UNIVERSITY OF CALIFORNIA

Los Angeles

Mining and Integrating Epigenomics
Big Data to Discover Novel
Mechanisms of Gene Regulation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Bioinformatics

by

Kai Fu

2018

ABSTRACT OF THE DISSERTATION

Mining and Integrating Epigenomics
Big Data to Discover Novel
Mechanisms of Gene Regulation

by

Kai Fu
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Matteo Pellegrini, Chair

Besides DNA sequences, the genes are regulated by epigenomic mechanisms. Advances in high-throughput sequencing technologies have enabled the generation of huge amount of epigenomic data sets. Those epigenomic big data then requires the application of sophisticated computational approaches and statistical algorithms. My dissertation then focuses on mining and integrating epigenomic big data to inform novel biological mechanisms behind those datasets. The first research project compares the binding patterns of pluripotent regulatory factors, i.e. Oct4, Sox2, Klf4 and c-Myc, between human and mouse in induced pluripotent stem cells. The result suggests the genome-wide regulatory mechanisms are conserved between those two species, but the detailed transcriptional mechanisms are diverged. The second research project analyzes the temporal expression data from embryonic stem cells to cardiomyocytes. The results in this project then identify regulators, including transcription factors and long intergenic non-coding RNAs, which are strongly associated with the cardiogenesis differentiation process. The third research project integrates datasets of DNA methylation and histone modification in 35 human cell types. The result shows histone modifications, especially for H3K4me3, are highly predictable of DNA methylation. As a summary, my dissertation analyzes and integrates epigenomic big data in biological context related with embryonic stem cells and induced pluripotent cells, provides and discoveries novel insights to understand epigenetic regulation of gene expression.

The dissertation of Kai Fu is approved.

Yi Xing

Jason Ernst

Kathrin Plath

Atsushi Nakano

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2018

To my father and mother,

Who give me the support and courage to achieve my Ph.D.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Chapter contributions

Chapter 2 is a version of an article has been submitted for publication:

Comparison of the binding of pluripotency factors between human and mouse in early iPSCs reprogramming

Kai Fu[1], Constantinos Chronis[2], Abdenour Soufi[3 , 5], Giancarlo Bonora[1], Miguel Edwards[4], Steve Smale[4], Kenneth S. Zaret[3], Kathrin Plath[2], and Matteo Pellegrini[1*]

[*]Correspondence: matteop@gmail.com

[1] University of California Los Angeles, Department of Molecular, Cellular and Developmental Biology, Bioinformatics Interdepartmental Program, Los Angeles, CA 90095, USA

[2] University of California Los Angeles, David Geffen School of Medicine, Department of Biological Chemistry, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, Molecular Biology Institute, Bioinformatics Interdepartmental Program, Los Angeles, CA 90095, USA

[3] University of Pennsylvania Perelman School of Medicine, Institute for Regenerative Medicine and Epigenetics Program, Department of Cell and Developmental Biology, Smilow Center for Translational Research, Philadelphia, PA 19104, USA

[4] University of California Los Angeles, Department of Microbiology, Immunology and Molecular Genetics, Los Angeles, CA 90095, USA

[5] Present address: University of Edinburgh, MRC Centre for Regenerative Medicine, SCRM Building, Edinburgh Bioquarter, Edinburgh, EH16 4UU, UK

Chapter 3 is a version of manuscript in preparation for publication:

A temporal transcriptome in human embryonic stem cell-derived cardiomyocytes identifies novel regulators of early cardiac development

Kai Fu[1*], Haruko Nakano[2*], Marco Moselli[2], Atsushi Nakano[2$] and Matteo Pellegrini[1$]

[*]Contribute equally to this work, [$]Correspondence: anakano@ucla.edu, matteop@gmail.com

[1] University of California Los Angeles, Department of Molecular, Cellular and Developmental Biology, Bioinformatics Interdepartmental Program, Los Angeles, CA 90095, USA

[2] University of California Los Angeles, Department of Molecular, Cellular and Developmental Biology, Los Angeles, CA 90095, USA

Chapter 4 is a version of manuscript in preparation for publication:

Integrative modeling of DNA methylation with core histone modifications in various human cells

Kai Fu[1], Giancarlo Bonora[1] and Matteo Pellegrini[1*]

[*]Correspondence: matteop@gmail.com

[1] University of California Los Angeles, Department of Molecular, Cellular and Developmental Biology, Bioinformatics Interdepartmental Program, Los Angeles, CA 90095, USA

Important people

Firstly, I would like to thank my advisor Matteo Pellegrini, who strongly supports my research during my Ph.D. He also gives me a lot of freedom to try different new things and show great patient and understanding during my research. Secondly, I would like to thank the members of my thesis committee Kathrin Plath, Atsushi Nakano, Yi Xing and Jason Ernst for their help and support. Thirdly, I would like to thank Yixiao Xu, who makes sure I have a life outside of this thesis. Lastly, I would like to thank Haruko Nakano and Marco Morselli who generated the data and validation results in chapter 3, and Giancarlo Bonora who carried out the initial analysis in chapter 4.

Funding

During my Ph.D, I received stipend and tuition from UCLA Bioinformatics Program and M.P. I also received a teaching assistant stipend from UCLA Molecular, Cellular and Developmental Biology Department.

Vita

2010	B.S. (Biomedical Engineering), Tongji University
2010–2013	Research Assistant, Yong Zhang Lab, Tongji University
2013	M.S. (Bioinformatics), Tongji University, Shanghai, China
2015	Teaching Assistant, Molecular, Cellular and Developmental Biology Department, UCLA.
2016	Internship, Bioinformatics Department, Genentech
2013–present	Graduate Student Researcher, Matteo Pellegrini Lab

PRESENTATIONS

Kai Fu, Constantinos Chronis, Giancarlo Bonora, Kenneth Zaret, Kathrin Plath, Matteo Pellegirni. Comparison of the binding of pluripotency factors between human and mouse in early iPSCs reprogramming. **2016**. UCLA

Kai Fu and Matteo Pellegrini. Applying computational approaches to discover novel epigenetic mechanisms. **2018**. UCSD

PUBLICATIONS

Fu K, Chronis K, Bonora G, Zaret K, Plath K, Pellegrini M. Comparison of the binding of pluripotency factors between human and mouse in early iPSCs reprogramming. ***BMC Genomics*** (Revision, First author)

Yu J*, Seldin M*, **Fu K***, Lam L, Li S, Wei B, Kulkarni R, Teitell M, Pellegrini M, Lusis A, Deb A. Topological arrangement of cardiac fibroblasts regulates cellular plasticity. ***Circulation Research***. (Co-first author)

Fu K*, Haruko Nakano*, Marco Moselli, Atsushi Nakano and Matteo Pellegrini. Identification of novel regulators during embryonic stem cells to cardiomyocytes. (Manuscript in preparation, First author)

Fu K and Pellegrini M. Integrative modeling of methylomes with core histone marks for various human cells. (Manuscript in preparation, First author)

Morselli M, Pastor W, Montanini B, Nee K, Ferrari R, **Fu K**, Bonora G, Rubbi L, Clark A, Ottonello S, Jacobsen S, Pellegrini M. In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. ***Elife*** .

Haruko Nakano, Itsunari Minami, Daniel Braas, Herman Pappoe, Xiuju Wu, Addelynn Sagadevan, Laurent Vergnes, **Fu K**, Marco Morselli, Xueqin Ding, Adam Stieg, James Gimzewski, Matteo Pellegrini, Peter Clark, Karen Reue, Aldon Lusis, Bernhard Ribalet, Siavash Kurdistani, Heather Christofk, Norio Nakatsuji, and Atsushi Nakano. Glucose inhibits cardiac muscle maturation through nucleotide biosynthesis. ***Elife***.

Ohashi M, Lee P, Allen D, **Fu K**, Vargas B, Cinkornpumin J, Salas C, Park, J, Germanguz I, Chronis K, Kuoy E, Wu T, Lin K, Xiao AZ, Chen L, Tran S, Xiao, G, Lin L, Jin P, Pellegrini M, Plath K and Lowry WE. Loss of MECP2 leads to telomere dysfunction and neuronal stress. (Under review)

Chapter 1

Introduction:

Epigenomics and Big Data

1.1 Epigenomics

The genome is the sum of genetic material within a cell. It contains all the genetic information that controls every aspect of biological processes of a living organism. In 2003, when the Human Genome Project was completed, scientists began to describe all the genes in the human genome [1]. This was a milestone in the history of science, and since then, these achievements have revolutionized biomedical science and research.

With the rapid development of high-throughput sequencing technologies, the overall time and expense to sequence a genome, such as human, has dramatically decreased during the last ten years. The major bottleneck is no longer sequencing DNA itself, but is the interpretation of the function of DNA sequences in a genome. Research in the post-genomics focuses on annotating sequenced genomes in new ways. Re-sequencing technologies, such as ChIP-Seq, to identify DNA-Protein interactions [2], RNA-Seq to quantify gene expression and search for novel transcripts [3], Bisulfite-Seq to measure DNA methylation levels [4], have become powerful tools for annotating a genome, and thus be able to interpret the function of regulatory elements in a genome.

International collaborative projects, such as ENCODE (Encyclopedia of DNA Elements) and Epigenomics Roadmap, have generated tens of thousands of such re-sequencing datasets and have used them to annotate the functional elements in hundreds of human tissues and cell lines [5, 6]. At the same time, biologists are generating vast numbers of genomics and epigenomics datasets to investigate various biological systems. To interpret these valuable datasets, we need not only sophisticated computational tools to analyze them, but also intelligent integration methods to generate novel biological insights.

The central dogma is the most important and fundamental concept in modern molecular biology [7]. It involves three major processes: DNA replication, DNA transcription into RNA, and RNA translation into proteins. My PhD thesis then focus on the epigenetic regulation of transcription, which is the key process that controls genomic information flow from DNA to RNA.

Epigenetic modifications are reversible modifications of a cell's DNA or histones that affect gene expression without altering the DNA sequence [8-10]. Recent progress in high-throughput sequencing technologies enables us to measure genome-wide epigenetic modifications in an unprecedented way. We can now decipher genomic DNA through DNA sequencing, transcribed RNA through RNA-Seq, and epigenetic modifications through ChIP-Seq or Bisulfite-Seq. However, measuring a single layer of information is not enough to reveal the hidden biological mechanisms, since cells are regulated in complex ways. To overcome this limitation and arrive at a more complete understanding of key cellular mechanisms, we need to systematically investigate the functionality of the genome from a multi-dimensional perspective.

Motivated by this goal, my thesis focuses on applying sophisticated statistical methods and development of novel computational algorithms to analyze and integrate different layers of high-throughput sequencing datasets. On the biology side, I focus on stem cell reprogramming and stem cell differentiation processes, since the understandings of these two biological systems provides key insights to address both basic biological questions and lead to potential applications in regenerative medicine.

1.2 Epigenomics of stem cell

Stem cells are able self-renew and differentiate into any cell type. In adult tissues, stem cells repair aging cells and replenish adult cells. In the developing embryo, stem cells differentiate into myriad specialized cells. As a result there is great interest in understanding biological mechanisms underlying the pluripotency of stem cells. The knowledge gained from the study of stem cells could have a significant impact in both increasing our understanding of basic biological questions and applications in regenerative medicine.

Although stem cells hold great promise in regenerative medicine, there is still a limited understanding of how stem cells are regulated. There are a number of ways to study stem cells: from a biological chemistry, cellular biology or developmental biology perspective. In my PhD dissertation, I focus on the study of epigenetic regulation of transcription in stem cells. By doing so, I address the following key questions in stem cell research: 1. How do the core transcription factors, Oct4, Sox2, Klf4, and c-Myc, control reprogramming to pluripotency, 2. How epigenetics changes affect stem cell differentiation or reprogramming, 3. What transcription factors control the differentiation process from stem cells to cardiomyocytes. 4. What is the quantitative relationship between DNA methylation and histone modifications. Answering these questions will reveal novel biological mechanisms underlying the regulation of stem cells and ultimately contribute to the transition of stem cell therapies.

I thus focus on three novel aims to study the biological mechanisms of stem cell regulation and epigenomics. In specific aim one, I study the stem cell early reprogramming process from a comparative genomics aspect. By comparing OSKM binding sites between human and mouse, I am able to reveal both the shared and divergent patterns of OSKM regulation. In specific aim two, I study a unique biological system where stem cells are differentiated into cardiomyocytes with very high efficiency. This system enables me to find cardiomyocyte specific novel genes and identify potential new driver transcription factors that mediate the differentiation process. In specific aim three, I integrate hundreds of ChIP-Seq and Bisulfite-Seq assays to study to crosstalk between DNA methylation and histone modifications for 35 human cell types. This result reveals that histone modifications are highly predictable of DNA methylation in a variety of human cell types.

1.3 Overview of research project

Chapter 2 is based on a manuscript in preparation for publication that compares OSKM regulation between human and mouse in early iPSC reprogramming process. By expressing the core transcription factors, Oct4, Sox2, Klf4 and c-Myc (abbreviated as OSKM), adult differentiated cells can be reprogrammed into induced Pluripotent Stem Cells (iPSCs) that have the ability to differentiate into any type of cells [11, 12]. Since transplants of iPSCs derived tissues or organs should not cause immune rejection to its donor, iPSCs technology holds great promise in regenerative medicine. However, there is still a limited understanding of the molecular mechanisms of iPSCs reprogramming. An important approach to understand reprogramming is to systematically investigate the mechanisms of the core transcriptional circuitries that underlie this process. Scientists have generated iPSCs from both human and mouse fibroblast cells. Previous studies show similar properties of either iPSCs from human or mouse [13-15]. However, it is unknown whether human reprogramming and mouse reprogramming are controlled by the same pattern of OSKM binding sites or not. This question motivates the first research project in my dissertation. Through a comprehensive comparison between human and mouse OSKM binding profiles, I am able to identify both shared and divergent patterns of OSKM binding in human and mouse on a genome-wide scale.

Chapter 3 is based on a manuscript in preparation that dissects the cardiomyocyte differentiation regulatory network. Stem cell-based cardiogenesis holds great promise for novel therapeutic approaches to heart diseases. However, we still have a limited understanding of the mechanisms associated tithe the differentiation of stem cells to cardiomyocytes. We thus established a differentiation protocol that yields about 90 percent cardiomyocytes from human embryonic stem cells [16, 17]. This powerful biological system provides a valuable tool to examine the mechanisms of cardiogenesis. Two main results are obtained in this research project. First, dissecting genome-wide gene expression changes during cardiogenesis. This helps us identify cardiogenesis associated genes or transcripts. Second, inferring potential driver transcription factors that control the cardiogenesis process by integrating expression profiles and genome-wide epigenetic profiles.

Chapter 4 is based on a manuscript in preparation that models the quantitative relationships between histone modifications and DNA methylation in human cells. Recently, hundreds of epigenomic landscape maps have become available though the Epigenome Roadmap Project [6]. These valuable maps provide unprecedented resources to study epigenetic regulation of cells. I thus integrate hundreds of ChIP-Seq and Bisulfite-Seq assays to interrogate the quantitative relationships between DNA methylation and histone modifications for 35 human cell types. As a result, I built a logistic regression model to link the two types of assay and predicted DNA methylation in high accuracy. This study provides the largest integration analysis of DNA methylation and histone modifications so far and reveals the close crosstalk between the two major epigenetic mechanisms in human.

References

1. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931–945.

2. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.** *Nat Protoc* 2012, **7**:1728–1740.

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

4. Krueger F, Kreck B, Franke A, Andrews SR: **DNA methylome analysis using short bisulfite sequencing data.** *Nat Methods* 2012, **9**:145–151.

5. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

6. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, et al.: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518**:317–330.

7. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227**:561–563.

8. Smith ZD, Meissner A: **DNA methylation: roles in mammalian development.** *Nat Rev Genet* 2013, **14**:204–220.

9. Esteller M: **Cancer epigenomics: DNA methylomes and histone-modification maps.** *Nat Rev Genet* 2007, **8**:286–298.

10. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nat Genet* 2007, **39**:311–318.

11. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126**:663–676.

12. Yamanaka S: **Induced pluripotent stem cells: past, present, and future.** *Cell Stem Cell* 2012, **10**:678–684.

13. Singh VK, Kalsan M, Kumar N, Saini A, Chandra R: **Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery.** *Front Cell Dev Biol* 2015, **3**:2.

14. Takahashi K, Yamanaka S: **A decade of transcription factor-mediated reprogramming to pluripotency.** *Nat Rev Mol Cell Biol* 2016, **17**:183–193.

15. Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, Borkent M, Apostolou E, Alaei S, Cloutier J, Bar-Nur O, Cheloufi S, Stadtfeld M,

Figueroa ME, Robinton D, Natesan S, Melnick A, Zhu J, Ramaswamy S, Hochedlinger K: **A molecular roadmap of reprogramming somatic cells into iPS cells.** *Cell* 2012, **151**:1617–1632.

16. Nakano H, Minami I, Braas D, Pappoe H, Wu X, Sagadevan A, Vergnes L, Fu K, Morselli M, Dunham C, Ding X, Stieg AZ, Gimzewski JK, Pellegrini M, Clark PM, Reue K, Lusis AJ, Ribalet B, Kurdistani SK, Christofk H, Nakatsuji N, Nakano A: **Glucose inhibits cardiac muscle maturation through nucleotide biosynthesis.** *Elife* 2017, **6**:025003.

17. Minami I, Yamada K, Otsuji TG, Yamamoto T, Shen Y, Otsuka S, Kadota S, Morone N, Barve M, Asai Y, Tenkova-Heuser T, Heuser JE, Uesugi M, Aiba K, Nakatsuji N: **A small molecule that promotes cardiac differentiation of human pluripotent stem cells under defined, cytokine- and xeno-free conditions.** *Cell Rep* 2012, **2**:1448–1460.

Chapter 2

Comparison of reprogramming factor targets reveals both species-specific and conserved mechanisms in early iPS cells

2.1 Abstract

Both human and mouse fibroblasts can be reprogrammed to pluripotency with Oct4, Sox2, Klf4, and c-Myc (OSKM) transcription factors. While both systems generate pluripotency, human reprogramming takes considerably longer than mouse. To assess additional similarities and differences, we sought to compare the binding of the reprogramming factors between the two systems. In human fibroblasts, the OSK factors initially target many more closed chromatin sites compared to mouse. Despite this difference, the intra- and intergenic distribution of target sites, target genes, primary binding motifs, and combinatorial binding patterns between the reprogramming factors are largely shared. However, while many OSKM binding events in early mouse cell reprogramming occur in syntenic regions, only a limited number is conserved in human. Our findings suggest similar general effects of OSKM binding across these two species, even though the detailed regulatory networks have diverged significantly.

2.2 Introduction

By expressing the transcription factors Oct4, Sox2, Klf4 and c-Myc (abbreviated as OSKM), differentiated cells can be reprogrammed into induced pluripotent stem cells (iPSCs) that have the ability to differentiate into any type of cell [1, 2]. iPSC technology holds great promise in regenerative medicine and for the modeling of diseases in a culture dish [3, 4]. However, there is still limited understanding of the essential mechanisms underlying reprogramming of somatic cells to iPSCs. Furthermore, there are marked differences in the reprogramming process for mouse and human cells, even though reprogramming can be accomplished by the same set of factors. Mouse cells reprogram within a week or two, whereas human cells take up to a month and the efficiency of the conversion is typically lower in the human system [5, 6]. Moreover, while mouse cells can be reprogrammed efficiently with OSK alone, ectopic c-Myc expression is more critical in the human process [2, 7, 8]. To understand universal features of reprogramming across species, we characterized the differences and similarities in the regulatory networks that were manifested at the onset of reprogramming of human and mouse somatic cells.

An important approach towards understanding the reprogramming process is to systematically investigate the binding of reprogramming factors in the genome. By investigating OSKM binding at 48 hours of reprogramming, previous studies have begun to elucidate the patterns and regulatory roles of OSKM in early reprogramming in the human and mouse systems [9-11]. Reprogramming typically is an inefficient process where only few cells in the culture dish induce the pluripotency program, yielding a highly heterogeneous cell population at the end of the process [12]. However, in the first 48 hours of reprogramming, the reprogramming culture is thought to react homogeneously [13, 14], enabling location studies of OSKM in the early reprogramming population. Moreover, for the 48-hour time point in mouse, we used fetal bovine serum containing media, which results in iPSC colonies within 2-3 weeks. In these conditions, the timing of reprogramming is similar to that found in human experiments. The early human and mouse cells are thus expected to be in a similar stage of reprogramming. However, the final iPSC stage between human and mouse is significantly different: the human cells are reprogrammed to a primed stage while the mouse cells are reprogrammed to a naïve stage [15]. For this reason, in this study we focused on the 48-hour comparison instead of the iPSC stage of reprogramming.

In this study, we compared the initial OSKM binding events between human and mouse fibroblasts to shed light on both conserved and species-specific mechanisms of OSKM-mediated processes early in reprogramming. By focusing on the binding events of OSKM early in reprogramming, we guaranteed minimal influence of the differences between human and mouse cell reprogramming that resulted in mouse iPSCs in the naïve pluripotent state and human iPSCs in the primed pluripotent state caused by the external culture conditions. We first show that general features of OSKM binding events, such as inter- and intragenic distribution, target genes, primary binding motifs, and combinatorial binding patterns between the reprogramming factors, are largely similar between human and mouse. However, when we compared the locations of OSKM binding events, we found that only a small fraction of binding sites in syntenic regions were conserved between human and mouse at 48 hours of reprogramming. This result indicates that the binding of the reprogramming factors is in large part distinct at the initial stage of the reprogramming process. We show that conserved binding events within syntenic regions often represent target sites that are also bound in the pluripotent end state and tend to occur in promoters and enhancers, suggesting that the engagement of pluripotency sites early in reprogramming is a conserved mechanism between mouse and human reprogramming. Lastly, we show that both motif usage and chromatin states contribute to the conservation of binding events in early human and mouse reprogramming.

2.3 Results

2.3.1 General features of OSKM binding events in early human and mouse reprogramming

In this study, we compare the binding of OSKM peaks in mouse and human at 48 hours post transfection. This is accomplished by analyzing previously published datasets [9, 11]. We note that there are some differences in the mouse and human datasets that are due to the difference in overexpression methodology and the starting cell type. While the mouse data was generated by overexpressing the pluripotency factors using a polycistronic cassette (ensuring that each cell expresses all four factors at comparable levels), the human data was generated using individual lentiviral vectors, which leads to more variability in the combination and level of expression of the factors. However, as we show below, these differences do not have a significant impact on our conclusions.

We first addressed the effects of overexpression between polycistronic and individual based approaches. While the primary results presented in Chronis *et al* were based on a polycistronic cassette [11], in the same study we also collected binding data generated by individually overexpressing factors using pmX. We showed that in mouse, individual retroviral based expression of Oct4, Sox2 and Klf4 (OSK) have strong signals in the polycistronic derived OSK peaks, indicating that the OSK signal from the two systems are enriched in similar genomic loci (Supp Fig 1). Moreover, we note that while the mouse experiments were carried out in embryonic fibroblasts, the human studies were done in fetal foreskin fibroblast. Since we did not have access to epigenomes from both embryonic fibroblasts and fetal foreskin fibroblasts in either human or mouse, we were not able to compare the potential differences between the two starting cell types. Nonetheless, we do have access to epigenomes for both human foreskin newborn and human lung fetal fibroblasts from Roadmap Epigenomics Project. To address the potential differences between different types of fibroblasts, we used DNaseI hypersensitive sites to represent the chromatin states and then compared their overlapping. Supplementary Fig 2 then shows that the two types of fibroblasts have a large overlapped number of DNaseI peaks, suggesting the overall similarities of chromatin states between those two types of fibroblasts. We thus argue that chromatin changes are modest between the two types of fibroblasts we used in this study.

Having shown that these two experimental strategies yield similar OSKM binding events, we chose to focus our analyses on the mouse polycistronic and human individual lentiviral cassette where all our ChIP-Seq and RNA-Seq data was collected. To further enable their comparison, we generated OSKM peaks for both human and mouse cells reprogramming using the same analysis pipeline for mapping and peak calling, setting the peak calling q-value cutoff of 0.05 (see Methods).

The human and mouse data sets generated a similar number of peaks for Oct4, while the early human reprogramming culture had about twice as many peaks for the other three factors compared to the mouse (Supp Fig 3). In both human and mouse, ChIP-seq for Myc generated fewer peaks than O, S, or K (Supp Fig 3). The average fragment size (average distances between plus strand reads and minus strand reads) was similar for all four reprogramming factors in the mouse and human data sets (Supp Fig 4, Supp Fig 5). We found that the human datasets for O, S, and K had a lower signal to noise ratio than the mouse data sets, whereas M binding events were slightly stronger in the human sample (Supp Fig 4, Supp Fig 5, Supp Fig 6).

We first asked whether OSKM peaks had a similar positional distribution with respect to transcriptional start sites (TSSs) in the two species (Fig 1a). Specifically, we classified the distances between peaks and TSSs into different groups, i.e. 0 to 5kb, 5 to 50kb etc. We found that O, S and K peaks were most abundant in the -500 to -50kb and 50 to 500kb bins in both human and mouse, indicating that O, S and K in both human and mouse predominantly bind regions distal to TSS. M peaks, however, were most abundant in -500 to -50kb and 50 to 500kb bins in human, while most abundant in the -5 to 0kb and 0 to 5kb bins in mouse. This result reveals that M has a different distribution between human and mouse: in humans M tends to bind distally to the TSS whereas in mouse it tends to bind proximal to TSS regions. In addition, we observed less overall binding of O, S and K in proximity to the TSS compared to distal sequences in human than in mouse cells (Fig 1a).

We next compared the target genes for each factor between human and mouse reprogramming. Targets were defined as a gene whose TSS is closest to the peaks for each factor irrespective of binding distance. Because there were tens of thousands of O, S, K, and M peaks, about 40% to 70% of all genes could be assigned to O, S, K or M peak. We calculated the number of overlapping target genes among the four factors in the two species and found that a large fraction of genes was targeted by the four factors in both species (Fig 1b). Furthermore, among the 8,433 OSKM co-targeted genes in human and 6,867 co-targeted genes in mouse, 3,919 of them were shared significantly (p-value < 10^{-16}, hypergeometric test), indicating a large fraction of OSKM co-targeted genes are conserved. Gene ontology enrichment analyses showed those shared co-targeted genes were enriched in the biological processes of regulation of transcription, in utero embryonic development and regulation of Wnt signaling pathway. This agrees with previous studies which showed that the Wnt signaling pathway modulated reprogramming efficiency when altered early in reprogramming [16]. When only considering orthologous genes between human and mouse, we also found a large overlap of target genes for each reprogramming factor (Fig 1c). The hypergeometric test showed that the number of overlapping target genes was also significant for each of the four factors (p-value<10^{-16}). Those results indicate that the factors tend to generally target the same genes in mouse and human fibroblasts. We also used another definition of target genes, requiring that the peak of each factor was within 20kb of the TSS and obtained a similar result (Supp Fig 7).

We next carried out *de novo* motif discovery in each factor's binding regions (see Methods). The DNA binding motifs we identified for each reprogramming factor was similar between human and mouse (Fig 1d). However, we observed minor motif differences in Oct4, which terminated with A/T AA in mouse but A/G C/T AT in human, as well as in c-Myc, which terminated with C G/A TG in mouse but C/T G T/C G in human. Moreover, *de novo* motifs of the four factors were largely consistent with their canonical motifs (obtained from Jaspar database) [17], indicating DNA binding preferences of O, S, K, and M are largely conserved between human and mouse.

To further characterize OSKM binding, we identified all possible combinations of binding events. If summits of peaks from different reprogramming factors were within 100 bp of each other, we considered them to be “co-“ binding events. If summits of peaks from one factor were at least 500 bp away from all other factors, we defined these as “solo” binding events. To gauge whether co-bound sites occurred more or less frequently than expected, we compared our counts to a synthetic null model for all possible combinations of factors (see Methods). We found that in both human and mouse, all co-binding events occurred more frequently than expected, whereas solo binding sites were observed less frequently than expected (Fig 1e). OSKM, OSK, OSM, OKM and SKM co-binding events were the most prevalent combinations in both human and mouse. Moreover, solo binding sites were more likely in human than in mouse and nearly all co-binding events (except KM and OKM) were more prevalent in mouse. Regardless of the differences, these results indicate that O, S, K, and M tend to bind together with similar combinatorial patterns in both human and mouse, suggesting that the factors often co-bind to exert their actions. Overall, we conclude that the general properties of O, S, K and M are similar, although there are some observable differences.

2.3.2 Comparison of OSKM binding to the chromatin state of starting cells

Next, we sought to compare the chromatin state in the starting cells for OSKM binding sites at 48 hours between human and mouse. This enabled us to see how

OSKM interacted with the initial chromatin states in fibroblasts. We analyzed H3K4me1, H3K4me3, H3K27me3, H3K27ac and H3K36me3 histone marks of human fibroblasts from the Roadmap Epigenome Project [18] and of mouse fibroblasts [11] to build a 15 chromatin state model using ChromHMM [19] with a concatenated human-mouse genome (see Methods). Based on the combinatorial probability of the five histone marks, we classified the mouse and human genomes into chromatin states such as active promoter and active enhancer. We chose a model with 15 chromatin states because these had a clearly distinct combination of histone marks and functional annotations based on prior expectations. The genomes of both human and mouse were segmented into non-overlapping 200 bp regions, and each bin associated with a specific chromatin state. Figure 2a shows the emission probabilities (signal enrichments) of each histone mark as well as the fractions of the genome (numbers in the brackets, human followed by mouse) that each chromatin state occupies in human and mouse fibroblasts. We noted primary differences between human and mouse chromatin states including the frequency of the two H3K9me3-containing chromatin states, weak repressed polycomb and quiescent chromatin state, where human fibroblasts had significantly more genomic regions annotated as ZNF/repeats and heterochromatin and less genomic regions annotated as the latter two states.

By intersecting OSKM peaks with chromatin states, we calculated the percentage of peaks within chromatin states in both human and mouse (Fig 2b).

As a result, about 40~50% of human O, S, and K peaks, and 20% of human M peaks were within low signal regions (states 14 and 15). This chromatin analyses agrees with a direct assessment of the individual histone modification states targeted by OSKM, which showed that O, S, and K predominantly target unmarked chromatin sites [9]. By contrast, in mouse, the percentage of low signal regions targeted decreased to 20% for O, S, and K peaks and 2% for M peaks. In addition, about 40~50% of mouse O, S, and K peaks and 30% of M peaks were within enhancers, consistent with the finding that mouse OSK efficiently target enhancers active in fibroblasts early in mouse reprogramming [11]. However, for human O, S and K peaks, this number dropped to about 10%~25% and M peaks showed a similar number of 30%. After correction for the genome percentage annotated as different chromatin states, the human peaks were still more enriched in low signal regions and less enriched in enhancer regions. Those results reveal a distinct distribution of OSKM in chromatin states of low signals and enhancers between human and mouse. This binding preference may also suggest pluripotent genes in human are more difficult to induce and thus, human reprogramming will take longer than in mouse. In addition, consistent with the genomic distribution analysis (Fig 1a), mouse c-Myc was more often associated with promoters compared to human.

Additionally, when considering the genomic region captured by each chromatin state, we calculated the enrichment of OSKM peaks in each chromatin state by calculating the log2 ratio between peak percentage and chromatin state percentage (Fig 2c). This reveals OSKM binding preferences in the various chromatin states. As a result, we observed a strong preference of mouse OSKM targeting promoters and enhancers, whereas in human, this preference still held but the extent was decreased.

2.3.3 OSKM binding events show limited conservation between human and mouse

To further compare OSKM occupancy in early mouse and human cell reprogramming, we mapped mouse peaks to the human genome based on synteny (see Methods). Mouse peaks were classified into three groups based on sequence conservation and binding conservation. Figure 3a shows a schematic illustration of the definition of the three groups: syntenic conserved peaks, syntenic unconserved peaks, and unsyntenic peaks. Syntenic conserved (SC) peaks had orthologous DNA sequences as well as binding events in both organisms. Syntenic unconserved (SU) peaks only had orthologous DNA sequences but no binding event detected in human. Unsyntenic (UN) peaks did not have orthologous DNA sequences between organisms and therefore could not be mapped between human and mouse.

We found that about 74, 80, 73 and 89 percent of mouse O, S, K, and M peaks, respectively, were syntenic with human, while the background ratio for the entire genome was about 40 percent (Fig 3b), indicating that elements bound by OSKM show much higher sequence conservation rates than the rest of the genome, consistent with OSKM bind to cis-regulatory events such as enhancers and promoters. However, for each reprogramming factor, we found that syntenic conserved peaks only represented a small fraction of peaks (Fig 3c). Specifically, 4%, 4.5%, 10.9% and 34.4% of mouse O, S, K, and M peaks, respectively, were syntenic conserved. O, S, and K, which mostly bind to enhancer regions in mouse (Fig 2b) [11], had a lower fraction of conserved peaks compared to M, which mostly binds to promoter regions in early mouse cell reprogramming (Fig 2b) [11]. We then asked whether the limited degree of conservation between mouse and human binding events could be solely explained by random background binding events between human and mouse. To address this we simulated both human and mouse background peaks (same number and length with the observed ones), then calculated the conservation rate and repeated the simulation 1,000 times. The simulation result showed a conservation rate for OSKM background peaks of approximately 1%, implying that although the fraction of conserved binding was relatively small, conserved binding events still occurred at a higher rate than expected by chance. Lastly, we also mapped mouse pMX peaks (individual retroviral based system) to human peaks. Consistent with the comparison between polycistronic peaks in mouse and lentiviral peaks in human, our result showed that there was a limited fraction of syntenic conserved peaks for Oct4, Sox2 and Klf4 (Supp Fig 8). This result also indicates that the divergence of binding between human and mouse is not affected by using different overexpression systems.

In a previous study, Cheng et al. showed that the degree of binding conservation varied markedly, from several percent to about 60 percent, between human and mouse among different transcription factors (TFs) [20]. In addition, promoter bound TF binding sites showed higher conservation rates than enhancer sites. Moreover, this trend held after adjusting the sequence conservation differences between promoters and enhancers, indicating that the TF binding sites in promoter regions are indeed more conserved than those in enhancer regions [20]. In another study, Schmidt et al. reported a 10 to 22 percent binding conservation rate between two of five mammals for liver-specific transcription factors [21]. In early reprogramming, we observed a low conservation rate for O, S, and K and a medium conservation rate for M, indicating the significance of binding divergence in early reprogramming system between human and mouse fibroblasts.

We next investigated whether peak binding strength (based on peak calling q-values) had an impact on conservation. We classified all mouse peaks into four groups based on their -log10 q-values (Fig 3d). For each reprogramming factor, we observed a clear trend where the strongest peaks (top 25%) had a higher percentage of syntenic conserved binding events compared to other three groups. This result suggests peak binding strength indeed is positively correlated with peak conservation rates and stronger peaks tend to be more conserved.

By analyzing the presence of repeat sequences within the three groups of peaks (see Methods) (SC, SU, and UN), we found that the unsyntenic peaks had a much higher percentage of repeat sequences compared to the other two groups, and, except for Sox2, syntenic conserved binding sites contained the fewest repeats (Fig 3e). Moreover, compared to peaks in syntenic regions, peaks in unsyntenic regions were more often associated with long terminal repeats (LTR) and short interspersed nuclear elements (SINE) and less often with simple repeats in the mouse genome (Supp Fig 9). These results are consistent with previous findings which showed that transposable elements are enriched in species-specific sequences and have rewired the transcriptional network during evolution [22, 23].

The analyses described above were carried out by mapping mouse OSKM peaks to the human genome, but we also performed the inverse analysis by mapping human OSKM peaks to the mouse genome (Supp Fig 10a). Approximately 60 percent of human peaks occurred in genomic regions syntenic with the mouse. The lower syntenic rate of human peaks mapping to the mouse genome compared with mouse peaks mapping to the human genome correlated with a higher proportion of repeats in human peak sequences (Fig 9). Among human OSKM peaks in syntenic regions, those also found in the mouse (syntenic conserved) constituted a small proportion as seen in the reverse mapping of mouse OSKM peaks to the human data (Supp Fig 10b). Interestingly, syntenic and unsyntenic human OSKM peaks showed a more similar distribution of certain types of repeats compared to mouse peaks (Supp Fig 11, Supp Fig 12).

We also investigated how human syntenic peaks and all peaks of mouse were distributed relative to each other. We first calculated the distances between human syntenic peak summits and mouse peak summits. We then categorized the distances into several groups of genomic ranges, i.e. within 200bp, 400bp, 600bp, 800bp etc. Lastly, to compare the observed distance distribution with simulated background, we calculated the background distance distribution, where the mouse peaks were shuffled and the human syntenic peaks were kept fixed. The result suggests that observed human syntenic peaks are indeed closer to observed mouse peaks than expected by chance (Fig 3f). Moreover, there was a clear trend showing that the log2 ratio between observed and simulated peaks declined with increased distance. Among the four factors, c-Myc showed the most dramatic trend. This is consistent with the fact that c-Myc is the most conserved factor compared to the other three.

