Dissecting cellular states and cell state transitions through integrative analysis of epigenetic dynamics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. Dipl.-Phys. Michael Johannes Ziller
aus Hamm

Tübingen
2014

Tag der mündlichen Qualifikation:	26.09.14
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Alexander Meissner
3. Berichterstatter:	Prof. Dr. Martin Vingron

Für Anne-Marie

“To know is to know something by causes.”

Aristotle

“Ich habe viel Mühe, ich bereite meinen nächsten Irrtum vor.”

Berthold Brecht

Abstract

Understanding how a single genome that is common to all cells in an organism can give rise to many different and highly specialized, cell types has been one of the major questions in biology over the past century and still many aspects remain unanswered. Over the last 15 years, incredible progress has been made in pinpointing the regulatory mechanisms that establish, maintain, and change cellular identities. In particular, the role of histone modifications and DNA methylation in the spatio-temporal control of gene expression and genome organization has been greatly appreciated. These histone and DNA modifications have been shown to be an integral part of *epigenetic* control mechanisms. They ensure stable silencing of not-required genes and gene regulatory elements as well as maintenance of active genes and gene regulatory elements that are required in a particular cellular context.

In addition to the identification and functional characterization of these mechanisms, the sequencing of many complex genomes and the advent of high-throughput sequencing technology has allowed us to precisely chart the location of all modified histones and methylated bases across the entire genome. In contrast to the genome sequence, the epigenome turns out to be highly variable between distinct cell types and reflective of their specific biology. Comprehensive mapping efforts of the epigenome provide a starting point for understanding the epigenetic basis of macroscopic phenotypes such as distinct cell types and states. Integrative analysis of many different types of histone modifications combined with gene expression data across several distinct cell types also revealed that certain histone modifications can be used to annotate and predict the activity of different types of gene regulatory elements such as promoters or enhancers.

This thesis takes advantage of these recent advances in the identifiability of gene regulatory elements and first establishes that the integration of epigenetic and transcriptional data on specific cellular states can be used to gain insights into the underlying regulatory logic maintaining and establishing these states. Second, it demonstrates the utility of this approach to generate experimentally testable hypothesis on the molecular mechanisms mediating specific cell state transitions.

While great progress has been made on mapping changes in histone modifications and identification of their demarcated gene regulatory elements, less is known about methylation of the DNA. In particular, it is still unclear to what extend and where DNA methylation changes over the course of human development and what the likely functions of these changes are. To fill this gap, we mapped DNA methylation patterns using whole-genome bisulfite sequencing across 30 distinct human cell and tissue types. Surprisingly, we find that only around 20% of all CpG dinucleotides, the main target of DNA methylation in mammals, change their methylation state across normal development. Interestingly, we find that most differentially methylated regions (DMRs) coincide

with gene regulatory elements, such as enhancers or transcription factor binding sites, that are relevant for the biology of a particular cell type. Most of these DMRs are constitutively hyper-methylated and only become hypo-methylated in a cell type where the underlying gene regulatory element becomes relevant. This study not only determined the extent to which DNA methylation is dynamic during normal development, but also established DNA methylation dynamics as an excellent marker of active, cell type-specific gene regulatory elements.

The first part of this thesis examined primarily differences in the DNA methylation landscape of distantly related stable steady state cell types and states. This experimental design is similar to the strategy chosen by most previous investigations of histone modification changes. However, much less is known about the epigenetic changes at high temporal resolution during cell fate transitions. To shed more light on the epigenetic remodeling events that establish and distinguish closely related cell types, we investigated the epigenetic and transcriptional dynamics during the *in vitro* differentiation of human embryonic stem cells to three distinct early-derived populations of each embryonic germ layer. This analysis revealed comprehensive epigenetic remodeling, particularly at enhancers and frequently converse dynamics between different germ layer populations. In addition, we identified an unexpected, epigenetic transition where a repressed locus first transiently gains H3K27me3 at an early stage of differentiation before becoming fully activated at a later time point.

While these experiments and analyses were very informative from an observational point of view, it still is unclear in many cases what are the key drivers of the cell state transitions and in particular of the epigenetic remodeling events. To shed light on this important problem, we developed a computational analysis strategy that leverages epigenetic information to identify key transcription factors (TFs) that are likely to mediate epigenetic changes and orchestrate particular cell state transitions. We then applied this approach to a 200-day time series of a novel, *in vitro* differentiation scheme of human embryonic stem cells towards the neural lineage. This analysis revealed a core network of transcription factors around PAX6 and OTX2 that is active during the entire process of *in vitro* neural differentiation. Furthermore, these core factors show evidence of differentiation stage-specific co-binding with other TFs as well as binding site redistribution across the genome. These results highlight the utility and general applicability of this framework for integrated epigenetic analysis. In particular, this approach provides a TF-centric perspective on the interpretation of epigenetic changes during cell state transitions.

In summary, this work has contributed to our understanding of the dynamic regulation of DNA methylation and provided a high-resolution, in-depth investigation of epigenetic dynamics during differentiation of embryonic stem cells to three distinct, differentiated cell populations. Furthermore, we provided a detailed view on the regulatory network

activity and architecture orchestrating *in vitro* neural differentiation. In addition, this work yielded novel tools and perspectives to interpret the rapidly rising number of epigenetic profiles.

Kurzfassung

Wie kann ein einziges Genom, das in nahezu allen Zellen eines Organismus gleich ist, eine derartige Vielfalt an hochgeradig spezialiserten Zelltypen hervorbringen wie sie in komplexen Organismen zu finden sind?

Obwohl diese Frage einen der Forschungsschwerpunkte der Biologie der letzten 100 Jahre darstellt, sind viele Aspekte nach wie vor ungeklärt. In den vergangenen 15 Jahren hat die Aufklärung der regulatorischen Mechanismen, die zur Generierung, Erhaltung und Veränderung verschiedener zellulärer Identitäten beitragen, enorme Fortschritte gemacht. Diese neuen Erkenntnisse untermauren die zentrale Rolle der räumlich-zeitlichen Kontrolle von Histonmodifikations und DNA-Methylierungsmustern entlang des Genoms. Histon- und DNA-Modifikationen bilden dabei eine entscheidende Komponente der epigenetischen Kontrollmechanismen der Genexpression und der Organisation des Genoms. Auf der einen Seite stellen sie die stabile Repression in einem spezifischen zellulären Kontext von nicht benötigten Genen und genregulatorischen Elementen sicher. Auf der anderen Seite sind diese Mechanismen wesentlich an der Erhaltung und Stabilisierung des aktiven Zustands von zell-relevanten Genen und genregulatorischen Elementen beteiligt. Neben der Entdeckung und Charakterisierung dieser Mechanismen hat vor allem die Sequenzierung zahlreicher komplexer Genome sowie die Einführung von Hochdurchsatz-Sequenzierungsverfahren es ermöglicht, die exakte genomische Position modifizierter und methylierter Basen zu kartieren. Im Gegensatz zur statischen Genomsequenz, variiert das Epigenom stark zwischen verschiedenen Zelltypen und Zellzuständen. Die umfangreichen Bemühungen zur Kartierung zahlreicher verschiedener Zelltypen liefern heute einen Startpunkt, um die epigenetische Basis der verschiedenen makroskopischen zellulären Phänotypen zu verstehen. Durch die kombinierte Analyse unterschiedlicher Histonmodifikationen sowie Genexpressionsmessungen ist es des Weiteren möglich, die Position und den Aktivitätszustand von genregulatorischen Elementen wie Promotoren und Enhancern vorherzusagen.

Diese Arbeit baut auf diesen, erst vor kurzem erzielten Fortschritten in der Interpretation des Genoms auf und demonstriert zunächst, dass die Integration von Epigenom- und Transkriptomdaten von eng verwandten zellulären Zuständen es gestattet, die den Zellzuständen und Zellzustandsübergängen zugrundeliegenden Mechanismen besser zu verstehen. Im nächsten Schritt der Arbeit wird dann aufgezeigt, wie sich ein derartiger integrativer Ansatz zur Generierung spezifischer, experimentell direkt überprüfbarer Hypothesen über die molekularen Mechanismen nutzen lässt, die den Zellzustandsübergängen zugrunde liegen .

Als Ausgangspunkt für dieses Unterfangen dienen Karten verschiedener Histonmodifikations- und DNA- Methylierungsmuster in diversen Zelltypen. Die umfangreichen Anstrengungen zur Kartierung von Histonmodifikationen finden jedoch keine Entsprechung

im DNA-Methylierungsfeld. Hier ist es vor allem unklar, wie stark sich die DNA-Methylierungsmuster im Zuge der menschlichen Organismusentwicklung in verschiedenen Zelltypen verändern. Weiterhin ist nicht klar, wo genau im Genom diese Veränderungen und welche Funktionen sie haben. Um diese Lücke zu schließen, haben wir die DNA- Methylierungsmuster in 30 verschiedene menschliche Zell- und Gewebeklassen genomweit kartiert. Überraschenderweise ändern lediglich etwa 20% aller CpG-Dinukleotide - das Hauptsubstrat von DNA-Methylierung in Säugetieren - ihren Methylierungszustand im Zuge der normalen Organismusentwicklung. Dabei stellt sich heraus, dass der Großteil der differentiell methylierten Regionen (DMRs) mit genregulatorischen Elementen wie Enhancern und Transkriptionsfaktor-Bindestellen überlappen, die für die Biologie des jeweiligen Zelltyps relevant sind.

Die grosse Mehrheit dieser DMRs ist dabei konstitutiv hypermethyliert und wird nur in dem Zelltyp hypomethyliert, in dem das unterliegende generegulatorische Element von Bedeutung ist. Diese Untersuchung hat nicht nur das Ausmaß der DNA-Methylierungsveränderungen während der menschlichen Entwicklung ermittelt, sondern auch demonstriert, dass DNA-Methylierungsveränderungen ein exzellenter Marker für aktive, zelltypspezifische generegulatorische Element sind.

Im vorangegangen Teil dieser Arbeit wurden primär Veränderungen von DNA-Methylierungsmustern über entfernt verwandte Zelltypen untersucht. Diese Art von experimentellem Design folgt dabei der Strategie der meisten vorausgegangen Studien zu Veränderungen in Histonmodifikationsmustern. Im Gegensatz zu den umfangreichen Kenntnissen zu epigenetischen Unterschieden zwischen statischen, entfernt verwandten und zeitlich stabilen Zellpopulationen, ist über die Natur der epigenetischen Dynamiken bei hoher zeitlicher Auflösung und eng verwandten, transienten Zelltypen deutlich weniger bekannt.

Um die epigenetischen Veränderungen die zur Etablierung neuer, eng verwandter Zelltypen genauer zu verstehen, haben wir daher die epigenetischen und transkriptionellen Veränderungen während der *in vitro* Differenzierung von menschlichen-, embryonalen Stammzellen zu drei verschiedenen frühen Zellpulationen der embryonalen Keimblätter, Endoderm, Mesoderm und Ectoderm, untersucht. Im Rahmen dieser Studie konnten wir umfangreiche epigenetische Remodellierungsprozesse, vor allem in Enhancern, beobachten. Weiterhin haben wir festgestellt, dass die epigenetischen Veränderungen an spezifischen genregulatorischen Elementen vielfach komplementär zwischen den verschiedenen Zellpopulationen ablaufen. Neben diesen generellen Trends konnten wir eine unerwartete Abfolge epigenetischer Veränderungen identifizieren, in der ein abgeschalteter genomischer Locus zunächst temporär mit H3K27me3 markiert wird, bevor dieser Locus in einem späteren Differenzierungsschritt H3K27me3 wieder verliert und aktiviert wird.

Obwohl diese Experimente äusserst informativ hinsichtlich der Beschreibung und Charakterisierung epigenetischer Veränderungen im Zuge von Zellzustandsübergängen sind, so

sind die treibenden Kräfte, welche diese Veränderungen auslösen und dirigieren, vielfach unklar. Um diesen Aspekt genauer zu beleuchten, haben wir eine computergestützte Methode entwickelt, die es gestattet, aus den genomweiten Verteilungen epigenetischer Modifikationen in verschiedenen Zelltypen die Transkriptionsfaktoren zu identifizieren, die mit hoher Wahrscheinlichkeit an der Organisation der epigenetischen Veränderungen beteiligt sind oder diese dirigieren. Im nächsten Schritt haben wir dann dieses Verfahren zur Analyse einer 200-tägigen *in vitro* Differenzierung humaner-, embryonaler Stammzellen in verschiedene neurale Zelltypen genutzt. Im Rahmen dieser Analyse ist es uns gelungen, ein Kernnetzwerk von Transkriptionsfaktoren wie OTX2 und PAX6 zu identifizieren, das während des gesamten Prozesses der *in vitro* Differenzierung aktiv ist. Weiterhin scheinen diese Kernfaktoren ihre Bindungspartnerfaktoren in Abhängigkeit zur Differenzierungsstufe zu verändern sowie eine allgemeine Relokalisation entlang des Genoms zu durchlaufen. Diese Resultate sind nicht nur biologisch interessant, sondern untermauern auch die Nützlichkeit und generelle Anwendbarkeit des präsentierten Analyseverfahrens.

Zusammenfassend hat dise Arbeit zum besseren Verständnis der dynamsichen Regulation von DNA-Methylierung beigetragen. Des Weiteren wurde eine hochauflösende, detaillierte Untersuchung von epigenetischen Veränderungen während der Differenzierung von embryonalen Stammzellen in drei verschiedenen, differenzierten Zellpopulationen der drei Keimblätter präsentiert. Schließlich wurde eine detaillierte Untersuchung der Aktivität und Architektur des genregulatorischen Netzwerks der neuralen Differenzierung dargelegt. Zusätzlich wurden im Rahmen dieser Arbeit verschiedene neue Verfahren vorgestellt, die zur Interpretation der rasant wachsenden Zahl epigenetischer Daten verwendet werden können.

Acknowledgements

Foremost, I thank my wife Anne-Marie for her continuous support during all these years and her willingness to leave Germany and join me in the U.S. Without her, I would hardly have made it this far. Thank you very much Anne for always being there for me, for tolerating my countless work-weekends and being with me in the lab until 5am when we again had to submit something last minute.

Also, I thank my parents for supporting me all this time, ideally and monetarily.

I am very grateful to my advisor and mentor Alex Meissner: I learned so much - not only scientifically, but also approaching and thinking about scientific problems. He is not only a great scientist, but also an awesome mentor. Thank you very much Alex for giving me a chance when I came to Harvard in 2009 and providing me with so much freedom and support to pursue my interest and develop my own scientific standing.

I am also greatly indebted to my other advisor Oliver Kohlbacher, who supported me with great advice and insights into the numerous computational problems I encountered during my thesis work. Thank you, Oliver for always helping me to move forward when I was stuck. Also, thank you very much for taking on the challenge of advising me remotely and being always available.

I am also very thankful for the support of Yechiel Elkabetz who taught me so much about neural development.

Last but not least I thank the current of former members of the Meissner lab for creating such a rich and stimulating environment. In particular, I thank Christoph Bock, Hongcang Gu, Fabian Müller, Julie Donaghey and Casey A. Gifford.

Contents

List of Figures

Abbreviations

CGI	CpG island
ChIP-Seq	Chromatin immunoprecipitation coupled with high-throuput sequencing
CpG	Cytosine dinucleotide
DE	Differential expression
dEC	Human embryonic stem cell (HUES64) derived ectoderm
dEN	Human embryonic stem cell (HUES64) derived endoderm
dME	Human embryonic stem cell (HUES64) derived mesoderm
DMR	Differentialy methylated region
DNAme	DNA methylation
DNAse HS	DNAse hypersensitive
ENCODE	The Encyclopedia of DNA Elements
E-RG	Early-radial glial
E-RGdN	Early radial glial derived neurons
FACS	Fluorescence activated cell sorting
FPKM	Fragments per kilobase per million reads sequenced
FDR	False discovery rate
FPR	False positive rate
FP	Footprint
GEV	Generalized extreme value
GO	Gene ontology
GRE	Gene regulatory element
GREAT	Genomic regions enrichment of annotations tool
GWAS	Genome-wide association study
hESC	Human embryonic stem cells
HUES64	Harvard University human embryonic stem cell line 64

H9	Human embryonic stem cell line H9
H3K4me1	Histone 3 lysine 4 tri-methylation
H3K4me2	Histone 3 lysine 4 di-methylation
H3K4me3	Histone 3 lysine 4 mono-methylation
H3K9me3	Histone 3 lysine 9 trimethylation
H3K27ac	Histone 3 lysine 27 trimethylation
H3K27me3	Histone 3 lysine 27 acetylation
HMM	Hidden Markov model
HMR	Highly methylated region
IDR	Irreproducible discovery rate
IMR	Intermediate methylated region
kb	Kilobase
LMR	Lowly methylated region
LNP	Late nerual precursor
L-RG	Late radial glial
L-RGdN	Late-radial glial derived neurons
L-RGdA	Late-radial glial derived astrocytes
L-RGdO	Late-radial glial derived oligodendrocytes
MGI	Mouse genome informatics
M-RG	Mid-radial glial
NE	Neuroectoderm
NEdN	Neuroectoderm derived neurons
NPC	Neural precursor cell
PLS	Partial least square
PWM	Position weight matrix
REMC	Epigenome Roadmap Projcet
RPKM	Reads per kilobase per million reads sequenced
RNA-Seq	RNA sequencing
RRBS	Reduced representation bisulfite sequencing
SD	Standard deviation
SNP	Short nucleotide polymorphism
TERA	Transcription factor epigenetic remodeling activity
TF	Transcription factor

TFBS	Transcription factor binding site
TPR	True positive rate
TSS	Transcription start site
UMR	Unmethylated region
WCE	Whole cell extract
WGBS	Whole genome bisulfite sequencing

Chapter 1

Introduction

One of the oldest questions in biology explores how living beings are created and develop. The first written attempt to address this question dates back to Aristotle who investigated this matter by collecting time course data on the development of chicken embryos (Aristotle and Lawson-Tancred, 1998, Gilbert, 2006). His observations gave rise to the theory of epigenesis (Aristotle and Peck, 1943), stating that organisms develop from seeds or eggs in a sequence of steps producing new structures and organs. This strategy that became the heart of the anatomical approach to developmental biology and dominated the perspective of the field for almost 2,000 years focused on observing morphological changes of the developing embryo and its structures (Gilbert, 2006). Among the key observations of this early era was the discovery and characterization of the three germ layers: ectoderm, mesoderm, endoderm. These three regions of the embryo are formed during gastrulation of almost all embryos when the single-layered blastula is re-organized into a three-layered structure. They give rise to distinct structures in the developed organisms such as the brain and spinal cord (ectoderm), liver and kidney (endoderm) and blood and muscle (mesoderm) (Gilbert, 2006). However, how these differentiation processes were functioning in an operational sense could not be addressed and was only at the periphery of the research agenda. This rapidly changed with the advent of the cell theory in the beginning of 19th century (Serafini, 1993) and a new perspective on organism development emerged: from simply observing and reconstructing the distinct steps of embryo development towards understanding the causal relationships and driving forces orchestrating the developmental process. One of the central goals of this research agenda was to uncover the "Entwicklungsmechanik" (causal embryology) of the developing embryo: to understand how the development of specialized cell types from a single cell occurs; or, in short, how differentiation works (Gilbert, 2006). To achieve these goals, this new field started to rely heavily on experimental approaches to functionally characterize distinct parts of a developing embryo.

One of the early key observations stated that cells do not change their morphology and functional properties instantaneously, but rather that differentiation happens gradually, starting with the commitment of a cell to a certain fate (Gilbert, 2006). The process of cell differentiation can be split up into two phases: specification and commitment (Harrison, 1933). While cells that are specified will develop into their respective fate when placed in a neutral environment devoid of differentiation-inducing cues, they remain capable of reverting back to their original unspecified state when confronted with proper environmental cues. When the latter property is lost and the cells will differentiate towards their specified state regardless of environmental signals, the cells are considered to be committed to a certain cell fate (Gilbert, 2006).

Subsequently, novel experimental approaches to characterize the underlying mechanics of this process led to the discovery of the organizing center of gastrulation by Spemann (Spemann and Mangold, 2001). At the same time, the field of embryology was vividly debating the question of what part of the egg - the cytoplasm or the nucleus - was responsible for the transmission of inheritable information (Gilbert, 1991). However, even after the rediscovery of Mendel's theory of inheritance and growing support for the existence of genes (Morgan, 1909), the field of causal embryology had little use for the gene theory. Instead, genetics evolved as a separate discipline (Gilbert, 1991), relying more and more on statistical approaches to understand inheritance (Fisher, 1930). In fact, the predominant opinion within the field of embryology stated that genetics was solely concerned with transmission of inherited traits, whereas embryology was focused on studying the expression of these traits (Gilbert, 1991, Morgan, 1926). This view was particularly prominent since according to the major embryologists of the time such as Spemann and Just, genetics failed to explain how presumably identical chromosomes (gene sets) within each cell could give rise to many different cell types and to provide evidence that genes are relevant for the early stages of embryo development (Gilbert, 2006). More drastically, many people believed that genes were only relevant for the adult organism (Gilbert, 2006). The latter question was addressed in the later 1930s by Salmon Gluecksohn-Schoenheimer and Conrad Waddington, who identified mutations in genes that affect early mouse development such as the Brachury gene - the genetic basis of the organizer - and genes causing malformations in the wings of the fruit fly (Gilbert, 1991, Gluecksohn-Schoenheimer, 1938, 1940, Waddington, 1940).

With proof of the genetic basis of embryonic development, one major obstacle for a grounding of embryology in gene theory remained: How is it possible that nuclear genes direct development when they are presumably all the same in every cell type (Gilbert, 2006)? From a locigal point of view it seemed possible that different cell types have different sets of genes and genomes. However, this hypothesis did not fit well with the fact that all cells arise from a single egg. This question of genomic equivalence was answered starting in the 1950s with somatic nuclear transfer experiments by Briggs and King (Briggs and King, 1952, Gilbert, 2006, King and Briggs, 1956). These early experiments showed that cell nuclei from blastula could direct the development of entire tadpoles when transferred into an activated enucleated frog egg. Despite these successes, Briggs and King were unable to reproduce these results with further developed somatic cells (Gilbert, 2006). It was John Gurdon and colleagues, who were able to demonstrate over a series of nuclear transplantation experiments in the 1960s and 1970s that a single somatic cell nucleus is capable of giving rise to all cell types of a young Xenopus (Gilbert, 2006, Gurdon et al., 1975, Wabl et al., 1975).

The final proof of genomic equivalence was achieved with the cloning of a mammalian organism by Ian Wilmut in 1997, creating the sheep Dolly (Wilmut et al., 1997). However, even with the problem of genomic equivalence resolved, the question of how exactly the same genome can give rise to many different cell types remained, leading to the formulation of the differential gene expression paradigm (Gilbert, 2006). This paradigm states that over the course of development as well as in fully differentiated cell types distinct combinations of genes are active or repressed at various levels, producing different sets and concentrations of proteins and regulatory RNA (Gilbert, 2006). On this background, all different cell types and physiological cell states can be interpreted in a unified framework as distinct cellular states (Huang et al., 2005), corresponding to distinct points in a high dimensional gene/RNA expression space. In the post genomic digital era, these distinct gene activity profiles are often referred to as molecular programs.

The characterization of cell types and physiological cell states through global profiling approaches such as entire transcriptome measurements starts to introduce a paradigm shift in the definition of cell types/states, extending and complementing morphological and functional approaches towards a molecular definition.

However, the molecular mechanisms of establishment and maintenance of distinct gene expression programs - despite having the same genomic background - are still an area of active investigation. The molecular programs are established and controlled by a variety of mechanisms discussed below.

1.1 Outline and key questions

Great progress has been made in understanding the molecular mechanisms underlying maintenance and changes of cellular identity. However, even though many general molecular principles of cell state changes have been uncovered, numerous open questions about **(1) the involvement of specific epigenetic marks such as DNA methylation, (2) the integration and coordination of transcriptional and epigenetic changes during cellular differentiation, and (3) the structure of the underlying transcription factor control networks orchestrating cell state maintenance or differentiation remain elusive.** The paradigm proposed in this work states that a causal understanding of cell state maintenance and dynamics can be achieved through a transcription factor-centric perspective, explaining epigenetic and (resulting) transcriptional changes through the differential binding of transcription factors to specific gene regulatory elements. This paradigm, in turn, forms the basis of this work's key hypothesis that the distribution of epigenetic states across DNA together with transcriptional information allows for inferences of cell-type-specific active regulatory networks and pathways, ultimately permitting the reconstruction of the regulatory logic mediating cell state maintenance and transitions.

The major goal of this work is therefore twofold: **First, establish that integration of epigenetic state and transcriptional information on specific cellular states can be used to understand the underlying regulatory logic mediating cell state transitions and second to apply this rationale to obtain a mechanistic understanding of specific cell state transitions and pinpointing the underlying regulatory networks of gene regulatory elements and transcription factors.**

The ultimate objective of this endeavor is to generate quantitative, predictive regulatory models of cellular states, integrating genetic, epigenetic, transcriptional, post-transcriptional, environmental and phenotypic information to explain differentiation into other cell types and the causes and consequences of deviations from the normal cell state trajectory, leading to diseases such as cancer.

To move closer to this goal it is crucial to **(1)** understand which epigenetic signatures can be used to gain insights into the functional characteristics (e.g. differentiation potential) and active regulatory networks of a cellular state. Several combinations of histone modifications such as H3K4me3, H3K4me2, H3K4me1, H3K27me3 and H3K27ac have been shown to be informative to identify key cell type specific regulatory transcription factors in distantly related cell types (Ernst and Kellis, 2010, Ernst et al., 2011, Heintzman et al., 2009, Maston et al., 2006) and are associated with differentiation capacity (Bernstein et al., 2006, Rada-Iglesias et al., 2011). However, less is known about the utility of DNA methylation in this context, the most widespread epigenetic modification. At this point it is unclear to what extent DNA methylation is dynamic over the course of development and disease, given that the vast majority of the genome is highly methylated. Furthermore, it remains open whether DNA methylation changes are in general associated with regulatory function. In particular, it is unclear whether or not DNA methylation can be used as a suitable epigenetic mark to gain insights into the active regulatory networks of a particular cellular state. These questions are addressed in **Chapter 4**. In the presented work we determine to what extend the DNA methylation landscape of the human genome changes between normal cell types and how these changes compare to DNA methylation dynamics observed in diseased cell types. Furthermore, this work establishes DNA methylation changes as an excellent marker of gene regulatory elements and demonstrates that genome wide methylation changes can be used to infer key cell type-specific transcription factors. To obtain insights into the cell type-specific regulatory networks, the new concept of DNA methylation footprints of transcription factor binding is introduced. In addition, this work highlights the utility of cell type specific DNA methylation signatures derived from a large cohort of samples with respect to classifying unknown samples and deconvoluting DNA methylation measurements of heterogeneous populations.

To understand these matters, it is essential to robustly determine differences in DNA methylation levels between two conditions given next generation sequencing (NGS) measurements. While most previous methods relied on off-the-shelf statistical methods such as Fisher's exact test or Student's t-test, this work puts forward a novel method for the detection of differentially methylated regions in bisulfite sequencing data, which is presented within the Methods section in **Chapter 4**. This method is specifically tailored to the statistical problems arising in single-base pair bisulfite sequencing datasets. The method overcomes several limitations of previous approaches by providing a statistical framework to integrate replicates, capturing not only variation occuring due to sampling but also due to normal, biological variation. The method therefore represents the underlying random process more accurately than for example the assumptions made by the Fisher's exact test. In addition, the novel DMR finder allows to determine confidence intervals on differential methylation levels between conditions and allows to integrate individual methylation measurements on a genomic region level.

In order to mechanistically dissect cell state transitions and determine key epigenetic remodeling events, it is important to **(2)** investigate cell state dynamics at a high temporal resolution. However, all previous studies, including our own in Chapter 4 focused on collections of distantly related cell types, frequently with different genetic background and not developmentally related.

To overcome these limitations, we devised an *in vitro* model system of differentiation of human embryonic stem cells into progenitor cells from the three germ layers, ectoderm, mesoderm, and endoderm. Using this unique model system, we are able to interrogate the mechanisms of three distinct cell fate decisions from the same starting point and perform a horizontal comparative analysis, focusing on dynamics in transcription, chromatin modifications, DNA methylation and transcription factor binding in **Chapter 5**. This study design for the first time allowed for a high-resolution view of epigenetic changes during cell state transitions. Furthermore, this high-resolution view allowed us to identify transient epigenetic changes characterized by a gain of the repressive mark H3K27me3 at regions of high DNA methylation, coinciding with a loss of DNA methylation and binding of the pioneering transcription factor FOXA2. Interestingly, many of the associated regions were located within open chromatin or localized in close proximity of gnees expressed at later stages of development, suggesting a multi-stage de-repression of the associated loci.

In addition, this study demonstrated that the majority of epigenetic changes during *in vitro* differentiation of hESC to these early germ layer populations were not associated with transcriptional changes and rather seemed to be related to the shutdown of alternative lineage programs or the absence of pluripotency factor binding. Finally, the study also emphasizes the usefulness of different epigenetic signatures such as H3K4me1 and H3K27ac which mark lineage specific gene regulatory elements and are highly dynamic during differentiation.

Given the encouraging findings from this study, we wanted **(3)** to expand on the possibility to dissect cell state transitions and determine key regulators and active regulatory modules from transcriptional and epigenetic data. To that end we took advantage of a novel differentiation paradigm of hESCs to the neural lineage giving rise to a time-course of six consecutive stages of *in vitro* neural differentiation. We devised a new computational methodology to understand the molecular basis of this process and interpret the transcriptional and epigenetic dynamics employing a partial least square approach. The methodology and insights into general principles of cell state transitions as well as the specifics of the *in vitro* neural differentiation are presented in **Chapter 6**. This study together establishes a novel platform to not only study neural development *in vitro* and dissect the underlying molecular mechanisms and gene regulatory networks of various stages of neural development. It is the unique contribution of this thesis to provide a framework to utilize the resulting complex, multidimensional datasets and extract the pan-neural and stage-specific regulatory networks. The proposed computational framework is of general applicability and provides a new perspective and approch to utilize and integrate epigenomic datasets. It can be used to dissect other differentiation trajectories or a collection of cellular steady states.

In particular, the presented study uncovers the underlying molecular driving force of the continous remodeling of the differentiation potential of the distinct stages of neural precursor development by linking it to specific, dynamic regulatory networks and predicted differential binding of key transcription factors. Furthermore the analysis suggests that a stably expressed core transcription factor network comprising PAX6 and OTX2 cooperates with stage specific factors and signaling pathways to regulate commitment and proper differentiation towards neuronal and glial cell types.

Finally, concluding remarks and next steps for further research are presented in **Chapter 7**.

Chapter 2

Background

2.1 Chromatin and chromatin modifications

Double strand DNA itself does not exist as a long, solitary molecule in the nucleus but is rather wrapped around the nucleosome core particle, a protein octamer consisting of four core histone proteins (H3, H4, H2A and H2B) (Zhou et al., 2011). This basic organizational unit of the chromatin comprises 147 bp of DNA and gives rise to several higher order structures. The formation of these structures is based on the compaction of "beads on a string" through physical linkage of the individual nucleosomes by the linker histone H1 (Zhou et al., 2011). The resulting condensed fibers achieve a 50-fold higher compaction level (Bell et al., 2011), that form the basis of large blocks of tightly packed heterochromatic regions of the genome (Zhou et al., 2011).

The nucleosome packaging density greatly influences the DNA's accessibility to protein binding, in particular to transcription factors and the transcriptional machinery. Therefore, modulating the local DNA accessibility harbors great regulatory potential to influence gene expression and silencing as well as DNA replication and repair (Bell et al., 2011). Studies over the last 15 years have shown that local chromatin structure is indeed highly dynamic and varies from cell type to cell type (Lee et al., 2004, Mikkelsen et al., 2007) thus providing the molecular basis for the differential expression paradigm. Multiple factors determine the level of DNA accessibility including sequence context, chemical modification of histone tails and in particular transcription factor binding (Bell et al., 2011). The tails of histone proteins can be subjected to a plethora of

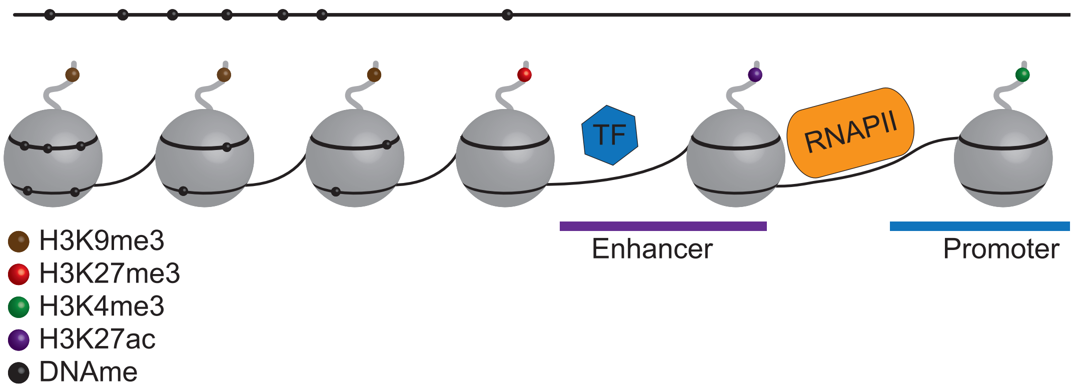


Figure 2.1: Schematic showing the principal organization of DNA around histones. Grey spehres represent nucleosomes with black lines indicating stretches of DNA. Black balls symbolyze methylated CpGs while colored balls on the tails of each histone representent distinct histone modifications. General gene regulatory elements coinciding with specific histone marks are indicated on the bottom. The two principal chromatin environments are displayed on the left for high packaging density/heterochromatin and on the right for low packaging density, corresponding to open or euchromatin.

distinct modifications such as acetylation, phosphorylation, ubiquitination, and methylation (2.1) (Strahl and Allis, 2000) . Some of these modifications physically affect DNA compaction through electrostatic forces when an acetyl group is added to histone tails, neutralizing the positive charge of the tail and therefore weakening the contact with the negatively charged backbone, reducing the energetic cost of nucleosome eviction (Dorigo et al., 2003, Shogren-Knaak et al., 2006).

However, most histone modifications exert their regulatory function through interaction with proteins capable of recognizing a particular modification, termed chromatin readers (Chen and Dent, 2013). This observation gave rise to the concept of a histone code, suggesting that distinct histone modifications, on one or more tails, act sequentially or in combination to form a histone code that is read by other proteins to bring about distinct downstream events (Strahl and Allis, 2000). More specifically, the histone modifications act in a combinatorial fashion forming binding modules (Ruthenburg et al., 2007) and, depending on the particular combination of modifications on a particular histone, attract or repel a specific set of proteins to a genomic locus (Kouzarides, 2007). Recent studies have provided strong evidence for the histone code concept, showing the coexistence and crosstalk between distinct histone modifications (Bartke et al., 2010) at the same genomic locus as well as the genome wide co-occurrence of chromatin regulator binding (Ram et al., 2011). There are many functions of these histone modifications, such as the recruitment of nucleosome remodeling complexes (e.g. NURF) (Wysocka et al., 2006), causing compaction or decompaction and providing a platform of open chromatin regions for transcription factor binding (Fig. 2.1). This process ultimately leads to the activation or modulation of transcription (Wysocka et al., 2006) or to the recruitment of repressive complexes such as the polycomb complex (Margueron and Reinberg, 2011), leading to gene silencing.

The addition of modifications to the histone tails are catalyzed by histone modifying enzymes that are frequently recruited by transcription factors or other chromatin modifying enzymes (Kouzarides, 2007). Despite enormous efforts, the exact mechanisms of action of many histone modifications and their associated readers and writers is poorly understood (Weiner et al., 2012). Recent years have seen dramatic progress in the genome- wide mapping of many of the histone marks, as well as transcription factor binding sites and DNA methylation in hundreds of different cell lines and tissues (e.g. by the ENCODE and Epigenome Roadmap projects). This has led to the widespread usage of chromatin modification pattern to identify many distinct classes of gene regulatory elements and the assignment of putative functions by means of correlation analysis (Ernst and Kellis, 2010, Ernst et al., 2011, Guttman et al., 2009, Heintzman et al., 2007, 2009, Mikkelsen et al., 2007, Roy et al., 2010). Distinct regulatory elements as well as distinct states of these elements can be distinguished based on the combinations of histone modifications and DNA methylation status (Ernst and Kellis, 2010, Mikkelsen et al., 2007). The totality of these modifications to the DNA/histone is commonly referred to as the epigenome.

It has been noted early on that tri-methylation of lysine 4 on histone 3 occurs in promoter regions of active genes or genes poised for rapid activation (Bernstein et al., 2006, Mikkelsen et al., 2007). To date, many chromatin and general epigenetic signatures demarcating numerous classes of classical gene regulatory elements such as promoters, enhancer or insulators have been described.

In addition to the identification of these gene regulatory elements in terms of their genomic locations, observations of the correlation of distinct chromatin signatures with transcription and genome compaction allows the assignment of distinct chromatin states to gene regulatory elements, summarizing our knowledge on the co-occurrence and likely function of these chromatin signatures (Ernst et al., 2011, Mikkelsen et al., 2007). These states, in turn, can change between cell types (Ernst et al., 2011, Rada-Iglesias et al., 2011). For example, while the promoter regions of many key developmental transcription factors are held in a repressed state by the polycomb group proteins in mouse embryonic stem cells, a subset of these genes becomes expressed upon differentiation to neural progenitor cells. Concordantly, the promoter region of these genes changes their chromatin signature from being polycomb-repressed (H3K27me3) to an open chromatin conformation characterized by H3K4me3 and H3K9ac (Wang et al., 2008).

2.2 Gene architecture and genomic regulatory elements

In eukaryotic organisms a gene encodes the amino acid sequence for a protein. However, in addition to the coding part of a gene, the DNA sequence that is transcribed into mRNA contains many non-coding elements such as the 5' and 3' untranslated region as well as intronic sequences that separate the coding sequence into distinct parts (exons). These intronic sequences are removed at the mRNA level in the splicing process (Gilbert, 2006, Lodish, 2008).

In addition to the removal of introns, many genes are also capable of coding for diverse proteins through the selective retainment of exons, a process called alternative splicing (Matlin et al., 2005).

While regulation of protein isoforms can be achieved on the exon level, protein abundance can be regulated on many stages. Prevalent mechanisms include regulation of mRNA expression level and mRNA stability e.g. through microRNA induced degradation by targeting of the 3'UTR (Fabian et al., 2010). In contrast, regulation of gene expression occurs primarily on the DNA level. Specialized gene regulatory elements encoded in the DNA such as the promoter and other regulatory regions further upstream/downstream of the transcription start site (TSS) function as regulatory platforms to initiate and regulate gene expresssion. While the promoter region in immediate vicinity of the TSS harbors the landing and assembly site of the transcriptional machinery with RNA-Polymerase II at its core (Lodish, 2008), parts of the promoter region more upstream (typically defined 2 kb upstream and 500 bp downstream) encompass many binding sites for transcription factors that frequently have the capacity to unlock transcription or modulate the rate of transcription when bound to the promoter region. This region is the key control element of an individual gene and the gateway to its expression, highly similar to the original operon model proposed by Jacob and Monod in 1961 (Jacob and Monod, 1961).

However, even genomic regions up to hundreds of kilobases away from the TSS can impact the transcription of a gene, although typically to a lesser extend than the promoter region (Ong and Corces, 2011). Regions capable of increasing the transcription of a

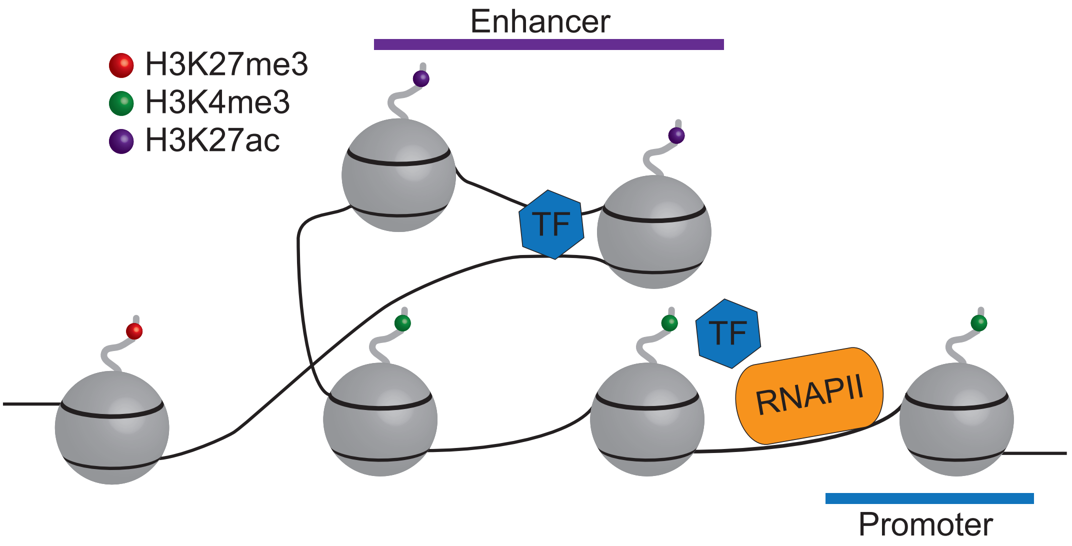


Figure 2.2: Schematic showing the general principal of enhancer-promoter looping interaction mediated by enhancer binding transcription factors.

gene when brought in physical contact with the promoter are called enhancers (Ong and Corces, 2011). By default, the physical interaction of promoter and enhancer does not take place and is mediated by enhancer specific transcription factors which in turn can recruit the mediator complex and cohesin (Kuras et al., 2003, Ong and Corces, 2011) leading to the formation of a DNA loop to the target promoter, ultimately increasing the expression level of the target gene (Fig. 2.2). Recent findings have shown that a single enhancer region can interact with multiple promoter regions (Gibcus and Dekker, 2013) when activated and a single promoter region can be the target of multiple enhancers (Gibcus and Dekker, 2013).

Analogous to enhancer regions, other gene regulatory elements have been described such as repressive and insulating elements. The former facilitates repression of a gene when bound by an appropriate transcription factor while the insulating elements are capable of blocking looping interactions such as between promoter and enhancers when occupied by a proper transcription factors such as CTCF (Maston et al., 2006).

Together, all these non-coding-regulatory elements give rise to a complex regulatory code allowing for spatially and temporally tightly regulated expression levels of genes and non-coding RNAs. However, key players initiating and maintaining these interactions are transcription factor proteins.

2.3 Transcription factors

Transcription factors (TFs) are proteins that bind DNA in order to activate or repress transcription, facilitate chromatin remodeling or direct a three-dimensional reorganization of the genome (Gilbert, 2006, Jin et al., 2013). In a simple model, activation of gene expression is achieved through TF binding to the promoter region of a particular gene or non-coding RNA, followed by recruitment of the transcriptional machinery (Gilbert, 2006, Spitz and Furlong, 2012). Similarly, increased gene expression can be achieved through TF binding to genomic regions distal to the TSS (inter- or intragenic) and a subsequent looping interaction with the target promoter mediated by the mediator complex, leading to enhanced transcription (Kuras et al., 2003, Ong and Corces, 2011). Distal regions with these properties are defined as enhancers (Ong and Corces, 2011). In contrast, TFs such as the REST protein (Schoenherr and Anderson, 1995) or lacZ are capable of directly repressing transcription (Jacob and Monod, 1961).

Based on their structural properties, TFs can be grouped into several families (Pabo and Sauer, 1992), which frequently participate in similar developmental processes such as the HOX proteins in axis formation and patterning (Gilbert, 2006). Many TFs have a high binding affinity to specific DNA sequences, typically between 8 and 20 bp long which is referred to as the binding motif of the TF (Jolma et al., 2013, Pabo and Sauer, 1992). Closely-related TFs frequently share a similar binding motif (Jolma et al., 2013). However, the DNA binding sequence preference of a particular TF can be modulated through protein-protein interactions with specific co-factors or other TFs, giving rise to homo- and heterodimeric complexes (Jolma et al., 2013, Slattery et al., 2011).

In addition to DNA sequence, the local chromatin environment and the DNA methylation status at a particular TF binding site (TFBS) can greatly influence the TF binding capacity. While some TFs such as MYC or TCF4 require a non heterochromatin environment in order to bind (Bartke et al., 2010), other factors such as the forkhead family members FOXA1 and FOXA2 are capable of binding heterochromatic regions and inducing chromatin de-compaction (Smale, 2010, Zaret and Carroll, 2011). Because of their ability to bind and open heterochromatic regions and the resulting opportunity for other factors to bind to newly created open chromatin, these factors have been named pioneering TFs (Smale, 2010, Zaret and Carroll, 2011).

Furthermore, TFs are capable of interacting with a large variety of different proteins that do not possess sequence specificity themselves and recruit them to specific genomic regions. Prominent interaction partners include chromatin modifying enzymes (Smith and Shilatifard, 2010) and DNA methyltransferases (Fuks et al., 2001, Puto and Reed, 2008), that will ultimately cause a change in the epigenetic state of a genomic locus, potentially leading to gene activation (e.g. recruitment of histone acetylases) or repression through recruitment of LSD1 (Whyte et al., 2012). However, recent evidence challenges the paradigm that a TF can function in a solely repressive or activating manner and rather suggests a model where a specific TF can indeed carry out either function, depending on the chromatin environment and available interaction partners (Jafari et al., 2012).

Interestingly, TFBSs in promoter, enhancer or silencer regions occur not in an isolated fashion, but are often flanked by functional binding sites of many other TFs that can be expressed at the same time (Davidson and Erwin, 2006, Remenyi et al., 2004). Frequently, several binding sites for the same factor occur within these gene regulatory elements (Hardison and Taylor, 2012). Furthermore, it has been shown that many gene regulatory elements such as promoters or enhancers can only carry out their function if multiple different TFs are bound to them, giving rise to the paradigm of combinatorial transcriptional control (Davidson and Erwin, 2006, Remenyi et al., 2004, Zinzen et al., 2009). In *Drosophila*, a set of thousands of particularly large clusters of gene regulatory elements, termed cis-regulatory modules (CRMs) were found to control mesoderm development in a combinatorial fashion (Hardison and Taylor, 2012, Negre et al., 2011, Zinzen et al., 2009). Recently, a similar class of CRMs carrying out essential cellular functions were also discovered in mammalian systems and were termed super enhancers (Hnisz et al., 2013).

Over the course of the last 20 years, TFs have been identified as key components directing and maintaining cellular fates. While knockdown of key TFs in different cell types induces de-differentiation (Holmberg and Perlmann, 2012, Mikkelsen et al., 2008), over-expression of a single factor such as MyoD is sufficient to convert human fibroblasts to myoblasts (Weintraub et al., 1989), which can then be further differentiated into muscle. Finally, combinations of TFs can dictate cellular identity and revert fully differentiated somatic cells back to a pluripotent state that is equivalent to embryonic stem cells (Takahashi and Yamanaka, 2006).

In summary, TFs appear to be the key control elements conferring specificity to a large set of epigenetic modifying enzymes. Combinatorial TF control and the tight mutual regulation of their expression (Davidson and Erwin, 2006, Gerstein et al., 2012) gives rise to a highly complex regulatory logic directing and maintaining cellular states.

2.4 DNA methylation

DNA can be subject to chemical modification through the addition of a methyl group to the 5' carbon atom of cytosine, giving rise to methyl-cytosine, sometimes called the 5th base (Bird, 2002, Holliday and Pugh, 1975, Jones, 2012, Riggs, 1975). In mammalian genomes, DNA methylation predominantly occurs in the context of CpG dinucleotides (Bird, 2002, Jones, 2012), although low levels of non-CpG methylation have been found in hESCs and neuronal cells (Lister et al., 2009, 2013, Ziller et al., 2011). CpG dinucleotides are heavily depleted in mammalian genomes (Bird, 2002, Lander et al., 2001, Venter et al., 2001) due to increased mutational burden of methylated CpGs through spontaneous deamination, causing progressive loss of this dinucleotide (Shen et al., 1994). Overall, there are only 28 million CpG dinucleotides as opposed to more than 400 million CpAs in the human genome.

In addition to their low abundance, the distribution of CpGs across the genome is highly asymmetric. Most of the genome is depleted of CpGs except for a set of small 500-2 kb regions that are highly enriched for CpGs (Bird, 2002, Gardiner-Garden and Frommer, 1987, Lander et al., 2001, Venter et al., 2001). The latter regions are termed CpG islands (CGIs) and are located near TSS of roughly half of all genes including almost all housekeeping genes, as well as key developmental regulators (Jones, 2012) . While CpG islands are constitutively hypo-methylated (Bird, 2002) the vast majority of the mammalian genome (80%) is highly methylated (Bird, 2002), except during a short period in pre-implantation development and primordial germ cell specification (Smith and Meissner, 2013).

DNA methylation is deposited by a specialized DNA methylation machinery that is traditionally grouped into two classes (Bestor, 2000). The first class is composed of DNA methyltransferases DNMT3A and DNMT3B, which catalyze *de novo* methylation of unmethylated regions. DNMT1 exhibits a strong preference for hemimethylated DNA, encountered during DNA replication. Therefore, this enzyme is considered the maintenance methyltransferase as it copies the DNA methylation pattern to the newly synthesized strand and follows the replication fork (Bestor, 2000). DNA methylation is essential for normal development (Bestor, 2000) and frequently misregulated in various diseases (Bergman and Cedar, 2013, Robertson, 2005) including different types of cancers (Baylin and Jones, 2011, Bergman and Cedar, 2013). In the latter case, particularly CpG islands located in the promoters of tumor suppressor genes become hyper-methylated while large domains of the entire genome lose methylation (Baylin and Jones, 2011, Bergman and Cedar, 2013).

Historically, the main function of DNA methylation was considered to be the repression of gene expression by blocking transcription initiation when occurring in the vicinity of the TSS and interference with TF binding either through direct physical repulsion or the recruitment of methyl-binding proteins (Bell and Felsenfeld, 2000, Deaton and Bird, 2011, Jones, 2012, Suzuki and Bird, 2008, Tate and Bird, 1993). However, recent evidence suggests additional functions for DNA methylation outside of the promoter context (Jones, 2012). These functions include stimulation of transcription elongation in the gene body, maintaining genome stability through silencing of transposable elements and highly repetitive DNA (Jones, 2012).

Additionally, recent studies paint a complex picture of the relationship between DNA methylation and TF binding (Lienert et al., 2011, Stadler et al., 2011). While early studies showed the capacity of DNA methylation to block binding of TFs and DNA binding proteins such as SP1 and CTCF (Tate and Bird, 1993), more recent findings suggest a strong factor dependence (Lienert et al., 2011, Stadler et al., 2011). In particular, pioneering TFs such as FOXA2 have the ability to bind DNA regardless of methylation state (Zaret and Carroll, 2011). Similarly, a recent study suggested that the transcriptional repressor REST (Stadler et al., 2011) as well as the glucocorticoid receptor (GR) exhibits a similar ability (Saluz et al., 1986, Wiench et al., 2011). Particularly the latter finding has received great attention in the context of DNA methylation at enhancer regions (Hon et al., 2013, Wiench et al., 2011). The binding of the GR to distal regulatory elements was able induce de-methylation at the respective loci and subsequent activation of the enhancer (Saluz et al., 1986, Wiench et al., 2011). More recent evidence further supports the involvement of DNA methylation in enhancer functions since it was shown that lowly methylated regions (LMRs) in mouse embryonic stem cells frequently coincide with putative enhancer elements (Stadler et al., 2011). In addition, another study in T cells showed that increased methylation in a subset of putative enhancer regions indeed resulted in lower enhancer activity in reporter assays (Schmidl et al., 2009). However, most evidence so far points towards a model in which DNA methylation of a particular genomic locus is not responsible for the initial silencing and repression of the locus, but is rather a second or third step in stably maintaining repression, locking the locus in a silent state (Bird, 2002, Jones, 2012). One particular model system that has greatly facilitated our understanding of chromatin regulation, DNA methylation and its critical role in cell fate determination are mouse and human pluripotent stem cells.

2.5 Pluripotent stem cells

In a functional sense, pluripotency is the capacity of a cell to give rise to differentiated derivatives that represent each of the three primary germ layers (Chambers and Tomlinson, 2009). Conceptually, pluripotency is understood as the potential ability of a cell to differentiate into all cell types of an organism (Ng and Surani, 2011). *In vivo*, pluripotency is encountered in the cells of the inner cell mass (ICM) of the developing blastocyst. These cells can be explanted *in vitro* to obtain ES (ES) cell lines when kept in proper culture conditions (Chambers and Tomlinson, 2009). These cells are capable of contributing to all germ layers of a mouse when injected into the blastocyst of developing mouse embryos, giving rise to chimeric animals (Ng and Surani, 2011).

While ES cells from mice have been derived as early as in 1981 using Leukemia Inhibitory Factor (LIF) and Bone Morphogenic Protein 4 (BMP4) as key cell culture ingredients (Evans and Kaufman, 1981, Martin, 1981, Smith et al., 1988), ES cells from humans (hESCs) were derived a decade later in 1998 relying on a distinct growth factor cocktail (Thomson et al., 1998). In contrast to mouse ES cells (mESCs), hESCs require the presence of Fibroblast Growth Factor 2 (FGF2) and Activin/Nodal signaling for establishment and maintenance (Ng and Surani, 2011). Despite distinct maintenance requirements and transcriptional profiles (Tesar et al., 2007), numerous studies have identified a shared transcriptional core network controlling the ES cell state in human and mouse with the TFs OCT4, SOX2 and NANOG at the center (Chambers and Tomlinson, 2009, Young, 2011).

Considering the ethically controversial derivation of pluripotent stem cells from human embryos, a recent landmark study uncovered that fully differentiated somatic mouse and human cell types can be reprogrammed to a pluripotent stem cell state using the over-expression of a defined set of TFs (Takahashi and Yamanaka, 2006, Takahashi et al., 2007). This study uncovered the molecular principles underlying John Gurdon's early transplantation experiments, and proved the enormous epigenetic plasticity of the genome. Induced pluripotent stem cells (iPSCs) are functionally and frequently molecularly identical to ES cells (Bock et al., 2011, Guenther et al., 2010) and can contribute to an entire organism (Zhao et al., 2009). Due to their self-renewing ability and capacity to differentiate into virtually all cell types, ES cells and iPS cells have become a valuable tool to study the molecular basis of cell fate commitment and differentiation. In addition, these cells hold great promise for regenerative medicine, if it is possible to identify suitable differentiation protocols to generate specific healthy cell types in sufficient numbers or even entire organs *ex vivo*.

Chapter 3

Materials and Methods

3.1 Introduction

The content of this chapter has in part been published previously in Ziller et al. (2013) (Ziller et al., 2013), Gifford and Ziller et al. (2013) (Gifford et al., 2013) and Ziller et al. (2014) (in revision). More detailed methods relevant to the individual studies are presented in the corresponding chapter.

3.2 Experimental methods

3.2.1 Measuring DNA methylation

Over the last 30 years, a plethora of detection methods for the DNA methylation state have been developed (Laird, 2010). In the context of this work, two bisulfite-based approaches were employed. The treatment of denatured genomic DNA with sodium bisulfite chemically deaminates unmethylated cytosine bases much faster than methylated cytosines (Laird, 2010). If properly timed, almost all unmethylated cytosine residues are converted to uracil while methylated cytosines remain unchanged. In a subsequent PCR step, uracil bases get converted to thymines, effectively converting unmethylated Cs to Ts. With the original DNA sequence known, the bisulfite-converted DNA can then be sequenced and the DNA methylation state determined (Krueger et al., 2012, Laird, 2010).

3.2.2 Whole-genome bisulfite sequencing

In order to determine the methylation state of all cytosines in the genome at single-base pair resolution, bisulfite treatment can be combined with genomic library production and next-generation sequencing (Lister and Ecker, 2009), giving rise to a whole-genome bisulfite sequencing (WGBS) library. Here, we employed the MethylC-Seq protocol ((Gifford et al., 2013, Lister and Ecker, 2009, Lister et al., 2008, 2009)) for WGBS library production coupled with next-generation sequencing using the Illumina HiSeq2000.

3.2.3 Reduced representation bisulfite sequencing

While WGBS libraries capture almost all bases of the genome, their generation is a very costly and inefficient process. In order to achieve sufficient genomic coverage for statistically robust single-base pair methylation calls, a genomic coverage of 30x is recommended (Consortium, 2011). However, a large fraction of the resulting sequenced reads (up to 60%) are not informative with respect to DNA methylation state since they do not contain any CpG dinucleotides due to their genome-wide depletion and asymmetric distribution (Bird, 2002, Ziller et al., 2013). However, this asymmetric distribution of CpG dinucleotides can also be exploited to enrich specifically for CpG rich regions. One of the most widespread approaches that takes advantage of this observation employs restriction enzymes to digest the genome coupled with fragment size selection and next-generation sequencing library generation to enrich for higher CpG content. This reduced representation bisulfite sequencing method (RRBS) allows to assay 8-10% (Gu et al., 2010, Meissner et al., 2008, Smith et al., 2009) of all genomic CpGs while sequencing only 1% of the entire genome. To achieve this high level of CpG enrichment, RRBS employs the restriction enzyme MspI that recognizes the DNA sequence CCGG, ensuring that each digested fragment and therefore each sequencing read is informative and contains at least one CpG (Gu et al., 2010, Meissner et al., 2008, Smith et al., 2009).

3.2.4 Chromatin immunoprecipitation coupled with bisulfite sequencing

In order to determine the co-occurrence of DNA methylation and histone modifications or TF binding is not sufficient to perform a DNA methylation and ChIP-Seq experiment separately, even if carried out on the same cell population. Due to potential heterogeneity in the assayed cell population, an observed co-localization of protein/histone modifications and DNA methylation based on next-generation sequencing data can simply arise due to the fact that half the cell population was bearing a histone modification or protein protein at a particular locus but being devoid of any DNA methylation while the other half exhibits high methylation levels but is depleted of DNA methylation. In order to reliably determine the co-localization of protein/histone modification and DNA methylation, we performed ChIP-bisulfite sequencing (Brinkman et al., 2012). This technique combines immunoprecipitation with bisulfite sequencing to assess the DNA methylation status of only those DNA fragments that were pulled down by the antibody against the protein/histone modification of interest (Brinkman et al., 2012).

3.2.5 Chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-Seq)

To goal of a chromatin immunoprecipitation experiment for DNA interacting proteins is to determine those DNA fragments that are enriched for binding of a particular protein (Furey, 2012, Park, 2009). The main protein classes of interest are TFs and histones. To obtain genome-wide maps of protein-DNA interaction, the DNA is cross linked in vivo by treating the cells with formaldehyde and the chromatin is subsequently sheared by sonication into small fragments in the range of 200-600bp (Park, 2009). Next, an antibody against the specific protein or chromatin modification of interest is used to immunoprecipitate the DNA-protein complex (Park, 2009). Finally, the protein-DNA crosslinking is reversed, the released DNA extracted and used for the construction of a next-generation sequencing library (Park, 2009). If the antibody was specific, only DNA fragments associated with the specific protein or histone modification should be present in the library. Mapping the resulting reads from the sequenced library to the genome then allows to assess the genome-wide distribution of the assayed TF or histone modification. Here, we employed several improved ChIP-Seq library construction protocols that are introduced in the subsequent methods section relevant to the individual chapters.

3.2.6 RNA-Seq library construction

RNA was extracted using the miRNeasy kit (Qiagen, 217004). Poly(A) RNA was isolated using Oligo d (T25) beads (NEB, E7490L). The poly(A) fraction was then fragmented (Invitrogen, AM8740). Fragments smaller than 200 bps were eliminated (Zymo, R1016) and the remaining fraction was treated with FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific, EF0652) and T4 Polynucleotide Kinase (NEB, M0201L). RNA was then ligated to a RNA adaptor using T4 RNA Ligase 1 (NEB, M0204L), which was then used to facilitate cDNA synthesis using Affinity Script Multiple Temperature Reverse Transcriptase (Agilent, 600105). RNA was then degraded and the cDNA was ligated to a DNA adaptor using T4 RNA Ligase 1. Final library amplification was completed using NEB Next High Fidelity 2X PCT Master Mix (M054L).

3.3 Computational methods

3.3.1 Basic methylation data processing

WGBS libraries were aligned using maq (Li et al., 2008) in bisulfite mode and BSMap (Xi and Li, 2009) to the hg19/GRCh37 reference assembly using default parameter values. Subsequently, CpG methylation calls were made excluding duplicate reads, and bases with a quality score ≥ 20 as well as reads with more than 10% mismatches. Here, a methylation call is defined as the computation of the number of reads overlapping a particular CpG harboring a C or a T at the cytosine coordinate of the CpG. Let m be the number of C's and u be the number of T's. The value $e = m/(m+u)$ then gives the methylation ratio of each CpG. The methylation calling pipeline was implemented in Python (http://python.org/) (Ziller et al., 2013) and mainly written by Fabian Müller. The accompanying code can be found in the accompanying zip archive in the folder Methylation_Pipeline.

3.3.2 Basic ChIP-Seq data processing and analysis

ChIP-Seq data was aligned to the hg19/GRCh37 reference genome using bwa version 0.5.7 (Li and Durbin, 2009) or maq version 0.7.1(Li et al., 2008) with default parameter settings. Subsequently, reads were filtered for duplicates and extended by 200 bp. Visualization of read count data was performed by converting raw bam files to .tdf files using IGV tools version 2.2.1 (Thorvaldsdottir et al., 2013) and normalizing to 1 million reads. To identify regions of chromatin modification enrichment we employed the approach suggested by Mikkelsen et al. (Mikkelsen et al., 2010), modeling the distribution of unique read sequences in a genomic region of size l with a Poisson model. To determine the enrichment over background, we used the whole-cell extract (WCE) to model our background distribution of read sequences. The WCE is generated in the same way as the ChIP-Seq library except for omitting the IP step. Using the WCE as background model has the advantage of implicitly accounting for shearing biases and read-mapping artifacts. We defined a nominal p-value for enrichment against the control within a given region i of length l in sample k harboring r_i midpoints of uniquely aligned fragment size extended sequencing reads as $P(C \geq r_{ik})$ with (Mikkelsen et al., 2010):

$$C \sim \text{Poisson}(\max[1, \epsilon_c] \times \lambda_k) \tag{3.1}$$

$$\epsilon_{ic} = \frac{r_{ic}}{\lambda_c} \tag{3.2}$$

$$\lambda_c = l \times T_c/G \tag{3.3}$$

$$\lambda_k = l \times T_k/G \tag{3.4}$$

with r_{ic} reads in the WCE control channel, T_c and T_k the total number of uniquely aligned reads in the WCE and ChIP library respectivly, and G the mappable size of the genome (for hg19: $G = 2.7 \times 10^9$ (Zhang et al., 2008)). The coverage-adjusted enrichment level of genomic region i of size l in sample k over WCE control sample c is then defined as

$$e_{ik} = \frac{\epsilon_{ik}}{\epsilon_{ic}}$$

with ϵ_{ik} corresponding to the reads per kilobase per million reads sequenced (RPKM) within the genomic region i in sample k or in the control c respectively as defined in equation (3.2) .

For differential enrichment analysis between two ChIP experiments, the second ChIP experiment is simply treated like the WCE control in (3.1) and (Mikkelsen et al., 2010). For genome-wide significance testing, the resulting p-values are corrected using the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method as implemented in the R function p.adjust. All ChIP-Seq analysis was conducted in R (http://www.r-project.org/)

In order to identify TFBS we employed traditional peak calling using MACS (Zhang et al., 2008) against the matching WCE, removing all duplicates and only retaining peaks with $p \leq 10^{-5}$. The accompanying code can be found in the accompanying zip archive in the folder ChIP-Seq_analysis.

3.3.3 Gene set enrichment analysis

Gene set enrichment analysis for genomic regions was carried out using the GREAT toolbox (McLean et al., 2010) and only categories with q-values $\leq$ 0.05 for both the Hypergeometric and the binomial test as well as a minimal region enrichment level greater than 2 were considered.

3.3.4 RNA-Seq data analysis

To assess expression levels of the entire transcriptome, we performed strand specific RNA-Sequencing for polyadenylated RNA, capturing coding as well as non-coding transcripts.

Reads were mapped to the human genome (hg19) using TopHat v2.0.6 (Trapnell et al., 2009)

(http://tophat.cbcb.umd.edu) with the following options: ”-library-type firststrand” and ”-transcriptome-index” with a TopHat transcript index built from RefSeq. Transcript expression was estimated with an improved version of Cuffdiff 2

(http://cufflinks.cbcb.umd.edu) (Trapnell et al., 2013). Cuffdiff was run with the following options: ”-min-reps-for-js-test 2 -dispersion-method per-condition” against the UCSC iGenomes GTF file from Illumina. The workflow used to analyze the data is described in detail in (Trapnell et al., 2012) (alternate protocol B).

Chapter 4

Charting a dynamic DNA methylation landscape of the human genome

4.1 Inroduction

The content of this chapter has previously been published in Ziller et al. 2013 (Ziller et al., 2013). The design of the analysis strategy as well as all analyses were carried out by M.J. Ziller. All wet lab experiments including next-generation sequencing library production were contributed by the co-authors. The manuscript was written by M.J. Ziller.

Progress in the genome wide mapping and analysis of chromatin states across many different cell types has not only broadened our understanding of epigenetic regulation, but also gave rise to the notion that specific chromatin signatures demarcate distinct classes and states of gene regulatory elements (Bernstein et al., 2012, Ernst and Kellis, 2010, Ernst et al., 2011, Heintzman et al., 2009, Maston et al., 2006, Mikkelsen et al., 2007). Furthermore, these studies suggested that the distribution of specific histone marks such as HK4me1,2,3, H3K27ac and H3K27me3 as well as DNAse I hypersensitive sites provide insights into the active regulatory network of transcription factors within a given cell type (Ernst et al., 2011, Neph et al., 2012). Therefore, the cell type-specific distribution of these epigenetic signatures permits insights into the distinct architecture of regulatory networks controlling each cellular state. While many different cell types have been investigated for histone modifications (Bernstein et al., 2012) and DNAse I hypersenstivity (Neph et al., 2012), the roles and dynamics of DNA methylation are far less clear. DNA methylation is a defining feature of mammalian cellular identity and essential for normal development (Bestor, 2000, Reik, 2007). Most cell types, except germ cells and pre-implantation embryos (Hackett and Surani, 2013, Seisenberger et al., 2012, Smith et al., 2012), display relatively stable DNA methylation patterns with 70-80% of all CpGs being methylated (Bird, 2002).

Despite recent advances we still have a too limited understanding of when, where, and how many CpGs participate in genomic regulation. Here we report the in-depth analysis of 42 whole-genome bisulphite sequencing data sets across 30 diverse human cell and tissue types. We observe dynamic regulation for only 21.8% of autosomal CpGs within a normal developmental context, most of which are distal to transcription start sites. These dynamic CpGs co-localize with gene regulatory elements, particularly enhancers and transcription factor-binding sites, which allow identification of key lineage-specific regulators. In addition, differentially methylated regions (DMRs) often contain single nucleotide polymorphisms associated with cell-type-related diseases as determined by genome-wide association studies. The results also highlight the general inefficiency of whole-genome bisulphite sequencing, as 70-80% of the sequencing reads across these data sets provided little or no relevant information about CpG methylation. To demonstrate further the utility of our DMR set, we use it to classify unknown samples and identify representative signature regions that recapitulate major DNA methylation dynamics. In summary, although in theory every CpG can change its methylation state, our results suggest that only a fraction does so as part of coordinated regulatory programs. Therefore, our selected DMRs can serve as a starting point to help to guide new, more effective reduced representation approaches to capture the most informative fraction of CpGs, as well as further pinpoint putative regulatory elements.

Changes in DNA methylation patterns and the resulting DMRs have been the focus of numerous studies in the context of normal development (Smith and Meissner, 2013) and disease (Bergman and Cedar, 2013). These studies have characterized many different DMR classes including partially methylated domains (Lister et al., 2009), condition-specific (Nazor et al., 2012), cell-type-specific (Laurent et al., 2010, Lister et al., 2009, Meissner et al., 2008, Weber et al., 2005) and tissue-specific DMRs (Irizarry et al., 2009, Varley et al., 2013), as well as DMRs arising in diseases such as cancer (Berman et al., 2012, Irizarry et al., 2009). Owing to the relatively small fraction of genomic CpGs assayed or small sample cohorts, the question of what fraction of genomic CpGs changes its methylation state in the context of normal development as well as their regulatory context remains underexplored.

4.2 Computational methods

4.2.1 Parametric description of CpG methylation levels

Given a single CpG at genomic coordinate i and two biological samples s_1 and s_2, we observe m_1, m_2 reads where the genomic CpG is seen as methylated and u_1, u_2 reads where the CpG is unmethylated. The ratio

$$e_{ij} = \frac{m_{ij}}{m_{ij} + u_{ij}}$$

gives the methylation level e_{ij} of genomic CpG i in replicate j.
At the heart of DNA methylation data analysis lies the desire to identify differentially methylated cytosines or DMRs between different conditions, for example distinct cell types or normal and disease samples. Previous approaches to the analysis of bisulfite sequencing relied heavily on statistical methods such as Fisher's exact test on a single CpG level or some variant of the t-test if a single CpG or a genomic region harboring multiple cytosines is compared across conditions (Becker et al., 2011, Bock, 2012, Lister et al., 2009). However, on the single cytosine level the use of Fisher's exact test might be to conservative as it does properly reflect the underlying sampling process giving rise to the cytosine methylation measurements. In addition, Fisher's exact test does not allow for the estimation of confidence intervals, a feature highly desirable in the context of differential methylation assessment. Similarly, the assumptions underlying the frequently used two proportion z-test are not met in the low coverage regime.

To overcome some of these drawbacks and in order to minimize the impact of noise on our estimates and improve dynamic CpG detection sensitivity, we devised a model based on the beta distribution to model single CpG methylation levels.

The underlying random process of the sampling procedure is modeled by a binomial distribution for the number of methylated reads per CpG i with methylation probability θ_{ij}:

$$m_{ij} \sim \mathrm{Bin}(n_{ij}, \theta_{ij}) \tag{4.1}$$

with the total number of reads covering CpG i in sample j $n_{ij} = m_{ij} + u_{ij}$. The true fraction of CpGs methylated in the profiled cell population θ_{ij} is unknown and we therefore follow a Bayesian approach treating θ_{ij} as random variable itself. To that end, we assume further that θ_{ij} follows a beta distribution with parameters α_i, β_i:

$$\theta_{ij} \sim \text{Beta}(\alpha_i, \beta_i) \tag{4.2}$$

The posterior distribution for θ_{ij} then has also beta distribution form:

$$P(\theta_{ij}|m_{ij}, n_{ij}) = \text{Beta}(\theta_{ij}|m_{ij} + \alpha_i, \beta_i + n_{ij} - m_{ij}) \tag{4.3}$$

The parameters of the prior distribution α_i, β_i are unknown and if only one replicate per sample is available we assume a uniform prior with $\alpha_i = \beta_i = 1$. If more than one replicate is available, we employ an empirical Bayes approach to estimate the unknown prior variables α_i, β_i. First, we reparameterize the beta prior distribution in terms of its mean $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$ and precision $M_i = \alpha_i + \beta_i$. The marginal distribution for the number of methylation events m_i is then given by the beta-binomial distribution. The parameters μ_i and M_i of the beta-binomial model are estimated by the method of moments (Martuzzi and Elliott, 1996):

$$\hat{\mu_i} = \frac{\sum\limits_{j=1}^{R} n_{ij} e_{ij}}{\sum\limits_{j=1}^{R} n_{ij}}$$

is the weighted mean of the observed methylation ratios e_{ij} across R replicates. An estimate of M_i can be obtained via (Martuzzi and Elliott, 1996):

$$\hat{M_i} = \frac{\hat{\mu_i}(1 - \hat{\mu_i}) - s^2}{s^2 - \frac{1}{\bar{n}}\hat{\mu_i}(1 - \hat{\mu_i})}$$

with the total weighted sampled variance s^2:

$$s_i^2 = \frac{\sum\limits_{j=1}^{R} n_{ij}(e_{ij} - \hat{\mu_i})^2}{\sum\limits_{j=1}^{R} n_{ij}}$$

Please note that $\hat{M_i}^2$ can be negative when the observed total variation is less than expected. In this case the inter-sample variation should be considered zero (Howley and Gibberd, 2003, Martuzzi and Elliott, 1996) and we set μ_i and M_i to zero and model the distribution of the methylation probability θ_i based on the replicate with the highest sampling frequency n_{ik} using the beta-posterior distribution (4.2):

$$P(\theta_i|m_{ik}, n_{ik}) = \text{Beta}(\theta_i|m_{ik} + 1, n_{ik} - m_{ik} + 1)$$

Transforming back to the original parameterization of the beta distribution using the definitions $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$ and precision $M_i = \alpha_i + \beta_i$ yields:

$$\hat{\alpha}_i = \hat{\mu}_i \hat{M}_i$$
$$\hat{\beta}_i = \hat{M}_i(1 - \hat{\mu}_i)$$

In terms of biology, the beta-distribution captures the natural biological variability encountered when comparing multiple biological samples of the same class and we subsequently use the estimated beta-distribution to describe the CpG of interest.

4.2.2 Estimating the difference in CpG methylation levels between two samples

In order to compute the difference in CpG methylation levels between two samples (groups) $\delta_i = \theta_{i1} - \theta_{i2}$ in a probabilistic framework, we take advantage of the parametric model of CpG methylation probability (4.3) to determine the distribution of the difference in CpG methylation probability between two samples. To that end we subtract the two random variables Θ_{i1} and Θ_{i2} from each other and obtain the new random variable $\Delta_i = \Theta_{i1} - \Theta_{i2}$ quantifying the difference in CpG methylation levels. Δ_i then follows a beta difference distribution (BD) (Phamgia et al., 1993): Wlg., let ($\delta_i > 0$):

$$\begin{aligned} BD(\delta_i|\alpha_{i1}, \beta_{i1}, \alpha_{i2}, \beta_{i2}) = & B(\alpha_{i1}, \beta_{i2})\delta_i^{\beta_{i1}+\beta_{i2}-1}(1-\delta_i)^{\alpha_{i2}+\beta_{i1}-1} \\ & F_1(\beta_{i1}, \alpha_{i1} + \beta_{i1} + \alpha_{i2} + \beta_{i2} - 2, 1 - \alpha_{i1}; \\ & \beta_{i1} + \alpha_{i2}, (1 - \delta_i), 1 - \delta_i^2)/(B(\alpha_{i1}, \beta_{i1})B(\alpha_{i2}, \beta_{i2})) \end{aligned}$$

With B is the beta function normalization factor and F_1 Appell's first Hypergeometric function in two variables, and

$$BD(\delta = 0|\alpha_1, \beta_1, \alpha_2, \beta_2) = B(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_2 - 1)/A$$

for $\delta = 0$ with $A = B(\alpha_1, \beta_1)B(\alpha_2, \beta_2)$ (Phamgia et al., 1993). Since the resulting distribution quickly approaches the normal distribution with increasing CpG coverage n_1, n_2, we distinguish two cases for the computation of the $100(1-\alpha)\%$ credible interval (Pham-Gia and Turkkan, 2003).

1. For low CpG coverage ($n_1 \leq 30$ or $n_2 \leq 30$) we estimate the probability density for δ by simulation, drawing R samples from the joint posterior distribution. The $100{*}(1-\alpha)\%$ credible interval is then given as $[p_{100\times\alpha/2}, p_{100\times(1-\alpha/2)}]$ with p_x being the respective percentile computed over all sample combinations $\{\theta^k_{i1}, \theta^k_{i2}\}_{k=1|\ldots,R}$. The nominal p-value for each one-sided test is then determined by the number of instances where δ_i was larger/smaller.

2. For high CpG coverage n_1, n_2 we approximate the BD distribution $\delta = \mu_1 - \mu_2$ by the normal distribution, with $\mu_l = \alpha_l/(\alpha_l + \beta_l)$ and variance (Pham-Gia and Turkkan, 2003):
$$\sigma_i^2 = \sum_{j=1}^{2} \frac{\alpha_j \beta_j}{(\alpha_j + \beta_j)^2(\alpha_j + \beta_j + 1)}$$

4.2.3 Determining CpG cluster effects

Given a genomic region of size l harboring N individual CpGs and measurements of the methylation level e_{ij} for all CpGs $i = 1, .., N_k$ in two distinct sample groups j, we want to determine the region methylation level difference $\bar{\delta}_k$, $k = 1, .., R$ and significance between the samples for R regions. To this end we use a classic random effects model (DerSimonian and Laird, 1986) to describe the region/CpG cluster methylation differences. Within the framework of this model, we define $\bar{\delta}_k$ as the abstract summary methylation difference of the entire region k. The observed CpG level methylation differences δ_{ki} are then considered to be a function of the region methylation difference $\bar{\delta}_k$. Let δ_{ki}, $i =, 1 \ldots, K$ be a collection of CpG level methylation differences for region k and let V_{ki} be the corresponding variance associated with each measurement. Furthermore, let Δ_k^2 be the population variance across all CpG difference measurements. The observed CpG level methylation differences δ_{ki} are then considered to be a function of the region methylation difference $\bar{\delta}_k$, deviating from this quantity by variations induced through sampling errors for the particular CpG ϵ_i as well as deviation from the mean region methylation level difference ζ_{ki}:

$$\delta_{ki} = \bar{\delta}_k + \zeta_{ki} + \epsilon_{ki} \tag{4.4}$$

A non-iterative estimate of this quantity can be obtained from the variance-weighted deviation of individual CpG methylation differences from the mean difference (DerSimonian and Kacker, 2007, DerSimonian and Laird, 1986):

$$\hat{\Delta}_k^2 = \frac{Q_k - df_k}{C_k},$$

with

$$Q_k = \sum_{i=1}^{N_k} w_{ki}\delta_{ki}^2 - \frac{\left(\sum_{i=1}^{N_k} w_{ki}\delta_i\right)}{\sum_{i=1}^{N_k} w_{ki}}$$

, $df_k = N_k - 1$ and

$$C_k = \sum_{i=1}^{N_k} w_{ki} - \frac{\sum_{i=1}^{N_k} w_{ki}^2}{\sum_{i=1}^{N_k} w_{ki}}$$

where each CpG is weighted according to its associated sampling variance w_{ik}:

$$w_i = \frac{1}{V_{ki}}.$$

δ_i and their associated variances V_i are determined based on the parametric model developed in the previous paragraphs. The total CpG difference variance per CpG is then given as:

$$\hat{V}_{ki} = V_{ki} + \hat{\Delta}_k^2$$

which in turn allows to compute the contribution of each CpG methylation difference $\hat{w}_i$ to the region level methylation difference:

$$\hat{w}_{ki} = \frac{1}{\hat{V}_{ki}}$$

Given these estimates, the region level methylation difference can be determined as (DerSimonian and Laird, 1986):

$$\bar{\delta}_k = \frac{\sum_{i=1}^{N_k} \hat{w_{ki}} \delta_{ki}}{\sum_{i=1}^{N_k} \hat{w_{ki}}}$$

and associated variance:

$$\bar{V}_k = \frac{1}{\sum_{i=1}^{N_k} \hat{w_{ki}}}$$

The estimated standard error then becomes:

$$\bar{\Delta}_k = \sqrt{\bar{V}_k}.$$

These values can then be used to compute the $(1 - \alpha)$ confidence intervals and p-values using the standard normal cumulative distribution function:

$$\begin{aligned} c_L &= \bar{\delta} - \Phi^{-1}(1 - \alpha) \\ c_U &= \bar{\delta} + \Phi^{-1}(1 - \alpha) \end{aligned}$$

and the associated p-value for a one-tailed test:

$$p_k = 1 - \Phi\left(\pm\frac{\bar{\delta}_k}{\bar{\Delta}_k}\right)$$

The methylation level of CpG clusters is determined by computing the coverage-weighted mean across all CpGs within a cluster, limiting the maximal overage contribution per CpG to 25:

$$\bar{\mu} = \sum_{i=1}^{n_j} w_{ij} * e_{ij}$$

with the coverage-based m_{ij} weights:

$$w_{ij} = \frac{n_{ij}}{\sum_{j=1}^{n_j} m_{ij}}$$

4.2.4 Methylation Specificity

The methylation specificity of a genomic region k is defined based on the Jensen-Shannon divergence (Endres and Schindelin, 2003):

$$JS(\underline{m}_1, \underline{m}_2) = H\left(\frac{\underline{m}_1 + \underline{m}_2}{2}\right) - \left(\frac{H(\underline{m}_1) + H(\underline{m}_2)}{2}\right)$$

with $\underline{m}_1, \underline{m}_2$ two discrete probability distributions fulfilling $\sum_{i=1}^{n} m_i = 1$ and $0 \leq m_i \leq 1$ and entropy H:

$$H(\underline{m}) = -\sum_{i=1}^{n} m_i \log_2(m_i)$$

Here, we exploit the well-known fact that the square root of the Jensen-Shannon divergence is metric to determine the methylation specificity (Cabili et al., 2011). For each genomic region k we first normalize the methylation level distribution across all N samples $\underline{m}_k$ to fulfill $\sum_{i=1}^{n} m_i = 1$. Next, we determine the Jensen-Shannon divergence between $\underline{m}_k$ and N prototypic methylation profiles representing the N possible extremes where the region k is completely methylated in only one of all samples and unmethylated in all other samples or vice versa:

$$JS(\underline{m}_k, \underline{r}_j)_{\text{sp}} = 1 - \sqrt{JS(\underline{m}_k, \underline{r}_j)}$$

with $r_{jl} = \delta_{jl}^K$ for hypermethylation specificity and $r_{jl} = 1 - \delta_{jl}^K$ for hypomethylation specificity, where δ_{jl}^K is the Kronecker delta. The hyper (HR)/hypomethylation (HO) specificity is then given as:

$$JS^{\text{HR}}(\underline{m}_k)_{\text{sp}} = \underset{j=1,\ldots,N}{\arg\max} JS(\underline{m}_k, \underline{r}_j^{\text{HR}})_{\text{sp}} \tag{4.5}$$

$$JS^{\text{HO}}(\underline{m}_k)_{\text{sp}} = \underset{j=1,\ldots,N}{\arg\max} JS(\underline{m}_k, \underline{r}_j^{\text{HO}})_{\text{sp}} \tag{4.6}$$

4.2.5 Estimation of saturation curve

To get an estimate of the rate of newly discovered dynamic CpGs as a function of number of samples, we randomly select sample subsets from our developmental cohort at different cell type coverage levels (5, 10, 15, 20, 24). For coverage levels 5 and 10 we create 10 distinct sample subsets, for 15 cell types five subsets and for 20 cell types two distinct subsets. Subsequently we run our standard dynamic CpG identification procedure on each of the defined subsets and average the number of dynamic CpGs found for each cell type coverage level. Based on these data points we compute a cubic spline fit and plot (Fig. 4.16).

4.2.6 Signature set derivation and classification

In order to derive cell group-specific signature regions, we utilized the mean and variance estimates derived for all dynamic regions across all samples. Assuming normal distribution of methylation levels for each region, we determined the average mean and variance across all samples within a group for each region separately, giving rise to the region's group methylation level and variance. Next, we computed all pairwise Euclidean differences between this region group methylation level and all samples. Based on this calculation, we computed the mean intra-group and inter-group distance for each region's group methylation level and all samples. Regions exhibiting mean intra-group distance smaller than 0.1 and a mean inter-group distance of larger 0.2 were classified as signature regions for each group of samples. The following groups were used: ectoderm (HUES64d EC, H1d NPC, fetal brain, hippocampus, substantia nigra), neural (H1d NPC, fetal brain, hippocampus, substantia nigra), hESd (hESd EC, hESd ME, hESd EN, H1d mesendo, H1d NPC), ES(HUES64), endoderm(HUES64d EN, liver), mesoderm (HUES64d ME, CD34, CD4, CD8, fetal heart, colonic mucosa, adipocyte nuclei), blood(CD34, CD4, CD8, Bcell, HSPC), adult stem cells (dMesenchyme, dNPC, CD34), fetal (fetal heart, fetal brain, fetal thymus, fetal muscle, fetal adrenal).

4.2.7 Sample deconvolution

To test the discriminatory performance of cell type-specific dynamic regions, we created a hybrid sample from scratch, merging two sequencing lanes from one of our hippocampus WGBS samples and two lanes from one of our HUES64 WGBS samples on the read level. Subsequently, we performed methylation calling using our standard pipeline. Next, we determined the methylation level in this particular hybrid sample for the previously identified dynamic regions. We filtered all dynamic regions according to sample specificity as described previously and used this set for our deconvolution approach. We assume that the hybrid sample's methylation profile h across the reduced dynamic region set is composed of a linear combination of the methylation profiles of all $N = 17$ samples:

$h = \sum_{i=1}^{N} a_i m_i$, with the coefficients a_i defining the scalar mixture proportions. Here, the coefficients a_i are required to fulfill $a_i \leq 0$ and $\sum_{i=1}^{N} a_i = 1$. To find the mixture proportions, we can reformulate the problem as a linear program. However, for simplicity's sake we decided to pose the deconvolution problem as a quadratic optimization problem and incorporate the constraint on the sum of the coefficients into the objective function and optimize this function directly using a quasi-Newton method for bound constrained optimization (L-BFGS-B) as implemented in the R optim (Team, 2012) package.

4.2.8 Genomic features

We obtained TSS, exon and intron coordinates for RefSeq genes from UCSC. Promoters were defined as 2 kb upstream and 500 bp downstream of RefSeq TSS. CpG islands were determined using CgiHunter (http://cgihunter.bioinf.mpi-inf.mpg.de/), requiring a minimum CpG observed vs. expected reatio of 0.6, a minimum GC content of 0.5 and a minimum length of 700 bp. CpG island shores were defined as 2 kb regions directly upstream and downstream of CpG islands. Experimentally determined CpG islands were taken from Illingworth et al. (Illingworth et al., 2010). HCP promoters were defined as RefSeq promoters containing a 500 bp window with a CpG observed over expected ratio > 0.8. Similarly, ICPs were defined as RefSeq promoters with a maximum CpG observed vs expected ratio between 0.45 and 0.8 within a 500 bp window and LCPs everything < 0.457 (Weber et al., 2007).

4.2.9 Transcription factor binding site, enhancer and DNAse I analysis

Processed TFBS .broadPeak files were downloaded from the UCSC ENCODE website (http://genome.ucsc.edu/ENCODE/downloads.html). Replicate experiments were merged conservatively, only retaining peaks present in both replicates. Subsequently, summary TFBS tracks for each individual transcription factor or chromatin modifier were created, merging all peaks from different experimental conditions/cell types. At this stage, we also included four additional TF-ChIP tracks for OCT4, SOX2, NANOG and FOXA2 from GSE46130. These tracks were then used for TFBS analysis. Additionally, we pooled all individual factor tracks into one consensus track, merging overlapping peaks giving rise to a consensus track containing 2,164,835 consensus peaks. In a similar fashion, we acquired and processed the DNAse I hypersensitive site data. For the global analysis of transcription factor binding site enrichment in DMRs we determined the overlap of our DMR set and 100 randomly drawn, size matched control regions sets to estimate the background distribution of TFBS.

Enrichment calculations for all transcription factors and chromatin modifiers (n=165) were performed on the basis of the individual summary tracks for each factor. To determine enrichment and significance of a particular factor within a given class of genomic regions, we generated 100 sets of size matched and mappability filtered control regions for the particular dynamic region class of interest, most comprising millions of regions per group. Next, we determined the number of TFBS per individual TF overlapping with the region class of interest and the control region sets. The ratio of the two is reported as the enrichment statistic shown in Fig. 4.9b and only TFBS results significant at $p < 0.05$ based on the empirical null distribution for the enrichment statistic and with an enrichment ratio greater than one are reported and set to 0 otherwise. In Fig. 4.9b, we report the top three TFBS discovered for all cell type-specific regions across the primary cell type/tissue data set. Enhancer regions were defined using H3K27ac as a proxy for enhancer activity. We obtained 31 H3K27ac profiles from the REMC and ENCODE project (http://www.roadmapepigenomics.org/data, http://genome.ucsc.edu/ENCODE/downloads.html) and used MACS (Zhang et al., 2008) to identify peaks within the individual, replicate merged H3K27ac libraries. We only retained peaks significant below 10^{-6} and merged the individual peak lists afterwards giving rise to $n = 285,344$ distinct putative enhancer regions.

4.2.10 Motif analysis

For each region class of interest, which in our case were primarily dynamic regions specific to a particular sample, we determined all human and mouse motifs contained in the Jolma et al. database (Jolma et al., 2013) (n=843) that are present within our regions. For Fig. 4.10 we additionally included all human and mouse motifs from the TransFac professional database 2009 (n=673) (Matys et al., 2006). Motif matching was performed using FIMO (Grant et al., 2011), retaining only motifs significant below 10^{-5}. As controls, we trained a 0^{th} order hidden Markov model on the input regions and generated ten size distribution-matched sets of control regions that were also subjeced to motif matching. Subsequently, we computed the ratio of motif instances in the region set of interest and the average over the 10 control region sets for each motif separately. With this procedure we control for base composition as well as sequence length. To determine not only motifs enriched over genomic background but also differentially enriched in dynamic regions specific to individual cell types, we determined the enrichment z-scores over all cell type specific regions across the primary cell type/tissiue data set. To correct for samples not enriched over background, we set the z-scores for motifs with an enrichment ratio below 1.2 to 0 in that particular sample. Finally, we rank the motifs by z-score for each of the nine samples individually and plot the top three motifs for each sample.

4.2.11 Cell type specific regions and cohort definition

Cell type-sepcific hypomethyalted regions are defined as exhibiting a Jensen-Shannon divergence (JS) of ≥ 0.15 across the sample set of interest. Each region fulfulling this criterion was assigned to the cell type where it exhibited the lowest methylation level, giving rise to the set of cell type-specific hypomethylated regions.

In our main analysis, we consider four different sample cohorts that divide our data set. The first cohort termed developmental or normal developmental cell types consists of all primary tissues, *in vitro* derived cell types and HUES64 (n=24). The long-term cell culture cohort harbors three cell lines (newborn foreskin fibroblasts, IMR90 and HepG2). The cancer cohort consists of one colon cancer sample. Dynamic CpGs were determined against a matching control of adjacent tissue. The Alzheimer's disease (AD) cohort comprises two AD frontal cortex brain samples derived from different patients with severe AD symptoms. Dynamic CpGs were determined against two control samples from patients of similar age but normal cognitive, molecular and physiological brain parameters.

4.2.12 Dynamic CpG identification algorithm

We identified dynamic CpGs in a multi-step procedure (see Fig. 4.1 for entire workflow).

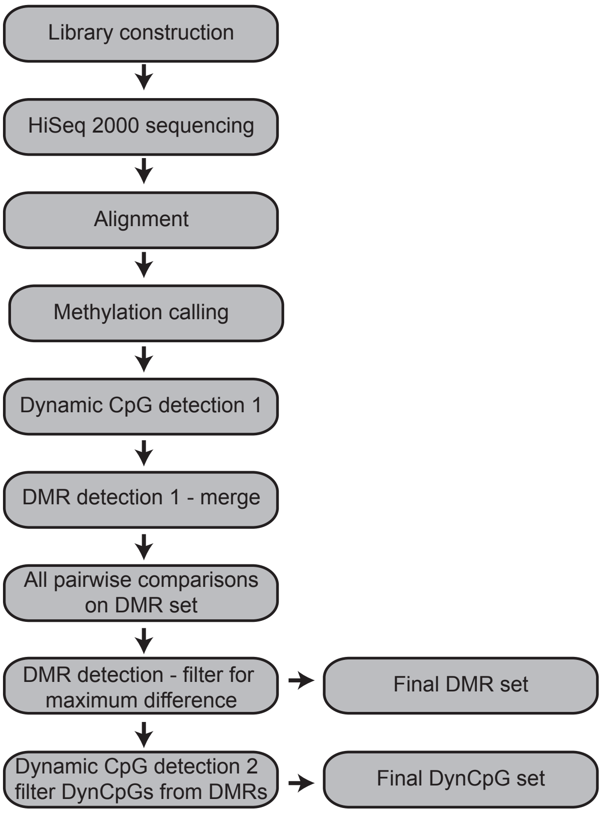


FIGURE 4.1: Complete workflow for methylation data generation and analysis

First, we use our probabilistic description of single CpG methylation levels to determine whether each CpG is differentially methylated based on the beta difference distribution. For this step, only methylation differences significant below ≤ 0.01 are retained. For the next comparisons ($n \geq 3$) we compare the CpG methylation level for each CpG of the new sample to the minimum and maximum methylation levels that have been observed so far based on the previous comparisons. If no comparison for a particular CpG has been significant yet, the initial values are still stored. If in the current comparison a significant deviation from the current minimal/maximal methylation level is found, the current minimum/maximum is replaced if the new estimated methylation difference exceeds the current difference between minimum/maximum. We initialize our comparison algorithm with the CpG methylation values in HUES64 and hippocampus and proceed in random order with the rest of the dataset.

In the next step we filter all dynamic CpGs identified so far and remove all CpGs overlapping HapMap SNPs (Abecasis et al., 2010) or exhibiting a mappability score below 50 (Lee and Schatz, 2012). Next, we merge all dynamic CpGs located within 500 bp of each other to give rise to a dynamic region set. The merging process is limited to a maximum region size of 10 kb, beyond that a new region is started. For CpGs without a neigbhor within 500 bp, we define a genomic interval by extending the dinucleotide region by 50 bp in either direction. Therefore our region catalog size is limited to 100-10000 bp.

In the last step, we compute the CpG cluster methylation differences across the dynamic region set for all pairwise comparisons in our dataset and retain only regions meeting our threshold criteria. The latter were set to minimum effect size ≥ 0.3 and significance level ≤ 0.01 for all main analysis.

4.2.13 Analysis of 450K and tDMR data

To gain further confidence in the identification methodology, we compared our dynamic region set to previously identified tDMRs (Irizarry et al., 2009) as well as DMRs obtained from publically available Illumina 450K array data for 23 distinct cell and tissue types (Nazor et al., 2012). Strikingly, we also recover between $50-75\%$ of previously identified dCpGs (Fig. 4.5a,b). We obtained preprocessed 450K Infinium Bead Chip data from GEO (GSE30654)(Nazor et al., 2012) and filtered for primary cells and tissues contained in this large data set. Next, we averaged individual CpGs over replicates leaving us 23 diverse human samples (Bladder Adult, Lymph Node Adult, Stomach Fetal, Stomach Adult, Blood Adult, Heart Fetal, Heart Adult, Tongue Fetal, Kidney Fetal, Liver Fetal, Brain Fetal, Brain Adult, Pancreas Adult, Thymus Fetal, Spleen Fetal, Spleen Adult, Ureter Adult, Lung Fetal, Lung Adult, Adrenal Fetal, Adrenal Adult, Skeletal Muscle Adult, Adipose Adult).

Subsequently, we determined the number of 450K CpGs contained in our final dynamic region set (n=153,454) and assessed the CpG-wise maximum difference observed across the 450K dataset. To determine to what extent the dynamics observed in the WGBS dataset are recapitulated based on the array data, we computed the fraction of overlapping 450K CpGs that exhibited a methylation level difference larger than a particular threshold and plotted the results in (Fig. 4.5b).

We obtained the coordinates of tDMRs from the publications supplementary website (Irizarry et al., 2009) and performed liftover to hg19 using the UCSC liftover tool prior to overlapping these regions with our dynamic region set.

4.2.14 Sensitivity and specificity calculation

We carried out sensitivity and specificity analysis as a function of coverage by directly downsampling the bam files for our highest coverage sample (hippocampus) to 2, 5, 10, 15, 20, 30 and 40x total coverage. Subsequently, we processed the resulting files with our standard pipeline and applied our canonical dynamic CpG identification approach by comparing the downsampled data sets to HUES64 samples. We defined true positives (TP) as those differentially methylated CpGs found at a given coverage threshold that were also detected at the maximum coverage level. False positive were defined similarly. The true positive rate or sensitivity was then defined as $TPR = TP/(TP + FN)$ and the false-positive rate FPR as $FPR = FP/(FP+TN)$. The results of this analysis are shown in (Fig. 4.3c-e).

4.2.15 FDR estimation

We estimated the false positive rate and false discovery rate based on our two deeply sequenced hippocampus replicates. First, we created ten pseudo replicates subsampling sets of reads from both samples giving rise to new pseudo samples. In order to reflect the coverage diversity in our dataset, we created pseudo samples at distinct total sequencing depths ranging from 10-40x. Next, we ran our standard dynamic CpG identification procedure on this hippocampus sample set of 12 and determined the number of differentially methylated CpGs at distinct minimal difference thresholds. Finally, we extrapolate the number dynamic CpGs discovered in the replicate set to the number of samples used in the main dataset (24), giving rise to (Fig. 4.3a).

4.2.16 Estimation of saturation curve

To get an estimate of the rate of newly discovered dynamic CpGs as a function of number of samples, we randomly select sample subsets from our developmental cohort at different cell type coverage levels (5, 10, 15, 20, 24). For coverage levels 5 and 10 we create 10 distinct sample subset, for 15 cell types five subsets and for 20 cell types two distinct subsets. Subsequently we run our standard dynamic CpG identification procedure on each of the defined subsets and average the number of dynamic CpGs found for each cell type coverage level. Based on these data points we compute a cubic spline fit and plot the result in (Fig. 4.16).

4.2.17 SNP and GWAS analysis

To determine overrepresentation of single nucleotide polymorphisms in the dynamic region set, we determined the number of CEPH HapMap SNPs (Abecasis et al., 2010) (Utah residents with ancestry from northern and western Eruope, CEU) overlapping with the dynamic regions set. First, we tested overrepresentation of SNPs with respect to genomic background. To that we determined the fraction of genomic space spanned by the dynamic regions assuming 89% of the human genome to be uniquely mappable (Lee and Schatz, 2012) and excluding X and Y chromosomes. Next, we used this fraction (19.2% of mappable genome) to determine the expected number of SNPs overlapping with the dynamic regions as $s = n \times p$ with n being the number of all autosomal SNPs and $p = 0.192$. We assessed significance of enrichment using a one-way binomial test using the R function *pbinom* in the standard *stats* package. In addition, we computed size and GC content matched randomly drawn control regions from the human genome, also taking the chromosomal distribution of dynamic regions into account. Fisher's exact test was then used to assess the hypothesis that the SNP counts in the dynamic region and control region classes were independent. In addition, we also generated 100 randomly drawn size-matched sets of control regions and determined the number of overlapping SNPs and the empirical distribution function of the background SNP overlap. All of the control sets contained fewer SNPs than the DMR set giving rise to an empirical p-value ($p < 0.01$) and median odds ratio of 1.48.

The genome wide association analysis (GWAS) was conducted using 5,726 SNPs originally published in the GWAS catalog (Hindorff et al., 2009) and annotated by Maurano et al.

(Maurano et al., 2012) (original release September 2010 and 71 additional liver-related SNPs from several newer GWAS studies (Chambers et al., 2011, Kawaguchi et al., 2012, Patin et al., 2012)). We grouped SNPs into 17 categories following the classification suggested by Maurano et al.

(Maurano et al., 2012). We first tested whether SNPs are overrepresented within dynamic regions compared to HapMap SNPs using a binomial test $b(x; n, p)$. To that end we set the expectation p to the fraction of autosomal HapMap SNPs present within our full dynamic region set, n equal to the number of autosomal GWAS SNPs and x to GWAS SNPs present in dynamic regions. The enrichment statistic is then reported as the fraction of observed vs. expected GWAS SNPs.

For the cell type-specific GWAS enrichment calculation we again focused on cell type-specific hypomethylated regions within our primary cell type/tissue cohort. Similar to the TFBS enrichment calculation, we employed a non-parametric approach based on 100 sets of randomly sampled, size-matched and mappability-filtered control region sets (each comprising mostly millions of regions). We then report the median ratio of GWAS SNPs in the cell type-specific dynamic regions over control regions for each disease class. Associated p-values were computed based on the resulting empirical null distribution for each cell type and each disease class. Our selection of 100 random draws represents a compromise between stringency and computational tractability.

4.3 Results

4.3.1 Identification and characteristics of differentially methylated regions (DMRs) in the human genome

In this study, we systematically investigated the DNA methylation state of most human autosomal CpGs to determine those that show dynamic changes and hence may participate in genome regulation in a developmental context (dynamic CpGs). In total, we included 42 whole-genome bisulphite sequencing (WGBS) data sets, comprising a range of human cell and tissue types ($n = 30$). The combined 40.4 billion reads enabled us to assay 25.71 million autosomal CpGs ($\geq 5\times$ coverage in at least $\geq 50\%$ of all samples; 96% of all hg19 autosomal CpGs). We organized the samples into four classes: human embryonic stem (ES) cells, human ES-cell-derived cell populations, normal somatic tissues, and disease conditions (Fig. 4.2a and the electronic table CH4_DataSet_Table located in the electronic archive accompanying this work within the folder corresponding to Chapter). On a global scale, human ES cells and their derivatives exhibit the highest DNA methylation levels, followed by primary tissues ($\approx 5\%$ less), which is in sharp contrast to the global hypomethylation observed in colon cancer ($\approx 10-15\%$ less) and long-term cultured cell lines (10-30% less).

Focusing initially on our developmental sample set ($n = 24$ total, ES cells, in-vitro-derived cell types and somatic tissues) we identified ≈ 5.6 million dynamic CpGs (minimum methylation difference ≥ 0.3, false discovery rate (FDR) $= 10.4\%$, 21.8% of captured autosomal CpGs; Fig. 4.2b,c and see Section 4.2.12) distributed across 716,087 discrete DMRs (19.2% of the mappable human genome). In addition to this moderately

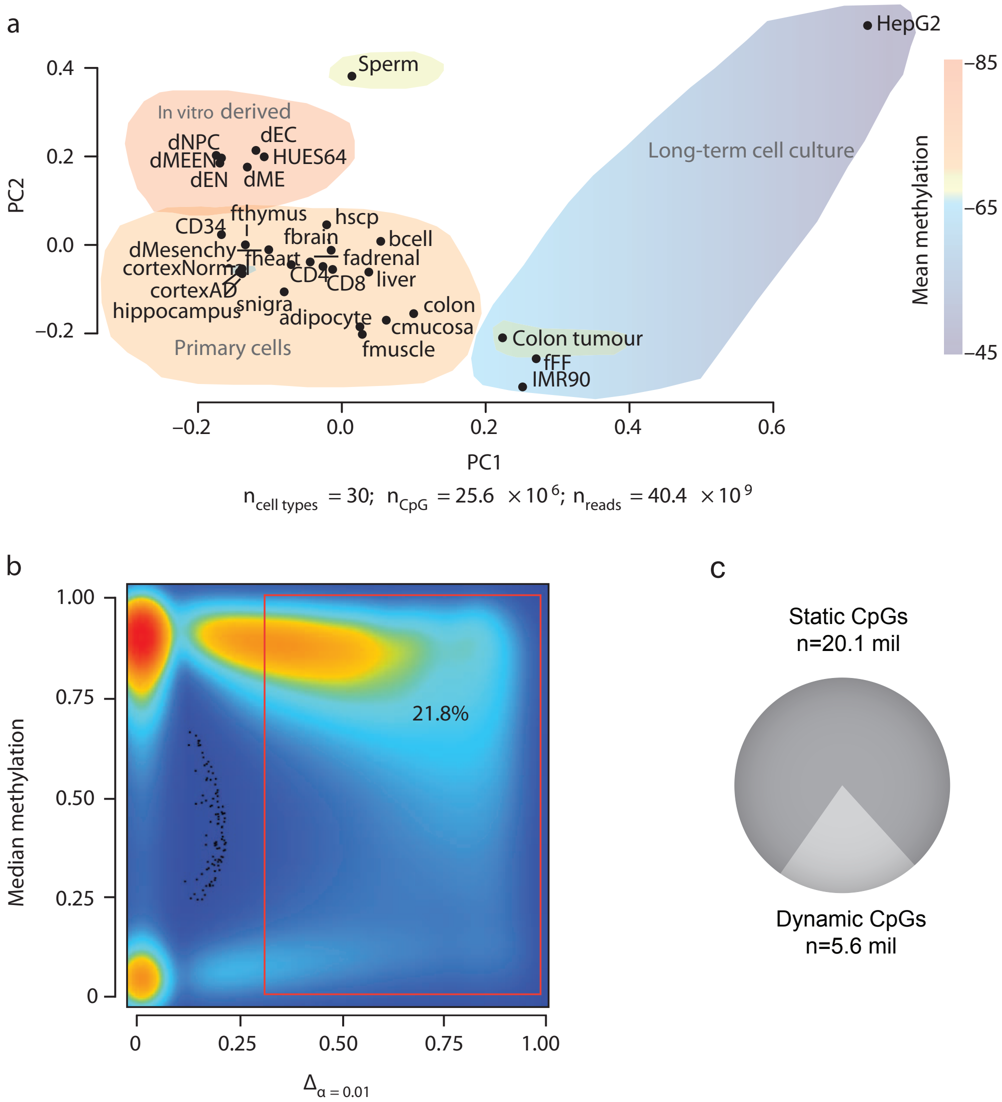


FIGURE 4.2: **a.** Principal component (PC) analysis based on CpG methylation levels for 1-kb tiles across 30 diverse human cell and tissue samples. Coloring indicates classification of samples into subgroups and group-wise mean DNA methylation. Detailed sample annotations are listed in the electronic table CH4_DataSet_Table located in the electronic archive accompanying this work within the folder corresponding to Chapter. Grey area indicates Alzheimer's disease (AD) samples. **b**. Density scatterplot of CpG-wise DNA methylation level differences (x axis, $P \leq 0.01$) and CpG median methylation (y axis) across the 24 developmental samples (excluding cancer and long-term culture). Coloring indicates CpG density from low (blue) to high (red). The red box highlights dynamic CpGs (≥ 0.3). **c.** Pie chart showing the fraction of static and dynamic CpGs.

stringent cut-off, we also tested thresholds as low as 10% methylation difference that may account for DNA methylation changes arising from relevant small subpopulations in heterogeneous tissue samples or noise, but still only find 10.4 million CpGs to be dynamic (Fig. 4.3a). To confirm the validity of our approach, we evaluated the sensitivity and false positive rate as a function of coverage and methylation difference (Fig. 4.3b-e). Focusing on the more stringent set ($\geq$ 0.3 difference), we find approximately

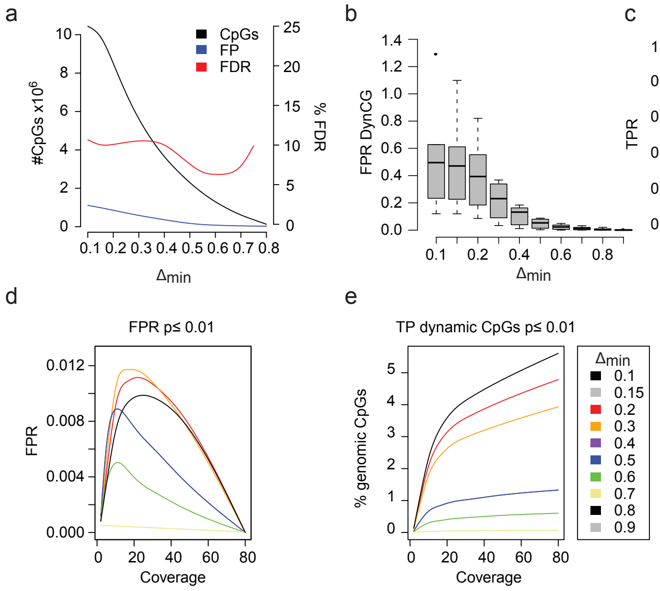


FIGURE 4.3: **a.** Number of detected dynamic CpGs across our developmental samples (n=24) as a function of minimum CpG cluster methylation difference (x-axis black line). **b.** Distribution of false positive rate as a function of minimum CpG cluster methylation difference (x-axis) for 7 individual samples with 2 replicates each. **c.** Sensitivity analysis of dynamic CpGs recovered as a function of coverage. Analysis is based on downsampling of a high coverage hippocampus sample. Differentially methylated CpGs were determined with respect to HUES64. **d.** False positive rate of dynamic CpGs as a function of coverage. Analysis is based on downsampling of a high coverage hippocampus sample. Differentially methylated CpGs were determined with respect to HUES64 (right). **e.** Fraction of genomic CpGs recovered that are differentially methylated between hippocampus and HUES64 as a function of coverage. Analysis for different minimum methylation differences between CpG clusters are shown as colored lines.

70% are on average highly methylated (> 75% methylation ratio), whereas less than 2% are on average unmethylated (< 10% methylation ratio); (Fig. 4.4a). In line with this observation, we find that hypomethylation of DMRs shows greater sample specificity than hypermethylation (Fig. 4.4b).

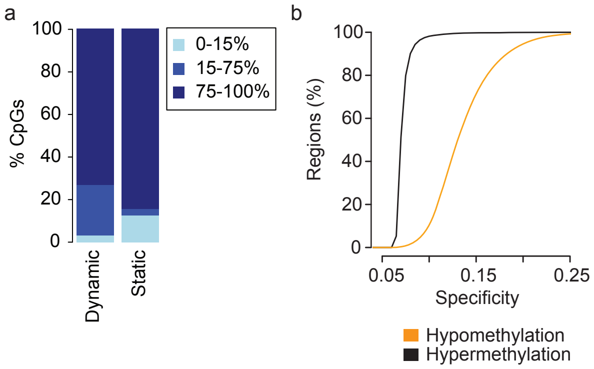


Figure 4.4: **a.** Fraction of median methylation levels across all samples grouped into unmethylated, intermediate and highly methylated for dynamic and static CpGs. **b.** Cumulative distribution of DMR specificity. High hypo/hypermethylation specificity indicates that a particular region is methylated/unmethylated in most tissues and deviates from this default state in only one or a few cases.

Interestingly, most of the DMRs are small ($> 75\%$ are smaller than 1 kb; Fig. 4.5a) and located distal to transcription start sites (Fig. 4.5b).

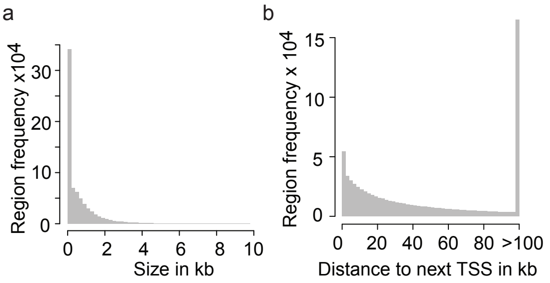


Figure 4.5: **a.** Size range distribution for all DMRs. **b.** Distance distribution of DMRs to the closest ENSEMBL TSS truncated at 100 kb.

However, the average variation in DNA methylation levels across all RefSeq promoters (n = 30,090) does still exhibit a clear increase specifically at the transcription start sites, with most of this variation occurring at intermediate and low CpG density promoters (Fig. 4.6). For CpG islands in general, we observe distinct dynamic regimes, highlighting that different classes of CpG islands are probably subject to different modes of regulation (Cohen et al., 2011, Lienert et al., 2011, Meissner et al., 2008)(Fig. 4.6, bottom). Consistent with previous reports (Irizarry et al., 2009), we find CpG island shores (regions within 2 kb of an island) to be among the most variable genomic regions (Fig. 4.6, bottom).

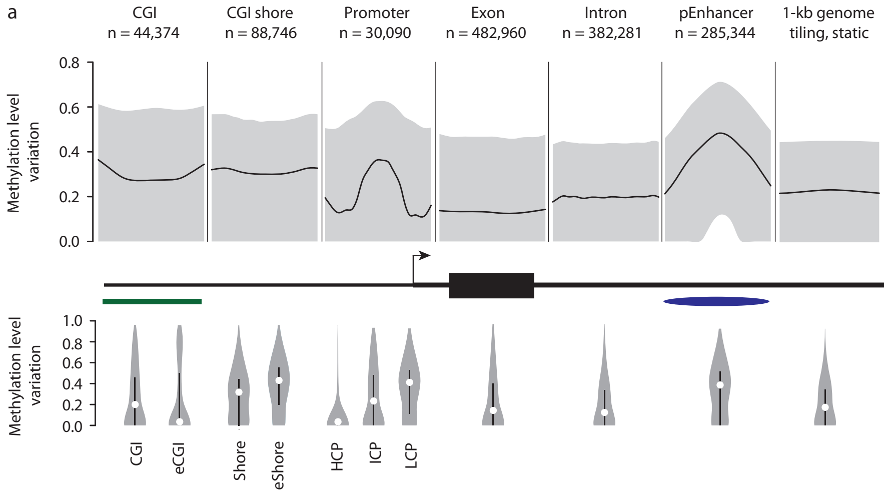


FIGURE 4.6: Top, composite plot of mean DNA methylation differences across various genomic features. Black lines indicate the median of the average DNA methylation difference across each feature. Grey areas mark twenty-fifth and seventy-fifth percentiles. Bottom, distribution of mean DNA methylation difference for each genomic feature. Black bars indicate twenty-fifth and seventy-fifth percentiles; white dots mark the median. For CGI islands, a smaller, experimentally determined set (eCGI; n=525,490) is also shown. Promoters are broken down into high CpG content (HCP; n=524,899), intermediate CpG content (ICP; n=510,920) and low CpG content (LCP; n=57,946) regions (n=543,765 total). Shore denotes regions within 2 kb of an island; eShore denotes experimentally determined shore. pEnhancer, putative enhancer.

These observations are exemplified at the OCT4 locus (also known as POU5F1), in which the promoter and large parts of the gene body exhibit high DNA methylation dynamics, whereas the strong downstream CpG island as well as the surrounding CTCF-binding sites remain static (Fig. 4.7a). Only 12.2% of our DMR set overlap with at least one of 568,430 annotated classic, gene-centric genomic features (promoter, exon, CpG island (CGI) , CGI-shore) (Fig. 4.7b). To gain insights into the role of the remaining set, we first investigated their co-localization with DNase I hypersensitive sites across 92 distinct cell type (Thurman et al., 2012) as well as a catalogue of putative enhancer elements for 31 cell and tissue types (Zhu et al., 2013). Notably, we found that 42.3% of our DMRs overlap with at least one DNase I hypersensitive site (Fig. 4.7b), and 26.1% co-localize with enhancer-like regions, which cover more than 50% of all H3K27ac regions in our catalogue (n = 285,344) and represent one of the most differentially methylated features (Fig. 4.6).

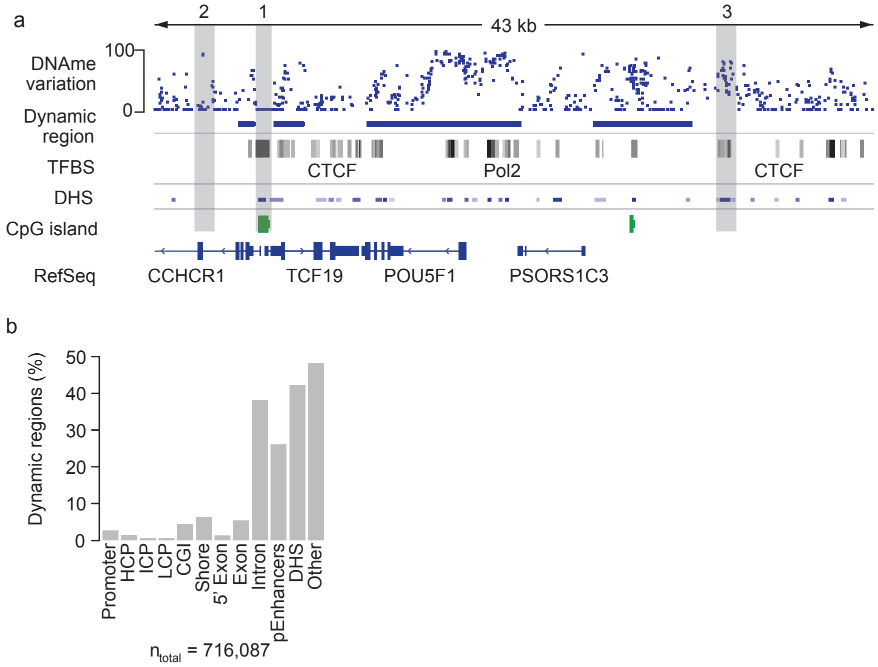


FIGURE 4.7: **a.** Methylation level variation across the OCT4 locus (chr6: 31,119,000-31,162,000) (top). Blue bars indicate significant DMRs at $P \leq 0.01$, and exhibit a minimum difference ≥ 0.3 across the 24 developmental samples. Grey boxes (1-3) are examples of regions that are static (1 and 2) or that do not meet the threshold of dynamic (3). For reference, ENCODE TFBS cluster track,DNase I hypersensitive sites, CpG islands and RefSeq genes are shown. DNAme, DNA methylation. **b.** Distribution of DMRs across various genomic features. Each region is assigned to only one of these genomic features according to the ranking promoter, CGI, CGI shore, 5' exon, exon, intron, putative enhancers, DNase I hypersensitive site (DHS) or other.

4.3.2 Dynamic CpG methylation regions frequently co-localize with transcription factor binding sites (TFBS)

Next, we examined DMR overlap with transcription-factor-binding site (TFBS) clusters compiled from 165 transcription factors profiled by the ENCODE project (Gerstein et al., 2012) and uncovered a highly significant overlap of the two feature classes (odds ratio 1.14, $p \leq 0.01$ empirical test). Interestingly, we find that more than 50% of all DMRs overlap with at least one and 25% with more than three TFBSs, accounting for an additional 13.0% of DMRs (Fig. 4.8a). Consistent with this, we find markedly increased variation in DNA methylation levels specifically across TFBSs (Fig. 4.8b).

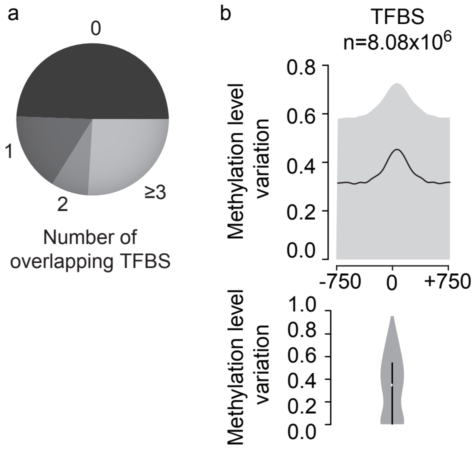


Figure 4.8: **a.** Overlap of DMRs with ENCODE TFBSs. **b.** Average of the maximal observed variation in DNAme levels across a 1.5 kb region centered at the middle of each ChIP-Seq peak (±750 bp) across 161 TFBS obtained from the ENCODE project (top). Distribution of the median methylation variation across the 1.5 kb region centered at the TF ChIP-Seq peak.

In summary, we were able to attribute 64.2% of all DMRs to at least one putative gene regulatory element or coding sequence, suggesting that they demarcate various classes of putative regulatory elements.

We determined all cell-type-specific hypomethylated regions (n = 396,995; see Section 4.2.4) and investigated the enrichment for 161 ENCODE factors (excluding MBD4, SETDB1, POL2P and HDAC2 from the previous set). Notably, we observe significant enrichment of cell-type-specific transcription factors that are known to be involved in the regulation of the respective cellular states (Fig. 4.9). For instance, the top three factors bound in HUES64-specific DMRs are OCT4, SOX2 and NANOG (Fig. 4.9).

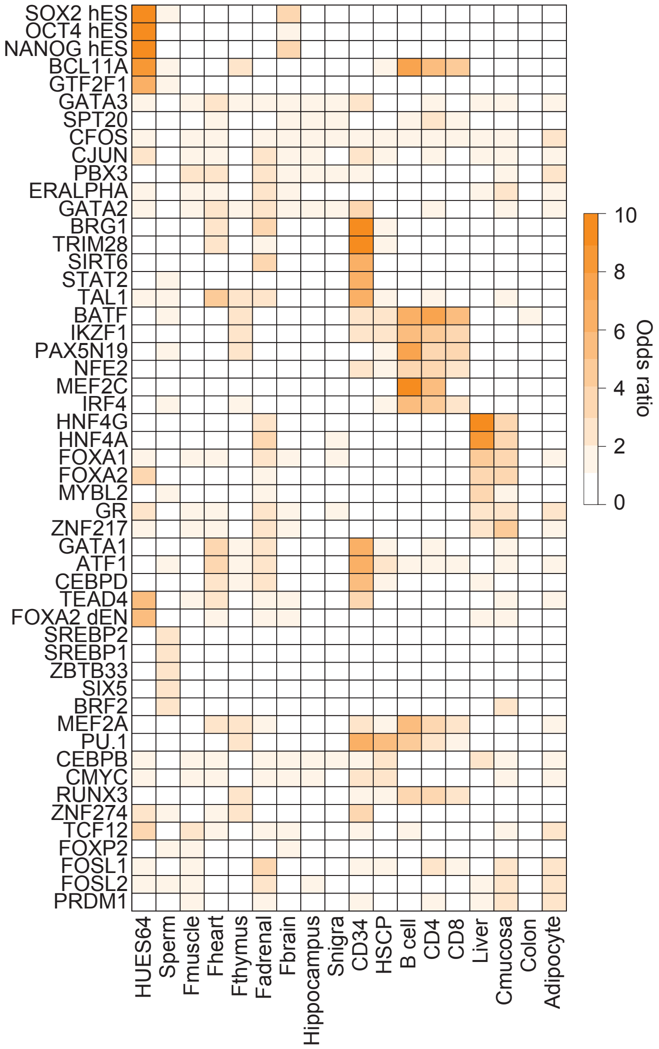


FIGURE 4.9: Enrichment of the top four TFBSs significantly overrepresented ($p < 0.01$, empirical test) in DMRs specific to the cell type indicated (specificity > 0.15). Colour code quantifies median enrichment odds ratio compared to size-matched random control regions.

Similarly, PU.1 and TAL1 are highly enriched in CD34 cells and hepatocyte nuclear factors in adult liver (Fig. 4.9). In further support of this, motif enrichment analysis revealed many more interesting cell-type-specific transcription factor associations, such as enrichment of distinct NKX factors in fetal heart and brain, and ESRRG in fetal adrenal cells (Fig. 4.10).

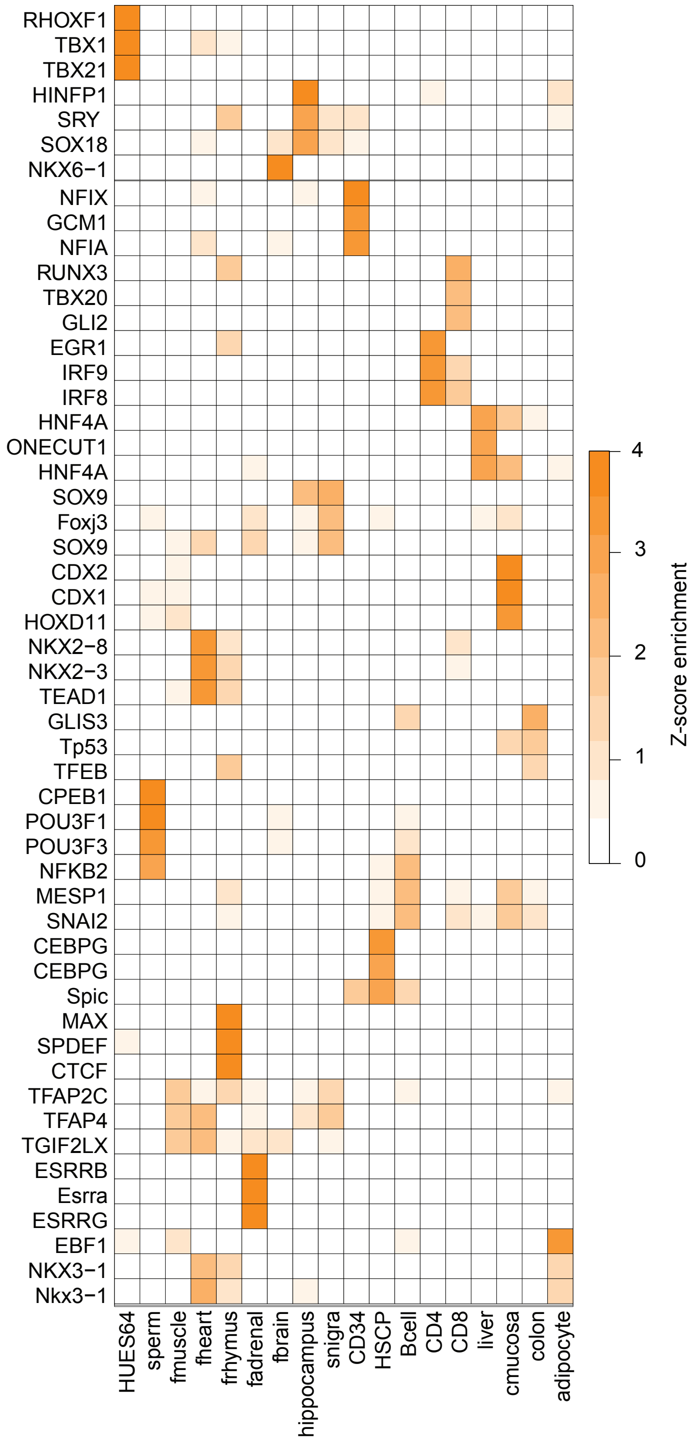


FIGURE 4.10: Normalized motif enrichment z-scores for the top three enriched motifs for DMRs specific to each of the selected somatic samples indicated on the left. For enrichment levels of full motif library was used (n=843).

Moreover, we tested whether the DMR set can be used to gain insights into the combinatorial control of cellular states by transcription factors. To that end, we determined all unmethylated (< 10% methylation) PAX5 motif instances (±100 base pairs (bp)) across the human genome in CD34 or fetal brain cells (Fig. 4.11a). Although both footprint sets show a large overlap (11,031 sites), regions exclusively unmethylated in CD34 or fetal brain are enriched for distinct sets of other known lineage-specific transcription factor motifs; such as PU.1 in CD34 and LMX1A or EN1 in fetal brain (Fig. 4.11a).

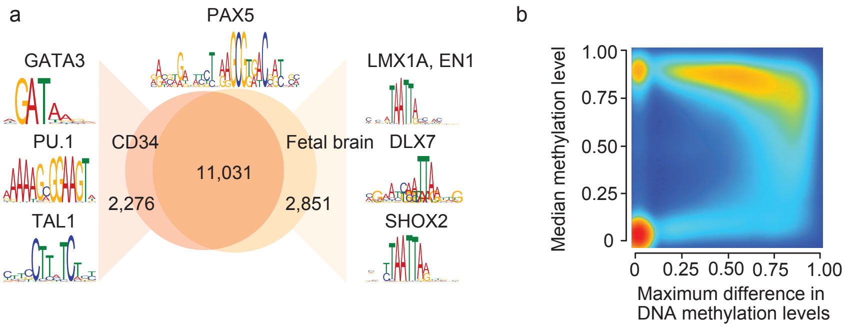


FIGURE 4.11: **a.** Overlap of PAX5 motifs (±100 bp top) unmethylated in CD34 cells or fetal brain across the entire human genome. Regions specifically unmethylated in CD34 or fetal brain were subjected to motif analysis, and top differentially co-occurring motifs are highlighted on the left for CD34 and on the right for fetal brain. **b.** Density scatterplot of maximum DNA methylation difference across 24 developmental samples for TFBS cluster track (n = 2.7 million) and median methylation level across all samples. Colour code indicates density of TFBSs from low (blue) to high (red).

Taken together, these findings highlight that cell-type-specific DNA methylation patterns can be used to detect footprints and infer potentially regulatory transcription factors. In fact, more than 60% of all ENCODE TFBSs are hypermethylated in most samples, but become hypomethylated very specifically in only one or two cell types (Fig. 4.11b), whereas 25% are constitutively unmethylated and never change (Fig. 4.11b). Breaking down this distribution of TFBSs reveals distinct patterns of variation for different types of transcription factor (Fig. 4.12a). More generally, we find that DNA methylation variation across TFBSs is strongly correlated with its median methylation level and therefore the (hypo-)methylation specificity (Fig. 4.12b), as well as the tissue specificity of transcription factor expression patterns (Ravasi et al., 2010) (Fig. 4.12c). These observations support the notion (Stadler et al., 2011) that selective transcription factor binding creates spatially highly constrained hypomethylated regions and confers cell type-specificity.

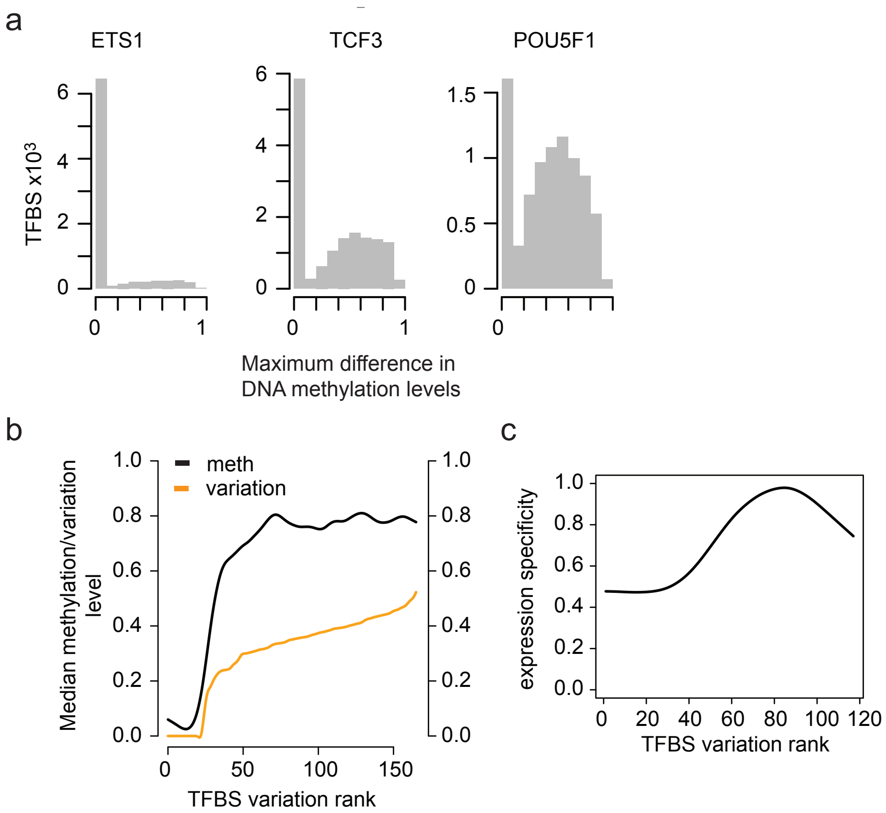


FIGURE 4.12: **a.** Median variation (y-axis left) and median methylation level (y-axis right) across all TFBS for 165 transcription factors profiled by ENCODE. Factors (x-axis) were ordered by increasing maximum median methylation level variation. Curves were determined by cubic-spline fit to the median variation and methylation levels. **b.** Cubic-spline fitted transcription factor expression specificity (see main text) as a function of TFBS methylation variation rank. **c.** Frequency distribution of maximum difference in DNA methylation levels for three selected TFs.

4.3.3 DMRs exhibit elevated SNP frequency and show non-random GWAS SNP enrichment

On the basis of these findings and previous reports (Maurano et al., 2012), we asked whether DMRs are more susceptible to point mutations that are functionally consequential. Even with strict filtering criteria, we found a significant enrichment of single nucleotide polymorphisms (SNPs) in DMRs compared to genomic background as well as different sets of random control regions (odds ratio 1.06, $P < 10^{-16}$, binomial test). We then determined the overlap of DMRs with recently evolved human-specific CpGs, termed CpG beacons (Bell et al., 2012), which shows a marked enrichment (odds ratio 1.37-1.6 compared to genomic background and random control regions, $P < 10^{-16}$). This suggests overall higher genetic intra-species variability specifically at regions that change their DNA methylation state. In accordance with the increased SNP frequency, DMRs are also significantly enriched for genome-wide association study (GWAS) SNPs from the GWAS catalog (Hindorff et al., 2009) (odds ratio 1.16, $P = 3.27 \times 10^{-10}$, binomial test). Similar to our observations on TFBSs, GWAS SNPs exhibit a non-random enrichment distribution across cell-type-specific DMRs (Fig. 4.13). For instance, we find DMRs specific to adult liver to be enriched for liver and serum metabolite-related GWAS SNPs, fetal heart DMRs enriched for cardiovascular disease SNPs and many of our blood cell type DMRs enriched for autoimmune disease and haematological parameter related SNPs.

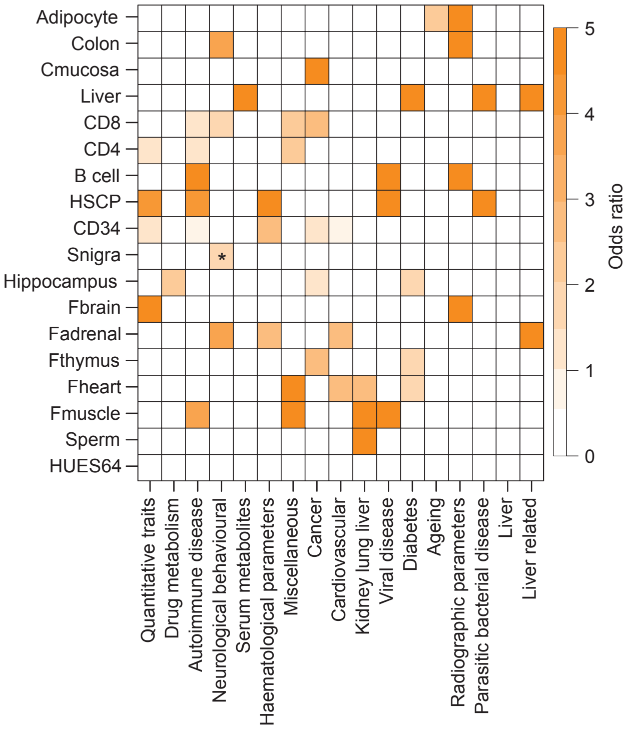


FIGURE 4.13: Odds ratio of significantly overrepresented ($p < 0.05$, empirical test, see Section 4.2.17) GWAS SNPs grouped into 16 categories in regions specifically hypomethylated within the sample indicated on the left. * $p < 0.1$

4.3.4 Effective classification and sample deconvolution using only the DMR set

It is well known that many cancers exhibit considerable DNA methylation changes (Ehrlich, 2009), we therefore compared a colon cancer to a matched control and found 532,665 differentially methylated CpGs. Forty % of these overlapped with the previously identified developmental dynamic set (Fig. 4.14a). Similarly, 36% of differentially methylated CpGs found in Alzheimer's disease samples compared to normal controls (n = 12,408) overlapped with our previous set of developmental CpGs. The most notable change in the number of dynamic CpGs occurs when comparing our developmental sample cohort to the long-term cell culture cohort, leading to the identification of 8.4 million additional dynamic CpGs (Fig. 4.14b). Importantly, this expanded set differs notably in terms of their sequence features, with cancer and Alzheimer's disease dynamic CpGs residing in less conserved regions that also exhibit lower motif complexity compared to the developmental and cell culture (Fig. 4.15a, b). The cell-culture-specific CpGs exhibit increased repeat content relative to developmental CpGs, a feature that is shared with Alzheimer's disease (Fig. 4.14c). Although the disease samples clearly add more dynamic CpGs, our analysis suggests a notable overlap with our previous set for CpGs that may participate in actual regulatory events.

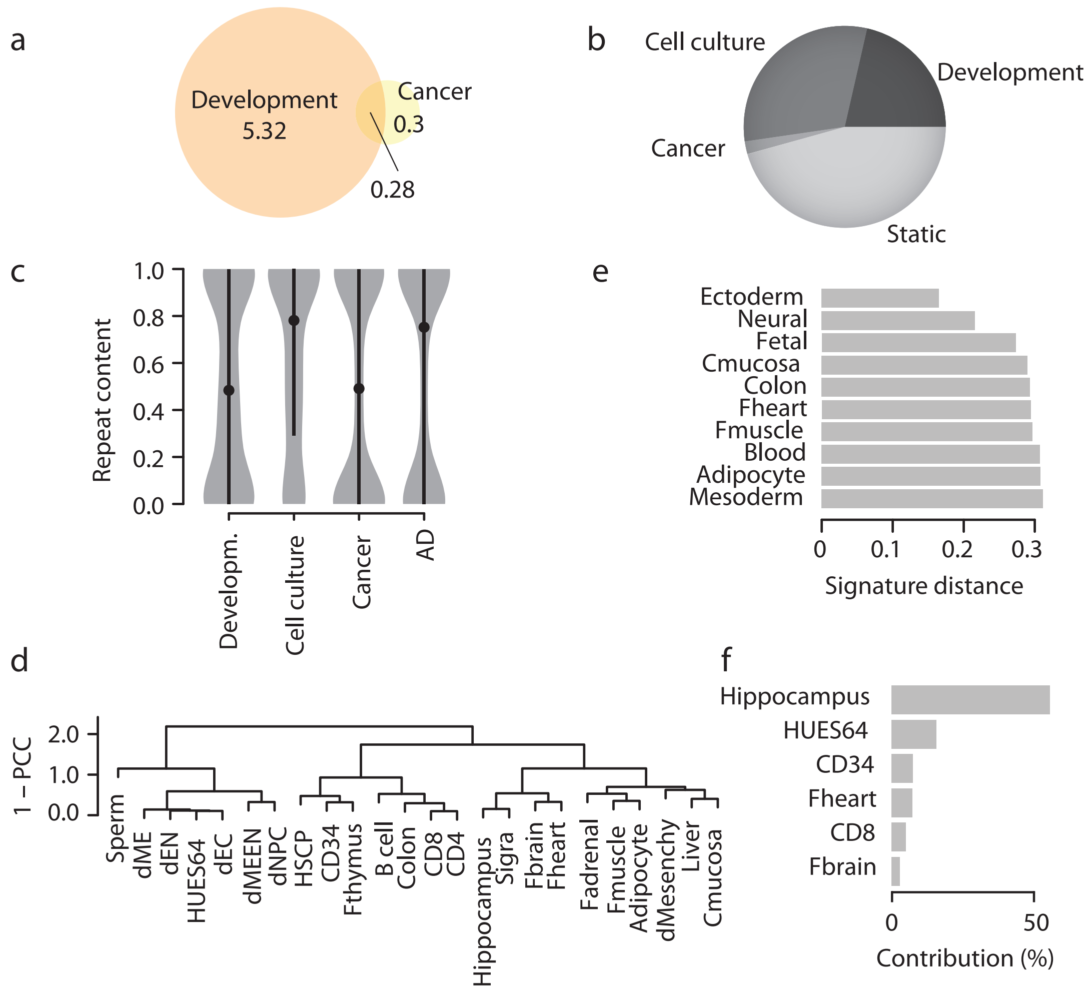


FIGURE 4.14: **a.** Overlap of dynamic CpGs (P≤ 0.01; |methylation| ≥ 0.3 in normal samples and between colon cancer and matching control CpG numbers (in millions). **b.** Distribution of autosomal CpGs across three conditions. Class name indicates sample group in which a CpG was observed dynamic (developmental (n = 24), cell culture (n = 3), cancer (n = 2)) or remained unchanged over the entire sample set (n = 30). **c.** Repeat content distribution of DMRs (sets as in b). AD, Alzheimer's disease. **d.** Hierarchical clustering using Pearson correlation coefficient (PCC) of the DMR values across the entire sample set (n = 30). **e.** Distance of the fetal brain sample to different sets of signature regions defined for sample classes or individual samples, but excluding regions identified by means of the fetal brain sample. **f.** Contribution of individual sample signature region sets to an in-silico-generated hybrid sample (HUES64 and hippocampus).

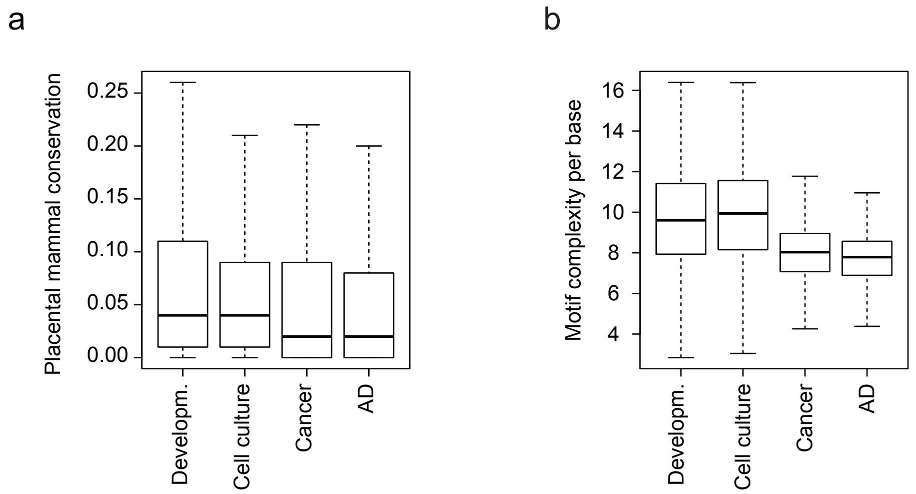


FIGURE 4.15: **a.** Distribution of average region conservation scores for dynamic CpG sets defined based on the developmental CpG, the cell culture cohort, cancer and the AD cohort. **b.** Distribution of motif complexity per base for dynamic CpG sets discovered based on the developmental CpG, the cell culture cohort, cancer and the AD cohort. Motif complexity is defined as sum over all motif occurrences within a region set. Each motif occurrence is thereby weighted by its complexity.

Finally, we investigated the utility and power of the reduced region set to accurately classify unknown samples or help to deconvolute a mixture of samples. We first clustered our developmental sample set based on the DMRs only (Fig. 4.14d) and found the result to be in excellent agreement with genome-wide 1-kb tiling-based clustering. To probe the potential of our DMR set to classify unknown samples accurately, we derived signature region sets for different sample groups. These signature regions turned out to be excellent classifiers of an unseen sample (Fig. 4.14e, fetal brain). Next, we tested as a proof of principle whether it is possible to use our DMR set to infer the different cell populations present within a heterogeneous sample. To that end, we deconvoluted an in silico mixture of HUES64 and hippocampus WGBS libraries using our DNA methylation signatures. Notably, the two top hits after application of a very simple deconvolution algorithm indeed proved to be hippocampus and HUES64 (Fig. 4.14f).

4.4 Discussion

Our study highlights and defines a relatively small subset of all genomic CpGs that change their DNA methylation state across a large number of representative cell types. Although we expect that number to increase with more diverse cell types as more WGBS data sets becoming available, our analysis suggests that the rate of newly discovered regulatory CpGs will drop rapidly once all major cell and tissue types have been mapped, mostly owed to the fact that between tissue variability exceeds within tissue variability by one order of magnitude (Fig. 4.16a, b). Future studies are likely to fine-map dynamics occurring in more specific subpopulations, giving rise to smaller changes in DNA methylation that we were unable to detect or include because of power constraints. Extreme conditions in vitro or in vivo such as loss or misregulation of DNMT1 may affect a larger subset including many intergenic CpGs that are generally static, but most of these additional CpGs are unlikely to overlap with functional elements such as TFBSs or enhancers. In combination with the fact that sequencing of WGBS libraries is very inefficient, as about 65% of all 101-bp reads in our set did not even contain any CpGs to begin with, were PCR duplicates or didn't pass the quality control, this amounts to an approximate combined loss of around 80% of sequencing depth on non-informative reads and static regions. Furthermore, once defined, it will probably be sufficient in most cases to profile only a representative subset of CpGs across a comprehensive set of DMRs using an array-based (Dedeurwaerder et al., 2011) or hybrid-capture-based (Gnirke et al., 2009) technology to recover representative dynamics and measure regulatory events. Using these results as a guiding principle, we expect further improved efficiencies in mapping DNA methylation and enhance its applicability as a marker for various regulatory dynamics in normal and disease phenotypes.

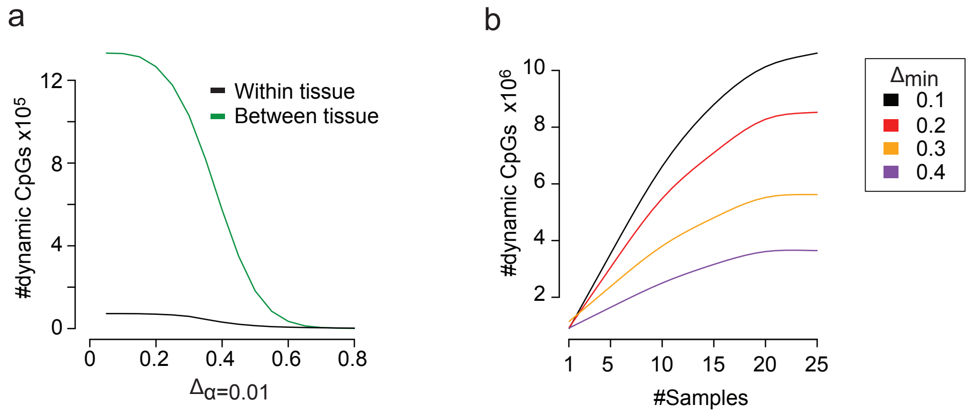


FIGURE 4.16: **a.** Number of differentially methylated CpGs ($p \leq 0.01$) between hippocampus and substantia nigra (black, within tissue variance) as well as betweenhippocampus and liver (green, between tissue variance) for different thresholds of minimal methylation difference. **b.** Estimated saturation curve for thediscovery of dynamic CpGs as a function of sample number. Estimates for different minimal methylation thresholds ($\delta_{\min}$) at a significance level of $p \leq 0.01$ are shown. See Section 4.2.5 for details.

However, DNA methylation changes are frequently accompanied by changes in overall chromatin and transcriptional landscapes. In order to understand the relationship between distinct epigenetic remodeling events as well as their impact on transcription it is essential to investigate the coordinated dynamics along these distinct dimensions in the same cell populations. In addition, it is important to not only examine distantly related cell types that reside in steady states but also closely related cell populations representing transient cellular states, a situation frequently encountered in development.

Chapter 5

Transcriptional and epigenetic dynamics during specification of human embryonic stem cells

5.1 Introduction

The content of this chapter has previously been published in Gifford & Ziller et al. 2013 (Gifford et al., 2013). This project was a joint research effort with equal contributions by Casey A. Gifford (leading wet lab scientist) and Michael J. Ziller (leading computational scientist). While most wet lab experiments including cell culture, FACS and next-generation sequencing library production (ChIP-Seq and RNA-Seq) were carried out by C.A. Gifford, computational analysis and design of analysis strategy was contributed by M.J. Ziller. Differential splicing analysis was contributed by C. Trapnell. M.J. Ziller and C.A. Gifford interpreted the data and wrote the paper together.

In order to gain a deeper understanding of the interplay between transcription, epigenetic remodeling and transcription factor binding, we implemented an *in vitro* differentiation system employing human ES cells. To overcome limitations of previous studies that either utilized distantly related cell types (Bernstein et al., 2012, Lister et al., 2009) or highly heterogeneous differentiation conditions (Laurent et al., 2010, Lister et al., 2011), we employed directed differentiation for five days into three distinct populations using defined culture conditions combined with FACS sorting. In contrast to earlier studies, this approach enabled us to investigate epigenetic and transcriptional dynamics that might only arise transiently in the differentiation process, since the examined populations represent transient cellular states at time of collection.

Coordinated changes to the epigenome are essential for lineage specification and maintenance of cellular identity. DNAme and certain histone modifications critically contribute to epigenetic maintenance of chromatin structures and gene expression programs (Smith and Meissner, 2013, Zhou et al., 2011). Genetic deletion of histone methyltransferases and the catalytically active DNA methyltransferases are embryonic or postnatally lethal (Li, 2002) providing evidence for their essential role in proper execution of developmental programs. Several groups have reported genome-wide maps of chromatin and DNAme in pluripotent and differentiated cell types. From these efforts, a global picture of the architecture and regulatory dynamics is beginning to emerge. For example, active promoters generally contain modifications such as H3K4me3 and H3K27ac, while active enhancers are generally enriched for H3K4me1 and H3K27ac (Creyghton et al., 2010, Ernst and Kellis, 2010, Heintzman et al., 2009, Rada-Iglesias et al., 2011). Repressed loci exhibit enrichment for H3K27me3, H3K9me2/3, DNAme, or a combination of the latter two modifications. The enrichment of repressive histone modifications, such as H3K27me3, which is initiated at CpG islands (CGI), is considered a facultative state of repression, while DNAme is generally considered a more stable form of epigenetic silencing (Smith and Meissner, 2013).

Recent studies have reported dynamics that suggest epigenetic priming such as the appearance of euchromatic histone modifications prior to gene activation during *in vitro* T-cell differentiation (Zhang et al., 2012) and cardiac differentiation (Wamstad et al., 2012). These results are reminiscent of changes that occur during the early stages of reprogramming towards the induced pluripotent state (Koche et al., 2011) and highlight possible similarities between differentiation and de-differentiation. In parallel to these advances, whole-genome bisulfite sequencing (WGBS) has been used to map DNAme genome-wide. Examination of WGBS data from murine ES cells (mESCs) and neural progenitor cells highlighted lowly methylated regions (LMRs) at distal sites that frequently overlap with DNAse I hypersensitive sites (HS) and/or displayed an enhancer signature defined by H3K4me1 and p300 enrichment (Stadler et al., 2011).

Studying the role of epigenetic modifications in the dynamic rewiring of human transcriptional programs in vivo is complicated by numerous technical and ethical limitations. However, models for in vitro differentiation of hES cells offer a unique opportunity to explore and characterize critical events that prepare, guide and possibly regulate cell fate decisions. Populations representing each embryonic germ layer have been produced from human embryonic stem cells (hESCs) (Chen et al., 2012, Kriks et al., 2011, Wei et al., 2012).

To dissect the early transcriptional and epigenetic events during hESC specification, we used two-dimensional, directed differentiation of hESCs to produce representative populations from the three germ layers, namely ectoderm, mesoderm, and endoderm (Evseenko et al., 2010, Hay et al., 2008, Lee et al., 2010) followed by fluorescence-activated cell sorting (FACS) to enrich for the desired differentiated populations. These three cell types, in addition to undifferentiated hESCs (HUES64), were then subjected to ChIP-Seq for four histone marks (H3K4me1, H3K4me3, H3K27me3, H3K27ac), WGBS and RNA-Sequencing (RNA-Seq). To complement this data, we also performed ChIP-Seq for three TFs (OCT4, SOX2, NANOG) in the undifferentiated hESCs, as well as ChIP-BS-Seq for FOXA2 in the endoderm population. The combined data sets provide a wealth of information, including holistic views of transcriptional and epigenetic dynamics that help further dissect the molecular events during human germ layer specification.

5.2 Experimental methods

Human ES cell line HUES64 for differentiated for five days into cell populations from the three proper germ layers by inhibiting TGFbeta, WNT and BMP signaling for ectoderm, addition of ACTIVIN A, BMP4, VEGF and FGF2 for mesoderm and ACTIVIN A and WNT3A for endoderm. Cells were isolated by FACS using germ layer specific surface markers. RNA-Seq, ChIP-Seq and WGBS was then carried out on the purified populations as described in detail in (Gifford et al., 2013) and Appendix A.1.3

5.3 Computational methods

5.3.1 RNA-Seq data analysis

To identify a gene or transcript as differentially expressed (DE), Cuffdiff 2 tests the observed log-fold change in its expression against the null hypothesis of no change (i.e. the true log-fold change is zero). Because of measurement error, technical variability, and cross-replicate biological variability might result in an observed log-fold-change that is nonzero, Cuffdiff assesses significance using a model of variability in the log-fold-change under the null hypothesis. This model is described in detail by Trapnell et al. (Trapnell et al., 2013). Briefly, Cuffdiff 2 constructs, for each condition, a table that predicts how much variance there is in the number of reads originating from a gene or transcript. The table is keyed by the average reads across replicates, so to look up the variance for a transcript using the table, Cuffdiff estimates the reads originating from that transcript, and then queries the table to retrieve the variance for that number of reads. Cuffdiff 2 then accounts for read mapping and assignment uncertainty by simulating probabilistic assignment of the reads mapping to a locus to the splice isoforms for that locus. At the end of the estimation procedure, Cuffdiff 2 obtains an estimate of the number of reads that originated from each gene and transcript, along with variances in those estimates. The read counts are reported along with fragments per kilobase per million rads sequenced (FPKM) values and their variances. Change in expression is reported as the log fold change in FPKM, and the FPKM variances allow the program to estimate the variance in the log-fold-change itself. Naturally, a gene that has highly variable expression will have a highly variable log-fold-change between two conditions.

The modifications made to Cuffdiff 2 improve sensitivity in calling DE genes and transcripts while maintaining a low false positive rate. They stem from the method used to calculate the variability in the log fold change in expression. In Trapnell et al. , Cuffdiff 2 used the delta method to estimate the variance of the log fold change estimate for a gene or transcript. This method yields a simple equation that takes as input the mean and variance of the transcript's expression in two conditions and produces a variance for the log fold change. However, the equation contains no explicit accounting for the number of replicates used to produce those estimates - they are assumed to be perfectly accurate.

The improved version of Cuffdiff 2 more accurately estimates the variance in the log-fold-change using simulated draws from the model of variance in expression for each of the two conditions. Imagine an experiment that has n replicates in condition A and m replicates in condition B. To estimate the distribution of the log-fold change in expression for a gene G under the null hypothesis, Cuffdiff first draws n times from the distribution of expression of G according to the algorithm's model of expression. Cuffdiff then takes the average of the n draws to obtain an expression ”measurement”. Cuffdiff then takes the log ratio of these averages, places this value in a list, and then repeats the procedure until there are thousands of such log-fold-change samples in the list. The software then makes a similar list, this time using the expression model for condition B - the null hypothesis assumes both sets of replicates originate from the same condition, but we do not know whether A or B is the better representative of that condition, so we must draw samples from both and combine them. To calculate a p-value of observing the real log-fold-change under this null model, we simply sort all the samples and count how many of them are more extreme than the log fold change we actually saw in the real data. This number divided by the total number of draws is our estimate for the p-value.

Cuffdiff 2 reports not only genes and transcripts that are significantly differentially expressed between conditions, but also groups of transcripts (i.e. the isoforms of a gene) that show significant changes in expression relative to one another. The test for this is similar to what is described in Trapnell et al. , but comparably modified along the lines described above for single genes or transcripts. Draws of expression are made for each transcript in a group according to the number of replicates in the experiment. These are averaged, and the shift in relative transcript abundance for the draw is made using the Jensen-Shannon metric. These draws are added to a list and used to calculate p-values for significance of observed shifts in relative abundance under the null hypothesis.

Clustering of gene expression profiles was achieved with the csDendro() function from CummeRbund (http://compbio.mit.edu/cummeRbund/). This function first transforms the FPKMs of all genes in each sample by adding one and then takes the logarithm. Next, it converts each genes transformed expression into a fraction of the total transformed expression. The distances between these transformed expression profiles are then measured by the Jensen-Shannon metric. The distances are then used to build a dendrogram via complete linkage hierarchical clustering using the R function hclust().

5.3.2 WGBS data analysis

To ensure comparability of region DNAme levels across all samples, only CpGs covered by $\geq 5x$ in 85% of the samples qualified for the computation of region DNAme levels. To assess the DNAme state of various genomic regions, we resorted to our previously published protocol estimating a genomic region's methylation state as the coverage weighted average across all CpGs within each region. Subsequently, we averaged a region's DNAme level over replicates. DMRs were defined as exhibiting significantly ($p \leq 0.05$, Fisher's exact test) different DNAme levels of at least 0.1.

Many gene regulatory elements (GREs) are marked by spatially highly constrained reduced DNAme levels. It has recently been suggested that besides CpG islands, which are mostly unmethylated (UMR) a second class of GRE is marked by low to intermediate DNAme (IMR) (Stadler et al., 2011). We reasoned that these regions might be of particular regulatory importance in our system and might be missed by looking at histone modification enrichments alone. Therefore we adopted a similar Hidden Markov model approach as proposed in Stadler et al. to identify regions of reduced DNAme level. Briefly, we utilized a three-state Hidden Markov Model operating on the methylation levels of each CpG in the human genome. Each state's emission probabilities for the DNAme levels were modeled by a normal distribution. The model was trained on all CpGs of chromosome 19 in the HUES64 dataset using an adaption of the well-known Baum-Welch algorithm to incorporate the normal distribution (Press, 2007). After initial parameter estimation, we utilized the approach reported by Stadler et al. (Stadler et al., 2011) to determine the FDR for IMR regions and adapted the initial parameter estimates for the IMR and HMR states to finally 0.01(UMR), 28.8 (IMR), 81.6 (highly methylated, HMR), yielding an FDR of 2%. This parameter set was subsequently used to segment all WGBS datasets. Finally, we used the Viterbi algorithm to compute the most probable path through each chromosome separately and assigned the CpG states accordingly to either unmethylated, intermediate or highly methylated. Subsequently, we merged neighboring CpGs residing in the same state and being less than 200 bp apart into unmethylated, intermediateor highly methylated regions. Only regions harboring more than three CpGs were retained for subsequent analysis. The resulting region set is more likely to pick up DMRs due to the highly spatially constrained nature of the marked GRE (often 200-400 bp) which easily gets masked by a coarse grained tiling based approach. The HMM inference framework was implemented as custom software in Python (http://python.org/) and extended to incorporate other state distribution types. To determine DMRs between two samples, we followed our previously established protocol (Bock et al., 2011).

5.3.3 ChIP-Seq data analysis

In order to identify regions enriched for chromatin modifications we employed a two step approach, first identifying all regions enriched for any chromatin modification. Next, using this comparatively small region set, we determined the quantitative enrichment level as well as significance of enrichment using a Poisson background model based on the WCE see Section 3.3.2. Finally, we utilize conservative enrichment and significance cutoffs to binarize our enrichment signal in order to increase robustness and simplify subsequent analysis. First, we segmented the genome into non-overlapping windows and classified each window into either enriched or no enriched. This analysis was conducted separately for two groups, 1. H3K27ac, H3K4me3 using 200 bp windows and H3K27me3, K3K4me1 using 400 bp windows according to Section 3.3.2 retaining only windows significant at $p \leq 0.05$ and enrichment above background greater than three. Next, enriched windows within a distance of 850 bp were merged into larger regions. Regions smaller than 400 bp (600 bp for broad marks) after merging were discarded as due to noise and regions greater than 10 kb were split. This procedure was carried out for three groups of histfour cell types. The resulting three lists of enriched regions were then merged in a hierarchical fashion: first regions identified based on H3K4me3 & H3K27ac and H3K4me1, retaining all H3K4me3 & H3K27ac regions but merging or splitting enriched H3K4me1 regions.

After completion of this initial processing step, regions were again filtered for minimal size discarding regions smaller than 400 bp. Next, the same procedure was repeated for the new H3K4me3, H3K27ac, H3K4me1 region set and the H3K27me3, H3K9me3 region list. Finally, the resulting list was merged with the regions classified as UMRs and IMRs, adding only regions not overlapping with any region identified so far. This procedure gave rise to the region catalog used in subsequent analysis.

In the second processing step, comparative analysis of ChIP-Seq experiments and assignment of chromatin states was carried out. First, for each region in the region catalog the significance and enrichment over WCE was determined using Poisson statistics (Section 3.3.2) applied to the duplicate filtered and insert size extended sequencing tag counts overlapping each identified region. Regions with tag counts deviating at a significance level of $p < 0.001$ from the WCE and exhibiting enrichment of WCE ≥ 3 were classified as enriched. We chose these moderately stringent thresholds in order also pick up chromatin state changes that occur only in a subset of the investigated cell population and therefore have lower signal. However, this comes at the expense of a higher false-positive rate. Next, we compared the enrichment levels for all four cell types (hESC, dEC, dME, dEN) for each epitope separately again using the method outlined in Sction 3.3.2. We defined regions deviating by more than three-fold at a significance level of $p \leq 0.05$ as being different. Next, we reconciled these differential enrichment calls with our enrichment over background classification. Since in our setting we were mostly concerned with incorrectly called differences between cell states (false positives) due to heterogeneity in the distinct populations and varying ChIP-Seq library complexity, we redefined regions that were classified as enriched in hESC and not enriched in one of the differentiated cell types but exhibiting no significant difference according to our differential analysis as being enriched in the differentiated cell type under study. This approach yields a lower false positive rate in terms of dynamics at the expense of a higher false negative rate. However, at this point it still remains to be determined what magnitudes of differences in chromatin modifications are actually meaningful. In this sense, our binary classification approach is rather conservative and relies on previously established observations. Subsequently, we classified each genomic region identified in this way into one of eleven epigenetic states based on the binary classification of enrichment levels for the various modifications. DNA methylation levels were not taken into account when histone modification based states were assigned. Only states devoid of significant enrichment for one of the histone modifications were classified based on DNA methylation levels. Genomic regions were associated with their nearest RefSeq gene using the R package ChIPpeakAnno (Zhu et al., 2010) and classified into promoter, intragenic, distal (≤ 50 kb from TSS and not promoter) and intergenic.

5.3.4 TF ChIP-Seq Analysis

For OCT4, SOX2, NANOG and FOXA2 aligned read files were processed with macs version 1.4 (Zhang et al., 2008) using the following parameters: -g 2.7e9 –tsize=36 –pvalue=1e-5 –keep-dup=1 and the HUES64 WCE as input control. All other parameters were left at their default setting. For our 25 bp libraries, tsize was set to 25. FDR was calculated using macs built-in function essentially comparing the original read count distribution with a randomly shuffled distribution. Following this initial peak calling, only peaks significant at an FDR of 0.05 and present in both replicates were retained. As a second replicate for our OCT4 ChIP-Seq experiment we took advantage of publically available OCT4 data (Kunarso et al., 2010).

5.3.5 ChIP Bisulfite Sequencing Analysis

For the FOXA2 ChIP-bisulfite sequencing experiment, the bisulfite treated ChIP library was processed similarly to the WGBS processing described above and subsequently overlaid with the peak calling results from the FOXA2-ChIP-Seq library that was not bisulfite treated.

5.3.6 Motif Analysis

Predefined sets of genomic regions were scanned for occurrences of motifs contained in the Transfac professional database (2009) using the FIMO program from the MEME suite (Grant et al., 2011). Only motifs with at least one known associated human TF and detected at a significance level of $p \leq 10^{-5}$ were used for further analysis. Next, the total number of occurrences was calculated for each motif. To correct for sequence composition, we trained a Hidden Markov Model on each set of input sequence sets and generated ten sets of number and size matched region sets using the inferred probabilities as controls. Subsequently, these sequence sets were also subjected to the same motif identification procedure and motif enrichment results were averaged over the 10 control runs. We defined the final motif enrichment score as the fraction of total motif occurrences in the region set of interest and the total number of motif occurrences in the averaged control region set. To determine differentially enriched motifs between region sets from different hESC-derived cell types, we calculated the fraction of motif scores between the two conditions, retaining only motifs with a differential enrichment ≥ 1.2. For the H3K27ac motif analysis, we computed overall motif enrichment scores for each region class separately as described above. Next, we correlated the motif enrichment scores only focusing on those motifs with scores ≥ 1.2. To that end we multiplied the motif enrichment score for the cell type of interest with the $\log_2$ fold change of the associated TF in that cell type, giving rise to a new combined motif score. If multiple TFs mapped to one motif, we took the average motif score. For each cell type we rank-ordered the motifs according to their enrichment scores and report the top 20 motifs with their raw motif score.

For the H3K4me1 analysis, we wanted to focus on all potential TFBS gaining H3K4me1 and not only those that also become expressed as in the H3K27ac analysis. First, we again determined the motif enrichment scores over background. To focus on motifs differentially enriched between the different cell types, we subtracted the mean motif enrichment across hESd cell types for each motif separately from the enrichment level and rank-ordered the motifs. For each cell type, we then report the top 20 enriched motifs.

For the analysis of potential upstream regulators of TFs that exhibit changes in their promoter region, we first scanned all distal and proximal regions associated with these TF genes for motif occurrences. Next, we determined whether any of the observed motifs were associated with TFs differentially expressed in any of the cell types and correlated the sign of the differential expression in each cell type with the sign of the epigenetic state change of the region the motif occurred in (gain of open chromatin mark: +1, loss of open chromatin mark or acquisition of a repressive state: -1). Next, we rank-ordered all observed TF motifs that were differentially expressed in at least one of the cell types based on their occurrence/epigenetic state change correlation for each cell type separately and reported the gene expression levels of top 30 motifs for each cell type.

5.4 Results

5.4.1 High resolution transcriptional measurements during directed differentiation of hESCs

To better understand the molecular dynamics involved in hESC differentiation, we produced populations representative of each embryonic germ layer, namely ectoderm (Lee et al., 2010), mesoderm (Evseenko et al., 2010) and endoderm (Hay et al., 2008) (see Appendix A.1.3). We chose the male hESC line HUES64, an NIH-approved line that readily differentiates into each of the three germ layers. These hESCs can be differentiated into a neuroectoderm-like progenitor population positive for SOX2 and PAX6 by inhibition of TGFb, WNT and BMP signaling (Fig, 5.1a, top). Alternatively, canonical mesoderm markers, such as GATA2 (Fig. 5.1 middle), can be induced using ACTIVIN A, BMP4, VEGF and FGF2 treatment. Lastly, differentiation towards a definitive endoderm-like fate, positive for markers such as SOX17 and FOXA2 (Fig. 5.1, bottom), is induced using ACTIVIN A and WNT3A.

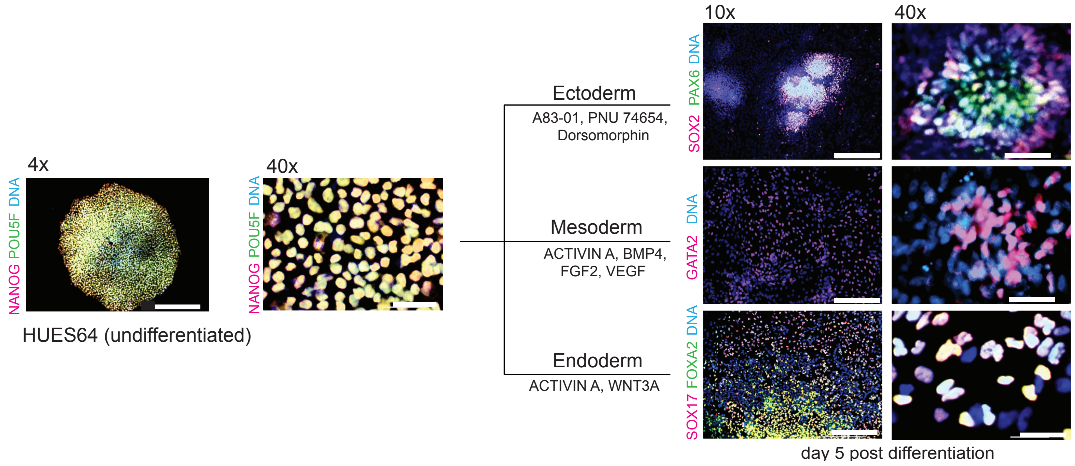


FIGURE 5.1: Left: Low (43) and high (403) magnification overlaid immunofluorescent images of the undifferentiated hESC line HUES64 stained with OCT4 (POU5F1) and NANOG antibodies. Right: Established directed (two-dimensional) differentiation conditions were used to generate representative populations of the three embryonic germ layers: hESC-derived ectoderm, hESC-derived mesoderm, and hESC-derived endoderm. Cells were fixed and stained after 5 days of differentiation with the indicated antibodies. Representative overlaid images at low (103) and high (403) magnification are shown. DNA was stained with Hoechst 33342 in all images. Scale bars, 200 mm (43), 100 mm (103), and 30 mm (403).

We began by measuring the expression of 541 selected genes, including many developmental TFs and lineage markers (Bock et al., 2011), at 24-hour intervals during differentiation towards each respective germ layer. We found that 268 of these genes exhibit expression changes (z-score $\log_2$ expression) during the first five days of differentiation (Fig. 5.2). Mesendodermal genes, such as EOMES, T, FOXA2 and GSC, are upregulated at 24 hours of mesoderm and endoderm induction, but not ectoderm differentiation (Fig. 5.2,5.3a). GSC expression decreases within 48 hours of differentiation in the mesoderm-like population, while the expression level is maintained in the endoderm population (Fig. 5.2, 5.3a). EOMES and FOXA2 expression is also maintained in the endoderm population accompanied by upregulation of GATA6, SOX17 and HHEX ((Fig. 5.2). After transient upregulation of mesendodermal markers, activation of mesodermal markers such as GATA2, HAND2, SOX9 and TAL1 is detected specifically in the mesoderm conditions (Fig. 5.2, 5.3a). None of these markers are detected during early ectoderm differentiation, which instead upregulates neural markers, such as PAX6, SOX10 and EN1 (Fig. 5.2, 5.3a).

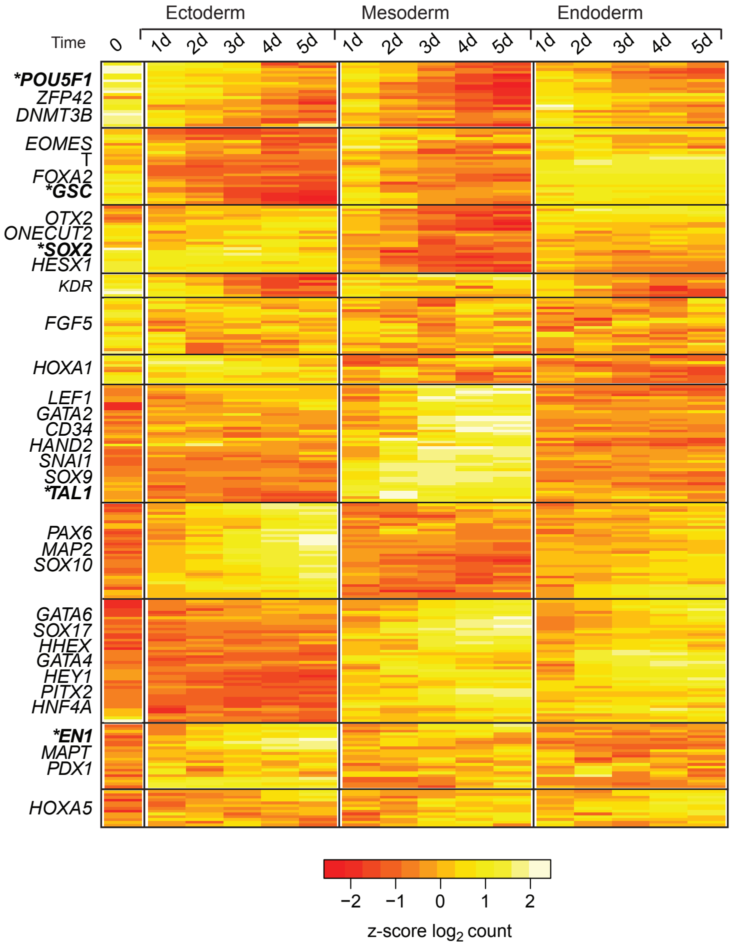


FIGURE 5.2: NanoString nCounter expression data (z-score $\log_2$ expression value of two biological replicates) for a time course of in vitro differentiation using the conditions shown in (Fig. 5.1) 541 genes were profiled, and 268 changing by more than 0.5 are displayed. Selected lineage-specific genes are shown on the left for each category that was identified based on hierarchical clustering. The average $\log_2$ expression value of two biological replicates is displayed. Error bars represent 1 SD.

We found that POU5F1 (OCT4), NANOG and to some extent SOX2 expression is maintained in our endoderm population (Fig. 5.2,5.3a). This is consistent with prior studies indicating that OCT4 and NANOG expression is detected during the course of early endoderm differentiation and supports NANOG's suggested role in endoderm specification (Teo et al., 2011). SOX2 expression is downregulated in mesoderm and to a lesser degree in endoderm, but maintained at high levels in the ectoderm population ($\log_2$ expression 10.9) (Fig. 5.3b), while ZFP42 (REX1) is similarly downregulated in all three lineages (Fig. 5.2).

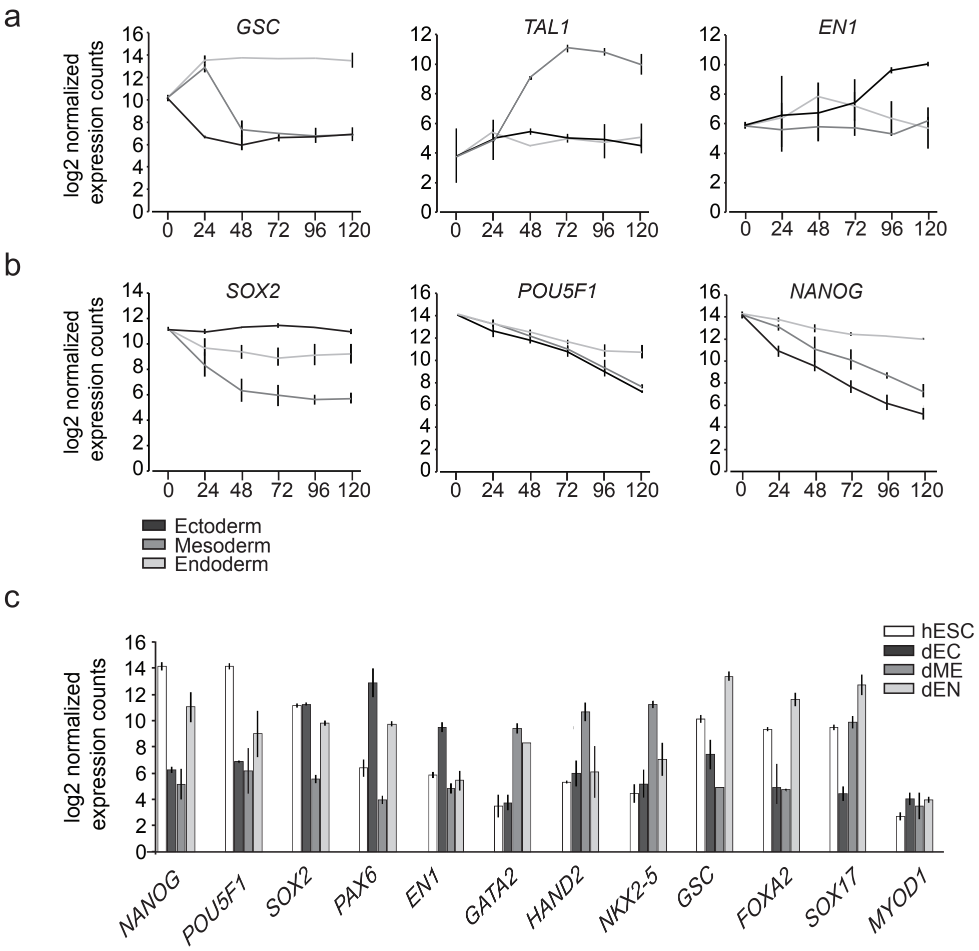


FIGURE 5.3: **a.** The average $\log_2$ expression values of two biological replicates of lineage-specific genes highlighted in a. are shown. Error bars represent 1 SD. If no error is evident, SD $< 0.5 \log_2$ expression units. **b.** The average $\log_2$ expression values of two biological replicates of pluripotent genes highlighted in **a.** are shown. Error bars represent 1 SD. If no error is evident, SD$< 0.5 \log_2$ expression units. **c.** NanoString nCounter profiling of FACS-isolated ectoderm (dEC), mesoderm (dME), and endoderm (dEN). Expression levels for MYOD1 (right) are included as a negative control. The average $\log_2$ expression value of two biological replicates is shown. Error bars represent 1 SD. If no error is evident, SD $< 0.5 \log_2$ expression units.

We confirmed that these populations indeed represent a precursor stage for each respective lineage by inducing them to differentiate further, which resulted in upregulation of genes such as OLIG2 and SST in the ectoderm (Chambers et al., 2012), TRPV6 in the mesoderm (Evseenko et al., 2010), and AFP and HGF in the later endoderm populations (Fig. 5.4a) (DeLaForest et al., 2011). Lastly, multidimensional scaling confirmed that at 24 hours the mesoderm population is very similar to the endoderm, while the ectoderm population has already moved in an alternative direction (Fig. 5.4b). These high temporal resolution gene expression signatures suggest that expression programs associated with the three unique cell populations, representing early stages of each germ layer are established within a similar timeframe of hESC differentiation.

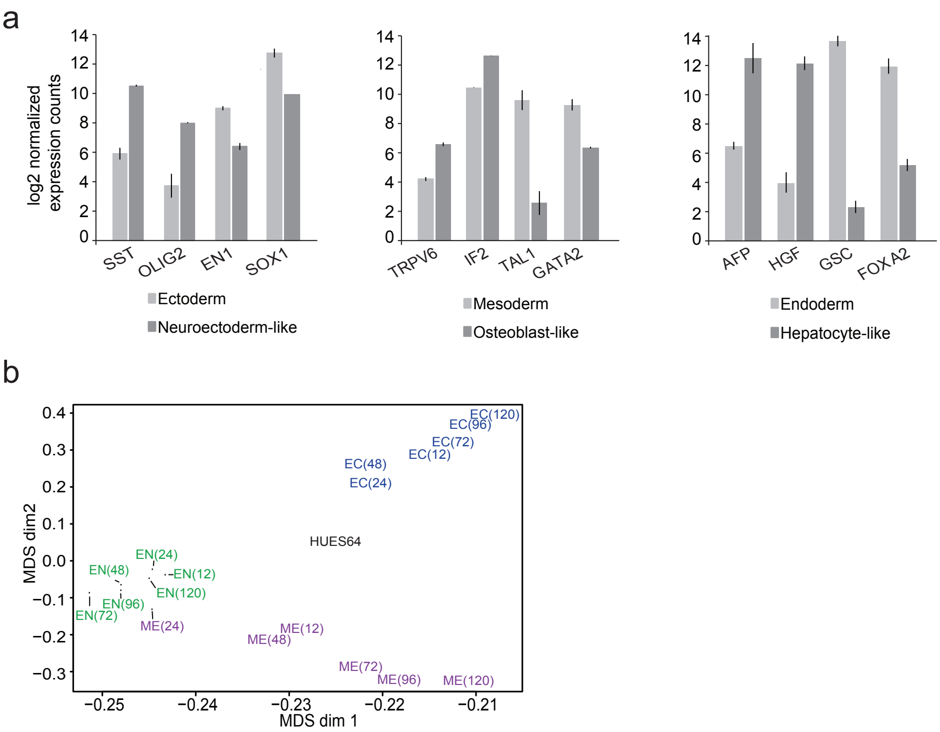


FIGURE 5.4: **a.** Median Nanostring expression values ($\log_2$) of populations derived from dEC, dME and dEN. **b.** Multidimensional scaling of populations included in differentiation time course.

5.4.2 Global transcriptional dynamics between hESCs and hESC-derived cell types

Based on these results, we selected day five as the optimal time point to capture early regulatory events in well-differentiated populations representing all three germ layers. To reduce heterogeneity, we used FACS to enrich populations based on previously reported surface markers (see Appendix (A.1.4)); populations isolated by FACS are referred to as dEC for the ectoderm, dME for the mesoderm and dEN for the endoderm. Expression analysis of the sorted populations confirms further enrichment for the desired populations (Fig. 5.3b). We next expanded on our selected gene signature profiles by performing strand specific RNA-Seq on poly-A fractions from each day 5 differentiated FACS-isolated populations and undifferentiated HUES64. Hierarchical clustering based on the global expression profiles of each cell type reveals that the dME population is the most distantly related cell type and that dEN and dEC are more similar to each other than to dME or hESCs (Fig. 5.5a). This was unexpected given that the dME and dEN populations are putatively derived through a common mesendoderm precursor stage (Fig. 5.2 5.3a) while the dEC does not upregulate markers associated with this stage (EOMES, T, GSC;). Overall, 14,196 RefSeq defined coding and non-coding transcripts (38% of defined transcripts) are expressed (FPKM $>$ 1) in at least one of the populations, with 11,579 (81.6% of the total number of transcripts detected within our cell types) being expressed in all three populations. Examining the overlap of genes expressed (FPKM $>$ 1) in each population reveals that the dME population exhibits expression of the largest number of unique genes (n=448, (Fig. 5.5b)), such as RUNX1 (FPKM: 3.4) and HAND2 (FPKM: 17.8). Examining genes unique to pairs of the differentiated cell types also reveals that dEC and dME have the least in common (n=37, Fig. 5.5b), while the dEC and dEN have the most number of transcripts in common (n=171, Fig. 5.5b) consistent with our clustering analysis. Genes such as PAX6 (dEC FPKM: 25.9, dEN FPKM: 5.6) and NKX6.1 (dEC FPKM: 2.3, dEN FPKM: 3.3), which are each required for both brain (Ericson et al., 1997) and pancreas development (Sander et al., 1997), are expressed in both the dEC and dEN. Canonical markers of embryonic development, such as FOXA2 (FPKM: 12.7) in the dEN and EN1 (FPKM: 5.8) in the dEC are restricted to their expected germ layers at our early stages.

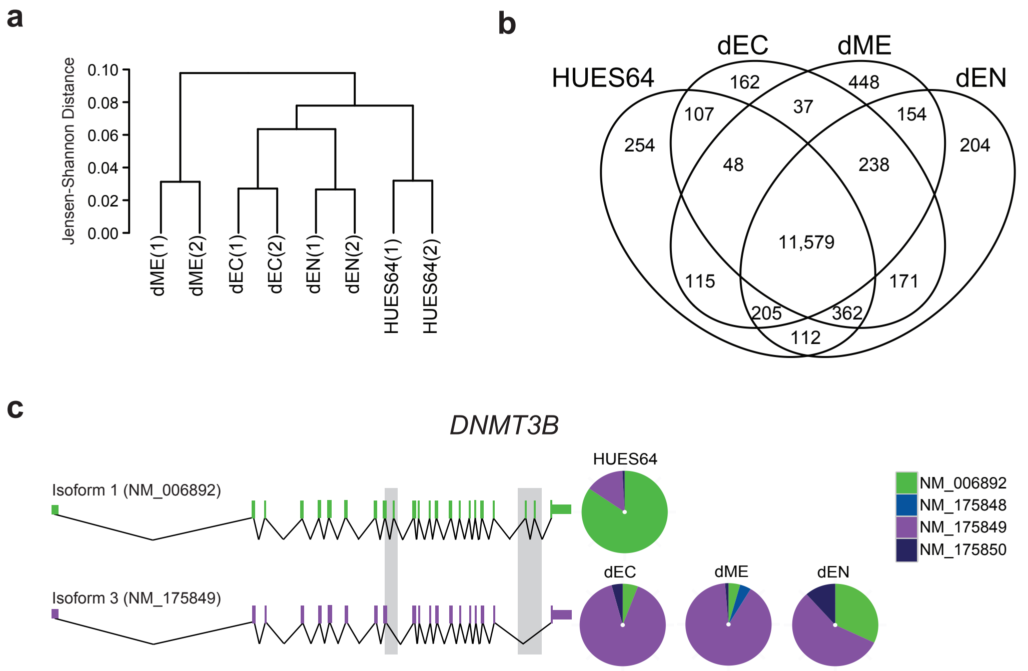


FIGURE 5.5: **a.** Hierarchical clustering of global gene expression profiles as measured by strand-specific RNA-seq for biological replicates of HUES64 and dEC, dME, and dEN is shown as a dendrogram. Pairwise distances between the replicates were measured using the Jensen-Shannon distance metric. **b.** Venn diagram illustrating unique and overlapping genes with expression (FPKM > 1) in HUES64 and the FACS-isolated directed differentiation conditions are shown. **c.** Differential splicing of DNMT3B in response to directed differentiation. Relative expression of isoforms 1 (NM006892, green) and 3 (NM175849, purple) as measured by RNA-seq are shown on the right.

Notably, we also identified 1,296 splicing events (FDR=5%) as well as alternative promoter usage within our populations (Trapnell et al., 2013). For example, we detected expression of multiple isoforms of DNMT3B ($p = 5 \times 10^{-5}$). Expression of DNMT3B isoform 1 (NM006892) was restricted to the undifferentiated hESCs (FPKM: 214.3), while the differentiated cell types predominantly express an alternative isoform, DNMT3B isoform 3 (NM175849) (dEC FPKM: 33.9, dME FPKM: 14.2, dEN FPKM: 20.0) (Fig. 5.5c). The presence of this isoform, as well as others, has previously been reported in more advanced stages of embryonic development as well as normal adult (Robertson et al., 1999) and cancerous tissues (Ostler et al., 2007). Our results suggest that this switch coincides with the exit from the pluripotent state, regardless of the specified lineage. We also identified expression of three PITX2 isoforms, with differential splicing leading to different isoform expression between the dEN and dME. In the Chick PITX2 is essential for heart looping and each isoform is responsible for executing distinct functions (Yu et al., 2001). Taken together, this suggests that both transcript levels and isoform expression contribute to cellular identity.

5.4.3 Generation of Comprehensive Reference Epigenome Maps

To gain a more complete picture of the underlying molecular mechanisms and investigate the regulatory events during the specification of the three germ layers, we collected approximately 12 million cells of the respective dEC, dME and dEN populations as well as HUES64. All samples were subjected to ChIP-Seq (H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K36me3, and H3K9me3) and WGBS (Fig. 5.6a), producing a total of 32 data sets with over 12 billion aligned reads (data are publicly available through the NIH Roadmap Epigenomics Project data repositories: http://www.roadmapepigenomics.org/, Gifford and Ziller et al. (Gifford et al., 2013), GSE46130 and in the electronic table accompanying this work in Table CH5_DataSet_Table in the subfolder corresponding to this chapter).

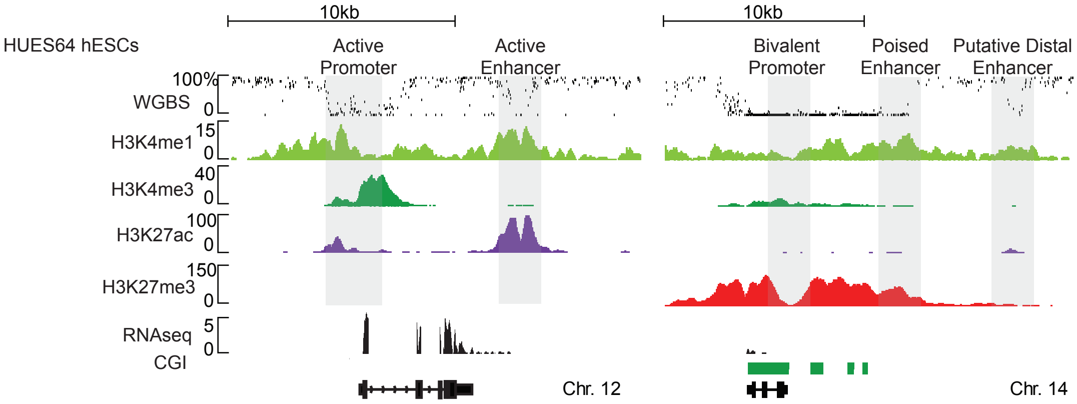


FIGURE 5.6: WGBS (% methylation), ChIP-seq (read count normalized to 10 million reads), and RNA-seq (FPKM, read count normalized) for the undifferentiated hESC line HUES64 at three loci: NANOG (chr12:7,935,038-7,957,818) and GSC (chr14:95,230,449-95,250,241). CGI are indicated.

5.4.4 Integrative analysis of epigenetic state transitions

We focused our analysis on previously identified informative chromatin states associated with various types of regulatory elements (Ernst et al., 2011, Rada-Iglesias et al., 2011), including the following specific combinations: H3K4me3+H3K27me3 (bivalent/poised promoter); H3K4me3+H3K27ac (active promoter); H3K4me3 (initiating promoter); H3K27me3+H3K4me1 (poised developmental enhancer); H3K4me1 (poised enhancer), H3K27ac+H3K4me1 (active enhancer); H3K27me3 (Polycomb-repressed). In addition, we segmented the WGBS data into three DNAme states: highly methylated regions (HMRs: $>$ 60%), intermediately methylated regions (IMRs: 11-60%) and unmethylated regions (UMRs: 0-10%). The latter differs from the highly methylated background of the genome and likely indicates functional importance as previously suggested (Stadler et al., 2011). We next assigned each genomic region to one of the resulting states (see Section 5.3.3) and determined all regions that change their assigned epigenetic state from hESC to at least one of the 3 hESC derived poplations ($n = 157,443$, Fig. 5.6a).

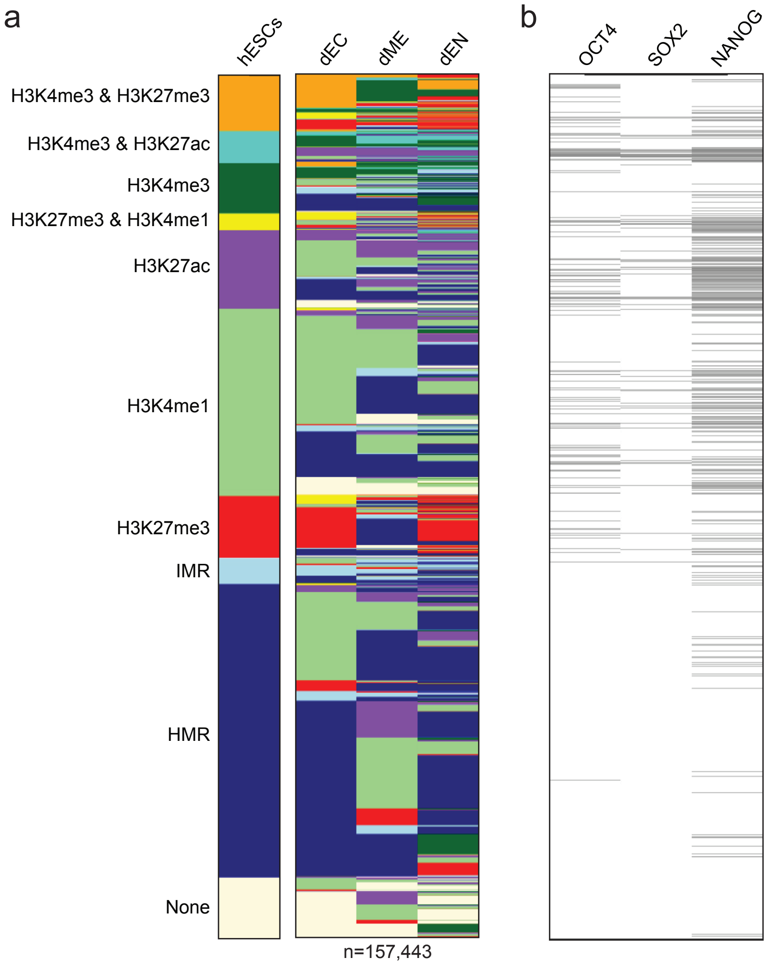


FIGURE 5.7: **a.** Epigenetic state map of regions enriched for one of four histone modifications in at least one cell type or classified as UMR/IMR in at least one cell type and changing its epigenetic state upon differentiation in at least one cell type (n = 157,433). **b.** Regions bound by OCT4, SOX2, and NANOG, as determined by ChIP-seq and organized using the chromatin states in b.

The majority of epigenetically dynamic regions are not located near promoters (6.8% +2 kb to -500 bp of the TSS- Promoters; 48.8% $>$ 50 kb upstream of TSS- Intergenic; 15.1% $>$ 500 bp downstream of TSS- Intragenic/Gene body). Correlating epigenetic changes with transcriptional dynamics, we find that overall the majority (62-67%) of all epigenetic remodeling events are not directly linked to transcriptional changes based on the expression of the nearest gene.

The loss of H3K4 methylation (me1 and me3) is commonly associated with a transition to high DNAme (Fig. 5.7a), which is most prominent in the dEN population and preferentially eliminated from genes involved in neural development (i.e. neural tube development $q = 9.6 \times 10^{-12}$). We identified 4,639 proximal bivalent domains in hESCs and observe that 3,951 (85.1%) of these domains resolve their bivalent state in at least one hESC-derived cell type (Fig. 5.6a and 5.8a). When we specifically investigated the promoters of TF-encoding genes, we found that 463 of these promoters are in a bivalent state in hESCs, and 400 of them change in at least one differentiated cell type (Fig. 5.8b). The majority transitions to H3K4me3-only or H3K27me3-only in a lineage-specific manner. In dME, H3K4me3 is gained at the ISL1 locus while H3K27me3 is lost, leading to expression (FPKM: 14.3). The lineage specific dynamics in this region are interesting given that this gene has known roles in all three germ layers, although at later time points (Ahlgren et al., 1997, Cai et al., 2003, Pfaff et al., 1996). Notably, in contrast to the limited overall association between many epigenomic dynamics and changes in expression, we found that a large proportion of these bivalent TFs (275) change their expression level during the differentiation (Fig. 5.8b).

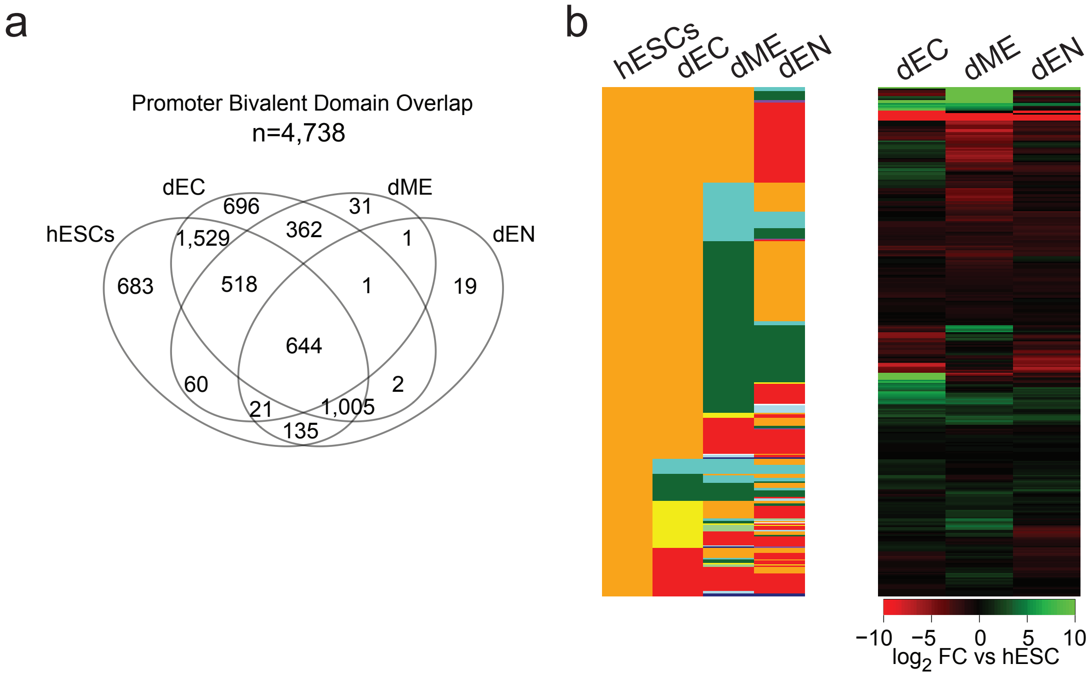


FIGURE 5.8: **a.** Venn diagram showing the overlap of identified proximal bivalent domains. **b.** Left: Chromatin state map for all TFs that are bivalent in hESCs and change their epigenetic state in at least one cell type (n=400). Right: Hierarchical clustering ordered heatmap of TF expression ($\log_2$FC relative to hESCs).

5.4.5 Pluripotent TF binding is linked to chromatin dynamics during differentiation

To further explore potential regulators of chromatin dynamics during the exit from pluripotency, we performed ChIP-Seq for OCT4, NANOG and SOX2 in HUES64 (Fig. 5.9a,b). We found that regions bound by all three factors (n=1,556), by SOX2-only (n=923) or by NANOG-only (n=14,531) are frequently associated with inter- and intragenic regions (Fig. 5.9c-e, top). In contrast, regions bound by OCT4-only (n=8,599) are more frequently associated with promoter regions (Fig. 5.9c). Examination of regions bound by OCT4, NANOG and SOX2 in hESCs showed H3K4me1 regions enriched for OCT4 binding sites frequently become HMRs in all three differentiated cell types whereas NANOG and SOX2 sites are more prone to change to an HMR state in dME (Fig. 5.9f). In general, many regions associated with open chromatin that are bound by NANOG are more likely to retain this state in dEN compared to dME and dEC (Fig. 5.9f). We also found that regions enriched for H3K27ac in hESCs that maintain this state in dEN or dEC are likely to be bound by SOX2 and NANOG. This is in agreement with the reported role of SOX2 during ectoderm development and differentiation (Wang et al., 2012), but also supports our observation that SOX2 expression is maintained in the dEN. Motif enrichment analysis detected the GATA3 motif in regions bound by OCT4 and SOX2 that transition to an active state in dEC. Furthermore, we found that regions bound by OCT4, NANOG and SOX2 that gain an active mark in dEC are enriched for the motifs PAX9, p63 and STATs. Examining epigenetic dynamics at sites of OCT4, NANOG and SOX2 binding further supports the observation that some pluripotency associated TFs are also involved in the downstream specification.

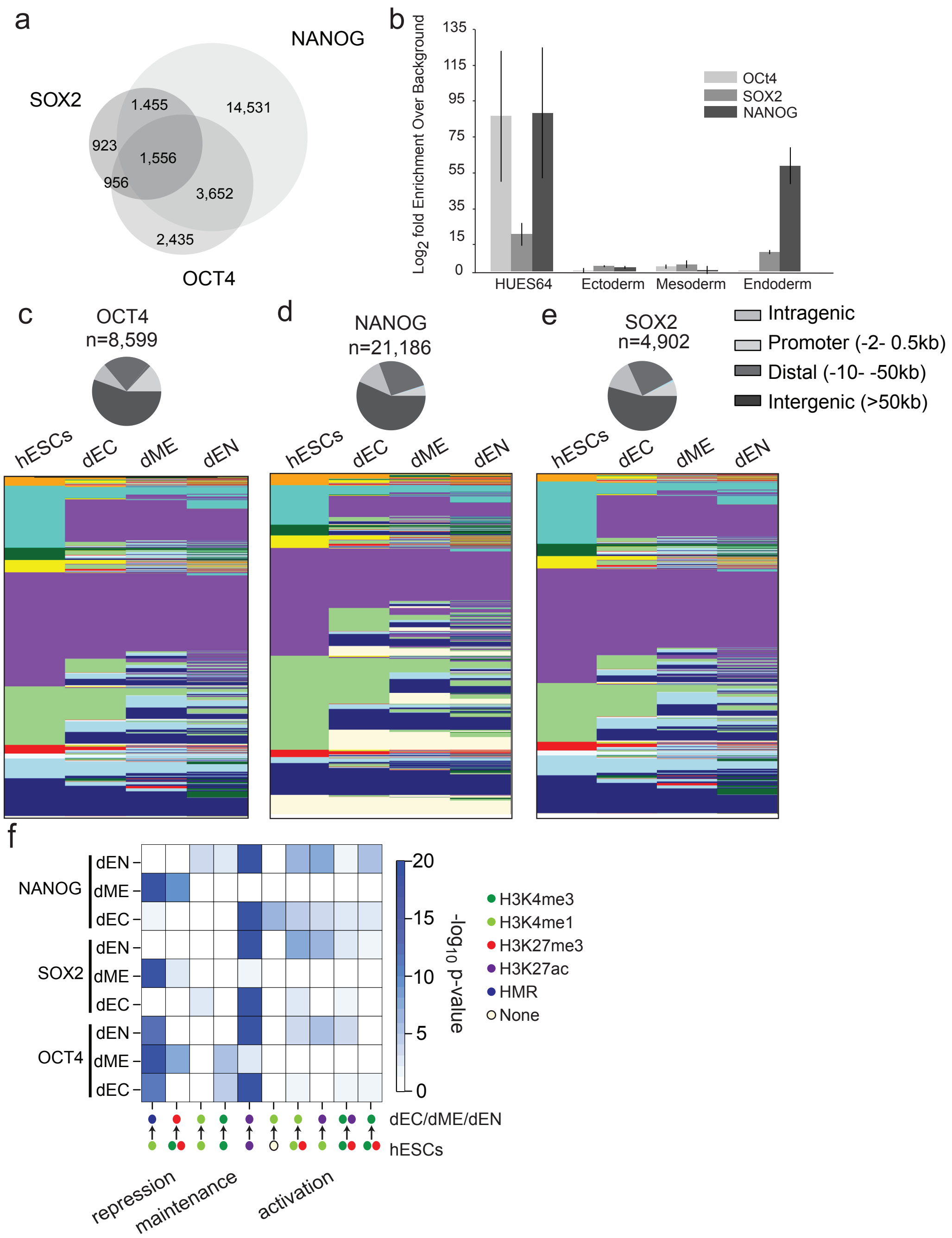


Figure 5.9: **a.** Venn diagram of the overlap between OCT4, NANOG and SOX2 binding sites identified in hESCs (total overlap = 1,556) **b.** Fold enrichment of OCT4, NANOG and SOX2 binding at the NANOG locus in hESCs, and each differentiated population on day 5 of differentiation. **c.** Genomic features of OCT4 binding sites (top) and the associated epigenetic states (bottom) (n=8,599). **d.** Genomic features of NANOG binding sites (top) and the associated epigenetic states (bottom) (n=21,186). **e.** Genomic features of SOX2 binding sites (top) and the associated epigenetic states (bottom) (n=4,902). **f.** Enrichment of OCT4, SOX2 and NANOG within various classes of dynamic genomic regions changing upon differentiation of hESC, computed relative to all regions exhibiting the particular epigenetic state change in other cell types. Epigenetic dynamics are categorized into three major classes: repression (loss of H3K4me3 or H3K4me1 and acquisition of H3K27me3 or DNAme), maintenance of open chromatin marks (H3K4me3, H3K4me1, H3K27ac) and activation of previously repressed states.

5.4.6 Gain of DNAme occurs at open chromatin enriched for TF motifs

We next utilized the WGBS data that cover approximately 26 Million CpGs (at ≥ 5 coverage) across all four cell types. Hierarchical clustering analysis of the WGBS data, which included human adult liver and hippocampus for comparison, revealed that the pluripotent hESCs and the hESC-derived cell types form a separate cluster arm with respect to the somatic tissues (Fig. 5.10a). We determined DMRs defined as exhibiting a significant ($p \leq 0.05$) minimal difference of CpG methylation level of 0.1 among our four cell types. The majority of all DMRs occur at CpG-poor intergenic regions in line with previous reports (Fig. 5.10b bottom) (Stadler et al., 2011). The dEN exhibits more than twice the number of regions that gain DNAme compared to dEC and dME (Fig. 5.10b, top).

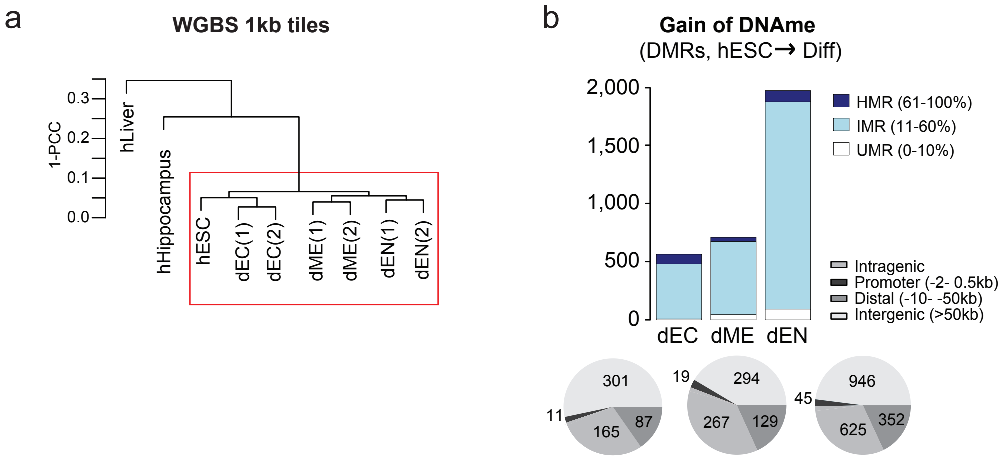


FIGURE 5.10: **a.** Hierarchical clustering of hESCs, hESC-derived populations (dEC, dME and dEN), human adult hippocampus and human adult liver based on mean DNAme levels of 1kb tiles across the human genome using Pearson Correlation Coefficient (PCC). Y-axis indicates sample distance in terms of 1 minus PCC. Red box indicates cell types interrogated in this study. **b.** Regions that significantly ($p \leq 0.05$) increase their DNAme levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNAme state found in hESCs. Bottom: Genomic features associated with DMRs gaining DNAme in each of the differentiated cell types based on RefSeq gene annotation and de novo discovered promoters by RNA-Seq.

Interestingly, only 65 of the total number of DMRs identified are shared between all three populations. However, reaffirming that our populations are depleted of pluripotent cells, this group of DMRs includes the regulatory region of OCT4. In line with the small number of shared regions, more than 60% of regions that gain DNAme are lineage specific (Fig. 5.11a) and include loci such as SMAD3 (dEC), CTNNA3 (dME) and FOXA2 (dEN). FOXA2 has an upstream CGI that exhibits gain of DNAme (Fig.

5.11b), and transcription in dEN is initiated downstream of this DMR at an alternative TSS, suggesting that TSS usage may be regulated, stabilized or reflected by DNAme (Maunakea et al., 2010).

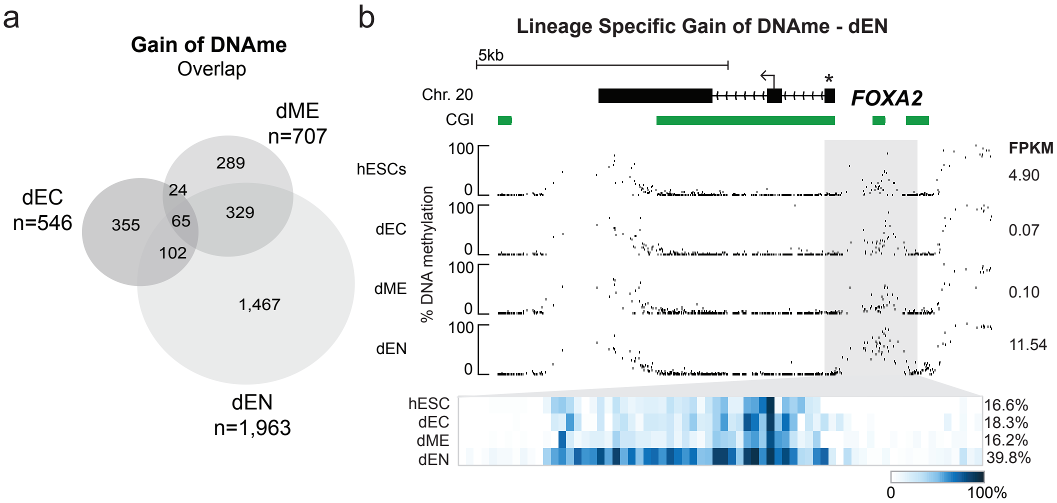


FIGURE 5.11: **a.** The overlap of these differentially methylated regions (DMRs) that increase their DNAme level in the three hESC-derived populations. **b.** DNAme levels and RNA-Seq expression values of FOXA2 (chr20:22,559,343-22,571,189) in hESCs and differentiated cell types. The heat map below shows the DNAme values of individual CpGs within the highlighted region. The average DNAme value for the entire highlighted region is shown on the right in red. CpG islands (CGI) are shown as green bars. Expression values (FPKM) are displayed on the right. The arrows indicate two known TSSs.

We find significant enrichment of various TF motifs as DNAme targets upon differentiation, which has some analogy to the gain of methylation observed at myeloid targets in the lymphoid lineage in vivo (Bock, 2012, Deaton and Bird, 2011, Ji et al., 2010). To extend this observation, we examined the DNAme state at regions bound by SOX2, OCT4 and NANOG in hESCs. For example, two regions 20 kb downstream of DBX1, a gene associated with early neural specification, are bound by all three TFs and gain DNAme in dME and dEN. In contrast, this region maintains low levels of DNAme in dEC, which has activated transcription of DBX1 (Fig. 5.12a). We generally find that co-bound sites gain DNAme in the dME and dEN, but not dEC. Further supporting the functional relevance of these dynamics, we find that regions with gain DNAme frequently coincide with DNAse I hypersensitive sites (Fig. 5.12b) (Thurman et al., 2012).

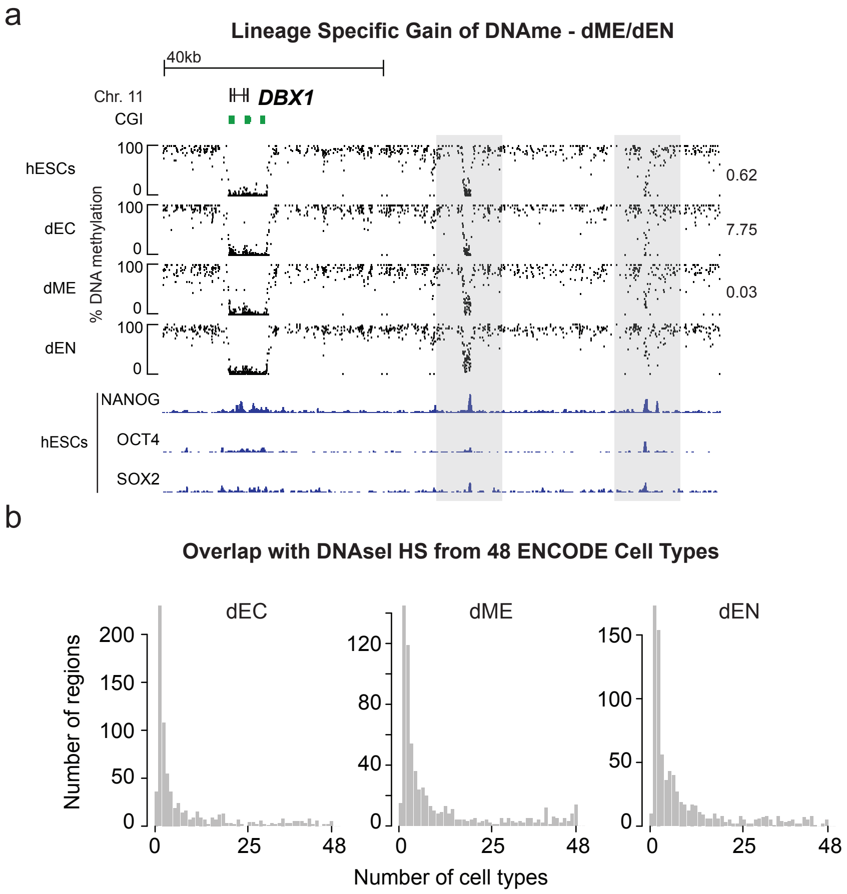


FIGURE 5.12: **a.** DNAme levels and OCT4, SOX2 and NANOG ChIP-Seq at the DBX1 locus (chr11:20,169,548-20,277,940). **b.** Frequency distribution of overlapping DMRs gaining DNAme in the differentiated populations with DNAse I hypersensitive sites across 48 cell ENCODE types (Thurman et al., 2012)

While transcriptional silencing was infrequently correlated with gain of DNAme at distal elements (Fig. 5.13a, left), the promoters that gain DNAme in dEC and dME, are associated with a decrease in expression as expected Fig. 5.13b, right). In examining the chromatin state of regions that gain DNAme during differentiation, we find that most regions exhibited enrichment of one or more histone modifications in hESCs (Fig. 5.13b). These results confirm that in particular distal regulatory elements show highly dynamic regulation of DNAme during specification.

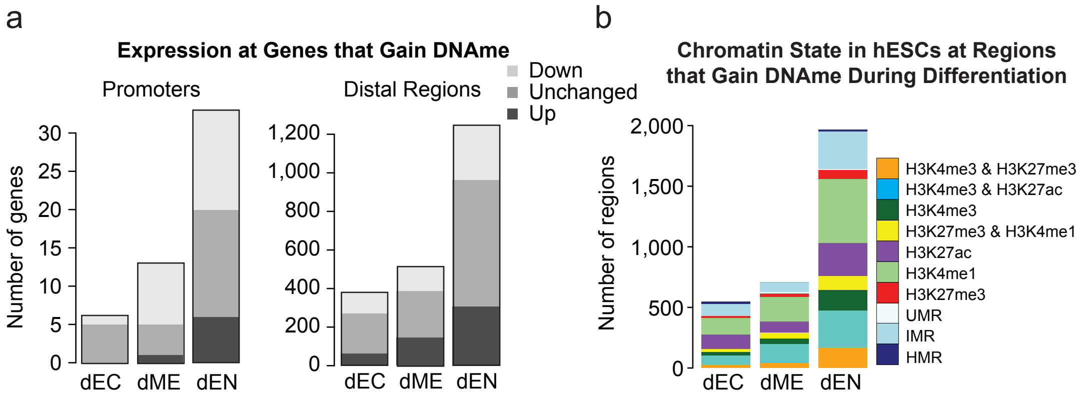


FIGURE 5.13: **a.** Distal elements (left) and Promoters (right) that gain DNAme separated by the changes in FPKM at associated genes. **b.** Chromatin state in hESCs at regions that gain DNAme during differentiation. Regions devoid of any detected chromatin marks are categorized according to their DNAme state in hESCs

5.4.7 Loss of DNAme is biased towards dEC

Loss of DNAme is asymmetric between the three populations (Fig. 5.14a, top) and occurs in a more lineage-specific fashion than gain (Fig. 5.14b). However, loss also occurs mainly at intergenic regions (Fig. 5.14a, bottom). Notably, the dEC has the most DMRs and many were associated with neuronal gene categories (for instance: neural tube development, $q = 3.13 \times 10^{13}$). This includes the ectodermal TF POU3F1, which has a bivalent promoter in hESCs, resolves to a H3K4me3-only state and exhibits transcriptional activation in dEC. Chromatin remodeling and activation at this locus coincides with specific loss of DNAme at a putative regulatory element downstream of the 3'UTR of this gene in dEC (Fig. 5.14c).

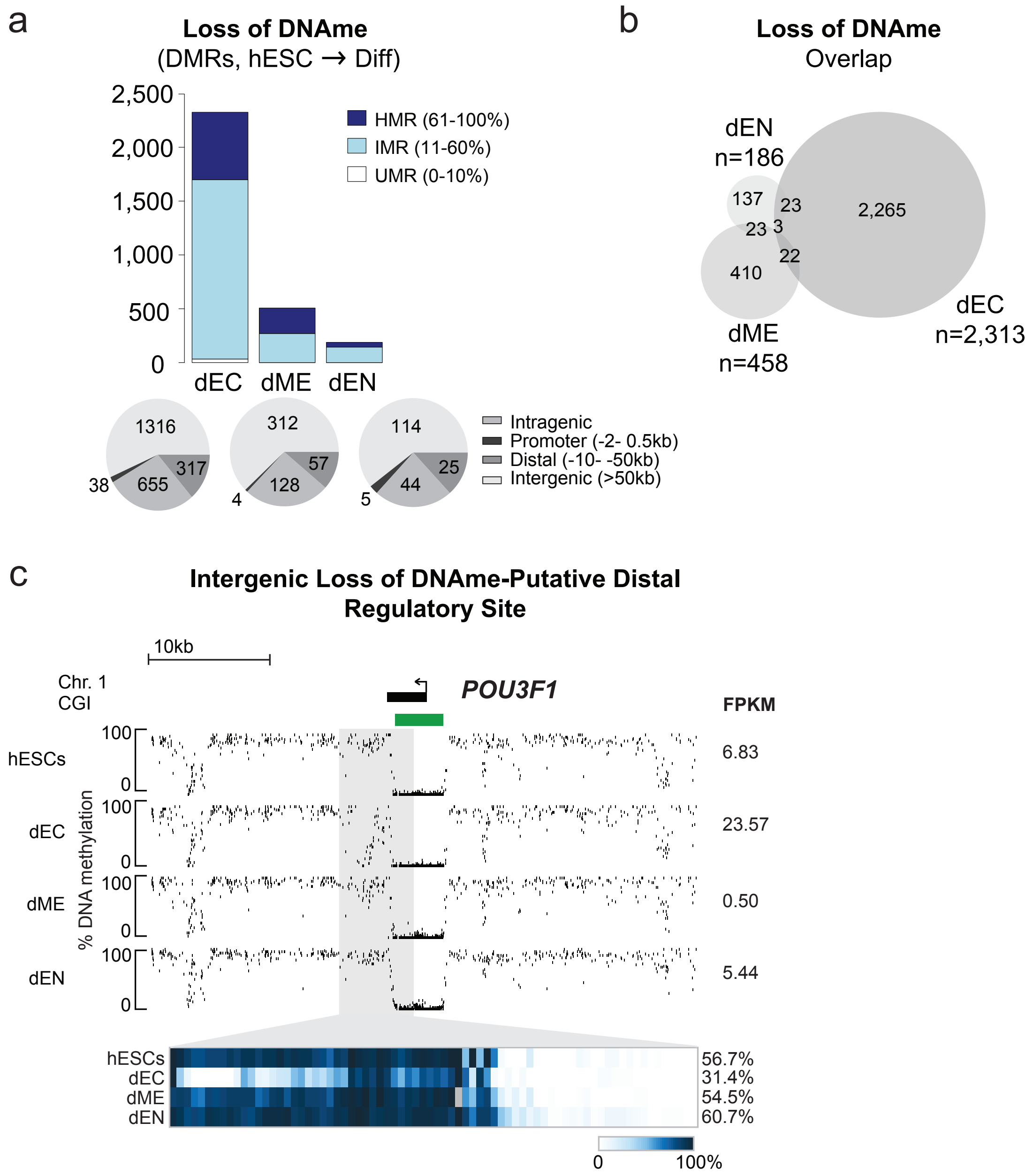


FIGURE 5.14: **a.** Regions that significantly ($p \leq 0.05$) decrease their DNAme levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNAme state distribution in the differentiated cell types, revealing that most regions reside in an IMR state after they lost DNAme (left). Genomic features (bottom) associated with DMRs losing DNAme in each of the differentiated cell types based on RefSeq gene annotation and de novo discovered promoters by RNA-Seq. **b.** Venn diagram of identified DMRs that decrease their DNAme level between the three hESC-derived populations. **c.** DNAme at the POU3F1 locus (chr1:38,493,152-38,532,618). The heat map below shows the DNAme values of individual CpGs within the grey region. The average DNAme value for the entire highlighted region is shown on the right in red. CGIs are shown as green bars. Expression values (FPKM) are displayed on the right.

On a global scale, an immediate correspondence between loss of DNAme and expression, such as that observed at POU3F1, occurs at about half the regions (Fig. 5.15a). More than 70% of DMRs that lose DNAme during differentiation are enriched for one of our profiled histone modifications in particular H3K4me1 or H3K27ac (Fig. 5.15b). Taken together, our hESC differentiation system reveals several interesting DNAme dynamics, including the lineage specific silencing of regulatory regions in default or alternative lineages. The asymmetric loss may also explain why our chromatin state analysis revealed fewer regions that gained H3K4me3 in the dEC population.

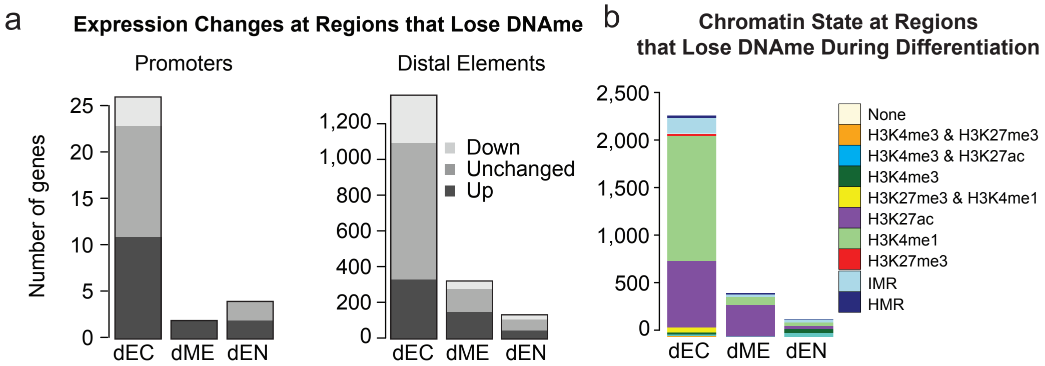


FIGURE 5.15: **a.** Promoters (left) and distal elements (right) that gain DNAme separated by the changes in FPKM at associated genes. **b.** Chromatin state in differentiated cell types at regions that loose DNAme during differentiation.

5.4.8 Gain of H3K27ac reveals putative regulatory elements

In addition to methylation on H3K4, open chromatin is also demarcated by enrichment of H3K27ac. It has also been suggested that the combination of H3K4me1 and H3K27ac at distal regions identifies active enhancer elements, while H3K4me1 and H3K27me3 corresponds to poised enhancer elements (Rada-Iglesias et al., 2011). To extend these observations, we focused specifically on regions that gain H3K27ac during differentiation and found that more than half of the identified regions are HMRs in hESCs (Fig. 5.16a), while another large fraction is enriched for H3K4me1 in hESCs (Fig. 5.16a). The majority of regions that gain H3K27ac are intergenic, as shown for the RUNX1 locus (Fig. 5.16a,b).

We next placed each region into one of three distinct categories (repressed, poised, open) based on their state in hESCs, and subsequently performed gene set enrichment analysis using the GREAT toolbox (Fig. 5.17a) (McLean et al., 2010). This analysis reveals enhancer dynamics in line with the lineage specific differentiation trajectory for dEC and dME (Fig. 5.17a). In contrast, the dEN population shows an unexpected enrichment for early neuronal genes (e.g. neural tube development, Fig. 5.17a). This

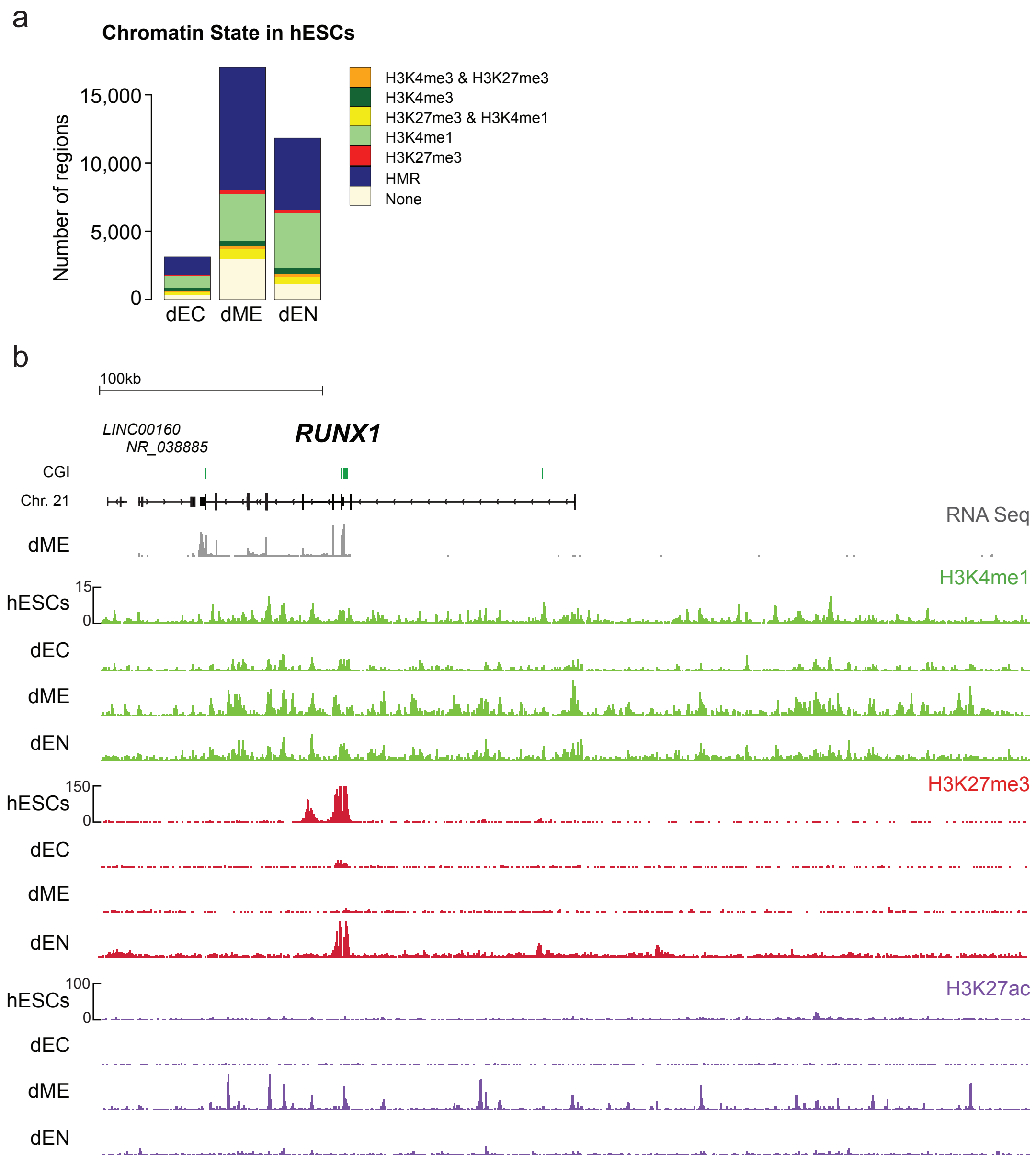


FIGURE 5.16: **a.** Number of regions and associated epigenetic state distribution in hESCs of regions that are transitioning to H3K27ac in the three populations. **b.** Normalized ChIP-Seq tracks (H3K4me1, H3K27me3 and H3K27ac) for the RUNX1 region (chr21:36,091,108-36,746,447) with corresponding RNA-Seq data in dME.

observation is consistent with the correlation that we reported between our dEC and dEN RNA-Seq data, suggesting that similar networks are induced in the early stages of both our ectoderm and endoderm specification (van Arensbergen et al., 2010). Moreover, we find strong enrichment of downstream effector genes of the TGFb, VEGF and BMP pathways in dME, directly reflecting the signaling cascades that were stimulated to induce the respective differentiation. In dEN we find enrichment of genes involved in WNT/b-CATENIN and retinoic acid (RA) signaling (Fig. 5.17a). While we did not use RA, this signaling cascade has previously been implicated in endodermal tissue development including pharyngeal and pancreatic cell types (Ostrom et al., 2008, Wendling et al., 2000). Concordantly, we also find high levels of SMAD3 motif enrichment in the repressed dME and dEN, particularly in the poised putative enhancer populations (Fig. 5.17b). Similarly, we observe enrichment of key lineage specific TF motifs such as the ZIC family proteins in dEC, TBX5 in dME and SRF in dEN. Interestingly, we also find the FOXA2 motif highly overrepresented in dEN where the factor is active, and also dEC where the factor is inactive but becomes expressed at a later stage of neural differentiation (Kriks et al., 2011), but not in dME (Fig. 5.17b).

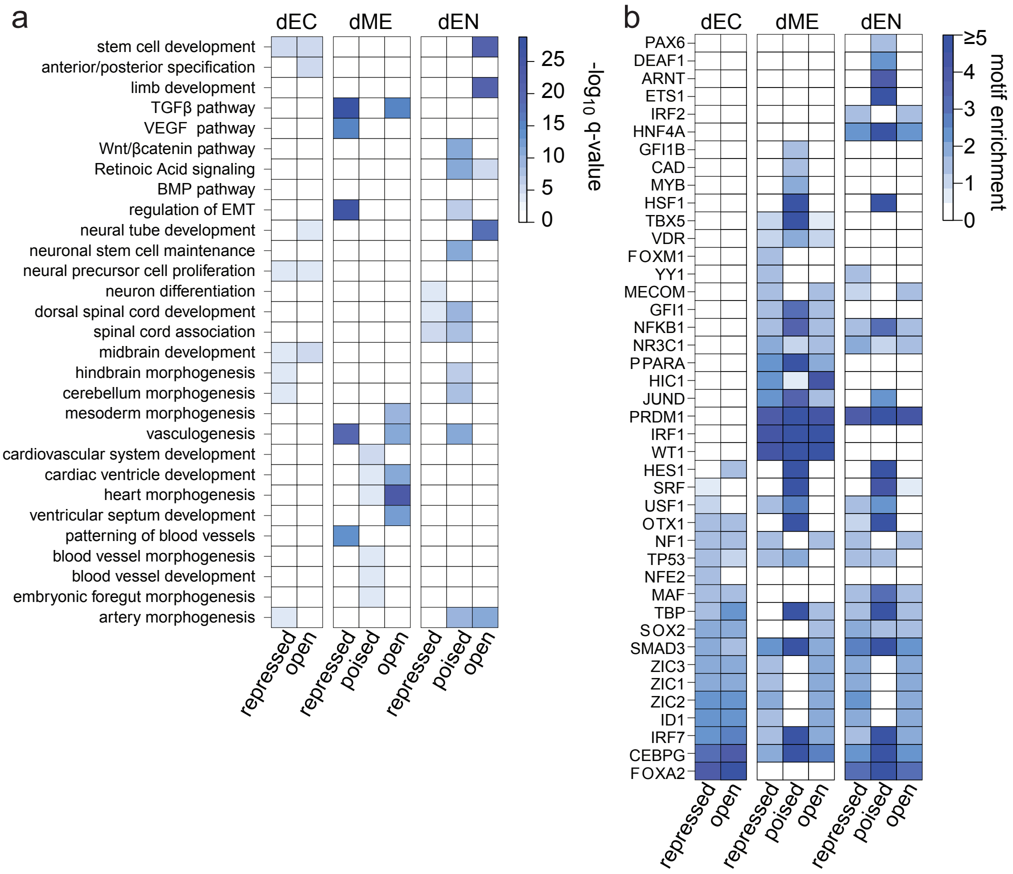


FIGURE 5.17: **a.** GO categories enriched in regions transitioning toH3K27ac in the cell type indicated on the right compared to hESCs as determined by GREAT analysis. Regions gaining H3K27ac were split up by state of origin in hESC into repressed (None, IMR, HMR, HK27me3), poised (H3K4me1/H3K27me3) and open (H3K4me3/H3K27me3, H3K4me3, H3K4me1). Color code indicates multiple testing adjusted q-value of category enrichment. **b.** TF motifs enriched in regions changing to H3K27ac in the cell type indicated on the right compared to hESCs. Color code indicates motif enrichment score incorporating total enrichment over background as well as differential expression of the corresponding TF in the respective cell type. Regions were split up by state of origin in hESCs similar to panel a. For each region class, the eight highest-ranking motifs are shown.

5.4.9 Acquisition of H3K4me1 without transcriptional activation suggests epigenetic priming

Many regions that exhibit high DNAme in hESCs and transition to H3K4me1 in one lineage remain HMRs in the two alternative cell types (Fig. 5.18a). Similar to the regions showing dynamic DNAme during differentiation, these regions are typically intergenic and $>$ 10 kb from the nearest TSS (Fig. 5.18b).

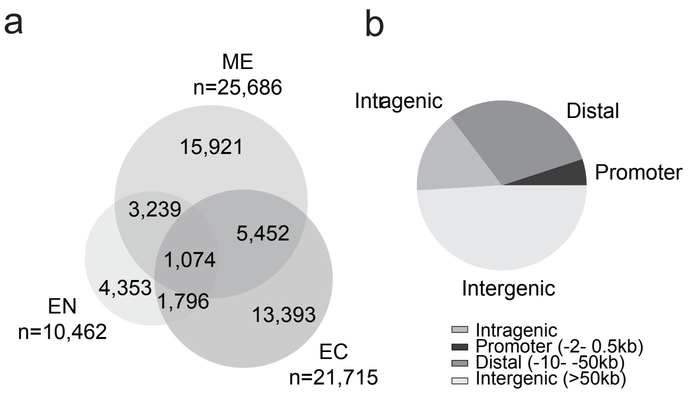


FIGURE 5.18: **a.** Overlap of regions gaining H3K4me1 in the three differentiated populations relative to hESCs. **b.** Genomic distribution of all regions gaining H3K4me1 compared to hESCs in at least one of the three differentiated populations.

GREAT analysis of these regions shows a strong enrichment for categories associated with brain development such as cerebellum morphogenesis in dEC ($q < 10^{-30}$) , TGbb pathway targets ($q < 10^{-10}$) in dME and suppression of EMT in dEN ($q < 0.0001$). To understand if regions that gain H3K4me1 in our system are associated with somatic identity, we took advantage of published microarray data for 24 human tissues and determined genes upregulated in these tissues with respect to hESCs (termed, Tissue Atlas, see Extended Experimental Procedures). Reaffirming the relevance of our dynamics, we found regions that gain H3K4me1 in dEC are associated with fetal brain and specific cell types found within the adult brain (Fig. 5.19a) based on region association with its nearest gene. The dME H3K4me1 pattern was associated with a range of interrogated tissues, such as heart, spinal cord and stomach, which may be due to heterogeneity of the tissues collected (Fig. 5.19c). The dEN associations were interesting given that, as with the RNA-Seq and H3K27ac trends, H3K4me1 was again associated with brain-related categories (Fig. 5.18c). Overall, less than half of the genes that gain H3K4me1 exhibit immediate transcriptional changes (Fig. 5.19b).

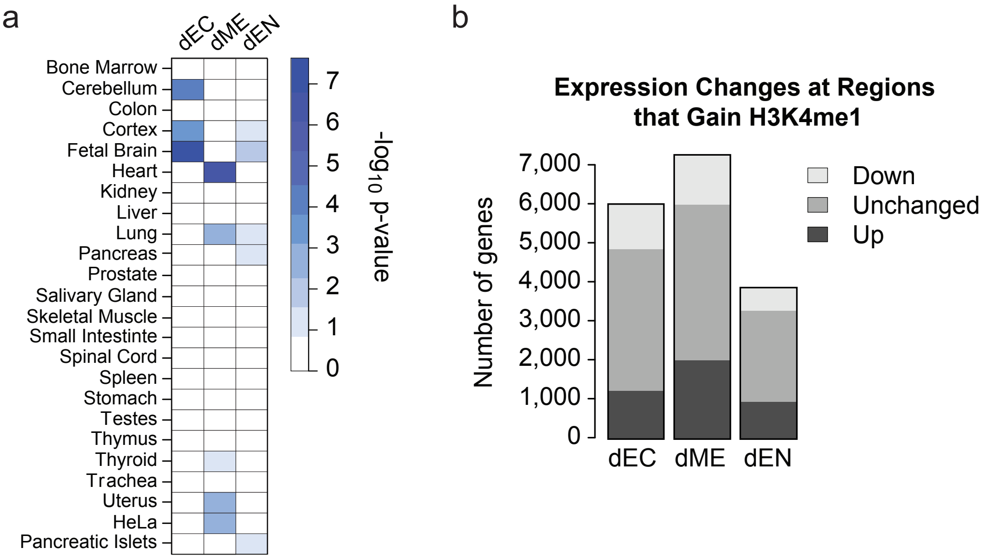


FIGURE 5.19: **a.** Tissue signature enrichment levels of genes assigned to regions specifically gaining H3K4me1 in the differentiated populations indicated on the bottom. **b.** Number and distribution of gene expression changes of genes assigned to regions gaining H3K4me1 in the differentiated populations. Associated genes were classified as either being up/down- regulated or unchanged relative to hESCs.

CYP2A6 and CYP2A7 (Fig. 5.20a) are representative examples that do not show a corresponding change in expression, while LMO2 does (Fig. 5.20b). To investigate these regions in more detail, we carried out motif enrichment analysis and found lineage specific enrichment of TF motifs near regions that gain H3K4me1. While the FOXA2 motif is enriched in all three cell types, the DBX1 motif is associated with the gain of H3K4me1 in dEC (Fig. 5.21a), which coincides with its transcriptional activation in this cell type (FPKM: 5.36). Conversely, the GLI3, HIC1 and CTF1 motifs are strongly enriched at regions that gain H3K4me1 in dEN (Fig. 5.21a). To further assess if this DNAme to H3K4me1 switch acts as a priming event, we differentiated the HUES64 endoderm population for five additional days in the presence of BMP4 and FGF2, leading to HNF4a positive hepatoblast-like (dHep) cells (data not shown). Interestingly, of the motifs enriched in dEN that gain H3K4me1, HIC1, KLF4 and CTF1 (Fig. 5.21a), several of these genes become expressed at the next stage of differentiation (Fig. 5.21b). Lastly, 1,346 of these putatively primed regions are enriched for the active enhancer mark H3K27ac in human liver (Fig. 5.21c).

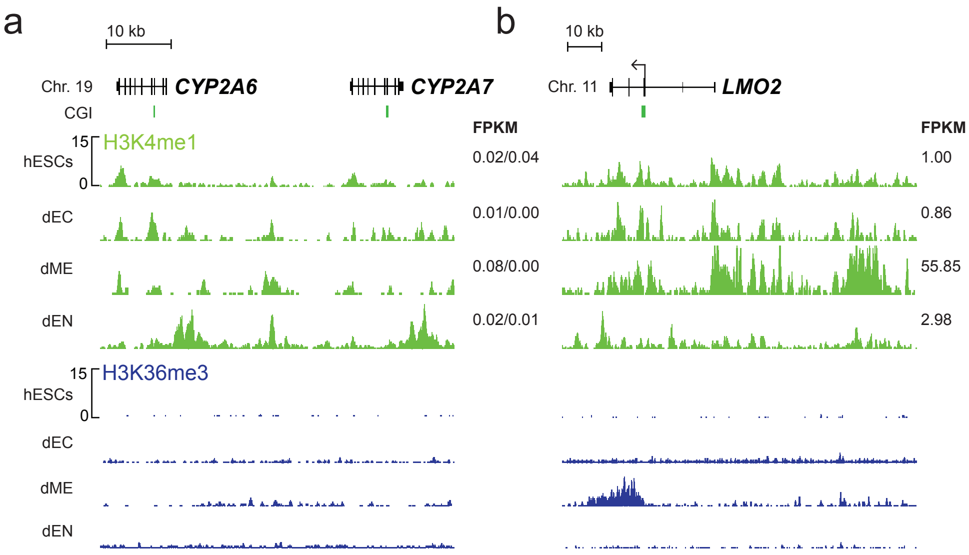


FIGURE 5.20: **a.** Normalized ChIP-Seq tracks (H3K4me1 and H3K36me3) for the LMO2 locus (Chr.11:33,865,134-33,977,858). Read counts on y-axis are normalized 10 million reads for each cell type. CGIs are indicated in green. **b.** Normalized ChIP-seq tracks (H3K4me1 and H3K36me3) for the CYP2A6/CYP2A7 region (Chr19: 41,347,260-41,395,599). Read counts on y-axis are normalized 10 million reads for each cell type. CGIs are indicated in green.

5.4.10 Loss of DNAme and acquisition of H3K27me3 at putative regulatory elements

More surprisingly, we observe intergenic regions that switch from high DNAme to H3K27me3 (n=3,985 in dEN) (Fig. 5.22a). This transition frequently occurred within CpG poor, distal regions, which is distinct from the common CpG island-centric targets of Polycomb Repressive Complex 2 and H3K27me3 (Fig. 5.22b). This switch is highly lineage specific and DNAme is generally retained in the alternative two cell types (Fig. 5.22c). Motif enrichment analysis, combined with the evaluation of publicly available TF binding site (TFBS) data from the ENCODE project, indicated that many regions exhibiting this transition in dEN were near binding sites of the pioneering factor FOXA2. This TF has putative roles in chromatin decompaction, but its distinct functions and limitations remain somewhat unclear (Li et al., 2012). To investigate this association, we performed ChIP-Seq for FOXA2 in the endoderm population. This analysis reveals that FOXA2 binding sites frequently overlap with regions that transition from HMR to H3K27me3 (Fig. 5.22d). We also confirmed that gain of H3K27me3 at dEN FOXA2 binding sites occurs predominantly in dEN, and not dEC or dME (Fig. 5.22e). A notable example of this transition can be seen at the ALB locus, where H3K27me3 is gained at

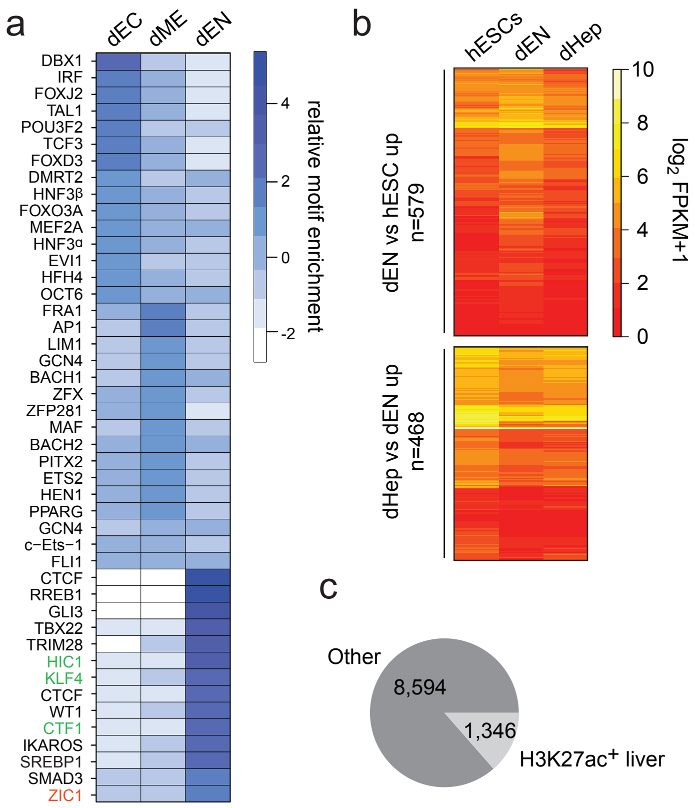


Figure 5.21: **a.** Normalized motif enrichment scores for the top 15 motifs enriched in regions specifically transitioning to H3K4me1 in the differentiated cell type indicated on the bottom. Motif highlighted in red corresponds to a TF that is upregulated at the next stage (hepatoblast) of endoderm differentiation while motifs highlighted in green are specifically upregulated in dEN but downregulated at the dHep stage. **b.** Gene expression levels of genes assigned to regions gaining H3K4me1 specifically in dEN compared to hESC and being upregulated in dEN but not hepatoblast (top). Gene expression levels of genes being upregulated between dEN and dHep (but not between hESC and dEN) and gaining H3K4me1 in dEN are shown on the bottom. **c.** Fraction of regions changing to H3K4me1 in dEN and being enriched for H3K27ac in human liver (n=1,346).

AFP and AFM, proximal to FOXA2 binding sites (Fig. 5.23a). This mark is not found in primary liver tissue, suggesting it represents a transient state (Fig. 5.23a). Many regions that exhibit this transition are required for later stages of development as with AFP and AFM or HBB1 in the dME. Importantly, the majority of these regions do not yet exhibit significant increases in expression (Fig. 5.23b).

A previous report found that FOXA1/FOXA2 could bind to regions exhibiting DNAme (Serandour et al., 2011), which is not a characteristic shared by all TFs (Zaret and Carroll, 2011). Regions bound by these factors subsequently lost DNAme and gained euchromatic histone modifications in our populations. We therefore compared DNAme at FOXA2 binding sites in hESCs to dEN and found a slight reduction in specifically in the dEN (Fig. 5.23c). To more directly assess this relationship, we interrogated the DNAme

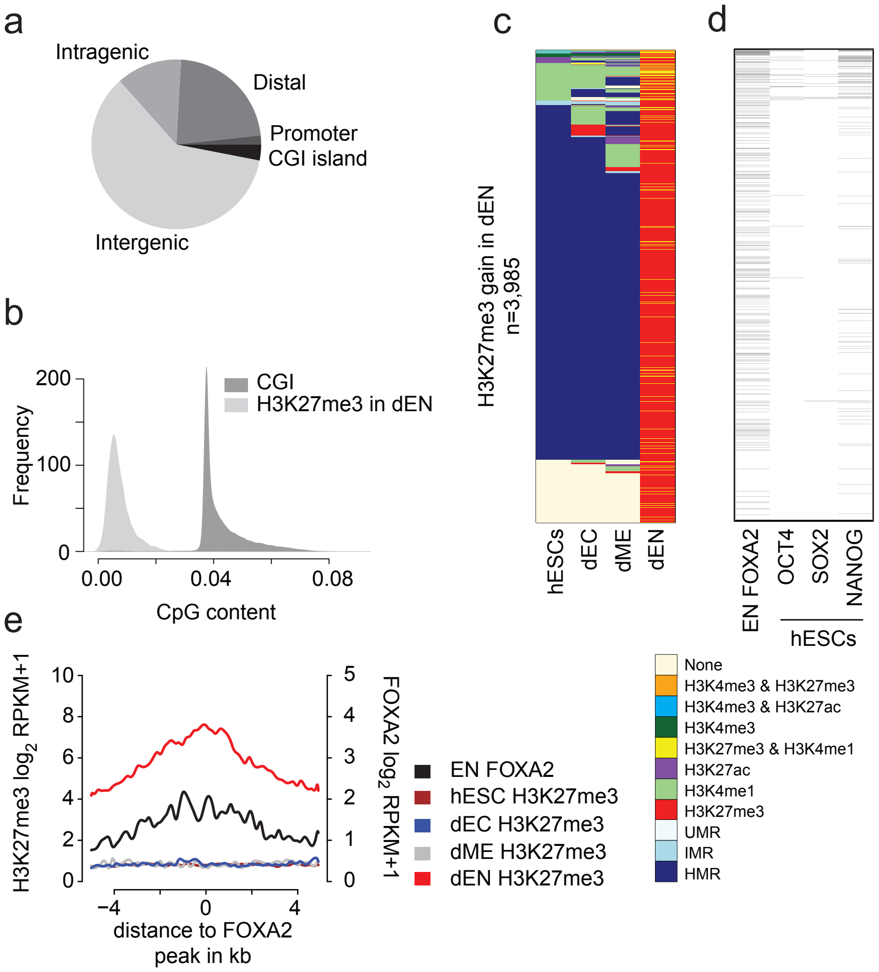


FIGURE 5.22: **a.** Distribution of genomic features associated with region gaining H3K27me3 (n=22,643) upon differentiation to any of the three hESC derived cell types compared to hESC. **b.** CpG content distribution of regions gaining H3K27me3 upon differentiation. For reference, the CpG content distribution of CpG islands is shown. **c.** Epigenetic state distribution in hESC, dEC and dME of regions that gain H3K27me3 in the dEN population compared to hESC. **d.** Binding profile of FOXA2 in dEN (n=357), OCT4 (n=32), SOX2 (n=12) and NANOG (n=124) in hESC across regions that gain H3K27me3 in dEN upon differentiation. **e.** Composite plot of median normalized tag counts (RPKM) of regions bound by FOXA2 in dEN and gaining H3K27me3 in dEN compared to hESC (n=357).

state of regions isolated by FOXA2-ChIP-Bisulfite sequencing in dEN (Brinkman et al., 2012). Interestingly, we saw a major depletion of DNAme at sites isolated by FOXA2-ChIP (Fig. 5.23c). To determine if these regions exhibit transcriptional activation after further differentiation we examined again our dHep RNA-Seq data and found that 50 genes, which were bound by FOXA2 and gained H3K27me3 in dEN, increased their expression (Fig. 5.23d) We also find H3K27ac enrichment at 197 loci in the human liver that had experienced the gain of H3K27me3 in dEN (Fig. 5.23e).

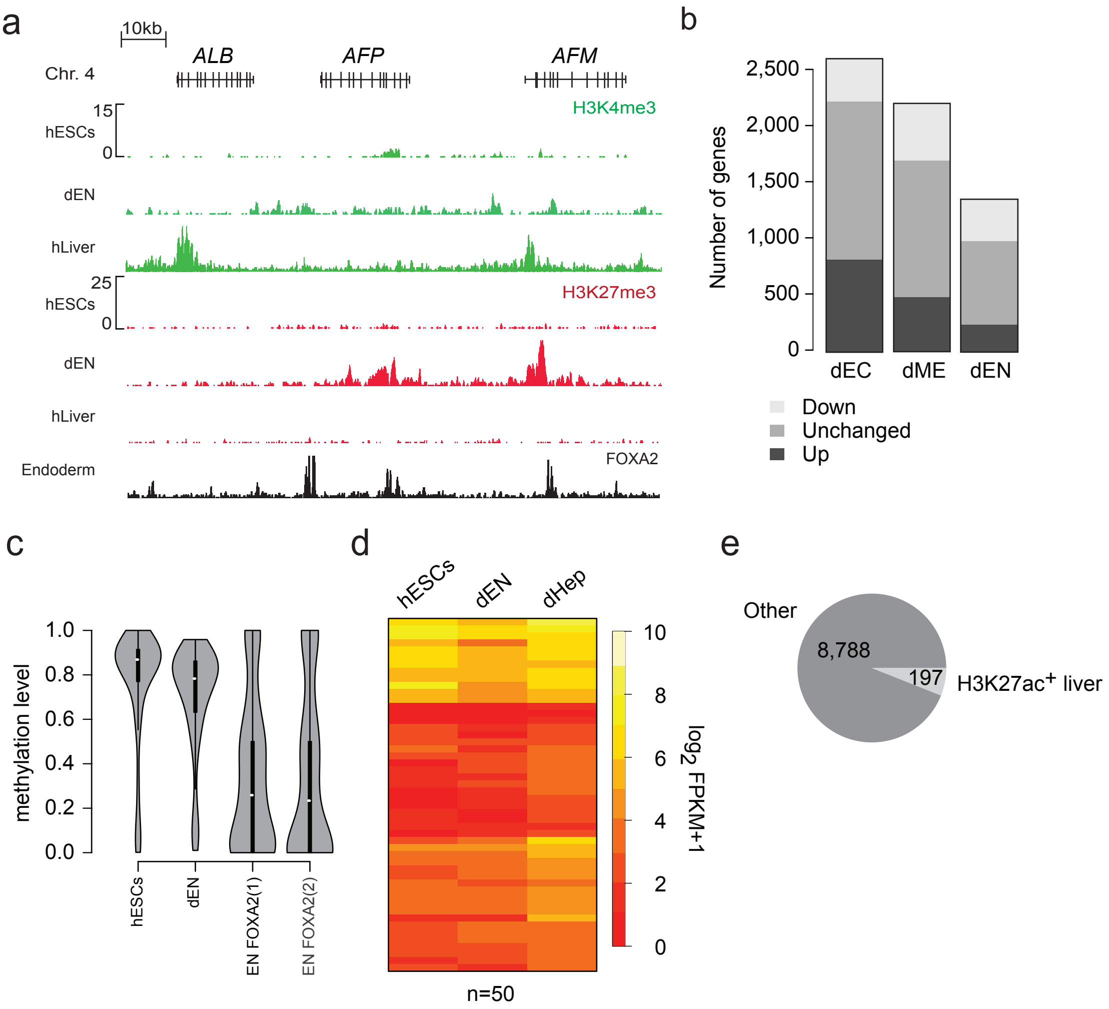


FIGURE 5.23: **a.** Normalized H3K27me3 and H3K4me3 ChIP-seq tracks for hESCs, dEN and human adult liver tissue at the ALB locus (chr4:74,257,882-74,377,753). Black bars (bottom) indicate TF binding of OCT4, SOX2 or NANOG in hESCs. Read counts on y-axis are normalized to 10 million reads. **b.** Classification of gene expression associated with regions gaining H3K27me3 in each germ layer into either up-regulated (FDR< 0.05), down-regulated (FDR< 0.05), or unchanged. **c.** Distribution of methylation levels of regions bound by FOXA2 and gaining H3K27me3 in dEN. DNAme information is depicted for hESC and dEN WGBS datasets and two biological replicates of FOXA2 ChIP-Bisulfite experiments in dEN (n=357). **d.** Gene expression profile of genes upregulated at the hepatoblast stage relative to dEN that are associated with regions bound by FOXA2 and gaining H3K27me3 in dEN (n=50). **e.** Fraction of regions gaining H3K27me3 in dEN and being enriched for H3K27ac in human liver (n=197).

5.5 Discussion

Using directed differentiation of hESCs to three distinct, FACS-enriched populations, representing early stages of embryonic development, we provide an extensive set of new data and many insights on the transcriptional and epigenetic dynamics that occur during human in vitro lineage specification.

Among other things we describe two very interesting, but distinct lineage specific dynamics from high DNAme to H3K4me1 or H3K27me3. These transitions occur at many sites that do not significantly change gene expression during our early stages of differentiation. Notably, we made similar observations for H3K4 methylation during the early stages of reprogramming to an iPS state (Koche et al., 2011), suggesting that this type of epigenetic priming event might be common. At this point, however, it is not clear whether these events reflect a regulatory mechanism to facilitate timely activation upon differentiation or indicate the absence of a critical co-factor necessary for complete transcriptional activation. We also cannot rule out that a subset of the observed priming events are due to heterogeneity in the cell population that are not detected by our RNA-Seq. Our observation that high DNAme switches to H3K27me3 enrichment in distal, CpG-poor regions is even more interesting. It remains to be tested whether targeted loss of DNAme at these regions causes a default gain of H3K27me3 in the absence of additional co-factors due to underlying sequence context (Mendenhall et al., 2010) or represents a more active recruitment event and regulatory mechanism. It is also possible that H3K27me3 gain at distal regions is due to genomic conformation changes and reflects H3K27me3 spreading in three dimensions. It was recently reported that the combination of H3K27me3 enrichment and a nearby nucleosome-depleted region creates sites amenable to TF binding (Taberlay et al., 2011). Based on these results, one may speculate that specific TFs, such as FOXA2, exert chromatin decompaction functions resulting in loss of DNAme leading to gain of H3K27me3, which creates a platform for subsequent binding of other TFs that cannot directly remodel a heterochromatic state, but instead function in transcription machinery assembly and transcriptional activation. In conclusion, our data provide new insights on transcriptional and epigenetic dynamics during hESC specification and represent valuable reference maps for many applications, including regenerative biology and the study of human developmental biology.

Chapter 6

Dissecting neural differentiation regulatory networks through epigenetic footprinting

6.1 Introduction

An extended and updated version of this chapter is part of Ziller & Edri et al. is currently in press at Nature. The paper and analysis strategy was conceived by M. J. Ziller (MJZ). All analysis and computer code development was carried by MJZ as well as the writing of the manuscript. R. Edri performed cell culture and FACS. ChIP-Seq library generation was performed by A. Goren, RNA-Seq library generation was carried out by C. Gifford and WGBS/RRBS library production was done by H. Gu. All data generated in this study is publically available udner GEO accession number GSE62193.

Chapter 3 established DNAme as a general marker that is suitable to gain insights into the regulatory logic of distinct cellular states. This observation was exploited in Chapter 4 where it was shown that the combination of different epigenetic marks and gene expression data can be used to dissect the differentiation process of human ES cells into three distinct but related cell populations. The power of this high-resolution profiling approach of related cell types in combination with a TF-centric analysis became apparent when we identified a novel epigenetic state switch that is likely mediated by the pioneering TF FOXA2. Furthermore we were able to identify numerous pathways activated and shutdown in the distinct cell populations using histone modification and

DNAme information, highlighting the utility of epigenetic information to dissect differentiation processes.

However, even though our previous work and that of others (Ernst and Kellis, 2010, Thurman et al., 2012) highlights the power of epigenetic information to provide insights into the underlying transcriptional networks, we are still lacking a coherent conceptual framework to interpret epigenetic data from regulatory network perspective. This should permit the extraction of TFs likely to drive specific epigenetic dynamics during cell state transitions and therefore be suitable to infer key regulators of these transitions. Furthermore, it is unclear what epigenetic marks are most informative for this purpose. Are certain histone modifications and DNAme complementary or redundant with respect to the information they provide with respect to the identification of key regulators? In Chapter 5 we showed that H3K4me1 seems to be associated with transcriptional priming, suggesting that at regions gaining H3K4me1 at a certain differentiation stage are likely to be enriched for TFBS of factors that gain expression at a later time point where the region switches from primed to active, e.g. through acquisition of H3K27ac. Here, we address these questions and investigate transcriptional and epigenetic dynamics across a differentiation time series along the neural lineage.

Human pluripotent stem cell-derived models that accurately recapitulate neural development *in vitro* and allow for the generation of specific neuronal subtypes are a major focus in the stem cell and biomedical research communities. Two of the key challenges are identifying the proper conditions to derive functional cell types as well as an in-depth molecular characterization of the regulatory events that establish and maintain the desired cellular states. Here, we report the transcriptional and epigenomic analysis of six consecutive stages of hESC differentiation along the neural lineage aimed at modeling key cell fate decisions including specification, expansion and patterning during the ontogeny of neural stem and progenitor cells. In order to dissect the regulatory mechanisms that orchestrate the stage-specific differentiation process we developed a computational framework to infer key regulators of each cell state transition based on the progressive remodeling of the epigenetic landscape. To do so, we exploited the fact that epigenetic dynamics help demarcate gene regulatory elements (Ernst et al., 2011, Gifford et al., 2013, Xie et al., 2013). The coordinated epigenetic changes are likely the result of differential TF activity and hence relevant to the respective differentiation process. To identify key regulators, we introduce the novel concept of TF epigenetic remodeling activity (TERA). This concept tries to explain epigenetic changes, e.g. H3K27ac or

DNAme dynamics during differentiation, through the differential activity of TF motifs present at footprints located within genomic regions that change their respective epigenetic state. Since the epigenetic remodeling activity of these motifs is unknown, we infer the most likely activity level combination at each time point for a large set of motifs and their corresponding TFs (Fig. 6.1). This idea allows to identify key regulatory factors likely to mediate distinct sets of epigenetic remodeling events at each cell state transition.

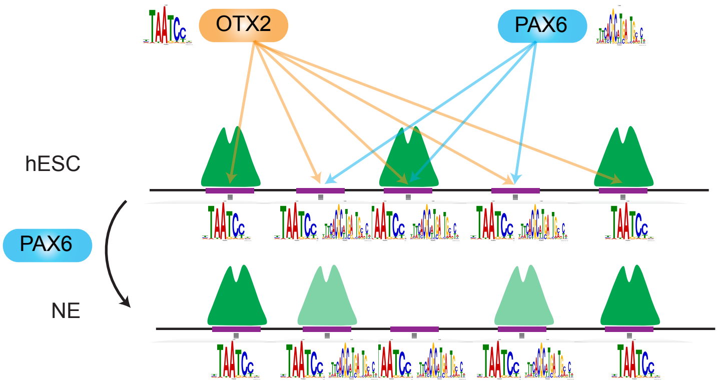


Figure 6.1: Illustration of the TERA principle. During the inference of TERA levels for all factors, the most likely activity levels are determined by trying to predict the epigenetic state at each time point from the TERA scores of all factors. The TERA scores minimizing the prediction error are then assigned. In the depicted toy example OTX2 is more likely to drive the epigenetic state in hESCs, whereas increased PAX6 activity can explain the change to NE better than increased OTX2 activity.

Our analysis suggests that a stably expressed core TF network comprising PAX6 and OTX2 cooperates with stage-specific factors and signaling pathways to regulate commitment and proper differentiation towards neuronal and glial cell types. Notably, gene regulatory elements involved in this process appear to be frequently disrupted by SNP's associated with different neurological pathologies such as Alzheimer's disease or schizophrenia, providing insights into potential causes and mechanisms. Taking advantage of our distinct differentiation stages, we are also able to refine our previous observation on epigenetic priming at TF binding sites that seems to be mediated by combinations of core and stage-specific factors. Additionally, we observe widespread DNAme changes from the pluripotent state to the neural lineage that are targeted to TFBS and gene regulatory elements involved in the early, but also subsequent stages of differentiation. This is in sharp contrast to stage-specific dynamics that occur in the consecutive populations of neural differentiation. Taken together, we demonstrate the utility of our reference maps and outline a general framework, not limited to the context of neural differentiation, to dissect regulatory circuits of differentiation.

6.2 Experimental methods

For a detailed description of the experimental methods see (Edri et al., 2014, Ziller et al., 2014) and Appendix A.2. Briefly, human ES cell line H9 (H9; WA-09; Wicell) expressing GFP under the HES5 promoter were directed towards neuroectodermal fate using dual SMAD inhibition (Chambers and Tomlinson, 2009). Neural rosettes were harvested beginning day 8-10 of differentiation. Rosettes were replated in on DMEM/F12 with N2 supplement as well as SHH, FGF8 and BDNF. In the following two weeks of differentiation, these factors were gradually replaced by FGF2 and EGF. Neuroectoderm/Neural progenitor cells (NPCs) were collected at day 12, 14, 35, 80 and 220 using FACS sorting for HES5:GFP and subjected to RNA-Seq, ChIP-Seq for histone modifications and/or WGBS/RRBS profiling.

6.3 Computational methods

6.3.1 Data processing

Data processing was conducted as described above for ChIP-Seq, RNA-Seq and WGBS data. Ascl1 ChIP-Seq data for Ascl1 overexpression in NHDF and NHEK was taken from GSE43916 and was processed in the same way. Subsequent to alignment, peaks were called using MACS (Zhang et al., 2008) against the matching whole cell extract (WCE), removing all duplicates and only retaining peaks significant below 10^{-5} and fold enrichment ≥ 5.

6.3.2 Differential expression analysis

Differential expression analysis was carried out using Cuffidff 2 (Trapnell et al., 2013) and only genes with an absolute minimum $\log_2$ difference above 1.5 and FDR ≤ 0.11 between any hESC, NE, RG or INP were retained for the NPC analysis. We chose this moderately stringet cutoff based on two considerations: First, we wanted to eliminate expression changes that are confined to the majority of the HES5+ population, eliminating spurious expression changes due to imperfect FACS isolation etc. Second, we determined the overlap of differentially expressed genes at different FDR levels with known key marker genes expressed at each stage as determined by qPCR and immunohistochemistry (see (Edri et al., 2014) for this data). The chosen values represent a compromise between these two considerations and give rise to the unusual FDR value. Due to the absence of replicates, we required a minimum log_2 difference above 2 and FDR ≤ 0.05 for differentially expressed genes across the terminally differentiated cells NEdN, E-RGdN, L-RGdO and L-RGdA. These resulting sets of differentially expressed genes were used in all subsequent analyses involving differentially expressed genes between two conditions.

6.3.3 ChIP Seq data analysis and normalization

Basic ChIP-Seq data processing and analysis was conducted as described earlier Section 3.3.2). For differential enrichment analysis between the different samples, only regions significant at a q-value ≤ 0.001 and minimum fold change greater than three were considered differentially enriched. For partial least square regression (PLS) (Boulesteix and Strimmer, 2007), we calculated the reads per kilobase per million reads sequenced values for each footprint region and applied quantile normalization to the $\log_2$(RPKM+1) transformed enrichment scores for each epigenetic mark separately. Finally, we subtracted the whole-cell extract $\log_2(\text{RPKM}+1)$ counts to normalize for distinct background enrichment levels. Regions with negative enrichment scores were subsequently set to 0. The resulting normalized gene expression and histone modification datasets were then used as input for the analysis pipeline outline in Figure 6.2 and described below in detail.

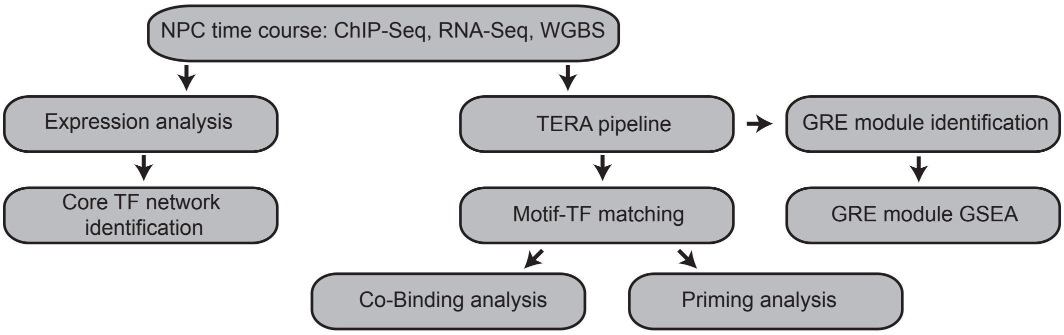


Figure 6.2: Outline of flow and key analysis modules used to generate the content of this chapter.

6.3.4 Gene set enrichment analysis

Gene set enrichment analysis for genomic regions was carried out using the GREAT toolbox (McLean et al., 2010) and only categories with q-values ≤ 0.05 for both the Hypergeometric and the binomial test as well as a minimal region enrichment level greater than 2 were considered. A selected subset of the resulting enriched GO categories is shown in Fig. 6.11a and Fig. 6.12b. In addition, we utilized the MGI expression gene set collection (Finger et al., 2011) throughout our analysis, only considering neural related gene sets that contain any of the following substrings: prosencephalon, forebrain, diencephalon, telencephalon, midbrain, mesencephalon, hindbrain, metencephalon, myelencephalon, rhombencephalon, rhombomere, spinal cord, cerebellum, cerebral cortex, hippocampus, subventricular zone. Since the resulting gene set collection for the latter structures is available for multiple developmental time points, we uncoupled the anatomical structure and temporal information and used the gene set with the lowest q-value in Fig. 6.11a and Fig. 6.12b. For expression-based gene sets, we also took advantage of the MGI database but used Fisher's exact test to determine enriched gene sets and uncoupled the resulting MGI gene sets in the same fashion for structures and time points. Only gene sets significant at p-values below 0.05 and odds ratio above 1.5 were considered. Gene set enrichment levels not meeting these criteria in any condition were set to 0 in the respective plots.

6.3.5 Footprinting detection

To determine small regions depleted of histone modifications but surrounded by regions of much greater enrichment, termed footprints, we extended an approach used for the analysis of DNAse I HS data (Neph et al., 2012). Our footprints identification algorithm consisted of three main phases: In the first phase, we used MACS (Zhang et al., 2008) to identify regions on a genome scale enriched against background for histone modification signal (peaks). We ran MACS for each track of histone modifications separately with the corresponding whole cell extract as control using the following parameter: ”-g 2.7e9 –tsize=36 –pvalue=1e-5 –keep-dup=1”. In the second phase, we identified footprints located within/around peak regions in the following manner:

1. For each peak, extend by 400 bp from apex in either direction.

2. Split entire resulting region into bins of size 20 bp.

3. Compute number of RPKM counts for a central sliding window across the entire region (shifting by increments of one bin) for different window sizes ranging from two bins to ten bins in increments of one.

4. For each position of the central window and for each window size, compute the following three quantities: C_{ij} - RPKM count for central window at current position i and window size j, R_{ij} - RPKM count for a 200 bp stretch directly to the right of the central window and L_{ij} - RPKM count for a 200 bp stretch directly to the left of the central window.

5. For each resulting position i and window size j compute the depletion score:

$$e_{ij} = f\frac{C_{ij}+1}{2L_{ij}} + f\frac{C_{ij}+1}{2R_{ij}}$$

   With the footprint size normalization factor $f = s/b$, with s the size of the central window and b the size of the border regions.

6. Identify non-overlapping, non-adjacent footprint candidates starting from small to larger central window sizes and recording footprint candidate iff $e_{ij} > 0$ AND $e_{ij} < 1$ AND $L_{ij} > Cij$ AND $R_{ij} > C_{ij}$, followed by removing all other potential footprints (central window+borders) of larger size overlapping the current candidate.

7. Finally, all resulting candidate footprints with a footprinting score $e_{ij} \leq 0.8$ were reported.

The outlined entire procedure was carried out for H3K27ac and H3K4me3 independently for each sample. Subsequently, we merged all footprints from individual samples into consensus footprints set for each epigenetic mark separately, collapsing overlapping footprints by taking the union of all regions with non-zero overlap.

6.3.6 DMR detection

DMR detection was carried out as described in Section 4.2.1. Pairwise comparisons of consecutive samples (hESC, NE, E-RG, M-RG, L-RG, LNP) were carried out on a single CpG level using a beta-binomial model and the beta difference distribution requiring a maximum p-value below 0.01 and an absolute methylation difference greater than

0.15. Subsequently, differentially methylated CpGs within 500 bp were merged into discrete regions. Differentially CpGs without neighbors were embedded into a 100 bp region surrounding each CpG. Next, differential methylation analysis was repeated on the region level using a random effects model. Only regions significant at p-value below 0.01 and an absolute methylation difference above 0.3 were considered differentially methylated and used for subsequent analysis.

6.3.7 Motif library construction and putative TF binding site identification

We combined the position weight matrices from Transfac professional database (Fogel et al., 2005) (2009) with the PWM collection reported in Jolma et al. (Jolma et al., 2013) only retaining motifs annotated for *Homo sapiens* or mouse. To eliminate redundant motifs, we clustered all resulting 1,886 PWMs using MATLIGN (Kankainen and Loytynoja, 2007) using the following parameters: ”-transv=-4 -transi=-4 -match=5 -gopen=-10 -gext=-1 -term=5 -norm=4 -spacer=1 -mode=1 -zscore=1 -random=0 -pseudo=0 -freqat=0.5 -freqcg=0.5”. A total of 551 motifs was retained after redundancy filtering and used for subsequent analysis. In order to determine putative binding sites in a given genomic region, we used a biophysical model of TF affinities to DNA (Manke et al., 2008, 2010) to determine putative binding to our footprint sets. This biophysical model requires the training of generalized extreme value distributions of binding affinities based on a PWM matrix for each TF and each set of genomic regions in order to generate a suitable background model. In order to take the distinct properties of footprints determined from different epigenetic marks as well as promoter and distal regions into account, we determined the GEV parameters for footprints arising from H3K27ac, H3K4me3 and DNAme as well as for promoters (±1 kbp of the TSS) and distal regions (everything except promoter) separately using the framework outlined by Manke et al. (Manke et al., 2008, 2010). The resulting six binding matrices were then filtered for minimal significant binding affinity at p-values below 0.05. All other entries with higher p-values were set to one. Next, we took the negative $\log_{10}$ of the entire matrix as a quantitative measure of binding affinity in subsequent analysis.

6.3.8 Inference of TF epigenetic remodeling activity based on epigenetic data

In order to infer the TF epigenetic remodeling activity (TERA) from epigenetic data, we first focused on motif activity analysis and associated each motif in a second step with its corresponding TF. For each epigenetic mark, we used the normalized epigenetic enrichment scores as input and only considered genomic regions that exhibited a score change of greater than 0.3 (H3K27ac), 0.2 (H3K4me1), or 0.5 (H3K4me3) and minimal DNAme difference of at least 0.3 to include only dynamic regions. To determine the unobserved epigenetic remodeling activity of a TF binding motif, we took advantage of recent developments in the microarray field (Boulesteix and Strimmer, 2005, 2007) and adapted this approach to epigenetic data. To that end we modeled the enrichment level y_{it} of a particular epigenetic mark at genomic region i and time point t as a linear function the unknown TF activities. Considering p predictor variables (TERA) and k time points we describe the unknown TERA X as a $p \times k$ matrix. For each epigenetic modification H3K27ac, H3K4me3 and DNAme separately, we determine the region set of size n that fullfills the above listed criteria of minimal change and assemble the observed epigenetic enrichment score matrix Y $(n \times k)$. We then employ the linear model:

$$Y = A + BX + E$$

With a constant offset matrix A $(n \times k)$, the connectivity matrix B $(n \times p)$, describing the filtered binding affinities for all TF motifs to all regions and an error term matrix E. Subsequently, we followed the approach outlined by Boulesteix and Strimmer (Boulesteix and Strimmer, 2005) and applied partial least square (PLS) regression and specifically the SIMPLs algorithm (Dejong, 1993) to determine the unknown TF motif activities. The idea in PLS is to employ a linear dimensionality reduction

$$T = BR$$

where the p predictors in X are mapped onto $c \leq \text{rank}(X) \leq \min(p, n)$ latent components T $(n \times c$ matrix) and to compute the weight matrix R not only based on the data matrix B but explicitly taking into account the response matrix Y. The latter strategy maximizes predictive power even for a small number of latent components.

In order to determine the number of latent components for each epigenetic mark and genomic context, we follow the procedure described by Boulesteix and Strimmer (Boulesteix and Strimmer, 2005). In this approach, the region set is randomly splitted into two sets, one for training purposes (2/3 of the regions) and one for test purposes comprising 1/3 of all regions. Next, the TERA values are determined for different numbers $p = 1, 2 \ldots, p_{\max}$ of hidden components based on the training dataset. Subsequently, we use each of the resulting inferred TERA values sets to predict the epigenetic enrichment scores on the test set. After repeating this entire process 20 times we then compute the mean squared prediction error for each tested number of hidden components p and choose p such that it minimizes the mean squared prediction error. The corresponding analysis methodology was implemented in the statistical programming language R adapting the implementation provided by Boulesteix and Strimmer (Boulesteix and Strimmer, 2005) and can be found in the code archive accompanying this work in the TERA folder.

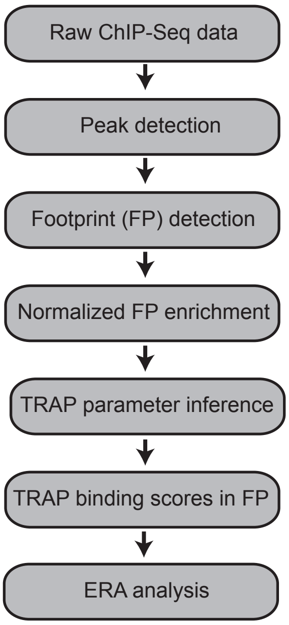


Figure 6.3: Outline of flow and key analysis steps in the computation of the TF epigenetic remodeling activities.

Finally, we determined the TF gene that best matches a given motif by correlating all gene expression measurements for TFs reported to be associated with a particular motif and its inferred motif activity score for each epigenetic mark and promoter/distal class separately across our differentiation time course. We then chose the factor with the highest absolute Pearson correlation. The raw associations of motifs with TFs can be found in Table 6.2 in the electronic archive accompanying this work. The TF-motif associations were derived based on the reported associations in either the TRANSFAC database or the original publication of the motif sets (Jolma et al., 2013). A summary of the entire pipeline workflow is given in Figure 6.3.

6.3.9 GRE-module detection

To identify global trends in epigenetic changes that are associated with distinct sets of TFs, we took advantage of the decomposition in terms of latent components that result from the PLS analysis. To that end, we first associated each motif with the latent component for which this motif exhibited the highest loading score b_i. Next, we took advantage of our motif to TF-mapping (see previous Section) and assigned each motif to one TF. To associate individual genomic footprint regions with one latent component, we computed the correlation of each latent component's activity with the H3K4me3 enrichment levels of H3K4me3 based footprints that exhibited a minimal estimated binding strength of ≥ 1.3 for at least one of the motifs that were associated with the latent component of interest. Finally, we assigned each footprint to the latent component for which it exhibited the highest absolute Pearson correlation. The resulting footprint subsets were then individually subjected to GREAT (McLean et al., 2010) analysis.

6.3.10 Co-binding analysis

Co-binding relationships were evaluated using an empirical approach with the entire set of footprints for each epigenetic mark as background. For a given factor i, we determined the footprint set F_i relevant for the current comparison (e.g. changing their epigenetic state in particular cell state transition) that were predicted to harbor a TFBS based on the binding model outlined above. Next, we computed the frequency of motif co-occurrence F_{ij} across F_i for all other motifs j in our database. To generate a proper null distribution, we randomly sampled $K = 100$ size standardized footprint sets G_k of cardinality $|F_i|$ from the entire footprint collection for the epigenetic mark under study and computed the same test statistic G_{ij}^k on these sets. Finally, we determined an empirical p-value and odds ratio based on these quantities by counting the number of instances for which $s_i^{G_k} j \geq s_{ij}^F$:

$$p_{ij} = \frac{\sum_k s_{ij}^{G_k} \geq s_{ij}^F}{K}$$

Only co-binding relationships significant at p-value ≤ 0.01 and odds ratio ≥ 1.5 were retained.

6.3.11 GWAS analysis

The GWAS analysis was conducted using 11,027 GWAS SNPs from the GWAS catalog (Hindorff et al., 2009) (August 2013). For each footprint set k, we sampled $K = 100$ randomly selected, size and GC-content matched control region sets of equal size and determined the overlap with GWAS SNPs for control and footprint sets of interest. Subsequently, we computed an empirical p-value for each trait/disease i in the catalog by determining the number of trait-associated SNPs s_{ij}^{k} overlapping with each control region set C_j^k and the number overlapping with the original footprint set k according to

$$p_{ki} = \sum_j \frac{s^{ki} \geq s_{ki}^{C_j^k}}{K}$$

6.3.12 Determination of core network

The core network was defined as those TFs that fulfilled the following criteria: Let x_i be the vector of normalized gene expression values for factor i across the entire time course except hESCs and let x'_i be the TF expression only in the NPCs (NE, E-RG, M-RG), then a factor was considered part of the core network iff:

1. $(\max(x'_i) - \min(x'_i))/(\bar{x}'_i) \leq 0.3$ and
2. $(\max(x_i) - \min(x_i))/\bar{x}_i \geq 1$ and
3. $x_{ij} \geq 4$ in all NPC states
4. Significantly upregulated (FDR ≤ 0.05 and $\log_2$ fold change ≥ 1.5) during neural induction from hESC to NE.

In addition to the factors identified by these means, we added PAX6 that failed to meet criterion 2 and was still expressed in a subset of the terminally differentiated populations. However, image-based analysis suggests absence of this gene in the truly differentiated populations and highly expressed in the NPCs, which is also supported by the literature. Since PAX6 was excluded from the core network because of technical reasons, we decided to manually add the factor to our core network.

6.3.13 TF binding site priming analysis

To determine TFs associated with TFBS priming prior to factor activation, we determined all TFs significantly upregulated (FDR$\leq$ 0.05 and $\log_2$ fold change ≥ 1.5) for the following pairwise comparisons: NE vs. E-RG, E-RG vs. M-RG, M-RG vs. L-RG, NE vs. NEdN, E-RG vs. E-RGdN, M-RG vs. L-RGdO, M-RG vs. L-RGdA. Only factors with expression levels ≥ 1 FPKM in the first condition of each pairwise comparison was considered. Next, we filtered all motifs associated with each candidate factor for those motifs with an increase in motif activity score based on H3K4me1 and DNAme during the cell state transition prior to the induction of the gene, e.g., hESC to NE for TFs gaining expression during NE to E-RG and NE to NEdN.

Factors with motifs fulfilling the latter condition were considered candidate primed factors, some of which are depicted in black in Fig. 6.15b. Finally, we determined which of the primed candidate factors exhibited a significantly increased predicted binding frequency (p-value ≥ 0.05, for both Fisher's exact test and permutation test) in DNAme and/or H3K4me1 footprints associated with genes changing their expression level ($\log_2$ FPKM change ≥ 1) during the transition in which the particular TF also does become induced. For this analysis, footprints were assigned to their closest gene.

6.4 Results

6.4.1 Validation on REMC data and ENCODE data

To validate the outlined strategy *in silico* we took advantage of publically available TF ChIP-Seq data in four cell lines from the ENCODE project (Bernstein et al., 2012) as well as H3K27ac and RNA-Seq data for 39 cell types from the REMC project. To that end, we investigate the following steps in detail:

1. Evaluation of footprinting based inference of TF binding.
2. Evaluation of distinct normalization strategies for histone modification ChIP-Seq data.
3. Evaluation of inferred TF epigenetic remodeling activities.

6.4.1.1 Evaluation of footprinting based inference of TF binding

In order to evaluate the performance epigenetic footprints - essentially dips in histone-modification peaks - we downloaded H3K27ac data as well as processed TF binding data from the ENCODE project (Bernstein et al., 2012) for the cell line K562 since abundant TF binding data based on ChIP-Seq was available for these lines. In addition, this dataset has been successfully used in several studies to benchmark TF binding predictions (Sherwood et al., 2014, Thurman et al., 2012). We then applied our TERA-pipeline to the H3K27ac datasets and computed the TF-binding affinities for a set of 557 distinct motifs. With these datasets at hand, we computed the true positive rate (TPR), the false positive rate (FPR) and the positive predictive values (PPV) for all TFs that could be matched to at least one motif with available binding affinities (46/117). In the event that one factor matched multiple motifs, we chose the motif with the highest AUC, similar to previous reports (Sherwood et al., 2014). While the overall performance is moderate with a median AUC of 0.73 and a PPV of 0.26 for $p = 0.05$ computed across all 46 TFs (Fig. 6.4), the results are within the range of current state of the art methods based on DNAse I hypersensitive sites (Sherwood et al., 2014), especially when considering the fact that we did not filter our TF ChIP-Seq peak sets for the presence of a PWM match within each TF ChIP-Seq binding site. However, only TF binding sites harboring a binding motif can be identified by motif based methods. Binding sites without motifs frequently correspond to indirect binding events indicating looping interactions which are frequently occurring at active enhancer. Earlier reports indicate that on average more than 50% of binding events measured by ChIP-Seq might be indirect as assessed by the absence of a corresponding TF motif (Thurman et al., 2012).as been successfully used in several studies to benchmark TF binding predictions (Sherwood et al., 2014, Thurman et al., 2012). We then applied our TERA-pipeline to the H3K27ac datasets and computed the TF-binding affinities for a set of 557 distinct motifs. With these datasets at hand, we computed the true positive rate (TPR), the false positive rate (FPR) and the positive predictive values (PPV) for all TFs that could be matched to at least one motif with available binding affinities (46/117). In the event that one factor matched multiple motifs, we chose the motif with the highest AUC, similar to previous reports (Sherwood et al., 2014). While the overall performance is moderate with a median AUC of 0.73 and a PPV of 0.26 for $p = 0.05$ computed across all 46 TFs (Fig. 6.4), the results are within the range of current state of the art methods based on DNAse I hypersensitive sites (Sherwood et al., 2014), especially when considering the fact that we did not filter our TF ChIP-Seq peak sets for the presence of a PWM match within each TF ChIP-Seq binding site. However, only TF binding sites harboring a binding motif can be identified by motif based methdos. Binding sites without motifs frequently correspond to indirect binding events indicating looping interactions which are frequently occuring at active enhancer. Earlier reports indicate that on average more than 50% of binding events measured by ChIP-Seq might be indirect as assessed by the absence of a corresponding TF motif (Thurman et al., 2012).

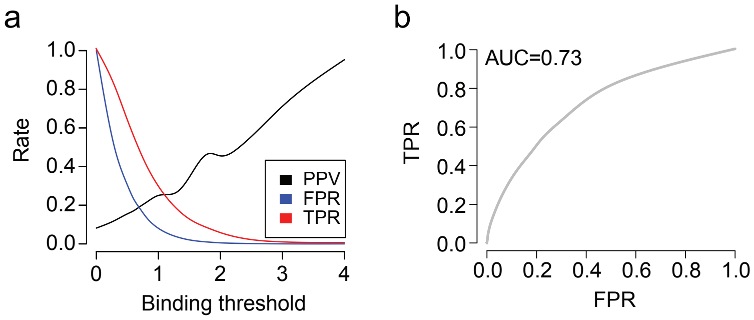


FIGURE 6.4: **a.** Median TPR (red), FPR (blue) and PPV (black) for n=46 TFs with matching motif for H3K27ac footprints (n=27,292) in K562 cells as a function of confidence in predicted binding ($-\log_{10}$ p-value). True positives were defined as predicted binding events overlapping with peaks determined by ChIP-Seq and false positives accordingly. The entire set of positives was defined as all TF ChIP-Seq peaks for a particular factor that overlapped with any H3K27ac footprint. **b.** ROC curve of the median TPR/FPR values from **a**.

It is very likely though, that the true PPV values at low to medium predicted binding strength threshold are significantly better than depicted in Figure 6.4. This hypothesis stems from the fact that TF binding is not a binary event but rather a continuum of different binding strengths (Biggin, 2011). However, stringent peak calling as done for example by the ENCODE project (Landt et al., 2012) not only binarizes the signal, but also tends to focus on the high to medium-high affinity binding sites. Therefore it is likely that this methodological problem causes a great inflation of the number of false positives and reduction of true positives in the low to intermediate binding strength regime, contributing to the low PPVs. However, we try to overcome this general drawback of a binarization of TF binding signal by using a continuous measure of predicted binding strength in our computation of the gene regulatory element to motif connectivity matrix and the estimation of TERA scores by taking advantage of the TRAP framework (Manke et al., 2008).

In addition to the reasonable performance of TF-binding inference, H3K27ac footprints are also more likely to be of functional importance: Genomic regions located under a dip in a H3K27ac region perform much better in transgenic reporter assays compared to the flanking regions (personal communication with T. S. Mikkelsen).

6.4.1.2 Evaluation of distinct normalization strategies for histone modification ChIP-Seq data

To identify the most suitable normalization strategy to obtain a robust, quantitative enrichment signal for ChIP-Seq data we evaluated 12 different approaches that test the impact of five distinct variables. More specifically, we evaluate the benefits of using no input control (raw), an input control specifically matching each dataset (WCE) as well as normalization to the average across the input controls of all cell types under investigation (meanWCE). In addition, we test whether it is advantageous to use a normalization strategy that does not only account for sequencing depth and differences in the background read distribution but also for differences in immunoprecipitation (IP) efficency. Finally, we also evaluate whether standard quantile normalization is beneficial to overcome differences in library quality. Since quantile normalization assumes comparable total enrichment levels, it is only reasonable to use this approach for open chromatin marks such as H3K4me3 or H3K27ac for which this is in general fullfilled.

In order to normalize for distinct IP efficiencies, we computed a sample specific normalization factor based on the number of non-duplicate reads located in regions identified as peaks compared to the number of reads present in the rest of the genome. The ratio of these two numbers can be interpreted as the signal to noise ratio, proportional to the IP efficiency. We then scaled the RPKM normalized values by this factor.

To evaluate the performance of each normalization approach, we use three distinct metrics:

1. Intra-group versus inter-group Pearson correlation: We compute the distribution of Pearson correlation coefficients between all replicate pairs (intra-group) after normalization as well as between all non-replicate pairs (inter-group) and determine the medians across the resulting distributions separately. The ratio of the two can be interpreted as a measure of the within-group variation eliminated over the between group variation preserved.

2. Intra-group versus inter-group euclidean distance: Similar to the approach listed above but using the Euclidean distance between samples as metric instead of the Pearson correlation.

3. Clustering of all samples based on the absolute pairwise Pearson correlation coefficients. Evaluation of the clustering was performed in a less rigorous fashion by determining whether biologically related samples would tend to cluster more together for each normalization strategy.

We applied each normalization strategy to a set of 70 distinct cell types from the REMC project, of which 38 had replicates. Prior to normalization, we determined the number of non-duplicate reads in regions identified as peaks using the IDR peak calling framework (Li et al., 2011) and normalized the counts in each region to reads per kilobase per million reads sequenced (RPKM) to account for differences in sequencing depth. Furthermore, we log_2 transformed the all RPKM counts after adding a pseudocount of one to each region.

The results for the different normalization schemes are summarized in Table 6.1, suggesting that normalization to the mean across all WCEs in combination with quantile normalization performs best. Evaluation of the clustering results confirmed this (Appendix Fig. A.1).

normStrategy	WGP	BGP	WGE	BGE	P-ratio	E-ratio
log_2 RPKM/mWCE quantile	0.88	0.33	183.72	424.75	2.63	2.31
soph log_2 RPKM/mWCE quantile	0.84	0.32	311.22	628.26	2.57	2.02
log_2 RPKM/mWCE	0.86	0.34	189.02	438.69	2.53	2.32
log_2 RPKM quantile	0.88	0.35	138.63	321.25	2.52	2.32
soph log_2 RPKM/mWCE	0.82	0.33	441.54	759.94	2.51	1.72
soph log_2 RPKM quantile	0.85	0.34	261.05	542.96	2.48	2.08
log_2 RPKM/WCE quantile	0.87	0.36	189.92	426.31	2.45	2.24
log_2 RPKM	0.83	0.34	363.28	638.49	2.41	1.76
soph log_2 RPKM	0.83	0.34	363.28	638.49	2.41	1.76
soph log_2 RPKM/WCE quantile	0.82	0.35	334.30	626.57	2.36	1.87
log_2 RPKM/WCE	0.81	0.34	465.70	761.73	2.35	1.64
soph log_2 RPKM/WCE	0.81	0.34	465.70	761.73	2.35	1.64

Table 6.1: Results for various normalization methods described in the text. WGP: average within-group Pearson correlation; BGP: average between-group Pearson correlation; WGE: average within-group Euclidean distance; P-ratio: ratio of within- to between-group Pearson correlation; P-ratio: ratio of within- to between-group Euclidean distance.

Finally, we also tested the presence of batch effects since our collection of ChIP-Seq datasets originates from two different labs. To that end we performed principal component analysis on the entire normalized dataset and computed the correlation and significance of association of all principal components with various surrogate variables such as lab of origin and project (roughly corresponding to time). While this analysis yielded significant association of several principal components with the the surrogate variable *lab of origin*, removal of distinct principle components lead to the close clustering of not-related cell types (data not shown). However, the original, normalized dataset showed excellent clustering performance, including correct clustering of biologically related cell types that were obtained from different labs. This observation in combination with the fact that several sub cohorts of samples generated in different labs are closely related, leads us to the conclusion that the surrogate variables are confounded to some extend and that the high variance components are not driven by batch effects.

6.4.1.3 Evaluation of inferred TF epigenetic remodeling activities

To test the validity of the idea to infer an epigenetic remodeling activity for each motif/TF, we inferred TERA scores for 557 distinct motifs across our 70 cell types REMC reference set for H3K27ac using all footprints identified across that dataset as input. We normalized the H3K27ac enrichment levels according to the optimal strategy outlined previously and then averaged across replicates. The resulting values were then used for PLS analysis, using our standard TERA pipeline (Fig. 6.3). If the TERA approach is valid, we expect a reasonable correlation of the TERA scores with the expression of corresponding TFs. However, we do certainly not expect a perfect linear correlation since it is unlikely that the total expression level of a TF is linearly related to its epigenetic remodeling activity as well as its overall binding profile. It is more likely that the TF gene expression shows a moderate correlation to TERA scores of corresponding motifs, following the idea that epigenetic remodeling events that can most likely be explained by the presence of a TF-motif in general requires that factor to be expressed at a minimal level. However, since it is unclear what a cutoff for a binary classification of expressed and not-expressed TFs in a particular cell type might be (in addition to probably being TF-specific), we choose the standard Pearson correlation coefficient to measure the association between TERA and TF gene expresssion.

To determine the correlation between these two measruements, we took advantage of RNA-Seq data from the Epigenome Roadmap Project (REMC) (Bernstein et al., 2010) for 39 distinct cell and tissue types for which also H3K27ac data was available (see EpigeneticDynamics folder for full RPKM normalized gene expression table in electronic archive accompanying this work). With this data at hand, we computed the Pearson correlation coefficient of each TF with all motifs that could be mapped to this factor and selected the motif with the highst absolute correlation coefficient. The results show moderate to high correlation for the vast majority of motifs (Fig. 6.5a), indicating that the TERA approach is valid.

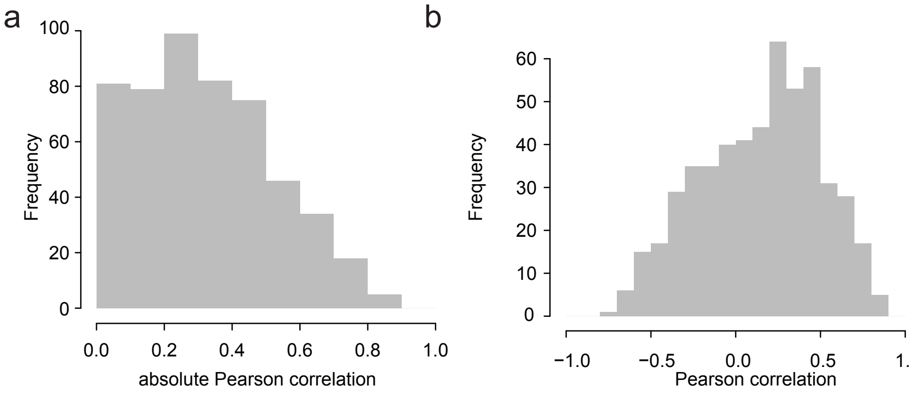


FIGURE 6.5: **a.** Distribution of absolute Pearson correlation coefficients for H3K27ac based TERAs and TF gene expression levels across 39 cell types from the REMC project. For each motif, we chose the highest Pearson correlation coefficient across all TFs mapping to the respective motif. **b.** Distribution of Pearson correlation coefficients for H3K27ac based TERAs and TF gene expression levels across 39 cell types from the REMC project. For each motif, we chose the highest Pearson correlation coefficient across all TFs mapping to the respective motif.

A large set of motifs shows anti-correlation with H3K27ac based TERA scores, suggesting a role in enhancer decommissioning or repression (Fig. 6.5b). Finally, several motifs show no association with any of the their corresponding TFs. This can be explained by multiple causes: 1. The motifs are of low quality; 2. the mapping to their corresponding TF genes is inaccurate; 3. These motifs and the corresponding TFs have no role in H3K27ac remodeling and therefore their inferred TERA levels are not informative. 4. The corresponding TF genes change their expression very little 5. The corresponding TFs' epigenetic remodeling activities are disconnected from their expression level as long as a minimal level of expression is maintained. Since the class of motifs to which this applies is relatively small and indeed encompasses many constitutively expressed factors, we leave a more detailed investigation of this point open for future studies.

In addition to the presented *in silico* validation, we are also carrying out a large scale shRNA knockdown screen against 250 TFs expressed across our time course at each stage of differentiation. The goal of this screen is to validate the ranking and identification of key TFs based on their TERA scores and co-binding patterns. However, the screen is still pending and will not be part of this work but is rather going to be part of the final journal publication.

6.4.2 Consecutive stages of in vitro neural specification are characterized by distinct transcriptional states and differentiation potential

While great progress has been made in the directed differentiation of neural precursors cells (NPCs) (Elkabetz et al., 2008) as well as regionally specified neuronal subtypes from human ES cells (Kirkeby et al., 2012, Maroof et al., 2013) the origin and relation of distinct NPC types is less well understood. In particular, the molecular mechanism underlying changes in NPC differentiation potential over time and the resulting capacity to differentiate into distinct neural subtypes remain elusive (Elkabetz and Studer, 2008, Temple, 2001). Using clearly defined stages along the neural lineage we created high-quality reference epigenome maps as part of the NIH Roadmap Epigenome Project (Bernstein et al., 2010).

The human ES cell line WA9 (or H9) was differentiated into neuroectoderm (NE), followed by the transition to early radial glial (E-RG) cells and then further into three additional distinct NPC populations (mid radial glial (M-RG), late radial glial (L-RG) and long term neural progenitor (LNP); Fig. 6.6a). The first three differentiated populations (NE, E-RG and M-RG) as well as the undifferentiated hESCs were subjected to strand-specific RNA-Seq, chromatin immunoprecipitation followed by sequencing (ChIP Seq) for H3K4me1 and me3, H3K27ac, and H3K27me3 as well as DNAme by whole-genome bisulfite sequencing (WGBS) while the last two stages (L-RG and LNP) were profiled by RRBS.

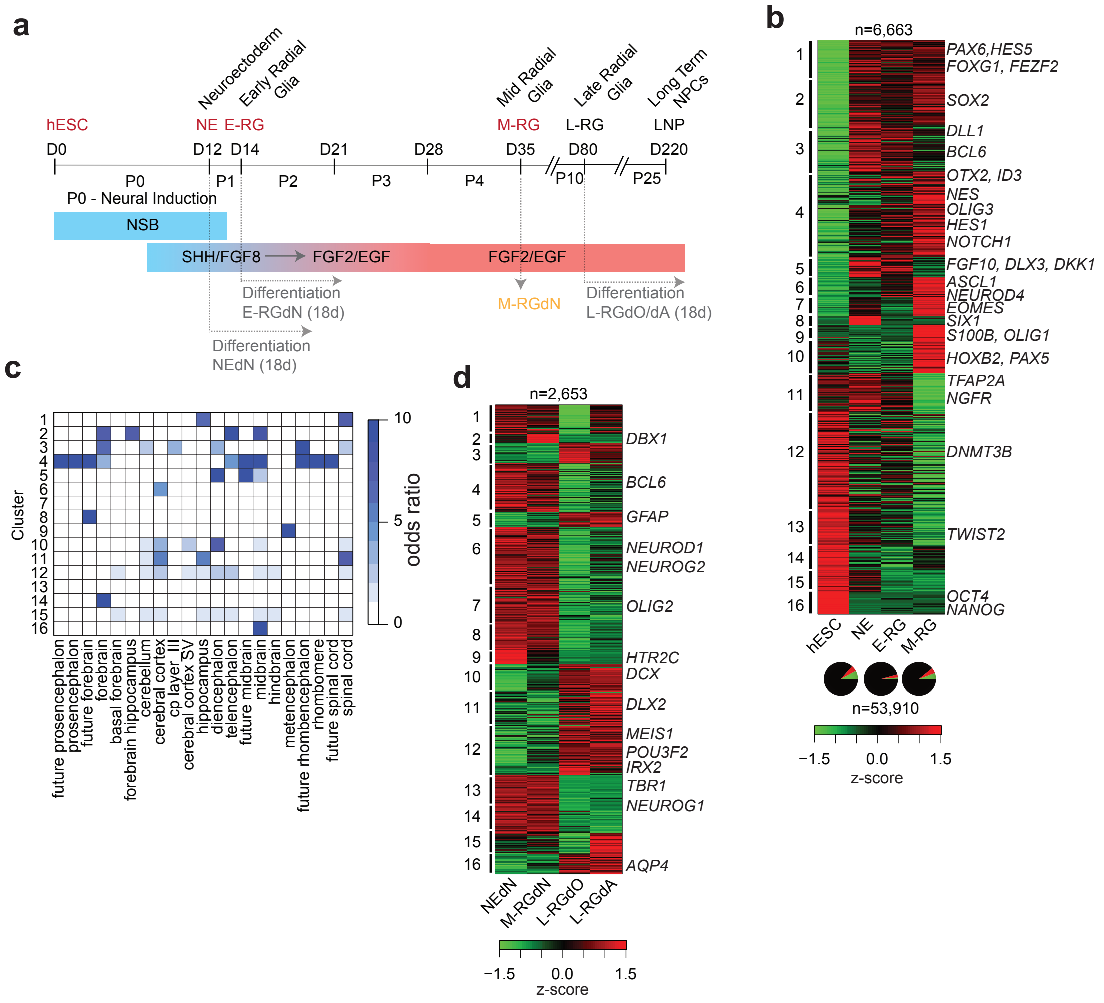


FIGURE 6.6: **a.** Schematic of our differentiation model including the times of sample collection. Human ESC were differentiated into neuroectoderm (NE) using dual inhibition of TGFb and BMP followed by the transition to neural base media. Subsequently, sonic hedgehog and FGF8, are used to transition to the early radial glial stage (E-RG). For the rest of the differentiation experiment the cells were constantly maintained in FGF2 and EGF2 neural base media to reach the mid radial glia (M-RG) stage after 35 days, the late radial glia (L-RG) stage after 80 and the long term neural progenitor (LNP) stage after 200 days of in vitro culture. Cell type names indicated in red were profiled for gene expression, histone modifications as well as DNAme by WGBS, while names shown in grey for gene expression only and names in black for DNAme by RRBS only. Cell type names in orange are just indicated as intermediates and were not profiled here. **b.** Gene expression patterns shown as z-scores for all differentially expressed genes (fold change $\geq$ 1.5) across hESCs and three neural precursor differentiation stages. Genes were grouped into 16 clusters using k-means clustering and Jensen-Shannon- based metric. Pie charts below indicate fraction of up (red) and down-regulated (green) genes during each transition out of all transcripts assayed (53,490). **c.** Gene set enrichment analysis for each cluster using MGI neural expression based genes sets. Only significantly enriched results are shown (at least $p \leq 0.05$ and odds ratio ≥ 2 (fisher's exact test)). **d.** Gene expression patterns shown as z-scores for all significantly differentially expressed genes (fold change $\geq$ 2) across four terminally differentiated cell types: NEdN- neurons derived from the NE stage, E-RGdN- neurons derived from the E-RG stage, L-RGdO/dA- oligodendrocytes and astrocytes derived from the L-RG stage. Genes were grouped into 16 clusters using k-means clustering and Jensen-Shannon-based metric.

We began with differential gene expression analysis to determine whether our differentiation system is capable of recapitulating expression changes observed during *in vivo* embryonic development and neurogenesis. To that end we investigated the dynamics of several well known marker genes for different stages of neural development and performed various types of gene set enrichment analysis to show activation of distinct neural expression programs during *in vitro* differentiation. We found 6,663 differentially expressed genes across the entire time course that can be grouped into 16 separate clusters (Fig. 6.6b). Genes that are upregulated upon exiting the pluripotent state and remain active in all subsequent stages include key neural TFs such as PAX6, HES5 and FOXG1 (Fig. 6.6b; cluster 1). Pluripotency associated genes such as OCT4 and NANOG are, as expected, rapidly downregulated (cluster 16), while SOX2 progressively increases in expression towards the M-RG stage (cluster 2). Clusters 1, 2 and 4 represent pan-neural genes, clusters 8 and 11 represent early NE fates (SIX1, TFAP2A, NGFR), cluster 5 and 6 correspond to the E-RG stage (FGF10, DKK1, LRP2, FOSB) and clusters 9 and 10 show the gliogenic bias that is already present at the M-RG stage (S100B, OLIG1). Gene set enrichment analysis using the MGI database (Finger et al., 2011) revealed a progressive change of NPC identity from more anterior regions (forebrain) to posterior (mid and hindbrain) (Fig. 1c) and a progression towards later developmental stages (Fig. 6.7a). To expand on this analysis we also profiled terminally differentiated populations derived from three distinct stages (Fig. 6.6a,d). We find 2,653 genes differentially expressed across the four subtypes, illustrating the distinct identity of the neuronal as well as glial populations (Fig. 6.7b,c) arising from each stage. While neuronal and glial cell types exhibit the most dramatic transcriptional differences including key markers such as NEUROD1 and AQP4, neurons derived from NE (NEdN) express the serotonin receptor HTR2C in contrast to neurons generated from E-RG (E-RGdN), which express the neuronal TF DBX1 (Fig. 6.6d). Notably, NE and E-RG can give rise to neurons (NEdN, E-RGdN) but are not yet capable of giving rise to glial cell types, while only the L-RG stage (and later NPCs) demonstrate the potential to generate both neuronal as well as glial cells (L-RGdA/dO, Fig. 6.6d). This is also compatible with the *in vivo* shift from neurogenic cell fate potential in neuroepithelial cells as well as early and mid-gestation radial glial cells during neurogenesis, towards gliogenic cell fate bias in late gestation radial glial cells (Ramon y Cajal, 1995). Taken together, these observations suggest a progressive change of neural precursor identity reminiscent of the sequence of neural cell populations and processes arising during *in vivo* brain development.

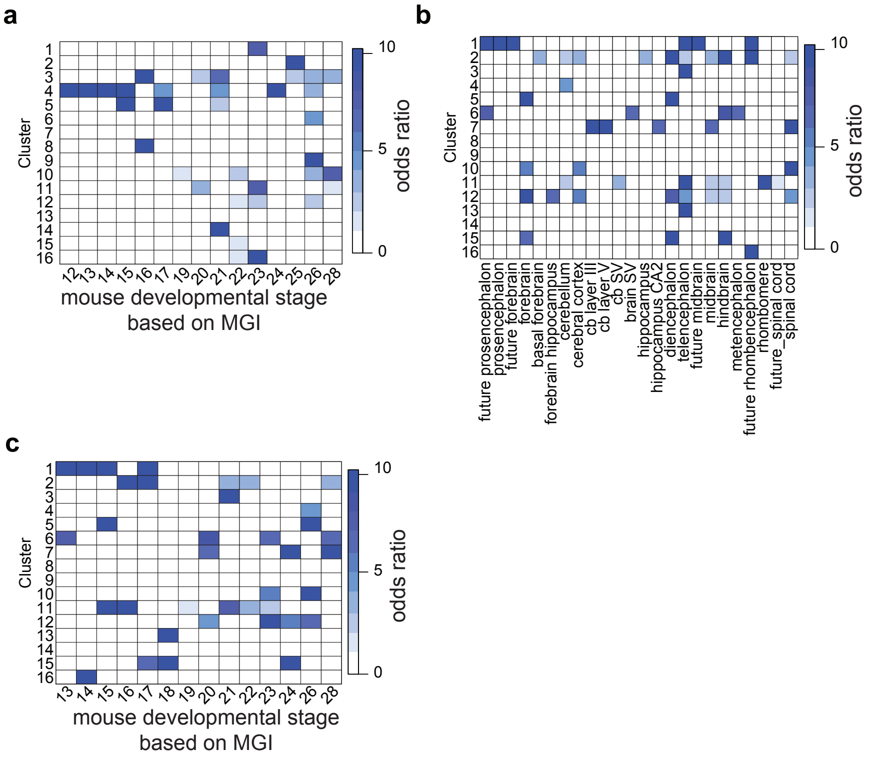


FIGURE 6.7: **a.** Enrichment analysis for RNA Seq based expression clusters for differentially expressed genes across hESC, NE, E-RG and M-RG determined by k-means clustering and Jensen-Shannon divergence based metric (Fig. 6.6b). Enrichment was determined for genes with mouse orthologs using the MGI expression database and neural related gene categories. For each murine developmental stage, we plotted the odds ratio of the highest significantly enriched gene set (p-value ≤ 0.01 and odds ratio ≥ 1.5 Fisher's exact test). **b.** Enrichment analysis for RNA Seq based expression clusters for differentially expressed genes across NEdN, E-RGdN, L-RGdO/dA determined by k-means clustering and Jensen-Shannon divergence based metric (Fig. 6.6d). Enrichment was determined for genes with mouse orthologs using the MGI expression database and selected neural structure related gene categories. For each structure, we plotted the odds ratio of the highest significantly enriched gene set across all developmental time points (p-value ≤ 0.01 and odds ratio ≥ 1.5 Fisher's exact test). **c.** Enrichment analysis for RNA+Seq based expression clusters for differentially expressed genes across NEdN, E-RGdN, L-RGdO/dA determined by k-means clustering and Jensen-Shannon divergence based metric (Fig. 1d). Enrichment was determined for genes with mouse orthologs using the MGI expression database and selected neural structure related gene categories. For each murine developmental stage, we plotted the odds ratio of the highest significantly enriched gene set (p-value ≤ 0.01 and odds ratio ≤ 1.5 fisher's exact test).

6.4.3 Developmental progression along the neural trajectory is accompanied by widespread epigenetic remodeling

In order to gain more insights into the underlying molecular mechanism we took advantage of our novel computational framework to infer TF epigenetic remodeling activities in combination with epigenetic footprinting. To that end, we exploited the substructure of histone tracks for H3K4me3 and H3K27ac and identified chromatin-based footprints, small dips between peaks demarcating a nucleosome-depleted region (NDR) more amenable to protein binding (Fig. 6.8a, top). In addition, we also included DNAme-based footprints that are known to frequently coincide with TF binding sites (TFBS) (Ziller et al., 2013). Next, we combined this set of typically 80-200 bp long stretches of the genome with a quantitative model of TF binding potential based on published motifs (Manke et al., 2008, 2010) to infer possible condition-specific TF binding (Fig. 6.8a, bottom).

We applied this strategy to our time course data and identified 408,416 distinct footprints (FPs).

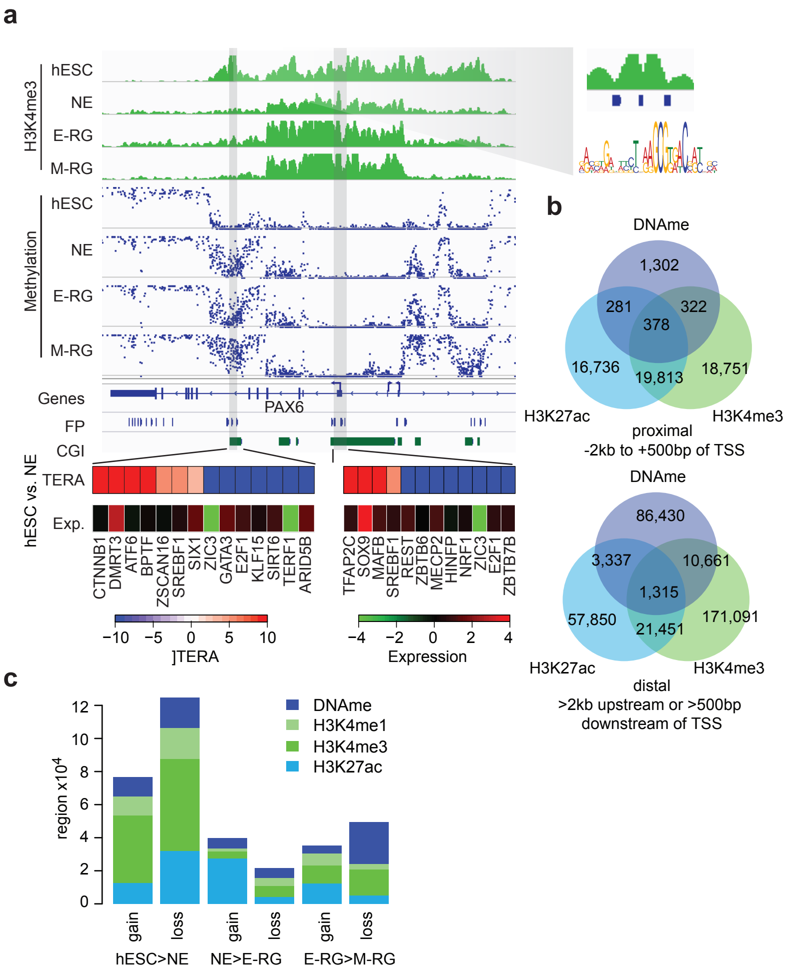


FIGURE 6.8: **a.** Illustration of genomic footprinting. Grey boxes highlight example footprints (FP) determined based on H3K4me3 methylation patterns. H3K4me3 methylation patterns across the PAX6 locus (chr11:31,809 kb - 31,852 kb) for four stages are shown on a scale from 0 to 1 and normalized to 1 million reads. Below, methylation levels for individual CpGs are shown in blue on a scale from 0 to 100% methylation (y-axis). The TERA labeled boxes show the inferred difference in epigenetic remodeling activity (TERA) for each TF between the hESC and NE stage for motifs found at this location. Below, the expression change of the corresponding TFs is shown. The magnification on the right shows a high-resolution view of the H3K4me3 peak structure and the corresponding footprints identified as blue boxes underneath as well as an example motif found at this location. **b.** Venn diagrams of the overlap of DNAme, H3K27ac and H3K4me3 based footprints for promoters (top) and distal regions (bottom). **c.** Barplot of the frequency and associated mark of epigenetic changes at footprints for all cell state transitions broken up into gain and loss.

To investigate whether it might be sufficient to solely focus on one epigenetic modification instead of three, we first computed the overlap of all identified footprints at proximal and distal sites. H3K4me3 and H3K27ac based FPs show over 50% overlap in promoter regions, but only about 12% for distal regions (Fig. 6.8b). Next, we sought to better understand whether the same TFs are identified as key regulators based on distinct epigenetic marks. To that end we investigated the fraction of the top 20 differentially expressed TFs between consecutive stages (n=57) that could be recovered based on the top 20 motifs ranked by differences in their TERA scores between the corresponding consecutive stages. This analysis revealed excellent recovery of distinct TFs (Fig. 6.9a) and indicated that distinct epigenetic marks indeed provide complementary information on key regulatory TFs as approximated by differential expression ranking (Fig. 6.9b).

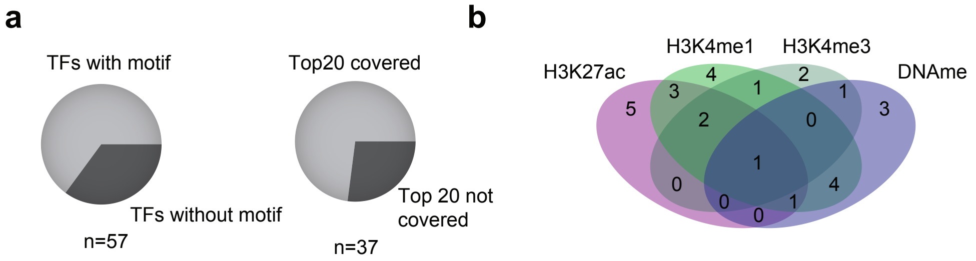


FIGURE 6.9: **a.**Left: Fraction of top 20 differentially expressed TFs in the transition from hESC to NE, NE to E-RG and E-RG to M-RG that have at least one reported motif. Right: Fraction of top 20 differentially expressed TFs and a reported binding motif that are recovered in the top 20 motifs predicted to be key regulators of the cell state transitions hESC to NE, NE to E-RG and E-RG to M-RG based on H3K27ac, H3K4me3, H3K4me1 and DNAme remodeling at distal footprints. **b.** Contribution of each epigenetic mark to the recovery of the top 20.

Interestingly, H3K4me3 footprints are enriched for schizophrenia and bipolar disorder related GWAS SNPs ($p < 0.05$, empirical test see Section 6.3.11 for details) and H3K27ac footprints for AD, t-tau protein and glioma associated SNPs ($p < 0.01$ empirical test). Among them is the prominent AD SNP rs157580 located upstream of the APOE gene (Fig. 6.9m). Using a biophysical model of TF binding (Manke et al., 2010), we identified putative binding sites for MEIS1 and beta-catenin as predicted to have a higher (2.2 and 1.2 fold) binding affinity for the minor allele. This observation suggests a possible mechanism for altered APOE expression levels through altered TF binding affinity within a putative enhancer region and demonstrates the power of high-resolution epigenomic footprinting. In particular, our analysis highlights an approach to identify key regulators and dissect their potential roles during cell state transitions.

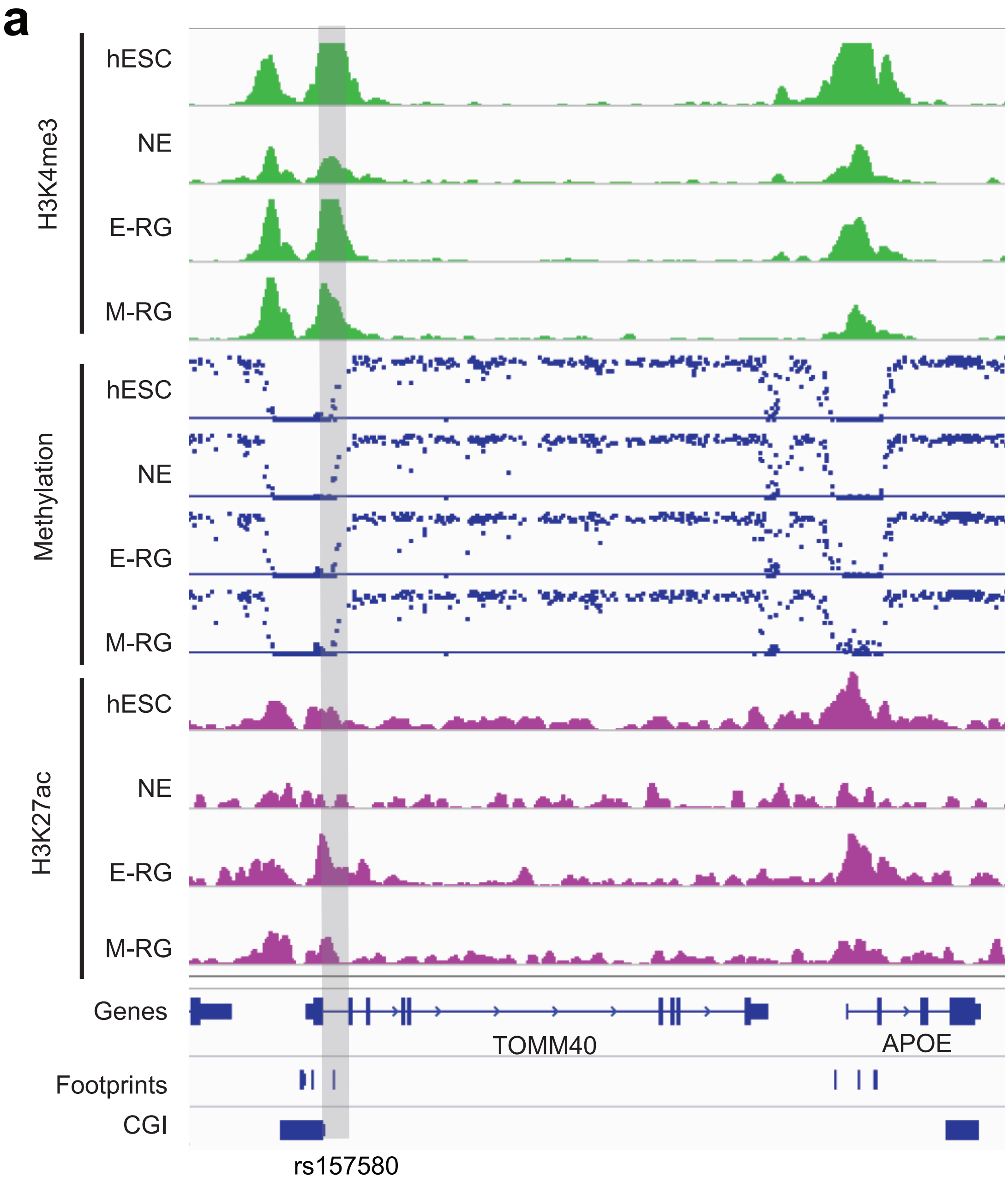
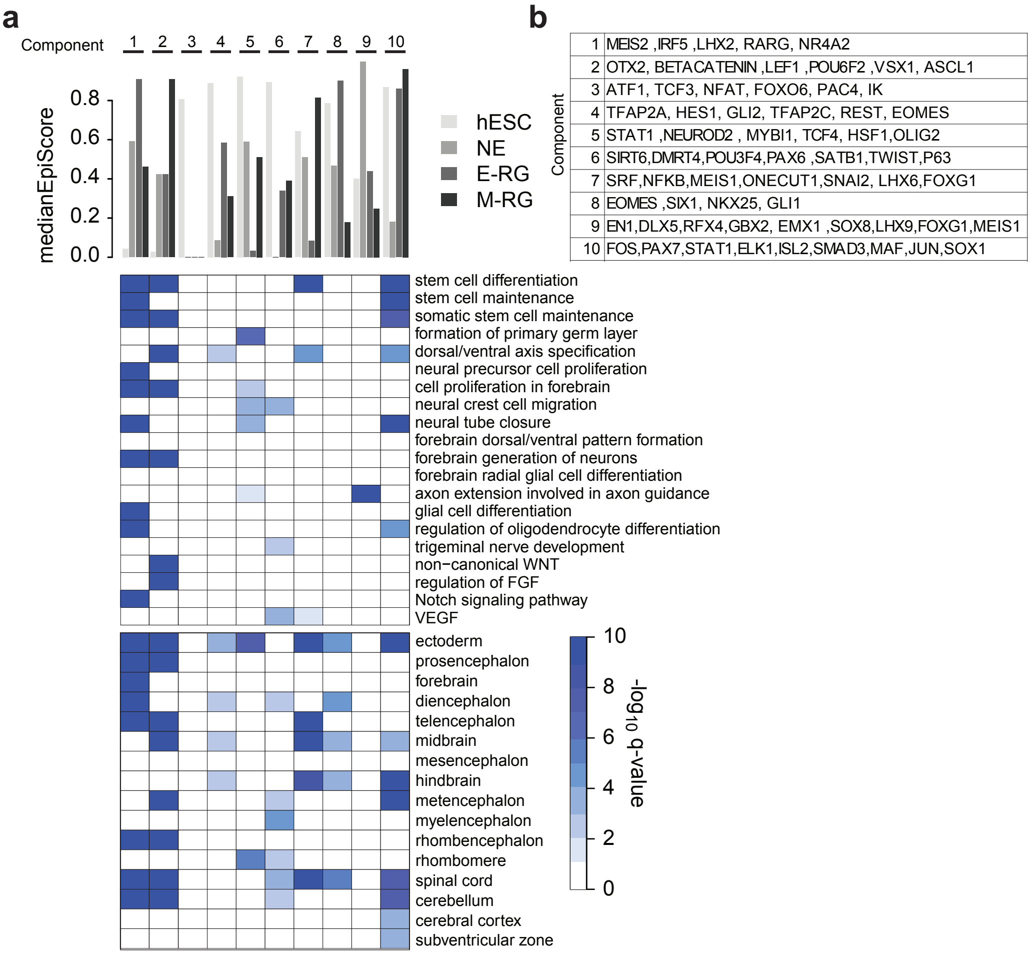


Figure 6.10: a. Epigenetic dynamics across the APOE locus (chr19:45,391 kb - 45,414 kb) for hESC and 3 stages of NSCs. H3K4me3 read counts normalized to 1 million reads are shown on a scale of 0 to 2 (green). DNAme levels for single CpGs are indicated as blue dots on a scale of 0 to 100% of methylation (y-axis). H3K27ac read counts normalized to 1 million reads are shown on a scale of 0 to 1 (purple). For reference footprints (FP) and CpG islands (CGIs) are indicated as blue boxes (bottom). Shaded gray box indicates the position of the putative enhancer element overlapping with the Alzheimer-related SNP rs157580.

6.4.4 TF modules drive stage specific epigenetic transitions

In order to dissect the epigenetic dynamics in more systematic and high resolution fashion, we identified ten distinct clusters of dynamic footprints based on H3K4me3, each associated with a specific subset of putative regulating TFs (Fig. 6.11a). Gene set enrichment analysis using the GREAT toolbox (McLean et al., 2010) highlighted two key neural clusters that exhibit widespread gain of H3K4me3 at the NE stage, are depleted of active marks in hESCs and remain partially active during the entire differentiation process (Fig. 6.11a, top 1,2). While both clusters are highly enriched for GREs involved in stem cell differentiation and somatic stem cell maintenance, cluster 1 is accordingly highly enriched for telencephalic fates. In contrast, cluster 2, while also active at the NE and E-RG stages, reaches peak activity at the M-RG stage and shows strong enrichment for telencephalon, midbrain and metencephalon related GREs. Interestingly, the former are also enriched for NOTCH targets, compatible with early NOTCH activation in the corresponding NE and E-RG stages in vivo. Consistent with their presumptive role, these clusters are regulated by *OTX2, POU5F2, LHX2* and *NR4A2* - key developmental genes widely expressed in telencephalic to metencephalic germinal zones. Furthermore, downstream targets of the WNT pathway are prominent (Fig. 6.11b), consistent with the key roles of this pathway at different stages of neural development (Ille and Sommer, 2005).



Component	
1	MEIS2 ,IRF5 ,LHX2, RARG, NR4A2
2	OTX2, BETACATENIN ,LEF1 ,POU6F2 ,VSX1, ASCL1
3	ATF1, TCF3, NFAT, FOXO6, PAC4, IK
4	TFAP2A, HES1, GLI2, TFAP2C, REST, EOMES
5	STAT1 ,NEUROD2 , MYBI1, TCF4, HSF1,OLIG2
6	SIRT6,DMRT4,POU3F4,PAX6 ,SATB1,TWIST,P63
7	SRF,NFKB,MEIS1,ONECUT1,SNAI2, LHX6,FOXG1
8	EOMES ,SIX1, NKX25, GLI1
9	EN1,DLX5,RFX4,GBX2, EMX1 ,SOX8,LHX9,FOXG1,MEIS1
10	FOS,PAX7,STAT1,ELK1,ISL2,SMAD3,MAF,JUN,SOX1

FIGURE 6.11: **a.** Top: Decomposition of H3K4me3 dynamics at distal footprints into 10 distinct components. Median RPKM level in each of the four cell states is shown as a bar in grey shading for each of the 10 components. Bottom: Gene set enrichment analysis for regions associated with each component. **b.** TFs associated with each component based on TF motif presence and expression, see Section 6.3.8 for details on motif-TF association.

In order to pinpoint upstream regulators potentially driving the differentiation related epigenetic remodeling events, we determined the top 15 TFs associated with H3K27ac (Fig. 6.12a), and H3K4me3 gain as well as loss of DNAme for each cell state transition (Extended Data Fig. 6.13a,b). Among the top TFs during the transition from hESCs to NE are various SOX motifs, e.g. SOX9/1 as well as MEIS3 known to play an important role in the spatial specification of the early neural plate (Elkouby et al., 2010, Ng et al., 1997, Pevny et al., 1998). In addition, we observe various proliferation related-genes with high H3K27ac remodeling activity, including MYC and MAF, of which the latter directly interacts with SOX9 (Huang et al., 2002). Interestingly, we also find differential activity of distinct downstream components of signaling pathways such as LEF1, TCF3 and betacatenin (WNT signaling) as well as ATF proteins downstream of FGF signaling (Fig. 6.12a, Fig. 6.13c), all integral parts of neural induction and neural development.

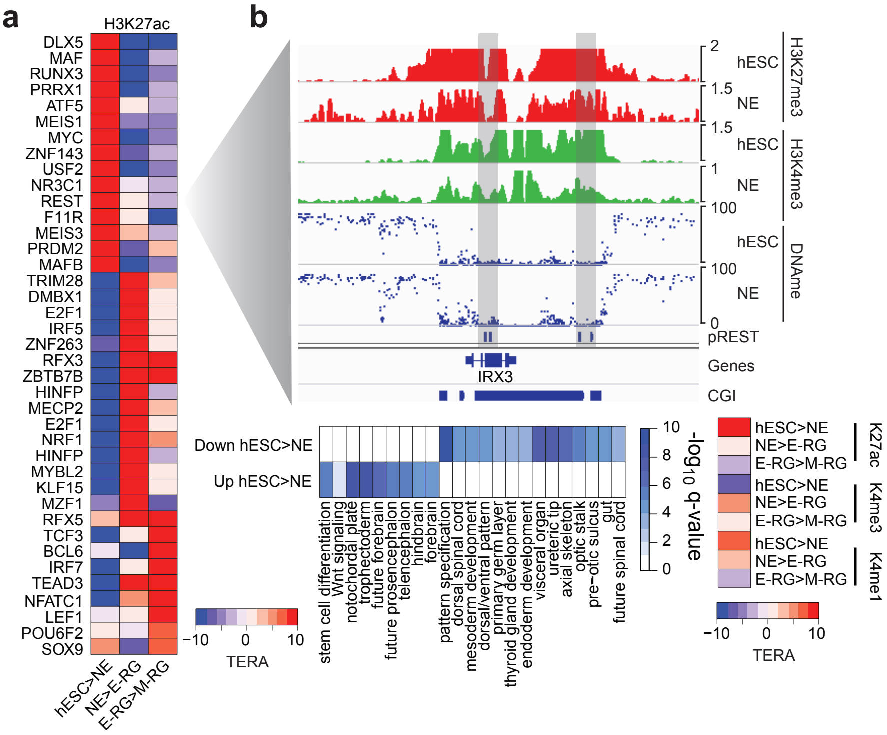


FIGURE 6.12: **a.** Inferred TERA changes based on distal H3K27ac footprints for three cell state transitions. **b.** Top: Epigenetic dynamics at the IRX3 locus (chr16:54,300kb - 54,340kb) during the hESC to NE transition. Chromatin enrichment levels are normalized to 1 million reads and indicated on the right. Methylation levels are shown for individual CpGs (blue dots) on a scale from 0 to 100% of methylation (y-axis). Grey boxes highlight REST binding sites. CpG islands (CGI) are indicated in blue on the bottom. Bottom: Selected gene set enrichment categories for REST motifs losing (upper part) and gaining (lower part) H3K4me3 during the transition to NE (left). REST motif activity changes inferred based on H3K27ac, H3K4me3 and DNAme for all three cell state transitions are shown on the right.

Furthermore, SMAD3 binding sites exhibit a decrease in H3K4m3 activity during the transition from hESC to NE consistent with inhibition of BMP and TGFb signaling promoting NE specification (Chambers and Tomlinson, 2009). In contrast, GLI binding sites show an increase in H3K4me3 as well as loss of DNAme during the transition from NE to E-RG, in line with the onset of SHH signaling (Fig. 6.12a, Fig. 6.13c). These observations highlight the possibility to infer the stimulation of specific signaling pathways solely based on epigenetic data as the temporal activity pattern of these pathways during the differentiation time course mimics our data-driven observations (Fig. 6.13c).

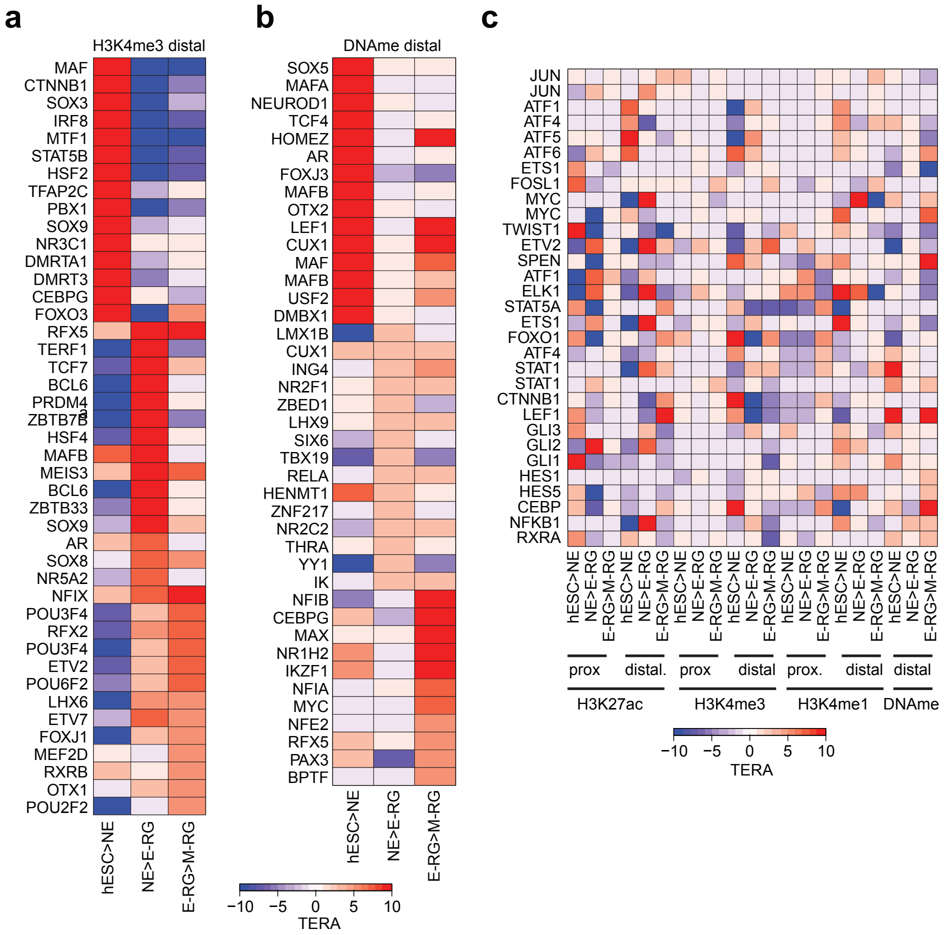


FIGURE 6.13: **a.** Inferred TF activity changes based on distal H3K4me3 footprints for three cell state transitions. **b.** Inferred TF activity changes based on distal DNAme footprints for 3 cell state transitions. **c.** ERA dynamics for key signaling factors across 3 cell state transitions. TERA changes based on four distinct epigenetic marks (H3K27ac, H3K4me3, H3K4me1, DNAme) and two genomic contexts (proximal and distal) are shown.

In addition to stage specific regulators, we also identified a core network of TFs that is up-regulated at the onset of neural induction (hESC to NE) and remains active in all stages of NPC progression. This network is comprised of *EMX2, FOXG1, FOS, OTX2, PAX3, SIX3, DMRT3, ARX, PAX6* and *LHX2*, some of which remain active in a subset of the terminally differentiated populations (Fig. 6.14a). Interestingly, these core factors show only moderate epigenetic dynamics individually (Fig. 6.14b,c) but exhibit widespread, stage-specific co-binding relationships (Fig. 6.14a,b). These observations suggest a model where NPC core factors dynamically rewire a subset of their binding sites in cooperation with distinct stage specific TFs to facilitate specific epigenetic changes modulating the differentiation propensity.

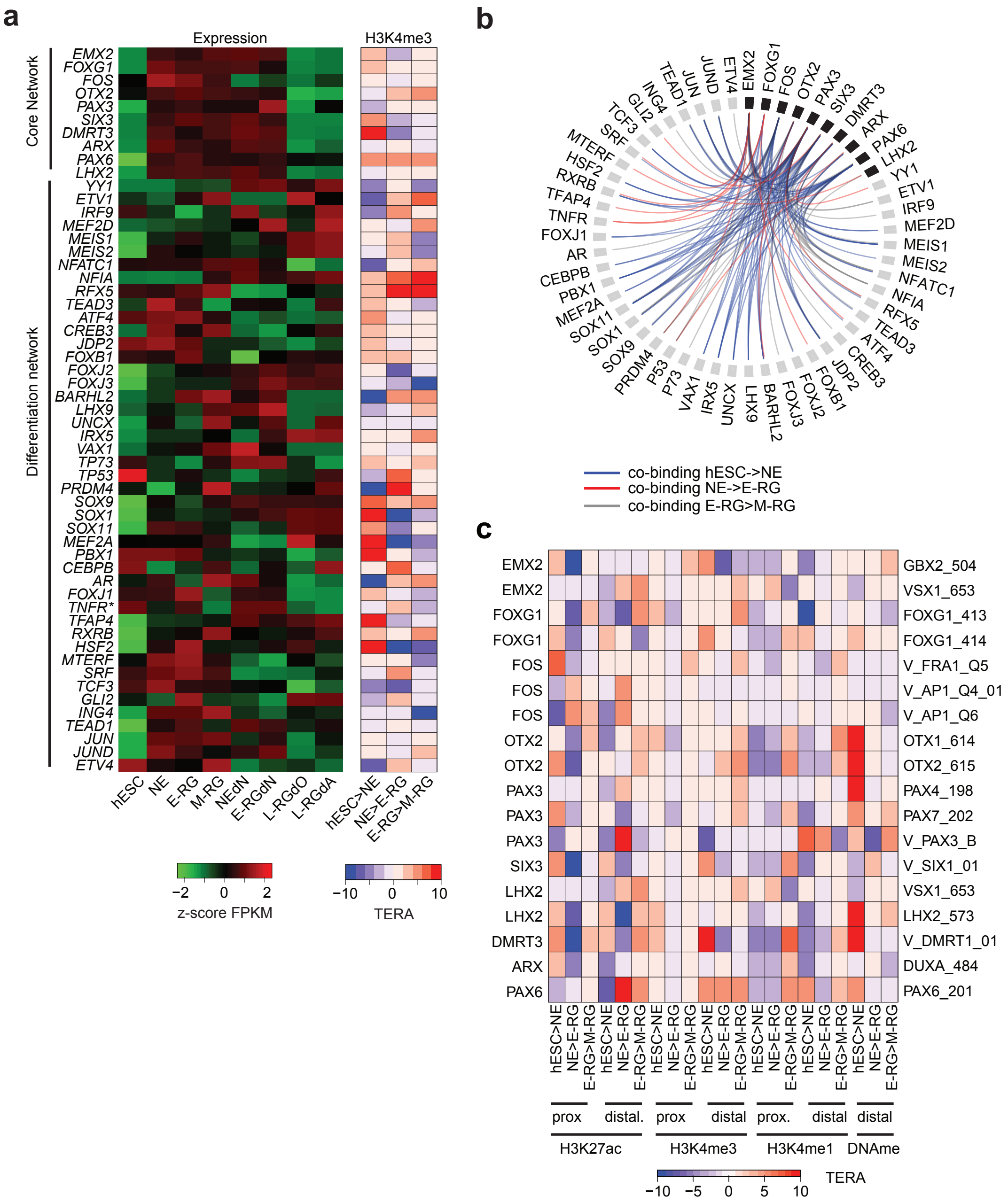


FIGURE 6.14: **a.** Gene expression levels for the NPC core network of TFs induced at the NE stage and maintaining high expression levels up to the INP stage are shown as z-scores. In addition, expression patterns for TFs exhibiting significant motif co-localization with the core factors at footprints gaining H3K4me3 during the transitions to NE, E-RG and M-RG (left). Inferred TERA based on H3K4me3 patterns are shown on the right. **b.** Motif co-occurrence relationships for core factors (black boxes) and TFs with significantly co-occuring motifs at footprints gaining H3K4me3 at the NE (blue), E-RG (red) and M-RG (grey) stage. **c.** TERA dynamics for all motifs associated TFs upregulated during neural induction and maintained expression at least up to the INP stage (core network). Name of core factor is indicated on the left while the motif identifier is shown on the right. TERA dynamics are displayed for each cell state transition and four distinct epigenetic marks (H3K27ac, H3K4me3, H3K4me1, DNAme) and two genomic contexts (proximal and distal).

6.4.5 Epigenome based inference identifies priming factors

Lastly, we repeatedly noticed an increase in inferred TERA for factors not yet expressed at the corresponding time point. To investigate the possibility and relevance of TFBS priming prior to factor activation, we identified TFs in each NPC population with increased TERA for H3K4me1 and DNAme but with low expression levels and induced at the next stage of differentiation (Fig. 6.15a). Surprisingly, more than half of the corresponding TF motifs were significantly enriched ($p \leq 0.05$ Fisher's exact and permutation test; Fig. 6.15b; green) in GREs associated with genes changing expression at the next stage of differentiation, strongly supporting the notion that the epigenetic state can predict the downstream activation of TFs. The top set of primed binding sites is associated with the Achaete-scute homolog TF ASCL1, a key pro-neural factor that is regulated by the NOTCH pathway (Bertrand et al., 2002, Guillemot et al., 1993) and the core factor for successful transdifferentiation of fibroblasts to neurons (Vierbuchen et al., 2010). Binding sites for this TF lose DNAme and gain H3K4me1 at the NE stage and gain H3K27ac and H3K4me3 during the transition to E-RG. Moreover, these putative GREs are significantly enriched at genes that become induced in neurons specifically derived from the NE stage (NEdN) as well as the E-RG stage (E-RGdN), concomitant with increased expression of ASCL1. In line with these observations, we find highly significant overlap of primed Ascl1 motifs with Ascl1 binding site in human NHEK and NHDF cells (Wapinski et al., 2013) ($p = 3.9 \times 10^{-5}$, odds ratio$= 2.91$ Fisher's exact test, Fig. 6.16).

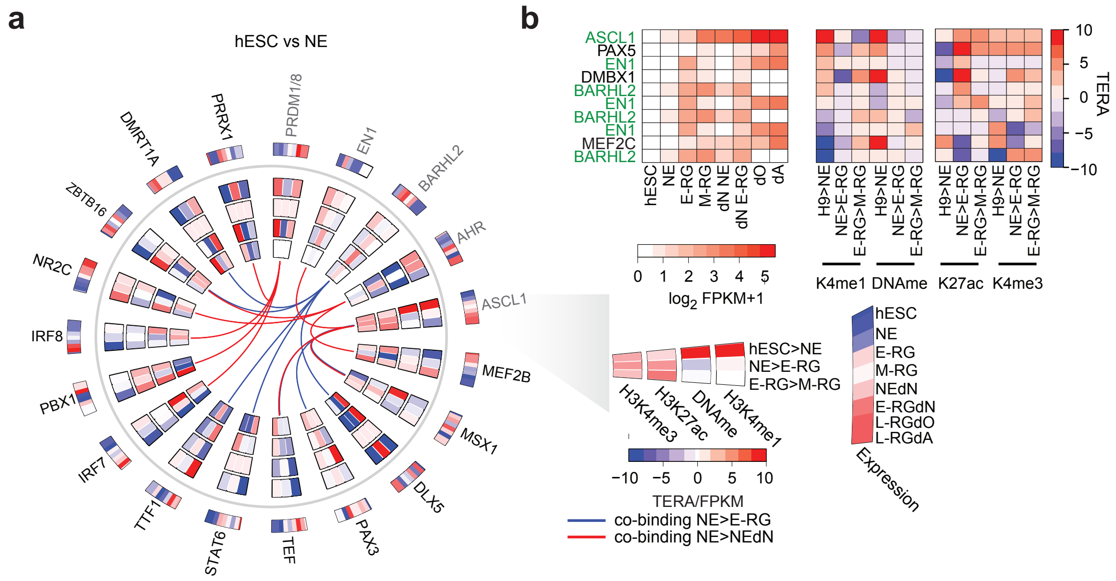


FIGURE 6.15: **a.** Characterization of TFs associated with motifs gaining TERA at the NE stage (grey factors) based on H3K4me1 and/or DNAme, not significantly upregulated/expressed at the hESC or NE stage, induced at the E-RG stage and/or in NEdN and significantly ($p \leq 0.05$) enriched at gene regulatory elements associated with genes changing expression (FC$\geq$ 1) at the E-RG stage and/or in NEdN. The latter TFs are defined as primed TFs. TFs shown in black exhibit significant co-occurrence of TF associated motifs at footprints gaining H3K4me1 and/or losing DNAme at the NE stage, are expressed (FPKM$\geq$ 3) at the NE or hESC stage and are associated with genes gaining expression (FC$\geq$ 1) at the E-RG stage (blue lines) and/or in NEdNs (red lines). Outermost heatmaps for each factor show expression as z-scores of particular factor in hESC, NE, E-RG, M-RG, NEdN, E-RGdN, L-RGdO/dA (right half of circle) and the other way around in the left half of the circle. Inner heatmaps show H3K4me1, DNAme, H3K27ac and H3K4me3 based inferred TFs activity changes for the transitions from hESC to NE, NE to E-RG and E-RG to M-RG (right half) and the other way around in the left half of the circle. Magnification on the right shows ASCL1 as an example **b.** Left: Absolute expression in FPKM for primed TF candidates identified based on gene expression and motif activity patterns at the NE stage but not tested for significant association with GREs associated with genes gaining expression at the E-RG stage and/or in NEdN. TF names associated with significantly ($p \leq 0.05$) enriched motifs are highlighted in green. Motif activity scores for primed candidates are highlighted on the right for H3K4me1, DNAme, H3K27ac and H3K4me3.

However, since the TFBS priming events cannot be caused by the primed factors themselves, we investigated significant ($p \leq 0.01$, odds ratio ≥ 1.5, permutation test) co-binding relationships with factors already active at the NE stage. This analysis revealed distinct co-binding relationships at binding sites associated with genes induced at the NEdN and E-RGdN stage (Fig. 6.15a). While ASCL1 shows significant motif co-localization for PAX3, MSX1 and TEF for genes gaining expression at the E-RG stage, only the co-occurrence of with TEF and MSX1 are conserved for genes that increase their expression in the NEdN populations (Fig. 6.15a). Likewise, we identified the bHLH factor BARHL2 which is involved in diencephalon development (Juraver-Geslin et al., 2011) as a second TF associated with binding site priming specifically at genes induced at the E-RG and M-RG stage (Fig. 6.15a). Similar to ASCL1, binding sites of BARHL2 frequently gain H3K4me1 at the NE stage as well as H3K27ac and H3K4me3 during the transition to E-RG. Interestingly, this factor co-localizes with various components of the previously defined NPC core network such as PAX3 and DMRT1A/3 as well as other key neural genes induced at the NE stage including DLX5 and IRF7 (Fig. 6.15a).

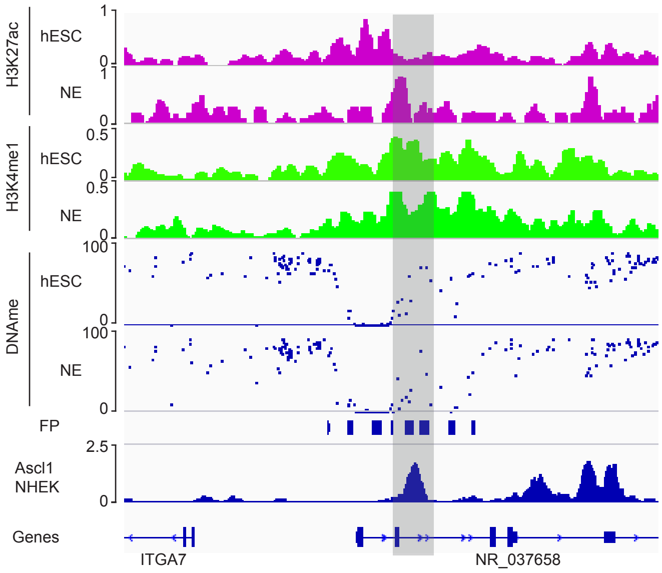


FIGURE 6.16: Epigenetic dynamics at the NR_037658 locus (chr12:56,104-56,116 kb) during the hESC to NE transition. Chromatin enrichment levels are normalized to 1 million reads and indicated on the right. Methylation levels are shown for individual CpGs (blue dots) on a scale from 0 to 100% of methylation (y-axis). Grey boxes highlight footprints that show evidence of priming for Ascl1 motif during hESC to NE transition. For reference Ascl1 binding sites in NHEK cells is shown on the below.

Following the same notion, we identified priming events at the E-RG and M-RG stages, yielding numerous prominent neural and glial factor motifs such as NEUROD4, OLIG3, TFAP2A and EOMES. These observations suggest, that the TERA framework can not only by utilized to identify active key regulators for each differentiation stage, but also provide insights into the priming or preperation of the epigenetic landscape for factors to be expressed at later stages of differentiation, potentially providing important guidelines for differentiation and transdifferentiation protocol optimization.

6.5 Discussion

We have generated extensive reference epigenome maps for several human hESC-derived neural stages and demonstrated the power of these data for dissecting the regulatory relationships that dictate and influence key cell fate decisions. The in depth molecular characterization also provides valuable insights about the identity of the hESC-derived cell types and will help improve the use of these as in vitro models. Furthermore, we have outlined a novel, universally applicable strategy to use epigenetic information to obtain insights into the regulatory logic governing distinct cellular state transitions and infer key drivers of the associated epigenetic dynamics. Our analysis revealed that different epigenetic marks carry complementary information with respect to identification of TFs relevant for the entire cell state transition. Furthermore, we show that H3K4me1 and DNAme changes frequently demarcate gene regulatory elements and TFs that become relevant at a later stage in the differentiation process. These insights and our novel computational framework can be directly extended to other differentiation systems as well as the study of disease-related cellular states such as the comparison of cancerous vs. normal cell types. In addition, the computational framework holds great promise to guide the interpretation of genetic variation in form of SNPs and form the basis of a more comprehensive network modeling approach to predict gene expression.

Chapter 7

Discussion and Conclusion

7.1 Discussion

The goal of this work was two-fold. First, we wanted to establish that the integration of epigenetic and transcriptional data on specific cellular states can be used to gain insights into the underlying regulatory logic maintaining and establishing these states. Second, we wanted to prove the utility of this approach and obtain a mechanistic understanding of specific cell state transitions.

To move closer to this goal, we first had to identify suitable epigenetic signatures that could be used for these purposes. While a lot of previous work examined the utility of various histone modifications (Ernst and Kellis, 2010, Heintzman et al., 2009, Mikkelsen et al., 2007) and demonstrated their power in annotating different classes of gene regulatory elements (GREs), the descriptive power of DNAme in this context was less clear. While earlier work focused on a role for DNAme predominantly in the context of promoter and repetitive element silencing (Bird, 2002), more recent evidence has suggested its involvement in distal regulatory elements such as enhancers as well (Stadler et al., 2011). Combining these observation with the fact that in somatic cell types most of the genome is highly methylated raised the questions of the extent to which DNAme participates in genome regulation and what the targets of this regulation are. Furthermore, we were interested whether it is possible to obtain insights into the regulatory logic of individual cell types. To address these questions, we employed the hypothesis that changes in DNAme between distinct cellular states suggest its participation in some regulatory process related to the differences between the compared cell states. In order to ask these questions on a global level, we set out to chart the dynamic DNAme landscape of the human genome, cataloging all regions that change their DNA methylaion state across a large cohort of normal human cell types.

We identified $\approx 20\%$ of all genomic CpGs that change their methylation status across this cohort in a highly coordinated and non-random fashion. In particular, we found that regions differentially methylated across our normal developmental cohort are highly enriched for gene regulatory elements such as putative enhancers, TF binding sites (TFBS), and DNAse I hypersensitive sites. Furthermore, regions hypo-methylated in a cell type-specific manner are highly enriched for TFBS associated with key TFs relevant to the biology of the particular cell state. In addition, these regions showed highly significant enrichment for GWAS SNPs associated with diseases relevant to the specific cell type. While these findings cannot determine the exact regulatory role of DNAme across these DMRs, they establish DNAme changes as a great marker for cell state-specific regulatory processes. In particular, these observations further establish DNAme as a universal epigenetic signature to obtain insights into cell state-specific TF and associated networks of gene regulatory elements.

Our findings also hold great promise to improve current profiling approaches for DNAme by focusing only on genomic regions that change their DNAme state instead of sequencing the entire genome using WGBS. While this method captures all DMRs, it is highly inefficient wasting at least 70% of all sequencing power to non-informative genomic regions. However, our presented analysis of the dynamic methylation landscape is only an initial draft. In order to complete the map of this landscape, we need to expand the number of distinct cell and tissue types beyond the 30 types covered in this study to identify more subtle dynamics between more closely related cell types. With the availability of close to 100 distinct cell types profiled by WGBS towards the end of this year (2014), it will be possible to greatly diversify the dataset and provide a more complete picture of the dynamic methylation landscape of the human genome. However, our inital analysis of the rate of *de novo* DMR discovery indicates that we already entered the saturation phase with our inital sample set, presumably due to its great diversity of sampled cell and tissue types. The second key question that still requires further investigation is the definition of the minimal methylation difference at a specific genomic locus that renders it dynamic. At his point, it is still unclear what range of DNAme changes can be considered functional. We chose a minimal, significant difference $\geq 30\%$, however it is likely that this threshold will not apply to all scenarios. Instead, the proper threshold is more likely to depend on the genomic context. In order to derive a context-dependent methylation variability level threshold, a new statistical model for each CpG or CpG class will be required. This model should take the sequence composition, repeat content, mappability, shearing bias as well as general variability of the CpG/CpG class into consideration. Furthermore, it will be necessary to test the functional impact of different levels of methylation at different classes of gene regulatory elements such as enhancers, promoters, and TF binding sites using reporter assays.

These questions directly connect to even broader questions on DNAme. Much work still needs to be done in order to dissect the specific function exerted by DNAme in distinct genomic contexts such as TF binding sites, promoters and enhancers. Under what circumstances does DNAme block gene reactivation or TF binding? In which contexts is it just a gap filler that is easily displaced without any particular function? Dissecting these and related questions promise an exciting future for DNAme research and will greatly benefit from the recent emergence of genome and epigenome editing technology. Furthermore, using these insights in combination with our profiling-based findings and extended profiling of larger tissue and individual cohorts harbors the potential to yield novel biomarkers and novel insights into disease biology through a network-centric analysis of the resulting data sets.

After we established the utility of DNAme to identify gene regulatory elements and associated TF networks, we next wanted to combine this power with chromatin modification and transcriptional information in an integrative analysis to dissect the cell state transitions that occur during the specification of hESCs (hESCs) into three distinct progenitor populations representing each embryonic germ layer. To that end, we established an *in vitro* model system that permits the generation of ectodermal, mesodermal, and endodermal progenitor populations from hESCs. While earlier work focused on distantly related, mostly somatic cell types, we were specifically interested in understanding the epigenetic changes that occur during differentiation, potentially capturing also transient dynamics, which are not detectable when only studying the far endpoints of differentiation trajectories. Our horizontal comparison of three hESC-derived cell populations revealed complementary pathways and gene regulatory elements that become silenced or activated across the three germ layers, while pluripotency-associated genes and gene regulatory elements frequently become silenced. Using our epigenetic signature approach, we identified a gene regulatory module set shared between early ectoderm and late endoderm/pancreas development. Furthermore, we uncovered two unexpected epigenetic changes, presumably only detectable in transient cell populations. First, the gain of H3K4me1 was found to be frequently associated with epigenetic priming and occurring in the vicinity of genes expressed at later stages of development. Second, the gain of the repressive mark H3K27me3 at regions of high methylation in the vicinity of genes also expressed at later stages of development. Detailed motif analysis and subsequent ChIP-Seq analysis revealed strong association with the binding of the pioneering factor FOXA2 that was also associated with the loss of DNAme. However, the exact molecular mechanisms of the pioneering factor action remains to be determined. In addition, it is unclear whether FOXA2 is indeed required and sufficient to mediate the observed dynamic and whether or not this remodeling event is required for further differentiation. While these findings highlight the power of an integrative transcriptional and epigenetic analysis of closely related cell types, this study lacked the ability to track individual genomic regions over more than one time point and define their trajectories. In addition, no coherent framework for a TF centric analysis of the dataset was available. Uncovering the details of the TF networks driving these transitions remains a key challenge. Currently two basic complementary approaches are available: While the straightforward strategy would be to identify all TFs exhibiting differential expression in this context and subsequently perform ChIP-Seq experiments for the selected factors, a less resource-intensive approach would be to try to utilize the epigenetic patterns in combination with motif analysis to obtain these insights.

Earlier studies, including the ones discussed here, have yielded already very promising results and a detailed motif based analysis of mammalian TF networks have been carried out using DNAse I footprinting (Neph et al., 2012, Thurman et al., 2012). However, using chromatin and DNAme information for this purpose has several advantages over the use of DNAse I. From a semantic perspective a lot is known about the functional context in which certain histone modifications occur. Therefore, occurrence of TF binding in the context of a specific histone modification already suggests a possible function for this binding. From a practical perspective, ChIP-Seq and WGBS experiments are by far easier to conduct and require much less input material (as low as 100,000 cells in contrast to 100,000,000 cells for DNAse footprinting assays). Finally, more than one laboratory is capable of conducting WGBS and ChIP-Seq as opposed to DNAse I assays. Furthermore, our studies and that of others (Ernst and Kellis, 2010, Ernst et al., 2011, Stadler et al., 2011) suggest that a large fraction of the information contained in DNAse

I HS sites is also captured by histone and DNAme information, although at lower resolution.

To overcome the discussed drawbacks, we developed a novel computational strategy to perform a TF centric analysis using motif detection to identify potentially driving factors of specific epigenetic state transitions. Using this method, we evaluated the information content and overlap of different epigenetic signatures during a 200-day differentiation time course of hESCs toward the neural lineage.

By combining the epigenetically driven key TF inference with transcriptional information on each of the differentiation stages, we identified a core TF network of PAX6 and OTX2 at its center which is active through the entire differentiation process. The members of this core network dynamically rewire a subset of their binding sites in a stage-specific fashion by associating with distinct co-binding partners that are expressed in a stage-specific fashion. In this way, the core factors are likely to orchestrate and establish distinct differentiation propensities while maintaining proliferative potential over the course of *in vitro* neural differentiation. Interestingly, many downstream effectors of key signaling pathways are among the stage-specific co-binding partners, frequently coinciding with the stimulation of the corresponding signaling pathway at the respective stage. This suggests that the activation of distinct signaling pathways can be directly inferred from its resulting footprint on the stage-specific epigenome.

Our analysis also revealed complementary changes in H3K27ac and H3K4me3 are much more associated with immediate changes in TFs activity at each particular stage. In contrast, regions changing their DNAme or H3K4me1 state frequently contain information not only on TF-activated at the stage of change but also at later stages of differentiation. These changes can be interpreted as priming events of neuronal TFBS reflecting the differentiation potential of each stage. In addition to the developmental analysis, we also utilize this framework to interpret non-coding variants associated with disease phenotypes related to the specific cell types under study. Specifically, we discuss the integration of an Alzheimer's disease-related SNP upstream of the APOE gene in the stage-specific TF network.

In summary, we have developed a novel computational strategy to identify key factors and networks of TFs, GREs and genes likely to be driving epigenetic remodeling during cell state transition and utilized this framework to dissect *in vitro* neural differentiation of hESCs. However, the presented work is only a first step to decipher the complex wiring logic and control structures of the regulatory networks governing these cell state transitions. From the computational perspective, it will be important to incorporate as many distinct cell and tissue types as possible to also understand which gene regulatory elements are being controlled by which factor combinations. An increase in sample number will also greatly increase the predictive power of this model. Furthermore, new statistical approaches are required to assess the likelihood of regulation and remodeling of a particular GRE by a specific TF or combination of TFs in a probabilistic fashion. Moreover, much work needs to be done to interpret the genomic regions controlled by a specific TF in terms of their biological processes and functions along the lines of a gene set enrichment analysis. These efforts would then ultimately lead to a probabilistic model, allowing for the prediction of gene expression based on inferred TF binding at gene associated gene regulatory elements. This would provide a significant conceptual and methodological advance in the field, since current approaches either rely on ChIP-Seq data for TF binding (Wilczynski et al., 2012) or on chromatin state data (Natarajan et al., 2012). Finally, the approach presented here does only consider TF binding as key driving events, ignoring all other layers of genomic and transcriptional regulation. Extending the current model to incorporate protein phosphorylation status as well as microRNA levels and the cells' external signaling environment will greatly improve the power of computational approaches to predict and explain cellular phenotypes.

Besides improvements on the computational end, we have only laid the groundwork for understanding the basis of *in vitro* neural differentiation. Our observations are so far mainly based on correlation analysis and motif occurrences and therefore indirect. It will be essential to functionally validate the key TFs predicted to orchestrate each cell state transition by RNAi-mediated knockdown. In addition, it will be crucial to confirm the predicted TF binding and rewiring events by ChIP-Seq for a selected subset of factors at each stage of the differentiation process. Furthermore, it will be very informative to assess the role and importance of individual gene regulatory elements and their associated TF binding by specifically deleting these elements from the genome in the relevant cell types. Besides these functional validations, it will be key to demonstrate more conclusively the relation of the dynamics observed in our *in vitro* system to dynamics occurring *in vivo*. To that end, a collection of micro-dissected brain regions from developing human fetuses would be required. Since the latter are not easy to obtain, the collection of a high-resolution time series in mouse and primates poses a reasonable alternative. Another dimension so far not touched on is the profund heterogeneity occurring during *in vitro* differentiation. Even though we use a HES5 reporter to isolate the neural stem cells, it is unclear what the heterogeneity at each time point within this population does look like. Within the mature neuronal populations the heterogeneity is likely to be substantially higher. However, so far we only investigated population averages with the assays at our disposal. To overcome this drawback and dissect the occurring cell state dynamcis at higher resolution, high-throughput single-cell assays such as single-cell RNA-Seq will be required.

Lastly, it will be of great clinical relevance to optimize the numbers, homogeneity, and maturation level of specific neuronal subtypes that can be obtained from the distinct stem cell populations. To that end, it will be required to systematically test distinct growth factor concentrations and differentiation protocols and evaluate their impact on the differentiation trajectory toward the desired mature neuronal cell types.

7.2 Outlook

The research presented in this work represents a step forward on the route to a predictive model of cell state maintenance and cell state transitions. However, many improvements are still necessary that can now be made on the foundation prepared by us and others.

First, a more powerful mathematical framework integrating genetic, epigenetic, and transcriptional information needs to be developed. In order to incorporate changes in the environmental conditions, an input-driven model would be ideal. In this context, dynamic Bayesian networks seem very promising.

This framework can then be applied to a large collection of epigenetic datasets for more than 100 cell types, generated by the ENCODE and REMC consortia, providing sufficient power to infer cell type specific gene regulatory networks.

The resulting flexible cell state models can then be used as a general tool to interpret molecular profiles of other cell states of interest, integrating novel cell state information as needed. Several interesting applications come to mind:

1. The interpretation of non-coding variants and the inference of their possible mechanisms of actions.

2. Re-interpretation of GWAS studies with improved power, focusing on SNPs located in gene regulatory elements, reducing the genome-wide significance level to a regulatory element significance level.

3. Dissecting complex traits by integrating non-coding variants with small effects on a network level.

4. Identifying pathways of action of disease-associated rare variants with large phenotypic effect in order to uncover potential druggable pathways and infer genetically and epigenetically informed drug dose-response curves.

5. Interpretation of electronic health record-based pheWAS studies.

All these suggestions arise from the fundamental opportunity that the integration of genetic, epigenetic, and transcriptional data provides: It structures the high-dimensional space of possible cellular states by providing constraints on the accessible phase space. However, in order to overcome several drawback associated with the population-induced heterogeneity and correlation-based nature of these measurements, the analysis should be complemented with single-cell measurements as well as focused functional validation using high-throughput genome and epigenome editing in the cell system under study.
The approach outlined above presents a unique resource allowing for a mechanistic, causal interpretation of normal cellular trajectories and the deviation from those in disease phenotypes.

Appendix A

Appendix

A.1 Methods related to chapter 4 and 5

A.1.1 WGBS

To that end Genomic DNA (1-5 μg) was fragmented to 100-500 bp using a Covaris S2 sonicator 9 times for 60 s at duty cycle 20%, intensity 5 and 200 cycles per burst. DNA fragments were cleaned up using a QIAGEN PCR purification kit. End-repair reactions (100 μl) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 unites Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min. at 19 °C and 15 min. For some libraries we used a dCTP-free dNTP mix instead of all four dNTPs during for the end-repair to avoid artificially unmethylated sites. Adenylation was performed for 30 min. at 37 °C in 50 μl 1x Klenow buffer containing 0.2 mM dATP and 15 units Klenow exo- (NEB). Adenylated DNA fragments and methylated paired-end adapters (purchased from ATDBio) were incubated overnight at 16 °C in a 50 μl reaction containing 5,000 units concentrated T4 DNA ligase (NEB) and 3 μM of adapters. Each enzymatic reaction was terminated and cleaned-up by phenol/chloroform extraction and ethanol precipitation as described above. To determine unmethylated cytosine conversion rates and methylated cytosine over-conversion rates by sodium bisulfite treatment, adapter-ligated fully methylated and fully unmethylated internal control DNA fragments, were spiked into WGBS library preparation at a molar ratio (spike-in to WGBS library) of 1:16,000 each. Adapter-ligated DNA of 270-370 bp, corresponding to DNA insert sizes of 150-250 bp, was size-selected on a 2.5% Nusieve (3:1) agarose gel (Lonza). Two consecutive bisulfite conversions were performed with an EpiTect Bisulfite Kit (QIAGEN) following the protocol specified for DNA isolated from FFPE tissue samples. One of 40 μl bisulfite-converted DNA was used in each of four 10-μl reactions to determine the minimal PCR cycle number for library amplification. PCR reactions contained 0.5 U of PfuTurboCx Hotstart DNA polymerase (Agilent technologies), 1 μl of 10x PCR buffer, 250 μM dNTPs, 1.5 μM of Primer 1.0 and 2.0 (Illumina). The thermocycling profile was 2 min. at 95 °C followed by 5-15 cycles of 30 s at 95 °C, 30 s at 65 °C, 1 min. at 72 °C, and a final 7-min. extension at 72 °C. Preparative library amplification using the empirically determined number of PCR cycles was performed in eight 25-μl aliquots, each containing 3 μl of bisulfite-converted DNA, 1.25 U of PfuTurboCx Hotstart DNA polymerase, 2.5 μl of 10x PCR buffer, 250 μM of dNTP, 1.5 μM of Primer 1.0 and 2.0. PCR products were pooled and purified twice using Agencourt AMPure XP SPRI Beads (Beckman Coulter) as per the manufacturer's instructions. The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen). The insert size was checked on a 4-20% non-denaturing polyacrylamide gel (Bio-Rad). Paired-end sequencing with 100 base reads was performed on an Illumina Hiseq 2000 followed the manufacturer's guidelines.

A.1.2 Bisulfite-ChIP

DNA was first subjected to end-repair in a 30-μl reaction containing 6 units T4 DNA polymerase, 2.5 units DNA Polymerase I (Large Klenow Fragment), 20 units T4 Polynucleotide Kinase (all New England Biolabs), dATP, dCTP, dGTP, and dTTP (0.125 mM each), and 1x T4 Ligase buffer with ATP for 30 min at 20 °C. DNA was then adenylated in a 20-μl reaction containing 10 units Klenow Fragment (3' to 5' exo-) (New England Biolabs), 0.5 mM dATP and 1x NEB buffer 2 for 30 min at 37 °C. DNA was then ligated to preannealed Illumina genomic DNA adapters containing 5-methylcytosine instead of cytosine (ATDBio) using T4 DNA ligase (New England Biolabs). Adapter-ligated DNA fragments were subsequently purified by phenol extraction and ethanol precipitation and size-selected on gel. 50 ng sheared and dephosphorylated Escherichia coli K12 genomic DNA was added to adapter-ligated DNA as carrier during size-selection and bisulfite conversion. DNA was run on 2.5% Nusieve 3:1 Agarose (Lonza) gels. Lanes containing marker (50 bp ladder; New England Biolabs) were stained with SYBR Green (Invitrogen), and size regions to be excised were marked with toothpicks and adapter-ligated DNA fragments from 200-400 and 400-550 bp were excised. DNA was isolated from gel using the MinElute Gel Extraction kit (QIAGEN). The low and high libraries were kept separate in subsequent steps.

Adapter-ligated and size-selected DNA was subjected to two subsequent 5-h bisulfite treatments using the EpiTect Bisulfite kit (QIAGEN) following the manufacturer's protocol for DNA isolated from FFPE tissue samples. PCR amplification was done with 1.25 units Pfu Turbo Cx Hotstart DNA Polymerase (Stratagene), primer LPX 1.1 and 2.1 (0.3 μM each), dNTPs (0.25 mM each), 1x Turbo Cx buffer. Amplified libraries were purified with the MinElute PCR Purification kit (QIAGEN) and subsequently purified from gel essentially as described above; whole gels were stained with SYBR Green, and no carrier DNA was added. Final libraries were analyzed on analytical 4%-20% TBE Criterion precast gels (BioRad), and measured by Quant-iT dsDNA HS Assays (Invitrogen).

A.1.3 Cell Culture

All in vitro derived cell types were derived from HUES64 (Chen et al., 2009). Human embryonic stem cells were expanded on murine embryonic fibroblasts (Global Stem) in KO-DMEM (Life Technologies) containing 20% Knockout serum replace (Life Technologies) and FGF2 (10 ng/mL) (Millipore). Cultures were passaged by enzymatic dissociation using Collagenase IV (1mg/mL) (Life Technologies). Prior to differentiation, cells were plated on matrigel-coated plates (BD Biosciences) and cultured in mTeSR1 (Stem Cell Technologies) for 3 to 4 days. Endoderm differentiation was induced in Advanced RPMI (Invitrogen), 0.5% FBS (Hyclone), Activin A (100ng/mL) (R&D) and WNT3A (50 ng/mL) (R&D). HUES64- derived hepatoblasts (dHep) were induced by culturing day 5 endoderm in RPMI media containing B27(1X), FGF2 (10ng/mL)(Millipore) and BMP4(20ng/mL)(R&D) for five days, and collected after 10 days total of differentiation. Hepatocyte-like cells were derived by culturing the HUES64-derived hepatoblasts in Lonza hepatocyte culture media containing 10ng/mL of HGF (R&D) for 5 additional days, or 15 days total. Mesoderm differentiation was induced by the addition of media consisting of in DMEM/F12 (Life Technologies), 0.5% FBS (Hyclone), Activin A (100ng/mL) (R&D) (for the first 24 hours only), BMP4 (100ng/mL) (R&D), VEGF

(100ng/mL) (R&D) and FGF2 (20ng/mL) (Millipore). To induce osteoblast differentiation, the day 5 mesoderm population was dissociated with accutase and replacted on matrigel coated plates (BD) in EGM-2 media (Lonza) for 7 days, or 12 days total. Ectoderm differentiation was induced using A83-01 (2um) (Tocris), PNU 74654 (2um) (Tocris) and Dorsomorphin (2um) (Tocris), DMEM/F12 (Life Technologies) containing 1% Knock serum replacer (Life Technologies). Neurectoderm differentiation was induced by switching the day 5 ectoderm population to media containing 3 μM CHIR99021 (TOCRIS), 10 μM SU5402 (TOCRIS), and 10 μM DAPT (TOCRIS), and collected after 6 more days, or 11 days total. N2-supplement (Life Technologies) was added to cells in 25% increments every other day beginning four days after the initiation of ectoderm differentiation. For all cell types, media was changed daily.

A.1.4 Antibodies

ChIP was performed using the following antibodies: H3K4me3 (Millipore, 07-473, Lot DAM1623866), H3K27ac (Abcam, ab4729, Lot 509313), H3K27me3 (Millipore, 07-449, Lot DAM1514011), H3K36me3 (Abcam, ab9050, Lot 499302), H3K4me1 (Abcam, ab8895, Lot 659352), H3K9me3 (Abcam, ab8898, Lot 484088), POU5F1 (Abcam, ab19857), SOX2 (Santa Cruz, sc-17320X), NANOG (R&D, AF1997) and FOXA2 (R&D, AF2400).

For live cell FACS isolation, cells were stained for 30 minutes on ice with the following antibodies directed towards extracellular surface proteins: CD326-PerCP-Cy5 (clone EBA1) (BD Biosciences), CD56-PE (clone NCAM16.2) (BD Biosciences), and CD184-PE-Cy5 (clone 12G5) (BD Biosciences).

Immunostaining was done with the following primary antibodies: FOXA2 (R&D, AF2400), GATA2 (Santa Cruz, sc-16044) SOX17 (R&D, AF1924), PAX6 (Covance, PRB-278P) and HNF4a (abcam, ab41989). Cells were fixed in 4% Formaldehyde, incubated in primary antibody overnight at 4 °C, and then incubated in secondary antibody for 1 hr at room temperature. DNA was detected using Hoechst 33342 trihydrochloride trihydrate (Invitrogen).

A.1.5 FACS Analysis

FACS was done on a BD FACSAria II using linear FSC and SSC scaling, followed by height and width-based doublet discrimination. The viability of the populations was assessed by Propidium Iodide staining, with the positively stained populations being excluded from the sorting gates. Compensation was calculated using FACS Diva autocompensation algorithms, and supplemented by manual compensation to correct for autofluorescence. Antibodies were used as described in the text.

A.1.6 Genomic DNA isolation

Flash-frozen human tissues or cell pellets were lysed at 55 °C overnight in 300-600 μl lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM EDTA, 10 mM NaCl and 0.5% wt/vol SDS) supplemented with 50 ng/μl DNase-free RNase (Roche) and 1 μg/μl proteinase K (NEB). After extraction with an equal volume of phenol:chlorofom:isopropanol alcohol (25:24:1; Invitrogen) and addition of 0.5 μl (20 μg/μl) glycogen (Roche) and 1/20 vol 5 M NaCl, DNA was precipitated with 2.5 vol ethanol, spun down (30 min/16,000 g) at °C and washed with 70% ethanol. DNA was re-suspended in 30-100 μl of TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and quantified using a Qubit fluorometer and a dsDNA BR Assay Kit (Life Technologies).

A.1.7 Chromatin immunoprecipitation (ChIP) and ChIP-sequencing library production

Cells collected by FACS were crosslinked in 1% formaldehyde for 15 minutes at room temperature, with constant agitation, followed by quenching with 125mM Glycine for 5 minutes at room temperature with constant agitation. Nuclei were isolated and chromatin was sheared using Branson sonifier until the majority of DNA was in the range of 200-700 base pairs. Chromatin was incubated with antibody overnight at 4 °C, with constant agitation.

Co-immunoprecipitation of antibody-protein complexes was completed using Protein A or Protein G Dynabeads for 1 hour 4 °C, with constant agitation. ChIPs were completed using previously reported methods Mikkelsen et al. (2010). Sequencing libraries were submitted for sequencing on the Illumina Hiseq 2000. Immunoprecipitated DNA was end repaired using the End-It DNA End-Repair Kit (Epicentre), extended using a Klenow fragment (3' to 5' exo)(NEB), and ligated to sequencing adapter oligos (Illumina). Each library was then PCR-amplified using PFU Ultra II Hotstart Master Mix (Agilent), and a size range of 300-600 was selected for sequencing. We confirmed binding of OCT4, NANOG and SOX2 at the NANOG promoter using qPCR.

A.2 Methods related to chapter 6

A.2.1 Culturing undifferentiated hESCs

HES5::eGFP BAC transgenic human ES cells (H9; WA-09; Wicell) expressing GFP under the HES5 promoter were cultured on mitotically inactivated mouse embryonic fibroblasts (MEFs) (Globalstem). Undifferentiated hES cells were maintained as described previously Elkabetz et al. (2008) in medium containing DMEM/F12, 20% KSR, 1mM Glutamine, 1% Penicillin/Streptomycin, non essential amino acids and beta-mercaptoethanol. Undifferentiated hES cells were purified with pluripotency markers Tra-1-60 and PE-conjugated SSEA-3(BD Pharmingen).

A.2.2 Neural induction and long-term propagation of NPCs

Neural differentiation of hES cells was performed as described previously Chambers and Tomlinson (2009), Elkabetz et al. (2008). Briefly, neuroectodermal cells were generated either by monolayer induction - with dissociated hES cells plated on Matrigel (BD biosciences), or by co-culture on MS5 stromal cells. In both cases neural fate was directed by dual SMAD inhibition protocol Chambers and Tomlinson (2009). Neural rosettes were harvested mechanically beginning on day 8-10 of differentiation. Rosettes were replated on culture dishes pre-coated with 15 μg/mL polyornithine (Sigma), 1 μg/mL Laminin (BD Biosciences) and 1ug/ml Fibronectin (BD Biosciences) (Po/Lam/FN) in N2 medium composed of DMEM/F12 and N2 supplement (Invitrogen). N2 supplement contained Insulin, Apo-transferin, Sodium Selenite, Putrecine and Progesterone. This medium was supplemented with SHH (200 ng/mL), FGF8 (100 ng/mL) and BDNF (20 ng/mL) (all from R&D Systems) to maintain early anterior regionalization of the neural plate. These factors were gradually replaced by FGF2 (20 ng/mL) and EGF (20 ng/mL) in the following two weeks of differentiation in order to maintain a proliferative (FGF and EGF responsive) NPC state. NPCs from all stages were collected at indicated days and FACS purified for HES5::GFP (NE to L-RG) or EGFR for LNPs to purify for the highest NPC state for each stage. Neuroectodermal cells were collected at day 12 of differentiation, Neuroepithelial/early radial glial cells were collected at day 14, mid neurogenesis radial glial cells were collected at day 35, late gliogenic radial glial cells were collected at day 80, and long term NPCs were collected at day 220. At each stage cells were either split for next passage or subjected to FACS purification for HES5::GFP as described. Cells were replated onto Po/Lam/FN culture dishes. For neuronal, astroglial or oligodendroglial differentiation, NPCs were seeded at high density and subjected to differentiated for 17 days in the presence of AA/BDNF, 5% Fetal Bovine Serum (FBS) (Invitrogen), or AA/BDNF/SHH/FGF8, respectively.

A.2.3 Chromatin Immunoprecipitation followed by sequencing (ChIP Seq)

We used similar approaches to Garber-Yosef et al. Garber et al. (2012). Particularly, 160.000 thousand Cells were crosslinked in formaldehyde (1%, 37 °C for 10 min), followed by quenching with glycine (5 min at 37 °C), washed with PBS containing protease inhibitor (Roche, 04693159001) and flash frozen in liquid nitrogen. To lyse the cells, we used 1% SDS, 10mM EDTA and 50mM Tris-HCl pH 8.1 complemented with protease inhibitor. The chromatin was then fragmented with a Branson Sonifier (model S-450D) at 4 °C, calibrated to a size range of 200 and 800 bp. For each antibody, 1 to 5 ug was conjugated to Protein-A and Protein-G Dynabeads mix (Invitrogen, 100-02D and100-07D, respectively) for 2 hours in blocking buffer (PBS containing 0.5% BSA and 0.5% TWEEN). Next, the conjugated antibody-beads were added to the sheared chromatin, and incubated overnight. Samples were washed 6 times with RIPA buffer (10mM Tris-HCl pH 8.0, 1mM EDTA pH 8.0, 14mM NaCl, 1% TritonX-100, 0.1% SDS, 0.1% DOC), twice with RIPA buffer containing 500mM NaCl, twice with LiCl buffer (10 mM TE, 250mM LiCl , 0.5% NP-40, 0.5% DOC), twice with TE (10Mm Tris-HCl pH 8.0, 1mM EDTA), and then eluted in elution buffer (10mM Tris-Cl pH 8.0, 5mM EDTA, 300mM NaCl, 0.1% SDS; pH 8.0) at 65 °C. Eluate was incubated in 65 °C over-night, and then treated sequentially with RNaseA (Roche, 11119915001) for 30 min and Proteinase K (NEB, P8102S) for two hours, and where then followed by library construction.

A.2.4 ChIP Seq library preparation and sequencing

To extract DNA and create the Illumina library we used AMPure XP beads (Agencourt) Solid-phase reversible immobilization (SPRI). SPRI beads were added to the samples, mixed 15 times, incubated at room temperature for 2 minutes. Supernatant was extracted from the beads from the beads 4 minutes on a magnet. We used 70% ethanol to wash the beads and then dried for 4 minutes. 40 ul EB buffer (10 mM Tris-HCl pH 8.0) was used to elute the DNA. The following steps of Illumina library construction (end-repair, addition of A-base, ligation of barcoded adaptors and PCR enrichment) we used a general SPRI cleanup procedure: addition of PEG buffer (20% PEG and 2.5 M NaCl), and extracted and washed as above. The enzymatic reactions were carried as follows: 1. DNA end-repair: T4 PNK and T4 polymerase (New England Biolabs) incubated at 12 °C for 15 min, 25 °C for 15 min; 2. A-base addition: Klenow (3'-¿5' exonuclease; New England Biolabs) incubated at 37 °C for 30 min. 3. Adaptor ligation: DNA ligase (New England Biolabs) and indexed oligo adaptors and incubated at 25 °C for 15 min, followed by 0.7X SPRI/reaction to remove non-ligated adaptors. 4. PCR enrichment: PCR mastermix (primer set, dNTP mix, Pfu Ultra Buffer (Agilent), Pfu Ultra-II Fusion (Agilent), water), for 20 cycles. The PCR amplified libraries we cleaned up using 0.7X SPRI/reaction (size selection mode) to remove excessive primers. Roughly 5 picomoles of DNA library was then applied to each lane of the flow cell and sequenced on Illumina HiSeq 2000 sequencers according to standard Illumina protocols.

A.2.5 Normalization evaluation

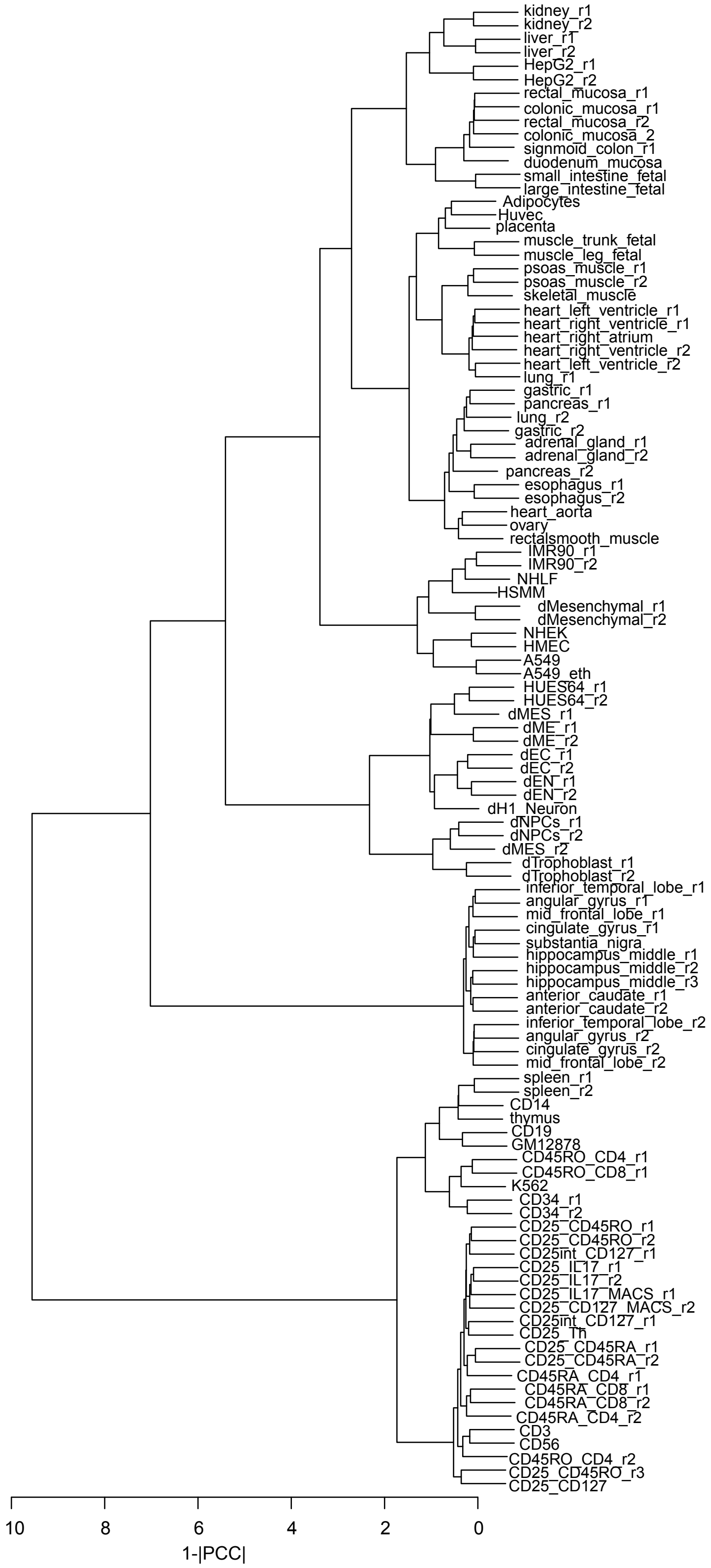


Figure A.1: Hierarchical clustering of H3K27ac enriched IDR called peaks across 106 distinct REMC samples after normalization using the optimal strategy described in the main text. Hierarchical clustering was performed using one minus the absolute Pearson correlation coefficient (PCC) as distance metric and Ward's method

Bibliography

G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

U. Ahlgren, S. L. Pfaff, T. M. Jessell, T. Edlund, and H. Edlund. Independent requirement for isl1 in formation of pancreatic mesenchyme and islet cells. *Nature*, 385(6613): 257–60, 1997.

Aristotle and H. Lawson-Tancred. *Metaphysics*. Penguin classics. Penguin Books, London ; New York, 1998.

Aristotle and A. L. Peck. *Generation of animals*. Loeb classical library. W. Heinemann ; Harvard University Press, London Cambridge, Mass., 1943.

T. Bartke, M. Vermeulen, B. Xhemalce, S. C. Robson, M. Mann, and T. Kouzarides. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*, 143(3):470–84, 2010.

S. B. Baylin and P. A. Jones. A decade of exploring the cancer epigenome - biological and translational implications. *Nature Reviews: Cancer*, 11(10):726–34, 2011.

C. Becker, J. Hagmann, J. Muller, D. Koenig, O. Stegle, K. Borgwardt, and D. Weigel. Spontaneous epigenetic variation in the arabidopsis thaliana methylome. *Nature*, 480 (7376):245–9, 2011.

A. C. Bell and G. Felsenfeld. Methylation of a ctcf-dependent boundary controls imprinted expression of the igf2 gene. *Nature*, 405(6785):482–5, 2000.

C. G. Bell, G. A. Wilson, L. M. Butcher, C. Roos, L. Walter, and S. Beck. Human-specific CpG "beacons" identify loci associated with human-specific traits and disease. *Epigenetics*, 7(10):1188–99, 2012.

O. Bell, V. K. Tiwari, N. H. Thoma, and D. Schubeler. Determinants and dynamics of genome accessibility. *Nature Reviews: Genetics*, 12(8):554–64, 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, 1995.

Y. Bergman and H. Cedar. Dna methylation dynamics in health and disease. *Nature Structural and Molecular Biology*, 20(3):274–81, 2013.

B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. Lange, C. M. van Dijk, R. A. Tollenaar, D. Van Den Berg, and P. W. Laird. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*, 44(1):40–6, 2012.

B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26, 2006.

B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The nih roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–8, 2010.

B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, 2012.

N. Bertrand, D. S. Castro, and F. Guillemot. Proneural genes and the specification of neural cell types. *Nature Reviews: Neuroscience*, 3(7):517–30, 2002.

T. H. Bestor. The DNA methyltransferases of mammals. *Human Molecular Genetics*, 9 (16):2395–402, 2000.

M. D. Biggin. Animal transcription networks as highly connected, quantitative continua. *Developmental Cell*, 21(4):611–26, 2011.

A. Bird. Dna methylation patterns and epigenetic memory. *Genes and Development*, 16 (1):6–21, 2002.

C. Bock. Analysing and interpreting DNA methylation data. *Nature Reviews: Genetics*, 13(10):705–19, 2012.

C. Bock, E. Kiskinis, G. Verstappen, H. Gu, G. Boulting, Z. D. Smith, M. Ziller, G. F. Croft, M. W. Amoroso, D. H. Oakley, A. Gnirke, K. Eggan, and A. Meissner. Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144(3):439–52, 2011.

A. L. Boulesteix and K. Strimmer. Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, 2:23, 2005.

A. L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings Bioinformatics*, 8(1):32–44, 2007.

R. Briggs and T. J. King. Transplantation of living nuclei from blastula cells into enucleated frogs eggs. *Proceedings of the National Academy of Sciences of the United States of America*, 38(5):455–463, 1952.

A. B. Brinkman, H. Gu, S. J. Bartels, Y. Zhang, F. Matarese, F. Simmer, H. Marks, C. Bock, A. Gnirke, A. Meissner, and H. G. Stunnenberg. Sequential chip-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Research*, 22(6):1128–38, 2012.

M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes and Development*, 25(18):1915–27, 2011.

C. L. Cai, X. Liang, Y. Shi, P. H. Chu, S. L. Pfaff, J. Chen, and S. Evans. Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Developmental Cell*, 5(6):877–89, 2003.

I. Chambers and S. R. Tomlinson. The transcriptional foundation of pluripotency. *Development*, 136(14):2311–22, 2009.

J. C. Chambers, W. Zhang, J. Sehmi, X. Li, M. N. Wass, P. Van der Harst, H. Holm, S. Sanna, M. Kavousi, S. E. Baumeister, L. J. Coin, G. Deng, C. Gieger, N. L. Heard-Costa, J. J. Hottenga, B. Kuhnel, V. Kumar, V. Lagou, L. Liang, J. Luan, P. M. Vidal, I. Mateo Leach, P. F. O'Reilly, J. F. Peden, N. Rahmioglu, P. Soininen, E. K. Speliotes, X. Yuan, G. Thorleifsson, B. Z. Alizadeh, L. D. Atwood, I. B. Borecki, M. J. Brown, P. Charoen, F. Cucca, D. Das, E. J. de Geus, A. L. Dixon, A. Doring, G. Ehret, G. I. Eyjolfsson, M. Farrall, N. G. Forouhi, N. Friedrich, W. Goessling, D. F. Gudbjartsson, T. B. Harris, A. L. Hartikainen, S. Heath, G. M. Hirschfield, A. Hofman, G. Homuth, E. Hypponen, H. L. Janssen, T. Johnson, A. J. Kangas, I. P. Kema, J. P. Kuhn, S. Lai, M. Lathrop, M. M. Lerch, Y. Li, T. J. Liang, J. P. Lin, R. J. Loos, N. G. Martin, M. F. Moffatt, G. W. Montgomery, P. B. Munroe, K. Musunuru, Y. Nakamura, C. J. O'Donnell, I. Olafsson, B. W. Penninx, A. Pouta, B. P. Prins, I. Prokopenko, R. Puls, A. Ruokonen, M. J. Savolainen, D. Schlessinger, J. N. Schouten, U. Seedorf, S. Sen-Chowdhry, K. A. Siminovitch, J. H. Smit, T. D. Spector, W. Tan, T. M. Teslovich, T. Tukiainen, A. G. Uitterlinden, M. M. Van der Klauw, R. S. Vasan, C. Wallace, H. Wallaschofski, H. E. Wichmann, G. Willemsen, P. Wurtz, C. Xu, L. M. Yerges-Armstrong, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genetics*, 43(11):1131–8, 2011.

S. M. Chambers, Y. Qi, Y. Mica, G. Lee, X. J. Zhang, L. Niu, J. Bilsland, L. Cao, E. Stevens, P. Whiting, S. H. Shi, and L. Studer. Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. *Nature Biotechnology*, 30(7):715–20, 2012.

T. Chen and S. Y. Dent. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews: Genetics*, 2013.

Y. F. Chen, C. Y. Tseng, H. W. Wang, H. C. Kuo, V. W. Yang, and O. K. Lee. Rapid generation of mature hepatocyte-like cells from human induced pluripotent stem cells by an efficient three-step protocol. *Hepatology*, 55(4):1193–203, 2012.

N. M. Cohen, E. Kenigsberg, and A. Tanay. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145(5):773–86, 2011.

E. R. Consortium. Standards and guidelines for whole genome shotgun bisulfite sequencing, 2011.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 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(50):21931–6, 2010.

E. H. Davidson and D. H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006.

A. M. Deaton and A. Bird. Cpg islands and the regulation of transcription. *Genes and Development*, 25(10):1010–22, 2011.

S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–84, 2011.

S. Dejong. Simpls - an alternative approach to partial least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

A. DeLaForest, M. Nagaoka, K. Si-Tayeb, F. K. Noto, G. Konopka, M. A. Battle, and S. A. Duncan. Hnf4a is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development*, 138(19):4143–53, 2011.

R. DerSimonian and R. Kacker. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28(2):105–114, 2007.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–88, 1986.

B. Dorigo, T. Schalch, K. Bystricky, and T. J. Richmond. Chromatin fiber folding: requirement for the histone h4 n-terminal tail. *Journal of Molecular Biology*, 327(1): 85–96, 2003.

R. Edri, Y. Yaffe, M. J. Ziller, O. Ziv, A. Zaritski, E. David, J. Jacob-Hirsch, G. Rechavi, I. Gat-Vics, A. Meissner, and Y. Elkabetz. Prospective isolation of distinct human es cell derived neural progenitor cells provides a cell culture model for cns establishment and cerebral development. *Currently submitted to Cell Stem Cell*, 2014.

M. Ehrlich. Dna hypomethylation in cancer cells. *Epigenomics*, 1(2):239–59, 2009.

Y. Elkabetz and L. Studer. Human esc-derived neural rosettes and neural stem cell progression. *Cold Spring Harbor Symposia on Quantitative Biology*, 73:377–87, 2008.

Y. Elkabetz, G. Panagiotakos, G. Al Shamy, N. D. Socci, V. Tabar, and L. Studer. Human es cell-derived neural rosettes reveal a functionally distinct early neural stem cell stage. *Genes and Development*, 22(2):152–65, 2008.

Y. M. Elkouby, S. Elias, E. S. Casey, S. A. Blythe, N. Tsabar, P. S. Klein, H. Root, K. J. Liu, and D. Frank. Mesodermal wnt signaling organizes the neural plate via meis3. *Development*, 137(9):1531–41, 2010.

D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *Ieee Transactions on Information Theory*, 49(7):1858–1860, 2003.

J. Ericson, P. Rashbass, A. Schedl, S. Brenner-Morton, A. Kawakami, V. van Heyningen, T. M. Jessell, and J. Briscoe. Pax6 controls progenitor cell identity and neuronal fate in response to graded shh signaling. *Cell*, 90(1):169–80, 1997.

J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–25, 2010.

J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9, 2011.

M. J. Evans and M. H. Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–6, 1981.

D. Evseenko, Y. Zhu, K. Schenke-Layland, J. Kuo, B. Latour, S. Ge, J. Scholes, G. Dravid, X. Li, W. R. MacLellan, and G. M. Crooks. Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 107 (31):13742–7, 2010.

M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA translation and stability by micrornas. *Annual Review of Biochemistry*, 79:351–79, 2010.

J. H. Finger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (gxd): 2011 update. *Nucleic Acids Research*, 39(Database issue):D835–41, 2011.

R. A. Fisher. *The genetical theory of natural selection*. The Clarendon press, Oxford,, 1930.

G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, A. M. Craven, H. B. Harlow, E. W. Su, J. E. Onyia, and C. Su. A statistical analysis of the transfac database. *BioSystems*, 81(2):137–54, 2005.

F. Fuks, W. A. Burgers, N. Godin, M. Kasai, and T. Kouzarides. Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. *EMBO Journal*, 20(10):2536–44, 2001.

T. S. Furey. Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nature Reviews: Genetics*, 13(12):840–52, 2012.

M. Garber, N. Yosef, A. Goren, R. Raychowdhury, A. Thielke, M. Guttman, J. Robinson, B. Minie, N. Chevrier, Z. Itzhaki, R. Blecher-Gonen, C. Bornstein, D. Amann-Zalcenstein, A. Weiner, D. Friedrich, J. Meldrim, O. Ram, C. Cheng, A. Gnirke, S. Fisher, N. Friedman, B. Wong, B. E. Bernstein, C. Nusbaum, N. Hacohen, A. Regev, and I. Amit. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular Cell*, 47(5):810–22, 2012.

M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–82, 1987.

M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and

M. Snyder. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

J. H. Gibcus and J. Dekker. The hierarchy of the 3d genome. *Molecular Cell*, 49(5): 773–82, 2013.

C. A. Gifford, M. J. Ziller, H. Gu, C. Trapnell, J. Donaghey, A. Tsankov, A. K. Shalek, D. R. Kelley, A. A. Shishkin, R. Issner, X. Zhang, M. Coyne, J. L. Fostel, L. Holmes, J. Meldrim, M. Guttman, C. Epstein, H. Park, O. Kohlbacher, J. Rinn, A. Gnirke, E. S. Lander, B. E. Bernstein, and A. Meissner. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, 153(5):1149–63, 2013.

S. F. Gilbert. Induction and the origins of developmental genetics. *Developmental Biology (N Y 1985)*, 7:181–206, 1991.

S. F. Gilbert. *Developmental biology*. Sinauer Associates, Inc. Publishers, Sunderland, Mass., 8th edition, 2006.

S. Gluecksohn-Schoenheimer. The development of two tailless mutants in the house mouse. *Genetics*, 23(6):573–584, 1938.

S. Gluecksohn-Schoenheimer. The effect of an early lethal (t(o)) in the house mouse. *Genetics*, 25(4):391–400, 1940.

A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–9, 2009.

C. E. Grant, T. L. Bailey, and W. S. Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–8, 2011.

H. Gu, C. Bock, T. S. Mikkelsen, N. Jager, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander, and A. Meissner. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods*, 7(2):133–6, 2010.

M. G. Guenther, G. M. Frampton, F. Soldner, D. Hockemeyer, M. Mitalipova, R. Jaenisch, and R. A. Young. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell*, 7(2):249–57, 2010.

F. Guillemot, L. C. Lo, J. E. Johnson, A. Auerbach, D. J. Anderson, and A. L. Joyner. Mammalian achaete-scute homolog 1 is required for the early development of olfactory and autonomic neurons. *Cell*, 75(3):463–76, 1993.

J. B. Gurdon, R. A. Laskey, and O. R. Reeves. The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs. *Journal of Embryology and Experimental Morphology*, 34(1):93–112, 1975.

M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, 458(7235):223–7, 2009.

J. A. Hackett and M. A. Surani. Dna methylation dynamics during the mammalian life cycle. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 368(1609):20110328, 2013.

R. C. Hardison and J. Taylor. Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews: Genetics*, 13(7):469–83, 2012.

R. G. Harrison. Some difficulties of the determination problem. *American Naturalist*, 67:306–321, 1933.

D. C. Hay, J. Fletcher, C. Payne, J. D. Terrace, R. C. Gallagher, J. Snoeys, J. R. Black, D. Wojtacha, K. Samuel, Z. Hannoun, A. Pryde, C. Filippi, I. S. Currie, S. J. Forbes, J. A. Ross, P. N. Newsome, and J. P. Iredale. Highly efficient differentiation of hescs to functional hepatic endoderm requires activina and wnt3a signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 105(34):12301–6, 2008.

N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–8, 2007.

N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–12, 2009.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, 2009.

D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–47, 2013.

R. Holliday and J. E. Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–32, 1975.

J. Holmberg and T. Perlmann. Maintaining differentiated cellular identity. *Nature Reviews: Genetics*, 13(6):429–39, 2012.

G. C. Hon, N. Rajagopal, Y. Shen, D. F. McCleary, F. Yue, M. D. Dang, and B. Ren. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, 45(10):1198–206, 2013.

P. P. Howley and R. Gibberd. Using hierarchical models to analyse clinical indicators: a comparison of the gamma-poisson and beta-binomial models. *International Journal for Quality in Health Care*, 15(4):319–329, 2003.

S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters*, 94 (12), 2005.

W. Huang, N. Lu, H. Eberspaecher, and B. De Crombrugghe. A new long form of c-maf cooperates with sox9 to activate the type ii collagen gene. *Journal of Biological Chemistry*, 277(52):50668–75, 2002.

F. Ille and L. Sommer. Wnt signaling: multiple functions in neural development. *Cellular and Molecular Life Sciences*, 62(10):1100–8, 2005.

R. S. Illingworth, U. Gruenewald-Schneider, S. Webb, A. R. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews, and A. P. Bird. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, 6 (9):e1001134, 2010.

R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, and A. P. Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–86, 2009.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–56, 1961.

S. Jafari, L. Alkhori, A. Schleiffer, A. Brochtrup, T. Hummel, and M. Alenius. Combinatorial activation and repression by seven transcription factors specify drosophila odorant receptor expression. *PLoS Biology*, 10(3):e1001280, 2012.

H. Ji, L. I. Ehrlich, J. Seita, P. Murakami, A. Doi, P. Lindau, H. Lee, M. J. Aryee, R. A. Irizarry, K. Kim, D. J. Rossi, M. A. Inlay, T. Serwold, H. Karsunky, L. Ho, G. Q. Daley, I. L. Weissman, and A. P. Feinberg. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338–42, 2010.

F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C. A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–4, 2013.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, 2013.

P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews: Genetics*, 13(7):484–92, 2012.

H. A. Juraver-Geslin, J. J. Ausseil, M. Wassef, and B. C. Durand. Barhl2 limits growth of the diencephalic primordium through caspase3 inhibition of beta-catenin activation. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (6):2288–93, 2011.

M. Kankainen and A. Loytynoja. Matlign: a motif clustering, comparison and matching tool. *BMC Bioinformatics*, 8:189, 2007.

T. Kawaguchi, Y. Sumida, A. Umemura, K. Matsuo, M. Takahashi, T. Takamura, K. Yasui, T. Saibara, E. Hashimoto, M. Kawanaka, S. Watanabe, S. Kawata, Y. Imai, M. Kokubo, T. Shima, H. Park, H. Tanaka, K. Tajima, R. Yamada, F. Matsuda, and D. Japan Study Group of Nonalcoholic Fatty Liver. Genetic polymorphisms of the human pnpla3 gene are strongly associated with severity of non-alcoholic fatty liver disease in japanese. *PLoS One*, 7(6):e38322, 2012.

T. J. King and R. Briggs. Serial transplantation of embryonic nuclei. *Cold Spring Harbor Symposia on Quantitative Biology*, 21:271–290, 1956.

A. Kirkeby, S. Grealish, D. A. Wolf, J. Nelander, J. Wood, M. Lundblad, O. Lindvall, and M. Parmar. Generation of regionally specified neural progenitors and functional neurons from human embryonic stem cells under defined conditions. *Cell Rep*, 1(6): 703–14, 2012.

R. P. Koche, Z. D. Smith, M. Adli, H. Gu, M. Ku, A. Gnirke, B. E. Bernstein, and A. Meissner. Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell*, 8(1):96–105, 2011.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

S. Kriks, J. W. Shim, J. Piao, Y. M. Ganat, D. R. Wakeman, Z. Xie, L. Carrillo-Reid, G. Auyeung, C. Antonacci, A. Buch, L. Yang, M. F. Beal, D. J. Surmeier, J. H. Kordower, V. Tabar, and L. Studer. Dopamine neurons derived from human es cells efficiently engraft in animal models of parkinson's disease. *Nature*, 480(7378):547–51, 2011.

F. Krueger, B. Kreck, A. Franke, and S. R. Andrews. Dna methylome analysis using short bisulfite sequencing data. *Nat Methods*, 9(2):145–51, 2012.

G. Kunarso, N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y. S. Chan, H. H. Ng, and G. Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, 42(7):631–4, 2010.

L. Kuras, T. Borggrefe, and R. D. Kornberg. Association of the mediator complex with enhancers of active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):13887–91, 2003.

P. W. Laird. Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews: Genetics*, 11(3):191–203, 2010.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22(9):1813–31, 2012.

L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring, and C. L. Wei. Dynamic changes in the human methylome during differentiation. *Genome Research*, 20(3):320–31, 2010.

C. K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36(8):900–5, 2004.

G. Lee, S. M. Chambers, M. J. Tomishima, and L. Studer. Derivation of neural crest cells from human pluripotent stem cells. *Nature Protocols*, 5(4):688–701, 2010.

H. Lee and M. C. Schatz. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16):2097–105, 2012.

E. Li. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews: Genetics*, 3(9):662–73, 2002.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–8, 2008.

Q. H. Li, J. B. Brown, H. Y. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, 2011.

Z. Li, P. Gadue, K. Chen, Y. Jiao, G. Tuteja, J. Schug, W. Li, and K. H. Kaestner. Foxa2 and h2a.z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, 151(7):1608–16, 2012.

F. Lienert, C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schubeler. Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genetics*, 43(11):1091–7, 2011.

R. Lister and J. R. Ecker. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Research*, 19(6):959–66, 2009.

R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–36, 2008.

R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462 (7271):315–22, 2009.

R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, 2011.

R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, 2013.

H. F. Lodish. *Molecular cell biology*. W.H. Freeman, New York, 6th edition, 2008.

T. Manke, H. G. Roider, and M. Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3):e1000039, 2008.

T. Manke, M. Heinig, and M. Vingron. Quantifying the effect of sequence variation on regulatory interactions. *Human Mutation*, 31(4):477–83, 2010.

R. Margueron and D. Reinberg. The polycomb complex prc2 and its mark in life. *Nature*, 469(7330):343–9, 2011.

A. M. Maroof, S. Keros, J. A. Tyson, S. W. Ying, Y. M. Ganat, F. T. Merkle, B. Liu, A. Goulburn, E. G. Stanley, A. G. Elefanty, H. R. Widmer, K. Eggan, P. A. Goldstein, S. A. Anderson, and L. Studer. Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. *Cell Stem Cell*, 12(5):559–72, 2013.

G. R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 78(12):7634–8, 1981.

M. Martuzzi and P. Elliott. Empirical bayes estimation of small area prevalence of non-rare conditions. *Statistics in Medicine*, 15(17-18):1867–73, 1996.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7:29–59, 2006.

A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–98, 2005.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue): D108–10, 2006.

A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, and J. F. Costello. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–7, 2010.

M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–5, 2012.

C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501, 2010.

A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–70, 2008.

E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. Gc-rich sequence elements recruit prc2 in mammalian es cells. *PLoS Genetics*, 6(12):e1001244, 2010.

T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, 2007.

T. S. Mikkelsen, J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner. Dissecting direct reprogramming through integrative genomic analysis. *Nature*, 454(7200):49–55, 2008.

T. S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen. Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143(1):156–69, 2010.

T. H. Morgan. Sex determination and parthenogenesis in phylloxerans and aphids. *Science*, 29(1):234–237, 1909.

T. H. Morgan. *The theory of the gene.* Yale University Mrs Hepsa Ely Silliman memorial lectures. Yale university press; etc., New Haven,, 1926.

A. Natarajan, G. G. Yardimci, N. C. Sheffield, G. E. Crawford, and U. Ohler. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, 22(9):1711–22, 2012.

K. L. Nazor, G. Altun, C. Lynch, H. Tran, J. V. Harness, I. Slavin, I. Garitaonandia, F. J. Muller, Y. C. Wang, F. S. Boscolo, E. Fakunle, B. Dumevska, S. Lee, H. S. Park, T. Olee, D. D. D'Lima, R. Semechkin, M. M. Parast, V. Galat, A. L. Laslett, U. Schmidt, H. S. Keirstead, J. F. Loring, and L. C. Laurent. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell*, 10(5):620–34, 2012.

N. Negre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A.

Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis, and K. P. White. A cis-regulatory map of the drosophila genome. *Nature*, 471(7339):527–31, 2011.

S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–86, 2012.

H. H. Ng and M. A. Surani. The transcriptional and signalling networks of pluripotency. *Nat Cell Biol*, 13(5):490–6, 2011.

L. J. Ng, S. Wheatley, G. E. Muscat, J. Conway-Campbell, J. Bowles, E. Wright, D. M. Bell, P. P. Tam, K. S. Cheah, and P. Koopman. Sox9 binds DNA, activates transcription, and coexpresses with type ii collagen during chondrogenesis in the mouse. *Developmental Biology*, 183(1):108–21, 1997.

C. T. Ong and V. G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews: Genetics*, 12(4):283–93, 2011.

K. R. Ostler, E. M. Davis, S. L. Payne, B. B. Gosalia, J. Exposito-Cespedes, M. M. Le Beau, and L. A. Godley. Cancer cells express aberrant dnmt3b transcripts encoding truncated proteins. *Oncogene*, 26(38):5553–63, 2007.

M. Ostrom, K. A. Loffler, S. Edfalk, L. Selander, U. Dahl, C. Ricordi, J. Jeon, M. Correa-Medina, J. Diez, and H. Edlund. Retinoic acid promotes the generation of pancreatic endocrine progenitor cells and their further differentiation into beta-cells. *PLoS One*, 3(7):e2841, 2008.

C. O. Pabo and R. T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61:1053–95, 1992.

P. J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews: Genetics*, 10(10):669–80, 2009.

E. Patin, Z. Kutalik, J. Guergnon, S. Bibert, B. Nalpas, E. Jouanguy, M. Munteanu, L. Bousquet, L. Argiro, P. Halfon, A. Boland, B. Mullhaupt, D. Semela, J. F. Dufour, M. H. Heim, D. Moradpour, A. Cerny, R. Malinverni, H. Hirsch, G. Martinetti,

V. Suppiah, G. Stewart, D. R. Booth, J. George, J. L. Casanova, C. Brechot, C. M. Rice, A. H. Talal, I. M. Jacobson, M. Bourliere, I. Theodorou, T. Poynard, F. Negro, S. Pol, P. Y. Bochud, and L. Abel. Genome-wide association study identifies variants associated with progression of liver fibrosis from hcv infection. *Gastroenterology*, 143 (5):1244–52 e1–12, 2012.

L. H. Pevny, S. Sockanathan, M. Placzek, and R. Lovell-Badge. A role for sox1 in neural determination. *Development*, 125(10):1967–78, 1998.

S. L. Pfaff, M. Mendelsohn, C. L. Stewart, T. Edlund, and T. M. Jessell. Requirement for lim homeobox gene isl1 in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation. *Cell*, 84(2):309–20, 1996.

T. Pham-Gia and N. Turkkan. Determination of exact sample sizes in the bayesian estimation of the difference of two proportions. *Journal of the Royal Statistical Society Series D-the Statistician*, 52:131–150, 2003.

T. Phamgia, N. Turkkan, and P. Eng. Bayesian-analysis of the difference of 2 proportions. *Communications in Statistics-Theory and Methods*, 22(6):1755–1771, 1993.

W. H. Press. *Numerical recipes : the art of scientific computing.* Cambridge University Press, Cambridge, UK ; New York, 3rd edition, 2007.

L. A. Puto and J. C. Reed. Daxx represses relb target promoters via DNA methyltransferase recruitment and DNA hypermethylation. *Genes and Development*, 22(8): 998–1010, 2008.

A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83, 2011.

O. Ram, A. Goren, I. Amit, N. Shoresh, N. Yosef, J. Ernst, M. Kellis, M. Gymrek, R. Issner, M. Coyne, T. Durham, X. Zhang, J. Donaghey, C. B. Epstein, A. Regev, and B. E. Bernstein. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, 147(7):1628–39, 2011.

S. Ramon y Cajal. *Histology of the nervous system of man and vertebrates.* History of neuroscience. Oxford University Press, New York, 1995.

T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegner, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–52, 2010.

W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–32, 2007.

A. Remenyi, H. R. Scholer, and M. Wilmanns. Combinatorial control of gene expression. *Nature Structural and Molecular Biology*, 11(9):812–5, 2004.

A. D. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics*, 14(1):9–25, 1975.

K. D. Robertson. Dna methylation and human disease. *Nature Reviews: Genetics*, 6(8): 597–610, 2005.

K. D. Robertson, E. Uzvolgyi, G. Liang, C. Talmadge, J. Sumegi, F. A. Gonzales, and P. A. Jones. The human DNA methyltransferases (dnmts) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic Acids Research*, 27(11):2291–8, 1999.

S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger,

S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, and M. Kellis. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330 (6012):1787–97, 2010.

A. J. Ruthenburg, H. Li, D. J. Patel, and C. D. Allis. Multivalent engagement of chromatin modifications by linked binding modules. *Nat Rev Mol Cell Biol*, 8(12): 983–94, 2007.

H. P. Saluz, J. Jiricny, and J. P. Jost. Genomic sequencing reveals a positive correlation between the kinetics of strand-specific DNA demethylation of the overlapping estradiol/glucocorticoid-receptor binding sites and the rate of avian vitellogenin mRNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 83(19):7167–71, 1986.

M. Sander, A. Neubuser, J. Kalamaras, H. C. Ee, G. R. Martin, and M. S. German. Genetic analysis reveals that pax6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes and Development*, 11(13):1662–73, 1997.

C. Schmidl, M. Klug, T. J. Boeld, R. Andreesen, P. Hoffmann, M. Edinger, and M. Rehli. Lineage-specific DNA methylation in t cells correlates with histone methylation and enhancer activity. *Genome Research*, 19(7):1165–74, 2009.

C. J. Schoenherr and D. J. Anderson. The neuron-restrictive silencer factor (nrsf): a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202):1360–3, 1995.

S. Seisenberger, S. Andrews, F. Krueger, J. Arand, J. Walter, F. Santos, C. Popp, B. Thienpont, W. Dean, and W. Reik. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular Cell*, 48(6):849–62, 2012.

A. Serafini. *The epic history of biology*. Plenum, New York, 1993.

A. A. Serandour, S. Avner, F. Percevault, F. Demay, M. Bizot, C. Lucchetti-Miganeh, F. Barloy-Hubler, M. Brown, M. Lupien, R. Metivier, G. Salbert, and J. Eeckhoute.

Epigenetic switch involved in activation of pioneer factor foxa1-dependent enhancers. *Genome Research*, 21(4):555–65, 2011.

J. C. Shen, r. Rideout, W. M., and P. A. Jones. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Research*, 22(6):972–6, 1994.

R. I. Sherwood, T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. van Hoff, V. Karun, T. Jaakkola, and D. K. