UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational approaches for elucidating the gene regulatory landscape of mammalian development

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

in

Bioinformatics and Systems Biology

by

Yuan Zhao

Committee in charge:

Professor Bing Ren, Chair
Professor Wei Wang, Co-Chair
Professor Joseph R. Ecker
Professor Christopher Glass
Professor Kun Zhang

2019

The Dissertation of Yuan Zhao is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2019

DEDICATION

I dedicate this dissertation to the men and women of science – past, present, and future – without whom this work would not be possible, and for whom, hopefully, this work will prove instructive.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Bing Ren for all of his support, guidance, and wisdom provided as the chair of my thesis committee. Dr. Ren has been without exaggeration a tremendous mentor, guiding me through learning to develop and research an independent project, write (and re-write) manuscripts, as well as providing invaluable career and professional advice. Without his generosity and mentorship, I would not be the man or the scientist I am today. Additionally, I would like to acknowledge my committee, who have also provided me with valuable guidance, scientific suggestions, and encouragement over the years.

I would also like to acknowledge the members of the Ren Lab at large, without whom my time as a graduate student would have been far less fulfilling. Besides the crucial scientific discussions, I am thankful for their support and the countless hours spent inside and outside of the lab. In particular, I want to acknowledge Dr. David Gorkin and Dr. Sebastian Preissl for their enduring support and mentorship, as well as genuine friendship. Additionally, I want to thank Dr. Yanxiao Zhang, Dr. Ramya Raviram, Dr. Naoki Kubo, Dr. Jian Yan, Rongxin Fang, Anugraha Raman, and Yunjiang Qiu. I thank Bernadeth Torres for always being helpful and available. Thank you all for your friendship.

Finally, I would like to acknowledge my good friends and family. My parents have been extremely supportive of my journey through academia as well as my sister and brother, both of whom have also gone through the trials of PhD life. To Dr. Justin Huang and Dr. Jenhan Tao, with whom I have had the good fortune of being roommates as well as fellow PhDs-in-training, I express gratitude for the years of camaraderie and useful discussions. I thank Tristin Liu for her love, patience, and kindness, in the best of times and the worst.

Chapters 2 and 3, in part, have been submitted for publication. DU Gorkin*, I Barozzi*, Y Zhao*, Y Zhang*, H Huang*, AY Lee, B Liu, J Chiou, A Wildberg, B Ding, B Zhang, M Wang, JS Strattan, JM Davidson, Y Qiu, V Afzal, JA Akiyama, I Plajzer-Frick, CS Novak, M Kato, TH Garvin, QT Pham, AN Harrington, BJ Mannion, EA Lee, Y Fukuda-Yuzawa, Y He, S Preiss, S Chee, JY Han, BA Williams, D Trout, H Amrhein, H Yang, JM Cherry, W Wang, K

Gaulton, JR Ecker, Y Shen, DE Dickel, A Visel, LA Pennacchio & B Ren. An atlas of dynamic chromatin landscapes in the developing mouse fetus. (* Authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, in part is currently being prepared for submission for publication of the material. Y Zhao, Z Cheng, Y Li, DU Gorkin & B Ren. The dissertation author was the primary investigator and author of this material.

VITA

2008	Neuqua Valley High School, Naperville, IL
2012	S.B. Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA
2019	Ph.D. Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA

PUBLICATIONS

DU Gorkin*, I Barozzi*, **Y Zhao***, Y Zhang*, H Huang*, AY Lee, B Liu, J Chiou, A Wildberg, B Ding, B Zhang, M Wang, JS Strattan, JM Davidson, Y Qiu, V Afzal, JA Akiyama, I Plajzer-Frick, CS Novak, M Kato, TH Garvin, QT Pham, AN Harrington, BJ Mannion, EA Lee, Y Fukuda-Yuzawa, Y He, S Preiss, S Chee, JY Han, BA Williams, D Trout, H Amrhein, H Yang, JM Cherry, W Wang, K Gaulton, JR Ecker, Y Shen, DE Dickel, A Visel, LA Pennacchio & B Ren. An atlas of dynamic chromatin landscapes in the developing mouse fetus. *Nature*. (2019)
* Authors contributed equally to this work

Y He, M Hariharan, DU Gorkin, DE Dickel, C Luo, RG Castanon, JR Nery, AY Lee, **Y Zhao**, H Huang, BA Williams, D Trout, H Amrhein, R Fang, H Chen, B Li, A Visel, LA Pennacchio, B Ren & JR Ecker. Spatiotemporal DNA methylome dynamics of the developing mammalian fetus. *Nature*. (2019)

I Juric*, M Yu*, A Abnousi, R Raviram, R Fang, **Y Zhao**, Y Zhang, Y Qiu, Y Yang, Y Li, B Ren & M Hu. MAPS: model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLOS Computational Biology*. (2019)

S Preissl*, R Fang*, H Huang, **Y Zhao**, R Raviram, DU Gorkin, Y Zhang, B Sos, V Afzal, DE Dickel, S Kuan, A Visel, LA Pennacchio, K Zhang & B Ren. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*. (2018)

F Zhao, **Y Zhao**, L Fa & M Zhao. Intrinsic relationship between dynamical resonance energy and decay rate. *Physical Review A*. (2013)

L Fa, L Wang, **Y Zhao**, L Liu, Y Zheng, N Zhao, M Zhao & G Li. Research Progress in Acoustical Application to Petroleum Logging and Seismic Exploration. *The Open Acoustics Journal*. (2013)

Y Zhao, N Zhao, L Fa & M Zhao. Seismic Signal and Data Analysis of Rock Media with Vertical Anisotropy. *Journal of Modern Physics*. (2013)

ABSTRACT OF THE DISSERTATION

Computational approaches for elucidating the gene regulatory landscape of mammalian development

by

Yuan Zhao

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Bing Ren, Chair
Professor Wei Wang, Co-Chair

Mammalian development including embryogenesis requires coordinated gene expression in order to control each cells eventual fate, a complex and finely tuned process driven by epigenetic information. Specifically, these developmental cues are interpreted by cells and lead to remodeling of the chromatin landscape, including changes in accessibility, conformation, and modifications to the chromatin. Thus, the state and accessibility of chromatin serve as key aspects of the cells epigenome, modulating and controlling the expression and function of the underlying genomic sequence. Understanding these developmental regulatory networks are of crucial importance as the dysregulation of developmental pathways has been implicated in

numerous disease phenotypes. To address this unmet scientific need, we systematically profiled a comprehensive array of mouse tissues spanning twelve tissues, over seven developmental time points (from 10.5 days after conception until birth), using histone ChIP-seq and ATAC-seq. We used these data to produce a catalog of putative cis-regulatory elements defined by chromatin accessibility (developmental regions of Tn5-accessible chromatin or d-TACs) and characterized their function with chromatin state annotations derived from the histone modification profiles. Within these regions of heightened accessibility, we analyzed the tissue-specific enrichments for human disease-associated sequence variation. Additionally, we studied the developmental dynamics of open chromatin regions over mouse embryogenesis, evaluating the tissue-specific temporal dynamic patterns as well as the relationship between chromatin state and accessibility. Finally, we applied novel machine learning techniques to elucidate the biology behind gene regulation, building a framework for comparing datasets using sequence-based models. Taken as a whole, this thesis provides a comprehensive study of and cutting-edge methods for understanding mouse fetal chromatin dynamics.

Chapter 1

Introduction

The central dogma of biology holds that organisms are defined the relationship between three types of molecules: DNA, RNA, and proteins. The genome, unique to every individual, contains the genes whose sequences are transcribed into RNA transcripts, which are then in turn read and translated into the proteins responsible for every cellular function. Though elegant in its simplicity, the central dogma is unfortunately, incomplete. Most critically, it fails to address the question of how those critical genes are selected for transcription and under what contexts they are active, be it developmental cues as an organism matures or environmental ones in response to external stresses. Sandwiched between the DNA and RNA layers of the central dogma is in fact a fourth layer, the epigenome, which can be defined as the chemical changes and modifications of the genome that control its regulation without modifying the genetic sequence itself.

Mammalian development including embryogenesis requires coordinated gene expression in order to control each cells eventual fate, a complex and finely tuned process driven by epigenetic information. Specifically, these developmental cues are interpreted by cells and lead to remodeling of the chromatin landscape, including changes in accessibility, conformation, and modifications to the chromatin. Epigenetic information comes in many forms including post-translational modifications of histone proteins and the positioning of nucleosomes[1]. Posttranslational histone modifications can influence gene expression by directly modulating DNA-histone interactions, or serving as binding sites for cognate reader proteins that function

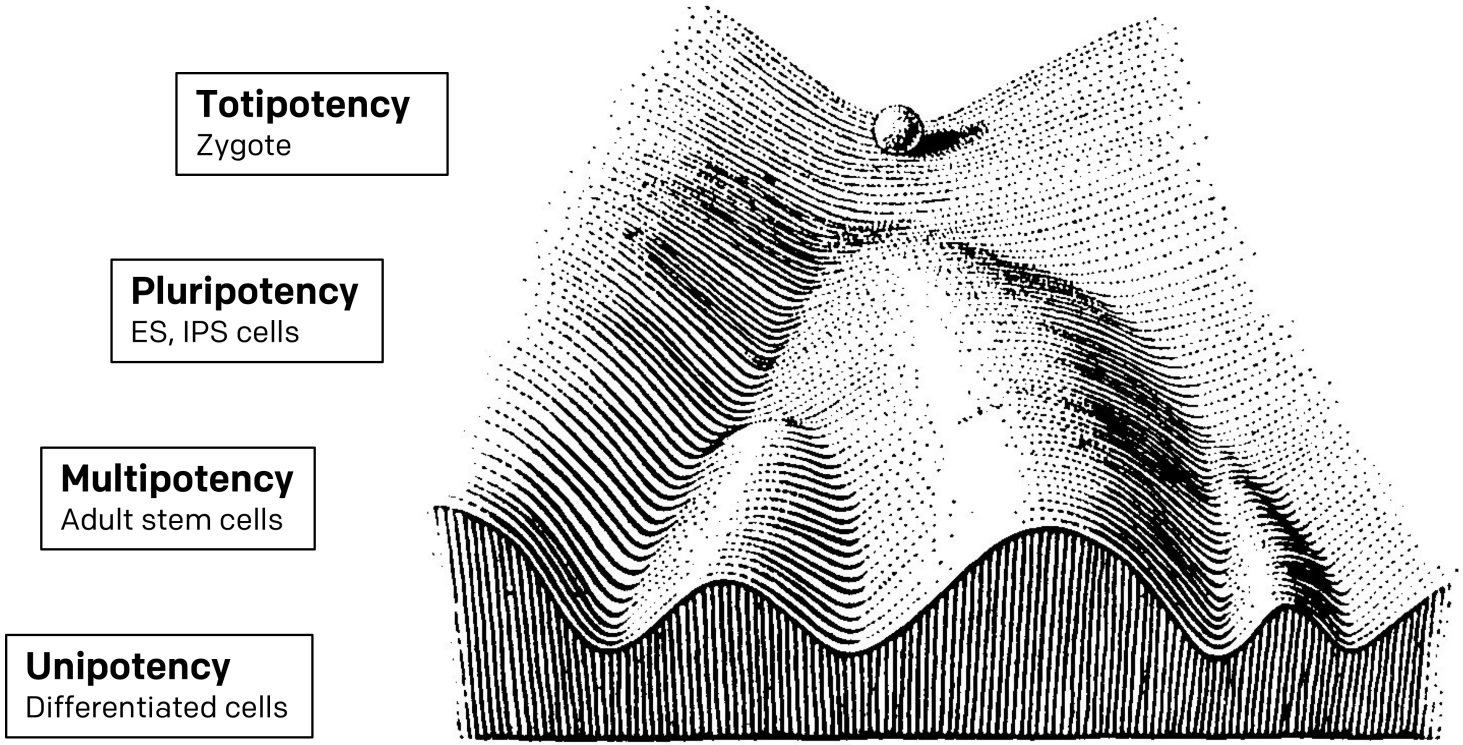

Figure 1.1. CH Waddington envisioned a cell's "epigenetic landscape" like a marble rolling into terminal valleys, where the cells developmental potential declines as it differentiates.
Adapted from "The Strategy of Genes," by CH Waddington, 1957, George Allen & Unwin.

downstream to enhance or repress transcription[2]. The importance of histone modifications is further demonstrated by the fact that mutations in genes encoding histones or histone modifying enzymes can lead to a variety of diseases including cancer, neurodevelopmental disorders, and congenital malformations[3][4]. Histone modifications are also valuable tools for genome annotation, because the modifications present at a given genome region often reflect the activity of the underlying sequence. While the profiles of individual histone modifications can be informative, the most comprehensive functional annotations are achieved when multiple histone modifications are integrated into a unified set of chromatin states based on combinatorial patterns[5][6]. Thus, the state and accessibility of chromatin serve as key aspects of the cells epigenome, modulating and controlling the expression and function of the underlying genomic sequence by silencing or enabling key regulatory elements. Ultimately, the epigenetic state (and remodeling thereof) of a given cell is responsible for the cells cellular context, or its outward phenotype.

In the case of development, these processes are responsible for the differentiation and emergence of new cell types and populations within heterogenous tissues, as the totipotent zygote eventually matures into a complex, diverse multicellular organism. In his famous 1957 book, The Strategy of Genes, philosopher and embryologist Conrad Hal Waddington famously proposed a topographical analogy for this developmental epigenetic landscape, imagining the process of differentiation as akin to a marble sliding down a landscape of branching troughs that represent cell lineages; his theories laid the foundation for epigenetic theory and evolutionary developmental biology (Figure 1.1)[7]. Less well known but no less impactful was a companion figure to the aforementioned landscape, showing the underside of this hypothetical topography (Figure 1.2). Waddington imagined that the troughs and contours of the landscape were pulled down by strings to anchor points, the combinations of which served as defining features for each cell lineage. Just as the landscapes topology represents the processes of development, these anchors are representative of the gene regulatory network behind said processes, namely the transcription factors that bind accessible sequences and the chromatin state enabling this to occur.

Understanding this complex biological landscape and the regulatory pathways driving

them are of critical scientific importance given the implications in human disease biology. Dysregulation of the epigenetics in a cell can lead to significant and wide-ranging alterations to normal gene regulation and lead to a variety of severe disorders including developmental defects, neurological disorders, and cancer. These regulatory programs are orchestrated, at least in part, by cis regulatory elements (cREs) that direct the expression of genes in response to specific developmental and environmental cues[6][8]. Beyond their normal role in development, cREs are important to human health because mutations in cREs can cause or contribute to a variety of diseases. Moreover, SNPs associated with disease by Genome-Wide Association Studies (GWAS) are highly enriched in putative cREs[10]. Active cREs have characteristic patterns of histone modifications including H3K4me3 and H3K27ac which mark chromatin around active promoters, and H3K4me1 and H3K27ac which mark chromatin around active enhancers[11][12][13]. Active cREs are also characterized by the presence of open or accessible chromatin that is relatively devoid of packaging nucleosome particles. This loose packaging makes the underlying DNA sequence more accessible to transcription factors, which can function in turn to recruit co-factors and stimulate transcriptional activity. Importantly, the level of accessibility at cREs can be modulated by chromatin remodeling factors and so-called pioneer transcription factors[14], highlighting the importance of chromatin accessibility in gene regulation.

There are several genomic tools available to the field for probing the mammalian epigenetic landscape. In addition to chromatin immunoprecipitation sequencing (ChIP-seq), which allows us to profile the functional histone modifications associated with various locations in specific cellular contexts, recent advances have enabled researchers to directly evaluate the accessible chromatin landscape using DNase I hypersensitive sites sequencing (DNase-seq) and the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). Indeed, these tools have been leveraged by previous researchers in several efforts to build catalogs and atlases of putative cREs[15][16][17][18]; however, there remains a large unmet need in the field for a catalog that comprehensively profiles the entire process of mammalian development in a diverse variety of cellular contexts (Figure 1.3). Most resources to date focus primarily on one of two

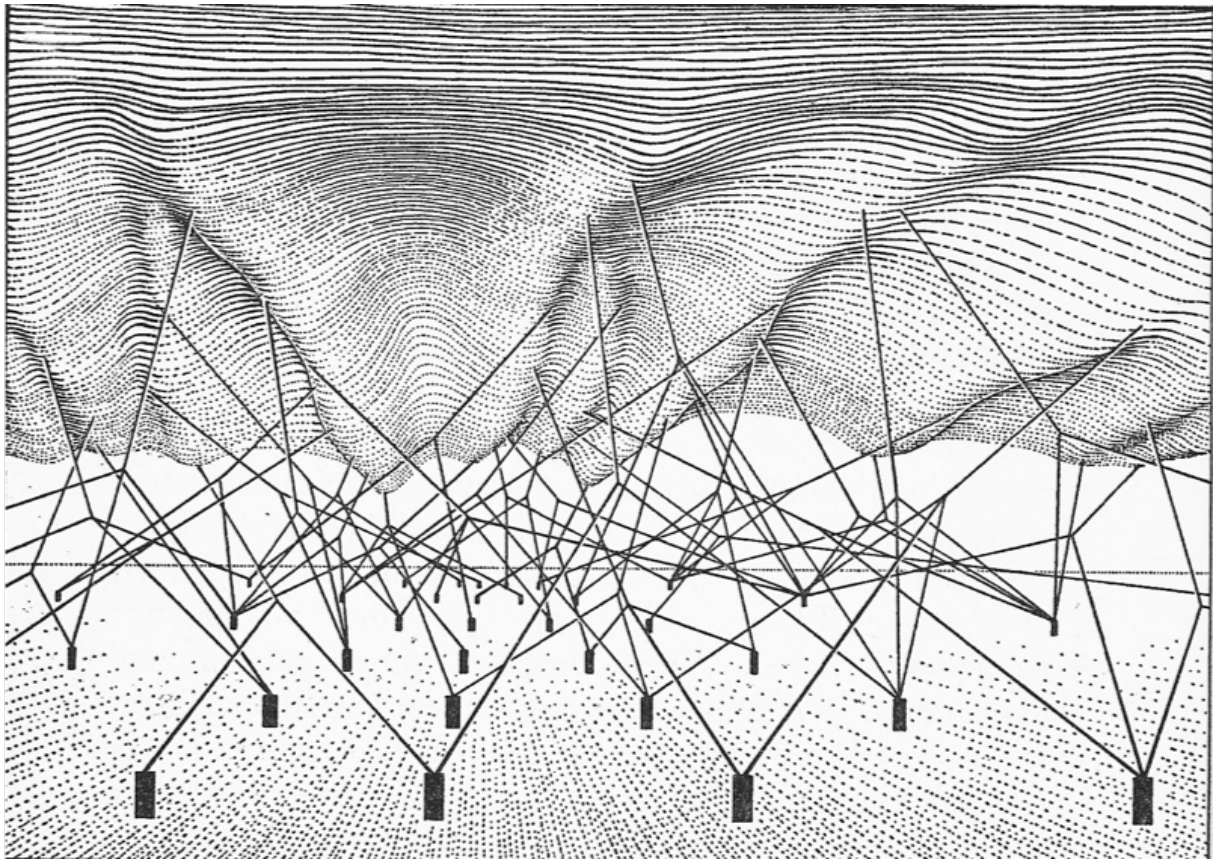

Figure 1.2. Waddington envisioned strings pulling down the landscape to create fate-determining valleys. These tethers are akin to transcription factors, and the combinations of their binding control a cells regulatory program.
Adapted from "The Strategy of Genes," by CH Waddington, 1957, George Allen & Unwin.

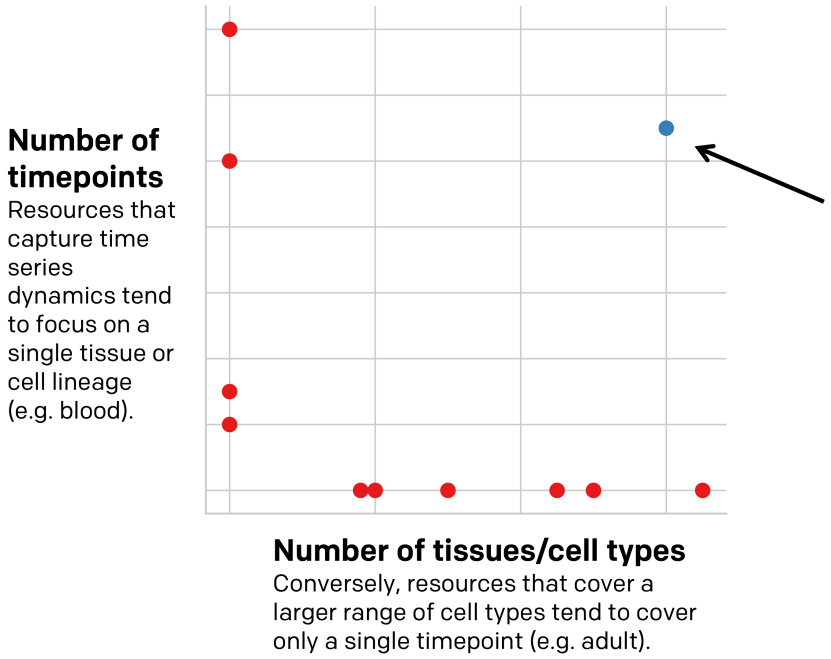

Figure 1.3. The majority of extant genomics data bases either focus on profiling the temporal (developmental) dimension or the cellular phenotype dimension, resulting in a dearth of resources that sample in both dimensions.

axes: either assessing a range of tissues or cell types at a single temporal stage (e.g. adult) or providing a temporal dataset but in a single, well-defined lineage (e.g. blood lineages). The availability of a resource that profiles samples in both dimensions would enable an unprecedented view of the dynamic chromatin landscape during development, helping us to better understand the way chromatin changes and affects gene regulation for different phenotypes.

In this thesis, we present a study in which we address this scientific need, profiling not two but in fact three dimensions of data: tissues, developmental times, and genomic assay. In this study, we profile both chromatin state and accessibility across a large panel of mouse fetal tissues and developmental stages. More specifically, we used chromatin Immunoprecipitation sequencing (ChIP-seq) to profile a set of eight histone modifications which in combination can distinguish multiple types of functional elements and activity levels. To assay chromatin accessibility, we used the Assay for Transposase-Accessible Chromatin using Sequencing

(ATAC-seq)[19]. Chromatin accessibility can also be mapped by DNase-seq, which has been integral to the identification of millions of candidate CREs in mammalian genomes[16][20]. However, ATAC-seq has emerged as a powerful alternative method for mapping open chromatin accessibility, offering a more streamlined workflow because fragmentation and incorporation of sequencing adapters are carried out in a single step by the Tn5 transposase. Furthermore, we apply cutting-edge machine learning techniques to better understand the epigenetic basis behind gene regulation using a sequence-based neural network strategy, providing a novel framework for comparing similarity between diverse samples regardless of genomic origin.

The resulting ATAC-seq maps of chromatin accessibility and neural network models provide deep insight into the genomic regions and processes that drive mouse fetal development. Highlights of our findings include:

- We identified more than 500,000 regions marked by accessible chromatin during mouse fetal developmental, including approximately 140,000 with dynamic temporal activity in at least one tissue.

- Human orthologs of accessible chromatin regions in fetal mouse tissues are enriched for human disease-associated sequence variation, with specific diseases showing tissue-restricted patterns of enrichment.

- Temporal changes in chromatin accessibility often coincide with changes in enhancer chromatin states, and tend to precede changes in nearby H3K27ac levels.

- We predict 21,142 enhancer-promoter interactions by measuring the correlation between enhancer-associated chromatin signals and gene expression across tissues-stages, and find enrichment for GWAS SNPs within enhancers predicted to regulate Mendelian disease genes.

- Neural networks produce highly predictive models for the accessible chromatin landscape in individual cell types and samples.

- Models trained on a combination known motifs and *de novo* motifs outperform those trained one just one or the other.

- Predictions made on a uniform set of sequences derived from pooled open chromatin regions can be projected onto a 2D space, enabling comparisons between diverse samples in terms of their genomic vocabulary and grammar.

Chapter 2

Systematic mapping of accessible chromatin during mouse development

2.1 Abstract

Mammalian development, including embryogenesis, requires coordinated gene expression in order to control each cells eventual fate, a complex and finely tuned process driven by epigenetic information. Specifically, these developmental cues are interpreted by cells and lead to remodeling of the chromatin landscape, including changes in accessibility, conformation, and modifications to the chromatin. To understand these processes, we systematically profiled a comprehensive array of mouse tissues using histone ChIP-seq and ATAC- seq. We used these data to produce a catalog of putative cis-regulatory elements and analyzed the tissue- specific enrichments for human disease-associated sequence variation.

2.2 Introduction

Traditionally, *cis* regulatory sequences are defined by genetic and comparative genomics, which can be limited in throughput and cellular specificity. In recent years, the development of high throughput sequencing methods to detect chromatin features and transcription factor binding sites has greatly accelerated genome-wide mapping of *cis* elements in diverse species. Among the various chromatin features, one of the commonly used is chromatin accessibility,

which could be monitored using DNase-I treatment followed by high throughput sequencing (DNase-seq). This approach has led to identification of millions of candidate *cis* regulatory sequences in the mammalian genomes of human, mouse and a number of other species. More recently, the Assay for Transposase accessible chromatin followed by sequencing (ATAC-seq) has been demonstrated to detect chromatin accessibility from as few as several hundred cells. For example, Wu et al. use this method to investigate the chromatin landscape in oocytes, zygotes and early embryos[22].

Variations in *cis*-regulatory elements among the human population have been linked to a growing number of common and rare diseases , presumably through dysregulation of gene expression[10]. Thus, identification of the *cis*-regulatory elements in a genome is of paramount importance to the study of developmental biology, evolution, and pathogenesis of human disease.

Despite the importance of chromatin states and accessibility in determining the functional output of the genome, a comprehensive survey of chromatin dynamics during mammalian fetal development has been lacking, except for very early stages of embryogenesis[21][22]. The mouse is an ideal system to study the developmental epigenome because it is the most widely used animal model in biomedical research, and enables experimental access to timed stages of development. Indeed, the biology of mice is largely representative of placental mammals including humans.

Previous experiments investigating open chromatin in the mouse genome have been focused largely on adult tissues and cell types, with only a handful of embryonic tissues profiled. While mouse models theoretically allow access to tissues at all developmental stages, studying embryonic development has been constrained by the limited quantities of tissue samples that can be obtained from early embryonic stages. To overcome this limitation, we have developed a modified protocol of ATAC-seq that could be applied to flash frozen primary mouse tissues at small quantities, and used this method to generate maps of open chromatin in up to 12 embryonic tissues across seven stages of mouse fetal development from 11.5 days post-conception (e11.5) to birth (p0). Integrative analysis of this dataset along with other DNA methylation,

chromatin modifications and transcriptome data from the same tissues allowed us to annotate the dynamic chromatin state of over half a million cis-regulatory elements in the mouse genome during embryogenesis, providing insights into the mammalian gene regulatory programs and the molecular basis of human diseases.

2.3 Results

2.3.1 Accessible chromatin profiled across a diverse cross-section of murine tissues and developmental stages

We harvested mouse tissues at closely spaced intervals from 11.5 days post conception (E11.5) until birth. At each stage, we dissected a diverse panel of tissues from multiple litters of C57BL/6N embryos and performed two replicates of ATAC-seq, and two replicates of ChIP-seq for each of eight histone modifications (H3K4me1, H3K4me2, H3Kme3, H3K27ac, H3K27me3, H3K9ac, H3K9me3, H3K36me3) (Figure 2.1a). These modifications were chosen to distinguish between different types of functional elements (e.g. promoters, enhancers, gene bodies), and activity levels (e.g. active, poised, repressed)[23]. We also collected 6 tissues at E10.5, but it was not feasible at this stage to harvest enough tissue for standard ChIP-seq, so we used a modified micro-ChIP-seq procedure designed to work on much smaller numbers of cells and restricted our scope to only six histone modifications.

The complete ChIP-seq data series includes more than 66 billion sequencing reads from 564 ChIP-seq experiments, each consisting of two biological replicates derived from different embryo pools (N=1,128 replicates total). The ATAC-seq data series includes more than seven billion sequencing reads from 66 experiments (N=132 replicates total). All ChIP-seq and ATAC-seq datasets were processed with a uniform pipeline, and subject to high quality standards. Whole genome bisulfite sequencing and RNA-seq were also performed on these same tissue samples by other groups in the ENCODE consortium. These data are presented in detail in companion manuscripts[25], but we make select use of these complimentary datasets here to

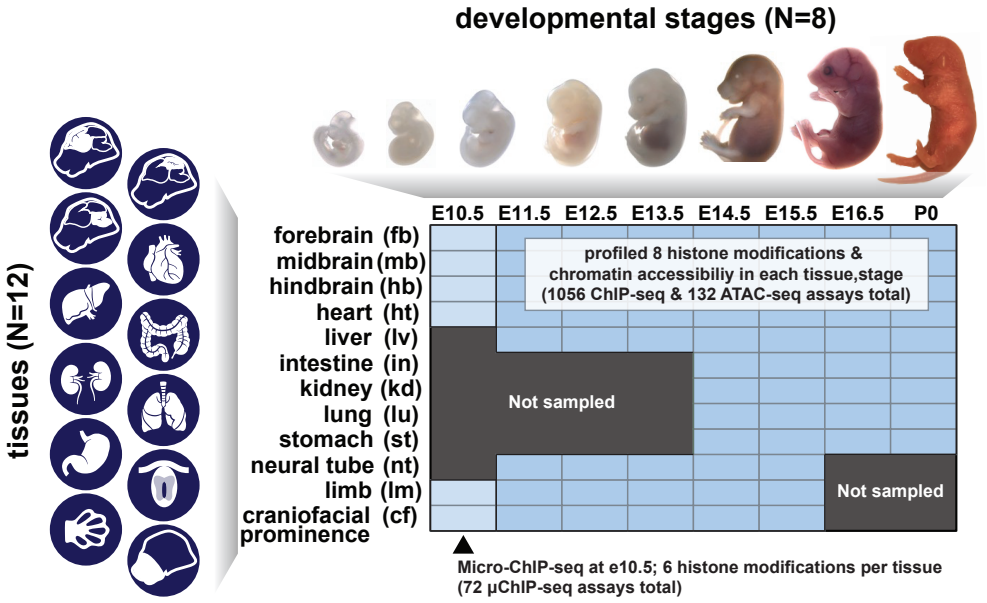

Figure 2.1. Overview of tissues and stages profiled. The ENCODE Consortium effort of which this thesis was a contributing component produced a comprehensive epigenomics resource featuring three dimensions of data: profiling up to 12 mouse embryonic tissues, over up to 8 developmental stages, and using genomics assays spanning 8 histone modifications, chromatin accessibility, the transcriptome, and the methylome.
Adapted with permission from "An atlas of dynamic chromatin landscapes in the developing mouse fetus," by DU Gorkin, I Barozzi, Y Zhao, Y Zhang, H Huang et al, 2019, Nature.[24]

perform integrative analysis.

A key novelty of the data produced for this resource is the profiling of a developmental (temporal) dimension in addition to the diversity of tissue samples. Crucially, these data are able to capture the biological differences inherent in each samples specific cellular context, both in terms of identifying tissue-specific peaks and patterns in the signal tracks, as well as temporal dynamics. This ability to discriminate unique sample-specific biology holds true for both the ChIP-seq datasets profiling histone modifications as well as the ATAC-seq for chromatin accessibility patterns (Figure 2.2).

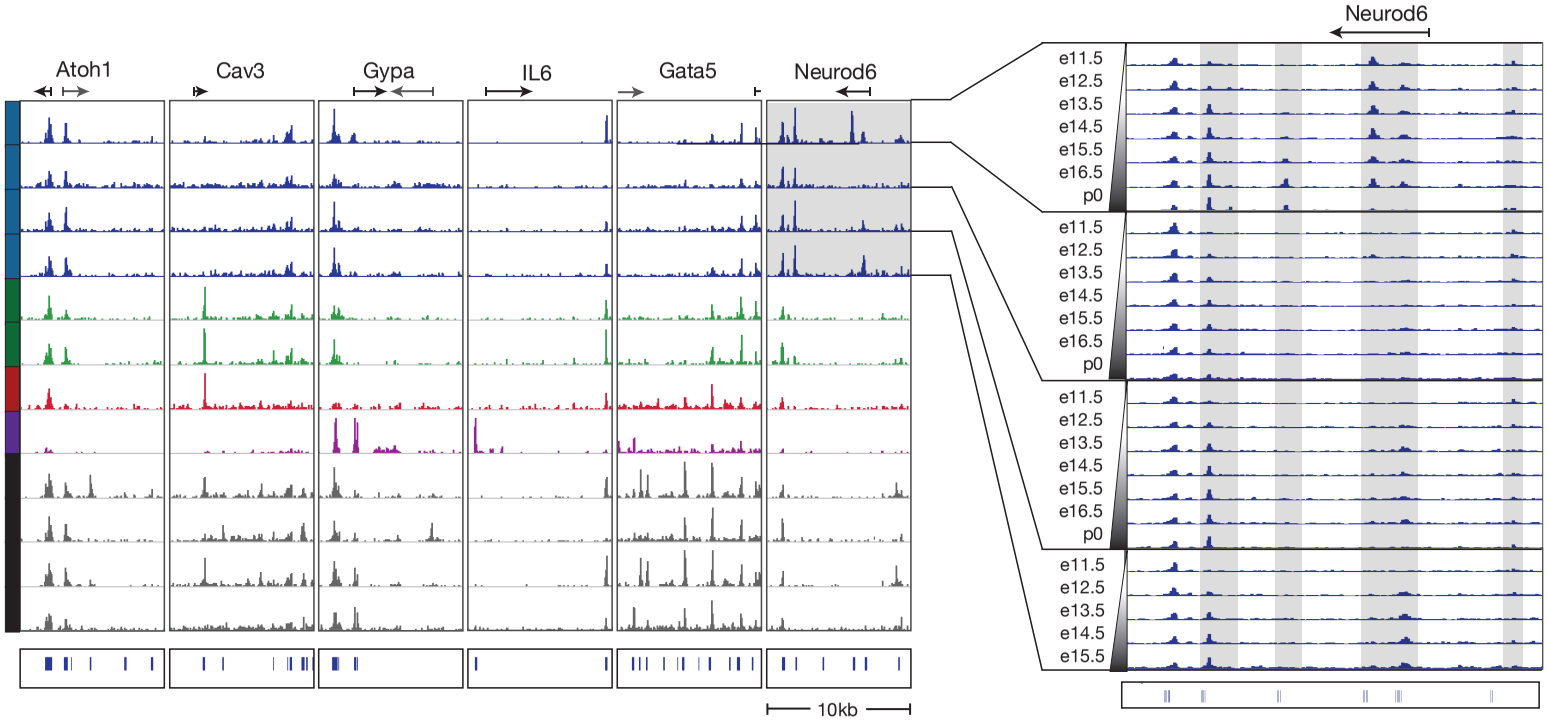

Figure 2.2. Example of signal tracks showing peaks and signal showing tissue-specific and temporal-specific accessibility and activity profiles.
Adapted with permission from "An atlas of dynamic chromatin landscapes in the developing mouse fetus," by DU Gorkin, I Barozzi, Y Zhao, Y Zhang, H Huang et al, 2019, Nature.

2.3.2 High-quality, reproducible datasets meet ENCODE Consortium standards

For the mouse ENCODE Consortiums third phase, we developed guidelines for stringent ATAC-seq data quality standards. Specifically, we required that each tissue and developmental stage be sequenced to a depth of at least 20 million usable reads (defined as sufficiently high-quality, non-duplicated, and non-mitochondrial) and meet a threshold of at least 10% of reads being represented in transcription start site regions (e.g. fraction of reads overlapping TSS, or FROT) (Figure 2.3). As compared to several reference datasets, the ATAC-seq generated compare favorably, showing sufficient enrichment of TSS-covering reads while being comparable to the Immunological Genome Project data series[27] and DNase-seq generated by our collaborators in the Stam Lab in the same ENCODE effort. The FROT score for our ATAC-seq data series also exceeds that of the corresponding H3K27ac histone ChIP-seq data.

