.1% of patient muscle cells *in vitro* [8,10]. The regulation of DUX4 expression is controlled by multiple epigenetic processes. In the case of SMCHD1, which binds to D4Z4 repeats, its depletion results in DUX4fl up-regulation through alterations of CpG methylation status [5-7]. Differences in CpG methylation at D4Z4 correlate with clinical variability in FHSD1 and FSHD2 [5], with SMCHD1 acting as a disease modifier in some FHSD1 patients [6]. Studies have shown that SMCHD1 binds to the H3K9me3 D4Z4 heterochromatic regions that are lost specifically in FSHD1 and FSHD2 that induce DUX4fl gene expression [2,10,13,14]. Within the D4Z4 heterochromatic regions, both SMCHD1 and cohesin proteins bind to D4Z4 repeats in an H3K9me3-dependent manner [2,13].

Although DUX4 is the most promising disease-causing gene in FSHD, studies have found that DUX4fl expression cannot be detected in some FSHD cases [8,15]. Furthermore, DUX4fl expression can sometimes observed even in unaffected individuals [8,15]. Studies have reported that the expression of DUX4fl induces the expression of DUX4 target genes in patient cells and even without detectable DUX4 expression, some of these DUX4 target genes can also be observed [4, 15-19]. Thus, the contribution of DUX4 expression and its target genes to the pathogenesis demands further careful investigation. One of the largest challenges for this investigation is the poorly-developed disease model for human FHSD studies. The D4Z4 repeat

array is not present in mouse [20]. While overexpression of human DUX4 does lead to muscular dystrophy, whether the cytotoxicity is caused by the same mechanism in human patient and experimental rodent muscles remains obscure. Therefore, patient muscle cells remain to be the essential resources for observing and assessing FSHD related pathogenesis and dysregulation. However, the variability of patient cells, including growth conditions, genetic variations, spontaneous differentiation may cause artifacts in observing gene expression and epigenetic profiles in FSHD and it is even more difficult to derive a reproducible conclusion on FSHD-specific patterns with extremely low DUX4 expression in only a small cluster of cells [21-26]. Researchers have tried to overcome this by overexpressing DUX4 in human muscle cells in order to identify DUX4 targets and defects in muscle cell differentiation [11,12]. However, overexpression may not be appropriate to derive solid physiological and cellular conclusion on the disease progression because of the much higher DUX4 expression compared to the endogenous level. Furthermore, previous population-based studies have found that DUX4 target genes are not consistently expressed across all FSHD patient cells though they are generally up-regulated by averaging population results [15-18, 27]. Also, given that DUX4 is only expressed in a small subset of cells, it is important to investigate the cellular heterogeneity in FSHD patient samples and to understand how DUX4 regulates target genes directly, as well as how they are involved in the disease dysregulation.

Here we focused on the SMCHD1-mutated FSHD2 subtype in order to characterize the heterogeneity of DUX4 and DUX4 target gene expression at the single-cell level by differentiating FSHD2 patient-derived myoblast to myotubes and comparing those to control myoblasts/myotubes *in vitro*. Using regular pooled RNA-seq, we profiled gene expression patterns during the differentiation time course and identified candidate disease-related key genes by comparing expression profiles between FSHD2 and control. We found that about 30% of differentially expressed genes (log2Fold Change>3) are known DUX4 target genes while the others may be potential candidates for the progression of the disease. We then used single-cell/single-nucleus RNA-seq [28] in myoblasts and 3-day post-differentiation myotubes to characterize the expression patterns of DUX4 and its target genes. We sucessfully detected the first DUX4+ single nucleus expression in FSHD and found these small subsets of DUX4+ nuclei did not express all DUX4 target genes simutanously whereas a much larger subset of nuclei

expressed the DUX4 targets. The single-cell and single-nuclei data agreed well with the regular (bulk) RNA-seq data. Using pseudotime analysis on the single-cell/single-nucleus data, we also found two interesting bifurcation time points that could separate FSHD2 myotube nuclei from control myotube nuclei and distinguish the DUX4 target negative FSHD2 nuclei from DUX4 positive or DUX4 negative but target gene positive FSHD2 nuclei which show the highest enrichment of differentialy expressed genes as observed in the population-based analysis.

3.3 Results

3.3.1 Up-regulation of the expression DUX4 target genes during FSHD2 myotube differentiation time-course

In order to comprehensively understandand the expression of DUX4 and its targets during myogenesis, we first differentiated FSHD2 patient-derived myoblast differentiation in vitro to measure the dynamics of gene expression in a 6 day time course using pooled RNA sequencing (RNA-seq) (Figure 3.1 & Table 3.1). The primary FSHD2 myoblasts with an SMCHD1 mutation (g.269799_2698003 deletion) and control myoblasts were derived from quadriceps muscle biopsies. We filter out lowly expressed genes across all samples and performed principal component analysis (PCA) with 9278 genes. We observed that FSHD2 and control were clearly separated by the first three principal components (73.9% varicance explained) but with the day of differentiaion aligned to each other (Figure 3.2). This result indicates that FSHD2 patient-derived myoblast can be distinguished at every stage of differentiation from control cells by profiling transcriptomes at population level.

In order to understand the differences between FSHD2 and control, we identified 168 differentially expressed genes (log2FoldChange>3) by comparing between FSHD2 and control per day (Methods). While we could not detect DUX4 during our time-course, many known DUX4 targets are up-regulated starting at day 2 in FSHD2, including MBD3L2, ZSCAN4, KHDC1L and LEUTX [27] and more known targets are stablely activated starting at day 3 (Figure 3.3A). We therefore focused on genes showing significant changes after day 3 and further classified 168 differentially expressed gene into 6 clusters using K-means clustering (Figure 3.3). We found that genes up-regulated in FSHD2 can be separated into early- (clusters 1 and 6) and late-induced expression patterns (cluster 2) (Figure 3.3A). Importantly, known DUX4

associated up-regulated genes represent over 50% of the genes induced in later stage, indicating that other genes within the same clusters could be novel candidates for DUX4 in FSHD2. We also detected genes downregulated in FSHD2 compared to control (cluster 3) (Figure 3.3B). We further extracted 109 genes (out of 168, 64.9%) that show differential expression across the entire time-course and found that 73 genes were highly expressed in FSHD2 (Figure 3.4) while 36 genes were downregulated in FSHD2 (Figure 3.5). Genes induced at the later stage (cluster 2) were highly enriched in negative regulation of cell differentiation ($P=1x10^{-11.40}$) and methylation-dependent chromatin silencing ($P=1x10^{-5.38}$) while genes in early stage (cluster 1) were enriched in leukocyte activation involved in immune response ($P=1x10^{-1.74}$) (Figure 3.4). For genes downregulated in FSHD compared to the control, we found that they were associated with epithelial cell apoptotic process ($P=1x10^{-3.40}$) and negative regulation of immune system process (P=1x10-1.75) (cluster 3) (Figure 3.5). In summary, our FSHD2 differentiation time-course shows robust expression of DUX4 target genes starting at day 3 of differentiation and suggests substantial changes in methylation-dependent chromatin silencing.

3.3.2 Detection of DUX4 positive and DUX4-target positive nuclei in FSHD2 myotube using single-nucleus RNA-seq

Although DUX4 was known to express at an extremely low level and DUX4-associated genes are detectably up-regulated in FSHD2 during myotube differentiation, we wondered whether this observation was true at the single-cell level in all cells or whether DUX4 and DUX4-target expression was only present in a subset of cells. We therefore performed single-cell RNA-seq on myoblast cells for both control primary myoblasts and FSHD2 primary myoblasts as well as single-nucleus RNA-seq using the C1 platform [28] for myotubes at day 3 of differentiation (Figure 3.1 & 3.6), based on the first day of robust detection of DUX4 target expression in section 3.3.1. As quality control that our single-cell data does matches our pooled time course, we first pooled reads from all single cells/nuclei for each cell type and performed incremental PCA with bulk time-course RNA-seq samples (Figure 3.7). As expected, pooled single-cell myoblast clustered with day 0 samples in both control and FSHD2. However, for pooled single-nucleus myotube, FSHD2 replicate 1 (FSHD2 R1) was aligned with day 3 FSHD2 pooled data in time-course data but the FSHD2 replicate 2 (FSHD2 R2) was located between control and FSHD2 days 3 in the time-course. Furthermore, we found that 3 out of 79 (3.8%)

nuclei in FSHD2 R1 showed high expression of DUX4 (the DUX4 expression level are 10.52 TPM , 33.37 TPM and 68.07 TPM) while we detected no DUX4 positive nuclei in FSHD2 R2, revealing the high level of heterogeneity in the FSHD2 cell population with DUX4 only expressed in a small fraction of cells, but expressed at a significant level in those nuclei when it is actively transcribed. We then analyzed the global profiles of the single-cell transcriptomes using both PCA (Figure 3.8A) and t-SNE (Figure 3.9A) analyses and found that all 3 DUX4 positive nuclei as well as most other FSHD2 replicate R1 nuclei clearly separated from FSHD2 R2 and control myotube nuclei. FSHD2 target genes specifically up-regulated during FSHD2 differentiation (Figure 3.3A) showed significantly higher enrichment in FSHD2 R1 myotube nuclei compared with control myotube nuclei ($P<2.2x10^{-16}$) with the highest enrichment in the group of nuclei clustering with the 3 DUX4+ nuclei and thus this group of nuclei are "DUX4 targets high" (Figure 3.8B). Besides, FSHD2 R2 myotube nuclei also showed significantly higher enrichment of FSHD2 target genes than control myotube nuclei ($P<2.2x10^{-16}$) but they have lower number of expressed targets than the one in "DUX4 targets high" group and therefore this group of nuclei are "DUX4 targets low" (Figure 3.8B). However, genes down-regulated during FSHD2 development (Figure 3.3B) were expressed across all types of cells/nuclei (Figure 3.8C). Furthermore, we found that DUX4 and DUX4 associated targets were not often expressed in the same nuclei (Figure 3.9B & C). By observing the co-expression between DUX4 and 6 known DUX4 targets, including KHDC1L, LEUTX, MBD3L2, MBD3L3, TRIM43 and ZSCAN4 (Figure 3.10A), we found that none of the myoblast cells and control myotube nuclei expressed these target genes. But for FSHD2 myotube nuclei, we found that more than half of the "DUX4 targets low" nuclei has no expression of these genes and the rest can mostly co-express with 3 genes. However, all "DUX4 targets high" nuclei express at least one DUX4 targets and they can co-express with at most 6 genes (Figure 3.10B). Therefore, we successfully detected a small number of DUX4 positive nuclei and a significantly larger set of DUX4 negative but target positive nuclei in FSHD2 myotubes but not in FSHD myoblasts or control myoblasts/myotubes.

3.3.3 DUX4 target-positive myotube nuclei form a distinct branch on a pseudo-time course of single-cell differentiation

We performed a pseudo time-course analysis using Monocle in order to understand whether the cells expressing DUX4 targets followed a distinct developmental trajectory (Methods). We reordered 317 single cells/nuclei (Figure 3.6) and as expected, the cells/nuclei formed a differentiation trajectory from myoblast to myotube and we found more heterogeneity in myotube nuclei compared with myoblast cells (Figure 3.11A). Interestingly, the FSHD R1 with 3 DUX4 positive nuclei formed a homogenous cluter (branch III) at one end of the pseudo time-course and showed up-regulation of the DUX4 target genes that we had identified in our bulk RNA-seq differentiation time-course (Figure 3.11B). However, nuclei from the DUX4 negative FSHD2 R2 replicate were mixed with many control nuclei (branch IV) and some of them were even located on the same branch as myoblast cells (branch I) in terms of pseudo temporal position (Figure 3.11A), indicating more heterogenity in this population might be caused by a less advanced differentiation status. We further observed that genes induced at late-stages (cluster 2, Figure 3.4) during FSHD2 differentiation time-course in single cells/nuclei (Figure 3.12A) and found that these genes show higher enrichment in FSHD2 myotube nuclei (44/52 (84.6%) for “DUX4 targets high” nuclei and 34/52 (65.4%) for “DUX4 tagets low” nuclei) compared with others (2/ 52 (3.8%) in control myoblast cells; 4/52 (7.7%) in FSHD2 myoblast cells; 3/52 (5.8%) in control myotube nuclei) (Figure 3.12B). Note that four genes expressed in control and FSHD2 myoblast (ZNF596, CCNA1, CHI3L1 and MUSTN1) were actually activated in day 1 and 2 in our time-course (clutster 2, Figure 3.4). In addition, we did not detect 7 potential DUX4 target genes based on our pooled data (HNRNPCL3, PRAMEF17, PRAMEF19, PRAMEF2, PRAMEF20, PRAMEF26, RBP7) in any of our myotube single nuclei. However, genes induced in the early stage (cluster 1, Figure 3.4) were expressed evenly across different cell types (16/21 (76.2%) in control myoblast; 18/21 (85.7%) in FSHD myoblast; 13/21 (61.9%) in control myotube; 14/21 (66.7%) in FSHD2 “DUX4 tagets high” nuclei; 20/21 (95.2%) in “DUX4 tagets low” nuclei) (Figure 3.13). These results further confirmed the importance of late-stage induced genes in distinugishing between control and disease phenotype. Prominent cluster 2 genes detected in the single myotube nuclei FSHD2 data include the known DUX4 targets such as TRIM43, ZSCAN4, MBD3L2, MBD3L3, MBD3L5, KHDC1L, LEUTX, and the DUX4 paralog DUXA, which indicates that these nuclei are most likely responding to the presense of DUX4 protein in their myotube even if DUX4 itself is not being transcribed in the same nucleus.

Although both FSHD2 “DUX4 targets high” and “DUX targets low” myotube nuclei have high enrichment of those late-induced genes compared with control myotubes, “DUX4 targets high” myotube nuclei show higher proportion than the one in “DUX targets low” myotube nuclei. To better understand the difference between these two population and discover novel targets of DUX4, we performed differential expression analysis between two populations in branch III and V, and found out that 3406 genes were up-regulated in branch III (“DUX4 targets high” population, FSHD2 myotube R1) while 2710 were up-regulated in branch V (“DUX4 targets low” population, FSHD2 myotube R2) (Figure 3.14A). We further detected that 84 TFs were up-regulated while 80 TFs were down-regulated in “DUX4 targets high” myotube nuclei (Figure 3.14B & 3.15A). In addition to already reported TFs activated in FSHD, such as LEUTX (20.3%, 16/79), ZSCAN4 (50.6%, 40/79), PITX (10.1%, 8/79), DUX4 (3.8%, 3/79), we observed several additional TFs associated with embryonic organ development and endocrine system development (Table 3.2) also up-regulated in many or all of the 79 “DUX4 targets high” FSHD2 nuclei, such as HES6 (100%,79/79 nuclei), TEAD4 (83.5%, 66/79), MBD3 (82.3%, 65/79), , FOXO1 (70.9%, 56/79), FOXH1 (34.2%, 27/79), and GATA3 (13.9%, 11/79) (Figure 3.14B & C). Besides, we also detected expression of DUX4 paralog gene DUXA (15.2%, 12/79), which turned up to be the top up-regulated TF in these nuclei (Figure 3.14B & C). However, we also detected MYOD1 (94.9%, 75/79) being more highly expressed in DUX4 target-positive nuclei. Cell cycle genes PCNA and CDK1 were down-regulated but myogenic differentiation makers CKM, TTN, MYH3, MYBPH, ACTA1 were up-regulated as expected in mature myotubes (Figure 3.16A). In addition, myogenic markers MYF5 and MYF6 were the top up-regulated TFs in Branch V “DUX4 targets low” population (Figure 3.15). Although DUX4 and target genes dysfunction have been known to cause developmental defects in FSHD2, the results above suggests that the maturation of myotube may also affect DUX4 and targets expression. Interestingly, we found that Desmin was also significantly more highly expressed in Branch III (“DUX4 targets high” myotube nuclei) than Branch V (“DUX4 targets low” myotube nuclei) (Figure 3.16A). By checking gene expression patterns across all myotube nuclei, we found that nuclei with DUX4 and targets expressed shown higher Desmin expression compared with others no matter the level of myogenic differentiation marker, such as CKM (Figure 3.16B) and cell cycle gene PCNA was relatively down-regulated in these nuclei (Figure 3.16C). In conclusion, “DUX4 targets high” and “DUX4 targets low” FSHD2 myotube were clearly seperated on our

pseudo timecourse. Although both of them expressed DUX4 target genes, a distinct set of TFs are up-regulated in "DUX4 targets high" myotube nuclei and their differential expression may play an important role in FSHD progression.

3.4 Discussion

In this study, we found that DUX4 targets are robustly detected in our FSHD2 differentiation time-course starting at day 3 even though we barely detect DUX4 transcript itself. We further show that we do detect DUX4 transcripts in a small number of FSHD2 myotube nuclei. We show that these nuclei cluster with a much larger set of FSHD2 nuclei expressing multiple DUX4 targets compared with either other FSHD2 nuclei expressing few or no targets or control myotubes. The cluster of DUX4 target-positive nuclei is also enriched with a specific set of the regulatory TFs that may play an important role in FSHD pathogenesis. Our data also suggests that the precise maturation status of the myotube may influence the actual expression of DUX4 and its targets.

While we detected DUX4 target genes in both populations of FSHD2 myotube nuclei, many of these targets were only detected in a subset of cells and showed heterogeneous expression patterns at the single-nucleus level . We failed to detect significant DUX4 expression in the FSHD2 pooled time-course and one single nucleus run, but DUX4 target genes were nevertheless highly expressed in these samples. Although we detected three DUX4+ nuclei in FSHD myotube, these nuclei did not express many of the known DUX4 target genes at highest level compared with others that did not have detectable DUX4 expression. Other nuclei from the same group actually contributed to the level of DUX4 targets expression in FSHD2 myotube. Our results agree with a previous study that proposed a model that the DUX4 transcript in a small number of myotube nuclei induces downstream targets through cascade diffusion of the translated DUX4 proteins into all the nuclei in the same myotube, where they activate target gene expression [29]. The target gene products could also diffuse through the cytoplasm of the myotubes and be imported into other nuclei further altering their epigenetic and expression profiles. We detected PITX1 as significantly activated in our "DUX4 targets high" nuclei compared with "DUX4 targets low" FSHD2 nuclei. Our single cell/nucleus shown the first high-resolution profile to observe the expression of DUX4 and target genes expression at the same

time and have found a set of transcription factors that were specifically more highly expressed in "DUX4 targets high myotube; these may be genes that would be normally silenced that are affected by the derepression caused by the expression of bona fide DUX4 targets such as the MDBL genes. It may be that the overexpression of these TFs may aggravate the perturbation of normal myotube differentiation thus leading to increased cell death or are a side-show to the main causes of DUX4 cytotoxicity. In either case, the rarity of DUX4 nuclei expression can account for why normal myotube forms, as the system is extremely sensitive to DUX4 expression in a critical time window.

As reported before [8,10], we only detected a small set of DUX4 transcript-positive nuclei but the proportion (3.8 %, 3/79) in this particular FSHD2 patient myotubes is higher than the reported 0.5% (1/200 in myotube nuclei) [29]. Furthermore, we found that the expression of DUX4 was not as low as stated in previous population studies [8,15] as we detected DUX4 expressed at 10.52 TPM , 33.37 TPM and 68.07 TPM. We further noticed that our DUX4 target-positive myotube nuclei had significant higher expression in markers of later myogenic differentiation. Considering that DUX4 target genes started expression at day 3 (later stage) of differentiation in our time-course, we conclude that the maturation status may have a substantial impact on the expression of DUX4 and target genes. Studies have already found that DUX4 expression caused defects in muscle development and even caused the death of cells [11,12]. One possible explanation is that there is a window of myotube maturation that allows for activation of DUX4 expression and that if the a myotube has not reached or passed that window DUX4 can no longer be expressed. DUX4 is a key regulator during human early embryonic development and dysregulation of DUX4 in muscles is thought to activate germline-specific genes in the process [4,19] that may not be tolerated by muscle cells in early differentiation status [30]. However, for mature myotubes that have matured beyond this critical window and whose cellular signaling and physiology features have been solidly established, this toxicity effects may be alleviated. Further experiments are required to validate this proposed mechanism. Interestingly, we detected some of the highest level of desmin in "DUX4 targets high" myotube nuclei. Desmin is known as one of the important myoblast marker. There is a reported FSHD2 clinical case with desmin accumulation and myopathic pattern observed in muscle biopsy samples [7]. Compared with other cases, this patient has apparent late onset at the age of 60

years and mild symptoms compared with early onset patients. Future studies are needed to measure and estimate the relationship between DUX4 expression and desmin accumulation, and their further influence in the severity of FSHD symptoms.

3.5 Figures

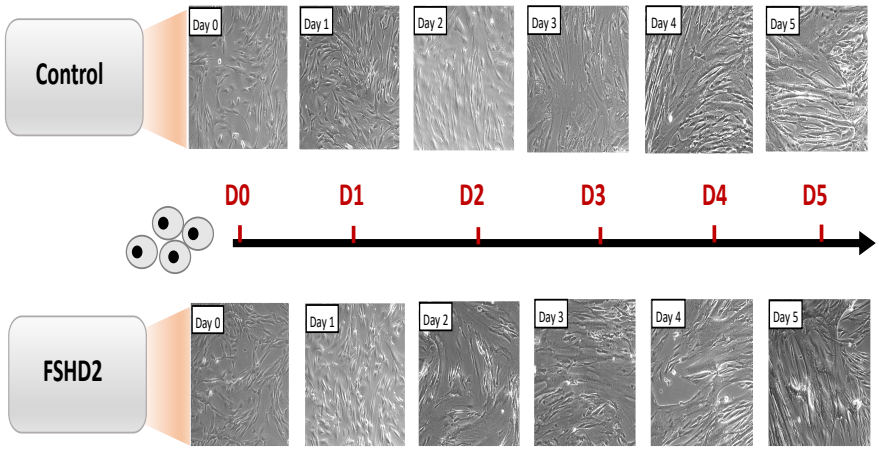


Figure 3.1. Differentiation time-course of control and FSHD2 patient-derived myoblast to myotube. Morphology changes are shown per day during differentiation.

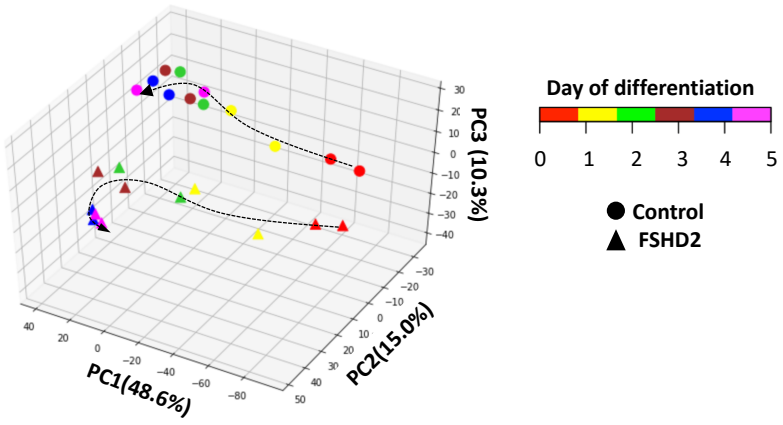


Figure 3.2. Principal component analysis (PCA) on control and FSHD2 myoblast differentiation time-course. Gene expression level is measured each day by using RNA-seq and duplicates are collected per day for both control and FSHD2. Cell types are labeled by distinct shape and time-points are labeled with distinct colored points.

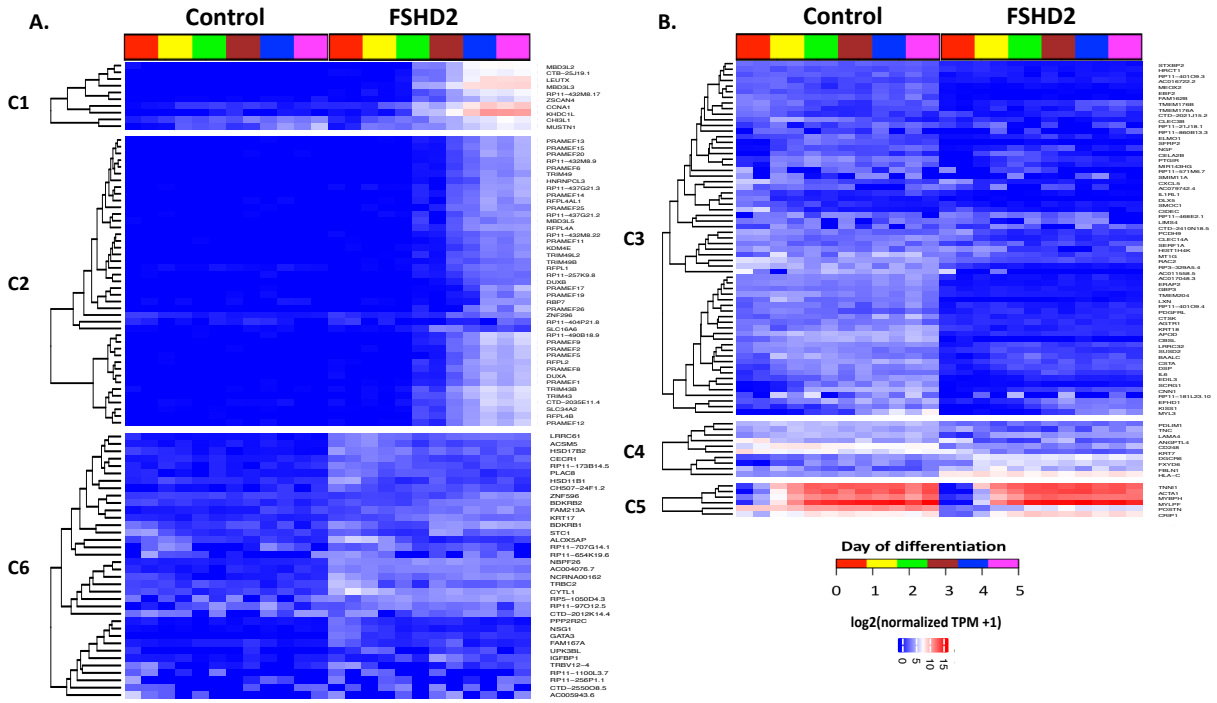


Figure 3.3. K-means clustering on 168 differentially expressed genes for each day in FSHD2 over control and they are clustered into 6 clusters. **(A)** Gene expression heatmap of clusters of genes that are up-regulated in FSHD2. Genes in cluster 1 and 6 are up-regulated at early stage of differentiation while genes in cluster 2 are up-regulated at late stage of differentiation. **(B)** Gene expression heatmap of clusters of genes that are down-regulated in FSHD2. Genes in cluster 3 are down-regulated through the differentiation time-course.

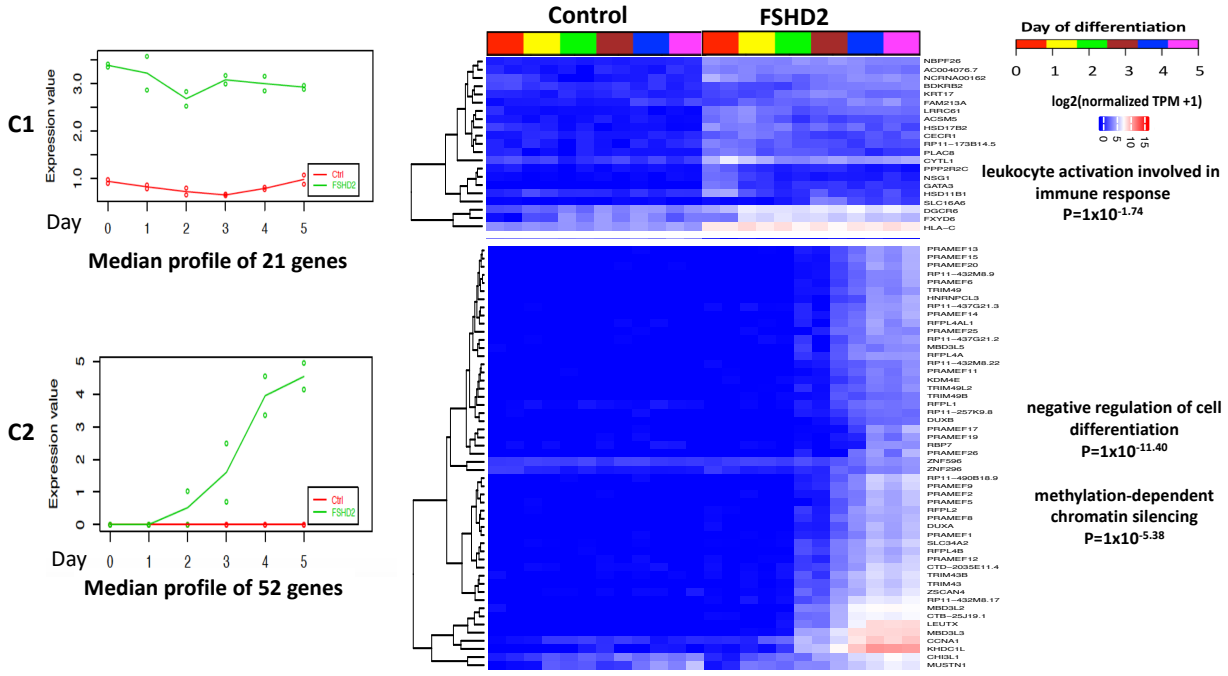


Figure 3.4. MasSigPro identifies 73 genes that are significantly up-regulated in FSHD2 during differentiation time-course compared with control. These 73 genes are further classified into 2 clusters. Genes in cluster 1 are up-regulated at early stage while genes in cluster 2 are up-regulated at late stage. The median expression profiles and gene expression heatmaps are shown for each cluster. The red line labels control and green line labels FSHD2. Heatmap are shown across time-course for both control and FSHD2.

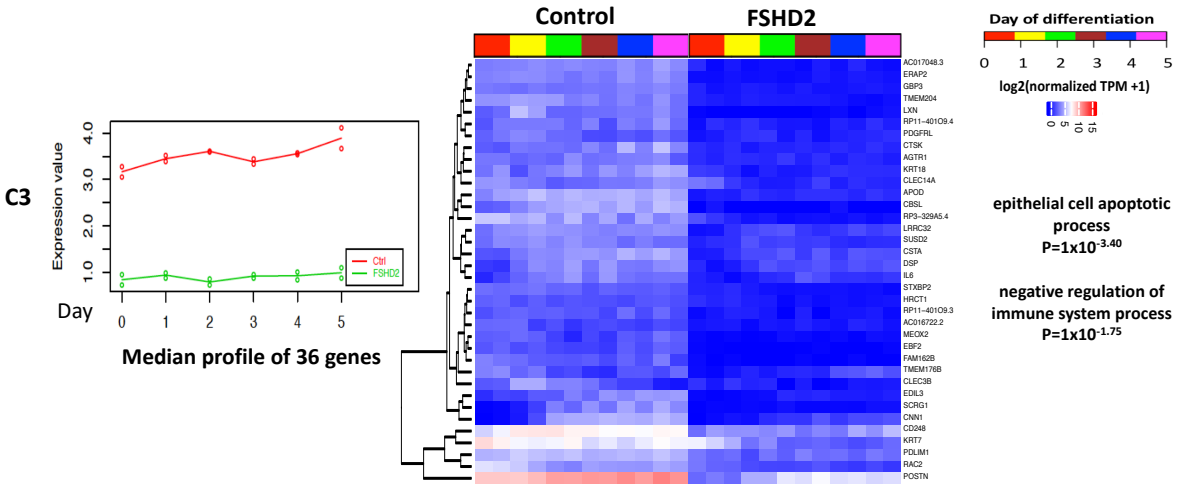


Figure 3.5. MasSigPro identifies 36 genes that are significantly down-regulated in FSHD2 during differentiation time-course compared with control. These 36 genes are classified into one cluster. Genes in cluster 3 are down-regulated at early stage of differentiation. The median expression profiles and gene expression heatmaps are shown for the cluster. The red line labels control and green line labels FSHD2. Heatmap are shown across time-course for both control and FSHD2.

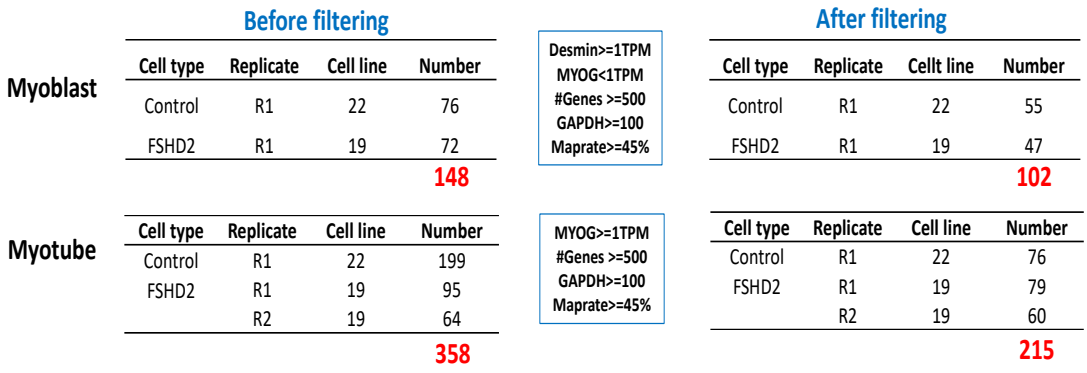


Figure 3.6. Single cells and nuclei are collected on day 3 during FSHD2 and control myoblast differentiation for RNA-seq. 317 cells/nuclei are left for downstream analysis after filtering by alignment quality, number of expressed genes and cell type specific markers.

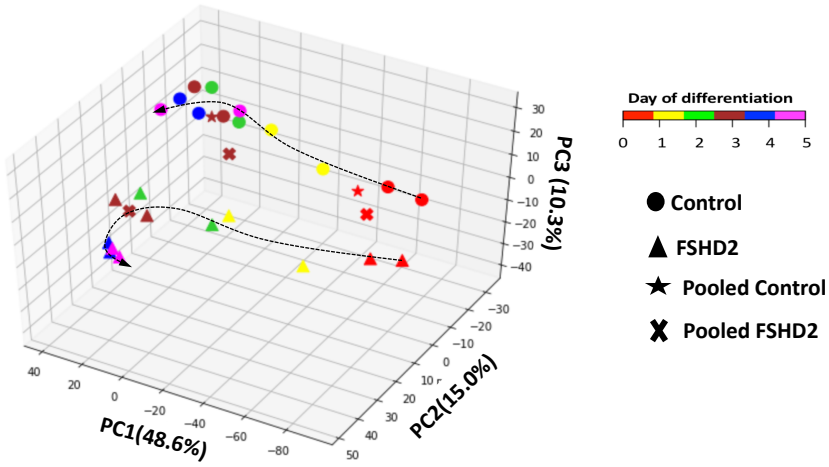


Figure 3.7. Incremental principal component analysis (IPCA) on control and FSHD2 myoblast differentiation time-course with pooled single cells/nuclei samples. In bulk time-course experiments, gene expression level is measured each day by using RNA-seq and duplicates are collected per day for both control and FSHD2. Single cells/nuclei RNA-seq data are pooled for each cell type and gene expression levels are calculated to aligned them with bulk time-course data. Cell types are labeled by distinct shape and time-points are labeled with distinct colored points.

