Title
Computational Methods for Analysis of Large-Scale Epigenomics Data

Author
Fiziev, Petko Plamenov

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Methods for Analysis of Large-Scale Epigenomics Data

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

by

Petko Plamenov Fiziev

2018

ABSTRACT OF THE DISSERTATION

Computational Methods
for Analysis of Large-Scale
Epigenomics Data

by

Petko Plamenov Fiziev
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2018
Professor Jason Ernst, Chair

Reverse-engineering and understanding the regulatory dynamics of genes is key to gaining insights into many biological processes on molecular level. Advances in genomics technologies and decreasing costs of DNA sequencing enabled interrogating relevant properties of the genome, collectively referred to as epigenetics, on very large scale. This work presents results from two collaborative projects with experimental biologists and two new general computational methods for analysis of high-throughput epigenomic data.

The first collaborative project is joint work with Dr. Kathrin Plath and members of her lab at UCLA on studying the epigenetics of somatic cell reprogramming in mouse. By generating and analyzing a large compendium of genomics datasets at four distinct stages during reprogramming, we discovered key properties of the regulatory dynamics during this process and proposed new ways to improve its efficiency.

The first computational method in this work, ChromTime, presents a novel framework for modeling spatio-temporal dynamics of chromatin marks. ChromTime detects expanding, contracting and steady domains of chromatin marks from time course epigenomics data. Applications of the method to a diverse set of biological systems show that predicted dynamic domains likely mark important regulatory regions as they associate with changes in gene expression and transcription factor binding. Furthermore, ChromTime enables analyses of

the directionality of spatio-temporal dynamics of epigenetic domains, which is a previously understudied aspect of chromatin dynamics. Our results uncover associations between the direction of expanding and contracting domains of several chromatin marks and the direction of transcription of nearby genes.

The second collaborative project is joint work with cancer researchers, Dr. Lynda Chin and Dr. Kunal Rai and members of their labs at MD Anderson Cancer Center in Houston, TX. Within this project we studied the epigenetics of melanoma cancer progression. Our collaborators generated genome-wide maps for a large number of histone modifications, DNA methylation and gene expression in tumorigenic and non-tumorigenic human melanocytes. By comparing these maps we discovered that loss of acetylation marks at regulatory regions is characteristic of tumorigenic melanocytes and that modulating acetylation levels can impact tumorigenic potential of cells. In addition, we developed a novel nanostring assay for interrogating the chromatin state at a small subset of genomic locations, which can potentially be used for diagnostic or prognostic purposes in future.

The second computational method presented in this work, CSDELTA, is designed to detect differential chromatin sites from genome-wide chromatin state maps in groups with multiple samples. Biological relevance of detected differential sites is supported by associations with changes in gene expression and transcription factor binding. Furthermore, CSDELTA models the functional similarity between chromatin states and improves upon the resolution of detection compared to existing methods, which enables more accurate downstream analyses to gain insights into the regulatory dynamics of biological systems.

The dissertation of Petko Plamenov Fiziev is approved.

Matteo Pellegrini

Siavash K Kurdistani

Eleazar Eskin

Jason Ernst, Committee Chair

University of California, Los Angeles

2018

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Jason Ernst, for introducing me to the field of epigenomics and providing me with guidance and support throughout my doctoral work. I have learned a great deal about so many topics in bioinformatics and machine learning during this process! I would like to thank the members of my doctoral committee: Matteo Pellegrini, Siavash Kurdinstani and Eleazar Eskin, for helpful discussions and feedback on many of my projects! Furthermore, I would like to thank all my collaborators without whom none of this work would be possible! In particular, Kathrin Plath, Constantinos Chronis, Giancarlo Bonora, Stefan Butz, Shan Sabri and Bernadett Papp for their work on our joint project about epigenetics of stem cell reprogramming! I would like to thank Kunal Rai, Kadir Akdemir and Lynda Chin for their work on our joint project about epigenetics of melanoma cancer! I am also grateful to Tonis Org, Ben Van Handel, Gabriel Ferguson and Denis Evseenko for their work on a joint project about human skeletogenesis! I was also fortunate to work with Pao-Yang Chen, Weilong Guo and Matteo Pellegrini on a computational method for calling DNA methylation levels from high-throughput bisulfite sequencing data. Thank you all!

I would like to thank all members of the Ernst lab at UCLA for being amazing colleagues!

I was partially supported by the CIRM Training Grant TG2-01169 and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Pre-doctoral Fellowship in Stem Cell Science. I was also supported by funds from my advisor, Jason Ernst, who was supported by NIH grants R01ES024995, U01HG007912, DP1DA044371, an NSF CAREER Award #1254200 and an Alfred P. Sloan Fellowship.

Chapter 2 was published as: 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. On this paper, I was co-first author with Constantinos Chronis. In addition, the contributions of individual authors to this project were as follows: Conceptualization, C.C. and K.P.; Methodology, C.C. and K.P.; Software, C.C., P.F., S.S., G.B., and

J.E.; Validation, C.C. and P.F.; Investigation, C.C., B.P., and S.B.; Formal Analysis, C.C., P.F., S.S., G.B., J.E., and K.P.; Data Curation, C.C., P.F., S.S., and G.B.; Writing Original Draft, C.C. and K.P.; Writing Review and Editing, C.C., P.F., B.P., G.B., J.E., and K.P.; Resources, C.C., J.E., and K.P.; Visualization, C.C., P.F., J.E., and K.P.; Supervision, J.E. and K.P.; Project Administration, K.P.; Funding Acquisition, J.E. and K.P.

Chapter 3 is currently submitted for publication. Jason Ernst proposed the project and contributed to the core ideas in the ChromTime method. I developed the ChromTime model, implemented the software and did all analyses related to this project. In this context, I would also like to thank Constantinos Chronis and all members of my lab for many helpful discussions!

Chapter 4 was published as: Fiziev P, Akdemir K, Miller J, Keung E, Samant N, Sharma S, Natale C, Terranova C, Maitituoheti M, Amin S, Martinez-Ledesma E, Dhamdhere M, Axelrad J, Shah A, Cheng C, Mahadeshwar H, Seth S, Barton M, Protopopov A, Tsai K, Davies M, Garcia B, Amit I, Chin L, Ernst J, Rai K. Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression. *Cell Reports*, 2017. On this paper, I was co-first author with Kadir Akdemir. In addition, the contributions of individual authors to this project were as follows: Conceptualization, L.C., J.E., K.R., K.C.A., and P.F.; Methodology, A.P., S. Seth, C.S.C., I.A., P.F., K.C.A., K.R., and J.E.; Investigation, P.F., K.C.A., K.R., J.E., J.P.M., E.Z.K., N.S.S., S. Sharma, C.A.N., C.J.T., M.M., S.B.A., E.M.L., M.D., J.B.A., A.S., H.M., and B.A.G.; Writing, J.E., L.C., K.R., K.C.A., and P.F.; Funding Acquisition, L.C., J.E., and K.R.; Resources, K.Y.T., M.A.D., J.E., L.C., and K.R.; Supervision, L.C., J.E., K.R., and M.C.B.

Chapter 5 is a manuscript that is currently in preparation. Jason Ernst proposed the project and contributed to the core ideas in the CSDELTA method. I contributed to developing the method and the procedure for assessing statistical significance, implemented the software and did all analyses related to this project. In this context, I would also like to thank all members of my lab for many helpful discussions!

Last, but not least, I would like to thank all my friends and my family who have supported

me throughout all these years!

VITA

1999-2000	Computer Science, University of Sofia "St. Kliment Ohridski", Sofia, Bulgaria.
2002	Computer Science (Vordiplom), Technical University Berlin, Germany.
2006	Computer Science (Diplom), Free University Berlin, Germany.
2011	Electronic Media (Master's degree), University of Sofia "St. Kliment Ohridski", Sofia, Bulgaria.

PUBLICATIONS

Hatzigeorgiou A, Fiziev P, Reczko M. DIANA-EST: a statistical analysis. Bioinformatics, 2001

Reczko M, Fiziev P, Staub E, Hatzigeorgiou A. Finding Signal Peptides in Human Protein Sequences Using Recurrent Neural Networks. Algorithms in Bioinformatics, WABI 2002.

Kiriakidou M, Nelson P, Kouranov A, Fiziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. A combined computational-experimental approach predicts human microRNA targets. Genes & Development, 2004.

Staub E, Fiziev P, Rosenthal A, Hinzmann B. Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. Bioessays, 2004.

Roepcke S, Fiziev P, Seeburg P, Vingron M. SVC: Structured Visualization of Evolutionary Sequence Conservation. Nucleic Acids Research, 2005.

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

Rai K, Akdemir K, Kwong LN, Fiziev P, Wu CJ, Keung E, Sharma S, Samant N, Williams M, Axelrad J, Shah A, Yang D, Grimm E, Barton M, Milton D, Heffernan T, Horner J, Ekmekcioglu S, Lazar A, Ernst J, Chin L. Dual Roles of RNF2 in Melanoma Progression. Cancer Discovery, 2015.

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. (* Co-first authors)

Fiziev P*, Akdemir K*, Miller J, Keung E, Samant N, Sharma S, Natale C, Terranova C, Maitituoheti M, Amin S, Martinez-Ledesma E, Dhamdhere M, Axelrad J, Shah A, Cheng C, Mahadeshwar H, Seth S, Barton M, Protopopov A, Tsai K, Davies M, Garcia B, Amit I, Chin L, Ernst J, Rai K. Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression. Cell Reports, 2017. (* Co-first authors)

Ferguson G, Van Handel B, Bay M, Fiziev P, Org T, Lee S, Wu L, Saitta B, Elphingstone J, Larson AN, Riester S, Pyle A, Bernthal N, Mikkola H, Ernst J, van Wijnen A, Bonaguidi M, Evseenko D. Transcriptional heterogeneity during lineage specification defines human skeletogenesis (submitted)

Fiziev P, Ernst J. Modeling Spatio-temporal Dynamics of Chromatin Marks. (submitted)

Fiziev P, Ernst J. Systematic detection of differential chromatin sites from group-wise comparisons of multiple ChIP-seq maps. (in preparation)

CHAPTER 1

Introduction

How genes are regulated in living cells is one of the most fundamental questions in molecular biology. In this process, the aggregate of DNA and proteins, collectively referred to as chromatin, acts as a complex and dynamic control system for the expression of genes. The chromatin consists of protein complexes, termed nucleosomes, around which DNA is wrapped. Nucleosomes, in turn, are made of proteins called histones, the tails of which can be modified by enzymes at specific sites in a multitude of ways. The joint presence or absence of such chemical marks can be associated with different regulatory elements that correlate with expression of genes[1]. In addition, DNA marks such as methylation[2] and hydroxymethylation[3] of cytosines in the genome can also play role in this regulatory process. By complex interactions with different chromatin regions in the genome, DNA binding proteins termed transcription factors control the expression of genes[4]. In general, the chemical modifications that encode relevant information for the organization of the genome are called epigenetic marks, and the field of study of all such marks across the genome is called epigenomics.

A number of high-throughput experimental protocols have been developed in the past decade to map genomic locations of different epigenetic marks. For example, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has been widely used to map histone modifications and transcription factor binding in many cell lines, primary tissues and conditions[5–7]. Bisulfite sequencing has enabled genome-wide mapping of cytosine methylation[8, 9]. Assays for transposase-accessible chromatin[10] (ATAC-seq) and DNaseI hypersensitivity[11] (DNase-seq) coupled with DNA sequencing have been utilized to map

transcription factor binding in more unbiased ways. In conjunction with protocols for measuring transcriptome-wide gene expression such as RNA-seq[12], these genomics technologies have enabled researchers to investigate on large scale the dynamics of gene regulation in many biological systems. However, genomics experiments typically yield tens to hundreds of gigabytes of raw data, which can be difficult to process in order to answer the relevant biological questions. In addition, these datasets contain substantial amounts of noise due to mostly unknown sources of technical or biological variation. For these reasons, bioinformatics analysis of genomics data is a non-trivial task, which often requires developing new computational methods.

In this work, I present two collaborative projects with experimental biologists and two novel computational methods that I developed together with my doctoral advisor, Dr. Jason Ernst, to address problems in studying epigenetic regulation. The first collaborative project was a joint work with Dr. Kathrin Plath, Dr. Constantinos Chronis and members of the Plath lab at UCLA aimed at understanding the principles governing somatic cell reprogramming. Somatic cell reprogramming is an artificial process that achieves dramatic changes in cell identity of a starting population of cells. During this process, differentiated cells are converted to induced pluripotent stem cells, which are indistinguishable in their properties from naturally occurring embryonic stem cells. In particular, induced pluripotent stem cells have the potential to further differentiate into any tissue type in the body. Studying somatic cell reprogramming has tremendous potential to elucidate phenomena in basic developmental and cell biology and to have impact on many biomedical areas including regenerative medicine, targeted drug development, disease

diagnostics and prognostics. However, our understanding of the mechanics of reprogramming still remains limited and current protocols yield only a small fraction of successfully converted cells from the starting cell population, which is a major problem in the field.

By using high-throughput experimental techniques, our collaborators from Dr. Plath's group have mapped the genomic locations of multiple epigenetic marks across four distinct stages of reprogramming corresponding to mouse embryonic fibroblasts (mEFs), mEFs at 48 hours after induction of reprogramming factors, pre-induced pluripotent stem cells and mouse embryonic stem cells. Among these features were nine post-translational histone modifications, one histone variant, DNA binding sites of a number of relevant transcription factors including Oct4, Sox2, Klf4 and cMyc that have been shown to lead to reprogramming upon induction, cytosine methylation and sites of transposase-accessible chromatin (ATAC-seq). In addition, gene expression levels were measured by RNA-seq assays. In this project, I performed the main part of the integrative computational analysis of these datasets to derive a comprehensive map of epigenetic changes and to identify key combinatorial transcription factor dynamics that drive reprogramming. Major analyses included deriving chromatin state maps for each stage during reprogramming, deriving chromatin state trajectories of epigenetic changes across all four stages by a novel application of software for chromatin state discovery, ChromHMM, in "stacked" mode, analyzing combinatorial transcription factor binding events upon induction of all or combinations of reprogramming factors, motif enrichment analysis to unravel relationships between transcription factor binding, underlying DNA sequence and chromatin states, and comparison of gene expression profiles across reprogramming stages. Together with our

collaborators, I worked to integrate all parts of the computational analysis and to interpret the findings in the study. As result, we discovered that reprogramming factors can exhibit dual roles in reprogramming which include both silencing of fibroblast specific genes and activation of stem cell specific genes. We further discovered combinatorial patterns of interactions between transcription factors that are required for successful reprogramming and a new additional factor, which facilitates the process. Chapter 2 contains a reprint of our joint publication with Dr. Plath's group, which describes our findings in detail[13].

Building on our experience during the collaboration with Dr. Plath's lab, together with my advisor Dr. Ernst, I developed ChromTime, a new computational method for analysis of time course data for chromatin marks. ChromTime is a general method for modeling spatio-temporal dynamics of epigenetic domains. The method detects domains of chromatin marks that expand, contract or hold steady in time course data from ChIP-seq experiments. Chapter 3 shows applications of the method to data for a diverse set of histone modifications and Pol2 in a range of biological systems. Domains of epigenetic marks that expand and contract likely mark important regulatory regions as their spatial dynamics correlate with changes in features captured by orthogonal genomics assays for measuring gene expression and transcription factor binding. Furthermore, ChromTime enables analysis of directionality of spatial dynamics of epigenetic peaks, which is a previously understudied aspect of chromatin regulation due to lack of suitable bioinformatics tools. Our directionality analysis showed that the direction of expanding and contracting peak boundaries of a subset of chromatin marks is well correlated with direction of transcription of genes in proximity of transcription start sites. ChromTime is implemented as a

software tool that can be used by researchers to study spatial dynamics of epigenetic peaks in time course ChIP-seq data. Chapter 3 contains a version of the manuscript describing the ChromTime method, which has been submitted and is under review at the time of writing of this thesis.

The second collaborative project was a joint work with Dr. Lynda Chin, Dr. Kunal Rai, Dr. Kadir Akdemir and their colleagues at MD Anderson Cancer Center in Houston, Texas. The goal of this project was to investigate epigenetic changes during melanoma cancer progression. Epigenetics are known to play important role in some cancers. However, systematic mapping of a large number of histone modifications during cancer progression remains largely uncharted territory. By utilizing a high-throughput ChIP-seq technology our collaborators mapped the genomic positions of 35 epigenetic marks in two human tumorigenic cell lines that give rise to melanoma cancer and their non-tumorigenic counterparts. In addition, they measured gene expression, DNA methylation and DNA hydroxymethylation in those cell lines. Within this project, I performed the main part of the bioinformatics analysis of the ChIP-seq data, which involved deriving and comparing chromatin state maps for non-tumorigenic and tumorigenic human melanocytes. These maps highlighted large changes in chromatin state including a loss of acetylations at regulatory elements. I also helped in integrating these results with analysis of gene expression and DNA methylation performed by our collaborators. In addition, I implemented a pipeline for selecting a small subset of genomic regions that are most differential with respect to specific histone modifications, which were later used to design a custom nanostring assay to differentiate tumorigenic from non-tumorigenic cells. The nanostring assay

was applied among others to samples from melanoma cancer patients and was able to differentiate in a proof-of-principle experiment benign nevi from tumor cells. Chapter 4 contains a reprint of our joint publication with Dr. Lynda Chin and Dr. Kunal Rai's groups, which describes our findings in detail[14].

Building on our experience during the collaboration with Dr. Kunal Rai and his colleagues, we explored more generally the problem of comparing chromatin state maps between conditions. As result, together with my advisor, Dr. Ernst, I developed CSDELTA, a general computational method for genome-wide comparison of chromatin state segmentations between groups with multiple samples. Our method can be applied to find epigenetic changes in an unbiased manner across the whole genome. Furthermore, CSDELTA operates at the resolution of the input chromatin segmentations, which is typically 200 base pairs or one nucleosome and spacer region, thus enabling detection at a fine resolution. CSDELTA models the functional similarity between chromatin states, which can increase the power to detect true chromatin changes. The software provides an easy to use command line interface and runs in reasonable time and memory. Chapter 5 shows applications of the method to a publicly available dataset, which showcases the advantages of CSDELTA over existing methods. The output of CSDELTA can be used to define group-specific chromatin state domains and to explore differences in regulatory dynamics in different cell types, tissues or conditions. Chapter 5 contains a version of the manuscript describing the CSDELTA method, which is in preparation for submission at the time of writing of this thesis.

REFERENCES

[1] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types.," *Nature*, vol. 473, no. 7345, pp. 43–9, May 2011.

[2] I. R. Henderson and S. E. Jacobsen, "Epigenetic inheritance in plants," *Nature*, vol. 447, no. 7143, pp. 418–424, May 2007.

[3] S. Kriaucionis and N. Heintz, "The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain," *Science (80-.).*, vol. 324, no. 5929, pp. 929–930, May 2009.

[4] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and microRNAs," *Nat. Rev. Genet.*, vol. 8, no. 2, pp. 93–103, Feb. 2007.

[5] A. Barski *et al.*, "High-resolution profiling of histone methylations in the human genome.," *Cell*, vol. 129, no. 4, pp. 823–37, May 2007.

[6] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome.," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[7] Roadmap Epigenomics Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015.

[8] R. Lister *et al.*, "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, no. 7271, pp. 315–322, Nov. 2009.

[9] S. J. Cokus *et al.*, "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning," *Nature*, vol. 452, no. 7184, pp. 215–219, Mar. 2008.

[10] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf,

"Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.," *Nat. Methods*, vol. 10, no. 12, pp. 1213–8, Dec. 2013.

[11] A. P. Boyle *et al.*, "High-resolution mapping and characterization of open chromatin across the genome.," *Cell*, vol. 132, no. 2, pp. 311–22, Jan. 2008.

[12] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[13] C. Chronis *et al.*, "Cooperative Binding of Transcription Factors Orchestrates Reprogramming.," *Cell*, vol. 168, no. 3, p. 442–459.e20, Jan. 2017.

[14] P. Fiziev *et al.*, "Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression.," *Cell Rep.*, vol. 19, no. 4, pp. 875–889, Apr. 2017.

CHAPTER 2

Cooperative Binding of Transcription Factors Orchestrates Reprogramming

Cooperative Binding of Transcription Factors Orchestrates Reprogramming

Constantinos Chronis,[1,2,3] Petko Fiziev,[1,2,3] Bernadett Papp,[1,2] Stefan Butz,[1,2] Giancarlo Bonora,[1,2] Shan Sabri,[1,2] Jason Ernst,[1,2,*] and Kathrin Plath[1,2,4,*]
[1]David Geffen School of Medicine, Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA 90095, USA
[2]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, Bioinformatics Program, Los Angeles, CA 90095, USA
[3]Co-first author
[4]Lead Contact
*Correspondence: jason.ernst@ucla.edu (J.E.), kplath@mednet.ucla.edu (K.P.)
http://dx.doi.org/10.1016/j.cell.2016.12.016

SUMMARY

Oct4, Sox2, Klf4, and cMyc (OSKM) reprogram somatic cells to pluripotency. To gain a mechanistic understanding of their function, we mapped OSKM-binding, stage-specific transcription factors (TFs), and chromatin states in discrete reprogramming stages and performed loss- and gain-of-function experiments. We found that OSK predominantly bind active somatic enhancers early in reprogramming and immediately initiate their inactivation genome-wide by inducing the redistribution of somatic TFs away from somatic enhancers to sites elsewhere engaged by OSK, recruiting Hdac1, and repressing the somatic TF Fra1. Pluripotency enhancer selection is a stepwise process that also begins early in reprogramming through collaborative binding of OSK at sites with high OSK-motif density. Most pluripotency enhancers are selected later in the process and require OS and other pluripotency TFs. Somatic and pluripotency TFs modulate reprogramming efficiency when overexpressed by altering OSK targeting, somatic-enhancer inactivation, and pluripotency enhancer selection. Together, our data indicate that collaborative interactions among OSK and with stage-specific TFs direct both somatic-enhancer inactivation and pluripotency-enhancer selection to drive reprogramming.

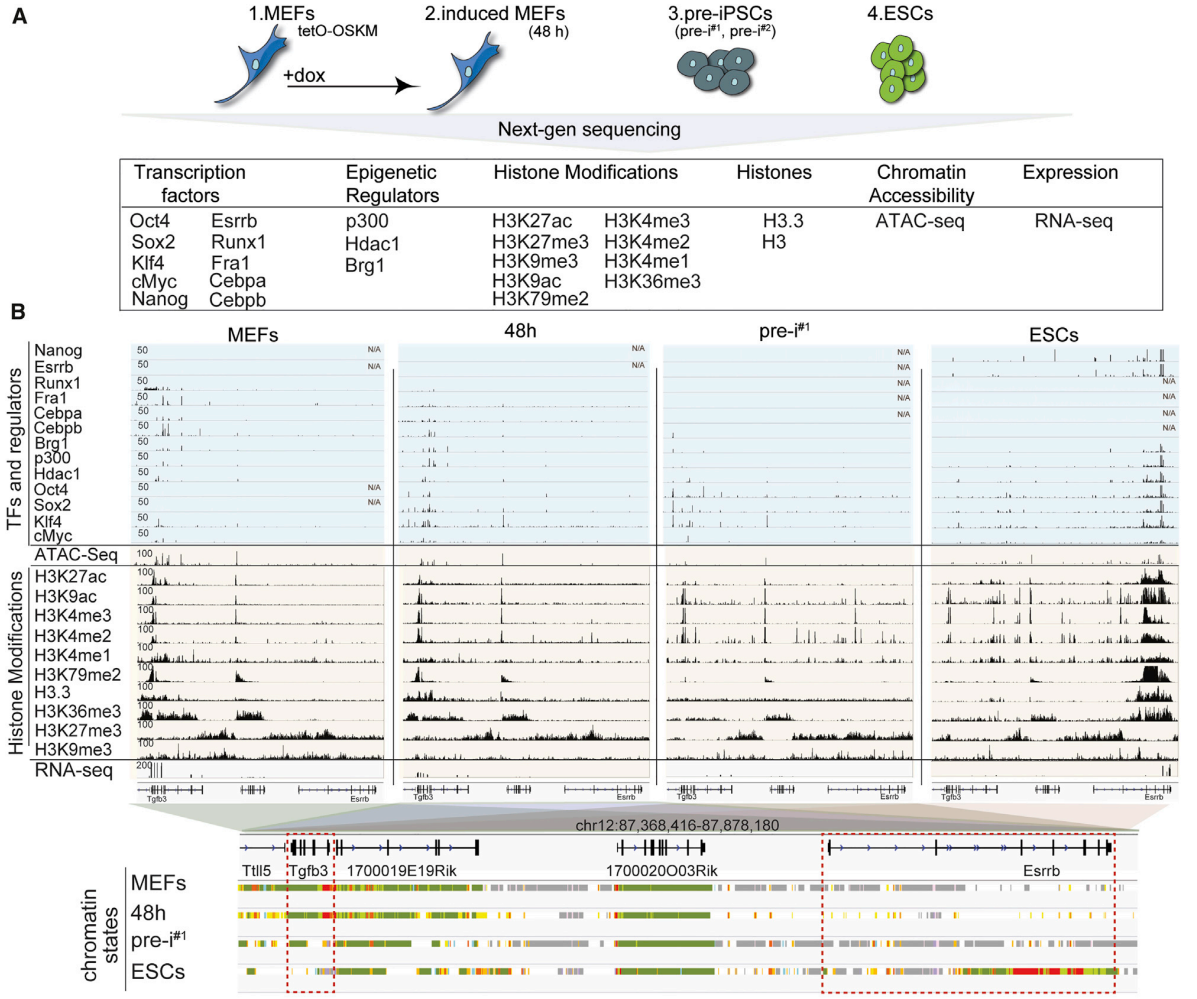
INTRODUCTION

Differentiated cells can be reprogrammed to pluripotency by overexpression of the four transcription factors (TFs) Oct4, Sox2, Klf4, and cMyc (OSKM) (Takahashi and Yamanaka, 2006). Successful reprogramming of somatic cells to induced pluripotent stem cells (iPSCs) leads to the faithful shutdown of the somatic program and activation of the target program. Conversely, in TF-induced conversions of one somatic cell type to another, incomplete extinction of the starting cell program represents a major barrier (Cahan et al., 2014). Hence, understanding the mechanisms by which OSKM inactivate the starting cell program and induce the pluripotency network will provide insights into the principles by which cell identity can be effectively manipulated.

The interaction of OSKM with chromatin has been primarily studied in ESCs, where O, S, and K preferentially bind enhancers and M primarily associates with promoters (Chen et al., 2008; Kim et al., 2008). In ESCs, enhancers are often occupied by additional pluripotency TFs including Nanog and Esrrb (Chen et al., 2008; Kim et al., 2008; Whyte et al., 2013), suggesting that complex regulatory interactions perpetuate the pluripotent state. Among the pluripotency TFs, O, S, and Nanog are thought to form a pivotal circuitry as they co-occupy enhancers with a higher frequency than other TFs (Chen et al., 2008), raising the questions of why K is an effective reprogramming factor when combined with O and S and how these factors interact during reprogramming. Moreover, it is unclear how and when pluripotency enhancer selection happens during reprogramming given that most pluripotency TFs are only available late in the process (Polo et al., 2012; Samavarchi-Tehrani et al., 2010). Since enhancers play a central role in driving cell-type-specific gene expression (Heinz et al., 2015), defining how the reprogramming factors control the reorganization of the enhancer landscape is critical for the mechanistic understanding of reprogramming.

A few studies reported that the target sites of the reprogramming factors change during reprogramming (Chen et al., 2016; Sridharan et al., 2009). In addition, it has been shown that O, S, and K each can act as pioneer factor since they can engage nucleosome-occluded sites in human fibroblasts and nucleosomal templates in vitro (Soufi et al., 2012, 2015). Whether these properties are relevant for their binding to pluripotency enhancers during the reprogramming process, however, remains elusive. Moreover, the pioneer factor model does not provide a mechanistic explanation for the silencing of the somatic program, and, therefore, it has remained unclear how the reprogramming factors would induce this process.

In our study, we delineated the interaction of the reprogramming factors with somatic and pluripotency enhancers. We uncovered that OSK mediate both somatic enhancer silencing and pluripotency enhancer selection through collaborative interactions among themselves and with stage-specific TFs.

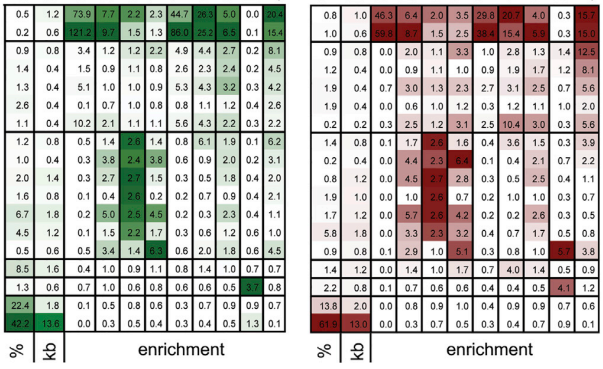


C

chromatin state		H3K27me3	H3K27ac	H3K4Me2	H3K4Me1	H3K9ac	H3K4me3	H3K36me3	H3K9me3	H3K79me2	H3.3	Input
Promoter	1_PromA	2	87	93	32	84	95	6	4	74	46	3
	2_PromP	18	47	91	32	77	87	0	4	1	4	2
Enhancer	3_EnhA	0	84	91	87	27	8	6	4	2	74	2
	4_EnhM1	1	56	96	96	7	1	1	1	1	3	3
	5_EnhM2	1	3	74	45	20	1	0	2	1	2	2
	6_EnhW	1	20	3	54	1	0	1	1	0	1	3
	7_EnhP	69	7	39	69	2	1	5	3	1	0	22
Transcribed Enhancer	8_TxEnhA	0	49	81	86	14	6	30	4	96	28	3
	9_TxEnhM	2	37	15	80	1	0	75	3	4	7	9
Transcribed	10_Tx	1	8	1	19	1	0	86	10	89	14	3
	11_Tx5'	0	4	3	19	3	0	11	2	83	2	2
	12_Tx3'	1	5	0	1	0	0	84	1	2	2	4
	13_TxWk3'	0	1	0	2	4	0	13	0	1	0	4
	14_Tx3'reg	3	19	5	16	5	2	34	38	3	75	5
Polycomb	15_ReprPC	56	0	0	1	0	0	1	3	0	0	7
Heterochromatin	16_Het	8	2	0	1	0	0	15	64	0	1	4
Low signal	17_Low	3	2	0	1	0	0	0	1	0	0	3
	18_LowL	0	0	0	0	0	0	0	1	0	0	0

emission probability 0 100

D



(legend on next page)

RESULTS

Comprehensive Mapping of TFs, Chromatin Features, and Expression at Defined Reprogramming Stages

To characterize the role of OSKM in reprogramming, we carried out chromatin immunoprecipitation for each reprogramming factor coupled to high-throughput sequencing (chromatin immunoprecipitation sequencing [ChIP-seq]) at four distinct stages of mouse embryonic fibroblast (MEF) reprogramming (Figure 1A). These stages included (1) MEFs carrying a tetracycline-inducible polycistronic OSKM expression cassette to capture the starting state; (2) the same MEFs induced for OSKM expression with doxycycline (dox) for 48 hr; (3) two independently generated pre-iPSC lines (pre-i$^{\#1}$ and pre-i$^{\#2}$); and (4) the pluripotent state represented by mouse ESCs for the end state (Figures S1A–S1C). The 48 hr time point represents an early reprogramming stage and was chosen to examine the initial interaction of OSKM with MEF chromatin. Importantly, within the first 48 hr, fibroblasts respond to OSKM activation in a homogeneous manner and with limited expression changes (Buganim et al., 2012; Koche et al., 2011; Polo et al., 2012). Since reprogramming cultures are heterogeneous at later time points (Pasque et al., 2014; Polo et al., 2012), we turned to pre-iPSC lines with closely related transcriptional, epigenetic, and OSKM binding profiles (Figures S1D, S1E, S2E, and S2F) that were isolated clonally from reprogramming cultures infected with OSKM-encoding retroviruses (Sridharan et al., 2009) for a proxy of a late intermediate stage. Since M and K are expressed endogenously in starting MEFs (Figures S1A–S1C), we mapped both in all four reprogramming stages, whereas O and S were profiled at 48 hr, in pre-iPSCs and ESCs.

Additionally, we determined the targets of endogenously expressed TFs (Cebpa, Cebpb, Fra1, Runx1, Esrrb, and Nanog) and chromatin regulators (p300, Hdac1, and Brg1) in relevant reprogramming stages to determine their interplay with OSKM, mapped histone H3 to assess nucleosome occupancy, and measured chromatin accessibility by an assay for transposase-accessible chromatin using ATAC sequencing (ATAC-seq) and gene expression by RNA sequencing (RNA-seq) (Figure 1A) (Tables S1 and S2). We also generated maps for nine histone modifications and the histone variant H3.3 for each reprogramming stage. The histone modifications included H3K4me3 and H3K9ac primarily associated with promoters; H3K4me1, H3K4me2, and H3K27ac characteristic of active promoters and enhancers; H3K79me2 and H3K36me3 associated with transcription, and the repressive marks H3K9me3 and H3K27me3 (Figure 1A) (Ernst et al., 2011). A snapshot of the various datasets is shown in Figure 1B. Data reproducibility was confirmed by correlating replicate experiments, experimental and imputed data (Ernst and Kellis, 2015), and through comparisons with published datasets (Table S3), leading to the merging of replicate datasets for downstream analyses. Additionally, for TFs, their known motifs were identified at occupied sites (Figure S1F), validating our datasets.

Identification of *cis*-Regulatory Elements at Each Reprogramming Stage

To enable a characterization of the chromatin environment at sites engaged by OSKM, we summarized the combinatorial and spatial patterns of histone modifications and H3.3 for each reprogramming stage by building a chromatin state model with 18 states using ChromHMM and assigned candidate functional annotations to each state based on marks present (Figure 1C) (Ernst and Kellis, 2012). The 18 states defined active and poised promoters, inter- and intragenic enhancers of varying activity levels, various transcribed regions, repressed regions, and genomic regions with minimal or no signal of any histone mark (Figures 1C and 1D). These chromatin state annotations were supported by associations with genomic landmarks such as CpG islands and transcriptional start sites (TSSs) of genes, chromatin accessibility and expression of nearby genes (Figures 1D, S1G, and S1H), and captured epigenetic states expected to occur at somatic and pluripotency loci during reprogramming (Figures 1B and S1I).

OSKM Predominantly Occupy Active and Poised Promoters and Enhancers at Each Reprogramming Stage

To understand OSKM action, we first investigated the characteristics of OSKM binding sites at each reprogramming stage. Regardless of reprogramming stage, O, S, and K predominantly bound in distal regions >2 kb away from the TSS, whereas M binding occurred more often in close proximity to the TSS (Figures 2A and S2A). Intersection of binding sites with chromatin states revealed that at each reprogramming stage, all four reprogramming factors bound both active and poised promoters and that the distal binding sites of OSK were predominantly located within active enhancers (Figures 2B and S2B). These binding preferences also applied when considering co-binding between the reprogramming factors, such that M in combinations with O, S, or K displayed strong promoter bias, whereas combinations of O, S, or K binding without M preferentially targeted active enhancers (Figure S2C). Sites occupied by O, S, K, or M displayed pronounced nucleosome depletion and chromatin accessibility at the respective stage (Figures 2C and S2D). Together, these results demonstrated that OSKM prefer to bind active and poised promoters and enhancers regardless of reprogramming stage.

Figure 1. Reprogramming Factor and Epigenome Maps in Four Reprogramming Stages

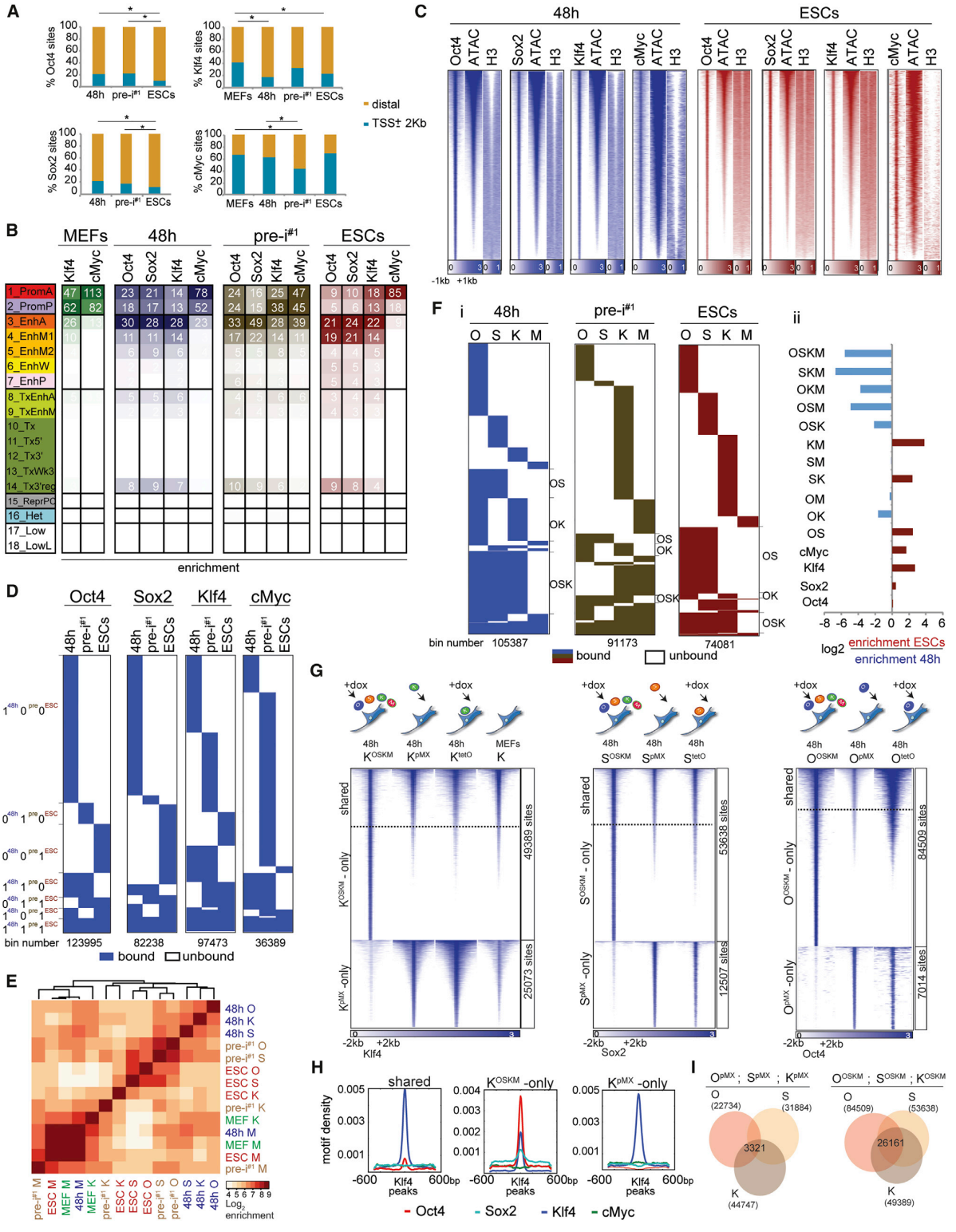
(A) Summary of reprogramming stages and data sets produced.

(B) Snapshot of indicated genomics data at a candidate genomic locus. N/A, no data produced. The color code represents the stage-specific chromatin states defined in (C). Red boxes mark the somatic gene *Tgfb3* and the pluripotency gene *Esrrb*.

(C) Rows represent chromatin states and their representative mnemonics, color coded and grouped based on their putative annotation. Cells show the frequency of each histone mark, H3.3, and input signal for each state (ChromHMM emission probabilities).

(D) Columns give the percentage of genome occupancy, median length in kilobases (kb), and fold enrichment of indicated features (TES, transcription end sites; TSS, transcription start sites; conservation [phastCons elements]; ERVK, endogenous retrovirus K elements; and ATAC-seq, transposase hypersensitivity) for each chromatin state described in (C) for MEFs and ESCs. Color code per column is from highest to lowest value.

See also Figure S1 and Tables S1 and S2.



(legend on next page)

OSKM Redistribution and Binding Partner Switch during Reprogramming

A comparison of binding sites between 48 hr, pre-iPSCs, and ESCs revealed that the genomic locations of each reprogramming factor differed dramatically between stages and that the majority of sites were stage-specific (coined "100," "010," and "001," where 1 represents presence and 0 absence of binding, and the digits from left to right binding at 48 hr, pre-$i^{\#1}$, and ESCs) (Figures 2D, S2E, and S2F). For instance, 48% of all Oct4 binding events occurred exclusively at 48 hr (100 sites) and 16% were specific for the pluripotent stage (001 sites). 48-hr-specific Oct4 binding events (100 sites) occurred close to genes with fibroblast functions based on gene ontology (GO) analysis, whereas pluripotency-specific sites (001 sites) were linked to genes that control stem cell function and early developmental decisions (Figure S2G; Table S4), suggesting that stage-specific binding events are associated with stage-specific gene functions. Together, these data revealed the predominant interaction of OSKM with somatic sites early in reprogramming and the redistribution to pluripotency-associated sites at later stages.

The remaining binding events were transient (110 and 011), absent in pre-iPSCs (101), or constitutive (111) (Figures 2D and S2E). Constitutively bound Oct4 sites, for instance, represented 8% of all 48-hr-bound sites and occurred in the vicinity of genes implicated in blastocyst formation, chromosome organization, and inhibition of MAPK signaling, which is closely tied to the maintenance of pluripotency (Ying et al., 2008) (Figure S2G; Table S4). Thus, the majority of sites associated with the pluripotent state become engaged by the reprogramming factors only late in the process, but certain sites are targeted within the first 48 hr. Motif analysis revealed lower densities of OSKM DNA binding sequences at 100 sites compared to 001 and 111 sites (Figure S2H), suggesting that temporal binding events differ in their regulation.

In addition, we found that M binding differed strongly from that of O, S, and K throughout reprogramming and, more surprisingly, that K sites coincided more with those of O and S at 48 hr but diverged from these in pre-iPSCs and the pluripotent state (Figures 2E, S2I, and S2J). Consequently, we observed significantly more co-binding of OSK and OK at 48 hr than in ESCs and, conversely, an increase in OS co-occupancy in ESCs relative to 48 hr (Figure 2F). Thus, co-binding preferences change from OSK/OK to OS during reprogramming, consistent with O and S composing the core pluripotency network in ESCs alongside Nanog instead of K (Chen et al., 2008).

OSK Co-occupancy at 48 hr Depends on Their Co-expression

By comparing K binding between MEFs and 48 hr, we found that many binding sites were gained, whereas others were lost at 48 hr, and only a subset maintained (Figures 2G and S3A). Upon overexpression of only Klf4 in MEFs for 48 hr, either retrovirally (K^{pMX}) or inducibly (K^{tetO}), without the other reprogramming factors, K predominantly engaged sites that were targeted by it in MEFs and not those newly accessible in the context of OSKM co-expression (Figure 2G), despite its higher expression level (Figure S3B). We conclude that O and S availability, and not the expression level of K per se, is responsible for the differential binding at 48 hr compared to MEFs. Moreover, whereas sites targeted by endogenous K in MEFs or upon individual overexpression of K carried only the K motif (Figures 2G, 2H, "K^{pMX}-only" and "shared" sites, and S3A), new locations bound by K at 48 hr of OSKM reprogramming were co-occupied by O and S and enriched for the motifs of all three factors (Figures 2G, 2H, K^{OSKM}-only sites, and S3A), revealing an unexpected dependence of K occupancy on O and S early in reprogramming. Conversely, the targeting of O and S, respectively, at 48 hr also strongly depended on the presence of the other reprogramming factors (Figure 2G). Specifically, when individually expressed, O and S bound many sites in open MEF chromatin that carried the motif of the respective reprogramming factor (Figures S3B–S3E), which did not overlap substantially between the factors (Figure 2I). Yet, when co-expressed in the context of OSKM for 48 hr, O and S co-occupied many new sites that also bound K and carried the motifs of all three factors (Figures 2G, 2I, and S3B–S3E). M was largely dispensable for the redistribution of K and OSK co-binding at 48 hr as co-expression of OSK, without M, led to engagement of largely the same sites at 48 hr as in OSKM-induced reprogramming (Figure S3F). We conclude that cooperative binding of O, S, and K is critical for the targeting of a vast number of genomic sites early in reprogramming and additionally restricts access to locations that carry the motif of only one reprogramming factor.

Figure 2. Characterization of OSKM Targets

(A) Fraction of TF binding sites within promoter-proximal (TSS ±2 kb) and -distal (>2 kb from TSS) regions. *$p < 0.0001$, two-sided binomial test.

(B) Fold enrichment of TF binding sites per chromatin state (Figure 1C) at the corresponding reprogramming stage, colored per column from highest to lowest.

(C) Heatmap of O, S, K, and M ChIP-seq signal for 48 hr and ESC peaks and corresponding signals for ATAC-seq and histone H3, ranked by ATAC-seq signal strength.

(D) Comparison of binding events of each reprogramming factor between 48 hr, pre-$i^{\#1}$, and ESCs (0/white = unbound, 1/blue = bound), at 100-bp resolution (bin).

(E) Hierarchical clustering of pairwise enrichments of O, S, K, and M binding events.

(F) (i) Clustering of O, S, K, and M binding events at 100-bp resolution (bin). OS, OK, and OSK co-binding events are marked. (ii) Differential enrichments of co-binding groups between ESCs and 48 hr.

(G) Heatmaps of ChIP-seq signal for K, S, or O peaks at 48 hr of OSKM or individual reprogramming factor expression (retrovirally [pMX] or inducibly [tetO]). Peaks were grouped based on presence/absence of peak calls comparing the OSKM and single TF expressing (pMX) samples. For K, binding events in MEFs were also plotted.

(H) Density plots of O, S, M, and K motifs in sets of K peaks defined in (G).

(I) Overlap of O, S, and K sites (number given) obtained from MEFs individually expressing O, S, and K for 48 hr (pMX, left) and MEFs co-expressing OSKM for 48 hr (right).

See also Figures S2 and S3 and Tables S2 and S4.

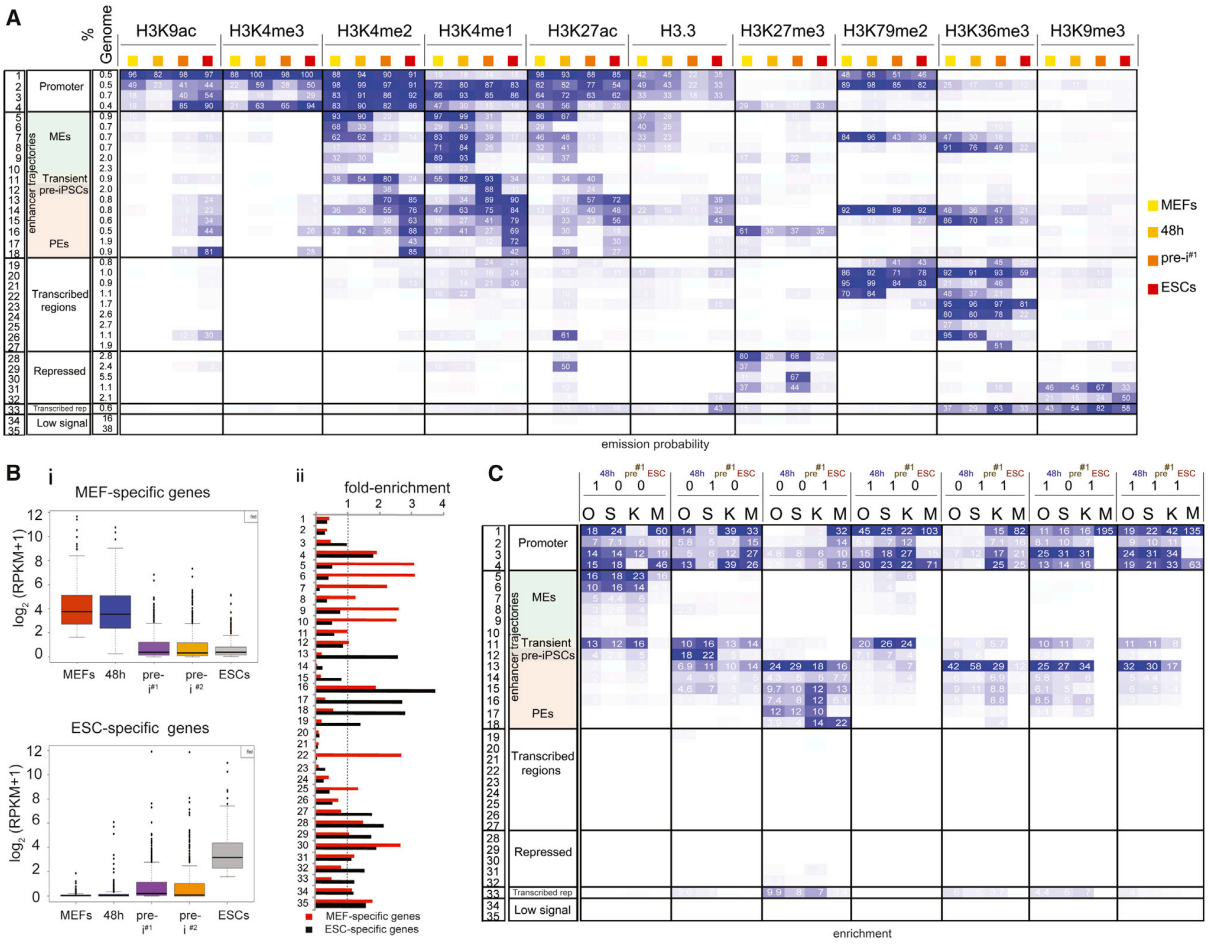


Figure 3. OSK Redistribution Mirrors Enhancer Reorganization

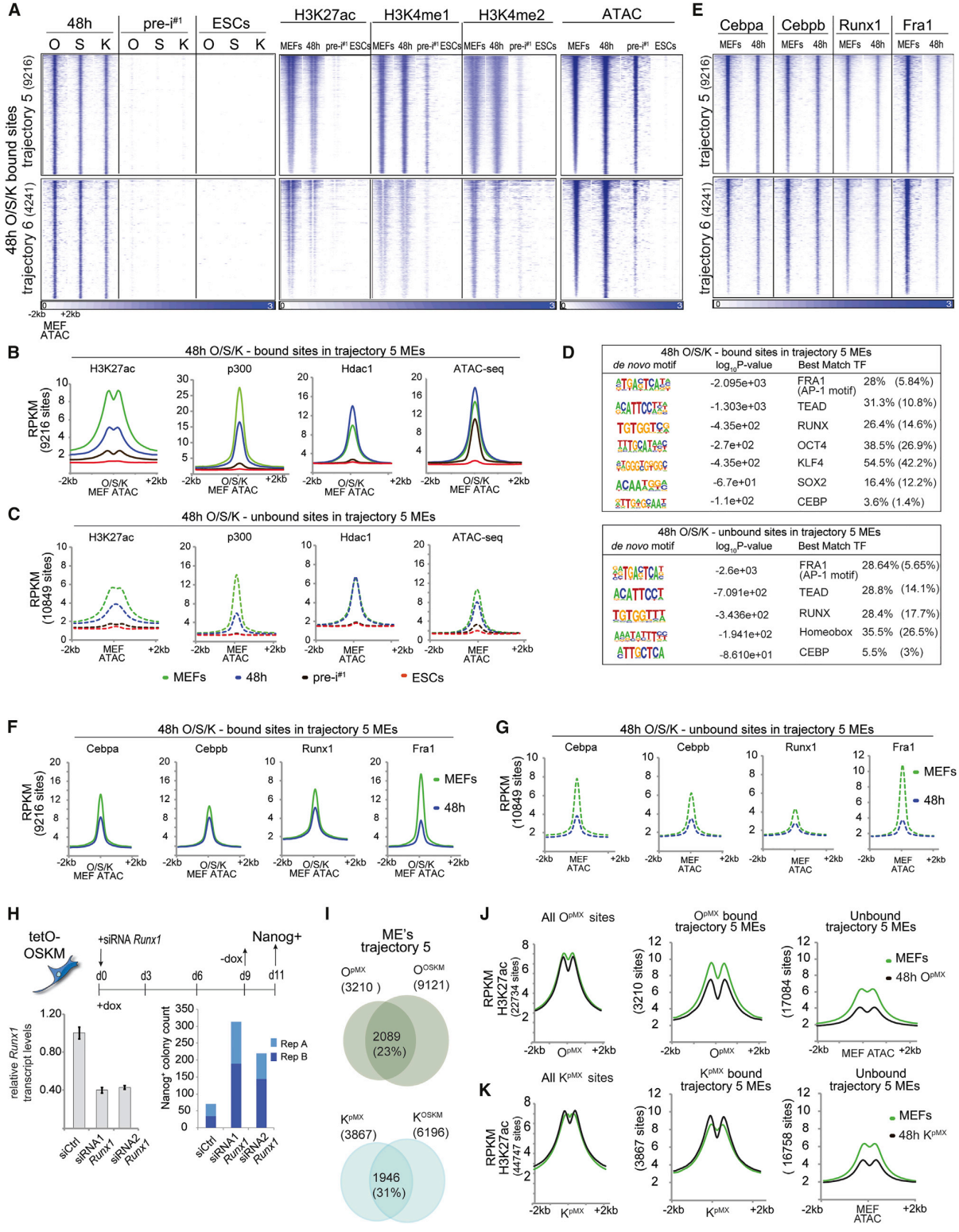
(A) Definition of the 35 chromatin trajectories that capture the major chromatin differences between our four reprogramming stages. The first three columns give the number, functional annotation, and genome fraction of each trajectory. Following columns are organized by histone mark and sub-ordered by reprogramming stage and display the frequency of each mark per reprogramming stage and trajectory, colored from 0 (white) to 100 (blue).

(B) (i) Boxplots of expression levels of MEF- and ESC-specific genes per reprogramming stage. (ii) Relative enrichment of each trajectory defined in (A) within ±20 kb of the TSS of MEF- and ESC-specific genes compared to ±20 kb of the TSS of all active genes. Values above the dashed line indicate higher enrichment in MEF- and ESC-specific genes, respectively.

(C) Fold enrichment of temporal O, S, K, and M binding events defined in Figure 2D for each trajectory in (A), colored per column from highest to lowest. See also Figure S4.

Enhancers Are Sites of Most Dramatic Chromatin Changes in Reprogramming

To examine the association between temporal OSKM binding events and chromatin changes during reprogramming, we derived an additional chromatin state model that took into consideration the combination of histone marks and H3.3 at any given genomic location within each reprogramming stage as well as the changes of these histone marks/H3.3 between the stages (Figure 3A) and defined 35 chromatin states that will be referred to as chromatin trajectories (tr.) hereafter. Based on the histone mark/H3.3 composition of each trajectory, we annotated genomic regions as candidate promoters (tr. 1–4), enhancers (tr. 5–18), units of transcription (tr. 19–27), repressed (tr. 28–32), transcribed repeats (tr. 33), and devoid of histone marks (tr. 34 and 35). These annotations were consistent with enrichments for genomic landmarks and expression of neighboring genes (Figures S4A and S4B). Differences in temporal histone marks/H3.3 composition between the reprogramming stages defined the stage-specific or constitutive chromatin character of each trajectory. We observed that the promoter states (tr. 1–4) did not carry a strong stage-specific identity (Figures 3A and S4B) consistent with promoter states being more conserved across cell types (Heintzman et al., 2009). Around 16% of the genome represented



(legend on next page)

enhancers, and, in contrast to promoters, the enhancer trajectories strongly differed in their histone mark composition between reprogramming stages and therefore likely in their activity and regulation (Figure 3A, tr. 5–18).

Based on the presence of the active enhancer mark H3K27ac in MEFs and its absence in ESCs, we defined MEF enhancers (MEs) (Figures 3A, S4A, and S4B). MEs were either inter- or intragenic (tr. 5, 6, 9, 10 and 7, 8, respectively) and typically located in the vicinity of genes with fibroblast-specific functions that tended to be expressed specifically early in reprogramming (Figures 3B, S4B, and S4C). Pluripotency enhancers (PEs) were defined based on the presence of H3K27ac in ESCs and near absence in MEFs (tr. 13–18). PEs of tr. 13 and 17 were intergenic, neighboring genes highly expressed in ESCs and implicated in stem cell maintenance, blastocyst formation, and developmental programs based on GO analysis (Figures 3A, 3B, and S4A–S4C). PEs associated with tr. 14, 15, 16, and 18 were predominantly intragenic or poised (carrying H3K27me3) and close to or within genes that tended to be either constitutively expressed or repressed during reprogramming (Figures 3A and 4B) and implicated in chromatin regulation and cell-fate specification.

One group of intergenic enhancers was marked by H3K4me1/2 at all four stages but displayed activity, defined by H3K27ac presence, in a transient manner at 48 hr and in pre-iPSCs (tr. 11, transient enhancers) (Figure 3A). These enhancers were linked to transiently expressed genes involved in various signaling pathways, most notably those acting in the bone morphogenetic protein (BMP) pathway (Figures S4B and S4C). Since BMPs have a positive role early in reprogramming (Samavarchi-Tehrani et al., 2010), activation of these enhancers may be critical for reprogramming progression. Other enhancers were active exclusively in pre-iPSCs (tr. 12) (Figure 3A), and their neighboring genes were enriched for neuronal ontologies (Figure S4C), consistent with the observation that neuronal genes can be ectopically induced during reprogramming (Ho et al., 2013). In summary, we identified enhancers as the most dynamic part of the epigenome during reprogramming and defined groups of enhancers that are selectively used at different reprogramming stages.

Changes in OSK Binding Mirror Enhancer Re-organization

To investigate how the redistribution of OSKM relates to the chromatin rearrangement during reprogramming, we intersected the genomic coordinates of temporal OSKM binding events (Figure 2D) with the chromatin trajectories (Figure 3A) and made several key observations (Figures 3C and S4D): first, we confirmed that OSK binding predominantly occurred in promoters and enhancers, whereas M preferred promoters throughout reprogramming. Second, the majority of O, S, and K binding events at 48 hr (100, 110 sites) occurred in promoters, MEs, and transient enhancers, indicating that early in reprogramming, O, S, and K predominantly target sites with open chromatin character in starting MEFs, unlike what has been reported for human cell reprogramming (Soufi et al., 2012). Third, O, S, and K binding at enhancers was typically observed when they were active (based on H3K27ac). For instance, pluripotency-specific O, S, K binding events (001 sites) were enriched specifically within PEs (tr. 13–18). Conversely, 48-hr-specific binding events (100 sites) enriched most in active MEs (tr. 5/6) and transient enhancers (tr. 11). These observations identified a dramatic shift of O, S, and K binding from MEs to PEs during reprogramming that accompanies their inactivation and selection/activation, respectively, and suggested that the reprogramming factors may directly control these two opposing processes. Fourth, we noted that a specific subset of PEs was targeted by O, S, and K early in reprogramming. Among all enhancers, constitutive binding by O, S, and K (111 sites) was most enriched in tr. 13 PEs, and occurred proximal to genes involved in stem cell maintenance, blastocyst formation (*Nanog*, *Lif*, *Esrrb*, *Stat3*, *Nodal*, etc.) and negative regulation of MAP kinase signaling (Figure S4E), supporting the conclusion that PE selection starts early in reprogramming and is finished in a stepwise manner throughout the process.

Since promoters displayed relatively little stage-specificity with respect to chromatin state and temporal reprogramming factor binding events, whereas enhancers were often stage specific for both (Figure 3C), we focused the rest of our study on the targeting and action of OSK at MEs and PEs to understand the regulation of ME silencing and PE selection as well as the regulation of distinct temporal binding patterns of OSK at enhancers.

Figure 4. ME Silencing Is Initiated Genome-wide Early in Reprogramming

(A) Heatmaps of O, S, K, H3K27ac, and H3K4me1/2 ChIP-seq signal and the ATAC-seq signal at all O, S, and K binding sites in tr. 5 and 6 MEs at 48 hr, ordered by the ATAC-seq signal strength. The total number of peaks is given in brackets.

(B) Metaplots of signal intensities for H3K27ac, p300, Hdac1, and ATAC-seq data in MEFs, 48 hr, pre-$i^{\#1}$, and ESCs at tr. 5 MEs occupied by O, S, or K at 48 hr, centered on ATAC-seq summits in MEFs.

(C) As in (B), except for tr. 5 MEs not bound by O, S, or K at 48 hr.

(D) De novo motifs identified at 48 hr O, S, or K- bound or unbound tr. 5 MEs. Last column: observed and expected motif frequencies (in parentheses).

(E) Heatmaps of somatic TF ChIP-seq signal at sites defined in (A).

(F) As in (B), except for somatic TFs in MEFs and at 48 hr.

(G) As in (C), except for somatic TFs in MEFs and at 48 hr.

(H) Schematic of the reprogramming experiment with Runx1 knockdown. *Runx1* transcript levels were determined at 48 hr (error bars represent SD) and Nanog-positive colonies were counted from two technical replicates (A and B).

(I) Comparison of O or K binding events at tr. 5 MEs in MEFs individually expressing the respective reprogramming factor (O^{pMX} or K^{pMX}) and MEFs co-expressing OSKM for 48 hr (O^{OSKM} or K^{OSKM}). Number of sites is given in brackets.

(J) Metaplots of signal densities for H3K27ac in starting MEFs and MEFs expressing only O for 48 hr (O^{pMX}) at all O^{pMX} bound sites and tr. 5 MEs bound or unbound by O^{pMX} at 48 hr.

(K) As in (J), but for K^{pMX}.

See also Figure S5.

MEs Are Suppressed Genome-wide Early in Reprogramming

Since it has remained unexplored how MEs become silenced during reprogramming and how the reprogramming factors contribute to this process, we examined active intergenic MEs captured by tr. 5/6 in more detail, approximately half of which were bound by O, S, or K at 48 hr. Considering tr. 5/6 MEs engaged by O, S, or K, we found extensive co-occupancy of these TFs at 48 hr, which was accompanied by an increased ATAC-seq signal (Figures 4A, 4B, S5A, and S5B). Later in reprogramming, in pre-iPSCs and ESCs, these MEs were depleted of active enhancer marks, OSK binding, and presented diminished chromatin accessibility defined by ATAC-seq (Figures 4A, 4B, S5A, and S5B), consistent with a predominant 100 OSK binding pattern at these enhancers.

Surprisingly, OSK-bound MEs displayed a lower level of the active enhancer mark H3K27ac at 48 hr compared to MEFs, which was corroborated by decreased binding of the H3K27 acetyltransferase p300 (Figures 4A, 4B, and S5B), indicating that somatic enhancer inactivation is initiated quite extensively very early in reprogramming. The enhancer marks H3K4me1/2 displayed smaller or no changes at 48 hr (Figures 4A, S5A, and S5B). The observation that the H3K27ac level was maintained or increased at other genomic locations (tr. 4, 9, and 11) at 48 hr (Figures S5C–S5E) argued against a global reduction of p300 activity and H3K27ac. The histone deacetylase Hdac1 was also present at OSK- targeted MEs and, unlike p300, its binding increased at 48 hr (Figure 4B), which was also seen in independent replicates, and, as for p300, occurred without alteration in its expression level (Figure S5F). We conclude that the change in balance of both p300 and Hdac1 observed at OSK-bound MEs at 48 hr likely accounts for the reduction in H3K27ac at these enhancers in the earliest phase of reprogramming. The completion of silencing of these enhancers occurred later indicating that ME inactivation is a stepwise process.

Unexpectedly, we observed that MEs of tr. 5/6 that were not engaged by OSK at 48 hr also had strongly reduced H3K27ac and p300 levels at 48 hr (Figures 4C, S5A, and S5B). These findings suggested that the disruption of the most active MEs takes place genome-wide early in reprogramming and extends beyond direct OSK targets. Interestingly, the increase in Hdac1 was specific to OSK-bound MEs and not observed at MEs that were not targeted by OSK (Figures 4B and 4C), potentially as a consequence of a direct action of O, S, or K.

Loss of Somatic TFs from OSK-Bound and Unbound MEs at 48 hr

To investigate how ME activity could be globally affected, we performed de novo motif scanning in OSK-bound and unbound tr. 5 MEs and identified DNA motifs of the Fra1 (AP-1 family), Tead, Runx, and Cebp families of TFs in both sets (Figure 4D). O, S, and K motifs were enriched specifically in the bound ME set (Figure 4D). We then performed ChIP-seq for the corresponding TFs Fra1, Cebpa, Cebpb, and Runx1, all highly expressed in MEFs (Figure S6J) and found that these TFs indeed occupied both OSK-bound and -unbound MEs in MEFs (Figures 4E–4G). At 48 hr, all four somatic TFs displayed reduced binding at OSK-bound and unbound MEs (Figures 4E–4G), which was independently supported by a reduction in ATAC-seq signal at MEs not targeted by OSK (Figures 4C, S5A, and S5B). These results suggested that the loss of somatic TFs from active MEs causes the reduction of p300 and H3K27ac at MEs at 48 hr genome-wide.

To test the functional significance of somatic TF loss, we performed siRNA-mediated knockdown of *Runx1* (Figure 4H) and *Cebpa/b* (Figure S5G) during reprogramming. Both treatments increased the number of Nanog-positive colonies indicating that the depletion of ME-bound somatic TFs represents a mechanism for improving reprogramming efficiency, likely by augmenting ME inactivation.

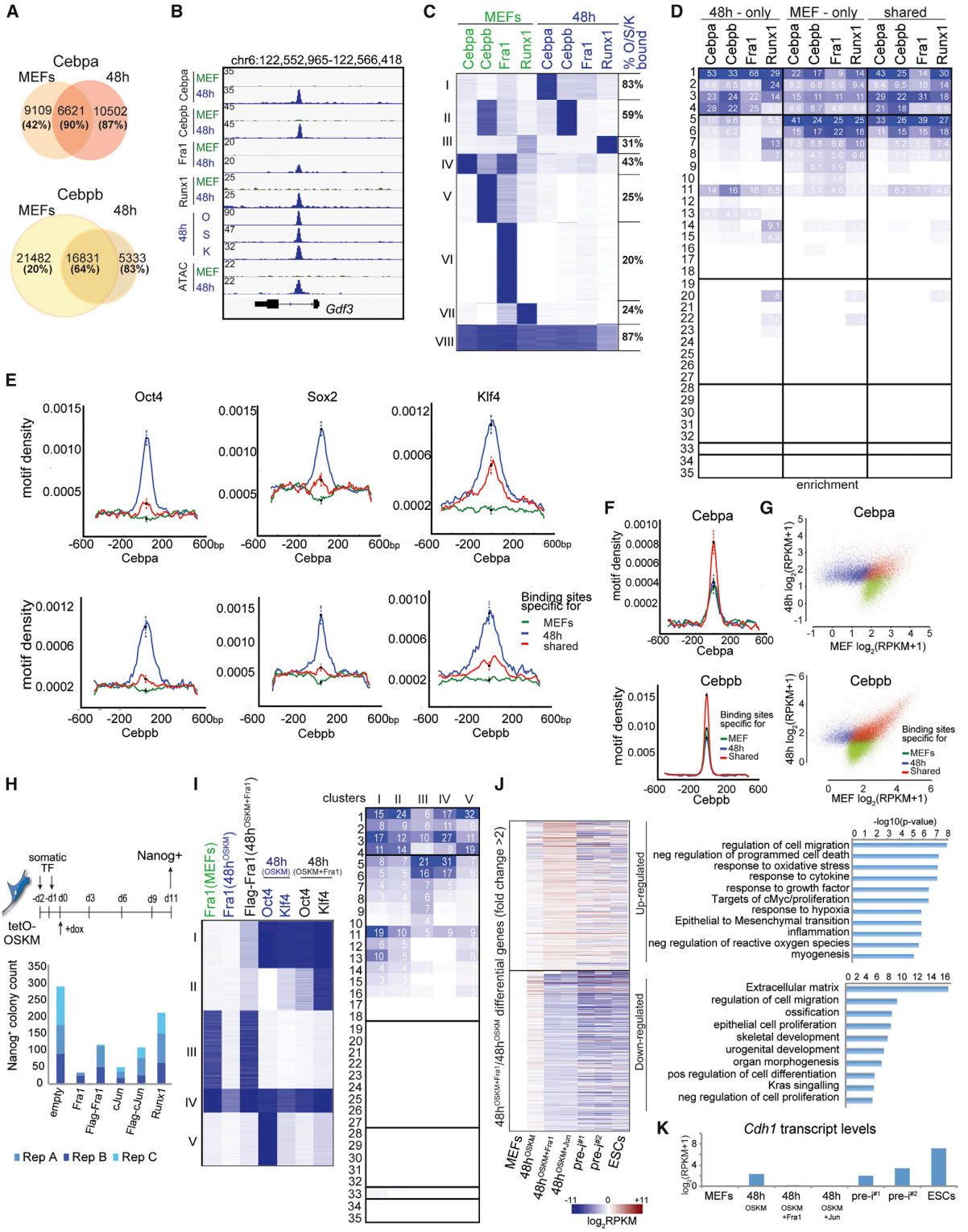
Reprogramming Factors Can Individually Induce ME Silencing

To determine whether OSK co-expression is required for global ME silencing, we analyzed O and K binding and H3K27ac levels at MEs in MEFs expressing only Oct4 or Klf4 for 48 hr. Only 23% and 31% of tr. 5 MEs bound by O and K, respectively, in the context of OSKM co-expression (O^{OSKM} and K^{OSKM}) were engaged by the single factors (O^{pMX} and K^{pMX}) (Figure 4I), emphasizing the importance of co-operative binding for the engagement of MEs. The H3K27ac level was reduced at tr. 5 MEs, but maintained over all binding sites of the individually expressed reprogramming factor (Figures 4J and 4K). Notably, individual reprogramming factors induced an H3K27ac drop at tr. 5 MEs comparable to that observed for OSKM-induced reprogramming (Figure S5H). Interestingly, for O, we observed a reduction of H3K27ac at tr. 5 MEs irrespective of its binding, but for K only at MEs not targeted by this reprogramming factor at 48 hr (Figures 4J and 4K), suggesting that Oct4, but not Klf4, may enhance silencing at its target MEs directly by increasing Hdac1 levels.

Somatic TF Redistribution at 48 hr Is Guided by OSK

A comparison of peak locations for Cebpa, Cebpb, Runx1, and Fra1 revealed the loss- and gain-of-binding events between MEFs and 48 hr as well as sites that were maintained (Figures 5A, 5B, S6A, and S6B). Gain and loss of binding occurred predominantly at sites occupied by only one of the somatic TFs (Figure 5C, clusters I–III and IV–VII), whereas binding sites maintained at 48 hr were more often co-occupied by the somatic TFs (Figure 5C, cluster VIII). Binding events lost or maintained at 48 hr were located predominantly in MEs (tr. 5–10), promoters (tr. 1–4), as well as transient enhancers (tr. 11) (Figure 5D, MEF-only and shared sites). Conversely, new binding sites of the somatic TFs at 48 hr were primarily enriched within promoters (tr. 1–4), transient enhancers (tr. 11), and tr. 13 PEs (Figure 5D, 48-hr-only sites). Together, these data revealed an unexpected redistribution of somatic TFs away from sites that include MEs toward new sites that include PEs.

48-hr-specific sites of Cepba, Cebpb, and Fra1, respectively, were extensively co-occupied by O, S, or K at 48 hr (>80%) and had a high density of OSK motifs, whereas MEF-specific sites displayed lower reprogramming factor occupancy (<42%) and lacked OSK motifs (Figures 5A, 5C, 5E, and S6B). Thus, somatic TFs relocate from MEs toward new sites that become available by binding of the reprogramming factors early in reprogramming,



(legend on next page)

suggesting that OSK directly guide this process, which, in turn, leads to the global destabilization of MEs. In support of an inter-dependency of somatic TFs and OSK, we found that somatic TF binding sites maintained at 48 hr were also targets of OSK early in reprogramming (Figures 5A, 5C, S6A, and S6B) and that Cebpb co-occupied many sites with OSK in pre-iPSCs (Figure S6C).