In addition to satisfying sequencing data quality controls, we required that our data be highly reproducible between biological replicates. Evaluating both the Pearson and Spearman correlations between replicates for each sample across the entire genome reveals a strong similarity between them (Figure 2.4). To assess the overall relationship between all of the samples and their similarity, we performed dimensionality reduction using multi-dimensional scaling (MDS) over read counts in replicated peak calls throughout the genome and projected the samples onto 2D and 3D spaces (Figure 2.5a-b).

2.3.3 Identification of developmental regions of transposase-accessible chromatin (d-TACs)

To build a catalog of candidate cREs contributing to mouse fetal development we merged ATAC-seq peaks across tissue-stages to identify a non-overlapping set of 523,159 regions, which we refer to hereafter as developmental regions of transposase accessible chromatin (d-TACs) (Figure 3a). Comparing chromatin state annotations to our catalog of d-TACs, we find that d-TACs are enriched in promoter and enhancer chromatin states, but generally depleted

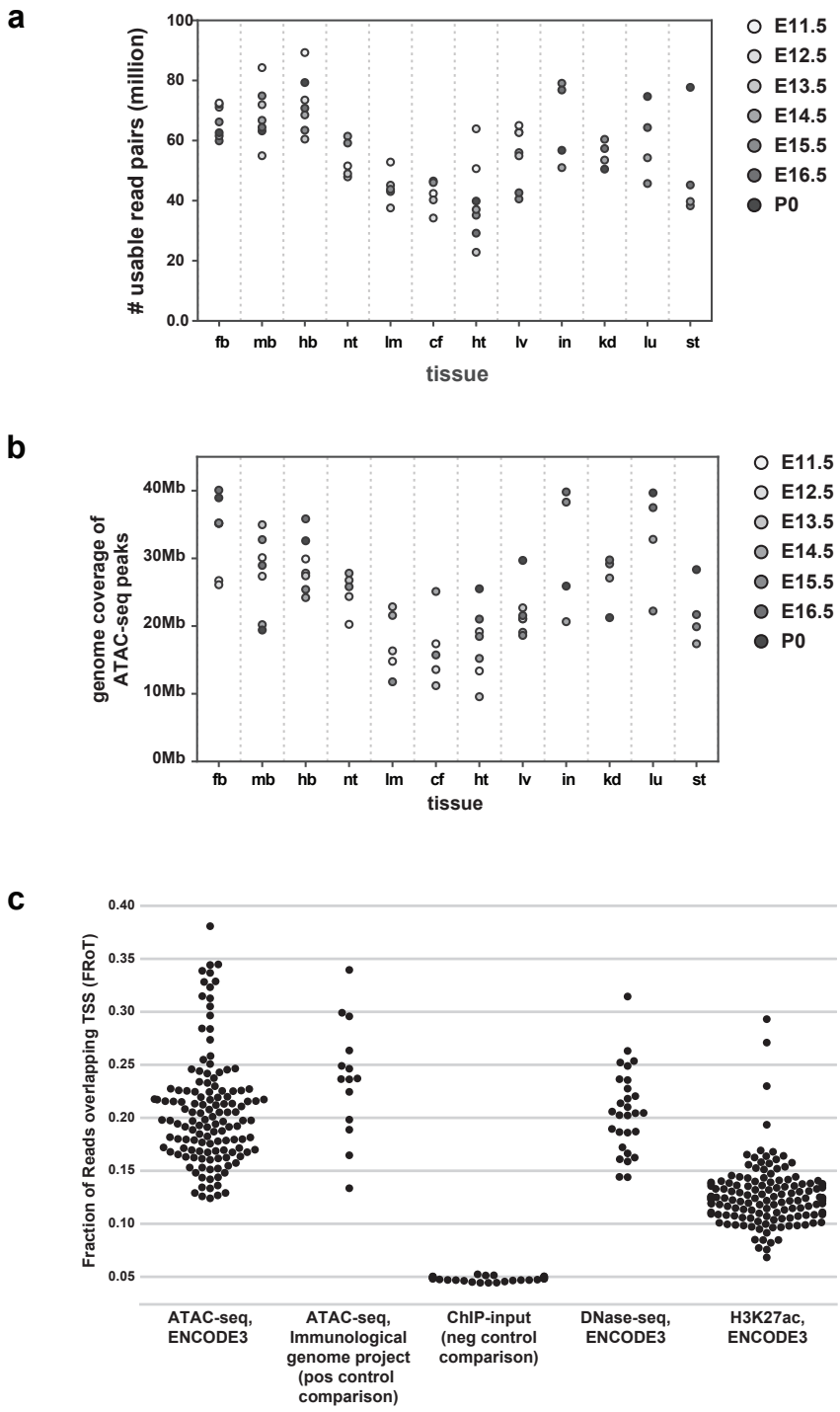

Figure 2.3. a. Number of usable read pairs in each sample. **b.** Genome coverage of ATAC-seq peaks called in each sample. **c.** Fraction of reads overlapping TSS (FROT) scores for mouse embryonic ATAC-seq data as well as ATAC-seq from the Immunological Genome Project (positive control), ChIP-input (negative control), a subset of matched DNase-seq, and matched H3K27ac histone ChIP-seq.

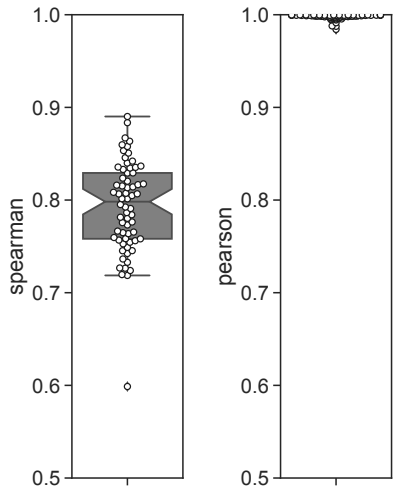

Figure 2.4. Pearson and Spearman correlation of read counts across replicated peak calls between replicates, indicating high levels of reproducibility in the ATAC-seq data.

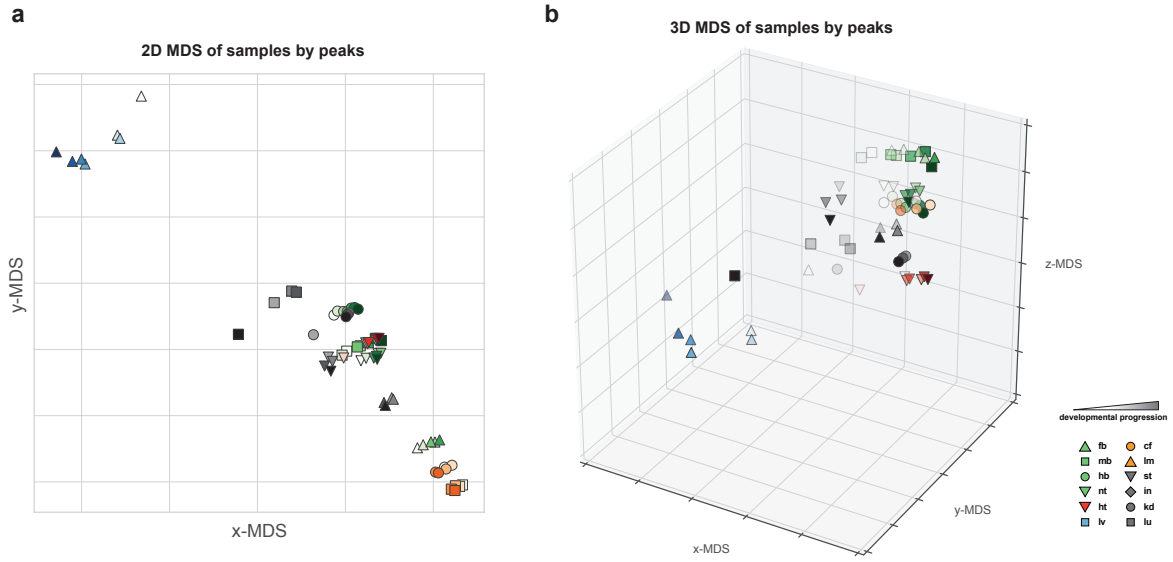

Figure 2.5. Multidimensional scaling projections showing the relationships between samples based on their total read counts sum(replicate 1, replicate 2) across replicated peak calls. Samples are color and shape-coded based on their tissue and shaded progressively from light to dark depending on chronological developmental stage. Note that for these figures, we used the ”zero” array in which peaks that are inaccessible in a given tissue are assigned a read count of 0 so that the values can be quantile normalized for batch effects within each tissue respectively. **a.** 2D MDS plot. **b.** 3D MDS plot.

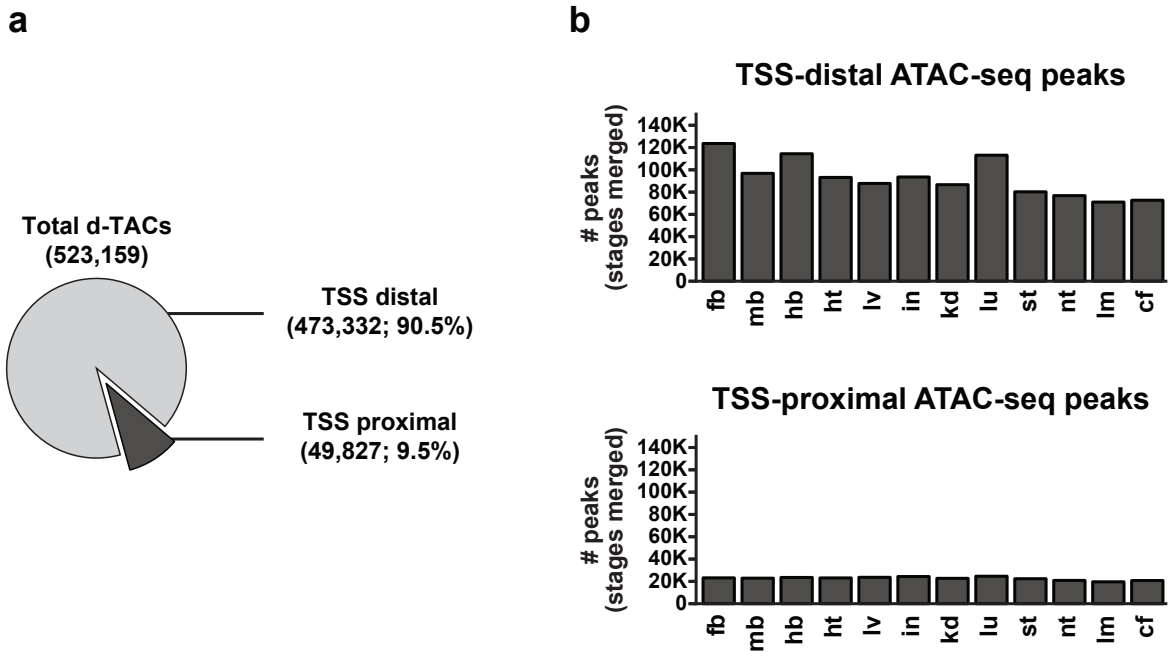

Figure 2.6. a. Summary of catalog of d-TACs (total number identified across all samples, number of TSS-distal, and number of TSS-proximal). **b.** Detailed breakdown of the number of TSS-distal and proximal d-TACs organized by tissue of interest, defined as TSS-proximal if found within 1kb of GENCODE TSS.

in chromatin states characteristic of gene bodies, heterochromatin, and regions showing no chromatin signature (Figure 3b). Despite the enrichment of d-TACs in promoter states, the vast majority of d-TACs are distal to annotated Transcription Start Sites (TSS), representing putative enhancers and other TSS-distal elements (473,332/523,159, 90.48% of d-TACs $>$1kb from TSS).

2.3.4 Predictions of accessible regions supported by orthogonal datasets

Comparison with the VISTA database of experimentally validated enhancers[28] shows that the d-TAC catalog has high sensitivity for identifying enhancers: 75.6-93.6% of sequences showing in vivo enhancer activity by reporter assay are d-TACs in the corresponding tissue at E11.5 (the stage at which embryos are collected in the VISTA reporter assays) (Figure 2.7b). The catalog has an even higher sensitivity for active promoters, as 87.6-95.1% of active promoters

Table 2.1. Number of replicated ATAC-seq peaks called per sample, for each by tissue and stage respectively. See Methods for description of how these peak sets were called and processed.

	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
Forebrain	60,632	55,198	76,369	77,855	86,322	107,709	102,783
Midbrain	64,291	63,382	80,207	48,166	85,460	71,634	49,033
Hindbrain	74,441	65,917	74,490	59,881	61,625	92,234	82,436
Neural Tube	49,696	59,006	65,048	72,611	65,402	-	-
Limb	39,322	52,210	62,776	60,989	34,407	-	-
Craniofacial	48,282	45,541	34,269	68,810	53,276	-	-
Heart	39,993	61,958	32,001	52,157	59,679	64,563	82,839
Liver	60,865	60,378	66,185	52,686	56,138	69,100	88,391
Intestine	-	-	-	52,058	98,614	103,437	74,031
Kidney	-	-	-	69,094	85,527	80,592	56,490
Lung	-	-	-	79,969	56,792	95,831	112,210
Stomach	-	-	-	47,660	64,047	59,879	76,844

Table 2.2. Number of d-TACs identified for each mouse embryonic tissue, categorized as TSS-distal or TSS-proximal based on distance to GENCODE annotated TSS sites.

	TSS-distal	TSS-proximal	Total
Forebrain	123,622	23,169	146,791
Midbrain	96,833	22,892	119,725
Hindbrain	114,409	23,518	137,927
Neural Tube	76,827	20,871	97,698
Limb	71,047	19,628	90,675
Craniofacial	72,707	20,775	93,482
Heart	93,172	23,107	116,279
Liver	87,771	23,652	111,423
Intestine	93,574	24,344	117,918
Kidney	86,642	22,727	109,369
Lung	113,120	24,609	137,729
Stomach	80,255	22,400	102,655

(i.e. TSS with transcripts present in available RNA-seq data from the corresponding tissue stage) are detected as d-TACs (Figure 2.7a). While these data suggest a high sensitivity of ATAC-seq in uncovering regulatory sequences in our samples, estimation of the specificity is more difficult due to the lack of a true negative dataset. However, using annotated promoters that lack DNase I hypersensitivity in matched samples as a negative set, we find that $<5\%$ of these regions are detected as d-TACs (1.7-4.5%), which suggests a low false positive rate.

2.3.5 Functional characterization of d-TAC catalog

Using the accessible regions derived from ATAC-seq, we were able to identify putative cis-regulatory elements cataloged in our map of developmental d-TACs; however, without further analysis, it is not immediately clear what types of elements are represented or what regulatory function is performed by these elements. To understand the constitution of our d-TAC catalog, we performed integrative analysis using the orthogonal ChIP-seq data to functionally characterize these developmentally accessible loci. Specifically, to leverage the chromatin state information captured by combinatorial patterns of histone modifications we applied ChromHMM[26], deriving a 15-state model that shows near-perfect consistency between biological replicates

Table 2.3. Number of d-TACs identified for each sample, for each by tissue and stage respectively.

	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
Forebrain	49,685	45,731	64,496	65,752	70,979	90,962	88,682
Midbrain	54,203	52,115	65,432	39,628	69,184	57,337	38,995
Hindbrain	61,540	51,909	58,554	50,165	50,777	77,417	69,990
Neural Tube	39,608	45,602	51,018	60,080	52,781	-	-
Limb	31,877	43,047	53,424	50,688	27,665	-	-
Craniofacial	38,715	36,872	27,484	57,326	43,734	-	-
Heart	31,613	50,362	26,447	43,438	49,407	53,090	69,568
Liver	50,077	49,571	54,977	42,996	44,039	56,630	73,732
Intestine	-	-	-	42,049	83,219	87,798	61,975
Kidney	-	-	-	58,481	71,424	68,248	47,638
Lung	-	-	-	68,452	48,566	82,994	97,699
Stomach	-	-	-	40,117	54,249	51,358	66,728

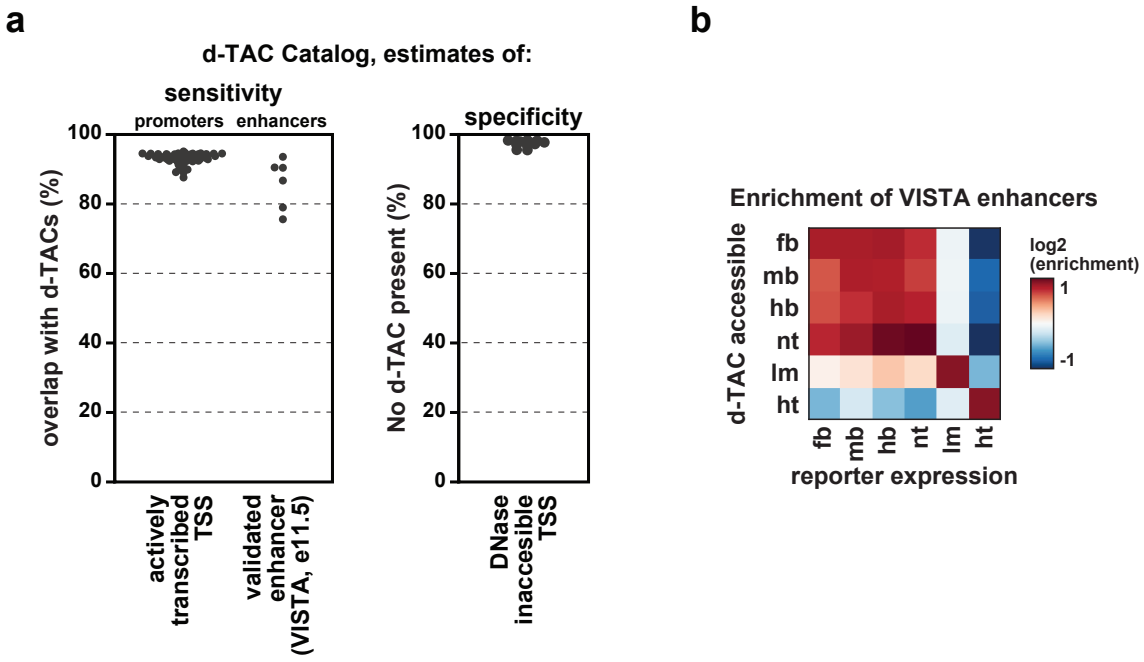

Figure 2.7. a. Estimates of the d-TAC catalogs sensitivity (left) and specificity (right). Each point represents one tissue-stage. For enhancers, 6 tissue-stages are plotted (E11.5 forebrain, midbrain, hindbrain, limb, heart, neural tube), because these 6 tissues match up with tissues and stages assayed and cataloged in the VISTA database. For DNase inaccessible TSS, only tissue-stages with matched DNase data available through the ENCODE portal are plotted here (N=18). **b.** VISTA-validated enhancer enrichment in accessible regions of the corresponding tissue.

and general agreement with previously published models. We segmented the genome for each tissue-stage with the full complement of eight histone modifications (N=66), excluding E10.5 to ensure a consistent approach. Each state was assigned a descriptive label based on similarity to known chromatin signatures[4][23][29], and genomic distribution. The resulting chromatin state maps allow for visualization of multiple functional predictions across a range of tissues and stages (Figure 2.8).

The 15 chromatin states fit into four broad functional classes: promoter, enhancer, transcriptional, and heterochromatin states. As expected, promoter states show the highest average levels of chromatin accessibility, followed by enhancer, transcriptional, and heterochromatin (Figure 2.8a). In total, we find that 33.4% of the genome shows a reproducible chromatin signature characteristic of one of these four functional classes in at least one tissue-stage. In this calculation we required that a region be called in the same state in both biological replicates and excluded states 15 (no signal) and 11 (permissive) that cover large swaths of the genome (Figure 2.8c). This does not necessarily imply that 33.4% of the genome sequence is functional during development, but rather that 33.4% of the genome sequence is mappable and packaged in chromatin with a reproducible signature in at least one tissue-stage profiled here. These chromatin signatures often reflect transcriptional and/or regulatory activity, but the underlying sequences may not contribute to fitness.

Comparing chromatin state annotations to our catalog of d-TACs, we find that d-TACs are enriched in promoter and enhancer chromatin states, but generally depleted in chromatin states characteristic of gene bodies, heterochromatin, and regions showing no chromatin signature (Figure 2.9).

Summary of characteristic enrichment patterns for histone modifications surveyed here. Note that modifications are generally categorized as narrow or broad depending on the typical breadth of enrichment. H3K9me3 is further distinguished from other broad marks because it shows very few regions of enrichment in non-repetitive sequence in primary tissues and cells[30] Metagene profiles were plotted with deeptools plotProfile[31], using data from E15.5 Heart

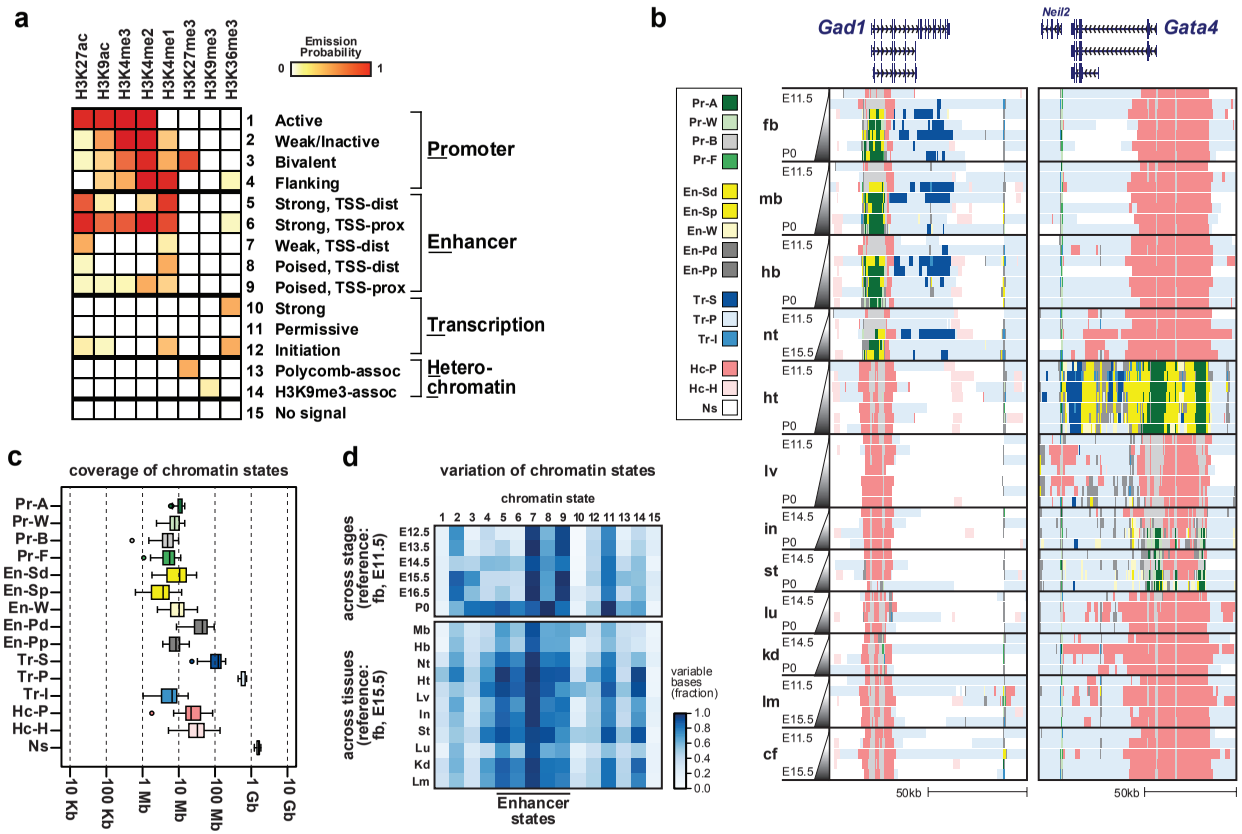

Figure 2.8. a. Heatmap on the left shows k-means clustering of dynamic forebrain enhancers based on H3K27ac signal at stages: E11.5, E12.5, e13.4, E14.5, E15.5, E16.5, P0 (k=4). **b.** Genome browser view showing predicted enhancers of Ascl1 (chr10:87,301,848-87,515,210; mm10). **c.** Enrichment of each chromatin state across the genome. **d.** Heatmap showing the variation of each chormatin state across tissues and stages.

Adapted with permission from "An atlas of dynamic chromatin landscapes in the developing mouse fetus," by DU Gorkin, I Barozzi, Y Zhao, Y Zhang, H Huang et al, 2019, Nature.

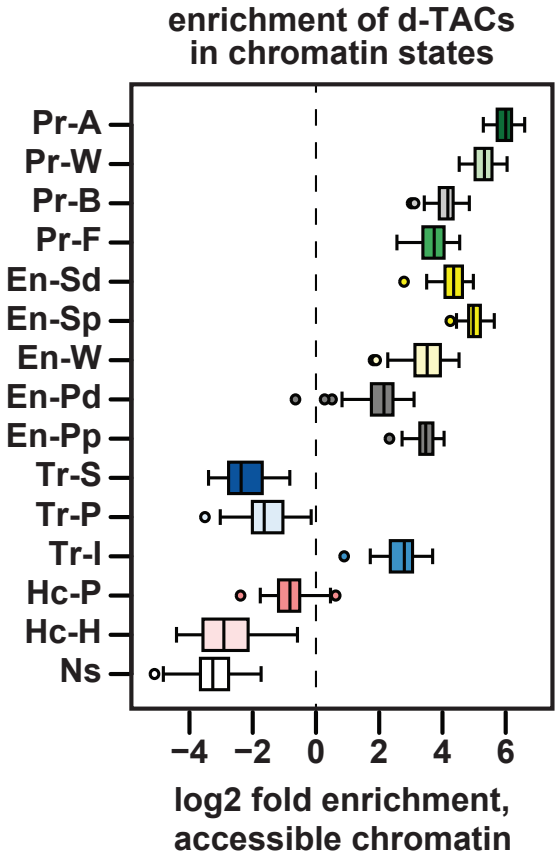

Figure 2.9. Enrichment of accessible chromatin within regions segmented into different chromatin states. Chromatin state labels same as Figure 2.8.

(Figure 2.10).

2.3.6 Orthologous regions of d-TACs in the human genome exhibit enrichment of GWAS traits

Catalogs of candidate cREs have proven to be valuable resources for the interpretation of non-coding genetic variation linked to disease, because such variation is highly enriched in regulatory sequence[32]. Thus, we sought to determine whether our catalog of mouse d-TACs could be leveraged to provide insight into human complex disease, as many developmental enhancers are believed to be conserved[33]. To this end, we used the liftover utility[34] to identify human orthologs of our mouse d-TACs (we use the term ortholog here to indicate that a

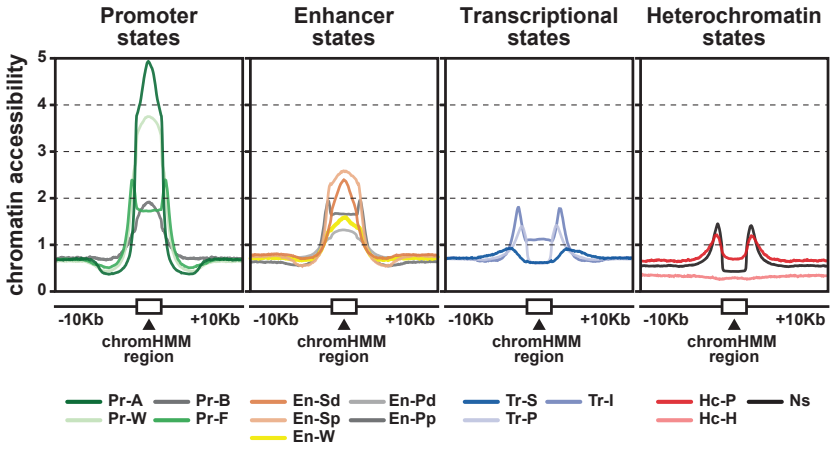

Figure 2.10. Metagene plots of chromatin states centered on accessible peaks of e15.5 forebrain. Adapted with permission from "An atlas of dynamic chromatin landscapes in the developing mouse fetus," by DU Gorkin, I Barozzi, Y Zhao, Y Zhang, H Huang et al, 2019, Nature.

given pair of mouse and human sequences are derived from a common ancestral sequence, not necessarily that the function of a mouse sequence is conserved in human). Of the total elements in the mouse TAC catalog, 172,879 (36.6%) had orthologous sequences in the human genome.

We discovered a significant enrichment (P=2.38 x 10-100) of the (European) GWAS loci within the lifted mouse ATAC sites as compared to the background loci (Figure 2.11, left). To demonstrate the novel knowledge contributed by our mouse embryonic data, we intersected the human d-TAC orthologs with a superset of regions from a combined data set found within ENCODE or Roadmap DNAse I hypersensitivity sites for around a total of 350 cell types. Approximately 89% (169,571/190,462) of the human-mappable orthologous sequences are also annotated as regions of accessible chromatin in human cells[10][20], suggesting conserved function. Though it is tempting to interpret the remaining, previously unannotated 11% of human orthologs to be previously undiscovered – likely embryonic – human regulatory elements, we must be cautious not to over-interpret their significance; the reason they were not previously identified in human screens is just as likely to be that they do not actually have any conserved or functional role in humans. Thus, to interrogate whether or not we identified putative human

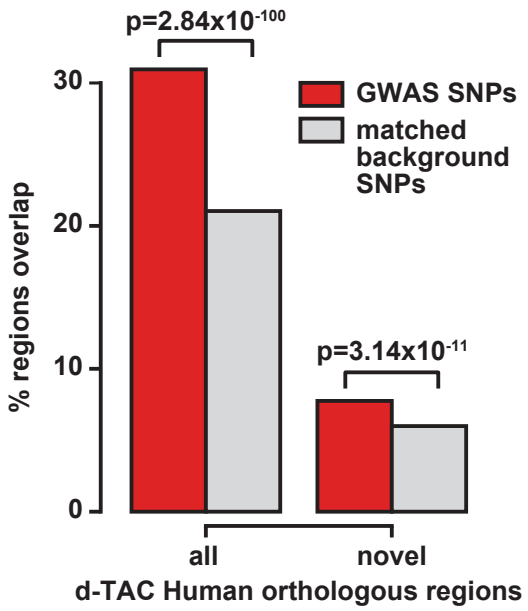

Figure 2.11. Enrichment of GWAS SNPs in human orthologs of d-TACs compared to background set of matched SNPs generated with SNPsnap[35] (p=2.84x10-100, hypergeometric). Novel regions (right) are defined as those that to our knowledge have not been previously described in catalogs of accessible chromatin regions in human cells (p=3.14x10-11, hypergeometric).

regulatory elements of embryonic origin, we performed the same GWAS enrichment analysis on this novel subset as the entire set of mappable human orthologs. For this subset of unique regions, we also observed a significant enrichment (P=3.14 x 10-11) of GWAS loci (Figure 2.11, right). This establishes that our resource generally captures novel information missed by current accessible chromatin data. Again, though not all of these novel mapped sequences are guaranteed to be functional in the human genome, the enrichment of functional variants in these loci suggests that they do encompass some human elements with embryonic activity.

Next, we sought to show specific examples of GWAS traits that are enriched within these lifted mouse developmental sites. We focused on enhancer sites because they are more likely to have tissue-specific effects. For each tissue, we lifted over collapsed chromHMM enhancer states

(En-Sd, En-Sp, En-W) that were merged across timepoints. We then defined active enhancer regions by overlapping enhancer regions with ATAC peaks. We tested for enrichment of human traits and diseases within active enhancers with polyTest (Figure 2.12a).

We find that human phenotype-associated genetic variation reveals relationships between specific phenotypes and particular developmental tissues and our results generally corroborate what is known in the literature. We observed significant enrichments of brain-related traits such as intelligence and sleep duration[37] as well as brain-related diseases such as biopolar disorder[38], schizophrenia, and ADHD[39] within brain annotations[40][41]. We also found significant enrichment of immune-related diseases such as rheumatoid arthritis[42] and primary sclerosing cholangitis[43] within liver annotations, which makes sense given that the liver is the primary site for mouse embryonic hematopoiesis. For Crohns Disease, ulcerative colitis, and IBD, we found enrichment within liver, intestine, and stomach active enhancers, which is concordant with the etiology of these diseases. For anthropometric traits related to development, we observe wide-ranging tissue enrichments[44]. Overall, we demonstrate how this unique resource can be used to break down genetic association information using tissue-specific regulatory sites.

We note that the ATAC-seq data generated here comes from heterogeneous tissue samples. However, our group recently published single-nucleus ATAC-seq (snATAC-seq) of mouse forebrain[45], which can further resolve chromatin accessibility profiles for distinct cell types within a complex tissue. Re-analyzing that data here, we detected enrichment of phenotype-associated genetic variation within regulatory sequence in specific forebrain cell types (Figure 2.12b). The enrichment of variation associated with late onset Alzheimer's disease within microglia regulatory sequence is particularly striking, and agrees with a growing body of literature linking microglia in the pathophysiology of Alzheimer's disease[46].

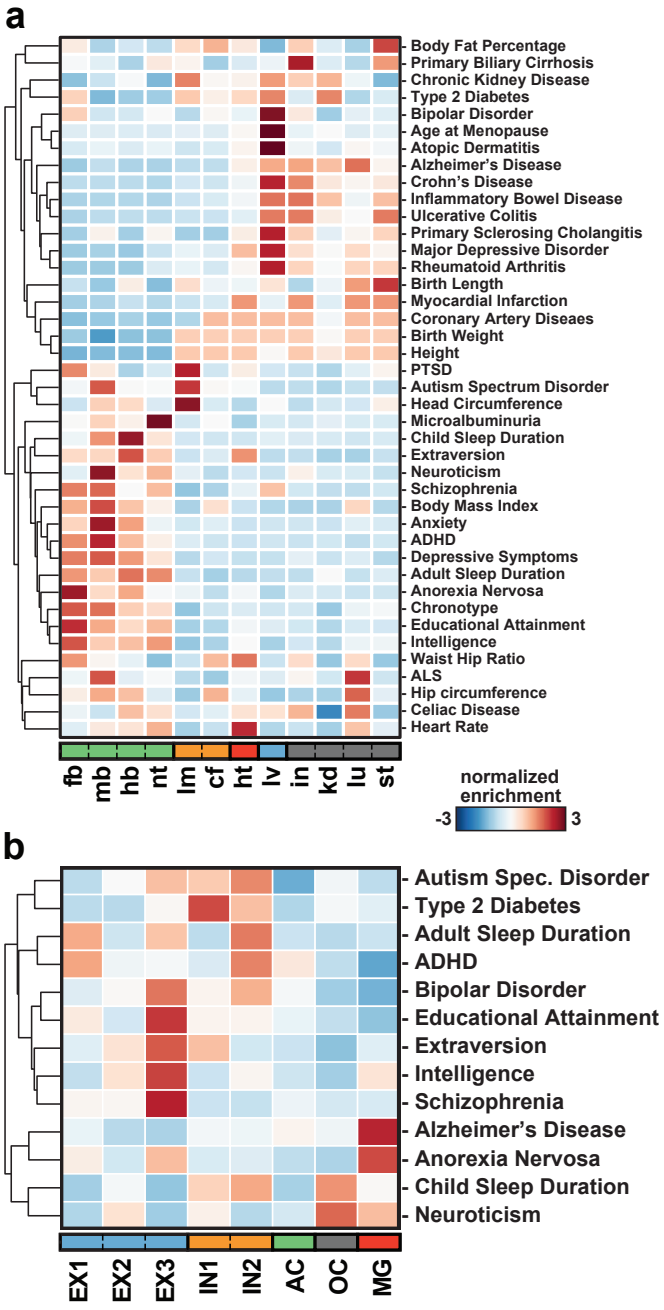

Figure 2.12. a. Enrichment of GWAS signal for complex traits and disease (y axis) within human orthologs of TSS-distal d-TACs from specific tissues (x axis) with polyTest[36]. Enrichment values plotted are signed -log10(p-values) which are z-score normalized within studies. **b.** Enrichment of GWAS signal for specific phenotypes (y axis) within TSS-distal accessible chromatin regions in specific cell types within mouse forebrain via scATAC-seq[45]. EX = Excitatory neurons, IN = Inhibitory neurons, AC = Astrocytes, OC = Oligodendrocyte, MG = Microglia.

2.3.7 Correlation-based network of d-TACs identifies potentially co-accessible regions

One application of an expansive map of cREs is the ability to infer potential associations and interactions based on correlations. Specifically, by using the read counts as a measure of degree of chromatin accessibility in each sample, we can build a correlative network of putative interactions across the catalog of d-TACs. We conclude that this map has some value as a potential map of enhancers to promoters (or enhancer to enhancer), since we know there is 3D conformation interaction where regions come into contact, so accessibility serves as an important proxy metric.

Constraining within embryonic TADs, we performed pairwise Pearson correlations between every potential pair of interacting d-TACs and took the ones with an r >= 0.7 to be a highly correlated interacting pair, resulting in a map of 2,700,062 total correlations (Figure 2.13). To assess the relationship between genomic distance and correlation, we plotted this and see that there is a distance decay, describing it.

We also observe that d-TACs close to each other in linear distance on a chromosome are more likely to have correlated activity across tissue-stages. These correlations are higher for elements in the same Topologically Associating Domain (TAD), providing further evidence of regulatory coordination during development at the level of TADs[47][48] and that they serve as a practical algorithmic constraint. A table of correlated d-TACs is provided in the supplemental material of Gorkin et al, 2019[24].

2.4 Discussion

In summary, we present a survey of the chromatin landscape in the developing mouse fetus that is unprecedented in its breadth. Our results describe a multi-tiered compendium of functional annotations for the developmental mouse genome, yielding valuable insight into key developmental processes and regulatory factors and expanding upon previous annotations. The

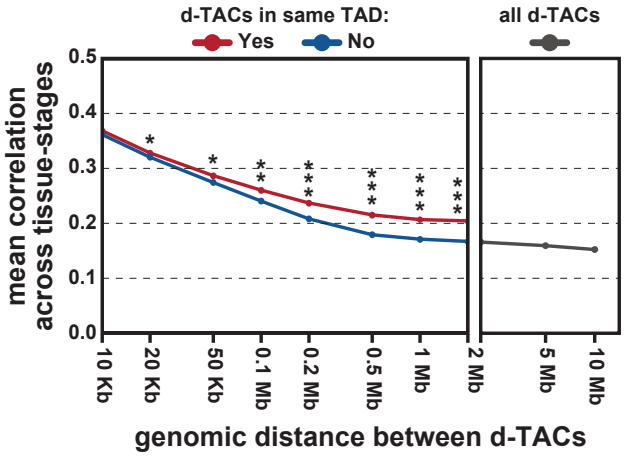

Figure 2.13. Correlation of ATAC-seq signal across tissue-stages is plotted as a function of genomic distance between d-TACs. In the left sub-panel, d-TACs are divided based on whether they are the same TAD (red line), or not (blue line). Statistical significant of the difference between those groups was measure by Wilcoxon signed rank test, and is indicated above each data point. *p<0.05, **p<0.005, ***p<0.0005.

resources detailed here include chromatin state maps for each tissue and stage, an extensive catalog of development candidate cREs, a genome-wide map of predicted enhancer target genes, and a collection of transgenic reporter assays that demonstrates a strong relationship between H3K27ac signal and likelihood of validation. Due to the uniquely critical role of the mouse as a model system in biomedical research, we believe that the tools and insights developed here will be a valuable resource to the biomedical research community.

Several key highlights of this chapter are listed below:

- Generated a comprehensive catalog of developmentally accessible chromatin in mouse.
- Developed standards and methods for processing and analyzing ATAC-seq data.
- Performed integrative, multiomics analysis to characterize the function of these regions.
- Explored the application of mouse developmental data for studying human disease.

2.5 Methods

2.5.1 Data collection

All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee. Tissue collection for all developmental stages was performed using C57BL/6N strain Mus musculus animals. For E14.5 and P0, breeding animals were purchased from both Charles River Laboratories (C57BL/6NCrl strain) and Taconic Biosciences (C57BL/6NTac strain). For all remaining developmental stages, breeding animals were purchased exclusively from Charles River Laboratories (C57BL/6NCrl strain). Wild-type male and female mice were mated using a standard timed breeding strategy. Embryos and P0 pups were collected for dissection using approved institutional protocols. Embryos were excluded if they were not the expected developmental stage. To avoid sample degradation, only one embryonic litter or p0 pup was processed at a time, and tissue was kept ice cold during dissection. Collection tubes for each tissue type were placed in a dry ice ethanol bath so that tissue samples could be flash frozen immediately upon dissection. Tissue from multiple embryos was pooled together in the same collection tube, and at least two separate collection tubes were collected for each tissue-stage for biological replication. Tissue was stored in a -80oC freezer or on dry ice until further processing. A step-by-step protocol for tissue collection, including detailed information about how embryonic stage was determined, can be found on the ENCODE Project website at: https://www.encodeproject.org/documents/631aa21c-8e48-467e-8cacd40c875b3913/@@download/attachment/StandardTissueExcisionProtoco_02132017.pdf.

2.5.2 Data generation protocol

Our full ATAC-seq protocol is available via the ENCODE data portal here: https://www.encodeproject.org/documents/4a2fc974-f021-4f85-ba7a-bd401fe682d1/@@download/attachment/RenLab_ATACseq_protocol_20170130.pdf. We required a minimum of 20 million usable

ATAC-seq read pairs per dataset (see methods), and a minimum Fraction of Read Overlapping TSS (FROT) of 0.1. We use FROT as a measure of signal-to-noise ratio in ATAC-seq datasets, because TSS are widely marked by open chromatin, even in tissues where the gene is not expressed. We calculate FROT for each library as the number of reads mapping within 1kb of a GENCODE v4 TSS[49], divided by the total number of usable reads. Our ATAC-seq data is highly reproducible between biological replicates of the same tissue stage as measured by Pearson and Spearman correlation. In addition, multidimensional scaling analysis of ATAC-seq enrichment across identified peaks confirms that the samples tend to cluster primarily by tissue types and then by developmental stage.

2.5.3 Data processing pipeline

ATAC-seq data were analyzed using a standardized software pipeline implemented by the ENCODE Data Coordinating Center (DCC)[50] for the ENCODE Consortium to perform quality-control analysis and read alignment. ATAC-seq reads were trimmed with a custom adapter script and mapped to mm10 using bowtie version 2.2.6[51] and samtools version 1.2[52] to eliminate PCR duplicates and mitochondrial reads. To center peaks on the Tn5 cut site, the paired-end read ends were converted to single-ended read ends and the read end was shifted 4bp towards the center of the fragment to account for the Tn5 insertion position by moving the read end towards the center of the fragment. MACS2 version 2.1.1.20160309[53] was used for generating signal tracks and peak calling with the following parameters: nomodel shift 37 ext 73 pval 1e-2 -B SPMR call-summits. To produce a set of replicated ATAC-seq peaks for analysis, the peak calling steps above were performed for each experiment on each pair of replicates independently as well as a pooled set of the two replicates. The intersectBed tool from the bedtools v2.27.1[54] suite was used to identify a set of replicated peaks which we define as the subset of peaks called in the pooled set, were also present in both of the replicate peak call sets.

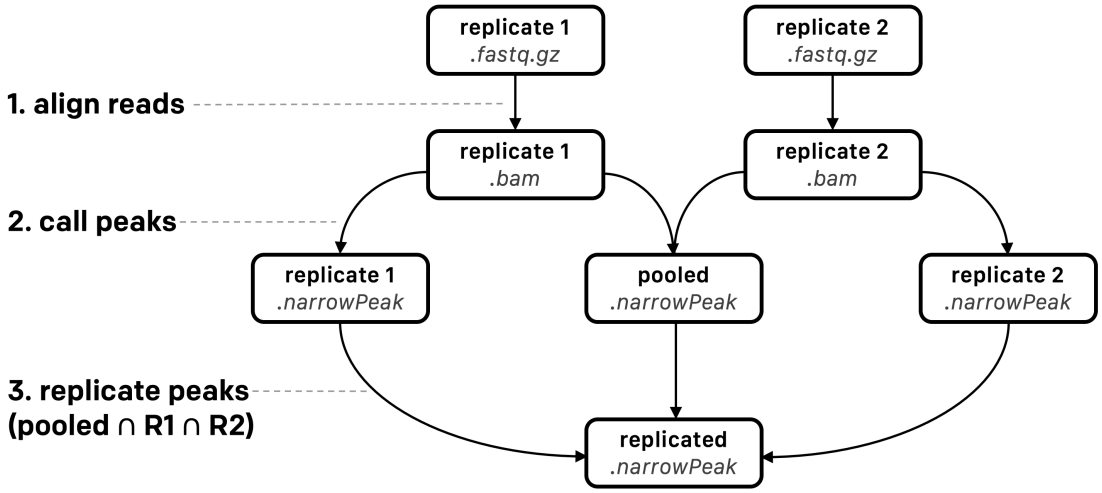

Figure 2.14. Flowchart overview of ATAC-seq data processing pipeline

2.5.4 Uniform d-TAC catalog generation

To obtain a uniform d-TAC catalog that can enable the multi-dimensional analysis across all 66 tissue-stages sampled over fetal development, the aforementioned replicated peak sets for each sample were concatenated, merged, sorted, and then labeled using the mergeBed and sortBed tools from the bedtools v2.27.1 suite. The intersectBed tool was used associate each uniform d-TAC with the original tissue-stages where its constituent peaks were accessible. The catalog was further categorized as being TSS distal or proximal based on a +/-1kb window around GENCODE v4 TSS. A flowchart diagram of this strategy is shown in Figure 2.14.

2.5.5 Sensitivity and specificity

To evaluate the sensitivity of our peak calls in detecting potential cis-regulatory elements, we calculated the true positive rate, or fraction of peaks recovered, for every applicable tissue-stage with respect to two reference sets using intersectBed from bedtools v2.27.1: (1) actively-transcribed promoters, and (2) enhancers from the VISTA enhancer database (accessed 7/22/17) with activity at e11.5. Using matched RNA-seq downloaded from www.encodedcc.org,

transcripts with counts of >=10 TPM were classified as actively transcribed for each tissue-stage. Catalog specificity was assessed by calculating the true negative rate of each tissue-stages d-TACs against the following two reference sets: (1) GENCODE v4 TSS that were not accessible to matched DNase-seq from www.encodedcc.org, and (2) enhancers from the VISTA enhancer database with no activity at e11.5. To further probe the tissue-specificity of the d-TAC catalog, the overlap between d-TACs for each tissue at e11.5 and enhancers showing activity in the matching tissue pattern was calculated, and compared to a background hit rate of enhancers with activity in any pattern, the enrichment significance of which was computed using a binomial test.

2.5.6 Enrichment of GWAS catalog variants in human orthologs of d-TACs

To enable comparison to GWAS of human phenotypes, we used liftOver with default settings to convert d-TACs from mm10 to hg19 genomic coordinates. We then defined novel d-TACs by removing those that overlapped DNaseI hypersensitivity sites from any cell line or tissue in two published datasets[10][20], one of which included embryonic tissues. We obtained index variants for all traits in the GWAS catalog (https://www.ebi.ac.uk/gwas/api/search/downloads/full) and retained a unique set of variants that were identified as genome-wide significant (P<5x10-8) in GWAS of European ancestry individuals. To obtain a background set of variants for enrichment testing, we used the filtered index variants as the input for SNPsnap[35], which matches based on (1) minor allele frequency, (2) distance to the nearest annotated gene, (3) gene density in the surrounding region, and (4) number of SNPs in linkage disequilibrium (LD), with the following parameters: European population, 10 matched SNPs, exclude HLA SNPs and input SNPs, and report clumping. As GWAS index variants are not necessarily causal and can be in LD with the true causal variant, we next defined loci for all index and matched background variants as all SNPs in high LD (r**2 >0.8) with the variant in European 1000 Genomes[55] samples using PLINK[56]. We then calculated the the number of GWAS and background loci with at least one variant overlapping either all d-TACs or novel d-TACs and used a hypergeometric test to assess

enrichment significance of GWAS loci compared to matched background loci.

To test for enrichment of complex phenotypes and diseases with publicly available summary statistics, we first defined sets of human orthologs of enhancer d-TACs. For each tissue, we collapsed all strong and weak enhancer chromatin states (En-Sd, En-Sp, En-W) across timepoints and used liftOver to convert genomic coordinates from mm10 to hg19. We then intersected orthologous enhancers with orthologous d-TACs to obtain a set of orthologous enhancer d-TAC for each tissue. We collected summary statistics for 41 human traits and diseases, converting odds ratios and confidence intervals to log odds ratios and standard errors for binary traits and estimating allele frequencies from the European subset of 1000 Genomes[55] where unavailable from the summary data. We used polyTest[36] to test for enrichment of variant effects on each phenotype within orthologous enhancer d-TAC annotations with the parameters –univariate –maf 0.05 –high-mem. We used hierarchical clustering on signed -log10(p-values) for enrichment that were z-score normalized within studies to group similar phenotypes.

2.5.7 Correlative d-TAC interaction map

Using the ATAC-seq read counts measured for the entire d-TAC catalog across all 66 tissue-stages normalized to RPKM and log2-transformed with a small pseudocount, a correlative map between d-TACs was generated for each chromosome by calculating the Pearson correlation coefficient (PCC) for each pair of d-TACs. Of these correlations, only those with a PCC >= 0.7 and represented a pair of d-TACs whose coordinates lie within the same TAD, as defined by mouse ES cells, were retained as a predicted interacting pair. The correlative map was binned into 100kb bins for the purpose of visualizing the interactions as a heatmap.

To assess the nature of distance decay in correlations as a function of genomic distance, we assigned each d-TAC to 10kb bins. For each bin A, the correlation was measured between its d-TACs and those of bin B, at various distances away ranging from 10kb to 2Mb. The average of these correlations across all chromosomes is plotted as a function of distance. Additionally, to investigate the validity of using mESC TAD boundaries as a constraint for the correlative map,

the mean correlation between d-TACs at various genomic distances were compared for pairs located within the same TAD to that of those not sharing a TAD. At several genomic distances, this mean correlation was computed for each chromosome, from which the significance of the difference in correlation between intra-TAD and inter-TAD d-TAC pairs was calculated using the Wilcoxon signed rank test.

2.5.8 Data access

Supplementary files associated with this study such as chromHMM calls, dynamic dTAC and enhancers, and ATAC-seq peak calls are available at:
http://renlab.sdsc.edu/renlab_website//download/encode3-mouse-histone-atac/.

The ATAC-seq pipeline for aligning reads and calling peaks is available at:
https://github.com/yuz207/atac-seq-pipeline.

Code notebooks for analysis performed in this thesis and generating figures can be found at: https://github.com/yuz207/encode3-atac.

2.6 Acknowledgments

We thank DU Gorkin for this guidance and assistance with study design and direction, data analysis, and manuscript and figure preparation. We thank Y Zhang and I Barozzi for their technical assistance with analysis and manuscript preparation. We thank AH Lee and H Huang for performing experiments. This study was funded by the National Human Genome Research Institute as part of the Encyclopedia of DNA Elements (ENCODE) project (U54HG006997), and was performed in compliance with all relevant ethical regulations. D.U.G. supported by the NIH Institutional Research and Academic Career Development Awards (IRACDA) program, and an A.P. Giannini Foundation fellowship. D.E.D, A.V., and L.A.P. were also supported by UM1HG009421, and research conducted at the E.O. Lawrence Berkeley National Laboratory was performed under Department of Energy Contract DE-AC02-05CH11231, University of California. I.B. is funded through an Imperial College Research Fellowship. Y.H. is supported by

the H.A. and Mary K. Chapman Charitable Trust. J.R.E. is an Investigator of the Howard Hughes Medical Institute. We thank Alex Oakenman for contributing to the organ-specific graphic icons. The embryo image second from the right in Figure 2.1 was adapted from Paudyal et al. 2010 (PMCID: PMC2930600), an Open Access article distributed under the terms of the Creative Commons Attribution License.

Chapter 2, in part, has been submitted for publication. DU Gorkin*, I Barozzi*, Y Zhao*, Y Zhang*, H Huang*, AY Lee, B Liu, J Chiou, A Wildberg, B Ding, B Zhang, M Wang, JS Strattan, JM Davidson, Y Qiu, V Afzal, JA Akiyama, I Plajzer-Frick, CS Novak, M Kato, TH Garvin, QT Pham, AN Harrington, BJ Mannion, EA Lee, Y Fukuda-Yuzawa, Y He, S Preiss, S Chee, JY Han, BA Williams, D Trout, H Amrhein, H Yang, JM Cherry, W Wang, K Gaulton, JR Ecker, Y Shen, DE Dickel, A Visel, LA Pennacchio & B Ren. An atlas of dynamic chromatin landscapes in the developing mouse fetus. (* Authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3

Integrative analysis of chromatin state and accessibility dynamics

3.1 Abstract

Transcriptional regulation plays a central role in animal development. Dynamic interactions between transcription factors and *cis* regulatory sequences drive spatiotemporal expression of genes during development and dictate how an organism grows, adapts to environment, and reproduces. Birth and turnover of *cis*-regulatory sequences, in addition to gains and losses of activity of genes encoding transcriptional factors, play a major role in evolution of species.

Leveraging the temporal dimension in our embryonic mouse data series, we performed integrative multiomics analysis to analyze the dynamics of developmentment, applying a combination of orthogonal datasets (histone ChIP-seq, transcriptomics, and chromatin accessibility data). We profiled the changes that occur over the development in our datasets, capturing the emergence and depletion of cell lineages as tissues develop and differentiate. With these findings, we are able to identify key regulators for each tissue and help elucidate the gene regulatory mechanisms during mammalian embryogenesis.

3.2 Introduction

The embryonic chromatin landscape is a complex, tightly regulated network of interacting transcription factors and regulatory sequences, dysregulation of which plays a direct role in the development of a variety of adult diseases and traits. Transcriptional regulation is at the heart of these processes, driven by the activity of transcription factor proteins that bind to regulatory elements in a sequence-specific manner[57]. Specifically, pioneering factors and so-called master regulators bind to critical loci within the genome, directing the cells to differentiate down their epigenetic landscapes to their final specified fates[58]. However, TFs do not act in a vacuum. Pioneering factors recruit other TFs and the collaborative binding and coordination of these diverse proteins – along with the vast number of putative enhancers, numbering in the hundreds of thousands – are ultimately responsible for the ability of a totipotent cell to become such a wide-range of lineages[59].

Naturally, the ability to dissect these dynamic processes over the time course of development is an extremely powerful biological tool. Since many adult diseases have origins that can be traced to early development and disruption in developmental regulatory processes can also lead to congenital disorders, there is tremendous translational value in these types of analyses. Unfortunately, in addition to the dearth of resources that profile the developmental time series, the availability of human samples at crucial time points such as fetal development remain limited for political and ethical reasons. Here, we present a number of computational strategies and integrative multiomics methods to extract relevant insights related to the chromatin remodeling during development by studying the differential changes that occur between sequential stages and interrogating the relationship between these changing regions and their functional chromatin state annotations.

3.3 Results

3.3.1 Identification of temporally dynamic regions of accessible chromatin between sequential stages

To more directly assess the temporal dynamics of chromatin accessibility during development we performed differential peak analysis, identifying d-TACs that show significant gain or loss of accessibility between sequential stages within a tissue. We defined sequential log fold change between sequential developmental stages for each active dT-AC in a given tissue type, and quantify the gain or loss of transposase accessibility for each d-TAC at every transition point. d-TACs exhibiting significant change in transposase accessibility in at least one such transition were classified as dynamic. Through this approach, we are able to describe each d-TAC in our catalog by its precise dynamic temporal pattern in each tissue.

In total, we identified 140,051 d-TACs that exhibit significant change of accessibility in at least one stage transition (27% of all d-TACs) (Figure 3.1), of which the overwhelming majority (127,354, or 90.9%) are TSS-distal regions (Figure 3.1a). We observe that most dynamic d-TACs show a significant change in accessibility at only one stage transition within a tissue in this developmental window (Figure 3.2, top), suggesting that these changes reflect enduring shifts in cell fate and/or composition rather than rapid on-off molecular switches. Thus, given that these data are bulk rather than single cell, one may interpret these patterns to represent the emergence of new cell lineages in concordance with increased tissue complexity over time. In other words, the temporally dynamic patterns of chromatin accessibility observed reflect highly tissue-specific developmental programs, rather than global tissue-agnostic changes.

3.3.2 Developmental pathways associated with tissue-specific patterns of accessibility

The tissue-specific patterns of transposase accessibility at the TACs revealed biological insights into mouse fetal development. Just as we examined the number of stage transitions

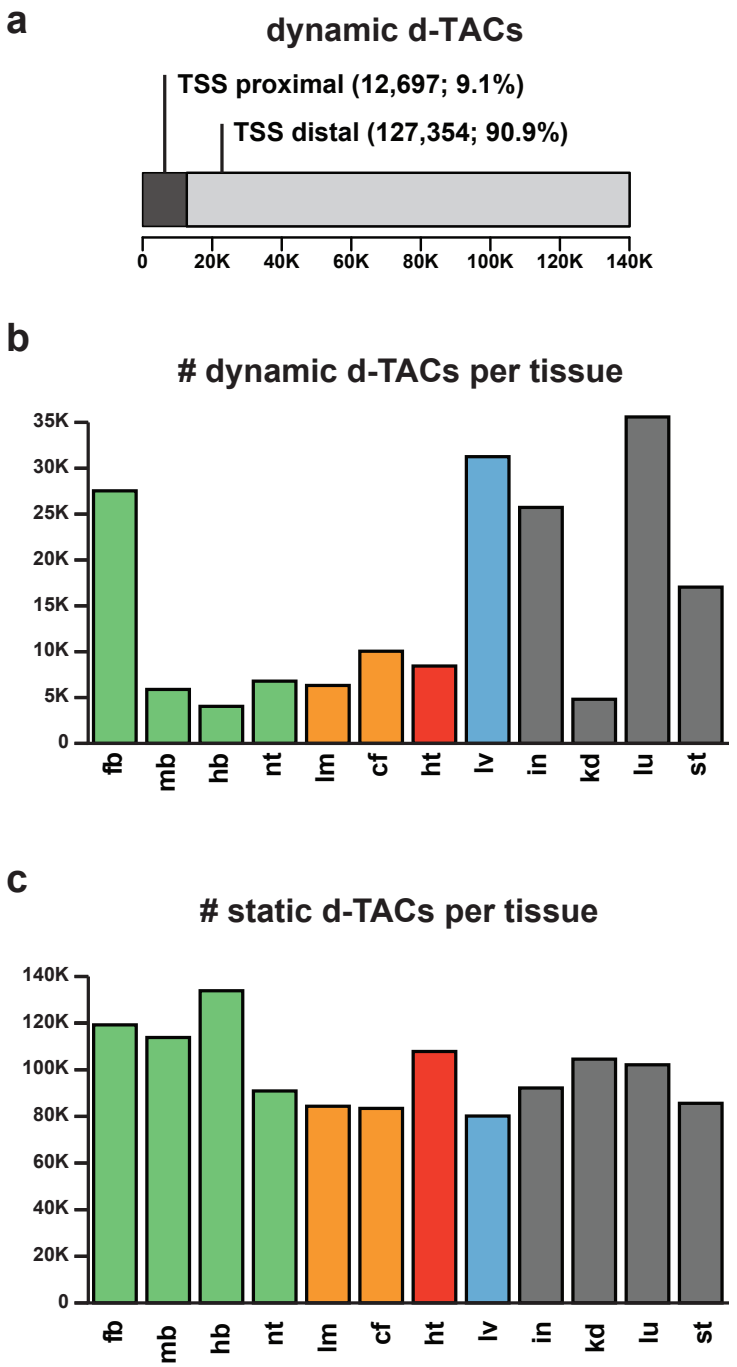

Figure 3.1. Overview of dynamic d-TACs identified in our catalog. If a d-TAC was called as significantly dynamic at any stage transition within it a tissue it was labeled as dynamic; otherwise it was labeled as static. **a.** The relative number of TSS-distal versus TSS-proximal dynamic d-TACs, which comprise the majority of dynamic elements. **b.** The breakdown of dynamic d-TACs categorized by tissue. **c.** The breakdown of static d-TACs categorized by tissue.

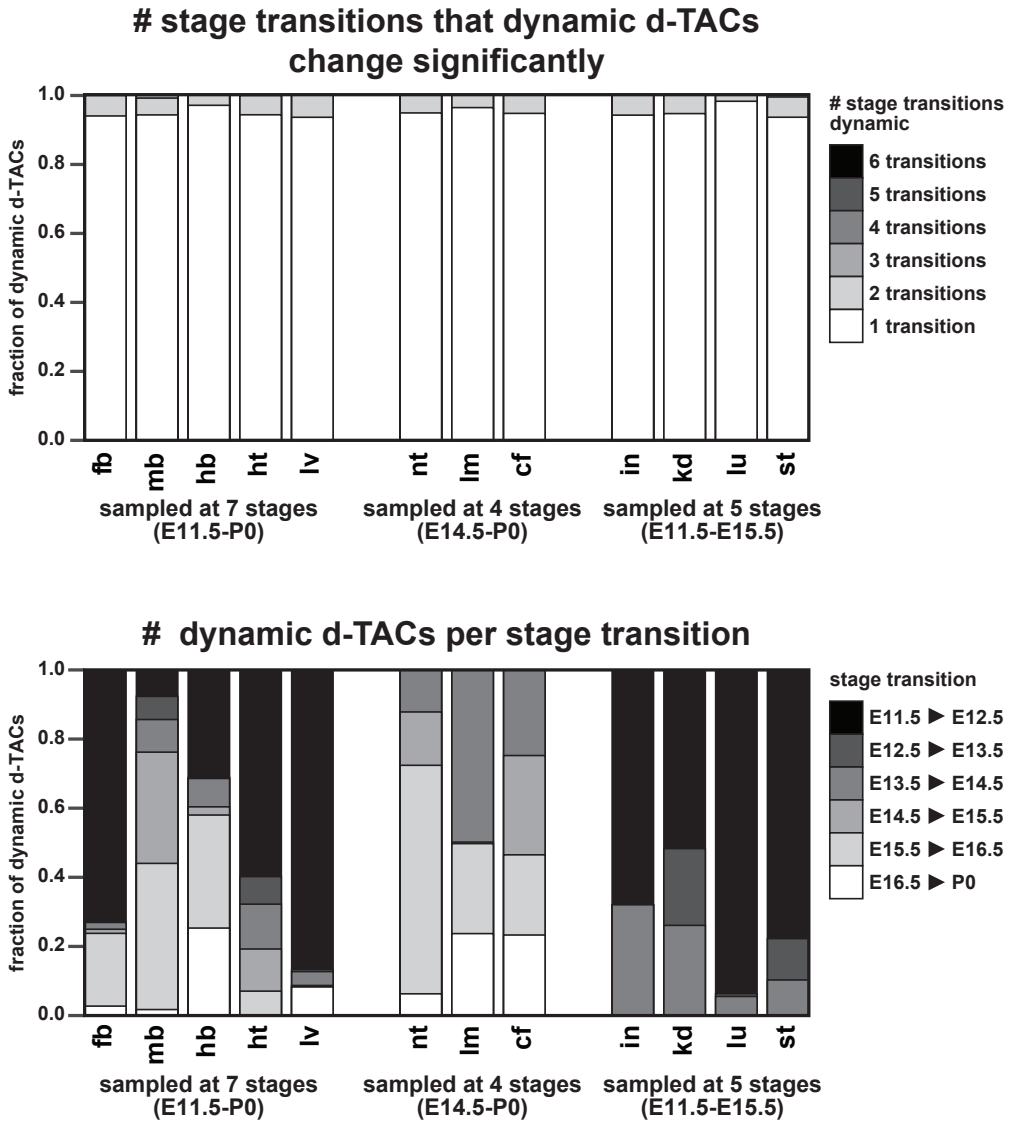

Figure 3.2. Summary of global trends observed in dynamic elements. **a.** Stacked bar plot shows the fraction of dynamic d-TACs in each tissue that are dynamic at one, two, three, four, five, or six stage transitions. Most dynamic d-TACs undergo significant changes in accessibility at only one stage transition within a tissue. **b.** The fraction of dynamic d-TACs within a tissue that undergo significant changes in accessibility at each stage transition.

Table 3.1. Number of dynamic d-TACs identified for each mouse embryonic tissue, categorized as gain of accessibility or loss of accessibility.

	Loss of accessibility	Gain of accessibility	Total dynamic d-TACs
Forebrain	12,986	16,307	27,537
Midbrain	2,893	4,311	6,678
Hindbrain	1,944	2,219	4,034
Neural Tube	3,536	3,617	6,788
Limb	2,361	4,199	6,319
Craniofacial	4,355	2,265	6,401
Heart	5,668	3,273	8,434
Liver	14,905	18,453	31,265
Intestine	10,120	17,133	25,730
Kidney	1,757	3,228	4,739
Lung	17,081	19,199	35,594
Stomach	7,000	9,607	15,494

in which a given d-TAC is exhibiting significant change in accessibility, the corrolary to this analysis to examine the stage transitions featuring the largest amount of turnover in terms of number of dynamic d-TACs. For most tissues, there is typically one stage transition when the majority of dynamic TACs exhibit a significant change in transposase accessibility (Figure 3.2, bottom). The precise stages of importance varied, with closely related tissues featuring more similar dynamic patterns. For example, forebrain development shows the most dramatic changes from e12.5 to e13.5 while the four endoderm derived tissues are more dynamic later at e16.5 to p0. Interestingly, this shift from e16.5 to p0 is a period during which a relatively high proportion of TACs exhibit turnover across several tissues. The heightened level of chromatin remodeling observed during this period can be attributed to both a larger temporal gap between sampling (˜2.5 days vs 1 day between other stages) as well as the considerable change to mouse metabolism and physiology at birth. The exact number of d-TACs gaining and losing accessibility respectively, and their totals, are shown for each tissue in Table 3.1.

Closer examination of dynamic d-TACs in each tissue reveals that these patterns resemble known developmental processes. For instance, the chromatin dynamics in embryonic forebrain

are dominated by remodeling in two particular stage transitions: from e12.5 to e13.5; and from e16.5 to p0 (Figure 3.3). These two periods coincide with major periods of growth in the brain, neurogenesis and gliogenesis respectively. Both require complex coordination of transcriptional regulators as the brain develops its adult structural organization and are characterized by a rapid expansion of newly differentiated cell types. During neurogenesis, neuronal stem cells and neural progenitor cells differentiate and give rise to post-mitotic neurons[64]. As a representative example, dynamic d-TACs upstream of Neurod2, a canonical marker gene of mature, adult neurons, exemplify the dramatic gain of accessibility from e12.5 to e13.5. Functional enrichment of loci characterized by a significant gain in accessibility during this stage transition using the GREAT tool[61] establishes associations between these d-TACs and biological processes related to the IGF-1 signaling pathway, a key growth factor in neurogenesis and synaptogenesis, and neuronal differentiation. Likewise, d-TACs characterized by a strong loss of accessibility are involved in the negative regulation of neuron development. Similarly, dynamic d-TACs from e16.5 to p0 show enrichment for processes promoting oligodendrocyte differentiation and are depleted for forebrain neuron development, reflecting a shift away from neurogenic activity and towards gliogenesis, which corresponds to the emergence of non-neuronal cell populations such as astrocytes and oligodendrocytes.

Interrogating heart-associated d-TACs also reveals insights into the signaling pathways driving to the development of the embryonic mouse heart. Unlike most tissues in this study, which re characterized by active remodeling throughout gestation, the heart is unique in that the majority of dynamic changes cease by e14.5 and heart-associated d-TACs are relatively static from e14.5 to p0 (Figure 3.9). This observation speaks to the earlier maturation of heart, which is mostly developed by birth. Pre-e14.5 heart d-TACs are functionally enriched for growth and morphogenesis of critical cardiac chambers, the atrium and ventricles, reflecting the development of vascular structure. Simultaneously, the cessation of dynamic accessibility at e14.5 is accompanied by the depletion of myeloid differentiation. Furthermore, dynamic heart-associated d-TACs associated with later stages fail to show enrichment for developmental

processes, only depletion of cardiac morphogenesis, offering further evidence that the heart is functionally matured quickly compared to other tissues.

The most dynamic of the tissues collected in this study is the liver, which undergoes the greatest transformation in cellular identity during embryogenesis (Figure 3.14). Early in gestation, the liver is primarily responsible for the generation of blood lineages and immune cells, and only late in development does the liver composition change to reflect the metabolic functions of the adult liver. This metamorphosis is reflected in the liver-associated d-TACs, a substantial fraction of which manifest dynamic accessibility behaviors from e16.5 to birth. These d-TACs are marked by a profound depletion of processes related to myeloid cell and erythrocyte differentiation, which being enriched for carbohydrate and lipid metabolism.

The corresponding heatmaps for all tissues profiled may be found in the Chapter 3 Appendix.

3.3.3 Clusters of stage-transition-specific dynamic d-TACs reveal developmental regulatory program

Using the catalog of mouse d-TACs and the annotations of their tissue-specific dynamic patterns, we describe the key regulators of the transcriptional network responsible for specifying cell fates, which is crucial to fully understand the regulatory landscape during embryogenesis. Since heightened ATAC-seq signal marks regions of increased availability to transcription factor binding, changes in the accessibility of regions containing TF binding sites can unveil potential key regulators those respective stages. Therefore, examination of binding motifs both the constantly accessible (non-dynamic) sets of d-TACs as well as those associated with dynamic behavior for specific tissues and stages can shed light on the role of various TFs and their coordination in each of these contexts. To elucidate these key regulators, we systematically enriched known binding motifs from the JASPAR database in the d-TAC catalog[60]. In addition to analyzing the non-dynamic d-TACs, each dynamic d-TAC was categorized with the tissue and stage transition in which it displayed its largest gain or loss of accessibility. By grouping

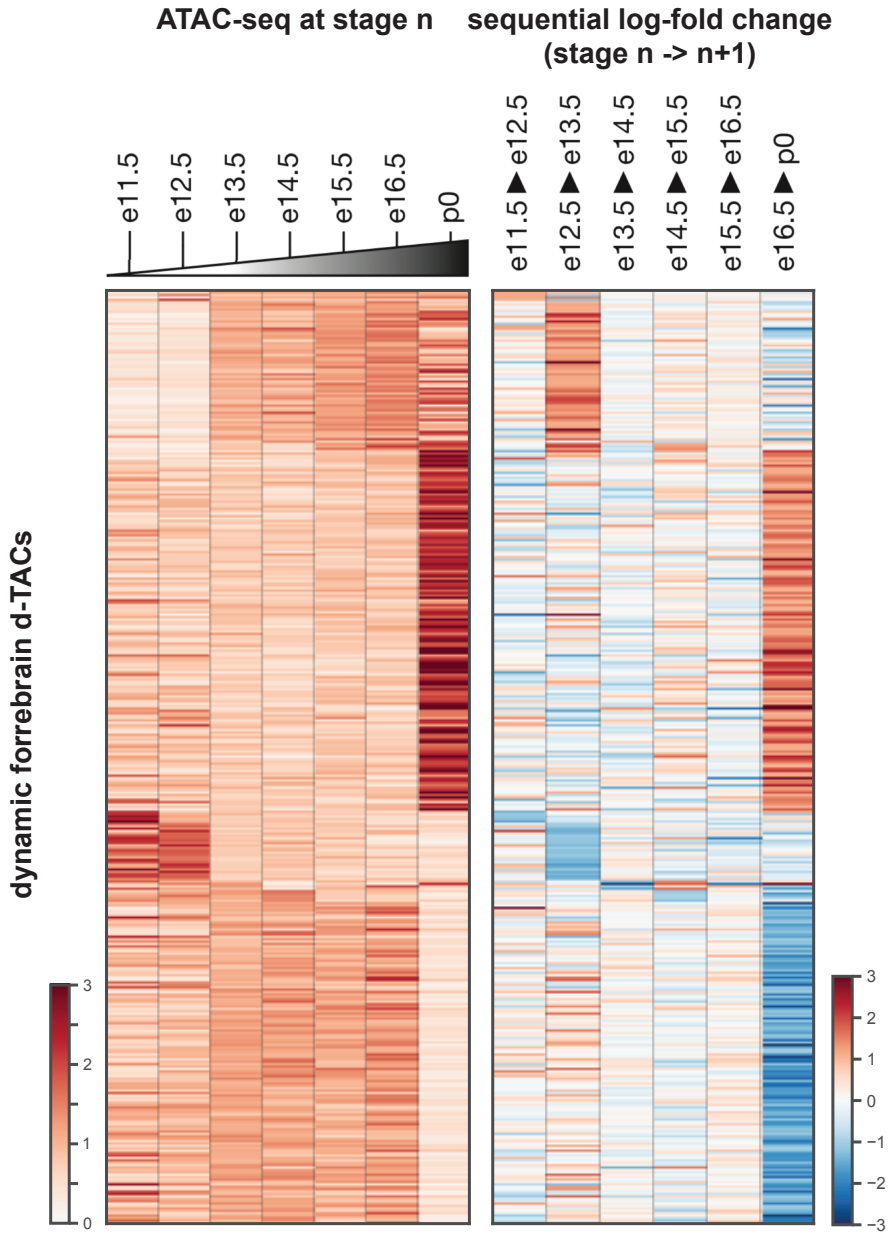

Figure 3.3. ATAC-seq and sequential logFC heatmaps for forebrain. d-TACs can be described by their pattern of dynamics over the course of a tissues development, reflecting the regulatory changes associated with the emergences or depletion of cell types. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in forebrain (e11.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition. For heatmaps of the other tissues, see the Chapter 3 Appendix.