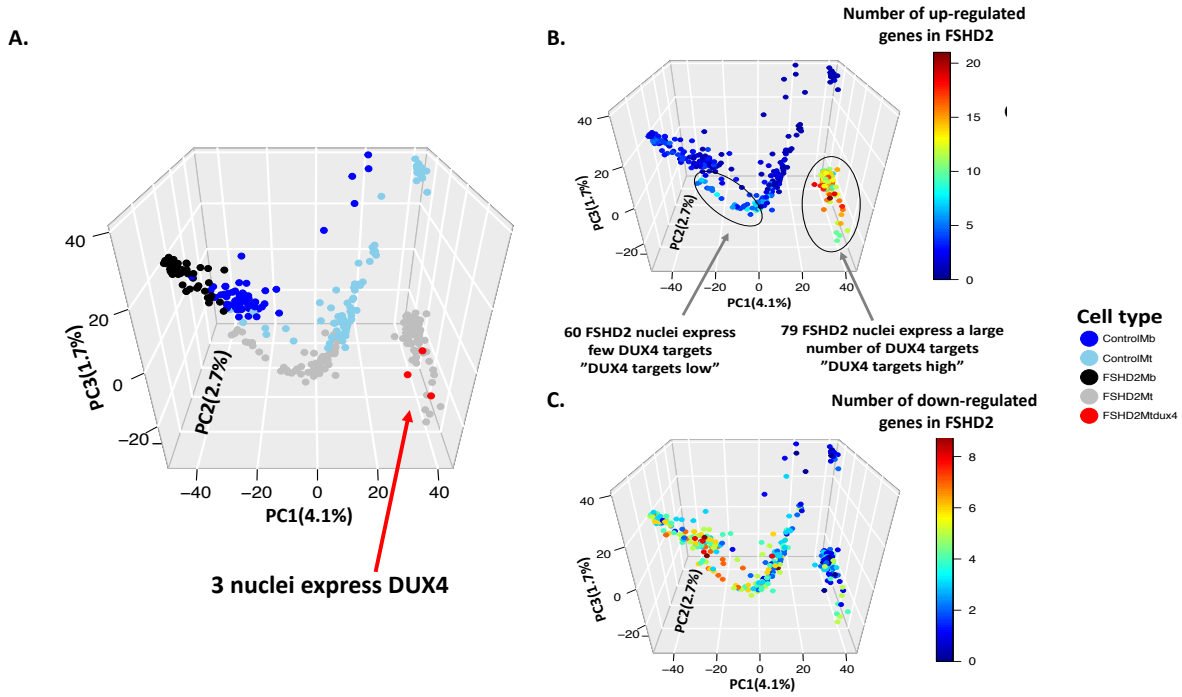


Figure 3.8. (A) PCA of single-cell (for myoblast) and single-nucleus (for myotube) RNA-seq data for both control and FSHD2. Cell types are labeled by distinct colored points and 3 DUX4 positive FSHD2 myotube nuclei are also pointed out by red arrow. **(B)** Same PCA as shown in panel **A**. Single cells/nuclei are colored by the number of genes up-regulated in FSHD2 at late stage of differentiation. **(C)** Same PCA as shown in panel **A**. Single cells/nuclei are colored by the number of genes down-regulated in FSHD2 during differentiation.

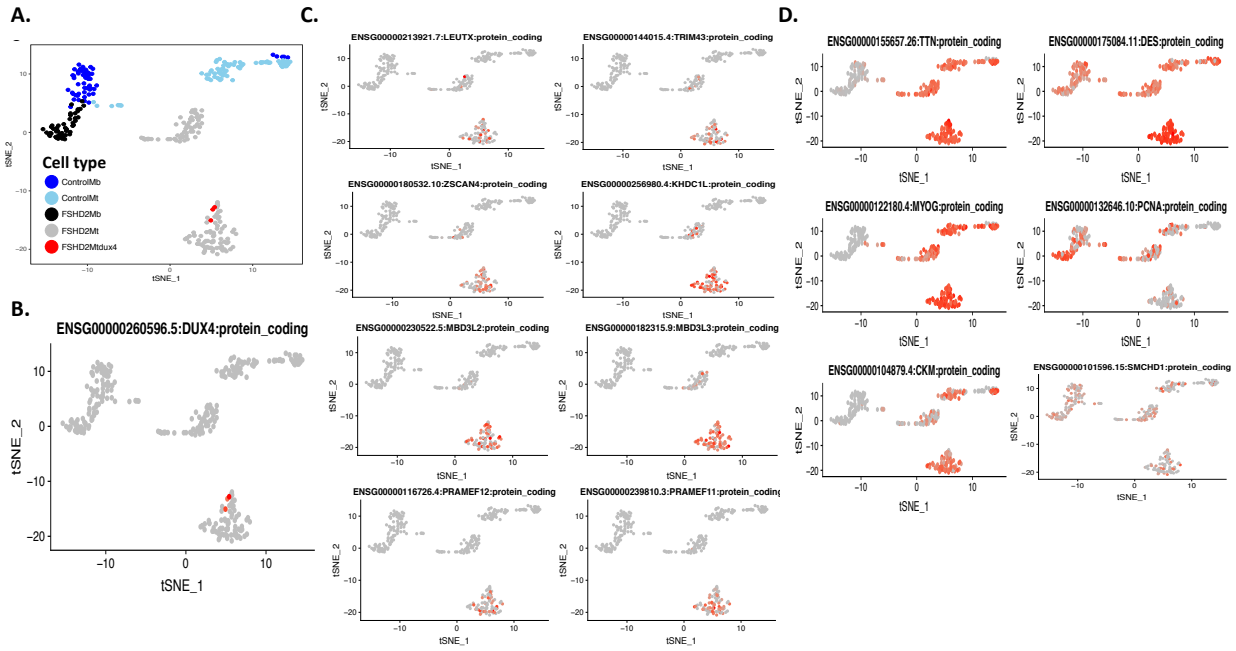


Figure 3.9. (A) t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of 317 single cells/nuclei RNA-seq data. Cell types are labeled by distinct colored points. **(B)** Cells/nuclei are labeled by DUX4 expression level. **(C)** Cells/nuclei are labeled by the gene expression level of known DUX4 target genes. **(D)** Cells/nuclei are labeled by the gene expression level of myogenic differentiation markers. Cells/nuclei with expression of specific genes are colored by red.

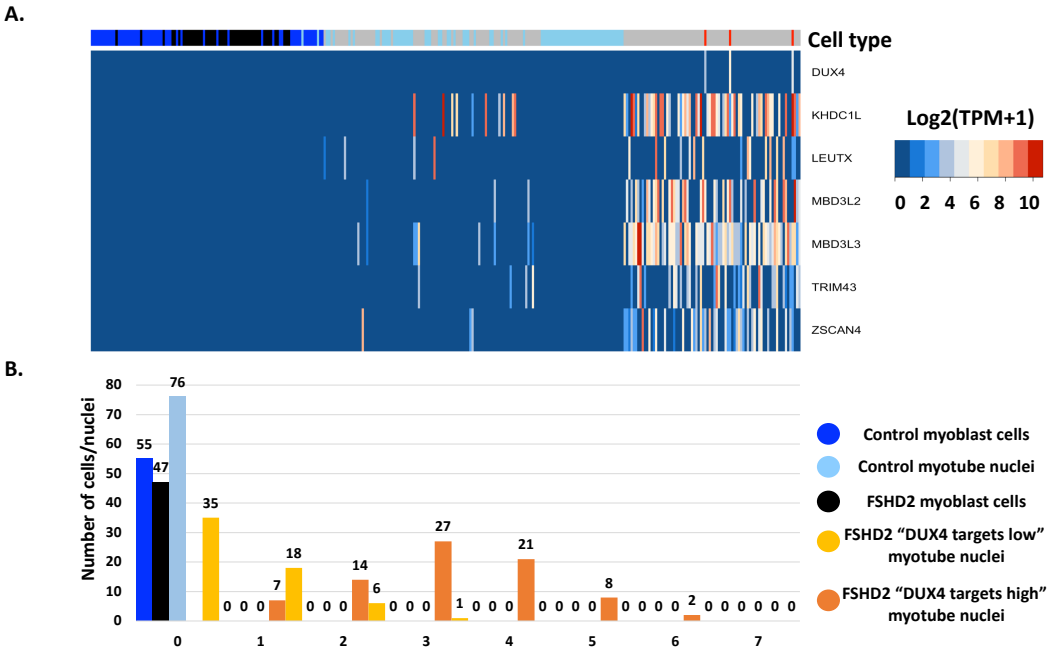


Figure 3.10. (A) Gene expression heatmap of DUX4 and 6 known target genes in single cells/nuclei. Cells and nuclei are ordered by the trajectory in PC1 in Figure 3.8A and cell types are labeled by distinct color in the annotation bar. High expression is shown in red and low expression is shown in blue. **(B)** Histogram of the number of cells/nuclei for each co-expressed targets group in different cell types.

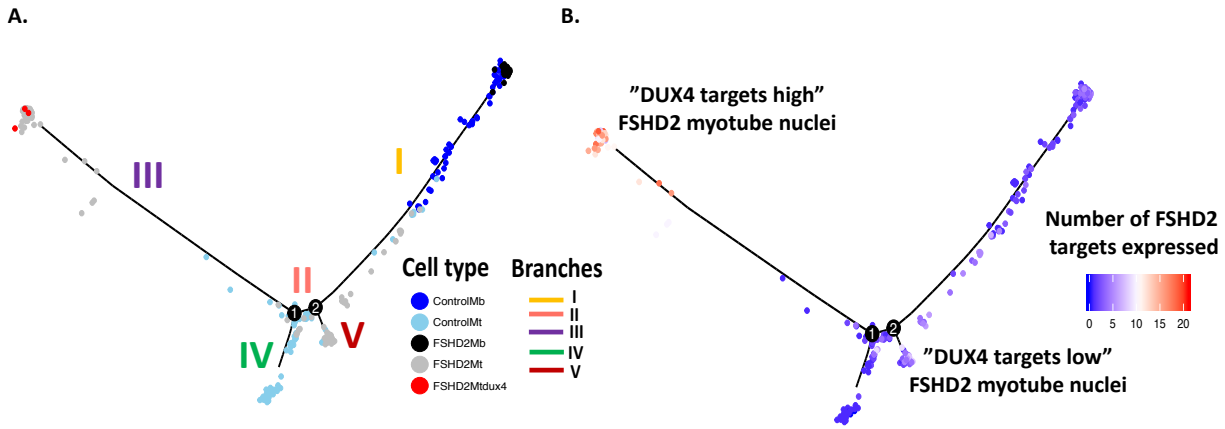


Figure 3.11. (A) Pseudo-temporal ordering of single cells/nuclei using independent component analysis by Monocle. Five branches (I-V) are identified to separate cells into subgroups. Cell types are labeled by distinct colored points. **(B)** Same pseudo ordering as shown panel **A** but cells and nuclei are labeled by the number up-regulated genes in FSHD2 (same genes sets in Figure 3.8B).

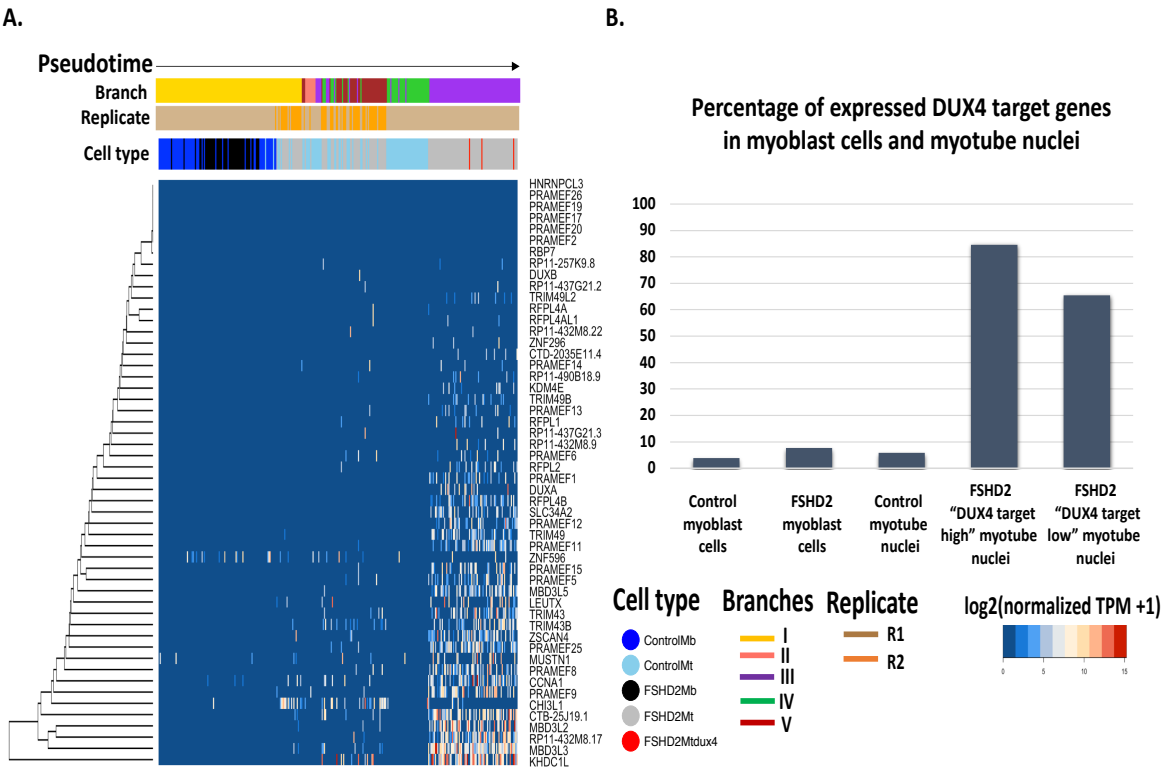


Figure 3.12. (A) Gene expression heatmap of 52 genes (Cluster 2 in Figure 3.4) following pseudo time-course of differentiation. Cells/nuclei are ordered by pseudo time-course and cell types are labeled in the annotation bar. **(B)** Histogram of the percentage of these 52 genes expressed in each cell type.

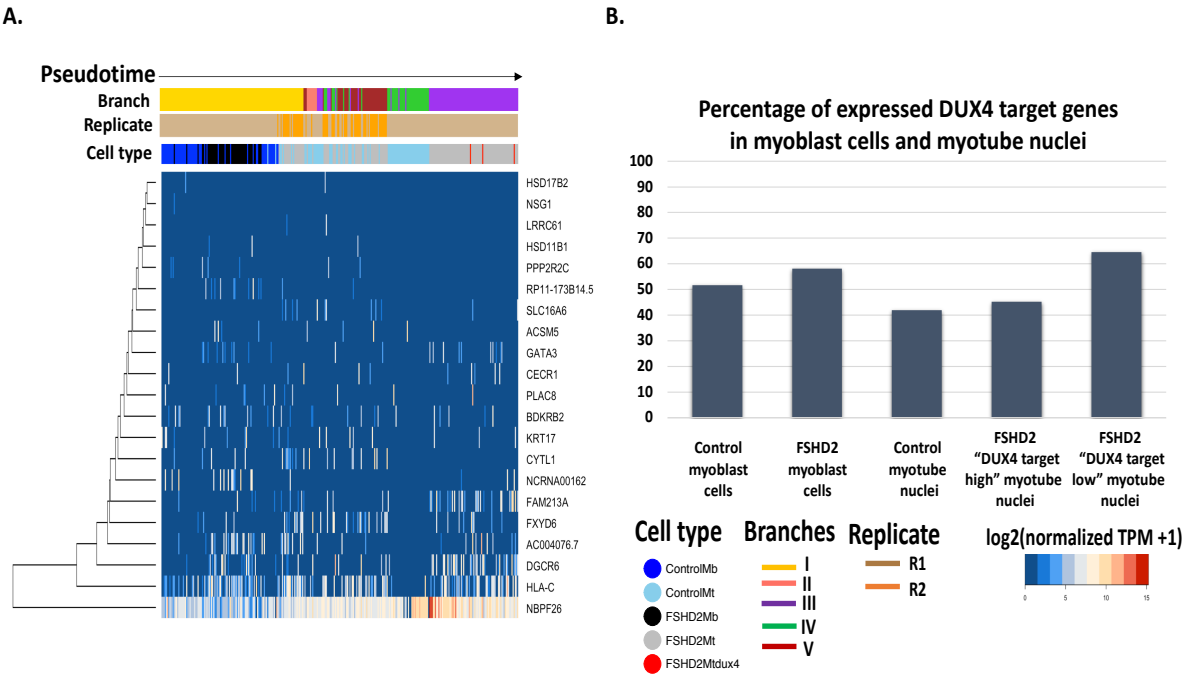


Figure 3.13. (A) Gene expression heatmap of 21 genes (Cluster 1 in Figure 3.4) following pseudo time-course of differentiation. Cells/nuclei are ordered by pseudo time-course and cell types are labeled in the annotation bar. **(B)** Histogram of the percentage of these 21 genes expressed in each cell type.

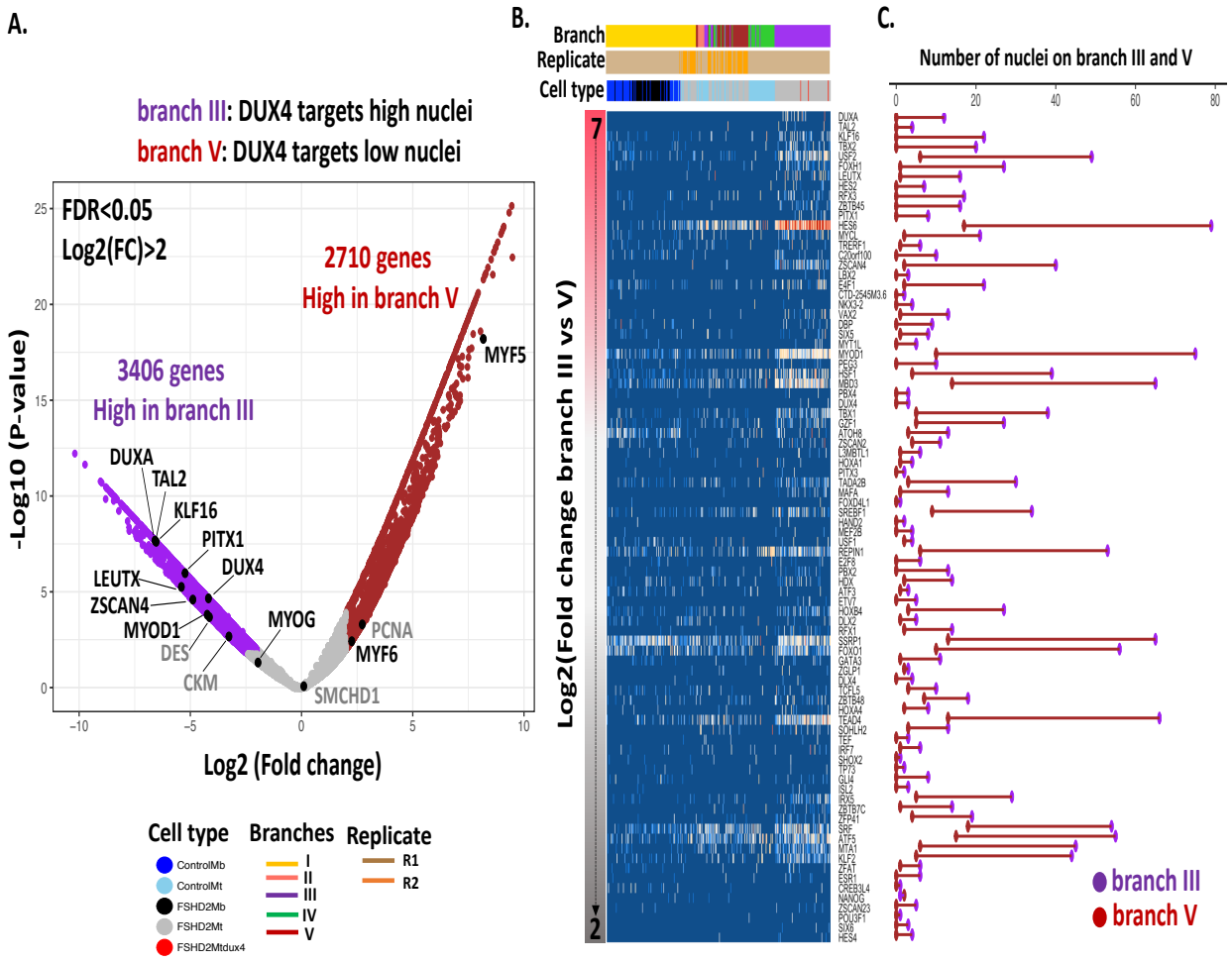


Figure 3.14. (A) Differential expression analysis between branch III (purple) and V (brown) in Figure 3.11A. Differentially expressed genes are colored for each branch (FDR<0.05, log2 (fold change)>2). **(B)** Gene expression heatmap of 84 up-regulated TFs in branch III. TFs are sorted by log2 (fold change) of expression between branch III and V. **(C)** Comparison of the number of nuclei expressing each TFs in both branch III and V.

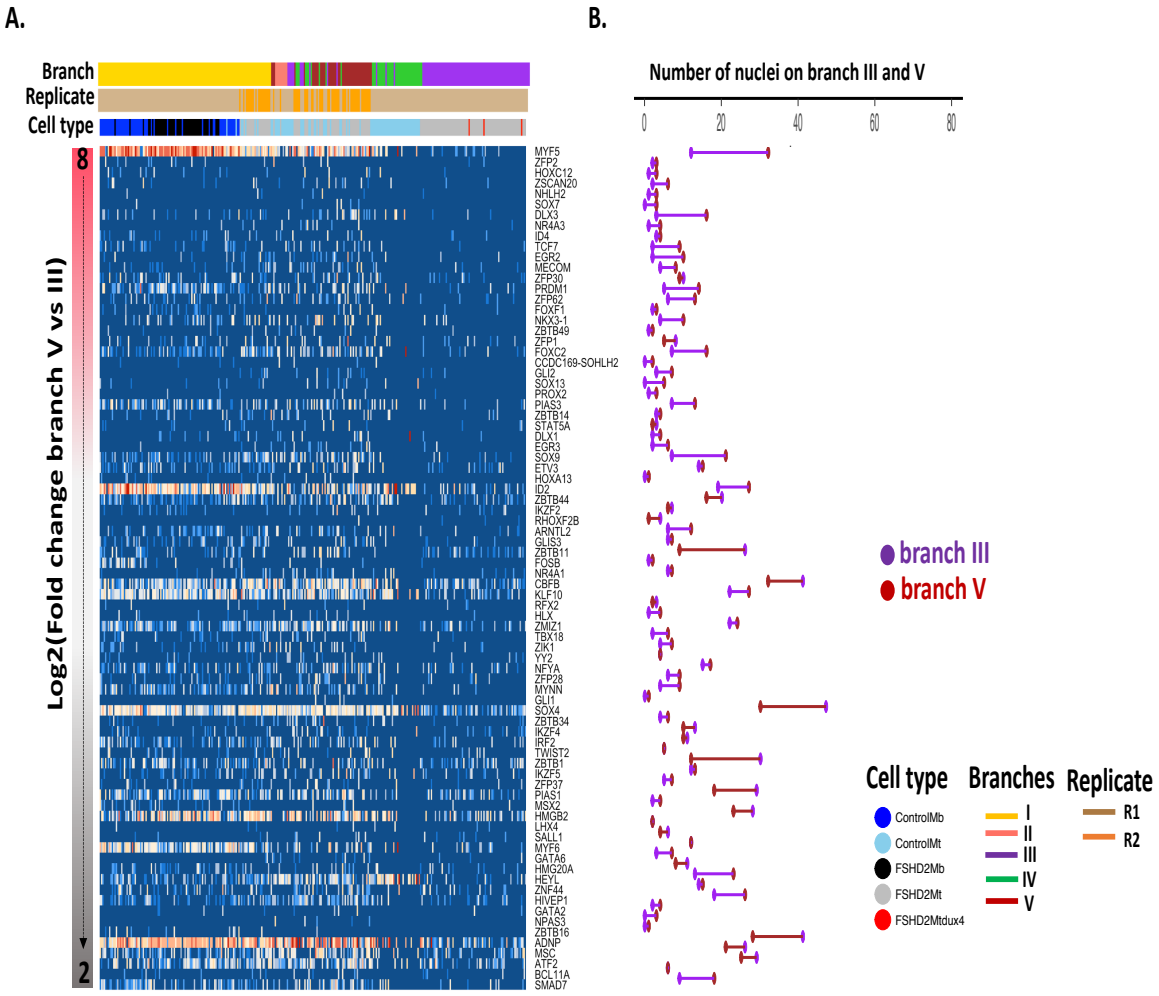


Figure 3.15. (A) Gene expression heatmap of 80 down-regulated TFs in branch III. TFs are sorted by log2 (fold change) of expression between branch III and V. **(B)** Comparison of the number of nuclei expressing each TFs in both branch III and V.

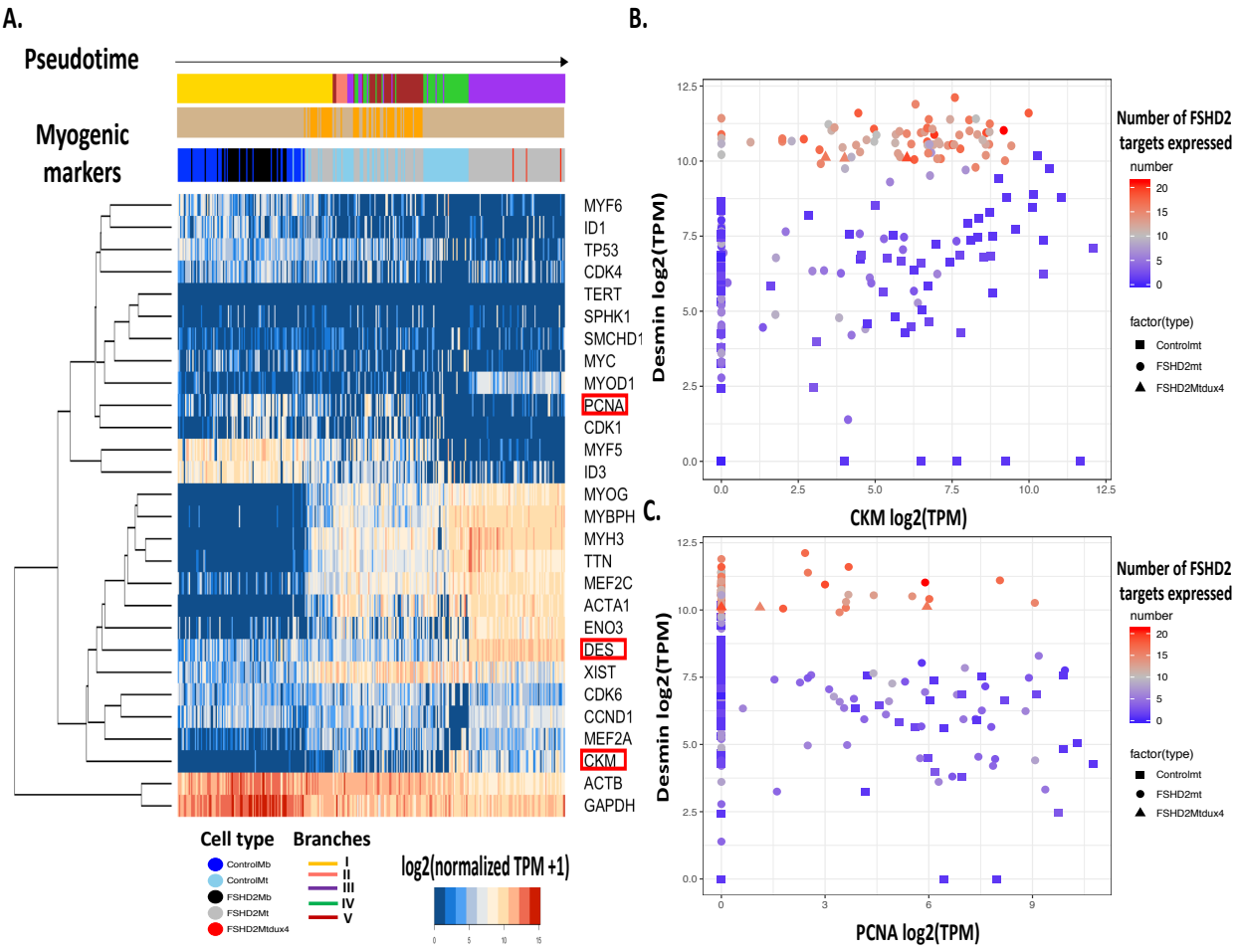


Figure 3.16. (A) Gene expression heatmap of myogenic markers in single cells/nuclei (ordered by pseudo time-course generated by Monocle). **(B)** Scatterplot of gene expression between CKM and desmin in myotube nuclei. Cell types are labeled by distinct shape. Single nuclei are colored by the number of up-regulated genes in FSHD2 during differentiation. **(C)** Scatterplot of gene expression between PCNA and desmin in myotube nuclei. Cell types are labeled by distinct shapes. Single nuclei are colored by the number of up-regulated genes in FSHD2 during differentiation.

Cell type	Sample	Input reads	Uniquely aligned reads	Uniquely aligned efficiency
FSHD2	D0R1	17326007	13506487	77.95%
FSHD2	D0R2	14456677	12254937	84.77%
FSHD2	D1R1	18503424	14466847	78.18%
FSHD2	D1R2	26344335	21869618	83.01%
FSHD2	D2R1	16427260	13006282	79.17%
FSHD2	D2R2	14100175	12083497	85.70%
FSHD2	D3R1	17349735	13757269	79.29%
FSHD2	D3R2	15296642	13077456	85.49%
FSHD2	D4R1	15248765	11784921	77.28%
FSHD2	D4R2	10802918	9234066	85.48%
FSHD2	D5R1	15360796	12174223	79.26%
FSHD2	D5R2	12657431	10992004	86.84%
Control	D0R1	11377692	8924129	78.44%
Control	D0R2	12896530	10687479	82.87%
Control	D1R1	10989815	8789990	79.98%
Control	D1R2	12992691	10919270	84.04%
Control	D2R1	15769963	12486011	79.18%
Control	D2R2	9038334	7800130	86.30%
Control	D3R1	14421611	11605843	80.48%
Control	D3R2	12550498	10718775	85.41%
Control	D4R1	15299123	12280000	80.27%
Control	D4R2	9406365	8060332	85.69%
Control	D5R1	13410795	10827968	80.74%
Control	D5R2	11826248	9945985	84.10%

Table 3.1 Quality control of RNA-seq samples collected during myoblast differentiation for control and FSHD2. Samples are assessed by the number of uniquely aligned reads and uniquely aligned efficiency.

GO ID	GO Terms	Log (P-value)	Genes
GO:0048568	embryonic organ development	-13.633593	NKX3-2,DLX2,GATA3,HOXA1,HOXA4,HOXB4,HSF1,SRF,TBX1,TBX2,TEAD4,FOXH1,IRX5,VAX2,MBD3,ZFAT,E2F8,KLF2,HES2,HES6,HES4,PBX2,PITX1,ATOH8,MYCL
GO:0007389	pattern specification process	-9.9056794	NKX3-2,DLX2,HOXA4,HOXB4,PBX2,RFX3,SRF,TBX1,TBX2,FOXH1,VAX2,HES2,HES6,HES4
GO:0007423	sensory organ development	-6.9894694	NKX3-2,DLX2,GATA3,HOXA1,MYCL,SIX6,SRF,TBX1,TBX2,IRX5,VAX2,SIX5
GO:0046016	positive regulation of transcription by glucose	-6.890547	SRF,USF1,USF2,ATF3,FOXO1,HSF1,MYOD1,SREBF1,MBD3
GO:0045165	cell fate commitment	-5.7121348	DLX2,GATA3,MYOD1,PITX1,TBX1,TBX2,MYT1L,ISL2,FOXO1,NKX3-2
GO:0007368	determination of left/right symmetry	-5.6428846	NKX3-2,RFX3,SRF,TBX1,TBX2,FOXH1,HOXA1,MYOD1,ATF3,PITX1,TEAD4,MEF2B,GATA3,HSF1,FOXO1,KLF2,E2F8,MBD3
GO:0035270	endocrine system development	-5.4771727	FOXO1,GATA3,PITX1,RFX3,SRF,TBX1,HSF1,KLF2,MBD3,ZFAT,E2F8,HOXB4,IRX5,IRF7,L3MBTL1,MYOD1
GO:0048732	gland development	-4.9784822	DBP,ESR1,GATA3,PITX1,SRF,TBX1,TBX2,USF2,E2F8,HOXB4,FOXH1,GZF1,TEAD4,ZSCAN4,DLX2
GO:0009299	mRNA transcription	-4.1520605	HSF1,SREBF1,SRF,ESR1,GATA3,RFX3,TRERF1,MAFA,TBX2,MBD3,KLF2,IRX5,FOXO1
GO:0021772	olfactory bulb development	-3.825355	DLX2,SRF,ATF5,PITX1,TAL2,VAX2,TBX1,MBD3,GATA3,MYCL,HES2,HES6,HES4,ISL2,NKX3-2
GO:0036003	positive regulation of transcription from RNA polymerase II promoter in response to stress	-3.7849687	ATF3,HSF1,KLF2,FOXO1,SREBF1,TBX1,USF1,ESR1,GATA3,SRF
GO:0048589	developmental growth	-3.7316995	ESR1,GATA3,HSF1,MYOD1,SRF,TAL2,TBX2,KLF2,ATF5,ATOH8,SIX5,ZBTB7C
GO:0071392	cellular response to estradiol stimulus	-3.7079458	ESR1,HSF1,MYOD1,MBD3,IRF7,SREBF1,TBX1,FOXH1,FOXO1,USF1,KLF2

Table 3.2. Gene ontology terms associated with 84 up-regulated TFs in branch III in Figure 3.14B.