We also noted that somatic TF binding sites maintained at 48 hr (shared sites) exhibited higher normalized tag counts and motif density of the respective TF than either MEF- or 48-hr-specific peaks (Figures 5F, 5G, S6D, and S6E). These results suggested that binding events maintained early in reprogramming display higher affinity for the somatic TF compared to those lost or gained and that OSK induce the relocation of somatic TFs from one set of lower affinity binding sites to another.

Runx1 also relocated early in reprogramming but the new sites at 48 hr occurred often in transcribed units and did not overlap as extensively with OSK binding as Fra1 and Cebpa/b (Figures 5C, 5D, and S6A), suggesting that a different mechanism controls the redistribution of this TF. In addition, we noticed that more sites were lost and fewer sites gained at 48 hr for Fra1 compared to Cebpa/b or Runx1 (Figures 5A, 5C, and S6B, also seen in independent replicates), which raised the question of whether the level of Fra1 was altered. Indeed, RNA-seq revealed limited transcriptional changes early in reprogramming (Figures S6F and S6G) (Koche et al., 2011) with *Cebpa*, *Cepbb*, and *Runx1* transcript levels remaining largely unchanged, whereas *Fra1* transcript levels decreased substantially (2.7-fold) (Figures S6J and S6K; Table S2). Hence, repression of *Fra1* appears to be an additional mechanism that contributes to the loss of somatic TFs from MEs. Loss of Fra1 binding at its own locus at 48 hr (Figure S6L) could enhance the downregulation of this TF via its known auto-regulation (Verde et al., 2007). Of note, genes upregulated early in reprogramming were enriched for 48-hr-specific somatic TF binding and downregulated genes for MEF-specific somatic TF peaks (Figures S6H and S6I), suggesting that the redistribution of somatic TFs also contributes to the few expression changes detected early in reprogramming.

Fra1 Repression Is Critical for Somatic Program Silencing and Reprogramming

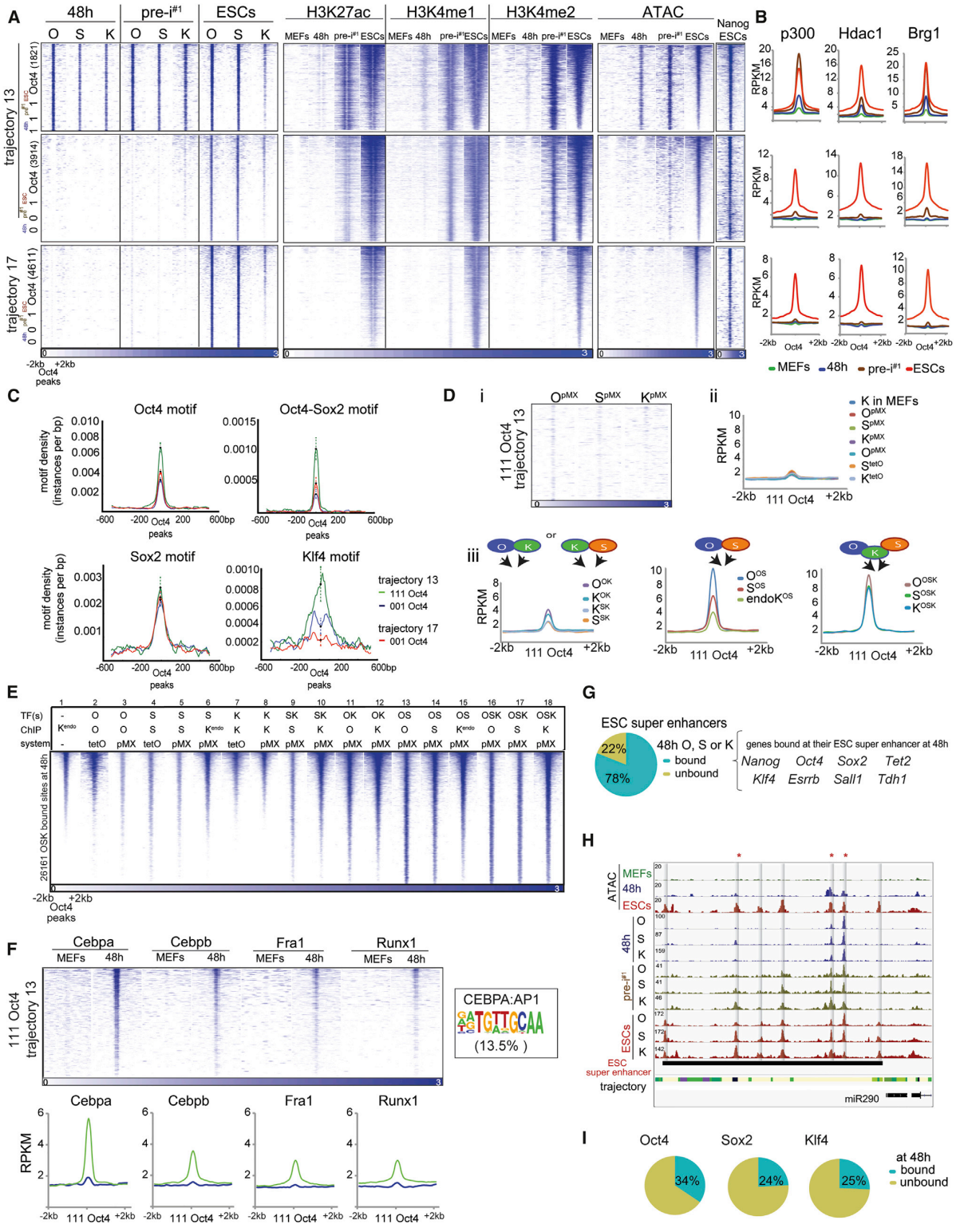
To test whether *Fra1* repression is critical for ME silencing, we ectopically expressed *Fra1* or a Flag-tagged version together with OSKM, which dramatically lowered the efficiency of reprogramming (Figures 5H and S6M). In comparison, *Runx1* overexpression had a more limited inhibitory effect on iPSC formation (Figures 5H and S6M), again hinting at differential control of reprogramming by Runx1. Ectopic expression of Flag-tagged Fra1 for 48 hr abrogated the loss of Fra1 from MEs that occurred early in OSKM-mediated reprogramming (Figures 5I, clusters III and IV, and S6L). Upon 48 hr overexpression together with OSKM, Fra1 also engaged new sites in promoters, PEs, and transient enhancers that were co-occupied by O and K (Figure 5I, cluster I, S was not tested here) and induced the targeting of these reprogramming factors to new sites (Figure 5I, clusters II and V) emphasizing the co-dependency of somatic TF and reprogramming factor binding events. Fra1 overexpression also reversed expression changes observed under standard reprogramming conditions at 48 hr and prevented the upregulation of the epithelial signature gene E-cadherin (Figures 5J and 5K). These data suggested that Fra1 loss from MEs is critical for their silencing and iPSC production. Overexpression of cJun, the binding partner of Fra1, was also detrimental for reprogramming (Figures 5H and S6M) (Liu et al., 2015) and produced similar expression changes as Fra1 overexpression (Figures 5J and 5K), suggesting that cJun may block reprogramming in synergy with Fra1.

Stepwise PE Selection Is Not Explained by Starting Chromatin State

Besides ME silencing, the selection of PEs is critical for reprogramming. The temporal differences in PE engagement, with a large number of PEs targeted by OSK only late in reprogramming and others first engaged at 48 hr or in pre-iPSCs (Figure 3C), prompted us to ask what distinguishes temporally different reprogramming factor binding at PEs. We focused this analysis on 111 and 001 O binding events in intergenic PEs of tr. 13 and 17 because of their association with genes involved in stem cell-related functions and high expression in ESCs (Figures S4B, S4C, and S4E).

Figure 5. Somatic TF Redistribution Early in Reprogramming

(A) Intersection of Cebpa or Cebpb binding sites between MEFs and 48 hr. The fraction of sites also bound by O, S, or K is given in brackets for each group.
(B) Genome browser view at the *Gdf3* locus of OSK and somatic TF binding and ATAC-seq data in MEFs and at 48 hr.
(C) K-means clustering of somatic TF binding events in MEFs and at 48 hr. The fraction of sites in each cluster also bound by O, S, or K is provided on the right.
(D) Fold enrichment of MEF-only, 48-hr-only, and shared binding sites of somatic TFs from (A) in chromatin trajectories defined in Figure 3A, colored per column from highest to lowest values.
(E) Density of O, S, or K motifs at MEF-only, 48-hr-only, and shared Cebpa (top) and Cebpb (bottom) sites from (A). Error bars = 95% confidence interval at summits.
(F) As in (E), but for Cebpa (top) and Cebpb (bottom) motifs.
(G) MEF and 48 hr input-normalized ChIP-seq signal for MEF-only, 48-hr-only, and shared binding events of Cebpa and Cebpb from (A).
(H) Schematic of the reprogramming experiment with retroviral overexpression of somatic TFs. Nanog-positive colony counts from three biological replicates are given.
(I) K-means clustering of Fra1 peaks in MEFs and Fra1, O, and K peaks at 48 hr of OSKM or OSKM+Fra1 co-expression. Right: fold enrichments of each cluster on the left in chromatin trajectories defined in Figure 3A, colored per column from highest to lowest values.
(J) Heatmap of differential gene expression between each of the reprogramming stages indicated at the bottom relative to MEFs, for genes 2-fold differentially expressed between 48 $hr^{OSKM+Fra1}$ and 48 hr^{OSKM}. Right, GO ontologies of these genes.
(K) E-cadherin (*Cdh1*) transcript level for indicated samples based on RNA-seq data.
See also Figure S6 and Table S2.



(legend on next page)

We first analyzed enhancer-associated histone marks at 111 and 001 O binding events in tr. 13 PEs and 001 O sites in tr. 17 PEs (Figure 6A). All sites existed in a closed chromatin conformation in MEFs lacking active histone marks and ATAC-seq signal (Figures 6A and S7A). For 111 sites in tr. 13 and 001 sites in tr. 17, O binding correlated with the gain of the enhancer marks H3K4me1/2 and H3K27ac and ATAC-seq signal (Figures 6A and S7A), suggesting that chromatin opening and selection of these sites is linked to reprogramming factor binding. At 001 sites in tr. 13 PEs, the gain of enhancer marks and chromatin accessibility preceded O binding (Figure 6A), implying a role for non-reprogramming TFs in the opening of these sites. Regardless, the transition of PEs from "closed" to "open" chromatin was associated with the recruitment of Brg1 (Figure 6B).

Interestingly, at 111 sites in tr. 13 PEs the level of H3K4me1/2 and H3K27ac was much lower at 48 hr than in pre-iPSCs and ESCs (Figures 6A and S7A). This pattern was recapitulated by p300 and Hdac1 binding (Figure 6B), demonstrating that reprogramming factor binding at 48 hr induced the selection of these PEs but not their full activation, which occurred at a later reprogramming stage (Figures 6A and 6B). One intriguing hypothesis is that reprogramming factor binding to PEs early in the process allows for the binding of additional TFs at later reprogramming stages, which is required for full enhancer activation. Regardless, these data showed that the stepwise selection and activation of PEs is largely controlled by parameters beyond chromatin state.

Motif Density and OSK Co-occupancy Distinguish Early and Late-Engaged PEs

Since the chromatin state in MEFs did not distinguish early (111) and late- (001) engaged PEs, we examined properties of the underlying DNA sequence. Early bound O sites in tr. 13 PEs carried significantly more Oct4, Oct4/Sox2 composite, and Klf4 consensus motifs compared to late-bound sites in tr. 13 and 17 (Figure 6C). De novo motif scanning also revealed a stronger enrichment of the Klf4 motif in 111 O sites in tr. 13 PEs compared to 001 sites in tr. 17 PEs (Figure S7B). Consistent with these differences in motif occurrence, 111 O sites in tr. 13 PEs were co-bound by S and K when they were first engaged at 48 hr, whereas tr. 13 and 17 PEs bound by O late (001) were predominantly co-occupied with S but not K in ESCs (Figures 6A and S7C). These data demonstrated that OSK co-occupancy is associated with PE selection early and OS co-binding with PE engagement late in reprogramming, which is driven by motif presence. Despite co-binding by OSK at 48 hr, 111 O sites in tr. 13 PEs were mostly bound by OS in ESCs (Figures 6A and S7C) in agreement with K binding being more distinct to O and S binding in ESCs (Figures 2E and 2F) and being influenced by other TFs in ESCs.

PE Engagement Early in Reprogramming Requires Collaborative Binding by OSK

To test mechanistically how the selection of PEs occurs early in reprogramming, we determined the independent ability of O, S, and K to engage these sites at 48 hr. We found that tr. 13 PEs were not targeted when O, S, and K were individually expressed (Figures 6Di, ii, S3B, S3C, and S3G), indicating that the ability of these reprogramming factors to act as pioneer factors is not at play for the opening of these sites. Retroviral co-expression of combinations of reprogramming factors and mapping of binding sites at 48 hr further demonstrated that OSK co-expression was sufficient for the selection of tr. 13 PEs at 48 hr, showing that ectopic M is not essential for PE selection, and additionally revealed lower occupancy when two reprogramming factors were expressed (OS, SK, OK) compared to three (OSK) (Figure 6Diii). Though these data were consistent with PE section requiring a collaborative mode of action by OSK, one exception was that OS co-expression resulted in binding levels close to those seen with OSK co-expression, particularly for O, despite the lack of ectopic K (Figure 6Diii). This result likely can be explained by the relocation of endogenously expressed K to these sites in OS-expressing MEFs (Figure 6Diii). Thus, we conclude that the selection of PEs early in reprogramming requires the collaborative action of O, S, and K and suggest that the necessity of OSK for reprogramming is linked to their ability to open a subset of PEs together.

We made similar observations when considering all sites that were co-occupied by O, S, and K at 48 hr of OSKM-induced reprogramming. Only ~30% were accessible to individually expressed O, S, or K, mainly at locations already engaged by endogenous Klf4 in MEFs (Figure 6E, columns 1–8). The number

Figure 6. Stepwise Selection of PEs and OSK Requirement

(A) Heatmaps of O, S, K, H3K27ac, H3K4me1/2, and Nanog ChIP-seq signal and ATAC-seq data in indicated reprogramming stages at 111 or 001 O binding sites within tr. 13 and 17 PEs, sorted by ESC ATAC-seq signal intensity. Number of peaks in each set is given in brackets.

(B) Metaplots of signal intensities of p300, Hdac1, and Brg1 for sites in (A).

(C) Motif density for sites in (A), with 95% confidence interval at the summits.

(D) (i) Heatmaps of O, S, and K ChIP-seq signal at 111 O sites in tr. 13 PEs in MEFs individually expressing O, S, or K for 48 hr. (ii) Metaplots of signal intensities of the indicated reprogramming factor individually expressed (pMX or tetO) in MEFs for 48 hr and of Klf4 in MEFs in tr. 13 111 O sites. (iii) As in (ii), except for MEFs expressing OK, SK, OS, or OSK for 48 hr. Binding of endogenous K in OS-expressing MEFs is also given.

(E) Heatmap of ChIP-seq signal for the factor indicated by ChIP, in MEFs ectopically expressing one or combinations of reprogramming factor(s) for 48 hr (TF) using retroviral (pMX) or inducible (tetO) expression (system), for sites co-bound by OSK at 48 hr of OSKM-induced reprogramming, sorted by Klf4 signal in MEFs. K^{endo} refers to targets of endogenously expressed K.

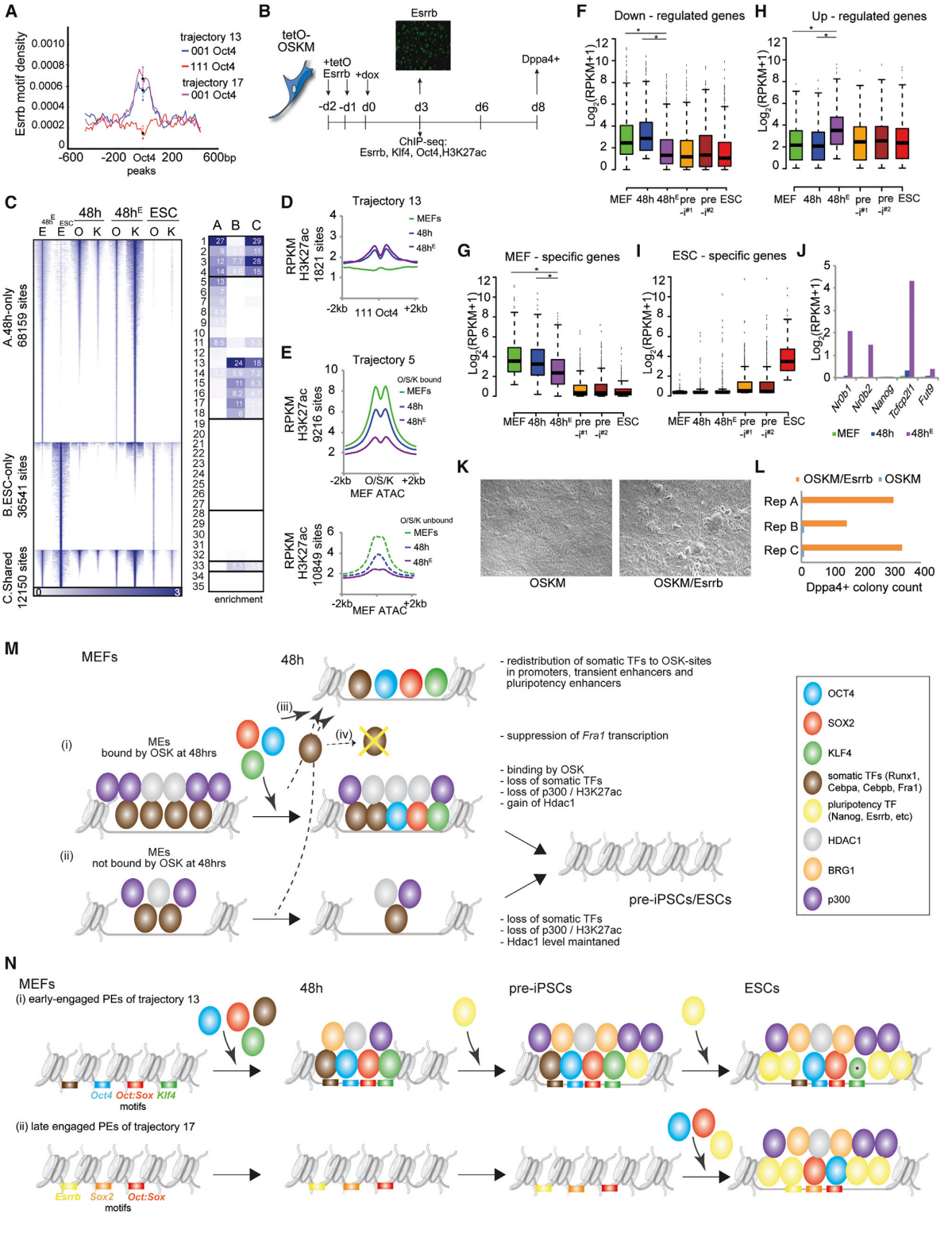
(F) As in (D), except for binding of somatic TFs in MEFs and at 48 hr. The given CEBPA:AP1 composite motif was identified in 13.5% of tr. 13 111 O sites.

(G) Fraction of ESC super enhancers occupied by O, S, or K at 48 hr and genes associated with OSK-targeted ESC super enhancers.

(H) Genome browser view of OSK ChIP-seq, ATAC-seq data, and chromatin trajectories (color coded as in Figure S4A) at the *mir290* ESC super enhancer. Gray bars indicate seven sub-elements engaged by OSK in ESCs and the asterisks mark those bound by OSK at 48 hr.

(I) Fraction of locations within ESC super enhancers that are bound by O, S, or K in ESCs and also engaged by the respective TF at 48 hr.

See also Figure S7 and Tables S5.



(legend on next page)

of accessible sites increased when double combinations of reprogramming factors were expressed, rising from SK, OK to OS (Figure 6E, columns 9–15).

We also noted that 13.5% of 111 O sites in tr. 13 PEs carried the CEBPA:AP1 composite motif, and that Cebpa, and to lesser extent Cebpb, Fra1, and Runx1, occupied these PEs with OSK at 48 hr (Figures 6F and 5B). Additionally, Fra1 extensively engaged tr. 13 PEs early in reprogramming upon overexpression together with the reprogramming factors (Figure 5I, clusters I, II, and V), further supporting a link between somatic TFs and OSK at PEs. At PEs, somatic TFs may be required for their selection, prevent their activation early in reprogramming, or may simply bind due to the open chromatin character, which will need to be studied further.

Early Engaged PEs Are Close to Core Pluripotency Genes

Recently, super enhancers, defined as dense clusters of enhancers with high activity, received attention as *cis*-regulatory elements of genes that control cell identity (Whyte et al., 2013). We found that ESC super enhancers gained enhancer marks gradually during reprogramming and that their neighboring genes were activated progressively (Figures S7D and S7E). ESC super enhancers were enriched most strongly in tr. 13 PEs (Figure S7F) and typically engaged by O, S, and K early in reprogramming (of 231 ESC super enhancers 78% were bound by O; 61% by S; 66% by K at 48 hr) including those at the *mir290*, *Pou5f1*, *Sox2*, *Klf4*, *Tdh1*, and *Nanog* loci (Figures 6G, S7G, and S7H). Thus, critical regulatory sites of the pluripotent state are among the PEs that are selected early in reprogramming. Interestingly, only a subset of sites within ESC super enhancers bound by O, S, and K in ESCs was engaged at 48 hr (Figures 6H, 6I, and S7I), suggesting that super enhancers do not act as a single entity and that the opening at specific sites early in reprogramming may be critical for full selection/activation later in the process. Notably, ESC super enhancers represented only a small fraction of all early engaged PEs as only ~4% of the 111 O sites in tr. 13 were located within them.

Esrrb Enhances Both ME Inactivation and PE Selection

The data described above argue in favor of a model where cooperative binding of TFs, including both reprogramming factors and endogenously expressed TFs, dictates their genomic targeting and thereby enhancer selection. For late-engaged PEs (001 sites), we therefore hypothesized that additional TFs that become available progressively during reprogramming are required for selection either prior to or together with OS. In support of this idea, we observed that in ESCs PEs were occupied by additional TFs that become expressed later in reprogramming, such as Nanog and Esrrb (Figure 6A; Table S5). Moreover, 001 O sites in PEs could be distinguished from 111 sites by the presence of the Esrrb motif (Figure 7A), establishing Esrrb, which is turned on very late in reprograming (Buganim et al., 2012; Pasque et al., 2014; Polo et al., 2012) (Figures S7J and S7K), as a unique candidate to test our hypothesis.

To this end, we expressed Esrrb alongside OSKM from an inducible lentivirus and profiled binding of Esrrb, O, K, and H3K27ac at 48 hr (48-hr^{E} samples) (Figures 7B and S7K). As for OSKM, most Esrrb binding sites at 48 hr differed from those in ESCs (Figure 7C). 48-hr-specific binding occurred predominantly in promoters, MEs, and transient enhancers (Figure 7C, group A), whereas ESC-specific sites enriched in PEs (Figure 7C, group B). 25% of ESC targets of Esrrb became engaged at 48 hr, many of which were located in promoters and, as seen for OSK, in tr. 13 PEs (Figure 7C, group C). The sites in group C were targeted by O and K in ESCs as well as upon OSKM/Esrrb co-expression at 48 hr, but only a third were engaged by O and K at 48 hr when merely OSKM were overexpressed (Figure 7C). Similarly, 2291 sites in PEs of tr. 13–18 normally engaged by O, S, or K only late (001 sites) (including 882 sites in tr. 13 and 415 in tr. 17), were targeted by O or K at 48 hr upon OSKM/Esrrb overexpression. Thus, PEs not accessible to the reprogramming factors early in reprogramming became accessible early when Esrrb was co-expressed. These data provide evidence for the cooperation of a pluripotency TF with OSK in the selection of PEs and highlight the need of additional pluripotency TFs for the reconstitution of the pluripotency network.

Figure 7. Control of ME Decommissioning and PE Selection by Esrrb

(A) Esrrb motif density in 111 and 001 O peaks in tr. 13 and 17 PEs. Error bars = 95% confidence interval at summits.

(B) Schematic of the reprogramming experiment with lentiviral overexpression of *Esrrb* (tetO-Esrrb). Image: Esrrb expression was confirmed by immunostaining at day 3 (in green).

(C) Heatmap of Esrrb (E) ChIP-seq signal for Esrrb peaks identified in ESCs and at 48 hr of co-expression of OSKM and Esrrb (48 hr^{E}). Peaks were divided into three groups (A–C) based on their reprogramming stage specificity. The O and K signal at 48 hr of OSKM (48 hr) or OSKM/Esrrb expression (48 hr^{E}) and in ESCs for the same sites is also shown. Right: fold enrichments of sites in groups A–C in chromatin trajectories defined in Figure 3A, colored per column from highest to lowest value.

(D) Metaplot of signal intensity of H3K27ac at 111 O sites in tr. 13 PEs for MEFs, 48 hr, and 48 hr^{E} (OSKM/Esrrb).

(E) As in (D), except for OSK-bound and unbound tr. 5 MEs, centered on ATAC-seq summits in MEFs.

(F) Boxplots of expression levels of genes downregulated at 48 hr of reprogramming with OSKM/Esrrb (48 hr^{E}) relative to OSKM alone (48 hr). Asterisks mark any significant differences between MEFs, 48-hr, and 48-hr^{E} samples (Wilcoxon test, adj. $p < 0.05$).

(G) As in (F), for MEF-specific genes.

(H) As in (F), for upregulated genes.

(I) As in (F), for ESC-specific genes.

(J) Expression of pluripotency genes known to be regulated by Esrrb.

(K) Bright-field image at day 6 of reprogramming with OSKM and OSKM/Esrrb.

(L) Count of Dppa4-positive colonies at day 8 of OSKM or OSKM/Esrrb expression from three biological replicates.

(M) Model for the functions of OSK at MEs.

(N) Model for the functions of OSK at PEs. Asterisk indicates reduced K binding in ESCs.

See also Figure S7 and Table S2.

Interestingly, at 48 hr, tr. 13 PEs reached a similar level of H3K27ac in the presence of Esrrb as reprogramming cells not exposed to Esrrb (Figure 7D). However, the average level of H3K27ac was much lower at tr. 5 MEs at 48 hr upon Esrrb expression indicating even more pronounced ME silencing (Figure 7E). Consistent with this observation, MEF signature genes were more strongly repressed at 48 hr with Esrrb (48 hr^{E} versus 48 hr; Figures 7F, 7G, and S7L). With the exception of a few pluripotency genes such as *Nr0b1/2*, *Tcfcp2l1*, and *Fut9*, Esrrb did not induce precocious expression of ESC-specific genes at 48 hr but instead induced genes involved in metabolic pathways ectopically (Figures 7H–7J and S7L). Last, we found that the molecular changes induced by Esrrb early in reprogramming correlated with a more than 100-fold increase in the number of $Dppa4^{+}$ colonies and shortened kinetics of iPSC-like colony formation (Figures 7K and 7L). Together, these data demonstrate a dramatic effect of the pluripotency TF Esrrb on MEF identity, promoting the inactivation of MEs, and, in parallel, on the induction of the pluripotency program by enhancing PE selection.

DISCUSSION

Our study provides a comprehensive analysis of OSKM occupancy at four reprogramming stages. Among the four reprogramming factors, M is distinct as it primarily targets promoters throughout reprogramming, whereas O, S, and K favor enhancers. We found that OSK switch from somatic to pluripotency enhancers during reprogramming and, unexpectedly, orchestrate both the inactivation of MEs and PE selection. Most importantly, our work revealed that the selection of genomic target sites of OSK and the opposing effects on MEs and PEs are controlled by the combinatorial interplay of O, S, and K with endogenously expressed TFs, many with a stage-specific expression.

The extensive silencing of MEs at 48 hr, affecting MEs bound by OSK and those not targeted by the reprogramming factors, was a particularly surprising finding and indicated the widespread interference with the somatic epigenetic network very early in reprogramming. We determined that OSK initiate the silencing of MEs by at least three distinct mechanisms: first, OSK bind to approximately 50% of the most active MEs and increase their Hdac1 levels, potentially attributable to the interaction of Hdac1 with Oct4 (Pardo et al., 2010) (Figure 7Mi). Second, OSK induce the removal of somatic TFs from OSK-bound and unbound MEs (Figure 7Mi and Mii). Mechanistically, the loss of somatic TFs from MEs is accomplished by their relocation away from MEs to new sites including PEs that become bound by OSK at 48 hr and carry the motifs for the reprogramming factors and the somatic TFs (Figure 7Miii). Third, OSK expression leads to a decrease in *Fra1* transcript levels, which contributes to the extensive loss of this TF from MEs (Figure 7Miv).

Binding sites in MEs contained fewer consensus DNA binding motifs of each reprogramming factor than those at PEs, suggesting that the interaction of OSK with MEs differs from that at PEs. Therefore, we propose that the targeting of MEs may involve non-consensus motifs that are accessible to the reprogramming factors due to the open chromatin state or protein-protein interactions with endogenously expressed factors. For instance, protein-protein interactions with Cebpa/b, Fra1, or Runx1 may contribute to the recruitment of OSK to MEs. Reciprocally, the same interactions could facilitate the redistribution of somatic TFs to new sites with OSK binding at 48 hr and contribute to the combinatorial binding between OSK and somatic TFs. The presence of fewer cognate OSK motifs at MEs could mediate a generally weaker binding that, in turn, facilitates the disengagement of OSK from these sites when somatic TFs become unavailable.

In addition to MEs, OSK engage a substantial number of PEs at 48 hr including sites in super enhancers neighboring critical pluripotency-associated genes (Figure 7Ni). However, the majority of PEs are bound only at later stages (Figure 7Nii), revealing that PEs selection is a stepwise process. The different kinetics of PE selection was not explainable by differences in the chromatin state in starting MEFs. Instead, we propose that the timing of PE selection is dictated by (1) the collaborative binding among O, S, and K, and with additional, endogenously expressed TFs, and (2) *cis*-encoded properties, i.e., the presence and combination of motifs at these sites (Figure 7N). O, S, and K together, potentially with somatic TFs, are required for the opening of PEs early in reprogramming (Figure 7Ni), which perhaps explains why this combination of reprogramming factors is so successful in establishing pluripotency. At early engaged PEs, opening by OSK does not result in strong enhancer activation at 48 hr, which likely requires other TFs that become available later. Conversely, late-engaged PEs are targeted by OS, without K, indicating that OS alone are not sufficient to effectively compete with nucleosomes at these sites early in reprogramming (Figure 7Nii). Here, additional TFs that only become available later, such as Esrrb, are required for their selection in concert with OS (Figure 7Nii).

Ectopic Esrrb not only influenced OSK binding at PEs, but also bound MEs and facilitated their silencing. Equally, Fra1 acted on both MEs and PEs when overexpressed. Thus, stage-specific TFs, including both somatic and pluripotency TFs, influence OSK binding, ME silencing, and PE selection, reinforcing the idea of the combinatorial control of TF binding during reprogramming. These observations and the fact that targeting of PEs in closed chromatin require binding of multiple TFs indicate that the pioneer factor model proposed for *human* somatic cell reprogramming (Soufi et al., 2012, 2015) does not act at enhancers in *mouse* cell reprogramming. Additional work will be required to understand the differences between reprogramming processes in different species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Cell lines, culture conditions, and reprogramming experiments
  - Immunofluorescence
  - Native ChIP-seq (N-ChIP)

- Cross-linked ChIP-seq (X-ChIP)
- ATAC-seq library construction and sequencing
- RNA-seq

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data Analysis and Visualization
  - ChIP-seq and ATAC-seq data validation
  - Correlation of our datasets with imputated datasets
  - Differential gene expression analysis
  - Defining combinatorial OSKM binding groups per reprogramming stage
  - Determination of temporal OSKM binding groups
  - Transcription factor clusters
  - Ontology Annotation
  - ChromHMM modeling parameters
  - TF enrichment in the vicinity of differentially expressed genes early in reprogramming (Figures S6H and S6I)
  - Positional expression plots (Figures S1H and S4B)
  - Calculations of fold-enrichment
  - Assigning peaks to TSS (+/−2Kb) regions
  - Motif analyses
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.12.016.

AUTHOR CONTRIBUTIONS

Conceptualization, C.C. and K.P.; Methodology, C.C. and K.P.; Software, C.C., P.F., S.S., G.B., and J.E.; Validation, C.C. and P.F.; Investigation, C.C., B.P., and S.B.; Formal Analysis, C.C., P.F., S.S., G.B., J.E., and K.P.; Data Curation, C.C., P.F., S.S., and G.B.; Writing – Original Draft, C.C. and K.P.; Writing – Review & Editing, C.C., P.F., B.P., G.B., J.E., and K.P.; Resources, C.C., J.E., and K.P.; Visualization, C.C., P.F., J.E., and K.P.; Supervision, J.E. and K.P.; Project Administration, K.P.; Funding Acquisition, J.E. and K.P.

ACKNOWLEDGMENTS

We are grateful to N.S.B. Thomas, K. Zaret, S. Smale, M. Pellegrini, S. Kurdistani, N. Davidson, W. Lowry, and K.P.'s lab for discussions and reading of the manuscript. C.C. was supported by CIRM (TG2-01169) and Leukemia and Lymphoma Research (#10040) fellowships; P.F. by CIRM (TG2-01169) and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research (BSCRC) at UCLA; G.B. by Whitcome, Dissertation Year and QCB fellowships from UCLA; J.E. by NIH (R01ES024995; U01HG007912), NSF (CAREER Award 1254200), and a Sloan Fellowship; and K.P. by the BSCRC, David Geffen School of Medicine, and Jonnson Comprehensive Cancer Center at UCLA, CIRM, and NIH (P01-GM099134).

Received: September 16, 2016
Revised: December 7, 2016
Accepted: December 14, 2016
Published: January 19, 2017

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bar-Joseph, Z., Gifford, D.K., and Jaakkola, T.S. (2001). Fast optimal leaf ordering for hierarchical clustering. Bioinformatics *17*(*Suppl 1*), S22–S29.

Beard, C., Hochedlinger, K., Plath, K., Wutz, A., and Jaenisch, R. (2006). Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis *44*, 23–28.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Buganim, Y., Markoulaki, S., van Wietmarschen, N., Hoke, H., Wu, T., Ganz, K., Akhtar-Zaidi, B., He, Y., Abraham, B.J., Porubsky, D., et al. (2014). The developmental potential of iPSCs is greatly influenced by reprogramming factor selection. Cell Stem Cell *15*, 295–309.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: Network biology applied to stem cell engineering. Cell *158*, 903–915.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Chen, J., Chen, X., Li, M., Liu, X., Gao, Y., Kou, X., Zhao, Y., Zheng, W., Zhang, X., Huo, Y., et al. (2016). Hierarchical Oct4 binding in concert with primed epigenetic rearrangements during somatic cell reprogramming. Cell Rep. *14*, 1540–1554.

Efron, B., and Tibshirani, R. (1991). Statistical data analysis in the computer age. Science *253*, 390–395.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

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

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. *33*, 364–376.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

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

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

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

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049–1061.

Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell *8*, 96–105.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

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

Liu, J., Han, Q., Peng, T., Peng, M., Wei, B., Li, D., Wang, X., Yu, S., Yang, J., Cao, S., et al. (2015). The oncogene c-Jun impedes somatic cell reprogramming. Nat. Cell Biol. *17*, 856–867.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell *1*, 55–70.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Morita, S., Kojima, T., and Kitamura, T. (2000). Plat-E: an efficient and stable system for transient packaging of retroviruses. Gene Ther. *7*, 1063–1066.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods *5*, 621–628.

Pardo, M., Lang, B., Yu, L., Prosser, H., Bradley, A., Babu, M.M., and Choudhary, J. (2010). An expanded Oct4 interaction network: Implications for stem cell biology, development, and disease. Cell Stem Cell *6*, 382–395.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. *37*, e123.

Pasque, V., Tchieu, J., Karnik, R., Uyeda, M., Sadhu Dimashkie, A., Case, D., Papp, B., Bonora, G., Patel, S., Ho, R., et al. (2014). X chromosome reactivation dynamics reveal stages of reprogramming to pluripotency. Cell *159*, 1681–1697.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., et al. (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell *151*, 1617–1632.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell *7*, 64–77.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994–1004.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. Cell *136*, 364–377.

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., et al. (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. Nat. Cell Biol. *15*, 872–882.

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

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-seq. Bioinformatics *25*, 1105–1111.

Tripathi, S., Pohl, M.O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D.A., Moulton, H.M., DeJesus, P., Che, J., Mulder, L.C., et al. (2015). Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. Cell Host Microbe *18*, 723–735.

Verde, P., Casalino, L., Talotta, F., Yaniv, M., and Weitzman, J.B. (2007). Deciphering AP-1 function in tumorigenesis: Fra-ternizing on target promoters. Cell Cycle *6*, 2633–2639.

Wagschal, A., Delaval, K., Pannetier, M., Arnaud, P., and Feil, R. (2007). Chromatin immunoprecipitation (ChIP) on unfixed chromatin from cells and tissues to analyze histone modifications. CSH Protoc. http://dx.doi.org/10.1101/pdb.prot4767.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rat monoclonal anti-Nanog	eBioscience	Cat#14-5761-80
Rabbit polyclonal anti-Nanog	cosmobio	Cat#REC-RCAB001P
Goat polyclonal anti-DPPA4	R&D	Cat#AF3730
Goat polyclonal anti-Oct4	RnD	Cat#AF1759
Goat polyclonal anti-Sox2	RnD	Cat#AF2018
Goat polyclonal anti-Klf4	RnD	Cat#AF3158
Goat polyclonal anti-cMyc	RnD	Cat#AF3158
Mouse monoclonal anti-Esrrb	RnD	Cat#H6705
Rabbit polyclonal anti-H3K9ac	Abcam	Cat#ab4441
Rabbit polyclonal anti-H3K4me3	Abcam	Cat#ab8580
Rabbit polyclonal anti-H3K4me2	Abcam	Cat#ab7766
Rabbit polyclonal anti-H3K4me1	Abcam	Cat#ab8895
Rabbit polyclonal anti-H3K27me3	Active Motif	Cat#39155
Rabbit polyclonal anti-H3K27ac	Abcam	Cat#ab4729
Rabbit polyclonal anti-H3K36me3	Abcam	Cat#ab9050
Rabbit polyclonal anti-H3K79me2	Active Motif	Cat#39143
Rabbit polyclonal anti-H3K9me3	Abcam	Cat#ab8898
Mouse monoclonal anti-H3K9me3	Millipore	Cat#05-1242
Rabbit polyclonal anti-H3	abcam	Cat#ab1791
Mouse monoclonal anti-H3.3	Abnova	Cat#H00003021-M01
Rabbit polyclonal anti-p300	SantaCruz	Cat#sc-585
Rabbit polyclonal anti-Runx1	Novus Biologicals	Cat#NBP1-61277
Rabbit polyclonal anti-Fra1	SantaCruz	Cat#sc-183X
Rabbit polyclonal anti-Cebpa	SantaCruz	Cat#sc-61X
Rabbit polyclonal anti-Cebpb	SantaCruz	Cat#sc-150X
Rabbit polyclonal anti-Hdac1	abcam	Cat#ab7028
Rabbit monoclonal anti-Brg1	abcam	Cat#ab110641
Mouse monoclonal anti-Gapdh	Fitzgerald	Cat#10R-G109A
Chemicals, Peptides, and Recombinant Proteins		
Micrococcal nuclease	Roche	Cat#10107921001
Formaldehyde	Fisher Scientific	Cat#F79-500
DSG	ThermoFisher Scientific	Cat#201593
Critical Commercial Assays		
TruSeq ChIP Sample Prep Kit	Illumina	Cat#IP-202-1012
TruSeq stranded mRNA sample preparation kit	Illumina	Cat#RS-122-2101
Nextera DNA library preparation kit	Illumina	Cat#FC-121-1030
RNeasy Mini kit	QIAGEN	Cat#74104
QIAGEN MinElute reaction clean up kit	QIAGEN	Cat#28204
Deposited Data		
ChIP-seq data, RNA-seq, ATAC-seq data:	This study	GEO: GSE90895
Experimental Models: Cell Lines		
Mouse embryonic fibroblasts isolated from 129SV/Jae mice	Laboratory of K. Plath (Sridharan et al., 2009)	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mouse embryonic fibroblasts isolated from 129SV/Jae/C57BL6J mice carrying Col1A:tetO-OSKM/wt Rosa26:M2rtTA/wt	Laboratory of K. Plath (Sridharan et al., 2013)	N/A
Mouse embryonic fibroblasts isolated from 129SV/Jae/C57BL6J mice carrying Col1A:tetO-Oct4/wt Rosa26:M2rtTA/wt	This paper	N/A
Mouse embryonic fibroblasts isolated from 129SV/Jae/C57BL6J mice carrying Col1A:tetO-Sox2/wt Rosa26:M2rtTA/wt	This paper	N/A
Mouse embryonic fibroblasts isolated from 129SV/Jae/C57BL6J mice carrying Col1A:tetO-Klf4/wt Rosa26:M2rtTA/wt	This paper	N/A
Pre-iPSC line 1 (12.1)	Laboratory of K. Plath (Sridharan et al., 2013)	N/A
Pre-iPSC line 2 (1A2)	Laboratory of K. Plath (Sridharan et al., 2009)	N/A
Mouse embryonic stem cell line V6.5	Laboratory of R. Jaenisch	N/A
PlatE cell line; 293T based	Laboratory of T. Kitamura (Morita et al., 2000)	N/A
293T cells	ATCC	Cat#CRL3216
Experimental Models: Organisms/Strains		
Mouse: 129SV/Jae/C57BL6J, Col1A: $OSKM^{tetO}$/wt R26: M2rtTA/wt	Laboratory of K.Plath (Sridharan et al., 2013)	N/A
Mouse: 129SV/Jae/C57BL6J, Col1A: O^{tetO}/wt R26: M2rtTA/wt	This paper	N/A
Mouse: 129SV/Jae/C57BL6J, Col1A: S^{tetO}/wt R26: M2rtTA/wt	This paper	N/A
Mouse: 129SV/Jae/C57BL6J, Col1A: K^{tetO}/wt R26: M2rtTA/wt	This paper	N/A
Recombinant DNA		
FUW-tetO Esrrb	(Buganim et al., 2012)	Addgene:#40798
pMX-RUNX1	This paper	N/A
pMX-Fra1	This paper	N/A
pMX-Flag-Fra1	This paper	N/A
pMX-cJun	This paper	N/A
pMX-Flag-cJun	This paper	N/A
Sequence-Based Reagents		
siRNA for Runx1-A	Dharmacon	Cat#D-048982-01
siRNA for Runx1-B	Dharmacon	Cat#D-048982-03
siRNA for Cebpa-A	Dharmacon	Cat#D-040561-03
siRNA for Cebpa-B	Dharmacon	Cat#D-040561-04
siRNA for Cebpb-A	Dharmacon	Cat#D-043110-06
siRNA for Cebpb-B	Dharmacon	Cat#D-043110-22
siRNA for Luciferase	Dharmacon	Cat#D-001210-02
Software and Algorithms		
ChromHMM v1.1.0	(Ernst and Kellis, 2012)	http://compbio.mit.edu/ChromHMM/
ChromImpute v1.0.0	(Ernst and Kellis, 2015)	http://www.biolchem.ucla.edu/labs/ernst/ChromImpute/
DESeq2	(Love et al., 2014)	https://bioc.ism.ac.jp/packages/3.1/bioc/html/DESeq2.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Metascape	(Tripathi et al., 2015)	http://metascape.org/gp/index.html#/main/step1
GREAT	(McLean et al., 2010)	http://bejerano.stanford.edu/great/public/html/
Bowtiev2	(Langmead and Salzberg, 2012)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Tophat	(Trapnell et al., 2009)	https://ccb.jhu.edu/software/tophat/index.shtml
MACS2 2.1.0	(Zhang et al., 2008)	https://github.com/taoliu/MACS

CONTACT FOR REAGENT AND RESOURCE SHARING

Please direct any requests for further information or reagents to the Lead Contact, Professor Kathrin Plath (kplath@mednet.ucla.edu), Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Chancellor's Animal Research Committee at University of California Los Angeles has approved our animal breeding and research protocols. Animals were used for isolating cultures of primary cells from mice. Mouse embryonic fibroblasts (MEFs) harboring the M2rtTA construct in the R26 locus (heterozygously) together with a single dox-inducible polycistronic cassette coding for OSKM in the *Col1A* locus (tetO-OSKM) (Ho et al., 2013; Sridharan et al., 2013) or a dox-inducible cassette encoding a single reprogramming factor (tetO-Oct4, tetO-Sox2, or tetO–Klf4) in the *Col1A* locus, or wild-type MEFs were used for reprogramming experiments, ChIP-seq, ATAC-seq, and RNA-seq assays. In addition, pre-iPSCs derived by retroviral overexpression of OSKM in MEFs and the mouse ESC line V6.5 were used to study different stages of reprogramming. All cell lines are described in the Key Resources Table.

METHOD DETAILS

Cell lines, culture conditions, and reprogramming experiments

The following cell lines were used for the comprehensive genomics analysis of the reprogramming process at discrete stages: primary MEFs harboring a heterozygous *R26-M2rtTA* allele and a single dox-inducible polycistronic cassette coding for OSKM in the *Col1A* locus (tetO-OSKM) (Ho et al., 2013; Sridharan et al., 2013), derived from day 13.5 embryos of timed mouse pregnancies; two independently generated male pre-iPSC lines (line 12-1 (pre-i$^{\#1}$) and 1A2 (pre-i$^{\#2}$) obtained upon retroviral (pMX-based) expression of Oct4, Sox2, Klf4, and cMyc in Nanog-GFP reporter MEFs (Sridharan et al., 2009, 2013); and the male ESC line V6.5 from the Jaenisch laboratory. All cell types were grown in standard mouse ESC media containing KO DMEM, 15% fetal bovine serum (FBS), recombinant leukemia inhibitory factor (Lif), β-mercaptoethanol, 1x penicillin/streptomycin, L-glutamine, and non-essential amino acids. Pre-iPSCs and ESCs were grown on irradiated MEFs (feeders), but feeder-depleted and grown overnight on gelatin for genomics experiment. For the 48h reprogramming time point, tetO-OSKM MEFs were cultured in ESC media containing 2 μg/ml doxycycline for 48 hr to induce the expression of OSKM. For all ChIP-seq, ATAC-seq and RNA-seq experiments, cells were grown in ESC medium.

For single reprogramming factor overexpression ChIP-seq experiments (Figures 2G, 4J, 4K, 6D, 6E, S3, and S5H), MEFs containing a dox-inducible cassette encoding a single reprogramming factor (tetO-Oct4, tetO-Sox2, or tetO–Klf4) in the *Col1A* locus and the tet-transactivator M2rtTA in the R26 locus (heterozygous) were generated as described (Beard et al., 2006) by targeting V6.5 ESCs carrying a FRT site in the Col1A locus, generating chimeric mice upon blastocyst injection, and breeding for germline transmission. These MEFs were induced with 2ug/ml doxycycline for 48h to assess the binding events of individually expressed reprogramming factors. Alternatively, wild-type 129SVJae MEFs were infected with a pMX retrovirus encoding an individual reprogramming factor (either pMX-Oct4, pMX-Sox2, or pMX-Klf4) for single factor overexpression, or with a combination of reprogramming factor baring retroviruses for double or triple reprogramming factor combinations (OK, SK, OS, or OSK). Briefly, the cDNAs of the three factors (Oct4, Sox2 or Klf4) were cloned into the pMX retroviral vectors and individually transfected into PlatE packaging cells (Maherali et al., 2007). Viral supernatants were harvested 48 hr post-infection and used to infect MEFs twice, for 8hrs continuously in the presence of 10 mg/ml polybrene. MEFs were harvested for genomics analyses 48 hr post infection.

The role of somatic TFs in the reprogramming process was tested via overexpression by infecting tetO-OSKM MEFs with pMX-retroviruses encoding the *Fra1*, *cJun*, or *Runx1* cDNA. N-terminally Flag-tagged versions of *cJun* and *Fra1* were also cloned into pMX vectors and tested for reprogramming efficiency. Viral supernatants were produced in platE cells as described above.

Subsequently, tetO-OSKM MEFs were infected twice for a span of 8 hr each time, followed by dox-induction of OSKM expression. For Esrrb overexpression, a lentiviral construct encoding the tet-inducible *Essrb* cDNA, obtained from (Buganim et al., 2014), was transfected alongside viral packaging vectors (pMDLg, pRSV-REV, pCMV-VSVG) into 293T cells using the CalPhos mammalian transfection kit (Clontech 062013) as per manufacturer's instructions. Lentiviral production was performed for 48h, and the harvested supernatant used to infect tetO-OSKM MEFs containing M2rtTA twice. To initiate reprogramming and *Esrrb* expression, the cells were cultured in ESC medium containing 2mg/ml doxycycline. The viral packaging vectors were a generous gift from Dr Zack laboratory in UCLA.

For the *Runx1, Cebpa and Cebpb* siRNA experiments, a set of four different siRNAs was purchased from Dharmacon and initially transfected into MEFs using lipofectamine–RNAi max (Life Technologies) according to manufacturer's instructions to assess knockdown efficiency. Of the four siRNAs, the two producing the most efficient knockdown were used in reprogramming experiments at a final concentration of 20uM. For *Runx1* these were D-048982-01 and D-048982-03, for Cebpa D-040561-03 and D-040561-04, and for Cebpb D-043110-06 and D-043110-22. For control siRNA treatment, we used the non-targeting Luciferase control (D-001210-02). siRNAs against *Runx1* were transfected into tetO-OSKM MEFs two times: first 12 hr before reprogramming was started by doxycycline addition, and second at the time of doxycycline addition, to induce depletion of *Runx1* only early in reprogramming. siRNAs against *Cebpa/b* were first transfected 12 hr before reprogramming was started by doxycycline addition and were re-transfected every three days to maintain knockdown throughout reprogramming.

In all experiments that assessed reprogramming efficiency, reprogramming cultures were shifted to reprogramming media, which is similar to ESC medium but contains 15% KSR instead of FBS, at day 3 of reprogramming. Reprogramming efficiency was scored by counting Nanog-positive colonies after immunostaining cultures with an anti-Nanog antibody (eBioscience 14-5761-80), 11 days post doxycycline induction. For the *Runx1* siRNA experiment, doxycyline was withdrawn from the cultures at day 9, for the last 48 hr, before fixation of the reprogramming cultures at day 11. For OSKM/Esrrb-induced reprogramming cultures reprogramming efficiency was calculated by counting DPPA4-positive colonies after immunostaining with an antibody directed against DPPA4 (R&D AF3730), 8 days post-OSKM/E induction with doxycycline. We have shown previously that DPPA4 is induced after Nanog expression during the final steps of reprogramming (Pasque et al., 2014).

Immunofluorescence

Cells were grown on coverslips pretreated with 0.3% porcine gelatin (Sigma G2500) in ESC medium for 48h. After fixation with 4% paraformaldehyde the cells were washed with 1xPBS-0.05% Tween, permeabilized with 1xPBS-0.5% Triton-X, and blocked with 5% donkey serum in 1xPBS-0.05% Tween. Primary antibody incubation was carried out at 4°C overnight, secondary antibody incubation was carried out at RT for 30min, each in blocking buffer. Between each incubation, cells were washed with 1xPBS-0.05% Tween for three times. Cells were then mounted using a mounting medium with DAPI (Vector Labs H-1200). Antibodies used for Nanog and DPPA4 to detect reprogrammed colonies are listed above. Antibodies for the detection of O,S,K,M or Esrrb were: anti-Oct4 (RnD; AF1759), anti-Sox2 (RnD AF2018), anti-Klf4 (RnD; AF3158), anti-cMyc (RnD; AF3158) and anti-Esrrb (RnD; H6705).

Native ChIP-seq (N-ChIP)

Native ChIP-seq was performed for as described in (Wagschal et al., 2007) for all histone modification except H3K79me2 and H3K9me3. Briefly, 50 × 10^6 Nuclei were isolated from non-crosslinked cells (MEFs, 48h, pre-$i^{\#1}$ and ESC) by incubation in 2 mL of a hypotonic solution (0.3M sucrose, 60mM KCl, 15mM NaCl, 5mM $MgCl_2$, 15mM Tris-HCl pH 7.5, 0.5mM DTT, 0.1% NP40, and protease inhibitor cocktail) followed by centrifugation through a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM $MgCl_2$, 0.1mM EGTA, 15 mM Tris-HCl pH 7.5, 0.5mM DTT, and protease inhibitor cocktail). Nuclei were then re-suspended in MNAse-digestion buffer (0.32M sucrose, 50mM Tris-HCl pH 7.5, 4mM $MgCl_2$, 1mM $CaCl_2$, and protease inhibitor cocktail) and digested with 3 units of MNase (Roche 10107921001) for 10 min at 37°C. The first soluble fraction (S1) was recovered by centrifugation for 10 min at 10,000 rpm. The pellet containing nuclei was then dialyzed overnight in 1l of dialysis buffer (1mM Tris-HCl pH7.5, 0.2mM EDTA, protease inhibitors) to more completely release the chromatin fraction (S2) from nuclei. 10 ug of soluble chromatin (S1 and S2) were then incubated with 5 ug of antibody targeting histone modifications-conjugated to magnetic beads (Active Motif; 53014) under constant stirring at 4°C for 16 hr. The antibodies used were: anti-H3K9ac (Abcam; ab4441), anti-H3K4me3 (Abcam; ab8580), anti-H3K4me2 (Abcam ab7766), anti-H3K4me1 (Abcam; ab8895), anti-H3K27me3 (Active Motif; 39155), antiH3K27ac (Abcam; ab4729), and anti-H3K36me3 (Abcam; ab9050). Beads were washed twice with wash buffer A (50mM Tris-HCl pH 7.5, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl pH 7.5, 10mM EDTA, 125mM NaCl), and wash buffer C (50mM Tris-HCl pH 7.5, 10mM EDTA, 175 mM NaCl). DNA was extracted using phenol:chloroform:iso-amylacohol and used for downstream library construction. DNA from fractions S1 and S2 was also isolated directly using phenol:chloroform:iso-amylacohol extraction and used as an whole genome input control (native Input). All protocols for Illumina/Solexa sequencing library preparation, sequencing, and quality control were performed as recommended by Illumina, with the minor modification of limiting the PCR amplification step to 10 cycles. All constructed libraries were sequenced using single-end 50 bp reactions.

Cross-linked ChIP-seq (X-ChIP)

Transcription factor and epigenetic regulator occupancy data generated in this study were acquired using ChIP after crosslinking cells (X-Chip). X-ChIP was also employed for mapping H3K79me2 (Active Motif, 39143), H3K9me3 (abcam, ab8898 or Millipore,

05-1242), H3 (abcam, ab1791) and H3.3 (Abnova, H00003021-M01). Briefly, cells were grown to a final concentration of $5x10^7$ cells for each ChIP-seq experiment. To stabilize HATs/HDACs (p300, Hdac1) and Brg1 on chromatin, cells were treated with 2 mM disuccinimidyl glutarate (DSG) for 10 min prior to formaldehyde crosslinking. For all other targets, cells were chemically cross-linked at room temperature by the addition of formaldehyde to 1% final concentration for 10 min and quenched with 0.125 M final concentration glycine. Cross-linked cells were re-suspended in sonication buffer (50mM HEPES-KOH pH 7.5, 140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS) and sonicated using a Diagenode Bioruptor for three 10 min rounds using pulsing settings (30 s ON; 1 min OFF). 10 ug of sonicated chromatin was then incubated overnight at 4°C with 5 ug of antibody conjugated to magnetic beads. The antibodies used were: anti-Esrrb (RnD; H6705), anti-Klf4 (RnD; AF3158), anti-cMyc (RnD; AF3696), anti-Nanog (cosmobio REC-RCAB001P), anti-Oct4 (RnD; AF1759), anti-Sox2 (RnD AF2018), anti-p300 (SantaCruz;sc-585), anti-Runx1 (Novus Biologicals NBP1-61277), anti-Fra1 (SantaCruz;sc-183X), anti-Cebpa (SantaCruz; sc-61X), anti-Cebpb (SantaCruz;sc-150X), anti-Hdac1(abcam; ab7028) and anti-Brg1(abcam; ab110641). Following the IP, beads were washed twice with RIPA buffer (50mM Tris-HCl pH8, 150 mM NaCl, 2mM EDTA, 1% NP-40, 0.1% Na-deocycholate, 0.1% SDS), low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% Na-deoxycholate, 1% NP-40), and 1xTE. Finally, DNA was extracted by reverse crosslinking at 60°C overnight with proteinase K (20ug/ul) and 1% SDS followed by phenol:chloroform:iso-amylacohol purification. Libraries were constructed as indicated above and sequenced using single-end 50 bp reactions.

ATAC-seq library construction and sequencing

ATAC-seq was done as previously described (Buenrostro et al., 2013). Briefly, 50000 cells (129SVJae MEFs, un-induced tetO-OSKM MEFs, $48h^{OSKM}$, pre-$i^{\#1}$, pre-$i^{\#2}$ or ESCs) were re-suspended in 50 μL lysis buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP40, and 1 × Complete Protease inhibitor (Roche)) and spun at 500*g* for 10 min at 4 °C to collect nuclei. Nuclei were washed in 1x PBS and subsequently re-suspended in 50 μl Transposase reaction (25 μl 2 × tagmentation buffer, 22.5 μl water, 2.5 μl Tn5 Transposase, following instructions by Illumina). Reactions were incubated for 30 min at 37 °C and DNA purified using QIAGEN MinElute columns (QIAGEN). The transposed DNA was subsequently amplified with custom primers as described (Buenrostro et al., 2013) for 7-9 cycles and libraries were visualized on a 2% TBE gel prior to sequencing with a single-end-sequencing length of 50 nucleotides.

RNA-seq

RNA from independent biological replicates of each un-induced MEFs, induced MEFs at 48hrs, pre-iPSCs (pre-$i^{\#1}$& pre-$i^{\#2}$) and ESCs, was isolated using the RNeasy Mini kit. RNA was treated on column with 0.5 kunitz units of DNase prior to elution according to manufacturers instructions. RNA from MEF cultures induced for 48h to express OSKM/Esrrb, OSKM/Fra1, OSKM/cJun or a single reprogramming factor (tetO-Oct4, tetO-Sox2, or tetO–Klf4, tetO–Myc) was also isolated. In all cases, messenger RNA was captured using oligodT Dynabeads (Life Technologies). Strand-specific RNA-seq libraries were constructed as described in (Parkhomchuk et al., 2009).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Analysis and Visualization

Reads from ChIP-seq experiments were mapped to the mouse genome (mm9) using Bowtie software (Langmead et al., 2009) and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the exact same location and strand in the genome were collapsed to a single read to account for clonal amplification effects. For ChIP-seq of TFs and ATAC-seq, peaks were called using MACS2 software (Zhang et al., 2008) using a bandwidth parameter of 150bp. Peaks with q-val cut-off < 0.005 and fold > = 4-fold were retained. Identified peak locations can be found in Table S1.

Reads from RNA-seq experiments were mapped to the mouse genome (mm9) using TopHat software (Trapnell et al., 2009) and only those reads that aligned with no more than two sequence mismatches were retained. Replicates were merged and RPKM values of mm9 RefSeq genes were calculated as described (Mortazavi et al., 2008) (Table S2). Prior to log2 transformation of RPKM values, a pseudo-count of 1 was added to all RPKM values (log2(RPKM+1).

Genome signal tracks of features (TFs, histone marks, ATAC-seq and RNA-seq) were calculated by partitioning the genome into non-overlapping bins of fixed size (100b for TFs, ATAC-seq and RNA-seq, and 25bp for the histone marks). RPKM values were calculated for each bin using the number of sequencing reads that overlap with the corresponding bin. For histone marks, each read was extended by 200 bp in the direction of the alignment. Tracks were visualized in the IGV genome browser (Thorvaldsdóttir et al., 2013).

To produce the heatmaps, in Figures 2C, 2G, S2D, S2E, S3A, S3D, S3F, 4A, 4E, 6A, 6D–6F, S6C, and 7C, we aligned the given feature (such as peaks of a TF) at their summit and tiled the flanking up- and downstream regions within ± 2kb in 100bp bins. For each location, we calculated RPKM values over all 100bp bins by using the number of sequencing reads that overlap each bin after extension by 200bp in the direction of the alignment. To control for input, we computed at each bin a $\log_2$ input-normalized RPKM value as $\log_2(RPKM_{FOREGROUND}) - \log_2(RPKM_{Input})$, where $RPKM_{FOREGROUND}$ denotes the RPKM of the corresponding TF or histone

dataset and $RPKM_{Input}$ denotes the RPKM value of the corresponding whole genome 'Input'. For visualization in figures, each 100 bp bin was displayed with JavaTreeview (Eisen et al., 1998). All metaplots were produced by computing the average input-normalized RPKM value for each 100bp bin across all locations in the given set.

The scatterplots in Figures 5G, S6D, and S6E were produced by first computing log_2(RPKM+1) values over 200bp windows centered at each binding site for the TF signal in MEFs and 48h. To control for the input, we computed log_2(RPKM+1) for the input signal in MEFs at each 200bp window and subtracted it from the values in MEFs and 48h to obtain an input-normalized log_2 RPKM value for each cell type: $log_2(RPKM_{TF\ in\ X}+1)$ - $log_2(RPKM_{MEF\_Input}+1)$, where $RPKM_{TF\ in\ X}$ is the RPKM value in MEF or 48h.

Figures S7D and S7I were generated with ngs.plot (Shen et al., 2014).

ChIP-seq and ATAC-seq data validation

Several external (published) datasets were used to validate our ChIP-seq data (Table S3). Moreover, the majority of ChIP-seq datasets in this study were generated in biological replicates (Table S3), and the correlation of replicate datasets demonstrated a high reproducibility of our data. Furthermore, to ensure that un-induced (starting) tetO-OSKM MEFs were not already representing a 'leaky' expression state for the reprogramming factors (already partially reprogrammed), we also profiled wild-type MEFs not carrying any reprogramming factor transgene for ATAC-seq. These ATAC-seq datasets correlated most closely with those of the un-induced tetO-OSKM MEFs.

Correlation of our datasets with imputated datasets

We created an imputed version of the H3, H3.3, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, p300, ATAC-seq and INPUT (Native Input) data for MEFs, 48h, pre-$i^{\#1}$, and ESCs, using ChromImpute v1.0.0 (Ernst and Kellis, 2015). In creating the imputed version of a dataset, we used all other datasets but the dataset being imputed. The imputed version of each dataset can be viewed as a pseudo-replicate for each dataset and can be used to assess reproducibility. The data put into ChromImpute were the RPKM normalized signals at 25bp resolution after removing reads that map to blacklisted regions in the mouse genome (https://sites.google.com/site/anshulkundaje/projects/blacklists; ENCODE Project Consortium, 2012) and excluding chrM. The signal files for all histone marks, H3, H3.3, and INPUT were generated by extending reads by 200bp in the direction of the alignment. The signal for p300 and ATAC-seq was generated without extension of the reads. ChromImpute was run with default options except the flag '-b 20 -tieglobal' was added to the GenerateTrainData command, the flag '-b 20' to the Train command, and the flag '-b 20 -tieglobal' to the Apply command. The imputed data were converted to a 1000bp resolution by averaging the signal for each 25bp within it. Signal tracks for the observed data were produced at 1000bp resolution in the same way as the signal at 25bp used as input for ChromImpute. Pairwise Pearson correlations were then computed based on the 1000-bp resolution data (Table S3). For each observed dataset, we also reported the maximum correlation with any of the other three observed datasets (for the other reprogramming stages) for the mark based on the 1000-bp resolution data (Table S3).

Differential gene expression analysis

HTSeq (Anders et al., 2015) was used to determine gene counts from replicate experiments, and DESeq2 (Anders and Huber, 2010) for differential analysis. Our quadruplicate datasets were used to identify differential genes between MEFs and OSKM-induced MEFs at 48h (48h), using an adjusted p value < 0.05. DESeq2 was also used to identify differential genes from the following comparisons 1) OSKM-induced MEFs at 48h against and OSKM/Esrrb-induced MEFs at 48h; 2) OSKM-induced MEFs at 48h against and OSKM/Fra1-induced MEFs at 48h; and 3) OSKM-induced MEFs at 48h against and OSKM/cJun-induced MEFs at 48h. In addition, genes were called MEF- or ESC- specific using the following criteria: 1) DESeq2 differential calls with an adjusted p value < 0.05 between ESCs and MEFs; 2) Fold-change of > = 5x between transcript levels in MEFs and ESCs; 3) low RPKM value of in the non-expressing type (typically < 1 RPKM).

Defining combinatorial OSKM binding groups per reprogramming stage

We generated sets of sites co-bound by the reprogramming factors at a given reprogramming stage by extending TF summits produced by MACS2 by 100 bp in each direction and intersecting the extended summits between Oct4, Sox2, Klf4, and cMyc per reprogramming stage (Figure 2F). In this case, we first defined sites bound by all four TFs by intersecting the extended summits of all four factors in any possible order and merging overlapping intersections. Analogously, we defined triply bound sites, and, subsequently, removed those regions that overlapped with the quadruply bound sites from them. Next, we defined doubly bound sites by intersecting the extended summits of every pair of TFs and removing regions that overlapped triply and quadruply bound sites. Finally, we defined solo bound sites as all sites that were not doubly, triply, or quadruply bound. To calculate the enrichment scores of the co-bound groups in Figure 2Fii, we used the middle point between the start and the end coordinates of quadruply, triply, and doubly bound sites. For solo sites, the coordinates of the original summits were used.

Determination of temporal OSKM binding groups

Seven co-binding groups (Figure 2D: '100', '010', '001', '110', '011', '101', '111') were generated in a similar manner as the combinatorial OSKM binding groups described above, by intersecting the extended TF summits (100bp) of a given TF among the three reprogramming stages: 48h, pre-$i^{\#1}$, and ESCs.

Transcription factor clusters

K-means clustering was employed to identify coherent groups of TF binding in Figures S2J, 5C, and 5I. To define these TF clusters, the genome was tiled into 500bp windows and the presence of TF peaks in each bin was determined. This procedure resulted in a vector of binary data for each TF reflecting its absence or presence within 500bp windows across the genome. The windows represented by these vectors were then clustered using R's k-means function applying the Hartigan-Wong method to obtain groups of windows exhibiting common combinatorial binding patterns across the genome. The number of clusters was chosen to reduce the number of potential combinatorial TF groups, while ensuring that each cluster was represented by a significant number of windows.

Ontology Annotation

To associate transcription factor peaks with the closest gene for Ontology analysis (Figures S2G, S4C, and S4E; Table S4) we used the GREAT tool (McLean et al., 2010) with default parameters. Differentially regulated genes defined by DESeq2 (Figures 5J, S6F, S6G, and S7L) were assigned to relevant GO ontology groups using the Metascape software (Tripathi et al., 2015).

ChromHMM modeling parameters

To derive chromatin state segmentations for each reprogramming stage (Figure 1C), we used ChromHMM (version V1.1.0) (Ernst and Kellis, 2012) with default parameters. First, we binarized the mapped reads for all chromatin marks and the native 'Input' indicated in Figure 1C with the ChromHMM's BinarizeBed procedure, using a p value cutoff of 1e-4. To reduce effects of artifacts, we removed redundancy in the input data by keeping only one sequencing read in cases where multiple reads mapped to the same genomic position and strand orientation. We examined models with different numbers of states ranging from 2 to 30 and chose a model with 18 chromatin states that is both interpretable and able to capture the combinatorial complexity of chromatin marks in each reprogramming stage.

In addition to the 18 chromatin state model shown in Figure 1C, which captures chromatin states per reprogramming stage, we used ChromHMM in the "stacked" mode to capture the chromatin changes between MEFs, 48h, pre-iPSCs (pre-i$^{\#1}$), and the pluripotent state, which yielded the 35 chromatin trajectories defined in Figure 3A. In particular, we constructed a single virtual cell type that has all datasets from MEFs, 48h, pre-i$^{\#1}$, and ESCs as individual marks by setting the label of each original dataset in the input file for ChromHMM to contain both the source cell type and the histone mark name. Then, we used ChromHMM to discover and annotate the genome for chromatin states in the virtual cell type. The rest of the preprocessing and ChromHMM parameters were the same as for the 18 state model described above. We considered models with different numbers of states ranging from 25 to 100 and chose 35 states, because it was the model with the minimum number of states that captured unique biological events. We termed the resulting 35 chromatin states (Figure 3A) "chromatin trajectories" to distinguish them from the chromatin states specific to each reprogramming stage (Figure 1C).

TF enrichment in the vicinity of differentially expressed genes early in reprogramming (Figures S6H and S6I)

OSKM binding combinations and groups of MEF-only, 48h-only, and shared somatic TF binding events were intersected with the genomic intervals encompassing TSS+/− 20kb regions of differentially expressed genes at 48h. To compute the fraction of bound upregulated genes for each type of TF set, we counted the number of upregulated genes between MEFs and 48h that have at least one such binding event within 20kb of their TSS and divided this number by the total number of upregulated genes between MEFs and 48h. Analogously, we computed the fraction of bound downregulated genes between MEF and 48h for each TF combination. We then divided the two fractions and plotted the ratio on log2 scale. The statistical significance of each log2 ratio was assessed by a chi-square test that compares the number of TF bound genes between the two groups given the total number of genes in each group.

Positional expression plots (Figures S1H and S4B)

For each chromatin state, we calculated average gene expression levels in MEFs, 48h, pre-i$^{\#1}$, and ESCs, conditioned on the state's distance from annotated transcription start sites. We restricted this analysis to 50kb up- or downstream of transcriptional start sites (TSS). We partitioned this region into non-overlapping bins of 200 bp. For each bin, we computed the average log2 (RPKM+1) value of genes that have a particular chromatin state at this distance relative to their TSS.

Calculations of fold-enrichment

Using the ChromHMM OverlapEnrichment function (Ernst and Kellis, 2012), we calculated enrichment scores for genomic features (TF binding events, conserved elements, repeats, exon, gene-bodies, TSS, TES, ESC super enhancers, etc) in the chromatin state of each corresponding reprogramming stage (18-state chromatin model) and for the 35 chromatin trajectories capturing the chromatin changes during reprogramming, respectively. The enrichment scores were calculated as the ratio between the observed and the expected overlap for each feature and chromatin state based on their sizes and the size of the mouse genome:

$$\frac{F \cap S}{F * S/G}$$

where F is the number of base pairs annotated for the feature F, S is the size of chromatin state S and G is the total length of the mouse genome.

To calculate log2 differential enrichments in Figure 2Fii, we used the following formula:

$$\log_2 \frac{\text{Enrichment in cell type A}}{\text{Enrichment in cell type B}}$$

where each enrichment is calculated based on a binomial background model that treats the corresponding TFs as independent in each cell type (48h and ESCs).

Coordinates for TSS, TES, CpG islands, Exon and Gene Body features used were part of the mm9 annotation included in the ChromHMM software (Ernst and Kellis, 2012). For the calculation of enrichments of conserved genomic regions in Figures 1D and S1G, we downloaded the 30-way Euarch phastCons elements from the UCSC genome browser for the mm9 genome that represent 30 vertebrate species (euarchontoglires) including human and mouse (Siepel et al., 2005).