2.3.4 Syntenic conserved peaks are associated with different genomic features compared with unconserved peaks

Since we observed that only a small fraction of syntenic peaks had conserved binding early in reprogramming in human and mouse cells, we sought to identify properties that distinguish conserved peaks from the others. We observed that syntenic conserved peaks had significantly higher ChIP enrichment (-log10 q-value) than the other two groups (Fig 4a), indicating the SC peaks tend to be bound more strongly. We then used the GREAT tool [24] to perform gene ontology enrichment analysis for the mouse SC, SU, and UN peaks, with all peaks as background (Supp Fig 13). For SC peaks of OSM, we found their target genes were enriched for fat pad, adipose tissue, and adrenal gland development. Surprisingly, for SU peaks no enriched gene ontology terms for any of the four factors were detected. UN peaks of OSKM were strongly enriched in immunity-related gene ontologies. These results suggest that the target genes of the three groups of peaks might be associated with distinct functions. When comparing the genomic locations of mouse SC peaks to all peaks with respect to the distance to the TSSs, we found that SC O, K, and M peaks more often occurred within the proximal TSS regions, while Sox2 was slightly more often within the distal TSS regions (Fig 4b).

We also compared binding of mouse OSKM at 48 hours with that in the pluripotent state, to define those mouse OSKM binding events that were bound both early in reprogramming and in the pluripotent state (based on mouse embryonic stem cell ChIP-seq data) versus those that only occur at 48 hours but not in pluripotent cells (Fig 4c,i) [11]. In our previous study, we described that many of these persistent binding events for OSKM were enriched in promoters and OSK were also highly enriched in pluripotency enhancers [11]. We calculated the percentage of SC peaks that were bound only early in reprogramming or persist throughout reprogramming. We found that compared with SU and UN peaks, mouse SC peaks of OKM at 48 hours had a higher fraction of persistent binding events (Fig 4c,ii-iv). Specifically, for Oct4, the percent of persistent bound events was 20, 9 and 14 for SC, SU and UN respectively. For Klf4, this percent was 56, 17 and 23, and for c-Myc, this percent was 59, 10 and 22. This result indicates that conserved binding events, especially for K and M, tend to be maintained during reprogramming and are therefore likely to be more functionally important than unconserved ones.

We next asked whether SC, SU and UN peaks had distinct patterns of chromatin states in mouse at 48 hours. A mouse 18 chromatin state model was generated with nine histone marks and described in our previous paper (Supp Fig 14) [11]. We therefore calculated the percentage of peaks within each chromatin state (Fig 4d). As a result, we found that SC peaks preferentially tended to occur within certain chromatin states compared to SU and UN peaks. Specifically, SC peaks of O, K and M had higher percentages within active promoters, bivalent promoters and certain groups of enhancers. By contrast, UN peaks of O, S and K had higher percentages within low signal regions. Those results indicate that different groups of peaks are likely to associate with different chromatin states.

To further investigate the chromatin states of syntenic peaks, we performed another comparison from a human-mouse transition perspective. We assigned each syntenic peak to the chromatin state in the concatenated human and mouse genome (Fig 2a) and compared the chromatin assignment of each SC peak between mouse and human (see Methods) (Fig 5a). The color in the heatmap reflects the percentage of SC peaks within that transition in chromatin state between the mouse and human syntenic genome. For example, the top left square in the heatmap is the transition from human TSS regions (state 1) to mouse TSS regions (state 1) and the bottom right is the transition from human quiescent regions (state 15) to mouse quiescent regions (state 15) (i.e. no changes in chromatin state), and any deviation from the diagonal represents a change in chromatin state. For SC peaks of O, S, and K, the most frequent transitions corresponded to human promoter to mouse promoter, human enhancer to mouse promoter, human enhancer to mouse enhancer, and human enhancer to mouse quiescent regions. By contrast, the majority of frequent transitions for c-Myc involved promoter to promoter states. We also asked whether the chromatin state transition patterns were different for unconserved peaks. When comparing the transition profiles between SC and SU peaks (Fig 5b), we found an enrichment in human promoter to mouse promoter, human enhancer to mouse promoter and human enhancer to mouse enhancer transitions, indicating that SC peaks are more often associated with certain regulatory sites in both species than SU peaks.

Another factor that may help maintain the conservation of peaks is the occurrence of binding motifs. Although we observed that SC peaks were preferentially found within promoters and enhancers, it was not clear whether motifs help maintain the conservation of peaks between mouse and human. To shed light on this question, we computed the motif frequency in each group of peaks (see Methods) (Fig 5c). We reasoned that if the conservation of peaks was strongly influenced by the presence of binding motifs between mouse and human, then SC peaks should have a different fraction of motifs compared to the other two groups. For Sox2, 53% of SC binding events had identifiable motifs within their peaks, compared to approximately 35% of SU and UN. However, for the other three factors, SC peaks contained more motifs but the differences among the three types of peaks were smaller, indicating the limited impact of sequence motifs in the determination of binding conservation.

2.3.5 Using transitions of regulatory motifs and chromatin states as predictors of conserved binding

To quantitatively assess the extent to which SC peaks are determined by motifs or chromatin states, we built a naïve Bayesian classifier to evaluate the prediction power for classifying syntenic peaks into the SC and SU groups (see Methods). This model was trained using different sources of information: motif only, chromatin state only, and the two combined. Area under the curve (AUC) values of receiver operator curves (ROC) were used to estimate the prediction power (Fig 5d). We found that except for Sox2, the chromatin state only model outperformed the motif only model. Moreover, when combining information from both motif and chromatin states, the AUC for O, S, K, and M were 0.63, 0.71, 0.90, and 0.71 respectively. Klf4 showed a strikingly high prediction power due to its strong motif preference in syntenic regions between human and mouse and its strong chromatin state preference for specific chromatin state transitions. Although the models for O, S, and M only predicted a fraction of conserved sites, these results demonstrate that conserved peaks are indeed associated with syntenic regions that contain strong motif sequences and preferred chromatin state transitions between mouse and human.

2.4 Discussion and conclusion

In this study, we systematically compared binding patterns of the four reprogramming factors OSKM between human and mouse at an early time point of reprogramming to the iPSC state. When analyzed in each genome separately, OSKM binding sites in human and mouse shared similar features: OSK tend to bind distal TSS regions, OSKM tend to target similar genes, have similar DNA binding motifs, and show similar combinatorial binding patterns among the reprogramming factors. This suggests that molecular properties of these factors are conserved between human and mouse. However, differences emerged when we investigated the chromatin state of target sites: OSKM targeted far more closed (low signal state) chromatin states in human cells than in mouse . Importantly, when we compared the binding sites across syntenic regions, we found that there was only a small percentage of sites that were bound in both genomes (i.e. syntenic conserved, SC). Altogether, our results suggest that the initial OSKM binding sites are largely distinct in these two species, even though the phenotypic consequences of these interactions ultimately lead to similar cell types.

We also observed that most early binding events do not persist in the later stages of iPSCs reprogramming [11]. However, we found that binding events that were conserved between mouse and human tended to persist more often throughout the reprogramming process compared to unconserved sites. Conserved binding sites also tended to have a higher proportion of conserved cis-regulatory elements associated with each factor. We also showed that binding sites were more likely to be conserved if the mouse and human chromatin states were similar and the motifs were conserved.

We recognize that there are certain limitations to our analysis. One is that human and mouse reprogramming was performed using slightly different experimental protocols. An inducible polycistronic cassette including all four reprogramming factors was used in mouse fibroblasts, ensuring homogeneous expression and stoichiometry across the cell population at 48 hours; whereas four separate lentiviral constructs were used in human, each expressing one factor. However, as we have shown by comparing mouse polycistronic to individual cassettes, these different overexpression methods lead to very similar binding peaks. Also, it is possible that at 48 hours, human and mouse cells might not be in the same reprogramming stages due to their different reprogramming kinetics. However, the time point we used corresponds to early events in the time series of both species, and should, therefore, identify the first interactions of these factors with chromatin. Moreover, we compared mouse embryonic fibroblasts and human fetal foreskin fibroblasts as starting cells of reprogramming. However, we believe that the epigenome changes from embryonic to fetal stages of fibroblasts are unlikely to have a dramatic effect on OSKM binding patterns. As a result, our conclusions drawn from the comparison of these two species should not be significantly affected by the differences in the experimental details of the human and mouse systems.

In conclusion, we have shown that while some general properties of OSKM binding are conserved between mouse and human, the detailed transcriptional network is vastly reorganized. A subset of the binding events are syntenic between the two species and this study has allowed us to identify these. We do not know if they represent key events that are distinct from the large fraction of other binding sites that are not conserved. However, several lines of evidence that we have presented, such as the fact that these sites tend to persist throughout the reprogramming process, do suggest that these may play a more significant role in reprogramming than the typical unconserved site. Nonetheless, the overall picture that emerges is that the OSKM regulatory networks have significantly diverged between the two species, and while the general properties of these networks are similar, the specific binding sites are generally distinct. This observation may suggest that reprogramming to pluripotency may be driven by global regulatory changes in cells that do not depend critically on a small set of specific interactions.

2.5 Materials and Methods

Cell culture and reprogramming

In the human reprogramming system, BJ fibroblasts were purchased from ATCC (CRL-2522) at passage 6 and cultured in the ATCC-formulated Eagle's Minimum Essential Medium supplemented with 10% fetal bovine serum at 37 C and 5% CO_2. The human H1-ES line [25] were purchased from ATCC and maintained as described [26]. More information about experimental details can be found in the supplementary documents of Soufi et al. 2012 [9]. In the mouse reprogramming system, the mouse embryonic fibroblasts were obtained from day 13.5 embryos of timed mouse pregnancies. In addition, mouse embryonic fibroblasts carrying a polycistronic, dox-inducible OSKM cassette in the Col1A locus and a heterozygous M2rtTA allele in the R26 locus, were grown in standard mouse ESC media containing knockout-DMEM, 15% fetal bovine serum, recombinant leukemia inhibitory factor (Lif), b-mercaptoethanol, 1x penicillin/streptomycin, L-glutamine, and non-essential amino acids. Repogramming was induced by the addition of 2ug/ml doxycycline. We generated mouse iPS cell lines as described [27, 28]. Briefly, BJ cells at passage 10 were infected with lentiviruses encoding for dox-inducible Oct4, Sox2, Klf4, and c-Myc, along with lentiviruses expressing rtTA2M2 in the presence of 4.5 mg/ml polybrene. Additional experimental details can be found in the supplementary documents of Chronis et al. 2017 [11].

Mapping and Peak Calling

The human OSKM ChIP-Seq datasets were downloaded from GEO with accession number of GSE36570, while mouse OSKM ChIP-Seq datasets were downloaded from GEO with accession number of GSE90895. Bowtie was used to map ChIP-Seq reads of both human and mouse to their respective genomes allowing two mismatches and keeping only uniquely mapped reads for further analysis [29]. MACS2 2.1.0 was used to identify ChIP-Seq peaks with a q-value cutoff of 0.05 [30].

Motif finding and motif occurrences within peaks

MEME-ChIP was used to perform de novo motif finding for OSKM binding peaks [31]. To identify the strongest motifs, the identified summits of peaks were ranked based on their enrichments and the top 10,000 summits, along with their surrounding 200bp, were used as the input regions. The enriched motifs were identified using the DREME algorithm in the MEME-ChIP software. Starting with the most significant motif for each factor, we then used the Position Weight Matrix of this motif to scan for peaks, and determined the peaks associated with this motif using a p-value cutoff of 0.001.

Combinatorial binding and solo binding

To identify combinatorial binding regions where multiple factors bind, peaks were merged if their summits were within 100 bp of each other. Then these different combinations of binding sites were broken down into their different combinations of factors. To identify solo binding regions where only one factor bound, we required that its summit be at least 500 bp away from all other factors. Note that this method is more stringent than that used by Soufi et al. [9]; the latter considered solo binding events as simply not falling within 100 bp of the peak center. Here, to estimate the background rates of combinatorial binding, the peaks of OSKM were first randomly shuffled in the genome (using the bedtools shuffle function) [32]. Secondly, the expected number of combinatorial binding events was re-calculated based on these shuffled peaks. Lastly, we compared the number of observed binding events versus the number of expected binding events for all possible combinations of factors.

RNA-Seq samples and analysis

The human fibroblasts and 48 hours of reprogramming cells for RNA-Seq were cultured and generated in the same condition with the samples for OSKM ChIP-Seq analyzed in this study. The total mRNA was then extracted and sequenced. The experimental details could be found within method section in Tong et al. [33]. The RNA-Seq samples for mouse fibroblasts cells were obtained from Chronis et al. [11]. The raw sequencing reads of both human and mouse were then mapped back to their corresponding genome using Tophat [34]. After this, HTSeq software was used to calculate the number of reads within each gene for both human and mouse [35]. Finally, the DESeq2 software was used to perform the differential expressed gene analysis with a q-value cutoff of 0.05 [36].

Mapping sequences between human and mouse

To map OSKM binding sites between human and mouse, the liftOver algorithm from the UCSC Genome Browser was used with a cutoff of 0.5. The LiftOver algorithm uses an alignment chain file to map genomic coordinates between different versions of assemblies, or different species. The algorithm searches for regions where the input sequences are in the same block with the converted assemblies or species. The cutoff of 0.5 requires that the mapped sequences share at least half of exactly same DNA sequences with the converted species. This cutoff is consistent with modENCODE project paper which compares transcription factor binding sites between human and mouse [20]. To confirm the reliability of our results, we also used another method named bnMapper and got very similar results [37].

Peaks associated with repeat sequences

Repeat sequences were downloaded from the RepeatMasker database. We extracted the genomic coordinates for the major repeat families including DNA (DNA transposon elements), LINE (Long interspersed nuclear elements), LTR (Long terminal repeats), Retroposon (Transposons via RNA intermediates), Satellite (Satellite DNA which belongs to tandem repeats), Simple (Simple repeats) and SINE (Short interspersed nuclear elements). A peak was considered to be associated with a repeat sequence if the genomic coordinate of this repeat was within this peak.

Chromatin states for concatenated human and mouse genomes

For mouse histone marks, we used the datasets for mouse fibroblast cells from our previously published paper [11]. For human histone marks, we used the datasets of IMR90 fibroblast cell line downloaded from RoadMap Epigenomics Project [18]. To learn the joint chromatin state for human and mouse, a pseudo chromosome size table was constructed by concatenating human and mouse genomes. Then the model was trained with the human fibroblast and mouse embryonic fibroblast histone data, producing a common set of emission probabilities. We then generated a 15 chromatin state model based on the combinatorial patterns of five histone marks, i.e, H3K4me1, H3K4me3, H3K27me3, H3K27ac and H3K36me3.

Chromatin state transitions between human and mouse

Each syntenic conserved peak in mouse and its corresponding orthologous peak in human was assigned a chromatin state as described above. We then calculated the number of peaks within each possible chromatin state transition. This leads to the generation of a 15 X 15 chromatin state transition matrix. For example, the top left of the matrix represents the fraction of syntenic peaks with state 1 of human and state 1 of mouse. We also performed the same calculation for syntenic unconserved peaks between human and mouse. To compare to relative enrichment of chromatin state transitions, the log2 ratio between the syntenic conserved and syntenic unconserved matrices was calculated.

Classification model

We built a Naïve Bayes model to classify syntenic peaks into a syntenic conserved and syntenic unconserved group, based on their chromatin state transition (see above) and motif occurrences transitions. The motif occurrence transition matrix was a 2 X 2 matrix that represents the frequency of motif occurrences for syntenic peaks between human and mouse. Log odds ratios were then calculated between syntenic conserved group and syntenic unconserved groups for both chromatin state transition and motif occurrence transition matrices. As a result, each peak was assigned two values: one was the chromatin state transition log odds ratio matrix to represent the chromatin state model, and another was the motif occurrences transition log odds ratio matrix to represent the motif model. The two values were added to represent both the chromatin state and motif occurrence model. Syntenic peaks were then ranked based on log odds ratio values from either the chromatin state transition matrix or motif occurrences transition matrix, or their sum. A syntenic conserved peak was labeled as 1 and a syntenic unconserved peak is labeled as 0. Lastly, the Area under the curve (AUC) values of the receiver operator curves (ROC) were calculated to represent the model performance for classifying syntenic peaks into 1 or 0 given the chromatin state transitions or motif occurrences transitions.

Availability of data and materials

The dataset analyzed in this study can be found at GEO with accession number of GSE90895 and GSE36570.

Figures

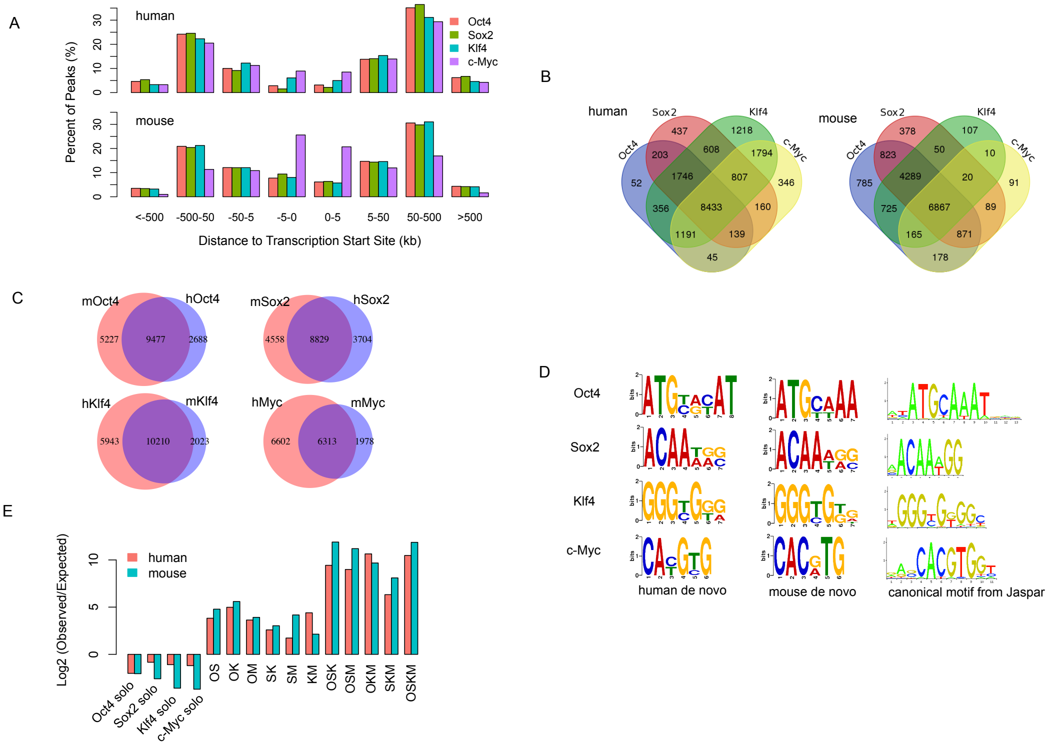
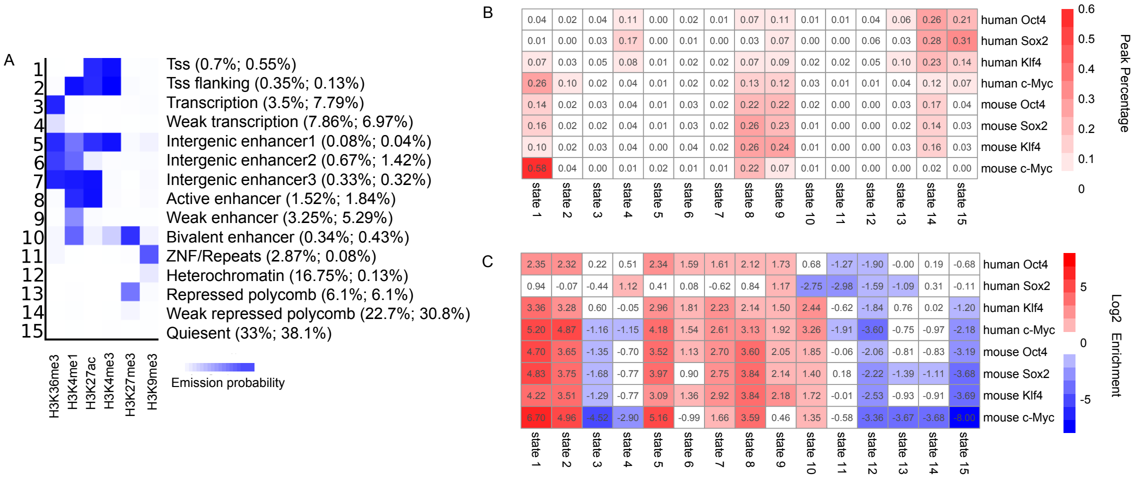


Figure 1. General feature comparison of OSKM ChIP-Seq peaks between human and mouse 48 hours fibroblast reprogramming.

A. Positional distribution of OSKM peaks with respect to Transcription Start Sites (TSSs). The top panel shows the peaks in human while the bottom panel shows that in mouse. B. Venn diagram of OSKM co-targeted genes in human (left panel) and in mouse (right panel). C. Venn diagram of OSKM co-targeted orthologous genes between mouse and human. D. De novo and canonical motifs of OSKM peaks. E. Log2 ratio of observed combinatorial binding events versus expected.



B

	state 1	state 2	state 3	state 4	state 5	state 6	state 7	state 8	state 9	state 10	state 11	state 12	state 13	state 14	state 15
human Oct4	0.04	0.02	0.04	0.11	0.00	0.02	0.01	0.07	0.11	0.01	0.01	0.04	0.06	0.26	0.21
human Sox2	0.01	0.00	0.03	0.17	0.00	0.01	0.00	0.03	0.07	0.00	0.00	0.06	0.03	0.28	0.31
human Klf4	0.07	0.03	0.05	0.08	0.01	0.02	0.02	0.07	0.09	0.02	0.02	0.05	0.10	0.23	0.14
human c-Myc	0.26	0.10	0.02	0.04	0.01	0.02	0.02	0.13	0.12	0.03	0.01	0.01	0.04	0.12	0.07
mouse Oct4	0.14	0.02	0.03	0.04	0.01	0.03	0.02	0.22	0.22	0.02	0.00	0.00	0.03	0.17	0.04
mouse Sox2	0.16	0.02	0.02	0.04	0.01	0.03	0.02	0.26	0.23	0.01	0.00	0.00	0.02	0.14	0.03
mouse Klf4	0.10	0.02	0.03	0.04	0.00	0.04	0.02	0.26	0.24	0.01	0.00	0.00	0.03	0.16	0.03
mouse c-Myc	0.58	0.04	0.00	0.01	0.02	0.01	0.01	0.22	0.07	0.01	0.00	0.00	0.00	0.02	0.00

C

	state 1	state 2	state 3	state 4	state 5	state 6	state 7	state 8	state 9	state 10	state 11	state 12	state 13	state 14	state 15
human Oct4	2.35	2.32	0.22	0.51	2.34	1.59	1.61	2.12	1.73	0.68	-1.27	-1.90	-0.00	0.19	-0.68
human Sox2	0.94	-0.07	-0.44	1.12	0.41	0.08	-0.62	0.84	1.17	-2.75	-2.98	-1.59	-1.09	0.31	-0.11
human Klf4	3.36	3.28	0.60	-0.05	2.96	1.81	2.23	2.14	1.50	2.44	-0.62	-1.84	0.76	0.02	-1.20
human c-Myc	5.20	4.87	-1.16	-1.15	4.18	1.54	2.61	3.13	1.92	3.26	-1.91	-3.60	-0.75	-0.97	-2.18
mouse Oct4	4.70	3.65	-1.35	-0.70	3.52	1.13	2.70	3.60	2.05	1.85	-0.06	-2.06	-0.81	-0.83	-3.19
mouse Sox2	4.83	3.75	-1.68	-0.77	3.97	0.90	2.75	3.84	2.14	1.40	0.18	-2.22	-1.39	-1.11	-3.68
mouse Klf4	4.22	3.51	-1.29	-0.77	3.09	1.36	2.92	3.84	2.18	1.72	-0.01	-2.53	-0.93	-0.91	-3.69
mouse c-Myc	6.70	4.96	-4.52	-2.90	5.16	-0.99	1.66	3.59	0.46	1.35	-0.58	-3.36	-3.67	-3.68	-8.00

Figure 2. OSKM peaks target of the chromatin states in starting cells.

A. Chromatin state model for concatenated human and mouse fibroblast cells based on five histone marks. The value in the heatmap represents the enrichment of that histone mark in that learned chromatin state. The values in the bracket represent the genomic percentage (human then followed by mouse) occupied by that chromatin state. B. Heatmap for percentages of OSKM peaks in each chromatin states from A. C. Heatmap for log2 enrichments between OSKM peaks percentages and chromatin state genomic percentages.

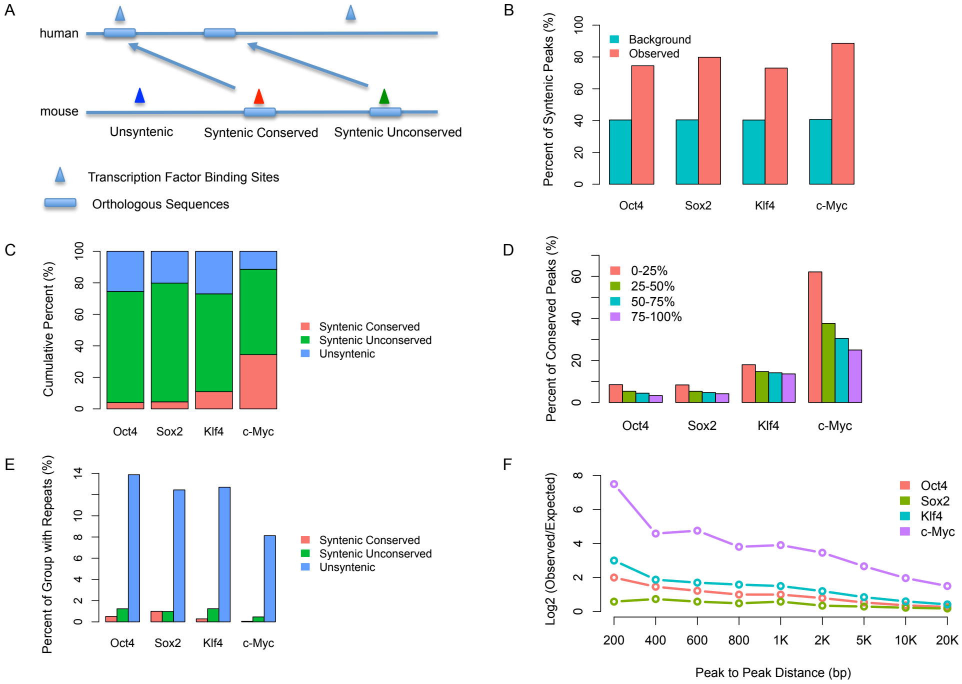


Figure 3. Map OSKM binding between human and mouse.

A. Schematic illustration of the three different groups of peaks, i.e. Syntenic Conserved (SC) binding group, Syntenic Unconserved (SU) binding group and UNsyntenic (UN) binding group. B. Percentage of mouse OSKM peaks that can be mapped to human. The background is calculated by the simulation of peaks that have the same size and same number as the real peaks, and are allowed to map anywhere on the genome. C. Fractional constitutions of SC, SU and UN peaks for each factor. D. Percentage of SC binding events with respect to all syntenic binding events. For each factor, syntenic peaks are classified into four groups based on their peak enrichments of -log10(q-value). 0-25% are the top 25 percent of peaks while 75-100 are the bottom 25 percent of peaks. E. Percentage of the three groups of peaks that contain repeat sequences. F. Log2 fold enrichment of distances between human syntenic peaks in mouse and mouse peaks compared to random background.

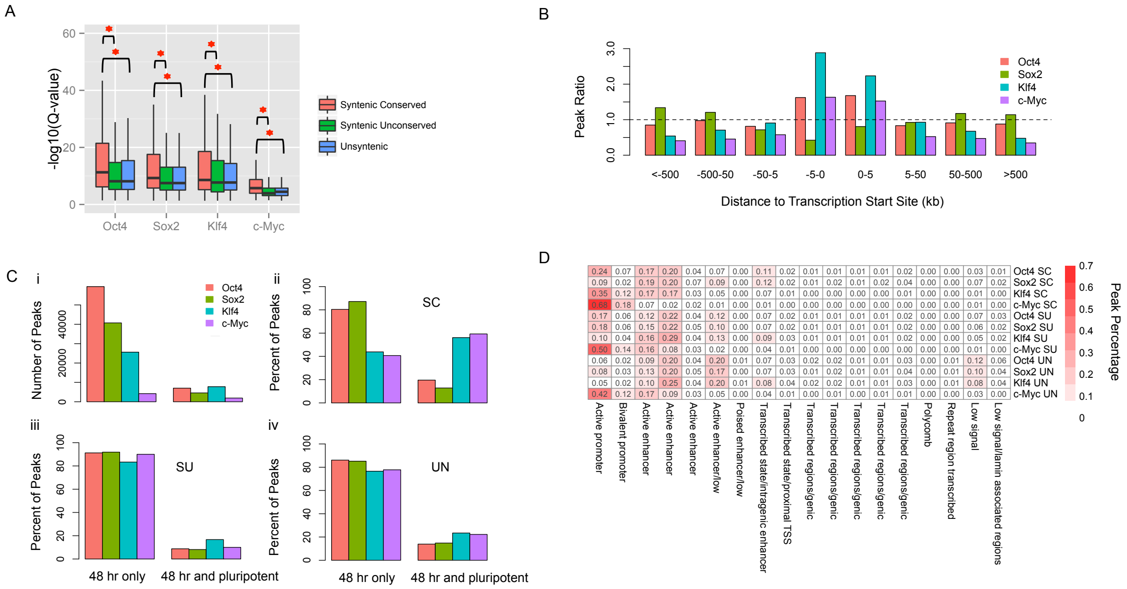


Figure 4. Comparisons of syntenic conserved peaks with syntenic unconserved peaks and unsyntenic peaks.

A. Box plot of peak calling q-values for the SC, SU and UN groups of peaks. B. Fold enrichment of positional distribution between SC peaks and all peaks around Transcription Start Sites. C. Percentage of SC, SU and UN peaks with consecutive bindings. 48 hr only represents the peaks that only bound in 48 hours of reprogramming, while 48 hr and pluripotent represents the peaks that are also bound in the reprogramming final stage. i represents the number of the two group of peaks. ii-iv represents the percentage of SC, SU and UN peaks that are either 48 hr only bound or 48 hr and pluripotent bound. D. Heatmap for percentages of mouse SC, SU and UN peaks in the mouse 18 chromatin states.

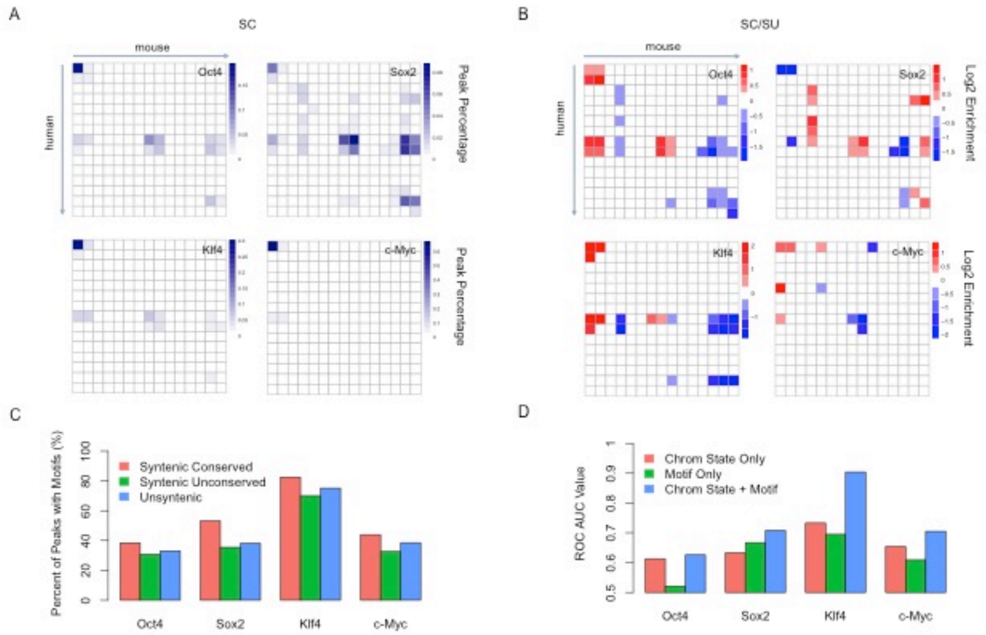
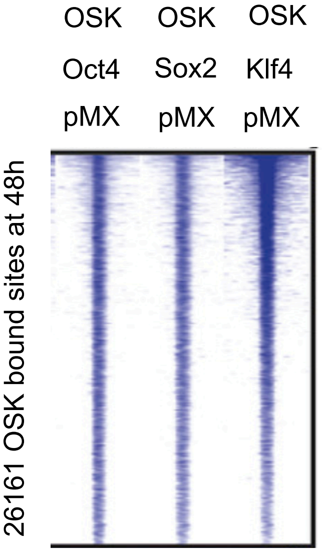


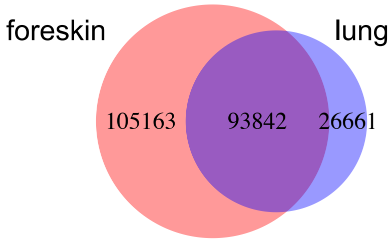
Figure 5. Chromatin state transitions, motif usages and their contributions in the maintaining of syntenic conserved peaks.

A. Chromatin state transitions of syntenic conserved peaks between human and mouse. The top left is state 1 of human to state 1 of mouse. The value in the heatmap represents the fraction of the number of syntenic conserved peaks in that square divided by the total number of all syntenic conserved peaks. B. Chromatin state transitions of the log2 ratio between syntenic conserved peaks versus syntenic unconserved peaks. The value in the heatmap represents the log2 ratio between the fraction of syntenic conserved peaks and the fraction of syntenic unconserved peaks in that square. C. Percentage of SC, SU and UN peaks that have canonical motifs. D. ROC AUC of a classifier to predict syntenic based on motif occurrences and chromatin state transitions.



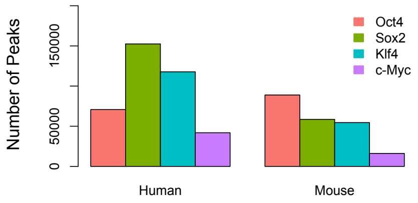
Supp Fig1. Similarities between individual retroviral based and poly-cistronic based system.

A. Heatmap of ChIP-Seq signal for Oct4, Sox2 and Klf4 using pMX (individual retroviral), for sites co-bound by OSK (polycistronic) at 48 hr of OSKM-induced reprogramming. The blue color represents ChIP-Seq signal. Each row represents an OSK co-bound peak. B. Venn diagram of mouse individual retroviral based (pMX) OSK peaks and polycistronic based OSK peaks. The numbers in the circle indicates the number of peaks.

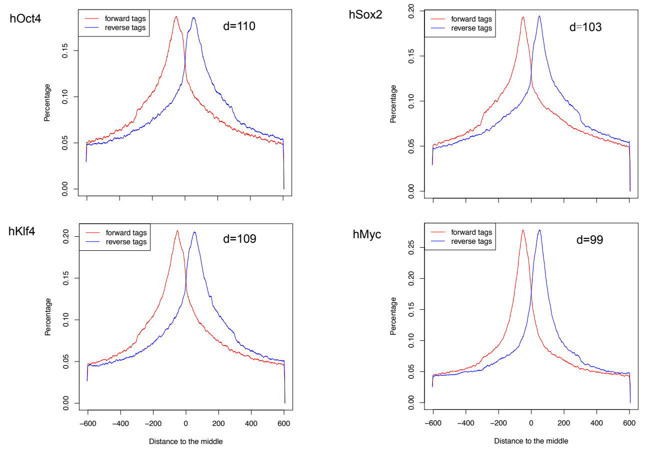


Supp Fig2. Venn diagram of DNaseI hypersensitive sites (broad peaks) between human foreskin newborn fibroblasts and human lung fetal fibroblasts.

The numbers in the circle indicates the number of peaks.