GO ID	GO Terms	Log (P-value)	Genes
GO:0001655	urogenital system development	-12.859604	FOXF1,FOXC2,GATA2,GLI1,GLI2,ID2,ID4,SMAD7,NKX3-1,SALL1,SOX4,SOX9,ZBTB16,TBX18,HEYL,DLX1,EGR2,MSX2,MYF5,MYF6,NR4A3,ATF2,PRDM1,HOXA13,KLF10,STAT5A,GATA6,MSC,FOSB
GO:1902105	regulation of leukocyte differentiation	-12.693537	PRDM1,CBFB,EGR3,GATA2,GLI2,HLX,ID2,SMAD7,KLF10,ZBTB16,SOX13,PIAS3,ZBTB1,ZMIZ1,HMGB2,NR4A3,SOX4,TCF7,FOXF1,FOXC2,DLX1,ID4,MSX2,SOX9,HMG20A,TWIST2,MYF5
GO:0007507	heart development	-12.47361	PRDM1,ATF2,FOXF1,FOXC2,GATA2,GATA6,GLI1,GLI2,ID2,SMAD7,MSX2,NKX3-1,SALL1,SOX4,SOX9,ZBTB14,HEYL,ZMIZ1,DLX3,EGR3,NR4A1,PROX2,NHLH2
GO:0048565	digestive tract development	-10.650868	PRDM1,FOXF1,GATA2,GATA6,GLI1,GLI2,HLX,ID2,SALL1,TCF7,FOXC2,MSX2,MYF5,SOX4,SOX9,ZBTB16,NR4A3,ID4,NKX3-1,STAT5A,ZBTB1,DLX1,MYF6,SOX13,ZMIZ1,ADNP,CBFB,DLX3,PROX2
GO:0001503	ossification	-9.4301818	CBFB,EGR2,FOXC2,GLI1,GLI2,ID2,ID4,MSX2,MYF5,SOX9,KLF10,ZBTB16,TWIST2,ATF2,GATA2,NR4A1,NR4A3,PIAS1,DLX1,SOX4,HMGB2,PINX1,NKX3-1,ADNP,HMG20A,SALL1
GO:0048608	reproductive structure development	-8.7979622	PRDM1,DLX3,GATA2,GATA6,GLI1,GLI2,HMGB2,ID4,NKX3-1,SALL1,SOX9,TCF7,LHX4,RFX2,ZBTB16,ATF2,MECOM,HMG20A,ZBTB1
GO:0060537	muscle tissue development	-8.0742405	EGR2,FOXC2,GATA6,GLI1,HLX,ID2,SMAD7,MYF5,MYF6,SOX9,MSC,HEYL,EGR3,FOXF1,PIAS1,NKX3-1,NR4A3
GO:0050673	epithelial cell proliferation	-6.7101467	ATF2,EGR3,GATA2,GLI1,HMGB2,NR4A1,ID2,NKX3-1,SOX9,STAT5A,NR4A3,FOXF1,FOXC2,EGR2,KLF10
GO:0035270	endocrine system development	-6.6519949	GATA2,GATA6,GLI1,GLI2,SALL1,SOX4,SOX9,PRDM1,ZBTB14,HEYL,KLF10,CBFB,TBX18,TCF7
GO:0048511	rhythmic process	-6.462465	EGR2,EGR3,ID2,ID4,NFYA,NHLH2,KLF10,ADNP,ARNTL2
GO:0001655	urogenital system development	-12.859604	FOXF1,FOXC2,GATA2,GLI1,GLI2,ID2,ID4,SMAD7,NKX3-1,SALL1,SOX4,SOX9,ZBTB16,TBX18,HEYL,DLX1,EGR2,MSX2,MYF5,MYF6,NR4A3,ATF2,PRDM1,HOXA13,KLF10,STAT5A,GATA6,MSC,FOSB
GO:1902105	regulation of leukocyte differentiation	-12.693537	PRDM1,CBFB,EGR3,GATA2,GLI2,HLX,ID2,SMAD7,KLF10,ZBTB16,SOX13,PIAS3,ZBTB1,ZMIZ1,HMGB2,NR4A3,SOX4,TCF7,FOXF1,FOXC2,DLX1,ID4,MSX2,SOX9,HMG20A,TWIST2,MYF5
GO:0007507	heart development	-12.47361	PRDM1,ATF2,FOXF1,FOXC2,GATA2,GATA6,GLI1,GLI2,ID2,SMAD7,MSX2,NKX3-1,SALL1,SOX4,SOX9,ZBTB14,HEYL,ZMIZ1,DLX3,EGR3,NR4A1,PROX2,NHLH2

Table 3.3. Gene ontology terms associated with 80 down-regulated TFs in branch III in Figure 3.15A.

3.6 Methods

3.6.1 Human myoblast culture and differentiation

Human control and FSHD2 patient-derived myoblast cells were grown on dishes coated with collagen in high glucose DMEM (Gibco) supplemented with 20% FBS (Omega Scientific, Inc.), 1% Pen-Strep (Gibco), and 2% Ultrasor G (Crescent Chemical Co.) [28]. Upon reaching 80% confluence, differentiation was induced by using high glucose DMEM medium supplemented with 2% FBS and ITS supplement (insulin 0.1%, 0.000067% sodium selenite, 0.055% transferrin; Invitrogen). Fresh differentiation medium was changed every 24hrs.

3.6.2 Bulk RNA-seq library construction

Total RNA was extracted by using RNeasy kit (QIAGEN). Between 19 and 38 ng of RNA were converted to cDNA using the SmartSeq 2 protocol [31]. Libraries were constructed by using the Nextera DNA Sample Preparation Kit (Illumina). Libraries were quality-controlled prior to sequencing based on Agilent 2100 Bioanalyzer profiles and normalized using the KAPA Library Quantification Kit (Illumina). The libraries were sequenced using paired-end 75bp mode on Illumina NextSeq500 platform with around 15 million reads per sample.

3.6.3 Single nucleus isolation and cell capture, RNA-seq library construction

Myoblast single cells were isolated from 6 cm dishes by washing with PBS then lifting the cells with trypsin. The myoblasts were spun down, resuspended in PBS and kept on ice.
Myotube single nuclei were isolated from mononucleated cells (MNCs) by washing a 6 cm dish once with trypsin then adding trypsin for about 5 min until myotubes lifted off the plate and MNCs were still attached. The cells were spun at 2000 rpm for 2 min and resuspended in 500 ul lysis buffer (0.02% IGEPAL CA-630, 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$). The cells were left at room temperature for 3 minutes then filtered through a 40 um cell filter to clear debris. The flow through was spun at 4000 rpm for 1 minute and resuspended in 100 ul of PBS and kept on ice. The single cells and nuclei were captured using the Fluidigm C1 on the large IFCs (17-25 um) and medium IFCs (10-17 um), respectively, at a density between 340 and 640 nuclei/ul in a volume of 10 ul. Each cell or nucleus was visually confirmed using the LIVE/DEAD kit (Thermo Fisher Scientific). Cell and nucleus loading, lysis, reverse transcription and preamplification were performed on the Fluidigm C1. Harvested cDNA was normalized to

approximately 0.1 ng/ul for tagmentation and library prep which was performed with the Nextera XT Library Prep Kit from Illumina according to Fluidigm's protocol. Libraries were base-pair selected based on Agilent 2100 Bioanalyzer profiles and normalized determined by KAPA Library Quantification Kit (Illumina). The libraries were sequenced using paired-end 75bp mode on Illumina NextSeq500 platform with around 1-3 million reads per sample.

3.6.4 Read alignment and expression analysis

Raw reads were mapped to hg38 using STAR (version 2.5.1b) [32] using defaults except with a maximum of 10 mismatches per pair, a ratio of mismatches to read length of 0.07, and a maximum of 10 multiple alignments. Quantitation was performed using RSEM (version 1.2.31) [33] with the defaults, and results were output in transcripts per million (TPM).

3.6.5 Differential expression analysis

Differential expression analysis per day during differentiation was done by using edgeR [34] with FDR < 0.05. Clustering of differentially expressed genes across the time-course was done by using maSigPro [35].

3.6.6 Quality control of single cell/nucleus

Myoblast cells were kept for downstream analysis if Desmin expression >=1TPM, MYOG <1TPM, number of expressed genes was more than 500 and expression level of GAPDH is higher than 100TPM. We only kept cells with uniquely mapped efficiency higher than 50%. Myotube nuclei were kept for downstream analysis if MYOG expression >=1TPM, number of expressed genes was more than 500 and expression level of GAPDH is higher than 100TPM. We also only kept cells with uniquely mapped efficiency higher than 50%.

3.6.7 Dimensionality reduction analysis

Incremental PCA analysis was performed by *IncrementalPCA* function from scikit-learn [36], python 2. t-SNE analysis was performed by using Seurat [37].

3.6.8 Pseudo-time analysis

Monocle (version 2.4.0) [38] was used to determine the pseudo-time trajectory with "DDRTree" methods to reduce the space to maximum three components.

3.6.9 Gene ontology analysis

Gene ontology analysis was done by using Metascape [39] with FDR<0.05.

3.7 References

1.Tawil, R. and S.M. Van Der Maarel, *Facioscapulohumeral muscular dystrophy.* Muscle Nerve, 2006. **34**(1): p. 1-15.

2.Zeng, W., Chen, Y. Y., Newkirk, D. A., Wu, B., Balog, J., Kong, X., ... & Mortazavi, A. (2014). Genetic and Epigenetic Characteristics of FSHD-Associated 4q and 10q D4Z4 that are Distinct from Non-4q/10q D4Z4 Homologs. *Human mutation*, *35*(8), 998-1010.

3.Young, J. M., Whiddon, J. L., Yao, Z., Kasinathan, B., Snider, L., Geng, L. N., ... & Tapscott, S. J. (2013). DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. *PLoS genetics*, *9*(11), e1003947.

4.Geng, L. N., Yao, Z., Snider, L., Fong, A. P., Cech, J. N., Young, J. M., ... & Tapscott, S. J. (2012). DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Developmental cell*, *22*(1), 38-51.

5.Lemmers, R. J., Tawil, R., Petek, L. M., Balog, J., Block, G. J., Santen, G. W., ... & Krom, Y. D. (2012). Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nature genetics*, *44*(12), 1370.

6.Sacconi, S., Lemmers, R. J., Balog, J., Van Der Vliet, P. J., Lahaut, P., Van Nieuwenhuizen, M. P., ... & Casarin, A. (2013). The FSHD2 gene SMCHD1 is a modifier of disease severity in families affected by FSHD1. *The American Journal of Human Genetics*, *93*(4), 744-751.

7.Larsen, M., Rost, S., El Hajj, N., Ferbert, A., Deschauer, M., Walter, M. C., ... & Müller, C. R. (2015). Diagnostic approach for FSHD revisited: SMCHD1 mutations cause FSHD2 and act as modifiers of disease severity in FSHD1. *European Journal of Human Genetics*, *23*(6), 808.

8.Snider, L., Geng, L. N., Lemmers, R. J., Kyba, M., Ware, C. B., Nelson, A. M., ... & Miller, D. G. (2010). Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS genetics*, *6*(10), e1001181.

9.Lemmers, R. J., Van Der Vliet, P. J., Klooster, R., Sacconi, S., Camaño, P., Dauwerse, J. G., ... & Miller, D. G. (2010). A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*, 1189044.

10.Himeda, C. L., Jones, T. I., & Jones, P. L. (2015). Facioscapulohumeral muscular dystrophy as a model for epigenetic regulation and disease. *Antioxidants & redox signaling*, *22*(16), 1463-1482.

11.Bosnakovski, D., Xu, Z., Gang, E. J., Galindo, C. L., Liu, M., Simsek, T., ... & Belayew, A. (2008). An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. *The EMBO journal*, *27*(20), 2766-2779.

12.Vanderplanck, C., Ansseau, E., Charron, S., Stricwant, N., Tassin, A., Laoudj-Chenivesse, D., ... & Belayew, A. (2011). The FSHD atrophic myotube phenotype is caused by DUX4 expression. *PloS one*, *6*(10), e26820.

13.Zeng, W., De Greef, J. C., Chen, Y. Y., Chien, R., Kong, X., Gregson, H. C., ... & Kimonis, V. E. (2009). Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *PLoS genetics*, *5*(7), e1000559.

14.Zeng W, Ball AR, Yokomori K (2012) The epigenetics of facioscapulohumeral muscular dystrophy. In: Appasani K, editor. Epigenomics: From Chromatin Biology to Therapeutics. Cambridge: Cambridge University Press. pp. 347-361.

15.Jones, T. I., Chen, J. C., Rahimov, F., Homma, S., Arashiro, P., Beermann, M. L., ... & Wagner, K. R. (2012). Facioscapulohumeral muscular dystrophy family studies of DUX4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Human molecular genetics*, *21*(20), 4419-4430.

16.Broucqsault, N., Morere, J., Gaillard, M. C., Dumonceaux, J., Torrents, J., Salort-Campana, E., ... & Ferreboeuf, M. (2013). Dysregulation of 4q35-and muscle-specific genes in fetuses with a short D4Z4 array linked to facio-scapulo-humeral dystrophy. *Human molecular genetics*, *22*(20), 4206-4214.

17.Ferreboeuf, M., Mariot, V., Bessieres, B., Vasiljevic, A., Attié-Bitach, T., Collardeau, S., ... & Rameau, P. (2013). DUX4 and DUX4 downstream target genes are expressed in fetal FSHD muscles. *Human molecular genetics*, *23*(1), 171-181.

18.Rahimov, F., King, O. D., Leung, D. G., Bibat, G. M., Emerson, C. P., Kunkel, L. M., & Wagner, K. R. (2012). Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. *Proceedings of the National Academy of Sciences*, 201209508.

19.Rickard, A. M., Petek, L. M., & Miller, D. G. (2015). Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. *Human molecular genetics*, *24*(20), 5901-5914.

20.Krom, Y. D., Thijssen, P. E., Young, J. M., den Hamer, B., Balog, J., Yao, Z., ... & Rijkers, T. (2013). Intrinsic epigenetic regulation of the D4Z4 macrosatellite repeat in a transgenic mouse model for FSHD. *PLoS genetics*, *9*(4), e1003415.

21.Tsumagari, K., Chang, S. C., Lacey, M., Baribault, C., Chittur, S. V., Sowden, J., ... & Ehrlich, M. (2011). Gene expression during normal and FSHD myogenesis. *BMC medical genomics*, *4*(1), 67.

22.van Overveld, P. G., Lemmers, R. J., Sandkuijl, L. A., Enthoven, L., Winokur, S. T., Bakels, F., ... & van der Maarel, S. M. (2003). Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nature genetics*, *35*(4), 315.

23.Ashe, A., Morgan, D. K., Whitelaw, N. C., Bruxner, T. J., Vickaryous, N. K., Cox, L. L., ... & Anderson, G. J. (2008). A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. *Genome biology*, *9*(12), R182.

24.Blewitt, M. E., Gendrel, A. V., Pang, Z., Sparrow, D. B., Whitelaw, N., Craig, J. M., ... & Kay, G. F. (2008). SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nature genetics*, *40*(5), 663.

25.Gendrel, A. V., Apedaile, A., Coker, H., Termanis, A., Zvetkova, I., Godwin, J., ... & Giannoulatou, E. (2012). Smchd1-dependent and-independent pathways determine developmental dynamics of CpG island methylation on the inactive X chromosome. *Developmental cell*, *23*(2), 265-279.

26.Gendrel, A. V., Tang, Y. A., Suzuki, M., Godwin, J., Nesterova, T. B., Greally, J. M., ... & Brockdorff, N. (2013). Epigenetic functions of smchd1 repress gene clusters on the inactive X chromosome and on autosomes. *Molecular and cellular biology*, MCB-00145.

27.Yao, Z., Snider, L., Balog, J., Lemmers, R. J., Van Der Maarel, S. M., Tawil, R., & Tapscott, S. J. (2014). DUX4-induced gene expression is the major molecular signature in FSHD skeletal muscle. *Human molecular genetics*, *23*(20), 5342-5352.

28.Zeng, W., Jiang, S., Kong, X., El-Ali, N., Ball Jr, A. R., Ma, C. I. H., ... & Mortazavi, A. (2016). Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic acids research*, *44*(21), e158-e158.

29.Tassin, A., Laoudj-Chenivesse, D., Vanderplanck, C., Barro, M., Charron, S., Ansseau, E., ... & Belayew, A. (2013). DUX 4 expression in FSHD muscle cells: how could such a rare protein cause a myopathy?. *Journal of cellular and molecular medicine*, *17*(1), 76-89.

30.Campbell, A. E., Belleville, A., Resnick, R., Shadle, S. C., & Tapscott, S. J. (2018). Facioscapulohumeral dystrophy: Activating an early embryonic transcriptional program in human skeletal muscle. *Human molecular genetics*.

31.Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, *9*(1), 171.

32.Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

33.Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12*(1), 323.

34.Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

35.Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096-1102.

36.Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

37.Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411.

38.Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., ... & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381.

39.Tripathi, S., Pohl, M. O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D. A., ... & Yángüez, E. (2015). Meta-and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. *Cell host & microbe*, *18*(6), 723-735.

CHAPTER 4

Comparative chromatin dynamics of definitive endoderm differentiation

Note: (1) Christina R. Wilcox cultured human embryonic stem cell.

(2) Xinyi Ma sequenced all samples.

(3) Dr. Ali Mortazavi conceived the idea and provide continued support and guidance throughout the project.

Chapter 4

Comparative chromatin dynamics of definitive endoderm differentiation

4.1 Abstract

The pluripotency of embryonic stem cells (ESCs) to differentiate into different germ layers makes it an attractive model to study cell-fate commitment and differentiation. However, the underlying gene regulation networks and their level of conservation in mammals are not fully understood. We focus on differentiating ESCs to definitive endoderm (DE) in human, mouse and rat *in vitro* when exposed to Activin A. We applied RNA-seq and ATAC-seq to observe the changes transcriptome and open chromatin region during the time-course DE differentiation in human, mouse, and rat. We observed both conserved and species-specific dynamic gene expression and chromatin accessibility changes during DE differentiation, which we used to define conserved regulatory modules enriched with highly expressed target genes and motifs of stage-specific transcription factors (TFs). We selected 14 key TFs in DE differentiation and further used chromatin accessibility footprints and gene expression to build gene regulatory network (GRN) of TF binding in the stem cell and definitive endoderm stages. Our study is the first global view of potential regulatory interactions between key endoderm genes across 3 mammalian species. We observed that the conservation of the ES GRNs were more highly conserved than those in DE.

4.2 Introduction

One of the key questions in developmental biology is how the differentiation and specification of different cell types are precisely controlled by the genome by gene regulatory networks (GRNs) of transcription factors and signaling pathways. Germ layer specification, one of the earliest developmental events preceding the formation of most cell types and tissues, is one attractive model for studying some of the first active GRNs. Endoderm formation was first defined as the innermost tissue or germ layer in all metazoan embryos. It gives rise to highly specialized epithelial cell types that encompass the respiratory and digestive tracts [1]. After gastrulation, these endodermal cells would differentiate further to form into the embryonic gut tube and organs such as the intestine, pancreas, liver, lungs and thyroid gland. Many studies have characterized the signaling and molecular pathways that specify the endoderm formation in

different species, including Xenopus [2,3], zebrafish [4], mouse [1,5], human [1,5], and even some invertebrate species like *C.elegans* [6], *Ciona intestinalis* [7] , sea urchin [8]. Comparative studies of vertebrates show that Nodal signaling plays an evolutionarily conserved player in the early initiation of endoderm differentiation in all species and plays a central role for inducing the endoderm GRN [5,9,10,11]. Nodal ligands are members of the TGFß family of secreted growth factors [1]. Nodal signaling is required for initiating both endoderm and mesoderm development with high concentrations of Nodal specifying endoderm formation while low concentrations specifying mesoderm formation [9,12,13]. One active area of research is to establish and to characterize the boundaries between endoderm and mesoderm, which involves a specific sets of core genes reinforcing endodermal fate while repressing mesodermal fate at the same time [1,5]. For example, in Xenopus, the T-box transcription factor Vegt is a master regulator of endoderm formation by activating Nodal genes and directly activates downstream TFs such as sox17a, sox7, cer1 and gsc during mesendoderm formation [14-16]. As Vegt is not conserved in mammalian species, another T-box gene EOMES is likely to play a similar role in early endoderm formation and to regulate a similar set of targets as Vegt does in Xenopus during human endoderm differentiation such as MIXL1, CER1, SOX17, FOXA2 [5,17-19]. Foxh1/Smad2/3 are another group of regulators of endoderm initiation and their targets are conserved in Xenopus [20,21], mouse [22,23], zebrafish [24] and human [25], including CER1, PITX2, GSC and MIXL1. The TFs Gata4, Gata6, Eomes and Foxa2 are also regulated by Foxh1-independent mechanism in Xenopus [3,26]. Previous studies have shown that Mix1 can further regulate endoderm lineage development by directly activating gsc and cer1 but repressing t genes expression in Xenopus [27,28,29], but whether this mechanism is conserved in mammalian species is unclear. Although a handful of TFs such as Mix, Sox17, and Foxa are known highly play highly conserved roles in multiple species, there are other TFs with species-specific regulations. For example, Gata4 and Gata6 regulate endoderm formation in Xenopus and zebrafish but their specific functions are poorly understood in human and mouse [5].

Previous functional genomics studies have shown that stage-specific TFs binding to *cis*-regulatory elements control the precise expression of genes temporally and spatially. And these regulatory interactions can be further organized into regulatory modules with hierarchical structure, gene regulatory network (GRN) [30,31], showing high conservation level across

species [32]. In the past decade, several groups have established protocols to efficiently differentiate human embryonic stem cells into pancreatic and hepatic cell *in vitro* and characterize changes in gene expression during the differentiation process [12,33,34]. However, these studies primarily focused on the later stage of differentiation and the generation of functional cells, thus leaving the initiation of endoderm formation understudied. Other mammalian species such as mouse have proved surprisingly difficult in generating robust endoderm layer *in vitro* [35,36,37].

To better understand the mammalian gene regulatory network controlling endoderm formation, we first optimize and differentiate embryonic stem cells into definitive endoderm on monolayer in human, mouse and rat. We provide the first global view of transcriptome complexity and cis-regulatory dynamics during 5 day (human) or 7 day differentiation (mouse and rat) using RNA-seq and ATAC-seq. We then compared the conservation level on both gene expression and chromatin accessibility, which we use to build the GRNs involving 14 key genes implicated in definitive endoderm development and compare the resulting GRNs across species. We recover some interaction already validated in Xenopus or other species and discover multiple potentially novel regulatory linkages between endodermal key genes.

4.3 Results

4.3.1 Generation of endodermal markers-positive DE cells in three species

Methods for differentiating human and mouse embryonic stem cell (ESC) to definitive endoderm (DE) *in vitro* robustly have been developed during the past decade [12,33,35,36,37,38]. The protocols for both species rely on the activation of Nodal signaling to induce the expression of TFs and target genes in DE formation. Protocols add exogenous Activin A (analog of Nodal) to mimic high concentration of Nodal signaling in order to produce successful differentiated definitive endodermal cells with expression of key marker genes like FOXA2, SOX17 and CXCR4. However, it turns out that human ESC and mouse ESC cells cannot be differentiated using the same medium. Human ESCs can be differentiated into DE on monolayer with high efficiency using ActivinA and Wnt3A/Fgf4 in 4 or 5 days. But the same cocktail terminates mouse ESCs differentiation too early and protocols from different labs for mESC monolayer differentiation add different cellular factors to push the cells to DE, which

only achieve it with low efficiency and poor reproducibility. There was no published protocol for rat. To observe DE establishment, we first adjusted and optimized protocols for all three species. In order to produce relatively pure DE populations, we required all the cells to be differentiated on monolayer without support from mouse embryonic fibroblast (MEF) cells. We performed a 5-day human differentiation with the STEMdiff Definitive Endoderm kit (Figure 4.1A; Methods) while for mouse and rat, we optimized one established protocol [38] that has been reported the best level of purity and efficiency in DE population so far. Rodent ESCs were successfully differentiated into DE cells within 7 days with around 70%-80% confluence cover layer on dish (Figure 4.1C; Methods). We performed immunostaining at the start and end of differentiation and detected the production of pluripotent marker POU5F1 (OCT4) and endodermal marker SOX17 and FOXA2 respectively (Figure 4.1B & D). To examine differentiation status globally, we performed RNA sequencing (RNA-seq) on each day during differentiation for three species (Figure 4.2 & Table 4.1) and found other key endodermal markers were activated, including CXCR4, MIXL1, EOMES, GSC, GATA4 and GATA6 (Figure 4.3). We observed that these genes were induced from day 2 in human with the high expression of mesendodermal marker Brachyury (T gene). FOXA2, CXCR4 and SOX17 were activated at day 4 in mouse while at day 3 in rat. T expression reached its highest level at day 4 with continuous high expression until day 7. Interestingly, we found that FOXA1 were only expressed in mouse and rat but completely missing during human DE differentiation. As previously reported [39,40,41,42], stemness markers POU5F1 and NANOG were still expressed at the end of differentiation in human but missing in rodents (Figure 4.3). In summary our protocols can successfully differentiate ESCs in DE in three species. While stage-specific markers were conserved across species, gene were turned on or off at different times in the three species.

4.3.2 Dynamic of gene expression and chromatin accessibility during endodermal differentiation in three species

In order to globally detect conserved gene expression module in three species, we performed differentially expression analysis during time-course using maSigpro [43] and detected 7,472 (60% out of 12,392) 1:1:1 orthologous differentially expressed genes during DE differentiation across three species. Genes were classified into 15 clusters, which included conserved stem cells-, mesendodermal cells and definitive endodermal cells- specific clusters

(Figure 4.4). Gene expression cluster 1 (911 genes) is enriched with pluripotent markers like POU5F1, NANOG and SOX2 in three species and they were significantly associated with chromosome segregation ($P=1x10^{-45.934}$) and DNA replication ($P=1x10^{-29.974}$). Cluster 4,7,9,12 were highly enriched with endodermal markers that showed gradually activation towards the end of DE differentiation. Genes in cluster 4 are associated with growth factor stimulus ($P=1x10^{-12.544}$), heart development ($P=1x10^{-11.775}$) and mesenchyme development ($P=1x10^{-7.869}$) while genes in cluster 9 were significantly involved in gastrulation ($P=1x10^{-13.062}$) and digestive system development ($P=1x10^{-6.064}$). We also found clusters that show reverse expression patterns in three species. Genes in cluster 5 show increased expression in human but decreased in mouse and rat were enriched in catabolic process ($P=1x10^{-8.453}$) while genes in cluster 15 were down-regulated during hESC differentiation but up-regulated in rodents were highly involved in RNA splicing ($P=1x10^{-18.097}$) and posttranscriptional regulation of gene expression ($P=1x10^{-8.767}$). These results indicate that there are groups of genes with conserved expression patterns across species. However, species-specific patterns may reflect the difference in growth condition and following generation of transcripts during DE differentiation.

Chromatin accessibility plays a critical role in transcriptional regulation by mediating the binding of stage-specific TFs to cis-regulatory elements controlling target genes. To further understand transcriptional control during DE differentiation, we performed ATAC-seq along with RNA-seq during DE differentiation each day in each species (Figure 4.2, 4.5 & Table 4.2). We collected about 270,000-310,000 consolidated open chromatin regions for three species respectively (310,465 in human, 273,685 in mouse and 278,937 in rat) and 28,700 of them (~10%) were conserved across three species. We then identified open chromatin regions that showed differentially chromatin accessibility during DE differentiation in three species (Methods) and found 23,232 (80% out of 28,700) regions that formed 20 clusters with distinct accessibility patterns (Figure 4.6). We further selected clusters based on their associated genes that were shared in multiple species. Corresponding to the gene expression patterns (Figure 4.4), we found that stem cell specific clusters (cluster 1,12,13) enriched with pluripotent markers while cluster 10,14,17 were enriched with endodermal markers. These result in a loss of 3,439 (~12%) regions during while a gain of 3,861 (~13%) regions during DE differentiation that were shared during DE differentiation. Interestingly, we found one cluster (cluster 7) that includes

CTCF and YY1, both of which have been known to involve in gene regulation by mediating the formation of chromatin loops and with increased accessibility in all three species. Genes within this cluster were significantly enriched in chromatin modification ($P=1x10^{-11.812}$), indicating their role in regulating chromatin state. Similar as gene expression patterns, we found clusters that show different accessibility patterns in different species. For example, cluster 18 regions show increased signal in rodents but decreased in human during DE differentiation. Genes associated with these regions were enriched in mesenchymal cell differentiation ($P=1x10^{-16.129}$) and Wnt signaling pathway ($P=1x10^{-8.991}$). For clusters that were only shown increased signal (cluster 15) or decreased signal (cluster 5) in mouse, nearly all genes were involved in nervous system development ($P=1x10^{-9.034}$), skeletal development ($P=1x10^{-5.713}$) and epithelial cell differentiation ($P=1x10^{-10.008}$). To further assess the quality and function of these open chromatin regions, we performed *de novo* motif analysis. Based on this, we selected "representative" TFs (Methods) for each cluster. We observed that stem cell specific clusters were not only enriched with the motifs of POU5F1 but also CTCF. Besides, as expected, endodermal specific clusters are enriched with the motifs of forkhead transcription factors (Figure 4.7). These results indicate that chromatin accessibility is complementary with gene expression to identify conserved regulatory modules during DE differentiation in three species. Also, the patterns of accessibility can be represented by stage-specific TFs. But compared with gene expression, chromatin accessibility is more sensitive in detecting species-specific regions to show potential heterogeneity in a differentiated population.

We have observed conserved stage-specific modules at both gene expression and chromatin accessibility level that shared in three species. To summarize these observations globally, we performed principal component analysis (PCA) on both RNA-seq (Figure 4.8A) and ATAC-seq samples (Figure 4.8B; Methods). For both of them, PC1 distinguished differentiation time-course with 24.2% explained variance in RNA-seq and 40.6% explained variance in ATAC-seq while PC2 was used to separate differentiation trajectory by different species with around 14% variance explained in both of the PCAs. We also found that trajectories were merged at the end of differentiation for rodents and days were aligned in those two species but were quite separated from the one for human. Although human and rodents had different lengths of differentiation, we could align their differentiation on PCA for early-, mid- and late-stages.

4.3.3 Building gene regulatory networks by using coupling gene expression and chromatin accessibility

Based on the results above, we have found that both gene expression and chromatin accessibility can distinguish stages during DE differentiation in three species. They also show the power in detecting conserved regulation genes and regions. Specifically, we observed shared gene markers that were up-regulated with the gain of open chromatin regions during differentiation. We also found that pluripotent genes were down-regulated with the loss of chromatin accessibility. To further understand the regulation under this dynamics, we deep-sequenced our ATAC-seq samples and merged duplicates for each time point to reach to 110-270 million reads per day for footprinting calling (Methods) and detected around 100,000-200,000 footprints per day (Table 4.3). We then constructed gene regulatory networks (GRNs) using gene expression, open chromatin regions and footprinting results (Figure 4.9; Methods). In brief, the connection between regulator and target genes were only built if the regulator was expressed and if its binding motif could be detected in footprints associated with target genes. We focused on 14 key genes that are known to be involved in DE differentiation time-course for different stages, including pluripotent markers POU5F1, NANOG and SOX2; mesendodermal markers Brachyury (T gene); and endoderm markers GSC, EOMES, FOXA1/2, GATA4/6, SOX17, CXCR4, MIXL1 and CER1. First, we built the GRNs for ESC and DE stages for each species and we found increased number of connections upon differentiation (Figure 4.10-4.12). Interestingly, genes that were not expressed at ES or DE often still exhibited footprints by other key genes. For examples, GSC, FOXA2, GATA4 and MIXL1 were all bound by the three ESC TFs although their expression were hardly detectable in human ESCs (Figure 4.10A). In the rodents' endoderm GRNs, although the three ES TFs genes were all silenced, they were actively regulated by endodermal markers (Figure 4.11B & 4.12B). Further experiments are needed to decide the function of these connections. As expected based on gene expression level (Figure 4.3), we detected human specific-connections from POU5F1 and NANOG to other endoderm genes, including POU5F1 to T, EOMES, FOXA2, GATA4, GATA6, SOX17, MIXL1 and CER1 as well as NANOG to GSC, EOMES, GATA4, CXCR4, MIXL1 and CER1 (Figure 4.10B). Furthermore, we also found that there was no regulatory connection to FOXA1 in both ES and DE stages in human but found several footprints in rodents, which may explain the silence of

FOXA1 only in human (Figure 4.3). Some other connections from T and FOXA1 were rodent-specific because of their high expression at the endoderm stage. Next, we compared the conservation level of GRNs across all three species. We detected 40 connections in ES GRNs and only 11 connections (27.5%) were conserved in three species. These conserved connections include regulation between the ES TFs, i.e. NANOG to NANOG, POU5F1 to NANOG, SOX2 to SOX2, POU5F1 and NANOG. Others were involved in connections to FOXA2, CXCR4 and GATA4. By doing comparison between any two species, we found that around 36% (13/36) of connections were conserved between human and mouse; 54% (20/37) were conserved between rodents, mouse and rat; 53% (21/39) were conserved between human and rat (Figure 4.13). However, for the conservation level of DE GRNs, only 7 out of 117, ~6% of connections were conserved across three species. We found that 24.7% (23/93) of the connections were shared between human and mouse (Figure 4.14); 26.6% (21/79) were shared in rodents (Figure 4.15); 16.7% (18/108) were shared in human and rat (Figure 4.16). Given the higher conservation level between mouse and rat, we found that the shared connections were primarily involved in the regulation of T. But for human and mouse or human and rat comparison, the shared connections were mainly from EOMES, FOXA2, GATA4/6, GSC and SOX17. Above all, we construct the gene regulatory networks for ES and DE stages, and provide the first global view of regulatory connections between keys genes during DE differentiation in mammalian species.

4.4 Discussion

We have shown the comprehensive gene expression and chromatin accessibility profiles of DE differentiation time-course in human, mouse and rat and observed conserved regulatory modules of endodermal key genes in both of them. We also identified many species-specific changes and some of them even showed regulation in the reverse direction in human and rodents. We found putative TFs binding using open chromatin footprinting, which we used to build species-specific GRNs focusing on 14 core genes in endoderm development.

As the first attempt to compare endodermal differentiation in three species at the same time and especially the first time to differentiate mouse and rat DE with the same methods, we observed gene expression dynamics across species. First, conserved regulators FOXA2 and SOX17 were all up-regulated during DE differentiation, with relatively higher expression in

rodents than in human. As reported before [12,37,38], CXCR4 is another key marker of mammalian endoderm differentiation and as expected, it was upregulated in all three species with highest expression in human. We also found that MIXL1 and EOMES starts expression early during DE differentiation, which has also been demonstrated in other organisms [18,28,29]. We also found GATA4/6 and GSC are actively expressed in our time-course. However, we found some genes showed unexpected expression differences between species. The first is POU5F1 (Oct4), which is one of the key TFs defining the pluripotent cell fate in stem cells. It stayed highly expressed in later DE differentiation in human but not in rodents. Previous studies have found similar results of high POU5F1 at the end of human DE expression and they suggested this is possibly the effects of culture conditions [39]. We also detected human-specific expression of NANOG in late DE stage and studies already shown that both of them induce DE markers during endoderm differentiation in human [44,45]. However, endoderm single-cell study shown the co-expression of POU5F1 and endoderm markers in human DE [41] and based on our previous experimental results, human DE differentiation with no POU5F1 showing increasing level of heterogeneity than the one having POU5F1 detected. Further experiments are required to understand the exact function of POU5F1 in DE differentiation and it may be one of the most important human-specific DE regulatory changes. In addition, we also detected expression of FOXA1 in mouse and rat but not in human during DE development. As a core member TF of the FOX family, expression of FOXA1 and 2 have been detected in both mouse and Xenopus endoderm. Although a study showed that FOXA1 is the direct target of EOMES in hDE and knock-down EOMES would decreased the FOXA1 expression [44], our results indicates that it is not an essential TF during human DE development and the equivalent function may be taken over by FOXA2 and 3. Furthermore, we found T expressed highly in both rodents but silenced in human, and the expression level is especially high in rat DE. T is known as a marker of mesendoderm stage, we expect it is silenced in all three species. However, given the active interaction between T and other genes in our data, the function of T in rodents needs further investigation.

We analyzed the footprint-predicted rewiring GRNs at both ES and DE stages in three species. Without additional validation from ChIP-seq and knock-down perturbation, it is challenging to identify the real interactions between genes. Thus, we checked our interactions

compared to previously published results and found that our GRNs recover many validated regulatory interactions. For examples, at ES stage, we detected conserved interactions between POU5F1, NANOG and SOX2 and their well-known auto-regulatory loops [46,47,48] in all three species. As previously reported, POU5F1 and SOX2 were co-binding partners that they not only bound around NANOG [48] but also endoderm marker [44]. POU5F1-SOX2 complexes repressed the expression EOMES in human ES stage [44]. However, we failed to detect this regulation in human but in both mouse and rat. Furthermore, we also found conserved regulation from ES TFs to FOXA2, GATA4, and CXCR4 across species, which have not been reported as direct targets of pluripotency genes. SOX17 also shared conserved regulation by NANOG and SOX2 in human and rat but not in mouse. To make sure these are real, further deep sequencing may be necessary on our ATAC-seq samples to increase our power of footprints detection.

In the endoderm stage, we recovered the validated regulatory linkages between Mix, Gsc and T in mouse and rat. Mix activates Gsc and represses the expression of T in Xenopus [27,28,29,49]. Although no direct regulation was reported for mammalian species, knock-down experiments in mouse shown that Mixl1 knock-down causes Gsc down-regulation but T up-regulation [50]. We found that Mixl1 actually bound around Gsc and that Gsc further regulates T in mouse and rat but not in human. In addition, we also detected the known connection from EOMES to GSC, CER1, MIXL1 in some of the three species but they are not always detected in all three. And we found a novel regulatory link from EOMES to GATA6 in rodents. SOX17 is another key TF during endoderm development and we found that the regulation from SOX17 to FOXA2 in rodents have been reported in human, mouse and Xenopus studies [51,52,53,54]. SOX17 footprints were also found in CER1 and GSC in human. Furthermore, both GATA4 and GATA6 showed many footprints in DE regulation in human and mouse. We found GATA6 regulation of SOX17 in our network, which has been reported in human, mouse, and Xenopus [55,56,57]. More interestingly, corresponding to the gene expression level, we found that there is no regulation from or to FOXA1 in human at both ES and DE stages whereas we detected multiple regulatory linkages to FOXA1 in rat, which expressed highest across all species. Similarly, we detected a significantly higher number of regulations to T in rodents than those in human. These results indicate that the number of regulatory linkages in GRNs may reflect gene

expression levels (at least when it involves activators), and TFs co-binding may be a common way in regulating and maintain gene expression.

Given the low conservation level (~6%) of regulatory linkages across three species, deeper sequencing of our ATAC-seq is necessary to confirm that we did not miss any footprints. Further validation experiments will be necessary for new regulatory linkages and confirming the "active" or "repressive" effects on target genes. For example, knocking-down EOMES may be the first step for validating the GRNs because it not only regulates GSC and MIXL1, which known as early onset markers of DE differentiation, but also regulate the expression of FOXA2, CXCR4, CER1 at a later stage in both human and mouse. Also, the POU5F1 knock-down may also be interesting to examine their roles in both repressing DE markers at ES stage and repressing (or activating) DE genes at endoderm stage.

4.5 Figures

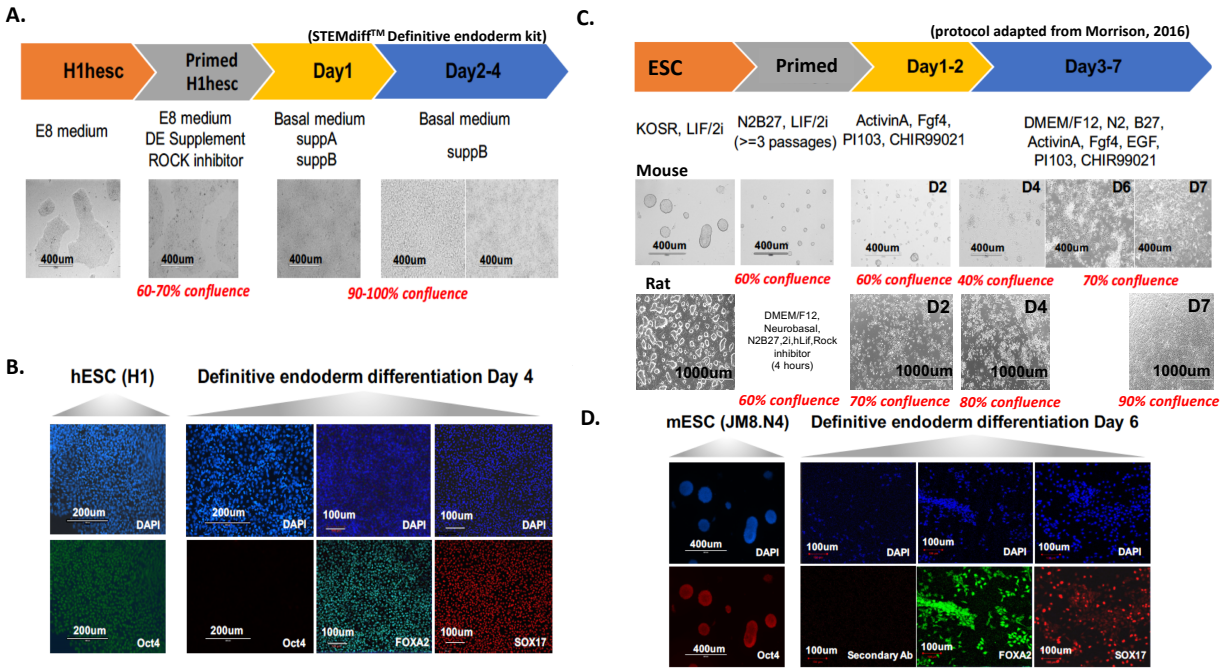


Figure 4.1. (A) Five day time-course of human definitive endoderm differentiation by using STEMdiff Definitive endoderm kit. **(B)** Immunostaining of stage-specific markers at the start and end points of human differentiation. Oct4 (Pou5f1) is a marker of stem cell stage while FOXA2 and SOX17 are markers of endoderm stage. **(C)** Seven day time-course of mouse and rat definitive endoderm differentiation by using adapted protocol [38]. **(D)** Immunostaining of stage-specific markers at the start and end points of mouse differentiation. Oct4 (Pou5f1) is a marker of stem cell stage while FOXA2 and SOX17 are markers of endoderm stage.

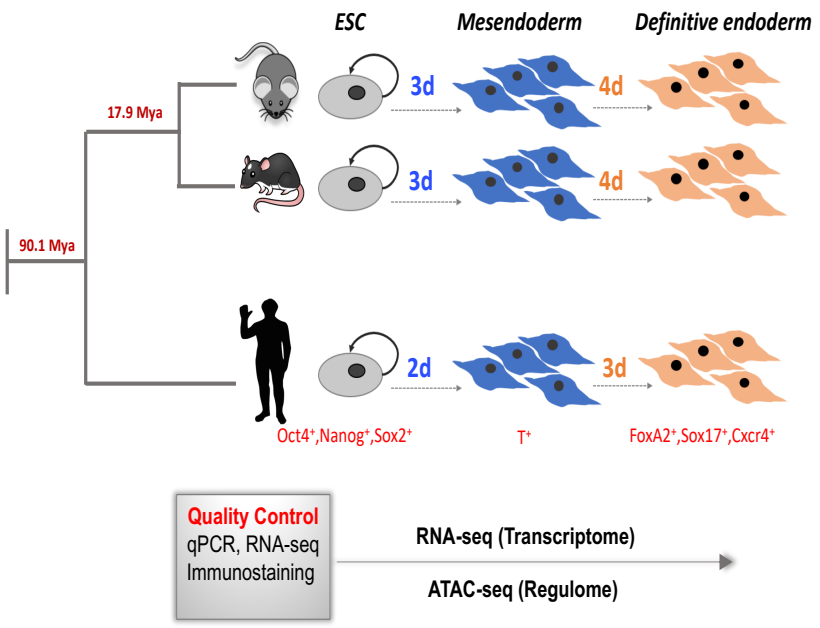


Figure 4.2. Phylogenetic tree of human, mouse and rat (adapted from TIMETREE; Mya, Millions of Years Ago). Experimental design of transcriptome and chromatin accessibility profiling from ESCs to definitive endoderm differentiation. Quality control for cell state is performed at the start and end points of differentiation. RNA-seq and ATAC-seq are performed during differentiation time-course.

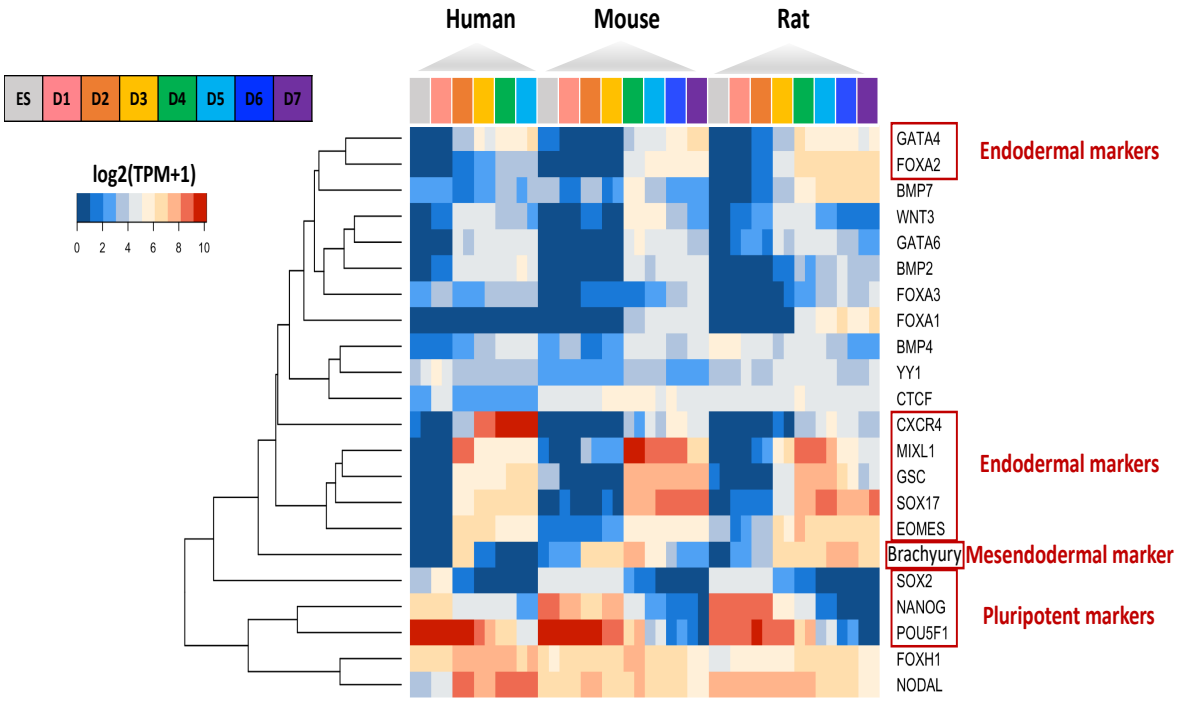


Figure 4.3. Expression heatmap of selected marker genes during DE differentiation time-course in three species (blue, low expression; red, high expression). These genes are stage-specific markers, including (1) POU5F1 (OCT4), NANOG and SOX2 for stem cell stage; (2) Brachyury (T) for mesendodermal stage; (3) CXCR4, MIXL1, GSC, SOX17, EOMES, GATA4 and FOXA2 for endodermal stage.

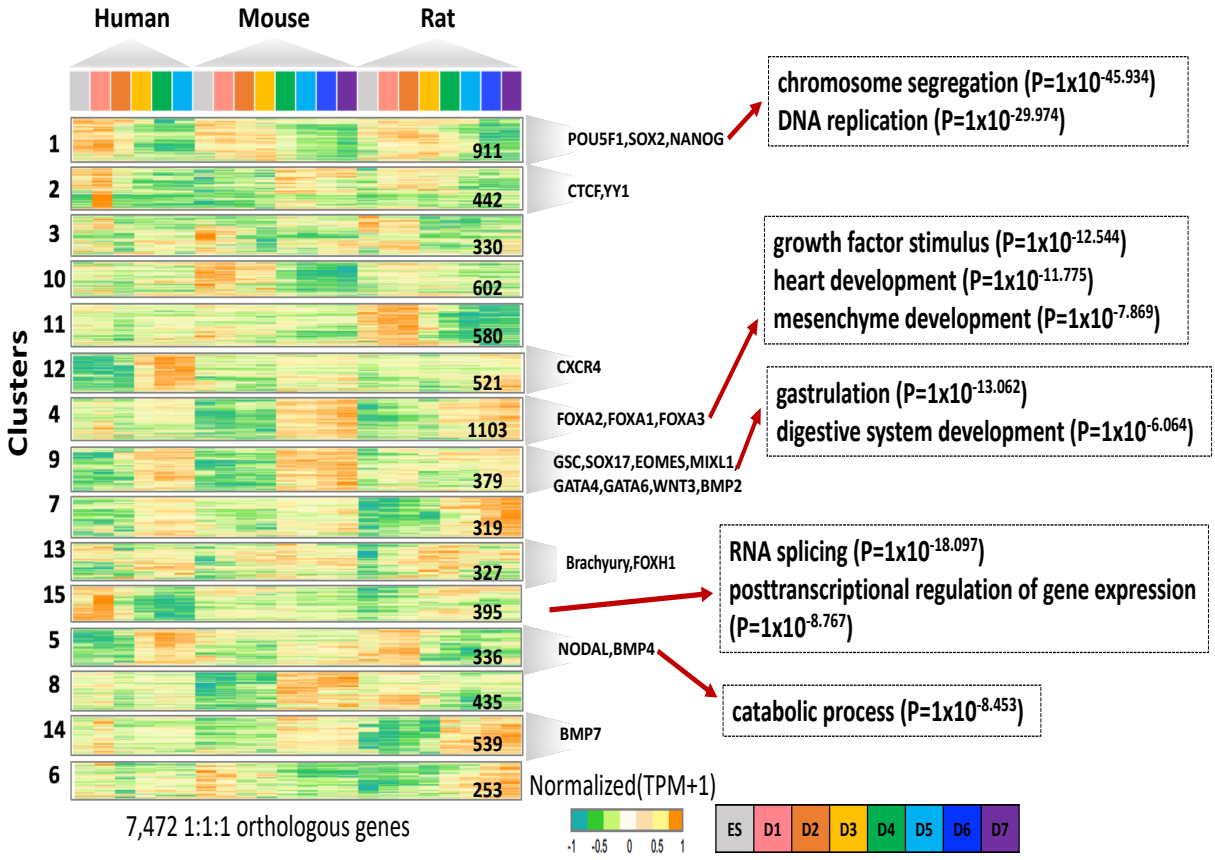


Figure 4.4. 7,472 differentially expressed gene during DE differentiation are identified across three species and genes are further clustered into 15 clusters by maSigPro (alpha=0.8, FDR<0.05; green, low expression; orange, high expression). Expression of these 7,472 differentially expressed genes are normalized between -1 and 1 by using z-score normalization and they are presented in the heatmap by cluster. Gene markers and gene ontology terms are labeled for stem cells-, mesendodermal cells and definitive endodermal cells-specific clusters.

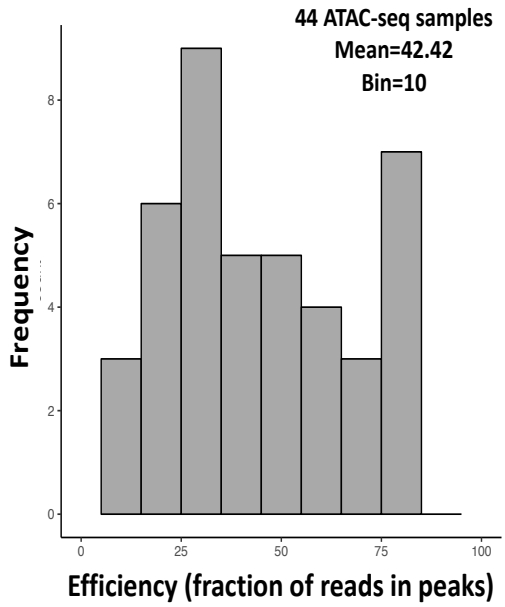


Figure 4.5. Distribution of ATAC-seq sample efficiency. Efficiency is calculated by the fraction of reads in peaks. Mean = 42.42%.

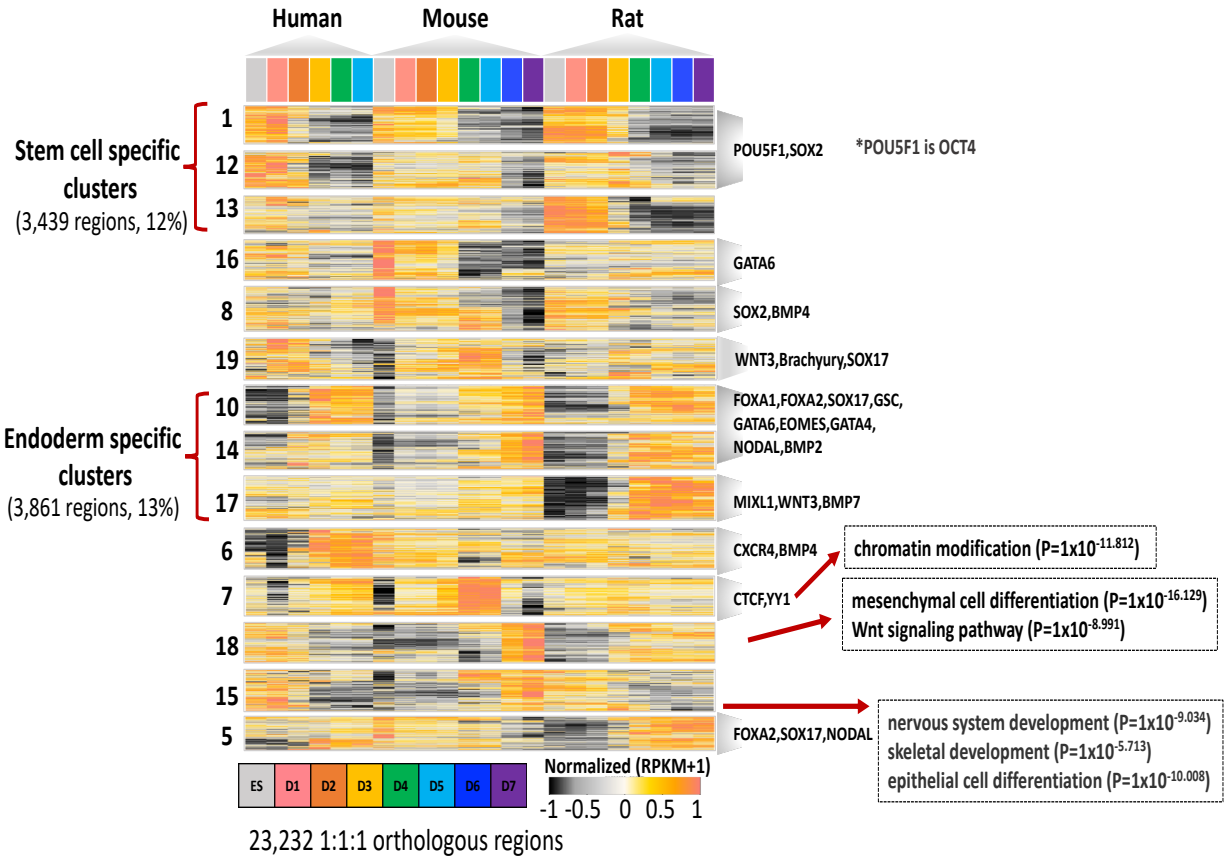


Figure 4.6. 23,232 differential open chromatin regions during DE differentiation are identified across three species and clustered into 20 clusters by maSigPro (alpha=0.7,FDR<0.05;black,low accessibility; orange, high accessibility). 14 clusters are selected to plot. Coverages over chromatin accessibility regions are normalized by sequence depth and length of regions to RPKM and they are further normalized between -1 and 1 by using z-score normalization for visualization. Stage-specific makers are labeled for stem cells-, mesendodermal cells and definitive endodermal cells-specific clusters based on the chromatin regions they associated with.

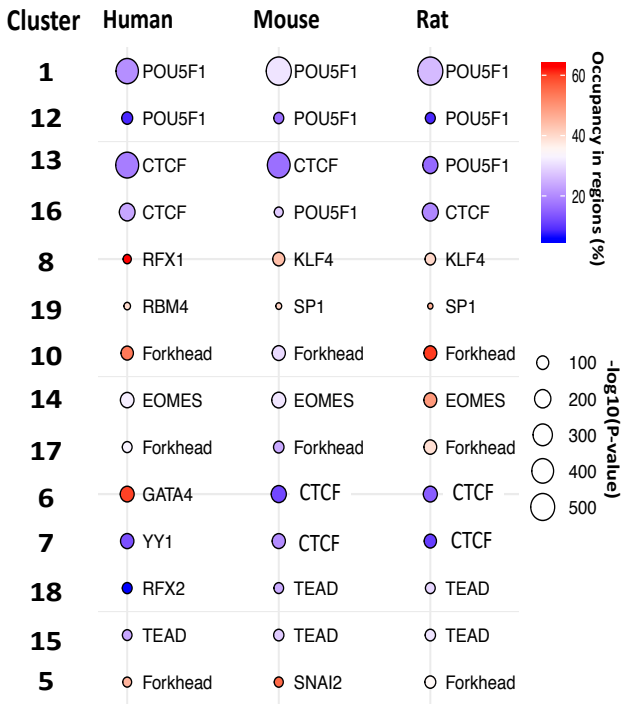


Figure 4.7. *De novo* motif calling in differentially open chromatin regions in three species. Motifs listed for each cluster in Figure 4.6 have the most significant p-value, high enrichment (%) and high motif similarity score. *de novo* motif calling is done by Homer.

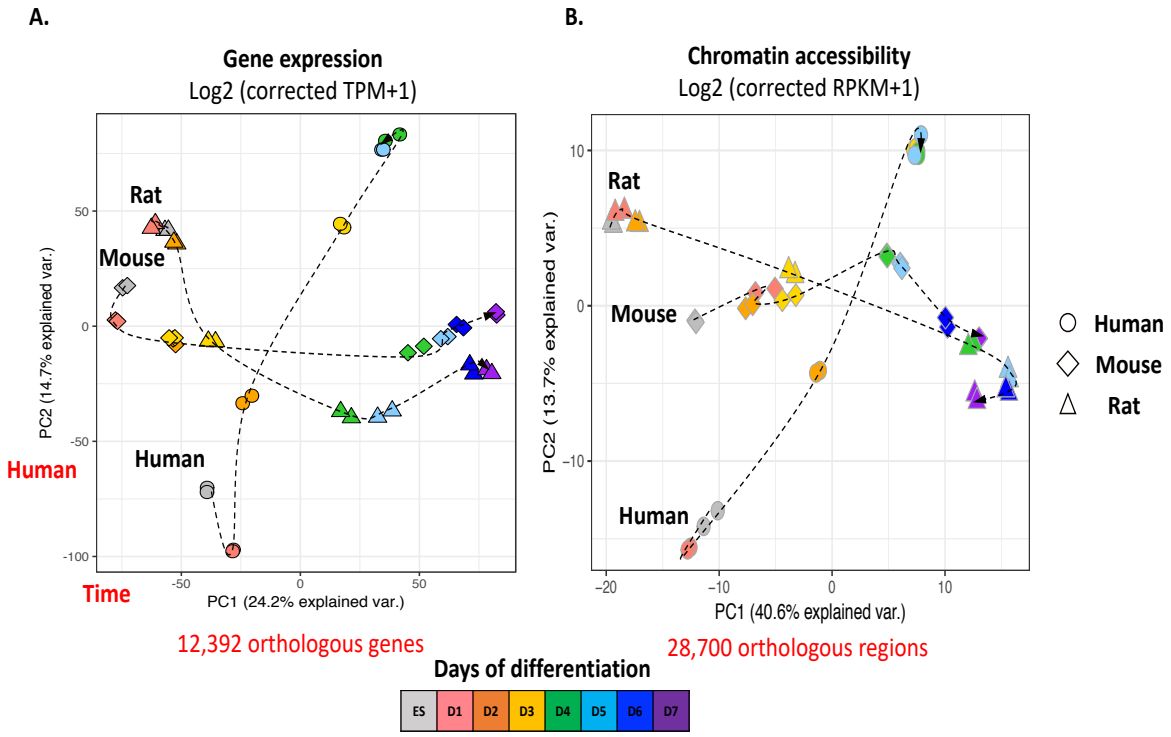


Figure 4.8. Principal component analysis (PCA) during DE differentiation for **(A)** gene expression and **(B)** chromatin accessibility in three species. Analyses are done on 1:1:1 orthologous genes and chromatin accessibility regions. Time points were connected serially to illustrate cell-specific trajectories. Species are labeled by distinct shape and cell types are labeled with distinct colored points.

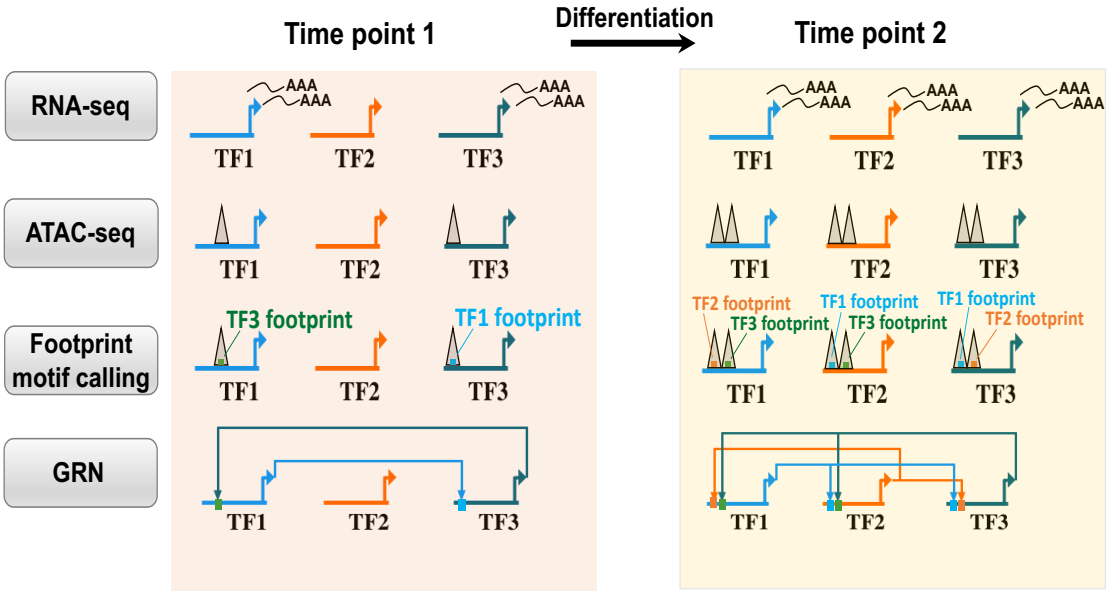


Figure 4.9. Strategy for building gene regulatory networks (GRNs). For each time point during differentiation, expression of key TFs are measured by RNA-seq and open chromatin regions around key TFs are detected by ATAC-seq. Then footprint and motif calling are performed in open chromatin regions within 20KB of target TFs gene body to decide which TFs potentially binding to regulate target TFs. If both target TFs and potentially binding TFs have detected gene expression level (>2TPM), the connections between them are constructed. GRNs are built with many layers of these connections.

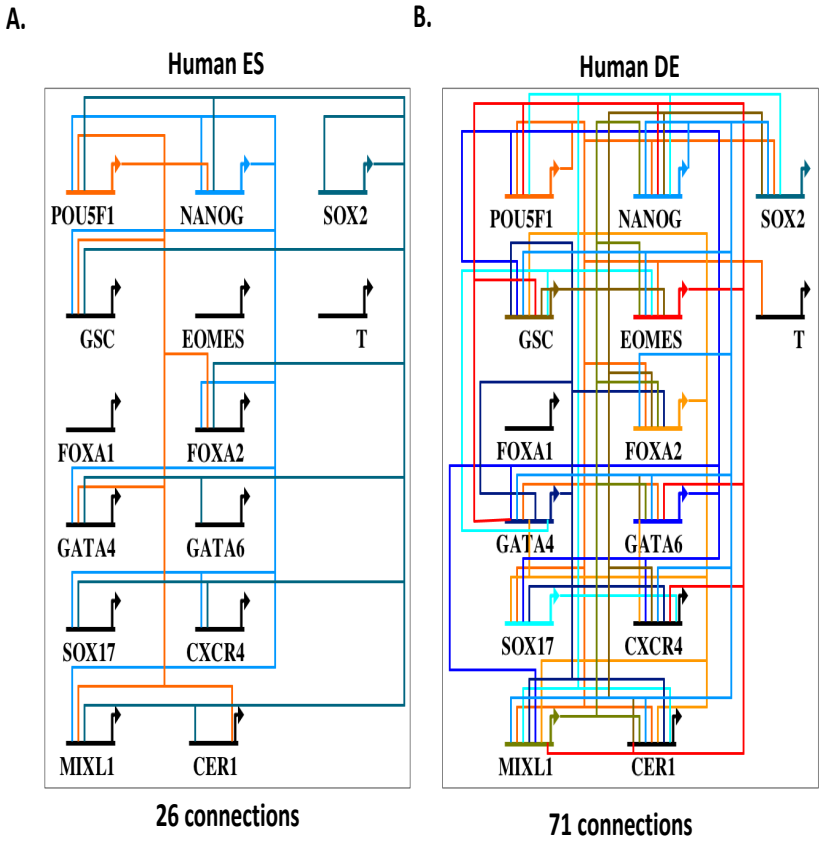


Figure 4.10. Gene regulatory network (GRN) of 14 key genes at ES and definitive endoderm stages, including pluripotent markers POU5F1, NANOG and SOX2; mesendodermal markers Brachyury (T gene); and endodermal markers GSC, EOMES, FOXA1/2, GATA4/6, SOX17, CXCR4, MIXL1 and CER1. **(A)** GRN at human ES stage. **(B)** GRN at human definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory interactions between ES and endoderm stages.

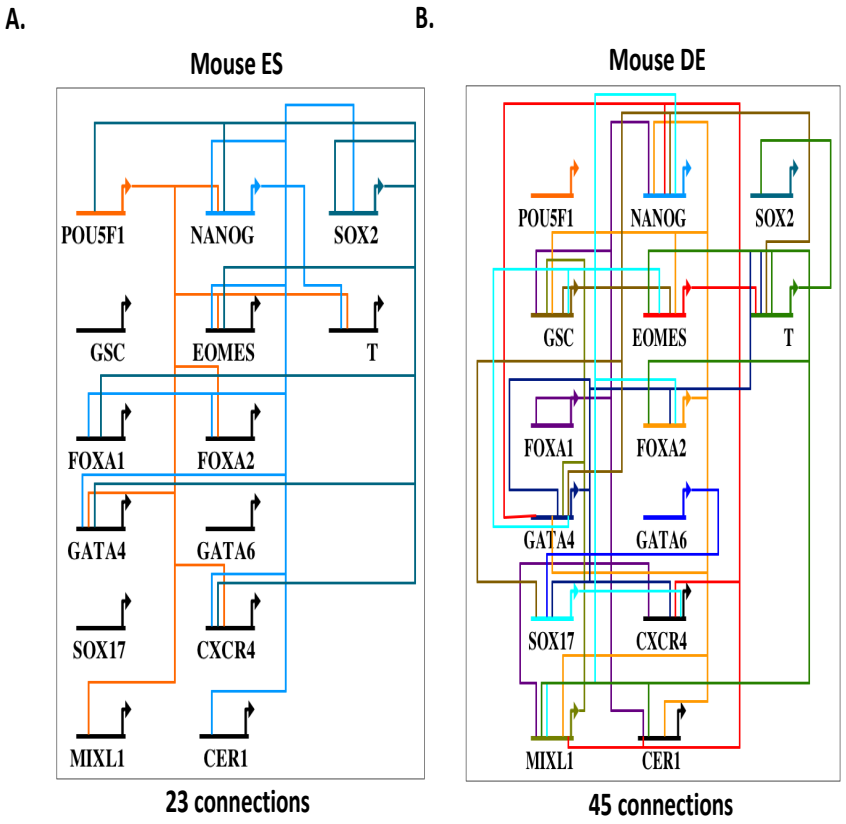


Figure 4.11. Gene regulatory network (GRN) of 14 key genes at ES and definitive endoderm stages, including pluripotent markers POU5F1, NANOG and SOX2; mesendodermal markers Brachyury (T gene); and endodermal markers GSC, EOMES, FOXA1/2, GATA4/6, SOX17, CXCR4, MIXL1 and CER1. **(A)** GRN at mouse ES stage. **(B)** GRN at mouse definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory interactions between ES and endoderm stages.

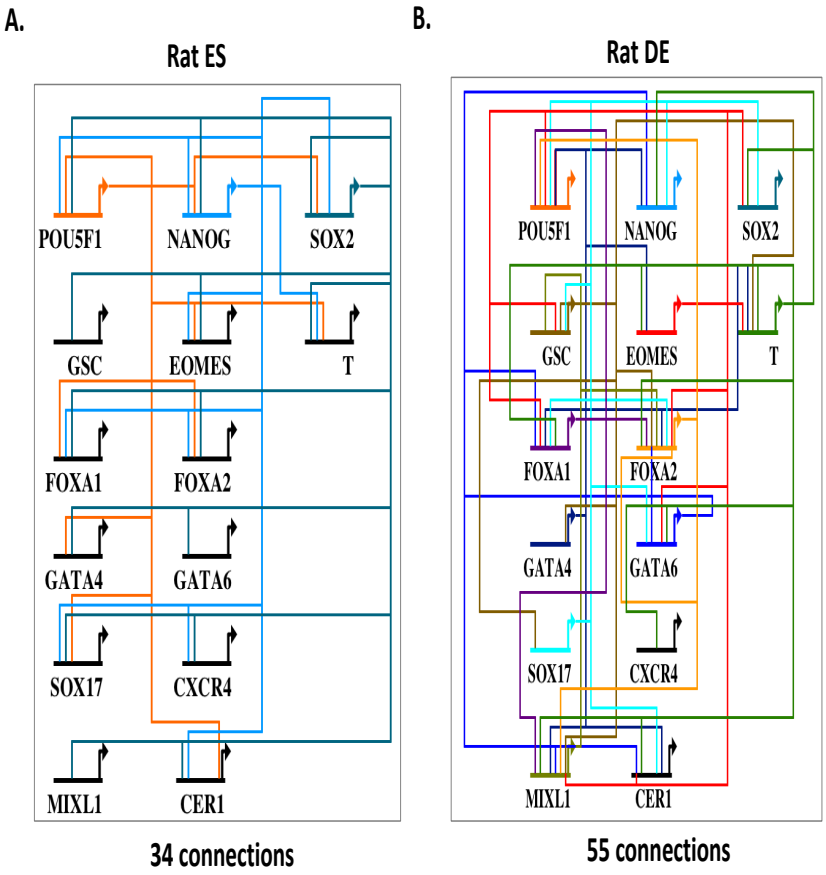


Figure 4.12. Gene regulatory network (GRN) of 14 key genes at ES and definitive endoderm stages, including pluripotent markers POU5F1, NANOG and SOX2; mesendodermal markers Brachyury (T gene); and endodermal markers GSC, EOMES, FOXA1/2, GATA4/6, SOX17, CXCR4, MIXL1 and CER1. **(A)** GRN at rat ES stage. **(B)** GRN at rat definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory interactions between ES and endoderm stages.

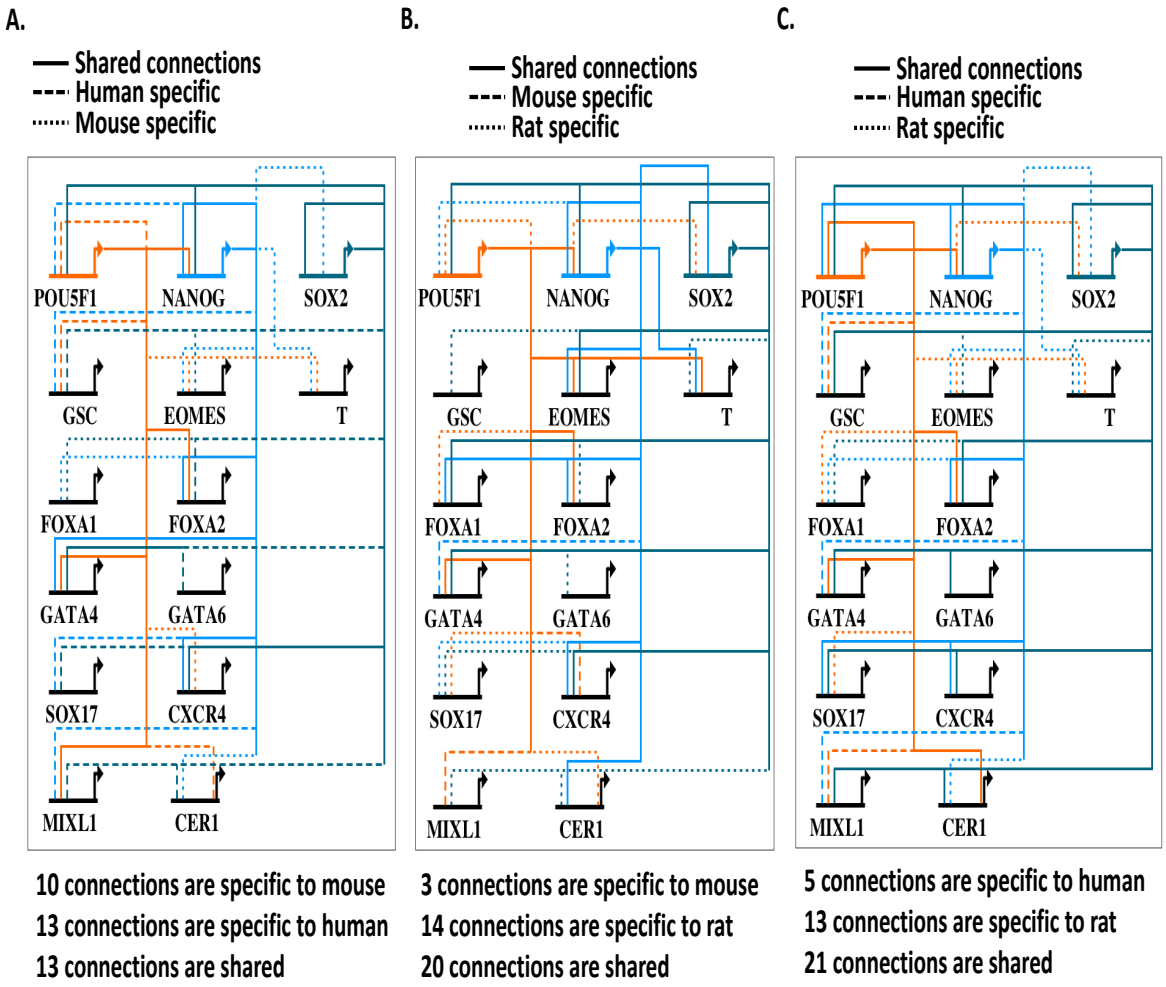


Figure 4.13. Gene regulatory network comparison between **(A)** human and mouse ES cells; **(B)** mouse and rat ES cells; **(C)** human and rat ES cells. Each gene is assigned a unique color to track changes in regulatory linkages between two genes. Solid line shows conserved regulatory linkages while dash line labels species-specific regulation between two species.

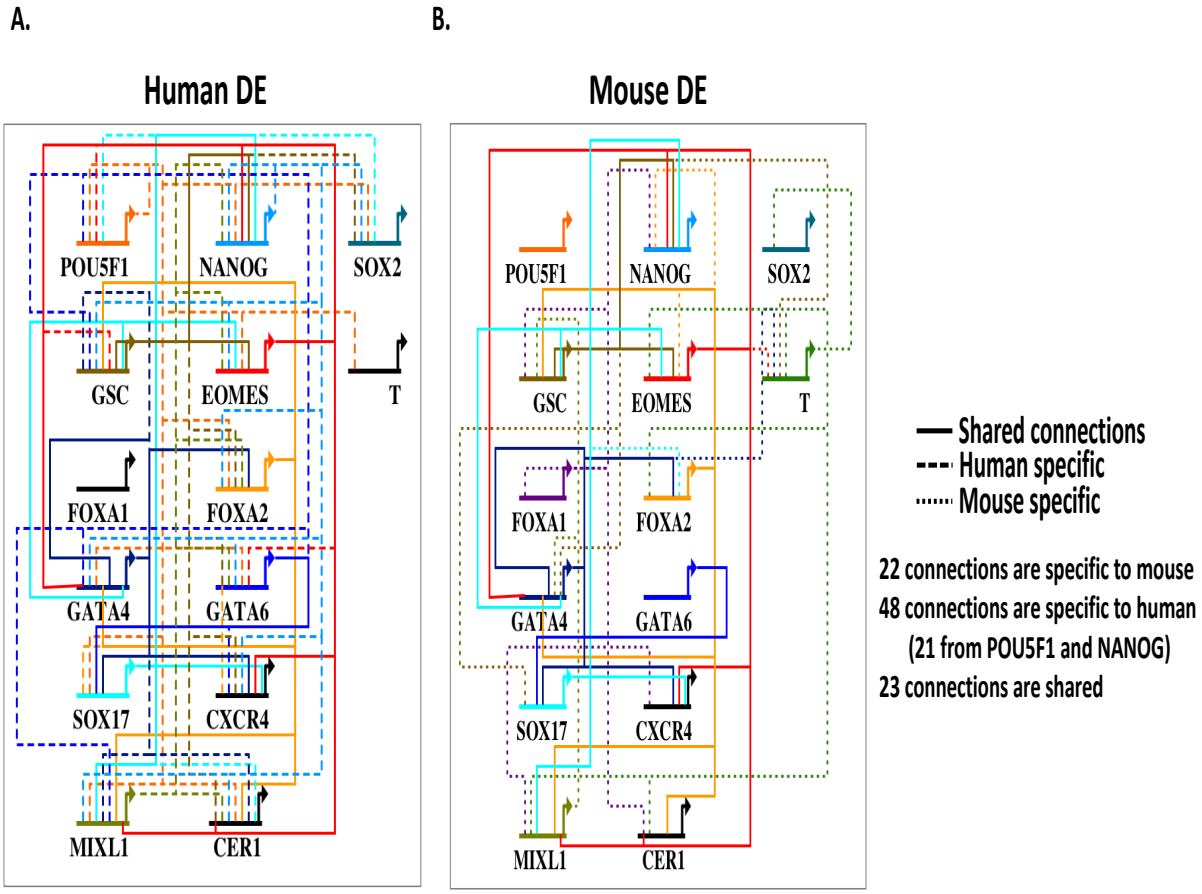


Figure 4.14. Comparison of gene regulatory networks between **(A)** human and **(B)** mouse at definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory linkages between two genes. Solid line shows conserved regulatory linkages while dash line labels species-specific regulation between two species.

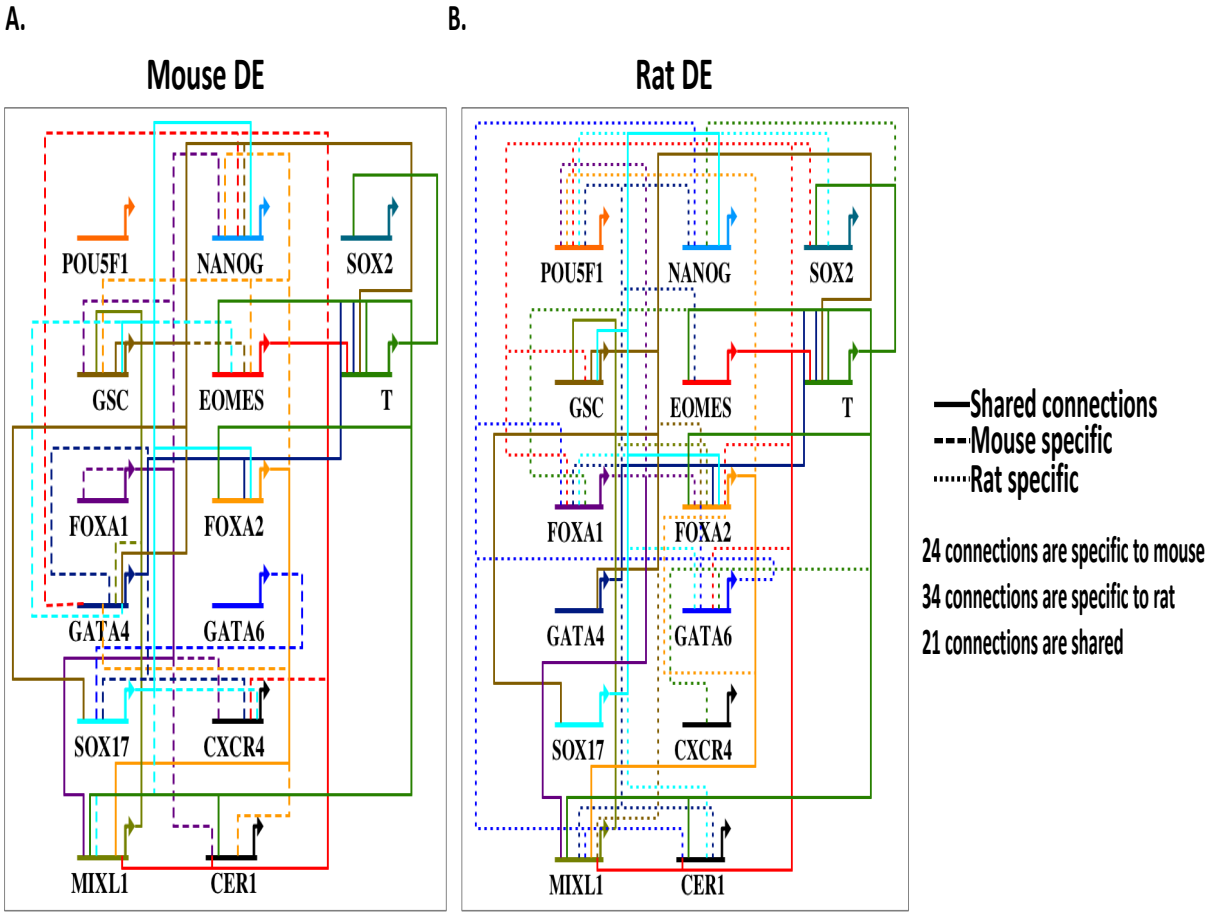


Figure 4.15. Comparison of gene regulatory networks between **(A)** mouse and **(B)** rat at definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory linkages between two genes. Solid line shows conserved regulatory linkages while dash line labels species-specific regulation between two species.

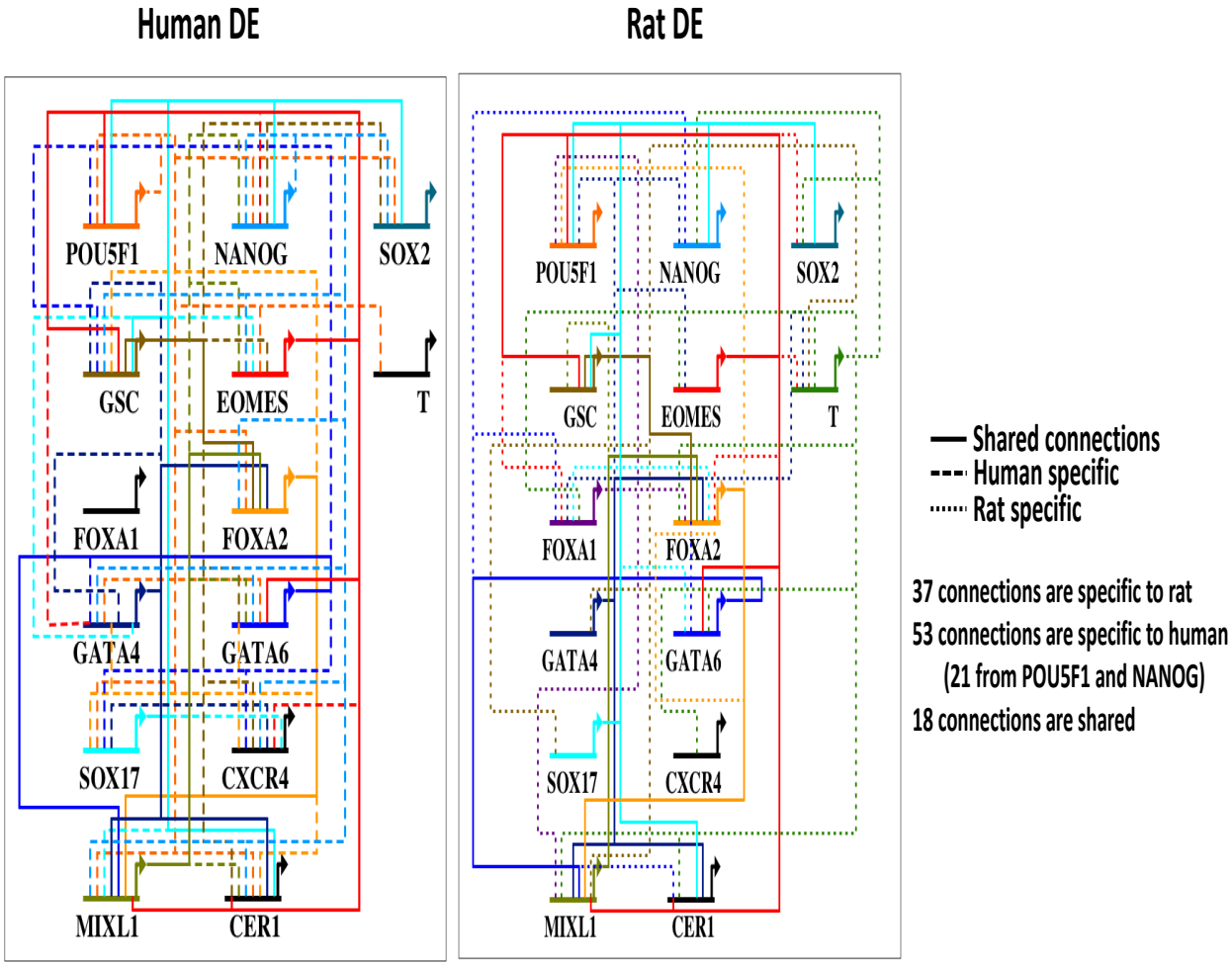


Figure 4.16. Comparison of gene regulatory networks between **(A)** human and **(B)** rat at definitive endoderm stage. Each gene is assigned a unique color to track changes in regulatory linkages between two genes. Solid line shows conserved regulatory linkages while dash line labels species-specific regulation between two species.

Species	Timepoint	Uniquely mapped reads of Rep1	Uniquely map efficiency of Rep1	Detected genes (>=1TPM) of Rep1	Uniquely mapped reads of Rep2	Uniquely map efficiency of Rep2	Detected genes (>=1TPM) of Rep2
Human	ES	11999017	76.28%	15155	9447848	76.43%	15209
	D1	7784268	82.97%	14742	5930764	82.73%	14947
	D2	8580086	80.91%	14195	9052857	80.62%	14121
	D3	10170322	80.72%	14783	9739162	80.82%	14961
	D4	8434648	81.71%	15377	9705908	80.98%	15393
	D5	8447074	82.02%	14725	8428818	82.77%	14940
Mouse	ES	10731526	73.76%	12444	14423022	72.77%	12473
	D1	10485397	72.65%	12470	10361873	72.83%	12526
	D2	9673945	71.98%	12672	10514836	71.77%	12702
	D3	5808820	69.76%	13621	5549428	67.98%	13442
	D4	10292464	76.08%	15862	10761733	76.42%	14626
	D5	10861103	74.40%	14527	8888845	75.03%	14684
	D6	10557251	75.05%	12952	10617300	76.33%	13145
	D7	10221795	76.17%	13105	10067411	75.87%	13043
Rat	ES	6177882	67.03%	11335	6152213	66.40%	11300
	D1	7152358	71.09%	13132	7665304	69.51%	11820
	D2	7738495	70.40%	12177	9015980	73.24%	12033
	D3	8742418	69.74%	11989	7173255	70.59%	12083
	D4	4206670	72.47%	11879	7087907	71.69%	11876
	D5	6440998	72.35%	11245	7785740	70.90%	11306
	D6	6198496	70.62%	11714	6693788	70.14%	11754
	D7	7045235	72.88%	11634	7996475	73.55%	11700

Table 4.1. RNA-seq samples are prepared into duplicates for each time point and the quality is assessed for each replicate by (1) the number of uniquely mapped reads; (2) uniquely mapped efficiency; (3) the number of detected genes.

Species	Timepoint	Uniquely mapped reads of Rep1	Uniquely map efficiency of Rep1	Mito% of Rep1	Number of peaks of Rep1	Uniquely mapped reads of Rep2	Uniquely map efficiency of Rep2	Mito% of Rep2	Number of peaks of Rep2
Human	ES	68011715	86.62%	3.65%	178399	72053670	85.97%	2.84%	126101
	D1	69243655	87.47%	7.90%	212793	82062531	87.62%	7.04%	221181
	D2	74820462	83.66%	3.13%	143624	86114404	83.65%	3.56%	133439
	D3	75745733	86.46%	8.02%	205728	94410199	86.43%	6.79%	217926
	D4	78460016	87.32%	6.10%	222771	92768265	87.28%	6.12%	228128
	D5	129912607	86.69%	5.26%	201400	98490122	88.54%	4.42%	228269
Mouse	ES	91645724	83.81%	3.70%	246440	107024951	83.73%	3.87%	254878
	D1	48785701	78.71%	6.04%	57664	67273043	77.73%	6.72%	65074
	D2	62064316	78.66%	7.60%	97821	71059299	78.88%	6.91%	94302
	D3	123415639	76.88%	9.67%	74010	100163677	77.53%	9.08%	61915
	D4	132608547	77.92%	4.33%	30684	137639833	77.20%	4.54%	30112
	D5	153163206	75.53%	4.07%	45901	92965158	75.15%	3.90%	38350
	D6	62717618	77.35%	5.68%	93014	57924080	77.07%	4.92%	60625
	D7	79898201	80.81%	3.49%	107649	97395057	80.47%	4.47%	129307
Rat	ES	62662252	80.95%	3.48%	211762	80655775	80.10%	3.64%	208060
	D1	94271542	75.51%	4.95%	128814	71196512	73.67%	7.87%	152582
	D2	78434520	77.06%	4.52%	165805	75991122	77.21%	4.99%	176434
	D3	94576355	78.09%	3.23%	105672	87644144	77.11%	4.16%	105934
	D4	118103133	79.57%	2.44%	131125	93749683	79.24%	4.73%	155244
	D5	107073870	81.44%	2.50%	145183	57966923	82.78%	2.69%	152535
	D6	83891664	82.82%	2.05%	155954	80282780	81.55%	2.01%	149409
	D7	79973852	80.99%	1.45%	119583	79266904	82.36%	1.91%	148405

Table 4.2. ATAC-seq samples are prepared into duplicates for each time point and the quality is assessed for each replicate by (1) the number of uniquely mapped reads; (2) uniquely mapped rate; (3) fraction of reads mapped to mitochondria DNA; (4) the number of peaks

Species	Timepoint	Input reads	Number of consolidated peaks	Number of footprints
Human	ES	140065385	119983	145343
	D1	151306186	191458	214514
	D2	160934866	116394	142523
	D3	170155932	190168	219494
	D4	171228281	200788	222488
	D5	228402729	187690	206684
Mouse	ES	198670675	220718	238327
	D1	116058744	44409	61732
	D2	133123615	76646	98498
	D3	223579316	52479	67493
	D4	270248380	17542	24474
	D5	246128364	23426	30891
	D6	120641698	54031	69431
	D7	177293258	95232	107990
Rat	ES	143318027	185514	220855
	D1	165468054	116020	151183
	D2	154425642	144598	179316
	D3	182220499	84234	111384
	D4	211852816	112884	137522
	D5	165040793	125068	150058
	D6	164174444	130594	156530
	D7	159240756	107044	127362

Table 4.3. Footprint calling during DE differentiation for three species. For footprint calling, uniquely mapped reads of duplicates are merged as input reads for each time point. Footprint calling are done by using Wellington algorithm with 1% FDR.

4.6 Methods

4.6.1 Embryonic stem cell maintenance

Human embryonic stem cell (H1, XY) was bought from WiCell and were maintained on matrigel in STEMCELL Technologies™ TeSR-E8 medium. Cell were routinely passaged every two days with EDTA. Mouse embryonic stem cell (JM8.N2, XY) was bought from KOMP Repository, UC-Davis. They were maintained on 0.1% Gelatin and were first cultured in JM8.N4 ES cell medium, supplemented with 1X KO DMEM, 15% FBS, 2mM Glutamine, 1mM NEAA, 1000U/ml LIF and 0.1 mM 2ß-ME, for 1-2 days until the cells reaching to over 70% confluency. Then mES cells were passaged with Accutase into KOSR+2i medium, supplemented with KO DMEM, 15% KO Serum replacement (KOSR), 4mM Glutamax, 1mM NEAA, 1mM Sodium pyruvate, 0.1mM 2ß-ME, 100U-ug/ml Pen/Strep, 200U/ml LIF, 5ug/ml Insulin, 1uM MEK inhibitor PD0325901 and 3uM GSK3 inhibitor CHIR99021. Cells were passaged every 3-5 days depending on the size of colony. Rat embryonic stem cells (DAc8, XY) was bought from Rat Resource and Research Center (RRRC), University of Missouri. Cells were first maintained on mouse embryonic fibroblast (MEF) feeders. MEF medium was made with GMEM, 10% FBS, 1% GlutaMAX and 1% Pen/Strep. Rat ES medium was supplemented with 100ml DMEM/F12, 1ml N2, 1.5ml HEPES (1M), 100ml Neurobasal medium, 2ml B27, 1ml GlutaMax-I, 100ul Inuslin, 66.7 CHIR99021 (3mM), 40ul PD0325901 (1mM), 2ml 100x2ß-ME, 200ul Y-27632 (5mM), 200ul hLIF (10ug/ml). MEF cells were plated on 0.1% Gelatin at least one day before plating rat ES cells. Rat ES cells were passaged every 4-6 days with Accutase. depending on the size of colonies.

4.6.2 Definitive endoderm differentiation on monolayer *in vitro*

Human ES cells were differentiated with E8-optimized Definitive Endoderm Kit from STEMCELL Technology. Differentiation was carefully conducted following the protocol. Mouse and rat ES cells were differentiated following the protocol optimizing for the stem cell lines. For mouse, cells were first passaged from KOSR+2i medium to NDiff N2B27 base medium supplemented with PD0325901 (1uM), CHIR99021 (3uM) and mLIF (100U/ml) (in 50ml, 49.939 ml of NDiff, 1ul mLIF, 10ul PD03, 50ul CHIR) on gelatin coated plate. After 2-3 days, cells were differentiated into NDiff medium supplemented with Activin A, Fgf4, Heparin, PI3 kinase inhibitor PI103, CHIR99021 (in 50ml, 49.88ml of NDiff medium, 10ul of 100ug/ml

ActivinA, 5ul of 100ug/ml Fgf4, 50ul of 1mg/ml of Heparin, 5ul of 1mM PI103, 50ul of 3uM CHIR). After 2 days of differentiation, medium was changed to DMEM/F12 with N2, B27-VA, L-glutamine, 2ß-ME, BSA, ActivinA, Fgf4, Heparin, Egf, PI103 and CHIR99021 (in 50ml, 48.8295ml of DMEM/F12,250ul of N2, 500ul of B27-VA, 250ul L-glutamin, 50ul 1000x 2ß-ME, 0.025g BSA, 10ul 100ug/ml ActivinA, 5ul 100ug/ml Fgf4, 50ul 1mg/ml Heparin, 0.5ul 100ug/ml Egf, 5ul 1mM PI103 and 50ul of 3uM CHIR). Cells were continuously differentiated for 5 days. For rat, cells were first transferred from MEFs to gelatin coated plate and they were seeded in rat ES medium for 4-6 hours. Then the medium was changed to the NDiff medium to differentiate following the same protocol using in mouse.

4.6.3 RNA-seq library construction

Total RNA was extracted on every day of differentiation by using RNeasy kit (QIAGEN). RNA were converted to cDNA using the SmartSeq 2 protocol [58]. Libraries were constructed by using the Nextera DNA Sample Preparation Kit (Illumina). Libraries were quality-controlled prior to sequencing based on Agilent 2100 Bioanalyzer profiles and normalized using the KAPA Library Quantification Kit (Illumina). The libraries were sequenced using paired-end 43 mode on Illumina NextSeq500 platform with around 10-15 million reads per sample.

4.6.4 ATAC-seq library construction

ATAC-seq samples were collected from the same pool cell population used for RNA-seq on every day of differentiation by using omni-ATAC protocol [59]. Around 50,000 cells were used per replicate and libraries were base-selected between 150-500bp for final construction. Libraries were normalized using the KAPA Library Quantification Kit (Illumina). The libraries were sequenced using paired-end 43 bp mode on Illumina NextSeq500 platform with around 60-100 million reads per sample.

4.6.5 Gene expression analysis

Raw reads were mapped to hg38 (human), mm10 (mouse) and rn6 (rat) using STAR (version 2.5.1b) [60] using defaults except with a maximum of 10 mismatches per pair, a ratio of mismatches to read length of 0.07, and a maximum of 10 multiple alignments. Quantitation was performed using RSEM (version 1.2.31) [61] with the defaults, and results were output in

transcripts per million (TPM). Batch effects were corrected by using limma removebatcheffect [62]. Clustering of differentially expressed genes across the time-course was done by using maSigPro [63] with alpha of 0.05 for multiple hypothesis testing and a false discovery rate of 0.05%. Gene ontology analysis was done by using Metascape [64].

4.6.6 ATAC-seq data processing and analysis

Raw reads were mapped to hg38 (human), mm10 (mouse) and rn6 (rat) using bowtie [65]. Reads mapped to ChrM were discarded and PCR duplicates were removed by using Picard [66]. HOMER/4.7 [67] was used to call open chromatin regions. It was first used for calling 200bp narrow peaks and then 500bp broad peaks. Then narrow and broad peaks were merged into a single peak list and they were further filtered by overlapping with ENCOE "blacklist" regions. Peaks that shown in both replicates were considered as reproducible open chromatin regions. Reads coverage was calculated using bedtools [68] for each region and they were further normalized by the size of library and the size of peaks. Normalized coverages were further batch corrected for the technical batch effects by using limma removebatcheffect [62]. The differential open chromatin regions through differentiation time-course were identified by using maSigPro [63] with alpha=0.05 and FDR<0.05.

4.6.7 *de novo* motif enrichment analysis

De novo motif calling was performed using HOMER/4.7 [67]. Open chromatin regions for each maSigPro cluster were converted to fasta using masked genome for each species. The size of motif was set as "-len 6,8,10,12,15,120" with at most 3 mismatches.

4.6.8 Footprints calling and GRN construction

ATAC-seq reads from duplicates were merged for each time-point to achieve around 120-200 million reads for footprints calling. Footprints were called by using Wellington algorithm from pyDNase [69] with parametes "-A -fp 6,31,1 -sh 7,36,4 -fdrlimit -2", 1% FDR. Then motifs were scanned within footprints regions by using FIMO [70] to identify transcription factor binding motifs from HOCOMOCO database [71]. Next, both gene expression and motif scanning results were used to build gene regulatory networks. The connection between two genes was constructed if (1) the binding motif of one gene (regulator) was detected within 20kb of TSS of the other

gene (target). (2) The expression level of regulator is higher than 2TPM in specific time-point. GRNs were constructed at stem cell and definitive endoderm stages. Networks were visualized by using Biotapestry [72].

4.6.9 Interspecies analysis

Interspecies pairwise comparisons were performed by aligned identified open chromatin regions between species in a reciprocal manner using UCSC liftOver [73] on genomic assemblies in three species. Each of the species was used anchor species and the open chromatin regions were mapped to the other 2 species with 50% minimum map ratio. Regions failing to be mapped in any of the other genomes were considered as unaligned regions. To identify conserved regions between two species, regions having orthologous regions overlapped in the second species with at least 1bp were collected. Final pairwise conserved regions were confirmed by doing this comparison reciprocally between species.

4.7 References

1. Zorn, A. M., & Wells, J. M. (2009). Vertebrate endoderm development and organ formation. *Annual Review of Cell and Developmental*, *25*, 221-251.
2. Sinner, D., Kirilenko, P., Rankin, S., Wei, E., Howard, L., Kofron, M., ... & Zorn, A. M. (2006). Global analysis of the transcriptional network controlling Xenopus endoderm formation. *Development*, *133*(10), 1955-1966.
3. Charney, R. M., Paraiso, K. D., Blitz, I. L., & Cho, K. W. (2017, June). A gene regulatory program controlling early Xenopus mesendoderm formation: network conservation and motifs. In *Seminars in cell & developmental biology* (Vol. 66, pp. 12-24). Academic Press.
4. Tseng, W. F., Jang, T. H., Huang, C. B., & Yuh, C. H. (2011). An evolutionarily conserved kernel of gata5, gata6, otx2 and prdm1a operates in the formation of endoderm in zebrafish. *Developmental biology*, *357*(2), 541-557.
5. Woodland, H. R., & Zorn, A. M. (2008). The core endodermal gene network of vertebrates: combining developmental precision with evolutionary flexibility. *Bioessays*, *30*(8), 757-765.
6. Wiesenfahrt, T., Osborne Nishimura, E., Berg, J. Y., & McGhee, J. D. (2016, July). Probing and rearranging the transcription factor network controlling the C. elegans endoderm. In *Worm* (Vol. 5, No. 3, pp. 483-91). Taylor & Francis.
7. Shi, W., Levine, M., & Davidson, B. (2005). Unraveling genomic regulatory networks in the simple chordate, Ciona intestinalis. *Genome research*, *15*(12), 1668-1674.
8. Peter, I. S., & Davidson, E. H. (2010). The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage. *Developmental biology*, *340*(2), 188-199.
9. Zorn, A. M., & Wells, J. M. (2007). Molecular basis of vertebrate endoderm development. *International review of cytology*, *259*, 49-111.

10. Stainier, D. Y. (2002). A glimpse into the molecular entrails of endoderm formation. *Genes & development*, *16*(8), 893-907.
11. Schier, A. F., & Shen, M. M. (2000). Nodal signalling in vertebrate development. *Nature*, *403*(6768), 385.
12. D'Amour, K. A., Agulnick, A. D., Eliazer, S., Kelly, O. G., Kroon, E., & Baetge, E. E. (2005). Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nature biotechnology*, *23*(12), 1534.
13. Shen, M. M. (2007). Nodal signaling: developmental roles and regulation. *Development*, *134*(6), 1023-1034.
14. Zhang, J., Houston, D. W., King, M. L., Payne, C., Wylie, C., & Heasman, J. (1998). The role of maternal VegT in establishing the primary germ layers in Xenopus embryos. *Cell*, *94*(4), 515-524.
15. Clements, D., Friday, R. V., & Woodland, H. R. (1999). Mode of action of VegT in mesoderm and endoderm formation. *Development*, *126*(21), 4903-4911.
16. Xanthos, J. B., Kofron, M., Wylie, C., & Heasman, J. (2001). Maternal VegT is the initiator of a molecular network specifying endoderm in Xenopus laevis. *Development*, *128*(2), 167-180.
17. Teo, A. K. K., Arnold, S. J., Trotter, M. W., Brown, S., Ang, L. T., Chng, Z., ... & Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes & development*.
18. Ryan, K., Garrett, N., Mitchell, A., & Gurdon, J. B. (1996). Eomesodermin, a key early gene in Xenopus mesoderm differentiation. *Cell*, *87*(6), 989-1000.
19. Conlon, F. L., Fairclough, L., Price, B. M., Casey, E. S., & Smith, J. C. (2001). Determinants of T box protein specificity. *Development*, *128*(19), 3749-3758.
20. Chiu, W. T., Le, R. C., Blitz, I. L., Fish, M. B., Li, Y., Biesinger, J., ... & Cho, K. W. (2014). Genome-wide view of TGFβ/Foxh1 regulation of the early mesendoderm program. *Development*, dev-107227.
21. Kofron, M., Puck, H., Standley, H., Wylie, C., Old, R., Whitman, M., & Heasman, J. (2004). New roles for FoxH1 in patterning the early embryo. *Development*, *131*(20), 5065-5078.
22. Hoodless, P. A., Pye, M., Chazaud, C., Labbé, E., Attisano, L., Rossant, J., & Wrana, J. L. (2001). FoxH1 (Fast) functions to specify the anterior primitive streak in the mouse. *Genes & development*, *15*(10), 1257-1271.
23. Yamamoto, M., Meno, C., Sakai, Y., Shiratori, H., Mochida, K., Ikawa, Y., ... & Hamada, H. (2001). The transcription factor FoxH1 (FAST) mediates Nodal signaling during anterior-posterior patterning and node formation in the mouse. *Genes & Development*, *15*(10), 1242-1256.
24. Slagle, C. E., Aoki, T., & Burdine, R. D. (2011). Nodal-dependent mesendoderm specification requires the combinatorial activities of FoxH1 and Eomesodermin. *PLoS genetics*, *7*(5), e1002072.
25. Yoon, S. J., Wills, A. E., Chuong, E., Gupta, R., & Baker, J. C. (2011). HEB and E2A function as SMAD/FOXH1 cofactors. *Genes & development*, *25*(15), 1654-1661.
26. Slagle, C. E., Aoki, T., & Burdine, R. D. (2011). Nodal-dependent mesendoderm specification requires the combinatorial activities of FoxH1 and Eomesodermin. *PLoS genetics*, *7*(5), e1002072.