In Figures 2E and S2I, we applied complete linkage hierarchical clustering with optimal leaf ordering to cluster the enrichments of all pairs of TFs (Bar-Joseph et al., 2001). The pairwise enrichments at base-pair resolution were calculated as the observed overlap divided by the expected overlap based on the binomial background model that treats both transcription factors as independent:

$$\text{Enrichment}(TF_A, TF_B) = \min\left(\frac{100 + TF_A \cap TF_B}{100 + TF_A * TF_B/G}, 500\right)$$

where the numerator is the size of the overlap between peaks of TF_A and TF_B and the denominator is the product between the total number of bp occupied by peaks of TF_A and TF_B divided by the size of the genome (G). A pseudo-count of 100 was added to both the numerator and the denominator to avoid instabilities due to division of small numbers and the maximum enrichment was set to 500.

In Figure 3Bii, the fold-enrichment of the 35 chromatin trajectories in the vicinity of MEF- and ESC-specific genes was calculated with the following formula:

$$\frac{\%\ \text{Cell type } A \text{ specific genes with trajectory } i \text{ within TSS} \pm 20kb}{\%\ \text{Cell type } A \text{ active genes with trajectory } i \text{ within TSS} \pm 20kb}$$

where the numerator is the percentage of genes (MEF- or ESC- specific; as described above and Table S2) carrying each trajectory i within 20kb from their TSS. As control, we divided by the percentage of all active genes in the same cell type (> 1 RPKM) carrying that trajectory.

Assigning peaks to TSS (+/−2Kb) regions

We computed the proportion of transcription factor binding summits of Oct4, Sox2, Klf4, and cMyc that are located within 2 kb of annotated transcription start sites (mm9 RefSeq TSS) and the rest (distal). p values in Figures 2A and S2A were calculated based on an exact two-sided Binomial test of the null hypothesis that the probability of TSS in one of the samples is given by the frequency in the other sample.

Motif analyses

We calculated motif densities at 10 bp resolution within 500 bp around ChIP-seq summits by using the annotatePeaks procedure from HOMER (Heinz et al., 2010) with the following command line arguments: annotatePeaks.pl mm9 -size -500,500 -hist 10

We used the positional weight matrices for the corresponding transcription factor binding motifs provided by HOMER with their default thresholds.

When we scanned regions that were bound by multiple transcription factors, we centered each region at the summit of the corresponding TF. For example, regions co-bound by OSKM were centered at the corresponding Oct4 summits when we scanned them for the Oct4 motif, then centered at the Sox2 summits for the Sox2 motif, then centered at the Klf4 summits for Klf4 motif, and, finally, centered at the cMyc summits for the cMyc motif.

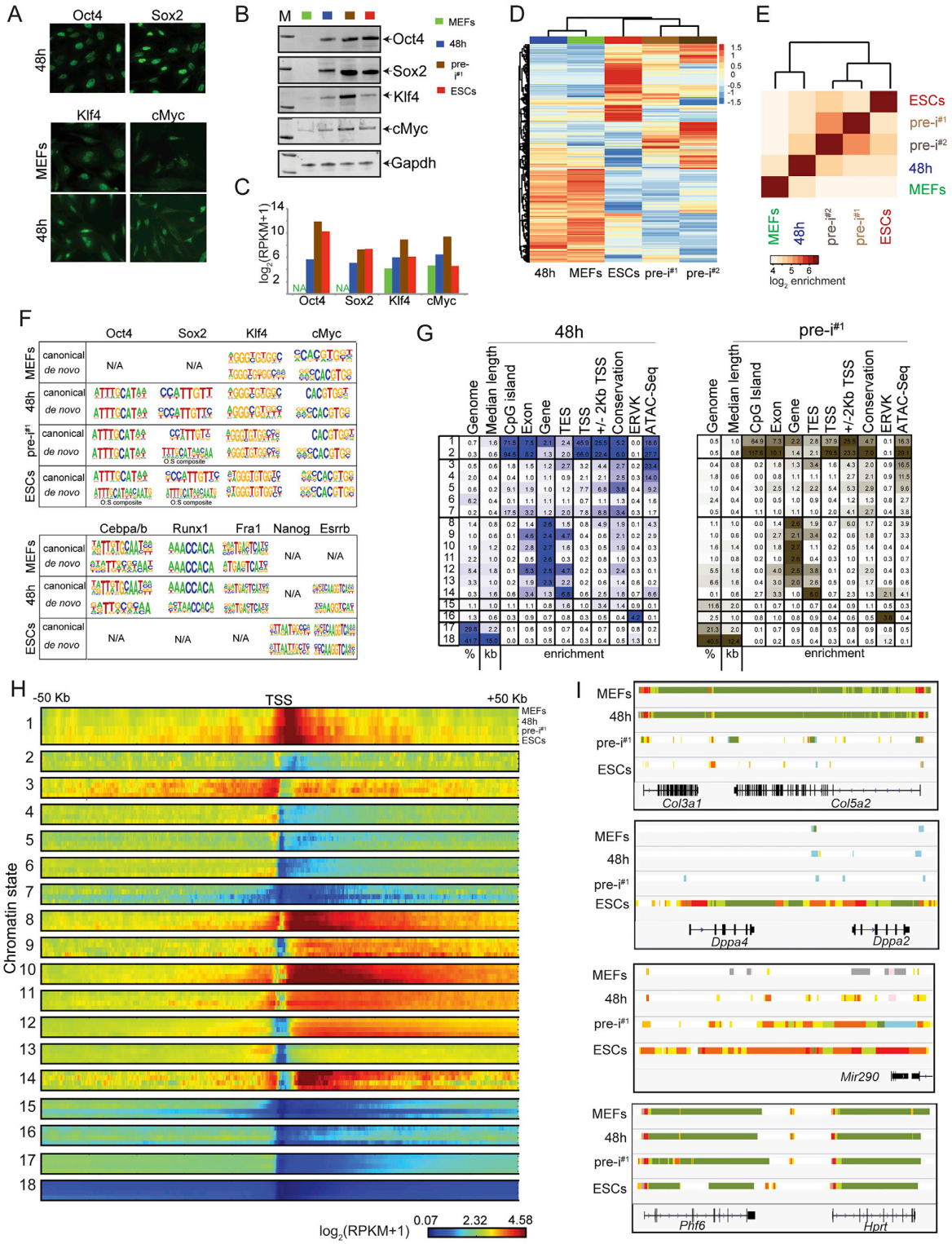
We subsequently smoothed the motif densities by applying a box kernel of length 5 bins centered at each bin. To calculate confidence intervals at the summit bin, we generated 1000 bootstrap samples within each group and calculated the 95% percentile bootstrap confidence intervals (Efron and Tibshirani, 1991).

For de novo motif discovery we used the findMotifsGenome.pl procedure from Homer using the following arguments. findMotifsGenome.pl mm9 -size 200 -mask -cache 1000

DATA AND SOFTWARE AVAILABILITY

The accession number for the genomics data reported in this paper is GEO: GSE90895.

Peak locations derived from ChIP-seq and ATAC-seq experiments are given in Table S1, and normalized expression measurements based on RNA-seq are given in Table S2.



(legend on next page)

Figure S1. Validation of Genomics Data and Characterization of Stage-Specific Chromatin States, Related to Figure 1

(A) Immunostaining for Oct4, Sox2, Klf4, and cMyc (green) in MEFs and at 48h of dox addition to MEFs carrying the polycistronic OSKM cassette, demonstrating endogenous expression of cMyc and Klf4 in MEFs and homogeneous induction of each of the four reprogramming factors across all cells upon dox treatment for 48h.

(B) western blot for Oct4, Sox2, Klf4 and cMyc in MEFs, at 48h, pre-i$^{\#1}$, and ESCs. Whole cell extracts of equal cell numbers were used and Gapdh protein levels served as a loading control.

(C) Transcript levels of the reprogramming factors in the four reprogramming stages (MEFs, 48h of dox-induction, pre-iPSCs and ESCs) based on RNA-seq data. Transcripts of *Oct4* and *Sox2*, unlike those of *cMyc* and *Klf4*, are not present in MEFs prior to induction of transgenic expression.

(D) Unsupervised hierarchical clustering of the top 10000 genes with most variant gene expression across MEFs, 48h, pre-i$^{\#1}$, pre-i$^{\#2}$, and ESCs. Scale is in $\log_2$RPKM. This heatmap demonstrates that the independently generated pre-iPSC lines pre-i$^{\#1}$ and pre-i$^{\#2}$ clustered together and that both lines are more similar to ESCs than to the early reprogramming states.

(E) Hierarchical clustering with optimal leaf ordering of the pairwise enrichment of ATAC-seq peaks in MEFs, 48h, pre-i$^{\#1}$, pre-i$^{\#2}$, and ESCs, at base pair resolution. The pre-iPSCs lines were more similar to each other followed by ESCs, while MEFs and 48h formed a separate node.

(F) Motif analysis of binding sites of OSKM, somatic TFs, and pluripotency TFs in MEFs, at 48h, in pre-i$^{\#1}$, and ESCs, as indicated. At 48h and in pre-iPSCs Oct4, Sox2, Klf4 and cMyc were ectopically expressed. Esrrb was ectopically expressed at 48h in OSKM-induced MEFs. N/A indicates that ChIP-seq data were not generated for the given TF at the indicated reprogramming stage. The Homer tool was used to scan for motif presence under the peaks of the corresponding TF. We scanned these peaks for all known motifs present in the Homer database and reported the top-scoring motif (canonical motif), which in all cases identified the respective known canonical motif. The same motifs were identified as the top represented by de novo motif analysis, with the exception of Oct4 and Sox2 in ESCs and Sox2 in pre-iPSCs, where the composite Oct4:Sox2 motif was most over-represented. For Cebpa and Cebpb similar motifs were identified.

(G) Genomic enrichments of chromatin states defined in Figure 1C at 48h of reprogramming and in pre-i$^{\#1}$. Columns represent percentage (%) of genome occupancy, median length of each state in kilo bases (kb), and fold-enrichments for CpG islands, exons, gene bodies, transcription end sites (TES), transcription start sites (TSS), promoters (defined as TSS ± 2kb), conserved elements (phastCons), ATAC-seq peaks, and endogenous retrovirus K elements (ERVK), colored within each column from highest (darkest) to lowest (white).

(H) Relationship between chromatin states and expression level of nearby genes. The average expression level of genes was plotted as a function of the position of the chromatin state relative to RefSeq-TSS up to 50 kb in both directions. Each larger row corresponds to a chromatin state (1-18) defined in Figure 1C. Within each larger row, smaller rows corresponding to each of our four reprogramming stages (MEFs, 48h, pre-i$^{\#1}$, and ESCs). Each small row shows for the presence of the given chromatin state at each position relative to the TSS, the average expression level of those corresponding genes at the given reprogramming stage. Red indicates higher expression, yellow intermediate expression, and blue low or no expression based on $\log_2$(RPKM+1) values from RNA-seq data. For instance, one can observe that the active promoter state (state 1) is present at the TSS of highly expressed genes, whereas the presence of the inactive/poised promoter state (state 2) around the TSS corresponds to a low or no expression. Also the strong enhancer state (state 3) is proximal to genes with higher expression than the weaker enhancer states (states 4-7).

(I) Validation of reprogramming stage-specific chromatin state annotations defined in Figure 1C by visualization of expected chromatin changes in reprogramming. A comparison of the chromatin states for each of the four reprogramming stages for genes known to be repressed during reprogramming (*Col3a1* and *Col5a2*), induced (*Dppa4*/*Dppa2* and *mir290* clusters), and constitutively expressed (*Hprt* and *Phf6*). The color code of chromatin states is given in Figure 1C. Notably, the *Dppa2*/*Dppa4* cluster is embedded in low signal chromatin states until the pluripotent state. Conversely, the genomic regions upstream the ESC-specific miR290 cluster gains enhancer marks (orange/yellow) as early as 48h post OSKM induction and forms a large enhancer domain in pre-iPSCs and ESCs.

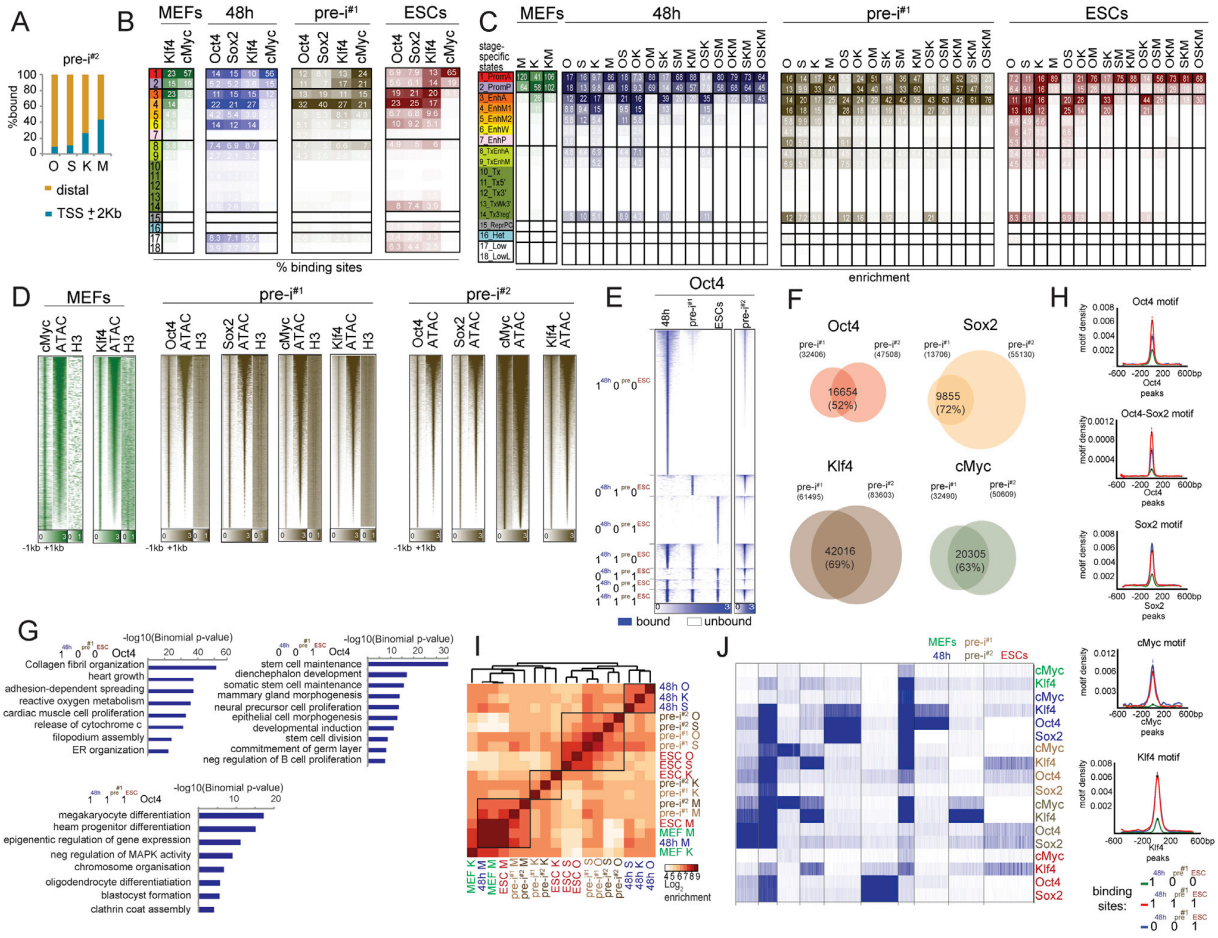


Figure S2. Additional Characterization of OSKM Binding Sites at Each Reprogramming Stage and OSKM Redistribution during Reprogramming, Related to Figure 2

(A) Percentage of O, S, K, and M binding events in promoter-proximal (TSS ± 2Kb) and distal genomic locations for pre-i$^{\#2}$. This figure accompanies Figure 2A.

(B) Percentage of O, S, K, and M binding events in each of the 18 chromatin states from Figure 1C, per reprogramming stage. Specifically, peaks of O, S, K, and M, respectively, in MEFs were analyzed with respect to the chromatin state in MEFs, 48h peaks to the chromatin state at 48h, pre-i$^{\#1}$ peaks against the chromatin state in these cells, and ESC targets to ESC chromatin state. This figure accompanies Figure 2B that shows the fold-enrichment for the same data.

(C) Fold-enrichment of OSKM co-binding groups defined in Figure 2Fi per chromatin state as defined in Figure 1C, for each reprogramming stage. Specifically, co-binding events of O, S, M, and K, respectively, at 48h were analyzed with respect to the chromatin state at 48h, those in pre-i$^{\#1}$ to the chromatin state in pre-i$^{\#1}$, etc.

(D) Heatmap of normalized tag densities (log_2RPKM) for O, S, K, and M binding events and the corresponding ATAC-seq and histone H3 signals at the same sites for MEFs and the two pre-iPSC lines pre-i$^{\#1}$ and pre-i$^{\#2}$. For each bound site, the signal is displayed within a 2 kb window centered on the peak summit for the respective reprogramming factor and peaks were ranked based on ATAC-seq signal strength.

(E) Heatmap of normalized tag densities for O binding events (log_2RPKM) for 48h, pre-i$^{\#1}$, and ESCs, for Oct4 binding groups shown in Figure 2D, depicting the actual signal at regions surrounding 2kb in either direction of the peak calls. In addition, the figure displays the normalized tag densities for O binding events for the same genomic locations in the independently derived pre-iPSC line pre-i$^{\#2}$.

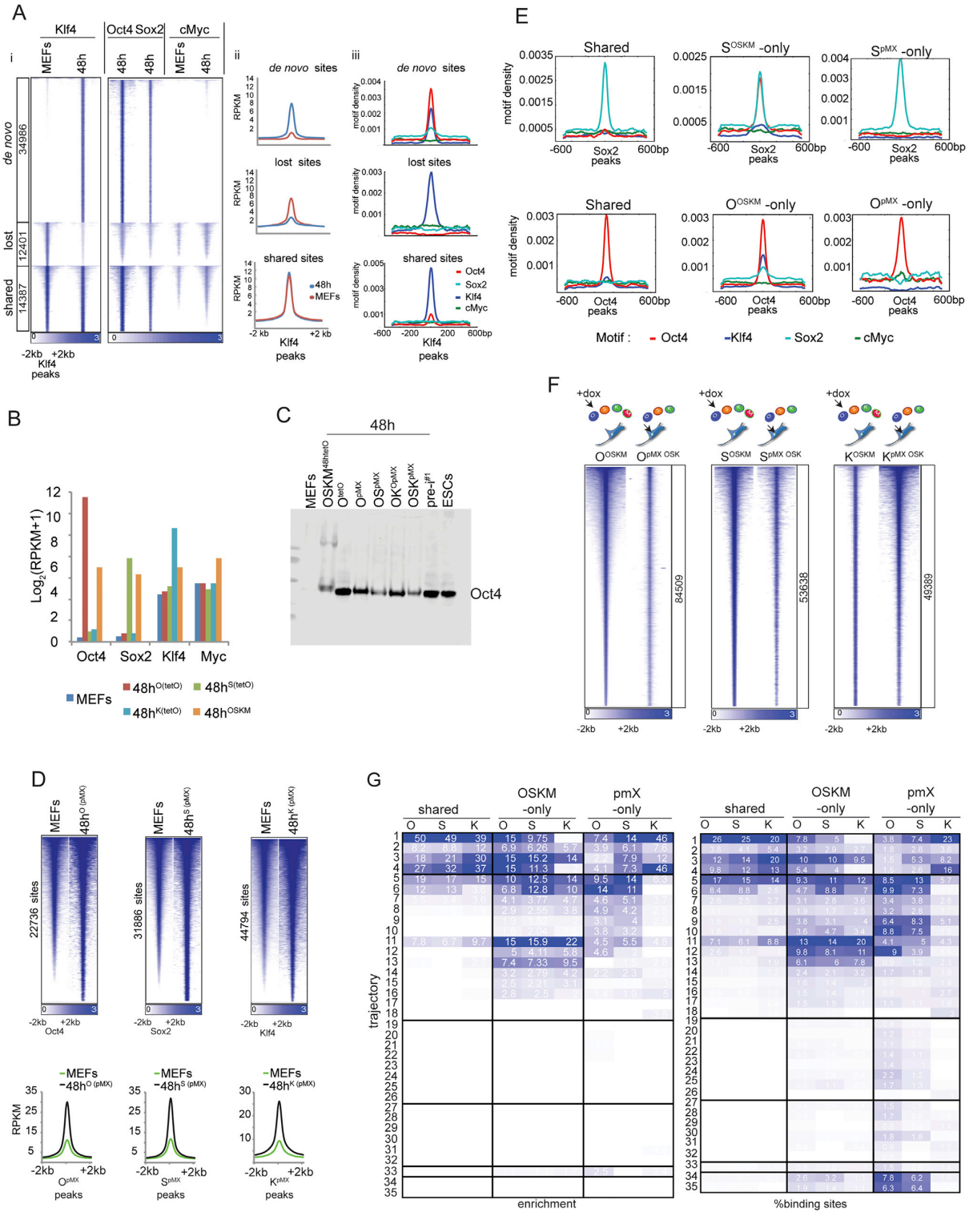
(F) Venn diagram depicting the overlap of O, S, K, and M binding events, respectively, between the pre-i$^{\#1}$ and pre-i$^{\#2}$ lines. The total number of binding events and the number of overlapping sites and their percentage (against the pre-i$^{\#1}$ events) are given.

(G) Ontology of genes associated with '111', '001', and '100' Oct4 sites defined in Figure 2D.

(H) Densities of the Oct4 and Oct4:Sox2 composite motifs at 48h-specific ('100'), constitutive ('111'), and ESC-specific ('001') binding events of Oct4, of the Sox2 motif within Sox2 peaks, the cMyc motif in cMyc peaks, and the Klf4 motif in Klf4 peaks. 95% confidence intervals at peak summits are indicated by the error bars.

(I) Hierarchical clustering with optimal leaf ordering of the pairwise enrichment of O, S, K, and M binding events in the four reprogramming stages and pre-i$^{\#2}$, at base pair resolution. Black boxes emphasize clusters of TFs. O and S bind similar targets in pre-i$^{\#1}$, pre-i$^{\#2}$ and ESC, and Klf4 binding events are more distinct at these stages, clustering away from OS and closer to Myc. At 48h, binding events of O, S, and K cluster together. Myc peaks are more similar to each other than to those of the other reprogramming factors.

(J) K-means clustering of O, S, K, and M peaks across MEFs, 48h, pre-i$^{\#1}$, pre-i$^{\#2}$, and ESCs. Extensive OSK and OK co-binding was observed at 48h, whereas OS co-binding was more prevalent in ESCs. Notably, a subset of sites co-bound by OSK at 48h remained bound throughout reprogramming (second cluster from left). This clustering approach of binding events supports the conclusions made in Figures 2E and 2F.



(legend on next page)

Figure S3. Additional Characterization of Binding Sites of Individually and Co-expressed Reprogramming Factors at 48 hr, Related to Figure 2

(A) Klf4 has relocated to new sites that are co-bound by Oct4 and Sox2 at 48h of reprogramming. (i) A comparison of Klf4 peaks in MEFs (endogenously expressed Klf4) and at 48h of reprogramming revealed sites bound at both stages (shared), sites that were bound in MEFs but not at 48h (lost sites), and sites that were targeted at 48h but not in MEFs (de novo sites). The heatmap shows normalized Klf4 ChIP-seq signal (log_2RPKM) at these sites. Each row shows the ± 2kb region around each Klf4 summit. The number of sites in each category is given. The normalized signal for Oct4, Sox2 and cMyc binding at 48h and in MEFs were added for the same genomic sites. (ii) The metaplots present the average normalized signal of Klf4 in MEFs and at 48h for the three binding groups defined in (i) demonstrating that shared sites have higher Klf4 signal strength than 'lost' and 'de novo' sites. Density plots of the Oct4, Sox2, cMyc, and Klf4 motifs for the three groups of Klf4 binding events defined in (i) are given in (iii). Oct4, Sox2, and Klf4 motifs can be found at de novo Klf4 sites, while only the Klf4 motif is present at lost and shared sites.

(B) Transcript levels (log_2(RPKM+1)) of the reprogramming factors in MEFs, at 48h of reprogramming with OSKM, and at 48h in MEFs overexpressing individual reprogramming factors from a dox-inducible cassette, based on RNA-seq data. Individually expressed reprogramming factors are 50x (Oct4), 2.5x (Sox2) and 8.8x (Klf4) upregulated compared to the corresponding factor at 48h of OSKM-induced reprogramming.

(C) western blot for Oct4 on starting MEFs and MEFs expressing the indicated individual reprogramming factor or combinations thereof either retrovirally (pMX) or inducibly (tetO) for 48h, pre-i$^{\#1}$, and ESCs. Whole cell extracts of equal cell numbers were used.

(D) Heatmap of normalized tag density for ATAC-seq data (log_2RPKM) at sites bound by the indicated reprogramming factor at 48h of individual overexpression in MEFs (O^{pMX}, S^{pMX}, or K^{pMX}). The MEF ATAC-seq signal at the same sites is also shown in each heatmap and the number of sites per reprogramming factor is given. Metaplots of the averaged normalized signal intensities of the ATAC-seq data are presented at the bottom.

(E) Density plots of Oct4, Sox2, Klf4, and cMyc motifs in Sox2 and Oct4 binding groups defined in Figure 2G (shared, OSKM-only, pMX-only). These data show that motif presence discriminates OSKM-only from shared and pMX-only sites.

(F) Heatmaps of normalized log_2RPKM signals for all Oct4, Sox2, and Klf4 binding events, respectively, at 48h of reprogramming with MEFs carrying all four reprogramming factors (OSKM). In addition, the figure displays the normalized tag densities for the binding events of the same reprogramming factor when only OSK were expressed together retrovirally for 48h in MEFs (OSK^{pMX}), without cMyc, for the same genomic locations. The number of peaks per reprogramming factor is given. These heatmaps demonstrate that the sites targeted by O, S, and K early in reprogramming in the context of OSKM co-expression are also largely targeted when only OSK are co-expressed in MEFs (without ectopic cMyc).

(G) Fold-enrichment for O, S, and K binding groups, defined in Figure 2G against the 35 chromatin trajectories described in Figure 3A, colored within each column from high (blue) to low (white) (left table). Percentage of binding events in each of the 35 chromatin trajectories is also given (right table; each column totals 100%) with each column colored from high (blue) to low (white).

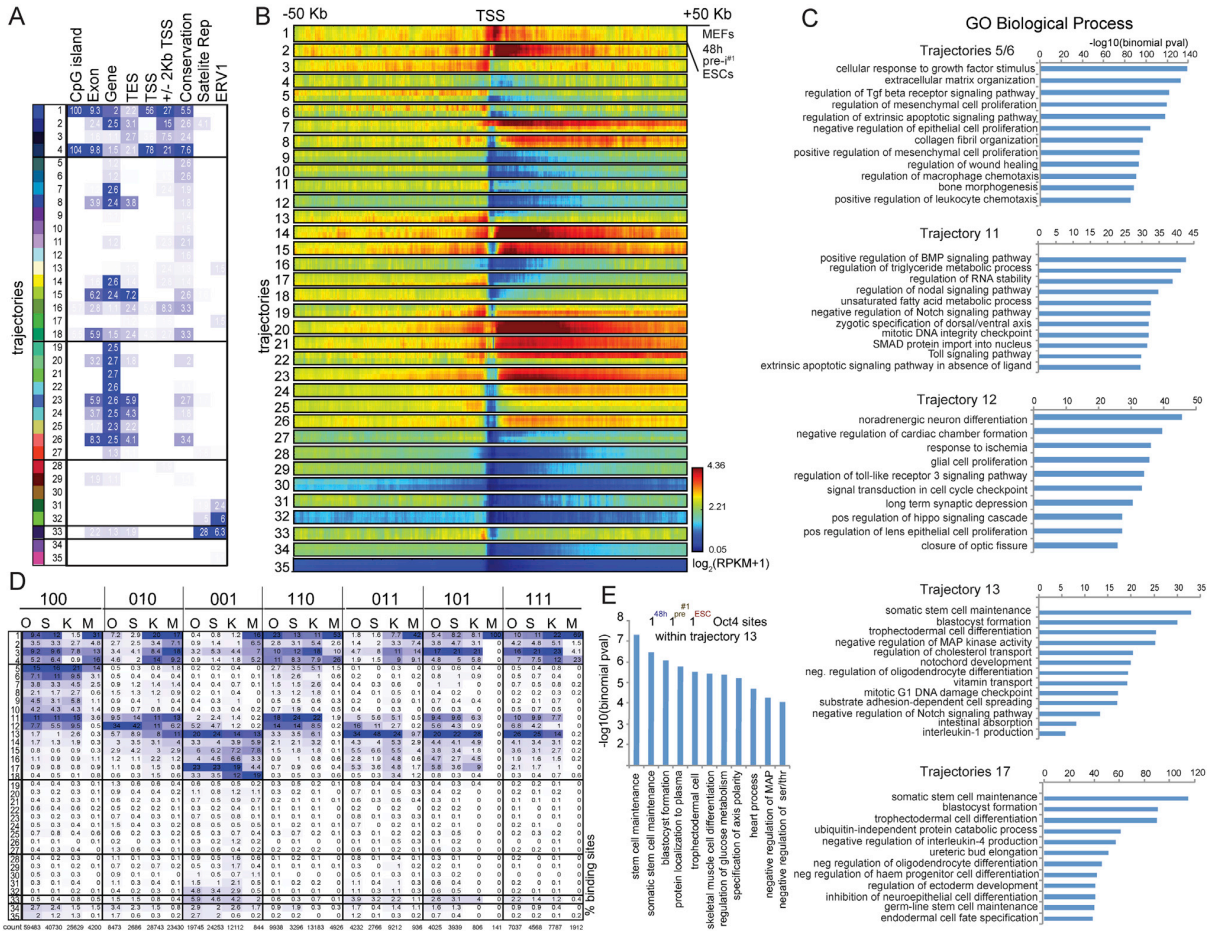


Figure S4. Additional Characterization of the 35 Chromatin Trajectories Describing the Major Chromatin Changes that Occur during Reprogramming, Related to Figure 3

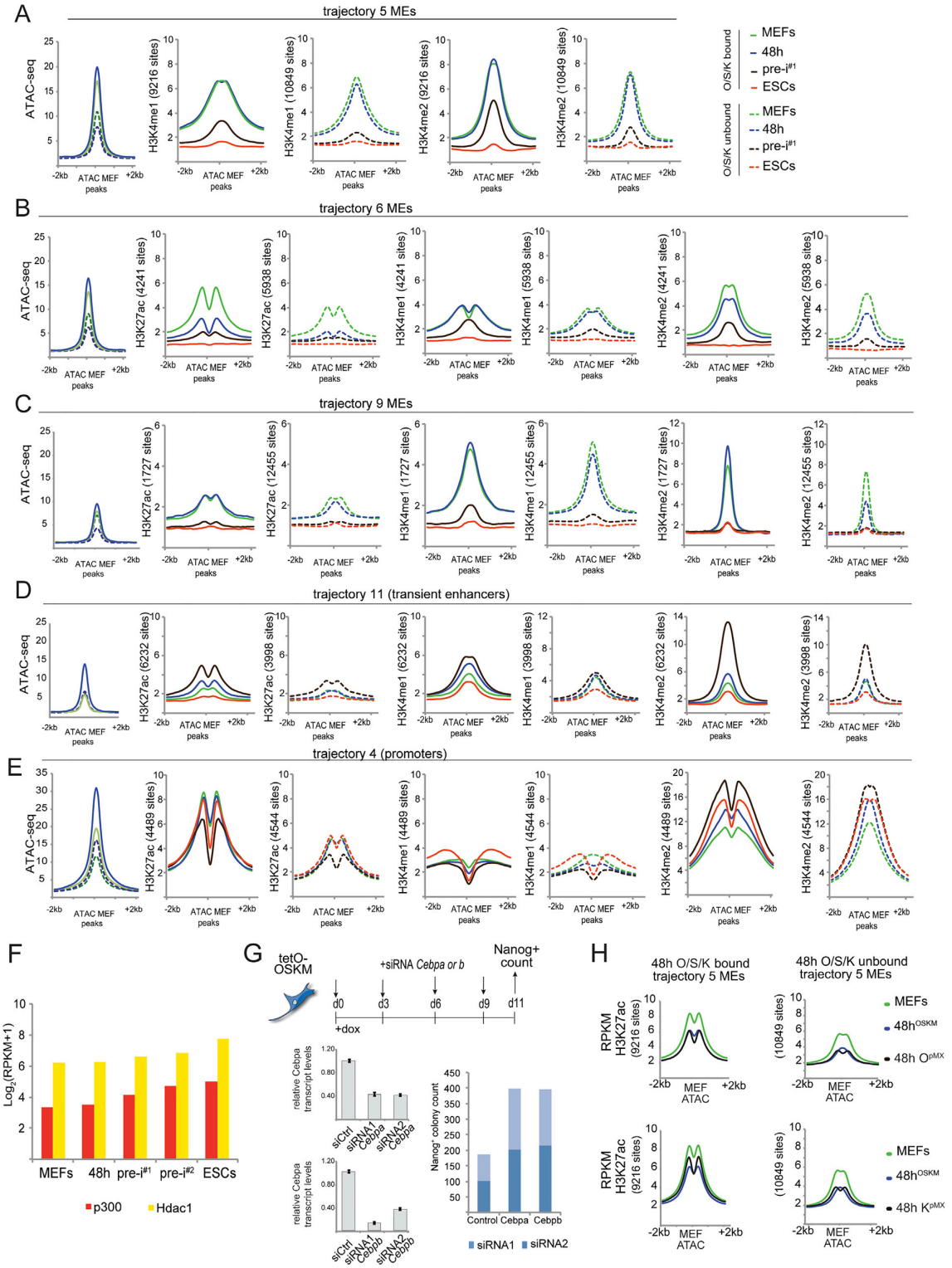
(A) Fold-enrichment of various genomic features for each of the 35 chromatin trajectories defined in Figure 3A. Columns represent fold-enrichment for CpG islands, exons, gene bodies, transcription end sites (TES), transcription start sites (TSS), promoters (defined as TSS ± 2kb), conserved elements (phastCons), satellite repeats as defined by RepeatMasker (RepeatMasker Open-4.0) and endogenous retrovirus 1 elements (ERV1). Enrichment scores were calculated as the ratio between the observed overlap and the expected overlap based on the state size, and colored within each column from high (blue) to low (white).

(B) Relationship between the 35 chromatin trajectories and the expression level of associated genes. The average expression level of genes is plotted as a function of the position of the chromatin state relative to RefSeq-TSS up to 50 kb in both directions. Each larger row corresponds to one of the 35 chromatin trajectories. Within each larger row are smaller rows corresponding to each of our four reprogramming stages (MEFs, 48h, pre-i$^{\#1}$, and ESCs). Each small row shows for the presence of the given chromatin trajectory at each position relative to the TSS, the average expression level of those corresponding genes at the given reprogramming stage. Red indicates higher expression, yellow intermediate expression, and blue low or no expression based on $\log_2$(RPKM+1) values from RNA-seq data. For instance one can observe that the pluripotency enhancer trajectory 13 is associated with a gradual increase in expression of associated genes from 48h to ESCs, while enhancer trajectory 17 is associated more clearly with ESC-specific gene expression. Conversely, the MEF enhancer states (trajectories 5 to 10) display higher expression in MEFs and at 48h than in pre-iPSCs and ESCs.

(C) Gene ontology analysis for enriched biological processes for the indicated chromatin trajectories based on the 35 chromatin states defined in Figure 3A.

(D) Percentage of stage-specific and constitutive O, S, K, and M binding events as defined in Figure 2D ('100', '001', '111' sites etc) for each of the 35 chromatin trajectories defined in Figure 3A. The total number of binding sites observed for each of the seven binding groups of O, S, K, and M, respectively, is given at the bottom of each column. Color scale within each column ranges from the highest (blue) to lowest (white).

(E) Gene ontology analysis for '111' Oct4 sites in trajectory 13.



(legend on next page)

Figure S5. Additional Characterization of Changes Occurring at MEs during Reprogramming, Related to Figure 4

(A) Metaplots of averaged normalized signal intensities of ATAC-seq data and ChIP-seq data for H3K4me1 and H3K4me2 at trajectory 5 MEs bound by O, S, or K (solid lines) and those not bound by any of the three reprogramming factors (dotted lines) in MEFs (green), 48h (blue), pre-i$^{\#1}$ (brown), and ESCs (red). The plots are centered on the summits of ATAC-seq peaks in MEFs.

(B) As in (A), except for trajectory 6 MEs and additional metaplots for H3K27ac.

(C) As in (B), except for trajectory 9 MEs.

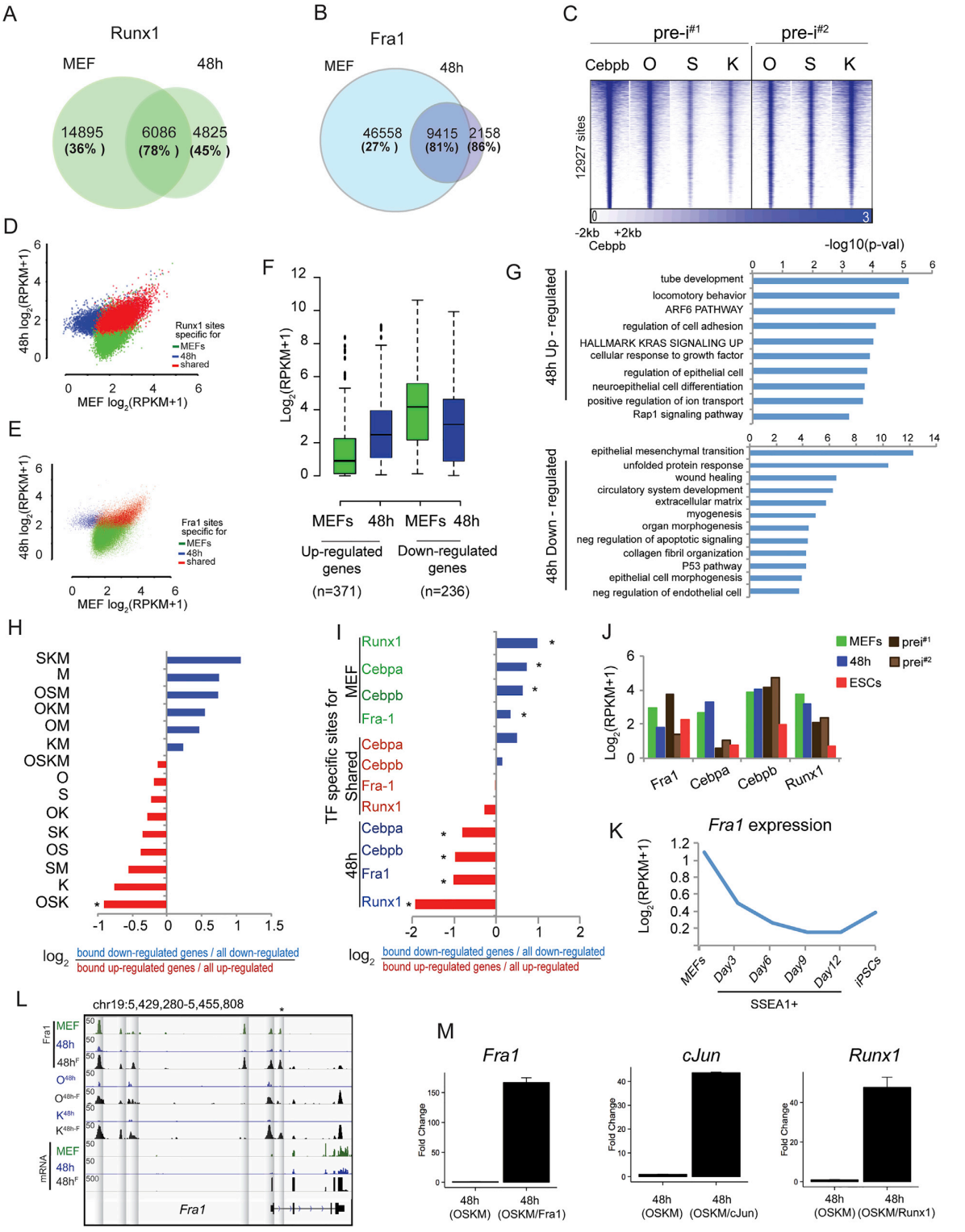
(D) As in (B), except for trajectory 11 elements (transient enhancers).

(E) As in (B), except for trajectory 4 (promoters).

(F) Normalized transcript levels of p300 and Hdac1 for the reprogramming stages indicated, based on RNA-seq data.

(G) Schematic of the reprogramming experiment testing the role of Cebpa/b in reprogramming. OSKM-inducible MEFs were transfected with siRNAs targeting *Cebpa* or *Cebpb* or with siCtrl every 3 days during the course of reprogramming. *Cebpa/b* transcript levels were determined 48h post dox-addition (error bars indicate standard deviation of duplicate qPCR measurements) and Nanog-positive colonies counted at day 11 post OSKM induction for two replicates. Each replicate was generated using different siRNA reagents (siRNA 1 and 2).

(H) Metaplots of averaged normalized tag densities (RPKM) of the enhancer mark H3K27ac at trajectory 5 MEs engaged by O, S, or K (left) and those not engaged by either O, S, or K (right) at 48h post OSKM induction (blue). For the same two sets of trajectory 5 MEs, H3K27ac levels in starting MEFs (green) and in MEFs individually expressing Oct4 (top panels) or Klf4 (bottom panels) for 48h (black) were plotted.



(legend on next page)

Figure S6. Additional Characterization of the Role of Somatic TFs in Reprogramming, Related to Figure 5

(A) Venn diagrams representing the overlap of Runx1 binding sites in MEFs and at 48h of reprogramming. The number of MEF-only, 48h-only and shared sites is given as well as the fractions of each set also bound by O, S, or K (in brackets).

(B) As in (A), for binding sites of Fra1.

(C) Heatmaps of normalized tag densities (log_2RPKM) of the Cebpb ChIP-seq signal at the 12927 Cebpb binding sites obtained in pre-i$^{\#1}$. In addition, the data for O, S, and K occupancy in pre-i$^{\#1}$ and the independent pre-iPSC line pre-i$^{\#2}$ are shown for the same sites, indicating extensive co-binding of O, S, and K with Cebpb in pre-iPSC lines.

(D) Scatterplot of input normalized ChIP-seq signal (log_2(RPKM+1) of Runx1 for MEF-only (green), 48h-only (blue), and shared (red) Runx1 binding events defined in (A).

(E) As in (D), except for Fra1 at sites defined in (B).

(F) Expression changes early in reprogramming. 609 genes (adjusted p value < 0.05) were differentially regulated within the first 48h of OSKM induction based on RNA-seq, with 372 genes induced and 237 genes downregulated (Table S2). Transcription levels of these up- and downregulated genes in MEFs and at 48h of reprogramming are represented as boxplots.

(G) Gene ontology groups associated with up and downregulated genes defined in (F).

(H) Differential enrichment of OSKM co-binding events in genes up- and downregulated early in reprogramming as defined in (F). We computed the log2 ratio between the fraction of bound downregulated genes out of all downregulated genes and the fraction of bound upregulated genes out of all upregulated genes for different combinations of OSKM binding. Bound genes were defined as genes that have at least one binding site of the corresponding combination within 20kb of their TSS. Blue and red coloring represent higher fractions in down- and upregulated genes, respectively. Only the enrichment of the OSK co-binding event was significant (*; p < 0.01 Chi-square test) indicating that sites co-occupied by O, S, and K are enriched in upregulated genes.

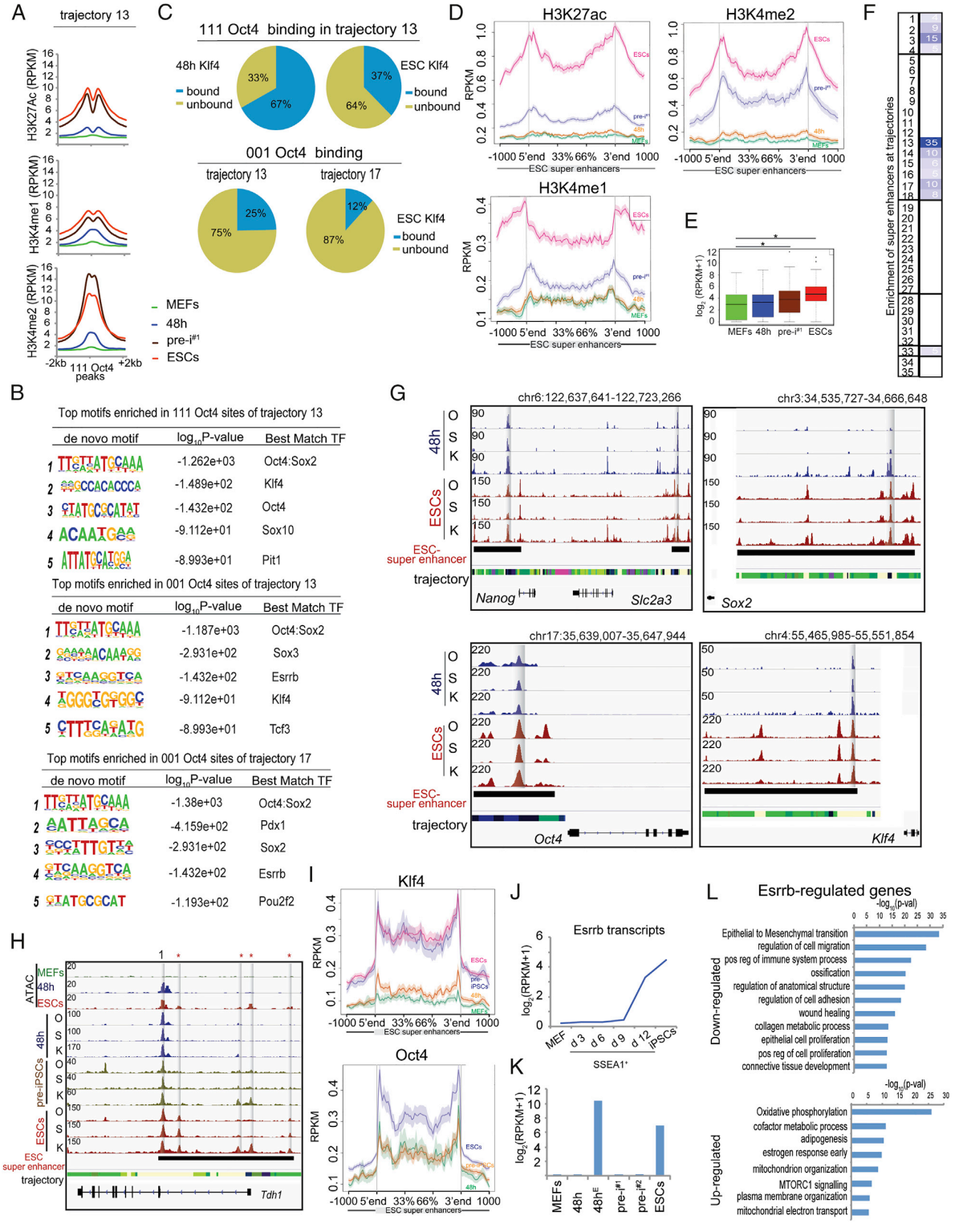
(I) As in (H), but showing the differential enrichment of MEF-only, 48h-only, and shared binding events of Cebpa, Cebpb, Fra1, and Runx1 as defined in Figures 5A, S6A, and S6B in genes up- and downregulated early in reprogramming. These data demonstrate that upregulated genes carry more 48h-only somatic TF binding events compared to downregulated genes whereas higher fractions of downregulated genes are occupied by MEF-only somatic TF binding relative to upregulated genes. * denotes significance (p < 0.01 Chi-square test).

(J) Transcript levels of the somatic TFs Fra1, Cebpa, Cepbb, and Runx1 in MEFs, 48h, pre-i$^{\#1}$, pre-i$^{\#2}$, and ESCs, based on RNA-seq data.

(K) Fra1 transcript level in MEFs, iPSCs, and days 3, 6, 9 and 12 sorted SSEA1$^+$ reprogramming populations, which are considered to be enriched for cells with higher reprogramming potential, as defined in (Polo et al., 2012).

(L) Genome browser view of the *Fra1* locus. RNA-seq reads and Fra1 ChIP-seq data (both in RPKM) in MEFs (green), at 48h of OSKM-induced reprogramming (48h; blue) and at 48h of reprogramming with OSKM in the presence of Fra1 overexpression (black, 48h^F) are shown. O and K binding in the locus for 48h and 48h^F reprogramming samples are also depicted. Shaded areas represent regions within the *Fra1* locus that lose Fra1 binding within the first 48h of reprogramming but have re-gained Fra-1 upon Fra1 overexpression in the context of OSKM/Fra1 (48h^F). The asterisk (*) denotes an intronic enhancer that is known to auto-regulate *Fra1* expression (Verde et al., 2007).

(M) Fold-increase of Fra1, cJun, and Runx1 transcript levels determined by RT-PCR at 48h of reprogramming with OSKM in combination with Fra1, cJun, or Runx1 overexpression, respectively, relative to 48h of reprogramming with OSKM only.



(legend on next page)

Figure S7. Additional Characterization of Reprogramming Factor Binding at PEs and the Esrrb Overexpression Effect, Related to Figures 6 and 7

(A) Metaplots of averaged normalized signal intensities (RPKM) of H3K27Ac, H3K4me1, and H3K4me2 at '111' Oct4 binding events within trajectory 13 PEs for MEFs (green), 48h (blue), pre-$i^{\#1}$ (brown), and ESCs (red).

(B) De novo scanning for motif identification in '001' and '111' Oct4 binding events in PEs of trajectories 13 and 17. The top enriched motifs per indicated set of peaks, the $\log_{10}$(P value) for each motif and the best matching TF are given.

(C) (top) Percentage of '111' Oct4 binding sites in trajectory 13 PEs that are also bound by Klf4 at 48h (left) or in ESCs (right), demonstrating prominent co-binding of Klf4 with Oct4 at these sites particularly early in reprogramming. (bottom) Percentage of '001' Oct4 sites in PEs of trajectories 13 or 17 also bound by Klf4 in ESCs.

(D) Metaplot of normalized signal intensities of H3K27ac, H3K4me1, and H3K4me2 for all ESC super enhancers defined by (Whyte et al., 2013), for each of the four reprogramming stages. 5′ and 3′ denote the start and stop coordinates for the super enhancers, and the shading represents one standard deviation from the mean.

(E) Boxplots of transcript levels for genes neighboring ESC super enhancer in each of our four reprogramming stages. Asterisks (*) mark significant change (p value < 0.007 and < 8.2e-12 for the MEF to pre-$i^{\#1}$ and MEF to ESC comparison, respectively, based on Wilcoxon test).

(F) Fold-enrichment of the 35 chromatin trajectory described in Figure 3A within ESC super enhancers colored within the column from highest (blue) to lowest (white).

(G) Snapshot of 48h and ESC O, S, and K ChIP-seq data (RPKM) at the ESC super enhancer regions associated with the *Nanog*, *Sox2*, *Oct4*, and *Klf4* genes. In addition, the chromatin changes of these region are given by the trajectory annotation (from the 35 chromatin state model) based on the color-code in Figure S4A. Sites bound by O, S, or K already at 48h are highlighted by the gray shading.

(H) Genome browser view of O, S, and K ChIP-seq data and ATAC-seq data at the *Tdh1* ESC super enhancer (RPKM) at the indicated reprogramming stages. In addition, the chromatin changes of this region are given by the trajectory annotation (from the 35 chromatin state model) based on the color code in Figure S4A. Of the five major sites in this super enhancer bound by O, S, or K in ESCs (highlighted by gray bars), one is engaged already at 48h (labeled with 1) and the others are bound only at later reprogramming stages (labeled with asterisks).

(I) Metaplot for normalized Klf4 (top) and Oct4 (bottom) ChIP-seq signal (RPKM) averaged across all ESC super enhancers for our four reprogramming stages. Oct4 data for MEFs were not available since it is not expressed in these cells. 5′ and 3′ denote the start and stop coordinates for ESC super enhancers and the shading indicates one standard deviation from the mean. Based on the comparison of Klf4 binding in MEFs and at 48hrs, we conclude that Klf4 already significantly binds ESC super enhancers at 48h.

(J) Transcript levels of *Esrrb* in MEFs, iPSCs, and days 3, 6, 9 and 12 sorted $SSEA1^{+}$ reprogramming populations, which are thought to enrich for cells with higher reprogramming potential, as defined in (Polo et al., 2012).

(K) Transcript levels of *Esrrb* in our reprogramming stages (MEFs, 48h of OSKM expression (48h), 48h of OSKM and Esrrb co-expression ($48h^{E}$), pre-iPSCs (pre-$i^{\#1}$,pre-$i^{\#2}$), and ESCs, based on RNA-seq.

(L) Gene ontology analysis for enriched biological processes for down- and upregulated genes defined comparing MEFs expressing OSKM/Esrrb for 48h ($48h^{E}$) versus MEFs expressing only OSKM for 48h (48h).

CHAPTER 3

ChromTime: Modeling Spatio-temporal Dynamics of Chromatin Marks

ABSTRACT

Spatial dynamics of chromatin mark peaks have been implicated as important feature of epigenetic regulation. We developed a computational method, ChromTime, that identifies regions for which peaks either expand or contract or hold steady in time course chromatin data. Peaks with predicted expanding and contracting boundaries likely mark regulatory regions associated with transcription factor binding dynamics and gene expression changes. Spatial dynamics of peaks are informative about gene expression changes beyond localized signal changes. In proximity of gene starts, peaks preferentially expand in the same direction as transcription and contract in the opposite direction.

KEYWORDS: epigenomics, time course ChIP-seq, spatial dynamics, histone modifications

BACKGROUND

Genome-wide mapping of histone modifications (HMs) and related chromatin marks using chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) has emerged as a powerful approach to annotate genomes and study cell states[1–3]. Through the efforts of large consortia projects such as the ENCODE[4], Roadmap Epigenomics[5] and BLUEPRINT[6] as well as individual labs[7–9] multiple different chromatin marks have been

mapped across more than a hundred different cell and tissue types. These maps have yielded numerous insights into gene regulation and genetic and epigenetic association with disease[10–14].

While many mapping efforts have largely focused on single or unrelated cell and tissue types[1, 4], a growing number of biological processes have been studied with temporal epigenomic data using time course ChIP-seq assays that map chromatin marks at consecutive stages during particular biological processes. Such datasets have been generated for a wide range of biological settings including T cell development[15], adipogenesis[16], hematopoiesis[17], macrophage differentiation[18], neural differentiation[10], cardiac development[19, 20], somatic cell reprogramming[21–24], embryogenesis[25] and many others[26–34]. The output of these experiments presents a unique opportunity to study the spatio-temporal changes of epigenetic peaks and associated regulatory elements. However, almost all computational methods designed or applied to epigenomic data have been developed based on single or multiple unrelated samples. For example, continuous regions of enrichments of single marks are detected by peak or domain calling methods[35–39]. In cases when multiple HMs are mapped in the same cell type, methods such as ChromHMM[40] and Segway[41] can be used to produce genome-wide chromatin state annotations. In addition, algorithms have been developed for pairwise comparisons of ChIP-seq signal data by differential peak calling[42, 43].

In the context of time course ChIP-seq data, only a few methods have been proposed that consider temporal dependencies between samples. One such method, TreeHMM[44], produces a chromatin state genome annotation similar to ChromHMM and Segway, while taking into account a tree-like structure that captures lineage relationships between the input cell types in order to potentially derive a more consistent annotation across samples. Another method, GATE[27], produces a genome annotation based on clustering fixed length genomic loci that can be modeled with the same switch from one chromatin state to another over time.

One important limitation of methods for pairwise comparison of ChIP-seq data and time course modeling of it is that they do not directly consider or model spatial changes in the genomic territory occupied by chromatin marks over time. Spatial properties of genomic peaks continuously marked by HMs have gained increasing attention as a potentially important characteristic of chromatin marks. For example, long peaks of H3K27ac have been associated with active cell type specific locus control regions termed super-enhancers or stretch enhancers in a number of cell types[45, 46]. Also, the length of H3K4me3 peaks has been associated with transcriptional elongation and consistency of cell identity genes[47]. In the context of cancer, long H3K4me3 peaks have been linked to transcriptional elongation and enhancer activity at tumor suppressor genes and have been observed to be significantly shortened in tumor cells[48]. Long H3K4me3 domains have been implicated to mark loci involved in psychiatric

disorders[49]. Expanded domains of H3K27me3 and H3K9me3 marks have been shown to be characteristic of terminally differentiated cells compared to stem cells[50]. These studies suggest that length of epigenetic peaks is a dynamic feature that can correlate with activity of putative functional elements regulating specific genes. Computational methods that do not explicitly reason about the spatial changes of HMs have significant limitations for studying the dynamics of these properties because they are unable to detect territorial changes that might be associated with redistribution of signal or identify asymmetric directional peak boundary movements.

In this work, we present ChromTime, a novel computational method for detection of expanding, contracting and steady peaks, which can detect patterns of changes in the genomic territory occupied by chromatin mark peaks from time course ChIP-seq data (**Fig 3.1A**). We applied ChromTime to a diverse set of data from different developmental, differentiation and reprogramming time courses. Expansions and contractions in general mark regulatory regions associated with changes in transcription factor binding and gene expression. ChromTime enables studying the directionality of spatial dynamics of chromatin mark peaks relative to other genomic features, which existing computational approaches do not directly address. Our results show that the direction of expansions and contractions correlates with direction of transcription near transcription start sites. ChromTime is a general method that can be used to

analyze time course ChIP-seq from a wide range of biological systems to gain insights into the dynamics of gene regulation.

RESULTS

Model for detecting expanding, contracting and steady peaks from temporal ChIP-seq data

We developed a computational method, ChromTime (https://github.com/ernstlab/ChromTime), designed for systematic detection of expansions, contractions and steady peaks from time course ChIP-seq of single chromatin marks (**Fig 3.1B**). ChromTime takes as input a set of genomic coordinates of aligned sequencing reads from ChIP-seq and, optionally, control experiments over the time course. The method consists of two stages – block finding and dynamics prediction. During the block finding stage, ChromTime determines continuous genomic regions (blocks) that may contain ChIP-seq peaks throughout the time course. To achieve this, ChromTime partitions the genome into fixed length bins and counts the number of ChIP-seq and control reads that map to each bin at each time point. Nearby bins that show significant enrichment are joined into continuous intervals, which subsequently are grouped into blocks if they overlap across time points. As a result, large portions of the genome that are likely to contain background noise at all time points are filtered out, so that peak boundary

dynamics are determined within a subset of the genome potentially enriched for the chromatin mark.

During the dynamics prediction stage, for each block ChromTime determines the most likely positions of the peak boundaries at each time point and whether the peak expands, contracts or holds steady at each boundary between consecutive time points. The method uses a probabilistic mixture model to partition the signal within each block at each time point into background and peak components (**Fig 3.1C**) by reasoning jointly about the data from all time points in the time course. The method assumes that central positions in blocks are more likely to be enriched for ChIP-seq reads and thus the peak component is flanked by the background components (see **Methods, Fig 3.S1D**). The number of sequencing reads in bins from each component at each time point is modelled with different negative binomial distributions that can account for the local abundance of control reads. Furthermore, between any two consecutive time points the boundaries of the peaks are assumed to follow one of three possible dynamics: steady, expand or contract. For steady dynamics, the peak boundaries are enforced to have the same genomic position. For expanding and contracting dynamics the number of genomic bins that the peak boundaries move between the two time points is modelled with different negative binomial distributions which depend on the pair of time points and the corresponding dynamic. Each dynamic is also assumed to have a prior probability which captures its genome-wide frequency at each time point. Within each block, except for peaks at

the first time point, the first appearance of a peak is modelled as expanding the peak boundaries from a zero-length peak at the previous time. Conversely, except for peaks at the last time point, completely removing a peak is modelled as contracting the peak boundaries to a zero-length peak at the next time point. ChromTime is thus flexible to model a wide range of spatial dynamics of peaks over time including peaks that are created later or removed earlier in the time course or those that appear only transiently.

All model parameters are learned jointly from the whole time course. As a result, ChromTime can adapt to different boundary movements, dynamics frequencies and noise levels across experiments and biological systems. The estimated parameters are used to make a prediction for each block for the most likely positions of the peak boundaries and the corresponding boundary dynamics that had generated the signal within the block. The final output contains predicted peak boundaries annotated and colored by their assigned dynamics, which can be used for downstream analysis with existing tools and visualized in genome browsers (**Fig 3.2, 3.S2,** https://github.com/ernstlab/ChromTime).

Reproducibility of ChromTime predictions and association with changes in gene expression, TF binding and DNaseI hypersensitivity sites

To investigate the reproducibility of ChromTime predictions, we applied ChromTime separately to two biological replicate datasets for the H3K4me2 and the H3K(9/14)ac marks in T cell development in mouse[15] and confirmed on average strong enrichment for the same ChromTime annotations co-localizing across replicates (**Fig 3.S3**). We then applied the method to data from pooled replicates for the H3K4me2 and H3K(9,14)ac marks from the mouse T cell development study[15] and to data for the H3K4me3 and H3K27ac marks from a study on stem cell reprogramming in human[21]. To investigate the biological relevance of ChromTime predictions, we examined changes in orthogonal genomic annotations for gene expression, transcription factor binding[4, 15] and DNaseI hypersensitivity sites (DHS) [5] at predicted ChromTime peaks. Peaks with predicted expanding and contracting boundaries that overlap annotated TSSs associated with increase and decrease, respectively, in gene expression (**Fig 3.3, 3.S4**). Predicted peaks with expanding and contracting boundaries enriched for sites bound by important regulators in each biological system as well as sites bound by generic TFs in a cell type specific manner. Furthermore, peaks with predicted steady boundaries throughout each time course associated with TF binding sites and DHS that are shared between the first and the last time point in each time course, which mark potentially stable regulatory elements.

Predicted spatial dynamics by ChromTime associate better with gene expression changes compared to boundary position changes of peaks called from individual time points in isolation

To investigate whether ChromTime's approach for reasoning jointly about the whole time course increases power to detect associations with gene expression compared to considering boundary differences of peaks at consecutive time points called in isolation, we analyzed gene expression changes of genes with TSSs overlapping ChromTime predictions in relation to posterior probabilities for expansions and contractions compared to boundary differences of peaks called with ChromTime from data from individual time points in isolation. We specifically investigated this in the context of H3K4me2 peaks in mouse T cell development[15] and for H3K4me3 peaks in stem cell reprogramming in human[21]. In most cases, ranking boundaries of blocks with at least one called peak during the time course by their predicted ChromTime posterior probabilities for expansions and contractions associated on average with larger gene expression changes compared to ranking block boundaries directly based on the change in the genomic positions of the boundaries of ChromTime peaks called at individual time points in isolation (**Fig 3.4**). These results also held when using peaks from two different peak callers, MACS2[35] and SICER[37] applied on data from individual time points (**Fig 3.S5**).

Spatial dynamics contain information about gene expression changes at consecutive time points not captured by corresponding pairwise HM signal changes

We next investigated whether there is additional information in ChromTime predictions beyond what can be captured by pairwise ChIP-seq signal density changes or by differential peak calls. For this analysis, we focused on H3K4me2 in mouse T cell development[15] and H3K4me3 in human stem cell reprogramming[21]. At each pair of consecutive time points from one before the first to one after the last time points with predicted peaks in a block, we computed the change in ChIP-seq signal density within the left most and right most predicted peak boundaries in the block. We associated the signal density changes with gene expression changes at the nearest TSS within 50kb of each block and computed the average gene expression change as a function of the ChIP-seq signal density change within blocks (**Fig 3.5**). We found that locations with the same ChIP-seq signal density change can associate with significantly different average gene expression changes of proximal genes depending on the predicted ChromTime dynamics. Notably, bidirectional expansions, expansions occurring on both sides of a peak, associated with greater average increase in gene expression than unidirectional expansions, those expansions occurring on one side with steady on the other. These unidirectional expansions in turn associated with greater expression change than steady regions, those regions with a steady call on both sides of a peak. We observed a similar relationship for contractions and decrease of gene expression. These results were replicated also after substituting ChIP-seq signal density changes with differential peak scores as outputted by two different differential peak calling methods (SICER[37] and MACS2[35], **Fig 3.S6**). Therefore, ChromTime predictions can provide additional information about gene

expression changes beyond what is contained in the corresponding ChIP-seq signal density changes as measured directly or by utilizing differential peak calling procedures.

Spatial dynamics are correlated between multiple histone marks

Previous studies have shown that the locations of different HMs can be correlated[1, 51]. In this context, we tested whether multiple HMs can also exhibit jointly the same type of spatio-temporal dynamics. For this purpose, we compared the genomic locations of predicted expansions, contractions and steady peaks for different HMs within the same time course. We focused on two previously published time courses – stem cell reprogramming in human[21] and adipogenesis in mouse[16], where multiple HMs were mapped (**Fig 3.6**). In both datasets, we observed that expansions for H3K4me1, H3K4me2, H3K4me3 and H3K27ac co-localized preferentially and similarly for contractions and steady. In contrast, different spatial dynamics for H3K36me3 and H3K27me3 tended to occupy distinct locations. These results suggest that spatial dynamics of HMs are coordinated at least at a subset of genomic locations.

Direction of expansions and contractions is correlated with direction of transcription

ChromTime can predict unidirectional expansions and contractions, which enables analysis of directionality of spatial dynamics of peaks, an aspect of chromatin regulation that has not been

previously systematically explored. To investigate this, we applied ChromTime on data from 11 previously published studies from a variety of developmental, differentiation and reprogramming processes (**Table 1**) for nine different HMs including narrow and broad marks and for Pol2. We observed that unidirectional expansions and contractions are predicted in most cases on average to be the majority of all expansions and contractions, respectively, at a given pair of consecutive time points (**Fig 3.S7**). One hypothesis for the prevalence of asymmetric boundary movements for the promoter associated chromatin marks is that the direction of boundary movements is associated with the asymmetry of transcription initiation in promoter regions. To test this hypothesis, for each dataset we compared the prevalence of each class of unidirectional dynamics as a function of its distance to the nearest annotated transcription start site (TSS) and the orientation of the corresponding gene (**Fig 3.7**). Consistent with our hypothesis, for H3K4me3, H3K4me2, H3K(9/14)ac, H3K79me2, and for Pol2, we found that unidirectional expansions that expand into the gene body (i.e. in the same direction as transcription) were substantially more frequently found in proximity of TSSs compared to unidirectional expansions in the opposite direction. Moreover, this difference was not observed for expansions distal from TSSs. Similarly, in most cases for these data unidirectional contractions that contract towards the TSS of the nearest gene (i.e. in the opposite direction of transcription) were substantially more frequently found compared to unidirectional contraction in the opposite direction in proximity of TSSs, whereas their frequencies at distal sites showed

much smaller differences. HMs H3K27ac, H3K4me1 and H3K27me3 exhibited the same trend, but to a smaller degree.

DISCUSSION

In this work, we presented ChromTime, a novel computational method for systematic detection of expanding, contracting and steady peaks of chromatin marks from time course ChIP-seq data. ChromTime employs a probabilistic graphical model that directly models changes in the genomic territory occupied by single chromatin marks over time. This approach allowed us to directly encode our modeling assumptions about dependencies between variables in an interpretable and extendable framework.

We applied ChromTime on datasets for broad and narrow HMs and for Pol2 from a variety of developmental, differentiation and reprogramming courses. Our results show that the method can identify sets of expanding and contracting peaks that are biologically relevant to the corresponding systems. In particular, expansions and contractions associate with up- and down-regulation of gene expression and differential transcription factor binding, supporting the biological relevance of ChromTime predictions.

ChromTime gains power by both reasoning jointly about all time points in a time course and by explicitly modeling the peak boundary movements. Supporting this we demonstrated that territorial changes identified by ChromTime had better agreement with gene expression changes compared to considering directly the boundary change of peaks called on data from individual time points in isolation. Additionally, we found that expanding and contracting peaks associated on average with greater change in gene expression compared to peaks with steady boundaries even after controlling for ChIP-seq signal density changes. Some of the power that ChromTime gains from considering spatial information might be explained by its ability to differentiate territorial expansions or contractions which can reflect changes in the number sites of TF binding in close vicinity from changes in ChIP-seq signal within steady peak boundaries. Changes in ChIP-seq signal without territorial expansions or contractions might reflect a change in proportion of cells with the chromatin mark without large changes in activity in any one cell. Additional power can come from the temporal and spatial information that allows the model to effectively smooth over noise in the data enabling more biologically relevant inferences.

ChromTime enables novel analysis of directionality of spatial epigenetic dynamics. In this context, we found that asymmetric unidirectional expansions and contractions for several marks correlate strongly with direction of transcription in promoter proximal regions, which suggests that spatial dynamics at such locations may be related to actions of the transcriptional

machinery. One possible explanation for the observed correlation between the direction of spatial dynamics of at least some HMs and transcription can be provided in part by previous studies that have shown that the Pol2 elongation machinery can recruit H3K4-methyltransferases such as members of the SET[52] and MLL[53] families at the promoters of genes. Our findings are consistent with such models where the Pol2 complex itself may be facilitating the attachment and removal of these marks[54].

The ChromTime software is also relatively efficient particularly when using its option to parallelize all computations during the parameter learning and prediction phases over multiple CPU cores. In our tests, processing ChIP-seq for the H3K4me2 mark and control input data from 5 time points in mouse T cell development[15] took 3 hours by using 4 CPU cores on a MacBook Pro laptop with 2.7GHz Intel Core i7 and 16 GB RAM.

One current limitation of the ChromTime method is that while the runtime of ChromTime still scales linearly with the number of time points, T, the number of observed combinations of dynamics can scale exponentially with T. This exponential growth can complicate downstream analyses that directly consider each combination of dynamics, as there will be 3^{T-1} possible sequences of dynamics at each side of a peak. Extensions of the ChromTime model could model the large number of combinations as being instances of a smaller number of more distinct dynamic patterns.

CONCLUSIONS

The availability of time-series HM data provides an opportunity to understand chromatin dynamics in many biological systems. To better leverage the information in these experiments we presented ChromTime, a method to detect expansions, contractions and steady peaks of single chromatin marks between time points from time course ChIP-seq data. Our method gains power by both reasoning about data from all time points in the time course and by explicitly modeling movements of peak boundaries. We showed that ChromTime predictions are reproducible across biological replicates and associate with relevant genomic features such as changes in gene expression and transcription factor binding. We demonstrated that territorial changes of peaks can contain additional information beyond ChIP-seq signal changes with respect to gene expression of proximal genes. ChromTime allows for novel analysis of directionality of spatial dynamics of chromatin marks. In this context, we showed for multiple chromatin marks that the direction of asymmetric expansions and contractions of peaks correlates strongly with direction of transcription in proximity of transcription start sites. ChromTime is generally applicable to modeling time courses of chromatin marks, and thus should be a useful tool to gaining insights into dynamics of epigenetic gene regulation in a range of biological systems.