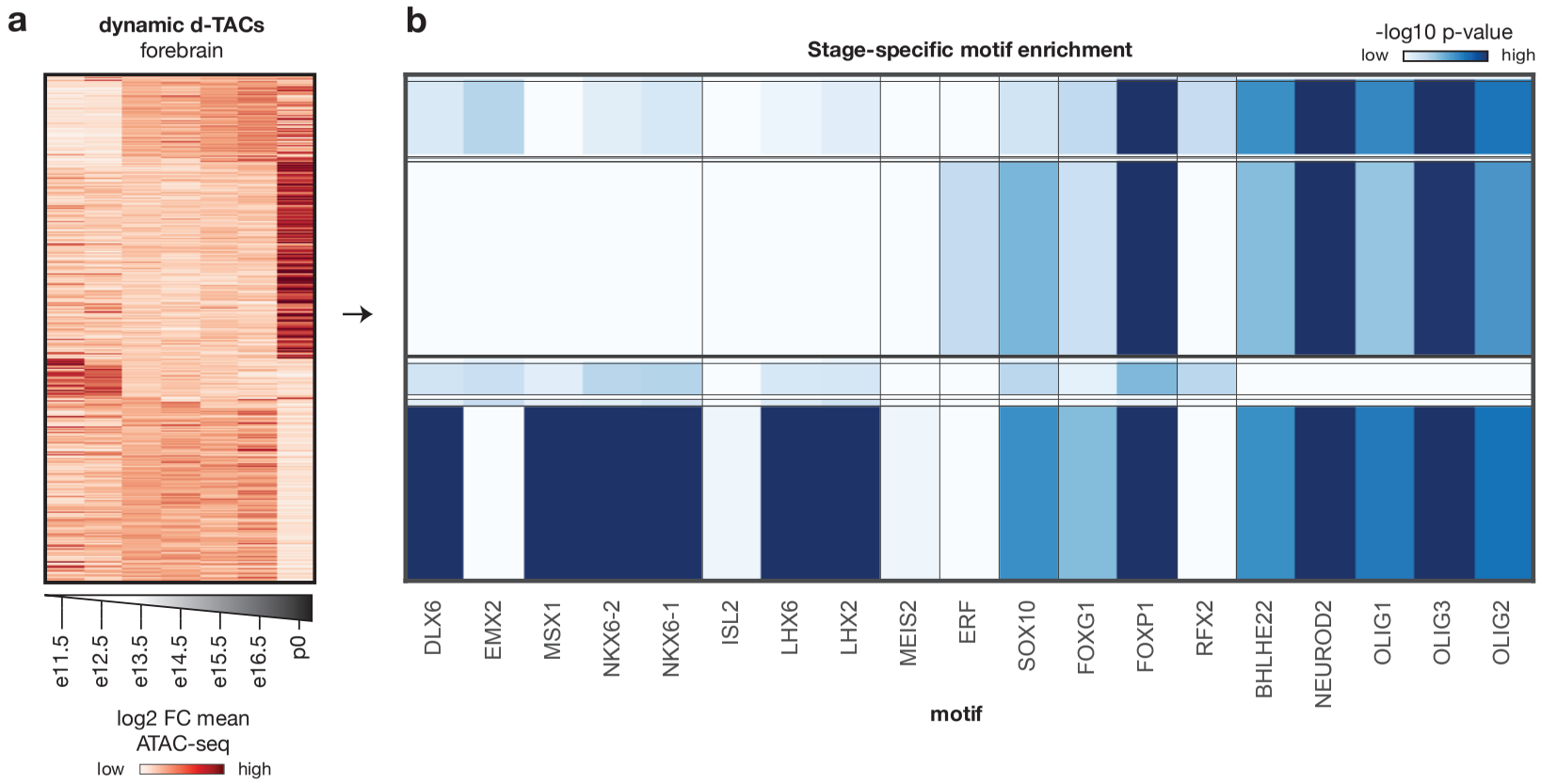

Figure 3.4. Motif analysis for dynamic d-TACs in forebrain. **a.** Heatmap of ATAC-seq accessibility signal of dynamic forebrain d-TACs, grouped by stage transition of largest absolute change. **b.** For each clusters of dTACs, grouped by gain or loss of accessibility in a given stage transition, motif analysis was performed to identify enriched motifs. Selected motifs are shown for several motif families. (Not shown) For each motif family, normalized transcriptomics data from matched RNA-seq is analyzed to identify potential matching TFs.

dynamic d-TACs in this manner, we are able to associate enriched motifs in each group with its context-specific accessibility profile.

Examining the binding motifs enriched in non-dynamic subset of the d-TAC catalog, or elements characterized by relatively constant and unchanging levels of transposase accessibility, reveals noteworthy trends across the panel of tissues characterized. For non-dynamic d-TACs, the patterns of motif enrichment are unsurprisingly relatively static between stages within a given tissue. However, when comparing TSS proximal d-TACs against TSS distal ones, it becomes apparent that motif family enrichments for proximal d-TACs are mostly invariant even across tissues. In contrast, while static temporally, non-dynamic TSS distal d-TACs do feature highly tissue-restricted patterns of motif enrichment, reflecting the prior observation that proximity to TSS (promoter-like) is associated with accessibility in a larger number of tissue types whereas distal d-TACs represent more distinct sets. Together, these organism-wide motif enrichment analyses describe the complex global regulatory map of TFs in non-dynamic d-TACs during this time series.

Similar analysis of dynamic d-TACs, both gain and loss of accessibility, uncovers the reciprocal component of the regulatory program in mouse development. Like the TSS distal non-dynamic elements, the motif enrichment patterns for dynamic d-TACs tend to be highly specific to each tissues respective regulatory programs. However, unlike the non-dynamic d-TACs, the motifs enriched can also be highly disparate between stages, presumably reflecting the TFs specific to each stage transition. For example, in the developing forebrain, we observe that the highly dynamic transition from e12.5 to e13.5, associated with neurogenesis, is highly enriched for basic helix-loop-helix (bHLH) family motifs (Figure 3.4b). These include Neurog1/2, Neurod2, Atoh1, Olig1-3, and Olig-related transcription factor Bhlhe22, with known functions in cell proliferation and lineage differentiation[64]. However, since closely related binding motifs may feature similar PWMs, motif enrichment analysis alone is insufficient for identifying the specific transcriptional regulators acting on each pattern. To address this problem, we integrated the TF gene expression data that were also generated for matched tissues and stages

in this phase of ENCODE, which allowed us to identify which TFs in each binding family were active and potentially driving the developmental pattern (Figure 3.4c). Gene expression patterns show downregulation of Olig2/3, Twist1/2, and Hes5 during this stage. Concurrently, Olig1 and Neurod2/6 are acutely upregulated. Of note, human orthologs Olig1/2 are located on Chromosome 21 in the Down syndrome critical region, and overexpression of Olig1/2 is observed in both Down Syndrome and the Ts65Dn mouse model[65]. Olig3 is likewise important in early neurogenesis and aberrant signaling has been associated with development of medulloblastoma[66]. Saethre-Chotzen syndrome, when caused by reduced Twist expression secondary to deletion[67][68], is likely to be associated with developmental delay and intellectual disability in addition to craniofacial and tissue abnormalities. On the other hand, Hes5, expressed on neural precursors and covering the developing embryo, inhibits the expression of proneural genes[69].

In addition to bHLH factors, the coordinated expression of TFs belonging to the Forkhead and Homeobox families appears responsible for distinct temporal patterns in other stages. For example, while both Rfx and Fox motifs are enriched during neurogenesis, during the transition from e16.5 to p0, these two Forkhead subfamilies tend to be diametrically associated with loss of accessibility and gain of accessibility respectively. FOXP1 is an outlier in that its motif is heavily represented in both groups of d-TACs from dynamic in e16.5 to p0, while its expression profile displays downregulation at e13.5 and a sharp increase in regulation at p0[70]. Notably, heightened expression of FOXP1 has been implicated in autism spectrum disorder pathologies[71][72]. Homeobox domain motifs such as Lhx6, deficits of which linked to neuropsychiatric disorders[73][74], are similarly co-associated with bHLH motifs during neurogenesis but primarily enrich in the differentially inaccessible motifs at p0. However, the factor most emblematic of this deeply interconnected regulatory landscape is Bhlha15, which despite being a member of the bHLH family, is enriched for dynamic elements shared with Forkhead and Homeobox factors but uniquely absent during neurogenesis. The differential expression of both proneural and inhibitory factors underscore the complex interplay of regulatory

factors during this critical period.

Motif enrichment of heart d-TACs reveals novel insights into the regulatory program governing cardiogenesis. The dynamic patterns of fetal heart are clearly bifurcated into two phases: a developmentally active phase through e14.5, and a relatively static period from the middle stages until birth. During this first phase, the early stages are accompanied by a previously undescribed switch from myocyte enhancer factor (MEF) proteins to a combination of GATA and C2H2 zinc finger transcription factors. Elements exhibiting loss of accessibility at each of these stages are highly enriched for members of the MEF2 family, peaking at e13.5 to e14.5, when the heart undergoes a substantial reduction in chromatin remodeling. These factors, particularly MEF2b/c, have been studied extensively given their wide-ranging roles in development across many tissues and are believed to be interact with a vast number of target genes. In the heart, MEF2 is highly conserved and plays a critical role in cardiac hypertrophy[75] and gene regulatory program of cardiomyocytes[76]. Furthermore, these proteins have been implicated in a variety of cancer types, potentially acting as oncogenes and tumor suppressor genes. Concomitantly, GATA motifs are enriched for d-TACs that gain accessibility throughout this first phase, also peaking by e14.5. GATA is a particularly well-studied TF in congenital heart disease (CHD)[**?**] that is believed to regulate MEF[77]. Also exclusive to this key developmental transition from e13.5 to e14.5 is the enrichment of basic leucine zipper (bZIP) motifs, such as JUNB and FOSL2, which appear to contribute to the regulation of cardiac morphogenesis along with GATA family TF. Thus, given the biological function of MEF and GATA, it is likely that this switch reflects the maturation and subsequent reprogramming of existing populations of cardiomyocytes rather than the expansion or differentiation of new cell types.

3.3.4 Significant changes in accessibility linked to changes in chromatin state annotations

Having identified distinct patterns of accessibility change in each tissue that correspond to tissue development, we examined the regions featuring dynamic chromatin remodeling to

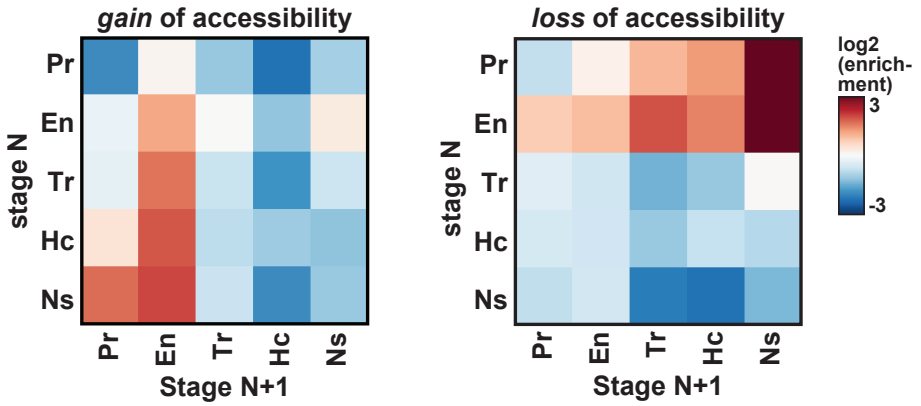

Figure 3.5. Heatmaps showing the chromatin state changes that occur at dynamic d-TACs that gain accessibility at a given stage transition (left), or lose accessibility at a given stage transition (right). Enrichment relative to coverage of each state in total d-TAC catalog. Pr is a superset composed of states Pr-A, Pr-W, Pr-B, Pr-F; En is a superset composed of En-Sp, En-Sd, En-W, En-Pp, En-Pd; Tr is a superset composed of Tr-S, Tr-I, Tr-P. Hc is a superset composed of Hc-A, Hc-H. Chromatin state definitions can be referenced in Figure 2.8.

determine if there was a link between accessibility changes and chromatin state annotations. To do this, we separately analyzed regions that gained accessibility as opposed to those that became less accessible and looked for an enrichment of certain state transitions before and after the gain or loss respectively. Analysis of chromatin state changes at dynamic d-TACs reveals that gain or loss of accessibility does in fact often correspond respectively to gain or loss of active enhancer chromatin states (Figure 3.5).

The most common state change pattern associated with a gain of accessibility was from No Signal to Enhancer-like. Within these broader categories, the most frequent originating states were the Transcription-Permissive, Enhancer-Poised, and No Signal states and the most frequent resultant states were Enhancer-Weak and Enhancer-Poised. That being said, the No Signal state did become other functional states with some level of frequency, in particular, promoters-like states. In contrast, the precise reciprocal trends were observed for dynamic loci featuring a loss of accessibility between stages.

3.3.5 Characteristic order of accessibility and histone dynamics observed in enhancer 'life cycle'

Given the coordination between chromatin state and accessibility dynamics, we sought to determine if there was a characteristic order to an enhancer becoming functional. To do this, we investigated the enhancers that were deemed strong, replicated enhancers based on their chromatin state that also overlap a d-TAC. Taking the subset of these strong enhancers that overlap a d-TAC, we examined the ones that featured dynamic H3K27ac (termed a dynamic enhancer), and evaluated the corresponding change in accessibility (ATAC-seq) at preceding and following stages. We note that most dynamic enhancers overlap d-TACs (67-88%, median 84%), though fewer overlap dynamic d-TACs (5-35%, median 14%) (Figure 3.6).

We found that there is a characteristic order in which accessibility and modification by H3K27ac appears to take place (Figure 3.7). For instance, chromatin accessibility increase tends to precede marking with H3K27ac by several stages, indicating that the chromatin around a regulatory sequence first becomes open before becoming functional as an enhancer. Conversely, the loss of accessibility follows the loss of H3K27ac. This trend was a global phenomenon and held true for all tissues.

3.4 Discussion

In this chapter, we characterized the changes in accessibility of this catalog, highlighting key tissue-specific and temporally-restricted patterns driving the respective regulatory programs of each tissue. We used an innovative approach to evaluating temporal dynamics by evaluating the specific changes that occur between sequences stages in the samples, rather than against an initial state or mean. By applying differential analysis strategies to sequence stage transitions, we are able to identify specific regions that undergo chromatin remodeling, reflecting developmental changes to the make up and composition of heterogeneous tissues. These characterizations also enable the study of transcription factor binding and the identification of potential regulators

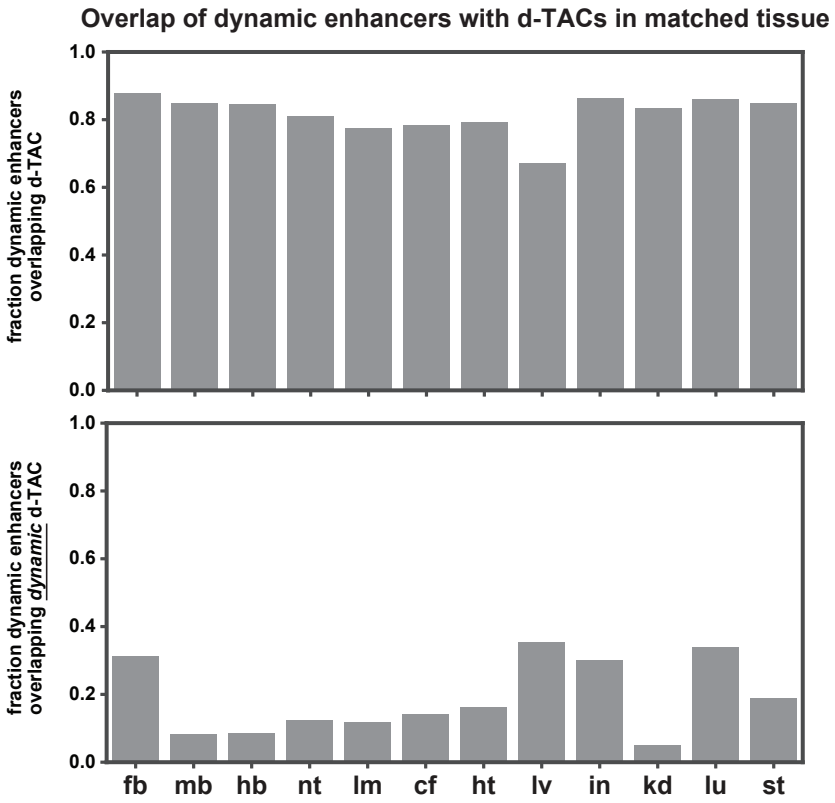

Figure 3.6. Barplot (top) shows the fraction of dynamic enhancers in each tissue that overlap d-TACs accessible in the matching tissue. Barplot (bottom) shows the fraction of dynamic enhancers in each tissue that overlap d-TACs that were also called as dynamic by ATAC-seq in the matching tissue.

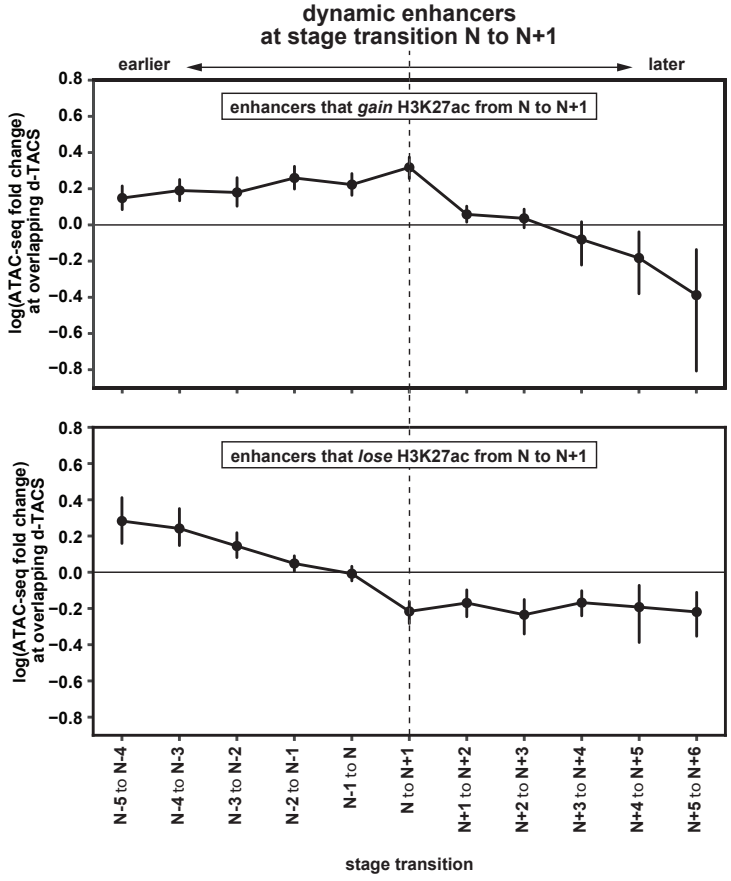

Figure 3.7. Top panel considers dynamic enhancers that gain H3K27ac at a given stage transition N to N+1. Line plots show the log2 fold change of ATAC-seq signal at d-TACs overlapping those enhancers at the matching stage transition, as well as before that stage transition (left half of plot) and after that stage transition (right half of plot). Dynamic enhancers that gain H3K27ac at a given stage transition tend to gain accessibility as measured by ATAC-seq either at or before that stage transition in question (sometimes preceding H3K27ac gain by as much as 5 stage transitions). Mean and standard deviation across all stage transitions are indicated by filled circles and vertical lines, respectively. Bottom panel is similar to top, but showing dynamic enhancers that lose H3K27ac at a given stage transition N to N+1. Dynamic enhancers that lose H3K27ac at a given stage transition tend to lose accessibility as measured by ATAC-seq either at or after that stage transition in question (sometimes proceeding H3K27ac loss by as much as 5 stage transitions).

in each developmental context as well as their specific, temporal activity profiles. Finally, we investigated the functional changes that occur in association with changes in chromatin accessibility, revealing a previously undescribed order in which regulatory sequences become functional enhancers.

Several key highlights of this chapter are listed below:

- Our developmental time series enables an unprecedented ability to analyze the developmental program of mouse embryogenesis.
- Different tissues feature unique, tissue-specific patterns of chromatin dynamics.
- These patterns, or regulatory programs, can be analyzed to better understand the biological changes driving or resulting from tissue differentiation.
- Analysis of chromatin state dynamics reveals multiple roles for cis-regulatory elements depending on cellular context.
- These techniques can be also used to study the regulatory differences between healthy and disease states and identify putative targets for therapy.

3.5 Methods

3.5.1 Read count arrays

To score the peaks by their read count intensity, the uniform d-TAC catalog was intersected against the read alignment (.bam) files for each sample and each each replicate using the 'bedtools coverage' command. These counts were then tabulated and added together for a pooled read count per peak in each sample.

3.5.2 Identification of differentially accessible d-TACs

For each d-TAC in the uniform catalog, the number of ATAC-seq reads overlapping the d-TAC was counted for each tissue-stage and replicate using the coverage function in bedtools

v2.27.1. For each tissue, d-TACs at any stage were classified as temporally dynamic if they were significantly differentially accessible (fold change >= 2, p-value <= 0.05) in sequential stages of development, using DEseq2.

3.5.3 GO ontology term enrichment

Functional enrichments through GREAT[61] were obtained through the greatBatchQuery.py script. The resulting lists were first filtered for the relevant ontologies. After that, only the terms showing a binomial FDR <= 0.05 and a regional enrichment equal or higher than 2-fold were considered.

3.5.4 chromHMM state enrichment

The d-TAC catalog was overlapped with autosomal chromHMM state calls for each tissue-stage (pooled or replicate call set, as indicated) using bedtools v2.20.1. Enrichment for a given state s in a particular tissue-stage was calculated as the observed number of base pairs of the d-TAC catalog that overlap state s, divided by the total number of base pairs expected to overlap state s based on its genome coverage (total bp coverage of d-TAC catalog * fraction of genome covered by state s).

3.5.5 Dynamic chromatin state enrichment

To investigate the relationship between changes in accessibility and changes in chromatin state, the dynamic d-TACs were classified as either gaining (positive log fold change) or losing (negative log fold change) accessibility. For each tissue-stage-transition (n to n+1), these sets of gain or loss of accessibility d-TACs were overlapped with chromHMM state calls for stages n and n+1 using intersectBed from bedtools v.2.27.1. Enrichment was calculated by taking the observed fraction of dynamic base pairs overlapping each combination of states (state at n, state at n+1) and dividing by the expected fraction of base pairs overlapping each state combination based on the dynamic and non-dynamic d-TACs.

3.5.6 Motif enrichment

For each tissue and dynamic stage transition, associated d-TAC regions were converted into fasta sequence format using the getfasta command in bedtools and shuffled using the fasta-shuffle-letters tool in the MEME suite for a background with the parameters '-kmer 2 -tag -dinuc -seed 1'. Then, each set of fasta sequences were enriched for motifs in the HOCOMOCO v11[62] core mouse database using AME in the MEME suite[63].

3.5.7 Coordination of H3K27ac and ATAC-seq in dynamic regions

To investigate the temporal relationship between H3K27ac and chromatin accessibility (e.g. the enhancer "life cycle"), dynamic strong-enhancers (replicated, chromHMM state U5) at each stage-transition were overlapped against d-TACs (using bedtools) for the respective tissue to identify matching enhancers and d-TACs. In cases where more than one d-TAC overlapped an enhancer, the d-TAC with the largest number of overlapping base pairs was selected. The sequential logFC in ATAC-seq signal was evaluated at every possible stage-transition for these matching d-TACs and a mean was taken. These stage-transitions were converted to offsets relative to the strong enhancers and the fold-changes averaged for the purpose of deriving a global trend (i.e. for dynamic enhancers at e11.5 ->e12.5; e11.5 ->12.5 is an offset of 0, e12.5 ->e13.5 is an offset of 1 and e16.5 ->p0 is an offset of 5). The inverse analysis was also performed to assess the log FC in H3K27ac at dynamic d-TACs.

3.6 Acknowledgments

We thank DU Gorkin for this guidance and assistance with study design and direction, data analysis, and manuscript and figure preparation. We thank Y Zhang and I Barozzi for their technical assistance with analysis and manuscript preparation. We thank AH Lee and H Huang for performing experiments. This study was funded by the National Human Genome Research Institute as part of the Encyclopedia of DNA Elements (ENCODE) project (U54HG006997),

and was performed in compliance with all relevant ethical regulations. D.U.G. supported by the NIH Institutional Research and Academic Career Development Awards (IRACDA) program, and an A.P. Giannini Foundation fellowship. D.E.D, A.V., and L.A.P. were also supported by UM1HG009421, and research conducted at the E.O. Lawrence Berkeley National Laboratory was performed under Department of Energy Contract DE-AC02-05CH11231, University of California. I.B. is funded through an Imperial College Research Fellowship. Y.H. is supported by the H.A. and Mary K. Chapman Charitable Trust. J.R.E. is an Investigator of the Howard Hughes Medical Institute.

Chapter 3, in part, has been submitted for publication. DU Gorkin*, I Barozzi*, Y Zhao*, Y Zhang*, H Huang*, AY Lee, B Liu, J Chiou, A Wildberg, B Ding, B Zhang, M Wang, JS Strattan, JM Davidson, Y Qiu, V Afzal, JA Akiyama, I Plajzer-Frick, CS Novak, M Kato, TH Garvin, QT Pham, AN Harrington, BJ Mannion, EA Lee, Y Fukuda-Yuzawa, Y He, S Preiss, S Chee, JY Han, BA Williams, D Trout, H Amrhein, H Yang, JM Cherry, W Wang, K Gaulton, JR Ecker, Y Shen, DE Dickel, A Visel, LA Pennacchio & B Ren. An atlas of dynamic chromatin landscapes in the developing mouse fetus. (* Authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

3.7 Appendix

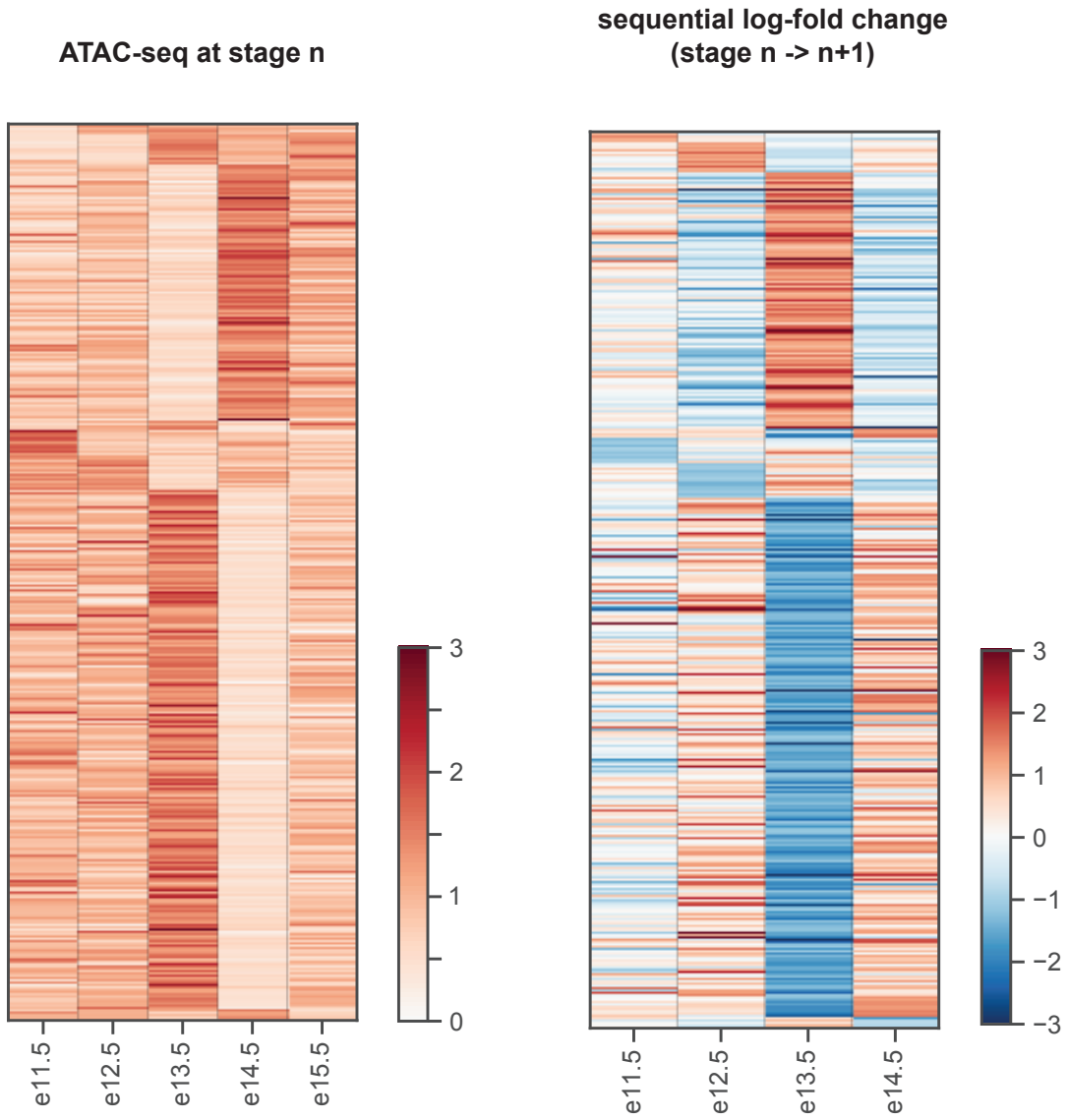

Figure 3.8. ATAC-seq and sequential logFC heatmaps for embryonic facial prominence. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in CF (e11.5 to e11.5). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

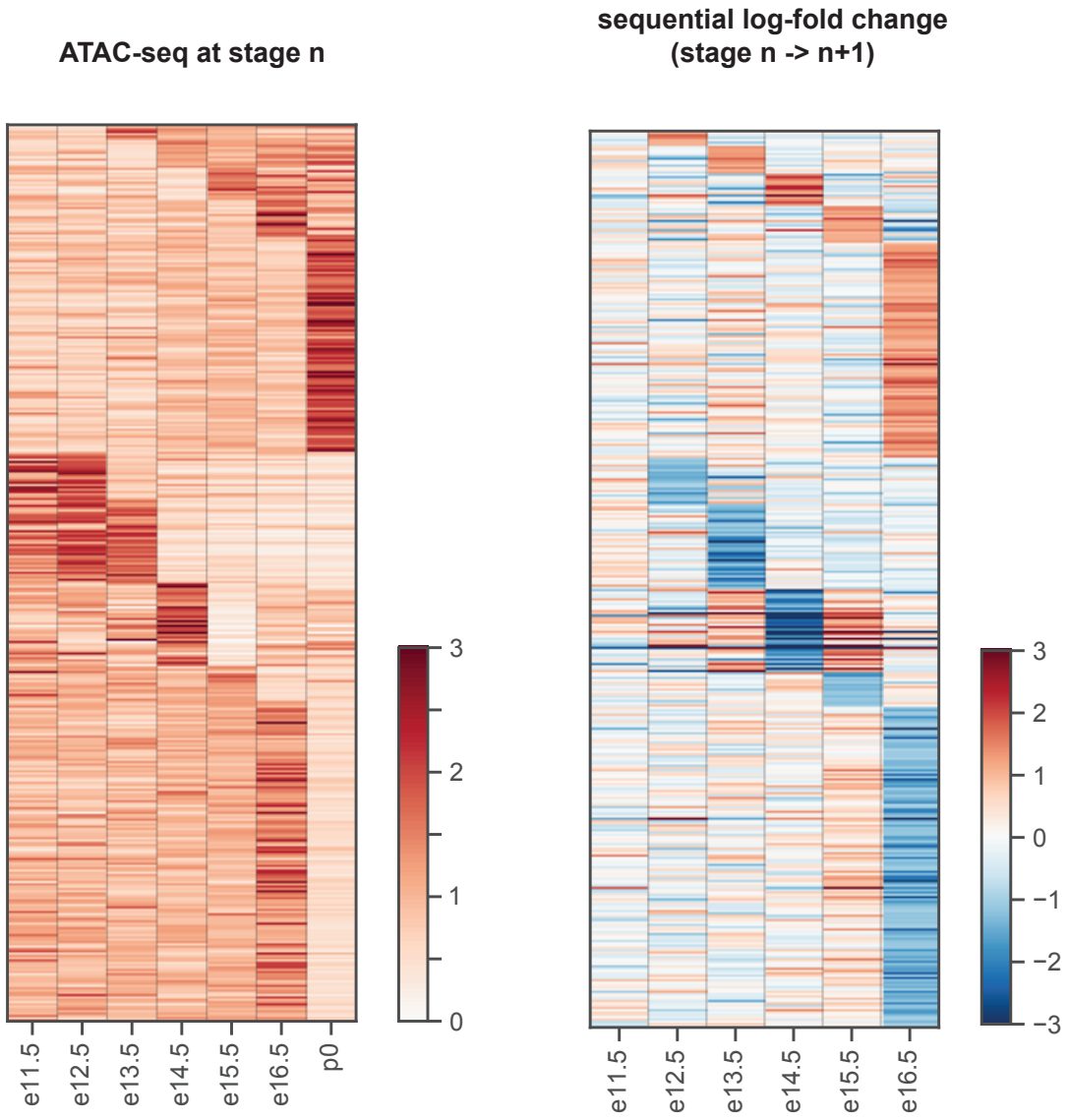

Figure 3.9. ATAC-seq and sequential logFC heatmaps for heart. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in heart (e11.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