27. Latinkic, B. V., Umbhauer, M., Neal, K. A., Lerchner, W., Smith, J. C., & Cunliffe, V. (1997). The Xenopus Brachyury promoter is activated by FGF and low concentrations ofactivinandsuppressed by high concentrationsof activin and by paired-type homeodomain proteins. *Genes & development*, *11*(23), 3265-3276.
28. Lemaire, P., Darras, S., Caillol, D., & Kodjabachian, L. (1998). A role for the vegetally expressed Xenopus gene Mix. 1 in endoderm formation and in the restriction of mesoderm to the marginal zone. *Development*, *125*(13), 2371-2380.
29. Colas, A., Cartry, J., Buisson, I., Umbhauer, M., Smith, J. C., & Riou, J. F. (2008). Mix. 1/2-dependent control of FGF availability during gastrulation is essential for pronephros development in Xenopus. *Developmental biology*, *320*(2), 351-365.
30. Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., ... & Otim, O. (2002). A genomic regulatory network for development. *science*, *295*(5560), 1669-1678.
31. Levine, M., & Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, *102*(14), 4936-4942.
32. Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., ... & Thurman, R. E. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, *515*(7527), 365.
33. D'Amour, K. A., Bang, A. G., Eliazer, S., Kelly, O. G., Agulnick, A. D., Smart, N. G., ... & Baetge, E. E. (2006). Production of pancreatic hormone–expressing endocrine cells from human embryonic stem cells. *Nature biotechnology*, *24*(11), 1392.
34. Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N. A., ... & Lam, D. K. (2015). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell stem cell*, *16*(4), 386-399.
35. Kim, P. T., Hoffman, B. G., Plesner, A., Helgason, C. D., Verchere, C. B., Chung, S. W., ... & Ong, C. J. (2010). Differentiation of mouse embryonic stem cells into endoderm without embryoid body formation. *PLoS One*, *5*(11), e14146.
36. Yasunaga, M., Tada, S., Torikai-Nishikawa, S., Nakano, Y., Okada, M., Jakt, L. M., ... & Nishikawa, S. I. (2005). Induction and monitoring of definitive and visceral endoderm differentiation of mouse ES cells. *Nature biotechnology*, *23*(12), 1542.
37. Mfopou, J. K., Geeraerts, M., Dejene, R., Van Langenhoven, S., Aberkane, A., Van Grunsven, L. A., & Bouwens, L. (2014). Efficient definitive endoderm induction from mouse embryonic stem cell adherent cultures: a rapid screening model for differentiation studies. *Stem cell research*, *12*(1), 166-177.
38. Morrison, G., Scognamiglio, R., Trumpp, A., & Smith, A. (2015). Convergence of cMyc and β-catenin on Tcf7l1 enables endoderm specification. *The EMBO journal*, e201592116.
39. Lu, J., Baccei, A., da Rocha, E. L., Guillermier, C., McManus, S., Finney, L. A., ... & Lerou, P. H. (2018). Single-cell RNA sequencing reveals metallothionein heterogeneity during hESC differentiation to definitive endoderm. *Stem cell research*, *28*, 48-55.
40. Wang, Z., Oron, E., Nelson, B., Razis, S., & Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell stem cell*, *10*(4), 440-454.
41. Chu, L. F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., ... & Thomson, J. A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, *17*(1), 173.

42. Norrman, K., Strömbeck, A., Semb, H., & Ståhlberg, A. (2013). Distinct gene expression signatures in human embryonic stem cells differentiated towards definitive endoderm at single-cell level. *Methods*, *59*(1), 59-70.
43. Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096-1102.
44. Teo, A. K. K., Arnold, S. J., Trotter, M. W., Brown, S., Ang, L. T., Chng, Z., ... & Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes & development*.
45. Wang, Z., Oron, E., Nelson, B., Razis, S., & Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell stem cell*, *10*(4), 440-454.
46. Jaenisch, R., & Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, *132*(4), 567-582.
47. Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., ... & Gifford, D. K. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *cell*, *122*(6), 947-956.
48. Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., ... & Wong, K. Y. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics*, *38*(4), 431.
49. Artinger, M., Blitz, I., Inoue, K., Tran, U., & Cho, K. W. (1997). Interaction of goosecoid and brachyury in Xenopus mesoderm patterning. *Mechanisms of development*, *65*(1-2), 187-196.
50. Pereira, L. A., Wong, M. S., Mossman, A. K., Sourris, K., Janes, M. E., Knezevic, K., ... & Elefanty, A. G. (2012). Pdgfrα and Flk1 are direct target genes of Mixl1 in differentiating embryonic stem cells. *Stem cell research*, *8*(2), 165-179.
51. Sinner, D., Rankin, S., Lee, M., & Zorn, A. M. (2004). Sox17 and β-catenin cooperate to regulate the transcription of endodermal genes. *Development*, *131*(13), 3069-3080.
52. Howard, L., Rex, M., Clements, D., & Woodland, H. R. (2007). Regulation of the Xenopus Xsox17α 1 promoter by co-operating VegT and Sox17 sites. *Developmental biology*, *310*(2), 402-415.
53. Séguin, C. A., Draper, J. S., Nagy, A., & Rossant, J. (2008). Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells. *Cell stem cell*, *3*(2), 182-195.
54. Shimoda, M., Kanai-Azuma, M., Hara, K., Miyazaki, S., Kanai, Y., Monden, M., & Miyazaki, J. I. (2007). Sox17 plays a substantial role in late-stage differentiation of the extraembryonic endoderm in vitro. *Journal of cell science*, *120*(21), 3859-3869.
55. Shimosato, D., Shiki, M., & Niwa, H. (2007). Extra-embryonic endoderm cells derived from ES cells induced by GATA factors acquire the character of XEN cells. *BMC developmental biology*, *7*(1), 80.
56. Afouda, B. A., Ciau-Uitz, A., & Patient, R. (2005). GATA4, 5 and 6 mediate TGFβ maintenance of endodermal gene expression in Xenopus embryos. *Development*, *132*(4), 763-774.
57. Tiyaboonchai, A., Cardenas-Diaz, F. L., Ying, L., Maguire, J. A., Sim, X., Jobaliya, C., ... & De Leon, D. D. (2017). GATA6 plays an important role in the induction of

human definitive endoderm, development of the pancreas, and functionality of pancreatic β cells. *Stem cell reports*, *8*(3), 589-604.
58. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, *10*(11), 1096.
59. Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., ... & Kathiria, A. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature methods*, *14*(10), 959.
60. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.
61. Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12*(1), 323.
62. Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420). Springer, New York, NY.
63. Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096-1102.
64. Tripathi, S., Pohl, M. O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D. A., ... & Yángüez, E. (2015). Meta-and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. *Cell host & microbe*, *18*(6), 723-735.
65. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25.
66. Picard http://broadinstitute.github.io/picard/
67. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, *38*(4), 576-589.
68. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.
69. Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., & Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, *41*(21), e201-e201.
70. Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017-1018.
71. Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., & Makeev, V. J. (2012). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, *41*(D1), D195-D202.
72. Longabaugh, W. J. (2012). BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. In *Gene Regulatory Networks* (pp. 359-394). Humana Press.
73. Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., ... & Harte, R. A. (2014). The UCSC genome browser database: 2015 update. *Nucleic acids research*, *43*(D1), D670-D681.

CHAPTER 5

Dynamics of NRSF/REST motif evolution favor the canonical NRSE/RE1 form

Notes: (1) Dr. Ricardo N. Ramirez cultured HL60 and made chromatin for ChIP-seq.
(2) Nicole El-Ali sequenced the samples.
(3) Dr. Ali Mortazavi conceived the idea and provided continued support and guidance throughout the project.

Chapter 5

Dynamics of NRSF/REST motif evolution favor the canonical NRSE/RE1 form

5.1 Abstract

The evolution of the binding repertoire of transcription factors is one of the key questions of comparative genomics. The transcription factor NRSF/REST represses many vertebrate neuronal genes in non-neuronal cells by binding to 3 distinct motif classes, which are the canonical 21bp NRSEs (C), longer non-canonical motifs (NC) and solo half-motifs (HF). We used ChIP-seq in four mammalian species to determine the evolution of the NRSF binding repertoire. We show that while some NRSEs are deeply conserved, genes with several NRSEs show evidence of compensatory binding turnover, suggesting that the association of the transcription factor to its target gene is more important than the specific binding instances. We also found that many newborn binding sites in human are associated with primate specific indels and transposable elements. Our analysis of motifs with conserved ChIP-binding in all 4 species demonstrates that both the non-canonical and solo half-motifs convert preferentially to canonical motifs. These findings support a model of dynamic conversion between different motif types that account for the preferential accumulation of the canonical NRSE during evolution.

5.2 Introduction

Non-coding elements selectively bound by specific transcription factors (TF) mechanistically regulate gene expression. Changes in these transcription factor binding sites, also known as *cis*-regulatory DNA elements, are believed to regulate the gain, loss or modification of traits across species [1-5].Transcription factor binding repertoires have expanded during vertebrate evolution [6], which provide more evolutionary raw material for *cis*-regulatory elements in rewiring gene regulatory networks. TF binding sites can be identified genome-wide using ChIP-seq [7] to assess the conservation and divergence of TF binding sites across close as well as distant species [8-13]. Tissue specific transcription factors showed surprising divergence in the rate and binding across closely related mammalian species [10-13]. Although individual TF binding sites are less constrained in both closely and distantly related mammals, TF regulatory networks are more highly conserved [14], which indicates strong evolutionary selection on TF regulatory networks rewired from unconstrained TF binding sites. Previous

comparative *cis*-regulatory studies have mainly focused on activators and the insulator CTCF. However, at least one-third of the transcription factors in mammals are thought to be repressors [6,15,16] and their evolutionary role could be significantly different, especially if they are involved in the silencing of specific repeat families [17].

The transcription factor Neuron Restrictive Silencer Factor NRSF [18], also known as REST (repressor element 1 silencing transcription factor) [19], primarily represses neuronal gene expression in non-neuronal cells and neuronal stem cells [20]. NRSF recruits several cofactor complexes to silence target genes by binding to a canonical 21bp NRSE/RE1 motif [21-27]. ChIP-seq and other genome-wide studies found that NRSF binds not only to its originally identified 21bp canonical (C) NRSE/RE1 motifs, but also to a smaller class of non-canonical (NC) NRSEs, which consist of two half-motifs (HF) separated by 10, 16-19 base pairs [7, 28]. Previous studies have found that canonical motifs are present in most but not all of the NRSF binding sites in non-neuronal cell types from different species [7, 29, 30]. However, there is no significant difference between canonical and non-canonical NRSEs in terms of binding with NRSF and repression of neuronal gene in non-neuronal cells [28] even though NRSF binding with either canonical or non-canonical motifs represses genes to a lower expression levels compared with half-motifs only or no motif [29]. Given that multiple non-canonical forms of the NRSE motifs would seem to provide more flexibility for evolving a functional binding site that is just as good as the canonical motif in terms of binding and repression, it is surprising that the canonical form is much more prevalent than the non-canonical forms.

A comparative study of the distribution of the canonical NRSE in vertebrate and invertebrate species found that the number of NRSE is relatively constant in mammalian genomes when compared with other vertebrates and that the canonical motif is absent in invertebrates [31]. Comparative analyses of canonical NRSEs in human [32] and *Xenopus tropicalis* [33] genomes with other vertebrate genomes demonstrated the existence of lineage-specific NRSEs enriched in neuronal genes. Another comparative analysis of NRSF binding between human and mouse showed significant expansion of NRSF binding in human embryonic stem cells, compared with mouse, in genes involved in human-specific neuronal regulatory functions [30]. However, these studies did not address the evolutionary relationship between the

three classes of canonical, non-canonical NRSEs and half-motifs. In order to determine the relationship between these classes, we performed a cross-species NRSF binding analysis across four mammalian species (human, mouse, dog and horse) to systematically examine the extent of NRSF binding birth, death, and motif conversion between classes. Based on our results, we propose a motif conversion model to explain how NRSF canonical motifs evolve from non-canonical motifs and solo half-motifs.

5.3 Results

5.3.1 The majority of NRSF binding instances are mediated by canonical motif

In order to use consistent terminology, we use the term *motif* when referring to canonical, non-canonical NRSEs and half-motifs. The NRSF bound DNA sequence detected by ChIP-seq is a *site* where we will search for the *motif* if present. A site could potentially be bound only in some cell types or be abolished due to a mutation. We also use the terms *peak* and *instance* to describe the observed binding between the NRSF protein and the *site* in our data. We performed ChIP-seq in biological replicates for NRSF in mouse (RAW.264) macrophage and dog (ML-2) myeloid cells as well as horse fibroblast (E.Derm) in conjunction with human NRSF ChIP-seq data from the ENCODE project [34] in the HL-60 myeloid cell line to identify and to compare NRSF binding instances across all 4 species. NRSF is known to repress the same genes in different non-neuronal cell types and we therefore do not expect substantial biological differences when comparing myeloid cells in human and dog and macrophage in mouse to horse fibroblasts. We identified 5260 reproducible NRSF peaks in human, 2357 in mouse, 2273 in dog and 2185 in horse (Figure 5.1A, 5.2 & 5.3). We also performed NRSF peak calling on published human embryonic stem cell (ESCs) [34] and mouse ESCs [35] as controls. We identified 10322 peaks in hESC and 3970 peaks in mESC. Our hESC peak calling result is close to the number (*n*=10141) reported by ENCODE project [34, 36] and found 2000 more peaks not identified in previous study of the same cells [30]. Canonical and non-canonical NRSE motifs are composed of two 10 bp half-motifs, separated respectively by 11 bp for the canonical and 10 or 16-19 bp for the non-canonical form (details in *Methods*). We found that 45% to 63% (*n*=2410 in human; *n*=1310 in mouse; *n*=1424 in dog; *n*=1146 in horse) of NRSF binding sites had canonical motifs while only about 10% to 14% (*n*=710 in human; *n*=331 in mouse; *n*=282 in dog; *n*=213 in horse) of sites had non-canonical motifs in each of the four species (Figure 5.1A). Consistent with

previous studies [29,30], we also found that a large fraction of binding sites have solo half-motifs only (Figure 5.1A) and that solo half-motifs are more frequent in human (*n*=1373, 26.10%, 95% CI [0.250, 0.273]) compared to other mammals (*n*=370, 15.70%, 95% CI [0.143, 0.172] in mouse; *n*=312, 13.73%, 95% CI [0.124, 0.151] in dog; *n*=224, 10.25%, 95% CI [0.090, 0.115] in horse) (Figure 5.1A). Mapping motifs back to NRSF binding sites revealed that NRSF ChIP-seq signal density changed significantly with the type of motifs in the binding sites. As expected, sites with high ChIP-seq binding density were enriched with canonical motifs while low binding density sites were enriched with solo half-motifs or no-motif (Figure 5.2 & 5.4). Non-canonical motifs were distributed throughout the signal range, confirming that some non-canonical NRSEs bind as well as or better than canonical NRSEs [7]. Gene ontology (GO) analysis on genes with canonical motifs found species-specific enrichments (Figure 5.5 & 5.6). For example, human-specific peaks with canonical motifs were enriched with genes involved in trans-synaptic signaling (n=161, $p=3.89\times10^{-46}$) and central nervous system development (n=149, $p=4.47\times10^{-19}$) while mouse-specific peaks were not only enriched with genes involved in chemical synaptic transmission (n=98, $p=2.24\times10^{-32}$) and regulation of nervous system development (n=109, $p=6.46\times10^{-23}$) but also in metal ion transport (n=95, $p=4.37\times10^{-21}$). We also found that peaks with non-canonical motifs in human were enriched in genes in chemical synaptic transmission (n=51, $p=1.58\times10^{-12}$) and regulation of nervous system development (n=44, $p=3.80\times10^{-6}$), but non-canonical mouse peaks were enriched for neuron projection development (n=28, $p=3.31\times10^{-6}$) and regulation of membrane potential (n=14, $p=1.66\times10^{-4}$). While solo half-motifs and no motif peaks show apparently weaker enrichment of genes with neuronal GO terms in mouse, the genes with such peaks in human still maintained substantial neuronal GO enrichment.

Taken together, our methods comprehensively identified and categorized NRSF binding sites based on motif class. More than 72% of NRSF binding sites were occupied by one or more instances of the 3 classes, with at least 60% of NRSF binding sites containing either canonical motifs or non-canonical motifs that are associated with higher NRSF ChIP-seq binding density in each of the four species. Furthermore, our results suggest that different motif classes might contribute to NRSF-associated pathways in a species-specific manner.

5.3.2 Canonical motifs are mostly enriched in deeply conserved binding peaks

While a gene such as PAX5 is associated with NRSF binding and has one perfectly conserved non-canonical binding motif inside the first exon in all 4 species, each species has an additional set of species-specific binding sites gains and losses in other parts of the gene (Figure 5.1B). We investigated conserved and divergent NRSF binding peaks across our four species. We categorized NRSF binding peaks as deeply conserved if bound in all species, partially conserved if bound in at least two and species-specific peaks based on the extent of peak conservation across species (Figure 5.7A). On average, 27% (*n*=3258 out of 12075) of NRSF binding peaks were shared between any two or three species and thus partially conserved and about 12% (*n*=1460 out of 12075) of total binding peaks were deeply conserved in all four species. Specifically, only 7% (*n*=365 out of 5260) of binding peaks in human were deeply conserved, with this fraction rising to about 16% (*n*=365 out of 2357 in mouse; *n*=365 out of 2273 in dog; *n*=365 out of 2185 in horse) in our three other mammals. Conversely, approximately 61% (*n*=7357 out of 12075) of total binding peaks were species-specific and had no orthologous regions overlapping with identified peaks in any other species. Species-specific peaks represent about 72% (*n*=3770 out of 5260) of NRSF binding in human while they occupied about 48%-60% (*n*=1441 out of 2357 in mouse; *n*=1088 out of 2273 in dog; *n*=1058 out of 2185 in horse) in other species. By investigating peak conservation in human and mouse embryonic stem cells instead, we found that about 70%-80% of NRSF peaks were species-specific in human and mouse while about 46% were species-specific in dog and horse (Figure 5.8).

Of the 599 human genes that have only one NRSF binding site, 271 genes (45.2%) show evidence of binding in 3 or 4 species; there are also 15 genes that show evidence of binding in mouse, dog, and horse but not human for a total of 286 genes with evidence of 1 binding site in at least 3 species for a total of 380 sites; there are more sites than genes, as some sites do not align. We observed that 215 of these 286 genes (75.17%) have their solo sites conserved in at least three species in multiple species alignments, while only 21 genes (7.34%) have no detectable site conservation (i.e. the sites do not align in multiple sequence alignments) (Table 5.1). Similarly, there are 606 genes having more than one associated binding sites in 3 or 4 species with 2282 sites and we found that 500 (82.51%) of those genes have at least one conserved binding site in 3 or more species while 29 (4.79%) genes have no conserved site. A

comparison of the number of binding sites turnover (i.e. sites alignable in only 1 or 2 species in genes with signal in at least 3 species) shows that 43.42% (165/380) of sites turnover for genes with only one binding site per gene while 73.93% (1687/2282) of sites turnover for genes with more than one binding sites ($p= 4.086x10^{-11}$, Chi-square test) in 3 or more species. This suggests that the presence of more than one binding site per gene increases the likelihood of any one site turning over while ensuring regulation by NRSF and is irrespective of the quality of the ChIP-seq data in any one species.

Motif type occupancy changes with the conservation level of peaks (Figure 5.7B & 5.9). The proportion of canonical motifs was higher while the proportion of solo half-motifs and no-motif was lower in conserved peaks when compared to non-conserved peaks. Motifs in deeply and three-species partially conserved peaks had at least one half-motif, but no-motif was found only highly enriched in species-specific peaks. Interestingly, compared with canonical NRSEs and solo half-motifs, the occupancy of non-canonical motifs did not change much according to the conservation level of peaks. Focusing on the deeply conserved peaks, we found that 93% (*n*=340) of peaks shared had the same type of motifs in four species and that the canonical motifs represent the highest proportion (77%, *n*=283) and solo half-motifs the least (3%, *n*=11) (Figure 5.7D).

A comparison of 4493 motifs found in human NRSF ChIP-seq peaks with their orthologous regions in other primate species, as well as mouse, dog and horse showed that a substantial subset of these motifs were primate-specific (Figure 5.7C). As expected, the overlapping fraction with human motifs in alignable regions decreased apparently with the earlier divergence from human. Whereas 93% of alignable human motifs were shared with chimp, 76% of alignable motifs were shared with rhesus macaque. However, we only found about 43-50% of alignable human motifs were shared with mouse, dog, horse and bushbaby, which is a non-anthropoid primate. Interestingly, anthropoid-specific motifs, i.e. motifs that are in anthropoid species but are missing from bushbaby and other mammals, were significantly enriched in genes with GO annotations for synaptic signaling ($n=95$, $p=8.51x10^{-17}$), regulation of nervous system development ($n=93$, $p=5.75x10^{-10}$), and neurotransmitter transport ($n=32$, $p=7.94x10^{-7}$) as well as neuronal specific transcription factors such as NEUROD4, POU3F1 and VSX1. Some of the TFs

with anthropoid-specific motifs like OLIG2, is negative regulator of transcription. In contrast, motifs specific for bushbaby, mouse, dog and horse were only enriched in negative regulation of multicellular organismal process (n=5, $p=6.03x10^{-4}$), developmental maturation (n=3, $p=6.61x10^{-4}$) and protein targeting (n=3, $p=1.35x10^{-2}$). Altogether, our observation reveals that canonical motifs are more likely to be enriched in conserved peaks whereas there is more motif diversity in the non-conserved peaks and that a large fraction of NRSEs are anthropoid primate-specific and have arisen more recently during primate evolution.

5.3.3 Rapid NRSF binding sites turnover is mediated by transposable elements and base sequence changes

Species-specific changes in the genomic sequence account for transcription factor binding instance divergence in both closely and distantly related species [10, 11, 13]. We identified peaks in alignable orthologous regions that were only present in one species and missing in the other three ("newborn") or bound in three species but missing in the fourth species ("dead"). We parsimoniously surmised that a peak found only in one species was more likely to be novel than "old" but lost in two or three species independently. We required that canonical or non-canonical motifs lie in peaks and had higher motif occurrence score than peak-free regions for both birth and death (details in Methods). We detected more newborn instances (543) than death instances (101) in four species (Figure 5.10). Gene ontology analysis revealed that genes associated with newborn binding instances are predicted to have neuronal functions. For example, synaptic signaling (n=29, $p=1.66x10^{-10}$) and regulation of nervous system development (n=27, $p=4.07x10^{-7}$) were significantly enriched in human newborn instances while mouse newborn instances were enriched in genes involved in neuronal action potential (n=3, $p=9.33x10^{-4}$) and forebrain development (n=7, $p=1.32x10^{-3}$). We found no regulatory function significantly enriched in genes with death instances. We also analyzed the gene ontology enrichment of genes with NRSF binding sites conserved in all four species and found that neuronal GO terms such as synaptic signaling (n=59, $p=3.72x10^{-31}$), regulation of membrane potential (n=34, $p=2.40x10^{-18}$), regulation of nervous system development (n=39, $p=1.86x10^{-11}$) had even higher enrichment and significance in conserved sites than in newborn sites. This suggests that conserved NRSF binding sites are likely to play a key role in neuronal development.

Nearly half of the genomes of human, mouse, dog and horse are composed of transposable elements (TEs) [37]. Previous studies have reported that TEs play an important role in the appearance of recently evolved *cis*-regulatory elements and remodel gene regulatory networks [11,38-40). We therefore studied the association of TEs in both newborn and dead instances for each species and found a strong association between TEs and species-specific newborn instances. Consistent with previous studies [30,31,41], human newborn instances are significantly enriched in LINE2. Moreover, the high enrichment of LINE2 was also shared in mouse and dog newborn instances (Figure 5.11). We also found that species-specific newborn instances were associated with additional families of TEs. MIR from the SINE family was only enriched in human whereas ERVL from the LTR family and CR1 from LINE family were specially enriched in dog and horse birth respectively. As expected, there is no TE class significantly associated with dead instances (Figure 5.12). Interestingly, we also detected TEs in deeply conserved sites and found exclusively high enrichment of MIR in four species (Figure 5.13). Our observation further confirmed that TEs might serve as ancestral sequence sources that can get converted into newborn NRSF binding instances. Further, species-specific birth can both arise from TEs that are shared between the four species but also arise from species-specific TEs classes.

To further understand how sequence changes drive species-specific birth of binding instances, we performed multispecies alignment analysis on 289 human newborn motifs (Figure 5.10) to mouse, dog , horse and six primate species (Chimpanzee, Gibbon, Rhesus macaque, Baboon, Squirrel monkey, Bushbaby). We found that insertions and deletions contributed to newborn instances and motif conversion in orthologous regions. In the comparison with the other mammals multi-species alignment, human insertion means that bases were inserted in orthologous regions while deletion means bases deleted in orthologous regions (details in Methods). We found 21 human newborn motifs that appeared during primate evolution. These motifs shared intact sequences with at least one primate species but had bases missing in orthologous regions for other mammalian species (Figure 5.14A). About (51/63) 81% of newborn NRSF motifs were only in the motif sequences when compared in orthologous regions, i.e. the bases missing happened for mouse, dog and horse that resulted in (45/63) 71% of motifs missing or only solo half-motifs left. A further 16 human newborn motifs were created by

primate-specific deletion, in which (41/48) 85% of motifs missing or only half-motif left in mouse, dog and horse (Figure 5.14B). Altogether, our observations reveal a rapid turnover process, in which human-specific NRSF binding instances were born during the last 70 million years of primate evolution driven by insertion and deletion.

5.3.4 Non-canonical and half-motifs show higher conversion rate to canonical motifs

Our results have shown that canonical motifs constitute the majority of conserved NRSF binding instances and they were significantly ($p<1.0x10^{-12}$) associated with high NRSF binding density in four species (Figure 5.1A, 5.2 & 5.4). Canonical motif occupancy is positively correlated with the conservation level of NRSF binding instances for each species, whereas half-motifs are negatively correlated. Interestingly, non-canonical motifs show a relatively stable proportion in conserved and species-specific binding instances. Canonical motifs are enriched not only in conserved sites but also in nonconserved sites for NRSF binding across species (Figure 5.7B & D). In order to understand why we observe more conserved canonical motifs than conserved non-canonical motifs or solo half-motifs, we focused on deeply conserved instances to observe the conversion dynamics between different motif forms (details in Methods). By focusing on alignable regions that had identical motifs in 3 species but changed in the fourth we could ascertain what the ancestral sequence was and the resulting change in the fourth (Figure 5.15A). We found that both non-canonical motifs ($p=7.96x10^{-49}$) and solo half-motifs ($p=2.82x10^{-17}$) had significantly higher conversion rate to canonical motifs compared with the reverse direction. We also found a significantly higher rate of motif conversion from solo half-motifs to non-canonical motifs compared with the reverse direction ($p=3.24x10^{-7}$) (Figure 5.15B-C & 5.16). In particular, we found no instances of canonical motifs converting into a non-canonical motif and only 0.35% (1/284) of them changed to half-motifs whereas 9.80% (5/51) of non-canonical motifs and 12.50% (2/16) of half-motifs were converted to canonical motifs. Thus half-motifs either tend to become canonical motifs or non-canonical motifs while non-canonical motifs in turn also convert into canonical motifs. Note that this result is robust, as it still holds if we restrict ourselves to comparisons between only the 3 myeloid cell lines in human, mouse, and dog (Figure 5.17) or when substituting human and mouse embryonic stem cells instead (Figure 5.18), which suggests that cell identity does not affect the results.

5.4 Discussion

We systematically compared the binding of NRSF/REST at canonical, non-canonical and solo half-motifs in four mammalian species to construct a comprehensive profile of conserved and evolving NRSE/RE1 motifs. Our data also shows that many of newborn sites are caused by transposable elements as well as individual base changes and small indels. The four-species analysis allows us to analyze the evolutionary relationship of three NRSE motif categories and to propose a motif conversion model of non-canonical and solo half-motifs converting preferentially to canonical forms over time, thus explaining the preponderance of the canonical NRSE.

The length of the canonical (21bp) and non-canonical NRSEs (20, 26-29bp) is probably too long to be generated easily from random local point mutations alone [41, 42]. The exaptation of transposable elements clearly accounts for some of the novel binding instances. We found that LINE/LINE2 enriched in human-, mouse-, and dog- specific instances. Other TEs families were also enriched in species-specific binding. Interestingly, we observed that SINE/MIR was the only TE family highly enriched in deeply conserved instances in all four species, particularly in dog and horse. As old TEs families, LINE2 and MIR propagated primarily before mammalian radiation [43,44] as opposed to the more recent expansion of LINE1, LTR and other SINE sub-families [37]. Our results suggest NRSE motifs might have been exapted in different waves, first with MIRs in early mammalian evolution, then LINE2 and more recently other LTR, SINE as well as LINE sub-families actively transposed in a more species-specific manner. Over time, sequences that are weakly NRSE-like would turn into stronger ones via a handful of point mutations.

The use of binding instances conserved in all four species allowed us to measure parsimoniously the rate at which non-canonical and half-motifs tend to become canonical over evolutionary time. This evolutionary dynamic between NRSE classes results in the accumulation of canonical NRSEs that we detect even in species-specific binding instances. While it is still not clear why the canonical form is ultimately preferred over the conserved one, given that noncanonical motifs seem to repress just as well as the canonical ones, it may be that the flexibility of NRSF in binding full motifs with non-standard linker spacing provides a greater set

of alternative paths to evolve from a solo half-motif to a full-fledge non-canonical motif, which can then be refined into canonical motifs by small indels that are relatively frequent in mammalian genomes. Interestingly, NRSF is one of many transcription factors that can bind to more than one motif class where the longest motif can be divided into two short half-motifs. For example, the 33/34 bp CTCF binding motif is composed of 20bp motif as M1 and 9bp motif as M2 with 20/21 spacer. Studies have found that binding instances with 33/34 bp motif shown stronger ChIP enrichment and higher occupancy than instances containing only M1 in deeply conserved CTCF binding across five mammalian species [11]. Based on our model, the full 33/34 bp CTCF motifs might accumulate by the rapid conversion from M1 motifs during evolution. This conversion process might be driven by repetitive element expansion to bring new 33/34 bp motifs to genomes [11]. Similar two half-motifs long motifs have also been found for p53 [45, 46]. However, the lack of comparative binding profile makes it difficult to confirm evolutionary relationship of p53 motifs during evolution. Additional studies of multi-class transcription factors binding to determine whether other factors have multi-step binding motifs evolutionary paths similar to that of NRSF will shed light on how binding instances arise and are refined during evolution.

5.5 Figures

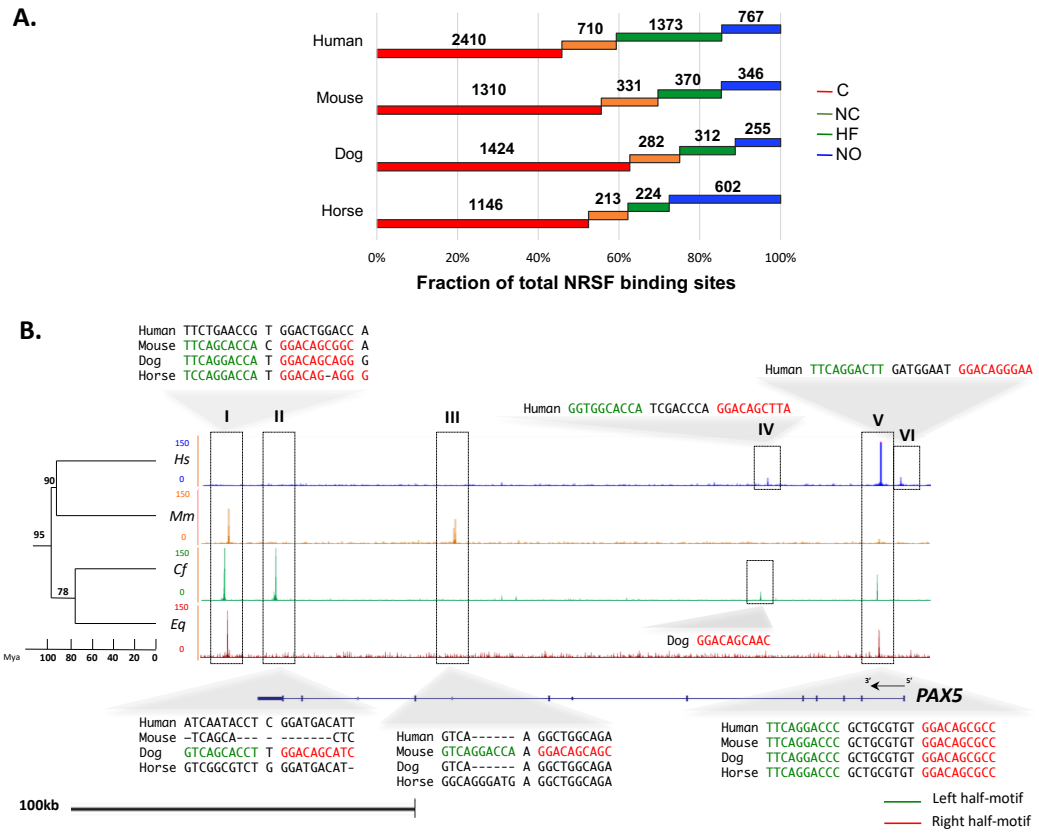


Figure 5.1. Genome-wide identification of NRSF binding sites across four species. **(A)** Motif distribution in NRSF binding sites of each species. **(B)** Multiple binding sites turnover in PAX5, which has three binding sites in human, three in mouse, four in dog and two in horse. Column I shows site death in human compared with orthologous regions in the other three species; II shown a site born in dog; III shown a site born in mouse. Column IV shows that NRSF binding only appears in human and dog in orthologous regions, but their motifs are not alignable. Column V shows binding shared in four species. Column VI shows human-specific binding that no aligned regions in other species. Evolutionary tree of four mammalian species was adapted from TIMETREE (median) [47].