METHODS

Overview of the ChromTime method

ChromTime takes as input a set of files in BED format with genomic coordinates of aligned sequencing reads from ChIP-seq experiments for a single mark over the time course and, optionally, from a set of control experiments. ChromTime consists of two stages (**Fig 3.1B-C**):

1) Detecting genomic intervals (blocks) potentially containing regions of ChIP-seq signal enrichment (peaks)

2) Learning a probabilistic mixture model for boundary dynamics of peaks within blocks throughout the time course and computing the most likely spatial dynamic and peak boundaries for each block throughout the whole time course

Detecting genomic blocks containing regions of ChIP-seq signal enrichment

The aim of this stage is to determine approximately the genomic coordinates of regions with potential ChIP-seq peaks at any time point in the time course. The signal within these blocks will be used as input to build the mixture model in the next stage of ChromTime. ChromTime

supports analysis of both narrow marks and broad marks in two different modes. The method partitions the genome into non-overlapping bins of predefined length, BIN_SIZE (200 bp in narrow mode, 500 bp in broad mode by default) and counts for each bin and time point the number of sequencing reads whose alignment starting positions after shifting by a predefined number of bases (100 bp in the direction of alignment by default) are within its boundaries. Next, each bin at each time point is tested for enrichment based on a Poisson background distribution at a predefined false discovery rate (5% by default). The expected number of reads for a bin at position *p* and time point *t*, $\lambda_{t,p}$, in the Poisson test is computed conservatively as the maximum of:

1) If control reads are provided: for each of windows of size w=1,000bp, 5,000bp and 20,000bp the average number of control reads in the window centered at the current bin, normalized by the ratio of total reads in ChIP-seq and control experiments, that is:

$$\lambda_{t,p,w} = \frac{\#\,[\text{Total Foreground Reads}]}{\#\,[\text{Total Control Reads}]}\frac{\text{BIN\_SIZE}}{w}\text{Ctrl}_{t,p,w}$$

where $\text{Ctrl}_{t,p,w}$ is the total number of control reads in each of window of size *w* around the bin at position *p* at time point *t*.

2) The average number of foreground ChIP-seq reads per genomic bin;

3) 1 read

Testing multiple different windows sizes for the background is a strategy we adopted from the MACS2 peak caller[35].

Within each time point, consecutive bins that are significantly enriched are merged into continuous intervals. The intervals are further extended in both directions to include continuous stretches of bins where each bin is significant based on a Poisson background distribution at a weaker P-value threshold (0.15 by default). Extended intervals within a predefined number of non-significant bins, MAX_GAP (3 bins by default), are further joined together. This joining strategy has been previously implemented by other peak callers for single datasets such as SICER[37]. Next, overlapping intervals across time points are grouped into blocks. To capture more of the potential background signal and to increase the likelihood that central bins within blocks contain higher ChIP-seq signal, the start and end positions of each block are extended additionally by a predefined number of bins, BLOCK_EXTEND, (5 by default) upstream from the left-most coordinate and downstream from the right-most coordinate of the intervals in the block, respectively, or up to the middle point between the current block and its adjacent blocks if they are within BLOCK_EXTEND bins apart. Restricting BLOCK_EXTEND to a relatively limited number of bins helps in keeping the running time of the method within reasonable bounds.

In narrow mode, blocks that contain multiple intervals at the same time point separated by gaps of non-significant bins longer than MAX_GAP are split into sub-blocks at each gap between those intervals. In particular, all gaps within a block are intersected across the time points that have gaps. For each gap intersection, the block is split at the position with the lowest average ChIP-seq signal across all time points. In broad mode, no such splitting is performed in order to avoid excessive peak fragmentation.

Probabilistic mixture model for boundary dynamics of peaks within blocks across the time course

The ChIP-seq signal within the blocks is used as input to build a probabilistic mixture model for the boundary dynamics of the peaks within blocks. One core assumption of the model is that each block contains zero or one peak at each time point. This implies that at each time point, the bins within a block can be partitioned into three continuous intervals: left-flanking background, foreground peak, and right-flanking background. Let $O_{t,p}$ denote the number of observed ChIP-seq reads in a bin at position p at time point t, and $V_{t,p}$ denote the label of that bin (one of Peak or Background). The conditional distribution of $O_{t,p}$ conditioned on $V_{t,p}$ is modeled with different negative binomial distributions depending on the value of $V_{t,p}$:

$$P\left(O_{t,p} = k \middle| V_{t,p} = \text{Peak}\right) = \text{NB}\left(k; \mu_{\text{Peak},t,p}, \delta_t\right)$$

$$= \frac{\Gamma(k + \delta_t)}{k!\,\Gamma(\delta_t)} \left(\frac{\delta_t}{\mu_{\text{Peak},t,p} + \delta_t}\right)^{\delta_t} \left(\frac{\mu_{\text{Peak},t,p}}{\mu_{\text{Peak},t,p} + \delta_t}\right)^{k}$$

and

$$P\left(O_{t,p} = k \middle| V_{t,p} = \text{Background}\right) = \text{NB}\left(k; \mu_{\text{Background},t,p}, \delta_t\right)$$

$$= \frac{\Gamma(k + \delta_t)}{k!\,\Gamma(\delta_t)} \left(\frac{\delta_t}{\mu_{\text{Background},t,p} + \delta_t}\right)^{\delta_t} \left(\frac{\mu_{\text{Background},t,p}}{\mu_{\text{Background},t,p} + \delta_t}\right)^{k}$$

Similarly to negative binomial regression models[55], ChromTime models the mean of each component through the log link as a linear combination of a two-dimensional vector of covariates, $X_{t,p}$,= (1, $\log \lambda_{t,p}$), which includes a constant term and the logarithm of the expected number of reads as computed in the previous section, $\lambda_{t,p}$:

$$\mu_{\text{Peak},t,p} = \exp[\alpha_t + \gamma_t \log \lambda_{t,p}]$$

$$\mu_{\text{Background},t,p} = \exp\left[\beta_t + \gamma_t \log \lambda_{t,p}\right]$$

where α_t, β_t and γ_t are time point specific scalar parameters. Negative binomial distributions have been successfully employed in a similar manner to capture the over-dispersion of

sequencing reads in ChIP-seq experiments in peak callers for single samples such as ZINBA[56]. Of note however, ChromTime requires that the dispersion parameter δ_t and the coefficient γ_t are shared between the two components at each time point. The first requirement ensures that the distribution with the smaller mean value has higher probabilities compared to the distribution with the larger mean value for the lowest values of the support domain of the negative binomial distribution, and that the opposite holds for the largest values of the support domain (See **Supplementary Methods**). Sharing the dispersion parameter here is analogous to sharing the variance parameter in Gaussian mixture models. The second requirement to share the γ_t parameter ensures that the input signal has equal importance in each component.

Formally, let $B_{L,t}$ and $B_{R,t}$ denote the bin indices relative to the beginning of the block of the first and the last bin in the peak partition at time *t*, respectively, and *N* be the length of the block (i.e. $1 \leq B_{L,t} \leq N+1$ and $0 \leq B_{R,t} \leq N$, with values of $B_{L,t} = N+1$ and $B_{R,t} = 0$ corresponding to the special cases of starting a peak after all positions and ending a peak before all positions in a block, respectively). For $B_{L,t}$ and $B_{R,t}$ to denote valid interval boundaries, ChromTime requires that $B_{L,t} \leq B_{R,t} + 1$ at each time point. This can be formally encoded by introducing one auxiliary binary variable for each time point in the model, Z_t, such that:

$$P(Z_t = 1 | B_{L,t} = l, B_{R,t} = r) = \begin{cases} 1 & \text{if } l \leq r+1 \\ 0 & \text{otherwise} \end{cases}$$

and thus also

$$P(Z_t = 0 | B_{L,t} = l, B_{R,t} = r) = \begin{cases} 0 & \text{if } l \leq r + 1 \\ 1 & \text{otherwise} \end{cases}$$

ChromTime treats all Z_t variables as observed with values equal to 1 for all blocks and time points. Then, the probability of a given partitioning of the ChIP-seq signal at time *t* under the model is:

$$\begin{aligned} P\left(B_{L,t} = l, B_{R,t} = r, Z_t = 1, O_t \middle| X_t\right) \\ = P\left(Z_t = 1 | B_{L,t} = l, B_{R,t} = r\right) \\ \times \prod_{p=1}^{l-1} \mathrm{NB}(O_{t,p}; \mu_{\mathrm{Background},t} = \exp[\beta_t + \gamma_t \log \lambda_{t,p}], \delta_t) \\ \times \prod_{p=l}^{r} \mathrm{NB}\left(O_{t,p}; \mu_{\mathrm{Peak},t} = \exp\left[\alpha_t + \gamma_t \log \lambda_{t,p}\right], \delta_t\right) \\ \times \prod_{p=r+1}^{N} \mathrm{NB}(O_{t,p}; \mu_{\mathrm{Background},t} = \exp[\beta_t + \gamma_t \log \lambda_{t,p}], \delta_t) \end{aligned}$$

An important special case of the above formulation when $B_{L,t} = B_{R,t} + 1$ corresponds to modelling the whole signal at time point *t* as background, which enables ChromTime to accommodate time points with no peaks.

ChromTime assumes uniform prior probabilities for the left and the right end boundaries at the first time point:

$$P(B_{L,1} = l) = Unif(1, N + 1)$$

and

$$P(B_{R,1} = r) = Unif(0, N)$$

where $Unif(a, b)$ denotes the uniform distribution of integer numbers in the closed interval $[a, b]$.

Between any two time points the ChromTime model allows for one of three possible dynamics at both the left and the right end boundaries of a peak: Steady, Expand or Contract. To capture the change of boundary positions between consecutive time points *t* and *t+1* we define the quantities $J_{L,t} = B_{L,t} - B_{L,t+1}$ and $J_{R,t} = B_{R,t+1} - B_{R,t}$ corresponding to the left and right boundaries respectively. Positive values of $J_{L,t}$ and $J_{R,t}$ indicate the number of bases a peak expanded, whereas negative values indicate the number of bases a peak contracted, and a value of 0 indicates the peak held steady on the left and the right side respectively. ChromTime models $J_{L,t}$ *and* $J_{R,t}$ with a different probability distribution for each of the three dynamics. Let $D_{S,t}$ denote the dynamic between time points *t* and *t+1* on boundary side *S*, where *S* is one of *L* (left side) or *R* (right side). For Steady dynamic, ChromTime uses the Kronecker delta function:

$$P\left(J_{S,t} \middle| D_{S,t} = \text{Steady}\right) = \begin{cases} 1, & \text{if } J_{S,t} = 0 \\ 0, & \text{otherwise} \end{cases}$$

For Expand and Contract, ChromTime employs negative binomial distributions to model the number of genomic bins a peak boundary moves relative to the minimal movement of one bin required for peak expansions and contractions:

$$P\left(J_{S,t} = j \middle| D_{S,t} = \text{Expand}\right) = NB(j-1; \mu_{\text{Expand,t}}, \delta_{\text{Expand,t}})$$

and

$$P\left(J_{S,t} = j \middle| D_{S,t} = \text{Contract}\right) = NB(-j-1; \mu_{\text{Conract,t}}, \delta_{\text{Contract,t}})$$

Furthermore, each distribution is parametrized with a mean and dispersion parameter depending on the dynamic and the time point, t: $\mu_{\text{Expand,t}}, \delta_{\text{Expand,t}}$ for expansions, and $\mu_{\text{Contract,t}}, \delta_{\text{Contract,t}}$ for contractions. Of note, in negative binomial distributions the probabilities for negative integers are defined to be 0. Therefore, the above parametrization enforces that boundary movements of negative or zero length (i.e. contracting or steady, respectively) are impossible for expansions and that boundary movements of positive or zero length (i.e. expanding or steady) are impossible for contractions.

The ChromTime model additionally assumes that there is a prior probability to observe each dynamic between time points *t* and *t+1*, $P(D_{S,t} = d) = \pi_{t,d}$, which is the same at each side (left and right). Thus, the joint probability to observe a particular partition of the data into peaks and background and dynamics $d_{L,t}$ and $d_{R,t}$ on the left and right side at two consecutive time points, respectively, can be expressed as:

$$P\left(B_{L,t} = l_t, B_{L,t+1} = l_{t+1}, D_{L,t} = d_{L,t}, B_{R,t} = r_t, B_{R,t+1} = r_{t+1}, D_{R,t} = d_{R,t}, Z_t = 1, Z_{t+1} = 1, O_t, O_{t+1} \middle| X_t, X_{t+1}\right) =$$
$$= P\left(B_{L,t} = l_t, B_{R,t} = r_t, Z_t = 1, O_t \middle| X_t\right) \times P\left(J_{L,t} = l_t - l_{t+1} \middle| D_{L,t} = d_{L,t}\right) \times P\left(D_{L,t} = d_{L,t}\right)$$
$$\times P\left(J_{R,t} = r_{t+1} - r_t \middle| D_{R,t} = d_{R,t}\right) \times P\left(D_{R,t} = d_{R,t}\right)$$
$$\times P\left(B_{L,t+1} = l_{t+1}, B_{R,t+1} = r_{t+1}, Z_{t+1} = 1, O_{t+1} \middle| X_{t+1}\right)$$

In a block $\boldsymbol{O}$ with covariates $\boldsymbol{X}$ in a time course with *T* time points, the total probability of a particular sequence of dynamics and boundary positions on the left side ($\mathfrak{D}_L$ and $\mathcal{B}_L$ respectively) and on the right side ($\mathfrak{D}_R$ and $\mathcal{B}_R$ respectively) can be expressed as:

$$P(\mathfrak{D}_L, \mathcal{B}_L, \mathfrak{D}_R, \mathcal{B}_R, \boldsymbol{O}|\boldsymbol{X}) =$$
$$P\left(B_{L,1} = l_1, B_{R,1} = r_1, Z_1 = 1, O_1 \middle| X_1\right) \prod_{t=2}^{T} \Big(P\left(B_{L,t} = l_t, B_{R,t} = r_t, Z_t = 1, O_t \middle| X_t\right)$$
$$\times P\left(J_{L,t-1} = l_{t-1} - l_t \middle| D_{L,t-1} = d_{L,t-1}\right) \times P\left(D_{L,t-1} = d_{L,t-1}\right)$$
$$\times P\left(J_{R,t-1} = r_t - r_{t-1} \middle| D_{R,t-1} = d_{R,t-1}\right) \times P\left(D_{R,t-1} = d_{R,t-1}\right)\Big)$$

where $\mathfrak{D}_S = (d_{S,t} | 1 \leq t \leq T-1)$ are *T-1* dimensional vectors of dynamics labels for each pair of consecutive time points on the left or the right side ($S = L$ or R), respectively, and $\mathcal{B}_S$ are the corresponding T dimensional vectors of boundary positions on each side at each time point. The total probability of the ChIP-seq signal in a block can be expressed as a sum over all possible sequences of dynamics and peak boundary positions that can generate the block across all time points. Thus, the probability of block $\boldsymbol{O}$ is:

$$P(\boldsymbol{O}|\boldsymbol{X}) = \sum_{\mathfrak{D}_L, \mathcal{B}_L, \mathfrak{D}_R, \mathcal{B}_R} P(\mathfrak{D}_L, \mathcal{B}_L, \mathfrak{D}_R, \mathcal{B}_R, \boldsymbol{O}|\boldsymbol{X})$$

where $\mathfrak{D}_L$ and $\mathfrak{D}_R$ each iterate over all possible 3^{T-1} combinations of peak boundary dynamics, and $\mathcal{B}_L$ and $\mathcal{B}_R$ each iterate over all possible ways to place left and right end boundaries across all time points that are consistent with the requirements that $B_{L,t} \leq B_{R,t} + 1$ at each time point. Let $\mathcal{O}$ be the total set of blocks in the data, $\mathcal{X}$ be the set of two-dimensional vectors containing the constant term and the log of the expected number of reads at each position and time point for each block, and M be the total number of blocks. Then the total likelihood of all blocks is:

$$P(\mathcal{O}|\mathcal{X}) = \prod_{i=1}^{M} P(\boldsymbol{O}_i|\boldsymbol{X}_i)$$

Model optimization

The total set of parameters of the model consists of:

1) Prior probabilities of each dynamic *d* at each time point *t*: $\pi_{t,d}$

2) Parameters of the negative binomial distributions that model the Peak and the Background components at each time point: α_t, β_t, γ_t and δ_t.

3) Parameters of the negative binomial distributions that model the boundary movements in Expand and Contract dynamics at each time point: $\mu_{\text{Expand,t}}, \delta_{\text{Expand,t}}$ and $\mu_{\text{Contract,t}}, \delta_{\text{Contract,t}}$, respectively.

The optimal parameter values are attempted to be estimated by Expectation Maximization (EM). In particular, ChromTime finds a local maximum of the expectation of the complete log-likelihood function (for details, see **Supplementary Methods**):

$$LL(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}) = \sum_{i=1}^{M} \log P(\boldsymbol{O}_i | \boldsymbol{X}_i)$$

Computing the most likely spatial dynamic and peak boundaries for each block across the whole time course

After the optimal values for all model parameters are estimated from the data, for each block the most likely positions of the peak boundaries at each time point are calculated. This procedure consists of two steps. First, ChromTime determines for each block all time points with significantly low probability of containing a false positive non-zero peak. Second, conditioned on those time points, ChromTime computes the most likely assignment of the peak boundary variables at each side and each time point (for details, see **Supplementary Methods**).

ChromTime parameters used in this study

In this work, we applied ChromTime in narrow mode on all data for H3K4me2, H3K4me3, H3K27ac and H3K(9,14)ac marks. We applied ChromTime in broad mode on all data for H3K79me2, Pol2, H3K4me1, H3K27me3, H3K9me3 and H3K36me3 marks. All other parameters were set to their default values.

SUPPLEMENTARY METHODS

ChromTime model optimization

As stated in the Methods, the total set of parameters of the model consists of:

4) Prior probabilities of each dynamic *d* at each time point *t*: $\pi_{t,d}$

5) Parameters of the negative binomial distributions that model the Peak and the Background components at each time point: α_t, β_t, γ_t and δ_t.

6) Parameters of the negative binomial distributions that model the boundary movements in Expand and Contract dynamics at each time point: $\mu_{\text{Expand,t}}, \delta_{\text{Expand,t}}$ and $\mu_{\text{Contract,t}}, \delta_{\text{Contract,t}}$, respectively.

The optimal parameter values are attempted to be estimated by Expectation Maximization (EM). In particular, ChromTime finds a local maximum of the expectation of the complete log-likelihood function:

$$LL(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}) = \sum_{i=1}^{M} \log P(\boldsymbol{O}_i | \boldsymbol{X}_i)$$

Furthermore,

$\log P(\boldsymbol{O}_i | \boldsymbol{X}_i) =$

$$= \sum_{l_1=1}^{N_i+1} \sum_{r_1=l_1-1}^{N_i} P\left(B_{L,1} = l_1, B_{R,1} = r_1 | \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \Bigg(\sum_{p=1}^{l_1-1} \log P\left(O_{i,1,p} | V_{1,p} = \text{Background}, X_{i,1,p}\right)$$

$$+ \sum_{p=l_1}^{r_1} \log P\left(O_{i,1,p} | V_{1,p} = \text{Peak}, X_{i,1,p}\right)$$

$$+ \sum_{p=r_1+1}^{N_i} \log P\left(O_{i,1,p} | V_{1,p} = \text{Background}, X_{i,1,p}\right)$$

$$+ \log P\left(B_{L,1} = l_1\right) + \log P\left(B_{R,1} = r_1\right) \Bigg)$$

$$+ \sum_{t=2}^{T} \sum_{d_{L,t-1} \in \mathbb{D}} \sum_{d_{R,t-1} \in \mathbb{D}} \sum_{l_t=1}^{N_i+1} \sum_{r_t=l_t-1}^{N_i} \sum_{l_{t-1}=1}^{N_i+1} \sum_{r_{t-1}=l_{t-1}-1}^{N_i}$$

$$\mathrm{P}\left(B_{L,t} = l_t, B_{L,t-1} = l_{t-1}, D_{L,t-1} = d_{L,t-1}, B_{R,t} = r_t, B_{R,t-1} = r_{t-1}, D_{R,t-1} = d_{R,t-1} | \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right)$$

$$\times \Bigg(\Bigg(\log P\left(B_{L,t} = l_t, B_{L,t-1} = l_{t-1} | D_{L,t-1} = d_{L,t-1}\right)$$

$$+ \log P\left(B_{R,t} = r_t, B_{R,t-1} = r_{t-1} | D_{R,t-1} = d_{R,t-1}\right) + \log P\left(D_{L,t-1} = d_{L,t-1}\right)$$

$$+ \log P\left(D_{R,t-1} = d_{R,t-1}\right) + \sum_{p=1}^{l_t-1} \log P\left(O_{i,t,p} | V_{t,p} = \text{Background}, X_{i,t,p}\right)$$

$$+ \sum_{p=l_t}^{r_t} \log P\left(O_{i,t,p} | V_{t,p} = \text{Peak}, X_{i,t,p}\right) + \sum_{p=r_t+1}^{N_i} \log P\left(O_{i,t,p} | V_{t,p} = \text{Background}, X_{i,t,p}\right) \Bigg) \Bigg)$$

where $O_{i,t,p}$ and $X_{i,t,p}$ denote the number of observed foreground reads and covariates, respectively, in block i at time point t at position p, and $\mathbb{D}$ denotes the set of all dynamics (Steady, Expand and Contract).

The complete log likelihood simplifies substantially, if we substitute in the above equation each of the following terms:

$$\sum_{l_1=1}^{N_i+1}\sum_{r_1=l_1-1}^{N_i} P\left(B_{L,1}=l_1, B_{R,1}=r_1|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\left(\sum_{p=1}^{l_1-1}\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Background}, X_{i,1,p}\right)\right.$$

$$+\sum_{p=l_1}^{r_1}\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Peak}, X_{i,1,p}\right)$$

$$\left.+\sum_{p=r_1+1}^{N_i}\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Background}, X_{i,1,p}\right)\right)$$

$$=\sum_{p=1}^{N_i}\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Peak}, X_{i,1,p}\right)\sum_{l_1=1}^{p}\sum_{r_1=p}^{N_i} P\left(B_{L,1}=l_1, B_{R,1}=r_1|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$$

$$+\sum_{p=1}^{N_i}\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Background}, X_{i,1,p}\right)$$

$$\times\left(\sum_{l_1=1}^{p-1}\sum_{r_1=l_1-1}^{p-1} P\left(B_{L,1}=l_1, B_{R,1}=r_1|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)+\sum_{l_1=p+1}^{N_i+1}\sum_{r_1=l_1-1}^{N_i} P\left(B_{L,1}=l_1, B_{R,1}=r_1|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\right)$$

$$=\sum_{p=1}^{N_i}\left(P\left(V_{1,p}=\text{Peak}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Peak}, X_{i,1,p}\right)\right.$$

$$\left.+\left(1-P\left(V_{1,p}=\text{Peak}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\right)\log P\left(O_{i,1,p}\middle|V_{1,p}=\text{Background}, X_{i,1,p}\right)\right)$$

$$=\sum_{p=1}^{N_i}\left(P\left(V_{1,p}=\text{Peak}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\log\left(\text{NB}\left(O_{i,1,p};\mu_{\text{Peak},1}=\exp\left[\alpha_1+\gamma_1\log\lambda_{i,1,p}\right],\delta_1\right)\right)\right.$$

$$\left.+P\left(V_{1,p}=\text{Background}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\log\left(\text{NB}\left(O_{i,1,p};\mu_{\text{Background},1}=\exp\left[\beta_1+\gamma_1\log\lambda_{i,1,p}\right],\delta_1\right)\right)\right)$$

and

$$\sum_{t=2}^{T}\sum_{d_{L,t-1}\in\mathbb{D}}\sum_{d_{R,t-1}\in\mathbb{D}}\sum_{l_t=1}^{N_i+1}\sum_{r_t=l_t-1}^{N_i}\sum_{l_{t-1}=1}^{N_i+1}\sum_{r_{t-1}=l_{t-1}-1}^{N_i}$$

$$\mathrm{P}\left(B_{L,t}=l_t, B_{L,t-1}=l_{t-1}, D_{L,t-1}=d_{L,t-1}, B_{R,t}=r_t, B_{R,t-1}=r_{t-1}, D_{R,t-1}=d_{R,t-1}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$$

$$\times\ \left(\log P\left(B_{L,t}=l_t, B_{L,t-1}=l_{t-1}|D_{L,t-1}=d_{L,t-1}\right)\right.$$

$$\left.+\log P(B_{R,t}=r_t, B_{R,t-1}=r_{t-1}|D_{R,t-1}=d_{R,t-1})\right)$$

$$=\sum_{t=2}^{T}\sum_{d_{L,t-1}\in\mathbb{D}}\sum_{l_{t-1}=1}^{N_i+1}\sum_{l_t=1}^{N_i+1}\mathrm{P}\left(B_{L,t}=l_t, B_{L,t-1}=l_{t-1}, D_{L,t-1}=d_{L,t-1}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$$

$$\times\log P\left(B_{L,t}=l_t, B_{L,t-1}=l_{t-1}|D_{L,t-1}=d_{L,t-1}\right)$$

$$+\sum_{t=2}^{T}\sum_{d_{R,t-1}\in\mathbb{D}}\sum_{r_{t-1}=0}^{N_i}\sum_{r_t=0}^{N_i}\mathrm{P}\left(B_{R,t}=r_t, B_{R,t-1}=r_{t-1}, D_{R,t-1}=d_{R,t-1}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$$

$$\times\log P\left(B_{R,t}=r_t, B_{R,t-1}=r_{t-1}|D_{R,t-1}=d_{R,t-1}\right)$$

$$=\sum_{t=1}^{T-1}\sum_{S\in\{L,R\}}\sum_{d_{S,t}\in\left\{\substack{\text{Expand}\\\text{Contract}}\right\}}\sum_{j=1}^{N_i}P\left(J_{S,t}=(-1)^{\omega(d_{S,t})}j, D_{S,t}=d_{S,t}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$$

$$\times\log\left(P\left(J_{S,t}=(-1)^{\omega(d_{S,t})}j|D_{S,t}=d_{S,t}\right)\right)$$

$$=\sum_{t=1}^{T-1}\sum_{S\in\{L,R\}}\sum_{d_{S,t}\in\left\{\substack{\text{Expand}\\\text{Contract}}\right\}}\sum_{j=1}^{N_i}P\left(J_{S,t}=(-1)^{\omega(d_{S,t})}j, D_{S,t}=d_{S,t}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)\log\left(\mathrm{NB}\left(j-1;\mu_{d_{S,t}},\delta_{d_{S,t}}\right)\right)$$

where

$$\omega(d_{S,t}) = \begin{cases} 1 & \text{if } d_{S,t} = \text{Contract} \\ 0 & \text{otherwise} \end{cases}$$

In the above we used the simplification that summing over all possible ways to place the peak boundaries on the left and on the right side at two consecutive time points is equivalent to summing over all possible distances between two left boundaries and all possible distances between two right boundaries, j.

Also, we can simplify:

$$\sum_{t=2}^{T} \sum_{d_{L,t-1} \in \mathbb{D}} \sum_{d_{R,t-1} \in \mathbb{D}} \sum_{l_t=1}^{N_i+1} \sum_{r_t=l_t-1}^{N_i} \sum_{l_{t-1}=1}^{N_i+1} \sum_{r_{t-1}=l_{t-1}-1}^{N_i}$$

$$\mathrm{P}\left(B_{L,t} = l_t, B_{L,t-1} = l_{t-1}, D_{L,t-1} = d_{L,t-1}, B_{R,t} = r_t, B_{R,t-1} = r_{t-1}, D_{R,t-1} = d_{R,t-1} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right)$$

$$\times \left(\log P\left(D_{L,t-1} = d_{L,t-1}\right) + \log P\left(D_{R,t-1} = d_{R,t-1}\right)\right)$$

$$= \sum_{t=1}^{T-1} \sum_{S \in \{L,R\}} \sum_{d_{S,t} \in \left\{ \begin{matrix} \text{Steady} \\ \text{Expand} \\ \text{Contract} \end{matrix} \right\}} P\left(D_{S,t} = d_{S,t} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log\left(P\left(D_{S,t} = d_{S,t}\right)\right)$$

$$= \sum_{t=1}^{T-1} \sum_{S \in \{L,R\}} \sum_{d_{S,t} \in \left\{ \begin{matrix} \text{Steady} \\ \text{Expand} \\ \text{Contract} \end{matrix} \right\}} P\left(D_{S,t} = d_{S,t} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log\left(\pi_{t,d_{S,t}}\right)$$

and

$$\sum_{t=2}^{T} \sum_{d_{L,t-1} \in \mathbb{D}} \sum_{d_{R,t-1} \in \mathbb{D}} \sum_{l_t=1}^{N_i+1} \sum_{r_t=l_t-1}^{N_i} \sum_{l_{t-1}=1}^{N_i+1} \sum_{r_{t-1}=l_{t-1}-1}^{N_i}$$

$$\mathrm{P}\left(B_{L,t} = l_t, B_{L,t-1} = l_{t-1}, D_{L,t-1} = d_{L,t-1}, B_{R,t} = r_t, B_{R,t-1} = r_{t-1}, D_{R,t-1} = d_{R,t-1} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right)$$

$$\times \left(\sum_{p=1}^{l_t-1} \log P\left(O_{i,t,p} \middle| V_{t,p} = \text{Background}, X_{i,t,p}\right) + \sum_{p=l_t}^{r_t} \log P\left(O_{i,t,p} \middle| V_{t,p} = \text{Peak}, X_{i,t,p}\right) \right.$$

$$\left. + \sum_{p=r_t+1}^{N_i} \log P\left(O_{i,t,p} \middle| V_{t,p} = \text{Background}, X_{i,t,p}\right) \right)$$

$$= \sum_{t=2}^{T} \sum_{p=1}^{N_i} \left(P\left(V_{t,p} = \text{Peak} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log P\left(O_{i,t,p} \middle| V_{t,p} = \text{Peak}, X_{i,t,p}\right) \right.$$

$$\left. + P\left(V_{t,p} = \text{Background} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log P\left(O_{i,t,p} \middle| V_{1,p} = \text{Background}, X_{i,t,p}\right) \right)$$

$$= \sum_{t=1}^{T} \sum_{p=1}^{N_i} \left(P\left(V_{t,p} = \text{Peak} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log \left(\mathrm{NB}\left(O_{i,t,p}; \mu_{\text{Peak},t} = \exp\left[\alpha_t + \gamma_t \log \lambda_{i,t,p}\right], \delta_t\right) \right) \right.$$

$$\left. + P\left(V_{t,p} = \text{Background} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log \left(\mathrm{NB}\left(O_{i,t,p}; \mu_{\text{Background},t} = \exp\left[\beta_t + \gamma_t \log \lambda_{i,t,p}\right], \delta_t\right) \right) \right)$$

With these substitutions, we can rewrite the complete log likelihood, $LL\left(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}\right)$, as

$$\sum_{i=1}^{M} \sum_{l_1=1}^{N_i+1} \sum_{r_1=l_1-1}^{N_i} P\left(B_{L,1} = l_1, B_{R,1} = r_1 | \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \left(\log P\left(B_{L,1} = l_1\right) + \log P\left(B_{R,1} = r_1\right) \right)$$

$$+\sum_{i=1}^{M}\sum_{t=1}^{T}\sum_{p=1}^{N_i}\Big(P\big(V_{t,p}=\text{Peak}\big|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\big)\log\Big(\text{NB}\big(O_{i,t,p};\mu_{\text{Peak},t}=\exp\big[\alpha_t+\gamma_t\log\lambda_{i,t,p}\big],\delta_t\big)\Big)$$

$$+\ P\big(V_{t,p}=\text{Background}\big|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\big)\log\Big(\text{NB}\big(O_{i,t,p};\mu_{\text{Bgr},t}=\exp\big[\beta_t+\gamma_t\log\lambda_{i,t,p}\big],\delta_t\big)\Big)\Big)$$

$$+\sum_{i=1}^{M}\sum_{t=1}^{T-1}\sum_{S\in\{L,R\}}\sum_{d_{S,t}\in\left\{\begin{smallmatrix}\text{Expand}\\\text{Contract}\end{smallmatrix}\right\}}\sum_{j=1}^{N_i}P\big(J_{S,t}=(-1)^{\omega(d_{S,t})}j,D_{S,t}=d_{S,t}\big|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\big)\log\Big(\text{NB}\big(j-1;\mu_{d_{S,t}},\delta_{d_{S,t}}\big)\Big)$$

$$+\sum_{i=1}^{M}\sum_{t=1}^{T-1}\sum_{S\in\{L,R\}}\sum_{d_{S,t}\in\left\{\begin{smallmatrix}\text{Steady}\\\text{Expand}\\\text{Contract}\end{smallmatrix}\right\}}P\big(D_{S,t}=d_{S,t}\big|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\big)\log\big(\pi_{t,d_{S,t}}\big)$$

Parameter initialization

The values of the model parameters before the first EM iteration are initialized as follows:

1) All dynamics priors, $\pi_{t,d}$, are set uniformly to $\frac{1}{3}$;
2) All parameters for the distributions modelling the Peak and Background components, α_t, β_t, γ_t and δ_t, are set to 1;
3) All dispersion and mean parameters for the distributions modelling the boundary movements in Expand and Contract dynamics, $\mu_{\text{Expand,t}}$, $\delta_{\text{Expand,t}}$, $\mu_{\text{Contract,t}}$, and $\delta_{\text{Contract,t}}$, are set to 1.

At each time point, ChromTime requires that the left boundary of each non-zero length peak is placed before its right boundary. This requirement in combination with the uniform priors over $B_{L,1}$ and $B_{R,1}$ induces non-uniform conditional probabilities $P\big(V_{t,p} = \text{Peak}|Z_t = 1\big)$, which depend on the position index p. As result, everything else being equal, bins in the middle of the block are more likely to be in the Peak component compared to flanking bins on each side. For example, in datasets with only one time point the conditional probability for a bin at position p $(1 \leq p \leq N)$ to be in a peak after marginalizing out the observed read counts in the model is:

$$
\begin{aligned}
&P\big(V_{1,p} = \text{Peak}|Z_1 = 1\big) \\
&\quad = \sum_{l=1}^{p}\sum_{r=p}^{N} P\big(B_{L,1} = l, B_{R,1} = r\big|Z_1 = 1\big) \\
&\quad = \sum_{l=1}^{p}\sum_{r=p}^{N} \frac{P(Z_1 = 1|B_{L,1} = l, B_{R,1} = r)P\big(B_{L,1} = l, B_{R,1} = r\big)}{P(Z_1 = 1)} \\
&\quad = \sum_{l=1}^{p}\sum_{r=p}^{N} \frac{P(Z_1 = 1|B_{L,1} = l, B_{R,1} = r)P\big(B_{L,1} = l\big)P(B_{R,1} = r)}{\sum_{l'=1}^{N+1}\sum_{r'=0}^{N} P(Z_1 = 1|B_{L,1} = l', B_{R,1} = r')P\big(B_{L,1} = l'\big)P(B_{R,1} = r')} \\
&\quad = \sum_{l=1}^{p}\sum_{r=p}^{N} \frac{1}{\sum_{l'=1}^{N+1}\sum_{r'=l'-1}^{N} 1} = \frac{2p(N-p+1)}{(N+1)(N+2)}
\end{aligned}
$$

The above equalities follow from $P(B_{L,1} = l') = P(B_{R,1} = r') = \frac{1}{N+1}$ for all $l' \in [1, N+1], r' \in [0, N]$ and $P(Z_1 = 1 | B_{L,1} = l', B_{R,1} = r') = 1$ for $l' \leq r' - 1$ and $P(Z_1 = 1 | B_{L,1} = l', B_{R,1} = r') = 0$, otherwise. Of note, $P(V_{1,p} = \text{Peak} | Z_1 = 1)$ is symmetric with respect to the center of the block and has its maximum of $\frac{N}{2(N+1)} < \frac{1}{2}$ at $p = \frac{N}{2}$. In datasets with more time points, $P(V_{t,p} = \text{Peak} | Z_t = 1)$ has shown in practice to have similar relationship to the position index p, which can be computed by marginalizing out all remaining latent variables and the observed read counts from the joint probability distribution defined by the model (**Fig 3.S1Di)**. These conditional probabilities play a role during the learning stage of ChromTime, because they direct the model during the initial iterations of the EM to correctly associate the Peak component with high ChIP-seq signal and the Background component with low ChIP-seq signal. The assumption that high ChIP-seq signal will more likely be located in the middle of blocks is motivated by the procedure that determines the block boundaries in the first phase of ChromTime, which naturally produces blocks with high signal in the middle compared to their flanking regions (**Fig 3.S1Dii-iv**). As result, no further efforts from the initialization or the training procedures are necessary in practice to identify correctly each component.

Expectation step

ChromTime provides an efficient implementation of the expectation step of EM based on a dynamic programming algorithm similar to the Baum-Welch algorithm for hidden Markov models. In brief, at each time point there are $O(N^2)$ ways to place the start and end positions of a peak, resulting in $O(N^4)$ combinations between any pair of consecutive time points. Thus, a standard forward-backward procedure that caches intermediate results can compute all expectations in $O(T*N^4)$ time and $O(T*N^2)$ memory. Since $O(T*N^4)$ time complexity can result in very long running times even for moderate N, ChromTime splits blocks that are longer than a predefined number of bins, MAX_BINS, (30 by default) into two halves (left and right) and estimates all sufficient statistics in each half independently. If block i is longer than MAX_BINS, the split is performed at the position with the highest average ChIP-seq signal across all time points in the block, K_i. This splitting procedure corresponds to imposing an additional constraint on the values of the boundary position variables that $B_{L,t} \leq K_i$ and $B_{R,t} \geq K_i - 1$ at each time point, t, while still having all bins between the left and the right boundaries annotated as peak bins (i.e. $V_{t,p} = \text{Peak}$ for all p such that $B_{L,t} \leq p \leq B_{R,t}$ and $V_{t,p} = \text{Background}$ for all other values of p). This heuristic reduces the time complexity to $O(T*N^2)$ and the memory footprint to $O(T*N)$, thus making the whole EM procedure run in feasible time and space. Since the $O(T*N^4)$ algorithm is applied only to blocks shorter than MAX_BINS bins, the total running time of ChromTime in a dataset of M blocks remains at most quadratic in the length of longer peaks in the data, $O(M*T*N^2)$.

For computational efficiency, if there are more than 10,000 blocks, ChromTime randomly selects 10,000 as input for the EM procedure.

Maximization step

The form of the complete log-likelihood implies that each set of model parameters can be optimized independently by solving for the roots of the respective partial derivatives. The dynamics prior probabilities are updated after each EM iteration as:

$$P(D_{L,t} = d) = P(D_{R,t} = d) = \frac{1}{2M}\sum_{i=1}^{M}\sum_{S\in\{L,R\}} P(D_{S,t} = d|\boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta})$$

Optimizing the peak and background signal components

The part of the total log likelihood that pertains to the peak and background signal components is:

$LL_{Signal}(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}) =$

$$= \sum_{i=1}^{M} \sum_{t=1}^{T} \sum_{p=1}^{N_i} P\left(V_{t,p} = \text{Peak} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log\left(\text{NB}\left(O_{i,t,p}; \mu_{\text{Peak},t} = \exp\left[\alpha_t + \gamma_t \log \lambda_{i,t,p}\right], \delta_t\right)\right) +$$

$$+ \sum_{i=1}^{M} \sum_{t=1}^{T} \sum_{p=1}^{N_i} P\left(V_{t,p} = \text{Background} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log\left(\text{NB}\left(O_{i,t,p}; \mu_{\text{Bgr},t} = \exp\left[\beta_t + \gamma_t \log \lambda_{i,t,p}\right], \delta_t\right)\right)$$

These equations are equivalent to the equations for finding the maximum likelihood estimates for the coefficients and the dispersion parameter of one weighted negative binomial regression for each time point and component (Peak and Background) that aims to predict the observed number of ChIP-seq reads, $O_{i,t,p}$ (as response), from the vector of covariates $X_{i,t,p} = \left[1, \log \lambda_{i,t,p}\right]$. The coefficients in our case are α_t and γ_t (for the Peak component) and β_t and γ_t (for the Background component). The weights for each regression correspond to the posterior probabilities $P\left(V_{t,p} = \text{Peak} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right)$ and $P\left(V_{t,p} = \text{Background} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right)$, respectively. In contrast to standard regressions, for each time point we have a pair of coupled negative binomial regressions that share the dispersion parameter δ_t and the coefficient γ_t. ChromTime implements a procedure that jointly optimizes each pair of coupled regressions, which is based on a modification of the glm.nb method from the MASS package[57] in R. In particular, we attempt to find the roots of the partial derivative of $LL_{Signal}\left(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}\right)$ with respect to the shared δ_t and γ_t. Each of these derivatives however is simply the sum of the partial derivatives with respect to each parameter of the two components. Therefore, the standard procedure of fitting weighted negative binomial regressions can be reused whereby the part that finds the

roots of the partial derivatives with respect to δ_t and γ_t, had they not been shared, is replaced by a routine that finds the roots of the sum of the partial derivatives across both components with respect to each parameter. On the other hand, the parts that find the roots of the partial derivatives with respect to α_t and β_t are the same as in the standard procedure for fitting weighted negative binomial regressions. The only other difference between our implementation and glm.nb is that ChromTime uses the HYBRD method from the MINPACK package[58] for finding roots of functions instead of Iterative Reweighted Least Squares (IRLS). In our tests, our optimization routine and glm.nb yielded very similar results for regular un-coupled weighted negative binomial regressions.

Optimizing the boundary movement components

The part of the total log likelihood that pertains to modelling the peak boundary movements is:

$$LL_{Movement}(\theta; \tilde{\theta}, \mathcal{O}, \mathcal{X}) =$$

$$= \sum_{i=1}^{M} \sum_{t=1}^{T-1} \sum_{S \in \{L,R\}} \sum_{d_{S,t} \in \left\{ \substack{\text{Expand} \\ \text{Contract}} \right\}} \sum_{j=1}^{N_i} P\left(J_{S,t} = (-1)^{\omega(d_{S,t})} j, D_{S,t} = d_{S,t} \middle| \boldsymbol{O}_i, \boldsymbol{X}_i; \tilde{\theta}\right) \log\left(\mathrm{NB}\left(j-1; \mu_{d_{S,t}}, \delta_{d_{S,t}}\right)\right)$$

Again, this equation is equivalent to the equation for finding the maximum likelihood estimates of the coefficients and the dispersion parameter of one weighted negative binomial regression

for each dynamic and time point that aims to predict the number of positions the left or the right boundary moves minus 1 ($j-1$, as response) from a single covariate which is the constant term equal to 1. The weights correspond to the posterior probability of moving the boundary by *j* positions, $P\left(J_{S,t}=(-1)^{\omega(d_{S,t})}j, D_{S,t}=d_{S,t}\middle|\boldsymbol{O}_i,\boldsymbol{X}_i;\tilde{\theta}\right)$. The procedure to find the maximum likelihood estimates is the same as the one used in the previous section, except that no sharing of parameters is enforced between any of the regressions.

Sharing dispersion parameter between negative binomial distributions

Sharing the dispersion parameter δ between two negative binomial distributions ensures that the distribution with the smaller mean value has higher probabilities compared to the distribution with the larger mean value for the lowest values of the support domain of the negative binomial distribution, and that the opposite holds for the largest values of the support domain. Here we will prove this claim. Let μ_1 and μ_2 be the means of two negative binomial distributions, *NB1* and *NB2*, respectively. Without loss of generality, we will assume that $0 \leq \mu_1 < \mu_2$. Dividing the probability mass functions of the two distributions gives:

$$\frac{NB1(k)}{NB2(k)}=\frac{\frac{\Gamma(k+\delta)}{k!\,\Gamma(\delta)}\left(\frac{\delta}{\mu_1+\delta}\right)^{\delta}\left(\frac{\mu_1}{\mu_1+\delta}\right)^{k}}{\frac{\Gamma(k+\delta)}{k!\,\Gamma(\delta)}\left(\frac{\delta}{\mu_2+\delta}\right)^{\delta}\left(\frac{\mu_2}{\mu_2+\delta}\right)^{k}}=\frac{\left(\frac{1}{\mu_1+\delta}\right)^{\delta}\left(\frac{\mu_1}{\mu_1+\delta}\right)^{k}}{\left(\frac{1}{\mu_2+\delta}\right)^{\delta}\left(\frac{\mu_2}{\mu_2+\delta}\right)^{k}}=\left(\frac{\mu_1}{\mu_2}\right)^{k}\left(\frac{\mu_2+\delta}{\mu_1+\delta}\right)^{\delta+k}$$

Since $\delta > 0$, substituting with $k = 0$, gives:

$$\frac{NB1(0)}{NB2(0)} = \left(\frac{\mu_1}{\mu_2}\right)^0 \left(\frac{\mu_2 + \delta}{\mu_1 + \delta}\right)^\delta = \left(\frac{\mu_2 + \delta}{\mu_1 + \delta}\right)^\delta > 1$$

Therefore, *NB1* has higher probability for $k = 0$ compared to *NB2*.

To prove that the opposite holds for the largest values of the support, we will take the limit of the above ratio for $k \to \infty$:

$$\lim_{k\to\infty} \frac{NB1(k)}{NB2(k)} = \lim_{k\to\infty} \left(\frac{\mu_1}{\mu_2}\right)^k \left(\frac{\mu_2 + \delta}{\mu_1 + \delta}\right)^{\delta+k} = \left(\frac{\mu_2 + \delta}{\mu_1 + \delta}\right)^\delta \lim_{k\to\infty} \left(\frac{\frac{\mu_1}{\mu_1 + \delta}}{\frac{\mu_2}{\mu_2 + \delta}}\right)^k = 0$$

The last equality holds, because:

$$\frac{\mu_1}{\mu_1 + \delta} - \frac{\mu_2}{\mu_2 + \delta} = \frac{\delta(\mu_1 - \mu_2)}{(\mu_1 + \delta)(\mu_2 + \delta)} < 0 \implies \frac{\frac{\mu_1}{\mu_1 + \delta}}{\frac{\mu_2}{\mu_2 + \delta}} < 1$$

Therefore, for sufficiently large k *NB2* has higher probability compared to *NB1*.

Computing the most likely spatial dynamic and peak boundaries for each block across the whole time course

After the optimal values for all model parameters are estimated from the data, for each block the most likely positions of the peak boundaries at each time point are calculated. This procedure consists of two steps. First, ChromTime determines for each block all time points with significantly low probability of containing a false positive non-zero peak. Second, conditioned on those time points, ChromTime computes the most likely assignment of the peak boundary variables at each side and each time point.

During the first step, for each block and each time point ChromTime computes the posterior probability that the whole time point is modelled as background, $\varphi_{t,i} = \prod_{p=1}^{N_i} P(V_{t,p} = \text{Background}|\boldsymbol{O}_i, \boldsymbol{X}_i)$. This probability can be interpreted as the probability of making a false positive non-zero length peak call as estimated by the model at time point *t* in block $\boldsymbol{O}_i$. To determine significant time points with low false positive probability, $\varphi_{t,i}$, ChromTime computes a time point specific threshold, τ_t, at a predefined false discovery rate (0.05 by default) by applying the standard Benjamini-Hochberg procedure[59] on all values of $\varphi_{t,i}$ from time point *t*.

In the second step, for each block ChromTime computes the most likely sequence of assignments of the boundary positions, conditioned on the event that all time points that failed to pass the FDR threshold in the previous step for the block are assigned to having no peaks. In particular, ChromTime executes a dynamic programming algorithm similar to the Viterbi algorithm for hidden Markov models, which uses the following recursive formula to find the most likely position for the peak boundaries at each side, *S* (Left or Right) and enforces that time points that failed the FDR test contain no peaks:

$$\mathrm{DP}_{t,l,r} = \begin{cases} \log P\left(B_{L,1} = l, B_{R,1} = r, O_1 \middle| X_1\right) & , t = 1 \\ \max\limits_{\substack{l_{t-1} \in [1,N+1] \\ r_{t-1} \in \begin{cases} [l_{t-1}-1, N] \text{ if } \varphi_{t,i} \le \tau_t \\ \{l_{t-1}-1\} \text{ otherwise} \end{cases}} \begin{pmatrix} \mathrm{DP}_{t,l_{t-1},r_{t-1}} + \\ \log P\left(B_{L,t} = l, B_{R,t} = r, O_t \middle| X_t\right) + \\ \log P\left(J_{L,t-1} = l_{t-1} - l \middle| D_{L,t-1} = \mathrm{DYN}(l_{t-1} - l)\right) + \\ \log P\left(D_{L,t-1} = \mathrm{DYN}(l_{t-1} - l)\right) + \\ \log P\left(J_{R,t-1} = r - r_{t-1} \middle| D_{R,t-1} = \mathrm{DYN}(r - r_{t-1})\right) + \\ \log P\left(D_{R,t-1} = \mathrm{DYN}(r - r_{t-1})\right) \end{pmatrix} & , t > 1 \end{cases}$$

$$\text{where } \mathrm{DYN}(j) = \begin{cases} \text{Steady} & \text{if } j = 0 \\ \text{Expand} & \text{if } j \ge 1 \\ \text{Contract} & \text{if } j \le -1 \end{cases}$$

and DP denotes the dynamic programing cube of size $T(N+1)^2$ that stores the log likelihood for the best assignment of the peak boundary variables up to time point *t*. Tracing the DP cube from the highest value on row *T* back to row 1 retrieves the best assignment of the peak end variables. Similarly to the expectation step of the EM phase, for blocks longer than

MAX_BINS bins the best Viterbi path is chosen among the splits at the top MAX_BINS positions in the block sorted by their average ChIP-seq signal across all time points. Since MAX_BINS is a predefined constant, the whole procedure has the same time and space complexity as computing the expectations in the EM phase of ChromTime. The dynamic between any two time points is determined from the direction of the movement of the optimal positions of the corresponding boundaries.

Transcription factor binding and DNaseI hypersensitivity data

In Figure 3.3A, TF binding data for GATA3 was used from the same study of mouse T cell development[15].

In Figure 3.S4A, OCT4, NANOG and P300 binding data for H1-hESC was downloaded from the ENCODE project [4]:

https://www.encodeproject.org/files/ENCFF002CJF/@@download/ENCFF002CJF.bed.gz

https://www.encodeproject.org/files/ENCFF002CJA/@@download/ENCFF002CJA.bed.gz

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/wgEncodeBroadHistoneH1hescP300kat3bPk.broadPeak.gz

In Figure 3.S4A IMR90 peaks for P300 generated from previously published data[60] were downloaded at 0.05 FDR from:

http://chip-atlas.org/view?id=SRX212184

Narrow peaks for all other TFs in Figure 3.S4A were downloaded from the ENCODE consortium[4] from the following URL:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

In Figure 3.S4A, DNaseI hypersensitivity peaks for H1 and IMR90 cells were downloaded from the Roadmap Epigenomics Consortium[5]:

http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/E003-DNase.macs2.narrowPeak.gz

http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/E017-DNase.macs2.narrowPeak.gz

Cell type specific and shared annotations were derived after subtracting one of the annotations from the other with the BedTools software[61] using the "*bedtools subtract –A"* command and by intersecting the two annotations with "*bedtools intersect"*, respectively.

Fold enrichments calculation

In Figures 3.3A and 3.S4A, three types of ChromTime blocks (T1-Tn Steady, Tx-Tn Expand and T1-Tx Contract) were defined to start and end from the left most to the right most boundary, respectively, of the non-zero length peaks across all time points within each block. "T1-Tn Steady" blocks have non-zero length peaks at all time points with both left and right boundaries predicted as steady across all time points. "Tx-Tn Expand" blocks have a non-zero length peak at the last time point, have no contracting boundaries and have an expanding boundary at at least one pair of consecutive time points. "T1-Tx Contract" blocks have a non-zero length peak at the first time point, have no expanding boundaries and have a contracting boundary at at least one pair of consecutive time points.

In Figure 3.S3, ChromTime peaks annotated with each dynamic class (e.g. E/E, E/S, etc) at each pair of consecutive time points were defined to start and end from the left most to the right most coordinate, respectively, of the peaks at the corresponding pair of time points.

Fold enrichments of base pair overlap in Figure 3.3A, 3.S3, 3.S4A and 3.6 were computed for each pair of genomic features by dividing the size of their observed overlap by the size of their expected overlap. For two genomic features *A* and *B*, the observed overlap was defined as the total number of bases in their intersection, $|A \cap B|$. The expected overlap was defined based on

a binomial null model that preserves the size of each feature and treats the two features as independently distributed within a certain set of eligible background genomic positions, *G*:

$$\mathrm{E}(A,B) = |A| * \frac{|B|}{|G|}$$

where |*G*| denotes the size of the set of all eligible positions. In Figures 3.3A and 3.S4A, the set of eligible background positions was defined for each time course as all genomic bases covered by ChromTime peaks for that time course. In Figure 3.S3, the set of eligible background positions at each time point was defined as all bases in the union of all predicted ChromTime peaks from both replicates at that time point. In Figure 3.6, the set of eligible background positions is defined to be all genomic bases in the corresponding genome. In addition, in Figure 3.S3 and 3.6 in order to avoid extremely high enrichments due to very rare predicted dynamics classes a pseudo-count of 200 bp (i.e. one genomic bin) was added to each overlap. In Figure 3.S3, the fold enrichments for each pair of consecutive time points were $\log_2$-transformed and the average of the log-transformed values across all time point pairs is shown. In Figure 3.6, geometric means were taken across enrichments at each pair of consecutive time points and the resulted average enrichments were capped at a maximum value of 50. Clustering with optimal leaf ordering[62] was performed Figure 3.6 after this procedure.

Gene expression

Gene expression data used in Figures 3.3B, 3.4, 3.S4B, 3.5, 3.S5 and 3.S6 was used from the same studies that provided the ChIP-seq data for the corresponding histone marks. Prior to all analysis, the gene expression values (RPKM) were transformed with log2(1+RPKM) and then Z-transformed within each time point. In Figures 3.3B and 3.S4B, we used all blocks with at least one non-zero length peak called with ChromTime by using data from all time points in each time course. Peaks that did not overlap a TSS were excluded from this analysis. For peaks that overlap multiple TSSs, the average gene expression change across all overlapping TSSs was used.

Average rank differences based on gene expression change

In Figures 3.4 and 3.S5 we compared whether ranking block boundaries by ChromTime posterior probabilities for Expand (left panels) and Contract (right panels) dynamics has better agreement with gene expression changes than ranking block boundaries by the number of genomic bases that peak boundaries move between consecutive time points when peaks are called at each time point by using input data only from that time point. The above comparison provides insights into whether ChromTime's posterior probabilities, which are computed by reasoning jointly about all time points in the time course, have benefits compared to analyzing boundary movements of peaks that are called at each time point in isolation. To compute the latter in Figure 3.4 and 3.S5A we applied ChromTime to call peaks at each time point by using

as input only data from that time point (ChromTime SINGLE). We then used the boundaries of those peaks that overlapped predicted ChromTime peaks called by using data from all time points in the time course (ChromTime ALL). In Figure 3.S5B, we used the boundaries of broad peaks called with MACS2[35] with the --broad option that overlap blocks with predicted ChromTime ALL peaks. To call MACS2 peaks in H3K4me2 in mouse T cell development[15] we additionally used "--nomodel --extsize 200" options. In Figure 3.S5C, we used the boundaries of peaks called with SICER[37] with parameters as recommended[63] that overlap blocks with predicted ChromTime ALL peaks.

To determine whether ranking based on ChromTime posteriors or ranking based on boundary movements between consecutive time points is better, we evaluated the consistencies of these rankings with respect to ranking all boundaries by the change in gene expression of genes whose TSS directly overlaps predicted peaks. Blocks with peaks that did not overlap a TSS were excluded from the analyses performed for Figures 3.4 and 3.S5. For blocks with peaks that overlap multiple TSSs, the average gene expression change across all overlapping TSSs was used.

Between a pair of consecutive time points each block is represented by two boundaries – one boundary on the left side and one on the right side. If M is the total number of blocks with predicted peaks from both methods in each pairwise comparison, then the total number of

boundaries is *2*M*. For a pair of consecutive time points, *t* and *t+1*, each of these boundaries can either stay steady, or expand or contract relative to time point *t*. Technically, ChromTime blocks have the same number of boundaries at all time points even if some time points are predicted as zero length peaks (i.e. all background). In the case of a zero length peak at time point *t*, the model sets the left and the right end boundaries so that $B_{L,t} = B_{R,t} + 1$ and posterior probabilities for the dynamics involving time point *t* are still estimated. The actual values of $B_{L,t}$ and $B_{R,t}$ of the boundaries of zero length peaks are determined by the model in the same way as boundary positions for non-zero length peaks.

Expanding block boundaries of H3K4me2 and H3K4me3 peaks are expected to be found near genes that increase in gene expression at time point *t+1* relative to time point *t*, and vice versa for contracting boundaries. To quantify the degree to which ranking by posteriors and ranking by boundary movements of peaks in isolation is more consistent with gene expression changes, for each pair of consecutive time points *t* and *t+1* we computed the following quantities for each boundary *i* and dynamic *D*:

1) $\mathrm{CT}(i, D)$ – rank of boundary *i* when boundaries are sorted by ChromTime posteriors of dynamic *D* (where *D* is one of *Expand* (left plots) or *Contract* (right plots)) in descending order from the highest to the lowest posterior.

2) $\mathrm{BM}(i, D)$ – rank of boundary i when boundaries are sorted by the number of genomic bases that the boundaries of the overlapping peaks called in isolation move. For D = *Expand* (left plots), this ranking is performed in descending order from the most expanding to the most contracting boundary, and vice versa for D = *Contract* (right plots). The number of genomic bases that a boundary moves is calculated as $(-1)^{\omega(S)}(B_{S,t+1} - B_{S,t})$, where $B_{S,t+1}$ and $B_{S,t}$ are the genomic positions of the peak boundary on side S (left or right) at times t+1 and t respectively, and $\omega(S) = 1$ for left end boundaries and $\omega(S) = 0$ for right end boundaries. To handle cases, where ChromTime SINGLE or MACS2 or SICER did not call a peak at a given time point within a block, a zero length peak for these methods were created artificially at the middle position of the non-zero length peak at the nearest time point to the time point with no peak. In this way, the first appearance in time of peaks for these methods within a block is effectively treated as a positive movement of both the left and the right boundary from a zero length peak at the previous time point that was placed in the middle of the new peak. Conversely, the complete removal of a peak is treated as a negative movement of both the left and the right boundary to a zero length peak at the next time point placed in the middle of the removed peak. Also, if two consecutive time points have no peaks, then the boundary movements between them of both the left and the right boundaries is set to 0. This procedure ensures that block boundaries at all time points can be associated with a boundary movement based peaks called in isolation.

3) $\Delta E(i, D)$ – rank of boundary i when boundaries are sorted by the change in gene expression at time point *t+1* relative to time point *t* of the overlapping TSS. The gene expression change at each TSS is quantified as $E_{t+1} - E_t$, where and E_{t+1} and E_t are the normalized gene expression levels at time points *t+1* and *t*, respectively. For *D=Expand* (left plots), this ranking is performed in descending order (i.e. most up-regulated genes rank first and most down-regulated genes rank last), and vice versa for *D=Contract* (right plots).

In all rankings, ties were broken randomly. Then for each rank k in rankings *CT* and *BM*, where $1 \leq k \leq 2 * M$, and dynamic D we computed the average $\Delta E(i, D)$ rank of all boundaries up to and including rank *k*:

$$\mu_{CT}(k, D) = \frac{1}{k} \sum_{k\prime=1}^{k} \Delta E(CT^{-1}(k', D), D)$$

and

$$\mu_{BM}(k, D) = \frac{1}{k} \sum_{k\prime=1}^{k} \Delta E(BM^{-1}(k', D), D)$$

where $CT^{-1}(k', D)$ and $BM^{-1}(k', D)$ denote the inverse functions of the rankings CT and BM, respectively, which return the boundary of rank k' according to the corresponding ranking. The two quantities, $\mu_{CT}(k, D)$ and $\mu_{BM}(k, D)$, measure the degree to which rankings CT and BM

associate with differential gene expression as measured by the ranking ΔE up to the first k boundaries ordered by each ranking. In particular, $\mu_{CT}(k, D) < \mu_{BM}(k, D)$ corresponds to the case where CT is more consistent with ΔE than BM is, because the first k boundaries according to CT on average rank higher in terms of gene expression changes compared to the first k boundaries according to BM, and vice versa for $\mu_{CT}(k, D) > \mu_{BM}(k, D)$. In Figures 3.4Aii, 3.4Bii and the bottom plots in 3.S5, for each pair of consecutive time points *t* and *t+1* we plot the difference $\delta(k, D) = \mu_{BM}(k, D) - \mu_{CT}(k, D)$ as a function of k. Thus, positive values correspond to ranks for which CT better associates with gene expression as measured by ΔE than BM, and vice versa for negative values. The shaded regions correspond to 95% confidence intervals. Finally, due to large fluctuations of the μ_{CT} and μ_{BM} quantities in the top ranks, the plots are shown for $k \geq 20$.

Gene expression changes as function of ChIP-seq signal changes for different predicted ChromTime dynamics

In Figures 3.5 and 3.S6, gene expression changes are plotted as function of ChIP-seq signal change for all peaks annotated with the same predicted ChromTime dynamics. First, each block *i* was associated with the difference of the standardized log2 gene expression (as defined in the **Gene expression** section) at the nearest TSS within 50kb at each pair of consecutive time points *t* and *t+1*, $\Delta e_{i,t}$.

In Figure 3.5, to compute the change of ChIP-seq density we first computed the $\log_2$ Control-normalized ChIP-seq density for each block *i* at each time point *t*, as:

$$D_{i,t} = \log_2(1 + RPKM_{M,t,i}) - \log_2\left(1 + \lambda_{t,i}\right)$$

where $RPKM_{M,t,i}$ denotes the number of reads per kilobase per million mapped reads (RPKM) from the histone mark ChIP-seq, *M*, and $\lambda_{t,i}$ denotes the average expected number of reads for all bins in block *i* at time point *t*. The RPKM and $\lambda_{t,i}$ values for each block at each time point were computed over the same genomic territory spanning from the left most to the right most coordinate of the predicted peaks across all time points in the block. To compute the change in signal density between consecutive time points, we computed the difference between the corresponding $D_{i,t}$ values:

$$\delta_{i,t} = D_{i,t+1} - D_{i,t}$$

To visualize the relationship between signal density changes and gene expression changes, the tuples $(\delta_{i,t}, \Delta e_{i,t})$ were pooled together across all time points and a Loess regression with linear polynomials was fitted with the loess function[64] in R with default parameters except for

degree=1. In cases with too many tuples the R package required excessive memory to compute the loess curves and, thus, 10,000 tuples were chosen at random as input for the loess function.

In Figure 3.S6, the same procedure was applied except that $\delta_{i,t}$ values were defined as differential peak scores of peaks called by different methods. Two independent differential peak callers were used, MACS2[35] and SICER[37], which were recommended by a previous study that evaluated a number of differential peak callers[63]. With each differential caller we called differential and common peaks between every pair of consecutive time points in each time course according to the instructions in the evaluation study[63]. For each peak caller, we intersected ChromTime blocks with all called peaks from the caller and performed the analysis only on peaks identified by both ChromTime and the corresponding differential peak caller.

For MACS2, the differential score was defined as the $\log_2$ fold change of the signal of each differential or common peak as computed by MACS2.

For SICER, the differential score for each peak was defined as:

$$(-1)^q(-\log_{10}\min(\text{FDR\_A\_vs\_B}, \text{FDR\_B\_vs\_A})$$

where

$$q = \begin{cases} 1 & \text{if FDR\_A\_vs\_B} < \text{FDR\_B\_vs\_A} \\ 0 & \text{otherwise} \end{cases}$$

The differential score for SICER has a negative sign for SICER peaks with enriched signal at the previous time point compared to the next time point, and a positive sign for peaks with enriched signal at the next time point compared to the previous time point. For peaks for which the FDR outputted by SICER was 0, the differential score was defined as the maximum differential score across all peaks with non-zero FDRs multiplied by $(-1)^q$ in order to take into account the direction of enrichment.

Analysis of directional preferences of spatial dynamics of chromatin marks

The average log-ratios in Figure 3.7 were computed across all tested datasets for the corresponding mark (see **Results**). For each time course, we split all ChromTime peaks into two groups, TSS+-1kb and TSS distal. The TSS+-1kb group contains all peaks whose distance to the nearest TSS is less than 1kb as measured with the "bedtools closest" software[61]. All other peaks were put in the TSS distal group. For each dataset and each group, we computed the log ratios for each pair of consecutive time points after adding a pseudo-count of 10 TSS +-1kb and 10 TSS distal peaks. Then, we averaged those log-ratios across all time points in the dataset. For marks mapped in at least six time courses, we then plotted the average across all

tested datasets as a solid black line in each subplot. A two-tailed Mann-Whitney test was performed for these marks to assess the statistical significance of the difference between the TSS+-1kb and TSS distal groups with the SciPy library[65]. To compute averages and test statistics in Figures 3.7 and 3.S7, all datasets were treated as independent, except in the case of the mouse hematopoiesis data[17]. For this time course, we applied ChromTime on data from each branch of the hematopoietic tree and computed a single average across all branches for the corresponding enrichment. The single average was then used in place of the values for each individual branch. This was done in order to avoid biasing statistics towards the Mouse hematopoiesis data, since branches in the hematopoietic tree overlap substantially and, thus, cannot be treated as independent datasets.

ABBREVIATIONS

DHS: DNase I Hypersensitive Sites

HM: Histone mark

TF: Transcription factor

TSS: Transcription start-site

ChIP-seq: Chromatin immunoprecipitation coupled with high throughput DNA sequencing

FDR: False Discovery Rate

EM: Expectation Maximization

DECLARATIONS

Availability of data and materials

ChromTime software is freely available at: https://github.com/ernstlab/ChromTime

No new experimental datasets were generated within this study. All ChIP-seq datasets used as input for ChromTime are publicly available from the references listed in **Table 1**. Links to all other datasets (gene expression, TF binding and DHS) used in this study are provided in the **Supplementary Methods**.

Funding

This work was supported by the CIRM Training Grant TG2-01169, the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA Training Program (P.F.); NIH grants R01ES024995, U01HG007912, DP1DA044371, an NSF CAREER Award #1254200 and an Alfred P. Sloan Fellowship (J.E.).

Acknowledgements

We are grateful to Constantinos Chronis, Kathrin Plath and members of the Ernst lab for useful discussions.

Figure 3.1

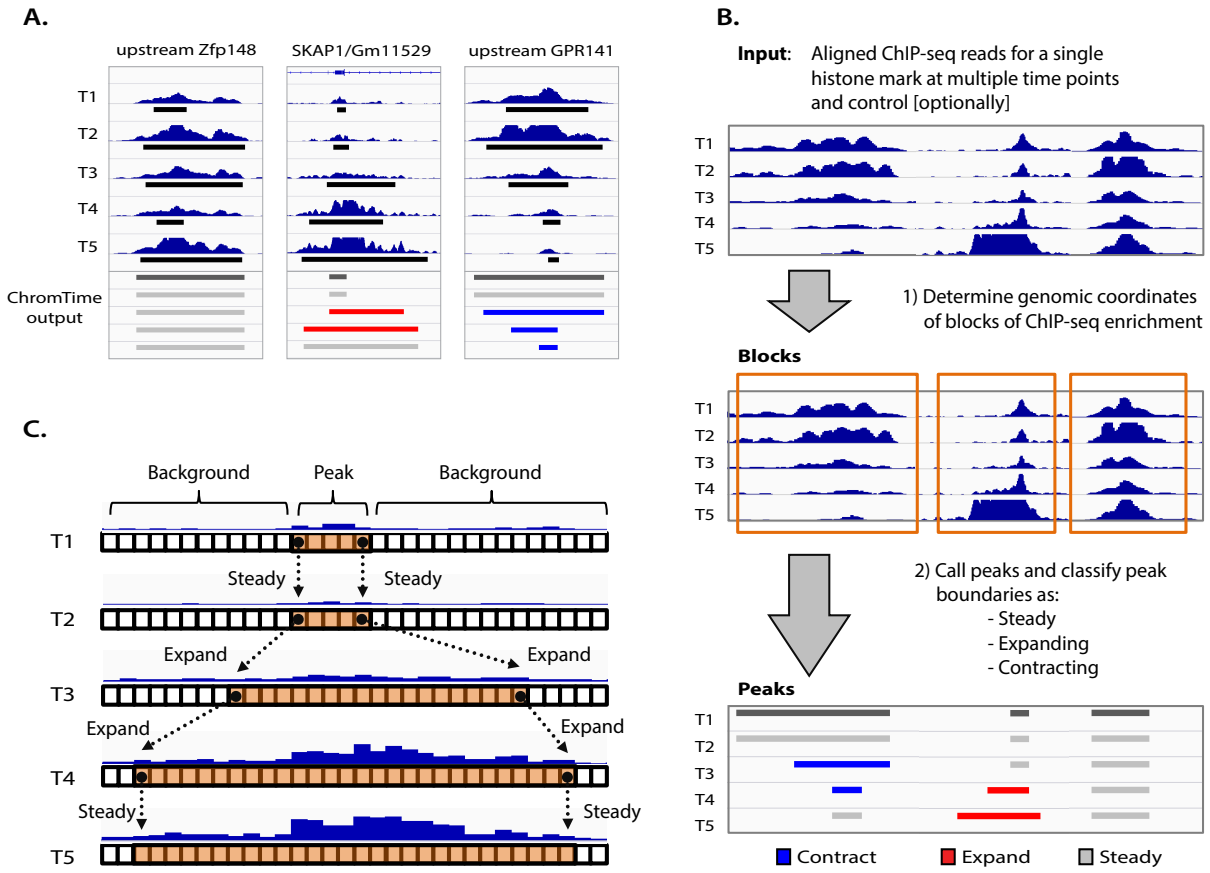

A.
upstream Zfp148
SKAP1/Gm11529
upstream GPR141
T1
T2
T3
T4
T5
ChromTime output
B.
Input: Aligned ChIP-seq reads for a single histone mark at multiple time points and control [optionally]
T1
T2
T3
T4
T5
1) Determine genomic coordinates of blocks of ChIP-seq enrichment
Blocks
T1
T2
T3
T4
T5
2) Call peaks and classify peak boundaries as:
- Steady
- Expanding
- Contracting
Peaks
T1
T2
T3
T4
T5
Contract
Expand
Steady
C.
Background
Peak
Background
T1
Steady
Steady
T2
Expand
Expand
T3
Expand
Expand
T4
Steady
Steady
T5

Figure 3.1: **Overview of the ChromTime method. (A)** Examples of H3K4me2 peaks with steady, expanding and contracting boundary dynamics, shown from left to right respectively, across five time points during mouse T cell development[15]. Time points 1, 2 and 3 correspond to in vitro differentiated T cell precursors (FLDN1, FLDN2a and FLDN2b), whereas time points 4 and 5 correspond to in vivo purified thymocytes (ThyDN3 and ThyDP). Normalized ChIP-seq signal, MACS2[35] peaks (black rectangles) and ChromTime output are shown for each time point. Peaks upstream of Zfp148 gene are called steady by ChromTime despite fluctuations of MACS2 peak boundaries. In contrast, ChromTime calls a peak at the Skap1/GM11529 promoter to expand after time points 2 and 3. Conversely, ChromTime calls a peak upstream of GPR141 gene to contract after time points 2, 3, and 4. **(B)** Overview of the ChromTime method. During the block finding stage, input ChIP-seq and, optionally, control reads are used to determine blocks of signal enrichment. In the dynamics prediction stage, for each block, peak boundary positions are predicted at each time point and peak boundary dynamics are predicted at each pair of consecutive time points. **(C)** Schematic of predicting dynamics for one block. Boxes represent genomic bins at each time point. ChIP-seq signal is depicted as blue bars for each bin whose height represents the number of reads mapped to the bin. ChromTime learns a probabilistic mixture model from the input data to partition each block at each time point into peak and background components. Bins in the peak component (orange) mark ChIP-seq peaks whereas those in the background component (white) mark flanking background signal. The movement of the boundaries on the left and the right side of

peaks between consecutive time points are estimated by reasoning jointly about the input data from all time points.

Figure 3.2

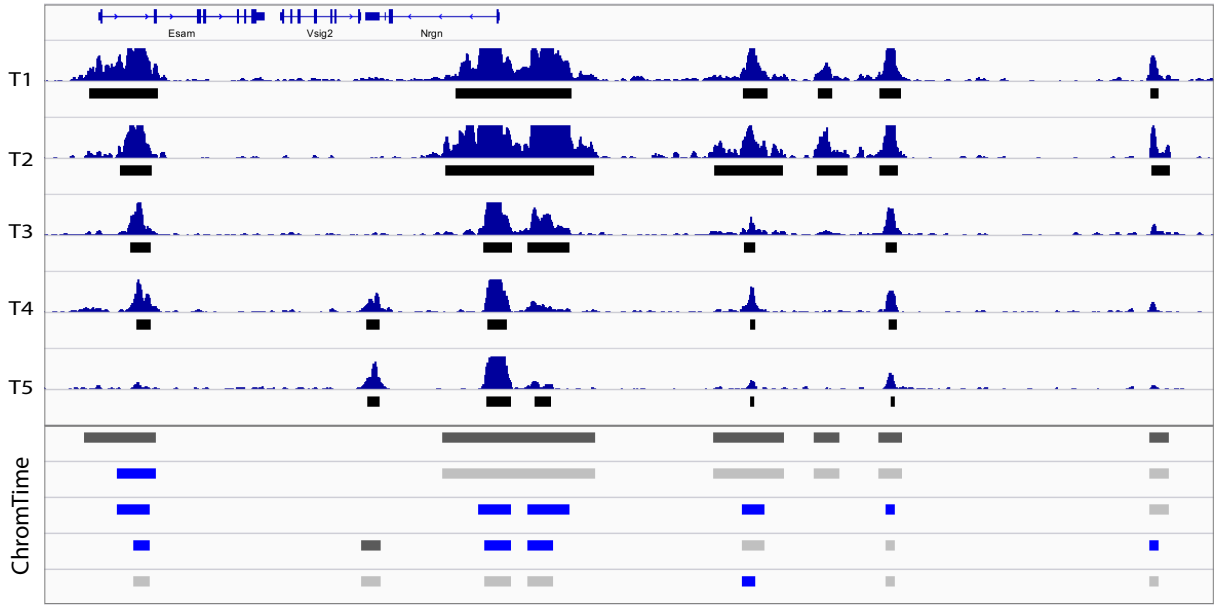


Figure 3.2: **Sample output from ChromTime with contracting peaks.** Genome browser screenshot with sample output of ChromTime for H3K4me2 from the T cell development time course in mouse[15] with 5 time points at the Esam/Vsig2/Nrgn locus. Time points 1, 2 and 3 correspond to in vitro differentiated T cell precursors (FLDN1, FLDN2a, and FLDN2b), whereas time points 4 and 5 correspond to in vivo purified thymocytes (ThyDN3 and ThyDP). The input ChIP-seq signal and MACS2[35] peaks (black boxes under each signal track) are shown in the upper panel of the screenshot. The ChromTime predicted peaks colored by their boundary dynamics for each block at each time point are shown in the bottom panel. The first peak in each block is colored in dark grey. Each subsequent peak is colored with respect to the predicted dynamic relative to its previous time point. Peaks with steady boundaries on both sides are shown in light grey, and those with at least one contracting boundary are shown in blue. Nearby peaks that touch boundaries are visualized as one peak by the genome browser. Not shown in the figure are expanding peaks, peaks at single time points and peaks with opposite dynamics (Expand on the left and Contract on the right, or vice versa), which would be colored in red, orange and black, respectively. See **Fig 3.S2** for examples of predicted expanding peaks.

Figure 3.3

A

Dynamic	% all block bases	FLDN1-ThyDP shared GATA3	ThyDP-specific GATA3	FLDN1-specific GATA3
T1-Tn Steady	28	1.2	1.1	0.5
Tx-Tn Expand	13	0.6	3.4	0.2
T1-Tx Contract	55	1	0.3	1.5
Base %		0.1	0.2	0.1

B

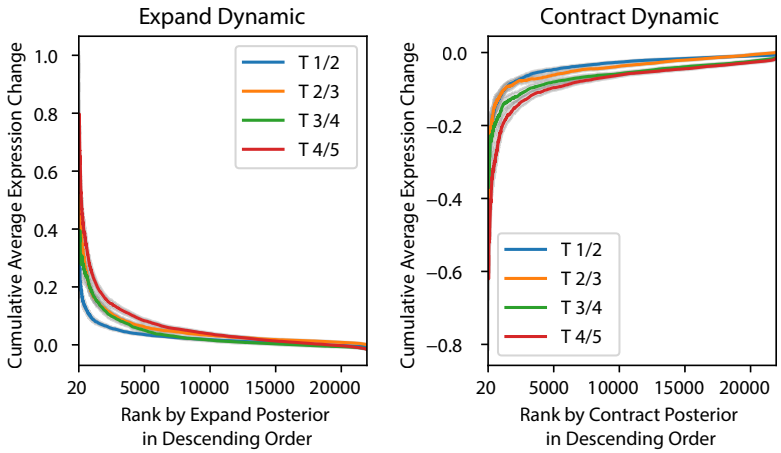


Figure 3.3: **Changes in GATA3 binding and gene expression at predicted H3K4me2 dynamics in T cell development.** (**A**) Fold enrichments of cell type specific and shared peaks of GATA3, which is a master regulator in T cell development[15], are shown for three sets of blocks with predicted H3K4me2 peaks: 1) blocks with peaks present at all time points whose boundaries hold steady on both sides throughout the whole time course (T1-Tn Steady); 2) blocks with non-contracting peaks whose boundaries expand between at least one pair of consecutive time points and have a peak at the last time point (Tx-Tn Expand); and 3) blocks with non-expanding peaks whose boundaries contract between at least one pair of consecutive time points and have a peak at the first time point (T1-Tx Contract). First column shows the percentage of bases out of all bases covered by peaks of the set. Last row shows the baseline percentage for each feature out of all bases covered by ChromTime peaks at any time point. Percentages are colored from 0 (white) to 100 (green). Fold enrichments in each column are colored from 1 (white) to the maximum value in the column (red). FLDN1 and ThyDP denote differentiated T cell precursors and purified thymocytes, which are the first and the last time point, respectively. **(B)** Boundaries of blocks with predicted H3K4me2 peaks overlapping annotated TSSs were sorted in decreasing order by their posterior probability for Expand dynamic (left plots) and Contract dynamic (right plots) at each pair of consecutive time points (see **Supplementary Methods**). Gene expression differences between consecutive time points were calculated as the average difference of all overlapping TSSs for each block. For each posterior rank (X-axis) the plot shows the cumulative average gene expression difference (Y-

axis). Expanding boundaries associated with increase of gene expression and contracting boundaries associated with decrease of gene expression. Shaded regions correspond to 95% confidence intervals.

Figure 3.4

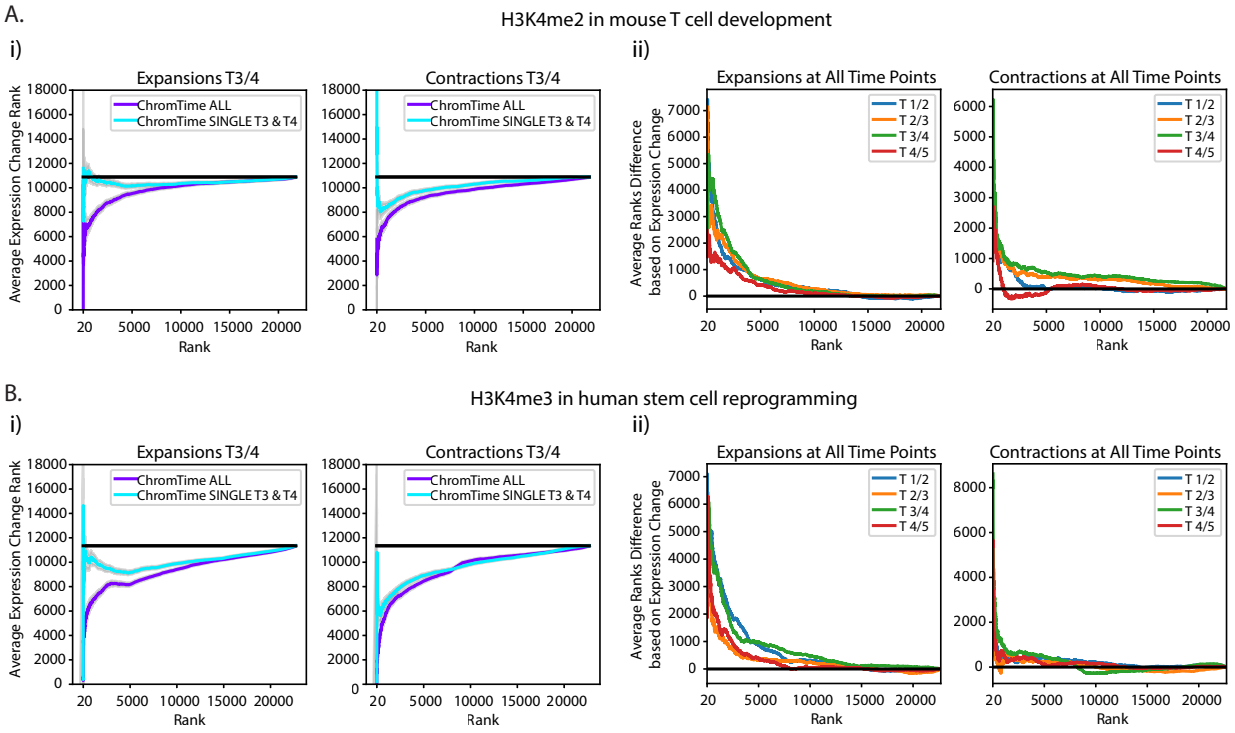

A.
H3K4me2 in mouse T cell development
i)
Expansions T3/4
Contractions T3/4
ChromTime ALL
ChromTime SINGLE T3 & T4
Average Expression Change Rank
Rank
ii)
Expansions at All Time Points
Contractions at All Time Points
T 1/2
T 2/3
T 3/4
T 4/5
Average Ranks Difference based on Expression Change
B.
H3K4me3 in human stem cell reprogramming

Figure 3.4: ChromTime predictions associate better with expression changes than boundary movements of peaks called in isolation. (A) For H3K4me2 in mouse T cell development[15] ChromTime was applied once with data from all time points (ChromTime ALL), and once with single time points in isolation (ChromTime SINGLE; see **Supplementary Methods**). Time points 1, 2 and 3 correspond to T cell precursors, whereas 4 and 5 to purified thymocytes. Peaks called by both procedures overlapping annotated TSSs were analyzed for their relationship with gene expression changes. (**i**) (Left) Comparison of agreement with expression for expansions when applying ChromTime ALL and ChromTime SINGLE for the change between time points 3 and 4. Block boundaries were sorted in decreasing order of their Expand posterior probabilities from ChromTime ALL and compared to sorting them in decreasing order of the difference of peak boundary positions in ChromTime SINGLE peaks with positive differences indicating peaks expanding with time. Each boundary was also ranked by gene expression difference of overlapping TSSs in decreasing order with positive expression differences indicating gain with time. The cumulative average boundary rank of expression change (Y-axis) is shown for the boundary change ranking for ChromTime ALL and ChromTime SINGLE (X-axis). Low Y-values indicate stronger association with expression changes. Black line shows expected average expression change rank. Shaded regions indicate 95% confidence intervals. Plots for other time points can be found in **Fig 3.S5**. (Right) Analogous to left plots for Contract posterior probabilities for ChromTime ALL, increasing order of the difference of boundary change positions for ChromTime SINGLE, and

increasing order of expression changes. (**ii**) Differences between ChromTime ALL and ChromTime SINGLE values shown in (**i**) between time points 3 and 4 as well as for each other pair of time points. Positive values correspond to boundaries ranked based on ChromTime ALL better associating with gene expression changes than ChromTime SINGLE. Black lines show expected difference of zero between random rankings. **(B)** As in (**A**) for H3K4me3 in human stem cell reprogramming[21]. Time points correspond to hiF-T fibroblasts, fibroblasts at 5, 10 and 20 days after induction, and hIPSC-T cells.

Figure 3.5

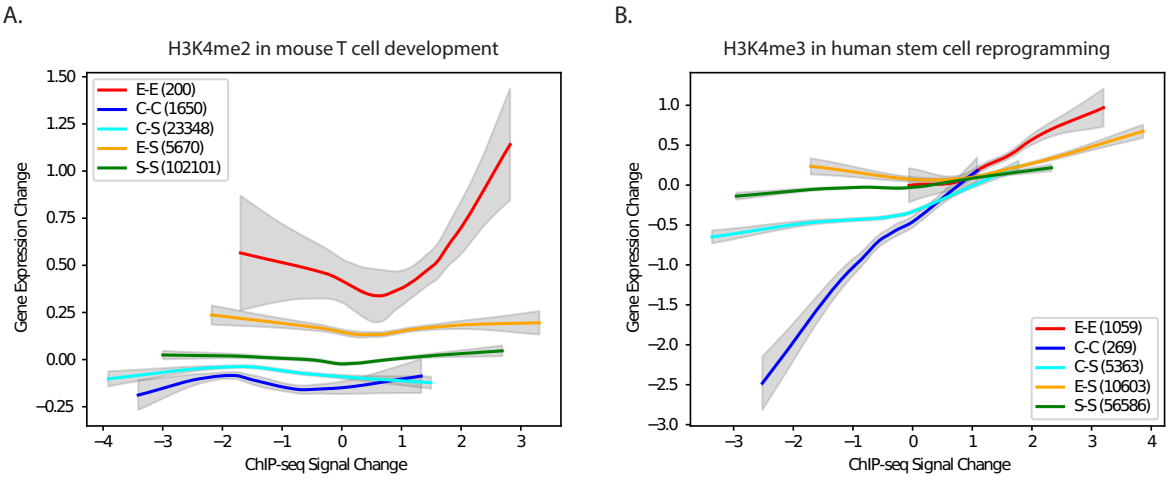
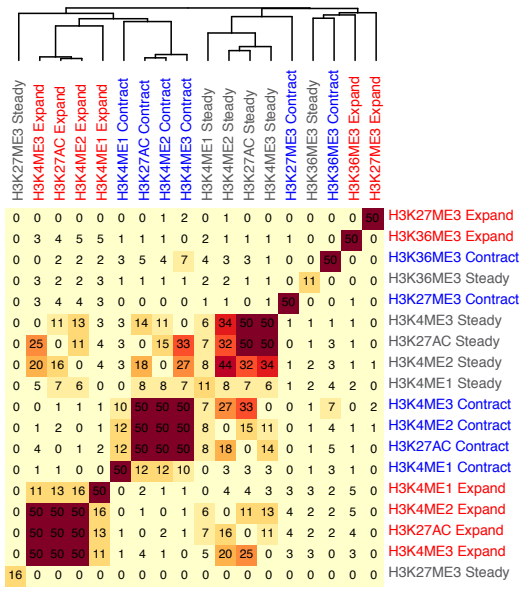

A.
H3K4me2 in mouse T cell development
E-E (200)
C-C (1650)
C-S (23348)
E-S (5670)
S-S (102101)
Gene Expression Change
ChIP-seq Signal Change
B.
H3K4me3 in human stem cell reprogramming
E-E (1059)
C-C (269)
C-S (5363)
E-S (10603)
S-S (56586)
Gene Expression Change
ChIP-seq Signal Change

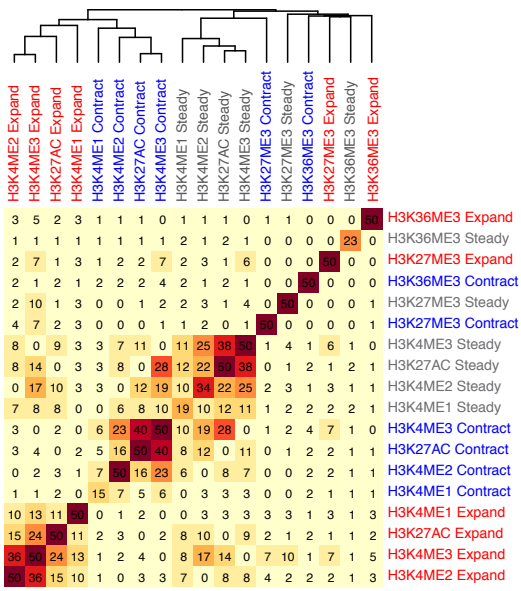
Figure 3.5: **Spatial dynamics can contain additional information about gene expression changes beyond ChIP-seq signal density changes.** Gene expression change is plotted as function of ChIP-seq signal density change after loess smoothing for each predicted ChromTime dynamic for **(A)** H3K4me2 dynamics in T cell development in mouse[15]; and **(B)** in H3K4me3 dynamics in stem cell reprogramming in human[21]. Peaks of each type of dynamics were pooled from all time points for this analysis. Peaks with asymmetric dynamics E/S and S/E were pooled together in the "E-S" group. Similarly C/S and S/C peaks were pooled in the "C-S" group. In both systems, peaks with the same signal density change can associate with different gene expression changes depending on the predicted spatial dynamic. Shaded regions represent 95% confidence intervals.

Figure 3.6

A.


H3K27ME3 Steady
H3K4ME3 Expand
H3K27AC Expand
H3K4ME2 Expand
H3K4ME1 Expand
H3K4ME1 Contract
H3K27AC Contract
H3K4ME2 Contract
H3K4ME3 Contract
H3K4ME1 Steady
H3K4ME2 Steady
H3K27AC Steady
H3K4ME3 Steady
H3K27ME3 Contract
H3K36ME3 Steady
H3K36ME3 Contract
H3K36ME3 Expand
H3K27ME3 Expand
0 0 0 0 0 0 0 1 2 0 1 0 0 0 0 0 0 50 H3K27ME3 Expand
0 3 4 5 5 1 1 1 0 2 1 1 1 1 0 0 50 0 H3K36ME3 Expand
0 0 2 2 2 3 5 4 7 4 3 3 1 0 0 50 0 0 H3K36ME3 Contract
0 3 2 2 3 1 1 1 1 2 2 1 1 0 11 0 0 0 H3K36ME3 Steady
0 3 4 4 3 0 0 0 0 1 1 0 1 50 0 0 1 0 H3K27ME3 Contract
0 0 11 13 3 3 14 11 0 6 34 50 50 1 1 1 1 0 H3K4ME3 Steady
0 25 0 11 4 3 0 15 33 7 32 50 50 0 1 3 1 0 H3K27AC Steady
0 20 16 0 4 3 18 0 27 8 44 32 34 1 2 3 1 1 H3K4ME2 Steady
0 5 7 6 0 0 8 8 7 11 8 7 6 1 2 4 2 0 H3K4ME1 Steady
0 0 1 1 1 10 50 50 50 7 27 33 0 0 1 7 0 2 H3K4ME3 Contract
0 1 2 0 1 12 50 50 50 8 0 15 11 0 1 4 1 1 H3K4ME2 Contract
0 4 0 1 2 12 50 50 50 8 18 0 14 0 1 5 1 0 H3K27AC Contract
0 1 1 0 0 50 12 12 10 0 3 3 3 0 1 3 1 0 H3K4ME1 Contract
0 11 13 16 50 0 2 1 1 0 4 4 3 3 3 2 5 0 H3K4ME1 Expand
0 50 50 50 16 0 1 0 1 6 0 11 13 4 2 2 5 0 H3K4ME2 Expand
0 50 50 50 13 1 0 2 1 7 16 0 11 4 2 2 4 0 H3K27AC Expand
0 50 50 50 11 1 4 1 0 5 20 25 0 3 3 0 3 0 H3K4ME3 Expand
16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 H3K27ME3 Steady

B.


H3K4ME2 Expand
H3K4ME3 Expand
H3K27AC Expand
H3K4ME1 Expand
H3K4ME1 Contract
H3K4ME2 Contract
H3K27AC Contract
H3K4ME3 Contract
H3K4ME1 Steady
H3K4ME2 Steady
H3K27AC Steady
H3K4ME3 Steady
H3K27ME3 Contract
H3K27ME3 Steady
H3K36ME3 Contract
H3K27ME3 Expand
H3K36ME3 Steady
H3K36ME3 Expand
3 5 2 3 1 1 1 0 1 1 1 0 1 1 0 0 0 50 H3K36ME3 Expand
1 1 1 1 1 1 1 1 2 1 2 1 0 0 0 0 23 0 H3K36ME3 Steady
2 7 1 3 1 2 2 7 2 3 1 6 0 0 0 50 0 0 H3K27ME3 Expand
2 1 2 1 2 2 2 4 2 1 2 1 0 0 50 0 0 0 H3K36ME3 Contract
2 10 1 3 0 0 1 2 2 3 1 4 0 50 0 0 0 1 H3K27ME3 Steady
4 7 2 3 0 0 0 1 1 2 0 1 50 0 0 0 0 1 H3K27ME3 Contract
8 0 9 3 3 7 11 0 11 25 38 50 1 4 1 6 1 0 H3K4ME3 Steady
8 14 0 3 3 8 0 28 12 22 50 38 0 1 2 1 2 1 H3K27AC Steady
0 17 10 3 3 0 12 19 10 34 22 25 2 3 1 3 1 1 H3K4ME2 Steady
7 8 8 0 0 6 8 10 19 10 12 11 1 2 2 2 2 1 H3K4ME1 Steady
3 0 2 0 6 23 40 50 10 19 28 0 1 2 4 7 1 0 H3K4ME3 Contract
3 4 0 2 5 16 50 40 8 12 0 11 0 1 2 2 1 1 H3K27AC Contract
0 2 3 1 7 50 16 23 6 0 8 7 0 0 2 2 1 1 H3K4ME2 Contract
1 1 2 0 15 7 5 6 0 3 3 3 0 0 2 1 1 1 H3K4ME1 Contract
10 13 11 50 0 1 2 0 0 3 3 3 3 3 1 3 1 3 H3K4ME1 Expand
15 24 50 11 2 3 0 2 8 10 0 9 2 1 2 1 1 2 H3K27AC Expand
36 50 24 13 1 2 4 0 8 17 14 0 7 10 1 7 1 5 H3K4ME3 Expand
50 36 15 10 1 0 3 3 7 0 8 8 4 2 2 2 1 3 H3K4ME2 Expand

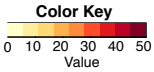
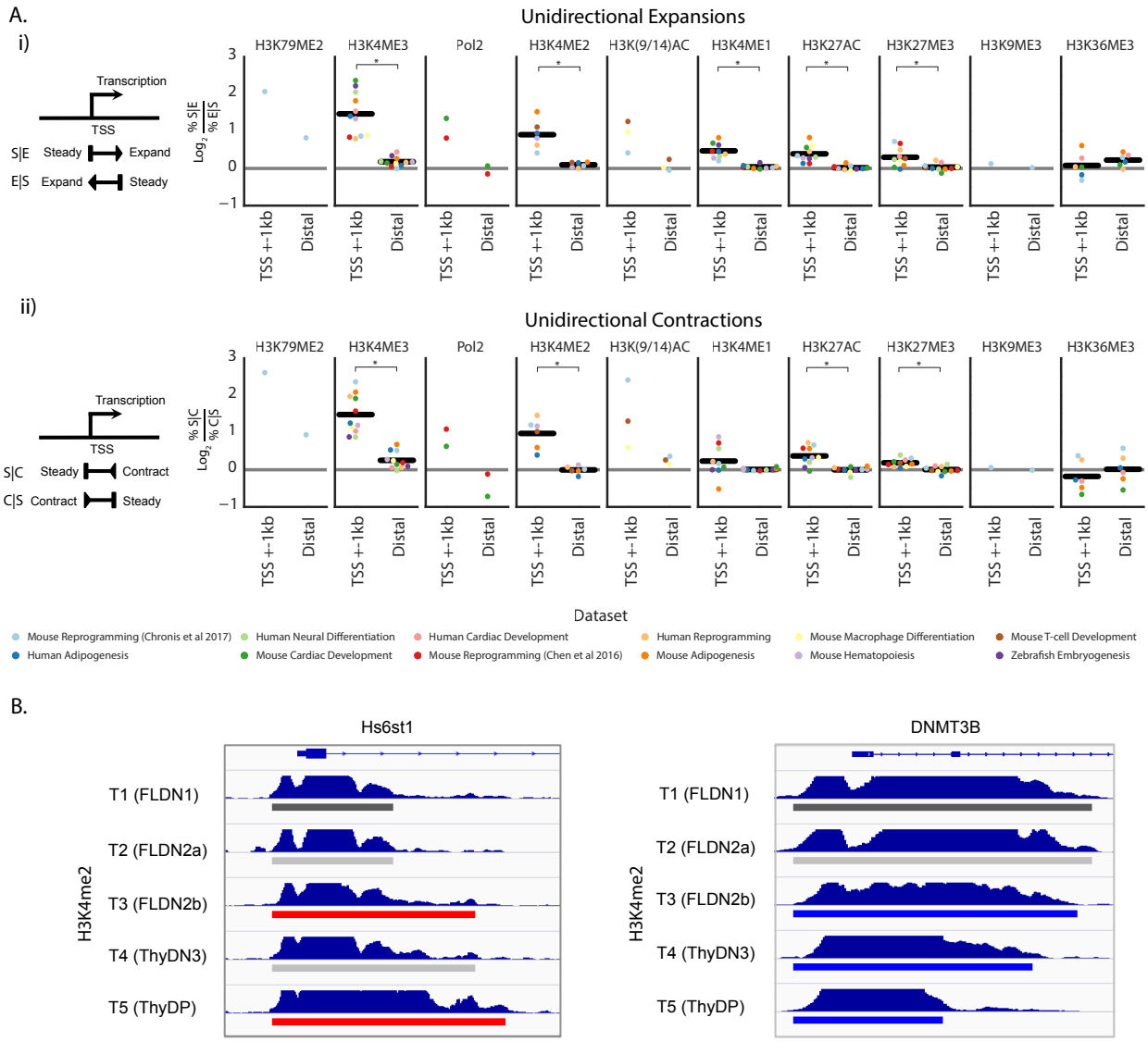

Color Key
0 10 20 30 40 50
Value

Figure 3.6: Spatial dynamics of multiple different HMs co-localize within a time course. Hierarchical clustering with optimal leaf ordering[62] of the geometric average fold enrichments taken across all time points of the overlap of every pair of predicted spatial dynamics for mapped HMs in **(A)** mouse adipogenesis[16] and **(B)** human stem cell reprogramming[21]. At each pair of time points, "Expand" and "Contract" dynamics are defined as all peaks that are predicted as either unidirectional or bidirectional expansions and contractions, respectively, whereas "Steady" dynamics are defined as all peaks that have predicted steady boundaries at both sides. Peaks with opposing dynamics at each side (i.e. E/C and C/E) were excluded from this analysis. In both datasets, expansions, contractions and steady peaks of H3K4me1, H3K4me2, H3K4me3 and H3K27ac tend to cluster together within each of the three classes, whereas spatial dynamics of H3K27me3 and H3K36me3 peaks tend to occupy different locations. All enrichments were capped at 50.

Figure 3.7

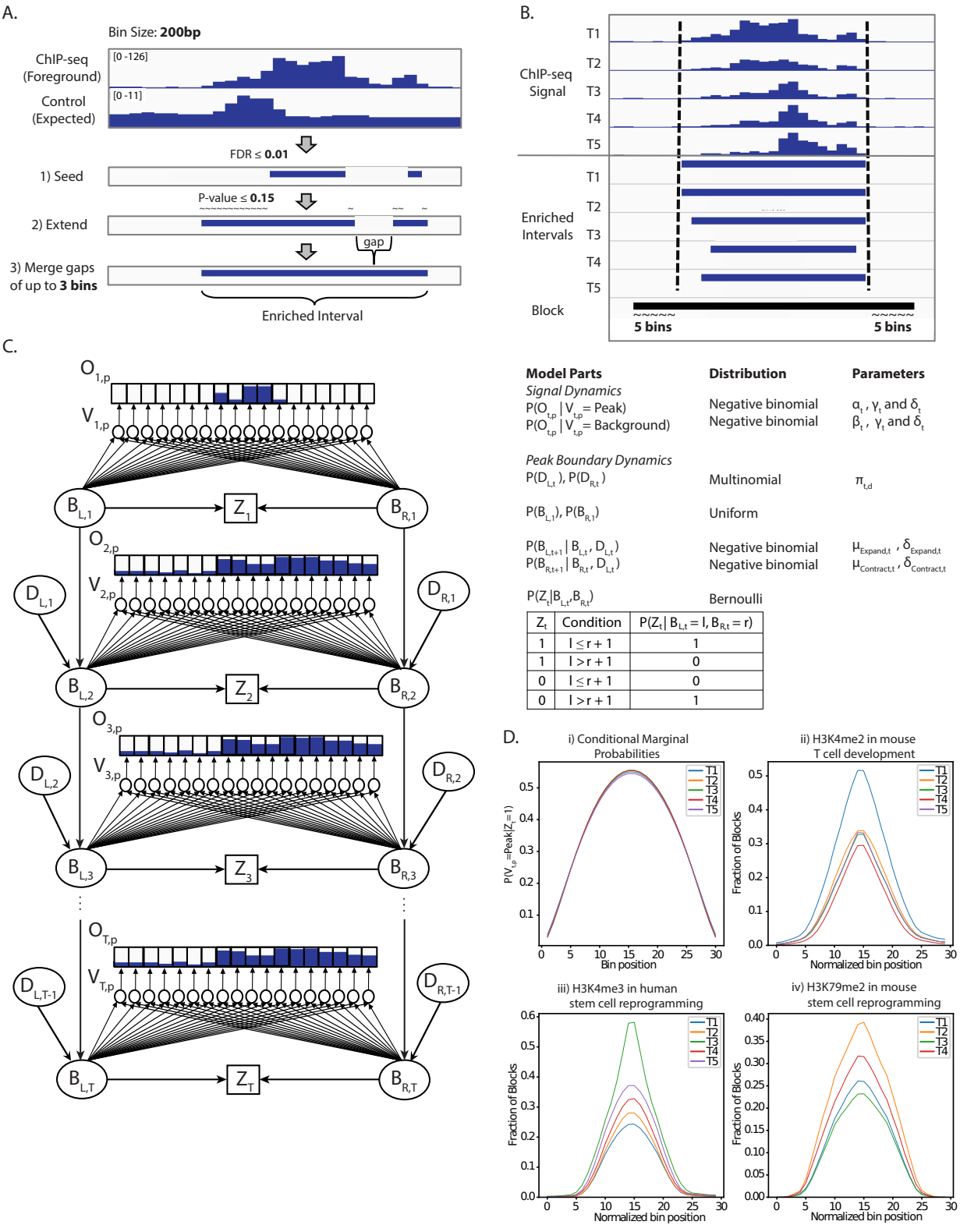

A.
i)
Unidirectional Expansions
Transcription
TSS
S|E Steady Expand
E|S Expand Steady
H3K79ME2
H3K4ME3
Pol2
H3K4ME2
H3K(9/14)AC
H3K4ME1
H3K27AC
H3K27ME3
H3K9ME3
H3K36ME3
Log2 % S|E / % E|S
TSS +-1kb
Distal
ii)
Unidirectional Contractions
Transcription
TSS
S|C Steady Contract
C|S Contract Steady
Log2 % S|C / % C|S
Dataset
Mouse Reprogramming (Chronis et al 2017)
Human Neural Differentiation
Human Cardiac Development
Human Reprogramming
Mouse Macrophage Differentiation
Mouse T-cell Development
Human Adipogenesis
Mouse Cardiac Development
Mouse Reprogramming (Chen et al 2016)
Mouse Adipogenesis
Mouse Hematopoiesis
Zebrafish Embryogenesis

B.
Hs6st1
DNMT3B
H3K4me2
T1 (FLDN1)
T2 (FLDN2a)
T3 (FLDN2b)
T4 (ThyDN3)
T5 (ThyDP)

Figure 3.7: **Direction of asymmetric dynamics correlates with direction of transcription.** **(A) (i)** Left panel shows a schematic representation of unidirectional expansions that expand in the same direction as transcription and in the opposite direction of transcription. The adjacent plots show, for each mark, the average $\log_2$ ratio between the fraction of unidirectional expansions that expand in the same directions as transcription of the nearest gene and the fraction of unidirectional expansions that expand in the opposite direction of transcription of the nearest gene for blocks that are within 1 kb of annotated TSSs and for more distal blocks. Positive values correspond to enrichment of unidirectional expansions in the same direction as transcription. For marks mapped in at least six time courses, a black line is plotted representing the average across all data sets and significant differences are denoted with asterisks based on a two-tailed Mann-Whitney test at a p-value threshold of 0.05. **(ii)** Left panel shows analogous schematic for unidirectional contractions. Likewise, adjacent plots show, for each mark, the average ratio between the fraction of unidirectional contractions that contract in the opposite direction of transcription of the nearest TSS and unidirectional contractions that contract in the same direction as transcription of the nearest TSS. **(B)** Left panel shows an example of unidirectional expansions between pairs of time points that expand in the same direction as transcription at the Hs6st1 gene of the H3K4me2 mark in the T cell development dataset[15]. Right panel shows an example of unidirectional contractions in the opposite direction of transcription at the DNMT3B gene. Time points 1, 2 and 3 correspond to in vitro differentiated T cell precursors, whereas time points 4 and 5 correspond to in vivo purified thymocytes. The

predicted ChromTime peaks colored by their boundary dynamics are shown under the signal track for each time point.

Figure 3.S1



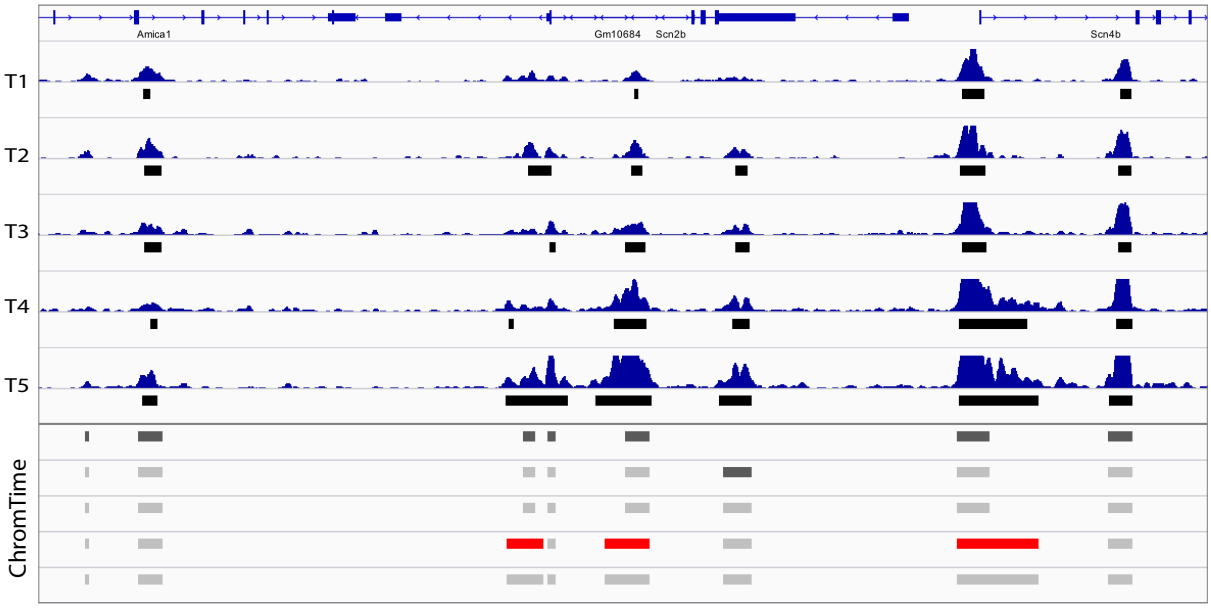
Model Parts	Distribution	Parameters
Signal Dynamics		
$P(O_{t,p} \mid V_{t,p} = \text{Peak})$	Negative binomial	α_t, γ_t and δ_t
$P(O_{t,p} \mid V_{t,p} = \text{Background})$	Negative binomial	β_t, γ_t and δ_t
Peak Boundary Dynamics		
$P(D_{L,t})$, $P(D_{R,t})$	Multinomial	$\pi_{t,d}$
$P(B_{L,1})$, $P(B_{R,1})$	Uniform	
$P(B_{L,t+1} \mid B_{L,t}, D_{L,t})$	Negative binomial	$\mu_{Expand,t}$, $\delta_{Expand,t}$
$P(B_{R,t+1} \mid B_{R,t}, D_{L,t})$	Negative binomial	$\mu_{Contract,t}$, $\delta_{Contract,t}$
$P(Z_t \mid B_{L,t}, B_{R,t})$	Bernoulli	

Z_t	Condition	$P(Z_t \mid B_{L,t} = l, B_{R,t} = r)$
1	$l \leq r + 1$	1
1	$l > r + 1$	0
0	$l \leq r + 1$	0
0	$l > r + 1$	1

Supplementary Figure 3.1: Details of the ChromTime method. (A) Calling enriched intervals at a given time point during the block finding stage of ChromTime. The number of ChIP-seq reads and its expectation are shown for each bin within a genomic region. Significantly enriched bins are called after correcting for multiple testing at FDR of 0.01 (seed step). Enriched bins are extended locally in both directions until a bin is found whose enrichment is not significant at a P-value of 0.15 (extend step). Extended bins are marked with "~" signs in the schematic. Continuous intervals are joined if they are separated by gaps of up to MIN_GAP bins (3 by default). **(B)** Enriched intervals from (**A**) that overlap across time are grouped in blocks. Blocks are further extended by BLOCK_EXTEND bins (5 by default, or up to the midpoint to their nearest neighbor block) up and down-stream from the left-most and the right-most position respectively of the enriched intervals within blocks. **(C)** Graphical model for the ChromTime mixture model with T time points. The learned model has T levels of $B_{L,t}$, $B_{R,t}$ $V_{t,p}$ and $O_{t,p}$ variables – one for each time point, and T-1 levels of $D_{L,t}$ and $D_{R,t}$ variables in between. Observed and latent variables are represented as boxes and circles, respectively. Example values for observed read counts are represented with blue bars inside each box. All conditional and prior probabilities from the model, their distribution type and their parameters estimated during the expectation maximization phase, are listed on the right. The probability mass function for $P(Z_t|B_{L,t}, B_{R,t})$ is given below all parameters. **(D)** (**i**) The conditional probabilities of the $V_{t,p}$ variables, which model the probability that bin at position p at time point t is annotated as Peak, conditioned on the requirement that the left

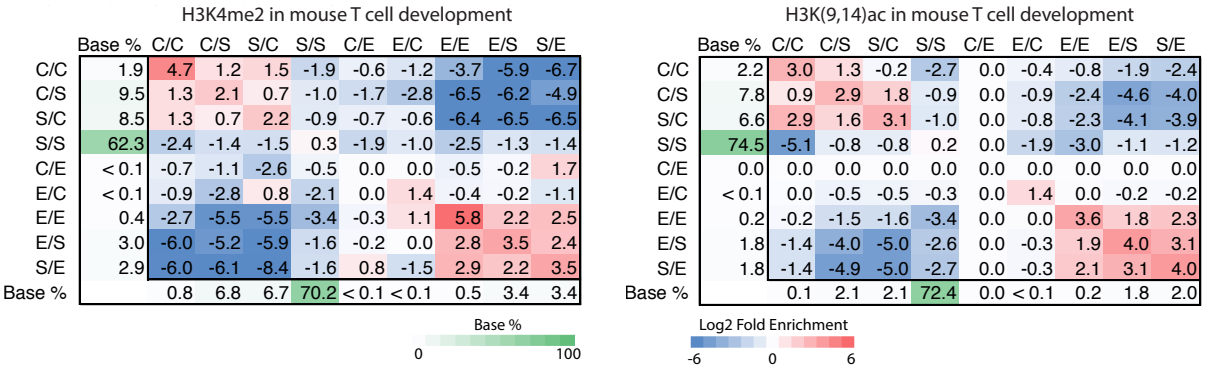
end boundary is placed before the right end boundary, $P(V_{t,p} = \text{Peak}|Z_t = 1)$, at each time point are plotted as a function of the bin position p. The values of $P(V_{t,p} = \text{Peak}|Z_t = 1)$ are shown for a block of length 30 bins (corresponding to default value of the MAX_BINS parameter, see **Supplementary Methods**) in a dataset with 5 time points computed after marginalizing out the observed read counts and with model parameters as initialized at the beginning of the EM procedure. Lines are largely overlapping and have their maximum at the center bin. (**ii**) The average fraction of blocks with significant bins at FDR 0.01 from the seed step during the block finding stage of ChromTime as a function of the bin position for H3K4me2 in mouse T cell development[15] for blocks of length up to 30 bins. Each line corresponds to one time point. X-axis corresponds to bin positions within blocks when blocks are rescaled uniformly to length of 30 bins. (**iii, iv**) as in (**ii**) for H3K4me3 in human stem cell reprogramming[21] and H3K79me2 in mouse stem cell reprogramming[24]. In all three cases, (**ii-iv**), the average fraction of blocks with significant bins has its maximum at the central positions of the blocks, which is consistent with the assumption of the method that the conditional probability $P(V_{t,p} = \text{Peak}|Z_t = 1)$ peaks at central positions.

Figure 3.S2



Supplementary Figure 3.2: Sample output from ChromTime with expanding peaks. Genome browser screenshot with sample output of ChromTime for H3K4me2 from the T cell development time course in mouse[15] with 5 time points at the Gm10684/Scn2b locus. Time points 1, 2 and 3 correspond to in vitro differentiated T cell precursors (FLDN1, FLDN2a, and FLDN2b), whereas time points 4 and 5 correspond to in vivo purified thymocytes (ThyDN3 and ThyDP). The input ChIP-seq signal and MACS2[35] peaks (black boxes under each signal track) are shown in the upper panel of the screenshot. The predicted ChromTime peaks colored by their boundary dynamics for each block at each time point are shown in the bottom panel. The first peak in each block is colored in dark grey. Each subsequent peak is colored with respect to the predicted dynamic relative to its previous time point. Peaks with steady boundaries on both sides are shown in light grey, and those with at least one expanding are shown in red. Not shown in the figure are contracting peaks, peaks at single time points and peaks with opposite dynamics (Expand on the left and Contract on the right, or vice versa), which would be colored in blue, orange and black, respectively. See **Fig 3.2** for examples of predicted contracting peaks.

Figure 3.S3


H3K4me2 in mouse T cell development
Base % C/C C/S S/C S/S C/E E/C E/E E/S S/E
C/C 1.9 4.7 1.2 1.5 -1.9 -0.6 -1.2 -3.7 -5.9 -6.7
C/S 9.5 1.3 2.1 0.7 -1.0 -1.7 -2.8 -6.5 -6.2 -4.9
S/C 8.5 1.3 0.7 2.2 -0.9 -0.7 -0.6 -6.4 -6.5 -6.5
S/S 62.3 -2.4 -1.4 -1.5 0.3 -1.9 -1.0 -2.5 -1.3 -1.4
C/E < 0.1 -0.7 -1.1 -2.6 -0.5 0.0 0.0 -0.5 -0.2 1.7
E/C < 0.1 -0.9 -2.8 0.8 -2.1 0.0 1.4 -0.4 -0.2 -1.1
E/E 0.4 -2.7 -5.5 -5.5 -3.4 -0.3 1.1 5.8 2.2 2.5
E/S 3.0 -6.0 -5.2 -5.9 -1.6 -0.2 0.0 2.8 3.5 2.4
S/E 2.9 -6.0 -6.1 -8.4 -1.6 0.8 -1.5 2.9 2.2 3.5
Base % 0.8 6.8 6.7 70.2 < 0.1 < 0.1 0.5 3.4 3.4
Base %
0
100
H3K(9,14)ac in mouse T cell development
Base % C/C C/S S/C S/S C/E E/C E/E E/S S/E
C/C 2.2 3.0 1.3 -0.2 -2.7 0.0 -0.4 -0.8 -1.9 -2.4
C/S 7.8 0.9 2.9 1.8 -0.9 0.0 -0.9 -2.4 -4.6 -4.0
S/C 6.6 2.9 1.6 3.1 -1.0 0.0 -0.8 -2.3 -4.1 -3.9
S/S 74.5 -5.1 -0.8 -0.8 0.2 0.0 -1.9 -3.0 -1.1 -1.2
C/E 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
E/C < 0.1 0.0 -0.5 -0.5 -0.3 0.0 1.4 0.0 -0.2 -0.2
E/E 0.2 -0.2 -1.5 -1.6 -3.4 0.0 0.0 3.6 1.8 2.3
E/S 1.8 -1.4 -4.0 -5.0 -2.6 0.0 -0.3 1.9 4.0 3.1
S/E 1.8 -1.4 -4.9 -5.0 -2.7 0.0 -0.3 2.1 3.1 4.0
Base % 0.1 2.1 2.1 72.4 0.0 < 0.1 0.2 1.8 2.0
Log2 Fold Enrichment
-6
0
6

Supplementary Figure 3.3: Reproducibility of ChromTime predictions across biological replicates. Average log2 fold enrichments (positive values, red) and depletions (negative values, blue) of base level overlap between peaks with predicted spatial dynamics in biological replicate experiments for H3K4me2 and H3K(9/14)ac in mouse T cell development[15]. The first column and the last row in each table show the average baseline fraction of bases covered by each dynamic out of all bases covered by ChromTime peaks. The averages are taken across all pairs of consecutive time points in each time course.

Figure 3.S4

A

i) H3K27ac in human stem cell reprogramming

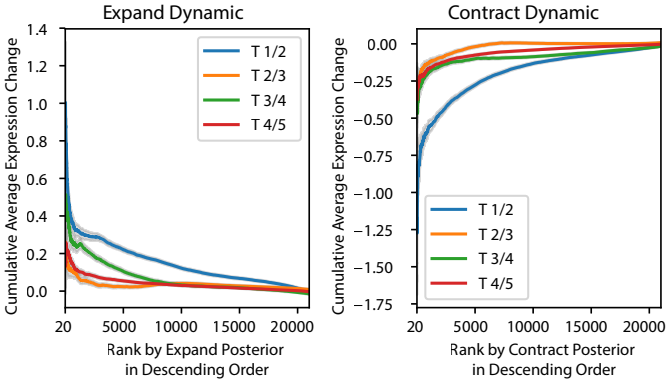
Dynamic	% all block bases	H1-IMR90 shared DHS	H1-IMR90 shared CEBPB	H1-IMR90 shared P300	H1-specific DHS	H1-specific CEBPB	H1-specific P300	H1 NANOG	H1 OCT4	IMR90-specific DHS	IMR90-specific CEBPB	IMR90-specific P300
T1-Tn Steady	2	3.6	3.7	1.4	0.8	1.2	1.3	1.6	1.3	0.6	1.2	1.7
Tx-Tn Expand	14	1.4	1.4	1	3	3.5	2	3.6	3.8	0.2	0.3	0.3
T1-Tx Contract	26	0.5	0.7	1	0.2	0.2	0.3	0.2	0.2	2.3	2	2.2
Base %		10	0.5	31	4.9	0.2	7.9	0.4	0.3	7.8	2.5	1.5

ii) H3K4me3 in human stem cell reprogramming

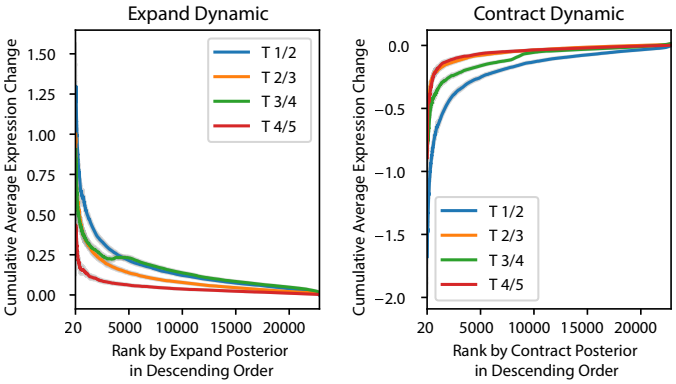
Dynamic	% all block bases	H1-IMR90 shared DHS	H1-IMR90 shared CEBPB	H1-IMR90 shared POL2	H1-IMR90 shared RAD21	H1-specific DHS	H1-specific CEBPB	H1-specific POL2	H1-specific RAD21	IMR90-specific CEBPB	IMR90-specific POL2	IMR90-specific RAD21	IMR90-specific DHS
T1-Tn Steady	26	1.5	1.5	1.8	1.3	0.6	0.9	1.2	0.9	1.2	1.2	1.2	0.8
Tx-Tn Expand	38	0.7	0.7	0.4	0.9	1.9	1.6	1.3	1.6	0.4	0.4	0.5	0.3
T1-Tx Contract	22	1.2	1.1	1.3	1	0.3	0.4	0.5	0.4	1.9	1.9	1.8	2.4
Base %		18	0.7	4.5	1.4	11	0.3	3	0.9	1.7	4.2	1	3.1

B

i) H3K27ac in human stem cell reprogramming



ii) H3K4me3 in human stem cell reprogramming



Supplementary Figure 3.4: **Changes in TF binding, DHS and gene expression at ChromTime predicted dynamics.** (**A**) Three sets of peaks were examined as defined in **Fig 3.3**: T1-Tn Steady, Tx-Tn Expand, and T1-Tx Contract. (**i**) For predicted H3K27ac dynamics in human stem cell reprogramming[21], fold enrichments are shown for DHS, CEBP, P300, OCT4 and NANOG. OCT4 and NANOG are key factors for maintaining pluripotency. IMR90 and H1 denote IMR90 human fetal lung fibroblast cells and H1 human embryonic stem cells, respectively, which resemble biologically the first and the last time point in this time course. **(ii)** As in **(i)** for DHS, CEBP, POL2 and RAD21 binding for different H3K4me3 dynamics during human stem cell reprogramming[21]. **(B)** For blocks with predicted peaks of **(i)** H3K27ac and **(ii)** H3K4me3 in human stem cell reprogramming[21] that overlap annotated TSSs, block boundaries were sorted in decreasing order by their posterior probability for Expand dynamic (left plots) and Contract dynamic (right plots) at each pair of consecutive time points (see **Supplementary Methods**). For each block boundary, gene expression differences were calculated between consecutive time points as the average difference of all TSSs overlapping the region spanning from the left-most to the right-most coordinate of peaks within the block, with positive values corresponding to increasing expression. For each posterior rank (X-axis) the plots show the cumulative average gene expression change (Y-axis). In both datasets, expanding boundaries associate with increase of gene expression and contracting boundaries associate with decrease of gene expression. Shaded regions correspond to 95% confidence intervals.

Figure 3.S5

A

ChromTime SINGLE

i) H3K4me2 in mouse T cell development

ii) H3K4me3 in human stem cell reprogramming

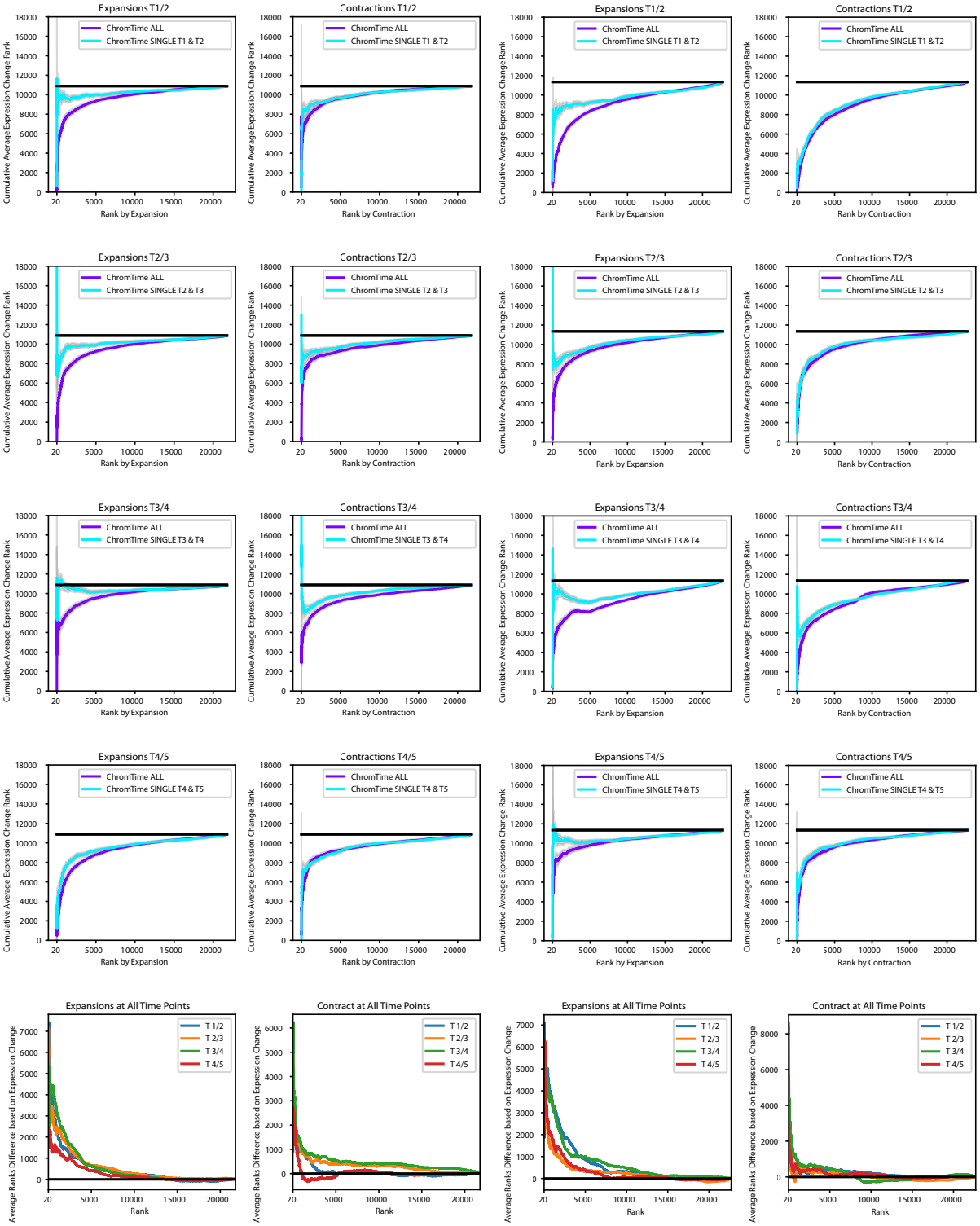


Figure 3.S5

B

MACS2

i) H3K4me2 in mouse T cell development

ii) H3K4me3 in human stem cell reprogramming

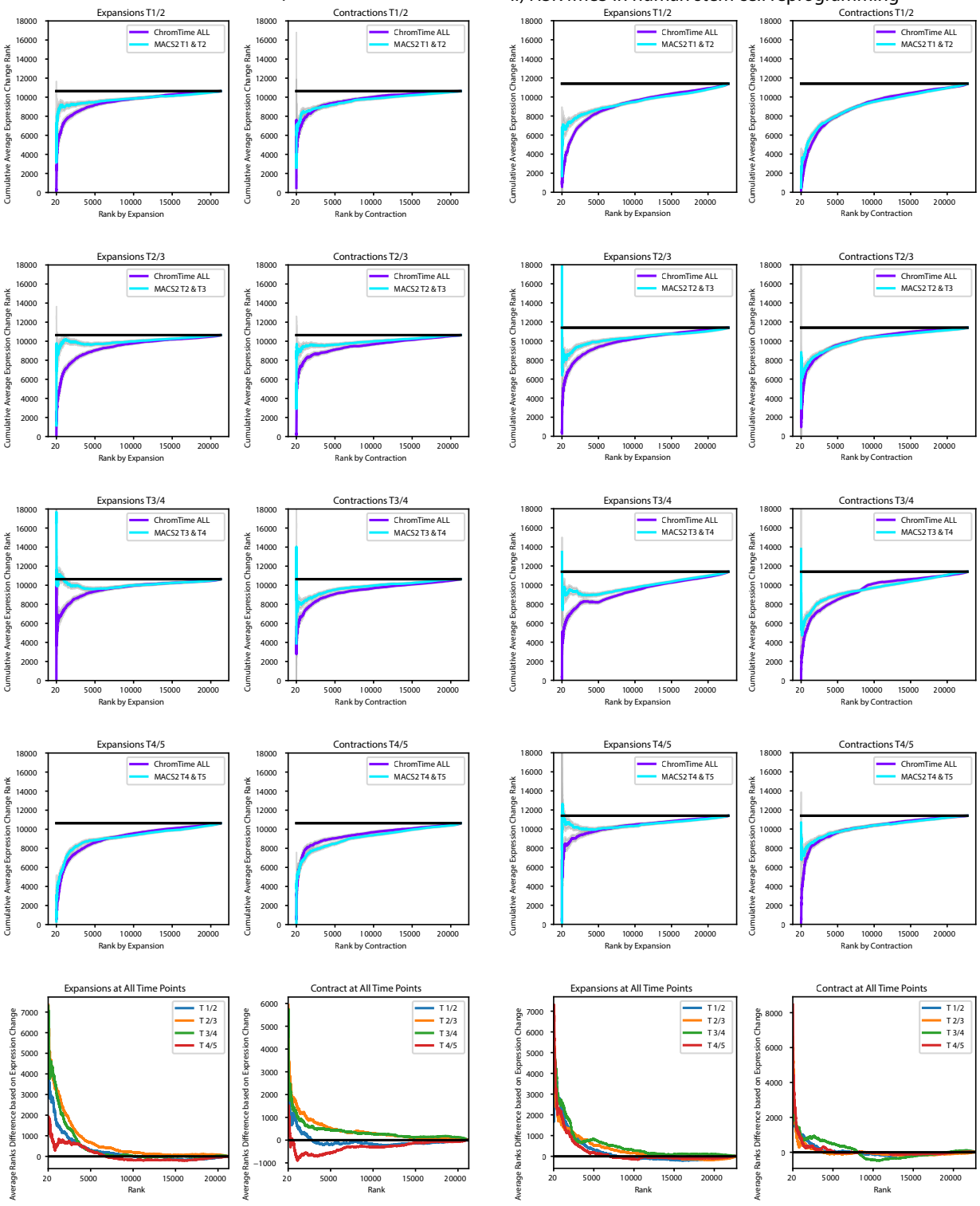


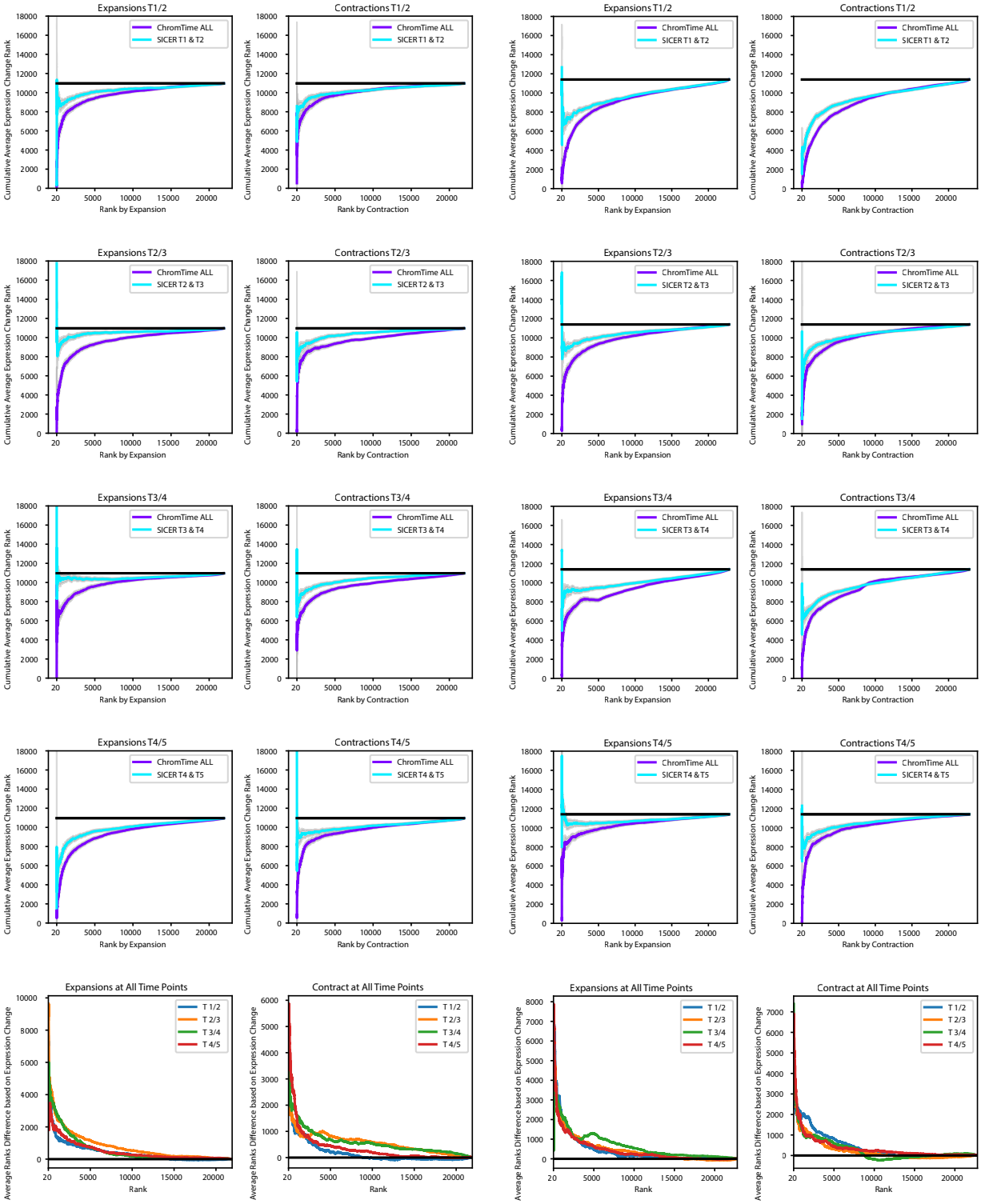
Figure 3.S5

C

SICER

i) H3K4me2 in mouse T cell development

ii) H3K4me3 in human stem cell reprogramming

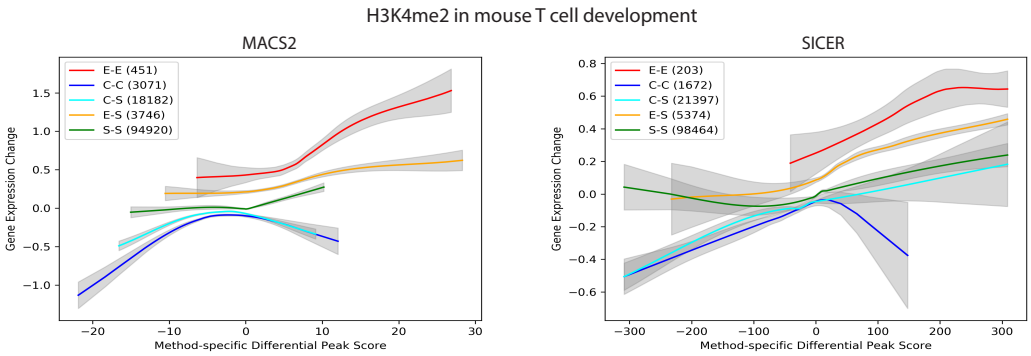


Supplementary Figure 3.5: Predicted spatial dynamics by ChromTime associate better with gene expression changes compared to boundary position changes of peaks called from single time points in isolation. This figure extends results presented in **Fig 3.4 (A) (i)** For H3K4me2 in mouse T cell development[15] ChromTime was applied once with data from all time points (ChromTime ALL), and once with single time points in isolation (ChromTime SINGLE; see **Supplementary Methods**). Time points 1, 2 and 3 correspond to T cell precursors, whereas 4 and 5 to purified thymocytes. Peaks called by both procedures overlapping annotated TSSs were analyzed for their relationship with gene expression changes. Left plots show comparisons of agreement with expression for expansions when applying ChromTime ALL and ChromTime SINGLE for the change between each pair of time points. Block boundaries were sorted in decreasing order of their Expand posterior probabilities from ChromTime ALL at each pair of consecutive time points and compared to sorting them in decreasing order of the difference of their positions in ChromTime SINGLE peaks with positive differences indicating peaks expanding with time. Each boundary was also ranked by gene expression difference of overlapping TSS in decreasing order with positive expression differences indicating gain with time. The cumulative average boundary rank of expression change (Y-axis) is shown for the boundary change ranking for ChromTime ALL and ChromTime SINGLE (X-axis). Low Y-values indicate stronger association with expression changes. Black line shows expected average expression change rank. Shaded regions indicate 95% confidence intervals. The last plot at the bottom shows differences

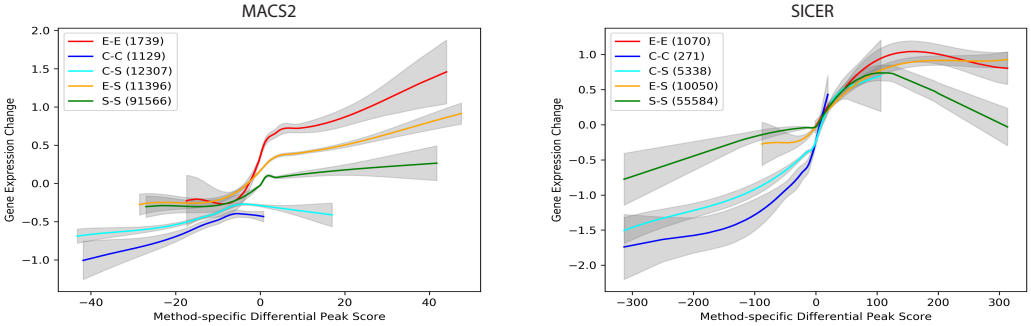
between ChromTime ALL and ChromTime SINGLE values shown in the individual plots above for each pair of time points. Positive values correspond to boundaries ranked based on ChromTime ALL better associating with gene expression changes than ChromTime SINGLE. Black lines show expected difference of zero between random rankings. Plots on the right show analogous comparisons for Contract posterior probabilities for ChromTime ALL, increasing order of the difference of boundary change positions for ChromTime SINGLE, and increasing order of expression changes. (**ii**) As in (**i**) for H3K4me3 in human stem cell reprogramming[21]. Time points correspond to hiF-T fibroblasts, fibroblasts at 5, 10 and 20 days after induction, and hIPSC-T cells. (**B**) as in (**A**) when ChromTime ALL posteriors are compared to boundary movements of MACS2[35] peaks called at single time points in isolation. (**C**) as in (**A**) when ChromTime ALL posteriors are compared to boundary movements of SICER peaks called at single time points in isolation.

Figure 3.S6

A.

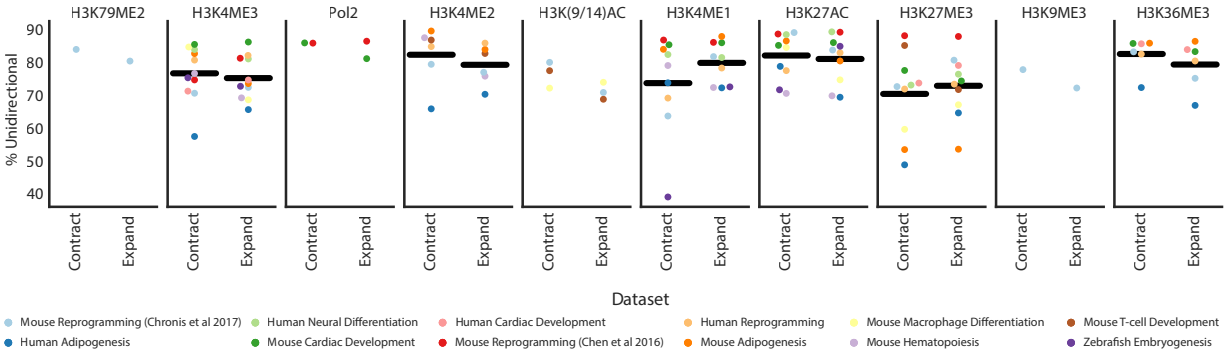


B.



Supplementary Figure 3.6: Spatial dynamics can contain additional information about gene expression changes beyond differential ChIP-seq peak calls. Similar to **Fig 3.5**, gene expression change is plotted as function of method-specific differential peak scores after loess smoothing from two differential peak calling methods, MACS2[35] (left) and SICER[37] (right) for each predicted ChromTime dynamic (see **Supplementary Methods**) for **(A)** H3K4me2 dynamics in T cell development in mouse[15]; and **(B)** in H3K4me3 dynamics in stem cell reprogramming in human[21]. Peaks of each type of dynamics were pooled from all time points for this analysis. Peaks with asymmetric dynamics E/S and S/E were pooled together in the "E-S" group. Similarly C/S and S/C peaks were pooled in the "C-S" group. In both systems, peaks with the same differential peak score can associate with different gene expression changes depending on the predicted spatial dynamic. Shaded regions represent 95% confidence intervals.

Figure 3.S7


H3K79ME2
H3K4ME3
Pol2
H3K4ME2
H3K(9/14)AC
H3K4ME1
H3K27AC
H3K27ME3
H3K9ME3
H3K36ME3
% Unidirectional
90
80
70
60
50
40
Contract
Expand
Dataset
Mouse Reprogramming (Chronis et al 2017)
Human Neural Differentiation
Human Cardiac Development
Human Reprogramming
Mouse Macrophage Differentiation
Mouse T-cell Development
Human Adipogenesis
Mouse Cardiac Development
Mouse Reprogramming (Chen et al 2016)
Mouse Adipogenesis
Mouse Hematopoiesis
Zebrafish Embryogenesis

Supplementary Figure 3.7: Average percentages of unidirectional expansions and contractions per pair of consecutive time points for each dataset. For each chromatin mark within each time course, the average percentage of unidirectional expansions and contractions out of all expansions and contractions, respectively, per pair of consecutive time points are calculated. For marks represented by six or more datasets, the average across all datasets is plotted as a black line.

Table 3.1: Datasets used for analysis with ChromTime

System	Chromatin Marks	Species	# Time Points	Reference
Adipogenesis	H3K4me2 H3K4me3 H3K27ac H3K4me1 H3K36me3 H3K27me3	Mouse Human	4	[16]
Blood Formation	H3K4me2 H3K4me3 H3K27ac H3K4me1	Mouse	5-7	[17]
Cardiac Development	H3K4me3 H3K27ac H3K4me1 H3K27me3 H3K36me3 Pol2	Mouse	4	[20]
Cardiac Development	H3K4me3 H3K27me3 H3K36me3	Human	5	[19]
Embryogenesis	H3K4me3 H3K27ac H3K4me1	Zebrafish	4	[25]
Macrophage Differentiation	H3K4me3 H3K9ac H3K27ac H3K27me3	Mouse	5	[18]
Neural Differentiation	H3K4me3 H3K27ac H3K27me3 H3K4me1	Human	5	[10]
Stem Cell Reprogramming	H3K4me2 H3K4me3 H3K27ac H3K4me1 H3K27me3 H3K36me3	Human	4-6	[21]
Stem Cell	H3K4me2	Mouse	4	[24]

Reprogramming	H3K4me3 H3K9ac H3K27ac H3K27me3 H3K36me3 H3K4me1 H3K79me2 H3K9me3			
Stem Cell Reprogramming	H3K4me3 H3K27ac H3K4me1 H3K27me3 Pol2	Mouse	9	[22]
T Cell Development	H3K4me2 H3K(9,14)ac H3K27me3	Mouse	5	[15]

REFERENCES

[1] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types.," *Nature*, vol. 473, no. 7345, pp. 43–9, May 2011.

[2] A. Barski *et al.*, "High-resolution profiling of histone methylations in the human genome.," *Cell*, vol. 129, no. 4, pp. 823–37, May 2007.

[3] T. S. Mikkelsen *et al.*, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.," *Nature*, vol. 448, no. 7153, pp. 553–60, Aug. 2007.

[4] ENCODE_Project_Consortium, "An integrated encyclopedia of DNA elements in the human genome.," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[5] Roadmap_Epigenomics_Consortium *et al.*, "Integrative analysis of 111 reference human

epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015.

[6] J. H. A. Martens and H. G. Stunnenberg, "BLUEPRINT: mapping human blood cell epigenomes.," *Haematologica*, vol. 98, no. 10, pp. 1487–9, Oct. 2013.

[7] F. D. Lay *et al.*, "Reprogramming of the human intestinal epigenome by surgical tissue transposition.," *Genome Res.*, vol. 24, no. 4, pp. 545–53, Apr. 2014.

[8] P. Fiziev *et al.*, "Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression.," *Cell Rep.*, vol. 19, no. 4, pp. 875–889, Apr. 2017.

[9] S. Mei *et al.*, "Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D658–D662, Jan. 2017.

[10] M. J. Ziller *et al.*, "Dissecting neural differentiation regulatory networks through epigenetic footprinting.," *Nature*, vol. 518, no. 7539, pp. 355–359, Dec. 2014.

[11] A. M. Tsankov *et al.*, "Transcription factor binding dynamics during human ES cell differentiation," *Nature*, vol. 518, no. 7539, pp. 344–349, Feb. 2015.

[12] K. K.-H. Farh *et al.*, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, no. 7539, pp. 337–343, Oct. 2014.

[13] E. Gjoneska *et al.*, "Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease," *Nature*, vol. 518, no. 7539, pp. 365–369, Feb. 2015.

[14] A. Gusev *et al.*, "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.," *Am. J. Hum. Genet.*, vol. 95, no. 5, pp. 535–52, Nov.

2014.

[15] J. A. Zhang, A. Mortazavi, B. A. Williams, B. J. Wold, and E. V Rothenberg, “Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity.,” *Cell*, vol. 149, no. 2, pp. 467–82, Apr. 2012.

[16] T. S. Mikkelsen *et al.*, “Comparative epigenomic analysis of murine and human adipogenesis.,” *Cell*, vol. 143, no. 1, pp. 156–69, Oct. 2010.

[17] D. Lara-Astiaso *et al.*, “Chromatin state dynamics during blood formation,” *Science (80-.).*, Aug. 2014.

[18] D. K. Goode *et al.*, “Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation,” *Dev. Cell*, Feb. 2016.

[19] S. L. Paige *et al.*, “A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development.,” *Cell*, vol. 151, no. 1, pp. 221–32, Sep. 2012.

[20] J. A. Wamstad *et al.*, “Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage.,” *Cell*, vol. 151, no. 1, pp. 206–20, Sep. 2012.

[21] D. Cacchiarelli *et al.*, “Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency,” *Cell*, vol. 162, no. 2, pp. 412–424, Jul. 2015.

[22] J. Chen *et al.*, “Hierarchical Oct4 Binding in Concert with Primed Epigenetic Rearrangements during Somatic Cell Reprogramming,” *Cell Rep.*, Jan. 2016.

[23] R. P. Koche *et al.*, “Reprogramming factor expression initiates widespread targeted

chromatin remodeling.," *Cell Stem Cell*, vol. 8, no. 1, pp. 96–105, Jan. 2011.

[24] C. Chronis *et al.*, "Cooperative Binding of Transcription Factors Orchestrates Reprogramming.," *Cell*, vol. 168, no. 3, p. 442–459.e20, Jan. 2017.

[25] O. Bogdanovic *et al.*, "Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis.," *Genome Res.*, vol. 22, no. 10, pp. 2043–53, Oct. 2012.

[26] N. Nègre *et al.*, "A cis-regulatory map of the Drosophila genome.," *Nature*, vol. 471, no. 7339, pp. 527–31, Mar. 2011.

[27] P. Yu *et al.*, "Spatiotemporal clustering of epigenome reveals rules of dynamic gene regulation.," *Genome Res.*, p. gr.144949.112-, Oct. 2012.

[28] P. Arnold *et al.*, "Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting.," *Genome Res.*, p. gr.142661.112-, Sep. 2012.

[29] N. Koike *et al.*, "Transcriptional architecture and chromatin landscape of the core circadian clock in mammals.," *Science*, vol. 338, no. 6105, pp. 349–54, Oct. 2012.

[30] R. Ostuni *et al.*, "Latent Enhancers Activated by Stimulation in Differentiated Cells," *Cell*, vol. 152, no. 1–2, pp. 157–71, Jan. 2013.

[31] S. K. Reilly *et al.*, "Evolutionary changes in promoter and enhancer activity during human corticogenesis," *Science (80-.).*, vol. 347, no. 6226, pp. 1155–1159, Mar. 2015.

[32] A. Weiner *et al.*, "High-Resolution Chromatin Dynamics during a Yeast Stress Response.," *Mol. Cell*, vol. 58, no. 2, pp. 371–386, Mar. 2015.

[33] J. Cotney *et al.*, "The evolution of lineage-specific regulatory activities in the human embryonic limb.," *Cell*, vol. 154, no. 1, pp. 185–96, Jul. 2013.

[34] J. Zhu *et al.*, "A time-series analysis of altered histone H3 acetylation and gene expression during the course of MMAIII-induced malignant transformation of urinary bladder cells," *Carcinogenesis*, vol. 38, no. 4, pp. 378–390, Apr. 2017.

[35] Y. Zhang *et al.*, "Model-based analysis of ChIP-Seq (MACS).," *Genome Biol.*, vol. 9, no. 9, p. R137, Jan. 2008.

[36] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.," *Mol. Cell*, vol. 38, no. 4, pp. 576–89, May 2010.

[37] S. Xu, S. Grullon, K. Ge, and W. Peng, "Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells.," *Methods Mol. Biol.*, vol. 1150, pp. 97–111, 2014.

[38] H. Xing *et al.*, "Genome-Wide Localization of Protein-DNA Binding and Histone Modification by a Bayesian Change-Point Method with ChIP-seq Data," *PLoS Comput. Biol.*, vol. 8, no. 7, p. e1002613, Jul. 2012.

[39] A. Harmanci, J. Rozowsky, and M. Gerstein, "MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.," *Genome Biol.*, vol. 15, no. 10, p. 474, 2014.

[40] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and

characterization.," *Nat. Methods*, vol. 9, no. 3, pp. 215–6, Mar. 2012.

[41] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation.," *Nat. Methods*, vol. 9, no. 5, pp. 473–6, May 2012.

[42] M. Allhoff, K. Seré, H. Chauvistré, Q. Lin, M. Zenke, and I. G. Costa, "Detecting differential peaks in ChIP-seq signals with ODIN.," *Bioinformatics*, vol. 30, no. 24, pp. 3467–75, Dec. 2014.

[43] H. Xu, C.-L. Wei, F. Lin, and W.-K. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data.," *Bioinformatics*, vol. 24, no. 20, pp. 2344–9, Oct. 2008.

[44] J. Biesinger, Y. Wang, and X. Xie, "Discovering and mapping chromatin states using a tree hidden Markov model," *BMC Bioinformatics*, vol. 14, no. Suppl 5, p. S4, 2013.

[45] D. Hnisz *et al.*, "Super-enhancers in the control of cell identity and disease.," *Cell*, vol. 155, no. 4, pp. 934–47, Nov. 2013.

[46] S. C. J. Parker *et al.*, "Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants," *PNAS*, p. 1317023110-, Oct. 2013.

[47] B. A. Benayoun *et al.*, "H3K4me3 breadth is linked to cell identity and transcriptional consistency.," *Cell*, vol. 158, no. 3, pp. 673–88, Jul. 2014.

[48] K. Chen *et al.*, "Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes.," *Nat. Genet.*, vol. 47, no. 10, pp.

1149–57, Oct. 2015.

[49] A. Dincer *et al.*, “Deciphering H3K4me3 broad domains associated with gene-regulatory networks and conserved epigenomic landscapes in the human brain,” *Transl. Psychiatry*, vol. 5, no. 11, p. e679, Nov. 2015.

[50] R. D. Hawkins *et al.*, “Distinct epigenomic landscapes of pluripotent and lineage-committed human cells.,” *Cell Stem Cell*, vol. 6, no. 5, pp. 479–91, May 2010.

[51] Z. Wang *et al.*, “Combinatorial patterns of histone acetylations and methylations in the human genome.,” *Nat. Genet.*, vol. 40, no. 7, pp. 897–903, Jul. 2008.

[52] H. H. Ng, F. Robert, R. A. Young, and K. Struhl, “Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity.,” *Mol. Cell*, vol. 11, no. 3, pp. 709–19, Mar. 2003.

[53] E. Smith, C. Lin, and A. Shilatifard, “The super elongation complex (SEC) and MLL in development and disease.,” *Genes Dev.*, vol. 25, no. 7, pp. 661–72, Apr. 2011.

[54] S. A. Lacadie, M. M. Ibrahim, S. A. Gokhale, and U. Ohler, “Divergent Transcription and Epigenetic Directionality of Human Promoters,” *FEBS J.*, Apr. 2016.

[55] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, Second edition. Cambridge University Press, 2013.

[56] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, “ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions,” *Genome Biol.*, vol. 12, no. 7, p. R67, 2011.

[57] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth. New York: Springer, 2002.

[58] Moré, B. Garbow, and K. Hillstrom, "User Guide for MINPACK-1," *ANL-80-74, Argonne Natl. Lab.*, 1980.

[59] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57. WileyRoyal Statistical Society, pp. 289–300, 1995.

[60] F. Jin *et al.*, "A high-resolution map of the three-dimensional chromatin interactome in human cells.," *Nature*, vol. 503, no. 7475, pp. 290–4, Nov. 2013.

[61] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

[62] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering.," *Bioinformatics*, vol. 17 Suppl 1, pp. S22-9, 2001.

[63] S. Steinhauser, N. Kurzawa, R. Eils, and C. Herrmann, "A comprehensive comparison of tools for differential ChIP-seq analysis.," *Brief. Bioinform.*, vol. 17, no. 6, pp. 953–966, Nov. 2016.

[64] E. G. and W. M. S. W. S. Cleveland, *Local regression models*. 1992.

[65] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001. [Online]. Available: http://www.scipy.org/. [Accessed: 26-Nov-2017].

CHAPTER 4

Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression

Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression

Petko Fiziev,[1,2,3,18] Kadir C. Akdemir,[4,18] John P. Miller,[4] Emily Z. Keung,[4] Neha S. Samant,[4] Sneha Sharma,[4] Christopher A. Natale,[5] Christopher J. Terranova,[4] Mayinuer Maitituoheti,[4] Samirkumar B. Amin,[4,6] Emmanuel Martinez-Ledesma,[4] Mayura Dhamdhere,[4] Jacob B. Axelrad,[4] Amiksha Shah,[4] Christine S. Cheng,[7,8] Harshad Mahadeshwar,[9] Sahil Seth,[9] Michelle C. Barton,[10] Alexei Protopopov,[9] Kenneth Y. Tsai,[11] Michael A. Davies,[12] Benjamin A. Garcia,[5] Ido Amit,[13] Lynda Chin,[4,9,14,*] Jason Ernst,[1,2,3,15,16,17,*] and Kunal Rai[4,19,*]

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA 90095, USA
[2]Department of Biological Chemistry, University of California, Los Angeles, CA 90095, USA
[3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, CA 90095, USA
[4]Division of Cancer Medicine, Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA
[5]Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104, USA
[6]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, USA
[7]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[8]Department of Biology, Boston University, Boston, MA 02215, USA
[9]Division of Cancer Medicine, Institute for Applied Cancer Science, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA
[10]Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[11]Division of Internal Medicine, Department of Dermatology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[12]Division of Cancer Medicine, Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[13]Weizmann Institute of Science, Rehovot 761001, Israel
[14]Institute for Health Transformation, The University of Texas System, Austin, TX 78701, USA
[15]Department of Computer Science, University of California, Los Angeles, CA 90095, USA
[16]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA
[17]Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA
[18]These authors contributed equally
[19]Lead Contact
*Correspondence: lchin@utsystem.edu (L.C.), jason.ernst@ucla.edu (J.E.), krai@mdanderson.org (K.R.)
http://dx.doi.org/10.1016/j.celrep.2017.03.078

SUMMARY

The extent and nature of epigenomic changes associated with melanoma progression is poorly understood. Through systematic epigenomic profiling of 35 epigenetic modifications and transcriptomic analysis, we define chromatin state changes associated with melanomagenesis by using a cell phenotypic model of non-tumorigenic and tumorigenic states. Computation of specific chromatin state transitions showed loss of histone acetylations and H3K4me2/3 on regulatory regions proximal to specific cancer-regulatory genes in important melanoma-driving cell signaling pathways. Importantly, such acetylation changes were also observed between benign nevi and malignant melanoma human tissues. Intriguingly, only a small fraction of chromatin state transitions correlated with expected changes in gene expression patterns. Restoration of acetylation levels on deacetylated loci by histone deacetylase (HDAC) inhibitors selectively blocked excessive proliferation in tumorigenic cells and human melanoma cells, suggesting functional roles of observed chromatin state transitions in driving hyperproliferative phenotype. Through these results, we define functionally relevant chromatin states associated with melanoma progression.

INTRODUCTION

Cancer cells acquire genetic and epigenetic alterations that increase fitness and drive progression through multiple steps of tumor evolution. However, the understanding of the roles of epigenetic alterations in cancer is lagging, in part due to challenges of generation of large-scale data for multiple epigenomes across tissues/time per individual and lack of ''germline normal'' equivalence.

The epigenome consists of an array of modifications, including DNA methylation and histone marks, which associate with dynamic changes in various cellular processes in response to stimuli. Although detailed profiles of specific epigenetic marks have been characterized in a number of normal tissues (ENCODE Project Consortium, 2012; Ernst et al., 2011; Kundaje et al., 2015) and some cancers including DNA methylation in human tumors, genome-wide profiles of multiple histone marks and

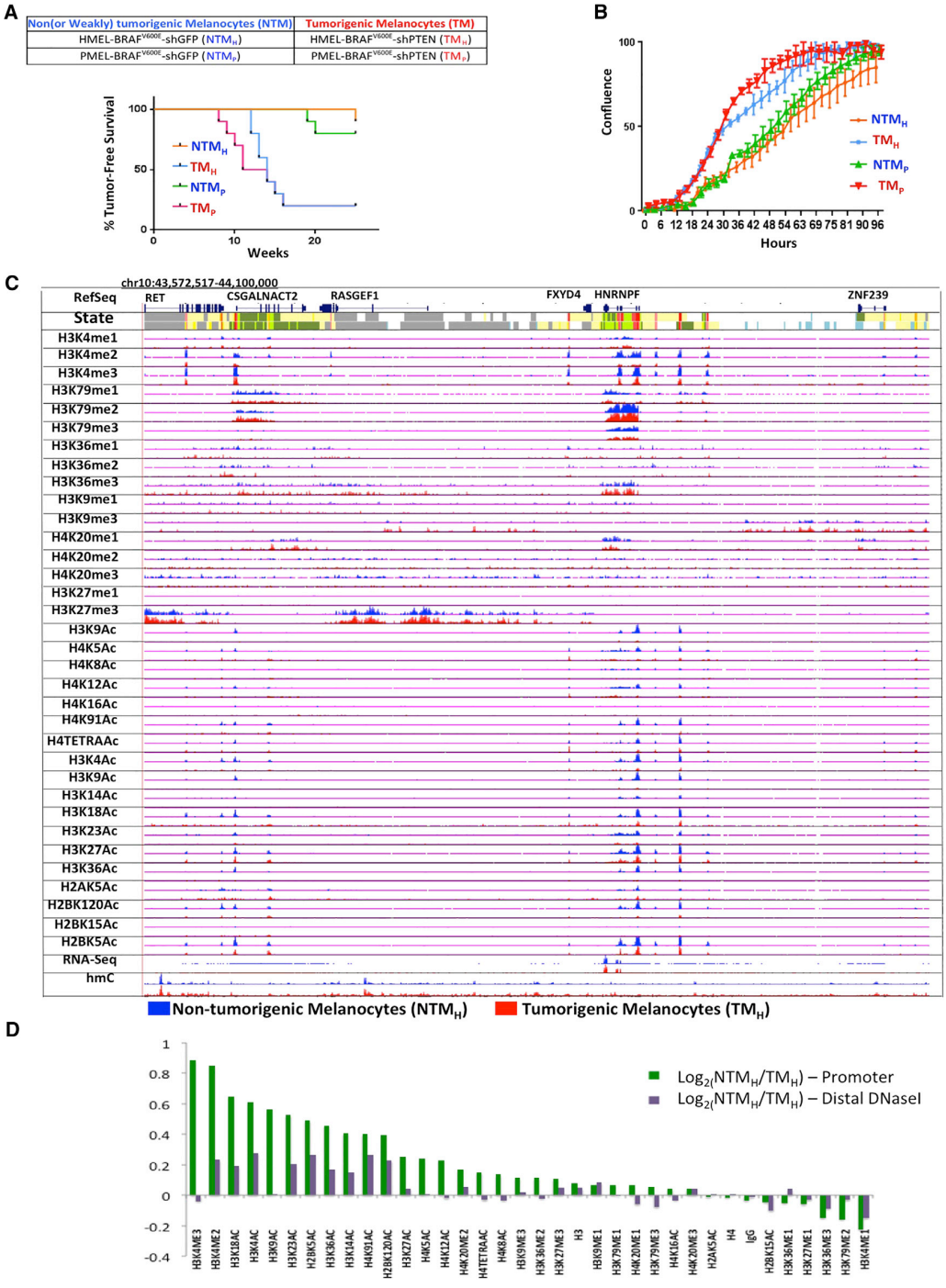


Figure 1. Cell-Line-Based Model of Melanoma Progression and Epigenome Profiling

(A) Brief description of the primary melanocyte-based model system that consists of two replicates of paired isogenic non (or weakly)-tumorigenic (NTM_H and NTM_P) and tumorigenic (TM_H and TM_P) cells. Kaplan-Meier curve showing tumor formation efficiency of NTM_H, NTM_P, TM_H, and TM_P cells. NTM_H and NTM_P cells display long latency, whereas TM_H and TM_P cells show shorter latency for tumor formation. Mantle-Cox test, $p = 0.0007$ for NTM_H versus TM_H and $p = 0.0016$ for NTM_P versus TM_P.

(legend continued on next page)

combinatorial chromatin states in cancer progression remain largely uncharacterized. Recently, enhancer aberrations were shown in diffuse large B cell lymphoma, colorectal, and gastric cancers by mapping H3K4me1/H3K27Ac (Akhtar-Zaidi et al., 2012; Chapuy et al., 2013; Muratani et al., 2014). Although these studies provide insight into the correlation of isolated epigenetic marks with cancer stage, more than 100 epigenetic modifications have been identified (Kouzarides, 2007; Tan et al., 2011) without clear understanding of their biological roles and interdependence. Furthermore, there is an even larger number of possible combinatorial patterns of these histone and DNA modifications, and it is these combinatorial patterns, not individual modifications, that dictate epigenetic states (Strahl and Allis, 2000). With the development of high-throughput chromatin immunoprecipitation (ChIP)-sequencing methodology (Garber et al., 2012), it is now possible to systematically and comprehensively profile many epigenetic marks with relative ease. Here, we profiled 35 epigenetic modifications in an isogenic cell system with distinct non-tumorigenic and tumorigenic phenotypes, and defined chromatin state alterations associated with transition to tumorigenesis. Further, we determined chromatin changes correlation with stable RNA-expression patterns, assessed their role in tumorigenesis, and established relevance to premalignant-to- malignant transition in human melanoma.

RESULTS

Systematic Epigenomic Profiling to Define Pro-tumorigenic Changes in Melanoma

To identify melanoma-associated changes, we leveraged a melanocyte cell model system with two characterized biological phenotypes, namely non (or weakly)-tumorigenic (NTM) and tumorigenic (TM) phenotypes (Figure 1A). The NTM phenotype is defined here as one poised to switch to the TM state but require additional cooperative driver alterations. Specifically, we used the well-characterized system of TERT-immortalized human primary foreskin melanocytes engineered with dominant-negative P53 and overexpression of $CDK4^{R24C}$ and $BRAF^{V600E}$ (Garraway et al., 2005). In two early passage ($n < 10$) clonal variants (HMEL and PMEL), isogenic cells were created with knockdown of either GFP (NTM) or PTEN (TM). NTM cells were confirmed to be inefficient in driving tumor formation (average tumor latency = 22 weeks) with low penetrance (10%–20%) in nude mice (Figure 1A). In comparison, TM cells expressing shRNA for PTEN (shPTEN) (~75% knockdown; Figure S1A) were able to drive tumorigenesis within 10–12 weeks with high penetrance (~80%) (Figure 1A). Similarly, TM cells showed aggressive behavior in proliferation, clonogenic, and invasion assays (Figures 1B and S1B–S1E). Hereafter, these two duplicate biological pairs are referred as: (1) NTM_H (HMEL-$BRAF^{V600E}$-shGFP [shRNA for GFP], NTM melanocytes) and TM_H (HMEL-$BRAF^{V600E}$-shPTEN, TM melanocytes); and (2) NTM_P (PMEL-$BRAF^{V600E}$-shGFP, NTM melanocytes) and TM_P (PMEL-$BRAF^{V600E}$-shPTEN, TM melanocytes). Unless specified otherwise, we have designated NTM_H and TM_H as the primary pair for discovery, and the NTM_P and TM_P as the pair for additional validation (Experimental Procedures). These two isogenic but phenotypically distinct melanocyte-derived cells provide a practical and relevant system for understanding epigenomic alterations that are associated with transition to tumorigenesis in melanoma.

To define the epigenome, we determined the status of 33 histone modifications, histone H3, H4, and IgG marks using a high-throughput ChIP-sequencing method (ChIP followed by next-generation sequencing) (Garber et al., 2012) (Figure 1C). We confirmed that the pairwise relationship between the occupancy patterns of different histone marks was consistent with known associations among the marks (Figure S3A). We also predicted combinatorial patterns of marks presented as "chromatin states" and annotated each cell type based on them (Ernst and Kellis, 2010). In addition, we profiled 5-methylcytosine using a 450K Illumina array and 5-hydroxymethylcytosine using 5-hydroxymethylcytosine DNA immunoprecipitation followed by sequencing (hMeDIP-seq). In total, we generated 3.08 billion uniquely aligned reads and produced 142 chromatin maps (Table S1). Furthermore, we performed RNA sequencing (RNA-seq) experiments to define the transcriptomes of these two biological states based on more than 1 billion uniquely aligned reads (Table S1; Supplemental Experimental Procedures).

Changes in Histone Marks during Transition of NTM Phenotypic State to TM Phenotypic State

We first compared the differences in occupancy of individual chromatin marks between cells in NTM and TM biological states. An analysis for relative enrichment of individual marks at all reference sequence database (RefSeq) annotated promoters revealed that multiple acetylation marks were consistently depleted in TM compared with NTM cells (Figures 1D and S1F) (Experimental Procedures). Similarly, a subset of acetylation marks was consistently depleted at a set of distal DNase I hypersensitive sites (in ENCODE melanocytes; Supplemental Experimental Procedures) in TM cells compared with NTM cells in both replicates (Figures 1D and S1F). We also identified H3K4me2/3 marks in promoter regions as higher in NTM biological state relative to TM biological state (Figures 1D and S1F). Interestingly, we did not observe any difference in global levels of histone acetylations or H3K4 methylations by mass-spectrometry-based quantitation (of the measurable marks) or by western blotting (Figures S1G–S1I). Overall, these data suggest that transition from NTM phenotype to TM phenotype is accompanied by switch to reduced acetylation and H3K4me2/3 methylation at nucleosomes in specific regions, but not at the global level.

(B) Proliferation curve showing differences in cell confluence (y axis) in NTM_H versus TM_H and NTM_P versus TM_P as a function of time (x axis).
(C) Normalized signal of all profiled chromatin marks, IgG control, and RNA-seq in an example region (chr10: 43,572,517–44,100,000) for NTM_H (blue) and TM_H (red) cells. Chromatin state tracks and gene annotations are also shown.
(D) Log_2 ratio between NTM_H and TM_H cells for the average signal strength of each chromatin mark in a window of 2 kb around annotated transcription start sites from RefSeq (green) and on DNase I hypersensitive sites from 'Melano' (purple) cell lines from ENCODE.
See also Figure S1 and Table S1.

To demonstrate human relevance of these observations from the cell model system, we assessed the status of representative marks in human benign nevi and melanoma samples representing pre-malignant (NTM) and malignant (TM) biological states, respectively. We first developed a validation strategy using a ChIP-string assay (Ram et al., 2011) and designed a 96-test probe ''codeset'' (Experimental Procedures; Table S2) that could be used to evaluate recapitulation of key epigenetic features observed in the isogenic cell models. This ChIP-string codeset was designed to measure enrichment of six selected histone marks (H3K27Ac, H2BK5Ac, H4K5Ac, H3K4me1, H3K27me3, and H3K4me3) that showed consistent differences between the NTM and TM cells and were part of three groups of epigenetic elements: promoters, enhancers, and Polycomb-repressed regions. To verify that the codeset for this ChIP-string assay performs as expected, we assayed them in both NTM_H and TM_H and showed that there was good correlation between the signal intensity from ChIP-string with ChIP-seq intensity ($R = 0.62–0.81$) for the tested probes (Figures S2A–S2F).

Chromatin immunoprecipitated DNA for two of the marks (H2BK5Ac and H4K5Ac) that provided sufficient yield after ChIP from nine melanoma tumors and four nevi samples (Table S3) was tested for enrichment on the designed codeset. As shown in Figures 2A and 2B, unsupervised hierarchical clustering analysis showed that nevi samples cluster with the NTM cells, whereas melanoma samples were more similar to the TM cells (Figures 2A and 2B). Average mark levels across all designated probes showed that enrichment for these marks in nevi and tumor samples was similar to those seen in NTM_H and TM_H cells, respectively (Figures 2C, 2D, and S2G–S2J). Further, principal component analysis (PCA) based on the enrichment of these chromatin marks at selected genomic regions showed that human nevi clustered tightly with NTM_H cells, but not the melanoma samples. Interestingly, although away from the NTM_H cells, the human melanoma samples do not cluster tightly around the TM_H cells, suggesting that the melanoma samples were more variable among themselves but collectively more different from the NTM cells and the nevi samples (Figures 2E and 2F). This may foretell that, like the genome, the epigenome is more heterogeneous and complex in tumors than benign neoplasms or normal cells, with the caveat that the assay was limited to 96 probes. In summary, a subset of chromatin changes observed in the NTM/TM melanocytic system displayed similar patterns as those observed in the benign/malignant human melanocytic lesions.

Chromatin State Changes between NTM and TM States

To assess whether and how the combinatorial and spatial patterns of chromatin marks (Ernst and Kellis, 2010; Ernst et al., 2011) may change with transition from NTM to TM biological states, we employed ChromHMM (Ernst and Kellis, 2012) to discover and define a set of chromatin states based on the 33 histone modification profiles, in addition to H3, H4, and IgG controls in both NTM and TM cells. In brief, by concatenating the chromatin maps for each mark, ChromHMM derived a common set of chromatin state definitions with cell-type-specific state assignments in both NTM and TM cells. A final model with 18 states was adopted for further downstream analyses (Figures 3A and S3B), based on the observations that it effectively achieved a balance of: (1) capturing important biological distinctions while (2) generating a manageable set of pairwise state combinations (Experimental Procedures). We found for some states of the model the assignments to be substantially recoverable by multiple different individual marks, found other states that required a specific mark in order to be able to recover their assignments, and found also some states that would need multiple marks to recover them (Supplemental Experimental Procedures; Table S4). By triangulating the defined chromatin states with known genome organization features (Figures S3C and S3D), we then grouped the 18 chromatin states by the following putative annotations: promoter regions (states 1–3), enhancers (states 4–6), transcribed enhancers (states 7–9), transcribed (states 10–12), active proximal (state 13), low signal (state 14), polycomb repressed (state 15), H3K9me3 heterochromatin (state 16), quiescent (state 17), and artifact/repetitive elements (state 18).

Within each of these groupings, enrichment of specific genomic structures was as expected (Figures S3C and S3D). For example, regions within 2 kb of RefSeq annotated transcription start sites (TSSs) were enriched specifically in chromatin states 1–3, corresponding to promoter regions and CpG islands in the genome. Consistently, 5-methylcytosine (5-MeC)-containing sites were weakly enriched in promoter-associated states, whereas 5-hydroxymethylcytosine (5-hMeC) showed complementary patterns to 5-MeC. RefSeq gene annotations were enriched in regions associated with chromatin states containing transcription marks H3K79me2/H3K79me1/H3K36me3, primarily states 1, 2, and 7–12. LaminB1-associated domain association (Guelen et al., 2008) was specifically seen in H3K9me3-enriched state 16. These enrichments support the biological relevance of this 18-state model and the annotation assigned to each state.

Next, we sought to define associations of chromatin states with NTM and TM cell phenotypes. To this end, we identified regions that transition to a different chromatin state in NTM and TM conditions. Calculation of coverage changes for each state in NTM_H and TM_H cells revealed that genome-wide occupancies of the most acetylated promoter state (State_1_TssA) and the most acetylated enhancer state (State_4_EnhA) were reduced from NTM_H to TM_H by 4.5- and 2.9-fold, respectively ($p < 1e{-}15$; Figure S4A). On the other hand, we noticed a 2.6-fold increase in the H3K9me3 repressive State_16_ReprK9me3 in TM_H cells when compared with NTM_H cells ($p < 1e{-}15$; Figure S4A).

To understand the global state transitions, we analyzed the pairwise state transition enrichments between NTM_H and TM_H relative to the same pair in the opposite direction, which controls for overall state similarity (Supplemental Experimental Procedures) (Figures 3B and S4B). This analysis revealed that, globally, there was a significant shift (transition) from strongly acetylated promoter and enhancer states to more weakly acetylated states accompanying the evolution from NTM to TM biological states (Figures 3B and S4B). For instance, the pairwise state transition from the strongly acetylated promoter state (State_1_TssA) in NTM_H to a more weakly acetylated promoter State_2_PromWkD or State_3_TssWkP in TM_H was 72 and 21 times, respectively, more enriched than observing a reverse transition from TM_H to NTM_H ($p < 1e{-}15$). Similarly, the pairwise state transition from

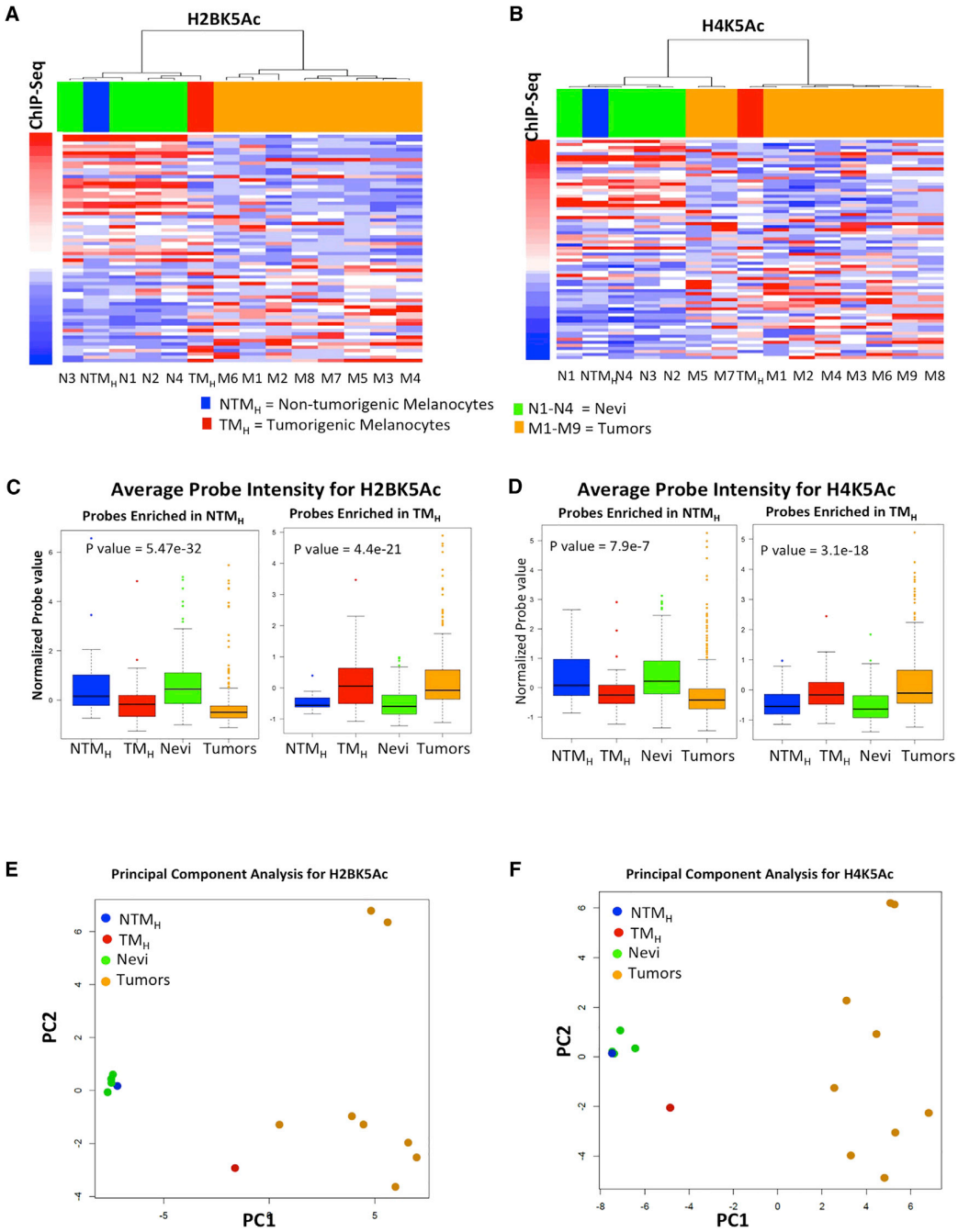


Figure 2. Chromatin Changes Are Reflected in Human Tumors

(A and B) Heatmap for H2BK5Ac (A) and H4K5Ac (B) showing enrichment in NTM$_H$, TM$_H$, four nevi samples (N1–N4), and up to nine melanoma tumor samples (M1–M9) as calculated by ChIP-string assay. Probes are ordered with increasing ChIP-seq signal in TM$_H$ cells. Columns are ordered based on hierarchical clustering.

(C and D) Boxplots showing average normalized intensity for ChIP-string probes across NTM$_H$, TM$_H$, nevi, and tumors (averaged over all enriched probes across all samples for nevi and tumors). Probes enriched in NTM$_H$ cells are on the left panel whereas those enriched in TM$_H$ cells are in the right panel.

(E and F) PCA plot for H2BK5Ac (E) and H4K5Ac (F) showing the relationship between NTM$_H$, TM$_H$, four nevi samples (N1–N4), and up to nine melanoma tumor samples (M1–M9) as calculated by ChIP-string assay.

*$p < 0.05$; **$p < 0.001$. See also Figure S2 and Tables S2 and S3.

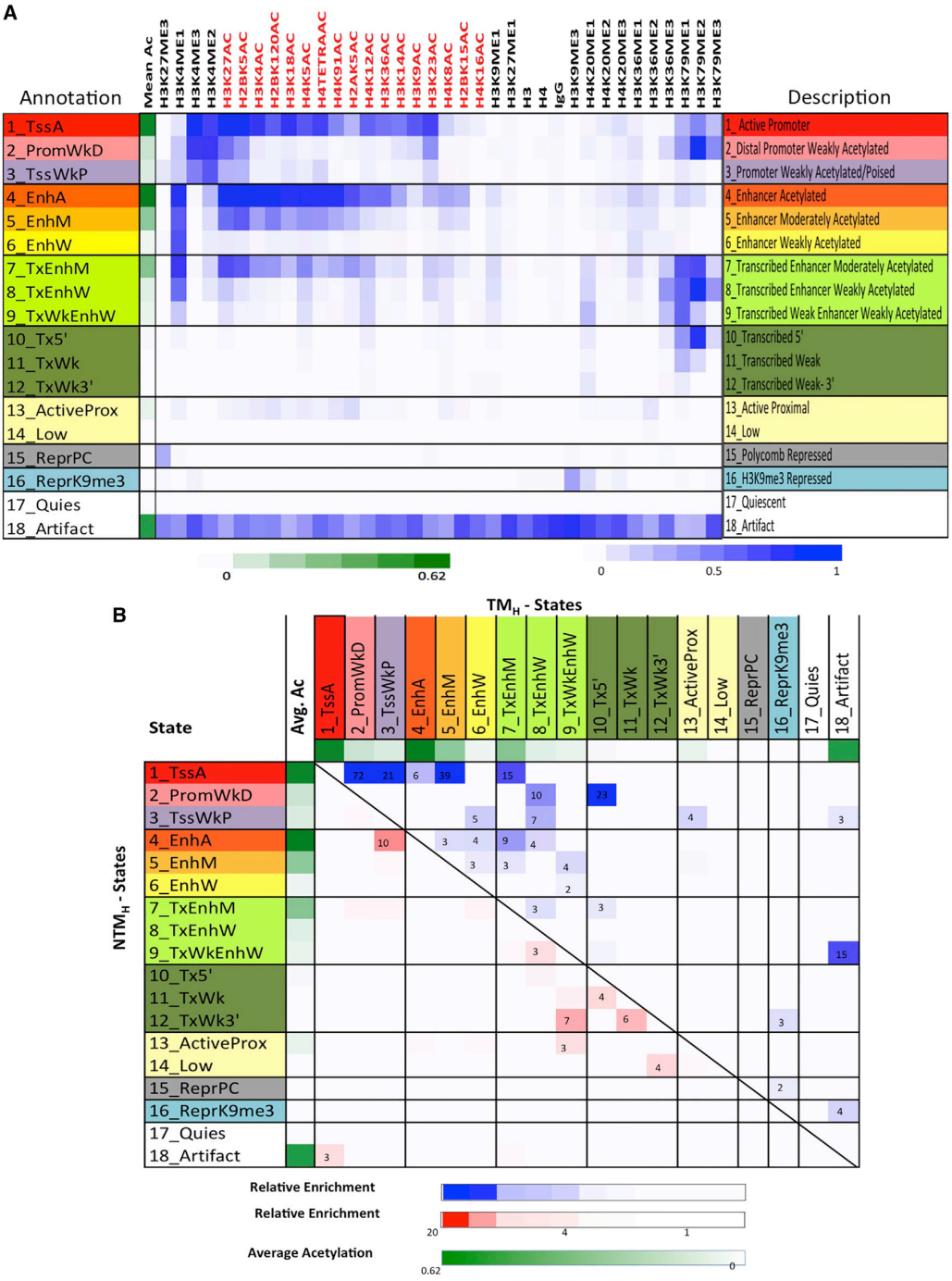


Figure 3. Chromatin State Predictions for NTM and TM Melanocytes

(A) Emission probabilities of the 18-state ChromHMM model (see Figure S3A for transition probabilities). Each row represents one chromatin state. First column gives state number and mnemonic, and last column gives the candidate state description. Second column indicates the intensity of mean acetylation from 0 (white) to 0.62 (green), which is the maximum mean acetylation across all states. Remaining columns each correspond to one chromatin mark with the intensity of the color in each cell reflecting the frequency of occurrence of that mark in the corresponding chromatin state on the scale from 0 (white) to 1 (blue).

(legend continued on next page)

the strongly acetylated enhancer state (State_4_EnhA) in NTM_H to a more weakly acetylated but transcribed enhancer state (State_7_TxEnhM) in TM_H was nine times more enriched, and to a weakly acetylated non-transcribed state 5_EnhM in TM_H was three times more enriched, than the reverse transition between the same pair of states from TM_H to NTM_H ($p < 1e{-}15$). The overall trends in chromatin state changes were similar after quantile normalization or downsampling to the same read depth (Figures S4C and S4D), as well as being replicated in NTM_P and TM_P cells (Figure S4E). Finally, we evaluated the correlation between mean histone acetylation and H3K4me2/3 changes on the same promoter and found them to be well correlated (Figures S5A and S5B). Together, these data suggest that during a NTM to TM phenotype switch, certain promoter and enhancer regions with specific chromatin states harboring higher acetylations and H3K4me2/3 transition to those with lower acetylation and H3K4me2/3 levels.

Chromatin State Changes Enrich on Genes Regulating Cancer-Associated Processes

To begin to explore the biological significance of prominent chromatin state transitions between NTM and TM biological states, we next performed pathway enrichment analysis [Gene Ontology (GO)] for genes associated with a specific pairwise transition in the promoter region (Supplemental Experimental Procedures) (Figure 4A; Table S5). We found specific enrichments for cancer-associated processes and metabolic processes. For example, promoters harboring highly acetylated State_1_TssA in NTM_H that transitioned to weakly acetylated State_2_PromWkD and State_3_TssWkP in TM_H were found preferentially by genes regulating cell cycle and apoptosis, as well as various cellular metabolic processes and protein modifications. These included important melanoma cell cycle inhibitors *CDKN1B* and *CDKN2A* (Bennett, 2016), as well as melanoma pro-apoptotic genes *BAD* and *APAF1* (Campioni et al., 2005; Sheridan et al., 2008) (Figures 4B, S5C, and S5D), suggesting increased proliferation and reduced apoptosis in TM cells. Interestingly, homophilic cell adhesion genes such as proto-cadherins were associated with the transition from a weakly poised promoter (State_3_TssWkP) or a quiescent state (State_17_Quies) to a more strongly H3K9me3-associated chromatin state (State_16_ReprK9me3) in TM cells (Figures 4A and 4C). This suggests that upon acquisition of a TM fate, genes promoting cell adhesion acquire a repressive chromatin signature, possibly contributing to loss of cell-cell adhesion in cancer.

Further, a pathway enrichment analysis of genes displaying chromatin state transition revealed additional association of cell signaling pathways with chromatin states (Figure 4D; Table S6; Supplemental Experimental Procedures). We found significant enrichments of important melanoma cell signaling pathways such as phosphatidylinositol 3-kinase (PI3K), interferon (IFN) γ-, LKB1-, TRAIL-, and platelet-derived growth factor (PDGF)-mediated signaling (Paluncic et al., 2016) in promoters transitioning from State_1_TssA to either State_2_PromWkD or State_3_TssWkP during NTM to TM phenotype switch (Figure 4D; Table S6).

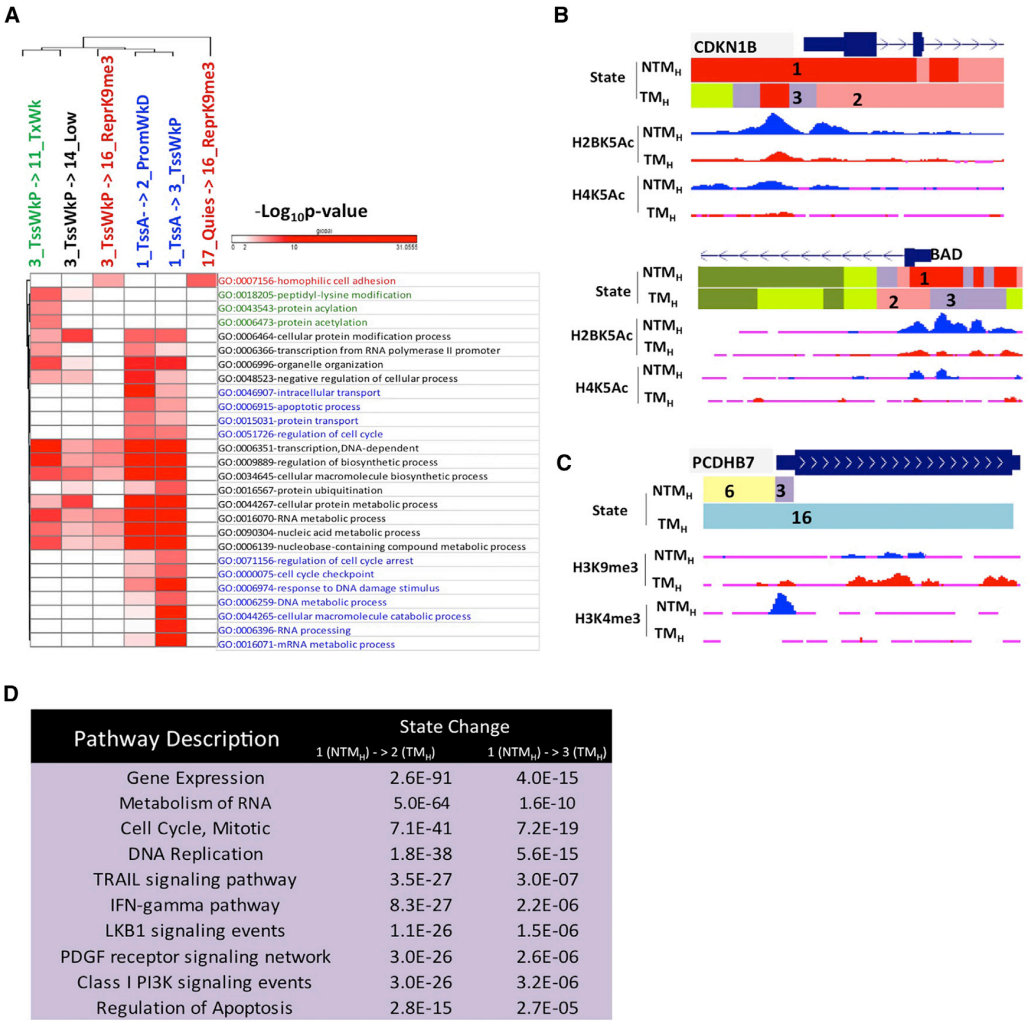
Similar analysis of enhancer regions in two of the most significant chromatin state transitions, State_4_EnhA in NTM_H to State_7_TxEnhM or State_5_EnhM in TM_H cells, showed enrichment of the nearest genes in important melanoma cell signaling events such as integrin, transforming growth factor (TGF) β, and mitogen-activated protein kinase (MAPK) signaling (Busse and Keilholz, 2011; Pinon and Wehrle-Haller, 2011; Sullivan and Flaherty, 2013) (Figures S5E and S5F; Table S6; Supplemental Experimental Procedures). Overall, these data suggest that chromatin state changes during transition to TM phenotype occur on promoters and enhancers of a large number of genes that are known to regulate relevant cancer processes such as proliferation, apoptosis, and adhesion.

Complex Relationship between Gene Expression and Chromatin States

To understand relationships between chromatin state and gene expression, we integrated RNA-seq profiles of NTM and TM cells with the chromatin states individually in each cell type. As expected, promoters of highly expressed genes (fragments per kilobase per million mapped reads [FPKM] > 5) displayed enrichment in chromatin states 1 and 2 that are marked by H3K4me3 and acetylations (Figure S3E). These promoters were depleted in repressed states 15–16, whereas their gene bodies were enriched in states 7–12 with transcription marks (H3K79me2/3 and H3K36me3) (Figure S3E). Furthermore, while comparing between different states within the same cell type, differences in acetylation content were associated with gene expression differences, particularly within the enhancer state group (Figure S3E).

Analysis of significant changes between NTM and TM states in the expression levels of known RefSeq transcripts revealed that changes in gene expression are bidirectional with similar numbers of genes upregulated or downregulated (Figure 5A). Next, we sought to determine associations of chromatin state transitions with gene expression changed between NTM and TM cells. To this end, we calculated the relative enrichment of all possible chromatin state transitions at the promoters +2 kb and −2 kb from TSS of genes that were upregulated, downregulated, or unchanged (Supplemental Experimental Procedures) (Figures 5B and S6A–S6F) with an expectation that genes downregulated in TM_H cells in comparison with NTM_H cells would show global switch from active chromatin states to repressed/low states on their promoters and vice versa. However, we observed overall similar patterns of chromatin state enrichments and with few exceptions did not see substantial chromatin state switches in upregulated, downregulated, or unchanged genes between NTM and TM cells (Figures 5B and S6A–S6F). This observation suggests that regulation of steady-state levels of

(B) Heatmap showing fold enrichment of transitions of chromatin states in NTM_H to TM_H cells controlling for the overall state size and similarity (Supplemental Experimental Procedures). The color intensities above (below) the main diagonal range from white (relative enrichment < 1) to blue (red) (relative enrichment > 20), thus indicating chromatin state transitions that lose acetylation marks from NTM_H to TM_H within the same category are more enriched compared with the reverse chromatin state transition (i.e., from TM_H to NTM_H) and the lack of those that gain acetylations.
See also Figures S3 and S4 and Table S4.



Pathway Description	State Change 1 (NTM_H) - > 2 (TM_H)	State Change 1 (NTM_H) - > 3 (TM_H)
Gene Expression	2.6E-91	4.0E-15
Metabolism of RNA	5.0E-64	1.6E-10
Cell Cycle, Mitotic	7.1E-41	7.2E-19
DNA Replication	1.8E-38	5.6E-15
TRAIL signaling pathway	3.5E-27	3.0E-07
IFN-gamma pathway	8.3E-27	2.2E-06
LKB1 signaling events	1.1E-26	1.5E-06
PDGF receptor signaling network	3.0E-26	2.6E-06
Class I PI3K signaling events	3.0E-26	3.2E-06
Regulation of Apoptosis	2.8E-15	2.7E-05

Figure 4. Chromatin State Changes during Transition to TM State Mark Specific Cancer Pathways

(A) Heatmap showing $-\log_{10}$(p value) for top GO terms enriched in specific promoter state transitions between NTM_H and TM_H cells.

(B) UCSC genome browser view of chromatin states as well as selected histone acetylation profiles (H2BK5Ac and H4K5Ac) for loci encompassing cell cycle regulator *CDKN1B* and apoptotic gene *BAD*, which showed loss from NTM_H to TM_H cells.

(C) UCSC genome browser view of chromatin states as well as selected histone mark H3K9me3 and H3K4me3 profiles for loci encompassing pro-adhesion *PCDHB7* in NTM_H and TM_H.

(D) Top 10 most significant pathways (pathway commons) associated with promoters displaying state transitions from State 1_TssA in NTM_H to states 2_PromWkD and 3_TssWkP in TM_H cells.

See also Figure S5 and Tables S5 and S6.

RNA transcripts in this system involves more than chromatin modification at the promoters.

Because changes in acetylation marks were prominent between NTM and TM cells, we quantitatively compared aggregate acetylation changes in promoter regions of all 17 acetylation marks profiled with gene expression changes. Here, we first identified a set of promoters for which the change in their average acetylation signal over all acetylation marks was statistically significant at a false discovery rate (FDR) of 1% (Supplemental Experimental Procedures) when comparing between NTM_H and TM_H (Figure 5C). In stark contrast to gene expression patterns, changes in acetylation levels were unidirectional, with most changed regions having a lower average acetylation in TM cells compared with NTM cells (Figures 5C, 5D, and S6G).

Given these differences between patterns of acetylation changes and stable gene expression changes, we explored the possibility that different subsets of genes are responding differently to acetylation changes on their promoters. To test

this directly, we systematically overlapped gene expression changes with changes in promoter acetylation to define nine possible subsets (Figures 5D and S6H; see Supplemental Experimental Procedures for definitions) and performed enrichment analysis of the genes to determine whether different cellular processes were enriched in different subsets (Figure 5E; Table S7). Indeed, we found that prominent cancer-relevant pathways were enriched among genes that were downregulated and showed loss of acetylation (LossAc_LossExp), including EGFR pathway targets, p53-regulated genes, and caspase-mediated apoptotic signaling genes (Figure 5E; Table S7). These genes are epigenetically regulated by specific chromatin alterations and may regulate melanoma growth. For example, *DUSP5* showed loss of acetylation on its promoter and concomitant downregulation in TM cells in comparison to NTM cells (Figure 5F). DUSP5 is a negative regulator of the MAPK pathway (Caunt and Keyse, 2013) that functions by reducing nuclear phosphorylated ERK; therefore, its loss can potentially provide positive feedback to MAPK signaling enhancing p-ERK levels. To test this hypothesis, we reduced DUSP5 levels in NTM_H cells by generating stable cell lines bearing two specific short hairpin RNAs (shRNAs) (Figure 5G) and tested p-ERK levels as well as proliferative capacity. Indeed, NTM_H cells bearing DUSP5 shRNAs showed increased p-ERK levels (Figure 5H) and proliferated faster in comparison with NTM_H cells harboring control shRNA (Figure 5I).

On the other hand, genes in certain signaling pathways such as aurora kinase and PLK signaling showed only acetylation loss in the promoters without expression change (LossAc_ConstExp group) (Figure 5E). For example, ATM, a critical mediator of the DNA damage checkpoint pathway (Shiloh and Ziv, 2013), was found to follow this pattern (Figure 5F). Possible deacetylation changes without accompanying alterations in steady-state RNA levels may reflect multi-level control of transcription requiring either upstream regulation such as promoter-enhancer interactions or downstream regulation by an additional event such as microRNA (miRNA)-mediated post-transcriptional regulation. Conversely, genes with differential expression but without acetylation changes (ConstAc_GainExp and ConstAc_LossExp) were enriched for various transport pathways, TCA cycle, and translation, raising the possibility that these pathways are less likely to be regulated on the epigenomic level through promoter acetylation during tumorigenesis.

Taken together, these integrative analyses showed that some well-characterized cancer signaling pathways exhibit promoter acetylation-correlated expression regulation, suggesting that these pathways can be regulated by epigenomic modifications. At the same time, it is intriguing that changes in expression of some pathway genes, such as those related to metabolism or transport, do not appear to show correlation with changes in their promoter acetylation.

Loss of CREB-Binding Protein Creates Pro-TM Chromatin Patterns and Accelerates TM Properties

Next, we asked whether decreased expression of a histone acetyltransferase or increased expression of a histone deacetylase (HDAC) in this system might be responsible for observed loss of acetylation peaks. We checked the expression differences of 32 known histone acetyltransferases and deacetylases between the NTM and TM models (Sammons et al., 2016) (Figure 6A). The expression of CREB-binding protein (CBP) acetyltransferase showed consistent patterns to observed acetylation loss in that its expression was downregulated >2-fold in both variants of TM cells compared with their counterpart NTM cells (Figures 6A and 6B). We knocked down CBP mRNA in NTM_H cells using two specific shRNAs (Figure 6C) and checked the levels of H2BK5Ac and H4K5Ac using the ChIP-string codeset that was utilized to validate acetylation changes in nevi/tumor samples. Indeed, stable cells harboring CBP shRNAs showed similar patterns of these two acetylations as seen in TM cells compared with NTM cells (Figures 6D–6G). Consistently, the NTM_H cells harboring CBP shRNAs showed significantly enhanced tumorigenesis compared with control shRNA-bearing cells (Figure 6H). These data argue that a TM phenotype might be associated with loss of acetylation irrespective of whether TM behavior was achieved by PTEN loss or by CBP loss in the same background (NTM background of *TERT*/*p53DD*/*CDK4*^{R24C}). This hypothesis was further supported by our observations that $NRAS^{G12D}$ overexpression in *TERT*/*p53*^{DD}/*CDK4*^{R24C} immortalized melanocytes created similar H4K5Ac and H2BK5Ac acetylation patterns to TM_H cells (overexpression of $BRAF^{V600E}$ along with shPTEN) (Figures 6D–G). NRAS has been previously shown to activate the MAPK pathway (result of BRAF activation) and PI3K pathway (result of PTEN loss) (Chudnovsky et al., 2005), thereby mimicking cellular phenotype of TM cells. Together, our data argue for a relatively uniform acetylation pattern of cells with TM behavior.

HDAC Inhibitors Specifically Reduce Proliferative Rate in TM Cells

Next, we sought to determine whether chromatin state changes seen during transition to tumorigenesis impart proliferative advantage to TM cells. Because loss of histone acetylation peaks was a consistent feature of all major chromatin state alterations, we tested the contribution of widespread acetylation loss to cell proliferation. Because steady-state acetylation loss seen in TM cells could be an outcome of aberrations in histone acetylation-deacetylation cycle in favor of accelerated deacetylation or reduced acetylation, we sought to alter the former by inhibition of HDACs, the primary driver enzymes of histone deacetylation in mammalian cells. We tested whether treatment of TM cells with HDAC inhibitors alters their acetylation levels toward those in NTM cells. Indeed, measurement of H2BK5Ac, H4K5Ac, and H3K27Ac levels in TM_H cells treated with vehicle or two different HDAC inhibitors (vorinostat and entinostat) by ChIP-string revealed that the levels of the histone acetylations on loci highly acetylated in NTM_H cells, but not in TM_H cells, were partially restored to the levels seen in NTM_H cells (Figures 7A, 7B, and S7A). However, this treatment had minimal impact on acetylation levels on the loci seen to harbor higher levels of acetylation in TM_H cells (Figures S7B–S7D). Next, we tested the impact of vorinostat and entinostat on the growth rate of NTM and TM cells in a time-course experiment. Indeed, both of these inhibitors showed preferential effect on abrogation of proliferation in TM cells TM_H and TM_P compared with NTM_H and NTM_P (Figures 7C, 7D, S7E, and S7F).

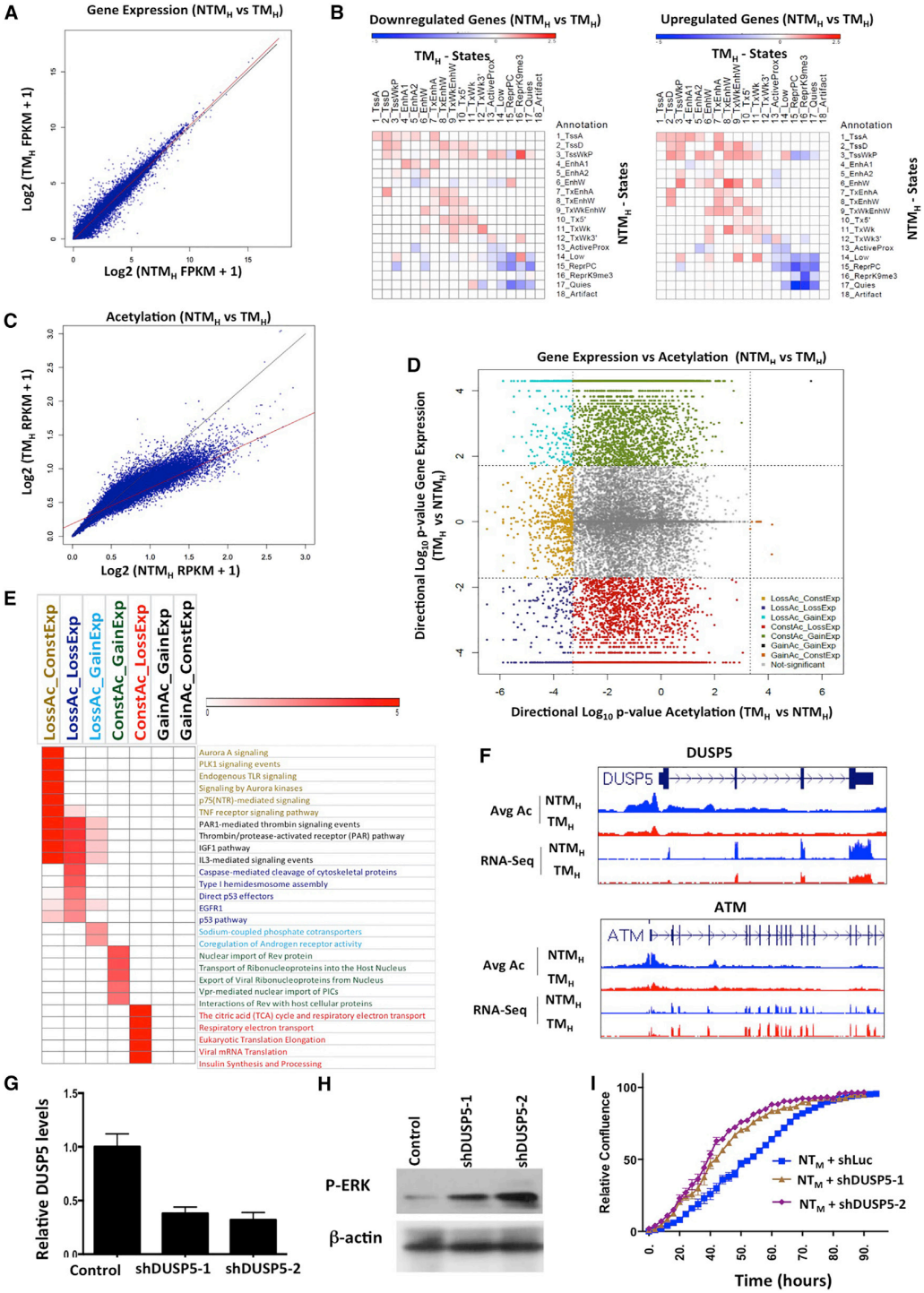


Figure 5. Correspondence of Chromatin State Changes with RNA Expression Changes during Transition to Tumorigenesis

(A) Scatterplot comparing gene expression values [$\log_2$ (FPKM+1)] in NTM$_H$ and TM$_H$ for RefSeq genes.

(B) Relative enrichment of chromatin state transitions at promoters of downregulated genes compared with upregulated genes (left panel) or upregulated gene promoters compared with downregulated (right panel) for all pairs of chromatin state transitions. Red shows enrichment, whereas blue is depletion.

(legend continued on next page)

Because these observations were made in an artificial model system that mimics melanoma progression, we next tested whether levels of histone acetylation in melanoma-derived cell lines could indicate vulnerability to HDAC inhibitors. To this end, we performed H3K27Ac ChIP-seq in five melanoma cell lines and measured relative acetylation levels on promoters deacetylated in TM_H cells. NTM_H and Hs839.T contained relatively higher levels, Skmel-28 moderate levels, and WM115, Skmel-5, and WM793B showed lower levels of H3K27Ac (Figure 7E). To determine whether the acetylation levels correlated with their response to HDAC inhibitors, we determined IC_{50} values and area under the curve (AUC) for vorinostat and entinostat in each cell line (Figures 7F–7H and S7G). Indeed, all three cell lines with lower acetylation levels (WM115, Skmel-5, and WM793B) showed substantially lower IC_{50}/AUC values compared with those that had higher acetylation levels (NTM_H and Hs839.T) to both treatments (Figures 7F, 7G, and S7G). Skmel-28, which harbored intermediate levels of acetylation, displayed intermediate IC_{50}/AUC values (Figures 7F, 7G, and S7G). Correlation score between average acetylation at TSS and AUC values was calculated to be 0.97 (Figure 7H). These data confirm disease relevance of our observations that lower acetylation levels in TM cells functionally contribute to the proliferative phenotype and suggest that responsiveness to HDAC inhibitors may associate with histone acetylation levels on specific genomic loci.

DISCUSSION

We have generated snapshots of the epigenome landscape at two phenotypically distinct biological states (e.g., NTM and TM) as a way to delineate changes that are associated with tumorigenesis by leveraging an isogenic cell model system. Although artificial, this system is well-suited for this study for the following reasons: (1) these cells are derived from primary melanocytes, the appropriate cells of origin for melanoma; (2) the cell system recapitulates known genetic alterations observed in human melanoma tumors, in particular in *P53*, *CDK4*, *BRAF*, and *PTEN* (Hodis et al., 2012; Krauthammer et al., 2012); and (3) NTM cells (shGFP) and TM cells (shPTEN) are otherwise isogenic. Finally, this melanocyte-based cell system has been previously used in other mechanistic studies related to regulation of melanomagenesis (Garraway et al., 2005; Rai et al., 2015).

We show that a predominant feature of chromatin state changes during progression to TM state in melanoma is lowering of frequency of detectable locations of acetylation modifications. Two independent observations suggest that these changes are relevant to human disease and could play a functional role: first, a selected subset of acetylation changes between NTM and TM cells was reproduced between human benign nevi and malignant tumors. Second, the treatment with HDAC inhibition, which restored acetylation patterns on deacetylated loci, was able to abrogate high proliferation rate of TM cells and melanoma cells that contained lower acetylation than NTM cells. Overall, our data suggest that a specific chromatin environment around certain loci in the genome can have pro-tumorigenic function.

One can hypothesize that such a state of chromatin can be established by one or multiple tumor-promoting genetic events such as PTEN deletion/mutation or other alterations in epigenetic machinery. This is supported by our observations that knockdown of CBP histone acetyltransferase in NTM_H cells or overexpression of NRAS (which recapitulates BRAF activation + PTEN loss because of its ability to activate MAPK and PI3K pathways; Chudnovsky et al., 2005) in *TERT*/*P53*/*CDK4*^{R24C} immortalized melanocytes showed similar histone acetylation profiles as in TM_H.

Functional characterization of the regions displaying altered chromatin states suggested that, consistent with observed phenotypes, promoters of genes with important roles in cancer progression show preferential deacetylation, such as cell cycle regulation and apoptosis, in TM cells. Further, we noted that a number of genes in important melanoma cell signaling pathways such as TRAIL, IFNγ, LKB1, PDGF, PI3K, ITGβ1, TGFβ, and cytokine signaling were associated with chromatin state changes involving histone acetylations (Figure 4). For example, TGFβ and INTGβ1 are known to regulate cell invasion and adhesion properties (Jakowlew, 2006; Trikha et al., 1997), consistent with observed invasive properties of TM cells. Enrichment of multiple such signaling events linked to observed TM phenotypes suggests that, in this model system, chromatin-associated changes are likely regulators of cancer progression, underscoring important roles of chromatin in tumorigenesis. This hypothesis is reinforced by abrogation of hyperproliferative phenotype by HDAC inhibitors, which restores the acetylation on deacetylated loci.

Interestingly, although acetylation intensity measurements based on ChIP-seq profiles revealed a loss of peaks of acetylation marks in TM cells, we did not observe any changes in total levels of histone acetylation marks either by western blotting in whole cell lysate and chromatin fraction or by mass spectrometry analysis of acid-extracted histones (Figure S1 and data not shown). Two independent observations in addition to ChIP-seq reinforced the results of loss of acetylation peaks in TM cells

(C) Scatterplot comparing promoter acetylations [$\log_2$(RPKM+1)] around ±2 kb of each RefSeq gene in NTM_H and TM_H. The line in red is a regression line, whereas in black it is the y = x line.
(D) Scatterplot displays directional $\log_{10}$(p value) for acetylation and gene expression changes between TM_H and NTM_H. Negative values represent genes with decreased expression or acetylation levels in TM_H compared with NTM_H cells. Dashed lines show the significance cutoff for acetylation or expression changes. Genes with significant gene expression and/or acetylation changes are colored based on grouping indicated.
(E) Heatmap represents enriched pathways (pathway commons) for each group identified in (D). Color scale represents $-\log_{10}$(HyperFdrQ corrected).
(F) UCSC genome browser view of average acetylation and RNA-seq for an example from each of the LossAc_LossExp (DUSP5) (top) and LossAc_ConstExp (ATM) groups (bottom).
(G) Graph showing relative levels of DUSP5 in NTM_H cells harboring either control or DUSP5 shRNAs.
(H) Western blot showing levels of p-ERK in NTM_H cells harboring either control or DUSP5 shRNAs.
(I) Growth curve showing proliferative capacity of NTM_H cells harboring control or DUSP5 shRNAs (shDUSP5-1 and shDUSP5-2).
See also Figure S6 and Table S7.

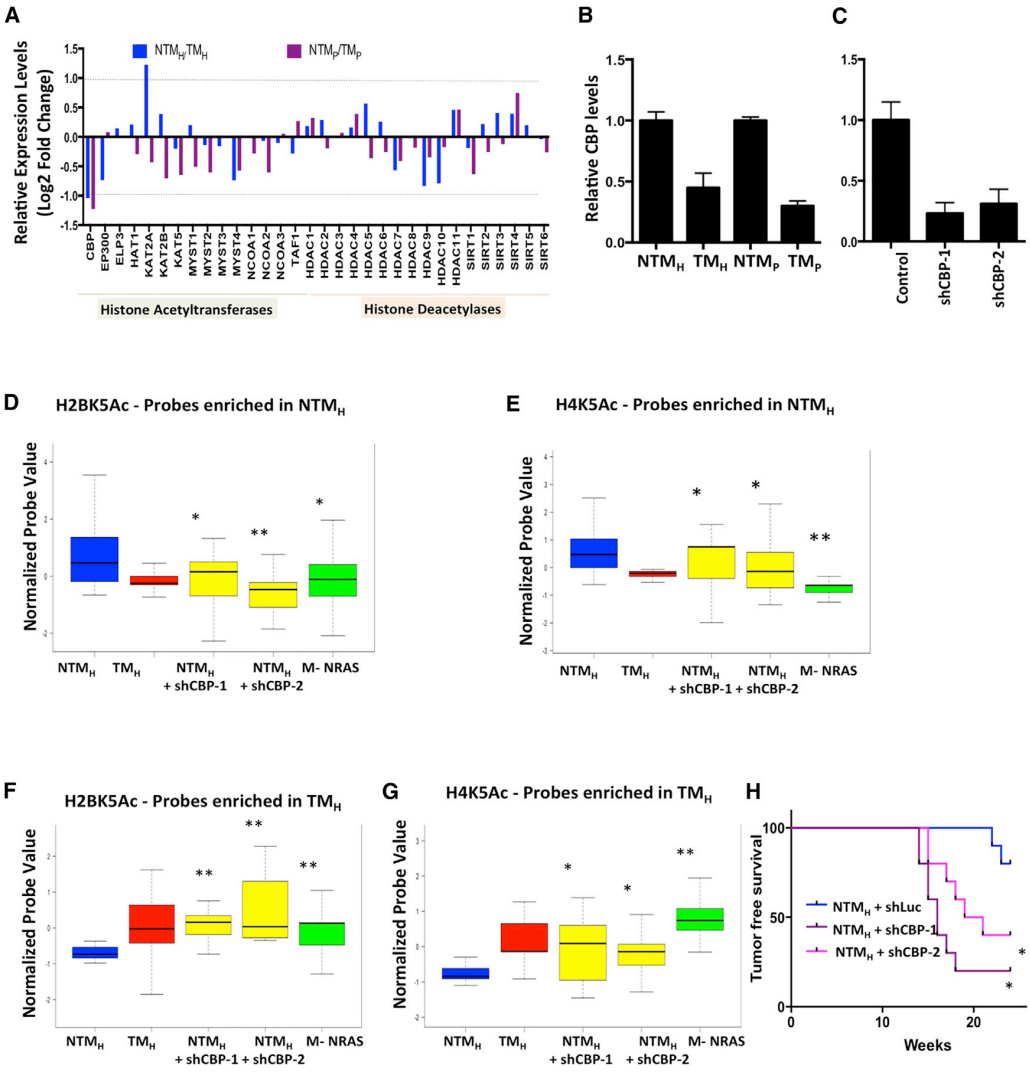


Figure 6. CBP Loss in NTM_H Cells Promotes Tumorigenesis and Mimics Acetylation Loss Seen in TM_H Cells

(A) Bar graph showing relative levels of 32 histone acetyltransferases and deacetylases between NTM_H/TM_H and NTM_P/TM_P cells. The y axis shows log_2 fold-change values. The dotted line shows the cutoff of 2-fold change.

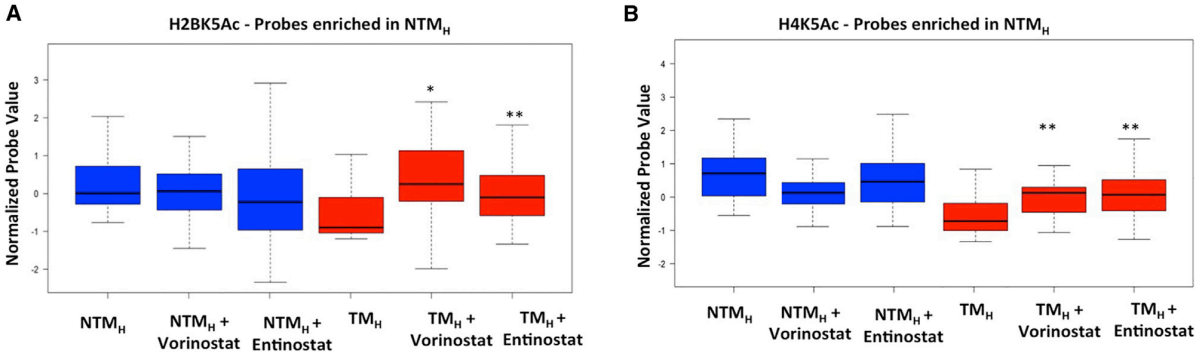
(B and C) Graph showing relative levels of CBP histone acetyltransferase in (B) NTM_H, TM_H, NTM_P, and TM_P cells and (C) NTM_H cells harboring either control or CBP shRNAs.