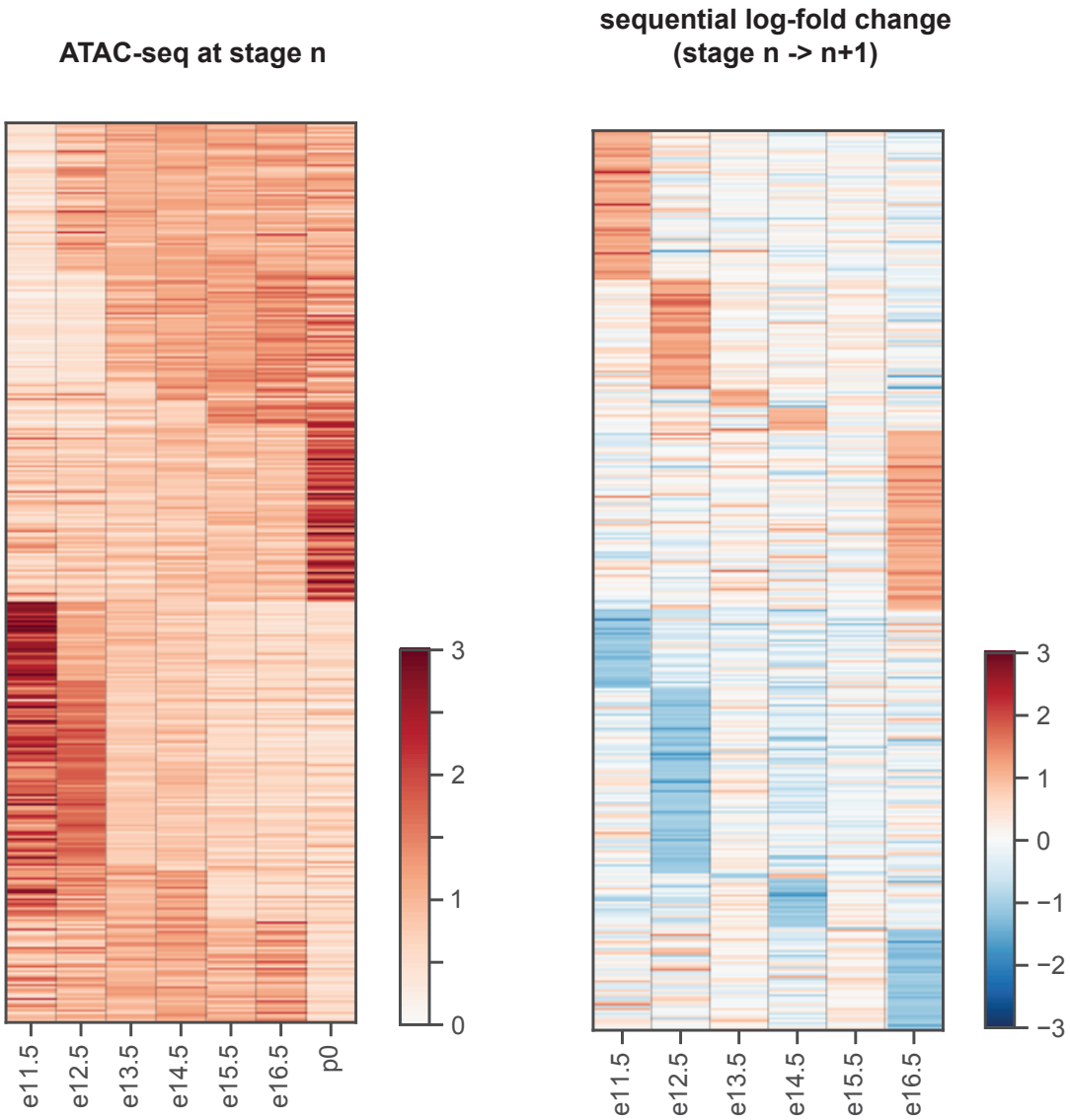

Figure 3.10. ATAC-seq and sequential logFC heatmaps for hindbrain. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in hindbrain (e11.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

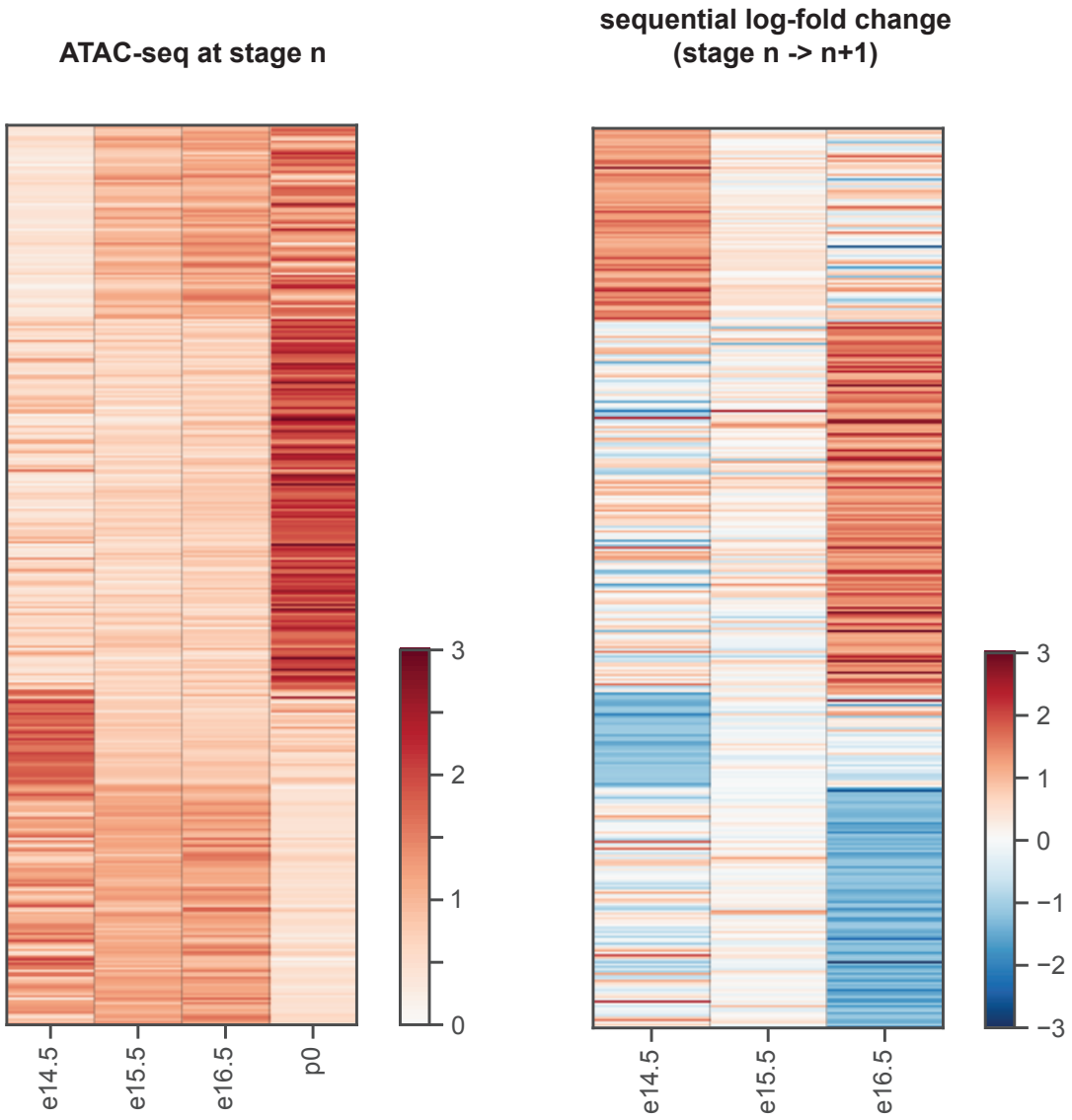

Figure 3.11. ATAC-seq and sequential logFC heatmaps for intestine. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in intestine (e14.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

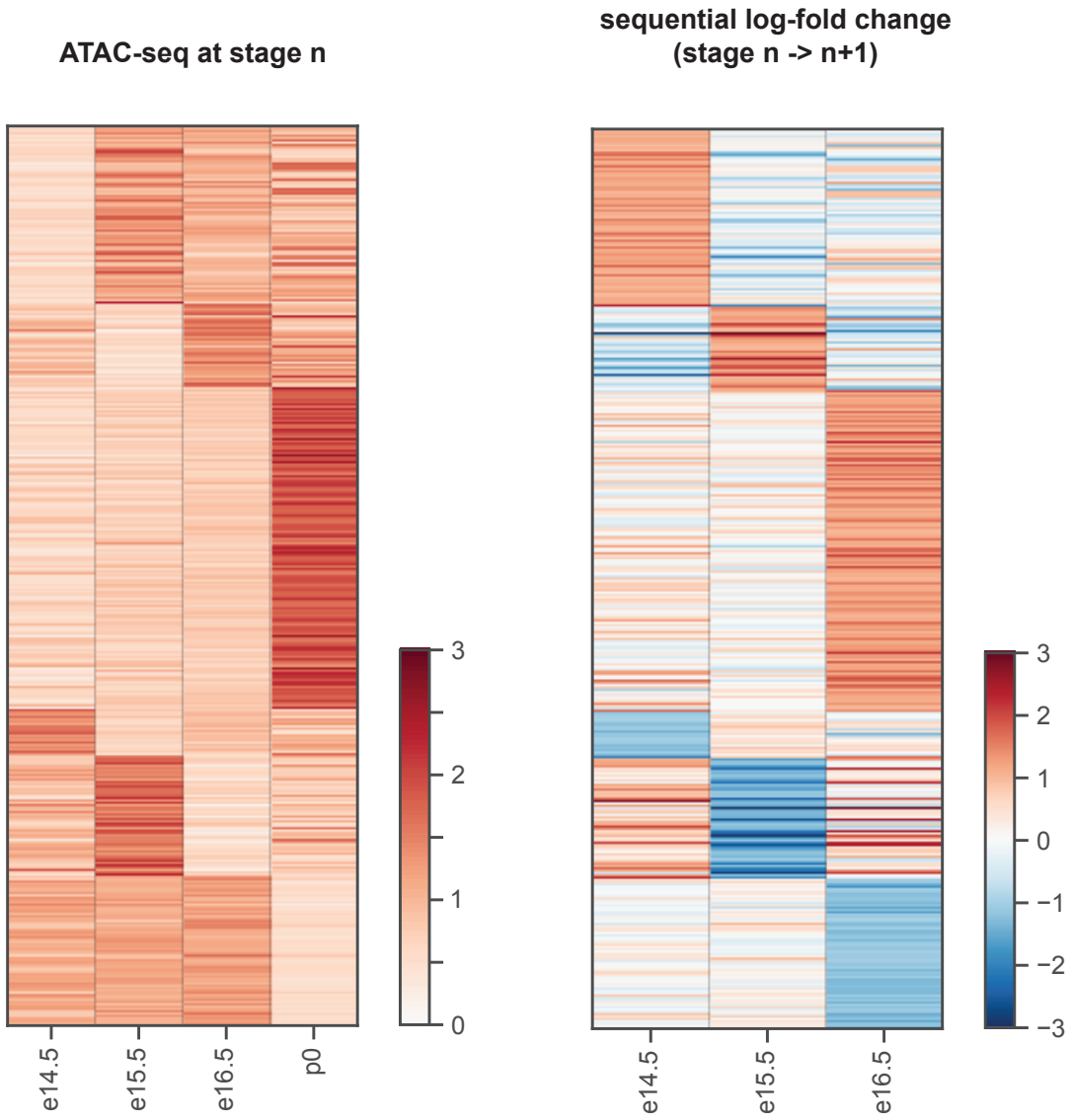

Figure 3.12. ATAC-seq and sequential logFC heatmaps for kidney. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in kidney (e14.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

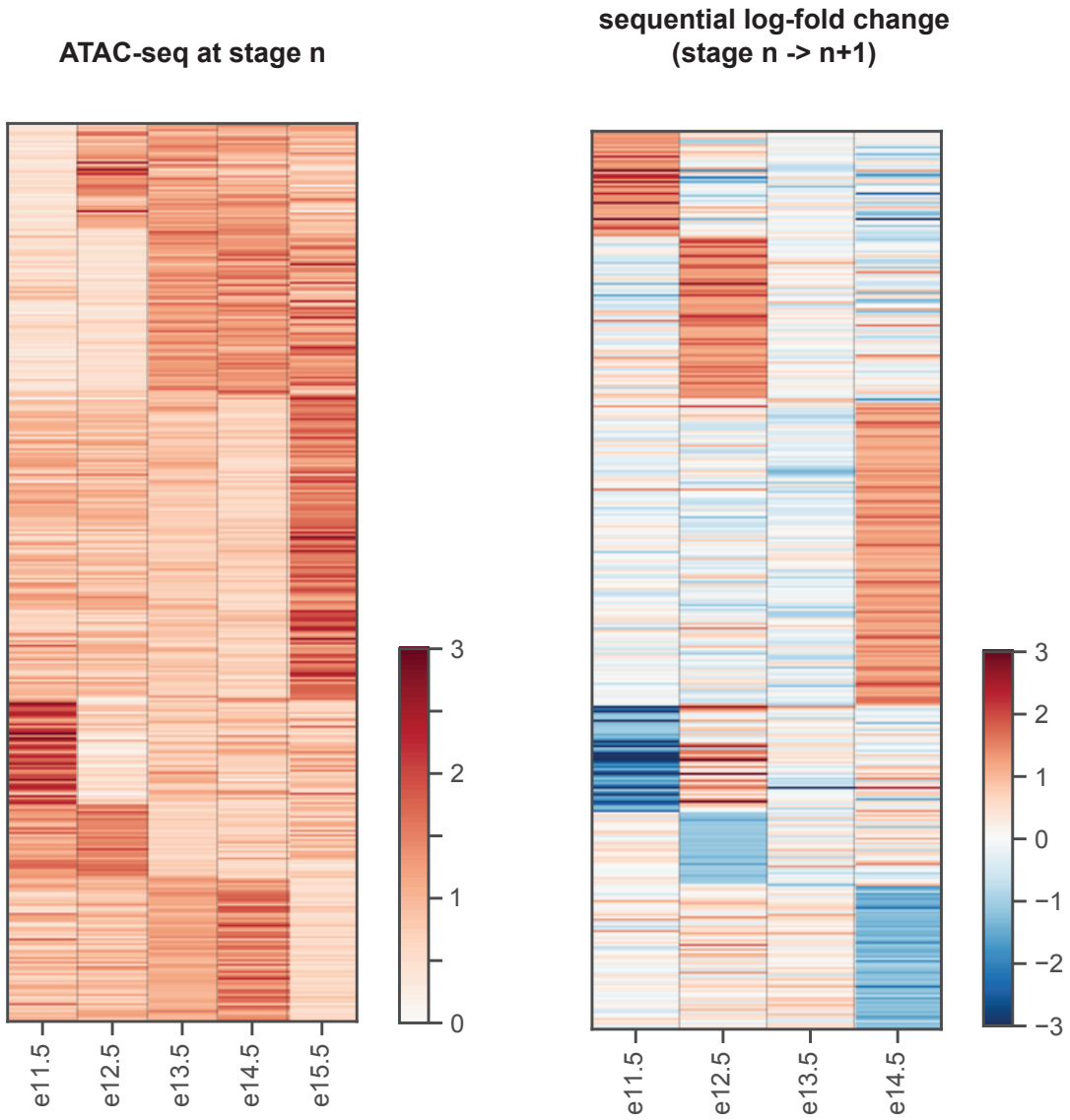

Figure 3.13. ATAC-seq and sequential logFC heatmaps for limb. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in limb (e11.5 to e15.5). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

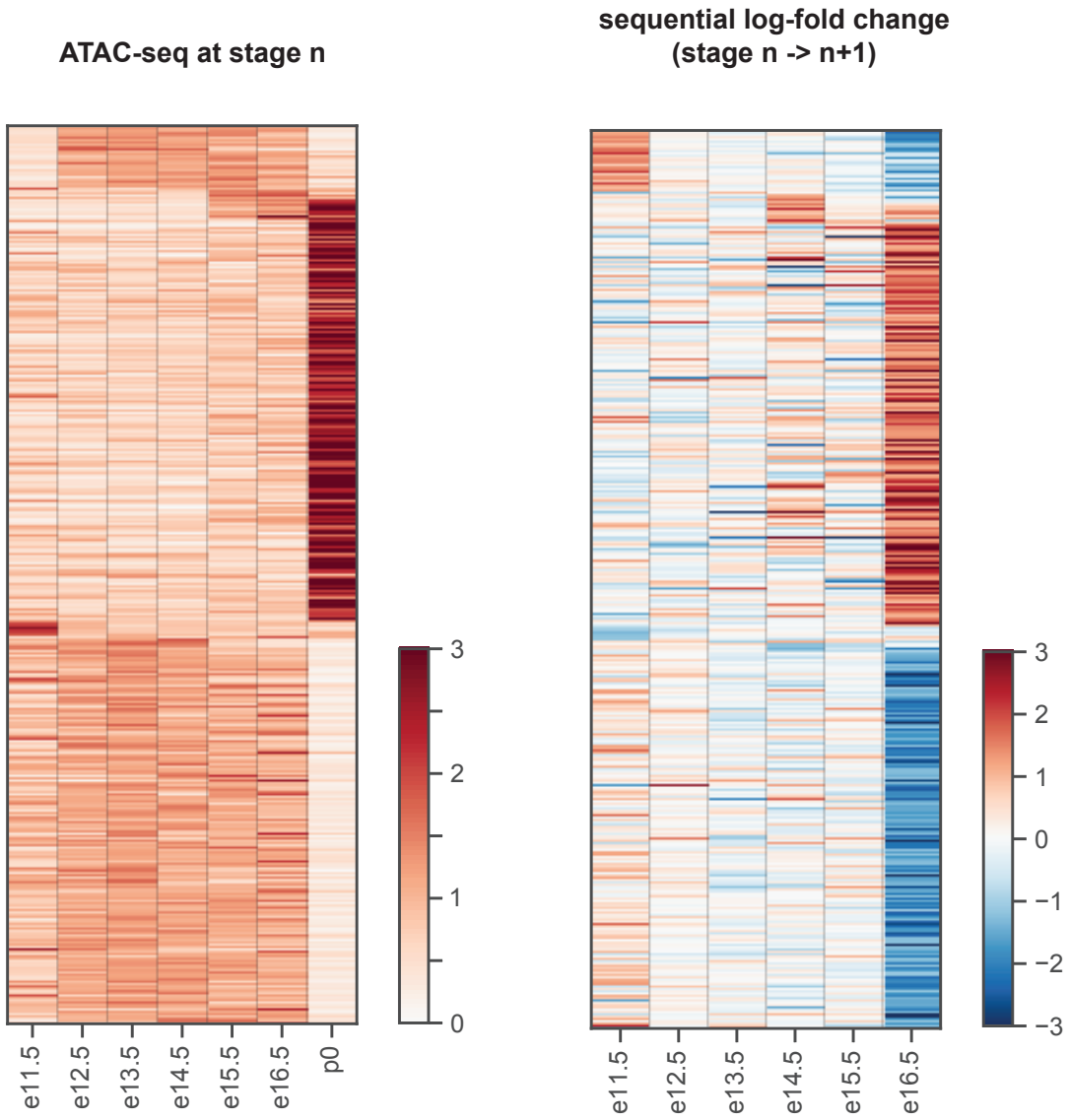

Figure 3.14. ATAC-seq and sequential logFC heatmaps for liver. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in liver (e11.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

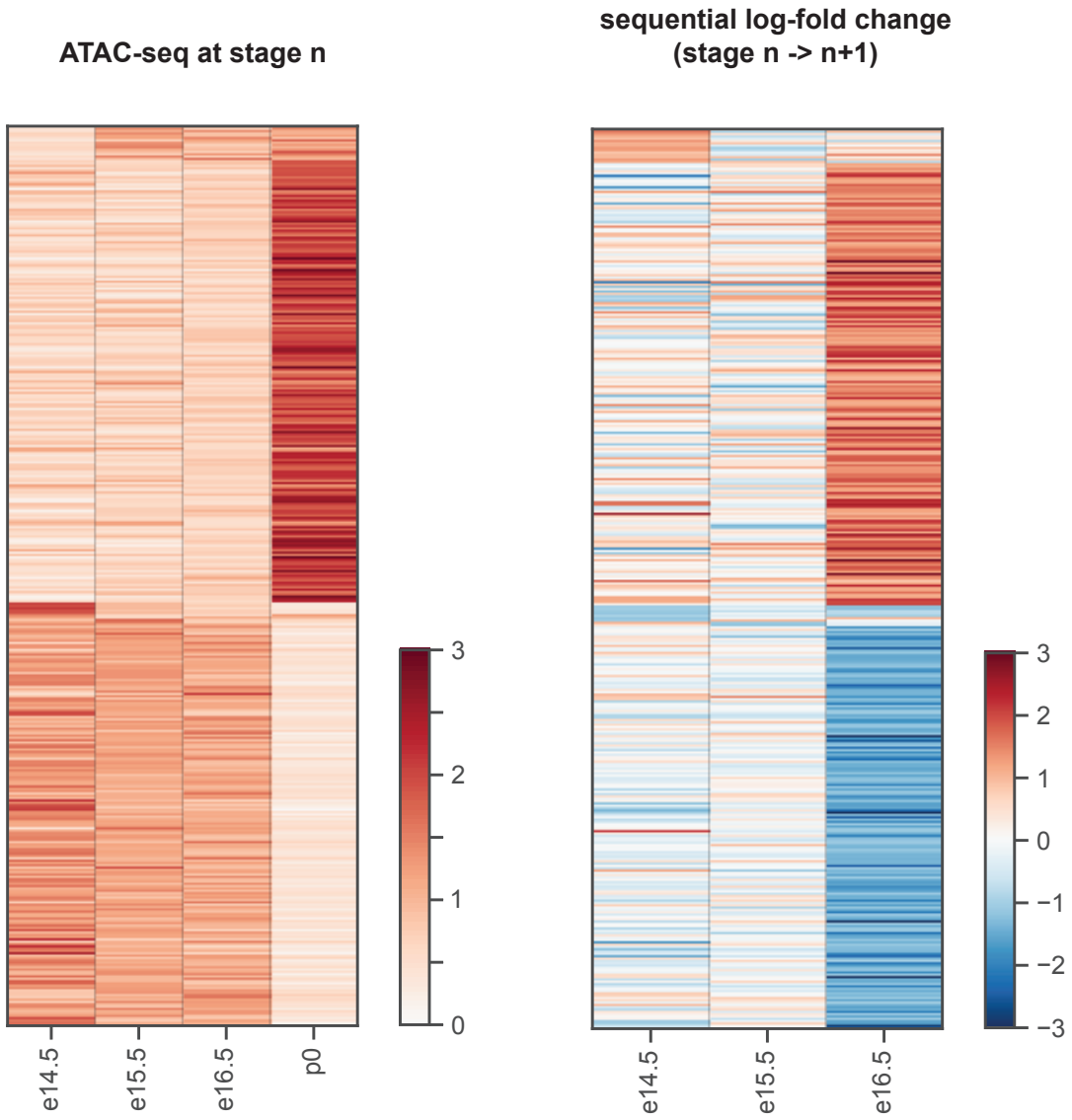

Figure 3.15. ATAC-seq and sequential logFC heatmaps for lung. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in lung (e14.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

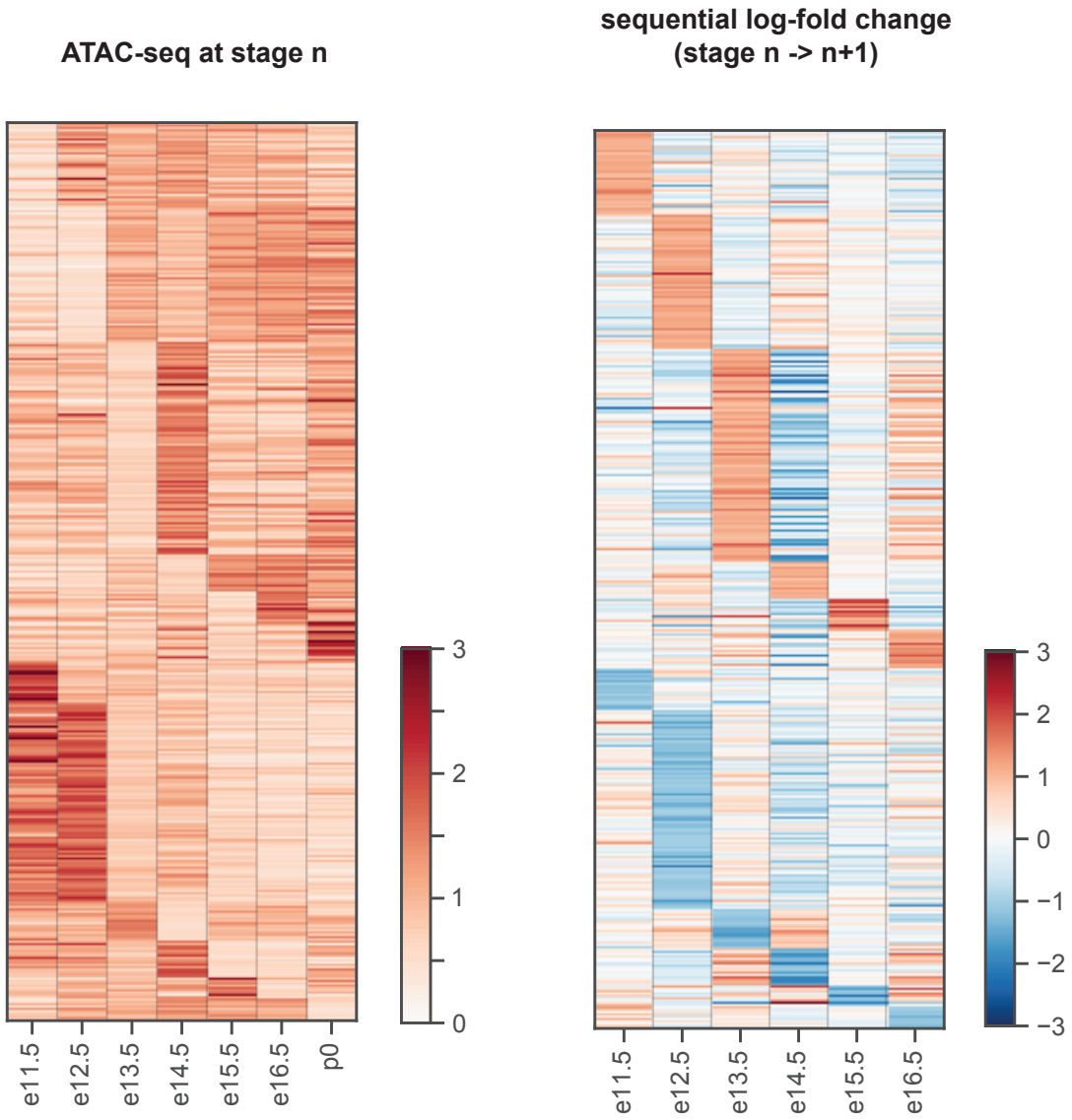

Figure 3.16. ATAC-seq and sequential logFC heatmaps for midbrain. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in midbrain (e11.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

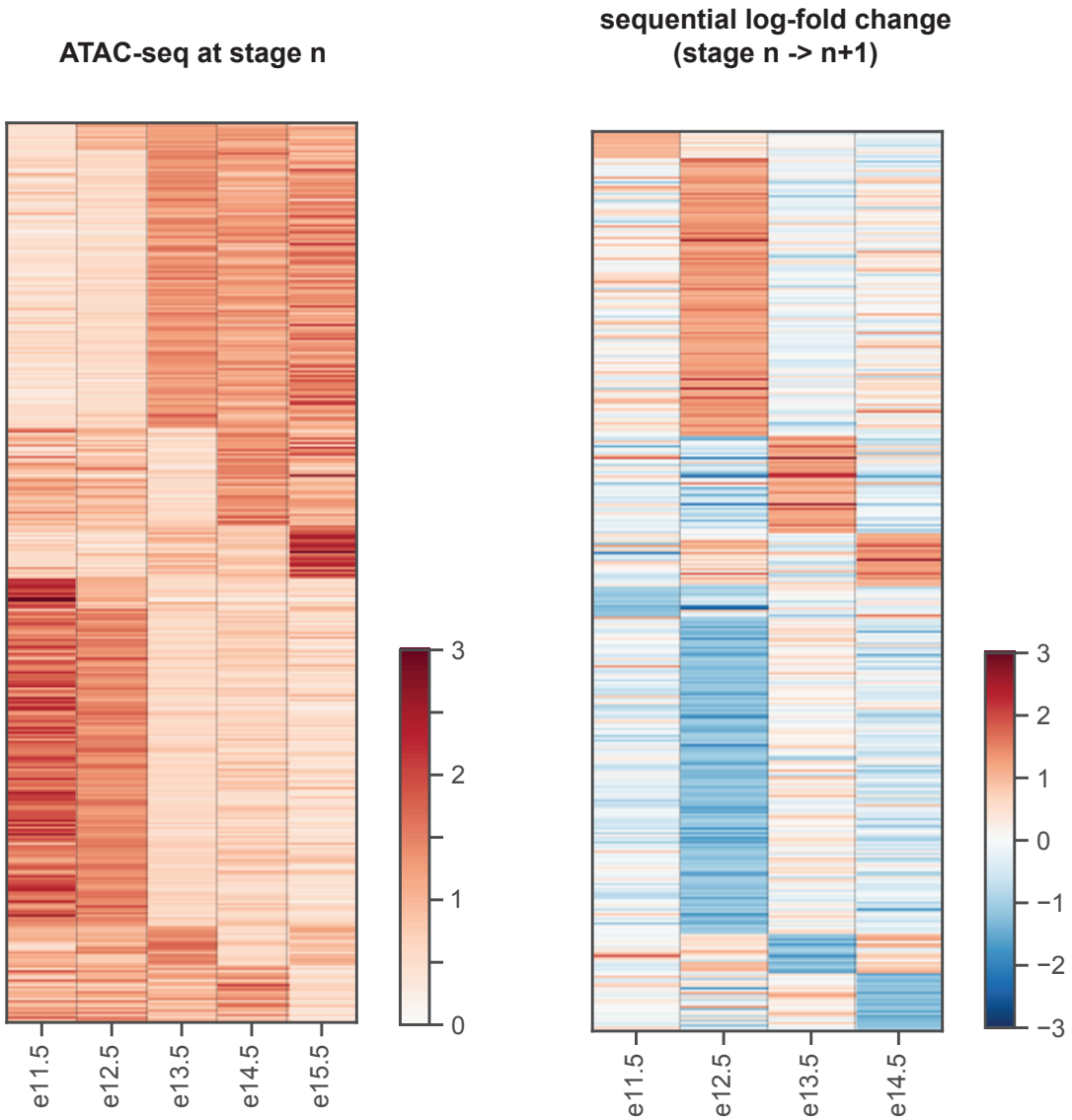

Figure 3.17. ATAC-seq and sequential logFC heatmaps for neural tube. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in NT (e11.5 to e15.5). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

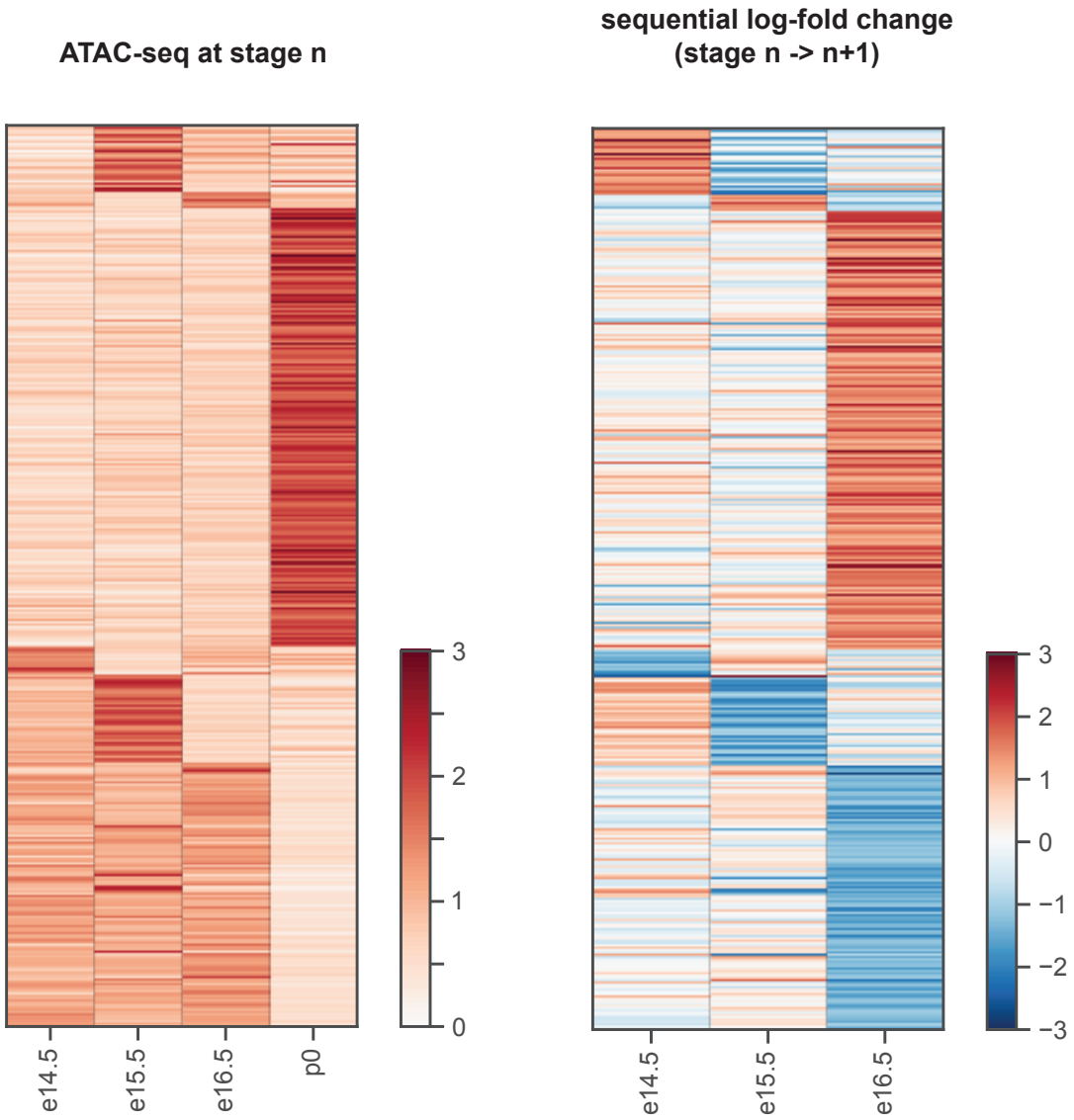

Figure 3.18. ATAC-seq and sequential logFC heatmaps for stomach. Heatmap (left) shows the normalized accessibility for each dynamic dTAC at every stage profiled in stomach (e14.5 to p0). Heatmap (right) shows the sequential log-fold-change in accessibility between each stage transition.

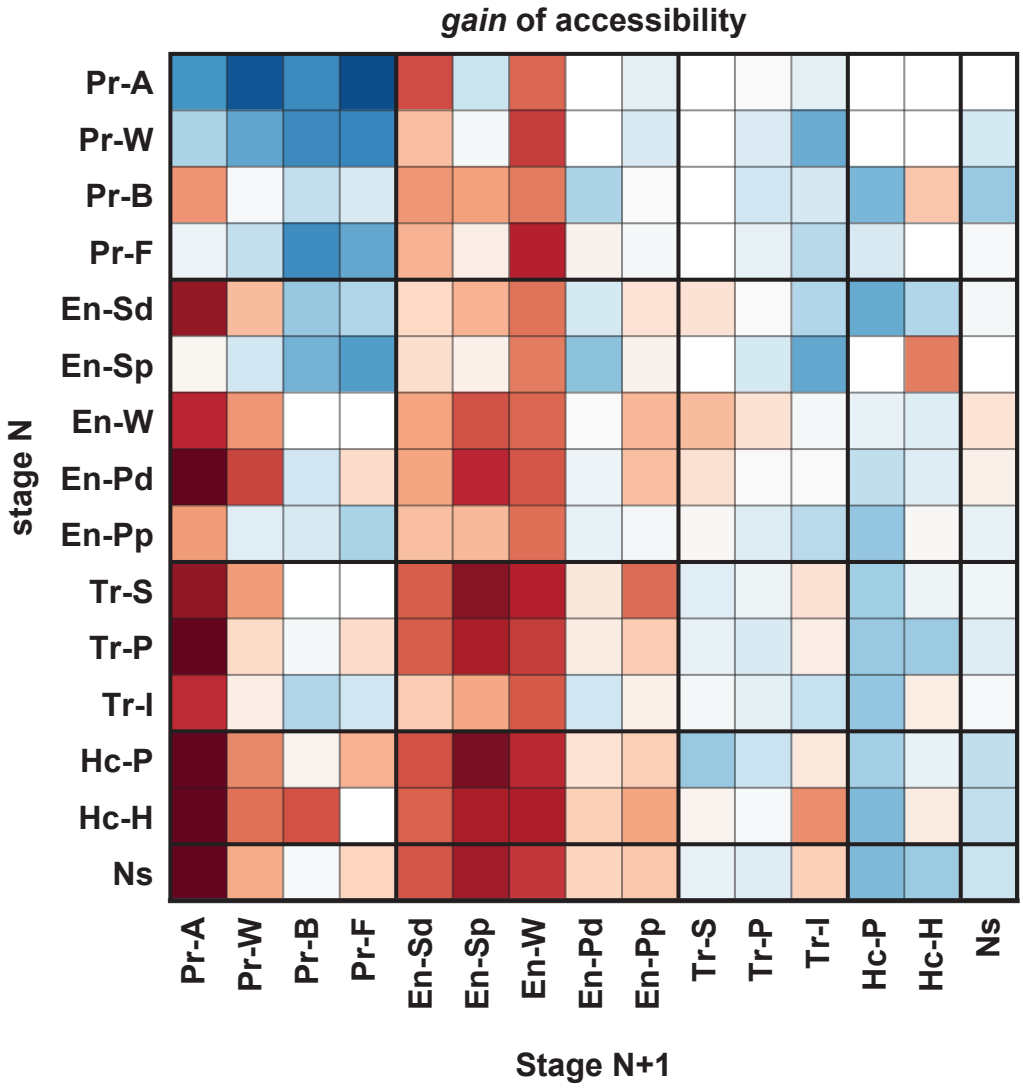

Figure 3.19. Similar schema to Figure 3.5 but showing each chromatin state separately instead of in supersets. The heatmap shows the chromatin state changes that occur at dynamic d-TACs which gain accessibility at a given stage transition. Enrichment is relative to the coverage of each state in total d-TAC catalog.