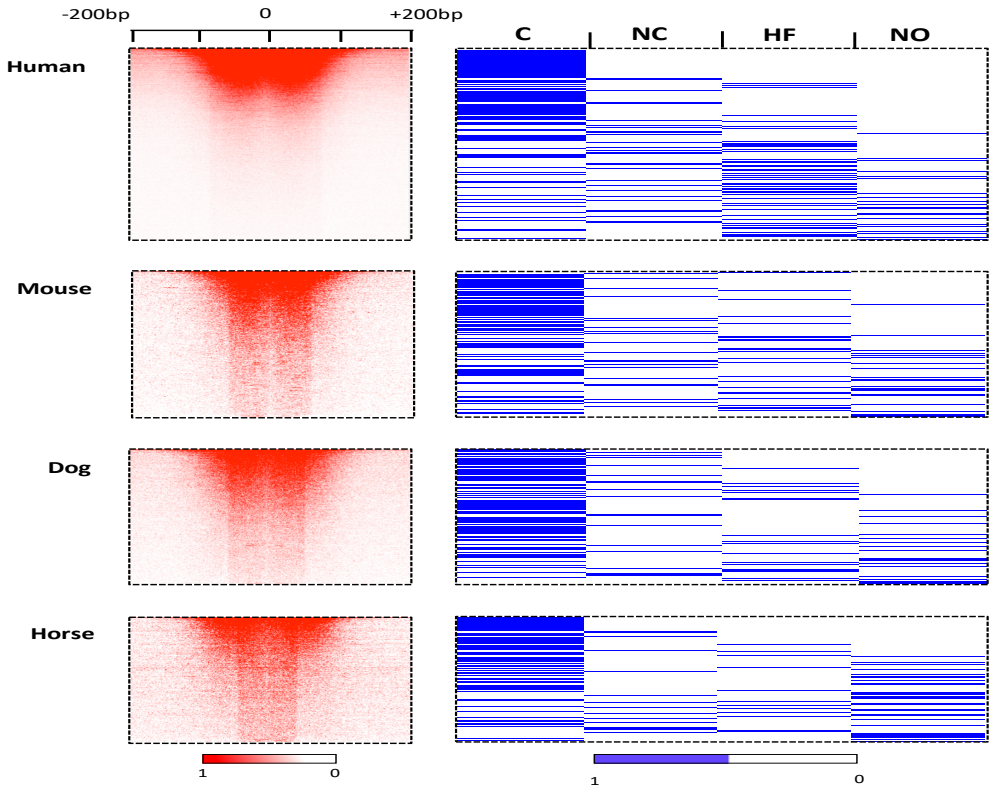


Figure 5.2. NRSF ChIP-seq binding signal in 200bp window around peak summit (left). Red, high binding density; White, low binding density ; Type of motifs in peaks (right). C, canonical motif; NC, non-canonical motif; HF, solo half-motif; NO, no motif.

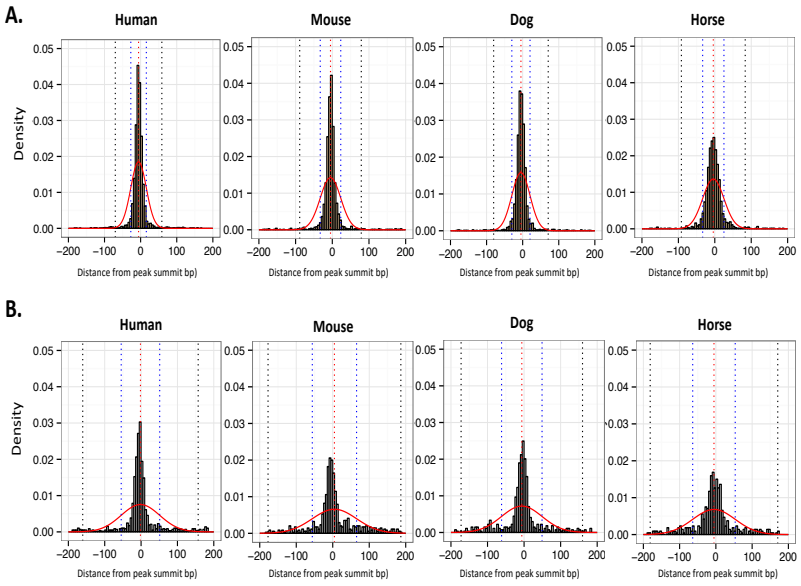


Figure 5.3. Histogram of distance between canonical motif **(A)** and non-canonical motif **(B)** to peak summit. Normal curve is shown in red; mean is shown in red dotted line; mean+s.d. and mean-s.d. are shown in blue dotted line; mean+3s.d. and mean-3s.d. are shown in black dotted line; bin=5bp.

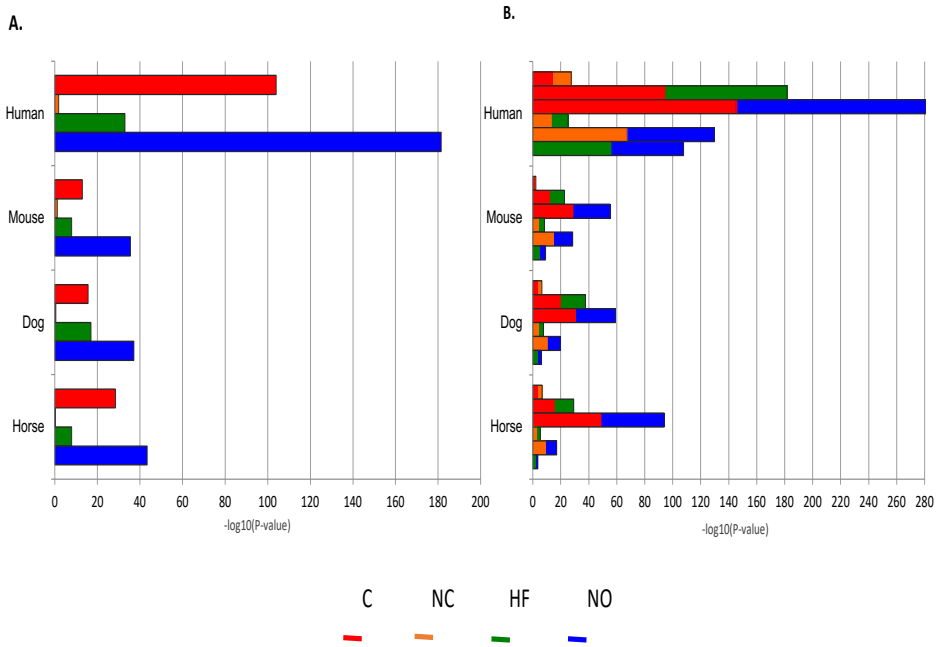


Figure 5.4. (A) Significance of canonical motifs enriched in high binding density sites while solo half-motifs and no-motif enriched in low density sites. Binding density is normalized to RPM (Mann-Whitney U test). **(B)** NRSF binding density divergents significantly with motif types. Binding density is normalized to RPM (Mann-Whitney U test).

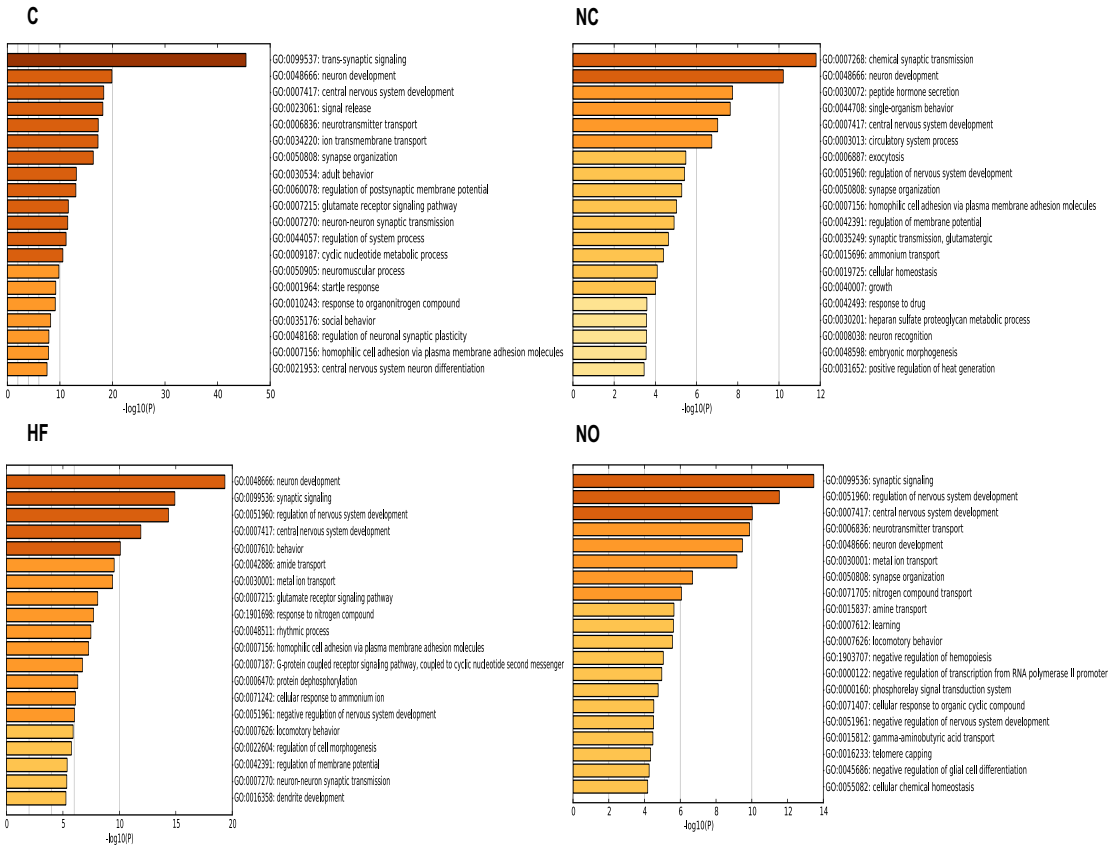


Figure 5.5. Gene ontology analysis on genes associated with peaks having canonical motifs (C), non-canonical motifs (NC), solo half-motifs (HF) and no motif (NO) in human (hypergeometric test p-value<0.05).

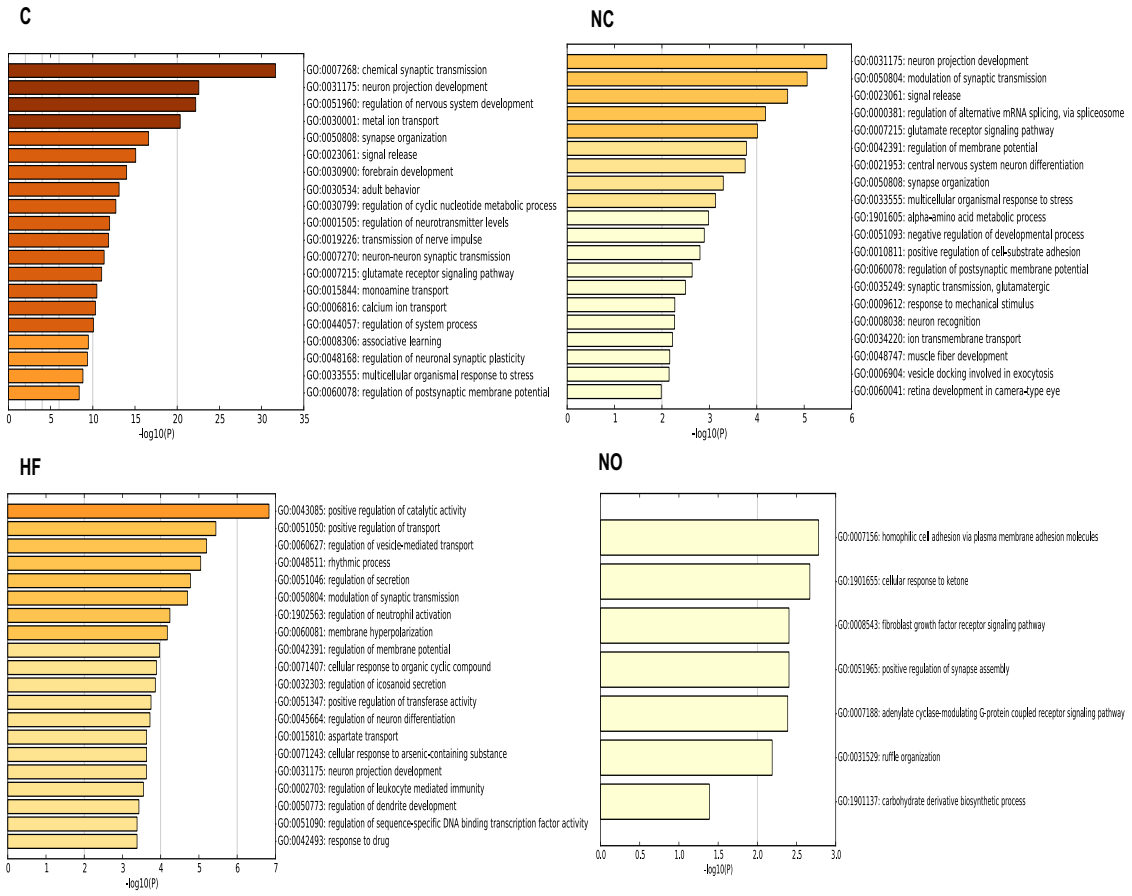


Figure 5.6. Gene ontology analysis on genes associated with peaks having canonical motifs (C), non-canonical motifs (NC), solo half-motifs (HF) and no motif (NO) in mouse. (hypergeometric test p-value<0.05).

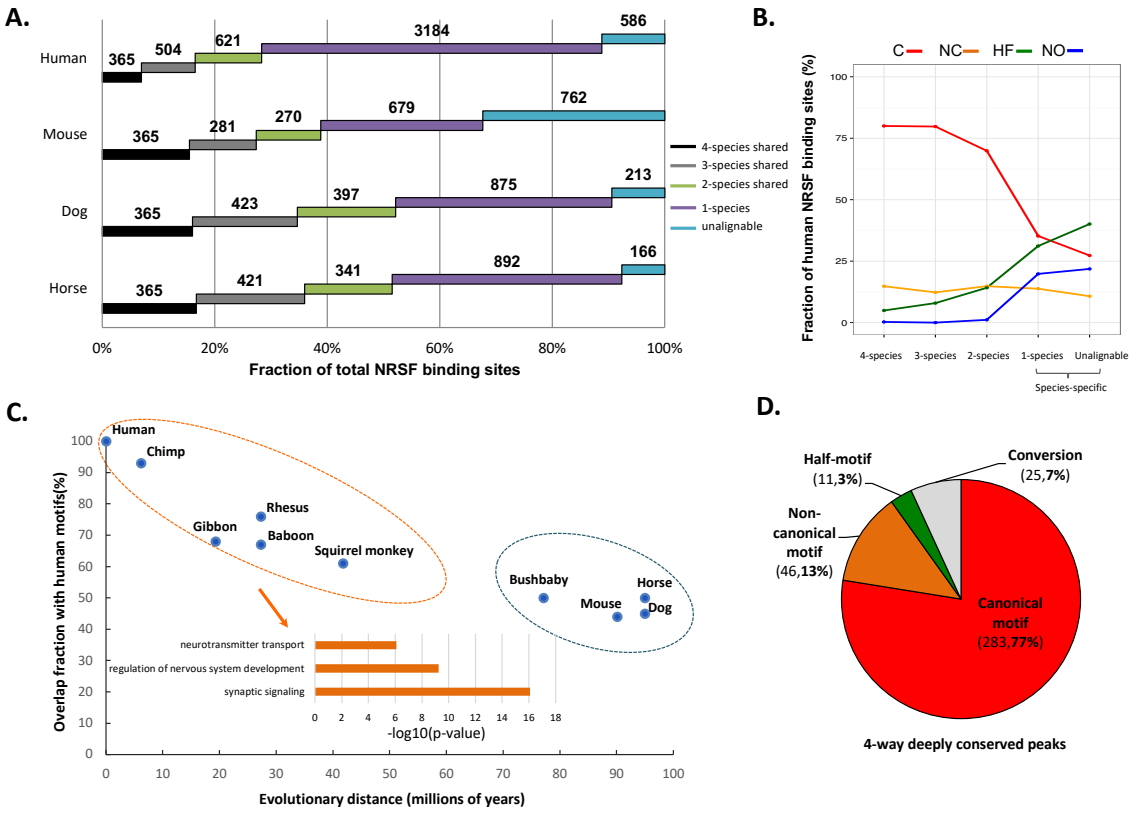


Figure 5.7. Divergence and conservation of NRSF binding across species. **(A)** Fraction of deeply conserved (4-species shared), partially conserved (3- and 2- species shared) and species-specific binding sites (1-species and unalignable) in each species. **(B)** Fraction of motifs in 4-species shared, 3-species shared, 2-species shared and species specific binding sites in human. (C, canonical motif, is composed of two 10bp half-motifs, separated by 11bp; NC, non-canonical motif, is composed of two 10bp half-motifs, separated by 10 or 16-19bp; HF, half-motif, is 10bp left or right half-motifs; NO, no motif). **(C)** The fraction of NRSF motifs overlapping with human motifs in each species. Evolutionary distances were adapted from TIMETREE (median) [47] **(D)** Fraction of peaks sharing the same type of motif in 365 deeply conserved binding sites. 'Conversion' are motif instances that have changed classes in orthologous regions in at least one of the four species.

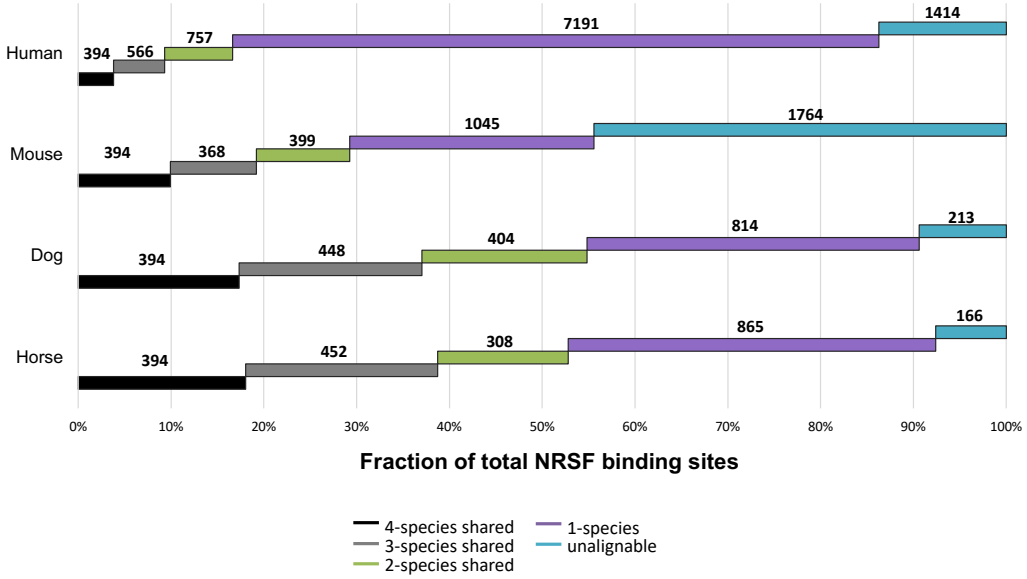


Figure 5.8. Fraction of deeply conserved (4-species shared), partially conserved (3- and 2-species shared) and species-specific binding sites (1-species and unalignable) in each species (cell lines for human and mouse are embryonic stem cells).

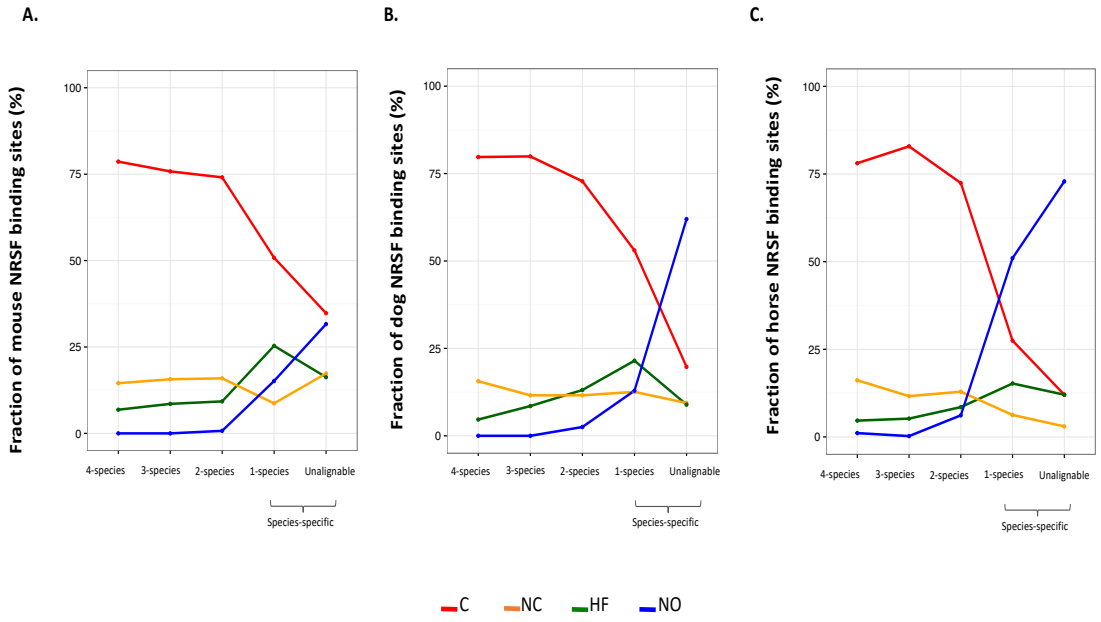


Figure 5.9. Fraction of motifs in 4-species shared, 3-species shared, 2-species shared and species specific binding sites in **(A)** mouse, **(B)** dog and **(C)** horse. C, canonical motif; NC, non-canonical motif; HF, solo half-motifs; NO, no motif.

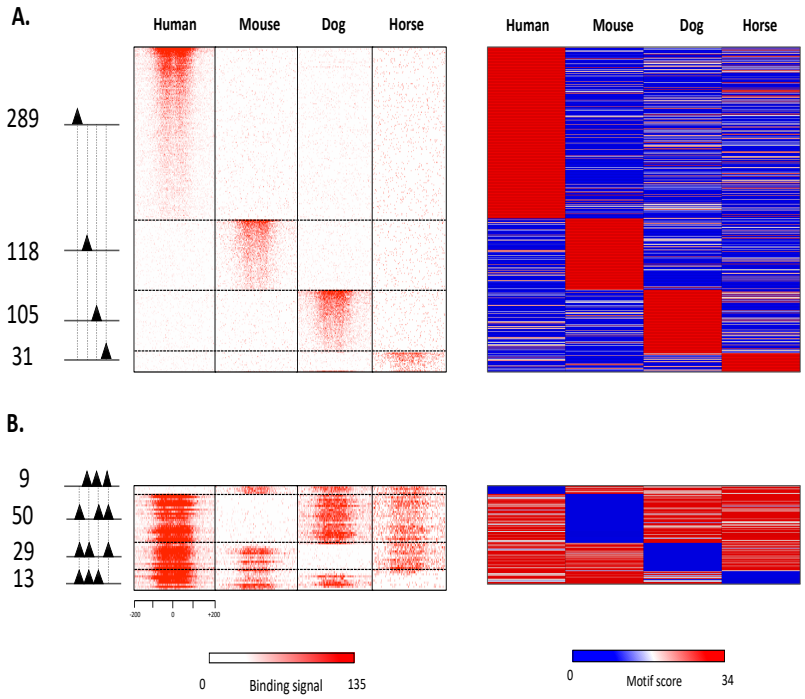


Figure 5.10. Birth and death of species-specific NRSF binding instances. **(A)** Heatmap of binding signal in 543 sites appeared ("born") in only one species with orthologous alignment (left). Heatmap of motif score in the corresponding sites (right). **(B)** Heatmap of binding signal in 101 orthologous regions where a site disappeared ("died") in only one genome (left). Heatmap of motif score in dead sites (right). Instances have strong NRSF binding are shown in red while weak binding shown in white. Species-specific newborn sites have strong NRSF motifs (shown in red) while dead sites have weak motif (shown in blue) in the affected species.

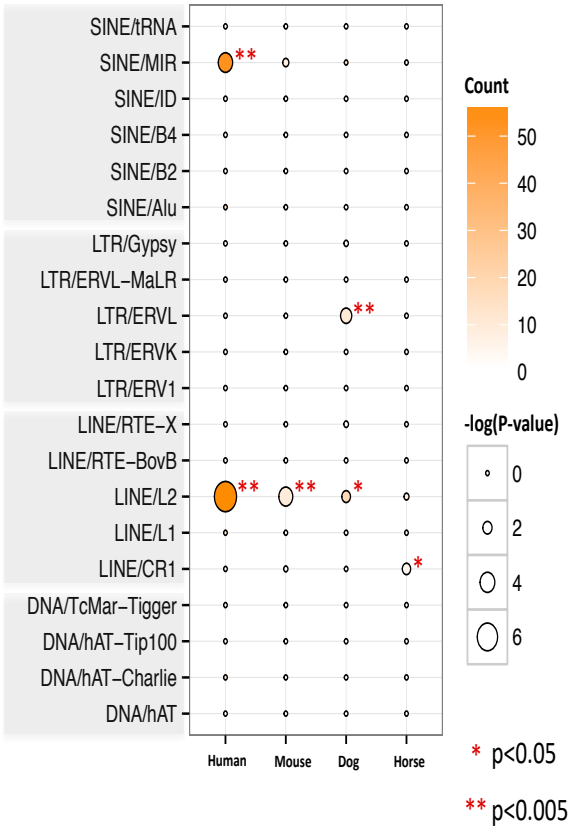


Figure 5.11. Transposable elements in species-specific NRSF birth sites. Significant NRSF birth instances associate with transposable elements families (SINEs, LTRs, LINEs and DNA transposons). Significance is evaluated using a binomial test corrected with Bonferroni correction.

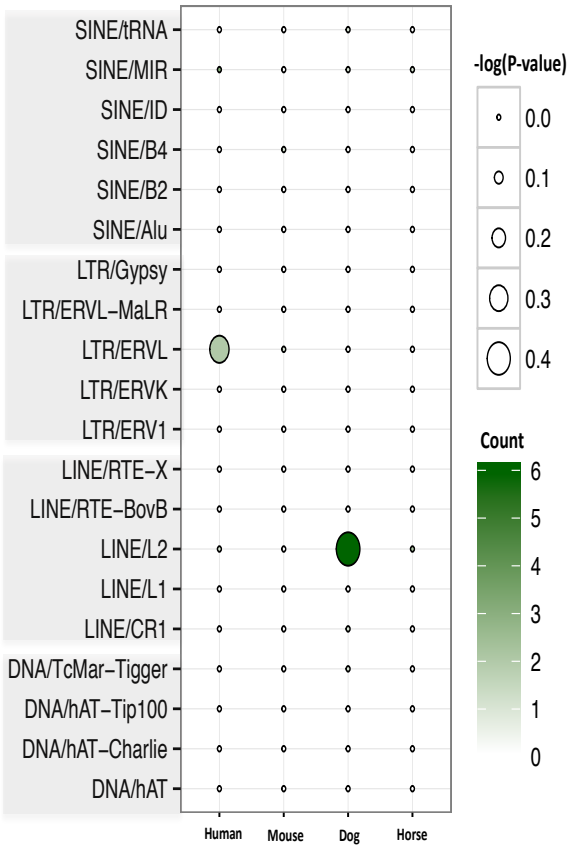


Figure 5.12. Transposable elements in species-specific NRSF death sites. Significant NRSF death instances associate with transposable elements families (SINEs, LTRs, LINEs and DNA transposons). Significance is evaluated using a binomial test corrected with Bonferroni correction.

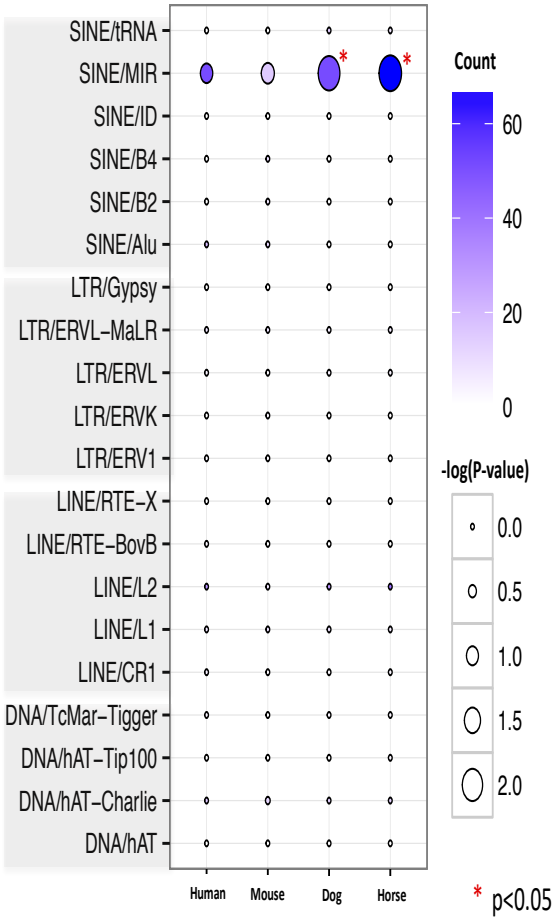
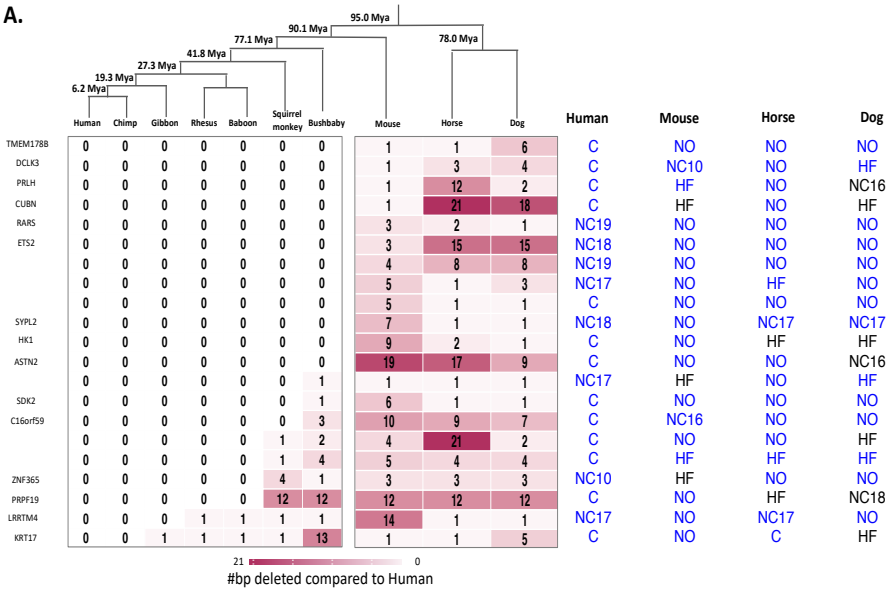


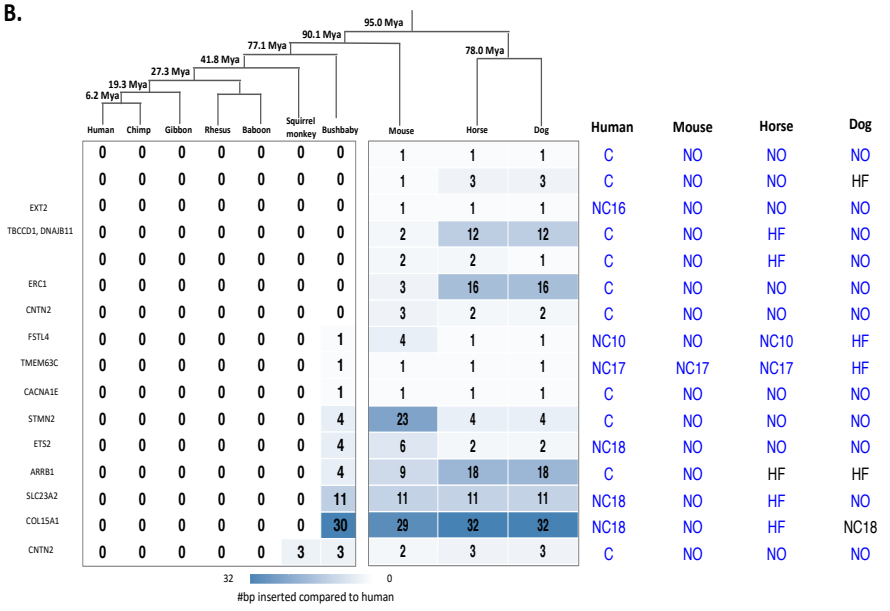
Figure 5.13. Transposable elements in deeply conserved NRSF binding sites. Significant conserved NRSF instances associate with transposable elements families (SINEs, LTRs, LINEs and DNA transposons). Significance is evaluated using a binomial test corrected with Bonferroni correction.

A.



	Human	Mouse	Horse	Dog
TMEM178B	C	NO	NO	NO
DCLK3	C	NC10	NO	HF
PRLH	C	HF	NO	NC16
CUBN	C	HF	NO	HF
RARS	NC19	NO	NO	NO
ETS2	NC18	NO	NO	NO
	NC19	NO	NO	NO
	NC17	NO	HF	NO
	C	NO	NO	NO
SYPL2	NC18	NO	NC17	NC17
HK1	C	NO	HF	HF
ASTN2	C	NO	NO	NC16
	NC17	HF	NO	HF
SDK2	C	NO	NO	NO
C16orf59	C	NC16	NO	NO
	C	NO	NO	HF
	C	HF	HF	HF
ZNF365	NC10	HF	NO	NO
PRPF19	C	NO	HF	NC18
LRRTM4	NC17	NO	NC17	NO
KRT17	C	NO	C	HF

B.



	Human	Mouse	Horse	Dog
	C	NO	NO	NO
	C	NO	NO	HF
EXT2	NC16	NO	NO	NO
TBCCD1, DNAJB11	C	NO	HF	NO
	C	NO	HF	NO
ERC1	C	NO	NO	NO
CNTN2	C	NO	NO	NO
FSTL4	NC10	NO	NC10	HF
TMEM63C	NC17	NC17	NC17	HF
CACNA1E	C	NO	NO	NO
STMN2	C	NO	NO	NO
ETS2	NC18	NO	NO	NO
ARRB1	C	NO	HF	HF
SLC23A2	NC18	NO	HF	NO
COL15A1	NC18	NO	HF	NC18
CNTN2	C	NO	NO	NO

Figure 5.14. Insertion and deletion are associated with NRSF birth in human. **(A)** 21 human birth sites are associated with primate specific insertion (left). Insertion associated motifs in human, mouse, dog and horse (right). Blue, motifs lie in insertion; White, motifs are not in insertion. C, canonical motif; NC, non-canonical motif; HF, solo half-motif; NO, no motif. **(B)** 16 human birth sites are associated with primate specific deletion (left). Deletion associated motifs in human, mouse, dog and horse (right). Blue, motifs lie in deletion; White, motifs are not in deletion. C, canonical motif; NC, non-canonical motif; HF, solo half-motif; NO, no motif.

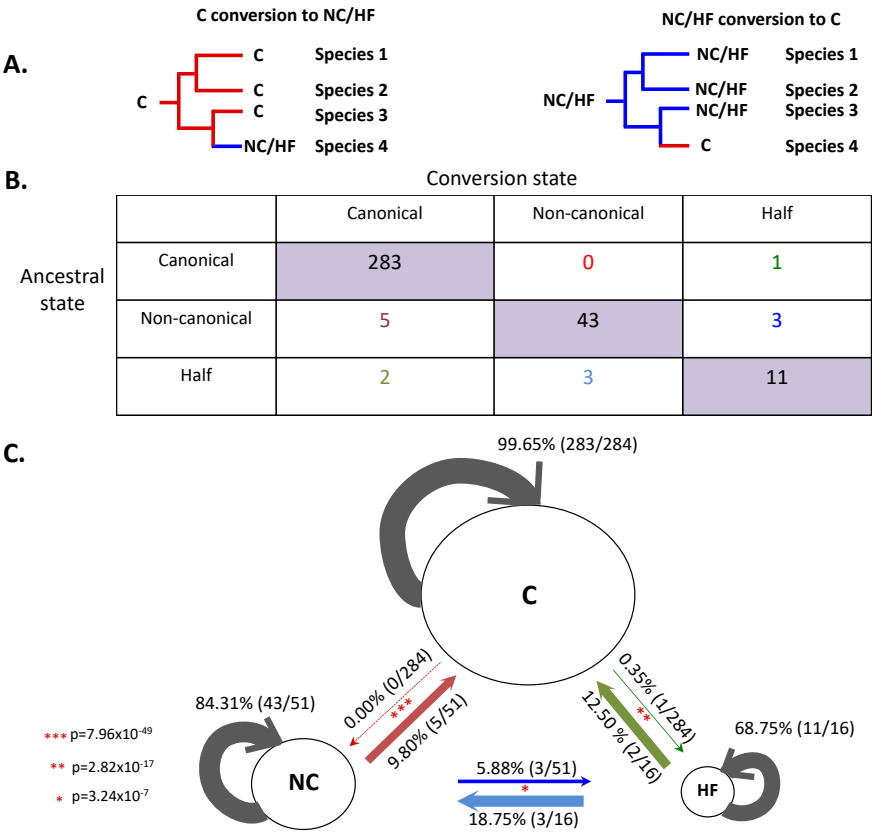


Figure 5.15. Motif conversion between canonical, non-canonical and half-motifs in deeply conserved NRSF binding sites. **(A)** Two examples show motif conversion between canonical motif and non-canonical motif/solo half-motif from ancestral state to conversion state. **(B)** Motif transition matrix shows conversion between canonical motifs, non-canonical motifs and half-motifs in 365 deeply conserved NRSF binding sites across four species. Conversion directions from ancestral state to the conversion state include: canonical motifs to non-canonical motifs (red); canonical motifs to half-motifs (green); non-canonical motifs to canonical motifs (dark red); non-canonical motifs to half-motifs (blue); half-motifs to canonical motifs (light green); half-motifs to non-canonical motifs (light blue). **(C)** Non-canonical motifs and half-motifs convert to canonical motifs at a significantly higher rate compared with reverse direction.

Conversion state

Ancestral state	Canonical	NC10	NC16	NC17	NC18	NC19	Half
Canonical	283	0	0	0	0	0	1
NC10	0	3	0	0	0	0	1
NC16	1	0	6	1	0	0	1
NC17	3	1	2	13	1	0	0
NC18	0	0	0	1	11	0	1
NC19	1	0	0	0	0	4	0
Half	2	2	0	1	0	0	11

Figure 5.16. Motif transition matrix shows conversion between canonical motifs, non-canonical motifs and solo half-motifs in deeply conserved NRSF binding sites across four species. Conversion directions from ancestral state to the conversion state include: canonical motifs to non-canonical motifs (red); canonical motifs to solo half-motifs (green); non-canonical motifs to canonical motifs (dark red); non-canonical motifs to solo half-motifs (blue); solo half-motifs to canonical motifs (light green); solo half-motifs to non-canonical motifs (light blue).

A.

Ancestral state \ Conversion state	Canonical	Non-canonical	Half
Canonical	365	2	7
Non-canonical	6	61	4
Half	2	7	28

B.

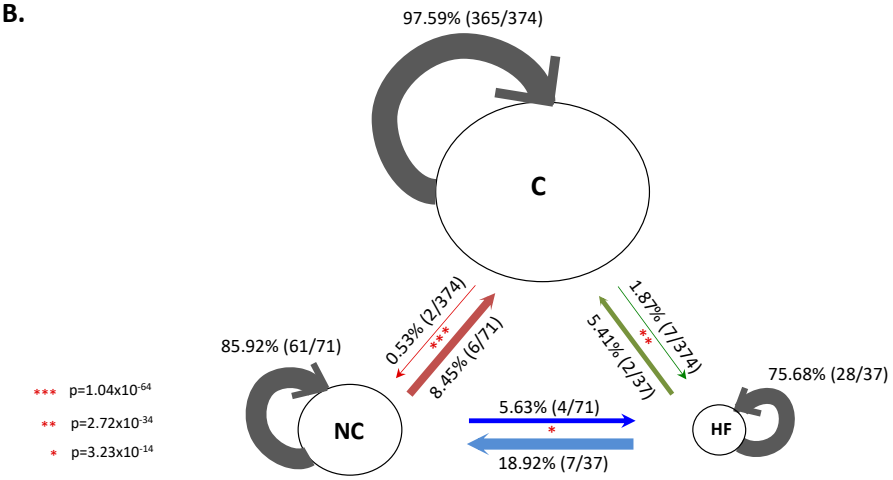


Figure 5.17. Motif conversion in conserved NRSF binding sites between human, mouse and dog. **(A)** Motif transition matrix shows conversion between canonical motifs, non-canonical motifs and solo half-motifs in conserved NRSF binding sites across human, mouse and dog. Conversion directions from ancestral state to the conversion state include: canonical motifs to non-canonical motifs (red); canonical motifs to solo half-motifs (green); non-canonical motifs to canonical motifs (dark red); non-canonical motifs to solo half-motifs (blue); solo half-motifs to canonical motifs (light green); solo half-motifs to non-canonical motifs (light blue). **(B)** Non-canonical motifs and solo half-motifs convert to canonical motifs at a significantly higher rate compared with reverse direction.

A.

Ancestral state \ Conversion state	Canonical	Non-canonical	Half
Canonical	303	0	3
Non-canonical	5	46	3
Half	3	3	12

B.

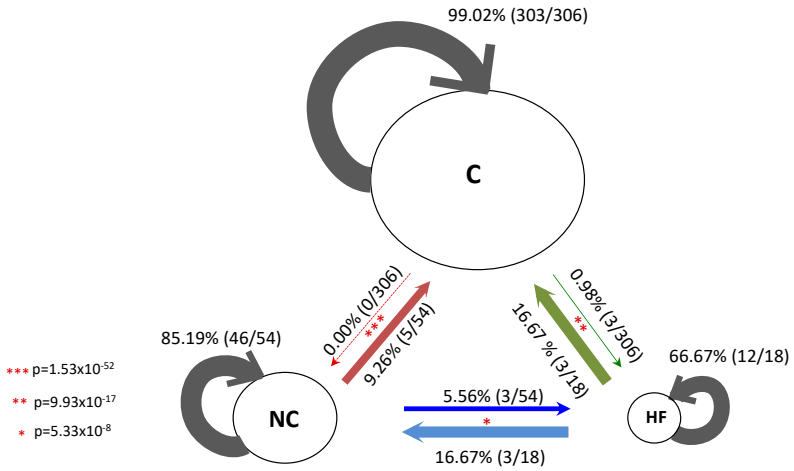


Figure 5.18. Motif conversion in conserved NRSF binding sites in four species by applying embryonic stem cells in human and mouse. **(A)** Motif transition matrix shows conversion between canonical motifs, non-canonical motifs and solo half-motifs in conserved NRSF binding sites across four species. Conversion directions from ancestral state to the conversion state include: canonical motifs to non-canonical motifs (red); canonical motifs to solo half-motifs (green); non-canonical motifs to canonical motifs (dark red); non-canonical motifs to solo half-motifs (blue); solo half-motifs to canonical motifs (light green); solo half-motifs to non-canonical motifs (light blue). **(B)** Non-canonical motifs and solo half-motifs convert to canonical motifs at a significantly higher rate compared with reverse direction.

Gene	birth/death	Function
ALK	hgbirth, mmdeath	Neuronal orphan receptor tryosine kinase
CAMTA1	hgbirth, hgbirth, mmbirth, cfbirth	Calmodulin binding transcription activator 1
CHST8	mmdeath, cfdeath	Carbohydrate sulfotransferase 8
CNR1	hgbirth, cfbirth	Cannabinoid Receptor 1
CNTN2	hgbirth, hgbirth	Axonal contactin 2
CRTAC1	mmbirth, mmdeath	Acidic secreted protein in cartilage
CTNNBL1	hgbirth, hgbirth, cfdeath	Catenin Beta Like 1
DCLK3	hgbirth, cfbirth	Doublecortin-Like Kinase 3
EXTL3	hgbirth, cfbirth	Exostosin Like Glycosyltransferase 3
FSTL4	hgbirth, mmbirth	Follistatin-like protein 4 Calcium ion binding protein
HIVEP3	hgbirth, hgbirth	Human Immunodeficiency Virus Type I Enhancer Binding Protein 3
IQSEC1	hgbirth, hgbirth	ADP-Ribosylation Factors Guanine Nucleotide-Exchange Protein 2
ISX	hgbirth, hgbirth	Pancreas-Intestine Homeodomain Transcription Factor
KCNIP1	hgbirth, cfdeath	Kv channel interacting protein 1
LRFN2	mmbirth, mmbirth	Synaptic Adhesion-Like Molecule 1
NEK11	mmbirth, hgdeath	Never in mitosis A (NimA) related kinase 11
NTM	mmbirth, eqdeath	Neurotrimin
PAX5	mmbirth, cfbirth, hgdeath	Paired box 5; B-cell lineage specific activator
SLC32A1	mmdeath, cfdeath	Solute Carrier Family 32 (GABA Vesicular Transporter), Member 1
SRRM4	mmbirth, mmbirth	Neural-Specific Serine/Arginine Repetitive Splicing Factor Of 100 KDa
TNR	hgbirth, eqbirth	Tenascin R

Table 5.1. 21 genes with multiple binding sites show site turnover in birth and death. These genes have at least two binding sites in at least one species.

5.6 Methods

5.6.1 Tissue culture

Mouse macrophage cells (RAW264.7, ATCC, Manassas, VA) were grown in DMEM medium (30-2002, ATCC, Manassas, VA) supplemented with 10% FBS (VWR, Radnor, PA) and penicillin/streptomycin (Life Technologies). Dog myelomonocytic leukemia cells (ML2, ATCC, Manassas, VA) were grown in RPMI-1640 medium (Life Technologies, Carlsbad, CA) supplemented with 1000 mM Hepes, 10mM MEM NEAA, 100mM Sodium pyruvate, 55mM β-mercaptoethanol, 10% FBS and penicillin/streptomycin. Horse skin fibroblast (E.Derm, ATCC, Manassas, VA) were grown in EMEM(30-2003, ATCC, Manassas, VA) supplemented with 10% FBS and penicillin/streptomycin. Trypan Blue (VWR) staining was performed before harvesting the cells for each assay to make sure that at least 90% of the cells were viable.

5.6.2 ChIP-seq

ChIP-seq experiments were performed on RAW264.7 (mouse, macrophage), ML2 (dog, myelomonocytic leukemia) and E.Derm (horse, skin fibroblast) following protocol adapted from Myer's Lab [48].We cross-linked about 20 million cells with 1% formaldehyde for each replicate. Cross-linked chromatin was sonicated into 200-300bp fragments. DNA fragments were incubated with Dynabeads protein G (Thermo Fisher Scientific) bound by NRSF monoclonal antibody (Caltech Protein Expression Center, Pasadena, CA, USA, 12C11-1B11) overnight to perform immunoprecipitation. Then chromatin was reverse cross-linked to extract NRSF-bound DNA. After performing real-time qPCR to check NRSF binding enrichment, DNA fragments recovered from ChIPs were end-repaired, ligated to adapters, size selected (200 to 300bp) and PCR-amplified to make libraries for sequencing. Each species had two NRSF ChIP replicates and one negative control, which were sequenced on Illumina Nextseq500 platform.

5.6.3 Published ChIP-seq experiments

The following published ChIP-seq data were used: Human HL60 (promyelocytic) and H1hesc (embryonic stem cell) NRSF ChIP-seq from HudsonAlpha ChIP-seq data sets,ENCODEproject.http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/ [34]. Mouse embryonic stem cell NRSF ChIP-seq data was downloaded from GSE27148.

5.6.4 Reads alignment and peak calling

ChIP and input sequencing reads were aligned using Bowtie v.0.12.8 [49] with parameters '-k 10 -m 10 --best –strata' to the following genome assemblies: human GRCh37/hg19, mouse GRCm38/mm10, dog CanFam3.1, horse EquCab2.0. After alignment, peaks were detected using MACS v.1.4.2 [50] with default parameters except ' –pvalue=1e-2 ', retaining all statistically enriched peaks ($p<10^{-4.5}$). Peaks were considered reproducible when they were identified in both replicates. Consensus peaks were then merged and called subpeaks using PeakSplitter_Cpp v.1.0 [51] with parameters '-c 10 -f -n 0 -l 0'. Subpeaks in 50bp were merged. Peaks having (1) higher signal in control than ChIP replicates in each species or (2) signal in control is higher than 1.4 RPM in each species or (3) in Blacklist [34] in human and mouse were filtered out as artificially high signal peaks.

5.6.5 Motif calling

Known canonical RE1 motif PSFM (position specific frequency matrix) [52] was divided into PSFMs for two half-motifs, i.e. RE1-left, from position 1 to 10 and RE1-right, from position 12-21. PSFMs for two half-motifs were used to scan within 200bp sequences centered on peak summit using MEME v.4.9.0_4 [53] with parameters 'fimo --thresh 5e-3 --parse-genomic-coord' respectively. Half-motifs have >50% motif occurrence score were accepted to detect canonical and non-canonical NRSF motifs. In order to find both types, we defined the distance between two half-motifs is between position 6 of RE1-left and position 16 of RE1-right (inclusive). This gives an 11bp distance in canonical RE1 motif and 10, 16-19bp distance in non-canonical RE1 motifs. Any canonical or non-canonical RE1 motifs were accepted for each peak. Motifs only with left half-motif or right half-motif were accepted if they were >60% motif occurrence score and located within 50bp sequences centered on peak summit. For interspecies analysis, identified motifs were further performed with unambiguous motif analysis in peaks. We used the following procedure to confirm only one motif in each peak: (1) if canonical RE1 lies in peak, the one with highest occurrence score was accepted; (2) if non-canonical RE1 lies in peak but no canonical RE1, the highest score non-canonical RE1 was accepted; (3) if peak only has half-motifs, the highest score half-motifs within 50bp centered around peak summit was accepted.

5.6.6 Interspecies analysis

Interspecies pairwise comparisons were performed by aligned identified binding instances between species in a reciprocal manner using UCSC liftOver [54] on genomic assemblies: human GRCh37, mouse GRCm38, dog CanFam3.1 and horse EquCab2.0. Each of the species was used as anchor species and the binding instances were mapped to the other three species with 50% minimum map ratio. Instances failing to be mapped in any of the other genomes were considered as unaligned instances and not used in comparisons. To identify conserved instances between two species, binding instances having orthologous regions overlapped a binding instance in the second species with at least 1bp were collected. Final pairwise conserved instances were confirmed by doing this comparison reciprocally, including human-mouse, human-dog, human-horse, mouse-dog, mouse-horse and dog-horse. Conserved binding instances shared in four species were collected as 4-species shared peaks. Conserved instances shared in any three species but not in 4-species shared peaks set were collected as 3-species shared peaks. Conserved instances shared between any two species but not in 4-species and 3-species shared peaks sets were considered as 2-species shared peaks. For each species, binding instances successfully aligning to any of the other three genomes but not overlapping with identified binding instances were collected as 1-species peaks. However, binding instances failed to align to all of the other three genomes were collected as unalignable peaks.

5.6.7 Birth and death of species-specific NRSF binding instances

Species-specific birth instances were identified in 1-species peaks set for each species. Peaks successfully aligning to all the other three genomes were collected as candidate list for further analysis: (1) candidate peaks were accepted when aligned regions in other genomes mapped to each other and overlap with minimum 1bp reciprocally. (2) candidate peaks were accepted when canonical or non-canonical RE1 were in target genome region and motif occurrence score was higher than the orthologous ones in the other three genomes. (3) peaks were further collected when mapping aligned regions in other genomes back to the target genome and overlap with candidate peaks. (4) peaks were filtered out when RPM of aligned regions in other species is higher than 1.

Species-specific death instances were identified in 3-species shared peaks set for each species (e.g. human death instances were identified from mouse, dog and horse shared peaks set). Peaks

from the other three genomes successfully aligning to the target genome were collected as candidate list for further analysis: (1) candidate death instances were accepted when canonical or non-canonical RE1 were in the other three genomes and motif occurrence scores were higher than the orthologous one in target genome. (2) death instances were further collected when mapping region of the target genome back to the other three genomes and overlap with peaks. (4) instances were further filtered out when RPM of death instances in the target species is higher than 1.

5.6.8 Transposable elements association analysis

Repeat elements associated with birth and death instances were detected using RepeatMasker [37] annotation in each species: human GRCh37/hg19, mouse GRCm38/mm10, dog CanFam3.1/canFam3 and horse EquCab2.0/equCab2. Enrichment of repeat element families was evaluated using a binomial test adjusted by bonferroni correction with all experimental defined peaks for each species as random background.

5.6.9 Indels association analysis

Indels associated with human birth instances were detected using MAF files from hg19/GRCh37, multiz100way alignment (UCSC genome browser). Selected species include Human (GRCh37/hg19), Chimp (CSAC2.1.4/panTro4), Gibbon (GGSC Nleu3.0/nomLeu3), Rhesus macaque (BGI CR_1.0/rheMac3), Baboon (Baylor Pham_1.0/papHam1), Squirrel monkey (Broad/saiBol1), Bushbaby (Broad/otoGar3), Mouse (GRCm38/mm10), Dog (Broad CanFam3.1/canFam3) and Horse (Broad/equCab2). Insertions were identified if bases were missed/deleted in other species compared to human birth sites in orthologous regions. Deletions were identified bases were inserted in the other species compared to human birth sites in orthologous regions. Insertions and deletions were considered as associated with birth instances if they were in the motifs of birth sites.

5.6.10 Associated genes and gene ontology analysis

Gene annotations from Ensembl [55] were used to associate genes with peaks. Peaks were annotated by using annotatePeaks.pl with default parameters in HOMER [56]. Gene ontology analyses were performed with Metascape [57] with hypergeometric test p-value lower than 0.05.

5.6.11 Motif conversion analysis

Motif conversion analysis was performed on deeply conserved (4-species shared) peaks. We define six types of motif conversion, including canonical motif to non-canonical motifs or half-motifs, non-canonical motifs to canonical motifs or half-motifs and half-motifs to canonical or non-canonical motifs (e.g. canonical motif converting to non-canonical motif if canonical motifs in three of four species but non-canonical motif in the 4th species) and summed all of the noncanonical conversions. The significance of each motif conversion was evaluated by performing chi-squared tests corrected by Fisher's exact test.

5.7 References

1. Davidson EH. 2006. *The regulatory genome: gene regulatory networks in development and evolution*. Academic press.
2. Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**: 206–216.
3. Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**: 786.
4. Shim S, Kwan KY, Li M, Lefebvre V, Šestan N. 2012. Cis-regulatory control of corticospinal system development and evolution. *Nature* **486**：74-79.
5. Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. 2015. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**: 68–83.
6. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252–263.
7. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.
8. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**: 730–732.

9. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavaré S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**: 434–438.
10. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–40.
11. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves,Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.
12. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**: 530–540.
13. Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* **15**: 221–33.
14. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**: 365–370.
15. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res* **16**: 669–677.
16. Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, Guo AY. 2012. AnimalTFDB: A comprehensive animal transcription factor database. *Nucleic Acids Res* **40**: 144–149.
17. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**: 242–5.
18. Schoenherr CJ, Anderson DJ. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**: 1360–1363.

19. Chong JA, Tapia-Ramírez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G. 1995. REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**: 949–957.
20. Chen ZF, Paquette a J, Anderson DJ. 1998. NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat Genet* **20**: 136–142.
21. Ballas N, Grunseich C, Lu DD, Speh JC, Mandel G. 2005. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* **121**: 645–657.
22. Coulson JM. 2005. Transcriptional regulation: Cancer, neurons and the REST. *Curr Biol* **15**: 668–670.
23. Schoenherr CJ, Paquette AJ, Anderson DJ. 1996. Identification of potential target genes for the neuron-restrictive silencer factor. *Proc Natl Acad Sci U S A* **93**: 9881–9886.
24. Andrés ME, Burger C, Peral-Rubio MJ, Battaglioli E, Anderson ME, Grimes J, Dallman J, Ballas N, Mandel G. 1999. CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proc Natl Acad Sci U S A* **96**: 9873–9878.
25. Grimes JA., Nielsen SJ, Battaglioli E, Miska EA, Speh JC, Berry DL, Atouf F, Holdener BC, Mandel G, Kouzarides T. 2000. The co-repressor mSin3A is a functional component of the REST-CoREST repressor complex. *J Biol Chem* **275**: 9461–9467.
26. Roopra A, Sharling L, Wood IC, Briggs T, Bachfischer U, Paquette AJ, Buckley NJ. 2000. Transcriptional repression by neuron-restrictive silencer factor is mediated via the Sin3-histone deacetylase complex. *Mol Cell Biol* **20**: 2147–57.
27. Battaglioli E, Andrés ME, Rose DW, Chenoweth JG, Rosenfeld MG, Anderson ME, Mandel G. 2002. REST repression of neuronal genes requires components of the hSWI.SNF complex. *J Biol Chem* **277**: 41038–45.
28. Otto SJ, McCorkle SR, Hover J, Conaco C, Han JJ, Impey S, Yochum GS, Dunn JJ, Goodman RH, Mandel G. 2007. A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J Neurosci* **27**: 6729–6739.
29. Rockowitz S, Lien WH, Pedrosa E, Wei G, Lin M, Zhao K, Lachman HM, Fuchs E, Zheng D. 2014. Comparison of REST Cistromes across Human Cell Types Reveals.

30. Rockowitz S, Zheng D. 2015. Significant expansion of the REST/NRSF cistrome in human versus mouse embryonic stem cells: potential implications for neural development. *Nucleic Acids Res*: gkv514. doi: 10.1093/nar/gkv514.
31. Mortazavi A, Thompson ECL, Garcia ST, Myers RM, Wold B. 2006. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res* **16**: 1208–1221.
32. Johnson R, Samuel J, Ng CKL, Jauch R, Stanton LW, Wood IC. 2009. Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor REST. *Mol Biol Evol* **26**: 1491–1507.
33. Saritas-Yildirim B, Childers CP, Elsik CG, Silva EM. 2015. Identification of REST targets in the Xenopus tropicalis genome. *BMC Genomics* **16**: 380.
34. The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
35. Arnold P, Schöler A, Pachkov M, Balwierz PJ, Jørgensen H, Stadler MB, Van Nimwegen E, Schübeler D. 2013. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res* **23**: 60–73.
36. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**: 1798–1812.
37. Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org*.
38. Villar D, Berthelot C, Flicek P, Odom DT, Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160**: 554–566.
39. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–4.
40. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24***: 1963-76.*

41. Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* **34**: 3862–3877.
42. Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. *Mol Biol Evol* **18**: 1764–70.
43. Smit A, Riggs A. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res* **23**: 98–102.
44. Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**: 657–663.
45. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**: 207–19.
46. Smeenk L, Van Heeringen SJ, Koeppel M, Van Driel MA, Bartels SJJ, Akkers RC, Denissov S, Stunnenberg HG, Lohrum M. 2008. Characterization of genome-wide p53-binding sites upon stress response. *Nucleic Acids Res* **36**: 3639–3654.
47. Hedges SB, Dudley J, Kumar S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* **22**: 2971–2972.
48. Pauli F, Myers RM. 2011. Myers Lab ChIP-seq Protocol. http://myers.hudsonalpha.org/documents/Myers%20Lab%20ChIP-seq%20Protocol%20v042211.pdf
49. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
50. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.
51. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. 2010. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11**: 415.
52. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and

updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142–7.

53. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res* **37**: 202–208.
54. Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161.
55. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. Ensembl 2016. *Nucleic Acids Res* **44**: D710–D716.
56. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.
57. Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, Moulton HM, Dejesus P, Che J, Mulder LCF, et al. 2015. Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding. *Cell Host Microbe* **18**: 723–735.

Chapter 6

Future directions

Chapter 6
Future directions

In this thesis, I have explored the versatility of transcriptional and epigenetic control during development at both cellular and whole-individual level and how it is affected in disease as well as how it changes between species.

In Chapter 2, I show that DNA methylation changes can be used to distinguish individual rats with different early life experiences. This discovery benefits from the novel strategy we used to examine individual differences, in which we identify environmental effects by looking at intra-individual differences. Epigenetic changes within a short period of time in the same individual that is affected by maternal care leads to long term consequences for the behavior of adult rats. Several studies have shown that DNA methylation can be used as a predictor of tissue aging in human body. We find that rats with different early life experiences cannot be distinguished by only comparing methylation changes by comparing them as two homogeneous groups. Instead these methylation changes can distinguish younger individuals from older ones even within a week of postnatal development. However, the intra-individual differences of these methylation changes, i.e methylation level changes between two time points for the same individual, can be a powerful signature to distinguish individuals with adverse experiences. Based on this, we expect that this identified DNA signature is commonly shared in rat individuals and should be predictive if it is robust. To achieve this goal, we have to increase the number of individuals by generating additional cohorts of rats (For now we only have 2 cohorts with 11 and 8 individuals respectively) and check whether the same signature are shared across all cohorts. We could then apply machine learning methods, such as support vector machines and logistic regression, to train on the original cohorts as training sets to extract important features of different experiences and then test the prediction on rats from the new and validation cohorts. The number of differential methylation regions may be reduced after training and predicting with machine learning methods that can hopefully identify key DNA regions that could suggest how methylation changes would influence downstream target gene expression and further affect the maturation of brain and other organs in rats. We have observed high enrichment of transcription factors on these differential methylation regions, indicating that these signature regions regulate

target gene expression in response to different early life experiences. For the future collection of cohorts, we could collect RNA and DNA methylation at the same time in order to examine how DNA methylation regulates gene expression within the same pool of cells. While it is interesting that rodents show these pronounced epigenetic changes in response to early life experiences, it would be exciting if this DNA methylation signature of early life maternal also applied to human newborns. In this chapter, we extract DNA methylation from buccal swap samples and it has been shown that methylation patterns in buccal swab samples are actually closer to those in hippocampus than blood and other tissues. Buccal swabs is the least harmful method to extract DNA samples from human newborns, and thus our approaches was designed up front to be also applied in humans. However, it may be more difficult to extract DNA signature from human than rat newborns because the early life experiences in human are more complicated. While controlling the bedding condition is the only factor that influences the maternal care passing from rat mothers to newborns, many other factors may influence the caring conditions in the human case, such as food, living condition, and the level of mothers' social involvement. Thus, a scoring system to evaluate each environmental factor is required for a human study and key features with highest variances can be also extracted by using computational methods. Given the complexity of the comparable human study, it will be necessary to collect more samples (n>100) to extract DNA and evaluate the predictive power of the differential methylation regions. It will be interesting to find whether a similar DNA methylation signature can be recovered in human samples and further distinguish human newborns with different early experiences. We expect to observe not only a conserved set of target genes regulated by these differential methylation regions in both human and rodent but a set of human-specific genes that are influenced by the adversity experiences drive the development of brain and other organs in human babies.

In Chapter 3, I present the first detection of DUX4 positive nuclei in FSHD2 myotube using single nucleus RNA-seq. These DUX4 positive nuclei share similar target gene expression to a larger set of nuclei with no DUX4 expression and further more than not all DUX4 and target genes are not always expressed together within the same nuclei. These results agree with a proposed model that DUX4 transcripts diffuse into cytoplasm and protein imported into adjacent nuclei to activate target gene expression. Then target genes that are TFs may also be imported into other nuclei and induce more "indirect" DUX4 target gene expression. Furthermore,

although only 2.1% of FSHD2 myotube nuclei express DUX4, we detect about 15% of FSHD2 myotube nuclei express high level of DUXA (not in control), a paralog gene of DUX4 in human, and they are not co-expressed with DUX4 within the same nuclei. The function of DUXA in FSHD and its interaction with DUX4 are still unclear, but we hypothesize that DUXA may be one substitute disease-causing gene of DUX4 in those DUX4 negative FSHD myotube nuclei. However, the low percentage of DUX4 positive FSHD2 myotube nuclei may be also caused by the incomplete annotation of DUX4 gene in reference genome and therefore, reads may be mis-aligned to other loci. Recent advances of the third-generation sequencing allow us to sequence genes with full-length transcripts and improve gene annotation without bias of assembling from short fragments. We can sequence FSHD2 myotube with third generation sequencing platforms, such as PacBio and Oxford Nanopore, to capture full-length transcripts not only for DUX4 but also for other interests genes like DUXA and DUX4 targets. By comparing with illumina annotated reference transcriptome, we can build a FSHD2-specific reference transcriptome with these updated transcripts and then use it to refresh the alignment of myotube nuclei to get more accurate expression levels of genes in FSHD pathogenesis. We also find that high DUX4 expression may only be expressed at a precise stage during the maturation of FSHD myotubes. We hypothesize that there is a specific window of maturation of myotube and if FSHD myotube survive this critical window would have no DUX4 up-regulated later on; otherwise, developmental defects would happen to FSHD myotube. Experiments are needed to test this hypothesis by examining DUX4 and DUX4 target expression in a further differentiated myotubes and the regulation between DUX4 and myogenic TFs needs careful investigation. One of the most straightforward methods is to check myogenic TFs binding around DUX4 and vice versa. Interestingly, we also detect very high expression of the myoblast marker, desmin, in DUX4 positive nuclei no matter the expression level of myogenic markers such as CKM and myogenin. To understand the function of desmin in DUX4 activation, RNA FISH and knockdown experiments are required to check how DUX4, myogenic markers and desmin regulate each other. We also identify a specific set of TFs that show significantly higher expression in DUX4 positive target positive nuclei and they could be further tested to confirm whether they are novel DUX4 downstream genes. Although DUX4 expression is detected in individual myotube nuclei, we fail to detect it at a substantial level in pooled RNA-seq data. One of the possibility is that DUX4 expression is transient and we missed the right time-point to

capture it or that the transcript half-time is short. Thus, more experiments are necessary for understanding the dynamics of DUX4 expression effectively. For examples, we can apply single molecule fluorescence in situ hybridization (smFISH) to quantitate transcription and post-transcription at the same time. In order to strengthen our conclusions, we should repeat our myoblast differentiation from other FSHD2 patients and examine whether they display similar DUX4 and targets expression patterns. Future studies should also focus on FSHD1, the major type of FSHD disease. It has been known that DUX4 is upregulated by different mechanisms in FSHD1 and 2, but DUX4 associated gene regulation is unclear in both. We hope that our understanding can be expanded by examining and comparing DUX4 related regulation in FSHD1 and FSHD2.

In Chapter 4, I study the role of epigenetic regulation during embryonic stem cell differentiation caused by changes in chromatin accessibility and its effect on gene regulation. I present the comprehensive view of gene expression and chromatin accessibility during endoderm differentiation in three mammalian species. Using open chromatin footprinting, we construct GRNs to understand gene regulatory networks controlling endoderm development. Although many TFs are involved in endoderm development and some of them are known to regulate each other, the direct targets of TFs and layer of gene regulation are still poorly understood in mammalian species. I have recovered many validated regulatory interactions based on the literature in other species in our GRNs but we still find more interactions that are novel and species-specific. These results may be caused by the limitations of sequencing depth in our current analysis or computation approaches, which may generate false-positive connections. Thus, experimental validation are necessary to refine the current version of GRNs. Experiments used for validating GRNs including ChIP-seq on specific TFs and perturbing GRNs by knock-down the same TFs. Another limitation of our GRNs is that it is built based on selected TFs and genes, and thus we did not attempt to find novel TFs in endodermal development. To build GRNs automatically, we can adapt machine learning methods, such as decision trees, to build predictive model from our footprints results. In brief, we can build connections between TFs and target genes by scanning all footprints with motif database. Then a subset of these connections can be trained by decision tree to extract features of mostly enriched TFs and their target genes. To improve the accuracy, many rounds of training are needed and subset of connections is

randomly picked every time. By using the trained model, we can predict which TFs are more likely to activate endoderm differentiation and which ones tend to be repressive. One of the biggest advantages of our cross-species system is that we can train the model by using the conserved connections across species as those connections may be more representative to show common features of important gene regulation in endoderm lineage. One of the highlights in this study is to compare the conservation level of GRNs at the same stage across mammalian species. We show a higher conservation level in gene expression than chromatin accessibility during endoderm differentiation between species, indicating that a conserved set of genes are required for mammalian endoderm differentiation but they may be regulated by divergent sets of cis-regulatory elements and corresponding TFs in different species. Previous studies have shown rapid turnover of cis-regulatory elements in orthologous regions between mammalian species. By only focusing on GRNs of stem cell and endoderm stages, we also show more extensive rewiring of GRNs in endoderm than the ones in stem cell stage. A technical explanation for this result may be the use of different media during endoderm differentiation in human and rodents. Although each medium can drive embryonic stem cells into endoderm commitment, the length of differentiation reflects that they activate different pathways during endoderm formation between human and rodents. In order to characterize this, we could try to apply human differentiation medium in rodents and vice versa. We know that the medium are not interchangeable between species as cells die in the other species' medium before they reach the definitive endoderm stage. However, we can examine the changes in expression and chromatin accessibility profiles during the first 2 or 3 day of differentiation and compare them with the data we have collected. Another way to understand the effect of the medium is to introduce another primate species into our comparison system by differentiating in the same medium as the one used in human. In this way, we can first extract the conserved regulation within primate and rodent groups respectively and then further extract divergent regulation by only comparing the conserved modules between primates and rodents. Another advantage of the fourth species is that we can identify which regulatory linkages are more ancestral (shared in three species) or more recently evolved (only found in one species) and further identify the core, necessary regulatory linkages for DE differentiation in mammalian species.

In Chapter 5, I propose a motif conversion model to explain the prevalence of canonical motifs for NRSF binding in mammalian genome. Although the repression of neuronal genes is conserved in mammalian species, NRSF binding is not always located at the same position for genes in different species. Our results reveal the rapid NRSF binding sites turnover in four species and that the canonical motif shows significantly high enrichment in conserved binding regions compared with other motif forms. By comparing NRSF ChIP-seq in four species, we can identify the ancestral motif form found in three of four species for each NRSF binding site and then observe the motif conversion from ancestral state to conversion state. We calculated a significantly higher conversion rate from non-canonical and half-motifs to canonical motifs, which explain the accumulation of canonical form for NRSF binding in genome. Although NRSF is known to repress neuronal genes in non-neural cells, cell types used in this study are not consistent in four species (immune cells for human, mouse and dog but skin fibroblast for horse) and this inconsistency may result in the divergent binding shown in this study. Besides, this cell type inconsistency may also restrict the number of deeply conserved binding regions in four species and further introduce bias into the motif conversion model. In a future study, the comparison system should be updated to use the same cell types, such as the four species stem cell comparison system in Chapter 4, in order to evaluate the reproducibility of the motif conversion model. Similar as NRSF, other TFs like CTCF and TP53 also have different form of binding motifs and it would be worthwhile to examine our motif conversion model in these TFs to see whether this motif selection is a common phenomenon during mammalian evolution. This motif conversion strategy may be related to the effectiveness of TFs regulation on target genes. Thus, further experiments are necessary to examine the difference of those motif forms in mediating TFs regulation. NRSF is known to regulate target genes through a series epigenetic control, including methylation, co-factor binding and histone modification. Future studies should also focus on checking how motif selection is related to this regulation by integrating with other functional genomics assays, such as RNA-seq and BS-seq.