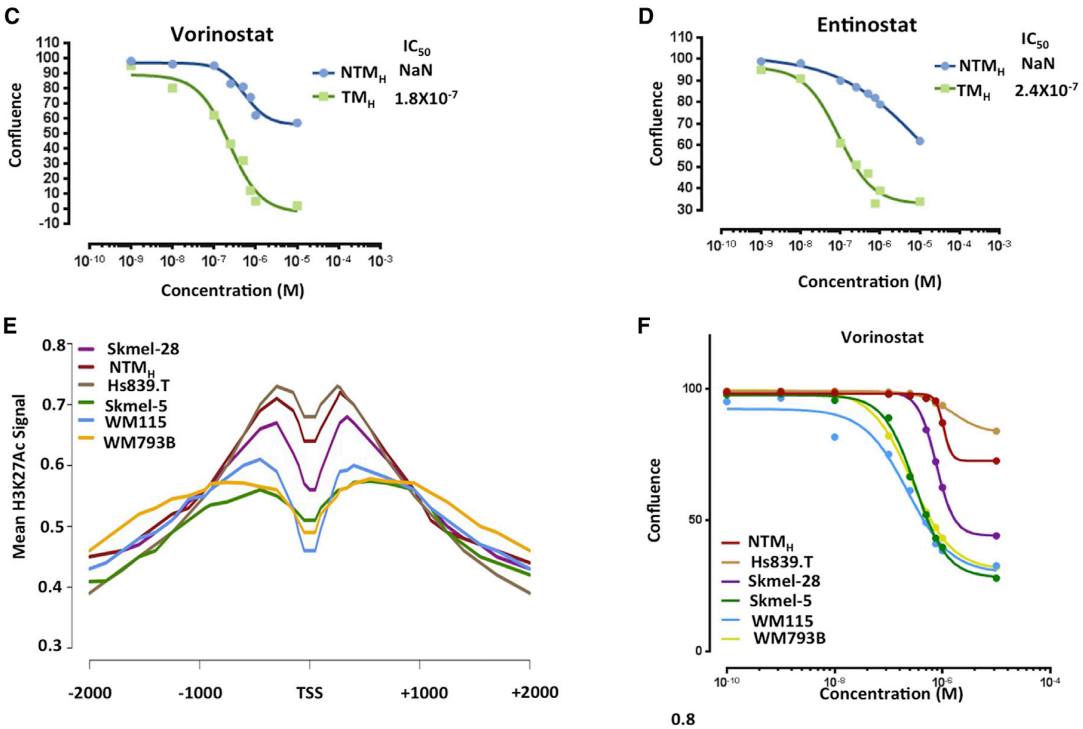
(D–G) Boxplots showing average normalized intensity for ChIP-string probes for (D and F) H2BK5Ac and (E and G) H4K5Ac in NTM_H, TM_H, and NTM_H cells harboring CBP shRNAs or $NRAS^{G12D}$-expressing transformed melanocytes (M-NRAS). The plot is limited to those probes that were originally enriched in (D and E) NTM_H cells or in (F and G) TM_H cells by ChIP-seq experiments and validated by ChIP-string in Figures S2A–S2F. *$p < 0.05$; **$p < 0.001$ (Wilcoxon rank test), when comparisons are made with NTM_H.

(H) Kaplan-Meier curve showing tumor formation efficiency of NTM_H cells harboring control or CBP shRNAs (shCBP-1 and shCBP-2).

and diminished the possibility of this being a technical or experimental bias. First, ChIP enrichment measurement by either nanostring at 96 probes or by qPCR at five loci revealed results consistent with the ChIP-seq signal (Figures 2 and S2 and data not shown). Second, biological replicates for ChIP followed by either nanostring or qPCR measurement revealed similar enrichment profiles (data not shown). Based on these observations, we speculate that, although both TM and NTM cells harbor the same levels of acetylated histones in the cell and on chromatin, acetylated histones are more diffusely incorporated throughout the genome in TM cells, including at locations where it is not present in normal cells.

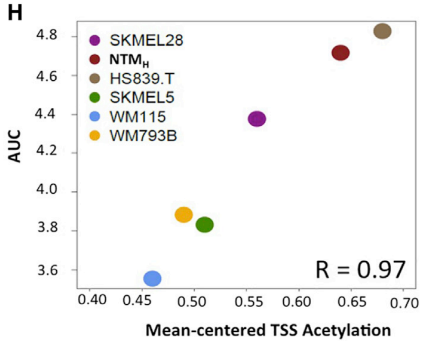
Overlap of epigenomic and transcriptomic data revealed that there was little correlation between chromatin changes and





G

	Vorinostat		Entinostat	
Cell-Line	IC50	AUC	IC50	AUC
NTM_H	NaN	4.7	NaN	4.6
Hs839	NaN	4.8	NaN	4.9
Skmel-28	1.74E-05	4.4	NaN	4.2
Skmel-5	5.34E-06	3.8	3.49E-06	3.6
Wm115	4.49E-06	3.6	1.71E-06	3.5
WM793B	6.15E-06	3.9	4.87E-06	3.7



(legend on next page)

gene expression changes at the global level between the TM and NTM cells in this system. Some other recent studies have also shown low correlation between gene expression and acetylation changes in specific systems (Sen et al., 2016; Sun et al., 2016). A plausible explanation for such an observation is that the steady-state RNA levels may not completely be reflective of all chromatin-associated gene expression changes during biological state switches in tumorigenesis process and are influenced by other post-transcriptional regulatory molecular processes. Nonetheless, systematic analysis of gene sets in different groups clearly suggests that the set of genes that had both lower acetylations and lower gene expression enriched for pathways with known roles in tumor progression underscoring the importance of chromatin-associated gene expression changes in cancer progression.

Taken together, our study provides a first systematic view of the epigenomic, as well as transcriptomic, landscape evolutions between two distinct biological states (e.g., NTM and TM) associated with melanoma tumorigenesis.

EXPERIMENTAL PROCEDURES

Cell Culture, Generation of Stable Cells, and Drug Treatment

HMEL-BRAF^{V600E}, PMEL-BRAF^{V600E} cells were obtained from Dr. David Fisher's laboratory (Garraway et al., 2005) and maintained in standard tissue-culture conditions in DMEM media with 10% fetal bovine serum (FBS). Stable knockdown of GFP (control) or PTEN (experimental) in early passage ($n < 10$) was performed using pMKO-shGFP or pMKO-shPTEN vectors (Addgene) to create NTM_H (HMEL-BRAF^{V600E}-shGFP), TM_H (HMEL-BRAF^{V600E}-shPTEN), NTM_P (PMEL-BRAF^{V600E}-shGFP), and TM_P (PMEL-BRAF^{V600E}-shPTEN) cells. Control and experimental cells were passaged together for the same time before harvesting cells for ChIP-seq experiments. Hs839.T, Skmel-28, Skmel-5, WM115, and WM793B cells were obtained from ATCC and grown according to the manufacturer's recommendation. Cells were treated with Vorinostat (Sigma), entinostat (MS-275; SelleckChem) or vehicle (DMSO) by direct addition to media.

ChIP-Seq

ChIP was performed as described earlier (Garber et al., 2012) with optimized shearing conditions and minor modifications for melanocytes. For more details, see Supplemental Experimental Procedures.

ChIP-Seq and Chromatin State Analysis

ChIP-seq reads were aligned using Bowtie (version 1.0.0) (Langmead et al., 2009) to human genome assembly NCBI Build 37 (University of California at Santa Cruz [UCSC] hg19) and uniquely mapped reads with one mismatch were retained. ChromHMM (Ernst and Kellis, 2012) was used with default parameters to derive genome-wide chromatin state maps for all cell types. We binarized the input data with ChromHMM's BinarizeBed method using a p value cutoff of 10^{-4}. Chromatin state models were learned jointly on all chromatin marks from NTM_H and TM_H ranging from 10 to 120 states. A model with 18 states was chosen for detailed analysis and is presented throughout the manuscript. Chromatin state annotations of NTM_H, TM_H, NTM_P, and TM_P were produced subsequently by applying this model to the original binarized, quantile normalized, or downsampled chromatin data from these cell types. For details, see Supplemental Experimental Procedures.

RNA-Seq

Strand-specific libraries were constructed using ScriptSeq Kit (Epicenter/Illumina). Reads were mapped to the human genome (hg19) using MapSplice algorithm version 2.1.4 (Wang et al., 2010). Transcript expression was estimated using Cuffdiff 2.11. Further details are in the Supplemental Experimental Procedures.

ChIP-String

Nanostring experiments were run on a custom ChIP-string array according to the manufacturer's recommendation using ChIP-DNA for shown marks (Figures 2 and S2) from NTM_H and TM_H cells and nevi and tumor cells. A custom ChIP-string array containing probes for 96 genomic locations (Table S2) was used. Details of the design are in the Supplemental Experimental Procedures. The analysis was done as previously described by Ram et al. (2011). The details are in the extended Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the datasets reported in this paper is GEO: GSE58953.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.03.078.

AUTHOR CONTRIBUTIONS

Conceptualization, L.C., J.E., K.R., K.C.A., and P.F.; Methodology, A.P., S. Seth, C.S.C., I.A., P.F., K.C.A., K.R., and J.E.; Investigation, P.F., K.C.A., K.R., J.E., J.P.M., E.Z.K., N.S.S., S. Sharma, C.A.N., C.J.T., M.M., S.B.A., E.M.L., M.D., J.B.A., A.S., H.M., and B.A.G.; Writing, J.E., L.C., K.R., K.C.A., and P.F.; Funding Acquisition, L.C., J.E., and K.R.; Resources, K.Y.T., M.A.D., J.E., L.C., and K.R.; Supervision, L.C., J.E., K.R., and M.C.B.

ACKNOWLEDGMENTS

We thank Siavash Kurdistani, Sharmistha Sarkar, Yonathan Lissanu Deribe, Ian R. Watson, and Lawrence Kwong for useful discussions or comments on the manuscript. We thank Marcus Coyle, Curtis Gumbs, and SMF core at MDACC for sequencing support. The work described in this article was supported by the NIH (grants 5U01 CA141508 and U01 CA168394 to L.C.; grants R01ES024995 and U01 HG007912 to J. E.; grant CA016672 to SMF Core; grants GM110174 and CA196539 to B.A.G.), NCI (grants 1K99CA160578

Figure 7. Acetylation Status on Deacetylated Promoters in T_H Correlates with Response to HDAC Inhibitors

(A and B) Boxplots showing average normalized intensity for (A) H2BK5Ac or (B) H4K5Ac on ChIP-string probes (that were enriched in NTM_H cells by ChIP-seq experiment) across NTM_H and TM_H cells that were either untreated or treated with vorinostat (200 nM) or entinostat (300 nM) for 72 hr. *$p < 0.05$; **$p < 0.001$ (Wilcoxon rank test), when comparisons are made with TM_H.

(C and D) Growth curves for NTM_H and TM_H cells grown under various concentrations of (C) vorinostat or (D) entinostat.

(E) Aggregate plot showing H3K27Ac levels around ±2 kb of deacetylated gene promoters (in T_H cells) in various melanoma cell lines.

(F) Growth curves for melanoma cell lines grown under various concentrations of vorinostat.

(G) Table showing IC_{50} values (the concentration at which 50% response is achieved) and area under the curve (AUC) for two HDAC inhibitors, vorinostat and entinostat, in melanoma cells lines. Immeasurable IC_{50} values are shown as NaN (not a number).

(H) Correlation plot between AUC and average H3K27Ac levels at TSS of gene promoters that showed loss of histone acetylation in TM_H cells compared with NTM_H cells.

See also Figure S7.

and R00CA160578 to K.R.), Cancer Prevention and Research Institute of Texas (grant R1204 to L.C.), National Science Foundation (grant 1254200 to J.E.), and Center for Cancer Epigenetics at MDACC (K.R.). The following fellowship support is acknowledged: Charles A. King postdoctoral fellowship (to K.R.), Alfred P. Sloan fellowship (to J.E.), California Institute for Regenerative Medicine Training Grant TG2-01169 (to P.F.), and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA Training Program (P.F.).

Received: October 13, 2016
Revised: February 18, 2017
Accepted: March 27, 2017
Published: April 25, 2017

REFERENCES

Akhtar-Zaidi, B., Cowper-Sal-lari, R., Corradin, O., Saiakhova, A., Bartels, C.F., Balasubramanian, D., Myeroff, L., Lutterbaugh, J., Jarrar, A., Kalady, M.F., et al. (2012). Epigenomic enhancer profiling defines a signature of colon cancer. Science *336*, 736–739.

Bennett, D.C. (2016). Genetics of melanoma progression: the rise and fall of cell senescence. Pigment Cell Melanoma Res. *29*, 122–140.

Busse, A., and Keilholz, U. (2011). Role of TGF-β in melanoma. Curr. Pharm. Biotechnol. *12*, 2165–2175.

Campioni, M., Santini, D., Tonini, G., Murace, R., Dragonetti, E., Spugnini, E.P., and Baldi, A. (2005). Role of Apaf-1, a key regulator of apoptosis, in melanoma progression and chemoresistance. Exp. Dermatol. *14*, 811–818.

Caunt, C.J., and Keyse, S.M. (2013). Dual-specificity MAP kinase phosphatases (MKPs): shaping the outcome of MAP kinase signalling. FEBS J. *280*, 489–504.

Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. (2013). Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell *24*, 777–790.

Chudnovsky, Y., Adams, A.E., Robbins, P.B., Lin, Q., and Khavari, P.A. (2005). Use of human tissue to assess the oncogenic activity of melanoma-associated mutations. Nat. Genet. *37*, 745–749.

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

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

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

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., et al. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117–122.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. Cell *150*, 251–263.

Jakowlew, S.B. (2006). Transforming growth factor-beta in cancer and metastasis. Cancer Metastasis Rev. *25*, 435–457.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Krauthammer, M., Kong, Y., Ha, B.H., Evans, P., Bacchiocchi, A., McCusker, J.P., Cheng, E., Davis, M.J., Goh, G., Choi, M., et al. (2012). Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat. Genet. *44*, 1006–1014.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

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

Muratani, M., Deng, N., Ooi, W.F., Lin, S.J., Xing, M., Xu, C., Qamra, A., Tay, S.T., Malik, S., Wu, J., et al. (2014). Nanoscale chromatin profiling of gastric adenocarcinoma reveals cancer-associated cryptic promoters and somatically acquired regulatory elements. Nat. Commun. *5*, 4361.

Paluncic, J., Kovacevic, Z., Jansson, P.J., Kalinowski, D., Merlot, A.M., Huang, M.L., Lok, H.C., Sahni, S., Lane, D.J., and Richardson, D.R. (2016). Roads to melanoma: key pathways and emerging players in melanoma progression and oncogenic signaling. Biochim. Biophys. Acta *1863*, 770–784.

Pinon, P., and Wehrle-Haller, B. (2011). Integrins: versatile receptors controlling melanocyte adhesion, migration and proliferation. Pigment Cell Melanoma Res. *24*, 282–294.

Rai, K., Akdemir, K.C., Kwong, L.N., Fiziev, P., Wu, C.J., Keung, E.Z., Sharma, S., Samant, N.S., Williams, M., Axelrad, J.B., et al. (2015). Dual roles of RNF2 in melanoma progression. Cancer Discov. *5*, 1314–1327.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., et al. (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628–1639.

Sammons, M.A., Zhu, J., and Berger, S.L. (2016). A chromatin-focused siRNA screen for regulators of p53-dependent transcription. G3 (Bethesda) *6*, 2671–2678.

Sen, D.R., Kaminski, J., Barnitz, R.A., Kurachi, M., Gerdemann, U., Yates, K.B., Tsao, H.W., Godec, J., LaFleur, M.W., Brown, F.D., et al. (2016). The epigenetic landscape of T cell exhaustion. Science *354*, 1165–1169.

Sheridan, C., Brumatti, G., and Martin, S.J. (2008). Oncogenic B-RafV600E inhibits apoptosis and promotes ERK-dependent inactivation of Bad and Bim. J. Biol. Chem. *283*, 22128–22135.

Shiloh, Y., and Ziv, Y. (2013). The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. Nat. Rev. Mol. Cell Biol. *14*, 197–210.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

Sullivan, R.J., and Flaherty, K. (2013). MAP kinase signaling and inhibition in melanoma. Oncogene *32*, 2373–2379.

Sun, W., Poschmann, J., Cruz-Herrera Del Rosario, R., Parikshak, N.N., Hajan, H.S., Kumar, V., Ramasamy, R., Belgard, T.G., Elanggovan, B., Wong, C.C., et al. (2016). Histone acetylome-wide association study of autism spectrum disorder. Cell *167*, 1385–1397.e11.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Trikha, M., Timar, J., Lundy, S.K., Szekeres, K., Cai, Y., Porter, A.T., and Honn, K.V. (1997). The high affinity alphaIIb beta3 integrin is involved in invasion of human melanoma cells. Cancer Res. *57*, 2522–2528.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Cell Reports, Volume *19*

Supplemental Information

Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression

Petko Fiziev, Kadir C. Akdemir, John P. Miller, Emily Z. Keung, Neha S. Samant, Sneha Sharma, Christopher A. Natale, Christopher J. Terranova, Mayinuer Maitituoheti, Samirkumar B. Amin, Emmanuel Martinez-Ledesma, Mayura Dhamdhere, Jacob B. Axelrad, Amiksha Shah, Christine S. Cheng, Harshad Mahadeshwar, Sahil Seth, Michelle C. Barton, Alexei Protopopov, Kenneth Y. Tsai, Michael A. Davies, Benjamin A. Garcia, Ido Amit, Lynda Chin, Jason Ernst, and Kunal Rai

SUPPLEMENTARY INFORMATION

EXTENDED SUPPLEMENTARY METHODS

Cell Culture and Generation of Stable Cells

HMEL-BRAF^{V600E}, PMEL-BRAF^{V600E}, HEK-293T cells were grown in 5% CO_2, at 37^{0}C in DMEM medium with 10% FBS. Cells were routinely tested for mycoplasma infection. Viral production was done in HEK-293T cells using pHIT60 and VSVg. Stable knockdown of GFP or PTEN in early passage (n < 10) HMEL-BRAF^{V600E}, PMEL-BRAF^{V600E} cells was performed using pMKO-shGFP or pMKO-shPTEN vectors (Addgene) to create NTM_H (HMEL-BRAF^{V600E}-shGFP), T_H (HMEL-BRAF^{V600E}-shPTEN), NTM_P (PMEL-BRAF^{V600E}-shGFP) and TM_P (PMEL-BRAF^{V600E}-shPTEN) cells.

ChIP-Seq

Chromatin immunoprecipitation was performed as described earlier (Garber et al., 2012) with optimized shearing conditions and minor modifications for melanocytes. Briefly, cells (5 million per antibody) were cross linked using 1% paraformaldehyde for 10mins at 37^{o}C. Reaction was quenched by 0.125M glycine for 5mins, and cells washed with PBS and stored at -80^{o}C. Next day cells were thawed on ice and lysed with RIPA buffer (10mM Tris-HCl pH 8.0, 1mM EDTA pH 8.0, 140mM NaCl, 1% Triton x-100, 0.2%SDS, 0.1% DOC) for 10min on ice. Sonication conditions were optimized for HMEL-BRAF^{V600E} cells and were performed using Branson Sonifier 250 to achieve shear length of 250-500bp. Extracts were then incubated overnight with respective antibody-dynabead mixtures that were incubated separately for 1hr at 4^{o}C earlier. Immunecomplexes were then washed in following order: 5 times with RIPA buffer, twice with RIPA-500 (RIPA with 500mM NaCl), twice with LiCl wash buffer (10mM Tris-HCl pH8.0, 1mM EDTA pH8.0, 250mM LiCl, 0.5% NP-40, 0.1% DOC) and once with TE (10mM Tris-HCl, 1mM

EDTA). Elution and decrosslinking was performed in direct elution buffer (10mM Tris-Cl pH8.0, 5mM EDTA, 300mM NaCl, 0.5% SDS) by incubating immunecomplexes at 65^{o}C overnight. Proteinase K (20mg/ml) and RNaseA treatment was performed and DNA cleaned up using SPRI beads (Beckman-Coulter). Library preparation was done as described in (Garber et al., 2012) using paired end adapters from IDT. Libraries were multiplexed together and sequencing was performed in Hiseq2000 (Illumina). Antibody details are below:

Mark	Company	Catalog Number
H2AK5ac	Abcam	ab45152
H2BK120ac	Active Motif	39119
H2BK15ac	Abcam	ab62335
H2BK5ac	Active Motif	39123
H3	Abcam	ab1791
H3K14ac	Millipore	07-353
H3K18ac	Abcam	ab1191
H3K23ac	Millipore	07-355
H3K27ac	Abcam	ab4729
H3K27me1	Millipore	07-448
H3K27me3	Abcam	ab6002
H3K36ac	Active Motif	39379
H3K36me1	Abcam	ab9048
H3K36me2	Abcam	ab9049
H3K36me3	Abcam	ab9050
H3K4ac	Millipore	07-539
H3K4me1	Abcam	ab8895
H3K4me2	Abcam	ab32356
H3K4me3	Abcam	ab8580
H3K79me1	Abcam	ab2886
H3K79me2	Abcam	ab3594
H3K79me3	Abcam	ab2621
H3K9ac	Abcam	ab4441
H3K9me1	Abcam	ab8896
H3K9me2	Abcam	ab1220
H3K9me3	Abcam	ab8898
H4	Millipore	05-858
H4ac4	Active Motif	39179
H4K12ac	Active Motif	39165
H4K16ac	Millipore	07-329

H4K20me1	Abcam	ab9051
H4K5ac	Millipore	07-327
H4K8ac	Abcam	ab15823
H4K91ac	Abcam	ab4627
5-hmC	Active Motif	39769
H4K20me2	Abcam	ab9052
H4K20me3	Abcam	ab9053

ChIP-Seq Data Analysis:

ChIP-Seq reads were aligned using Bowtie (version 1.0.0) (Langmead et al., 2009) to human genome assembly NCBI Build 37 (UCSC hg19) with the following parameters: -n 1 -m 1 --best –strata (uniquely mapped reads with one mismatch were retained). First 36bp from 5' end of the reads were retained in case read lengths are longer than 36bp for any given histone modifications. To avoid biases due to PCR artifacts, sequencing reads that map to the same genomic location and strand were counted once in the input data.

Peak calling was performed using MACS algorithm (Zhang et al., 2008) with default parameters except a p-value cut-off 10E-8 applied. DiffBind bioconductor package was used to cluster histone marks by using identified peaks with MACS algorithm.

We generated signal tracks at 200bp resolution, by partitioning the genome into non-overlapping bins at that resolution. We calculated signal values over all bins for each histone mark using the following formula:

$$Signal_i = \frac{K_i * 10^9}{L_i * N}$$

where $Signal_i$ is the signal value of a given histone mark at bin *i*, K_i is the raw number of sequencing reads for that mark that span bin *i* after extending each read by 200bp from the start in the direction of the alignment, L_i is the length of bin *i*, and *N* is the total number of sequencing reads for that mark.

ChromHMM Analysis

We used ChromHMM (Ernst and Kellis, 2012) with default parameters to derive genome-wide chromatin state maps for all cell types. We binarized the input data with ChromHMM's BinarizeBed method using a p-value cutoff of 1e-4. We observed that the total number of binary presence calls was very similar between NTM_H and TM_H. However, total number of calls was higher in NTM_P compared to T_P. Thus, to reduce the effect of potential technical confounders, we normalized each chromatin mark in NTM_P and TM_P to have the same number of binary presence calls across these two cell types. To achieve that, we first used the BinarizeBed option of ChromHMM on all datasets from NTM_P and TM_P. Then, we sorted the binary calls for each chromatin mark in each bin according to the number of reads assigned and kept only the top N binary calls, where N is the smaller of the total numbers of binary calls for the corresponding chromatin mark in NTM_P and TM_P. In the case of ties, we dropped randomly binary calls that had the same number of sequencing reads assigned in order to arrive at equal number of binary calls across the two cell lines for the corresponding chromatin mark.

We considered chromatin state models learned jointly on all chromatin marks from NTM_H and TM_H ranging from 10 to 120 states. Two models were considered for additional analysis: 18-state model (with the minimum number of states that had a separate state containing likely artifactual signal locations) and 45-state model (with the minimum number of states that contains a clear poised/bivalent state). We chose to focus on a model with 18 states for our main analysis to balance capturing informative state distinctions while maintaining interpretability and having a manageable number of pairwise state transitions. In particular the model with 18 states was the model with the

minimum number of states that had a separate state containing likely artifactual signal locations. The chromatin state annotations of NTM_H, TM_H, NTM_P and TM_P was produced subsequently by applying this model to the chromatin data from these cell types.

Analysis of Chromatin State Changes

To find important chromatin state changes between non-tumorigenic and tumorigenic cell lines, we intersected the chromatin state annotations of NTM_H and TM_H, and of NTM_P and TM_P, respectively. In each case, we counted the number of 200bp bins that is occupied by each of the 18 by 18 possible chromatin state transitions. To calculate enrichment scores, we divided this number by the expected number of such bins assuming a null model that treats the two chromatin states involved in each transition as independently distributed. Finally, to control for state similarity between each pair of chromatin states *i* and *j*, we divided the enrichment score of transitioning from state *i* in non-tumorigenic cells to state *j* in tumorigenic cells by the enrichment score of transitioning from state *j* in non-tumorigenic cells to state *i* in tumorigenic cells. In order to avoid division by 0 in cases where no overlap was detected between pairs of chromatin states, we added a pseudo-count of 1 bin to each intersection before we computed all enrichments, enrichment ratios and p-values.

Besides our main analysis, we performed the above computations under two other normalization schemes. First, we downsampled randomly the number of sequencing reads for each chromatin mark to the minimum number across NTM_H and TM_H, and across NTM_P and TM_P, respectively. We applied the previously learned 18 states model on the downsampled data and ran the above analysis pipeline on the produced chromatin state annotations. In our second normalization scheme, we downsampled the number of binary calls from ChromHMM's BinarizeBed routine for each chromatin mark

to the minimum number across all four cell types in the same way we did previously for NTM_P and TM_P. Again, we applied the 18 state model to this data and ran the above analysis pipeline.

Analysis of Chromatin State Recovery with Subsets of Marks

For the analysis of the chromatin state recovery with subset of marks relative to using all marks we used the EvalSubset of ChromHMM command(Ernst and Kellis, 2012) (Ernst and Kellis, 2015) applied to the chromatin state annotations of NTM_H and TM_H. For this analysis we separately evaluated for each mark, recovery based on only that mark and using all marks except that mark.

Analysis of Individual Mark Enrichments at Promoters and DNaseI hypersensitive sites

Promoter regions were defined as 4kbp regions centered at annotated transcription start sites from RefSeq (as downloaded on March 2014 from UCSC Genome Browser). As for DNaseI hypersensitive sites (DHS), we downloaded a data set with DNaseI peaks for the Melano cell type from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/wgEncodeOpenChromDnaseMelanoPk.narrowPeak.gz). To define distal sites, we further excluded peaks whose midpoints are within 4kb of annotated transcription start sites in hg19. To compute the histone mark signal over the remaining sites, we extended them by 2kb from their midpoints in both directions. For each promoter region or distal DHS *i*, chromatin mark in cell type *c*, we calculated the signal strength in RPKM as:

$$\mathrm{RPKM}_{i,c} = \frac{K_{i,c} * 10^9}{L_i * N_c}$$

where $K_{i,c}$ is the number of sequencing reads from that mark in cell type *c* whose center position overlaps with region *i* after extending each read by 200bp in the direction of the alignment, L_i is the length of region *i*, and N_c is the total number of reads for the mark. We then calculated the average fold change of each mark at promoters and DHS separately by summing over all regions for them as:

$$\Delta = \log_2 \left(\frac{\sum_i \mathrm{RPKM}_{i,c_1}}{\sum_i \mathrm{RPKM}_{i,c_2}} \right),$$

Where c_1 and c_2 are NTM_H and TM_H or NTM_P and TM_P respectively.

Differentially Acetylated Promoters

To identify statistically significant differences of acetylations in aggregate at promoters, we compared to a null model in which the non-tumorigenic and tumorigenic label for each acetylation was randomly permuted. Specifically, for both promoters in each cell type separately we calculated an average acetylation level in each region *i* by taking the mean RPKM value across all acetylation marks for the given cell type *c*, denoted by $\mathrm{RPKM(Ac)}_{i,c}$. Next, we calculated the change in the average acetylation levels at region *i*:

$$\Delta(\mathrm{Ac})_i = \log_2 \left(\frac{\mathrm{RPKM(Ac)}_{i,c_1} + 1}{\mathrm{RPKM(Ac)}_{i,c_2} + 1} \right),$$

where c_1 and c_2 are NTM_H and TM_H, or NTM_P and TM_P, respectively. To determine significant changes at a FDR of 1% we used a null model based on 100 randomized pairs of cell types for each system (NTM_H / TM_H and NTM_P / TM_P). Each randomized pair was generated by iterating through all acetylation datasets from NTM_H and TM_H (or NTM_P and TM_P) and randomly switching their labels with probability of 0.5. Based on the randomized data we constructed a background distribution of the $\Delta(\mathrm{Ac})_i$ values across all intervals and randomizations, which we used to calculate two-sided P-values for all

observed $\Delta(\mathrm{Ac})_i$. We then applied the Benjamini–Hochberg procedure on these P-values to derive cutoffs at FDR of 1%.

Promoter State Analysis

Every RefSeq gene was assigned to one chromatin state based on the state call on the gene's TSS in non-tumorigenic (NTM_H, NTM_P) and tumorigenic (TM_H, TM_P) cell types by using the 18 state ChromHMM genome annotation output. STEM software (Ernst and Bar-Joseph, 2006) was used to analyze enriched gene ontology (GO) terms for the genes that are changing their promoter states between non-tumorigenic (NTM_H, NTM_P) and tumorigenic (TM_H, TM_P) cells. Default settings changed to only reporting GO-Biological Process (BP) terms, with equal or below level 5 according to GO taxonomy. STEM software output is processed as following: BP-terms that are enriched for a state-transition with a p-value of less than 10^{-4} and at least 3 genes was assigned for that specific state-transition regarding that particular term is retained. To estimate an overall false discovery rate, we generated random gene sets by keeping the number of genes per state-transition constant but randomly assigning genes from RefSeq annotation table. We did not identify any enriched GO-terms for randomized promoter state-transition pairs with the explained filtering steps. Identified state-transitions and BP-terms were used for heat map generation (Figure 4A).

Pathway Analysis

Pathway Commons analysis on the enriched genomic regions was done using GREAT tool (McLean et al., 2010) (www.great.stanford.edu). For promoter state regions, we used the basal + extension option with -2Kb to +2Kb proximal to TSS and 20Kb extension. For enhancers we used the option of 'single nearest gene" with 1000Kb extension.

RNA-Seq Analysis

Strand specific libraries were constructed using a strand specific method (Levin et al., 2010). Reads were mapped to the human genome (hg19) using Mapsplice algorithm version 2.1.4 (Wang et al., 2010). We first merged the annotations of UCSC gene annotation in Illumina's iGenomes (available at http://cole-trapnell-lab.github.io/cufflinks//igenome_table/index.html) gtf file with the recent long non-coding RNA annotation file (Kelley and Rinn, 2012) using GFFRead tool as part of Cufflinks suite (Trapnell et al., 2013) http://cole-trapnell-lab.github.io/cufflinks. Transcript expression was estimated using Cuffdiff 2.11 with the following option: ''--library-type firststrand'' against the merged annotation file. We then applied Cuffmerge 2.11 based on published protocol (Trapnell et al., 2012) for merging all identified transcripts in each replicates and generated master GTF file used for differential expression analysis. Cuffdiff 2.11 was run with the following options: ''--library-type firststrand, --min-reps-for-js-test 2, –dispersion-method per condition'' and transcripts with less than 0.05 q-values called as differentially expressed (snoRNAs removed from the differentially expressed transcripts list). A transcript was designated as protein coding if it could be assigned to a protein ID using UCSC table browser, rest of the transcripts referred as non-coding. For the up-regulated, down-regulated or unchanged genes, we calculated occurrence of every possible combination of state transitions on TSS, or within -2Kb or +2Kb range. Log2 fold changes calculated based on observed versus expected number of state transitions. Expected number of state transitions was calculated by multiplying all observed transitions within each range (TSS, -2Kb and +2Kb) with the number of up-regulated, down-regulated or unchanged genes then diving with the total RefSeq gene number.

Overlap of average acetylation and transcriptomic data

We systematically overlapped gene expression changes with changes in promoter acetylation to define the nine possible subsets (Figure 5D, S6H): (1) deacetylated-promoters with no corresponding gene-expression changes (LossAc_ConstExp), (2) deacetylated-promoters accompanied with corresponding downregulated gene-expression changes (LossAc_LossExp), (3) deacetylated-promoters accompanied with corresponding upregulated expression changes (LossAc_GainExp), (4) promoters that do not change their acetylation levels but are downregulated at the expression level (ConstAc_LossExp), (5) promoters that do not change their acetylation levels but are upregulated at the expression level (ConstAc_GainExp), (6) acetylation gaining promoters with no corresponding gene-expression changes (GainAc_NoExp), and (7) acetylation gaining promoters accompanied with corresponding upregulated gene-expression changes (GainAc_GainExp). Of the remaining two subsets, one (GainAc_LossExp) was an empty set, while the other set contained only unchanged loci (ConstAc_ConstExp).

DNA Methylation Analysis

We utilized Illumina Infinium HumanMethylation450 BeadChip arrays to profile DNA methylation profiles in NTM_H and TM_H cell lines. The Illumina Infinium HumanMethylation450 BeadChip covers over 450,000 CpG sites in the human genome. We processed the HumanMethylation450 images by Illumina's GenomeStudio Methylation Module software to calculate average beta values for each probes. Later, we used IMA (Illumina Methylation Analyzer) Bioconductor package (Wang et al., 2012) to identify average methylation of CpGs in triplicates of NTM_H and TM_H cells. We removed

the sites with missing beta values and performed quantile normalization and peak correction (Dedeurwaerder et al., 2011).

In addition, we utilized 5-hmCDIP-Seq assay (performed at Active Motif) to identify enriched locations for 5-hydroxymethyl cytosines for non-tumorigenic (NTM_H) and tumorigenic (TM_H) cell lines. Libraries were sequenced as 50bp single-end reads and mapped to the genome using bowtie as mentioned earlier. Peak calling was performed using MACS algorithm with whole cell extract as negative control, and a p-value cut-off of 10^{-10}.

ChIP-String Experiments

We conducted ChIP-string experiments for H2BK5ac, H4K5ac, H3K27ac, H3K4me1, H3K4me3 and H3K27me3 marks in 4 nevi and up to 9 melanoma tumors as well as in NTM_H and TM_H cells on a custom ChIP-string array. These histone marks were chosen to test representative regions from three groups: promoters, enhancers and Polycomb-repressed regions. Since space on the array was limited to 96 probes, we aimed to prioritize marks and regions that are most differential based on the ChIP-seq signal between the tumorigenic and non-tumorigenic cells within each of the three groups. The tested marks were selected based on a combination of prior knowledge about their association with each type of regulatory region and findings in our ChIP-seq data. Initially, we selected H3K27me3 to test Polycomb repressed regions, H3K4me1 to test enhancers, H3K4me3 to test promoters, and H3K27ac to test both enhancers and promoters as these marks are known to correlate with the respective regulatory types. To increase our mark coverage, we further sought to select additional marks that could be tested on the same probes for differential enrichment between NTM and TM cells. By inspecting the top differential regions for pairs of marks, we identified H2BK5ac as a candidate mark that can differentially enrich with H3K27ac, and H4K5ac with H3K4me1.

The 96 probes were split equally in four parts to test regions for differential enrichment of H3K27me3, H3K4me3, H3K27ac together with H2BK5ac, and H3K4me1 together with H4K5ac. In particular, 24 probes were designed for each mark or pair of marks, half of which (12 probes) were chosen to be consistently differentially enriched in both non-tumorigenic cell lines (NTM_H and NTM_P) for the histone mark or pair of marks, and the other half were chosen to be consistently differentially enriched in both tumorigenic cell lines (TM_H and TM_P). This symmetric design allows for a natural positive and negative control of each experiment, because a properly classifiable sample would show positive ChIP-string signal in precisely one of the two groups and no signal in the other group.

To select genomic regions for each mark or pair of marks, we first divided the genome into non-overlapping bins of 200 bp and computed RPKM values for each bin. We then sorted in ascending order all bins by the ratios in their ChIP-seq signal between NTM and TM cells (the smaller of (NTM_H / TM_H) and (NTM_P / TM_P)). For regions tested on pairs of marks, we sorted the bins by the smaller of the ratios of the two marks. We further required that selected bins undergo a transition from one chromatin state to a sufficiently different chromatin state (e.g. bins annotated as promoters in NTM cells transitioning to low signal or to Polycomb repressed in TM cells were allowed, but promoter bins transitioning to other types of promoters were excluded). The chromatin states were defined based on a ChromHMM model learned from a subset of our final ChIP-seq data, which was available at the time the ChIP-String array was commissioned (in the final dataset, a file for H3K4me3 in NTM_P cells was replaced due to a mislabeling issue). Additionally, we required that a binary presence call was made by ChromHMM's BinarizeBed procedure in the cell type the signal was considered enriched in and no binary presence calls were made within 2 kb of the bin for the same mark in the other

cell type. Finally, bins within 2 kb of higher scoring bins were excluded. The probes for the ChIP-string array were designed from the top and bottom parts of the sorted list for either gain or loss of signal, respectively, between NTM and TM cells that pass all of the above criteria. Bins that presented technical problems for probe design were replaced with the next possible bin from the corresponding sorted list. The genomic coordinates of all selected regions for each mark used in the final design of the ChIP-string array are listed in Table S2.

ChIP-String Data Analysis

Raw probes values were first normalized by the same method as the one used by Ram et al. (Ram et al., 2011). Counts for all probes of each sample were then compared to negative controls, and samples in which greater than 90% of probes were at or below background level counts (based on inbuilt negative controls) were omitted from further analysis. Counts derived from each ChIP sample were then normalized as follows. Probes for each individual sample were divided by the median count within the sample, then each probe was divided by the median value of that probe across all samples. The mean and standard deviation were calculated per sample. The mean value per sample was subtracted from each probe and then each probe was divided by the standard deviation. The resulting values were subsequently used for the analysis.

Aggregate Plots

Genome-wide coverage files (bigWig) for each H3K27Ac experiment was generated by using bamCoverage function of deepTools (Ramirez et al., 2016), with Reads Per Kilobase per Million mapped reads (RPKM) normalization. Then, obtained coverage tracks used for aggregate plots of H3K27Ac levels around -/+ 2Kb of de-acetylated promoters with visualization tool – ChAsE (Younesy et al., 2016).

Calculation of IC_{50} Values

Cells were plated in 96-well plates and treated in six replicate wells. The images were obtained and confluence was calculated by the Incucyte machine (Essen biosciences) and associated software. The confluence data was then used for calculation of drug response. Drug-response data was adjusted to a four-parameter log-logistic function using the R package drc. IC_{50} were predicted using the derived model. The area under the curve (AUC) was obtained by numerical integration of cell viability in function of dose (log10 scale) using Bolstad R package.

Mass Spectrometry

Total histones were prepared and subject to mass spectrometry analysis as previously described (Karch et al., 2014) using the LTQ-Orbitrap Velos Pro (Thermo Scientific).

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Cell line based model of melanoma progression and epigenome profiling, Related to Figure 1. (A) Validation of PTEN loss by PTEN shRNA by western blot**. (B-C)** Soft agar colony formation ability of NTM_H, TM_H, NTM_P and TM_P cells. Panel B shows representative image and panel C shows the quantitation of soft-agar colonies. **(D-E)** Matrigel-invasion ability of NTM_H, TM_H, NTM_P and TM_P cells. Panel D shows representative image of invaded cells post Boyden chamber assay and panel E shows the quantitation of invaded cells. **(F)** Log_2ratio between NTM_P and TM_P cells for the average signal strength of each chromatin mark in a window of 2kb around annotated transcription start sites from RefSeq (Blue) and on distal DNaseI hypersensitive sites from 'Melano' cell lines (Red, See Supplementary Methods) from ENCODE. **(G-I)** Measurement of global levels of histone modification marks in NTM_H and TM_H cells. (G-H) Mass Spectrometry based quantitation of various histone marks on histone H3 (G) or histone H4 (H). X-axis shows peptide identity whereas Y-axis shows relative abundance. (I) Western blot analysis for indicated histone marks from acid-extracted histones from NTM_H and TM_H cells.

Figure S1

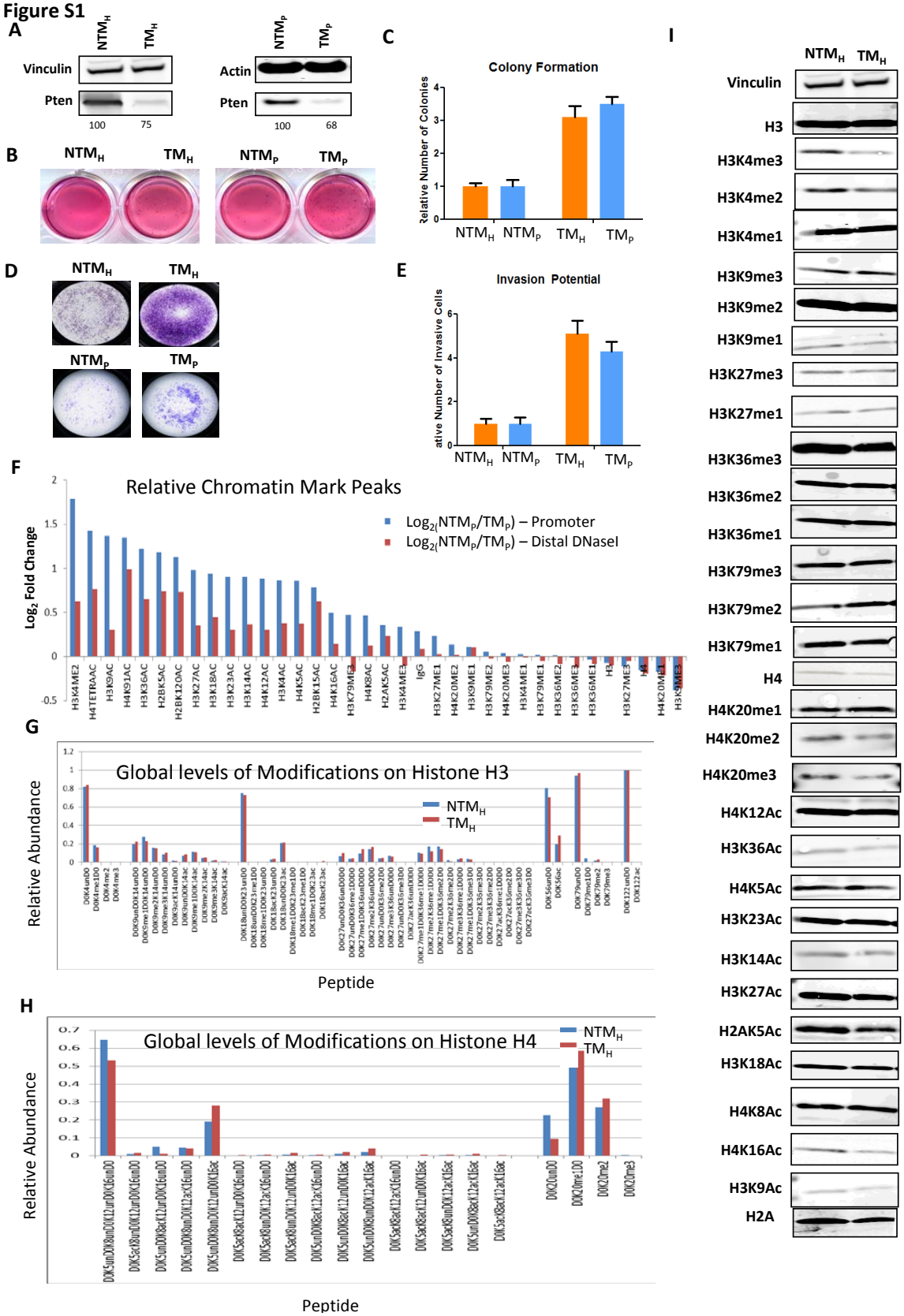

A
NTM_H
TM_H
Vinculin
Pten
100
75
NTM_P
TM_P
Actin
Pten
100
68
B
NTM_H
TM_H
NTM_P
TM_P
C
Colony Formation
Relative Number of Colonies
NTM_H
NTM_P
TM_H
TM_P
D
NTM_H
TM_H
NTM_P
TM_P
E
Invasion Potential
ative Number of Invasive Cells
NTM_H
NTM_P
TM_H
TM_P
F
Relative Chromatin Mark Peaks
Log2(NTMP/TMP) – Promoter
Log2(NTMP/TMP) – Distal DNaseI
Log2 Fold Change
H3K4ME2
H4TETRAAC
H3K9AC
H4K91AC
H3K36AC
H2BK5AC
H2BK120AC
H3K27AC
H3K18AC
H3K23AC
H3K14AC
H4K12AC
H3K4AC
H4K5AC
H2BK15AC
H4K16AC
H3K79ME2
H4K8AC
H2AK5AC
H3K4ME3
IgG
H3K27ME1
H4K20ME2
H3K9ME1
H3K79ME2
H4K20ME3
H3K4ME1
H3K79ME1
H3K36ME2
H3K36ME3
H3K36ME1
H3
H3K27ME3
H4
H4K20ME1
H3K9ME3
G
Global levels of Modifications on Histone H3
NTM_H
TM_H
Relative Abundance
Peptide
H
Global levels of Modifications on Histone H4
NTM_H
TM_H
Relative Abundance
Peptide
I
NTM_H
TM_H
Vinculin
H3
H3K4me3
H3K4me2
H3K4me1
H3K9me3
H3K9me2
H3K9me1
H3K27me3
H3K27me1
H3K36me3
H3K36me2
H3K36me1
H3K79me3
H3K79me2
H3K79me1
H4
H4K20me1
H4K20me2
H4K20me3
H4K12Ac
H3K36Ac
H4K5Ac
H3K23Ac
H3K14Ac
H3K27Ac
H2AK5Ac
H3K18Ac
H4K8Ac
H4K16Ac
H3K9Ac
H2A

Figure S2. Validation of chromatin changes in human tumors, Related to Figure 2.

(A-F) Correlation plots between ChIP-Seq and ChIP-String. Plots showing correlations of normalized mark intensity in ChIP-Seq experiment (Y-axis) and ChIP-String experiment (X-axis) in NTM_H and TM_H cells for H2BK5Ac (A), H4K5Ac (B), H3K27me3 (C), H3K27Ac (D), H3K4me1 (E) and H3K4me3 (F). **(G-J)** Boxplots showing average normalized intensity for ChIP-string probes across NTM_H, TM_H, nevi and tumors individually for H2BK5Ac probes high in NTM_H cells (G), H2BK5Ac probes high in TM_H cells (H), H4K5Ac probes high in NTM_H cells (I) and H4K5Ac probes high in T_H cells (J).

Figure S2

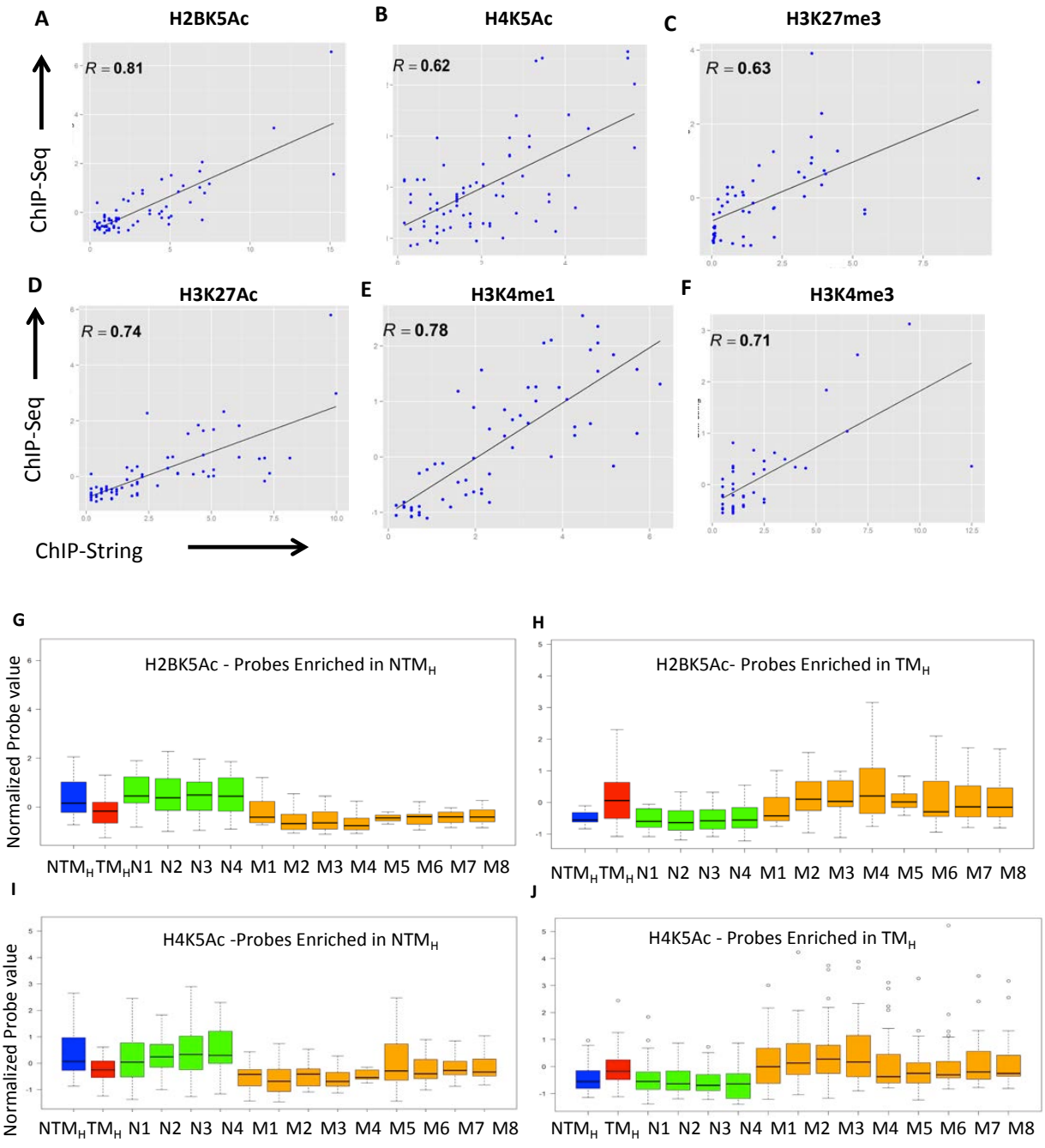

A
H2BK5Ac
R = 0.81
B
H4K5Ac
R = 0.62
C
H3K27me3
R = 0.63
D
H3K27Ac
R = 0.74
E
H3K4me1
R = 0.78
F
H3K4me3
R = 0.71
ChIP-Seq
ChIP-Seq
ChIP-String
G
H2BK5Ac - Probes Enriched in NTM_H
H
H2BK5Ac- Probes Enriched in TM_H
I
H4K5Ac -Probes Enriched in NTM_H
J
H4K5Ac - Probes Enriched in TM_H
Normalized Probe value
Normalized Probe value
NTM_H TM_H N1 N2 N3 N4 M1 M2 M3 M4 M5 M6 M7 M8

Figure S3: Chromatin state profiles, Related to Figure 3 (A) Correlation plot showing pearson correlations of histone modification peaks between the histone marks profiled in our study in NTM_H cells computed based on encoding the presence of a peak with a value of 1 and the absence of a peak with a value of 0. Peak calling was performed using MACS algorithm with default parameters except p-value < $1x10^{-8}$. 'DiffBind' bioconductor package was used to cluster correlation values. **(B)** Transition parameters for 18-state model derived by ChromHMM for NTM_H and TM_H cells. **(C-D)** Overlap of different genomic features (CpG island, RefSeq TSS, RefSeq TES, laminB lads (Guelen et al., 2008), 5-hMeC enriched and 5-MeC enriched regions) with chromatin state calls in NTM_H (C) and TM_H (D) cells. The fold enrichments are calculated as the ratio between observed and expected number of genomic bins for each overlap. The color intensities are normalized within each column between its minimum value (white) and its maximum value (blue). The last column shows the mean DNA methylation level for each chromatin state on the scale from completely unmethylated (white) to fully methylated (red). **(E)** Overlap enrichment of TSS coordinates and then gene body of highly expressed (FPKM >5) and low/not expressed (FPKM <5) genes in NTM_H cells with chromatin states. Next to them is a gene expression positional plot that shows the average gene expression per chromatin state and cell type at a given distance within 50kb of annotated transcription start sites.

Figure S3

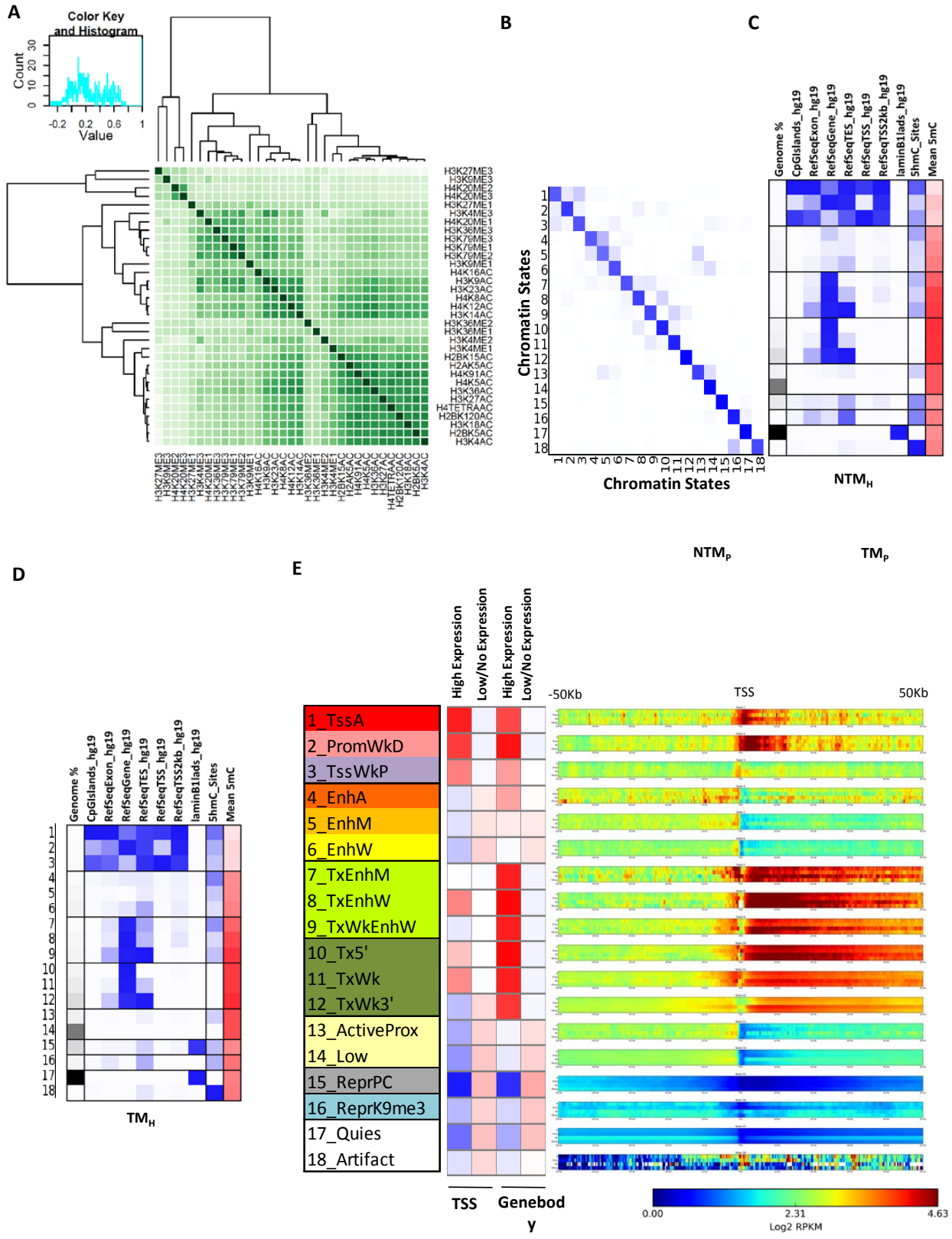

A
Color Key and Histogram
Count
Value
B
Chromatin States
C
Genome %
CpGIslands_hg19
RefSeqExon_hg19
RefSeqGene_hg19
RefSeqTES_hg19
RefSeqTSS_hg19
RefSeqTSS2kb_hg19
laminB1lads_hg19
5hmC_Sites
Mean 5mC
NTM_H
NTM_P
TM_P
D
TM_H
E
High Expression
Low/No Expression
1_TssA
2_PromWkD
3_TssWkP
4_EnhA
5_EnhM
6_EnhW
7_TxEnhM
8_TxEnhW
9_TxWkEnhW
10_Tx5'
11_TxWk
12_TxWk3'
13_ActiveProx
14_Low
15_ReprPC
16_ReprK9me3
17_Quies
18_Artifact
-50Kb
TSS
50Kb
TSS
Genebody
0.00
2.31
4.63
Log2 RPKM

Figure S4. Chromatin state transitions between non-tumorigenic and tumorigenic cells, Related to Figure 3. (A) Log_2 ratios between the total number of genomic bins occupied by each chromatin state in non-tumorigenic (NTM_H) and tumorigenic (TM_H) cells. **(B)** Heat map showing state transitions in NTM_H and TM_H cells with raw enrichment scores. **(C-E)** Heat maps showing under different normalization schemes fold enrichment of transitions of chromatin states in non-tumorigenic to tumorigenic cells controlling for the overall state size and similarity (**see Supplementary Methods**). The color intensities above the main diagonal range from white (relative enrichment <1) to blue (relative enrichment > 20), thus indicating chromatin state transitions that lose acetylation marks from non-tumorigenic to tumorigenic cells within the same category are more enriched compared to the reverse chromatin state transition (i.e. from tumorigenic to non-tumorigenic). Similarly, the colors below the main diagonal range from white (relative enrichment < 1) to red (relative enrichment > 20), thus indicating the lack of chromatin state transitions that gain acetylation marks from non-tumorigenic to tumorigenic cells within each category that are more enriched compared to the reverse chromatin state transition (i.e. from tumorigenic to non-tumorigenic). **(C)** Relative enrichments for NTM_H vs. TM_H with binary peak calls normalized to the same number. **(D)** Relative enrichments for NTM_H vs. TM_H with sequencing reads downsampled to same number. **(E)** Relative enrichments for NTM_P vs. TM_P with binary peak calls normalized to the same number.

Figure S4

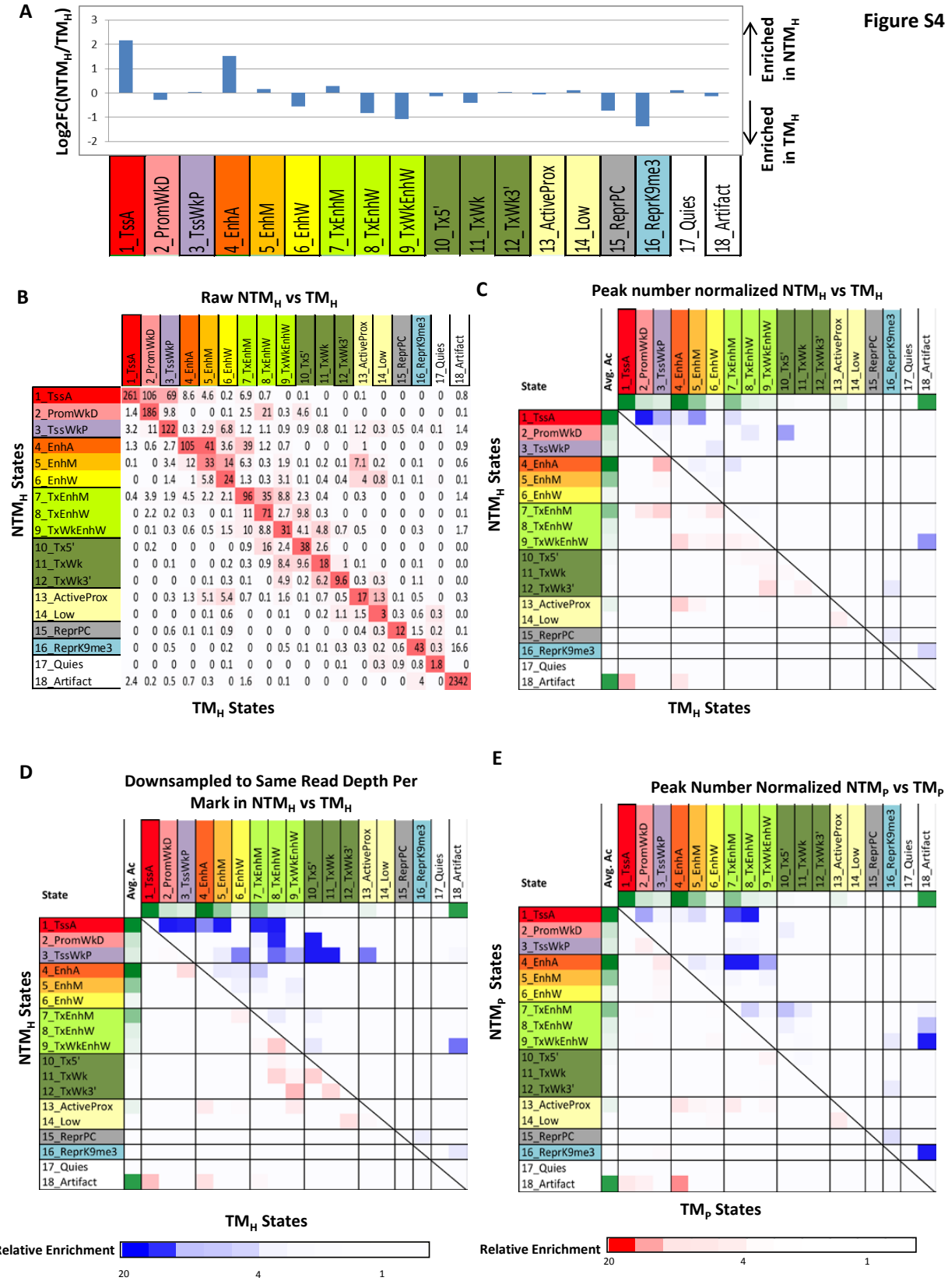
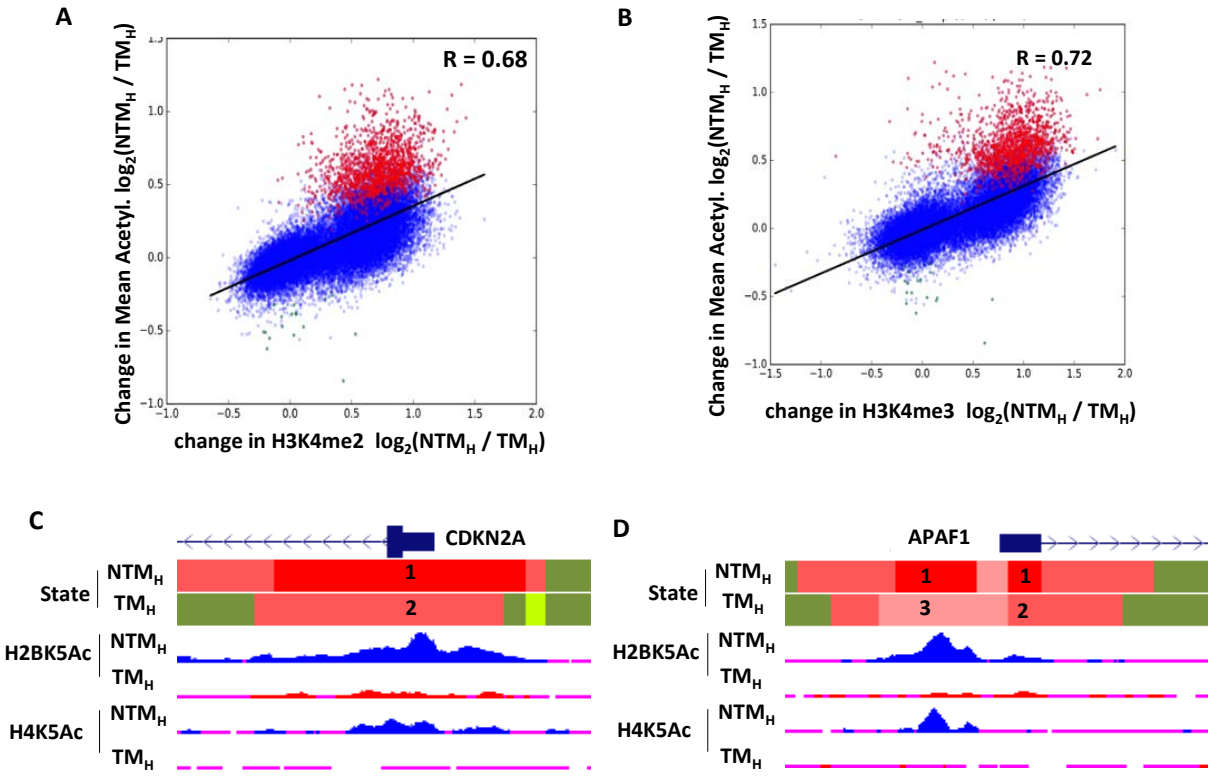

A
Log2FC(NTM_H/TM_H)
Enriched in NTM_H
Enriched in TM_H
1_TssA
2_PromWkD
3_TssWkP
4_EnhA
5_EnhM
6_EnhW
7_TxEnhM
8_TxEnhW
9_TxWkEnhW
10_Tx5'
11_TxWk
12_TxWk3'
13_ActiveProx
14_Low
15_ReprPC
16_ReprK9me3
17_Quies
18_Artifact
B
Raw NTM_H vs TM_H
NTM_H States
TM_H States
C
Peak number normalized NTM_H vs TM_H
NTM_H States
TM_H States
D
Downsampled to Same Read Depth Per Mark in NTM_H vs TM_H
NTM_H States
TM_H States
Relative Enrichment
20
4
1
E
Peak Number Normalized NTM_P vs TM_P
NTM_P States
TM_P States
Relative Enrichment
20
4
1

Figure S5. Chromatin state changes mark specific cancer pathways, Related to Figure 4. (A-B) The graphs show correlation between log_2 fold change in mean acetylation and (A) H3K4me2 or (B) H3K4me3. Log_2 fold change in mean acetylation (Y-axis) for a particular RefSeq promoter was plotted against log_2 fold change in H3K4me2 or H3K4me3 signal (X-axis) in that promoter for NTM_H vs. TM_H. Log_2 fold changes were calculated as $log_2((1 + \text{signal in } NTM_H) / (1 + \text{signal in } TM_H))$. Overall these changes in H3K4me2/3 correlate highly (R = 0.68 for H3K4me2 and 0.72 for H3K4me3) with alteration in mean acetylation suggesting that these marks function as coregulators. Points in red indicate promoters that were called as significantly deacetylated in TM_H at FDR of 1% by the permutation test in our analysis. **(C-D)** UCSC genome browser track for chromatin state and histone acetylations H2BK5Ac and H4K5Ac on genomic loci encompassing CDKN2A (A) and APAF1 (B) in NTM_H and TM_H cells. **(C-D)** Top enriched pathways (pathway commons) associated with genes closest to enhancers displaying state transitions from State 4_EnhA in non-tumorigenic cells (NTM_H) to States 7_TxEnhM and 5_EnhM in tumorigenic (TM_H) cells.

Figure S5



E

State 4_EnhA (NTM_H) to 7_TxEnhM (TM_H)

Pathway	Hyper FDR
Integrin family cell surface interactions	2.54E-29
VEGF and VEGFR signaling network	1.65E-29
Beta1 integrin cell surface interactions	1.73E-29
ErbB receptor signaling network	2.22E-29
LKB1 signaling events	2.86E-29
Signaling events mediated by VEGFR1 and VEGFR2	2.66E-29
Glypican 1 network	3.48E-29
Proteoglycan syndecan-mediated signaling events	3.55E-29
PDGF receptor signaling network	3.29E-29
IFN-gamma pathway	4.51E-29

F

State 4_EnhA (NTM_H) to 5_EnhM(TM_H)

Pathway	HyperFDR
Beta1 integrin cell surface interactions	8.48E-31
Integrin family cell surface interactions	4.98E-31
Proteoglycan syndecan-mediated signaling events	7.29E-31
Glypican pathway	2.83E-30
ErbB receptor signaling network	7.80E-30
VEGF and VEGFR signaling network	8.29E-30
TRAIL signaling pathway	1.16E-29
Glypican 1 network	1.34E-29
LKB1 signaling events	1.58E-29
PAR1-mediated thrombin signaling events	2.15E-29

Figure S6. Comparative analysis of chromatin changes with RNA expression changes, Related to Figure 5. (A) Relative enrichment (in log space) of number of steady genes that do not change expression between NTM_H and TM_H cells for all pairs of chromatin state transitions in their promoters. **(B)** Difference in enrichment of downregulated genes and upregulated genes on all pairs of chromatin state transitions between NTM_H and TM_H cells. **(C-E)** Relative enrichment of number of downregulated genes (C) or upregulated genes (D) or steady genes that do not change expression between NTM_P and TM_P cells (E) for all pairs of chromatin state transitions. **(F)** Difference in enrichment of downregulated genes and upregulated genes on all pairs of chromatin state transitions between NTM_P and TM_P cells. **(G)** Percent of genes showing down- (orange) or up- (blue) regulated gene expression change in top 100, 250, 500, 1000 and 2000 genes with differential levels in either direction between NTM_H and TM_H cells. Similarly, percent of the promoters showing gain (red) or loss (green) of acetylation in top 100, 250, 500, 1000 and 2000 promoter regions with differential levels between NTM_H and TM_H cells. **(H)** Scatter plot displays $\log_2$(fold change + 1) for acetylation and gene expression changes between NTM_H and TM_H. The color scheme is same as that in Figure 5D.