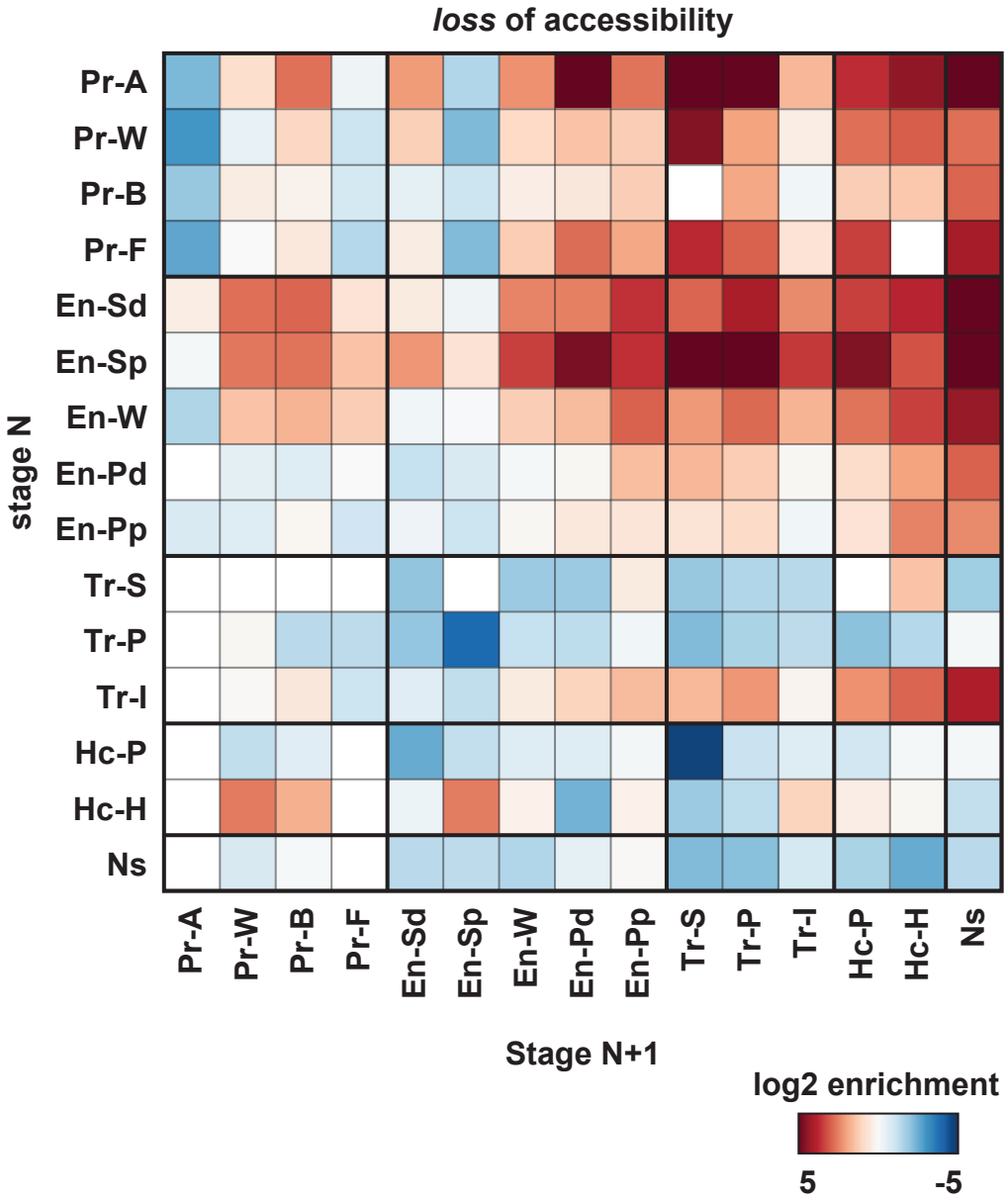

Figure 3.20. Same as Figure 3.19 but for d-TACs that lose accessibility at a given stage transition.

Chapter 4

Understanding genomic vocabulary and grammar with deep neural networks

4.1 Abstract

Mammalian gene regulatory networks are controlled by transcription factor binding to accessible regulatory sequences in the genome, such as enhancers[78]. Though there are many effective strategies for extracting biology from compendiums of these regulatory elements, the most fundamental and biological interpretable approach would be to study the regulatory logic encoded in the sequences themselves to which transcription factors bind. In this chapter, we propose a novel paradigm for thinking about the functional "unit" of regulation as a sequence rather than a peak and describe a neural network-based machine learning strategy for predicting accessibility of putative sequences in a cell-type-specific manner. We leverage these models to identify similarities between samples based on the composition of their accessible regulatory sequences, which can be utilized in a variety of clinical applications (ie. identifying model systems or comparing samples after external stimuli perturbations).

4.2 Introduction

Gene regulation is controlled by the coordinated binding and activity of hundreds of transcription factors, each with their individual specific binding motifs[79]. These motifs are

Table 4.1. Toy example of regulatory vocabulary scores for several samples. For each sequence (or combination of sequences), a score of some kind (i.e. probability, importance score, etc.) is calculated to reflect its importance or predictiveness in describing the biology of each sample.

k-mers	sample-1	sample-2	...	sample-66
ACTGTG	4.12	1.48	...	-0.1
ATCTTA	7.31	0.00	...	2.51
...	...	...	...	...
TCGAAA	2.78	-0.81	...	3.15
GGAATA	0.00	0.00	...	-0.61

heavily enriched in *cis*-regulatory elements, such as enhancers and promoters, as binding at these key loci is a crucial aspect of a cell's regulatory processes[80], including initiation of transcription, recruitment of additional regulatory machinery, coordination of lineage-specifying TFs by pioneering factors[81], and even remodeling of the 3D conformation of the genome via chromatin looping[82]. Since these vital regulatory sequences must be accessible for TFs to bind, the availability of a map of putative cREs in our d-TAC catalog represents a unique opportunity to investigate the cell-type-specific regulatory networks driving gene expression.

However, genomics analysis tends to revolve around comparison or analysis of peaks. While analysis peaks can provide useful insights, the idea of a peak as a "unit" of genomics is less than ideal. Peak sets are problematic because the are fundamentally arbitrary and non-standard, derived empirically from each sample and subject to intrinsic variability. This includes biases due to data quality as well as algorithmic factors, such as data processing methods and peak calling parameters. Peaks are not even necessarily standardized in length, requiring further analysis to create a uniform set. Most limiting, any analysis involving comparison of peaks requires that samples of interest share a uniform genomic coordinate system, complicating cross-species comparisons (without the use of additional tools such as liftover). Furthermore, the biological meaning of a peak is hard to interpret; it is basically a locus in the genome, but what one peak represents and even how important a peak is is unknown.

We propose the use of a *genomic vocabulary*[83][84] as a superior alternative to peaks

for the purposes of evaluating sample similarity. As mentioned above, coordinated transcription factor binding is central to define a cell's phenotype, so being able to describe a cell based on the availability of binding motifs in its accessible sequences is a biologically interpretable and innovative approach. By distilling a sample's regulatory network into a weighted combination of sequences, we can describe the importance of each sequence to a given cellular context. Furthermore, by characterizing the importance of various co-occurring patterns of sequences, we can ascertain the rules of gene regulation, or the *genomic grammar*[85]. Just as linguistic grammar defines the proper arrangement of words to create a sentence with semantic meaning, the specific order, combinations, orientations, and intervening distances of key regulator sequences reflects the molecular interactions of TF binding.

Under this framework, it becomes trivial to compare samples based on their distinct biological patterns. The applications of this capability are significant, including the ability to identify ideal animal models for disease research, categorize undefined clinical samples (including single cell data), and characterize drug responses under different contexts (e.g. identifying which cancer cell lines respond to treatment).

4.3 Results

4.3.1 Sequence-based regulatory vocabulary for analyzing accessible chromatin regions

To fulfill our objective of developing a biologically meaningful, sequence-based *unit* of gene regulation, we leveraged the extant knowledge base of binding motifs from the mammalian JASPAR database[60]. These motifs were manually curated and merged to combine similar or degenerative, non-specific motifs, resulting in a set of 390 curated motif PWMs [86]. However, despite covering a large range of mammalian organisms, motif databases including JASPAR are incomplete and numerous binding motifs remain uncharacterized. Furthermore, there may be sequences with predictive regulatory function that are not cleanly represented by a single motif,

so to capture these potential sources of unknown biology, we also allowed our models to train *de novo* sequence filters (e.g. pseudo-motifs) as described below.

4.3.2 Neural network model identifies sample-specific sequence features

There are numerous machine learning techniques that have been successfully used to explore the sequence-composition of regulatory elements, including random forests[87], support vector machines (SVMs)[83], and neural networks[88]. From computer vision (i.e. Google Image Search, self-driving cars) to natural language processing (i.e. Siri, Alexa) to even video game e-sports (OpenAI Five DotA[90]), (of 2019) convolutional neural networks in particular have captured the public imagination, and emerged as the future of artificial intelligence and machine learning as, finding its way into virtually every area of research (and even aspects of everyday life). Acknowledging the similarity between the encoding of genetic sequence information (as an 4x*n* array) and digital images (essentially a matrix of pixel color values), we leveraged advances in deep learning and applied these techniques to interface neural networks with our biological objectives. As we will describe in detail below, we found convolutional neural networks to be ideally suited to addressing regulatory vocabulary and grammar (Figure 4.1), though other high-performing grammar methods exist such as attention networks[91][86] and recurrent neural networks with word embeddings[92].

We developed our neural network models to differentiate between open chromatin regions from GC-matched, dinucleotide shuffled genomic background, in a given genomic context or condition. For each sample, a model was trained independently to predict on that sample's specific sequence features, allowing for superior training speed and predictive performance compared to a multi-class singular model. In addition, using this modular approach for each additional sample allows for rapid scaling of our models without the need to re-train previous samples each time a new sample is to be considered.

Critically, the distillation of the biological information contained in peak calls into a systematic model is an important innovation that will reduce the impact of variable data quality.

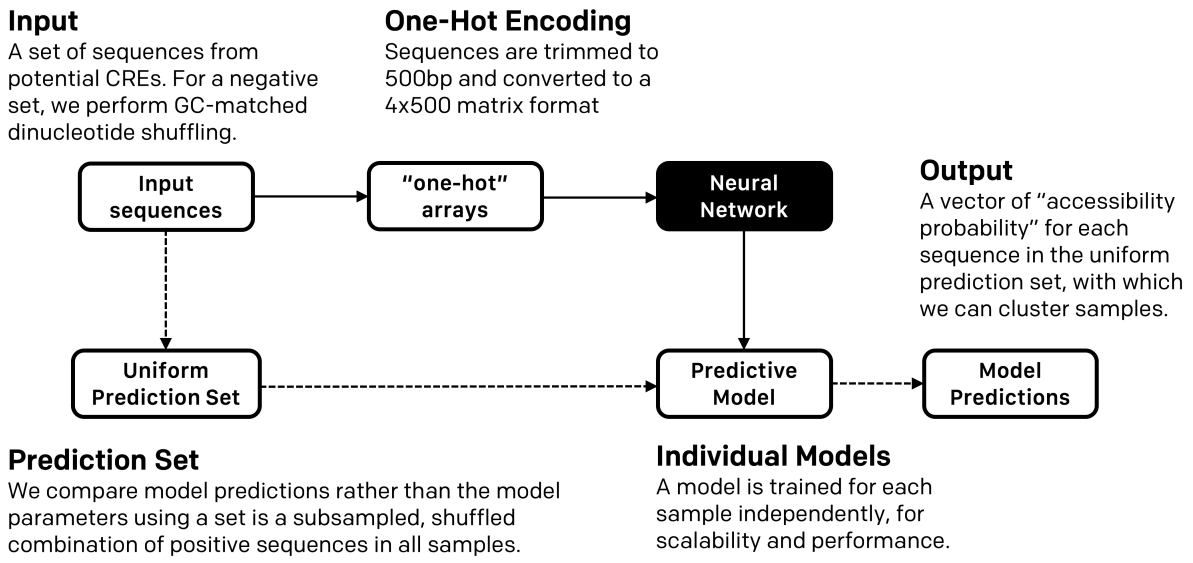

Figure 4.1. Overview of the neural network comparison method. Input sequences are converted into one-hot encoded array format and fed into the neural network for training. A uniform set of prediction sequences are then used to generate model predictions for each model, the output of which is used for model comparisons. See Figure 4.1 for an overview of the neural network architecture.

Given the same parameters and pipeline, generally speaking, having more and better quality reads results in a larger number of peak calls that tend to be less noisy. As our neural network models are still trained from the peak centers, the quality of the source dataset is a factor; however, by reducing the output for each sample to a uniform model, each of which share the exact same architecture and number of parameters, it becomes more straightforward to compare the regulatory information extracted by these models.

Our architecture revolves around two convolutional layers, which provide a perfect analogy for (1) the enrichment of the genomic vocabulary sequences and (2) genomic grammar patterns respectively. The first layer searches for and scores the motif PWMs across a sliding window in each input sequence, and the second layer looks for combinations of these motif scores, indicating the positions, orientations, and combinations of motifs (e.g. genomic grammar) that are predictive for a sample's regulatory landscape in training the model (Figure 4.2).

To build an initial reference set of models as a proof-of-concept, we trained models on all

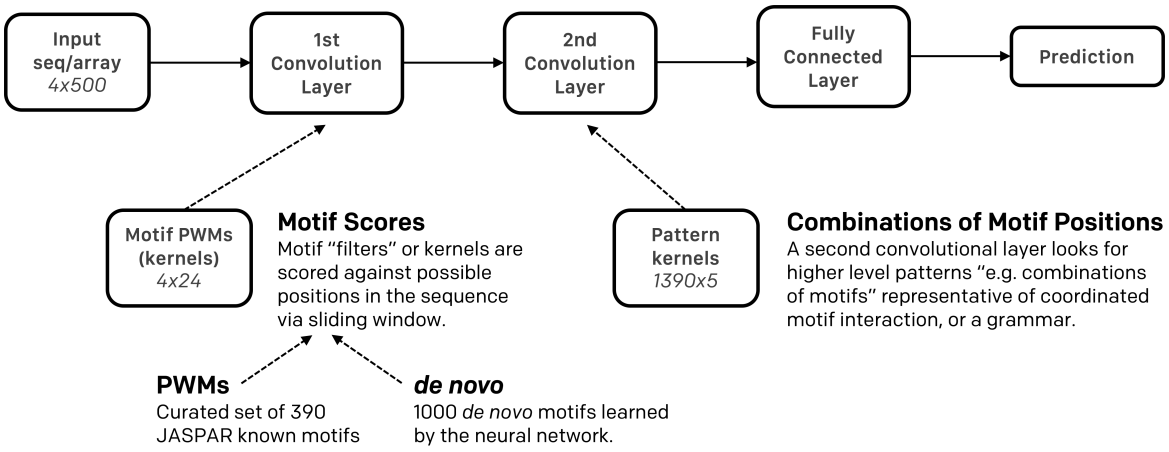

Figure 4.2. Overview of neural network architecture. The architecture features two convolution filters that generate models based on the predictive potential of motif sequence scores and combinations of motif scores respectively. The motif kernel filters are a combination of curated PWMs from the JASPAR database[86] and *de novo* motifs learned by the neural network.

of the mouse embryonic ATAC-seq samples generated in the previous study, as well as various human DNase-seq samples from corresponding tissues from the ENCODE database (Table 4.2).

4.3.3 Model predictions and performance

Models were trained using ten-fold cross-validation and repeated ten times for each sample to verify performance metrics. In general, all of the models trained with our neural network architecture had strong performance, as measured by area under the Receiver Operating Characteristic curve (auROC) and Precision-Recall Curve (auPRC), ranging from 0.95-0.98) (Table 4.4). Shown below is a characteristic curve, specifically from the model trained on accessible sequences in e11.5 forebrain (Figure 4.3). The model performance depended on the sample of origin including the number of peaks available for training, with better performance correlated with a larger training set; however, given the generally strong and comparable performance metrics, we believe using these models to be a superior comparative tool to peak lists.

Table 4.2. Table containing the description and number of samples on which neural network models were trained. Samples are grouped by species, tissue classification, and temporal stage of development.

	Mouse							Human			
tissue vs stage	**e11.5**	**e12.5**	**e13.5**	**e14.5**	**e15.5**	**e16.5**	**p0**	**<=91d**	**108d**	**120/121d**	**>=15w**
Forebrain*	1	1	1	1	1	1	1	-	-	-	6
Midbrain	1	1	1	1	1	1	1	-	-	-	-
Hindbrain	1	1	1	1	1	1	1	-	-	-	-
Neural tube	1	1	1	1	1	-	-	-	-	-	-
Limb	1	1	1	1	1	-	-	-	-	-	-
Craniofacial	1	1	1	1	1	-	-	-	-	-	-
Heart	1	1	1	1	1	1	1	-	-	-	-
Liver	1	1	1	1	1	1	1	-	-	-	-
Intestine**	-	-	-	1	1	1	1	2	-	2	-
Kidney	-	-	-	1	1	1	1	1	-	1	-
Lung	-	-	-	1	1	1	1	1	1	1	-
Stomach	-	-	-	1	1	1	1	1	-	1	-

Notes:

1. *Human neocortex samples grouped with mouse forebrain.
2. **Human small, large intestine samples grouped with mouse intestine.

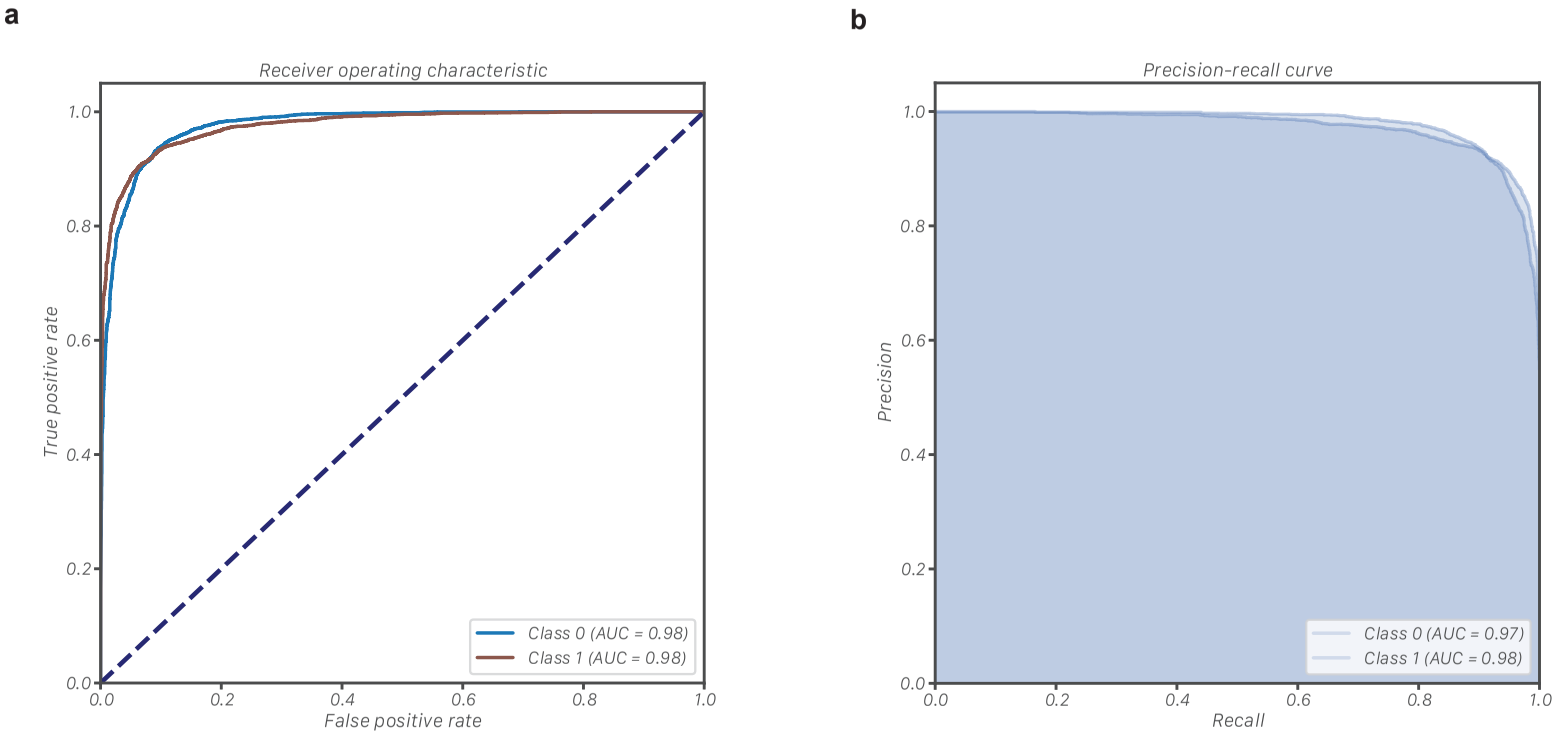

Figure 4.3. Representative example of model predictive performance for e11.5 forebrain. Class 0 and class 1 refer to the positively identified sequences and the shuffled, genomic background respectively. **a.** Area under the Receiver Operative Characteristic curve (auROC). **b.** Area under the Precision-Recall Curve (auPRC).

Table 4.3. Training accuracy and loss metrics for various numbers of *de novo* motif kernel filters at up to 15 training epochs, to optimize training parameters.

Training Accuracy

	epochs														
filters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
50	0.855	0.909	0.917	0.923	0.927	0.931	0.934	0.936	0.940	0.943	0.946	0.949	0.952	0.956	0.959
100	0.869	0.917	0.925	0.933	0.939	0.943	0.948	0.953	0.957	0.961	0.964	0.969	0.972	0.974	0.975
200	0.875	0.919	0.928	0.937	0.943	0.950	0.955	0.962	0.966	0.970	0.975	0.977	0.979	0.981	0.984
300	0.877	0.919	0.930	0.938	0.946	0.953	0.960	0.967	0.972	0.976	0.979	0.983	0.982	0.984	0.985
400	0.876	0.916	0.929	0.938	0.947	0.954	0.962	0.970	0.972	0.979	0.980	0.983	0.984	0.985	0.987
500	0.877	0.920	0.933	0.939	0.950	0.959	0.967	0.971	0.976	0.978	0.982	0.982	0.985	0.986	0.987
750	0.879	0.920	0.933	0.943	0.951	0.960	0.968	0.973	0.978	0.979	0.984	0.986	0.985	0.987	0.988
1000	0.876	0.921	0.933	0.944	0.953	0.962	0.971	0.976	0.980	0.983	0.986	0.985	0.989	0.986	0.990

Training Loss

	epochs														
filters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
50	0.323	0.228	0.207	0.192	0.181	0.173	0.166	0.160	0.150	0.143	0.136	0.128	0.120	0.110	0.105
100	0.298	0.209	0.186	0.168	0.154	0.147	0.132	0.122	0.111	0.100	0.092	0.083	0.072	0.068	0.065
200	0.289	0.201	0.179	0.161	0.145	0.128	0.115	0.099	0.088	0.079	0.066	0.059	0.055	0.051	0.045
300	0.284	0.202	0.177	0.156	0.137	0.119	0.103	0.086	0.072	0.061	0.056	0.046	0.045	0.043	0.041
400	0.289	0.207	0.179	0.159	0.136	0.117	0.097	0.078	0.072	0.055	0.053	0.046	0.041	0.038	0.034
500	0.285	0.201	0.170	0.152	0.127	0.105	0.086	0.075	0.062	0.055	0.046	0.047	0.039	0.035	0.036
750	0.282	0.199	0.170	0.147	0.124	0.102	0.083	0.072	0.056	0.054	0.042	0.039	0.038	0.033	0.033
1000	0.287	0.197	0.167	0.143	0.118	0.096	0.077	0.061	0.051	0.046	0.039	0.039	0.031	0.036	0.028

4.3.4 Uniform set of sequences for model prediction and comparison

Neural networks are non-deterministic in nature, meaning that for the same input training data, different training iterations can result in models with wildly different parameters. These diverse models can still produce comparable if not identical prediction results, but consequently, this complicates model comparisons as evaluating the differences in model parameters – even for the same architecture – is not informative of how similarly the models perform.

Therefore, if the goal is to compare two samples' similarity based on their respective neural network models, an alternative approach is to avoid comparing the models in favor of their predictions on a shared set of input sequences. Put another way, if the objective is to see how well two models represent the underlying biology of their original samples, their ability to discriminate between which sequences out of a shared set are representative of said sample can be informative as to their biological similarity. The output of this approach would be a vector of length n, where n is the number of sequences in the prediction set, and vector values representing the probability each sequence is accessible in a given sample.

To determine the composition of this uniform set of sequences is a non-trivial task and there are numerous opportunities to explore regulatory biology through this method. One potential set would be a series of sequences comprised of various permutations and orientations of motifs, the predictions upon which would be directly interpretable as the biological relevance of that particular genomic grammar to a sample. Alternatively, generative adversarial networks (GANs)[93] have recently been featured prominently in literature and mainstream media for their ability to generate "fake" images from trained models, such as the artificial generation of realistic-looking human faces. In this context, GANs could be used to generate "fake" regulatory elements, or sequences potentially representative of enhancers or other *cis*-regulatory elements but not actually derived from any real endogenous cREs. For simplicity, however, we simply took a superset of all positive input sequences used in training and shuffled them to generate a new, uniform test sequences that would capture the diversity and content of all the originating

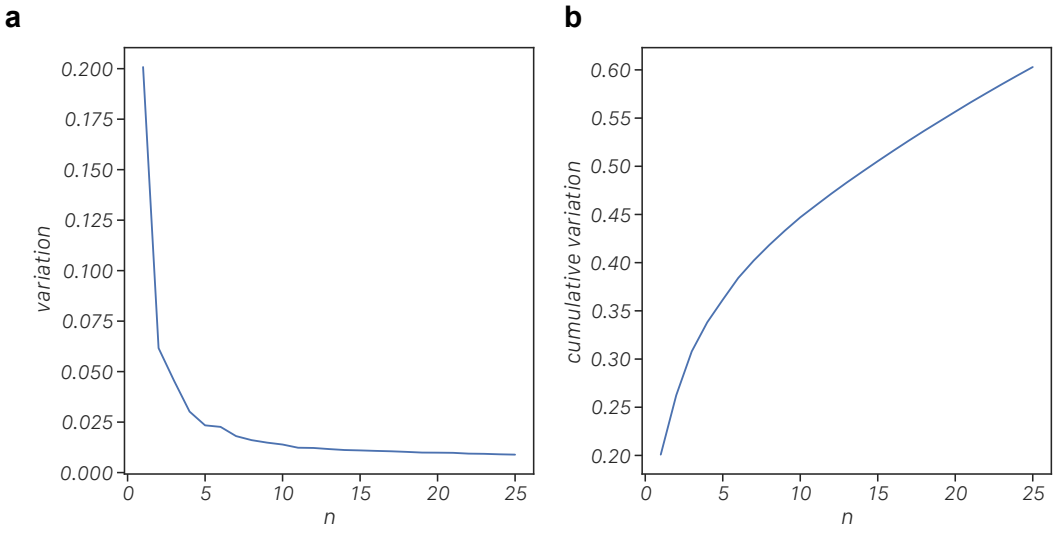

Figure 4.4. Curves exploring the level of variation explained by principal components. **a.** Variation explained by each principal component. **b.** Cumulative variation explained by the first 25 principal components, showing that the majority of the variation can be explained within the first 25.

samples.

4.3.5 Neural network framework enables genome-agnostic sample clustering and comparison

Given the large number of features in the uniform prediction set (100K input sequences), we performed dimensionality reduction in order to visualize the samples by projecting onto a lower dimensional space. To evaluate the validity of using dimensionality reduction with our specific data matrix, we used principal component analysis (PCA) to calculate the variation explained by the first m principal components and the contribution of each of the original sequences to each component. We found that despite the initially large number of sequences to be compressed, a majority of the overall cumulative variance could be explained within the first 25 components (Figure 4.4).

To visualize the relationship between samples, we first reduced the number of dimensions with PCA to 25 and then projected these principal components onto a 2D space with two

different commonly used visualization algorithms: t-SNE and UMAP. While t-SNE has been used extensively in the field, UMAP has a distinct advantage in that its projections retain meaningful distance information whereas in t-SNE, the distance between two points is not interpretable.

Visually, samples cluster by their tissue of origin primarily, and then where applicable, the stages of development (Figure 4.5). Zooming out slightly further, samples are found to be associating based on their tissue layer of origin (i.e. endoderm, mesoderm, or ectoderm). For instance, given the similarity between the four central nervous system tissues in the mouse ENCODE data series (forebrain, mindbrain, hindbrain, neural tube), it is unsurprisingly that these samples are closely grouped together and most distant to heart and liver.

Crucially, we find that based on these projections, our proposed neural network strategy is in fact able to meet our objective of classifying and grouping samples based on biological similarity, regardless of genomic structure. In these plots, mouse ATAC-seq samples are plotted as red dots while human DNase-seq samples are plotted as blue ones. We find that the six adult human neocortex samples (from three individuals) are clustered with forebrain samples, reflecting the analogous structures of the mouse forebrain and the cerebrum, of which the neocortex is a constituent component. (Notably, two human samples are farther away from the others, suggesting either additional biological differences inherent to the individual donors or possible batch effects in collection or processing.) The neocortex samples are particularly striking in that they come from adult human donors, yet they still positively associate with the mouse CNS tissues of embryonic origin. Likewise, the embryonic human lung samples most closely associate the mouse lung. Finally, we found that the human digestive tissues (intestine, stomach) were close in similarity to the corresponding mouse tissues.

These general trends hold true for both projection methods (Figure 4.7), but as mentioned previously, UMAP is the preferred algorithm as sample similarity can then be calculated as a Euclidean distance (Figure 4.6). In the case of using t-SNE as the projection algorithm, one's choice of perplexity parameter will have a large effect on the ultimate visualization. Additionally,

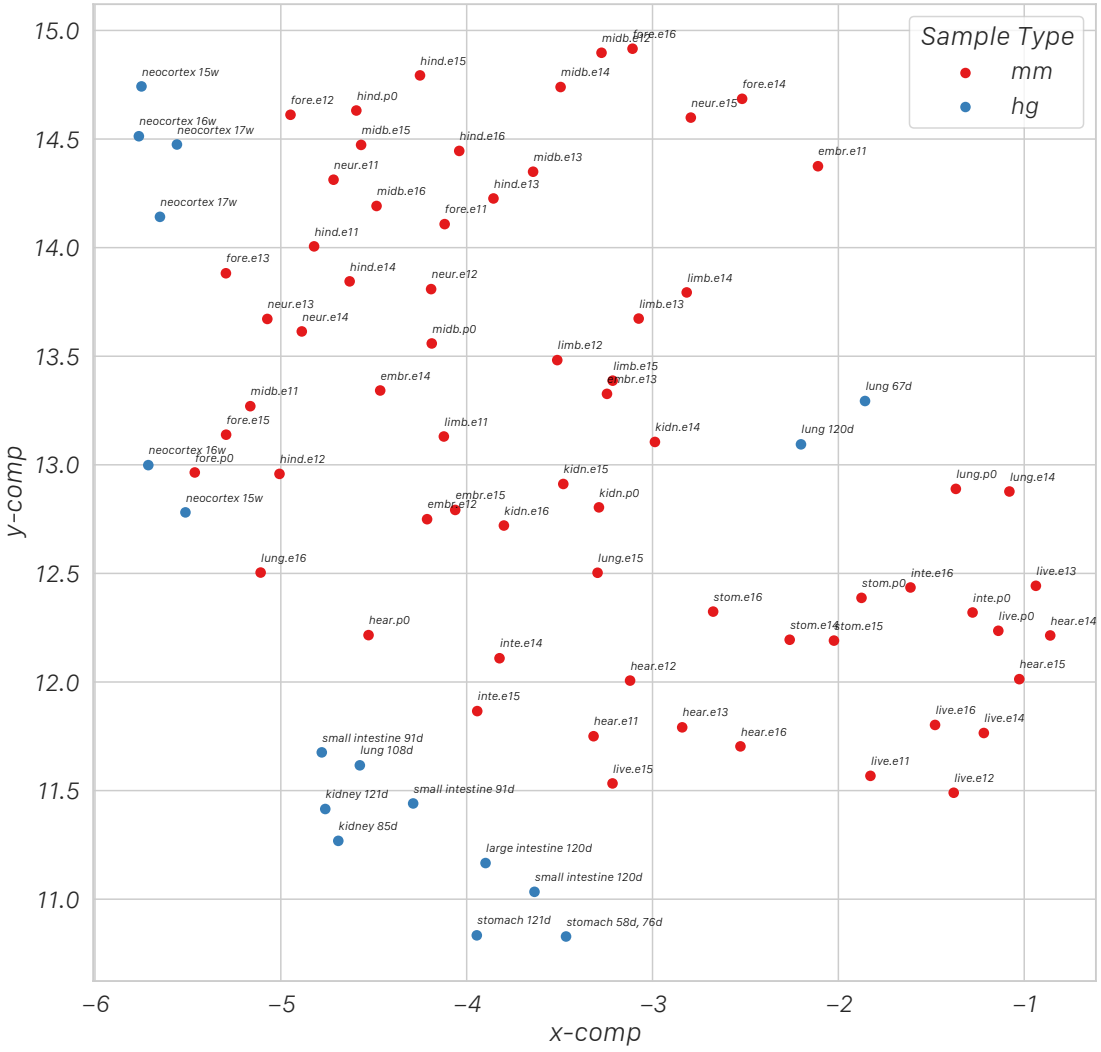

Figure 4.5. 2D projection of model predictions of selected samples using UMAP. Human samples are color-coded as blue dots and mouse samples re color-coded as red dots, revealing clusters of samples based on biological tissue similarity.