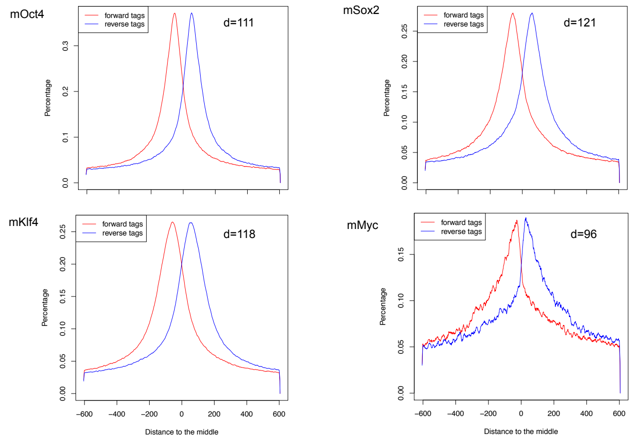


Supp Fig3. Bar plot of the number of the identified OSKM ChIP-Seq peaks in human and mouse

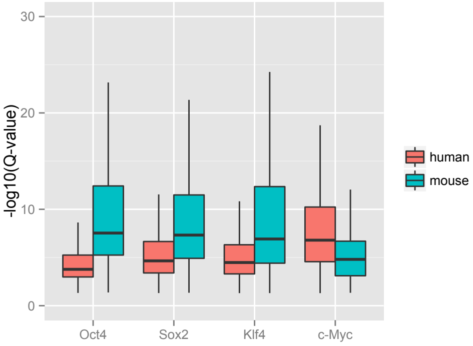


Supp Fig4. MACS2 peak calling models for human OSKM.

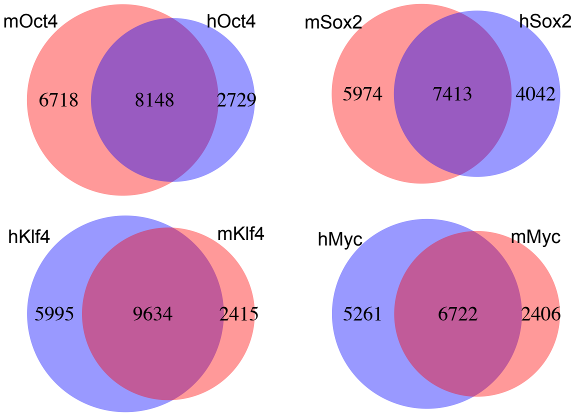
The model represents the average signal profiles for forward strand reads and reverse strand reads within top 1000 peaks. The d value in the figure represents the average fragment size between forward strand and reverse strand.



Supp Fig5. Same as Supp Fig4, except this is for mouse.

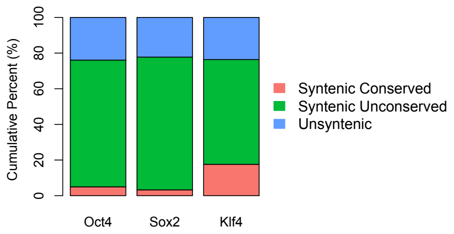


Supp Fig6. Boxplot of peak calling q-values for OSKM peaks in human and mouse. Q-value is calculated by MACS2 software to measure the false discovery rate of an identified peak.

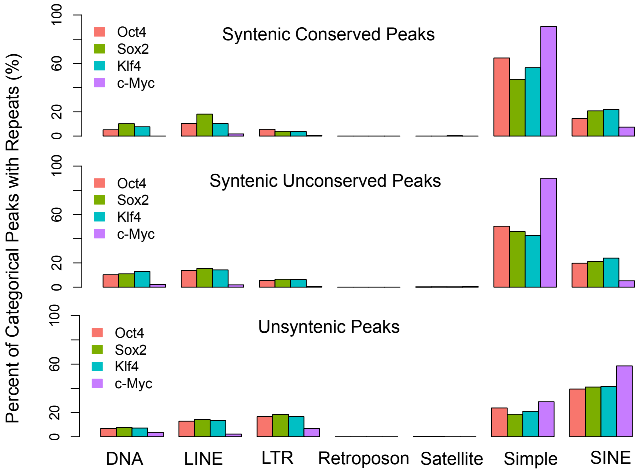


Supp Fig7. Venn diagram of OSKM target orthologous genes between mouse and human using 20 kb as a target gene cutoff.

Hypergeometric test shows that the number of shared orthologs is significant (p-value<10^{-16}) for all the four factors.

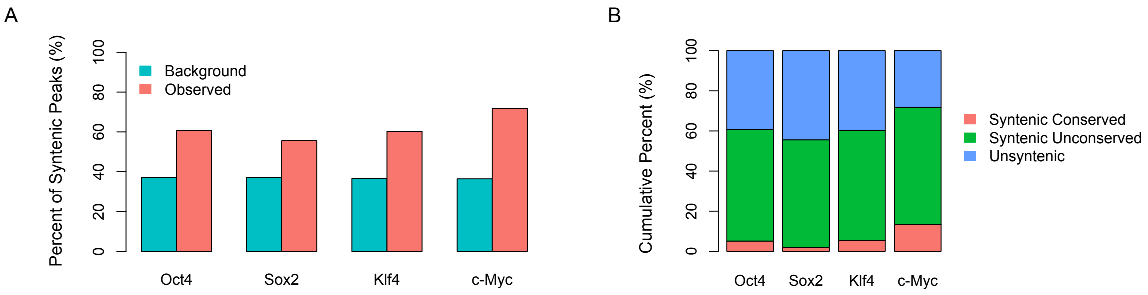


Supp Fig8. Distribution of mouse individual retroviral based (pMX) syntenic conserved, syntenic unconserved and unsyntenic peaks for Oct4, Sox2 and Klf4.



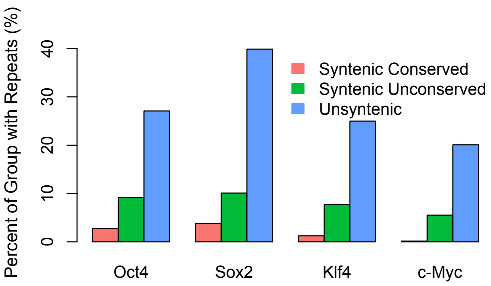
Supp Fig9. Percentages of mouse peaks that contain specific type of mouse repeat sequences.

Seven major types of repeat, i.e. DNA (DNA transposon elements), LINE (Long interspersed nuclear elements), LTR (Long terminal repeats), Retroposon (Transposons via RNA intermediates), Satellite (Satellite DNA which belongs to tandem repeats), Simple (Simple repeats) and SINE (Short interspersed nuclear elements) are calculated.

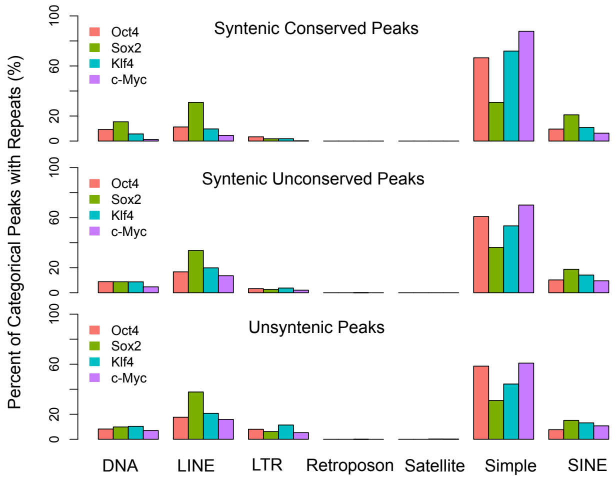


Supp Fig10. Map human OSKM peaks to mouse show limited conservation.

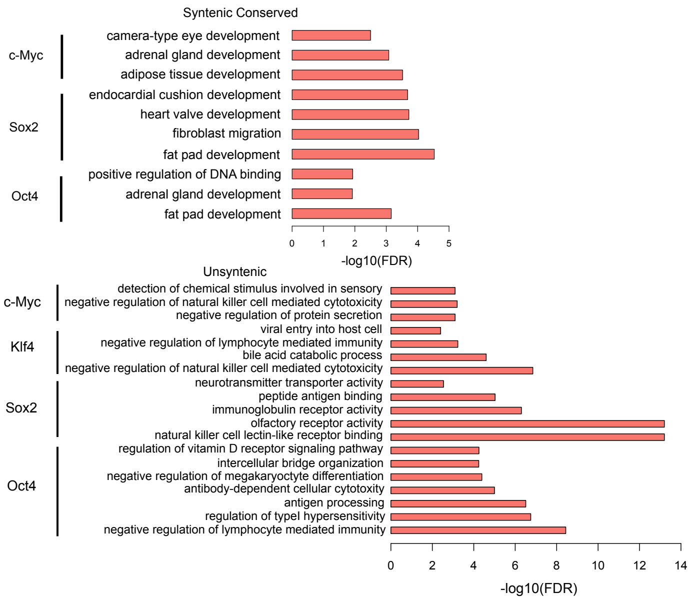
A. Percentage of human OSKM peaks that can be mapped to mouse. B. Distribution of human syntenic conserved, syntenic unconserved and unsyntenic peaks for each factor.



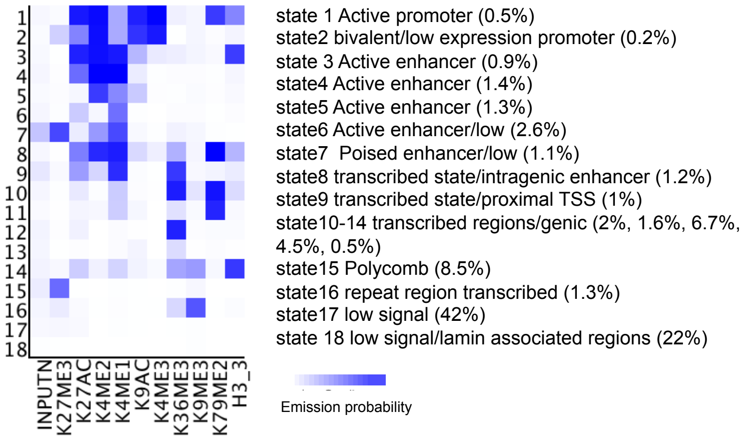
Supp Fig11. Percentage of the human syntenic conserved, syntenic unconserved and unsyntenic group of peaks that contain human repeat sequences



Supp Fig12. Same as Supp Fig9, except this is for human



Supp Fig13. Enriched gene ontology terms for genes near mouse SC and UN OSKM peaks



Supp Fig14. 18 chromatin state model for mouse 48 hours post induction of OSKM based on nine histone marks.

This figure is taken from Chronis et al. The value in the heatmap represents the enrichment of that histone mark in that learned chromatin state. The value in the brackets represents the percentage of genome that is occupied by that specific chromatin state.

References

1. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126**:663–676.

2. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S: **Induction of pluripotent stem cells from adult human fibroblasts by defined factors.** *Cell* 2007, **131**:861–872.

3. Hirschi KK, Li S, Roy K: **Induced pluripotent stem cells for regenerative medicine.** *Annu Rev Biomed Eng* 2014, **16**:277–294.

4. Singh VK, Kalsan M, Kumar N, Saini A, Chandra R: **Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery.** *Front Cell Dev Biol* 2015, **3**:2.

5. Takahashi K, Yamanaka S: **A decade of transcription factor-mediated reprogramming to pluripotency.** *Nat Rev Mol Cell Biol* 2016, **17**:183–193.

6. Yamanaka S: **Induced pluripotent stem cells: past, present, and future.** *Cell Stem Cell* 2012, **10**:678–684.

7. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126**:663–676.

8. Malik N, Rao MS: **A review of the methods for human iPSC derivation.** *Methods Mol Biol* 2013, **997**:23–33.

9. Soufi A, Donahue G, Zaret KS: **Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome.** *Cell* 2012, **151**:994–1004.

10. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS: **Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming.** *Cell* 2015, **161**:555–568.

11. Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, Ernst J, Plath K: **Cooperative Binding of Transcription Factors Orchestrates Reprogramming.** *Cell* 2017, **168**:442–459.e20.

12. Liu LL, Brumbaugh J, Bar-Nur O, Smith Z, Stadtfeld M, Meissner A, Hochedlinger K, Michor F: **Probabilistic Modeling of Reprogramming to Induced Pluripotent Stem Cells.** *Cell Rep* 2016, **17**:3395–3406.

13. Koche RP, Smith ZD, Adli M, Gu H, Ku M, Gnirke A, Bernstein BE, Meissner A: **Reprogramming factor expression initiates widespread targeted chromatin remodeling.** *Cell Stem Cell* 2011, **8**:96–105.

14. Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, Borkent M, Apostolou E, Alaei S, Cloutier J, Bar-Nur O, Cheloufi S, Stadtfeld M, Figueroa ME, Robinton D, Natesan S, Melnick A, Zhu J, Ramaswamy S,

Hochedlinger K: **A molecular roadmap of reprogramming somatic cells into iPS cells.** *Cell* 2012, **151**:1617–1632.

15. Nichols J, Smith A: **Naive and primed pluripotent states.** *Cell Stem Cell* 2009, **4**:487–492.

16. Ho R, Papp B, Hoffman JA, Merrill BJ, Plath K: **Stage-specific regulation of reprogramming to induced pluripotent stem cells by Wnt signaling and T cell factor proteins.** *Cell Rep* 2013, **3**:2113–2126.

17. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C-Y, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2014, **42**(Database issue):D142–7.

18. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, et al.: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518**:317–330.

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

20. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, Euskirchen G, Lin S, Lin Y, Visel A, Kawli T, Yang X, Patacsil D, Keller CA, Giardine B, Mouse ENCODE Consortium, Kundaje A, Wang T, Pennacchio LA, Weng Z, Hardison RC, Snyder MP: **Principles of regulatory information conservation between mouse and human.** *Nature* 2014, **515**:371–375.

21. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.** *Science* 2010, **328**:1036–1040.

22. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T: **Widespread contribution of transposable elements to the innovation of gene regulatory networks**. *Genome Res* 2014, **24**:gr.168872.113–1976.

23. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells**. *Nat Genet* 2010, **42**:631–634.

24. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nat Biotechnol* 2010, **28**:495–501.

25. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM: **Embryonic stem cell lines derived from human blastocysts.** *Science* 1998, **282**:1145–1147.

26. Lerou PH, Yabuuchi A, Huo H, Miller JD, Boyer LF, Schlaeger TM, Daley GQ: **Derivation and maintenance of human embryonic stem cells from poor-quality in vitro fertilization embryos.** *Nat Protoc* 2008, **3**:923–933.

27. Hockemeyer D, Soldner F, Cook EG, Gao Q, Mitalipova M, Jaenisch R: **A drug-inducible system for direct reprogramming of human somatic cells to pluripotency.** *Cell Stem Cell* 2008, **3**:346–353.

28. Park I-H, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, Daley GQ: **Reprogramming of human somatic cells to pluripotency with defined factors.** *Nature* 2008, **451**:141–146.

29. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.

30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.

31. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets.** *Bioinformatics* 2011, **27**:1696–1697.

32. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841–842.

33. Tong A-J, Liu X, Thomas BJ, Lissner MM, Baker MR, Senagolage MD, Allred AL, Barish GD, Smale ST: **A Stringent Systems Approach Uncovers Gene-Specific Mechanisms Regulating Inflammation.** *Cell* 2016, **165**:165–179.

34. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105–1111.

35. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31**:166–169.

36. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15**:550.

37. Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J: **Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution.** *BMC Genomics* 2015, **16**:87.

Chapter 3

A temporal transcriptome in human embryonic stem cell-derived cardiomyocytes identifies novel regulators of early cardiac development

3.1 Abstract

Stem cell based cardiogenesis has become a powerful tool to enhance our understanding of cardiac development and test novel therapeutics for cardiovascular diseases. However, this approach usually yields a high heterogeneity of cardiac cells that impede accurate research discoveries. We thus established a robust protocol that yields human cardiomyocytes (hCM) with more than 90% purity from human Embryonic Stem Cells (hESC). To take advantage of our protocol, we systematically examined how gene expression and epigenetic program changes at temporal developmental stages during cardiogenesis. Our results then provide a comprehensive view of expression changes during cardiogenesis, allowing us to identify key transcription factors as well as lincRNAs that are strongly associated with cardiac differentiation. Moreover, we incorporated a simple but powerful method to screening for novel regulators of cardiogenesis solely based on expression changes. As a result, we found four novel cardiac-related transcription factors, i.e. SORBS2, ZNF436, DPF3 and MITF, which have no or few literature reports. Our strategy of identifying novel regulators of cardiogenesis can also be easily implemented in other stem cell based systems. In summary, our results provide a valuable resource for understanding cardiogenesis and identify four novel cardiac-related transcription factors.

3.2 Introduction

Recent advances in the method of directed differentiation using human pluripotent stem cells (hPSCs) to generate enriched cardiomyocytes in a dish have provided a great platform for not only studying human development but also disease mechanism and translational research [1-4]. Yet, transcriptional and epigenetic regulation of multiple transitional stages from pluripotent cells to committed cardiomyocytes has not been fully characterized with hPSC-CMs differentiated in a chemically defined manner, which can be the most suitable model for the future studies in regard to the scalability and the cost-effectiveness.

Formation of the mature mammalian heart is governed by intricate gene regulatory network: precise temporal and spatial gene expression dictates cell fate. Transcription factors (TFs) regulate both activation and repression of the genes in cell lineage specification and differentiation [5]. A zinc finger transcription factor GATA4 is one of the major TFs that determines cardiac cell fate and its mutation results in congenital heart disease [3].

Here, we performed systematic analysis of transcriptome and histone modification of five stages of directed cardiac differentiation of human embryonic stem cells (hESCs). We utilized chemically defined directed differentiation of cardiomyocytes from hESCs, which reproducibly yields over 90% of cardiomyocytes [6-8]. That enabled us to profile the stage specific changes in global transcriptome and epigenetic changes during cardiac differentiation without further purification step of differentiated cell population. The direct comparison of our data to the ones from past reports [1, 9] validated the highly efficient cardiac differentiation system by demonstrating higher enrichment of cardiac gene expressions in our system. By using the induced rate scoring model, we discovered previously understudied transcription factors: MITF, SORBS2, DPF3, and ZNF436, which regulatory functon of cardiac genes are further validated by RNAi gene knockdown. This work provided a comprehensive view of gene expression and epigenetic changes during cardiac differentiation and proposed a simple but powerful screening method of novel gene regulators, which is easily transferrable to other organ development.

3.3 Results

3.3.1 Differentiation of high purity Cardiomyocytes from Embryonic Stem Cells

Heterogeneity is a common bottleneck for research using stem cell derived cell lines. To overcome this problem in the in-vitro modeling of cardiogenesis, we utilized a chemically defined differentiation protocol, which reproducibly yields ~90% of cardiomyocytes [6, 8]. The purity of cardiomyocytes generated with our protocol thus out-performs previous published studies. To characterize the differentiation process, we performed serially mRNA expression profiling with RNA-Seq from H9 hESC-CMs at five distinct stages: undifferentiated stage (hESC, day 0); mesodermal precursor stage (hMP, day 2); cardiac progenitor stage (hCP, day 5);

immature cardiomyocyte (hCM15, day 15); and hESC-CM differentiated for 15 additional days (hCM30, day 30). Replicates were highly correlated and we thus merged them for the downstream analysis.

To further prove the purity and advantages of our system, we compared fold enrichments between cardiac signature gene expression and average gene expression at CM stage in our data and data obtained from Paige et al and Liu et al. [1, 9]. As a result, for all the signature genes we tested, our system showed higher expression enrichments, indicating our data is composed with stronger cardiac signals. Inspired by those observations, we then sought to perform a comprehensive analysis to illustrate a detailed cardiogenesis process and screening for novel cardiac regulators using our system.

3.3.2 Gene expression changes during ESCs to Cardiomyocytes

We first investigated the general pattern of genome-wide gene expression changes during cardiomyocyte differentiation. For all the samples, the normalized read counts were used to represent the expression values for each gene (see Methods for more details). To characterize the expression changes, expression values were subtracted by their means across the five stages. Figure 1b then shows the heatmap for the k-means clustering of all the genes with their expression changes. Genes in cluster 1 (with 1,755 genes) show a clear pattern of decreasing gene expression, while genes in cluster 3 (with 1,005 genes) show the opposite.

Gene Ontology (GO) enrichment analysis showed that cluster 1 genes were significantly enriched in GO terms related with cell proliferation and stem cell population maintenance, and cluster 3 genes were significantly enriched in GO terms of cardiac muscle contraction and heart development. Interestingly, we also found genes in cluster 2 (with 1,331 genes) showed an increasing pattern but remained largely unchange from day 5 to day 30. Those genes were significantly enriched in GO terms of anterior/posterior pattern and heart development. The remaining genes in cluster 4 and cluster 5, which represented 81 percent of all genes, showed limited or no changes during cardiomyocytes differentiation.

An example cluster 1 gene is POU5F1 (also known as OCT4), which is a key signature gene for stem cells (9). The expression of POU5F1 dramatically decreased at day 5 as the cells developed into cardiac progenitor stage (Fig 1d). In contrast, MYL7 in cluster 3, a key signature gene for cardiac muscle cells, was expressed at day 5 and further induced at later stages (Fig 1d). Our data thus captured both expression changes and exon information of those genes, providing an opportunity to expand our analysis in more details.

We next sought to identify upstream regulators that bound to gene promoter regions for different clusters of genes. We thus used Ingenuity software [10], which collected empirical information of upstream regulators for a large set of genes, to searching for significantly enriched regulators. Figure 1c shows the identified regulators and their statistical p-values. For cluster 1 genes, we found POU5F1, SOX2, IFNG and CREB1. For cluster 2 genes, TGFB1 and WNT3A were identified. For genes in cluster 3, we found TBX5, MEF2C, HNF1A, MYOCD and DNMT3B.

Besides the clustering of all genes, we performed differential expressed genes analysis between each neighboring stages (Fig 1e)(see Methods). Most differential expressed genes occurred within day2-day5 and day5-day15 comparisons. Gene ontology enrichment analysis confirmed that there were increasing expressions of cardiac-related genes and decreasing expressions of stem cell genes. Notably, gene expression changes for day15-day30 were much smaller than the other three comparisons, indicating the limited degree of expression changes in cells from day 15 to day 30. All together, our results revealed the general characteristics of how gene expression changes during cardiogenesis.

3.3.3 Transcription factors expression changes during ESCs to Cardiomyocytes

Transcription factors (TFs) are the drivers of gene expression program. Having shown the global pattern of gene expression changes, we then focused our analysis on the transcription factors. By doing so, we investigated the expression changes of 1,835 human TFs from a comprehensive curated list [11]. We first asked what transcription factors varied most during cardiogenesis. Figure 2A shows the top 50 TFs with largest variances of normalized expression values among the five stages. Canonical pluripotency TFs, such as POU5F1, SOX2, NANOG, were gradually repressed, while canonical cardiac TFs, such as HAND2, TBXs, NK2s, and MEF2C, were gradually induced. Mesodermal TFs including T, MIXL1, MESP1, and EOMES were peaked at day 2. This result provides a set of most dynamic TFs during cardiogenesis, indicating their potentially important regulatory functions.

Since the differentiation process includes multiple developmental stages of cardiogenesis, we sought to identify stage-specific TFs that associate with each individual time point. By doing so, we calculated the ratios between normalized expression values of that stage and the sum values of all stages for each TF. The proportions then represented the stage-specific gene expression ratios. A TF was considered to be stage-specific only if it showed a proportion of at least 0.6, which is 3 fold of the average proportion 0.2 across the five stages. Under our criteria, we identified 16 ESC-specific, 23 MP-specific, 12 CP-specific, 2 CM15-specific and 8 CM30-specific TFs (Fig. 2B). The gene sets of ESC- and MP-specific genes are similar to the ones identified in Fig. 2A. However, this analyses revealed set of genes that were not identified in Fig. 2A, such as HOXBs at day 05 and PRDM16 and NFIX at day 30. This result thus defined a number of cell type specific TFs associating with distinct developmental stages from ESCs to Cardiomyocytes.

3.3.4 Screening for novel regulatory transcription factors during Cardiogenesis

To identify transcription factors regulating the activation of cardiac-related pathways, we hypothesized that those TFs would show significant gene expression inductions during cardiogenesis. With this hypothesis, we thus used a scoring system based on induced rates between day 30 and day 0 to sort the 1,835 curated TFs (see Methods). Figure 3A then shows the top 25 TFs with their induced rate scores. For example, HAND2, as the top 1 factor based on our ranking, plays an essential role in cardiac morphogenesis. Genome browser view of this gene revealed enormous expression at CM30 stage while limited expression at ESC stage (Fig. 3B). It then showed an induced rate of 139 during cardiogenesis.

Surprisingly, among the top 25 factors in Figure 3A, 21 of them are cardiac-related TFs with extensive literature supports. In Paige et al., the authors used a combinatorial model of histone modifications and gene expression changes to predict cardiac regulatory transcription factors [1]. From the ranking list based on their scoring model, 20 of top 25 TFs were supported by extensive literatures (Fig 3C). In addition, the two ranking lists had an overlaps of 11 factors among the top 25 hits (Fig 3d), which is highly significant considering of the total number of TFs. However, our method only required cardiomyocytes with high purity and RNA-Seq expression profiling of undifferentiated and fully differentiated cells. By contrast, serial ChIP-Seq of several histone modifications and RNA-Seq data were the prerequisite in order to use the model proposed by Paige et al. Our method for screening novel regulators thus is much more implementable and cost-efficient, while enabling to retain a high sensitivity and high specificity.

3.3.5 Identification of lincRNAs associated with cardiogenesis

Long intergenic non-coding RNAs (lincRNAs) play important regulatory roles in various cell differentiation processes by activating or repressing their neighboring genes. We then extended our analysis from protein coding genes to lincRNAs. We used a curated list of human lincRNAs and re-mapped RNA-Seq reads to each lincRNA. As the same with protein-coding genes, the normalized read counts falling in each lincRNA region were then used to represent its expression value.

We next computed the variance of lincRNA normalized expression values among the five developmental stages. Figure 4a then shows the top 50 lincRNAs that have the largest variances within our list. Moreover, we also calculated an induced rate for each lincRNA between CM30 and ESC. Figure 4b shows the ones with an induced rate of 10 or higher. Genome browser viewing of those lincRNAs revealed dramatic expression activation in later developmental stages during cardiogenesis. All together, this result provided a valuable resource for the communities to further validate the functions of those lincRNAs during cardiogenesis.

3.3.6 Genome-wide methylation changes during cardiogenesis

To measure epigenetic changes during cardiogenesis, we performed reduced-representative bisulfite sequencing (RRBS) for four stages of hESC, hMP, hCP and hCM15. We then identified 133,912 RRBS fragments covering at least 3 CpG sites across samples (see Methods for more details). To obtain the differential methylated fragments, we further selected fragments that showed delta methylation changes of 0.2 or higher relative to the average methylation levels. As a result, 3,890 of differential RRBS fragments were identified for downstream analysis.

We next performed hierarchical clustering on the 3,890 RRBS fragments (Fig 5a). In clusters 1 and cluster 2, we observed a clear decrease of methylation levels from ESC to CM. While in cluster 3, we observed an opposite trend. We next asked what genes were close to the differential methylated regions within each cluster. As we expected, gene ontology analysis revealed that cluster 1 regions were enriched in heart morphogenesis, cluster 2 regions were enriched in cardiac muscle development, while cluster 3 regions were enriched in stem cell related terms. Notably, these differential methylated fragments tend to occur in distal transcription start sites (TSS) regions, instead of proximal TSS regions, indicating the methylation of enhancers might be affected.

DNA methylation is maintained by DNMT3a, which may affect the binding of various transcription factors. We thus asked whether there were enriched TF binding sites in the three groups of the differential methylated regions. By performing motif discoveries in these regions, we identified a total of 14 TFs and their motif sequences with a stringent q-value cutoff of 10^{-15} (Fig 5b)(see Methods). Among them, we were able to re-capture well-characterized TFs, such as GATA4 in cluster 1, GATA4, TBX20, and MEIS1 in cluster 2, and OCT4 in cluster 3. Thus, there is a clear trend that DNA methylation status is associated with the gene expression genome-wide.

3.3.7 Identification of novel regulators in cardiac development

Four remaining factors from the top 25 of our list in Figure 3A, i.e. SORBS2, ZNF436, DPF3 and MITF, had no or limited literature support showing they were cardiac regulators. We thus decided to perform experimental validations by knocking down the four TFs with siRNAs and see whether the expression of cardiac marker gene will be affected or not. Knockdown efficiency was 70 % for MITF, 90 % for SORBS2, 50 % for DPF, and 56 % for ZND436. As a result, cardiac genes were downregulated significantly by their knockdown (Figure 6). Majority of the cardiac sarcomeric genes were all down-regulated especially by MITF and SORBS2, validating their roles as cardiac regulators. The cardiac transcription factors were downregulated but in lesser extent than sarcomeric genes also varied in each knockdown, suggesting some gene specificities existed. For example, HAND2 was significantly downregulated by DPF3 knockdown but not by MITF even though knockdown efficiency of MITF is higher than that of DPF3. The relation of these novel factors with the known regulatory gene network needs to be addressed in the future.

3.4 Discussion and conclusion

In this work, we utilized a high purity ESC-based cardiomyocytes to study the early development of cardiac. Our analysis focused on the identification of transcription factors and lincRNAs that are strongly associated with the temporal development of cardiogenesis. As a result, we were able to capture both well-characteristic regulatory factors as well as identify novel ones. Therefore, our results provide a comprehensive picture of expression changes for important cardiac regulators.

There are two major factors influencing the study of hPSC-derived systems. One is which assay, e.g. RNA-Seq, ATAC-Seq, BS-Seq, ChIP-Seq, to choose in order to study the biological mechanisms of the differentiation process. Another is the efficiency of the differentiation protocol. In this study, we showed that high purity differentiated cells plus RNA-Seq expression changes were able to capture the regulatory factors in a large extent. This strategy thus is highly implementable and cost-efficient to screening for novel regulators.

With the above reasons, we expect this differentiation protocol would be a valuable approach to further investigates the early development of cardiac cells. Although we found four novel transcription factors associated with cardiogenesis, in vivo experiments would be needed to further address their functions. Moreover, how transcription factors, epigenetic enzymes, non-coding RNAs interact with each other in a regulatory network would be the next step to understand cardiogenesis in a system biology way. This knowledge will ultimately lead to novel therapeutic and drug target development in the diagnosis and treatment of heart related diseases.

3.5 Materials and Methods

Cell cultures from Embryonic Stems Cells to Cardiomyocytes

H9 (WA09) hESC lines were maintained as described before [7]. Authentication of hESCs was achieved by confirming the expression of pluripotency genes and protein markers. hESCs were routinely verified as mycoplasma-free using a PCR-based assay. hESCs were grown and differentiated in a chemically defined condition [6]. Usage of all the human embryonic stem cell lines is approved by the UCLA Embryonic Stem Cell

Library preparations and sequencing

RNA was extracted from the cells of five stages namely, hESC, hMP, hCP, hCM14, and hCM28, using TRIZOL (TheroFisher) and RNeasy kit (QIAGEN) according to manufacturer’s protocol. 500 ng of DNaseI-treated RNA was used as input material for library preparation using the Illumina TruSeq mRNA kit (Illumina, RS-122–2001), according to manufacturer’s instructions. Final libraries were sequenced as Sequencing was performed on an Illumina HiSeq 3000 for a paired end 2x150 run.

RNA-Seq analysis

We first used tophat to map the RNA-Seq reads back to the human genome (hg19) [12]. After that, we utilized HTseq software to calculate the number of reads falling in each gene [13]. With the RNA-Seq read counts matrix, we then used DESeq2 package of R to perform data normalization (rlog function) and differential expressed genes (DEG) analysis [14]. For DEG analysis, we set a cutoff of FDR <= 0.01 and fold change >= 2. Replicate samples of each stage were highly correlated. We thus merged them for downstream analysis.

Induced rate calculation

To search for the most changed genes from ESCs to CMs, we introduced the calculation of induced rate. An induced rate (IR) was calculated as the ratio of Reads Per Million between day 30 (y_i) and day 0 (x_i) for each gene *i*. We also added a small pseudo read count to day 0, resulting:

$$IR = y_i/(x_i+1)$$

We then ranked all the transcription factors and lincRNAs based on their induced rates, enabling us to re-capturing both known cardiac factors and screening for novel ones.

siRNA knock-down and functional validation of novel regulators

hESC-derived CMs were transfected with siRNA Negative Control (Qiagen) or human MITF, SORBS2, DPF3, and ZNF436 targeting siRNA 40 nM (MITF and SORBS2; Qiagen, DPF3 and ZNF436; ThermoFisher) using lipofectamine RNAi MAX reagent (ThermoFisher) according to the manufacturer’s instructions. The medium was changed 48 hr after transfection, and cells were then incubated for an additional 7 days.

Reduced-reprehensive bisulfite sequencing (RRBS) analysis

We used BS-Seeker2 software to mapping the RRBS data back to human genome (hg19) and calculating methylation levels for each CpG [15]. The RRBS data then covered 1,044,850 of CpG sites across the samples. To robustly estimate the methylation level for each CpG, we filtered the CpG sites by requiring they were covered by at least 10 reads. For each RRBS fragment, we then calculated the average methylation levels of CpG within this fragment. The average value was thus assigned to this fragment to represent the fragment methylation level. Lastly, we further filtered the fragments that had less than 3 CpG sites. This analysis pipeline then allowed us to identify 133,912 RRBS fragments covering at least 3 CpG sites across samples.

Motif discoveries within differential methylated regions

The RRBS fragments had a medium size around 500 base pairs. Based on the hierarchical clustering result, we performed motif discovery analysis using HOMER software to RRBS fragment for regions in each cluster [16]. HOMER calculated the statistical significance of motif occurrences for observed regions to a large set of known TF motif usages. Lastly, we set a stringent cutoff of 10^{-15} FDR to filter the identified motifs.

Figures

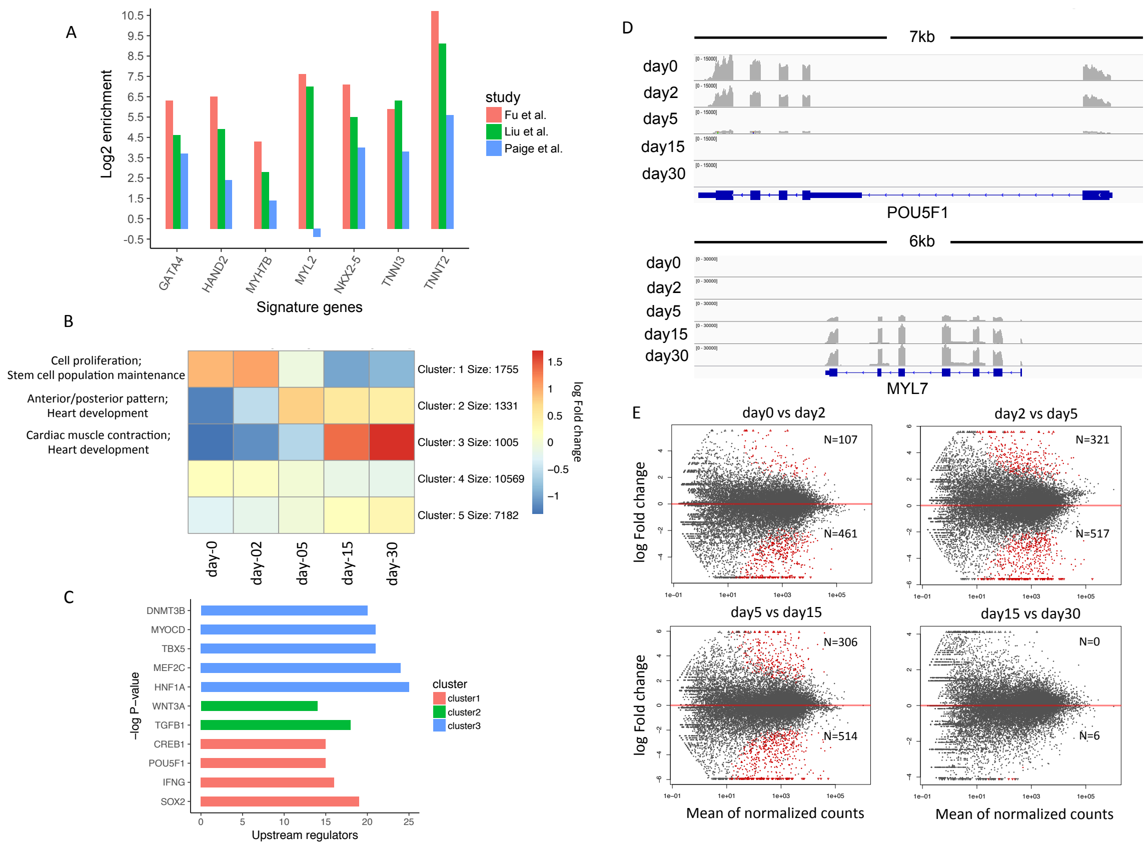


Figure 1. Global gene expression changes during cardiogenesis.

A. Comparison of signature gene expression enrichments for cardiomyocytes between our study and Paige et al. The y-axis represents the log 2 fold change between expression values of signature genes and average expression values of all genes. B. K-means clustering of normalized gene expression values for samples among the five developmental stages. The color in the heatmap represents the log 2 fold change of expression values. The red color represents a higher expression value than average expression across samples, while the blue color represents the opposite. Text on the left of heatmap shows the enriched gene ontology terms for each cluster of genes. C. Ingenuity analysis identifies statistical significant upstream regulators for cluster of genes showed in B. D. Genome browser view of RNA-Seq data for POU5F1 gene and MYL7 gene. Each track shows the expression profile for a different stage. E. MA-plot of differential expressed genes for neighboring stage comparisons (A VS B). The number N shows the number of up and down regulated genes. The number above the red line represents the number of DEGs that is up regulated in A.

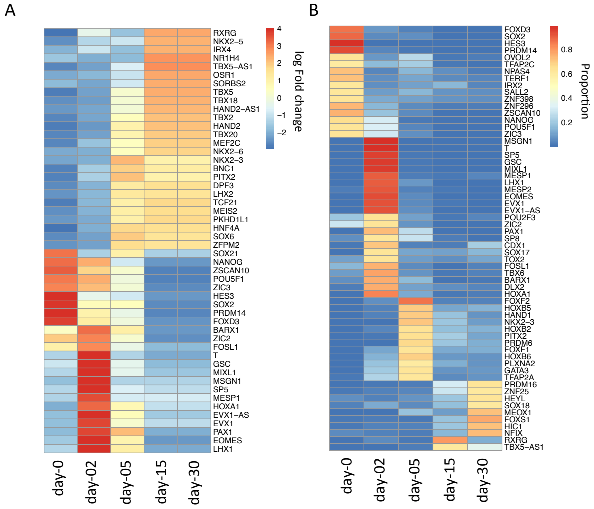


Figure 2. Transcription factors expression changes during cardiogenesis.

A. Heatmap for top 50 TFs showing the largest expression variations from ESCs to Cardiomyocytes. Each column represents a different sample and each row represents a different gene. The color in the heatmap again represents the log 2 fold change of normalized expression values. B. Heatmap for stage-specific TFs from ESCs to Cardiomyocytes. The color in the heatmap represents the proportion of expression value for that sample to the sum of all samples.

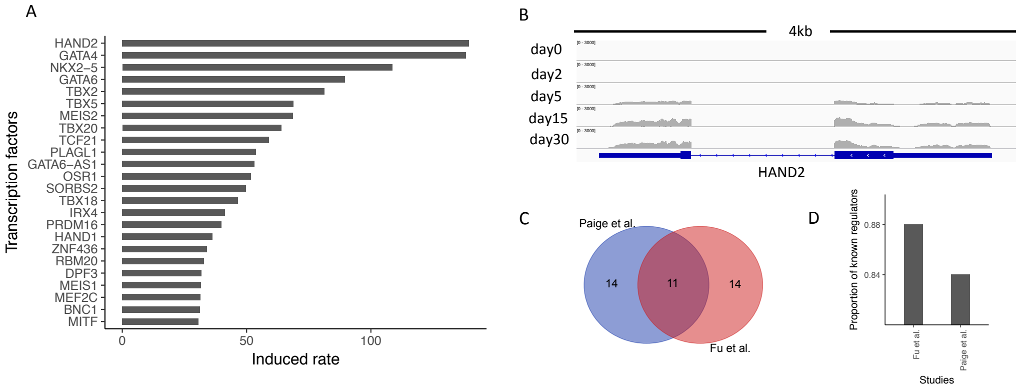


Figure 3. Screening for novel cardiac related transcription factors.

A. Bar-plot of top 25 transcription factors (y-axis) ranked by their induced rate (x-axis). B. Genome browser view of RNA-Seq data for HAND2. C. The number of known cardiac TFs in top 25 hits based on the ranking calculation by our study and by Paige et al. D. Venn-diagram of top 25 transcription factors identified by our ranking approach and by Paige et al.

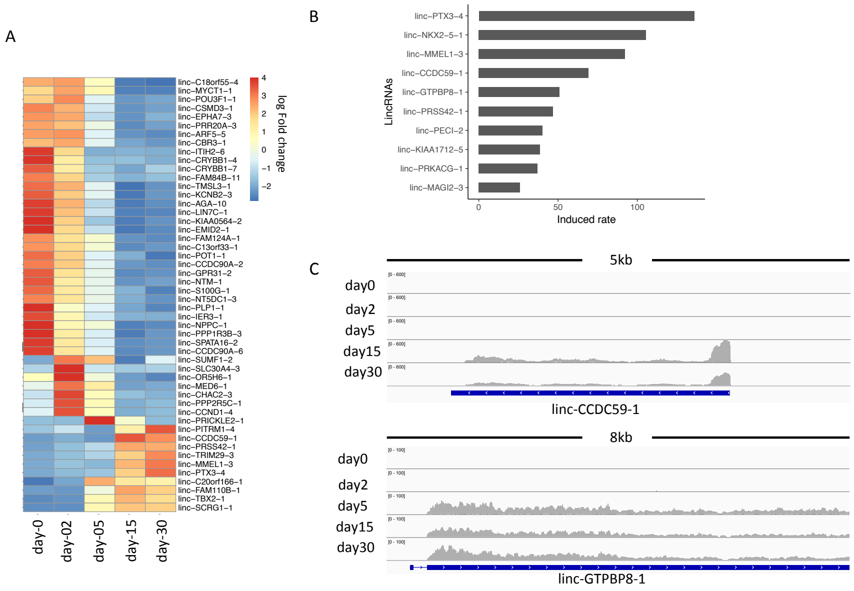


Figure 4. LincRNA expression changes during cardiogenesis.

A. Heatmap for top 50 lincRNAs showing the largest variations from ESCs to Cardiomyocytes. The color in the heatmap represents the log 2 fold change of normalized lincRNA expressions. B. Bar plot of lincRNAs that show at least 10 fold of induced rates. C. Genome browser view of RNA-Seq data for linc-CCDC59-1 and linc-GTPBP8-1 during cardiogenesis.

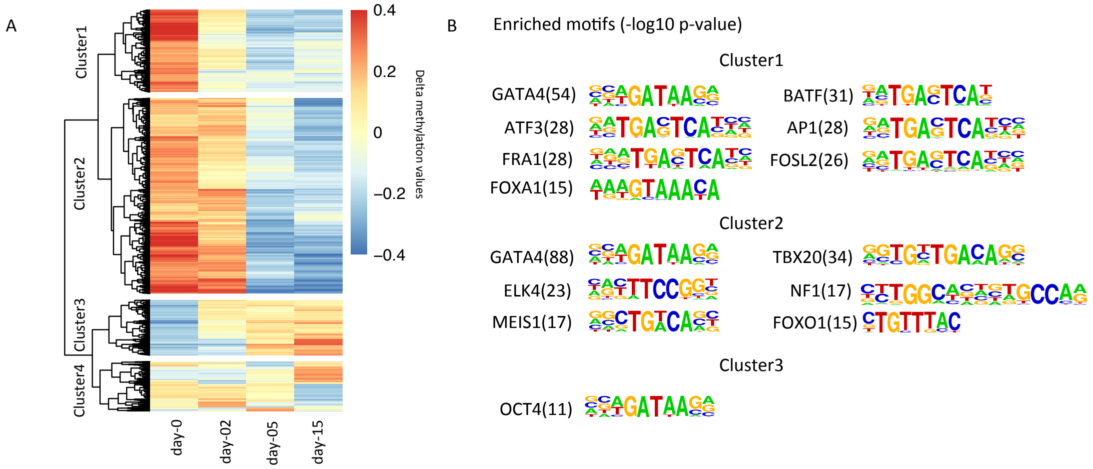


Figure 5. RRBS-based DNA methylation changes during cardiogenesis.

A. Hierarchical clustering of RRBS fragments that show at least 0.2 delta methylation changes compared to average methylation levels across samples. The color in the heatmap represents the delta methylation value. The red color represents hypo methylation while blue color represents hyper methylation. B. Enriched transcription factors and their motif sequences found in each cluster of RRBS fragments. The number in the bracket shows the –log10 p-value of the statistical significance for motif discovery.

References

1. Paige SL, Thomas S, Stoick-Cooper CL, Wang H, Maves L, Sandstrom R, Pabon L, Reinecke H, Pratt G, Keller G, Moon RT, Stamatoyannopoulos J, Murry CE: **A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development.** *Cell* 2012, **151**:221–232.

2. Lan F, Lee AS, Liang P, Sanchez-Freire V, Nguyen PK, Wang L, Han L, Yen M, Wang Y, Sun N, Abilez OJ, Hu S, Ebert AD, Navarrete EG, Simmons CS, Wheeler M, Pruitt B, Lewis R, Yamaguchi Y, Ashley EA, Bers DM, Robbins RC, Longaker MT, Wu JC: **Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells.** *Cell Stem Cell* 2013, **12**:101–113.

3. Ang Y-S, Rivas RN, Ribeiro AJS, Srivas R, Rivera J, Stone NR, Pratt K, Mohamed TMA, Fu J-D, Spencer CI, Tippens ND, Li M, Narasimha A, Radzinsky E, Moon-Grady AJ, Yu H, Pruitt BL, Snyder MP, Srivastava D: **Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis.** *Cell* 2016, **167**:1734–1749.e22.

4. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG: **Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage.** *Cell* 2012, **151**:206–220.

5. Luna-Zurita L, Stirnimann CU, Glatt S, Kaynak BL, Thomas S, Baudin F, Samee MAH, He D, Small EM, Mileikovsky M, Nagy A, Holloway AK, Pollard KS, Müller CW, Bruneau BG: **Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis.** *Cell* 2016, **164**:999–1014.

6. Nakano H, Minami I, Braas D, Pappoe H, Wu X, Sagadevan A, Vergnes L, Fu K, Morselli M, Dunham C, Ding X, Stieg AZ, Gimzewski JK, Pellegrini M, Clark PM, Reue K, Lusis AJ, Ribalet B, Kurdistani SK, Christofk H, Nakatsuji N, Nakano A: **Glucose inhibits cardiac muscle maturation through nucleotide biosynthesis.** *Elife* 2017, **6**:025003.

7. Zhu H, Scharnhorst KS, Stieg AZ, Gimzewski JK, Minami I, Nakatsuji N, Nakano H, Nakano A: **Two dimensional electrophysiological characterization of human pluripotent stem cell-derived cardiomyocyte system.** *Sci Rep* 2017, **7**:43210.

8. Minami I, Yamada K, Otsuji TG, Yamamoto T, Shen Y, Otsuka S, Kadota S, Morone N, Barve M, Asai Y, Tenkova-Heuser T, Heuser JE, Uesugi M, Aiba K, Nakatsuji N: **A small molecule that promotes cardiac differentiation of human pluripotent stem cells under defined, cytokine- and xeno-free conditions.** *Cell Rep* 2012, **2**:1448–1460.

9. Liu Q, Jiang C, Xu J, Zhao M-T, Van Bortle K, Cheng X, Wang G, Chang HY, Wu JC, Snyder MP: **Genome-Wide Temporal Profiling of Transcriptome and**

Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. *Circ Res* 2017, **121**:376–391.

10. Krämer A, Green J, Pollard J, Tugendreich S: **Causal analysis approaches in Ingenuity Pathway Analysis.** *Bioinformatics* 2014, **30**:523–530.

11. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R: **TFCat: the curated catalog of mouse and human transcription factors.** *Genome Biol* 2009, **10**:R29.

12. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105–1111.

13. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31**:166–169.

14. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15**:550.

15. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen P-Y, Pellegrini M: **BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.** *BMC Genomics* 2013, **14**:774.

16. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38**:576–589.

Chapter 4

Integrated modeling of DNA methylation with core histone modifications in various human cells

4.1 Abstract

DNA methylation and histone modifications are the major two epigenetic mechanisms in mammalian cells. Previous studies have revealed those two mechanisms exhibit a crosstalk in the regulation of gene expression. However, those evidences are in a descriptive and qualitative way. In this project, we thus sought to systematically evaluate the quantitative relationship between DNA methylation and the major histone modification marks in human. Our analysis integrated 35 whole genome bisulfite sequencing (about 800 million CpG sites) and 175 ChIP-Seq histone modification assays for 35 human cell types. The logistic regression model we built show that there is an universal quantitative relationship between DNA methylation and histone modification in human. Importantly, we find that H3K4me3 is a dominant predictor of DNA methylation. Interestingly, our result suggests that the power for histone modification based prediction of DNA methylation varies among different type of cells, where pluripotent cells tend to have higher predictive power. Lastly, we show that two H3K27me3 associated chromatin states, i.e. bivalent enhancer and repressed polycomb show distinct residual predicted DNA methylation values compared with other states. In summary, our results provide a comprehensive evaluation of the quantitative crosstalk between DNA methylation and histone modification in a variety of human cell types.

4.2 Introduction

DNA methylation is a major epigenetic mechanism of gene regulation [1]. This epigenetic process plays an important role in silencing of transposon, X chromosome inactivation and regulation of gene expression. In mammalian cells, DNA methylation are regulated by DNA methyltransferases (DNMTs) [2]. Specifically, DNMT1 is a maintenance DNMT, DNMT3a/3b are de novo DNMTs and DNMT3L is an inactive member that increase the catalytic activity of DNMT3a/b. There are three different domains in active de novo DNMTs: the catalytic domain, an ADD domain and a PWWP domain [3]. Previous studies have found ADD domains preferentially bind histone 3 tails lacking methylation at lysine 4 (H3K4me), while the PWWD domain bind to histone 3 tails lacking methylation at lysine 36 (H3K36me). The functional unit within DNMTs then forms the basis of interaction between histone modifications and DNA methylation.

Over the past decade, whole-genome bisulfite sequencing (WGBS) technology has been developed to capture the methylation level for CpG sites across the genome [4, 5]. This technology has been widely applied to map methylome in a variety of cellular stages for a number of species. Besides studying of epigenetic mechanisms, DNA methylation can also be used as a robust biomarker in human complex diseases such as cancer. One example of this is methylation profiling of cell-free circulating DNA enables the identification of cancer and its tissue of origin in a non-invasive way [6]. Another example of using methylome in clinics is to classify brain tumor patients into groups that have different treatment options and prognosis progress [7]. Despite those advances, it is still expensive for a typical laboratory to perform whole-genome bisulfite sequencing since the estimation of CpG methylation value requires re-sequencing the genome in a high coverage (usually more than 20x). This has limits the application of WGBS assay as a routine method in study of epigenetic changes for the wide biomedical research community. Thus a method that is capable of predicting genome-wide methylation can be highly valuable.

In this study, we thus sought to systematically evaluate the relationships between histone modifications and DNA methylation from a quantitative perspective. We chose five histone marks: H3K4me3, H3K4me1, H3K27me3, H3K36me3 and H3K9me3 as the core histone marks. Those markers have been shown to represent a large part of histone code and can be used to identify the major chromatin states for genomic annotation. We then use a multiple logistic regression model to predict DNA methylation with the core histone marks in 35 human cells and tissues. We show that histone modifications are highly predictable of methylome and H3K4me3 is a predominant predictor. In addition, the significant predictive power exists in all human cell lines or tissues we examned, indicating there is a universal relationship between histone modifications and DNA methylation in human. Interestingly, we observed a variation of predictive power where pluripotent cells are generally having a stronger values compared with cells from tissues. Lastly, we found two H3K27me3 associated chromatin states, i.e. bivalent enhancers and repressed polycomb showing distinct predictive methylation values compared with other chromatin states. Our work thus provides a comprehensive and quantitative evaluation between histone modifications and DNA methylation in a variety type of human cells.

4.3 Results

4.3.1 Materials and data integration

To systematically evaluate the quantitative relationship between core histone modifications and DNA methylation, we used datasets from Roadmap Epigenomics Project of 35 human cells lines and tissues [8]. This enables us to perform an extensive inspection. We chose to use datasets from Roadmap Epigenomics Project for two reasons. First, the consortium produced a large number of high quality whole genome bisulfite sequencing (WGBS) as well as corresponding core ChIP-Seq histone marks for various of human cell lines and tissues. Second, the datasets were generated in a consistent way, allowing a robust integrated analysis and limiting the technical variations between data samples.

We thus used 35 WGBS samples generated by the project to represent the methylome for a variety of cells. Each sample contains an average of about 25 million CpG sites with methylation calls, leading to a total of about 800 million CpG sites. To compare between different histone marks, we used the normalized fold change values, i.e. normalized fold change between specific histone ChIP-Seq samples and control, to represent the histone modification signals for each genomic coordinates. As a result, 175 normalized ChIP-Seq samples were used to represent the histone modifications. In addition, we also used chromatin states datasets to characterize the genomic annotation of methylated CpG sites. The chromatin states were learnt from a hidden markov model to represent the combinatorial pattern of histone marks.

We next sought to integrate those datasets together. Figure 1 shows the schematic illustration of our integrative approach. For each CpG sites with methylation calls, we first calculated its neighboring 200 bp (centered on CpG site) of average ChIP-Seq signal for each histone marks. Each CpG site then had five normalized histone modification values. We then assigned each CpG site to its nearest chromatin state of genomic annotation. As a result, this strategy allows us to generate an integrated matrix whose rows represent all the CpG sites and columns represent chromosome, genomic location, methylation level, histone modification values, chromatin state annotation and cell of origin for each CpG site.

4.3.2 Characteristics of input variables in the integrated matrix

To understand the characteristics of variables in the matrix, we first investigated their value distributions. As expected, DNA methylation levels show a bimodal distribution that peaks at no methylation (0) and fully methylation (1) (Fig 2A), and histone modifications also show a bimodal distribution that peaks at no signal (0) and similar to background (~1 fold) (Fig 2B and Supp Fig1). We next calculated the genome-wide Pearson correlation between DNA methylation and the core histone marks (Fig 2C). As a result, K4me3, K4me1 and K27me3 showed a significant anti-correlation (-0.6, -0.3 and -0.2 respectively) with DNA methylation, while K36me3 showed a positive-correlation (0.1) with DNA methylation and K9me3 showed neutral-correlation (0). This result then suggests a distinct pattern of relationships between the different histone marks and DNA methylation.

We next inspected the relationship between DNA methylation and chromatin states. Figure 2D then shows the distribution of DNA methylation in each chromatin state. As expected, chromatin states associated with transcription, ZNF/Repeats, quiescent and heterochromatin showed a high level of DNA methylation, while K4me3 associated chromatin states, including bivalent regions (with both K4me3 and K27me3) and TSS regions showed hypomethylation signal. Interesting, we also found bivalent enhancer (with K4me1 and K27me3) tend to be depleted of DNA methylation. Moreover, enhancers and repressed polycomb regions showed a highly variable level of DNA methylation values, indicating there are multiple methylation mechanisms within those chromatin states.

4.3.3 Modeling of DNA methylation with multiple logistic regression

With the integrated matrix above, we then used multiple logistic regression to model the quantitative relationship between histone modifications and DNA methylation. We chose logistic regression model instead of others for two major reasons. First, DNA methylation value is between 0 and 1, this fits naturally for logistic regression model. Second, logistic regression, as a class of linear regression, makes it easier to interpret the learnt parameters and relationships between variables. The following equation then represents the modeling process:

$$ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

Where x1 to x5 is the log2 fold change for different histone modification values, and Pi is the DNA methylation values for each CpG site i.

To evaluate the model performance, we then used R_{McFadden} to represent the correlation between observed methylation value and predicted methylation value:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

Where Lc is the maximum likelihood value from the fitted model, and Lnull is the likelihood value from the model with only an intercept but no covariates. The resulted R value then captures fraction of responsive values that could be explained by the input variables.

We thus evaluate the model performance with different combinations of histone marks as input variables. When single histone mark was used in the model, we observed a superior performance of K4me3 (0.52) compared to the other four histone marks (K9me3: 0.02, K36me3: 0.07, K27me3: 0.18, K4me1: 0.26) (Fig 3A). This result then suggests K4me3 is a strong predictor of DNA methylation, while K4me1, K27me3 and K36me3 shows a medium but also significant prediction power.

We next sought to include multiple histone marks in the model. As a result, when used the four histone marks beside K4me3, we got a model performance of 0.32 (Fig 3A), which was 40 percent lower than using K4me3 alone, again highlighting the evident predictive power of K4me3. More importantly, when incorporated all five histone marks in the model, we got an overall predictive power of 0.56 (Fig 3A). This power value thus indicates that there is a universal quantitative relationship between histone modifications and DNA methylation for different types of human cells, where more than half of methylation variations could be explained by the five histone marks we used in the model.

To have a clearly genome-wide view of the model performance, we then plot the predicted methylation values against observed methylation values for each CpG sites (Fig 3B). The scatterplot shows two enriched groups (color in red) where low observed methylation values are also predicted to be low and high observed methylation values are also predicted to be high, indicating the predicted model performs well for those CpG sites. To look at the model performance in specific genomic regions, we chose a random region in chromosome 1 for an ESC cell line (HUES64 cells). Figure 3D then shows the genomic view of the observed and predicted methylation signal, the five histone modification signals and the refseq genes. It is obvious that the predicted methylation track agrees well with the observed methylation track. In summary, we showed that histone modifications are highly predictable of DNA methylation and K4me3 is the dominant predictor of DNA methylation.

4.3.4 Model performance in 35 human cells or tissues

In the above section, we showed that our model was able to capture a large variation of DNA methylation for the integrated 800 million CpG sites for a variety of human cells or tissues. The prediction power then represents the average predictive power for 35 human cells types. We next asked whether this prediction power varied in different types of human cells. We thus used the same approach for CpG sites and their histone signals in each individual cell lines or tissues we have collected. As a result, we indeed observed some differences of model performance, ranging from 0.43 (fetal intestine small cells) to 0.70 (HUES64 stem cells). Strikingly, we also found a clear pattern that pluripotent cells tend to have a higher methylation predictive power compared with tissue cells. One possible explanation of this is that tissue cells are highly heterogeneity, making it less accurate to do the methylation prediction based on histone modifications. Another explanation is the quantitative relationships between DNA methylation and histone modifications decrease when the cells differentiated into developed cells. Further investigations will need to be carried out to address which hypothesis plays a major role in the differences of methylation prediction power.

4.3.5 Characteristics of mis-predicted methylation loci

Although our model was able to capture a significant fraction of methylation signal, there was still a considerably degree of mis-prediction. We next sought to examine the characteristics of those mis-predicted methylation loci. Figure 5A then shows the distribution of residual values between predicted and observed methylation. Residual distribution generally followed a normal distribution. However, we also observed a longer tail towards 1 where predicted values tend to be hypermethylated compared with actual values. This result suggests that there are additional de-methylation mechanisms that could be explained by the core histone modifications.

Moreover, we found that there were less than 5 percent of CpG sites showing a residual methylation values larger than 0.5. This result then reveals that the model would perform extremely well if the task is classify the CpG methylation level into a binary event, either un-methylated (0) or methylated (1). To quantitative evaluate this finding, we then carried out Receive Operation Curve analysis (Fig 5B). The area under curve (AUC) thus represents the power of the model to correctly classify the CpG sites into methylated sites or un-methylated sites. As a result, we observed a AUC of 0.97, indicating the core histone modifications have a extremely high performance to classify CpG sites into a binary event.

To further characterize the residual methylation, we next investigated the distribution of those values in each chromatin states (Fig 5C). In most cases, the residual distribution followed a similar distribution as shown in Fig 5A. However, residual values in bivalent enhancer (with K4me1 and K27me3) and repressed polycomb (with K27me3) showed a distinct pattern compared with others. Specifically, predictions in both bivalent enhancers and repressed polycomb tend to be hypermethylated, indicating there are additional de-methylation mechanisms associated with those two K27me3 chromatin states. Therefore, it appears that K27me3 can influence DNA methylation in those chromatin states.

One explanation of the above finding for bivalent enhancers is that there are some factors, e.g. transcription factors, bound to those regions that are inhibiting DNA methylation. To test this hypothesis, we next performed gene ontology annotation and motif discovery analysis for the top mis-predicted CpG regions.

4.4 Discussion and conclusion

In this study, we built a predictive model of DNA methylation based on core histone marks. This model integrates about 800 million CpG sites and achieved a reasonably high predictive power. We found H3K4me3 is a dominant predictor and DNA methylation, where regions with H3K4me3 are generally depleted of DNA methylation. Moreover, we also showed that our model has a less than five percent of error when classifying CpG methylation level into a binary value. This result then can be widely applied in the circumstances where researchers only care about hypermethylation or hypomethylation. In addition, previous studies have revealed relationships between histone modifications and DNA methylation in a descriptive and qualitative way. Our work then integrates data from 35 types of human cells and provide a comprehensive and quantitative interrogation of the crosstalk between the two major epigenetic mechanisms.

We observed a variation of power when examined the predictive values for each cell type. We think this might be caused by either cells having different heterogeneity or under different developmental stages. It is then interesting to further study whether the crosstalk between the two major epigenetic mechanisms decrease or not during cell differentiation. Another interesting observation is the H3K27me3-associated genomic regions show distinct pattern of residual methylation distribution. This result then suggests there is other de-methylation mechanisms that can not be explained solely by histone modifications. This mechanism can be other regulators, such as transcription factors, where they bind to those regions and initiate the de-methylation activities.

Figures

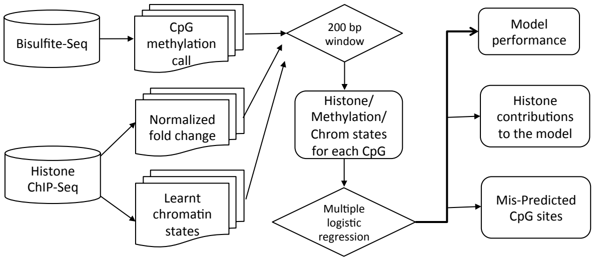


Figure 1. Schematic illustration of the integrative approach.

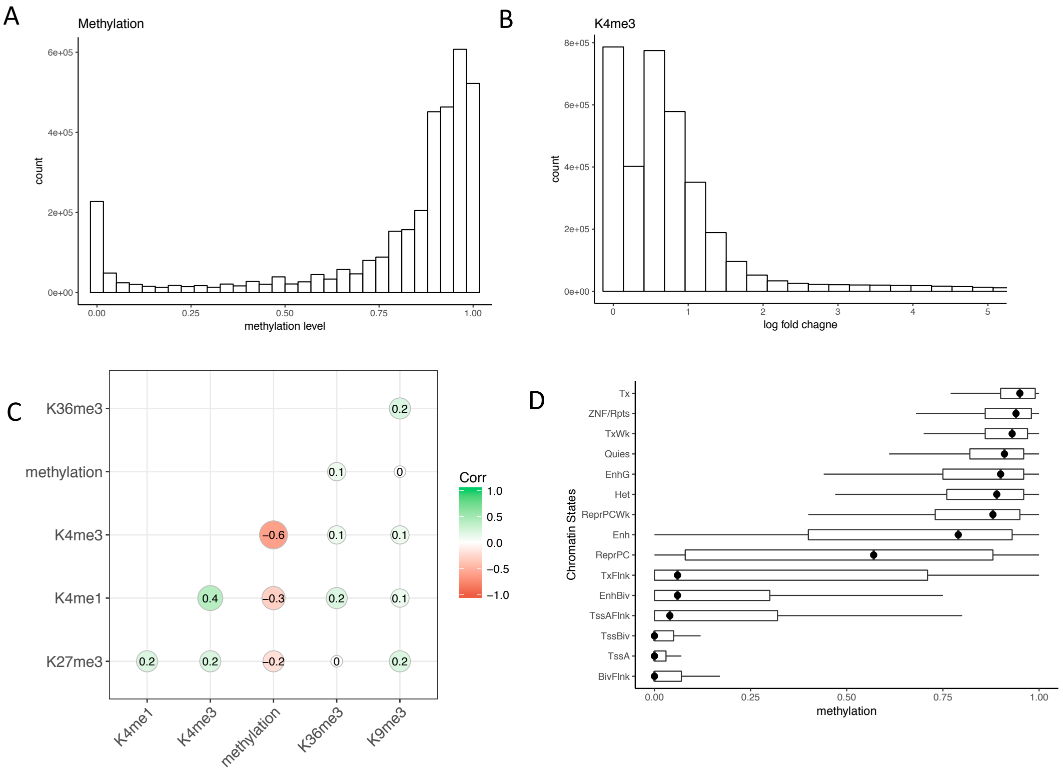


Figure 2. Relationships among input variables in the integrated matrix.

A. Distribution of the DNA methylation values in the integrated matrix. B. Distribution of the H3K4me3 normalized log fold change values in the integrated matrix. C. Genome-wide pearson correlation between the five histone modifications and DNA methylations. D. Distribution of DNA methylation in each learnt chromatin states (y axis).

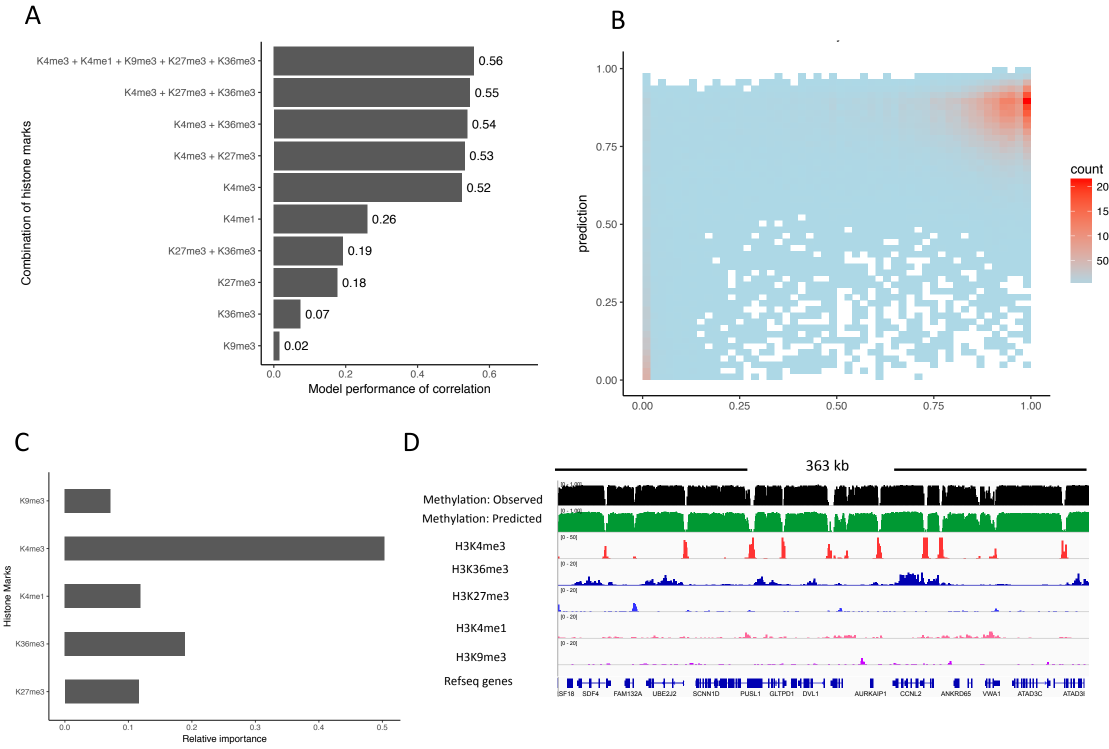


Figure 3. Performance of the predictive model.

A. Model performance based on different combinations of histone modifications as input variables. The X-axis represents the model performance evaluated by $R_{McFadden}$. B. Scatterplot of observed DNA methylation and predicted DNA methylation. C. Relative importance of the histone modification in the core 5 histone modification model. D. An example genomic region of the predicated DNA methylation and its corresponding histone modifications.

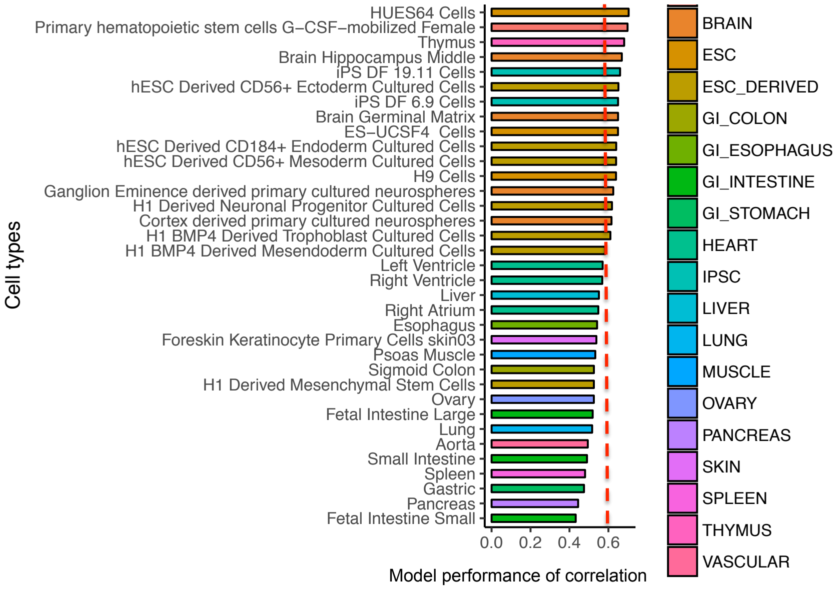


Figure 4. Performance for 35 different types of human cell lines/tissues.

The x-axis represents the model performance for each human cell type. The y-axis represents the name for each human cell type. The legend represents the corresponding tissue for each cell type.

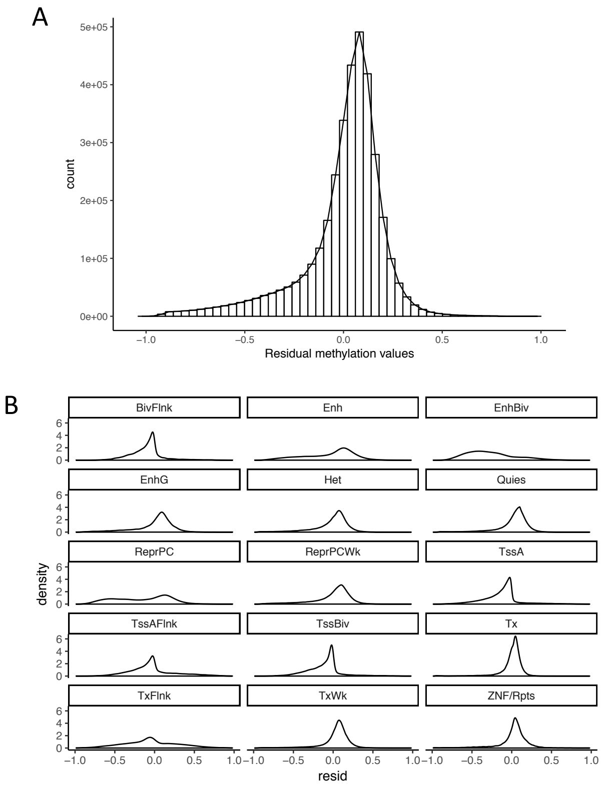


Figure 5. Characteristics of residual DNA methylation values.

A. Distribution of residual DNA methylation values. B. Density plot of residual DNA methylation values in each chromatin states.

References

1. Iyer LM, Abhiman S, Aravind L: **Natural history of eukaryotic DNA methylation systems.** *Prog Mol Biol Transl Sci* 2011, **101**:25–104.

2. Lyko F: **The DNA methyltransferase family: a versatile toolkit for epigenetic regulation.** *Nat Rev Genet* 2018, **19**:81–92.

3. Law JA, Jacobsen SE: **Establishing, maintaining and modifying DNA methylation patterns in plants and animals.** *Nat Rev Genet* 2010, **11**:204–220.

4. Krueger F, Kreck B, Franke A, Andrews SR: **DNA methylome analysis using short bisulfite sequencing data.** *Nat Methods* 2012, **9**:145–151.

5. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning**. *Nature* 2008, **452**:215–219.

6. Guo S, Diep D, Plongthongkum N, Fung H-L, Zhang K, Zhang K: **Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.** *Nat Genet* 2017, **49**:635–642.

7. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE, Kratz A, Wefers AK, Huang K, Pajtler KW, Schweizer L, Stichel D, Olar A, Engel NW, Lindenberg K, Harter PN, Braczynski

AK, Plate KH, Dohmen H, Garvalov BK, Coras R, Hölsken A, Hewer E, Bewerunge-Hudler M, Schick M, Fischer R, et al.: **DNA methylation-based classification of central nervous system tumours.** *Nature* 2018, **555**:469–474.

8. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, et al.: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518**:317–330.