Figure S6

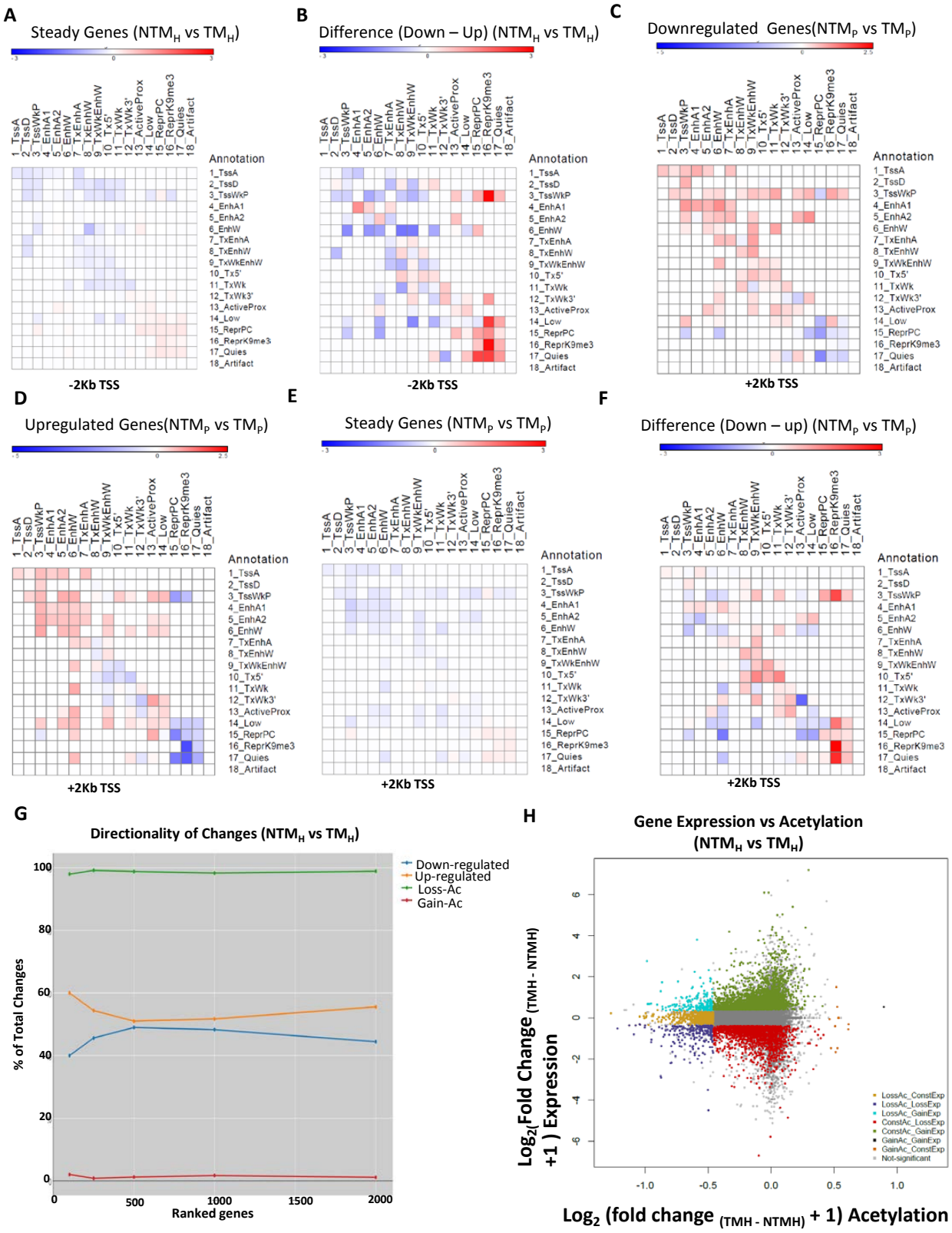
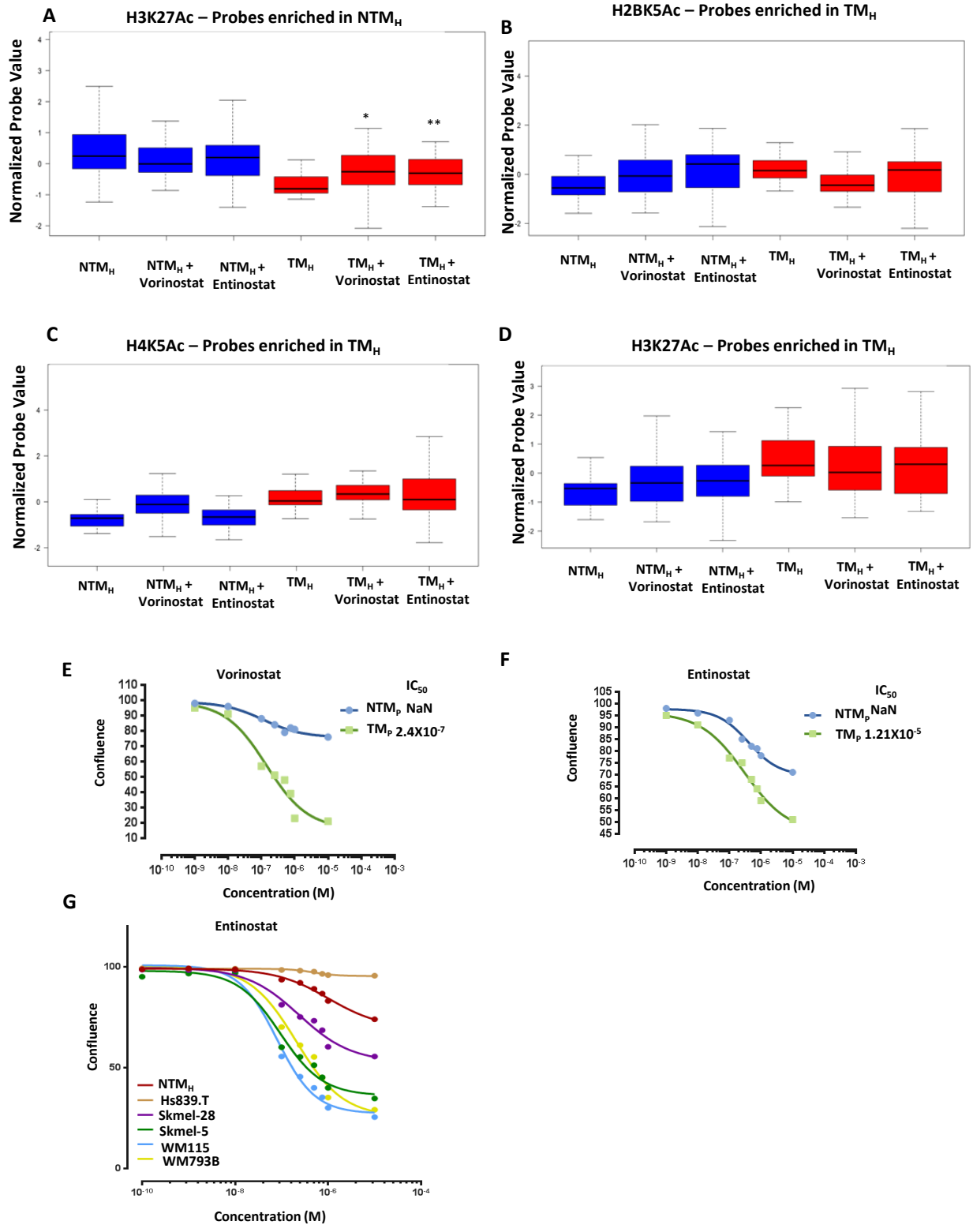

A
Steady Genes (NTM$_H$ vs TM$_H$)
B
Difference (Down – Up) (NTM$_H$ vs TM$_H$)
C
Downregulated Genes(NTM$_P$ vs TM$_P$)
D
Upregulated Genes(NTM$_P$ vs TM$_P$)
E
Steady Genes (NTM$_P$ vs TM$_P$)
F
Difference (Down – up) (NTM$_P$ vs TM$_P$)
Annotation
1_TssA
2_TssD
3_TssWkP
4_EnhA1
5_EnhA2
6_EnhW
7_TxEnhA
8_TxEnhW
9_TxWkEnhW
10_Tx5'
11_TxWk
12_TxWk3'
13_ActiveProx
14_Low
15_ReprPC
16_ReprK9me3
17_Quies
18_Artifact
-2Kb TSS
+2Kb TSS
G
Directionality of Changes (NTM$_H$ vs TM$_H$)
Down-regulated
Up-regulated
Loss-Ac
Gain-Ac
% of Total Changes
Ranked genes
H
Gene Expression vs Acetylation
(NTM$_H$ vs TM$_H$)
Log$_2$(Fold Change (TMH - NTMH) +1) Expression
Log$_2$ (fold change (TMH - NTMH) + 1) Acetylation
LossAc_ConstExp
LossAc_LossExp
LossAc_GainExp
ConstAc_LossExp
ConstAc_GainExp
GainAc_GainExp
GainAc_ConstExp
Not-significant

Figure S7. Acetylation status and proliferation changes in response to HDAC inhibitors, Related to Figure 7. (A) Boxplots showing average normalized intensity for H3K27Ac levels on ChIP-string probes (that are enriched in NTM_H cells in ChIP-seq studies) across NTM_H and TM_H cells that were either untreated or treated with vorinostat (200nM) or entinostat (300nM) for 72 hrs. **(B-D)** Boxplots showing average normalized intensity for (B), H2BK5Ac, (C), H4K5Ac, or (D) H3K27Ac levels on ChIP-string probes that are enriched in T_H cells across NTM_H and TM_H cells treated with vorinostat (200nM) or entinostat (300nM) for 72hrs. Asterisk (*) represents $p<0.05$ and double asterisk (**) represents $p<0.001$ (Wilcoxon Rank test) when comparisons are made to TM_H. **(E-F)** Growth curves for NTM_H and TM_H cells grown under various concentrations of (E) vorinostat or (F) entinostat. IC_{50} values are also shown. NaN refers to 'not a number'. **(G)** Growth curves for melanoma cell lines grown under various concentrations of entinostat. IC_{50} values are in Figure 7G.

Figure S7


A
H3K27Ac – Probes enriched in NTM_H
Normalized Probe Value
NTM_H
NTM_H + Vorinostat
NTM_H + Entinostat
TM_H
TM_H + Vorinostat
TM_H + Entinostat
B
H2BK5Ac – Probes enriched in TM_H
C
H4K5Ac – Probes enriched in TM_H
D
H3K27Ac – Probes enriched in TM_H
E
Vorinostat
IC50
NTM_P NaN
TM_P 2.4X10-7
Confluence
Concentration (M)
F
Entinostat
NTM_P NaN
TM_P 1.21X10-5
G
Entinostat
NTM_H
Hs839.T
Skmel-28
Skmel-5
WM115
WM793B

SUPPLEMENTARY TABLES

Table S1. Details of sequencing data generated in this study, Related to Figure 1. Total read numbers for each mark in each of the 4 cell types used in this study, NTM_H, TM_H, NTM_P and TM_P.

Table S2. Details of probe locations used for ChIP-String, Related to Figure 2. Genomic location (hg19) of the 96-probes used in the nanostring codeset.

Table S3. Details of the nevi and tumor samples, Related to Figure 2. Clinical and genetic data for the nevi and tumor samples that were used for validation of histone modification levels used in Figure 2.

Table S4. Chromatin state recovery with subsets of marks. Related to Figure 3. Top panel shows fraction of state assignments recovered of the state of the row with only the mark of the column compared to using all the marks in NTM_H and TM_H cells. We observed cases of high recovery (>60%) of acetylated enhancer or promoter states with a single acetylation mark. Bottom panel shows fraction of state recovery with all marks except the mark of the column in NTM_H and TM_H cells compared to using all marks. We observed four cases of low recovery (<60%) of a chromatin states when all marks except one mark were included highlighting the existence of many locations uniquely marked by one mark. These four cases of low recovery were the chromatin states 6_EnhW, 10_Tx5', 15_ReprPC, and 16_ReprK9me3 when excluding the H3K4me1, H3K79me2, H3K27me3, and H3K9me3 marks respectively (Table S4).

Table S5. GO terms for all state changes from NTM_H to TM_H, Related to Figure 4.

Lists of GO-terms for the regions that belong to all significant chromatin state changes between NTM_H and TM_H cells.

Table S6: Pathways enriched for top 2 promoter and enhancer state transitions from NTM_H to TM_H, Related to Figure 4.

Table S7: Pathway analysis for different groups in overlap of gene expression and average acetylation, Related to Figure 5.

REFERENCES

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. Epigenomics *3*, 771-784.

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics *7*, 191.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature methods *9*, 215-216.

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature biotechnology *33*, 364-376.
Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z.*, et al.* (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Molecular cell *47*, 810-822.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W.*, et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948-951.
Karch, K.R., Zee, B.M., and Garcia, B.A. (2014). High resolution is not a strict requirement for characterization and quantification of histone post-translational modifications. J Proteome Res *13*, 6152-6159.

Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome biology *13*, R107.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods *7*, 709-715.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M.*, et al.* (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628-1639.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research *44*, W160-165.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology *31*, 46-53.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., and Liu, S. (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics *28*, 729-730.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M.*, et al.* (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research *38*, e178.
Younesy, H., Nielsen, C.B., Lorincz, M.C., Jones, S.J., Karimi, M.M., and Moller, T. (2016). ChAsE: chromatin analysis and exploration tool. Bioinformatics.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

CHAPTER 5

CSDELTA: Systematic detection of differential chromatin sites from group-wise comparisons of multiple ChIP-seq maps

ABSTRACT

Comparing epigenetic marks between samples has emerged as a useful way to characterize regulatory elements in the genomes of living cells from high-throughput chromatin data. Yet, detecting differential chromatin sites in large epigenomics datasets is currently challenging due to the lack of suitable bioinformatics tools that can capture the combinatorial complexity of epigenetic marks at the resolution of single nucleosomes. In this work we present CSDELTA, a general method for genome-wide comparison of epigenomic maps between groups with multiple samples. CSDELTA models the functional similarity of different types of chromatin state domains, which can increase the power to detect real changes. Moreover, the method can detect chromatin changes at nucleosome level resolution. We show applications of the method on comparing human embryonic stem cells and brain tissues, which demonstrate the biological relevance of predicted differential sites and the superior performance of CSDELTA relative to existing methods.

INTRODUCTION

Epigenetic regulation of genes as manifested by dynamic patterns of post-translational histone modifications plays important role in normal development [1–7] and disease [8–12]. Experimental protocols such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) have enabled mapping histone modifications on a genome-wide scale in a large number of cell types and conditions [13]. Large consortia efforts [14–16] and individual studies [17, 11, 18] have produced an unprecedented amount of histone mark data that can be leveraged to uncover the underlying mechanisms and principles of epigenetic control in biological processes.

While early on single histone marks have been found to correlate with specific genomic features such as gene expression [19, 20], combinatorial presence of multiple histone marks has been implicated to be important in a variety of biological processes across a large spectrum of cell

types and tissues [13–15, 21–23]. For instance, the combination of H3K4me3 and H3K27ac marks is found at active promoters, whereas the joint presence of H3K4me3 and H3K27me3 associates with poised or bivalent promoters [24]. As another example, the H3K36me3 mark is known to be deposited within the gene bodies of actively transcribed genes [13], whereas H3K36me3 in combination with H3K9me3 is preferentially found within the bodies of zinc finger genes and at repetitive regions [13, 25]. In contrast, broad H3K9me3 domains without H3K36me3 are found at constitutive heterochromatic regions that lack gene expression [26].

To summarize and understand the combinatorial complexity of the histone code, the concept of chromatin states has been introduced and computational methods such as ChromHMM [27] and Segway [28] have been developed and applied for chromatin state discovery and genome-wide annotation. Subsequently, methods have been proposed that leverage dependencies between histone marks not only within each cell type, but also across cell types [29, 30]. The output of these methods enables comparisons of chromatin datasets between cell types, tissues and conditions that can capture combinatorial changes of histone marks and thus improve our understanding about the dynamics of regulatory mechanisms of gene expression.

Pairwise comparisons between epigenomic maps are complicated both by technical noise in the data and by unknown relationships between chromatin marks. Currently, only few methods exist that model joint changes of multiple marks between conditions. One such method, dPCA [31], performs Principal Component Analysis on a de-noised version of the difference in the signal for each chromatin mark between two conditions. The method can be applied only on a predefined set of regions such as promoters or previously mapped transcription factor binding sites. While useful in such settings, individual differential principle components can be difficult to interpret and applying dPCA genome-wide at higher resolutions is not computationally feasible. A second method, ChromDiff [32], was proposed for pairwise comparison of chromatin maps between pairs of conditions with multiple samples. In contrast to dPCA, ChromDiff operates on chromatin state segmentations produced by methods such as ChromHMM by pooling information across a predefined set of relatively broad genomic regions such as gene

bodies. While useful to detect differences on the level of genes, ChromDiff is prone to missing important epigenetic changes at sites not included in the input set or at finer resolution (e.g. at the level of individual enhancers or even nucleosomes). Furthermore, this method does not model the similarity between different chromatin states, which can impact its power to detect real changes. Another tool, EpiCompare [33], was developed to detect tissue-specific enhancers and promoters at higher resolution. While useful in this setting, EpiCompare is limited only to these two types of regulatory elements and out-of-the-box can process only data from the Roadmap Epigenomics Consortium, thus requiring substantial manual adaptation of the source code to be applicable in more general settings.

Here we present CSDELTA, a general method for genome-wide comparison of chromatin state segmentations between pairs of conditions with multiple samples. Our method presents several improvements over limitations of different existing methods. First, CSDELTA can be applied to find epigenetic changes in an unbiased manner across the whole genome without the need to specify a predefined set of input regions. Second, CSDELTA operates at the resolution of the input chromatin segmentations, which is typically 200 base pairs or one nucleosome and spacer region, thus enabling detection at much finer resolution. Third, CSDELTA models the functional similarity between chromatin states, which can increase the power to detect true changes. Fourth, the software is not tailored to specific datasets, provides an easy to use command line interface and runs in reasonable time and memory.

CSDELTA takes as input two groups of chromatin state segmentations and produces a ranked list of differential chromatin locations with respect to each chromatin state. Furthermore, CSDELTA provides an estimate of the false discovery rate (FDR) for each differential site, that can be derived from one of two possible background models. In cases where at least four samples are available in each group, a background model with sufficient power can be built by randomizing chromatin state maps between the groups. This procedure makes the assumption that samples are independent representatives of their groups. In cases where fewer samples are available per group (as few as two samples per group) and the chromatin maps were derived with

ChromHMM[27], a background model can be built based on the additional assumption that histone mark datasets used to derive the chromatin state segmentations are independent measurements of the chromatin states in each sample.

METHODS

CSDELTA is designed for genome-wide comparisons of chromatin state segmentations produced by ChromHMM or similar methods between two groups with multiple samples at the resolution of the input segmentations (**Fig 5.1**). The output of CSDELTA is a ranked list of genomic loci ordered by a quantitative measure of how differential each location is between the two groups with respect to each chromatin state. In addition, CSDELTA builds a background model for the distribution of differential scores under the null hypothesis that there are no differences between the two groups and estimates the FDR for each locus with respect to each state differential score.

Differential scores

CSDELTA ranks all genomic locations by the degree of change between two input groups with respect to each chromatin state as quantified by a single number, further referred to as *state differential score*. Intuitively, the state differential score captures the likelihood that particular genomic location is annotated with a chromatin state and this state is over- or under-represented among the samples from the first group compared to the second group. Furthermore, the score attempts to account for functional similarity between chromatin states so that transitions between similar states rank lower than more substantial chromatin changes.

Computing the state differential scores requires a formal notion of functional similarity between chromatin states. CSDELTA leverages information from multiple samples within each group to learn a distance function between states. The basic idea is that two states are likely to be more functionally similar if genomic locations annotated with one of them are annotated frequently

with the other state in samples from the same group. In contrast, two chromatin states are more different when genomic locations annotated with one of them are rarely annotated with the other in samples from the same group. Formally, this can be expressed with the help of a graphical model (**Fig 5.1Aii**), which models the conditional probability distribution of observing a chromatin state at a given location in a sample, conditioned on observing another chromatin state at that location in another sample from the same group. Let *G* and *S* be variables that denote the group and the sample label, respectively. Let *B* denote the genomic position of a bin, and let *O* denote the observed chromatin state in sample *S* at position *B*. Then, let *R* denote the chromatin state in another randomly selected sample from the same group at the same genomic position. CSDELTA models the probability $P(R|S,O)$ directly without any further assumptions about the mechanics of the process that generated *O* and *R*. In particular, CSDELTA estimates $P(R|S,O)$ from the input data by computing for each sample the average co-occurrence frequencies for each pair of states conditioned on the total fraction of the genome occupied by the states in that sample, across all samples in the same group:

$$P(R=y|S=s,O=x)=\frac{1}{N-1}\sum_{\substack{r\in\mathbb{G}\\ r\neq s}}\frac{\#[y \text{ and } x \text{ are assigned to the same bin in } s \text{ and } r]}{\#[\text{bins annotated with } x \text{ in } s]}$$

where $\mathbb{G}$ denotes the set of all samples from the group and *N* denotes the number of samples. Intuitively, if x and y are more closely related states, then $P(R=y|S=s,O=x)$ will be different from zero, and vice versa for more distinct states. Next, CSDELTA computes the probability distribution of chromatin states at each position i in each group, g:

$$\begin{aligned}P(R=y|G=g,B=i)&\\ =\sum_{s\in\mathbb{G}}P(R=y,S=s|G=g,B=i)&=\sum_{s\in\mathbb{G}}P(R=y|S=s,G=g,B=i)P(S=s|G=g)\\ =\frac{1}{N_g}\sum_{s\in\mathbb{G}}P\big(R=y|S=s,O=x_{s,i}\big)&\end{aligned}$$

where $\mathbb{G}$ is the set of samples in group g, N_g is the number of samples in group g, and $x_{s,i}$ denotes the chromatin state in dataset s at position i. The above equality follows from the model assumption that every sample is equally likely to be generated in group g,and thus $P(S = s|G = g)$ is uniform. Importantly, $P(R = y|G = g, B = i)$ captures both the uncertainty about the chromatin state assignments at position i across all samples in group g and the similarity of chromatin states as measured by their genome-wide co-occurrence frequencies within group g. Then, the state differential score between two groups, g_1 and g_2, for each position i and state y is computed as:

$$\delta_{y,i} = P(R = y|G = g_1, B = i) - P(R = y|G = g_2, B = i)$$

State differential scores captures the degree to which the annotations at position i differ with respect to chromatin state y between the two groups. Furthermore, the sign of $\delta_{y,i}$ indicates which of the two groups exhibits higher frequency of state y, which allows to call group-specific chromatin state regions.

False-discovery rate computation

CSDELTA computes an estimation of the false discovery rate of the state differential scores for each genomic location based on permutation tests. The software provides two options to construct random groupings depending on the number of available samples and histone mark datasets in each group.

Permuting chromatin state segmentations between groups

This background model keeps the number of samples in each group and generates every permutation of the group assignments of individual samples. The power to detect significant changes with this procedure depends on the number of samples in each group, which in turn determines the total number of possible unique permutations. For example, significant changes

can be detected at FDR level of 0.01 if there are at least five samples in each group, whereas an FDR threshold of 0.05 would require each group to have at least four samples. Moreover, in cases where it is not computationally feasible to inspect all possible permutations, a random subset of 100 permutations is generated. For each permutation, the conditional distributions $P(R = y|S = s, O = x)$ for each sample s are computed by using the samples within its shuffled group. Then, differential scores between the two shuffled groups are computed for a random subset of 10 continuous regions of $5*10^6$ base pairs each.

Permuting histone marks across the two groups

For chromatin state segmentations derived with ChromHMM[27], in cases where only few samples are available in each group, CSDELTA provides an option to estimate the FDR based on a background model that attempts to capture the technical rather than the biological reproducibility of the findings. This procedure assumes that the technical noise is not correlated between ChIP-seq experiments and individual histone mark experiments can be treated as independent measurements of the epigenetic state of the corresponding samples. CSDELTA generates mock samples by shuffling the histone mark datasets across all samples from both groups. After each shuffling, the mock samples are partitioned into two groups so that the number of samples in each group is the same as in the original groups. Next, ChromHMM is applied with the "MakeSegmentation" option to produce chromatin state segmentations for the mock samples by using the model parameters learned from the original data. This procedure preserves the original chromatin state definitions and thus allows for computation of the background distribution of scores over the same set of chromatin states. For computational reasons, a random subset of 100 permutations and 10 randomly selected continuous regions of $5*10^6$ base pairs is used.

FDR estimation

Since state differential scores range from -1 to 1, where scores equal to 0 indicate no difference between the groups, CSDELTA tests whether the absolute value of $\delta_{y,i}$ is significantly different from 0:

$$\mathrm{FDR}\left(|\delta_{y,i}| = q\right) = \frac{\mathrm{Bgr}(|\delta_{y,i}| \geq q)}{\mathrm{Fgr}(|\delta_{y,i}| \geq q)}$$

where $\mathrm{Bgr}(|\delta_{y,i}| \geq q)$ denotes the fraction of the background scores for state y whose absolute values are greater than or equal to q, and $\mathrm{Fgr}(|\delta_{y,i}| \geq q)$ denotes the fraction of foreground scores for state y (i.e. scores from the true grouping) whose absolute values are greater than or equal to q. To ensure monotonicity, the reported FDR values are adjusted by using the Yekutieli-Benjamini procedure [34].

RESULTS

Comparison with ChromDiff and Fisher's exact test methods

We evaluated the performance of CSDELTA, ChromDiff [32] and a baseline method based on Fisher's exact test (FET) on a pairwise comparison of chromatin state maps of human embryonic stem (ES) cells (n=8) and brain cells (n=10) downloaded from the Roadmap Epigenomics Project [14].

In this comparison, ChromDiff was applied without its procedure for external covariate correction, which produced better results in our tests compared to using covariate correction. Since ChromDiff was designed to detect chromatin changes on more coarse resolutions, we applied this method genome-wide on bins of length 1mb, 100kb and 10kb and on 200bp bins from chromosome 19, further referred to as ChromDiff_1mb, ChromDiff_100kb, ChromDiff_10kb and ChromDiff_200bp/chr19, respectively. We chose chromosome 19 for

computational reasons and because it contains most zinc finger genes in the human genome, which are used as feature in our evaluation. For each state y and window length $w \in$ {1mb, 100kb, 10kb, 200bp}, we assigned to each 200bp bin, $j \in [i * w, (i+1) * w]$, the same score derived from ChromDiffs's output:

$$\delta_{\text{ChromDiff},y,w,j} = -\log_2\left(\text{qvalue}_{y,w,i}\right) * \text{sign}(\mu_{y,w,i,g_1} - \mu_{y,w,i,g_2})$$

where $\text{qvalue}_{y,w,i}$ denotes the FDR corrected P-value for state y at window i of length w from the Mann-Whitney test as outputted by ChromDiff and μ_{y,w,i,g_1} and μ_{y,w,i,g_2} denote the average ranks with respect to the frequency of state y in that window in group g_1 and g_2, respectively.

The FET method closely resembles one of the procedures for chromatin state comparison in EpiCompare [33], but generalizes to any chromatin state and not just promoters and enhancers as implemented in EpiCompare. This method constructs a two-by-two contingency table for each chromatin state at each 200bp genomic bin and computes a P-value by Fisher's exact test, which then is converted to a score to rank bins by the degree they differ between the two groups:

$$\delta_{\text{FET},y,i} = -\log_2\left(\text{pvalue}_{y,i}\right) * \text{sign}(\text{freq}(y,i,g_1) - \text{freq}(y,i,g_2))$$

where $\text{pvalue}_{y,i}$ denotes the P-value from the Fisher's exact test at bin i and $\text{freq}(y,i,g_1)$ and $\text{freq}(y,i,g_2)$ denote the frequency of state y at bin i in group g_1 and g_2, respectively.

Conditional probabilities learned by CSDELTA reflect expected chromatin state similarities and capture more chromatin state variability than ChromHMM posterior probabilities

Conditional probabilities learned by CSDELTA when comparing ES cells and brain tissues, $P(R = y|S = s, O = x)$, reflected expected chromatin state similarities (**Fig 5.2A**). For example,

on average, the majority of the states are most similar to themselves. In addition, promoter and enhancer states exhibited higher similarities among each other than to other chromatin states. Furthermore, broad weakly transcribed (5_TxWk), heterochromatin (9_Het) and weakly Polycomb repressed (14_ReprPCWk) states showed similarity with the low signal state (15_Quies). Moreover, we investigated whether co-occurrence frequencies across samples from the same group provide more information about chromatin state similarities than statistics that do not explicitly consider variability at the same genomic location across samples. For example, in ChromHMM, posterior probabilities of chromatin state assignments can be outputted for each genomic bin, which give the uncertainty of the model about state assignments at that bin. Based on these uncertainty estimates, the expected co-occurrence frequency of chromatin state assignments by ChromHMM at that bin in a replicate experiment is given by the outer product of the vector of posterior probabilities. In our results, the amount of chromatin state variability captured by the average conditional probabilities in CSDELTA was substantially more than the variability captured by the average outer product of ChromHMM posterior probabilities across the genome in all samples as measured by the Shannon entropy of these distributions. Therefore, this suggests that the co-occurrence frequencies of chromatin states in samples from the same group provide additional information about chromatin state similarities that is not captured by the uncertainty about chromatin state assignments estimated by ChromHMM from data in individual samples.

Comparison of the genome territory called as significantly differential by each method

We next computed for each chromatin state the fraction of genome that showed statistically significant differences at FDR of 0.05 for each method (**Fig 5.2B**). CSDELTA, ChromDiff_1mb, ChromDiff_100kb and ChromDiff_10kb called significant differences for all states, whereas FET and ChromDiff_200bp/chr19 called much fewer locations only for a small subset of the chromatin states. As ES and brain cells are expected to have many epigenetic differences, this suggests that FET and ChromDiff_200bp are substantially underpowered to detect differences on nucleosome level resolution likely due to insufficient number of samples in the two groups.

Furthermore, using different background models for computing FDR estimates in CSDELTA had very little effect on the amount of genomic territory called as significant.

Differential regions from CSDELTA and other methods associate with gene expression changes

We compared how rankings produced by CSDELTA, ChromDiff and FET correlate with gene expression changes between ES and brain cells (**Fig 5.3**). Since chromatin states have been shown to correlate with gene expression[35], rankings with respect to differential chromatin states are expected to reflect this association. In this context, we computed for each ranking the cumulative average gene expression change for bins within 2kb of all annotated transcription start sites. As a quantitative assessment of the overall association we computed the normalized area under the curve for each ranking. Since gene expression changes are typically reflected in changes of epigenetic marks in windows much larger than 200bp, we also included in this comparison smoothed CSDELTA scores by sliding a window of length 21 bins centered at each bin (i.e. 10 bins upstream and 10 bins downstream from the current bin) and convolving the signal in the window with a Gaussian density function with $\sigma = 21/6$. This procedure allows for leveraging information from CSDELTA scores in nearby bins. All methods correctly associated changes in active promoter states with changes in gene expression to different degrees. CSDELTA, with and without smoothing, outperforms both FET and ChromDiff_200bp/chr19 on chromosome 19 with respect to the normalized area under the curve for states associated with TSS proximal regions including 1_TssA, 2_TssAFlnk, 3_TxFlnk, 11_BivFlnk and for the Polycomb repressed state, 13_ReprPC. Genome-wide, ChromDiff_10kb and ChromDiff_100kb outperform non-smoothed CSDELTA scores for some states including 1_TssA. Smoothed CSDELTA scores overall show the best performance for this task across all states associated with TSS proximal regions. These results suggest that: 1) CSDELTA's model of similarity between chromatin states can increase the overall association of the detected changes with gene expression at 200bp resolution, and 2) pooling information across

neighboring bins as implemented in ChromDiff or by Gaussian smoothing can improve the strength of this association.

CSDELTA scores associate better with differential DNaseI hypersensitivity sites than ChromDiff and FET

We evaluated the performance of all three methods based on their ability to detect differential DNaseI hypersensitivity sites (DHS) (**Fig 5.4**). DHS are a proxy for mapping transcription factor binding activity, which is known to correlate with cell type specific gene regulation, particularly at enhancers [36]. For this comparison, we computed Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the task of discriminating distal ES-specific DHS from distal Brain-specific DHS based on the state differential scores for enhancer states (6_EnhG, 7_Enh and 12_EnhBiv) produced by each method. Distal sites were defined as sites that are farther than 2kb from annotated TSSs. In all cases, CSDELTA showed a substantial advantage over ChromDiff and FET as measured by the area under the ROC and PR curves. Smoothing of CSDELTA scores presented no advantage for this task and in fact slightly decreased the strength of association between the scores and distal differential DHS, thus implying that pooling chromatin state changes across broader regions is not beneficial in this case. Consequently, ChromDiff's poor performance on this task can be explained by the fact that this method was designed for broader domains and is thus not suitable for detection of changes in more localized genomic features such as DHS. In addition, both FET and ChromDiff do not model chromatin state similarity, which can further affect their power to detect true changes. Overall, these results imply that CSDELTA's model for chromatin state similarity, which leverages state co-occurrence patterns from samples within each group, can lead to improvements in accuracy when detecting localized chromatin changes at nucleosome level resolution.

CSDELTA scores associate better zinc finger genes than ChromDiff and FET

Finally, we evaluated the three methods on their ability to detect changes with respect to state 8_Znf/Rpts, which is associated with zinc finger genes and repeats (**Fig 5.5**). State 8_Znf/Rpts is defined by the joint presence of H3K9me3 and H3K36me3 and as such is difficult to detect from inspecting any of these histone marks in isolation. Here we computed ROC and PR curves for the task of discriminating zinc finger genes from the rest of the genome based ranking genomic bins by the absolute value of each score. Based on the area under the ROC and PR curves, CSDELTA outperformed both FET and ChromDiff genome-wide and on chromosome 19, where most zinc finger genes reside. As zinc fingers span broader domains, smoothing of CSDELTA scores slightly improves the strength of the association.

CSDELTA run time and memory usage

Comparing chromatin state maps of human embryonic stem (ES) cells (n=8) and brain cells (n=10) downloaded from the Roadmap Epigenomics Project [14] and computing statistical significance by permuting chromatin state maps between groups took 1 hour and 57 minutes on a 2.7 GHz Intel Core i7 quad-core MacBook Pro laptop with 16GB RAM by using all four CPU cores.

DISCUSSION

In this work we presented a new computational method, CSDELTA, for group-wise comparison of chromatin state segmentations. We benchmarked CSDELTA against two existing methods, ChromDiff [32] and Fisher exact test, which resembles a procedure implemented in EpiCompare [37]. CSDELTA outperformed the other two methods at detecting differential sites at nucleosome resolution, which correlate with transcription factor binding activity and associate with zinc finger genes. CSDELTA was able to detect chromatin state changes associated with gene expression changes at comparable accuracy to the other methods.

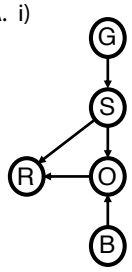
CSDELTA provides two options to estimate the statistical significance of differential sites based on permutation tests, which can be useful in different scenarios. The first option shuffles the chromatin state segmentations between the groups. This mode is preferable in cases where groups have sufficient number of samples, because it makes fewer assumptions about the input data and thus FDR estimates can be more accurate. Specifically, this procedure only assumes that the variability of chromatin state segmentations due to technical reasons is not correlated between samples, and thus each sample can be treated as an independent representative of its original group. While useful, this strategy can have a potential drawback as each group is required to have a sufficient number of samples in order to compute accurate FDR estimates in practice. For example, in cases where fewer than four samples per group are available, the number of possible permutations is too small and computed estimates can become overly conservative.

The second option in CSDELTA to estimate FDR is to randomize the input histone mark datasets used to derive the original chromatin state segmentations. Since there are typically multiple histone mark datasets per sample, this procedure can generate a sufficient number of randomized samples for the permutation test. However, this gain in power comes at the cost of additional assumptions, namely that the technical variability between individual histone mark datasets within each sample is not correlated. This assumption can be violated in practice, for example, if ChIP-seq quality of datasets from the same biological sample is affected either by genomic features unique to that sample such as chromatin accessibility or by technical artifacts due to batch effects. As result, in cases of large batch effects or other unaccounted confounders, FDR estimates computed by shuffling histone mark datasets can produce inflated false positive rates. In such cases, it is generally preferable to flag problematic datasets beforehand and exclude them from further analysis if possible or correct for the corresponding confounders before deriving chromatin state segmentations. Overall, shuffling ChIP-seq datasets can still provide useful estimates of the FDR in cases where only few samples are available for each group. Future work entails comprehensive comparison of the performance of each background model.

In addition, I plan on extending the CSDELTA method to comparisons of more than two groups of samples. In particular, the state differential scores can be extended to compare one group against multiple other groups, which can be used to detect for example cell type and tissue specific regulatory regions from large compendiums of data such as the Roadmap Epigenomics Project [14] and others [16, 23]. In this context, the FDR estimation procedures can also be adapted to handle multi-group comparisons.

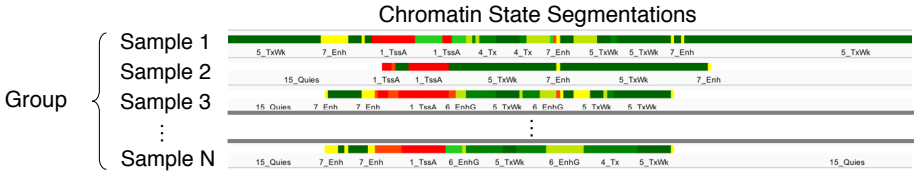
Figure 5.1

A. i)



G: Group
S: Sample
O: State in current sample
R: State in another sample
B: Bin index

ii)



Compute state co-occurrence frequencies within the group:

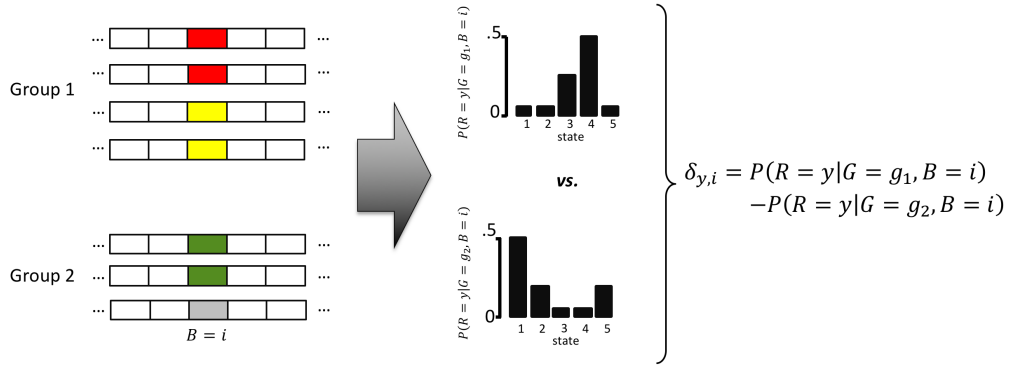
$P(R=y \mid S=s, O=x)$
$s \in \text{Samples(Group)}$

R=y

O=x \ State	1_TssA	2_TssAFlnk	3_TxFlnk	4_Tx	5_TxWk	6_EnhG	7_Enh	8_ZNF/Rpts	9_Het	10_TssBiv	11_BivFlnk	12_EnhBiv	13_ReprPC	14_ReprPCWk	15_Quies
1_TssA	69	9	0	0	4	0	4	0	0	5	1	0	1	1	5
2_TssAFlnk	22	37	1	0	6	1	21	0	0	2	2	1	1	1	5
3_TxFlnk	11	12	35	6	11	11	8	1	0	1	1	0	1	1	2
4_Tx	0	0	0	64	30	2	0	1	0	0	0	0	0	0	3
5_TxWk	0	0	0	8	57	0	5	0	1	0	0	0	0	2	25
6_EnhG	0	1	3	22	27	27	15	0	0	0	0	0	0	1	3
7_Enh	1	3	0	1	20	1	43	0	0	0	0	1	1	4	25
8_ZNF/Rpts	1	0	0	8	13	0	1	47	20	0	0	0	0	1	8
9_Het	0	0	0	0	3	0	0	3	42	0	0	0	1	4	47
10_TssBiv	22	3	0	0	1	0	1	0	0	40	12	6	9	4	2
11_BivFlnk	12	8	0	0	1	0	4	1	0	24	27	12	7	2	1
12_EnhBiv	3	2	0	0	4	0	12	0	0	6	7	24	24	12	6
13_ReprPC	1	0	0	0	3	0	2	0	2	2	1	5	47	26	11
14_ReprPCWk	0	0	0	0	5	0	2	0	2	0	0	0	5	38	47
15_Quies	0	0	0	0	4	0	1	0	2	0	0	0	0	4	88

iii)

$$P(R=y|G=g_1,B=i)=\frac{1}{N_1}\sum_{s\in G_1}P(R=y|S=s,O=x_{s,i})$$



$$\delta_{y,i}=P(R=y|G=g_1,B=i)-P(R=y|G=g_2,B=i)$$

$$P(R=y|G=g_2,B=i)=\frac{1}{N_2}\sum_{s\in G_2}P(R=y|S=s,O=x_{s,i})$$

B.

Figure 5.1: **Overview of CSDELTA method**. **(A) (i)** Graphical model for the CSDELTA method. Variable explanations are shown under the model. The probability of observing a state in another sample from the same group, $P(R|S,O)$, is modelled directly without additional assumptions about the process that generates the chromatin states. **(ii)** In the first step of the method, $P(R|S,O)$ is estimated from co-occurrence frequencies for each pair of chromatin states among samples within each group. **(iii)** In the second step, for each genomic bin, *B*, and each group, *G*, $P(R|G,B)$ is computed as the average across the conditional probabilities given the observed states in the samples from the group. The state differential scores are defined as the difference between the $P(R|G,B)$ probabilities for each state in each group. **(B)** Example CSDELTA output from comparing ES cells and brain tissues from the Roadmap Epigenomics Project[14] at the NANOG locus. ChromHMM segmentations are shown for ES cells at the top and brain tissues at the bottom. The tracks in the middle show the mnemonics, color codes and the state differential scores for all states. Several genes in this locus are differentially active in ES cells, which is reflected in the differential scores for specific chromatin states. For example, the promoter and the gene body of the NANOG gene, which is a key ES-specific regulator, show high positive differential scores for the promoter state, 1_TssA, and the transcribed states, 4_Tx and 5_TxWk, respectively.

Figure 5.2

A.

Average P(R = y | S = s, O = x)

State	1_TssA	2_TssAFlnk	3_TxFlnk	4_Tx	5_TxWk	6_EnhG	7_Enh	8_ZNF/Rpts	9_Het	10_TssBiv	11_BivFlnk	12_EnhBiv	13_ReprPC	14_ReprPCWk	15_Quies	Entropy
1_TssA	69	8.7	0.5	0.1	4.2	0.1	3.9	0.4	0.2	4.6	1.1	0.5	0.6	1.4	4.8	1.80
2_TssAFlnk	22	37	1	0.1	5.9	0.5	21	0.2	0.2	2.1	2.2	1	0.6	1.5	5.1	2.53
3_TxFlnk	11	12	35	5.6	11	11	8	0.9	0.3	0.7	0.5	0.4	0.6	0.9	2.1	2.90
4_Tx	0	0	0.1	64	30	1.5	0.4	0.8	0.2	0	0	0	0.1	0.2	2.8	1.32
5_TxWk	0.3	0.2	0	8.2	57	0.4	4.8	0.3	0.7	0	0	0.1	0.2	2.2	25	1.77
6_EnhG	0.4	1.2	2.6	22	27	27	15	0.3	0.1	0	0	0.1	0.2	0.6	3.1	2.43
7_Enh	1.2	3	0.1	0.5	20	1.1	43	0.1	0.3	0.1	0.1	0.7	0.6	4	25	2.17
8_ZNF/Rpts	1.4	0.2	0.1	7.7	13	0.2	0.5	47	20	0.2	0.2	0.1	0.6	0.7	8.2	2.24
9_Het	0.1	0	0	0.2	3.1	0	0.3	2.7	42	0	0	0	0.5	3.8	47	1.61
10_TssBiv	22	3.2	0.1	0	0.9	0	1.5	0.3	0.2	40	12	5.9	9	3.7	1.7	2.57
11_BivFlnk	12	7.5	0.1	0	0.9	0	3.8	0.7	0.2	24	27	12	7.5	2.4	1.4	2.84
12_EnhBiv	2.6	2.5	0.1	0.1	3.5	0.1	12	0.1	0.3	6.3	7.4	24	24	12	5.5	2.98
13_ReprPC	0.7	0.3	0	0.1	2.6	0	1.7	0.3	2.1	2.1	1	4.7	47	26	11	2.22
14_ReprPCWk	0.2	0.1	0	0.1	5.4	0	2	0.1	1.9	0.1	0	0.4	5.2	38	47	1.81
15_Quies	0.1	0	0	0.1	4.4	0	1	0	1.9	0	0	0	0.2	3.9	88	0.76

Average Outer Product of ChromHMM Posterior Probabilities

1_TssA	2_TssAFlnk	3_TxFlnk	4_Tx	5_TxWk	6_EnhG	7_Enh	8_ZNF/Rpts	9_Het	10_TssBiv	11_BivFlnk	12_EnhBiv	13_ReprPC	14_ReprPCWk	15_Quies	Entropy
90	4.6	0.2	0	0.9	0	1.2	0.1	0.1	1.5	0.1	0	0	0.3	0.9	0.71
12	76	0.7	0	0.2	0.2	10	0	0	0.2	0.7	0.1	0	0.1	0.1	1.19
6.3	11	73	0.9	0.5	7.3	0.9	0.7	0	0	0.1	0	0	0	0	1.43
0	0	0	83	16	0.8	0	0.3	0	0	0	0	0	0	0.1	0.75
0.1	0	0	3.3	76	0.1	2.7	0.1	0.2	0	0	0	0	1.2	16	1.14
0.1	0.4	1	11	8.4	65	13	0.2	0	0	0	0	0	0	0.1	1.58
0.3	1.1	0	0	12	0.8	78	0	0.1	0	0	0.3	0	1.2	6.7	1.17
0.3	0	0	2.7	4.1	0.1	0.1	81	11	0	0.1	0	0.1	0.1	0.5	1.07
0	0	0	0	1	0	0.1	1	72	0	0	0	0.1	1.7	25	1.09
9	0.6	0	0	0	0	0.1	0.1	0	77	7.6	1.9	3.1	0.8	0.1	1.28
1.4	4	0	0	0	0	0.2	0.3	0	18	72	4.3	0.2	0.1	0	1.33
0.2	0.3	0	0	0.3	0	5.7	0.1	0	2.1	1.8	69	15	5.3	0.2	1.57
0	0	0	0	0.2	0	0.1	0.1	0.4	0.5	0	2.3	78	18	0.5	1.00
0	0	0	0	2.5	0	0.6	0	0.7	0	0	0.1	2.2	67	27	1.27
0	0	0	0	2.7	0	0.3	0	0.9	0	0	0	0	2.4	94	0.44

B.

% Genome significant at FDR of 0.05

State	ESC Average Genome %	Brain Average Genome %	CSDELTA/ shuffled marks	CSDELTA/ shuffled segmentations	FET	ChromDiff_200bp / chr19	ChromDiff_10kb	ChromDiff_100kb	ChromDiff_1mb
1_TssA	0.85	0.86	0.53	0.51	0.00	0.04	0.95	10.61	22.58
2_TssAFlnk	0.23	0.47	0.56	0.48	0.00	0.02	1.10	14.66	40.37
3_TxFlnk	0.04	0.02	0.05	0.03	0.00	0.00	0.04	0.72	7.64
4_Tx	3.27	2.55	0.88	0.85	0.00	0.00	0.44	4.15	9.96
5_TxWk	11.65	12.48	7.16	7.48	0.18	0.13	2.45	15.55	32.21
6_EnhG	0.13	0.22	0.21	0.12	0.00	0.00	0.13	3.17	16.61
7_Enh	2.93	2.61	4.19	4.57	0.09	0.03	5.43	23.53	28.48
8_ZNF/Rpts	0.36	0.17	0.03	0.04	0.00	0.00	0.14	2.89	21.58
9_Het	3.97	1.85	1.14	1.16	0.00	0.00	0.96	7.07	9.79
10_TssBiv	0.27	0.10	0.35	0.37	0.00	0.00	0.49	7.28	38.10
11_BivFlnk	0.10	0.06	0.01	0.08	0.00	0.00	0.09	1.79	7.52
12_EnhBiv	0.20	0.09	0.30	0.40	0.00	0.00	0.38	4.37	22.90
13_ReprPC	0.94	0.71	0.60	0.99	0.00	0.04	0.84	7.25	25.31
14_ReprPCWk	4.00	6.72	5.19	5.84	0.05	0.02	0.94	8.78	27.24
15_Quies	71.07	71.08	10.32	11.70	3.24	0.11	4.40	16.41	28.75

Figure 5.2: **CSDELTA conditional distributions and differential regions from comparison with CSDELTA, ChromDiff and FET.** ES cells and brain tissues from the Roadmap Epigenomics Project[14] were compared with CSDELTA, ChromDiff and FET. **(A)** Left table shows the average estimated conditional distributions, $P(R|S,O)$, across all samples in both groups. Right table shows the average outer product of ChromHMM posterior probabilities across all bins in the genome and all samples in both groups. The average outer product of ChromHMM posterior probabilities estimates the chromatin state similarities from the uncertainty about chromatin state assignments in individual samples without considering data from other samples at the same location. Probabilities in both tables were converted to percentages and color coded from 0 (white) to 100 (red). Last column in each table shows the Shannon entropy for the corresponding distributions with cells color coded from lowest value (white) to the highest value across all entropies from both tables (green). Distributions learned by CSDELTA have higher entropies and thus capture more of the variability of chromatin state assignments compared to the average outer product of ChromHMM posterior probabilities. **(B)** First two columns show the average percentage of genomic territory occupied by each state in ES cells and in brain tissues. Subsequent columns show for each method the percentage of genomic territory called as significantly differential with respect to each state at FDR of 0.05. Cells are color coded from 0 (white) to the highest value in the whole table (green).

Figure 5.3

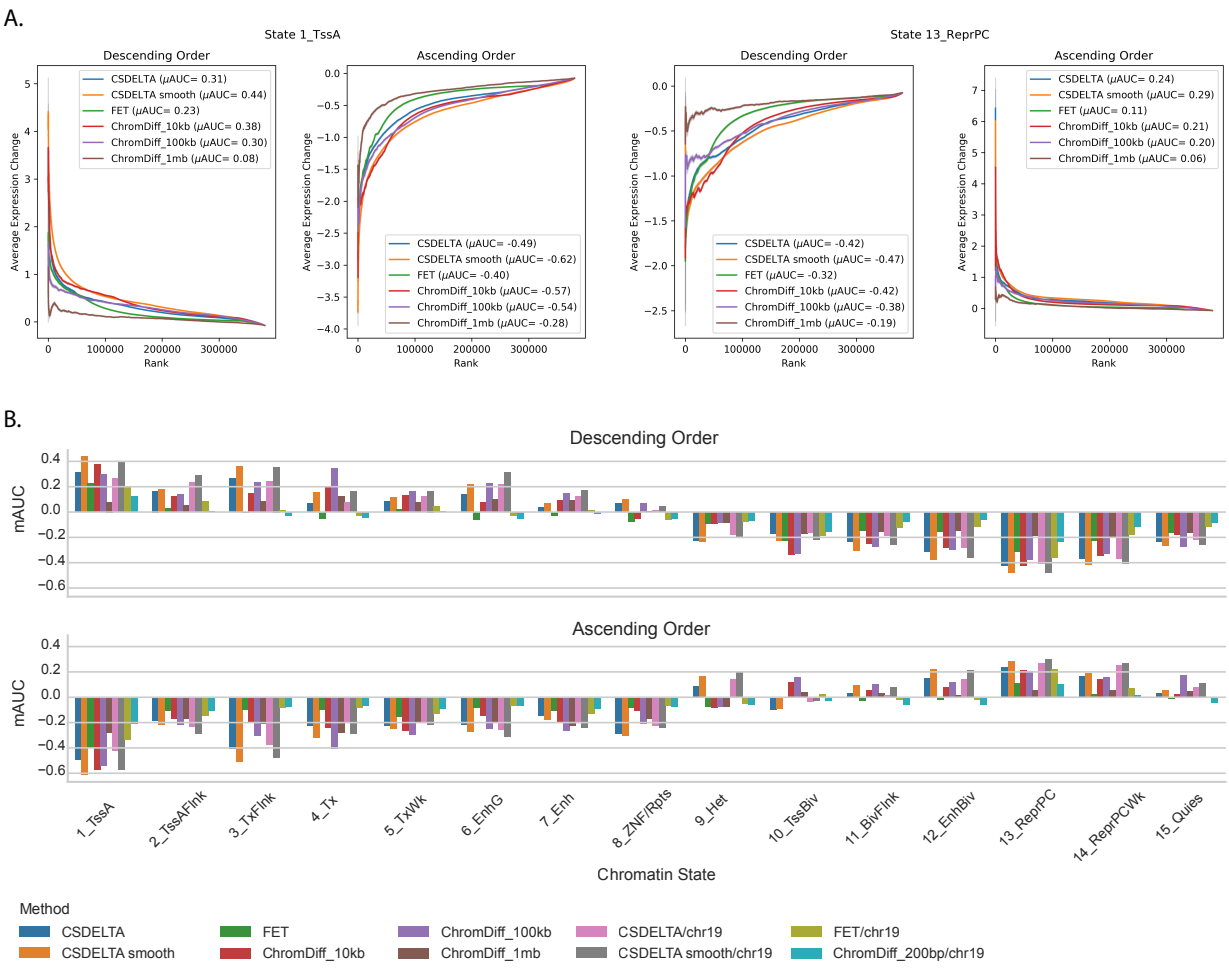

A.
State 1_TssA
State 13_ReprPC
Descending Order
Ascending Order
Average Expression Change
Rank
CSDELTA (μAUC= 0.31)
CSDELTA smooth (μAUC= 0.44)
FET (μAUC= 0.23)
ChromDiff_10kb (μAUC= 0.38)
ChromDiff_100kb (μAUC= 0.30)
ChromDiff_1mb (μAUC= 0.08)
CSDELTA (μAUC= -0.49)
CSDELTA smooth (μAUC= -0.62)
FET (μAUC= -0.40)
ChromDiff_10kb (μAUC= -0.57)
ChromDiff_100kb (μAUC= -0.54)
ChromDiff_1mb (μAUC= -0.28)
CSDELTA (μAUC= -0.42)
CSDELTA smooth (μAUC= -0.47)
FET (μAUC= -0.32)
ChromDiff_10kb (μAUC= -0.42)
ChromDiff_100kb (μAUC= -0.38)
ChromDiff_1mb (μAUC= -0.19)
CSDELTA (μAUC= 0.24)
CSDELTA smooth (μAUC= 0.29)
FET (μAUC= 0.11)
ChromDiff_10kb (μAUC= 0.21)
ChromDiff_100kb (μAUC= 0.20)
ChromDiff_1mb (μAUC= 0.06)
B.
Descending Order
Ascending Order
mAUC
1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
Chromatin State
Method
CSDELTA
CSDELTA smooth
FET
ChromDiff_10kb
ChromDiff_100kb
ChromDiff_1mb
CSDELTA/chr19
CSDELTA smooth/chr19
FET/chr19
ChromDiff_200bp/chr19

Figure 5.3: **State differential scores associate with gene expression changes.** ES cells and brain tissues from the Roadmap Epigenomics Project [14] were compared with CSDELTA, ChromDiff and FET. **(A)** For chromatin states 1_TssA and 13_ReprPC, the cumulative average gene expression change between ES and brain cells is plotted as a function of the rank of bins within 2kb of annotated TSSs based on the differential score from each method sorted in descending order (left panels) and ascending order (right panels). The area under the curve normalized by the number of bins (μAUC) is shown for each method. μAUC measures the overall association of rankings with gene expression changes with positive values corresponding to increase of gene expression compared to brain cells and vice versa for negative values. **(B)** μAUC is shown for each chromatin state and method for scores sorted in descending order (top panel) and ascending order (bottom panel).

Figure 5.4

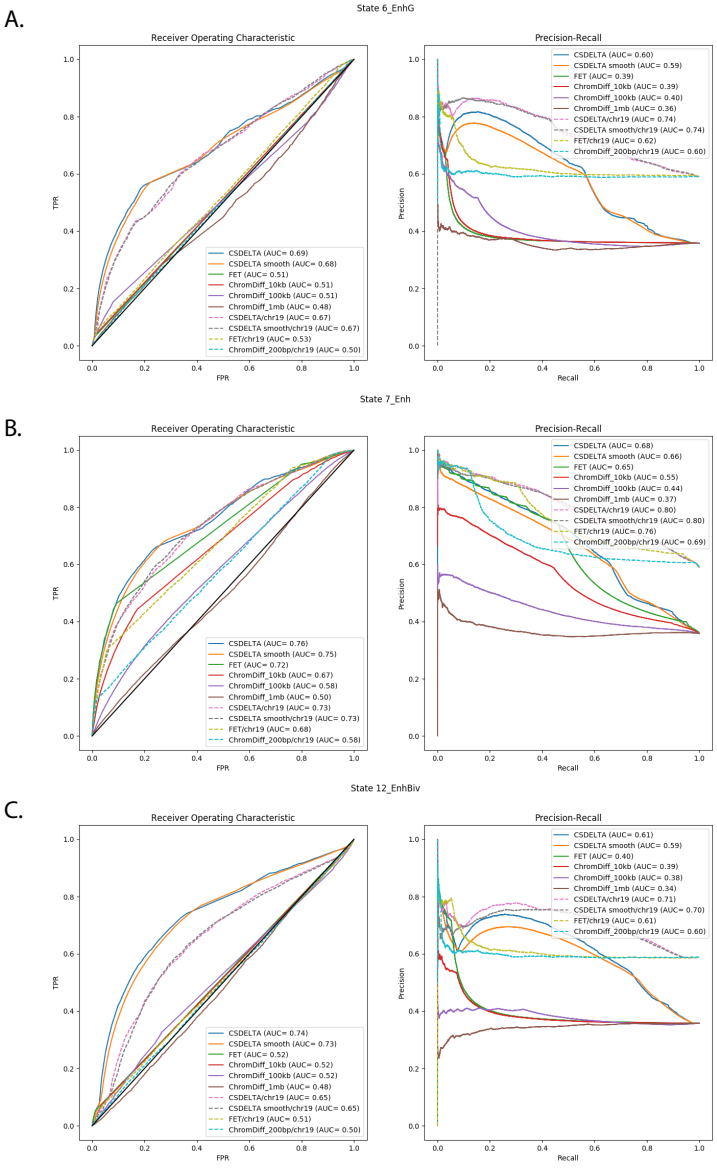

A.
State 6_EnhG
Receiver Operating Characteristic
TPR
FPR
CSDELTA (AUC= 0.69)
CSDELTA smooth (AUC= 0.68)
FET (AUC= 0.51)
ChromDiff_10kb (AUC= 0.51)
ChromDiff_100kb (AUC= 0.51)
ChromDiff_1mb (AUC= 0.48)
CSDELTA/chr19 (AUC= 0.67)
CSDELTA smooth/chr19 (AUC= 0.67)
FET/chr19 (AUC= 0.53)
ChromDiff_200bp/chr19 (AUC= 0.50)
Precision-Recall
Precision
Recall
CSDELTA (AUC= 0.60)
CSDELTA smooth (AUC= 0.59)
FET (AUC= 0.39)
ChromDiff_10kb (AUC= 0.39)
ChromDiff_100kb (AUC= 0.40)
ChromDiff_1mb (AUC= 0.36)
CSDELTA/chr19 (AUC= 0.74)
CSDELTA smooth/chr19 (AUC= 0.74)
FET/chr19 (AUC= 0.62)
ChromDiff_200bp/chr19 (AUC= 0.60)
B.
State 7_Enh
Receiver Operating Characteristic
TPR
FPR
CSDELTA (AUC= 0.76)
CSDELTA smooth (AUC= 0.75)
FET (AUC= 0.72)
ChromDiff_10kb (AUC= 0.67)
ChromDiff_100kb (AUC= 0.58)
ChromDiff_1mb (AUC= 0.50)
CSDELTA/chr19 (AUC= 0.73)
CSDELTA smooth/chr19 (AUC= 0.73)
FET/chr19 (AUC= 0.68)
ChromDiff_200bp/chr19 (AUC= 0.58)
Precision-Recall
Precision
Recall
CSDELTA (AUC= 0.68)
CSDELTA smooth (AUC= 0.66)
FET (AUC= 0.65)
ChromDiff_10kb (AUC= 0.55)
ChromDiff_100kb (AUC= 0.44)
ChromDiff_1mb (AUC= 0.37)
CSDELTA/chr19 (AUC= 0.80)
CSDELTA smooth/chr19 (AUC= 0.80)
FET/chr19 (AUC= 0.76)
ChromDiff_200bp/chr19 (AUC= 0.69)
C.
State 12_EnhBiv
Receiver Operating Characteristic
TPR
FPR
CSDELTA (AUC= 0.74)
CSDELTA smooth (AUC= 0.73)
FET (AUC= 0.52)
ChromDiff_10kb (AUC= 0.52)
ChromDiff_100kb (AUC= 0.52)
ChromDiff_1mb (AUC= 0.48)
CSDELTA/chr19 (AUC= 0.65)
CSDELTA smooth/chr19 (AUC= 0.65)
FET/chr19 (AUC= 0.51)
ChromDiff_200bp/chr19 (AUC= 0.50)
Precision-Recall
Precision
Recall
CSDELTA (AUC= 0.61)
CSDELTA smooth (AUC= 0.59)
FET (AUC= 0.40)
ChromDiff_10kb (AUC= 0.39)
ChromDiff_100kb (AUC= 0.38)
ChromDiff_1mb (AUC= 0.34)
CSDELTA/chr19 (AUC= 0.71)
CSDELTA smooth/chr19 (AUC= 0.70)
FET/chr19 (AUC= 0.61)
ChromDiff_200bp/chr19 (AUC= 0.60)

Figure 5.4: **State differential scores for enhancers associate with differential DHS.** ES cells and brain tissues from the Roadmap Epigenomics Project [14] were compared with CSDELTA, ChromDiff and FET. Receiver operating characteristic and Precision-recall curves for enhancer states **(A)** 6_EnhG, **(B)** 7_Enh and **(C)** 12_EnhBiv for the task of discriminating distal ES-specific DHS from distal Brain-specific DHS based on the ranking of the scores from each method. Distal DHS sites are defined as DHS sites that are farther than 2 kb from annotated TSSs. The areas under each curve is shown for each method. Solid lines correspond to predicting tissue-specific DHS sites in the whole genome and dashed lines correspond to predicting tissue-specific DHS sites only on chromosome 19. CSDELTA without smoothing performed best compared to the rest of the methods on this task.

Figure 5.5

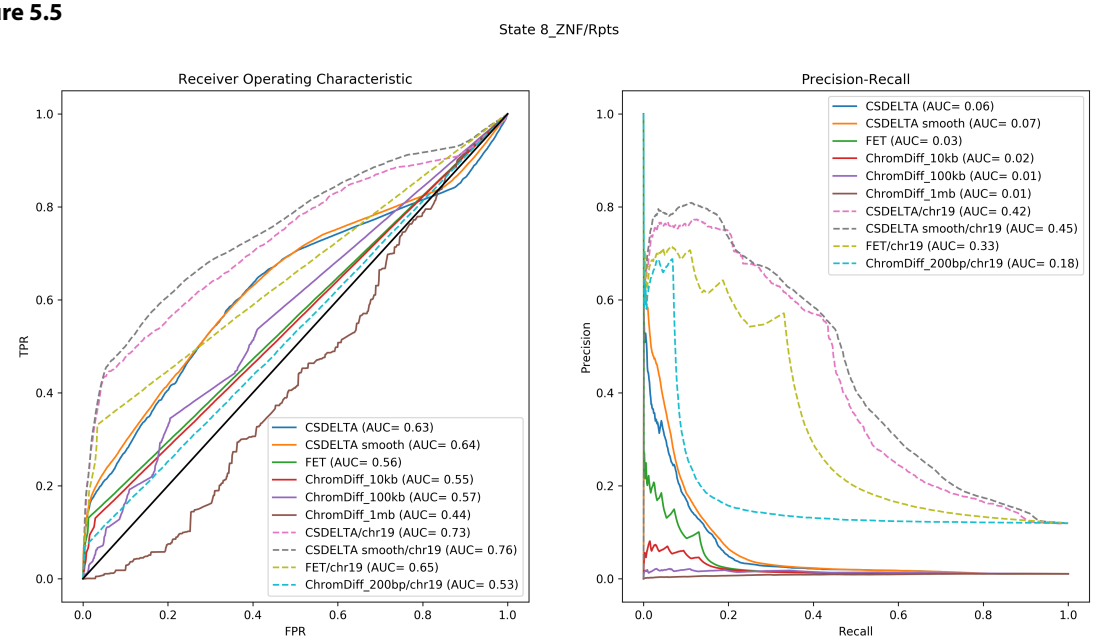

State 8_ZNF/Rpts
Receiver Operating Characteristic
TPR
FPR
CSDELTA (AUC= 0.63)
CSDELTA smooth (AUC= 0.64)
FET (AUC= 0.56)
ChromDiff_10kb (AUC= 0.55)
ChromDiff_100kb (AUC= 0.57)
ChromDiff_1mb (AUC= 0.44)
CSDELTA/chr19 (AUC= 0.73)
CSDELTA smooth/chr19 (AUC= 0.76)
FET/chr19 (AUC= 0.65)
ChromDiff_200bp/chr19 (AUC= 0.53)
Precision-Recall
Precision
Recall
CSDELTA (AUC= 0.06)
CSDELTA smooth (AUC= 0.07)
FET (AUC= 0.03)
ChromDiff_10kb (AUC= 0.02)
ChromDiff_100kb (AUC= 0.01)
ChromDiff_1mb (AUC= 0.01)
CSDELTA/chr19 (AUC= 0.42)
CSDELTA smooth/chr19 (AUC= 0.45)
FET/chr19 (AUC= 0.33)
ChromDiff_200bp/chr19 (AUC= 0.18)

Figure 5.5: **State differential scores associate with zinc finger genes.** ES cells and brain tissues from the Roadmap Epigenomics Project [14] were compared with CSDELTA, ChromDiff and FET. Receiver operating characteristic and Precision-recall curves for chromatin state 8_ZNF/Rpts for the task of discriminating zinc finger genes from the rest of the genome. Solid lines correspond to classifying zinc fingers genome-wide and dashed lines correspond to classifying zinc fingers only on chromosome 19. CSDELTA performed best compared to the rest of the methods on this task. Smoothing CSDELTA scores provides a slight advantage in this case.

REFERENCES

[1] J. A. Zhang, A. Mortazavi, B. A. Williams, B. J. Wold, and E. V Rothenberg, "Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity.," *Cell*, vol. 149, no. 2, pp. 467–82, Apr. 2012.

[2] J. Cotney *et al.*, "The evolution of lineage-specific regulatory activities in the human embryonic limb.," *Cell*, vol. 154, no. 1, pp. 185–96, Jul. 2013.

[3] S. K. Reilly *et al.*, "Evolutionary changes in promoter and enhancer activity during human corticogenesis," *Science (80-.).*, vol. 347, no. 6226, pp. 1155–1159, Mar. 2015.

[4] T. S. Mikkelsen *et al.*, "Comparative epigenomic analysis of murine and human adipogenesis.," *Cell*, vol. 143, no. 1, pp. 156–69, Oct. 2010.

[5] D. Lara-Astiaso *et al.*, "Chromatin state dynamics during blood formation," *Science (80-.).*, Aug. 2014.

[6] S. L. Paige *et al.*, "A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development.," *Cell*, vol. 151, no. 1, pp. 221–32, Sep. 2012.

[7] J. A. Wamstad *et al.*, "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage.," *Cell*, vol. 151, no. 1, pp. 206–20, Sep. 2012.

[8] K. K.-H. Farh *et al.*, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, no. 7539, pp. 337–343, Oct. 2014.

[9] E. Gjoneska *et al.*, "Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease," *Nature*, vol. 518, no. 7539, pp. 365–369, Feb. 2015.

[10] A. Gusev *et al.*, "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.," *Am. J. Hum. Genet.*, vol. 95, no. 5, pp. 535–52, Nov. 2014.

[11] P. Fiziev *et al.*, "Systematic Epigenomic Analysis Reveals Chromatin States Associated with Melanoma Progression.," *Cell Rep.*, vol. 19, no. 4, pp. 875–889, Apr. 2017.

[12] K. Chen *et al.*, "Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes.," *Nat. Genet.*, vol. 47, no. 10, pp. 1149–57, Oct. 2015.

[13] A. Barski *et al.*, “High-resolution profiling of histone methylations in the human genome.,” *Cell*, vol. 129, no. 4, pp. 823–37, May 2007.

[14] Roadmap_Epigenomics_Consortium *et al.*, “Integrative analysis of 111 reference human epigenomes,” *Nature*, vol. 518, no. 7539, pp. 317–330, Feb. 2015.

[15] ENCODE_Project_Consortium, “An integrated encyclopedia of DNA elements in the human genome.,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[16] J. H. A. Martens and H. G. Stunnenberg, “BLUEPRINT: mapping human blood cell epigenomes.,” *Haematologica*, vol. 98, no. 10, pp. 1487–9, Oct. 2013.

[17] F. D. Lay *et al.*, “Reprogramming of the human intestinal epigenome by surgical tissue transposition.,” *Genome Res.*, vol. 24, no. 4, pp. 545–53, Apr. 2014.

[18] S. Mei *et al.*, “Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D658–D662, Jan. 2017.

[19] H. Santos-Rosa *et al.*, “Active genes are tri-methylated at K4 of histone H3,” *Nature*, vol. 419, no. 6905, pp. 407–411, Sep. 2002.

[20] R. Cao *et al.*, “Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing,” *Science (80-.).*, vol. 298, no. 5595, pp. 1039–1043, Nov. 2002.

[21] Z. Wang *et al.*, “Combinatorial patterns of histone acetylations and methylations in the human genome.,” *Nat. Genet.*, vol. 40, no. 7, pp. 897–903, Jul. 2008.

[22] N. Nègre *et al.*, “A cis-regulatory map of the Drosophila genome.,” *Nature*, vol. 471, no. 7339, pp. 527–31, Mar. 2011.

[23] J. A. Stamatoyannopoulos *et al.*, “An encyclopedia of mouse DNA elements (Mouse ENCODE).,” *Genome Biol.*, vol. 13, no. 8, p. 418, Aug. 2012.

[24] B. E. Bernstein *et al.*, “A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells,” *Cell*, vol. 125, no. 2, pp. 315–326, Apr. 2006.

[25] J. Ernst and M. Kellis, “Discovery and characterization of chromatin states for systematic annotation of the human genome,” *Nat. Biotechnol.*, vol. 28, no. 8, pp. 817–825, Aug. 2010.

[26] M. Lachner, D. O’Carroll, S. Rea, K. Mechtler, and T. Jenuwein, “Methylation of histone

H3 lysine 9 creates a binding site for HP1 proteins," *Nature*, vol. 410, no. 6824, pp. 116–120, Mar. 2001.

[27] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization.," *Nat. Methods*, vol. 9, no. 3, pp. 215–6, Mar. 2012.

[28] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation.," *Nat. Methods*, vol. 9, no. 5, pp. 473–6, May 2012.

[29] J. Biesinger, Y. Wang, and X. Xie, "Discovering and mapping chromatin states using a tree hidden Markov model," *BMC Bioinformatics*, vol. 14, no. Suppl 5, p. S4, 2013.

[30] Y. Zhang, L. An, F. Yue, and R. C. Hardison, "Jointly characterizing epigenetic dynamics across multiple human cell types.," *Nucleic Acids Res.*, p. gkw278-, Apr. 2016.

[31] H. Ji, X. Li, Q.-F. Wang, and Y. Ning, "Differential principal component analysis of ChIP-seq.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 17, pp. 6789–94, Apr. 2013.

[32] A. Yen and M. Kellis, "Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type," *Nat. Commun.*, vol. 6, p. 7973, Aug. 2015.

[33] Y. He and T. Wang, "EpiCompare: an online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features," *Bioinformatics*, vol. 33, no. 20, pp. 3268–3275, Oct. 2017.

[34] D. Yekutieli and Y. Benjamini, "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics," *J. Stat. Plan. Inference*, vol. 82, no. 1–2, pp. 171–196, Dec. 1999.

[35] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types.," *Nature*, vol. 473, no. 7345, pp. 43–9, May 2011.

[36] R. E. Thurman *et al.*, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, pp. 75–82, Sep. 2012.

[37] Y. He and T. Wang, "EpiCompare: An online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features," *Bioinformatics*, Jun. 2017.