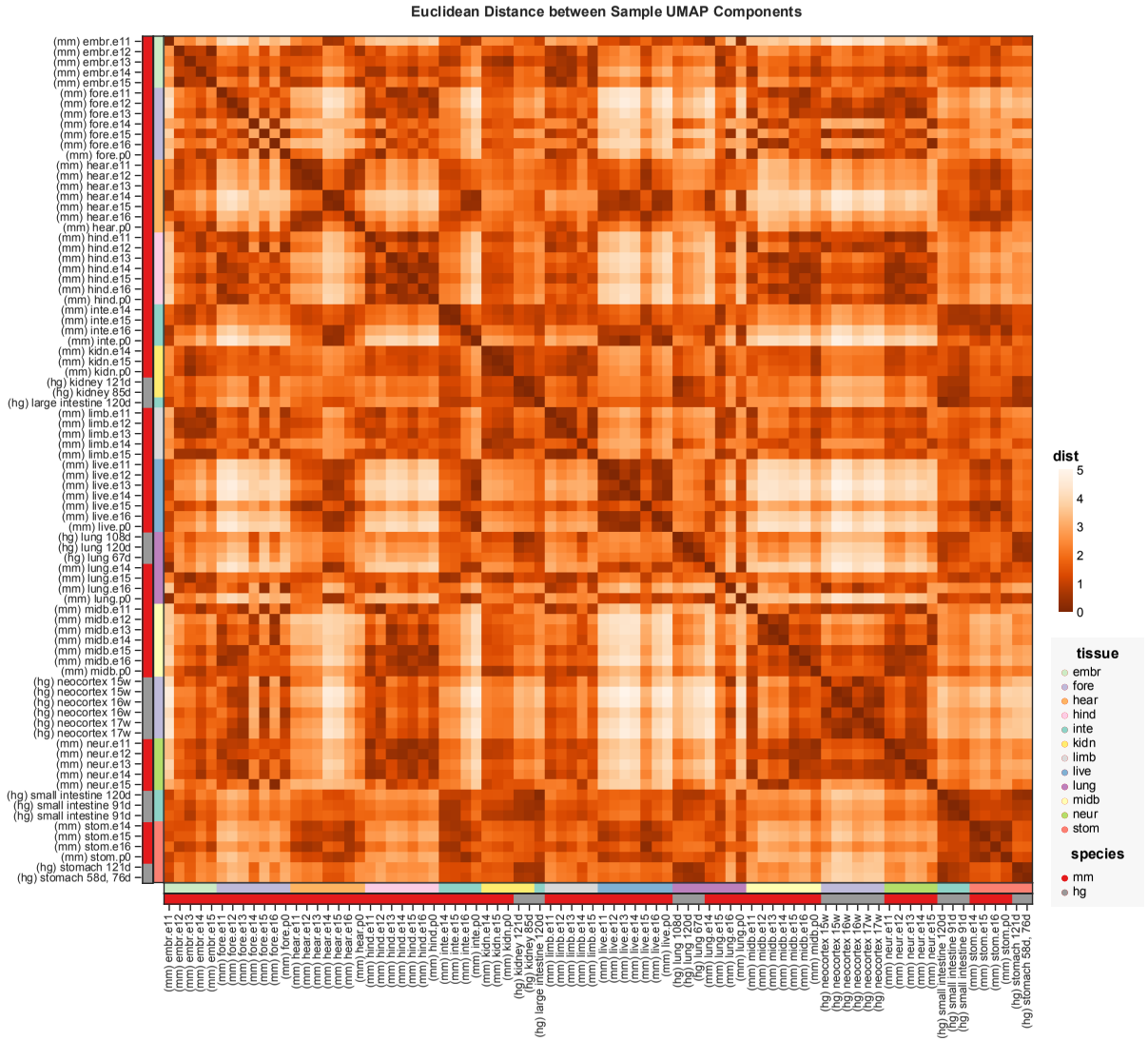

Figure 4.6. Pairwise Euclidean distance matrix between samples in the initial UMAP projected ”seed map”. Samples are hierarchically clustered along each axis. These distances serve as a quantitative measure of sample similarity observed in the UMAP projection.

as t-SNE is better used simply for grouping samples together as similar clusters, the human samples tended to be more closely associated with each other. Though the neocortex clearly associated with forebrain in the CNS-like super cluster for instance, t-SNE would be ill-suited for more granular analysis within this cluster.

Finally, we compared the projected maps of our neural network predictions against a comparable projection derived from using peaks. Notably, the peaks method requires a uniform set of peaks for each sample and cannot be used to compare samples on dissimilar genomic coordinate systems, so we could only project the mouse embryonic tissue samples. Using the catalog of d-TACs, we calculated the read counts within each d-TAC for each sample and used t-SNE (Figure 4.9 and UMAP (Figure 4.10 as described above. We note that there is a tighter clustering with peaks compared to the neural network model predictions, but that is not unexpected since using peaks for this type of map is somewhat cyclical in logic; if you already have a uniform set of peaks, there is no need to do a 2D projection to calculate correlation or distance between them in the first place. The important takeaway is that using the neural network predictions works similarly well for the NN models in terms of capturing biological similarity, while facilitating powerful comparisons in situations where peaks simply cannot be used. In fact, in the case of t-SNE, an argument can be made that the model predictions are actually *better* than the peaks-based approach, as t-SNE often created false super-clusters with the peaks data (e.g. hindbrain and forebrain being grouped with stomach and lung) depending on the perplexity and iteration count whereas the model t-SNE projections were more robust and stable.

4.4 Discussion

In this chapter, we proposed the use of a neural network architecture to discriminate between the accessible and inaccessible sequences in a given cellular context, based on the content of their sequence in terms of TF-relevant regulatory vocabulary. We show that models trained by this novel approach are highly predictive and specific to their sample of origin. Our

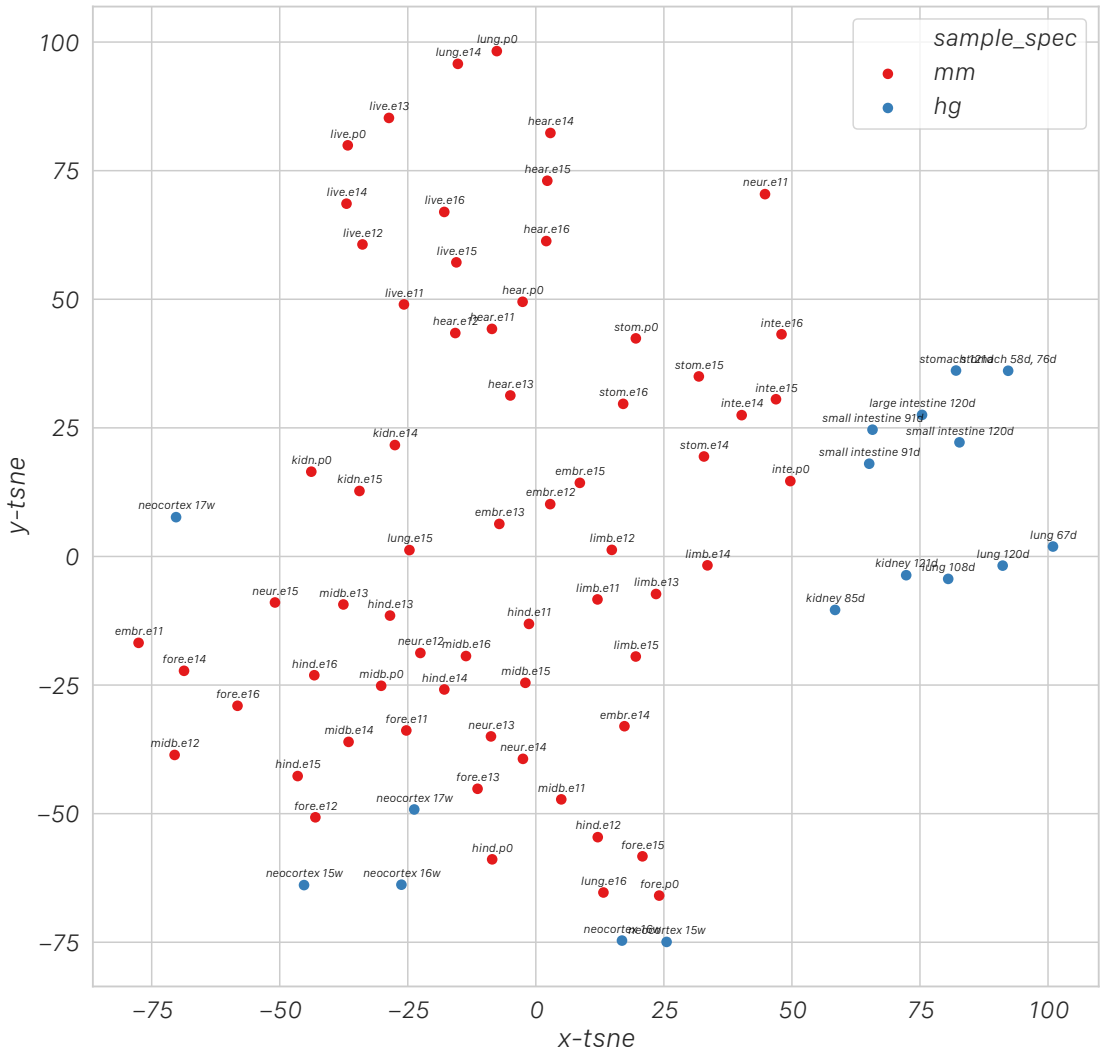

Figure 4.7. 2D projection of model predictions of selected samples using t-SNE. Human samples are color-coded as blue dots and mouse samples re color-coded as red dots, revealing clusters of samples based on biological tissue similarity.

neural networks have a further advantage in reducing the biases inherent in peak calls from samples of disparate quality and protocols. Furthermore, we provide a strategy for comparing samples of varied genomic background using these neural network models by predicting on a uniform set of diverse, representative sequences and performing dimensionality reduction on the resulting model x sequence matrix. These similarity representations are able to recapitulate known biological similarity as compared to a peaks-based read scoring strategy while enabling comparisons across different genomes. Based on these findings, we believe this sequence-based approach is a compelling alternative to peaks-based analysis for answering certain elusive biological questions in the field of gene regulation.

Several key highlights of this chapter are listed below:

- Peak calls are an imperfect unit for genomics analysis, especially sample comparison.
- Sequence-based strategies can better describe the regulatory network in a biologically interpretable manner.
- Using neural networks, we trained models for a wide range of samples and species, and projected their similarity in a graphical format.
- Being able to compare samples has numerous applications, including the identification of appropriate model systems and evaluating drug responses in disease research.

4.5 Methods

4.5.1 Extraction and shuffling of sequences from accessibility data

Peak calls were trimmed to their respective central 500bp using a custom script and only the TSS-distal peaks were taken (defined as being >=1kb away from either GENCODE v4 mouse TSS or human Ensembl hg38 promoters). From these trimmed distal peak centers, the corresponding fasta sequences were extracted using the 'bedtools getfasta' command. Finally,

shuffled genomic background sequences to be used as the negative control set were generated with the fasta-shuffle-letters command in the MEME tookit with parameters '-kmer 2 -tag -dinuc -seed 1'. These fasta sequences were then loaded into the neural network training program where they were converted into a 4x500 "one-hot" encoded array format.

4.5.2 Neural network architecture

The architecture of our convolutional neural network model is shown in Figure 4.8. To score the existing motif PWMs from the curated JASPAR library ([86]) as well as learn *de novo* motif filters with predictive potential, our neural network applies a 1-dimensional convolution over the input sequence (in one-hot array form) via sliding windows at each possible position. These filters, padded to a uniform 4x24 size, are fixed in the case of the JASPAR library motifs but allowed to be trained for the *de novo* sequences. In our final models, we use a combination of the 390 JASPAR motifs and 1000 *de novo* filters. The output from this layer is then padded and pooled in bins of 10bp (to a width of 50 pooled positions) and batch normalized.

The second convolutional layer is also a 1-dimensional convolution, but this time, over the pooled output from the first convolution. These kernel filters can be interpreted as the combination of motif scores from the first layer and represent the genomic grammar controlling the sample's regulatory network. A total of 500 kernels of size 1390x5 are trained in this layer. Notably, the width of these kernels is a parameter that can be tuned depending on how close or distant one wishes to constrain combinations of motifs. For instance, if one wishes only to identify combinations of motifs at a single pooled position, one should use a filter size of 1390x1. Conversely, to capture potentially long-ranging interactions across the entirety of the input sequence, one would select a filter size of 1390x50. For practical reasons and based on an empirical evaluation of performance, we chose to examine the genomic grammar covering short to medium-ranged distances, though this is a flexible parameter as are the number of kernels depending on the design goals of the neural network. The output of this layer is also pooled and batch normalized.

Finally, the a fully-connected layer of dense neurons connects to output neurons with a sigmoid activation function for binary classification. Our neural networks were implemented using the keras 2.2.4 front-end package and TensorFlow 1.10.0 as the back-end.

4.5.3 Model training and validation

Models were trained on the entire set of TSS-distal peaks for each sample, though optionally, TSS-proximal peaks can be included or the training set can be constrained artificially to a top cutoff of highest expressed peaks. We employed ten-fold cross-validation and repeated the model training ten times for each sample to verify performance metrics. Models were trained by optimizing the performance of the binary entropy metric as the model's validation loss function with the Adam optimizer. They were trained for 15 epochs to minimize the above validation loss metric while avoiding overfitting and a batch size of 64 to balance runtime against risk of sharp minima and poor model generalization. To assess the impact of number of kernel filters, we tested model training accuracy and loss (Table 4.3 at various epochs.

4.5.4 Generating a uniform prediction set

To generate a uniform set of sequences upon which to predict accessibility, we pooled the entire set of all positive accessible sequences in every sample, regardless of sample classification or origin. Then, we shuffled these sequences and randomly subsampled this superset down to 100K sequences using the 'fasta-subsample' command in the MEME tookit. We repeated this process twenty times to control against sampling bias when generating the similarity projection maps in addition to cross-validating with the unselected sequences from the subsampling.

4.5.5 Dimensionality reduction and clustering

Dimensionality reduction was performed using principal component analysis (PCA). After reducing the dimensions of the prediction array to 25, the data were projected onto a 2D plane with t-SNE (using perplexity of 10 and 1000 iterations) and with UMAP using the

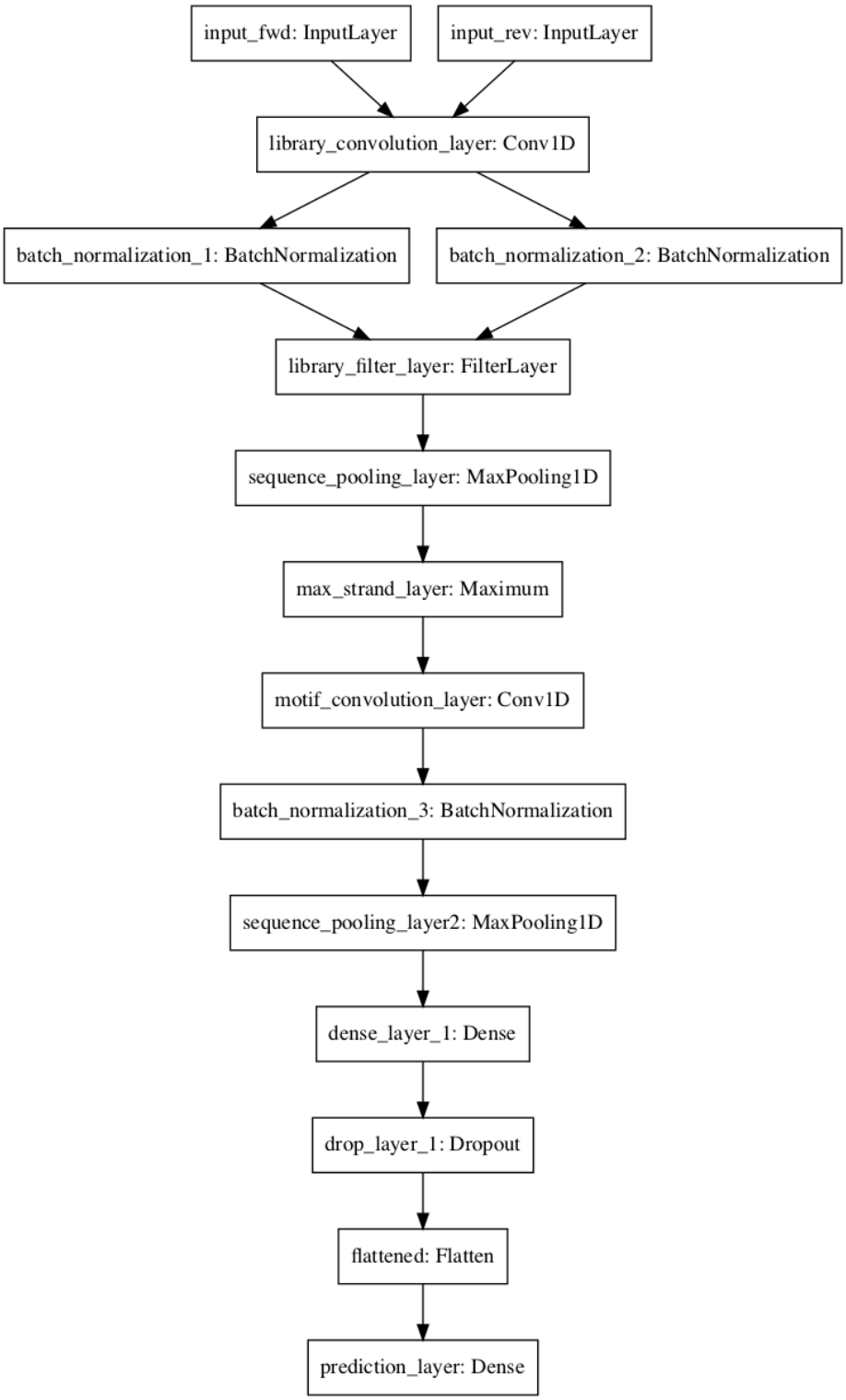

Figure 4.8. Neural network architecture, featuring two convolutional layers and a dense neuron layer. Max pooling and batch normalization steps are taking after each convolution.

'euclidean' metric. For the comparison figures based on read counts within a set of peaks (here, uniform d-TACs), we used the same parameters for t-SNE but the 'correlation' metric for UMAP.

4.5.6 Software availability

The source code for our neural network model implementation is available at: https://github.com/yuz207/atacgps.

4.6 Acknowledgments

We thank DU Gorkin for this guidance and assistance with study design and direction in this project. We also thank Y Li and Z Cheng for their technical contributions and discussions. We thank the UCSD Center of Epigenomics for their support and expertise.

Chapter 4, in part is currently being prepared for submission for publication of the material. Y Zhao, Z Cheng, Y Li, DU Gorkin & B Ren. The dissertation author was the primary investigator and author of this material.

4.7 Appendix

Table 4.4. Mean model performance for each sample used in the seed map, given by area under the receiver operating characteristic (ROC) and precision-recall (PRC) curves respectively.

Sample	auROC	auPRC
fore.e11.5	0.98	0.97
fore.e12.5	0.98	0.98
fore.e13.5	0.98	0.98
fore.e14.5	0.98	0.98
fore.e15.5	0.98	0.98
fore.e16.5	0.98	0.98
fore.p0	0.98	0.98
midb.e11.5	0.98	0.97
midb.e12.5	0.98	0.98
midb.e13.5	0.97	0.97
midb.e14.5	0.97	0.97
Continued on next page		

Table 4.4. Mean model performance for each sample used in the seed map, given by area under the receiver operating characteristic (ROC) and precision-recall (PRC) curves respectively.

Table 4.4 – continued from previous page

Sample	auROC	auPRC
midb.e15.5	0.98	0.98
midb.e16.5	0.98	0.98
midb.p0	0.98	0.98
hind.e11.5	0.98	0.98
hind.e12.5	0.98	0.98
hind.e13.5	0.98	0.98
hind.e14.5	0.98	0.98
hind.e15.5	0.98	0.98
hind.e16.5	0.98	0.97
hind.p0	0.98	0.98
neur.e11.5	0.98	0.98
neur.e12.5	0.98	0.98
neur.e13.5	0.98	0.98
neur.e14.5	0.98	0.98
neur.e15.5	0.98	0.98
limb.e11.5	0.97	0.97
limb.e12.5	0.97	0.97
limb.e13.5	0.98	0.98
limb.e14.5	0.98	0.98
limb.e15.5	0.98	0.98
embr.e11.5	0.97	0.97
embr.e12.5	0.97	0.97
embr.e13.5	0.97	0.97
embr.e14.5	0.98	0.98
embr.e15.5	0.98	0.98
hear.e11.5	0.98	0.97
hear.e12.5	0.98	0.98
hear.e13.5	0.97	0.97
hear.e14.5	0.97	0.97
hear.e15.5	0.98	0.98
hear.e16.5	0.98	0.98
hear.p0	0.98	0.98
live.e11.5	0.98	0.98
live.e12.5	0.98	0.98
live.e13.5	0.98	0.98
live.e14.5	0.98	0.98
live.e15.5	0.98	0.98
live.e16.5	0.98	0.98

Continued on next page

Table 4.4. Mean model performance for each sample used in the seed map, given by area under the receiver operating characteristic (ROC) and precision-recall (PRC) curves respectively.

Table 4.4 – continued from previous page

Sample	auROC	auPRC
live.p0	0.98	0.98
inte.e14.5	0.98	0.98
inte.e15.5	0.98	0.98
inte.e16.5	0.98	0.98
inte.p0	0.98	0.98
kidn.e14.5	0.98	0.98
kidn.e15.5	0.98	0.98
kidn.e16.5	0.98	0.97
kidn.p0	0.98	0.98
lung.e14.5	0.98	0.97
lung.e15.5	0.98	0.97
lung.e16.5	0.98	0.98
lung.p0	0.98	0.98
stom.e14.5	0.98	0.97
stom.e15.5	0.98	0.98
stom.e16.5	0.98	0.97
stom.p0	0.98	0.98
neocortex 15w	0.96	0.95
neocortex 15w	0.95	0.95
neocortex 16w	0.96	0.96
neocortex 16w	0.96	0.96
neocortex 17w	0.96	0.96
neocortex 17w	0.96	0.96
lung 67d	0.98	0.98
lung 108d	0.97	0.97
lung 120d	0.98	0.98
kidney 85d	0.97	0.97
kidney 121d	0.98	0.98
small inte 91d	0.98	0.98
small inte 120d	0.98	0.98
large inte 120d	0.98	0.98
stomach 58/76d	0.98	0.98
stomach 121d	0.98	0.98
kidney 85d	0.98	0.98

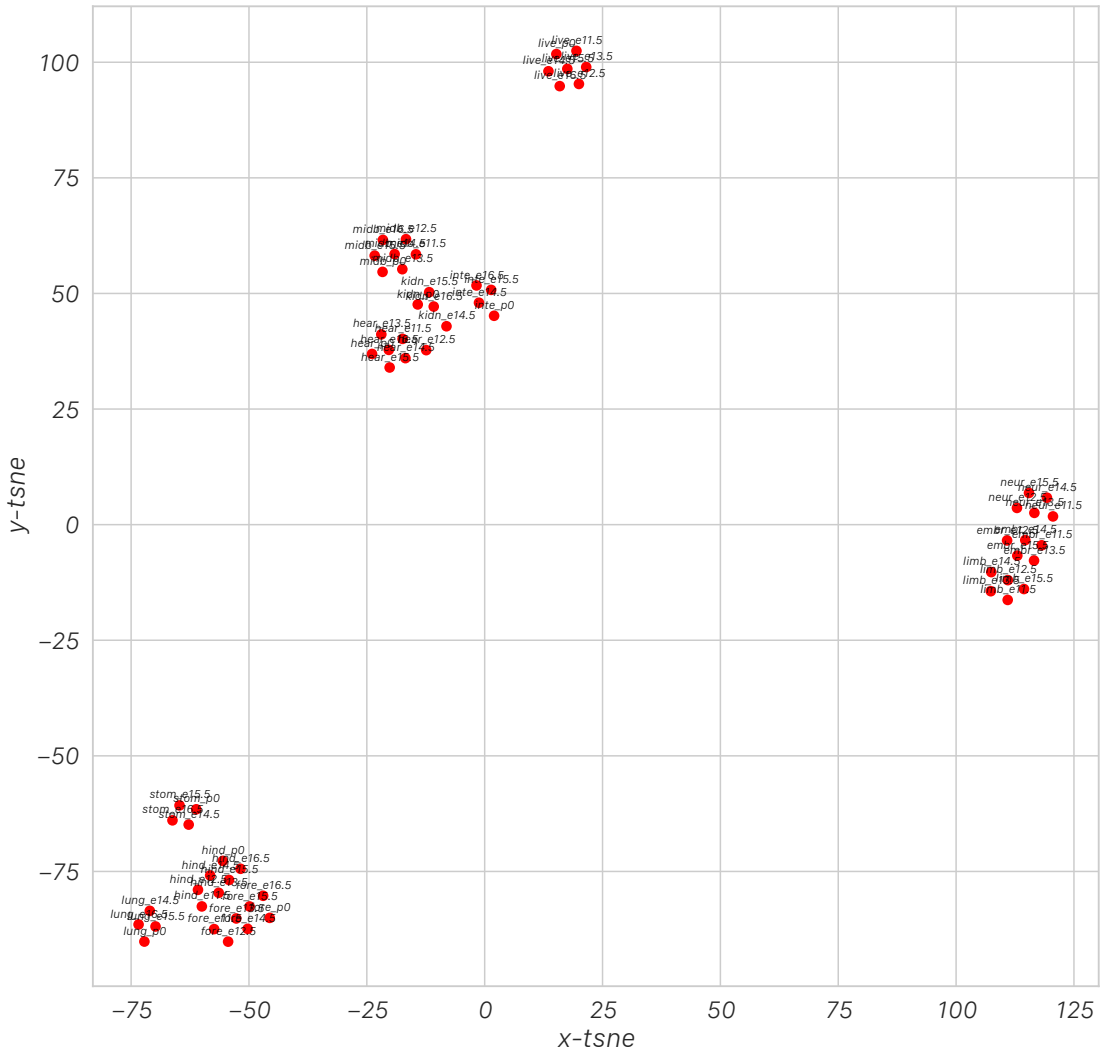

Figure 4.9. 2D projection of quantile normalized read counts over the uniform d-TAC catalog for mouse samples using t-SNE. This map serves as an comparable peak-based companion figure to the neural network model t-SNE projection in Figure 4.7, with the limitation that only the mouse embryonic samples could be projected due to inability to compare samples across different genomes using peaks.

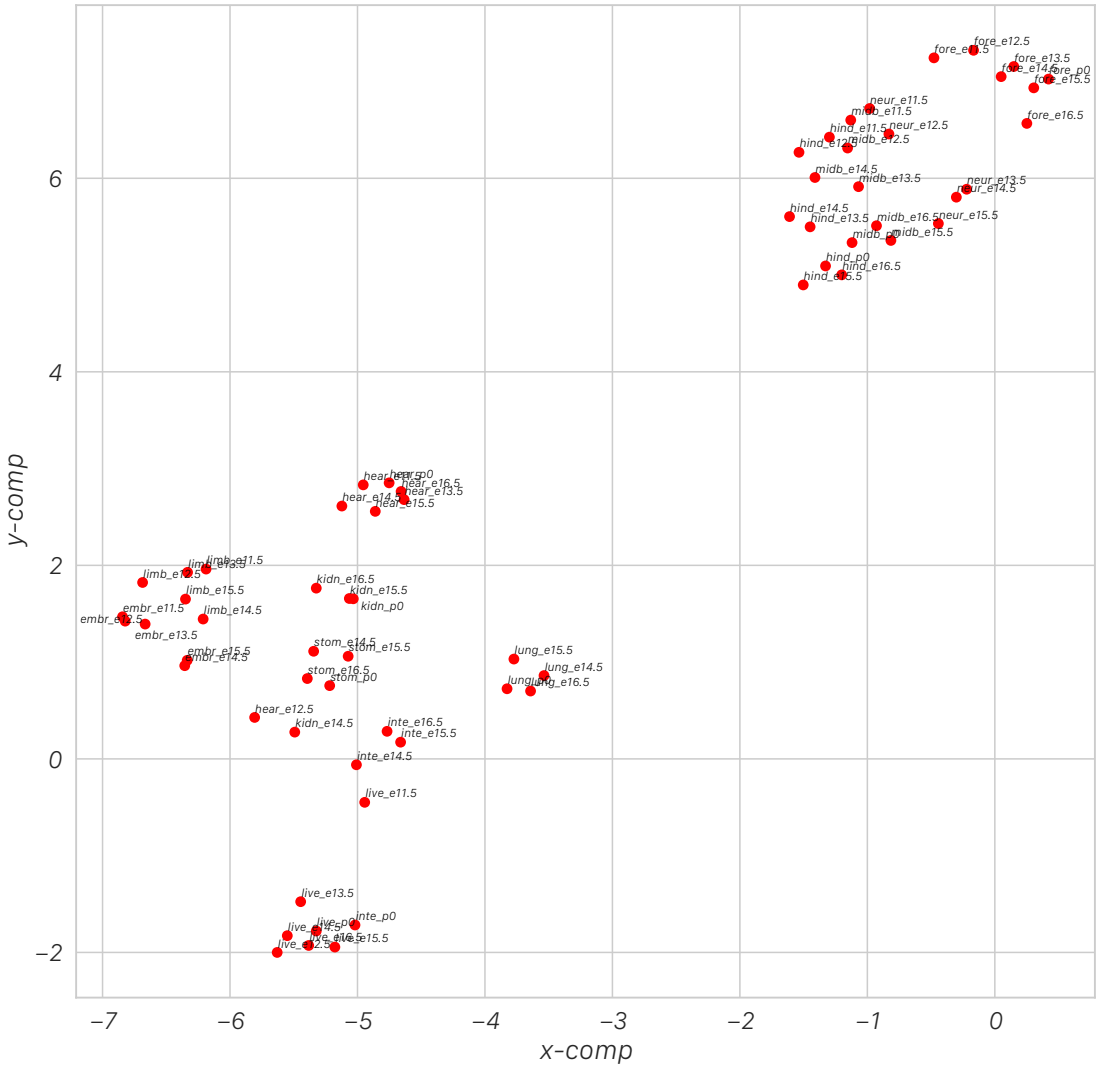

Figure 4.10. 2D projection of quantile normalized read counts over the uniform d-TAC catalog for mouse samples using UMAP. This map serves as an comparable peak-based companion figure to the neural network model UMAP projection in Figure 4.5, with the limitation that only the mouse embryonic samples could be projected due to inability to compare samples across different genomes using peaks. Note that this UMAP metric used correlation, which is better suited to the data type, rather than euclidean, which clustered tissues well but formed unexpected super-clusters of unrelated tissues.

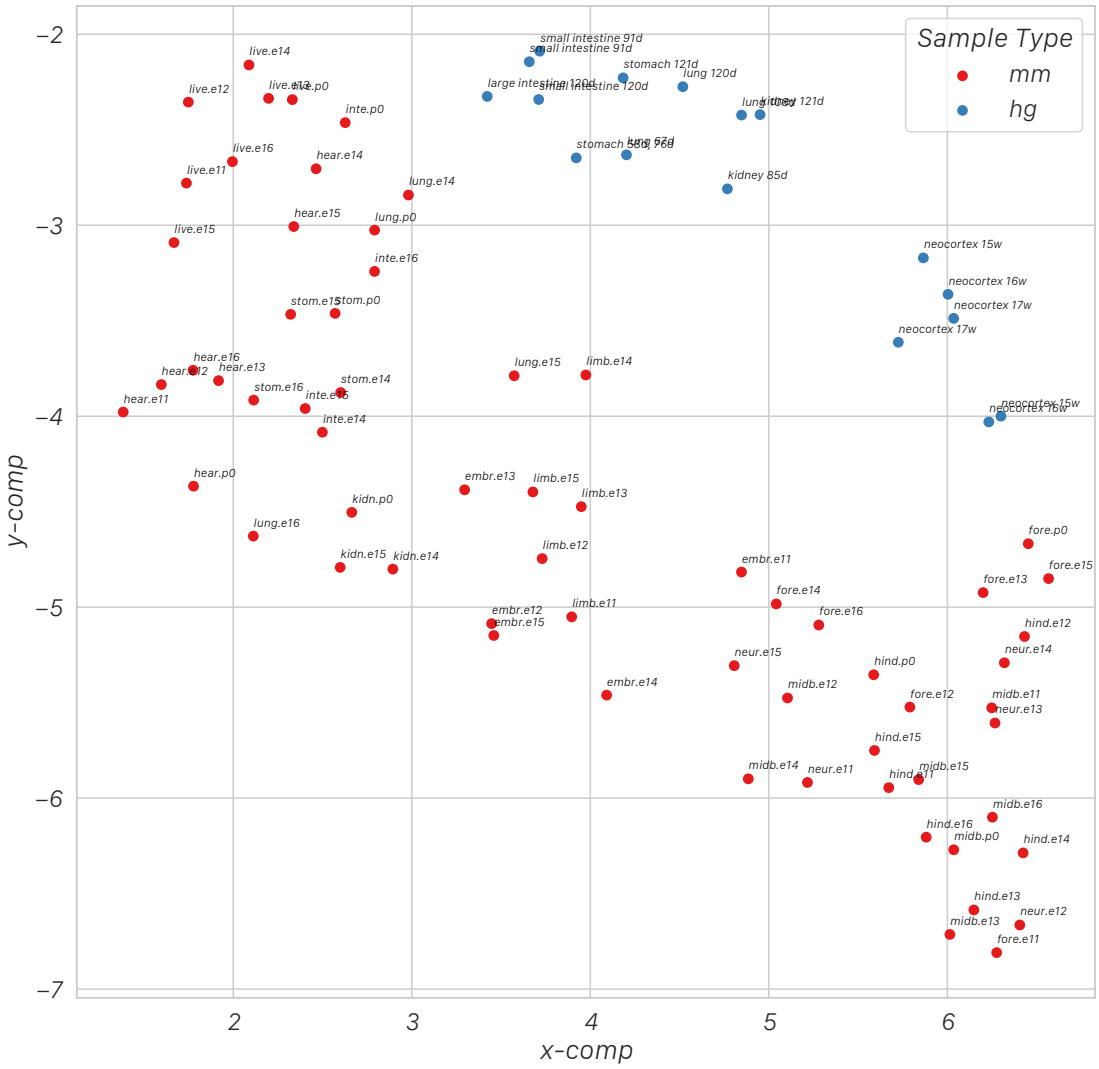

Figure 4.11. Included for comparison with Figure 4.10, this UMAP projection contains the same data as in Figure 4.5 but using the 'correlation' metric instead of the 'euclidean' metric. With either metric, the conclusions are unchanged and the choice of metric ultimately will depend on specific datasets and clustering priorities. With 'correlation', there is a larger species-orientated grouping than with 'euclidean' for the samples modeled.

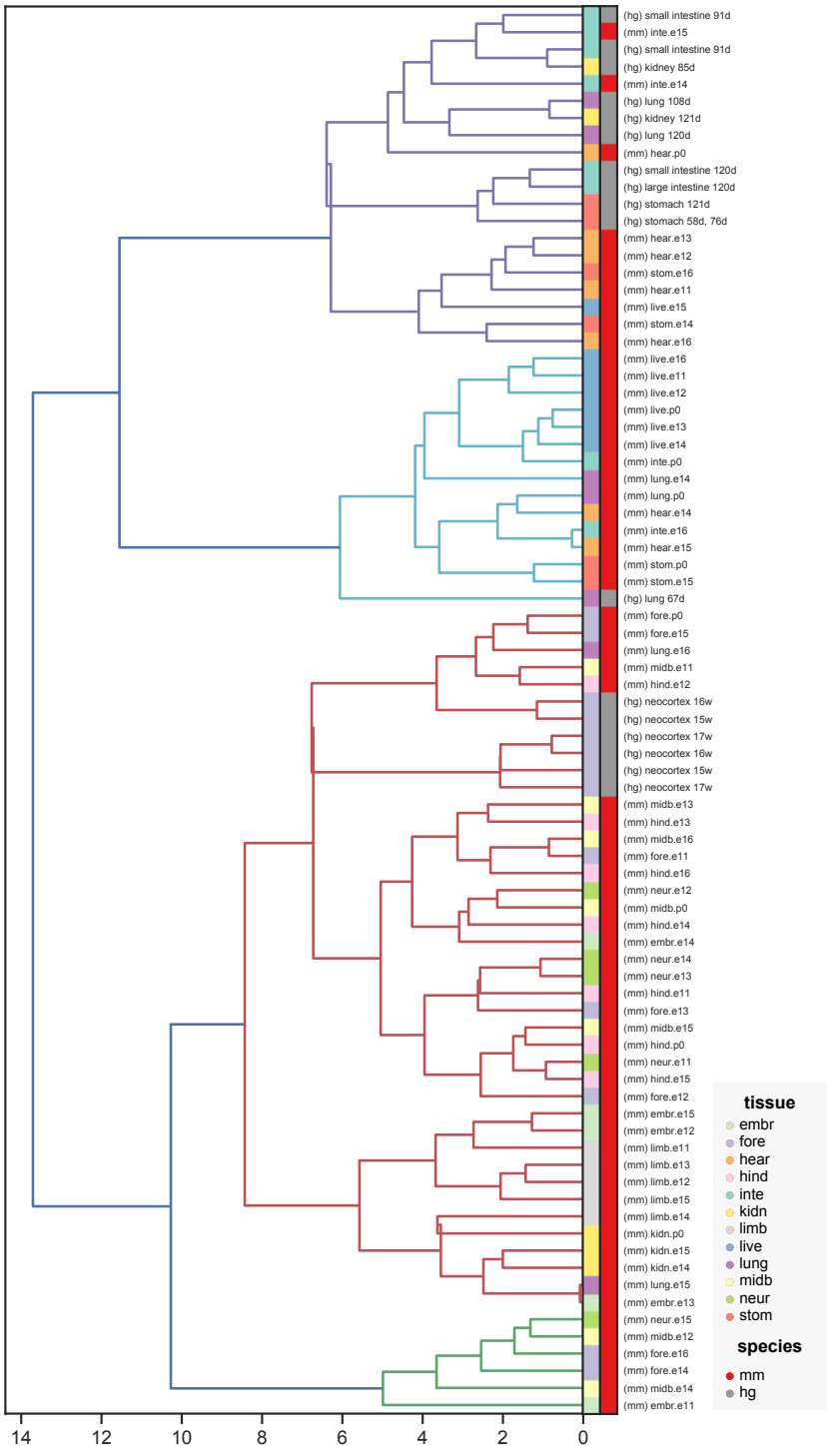

Figure 4.12. Dendrogram of hierarchical clustering of the UMAP euclidean distance matrix in Figure 4.6 to organize samples according to their relative similarity to all other samples. Note the general association of samples from similar tissue origins regardless of species.

Chapter 5

Conclusions

In summary, our results describe a multi-tiered compendium of functional annotations for the developmental mouse genome, including chromatin state maps for 66 distinct tissue-stages, an extensive catalog of candidate cREs, many with dynamic temporal activity, and enhancer target correlative interaction predictions. By systematically profiling tissues across sequential stages in late development, our catalog provides an unprecedented examination of the differentially dynamic chromatin patterns involved in this highly dynamic period of growth and maturation. Additionally, the breadth of tissues examined presents an illustrative, global view of mouse development at the organism level. These two comprehensive dimensions of data enable the genome-wide study of dynamic chromatin and TF binding through embryogenesis.

We characterized the changes in accessibility of this catalog as well as the changes in functional state annotations, highlighting key tissue-specific and temporally-restricted patterns driving the respective regulatory programs of each tissue. These characterizations also enable the study of transcription factor binding and the identification of potential regulators in each developmental context as well as their specific, temporal activity profiles. Finally, we extend the application of this mouse catalog to conserved loci in human, identifying orthologous regions that are functionally enriched for relevant human traits. These findings not only expand the annotation of human regulatory sequences with putative embryonic-active regions, but also provide support for the use of the mouse as a model organism in studying human gestational development. In

sum, this report provides a crucial and unprecedented perspective on the dynamically fluid epigenetic architecture of mouse genome, and novel insights into the processes that govern it during embryogenesis.

Despite the broad scope of this study, we note some important limitations. First, there are multiple developmental tissues that were not surveyed here (e.g. skeleton, gonads, pancreas). Second, as noted above, the tissues examined here are heterogenous, and future efforts to examine the epigenomes of single cells during development will be critical to achieve a deeper understanding of developmental gene regulation. Nonetheless, to our knowledge the survey of fetal chromatin landscapes reported here is unprecedented in its breadth. Given the uniquely critical role of the mouse as a model system in biomedical research, we believe that these data and insights will be a valuable resource to the biomedical research community.

Finally, we used these valuable mouse data sets in conjunction with a number of human datasets to explore novel methods for understanding biological similarity. By analyzing the sequence content of potential regulatory elements rather than using the arbitrary definition of peak coordinates, we were able to train powerful and highly predictive models for individual samples and cell-types using convolutional neural networks. There are further advances based on these machine learning strategies currently being explored as follow-up work. First, we plan to extract the patterns contained within the second, ”grammar filter” layer of the convolutional networks and use these information to cluster samples, providing access not only to a similarity matrix of samples but also to the specific, biologically relevant ways their regulatory programs are similar. Likewise, the intermediate prediction matrices can be extracted from the models as well to cluster the sequences of each sample, grouping them within a sample, with applications such as identifying clusters within single-cell data. Finally, we can generate a new uniform input sequence based on the content of the aforementioned ”grammar filters” to remove the need for the convolutional layers themselves to be biologically interpretable, enabling deeper and more advanced neural network models. This study therefore serves as an important proof-of-concept for leveraging neural networks and sequence-based strategies in general for understanding

biological regulatory networks and facilitating sample comparison.

Bibliography

[1] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet 17, 487-500, doi:10.1038/nrg.2016.5(2016).

[2] Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. Nature reviews. Molecular cell biology 15, 703-708, doi:10.1038/nr890 (2014).

[3] Roy DM, Walsh LA, Chan TA. Driver mutations of cancer epigenomes. Protein & cell 5, 265-296, doi:10.1007/s13238-014-0031-6 (2014).

[4] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49, doi:10.1038/nature09906 (2011).

[5] Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology 4, e1000201, doi:10.1371/journal.pcbi.1000201 (2008).

[6] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272-286, doi:10.103nrg3682 (2014).

[7] CH Waddington. The Strategy of Genes. George Allen & Unwin. (1957)

[8] Visel A, Rubin E. M, Pennacchio L. A. Genomic views of distant-acting enhancers. Nature 461, 199-205, doi:10.103nature08451 (09).

[9] Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol 21, 210-219, doi:10.1038/nsmb.2784 (2014).

[10] Maurano MT,, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-1195, doi:10.1126/science.1222794 (2012).

[11] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B.. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics 39, 311-318, doi:10.1038/ng1966 (2007).

[12] Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-283, doi:10.1038/nature09692 (2011).

[13] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America 107, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

[14] Zaret KS, Mango E. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr Opin Genet Dev 37, 76-81, doi:10.1016/j.gde.2015.12.003 (2016).

[15] Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, Rubenstein JL, Rubin EM, Pennacchio LA, Visel A. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell 155, 1521-1531, doi:10.1016/j.cell.2013.11.033 (2013).

[16] Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116-120, do10.1038/natur1243 (2012).

[17] Cotney J, Leng J, Oh S, DeMare LE, Reilly SK, Gerstein MB, Noonan JP. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. Genome research 22, 1069-1080, doi:10.1101/gr.129817.111 (2012).

[18] Dickel DE, Barozzi I, Zhu Y, Fukuda-Yuzawa Y, Osterwalder M, Mannion BJ, May D, Spurrell CH, Plajzer-Frick I, Pickle CS, Lee E, Garvin TH, Kato M, Akiyama JA, Afzal V, Lee AY, Gorkin DU, Ren B, Rubin EM, Visel A, Pennacchio LA. Genome-wide compendium and functional assessment of in vivo heart enhancers. Nature communications 7, 12923, doi:10.1038/ncomms12923 (2016).

[19] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213218, doi:10.18/nmeth.2688 (2013).

[20] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen

RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature 489, 75-82doi:10.1038/nure11232 (2012).

[21] Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B; Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515,55-364, doi:11038/nature13992 (2014).

[22] Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, Li W, Li Y, Ma J, Peng X, Zheng H, Ming J, Zhang W, Zhang J, Tian G, Xu F, Chang Z, Na J, Yang X, Xie W. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature 4, 652-657, d:10.1038/nature18606 (2016).

[23] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nate 518, 317-33 doi:10.1038/nature14248 (2015).

[24] Gorkin DU, Barozzi I, Zhao Y, Zhang Y, Huang H, Lee AY, Liu B, Chiou J, Wildberg A, Ding B, Zhang B, Wang M, Strattan JS, Davidson JM, Qiu Y, Afzal V, Akiyama JA, Plajzer-Frick I, Novak CS, Kato M, Garvin TH, Pham QT, Harringon AN, Mannion BJ, Lee EA, Fukuda-Yuzawa Y, He Y, Preissl S, Chee S, Han JY, Williams Ba, Trout D, Amrhein H, Yang H, Cherry JM, Wang W, Gaulton K, Ecker JR, Shen Y, Dickel DE, Visel A, Pennacchio LA, Ren B. An atlas of dynamic chromatin landscapes in the developing mouse fetus. Nature (2019, in press).

[25] He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon RG, Nery JR, Lee AY, Zhao Y, Huang H, Williams BA, Trout D, Amrhein H, Fang R, Chen H, Li B, Visel A, Pennacchio LA, Ren B, Ecker JR. Spatiotemporal DNA methylome dynamics of the developing mammalian fetus. Nature (2019, in press).

[26] Ernst J, Kellis M. in Nature methods Vol. 9 215-216 (2012).

[27] Shay T, Kang J. Immunological Genome project and systems immunology. Trends Immunol 34, 602-609, doi: 10.1016/j.it.2013.03.004 (2015).

[28] Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic acids research 35, D88-92, doi:10.1093/nar/gkl822 (2007).

[29] McCormick MB, Tamimi RM, Snider L, Asakura A, Bergstrom D, Tapscott SJ. NeuroD2 and neuroD3: distinct expression patterns and transcriptional activation potentials within the neuroD gene family. Molecular and cellular biology 16, 5792-5800 (1996).

[30] Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, Bennett DA, Houmard JA, Muoio DM, Onder TT, Camahort R, Cowan CA, Meissner A, Epstein CB, Shoresh N, Bernstein BE. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell 152, 642-654, doi:10.1016/j.cell.2012.12.033 (2013).

[31] Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research 42, W187-191, doi:10.1093/nar/gku365 (2014).

[32] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

[33] Dickel DE, Ypsilanti AR, Pla R, Zhu Y, Barozzi I, Mannion BJ, Khin YS, Fukuda-Yuzawa Y, Plajzer-Frick I, Pickle CS, Lee EA, Harrington AN, Pham QT, Garvin TH, Kato M, Osterwalder M, Akiyama JA, Afzal V, Rubenstein JLR, Pennacchio LA, Visel A. Ultraconserved Enhancers Are Required for Normal Development. Cell 172, 491-499.e415, doi:10.1016/j.cell.2017.12.017 (2018).

[34] Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34, D590-598, doi:10.1093/nar/gkj144 (2006).

[35] Pers TH, Timshel P, Hirschhorn JN. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics 31, 418-420, doi:10.1093/bioinformatics/btu655 (2015).

[36] Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. RNA splicing is a primary link between genetic variation and disease. Science 352, 600-604, doi:10.1126/science.aad9417 (2016).

[37] Jones SE, Tyrrell J, Wood AR, Beaumont RN, Ruth KS, Tuke MA, Yaghootkar H, Hu Y, Teder-Laving M, Hayward C, Roenneberg T, Wilson JF, Del Greco F, Hicks AA, Shin C, Yun CH, Lee SK, Metspalu A, Byrne EM, Gehrman PR, Tiemeier H, Allebrandt KV, Freathy RM, Murray A, Hinds DA, Frayling TM, Weedon MN. Genome-Wide Association Analyses in 128,266 Individuals Identifies New Morningness and Sleep Duration Loci. PLOS Genet. 12, e1006125 (2016).

[38] Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landn M, Adli M, Alda M, Ardau R, Arias B, Aubry JM, Backlund L, Badner JA, Barrett TB, Bauer M, Baune BT, Bellivier F, Benabarre A, Bengesser S, Berrettini WH, Bhattacharjee AK, Biernacka JM, Birner A, Bloss CS, Brichant-Petitjean C, Bui ET, Byerley W, Cervantes P, Chillotti C, Cichon S, Colom F, Coryell W, Craig DW, Cruceanu C, Czerski PM, Davis T, Dayer A, Degenhardt F, Del Zompo M, DePaulo JR, Edenberg HJ, tain B, Falkai P, Foroud T, Forstner AJ, Frisn L, Frye MA, Fullerton JM, Gard S, Garnham JS, Gershon ES, Goes FS, Greenwood TA, Grigoroiu-Serbanescu M, Hauser J, Heilbronner U, Heilmann-Heimbach S, Herms S, Hipolito M, Hitturlingappa S, Hoffmann P, Hofmann A, Jamain S, Jimnez E, Kahn JP, Kassem L, Kelsoe JR, Kittel-Schneider S, Kliwicki S, Koller DL, Knig B, Lackner N, Laje G, Lang M, Lavebratt C, Lawson WB, Leboyer M, Leckband SG, Liu C, Maaser A, Mahon PB, Maier W, Maj M, Manchia M, Martinsson L, McCarthy MJ, McElroy SL, McInnis MG, McKinney R, Mitchell PB, Mitjans M, Mondimore FM, Monteleone P, Mhleisen TW, Nievergelt CM, Nthen MM, Novk T, Nurnberger JI Jr, Nwulia EA, sby U, Pfennig A, Potash JB, Propping P, Reif A, Reininghaus E, Rice J, Rietschel M, Rouleau GA, Rybakowski JK, Schalling M, Scheftner WA, Schofield PR, Schork NJ, Schulze TG, Schumacher J, Schweizer BW, Severino G, Shekhtman T, Shilling PD, Simhandl C, Slaney CM, Smith EN, Squassina A, Stamm T, Stopkova P, Streit F, Strohmaier J, Szelinger S, Tighe SK, Tortorella A, Turecki G, Vieta E, Volkert J, Witt SH, Wright A, Zandi PP, Zhang P, Zollner S, McMahon FJ. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. Hum. Mol. Genet. 25, 33833394 (2016).

[39] Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, Baldursson G, Belliveau R, Bybjerg-Grauholm J, Bkvad-Hansen M, Cerrato F, Chambert K, Churchhouse

C, Dumont A, Eriksson N, Gandal M, Goldstein JI, Grasby KL, Grove J, Gudmundsson OO, Hansen CS, Hauberg ME, Hollegaard MV, Howrigan DP, Huang H, Maller JB, Martin AR, Martin NG, Moran J, Pallesen J, Palmer DS, Pedersen CB, Pedersen MG, Poterba T, Poulsen JB, Ripke S, Robinson EB, Satterstrom FK, Stefansson H, Stevens C, Turley P, Walters GB, Won H, Wright MJ; ADHD Working Group of the Psychiatric Genomics Consortium (PGC); Early Lifecourse Genetic Epidemiology (EAGLE) Consortium; 23andMe Research Team, Andreassen OA, Asherson P, Burton CL, Boomsma DI, Cormand B, Dalsgaard S, Franke B, Gelernter J, Geschwind D, Hakonarson H, Haavik J, Kranzler HR, Kuntsi J, Langley K, Lesch KP, Middeldorp C, Reif A, Rohde LA, Roussos P, Schachar R, Sklar P, Sonuga-Barke EJS, Sullivan PF, Thapar A, Tung JY, Waldman ID, Medland SE, Stefansson K, Nordentoft M, Hougaard DM, Werge T, Mors O, Mortensen PB, Daly MJ, Faraone SV, Brglum AD, Neale BM. Discovery Of The First Genome-Wide Significant Risk Loci For ADHD. Nat Genet 51(1):63-75. doi: 10.1038/s41588-018-0269-7 (2019).

[40] Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, Byrne EM, Blackwood DH, Boomsma DI, Cichon S, Heath AC, Holsboer F, Lucae S, Madden PA, Martin NG, McGuffin P, Muglia P, Noethen MM, Penninx BP, Pergadia ML, Potash JB, Rietschel M, Lin D, Mller-Myhsok B, Shi J, Steinberg S, Grabe HJ, Lichtenstein P, Magnusson P, Perlis RH, Preisig M, Smoller JW, Stefansson K, Uher R, Kutalik Z, Tansey KE, Teumer A, Viktorin A, Barnes MR, Bettecken T, Binder EB, Breuer R, Castro VM, Churchill SE, Coryell WH, Craddock N, Craig IW, Czamara D, De Geus EJ, Degenhardt F, Farmer AE, Fava M, Frank J, Gainer VS, Gallagher PJ, Gordon SD, Goryachev S, Gross M, Guipponi M, Henders AK, Herms S, Hickie IB, Hoefels S, Hoogendijk W, Hottenga JJ, Iosifescu DV, Ising M, Jones I, Jones L, Jung-Ying T, Knowles JA, Kohane IS, Kohli MA, Korszun A, Landen M, Lawson WB, Lewis G, Macintyre D, Maier W, Mattheisen M, McGrath PJ, McIntosh A, McLean A, Middeldorp CM, Middleton L, Montgomery GM, Murphy SN, Nauck M, Nolen WA, Nyholt DR, O'Donovan M, Oskarsson H, Pedersen N, Scheftner WA, Schulz A, Schulze TG, Shyn SI, Sigurdsson E, Slager SL, Smit JH, Stefansson H, Steffens M, Thorgeirsson T, Tozzi F, Treutlein J, Uhr M, van den Oord EJ, Van Grootheest G, Vlzke H, Weilburg JB, Willemsen G, Zitman FG, Neale B, Daly M, Levinson DF, Sullivan PF. A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry 18, 497511 (2013).

[41] Robinson EB, St Pourcain B, Anttila V, Kosmicki JA, Bulik-Sullivan B, Grove J, Maller J, Samocha KE, Sanders SJ, Ripke S, Martin J, Hollegaard MV, Werge T, Hougaard DM; iPSYCH-SSI-Broad Autism Group, Neale BM, Evans DM, Skuse D, Mortensen PB, Brglum AD, Ronald A, Smith GD, Daly MJ. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat. Genet. 48, 552555 (2016).

[42] Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR, Manoharan A, Ortmann W, Bhangale T, Denny JC, Carroll RJ, Eyler AE, Greenberg JD, Kremer JM, Pappas DA, Jiang L, Yin J, Ye L, Su DF, Yang J, Xie G, Keystone E, Westra HJ, Esko T, Metspalu A, Zhou X, Gupta N, Mirel D, Stahl EA, Diogo D, Cui J, Liao K, Guo MH, Myouzen K, Kawaguchi T, Coenen MJ, van Riel PL, van de

Laar MA, Guchelaar HJ, Huizinga TW, Dieud P, Mariette X, Bridges SL Jr, Zhernakova A, Toes RE, Tak PP, Miceli-Richard C, Bang SY, Lee HS, Martin J, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Rantap-Dahlqvist S, Arlestig L, Choi HK, Kamatani Y, Galan P, Lathrop M; RACI consortium; GARNET consortium, Eyre S, Bowes J, Barton A, de Vries N, Moreland LW, Criswell LA, Karlson EW, Taniguchi A, Yamada R, Kubo M, Liu JS, Bae SC, Worthington J, Padyukov L, Klareskog L, Gregersen PK, Raychaudhuri S, Stranger BE, De Jager PL, Franke L, Visscher PM, Brown MA, Yamanaka H, Mimori T, Takahashi A, Xu H, Behrens TW, Siminovitch KA, Momohara S, Matsuda F, Yamamoto K, Plenge RM. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376381 (2014)

[43] Ji SG, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, Kumasaka N, Atkinson EJ, Schlicht EM, Liu JZ, Shah T, Gutierrez-Achury J, Boberg KM, Bergquist A, Vermeire S, Eksteen B, Durie PR, Farkkila M, Mller T, Schramm C, Sterneck M, Weismller TJ, Gotthardt DN, Ellinghaus D, Braun F, Teufel A, Laudes M, Lieb W, Jacobs G, Beuers U, Weersma RK, Wijmenga C, Marschall HU, Milkiewicz P, Pares A, Kontula K, Chazouillres O, Invernizzi P, Goode E, Spiess K, Moore C, Sambrook J, Ouwehand WH, Roberts DJ, Danesh J, Floreani A, Gulamhusein AF, Eaton JE, Schreiber S, Coltescu C, Bowlus CL, Luketic VA, Odin JA, Chopra KB, Kowdley KV, Chalasani N, Manns MP, Srivastava B, Mells G, Sandford RN, Alexander G, Gaffney DJ, Chapman RW, Hirschfield GM, de Andrade M; UK-PSC Consortium; International IBD Genetics Consortium; International PSC Study Group, Rushbrook SM, Franke A, Karlsen TH, Lazaridis KN, Anderson CA. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nat. Genet. 49, 269273 (2017).

[44] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, Lo KS, Locke AE, Mgi R, Mihailov E, Porcu E, Randall JC, Scherag A, Vinkhuyzen AA, Westra HJ, Winkler TW, Workalemahu T, Zhao JH, Absher D, Albrecht E, Anderson D, Baron J, Beekman M, Demirkan A, Ehret GB, Feenstra B, Feitosa MF, Fischer K, Fraser RM, Goel A, Gong J, Justice AE, Kanoni S, Kleber ME, Kristiansson K, Lim U, Lotay V, Lui JC, Mangino M, Mateo Leach I, Medina-Gomez C, Nalls MA, Nyholt DR, Palmer CD, Pasko D, Pechlivanis S, Prokopenko I, Ried JS, Ripke S, Shungin D, Stanckov A, Strawbridge RJ, Sung YJ, Tanaka T, Teumer A, Trompet S, van der Laan SW, van Setten J, Van Vliet-Ostaptchouk JV, Wang Z, Yengo L, Zhang W, Afzal U, Arnlv J, Arscott GM, Bandinelli S, Barrett A, Bellis C, Bennett AJ, Berne C, Blher M, Bolton JL, Bttcher Y, Boyd HA, Bruinenberg M, Buckley BM, Buyske S, Caspersen IH, Chines PS, Clarke R, Claudi-Boehm S, Cooper M, Daw EW, De Jong PA, Deelen J, Delgado G, Denny JC, Dhonukshe-Rutten R, Dimitriou M, Doney AS, Drr M, Eklund N, Eury E, Folkersen L, Garcia ME, Geller F, Giedraitis V, Go AS, Grallert H, Grammer TB, Grler J, Grnberg H, de Groot LC, Groves CJ, Haessler J, Hall P, Haller T, Hallmans G, Hannemann A, Hartman CA, Hassinen M, Hayward C, Heard-Costa NL, Helmer Q, Hemani G, Henders AK, Hillege HL, Hlatky MA, Hoffmann W, Hoffmann

P, Holmen O, Houwing-Duistermaat JJ, Illig T, Isaacs A, James AL, Jeff J, Johansen B, Johansson , Jolley J, Juliusdottir T, Junttila J, Kho AN, Kinnunen L, Klopp N, Kocher T, Kratzer W, Lichtner P, Lind L, Lindstrm J, Lobbens S, Lorentzon M, Lu Y, Lyssenko V, Magnusson PK, Mahajan A, Maillard M, McArdle WL, McKenzie CA, McLachlan S, McLaren PJ, Menni C, Merger S, Milani L, Moayyeri A, Monda KL, Morken MA, Mller G, Mller-Nurasyid M, Musk AW, Narisu N, Nauck M, Nolte IM, Nthen MM, Oozageer L, Pilz S, Rayner NW, Renstrom F, Robertson NR, Rose LM, Roussel R, Sanna S, Scharnagl H, Scholtens S, Schumacher FR, Schunkert H, Scott RA, Sehmi J, Seufferlein T, Shi J, Silventoinen K, Smit JH, Smith AV, Smolonska J, Stanton AV, Stirrups K, Stott DJ, Stringham HM, Sundstrm J, Swertz MA, Syvnen AC, Tayo BO, Thorleifsson G, Tyrer JP, van Dijk S, van Schoor NM, van der Velde N, van Heemst D, van Oort FV, Vermeulen SH, Verweij N, Vonk JM, Waite LL, Waldenberger M, Wennauer R, Wilkens LR, Willenborg C, Wilsgaard T, Wojczynski MK, Wong A, Wright AF, Zhang Q, Arveiler D, Bakker SJ, Beilby J, Bergman RN, Bergmann S, Biffar R, Blangero J, Boomsma DI, Bornstein SR, Bovet P, Brambilla P, Brown MJ, Campbell H, Caulfield MJ, Chakravarti A, Collins R, Collins FS, Crawford DC, Cupples LA, Danesh J, de Faire U, den Ruijter HM, Erbel R, Erdmann J, Eriksson JG, Farrall M, Ferrannini E, Ferrires J, Ford I, Forouhi NG, Forrester T, Gansevoort RT, Gejman PV, Gieger C, Golay A, Gottesman O, Gudnason V, Gyllensten U, Haas DW, Hall AS, Harris TB, Hattersley AT, Heath AC, Hengstenberg C, Hicks AA, Hindorff LA, Hingorani AD, Hofman A, Hovingh GK, Humphries SE, Hunt SC, Hypponen E, Jacobs KB, Jarvelin MR, Jousilahti P, Jula AM, Kaprio J, Kastelein JJ, Kayser M, Kee F, Keinanen-Kiukaanniemi SM, Kiemeney LA, Kooner JS, Kooperberg C, Koskinen S, Kovacs P, Kraja AT, Kumari M, Kuusisto J, Lakka TA, Langenberg C, Le Marchand L, Lehtimki T, Lupoli S, Madden PA, Mnnist S, Manunta P, Marette A, Matise TC, McKnight B, Meitinger T, Moll FL, Montgomery GW, Morris AD, Morris AP, Murray JC, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Ouwehand WH, Pasterkamp G, Peters A, Pramstaller PP, Price JF, Qi L, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ritchie M, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schwarz PE, Sebert S, Sever P, Shuldiner AR, Sinisalo J, Steinthorsdottir V, Stolk RP, Tardif JC, Tnjes A, Tremblay A, Tremoli E, Virtamo J, Vohl MC; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study, Amouyel P, Asselbergs FW, Assimes TL, Bochud M, Boehm BO, Boerwinkle E, Bottinger EP, Bouchard C, Cauchi S, Chambers JC, Chanock SJ, Cooper RS, de Bakker PI, Dedoussis G, Ferrucci L, Franks PW, Froguel P, Groop LC, Haiman CA, Hamsten A, Hayes MG, Hui J, Hunter DJ, Hveem K, Jukema JW, Kaplan RC, Kivimaki M, Kuh D, Laakso M, Liu Y, Martin NG, Mrz W, Melbye M, Moebus S, Munroe PB, Njlstad I, Oostra BA, Palmer CN, Pedersen NL, Perola M, Prusse L, Peters U, Powell JE, Power C, Quertermous T, Rauramaa R, Reinmaa E, Ridker PM, Rivadeneira F, Rotter JI, Saaristo TE, Saleheen D, Schlessinger D, Slagboom PE, Snieder H, Spector TD, Strauch K, Stumvoll M, Tuomilehto J, Uusitupa M, van der Harst P, Vlzke H, Walker M, Wareham NJ, Watkins H, Wichmann HE, Wilson JF, Zanen P, Deloukas P, Heid IM, Lindgren CM, Mohlke KL, Speliotes EK, Thorsteinsdottir U, Barroso I, Fox CS, North KE, Strachan DP, Beckmann JS, Berndt SI, Boehnke M, Borecki IB, McCarthy MI, Metspalu A, Stefansson K, Uitterlinden AG, van Duijn CM, Franke L, Willer CJ, Price AL, Lettre G, Loos RJ, Weedon MN, Ingelsson E, O'Connell JR, Abecasis

GR, Chasman DI, Goddard ME, Visscher PM, Hirschhorn JN, Frayling TM. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46, 11731186 (2014).

[45] Preissl S, Fang R, Huang H, Zhao Y, Raviram R, Gorkin DU, Zhang Y, Sos BC, Afzal V, Dickel DE, Kuan S, Visel A, Pennacchio LA, Zhang K, Ren B. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat Neurosci 21, 432-439, doi:10.1038/s41593-018-0079-3 (2018).

[46] ElAli A, Rivest S. Microglia in Alzheimer's disease: A multifaceted relationship. Brain Behav Immun 55, 138-150, doi:10.1016/j.bbi.2015.07.021 (2016).

[47] Dixon JR., Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. Mol Cell 62, 668-680, doi:10.1016/j.molcel.2016.05.018 (2016).

[48] ixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331-336, doi:10.1038/nature14222 (2015).

[49] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guig R, Hubbard TJ. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res, 22(9): 17601774, doi: 10.1101/gr.135350.111, (2012).

[50] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57-74. doi: 10.1038/nature11247 (2012).

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

[52] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16):2078-9. doi: 10.1093/bioinformatics/btp352, (2009).

[53] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137.

[54] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

[55] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 526, 6874 (2015).

[56] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7 (2015).

[57] Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nature Reviews Molecular Cell Biology, 16(3):144154, (2015).

[58] Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK. Effect of natural genetic variation on enhancer selection and function. Nature, 503(7477):487-92. doi: 10.1038/nature12615 (2013).

[59] Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, Majeti R, Chang HY. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature Genetics, 48(10):11931203, (2016).

[60] Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Chneby J, Kulkarni SR, Tan G, Baranasic D, Arenillas DJ, Sandelin A, Vandepoele K, Lenhard B, Ballester B, Wasserman WW, Parcy F, Mathelier A. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 46:D260D266, doi: 10.1093/nar/gkx1126 (2018).

[61] McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology 28, 495-501, doi:10.1038/nbt.1630 (2010).

[62] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, Fedor A Kolpakov, Vsevolod J Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res, 46(Database issue): D252D259, doi: 10.1093/nar/gkx1106 (2018).

[63] Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res, 43(Web Server issue): W39W49, doi: 10.1093/nar/gkv416 (2015).

[64] Sun Y, Nadal-Vicens M, Misono S, Lin MZ, Zubiaga A, Hua X, Fan G, Greenberg ME. Neurogenin Promotes Neurogenesis and Inhibits Glial Differentiation by Independent Mechanisms. Cell. 2001; 104(3):365-376, doi: 10.1016/S0092-8674(01)00224-0.

[65] Chakrabarti L, Best TK, Cramer NP, Carney RS, Isaac JT, Galdzicki Z, Haydar TF. Olig1 and Olig2 triplication causes developmental brain defects in Down syndrome. Nature Neuroscience volume 13, pages 927934 (2010).

[66] Ligon KL, Fancy SP, Franklin RJ, Rowitch DH. Olig gene function in CNS development and disease. Glia. 2006 Jul;54(1):1-10. DOI: 10.1002/glia.20273.

[67] Chun K, Teebi AS, Jung JH, Kennedy S, Laframboise R, Meschino WS, Nakabayashi K, Scherer SW, Ray PN, Teshima I. Genetic analysis of patients with the SaethreChotzen phenotype. Medical Genetics. 2002; 110(1): 136-143. doi: 10.1002/ajmg.10400.

[68] Johnson D, Horsley SW, Moloney DM, Oldridge M, Twigg SR, Walsh S, Barrow M, Njlstad PR, Kunz J, Ashworth GJ, Wall SA, Kearney L, Wilkie AO. A Comprehensive Screen for TWIST Mutations in Patients with Craniosynostosis Identifies a New Microdeletion Syndrome of Chromosome Band 7p21.1. AJHG 63(5): 1282-1293. (1998)

[69] Fior R, Henrique D. A novel hes5/hes6 circuitry of negative regulation controls Notch activity during neurogenesis. Developmental Biology 281(2), 318-333. doi:10.1016/j.ydbio.2005.03.017 (2005).

[70] Meerschaut I, Rochefort D, Revenu N, Ptre J, Corsello C, Rouleau GA, Hamdan FF, Michaud JL, Morton J, Radley J, Ragge N, Garca-Miar S, Lapunzina P, Bralo MP, Mori M, Moortgat S, Benoit V, Mary S, Bockaert N, Oostra A, Vanakker O, Velinov M, de Ravel TJ, Mekahli D, Sebat J, Vaux KK, DiDonato N, Hanson-Kahn AK, Hudgins L, Dallapiccola B, Novelli A, Tarani L, Andrieux J, Parker MJ, Neas K, Ceulemans B, Schoonjans AS, Prchalova D, Havlovicova M, Hancarova M, Budisteanu M, Dheedene A, Menten B, Dion PA, Lederer D, Callewaert B. FOXP1-related intellectual disability syndrome: a recognisable entity. J Med Genetic, 54(9):613-623, doi: 10.1136/jmedgenet-2017-104579 (2017).

[71] Jay K, Mitra A, Harding T, Matthes D, Van Ness B. Identification of a de novo FOXP1 mutation and incidental discovery of inherited genetic variants contributing to a case of autism spectrum disorder and epilepsy. Mol Genet Genomic Med, 20:e751, doi:10.1002/mgg3.751 (2019).

[72] Ayhan F, Konopka G. Regulatory genes and pathways disrupted in autism spectrum disorders. Prog Neuropsychopharmacol Biol Psychiatry, 89:57-64, doi: 10.1016/j.pnpbp.2018.08.017 (2019)

[73] Volk DW, Lewis DA. Cortical inhibitory neuron disturbances in schizophrenia: role of the ontogenetic transcription factor Lhx6. Schizophr Bull, 40(5):1053-61, 10.1093/schbul/sbu068 (2014).

[74] Volk DW, Chitrapu A, Edelson JR, Lewis DA. Chemokine receptors and cortical interneuron dysfunction in schizophrenia. Schizophr Res, 167(1-3):12-7 doi: 10.1016/j.schres.2014 (2015).

[75] Wei J, Joshi S, Speransky S, Crowley C, Jayathilaka N, Lei X, Wu Y, Gai D, Jain S, Hoosien M, Gao Y, Chen L, Bishopric NH. Reversal of pathological cardiac hypertrophy via the MEF2-coregulator interface. JCI Insight. 2(17). pii: 91068. doi: 10.1172/jci.insight.91068 (2017)

[76] Desjardins CA, Naya FJ. The Function of the MEF2 Family of Transcription Factors in Cardiac Development, Cardiogenomics, and Direct Reprogramming. J Cardiovasc Dev Dis, 3(3). pii: 26, doi:10.3390/jcdd3030026 (2016).

[77] Morin S, Charron F, Robitaille L, Nemer M. GATA-dependent recruitment of MEF2 proteins to target promoters. EMBO J, 19(9):2046-55, doi: 10.1093/emboj/19.9.2046 (2000). Rana Rana MS, Christoffels VM, Moorman AF. A molecular and genetic outline of cardiac morphogenesis. Acta Physiol (Oxf), 207(4):588-615, doi: 10.1111/apha.12061 (2013).

[78] Noonan JP, McCallion AS. Genomics of long-range regulatory elements. Annual review of genomics and human genetics 11, 1-23, doi:10.1146/annurev-genom-082509-141651 (2010).

[79] Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Research, 20(5):565577, (2010).

[80] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell, 38(4):57689, (2010).

[81] Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. Genes Dev, 28(24):26792692, (2014).

[82] Heinz S, Glass CK. Roles of lineage-determining transcription factors in establishing open chromatin: lessons from high-throughput studies. Curr Top Microbiol Immunol, 498(7455):511-5. doi: 10.1038/nature12209 (2013).

[83] Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics, 32(14): 22052207, doi: 10.1093/bioinformatics/btw203, (2016).

[84] Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. A method to predict the impact of regulatory variants from DNA sequence. Nature Genetics volume 47, pages 955961 (2015).

[85] Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proceedings of the National Academy of Sciences, 113(23):65086513, (2016).

[86] J Tao. Machine learning approaches for relating genomic sequence to enhancer activity and function. UCSD Library, (2019).

[87] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. RFECS: A Random-Forest Based Algorithm for Enhancer Identification

from Chromatin State. PLOS Computational Biology, doi: 10.1371/journal.pcbi.1002968, (2013).

[88] Kelley D, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7):9909, (2016).

[89] Kelley D, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research, 28: 739-750, doi: 10.1101/gr.227819.117, (2018).

[90] OpenAI. OpenAI Five. Accessed: https://blog.openai.com/openai-five/ (2018).

[91] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L. Attention is all you need. arXiv, arXiv:1706.03762v5, (2017).

[92] Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A comparison of word embeddings for the biomedical natural language processing. J Bomed Inform, 87:12-20, doi: 10.1016/j.jbi.2018.09.008, (2018).

[93] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv, arXiv:1812.04948v3, (2019).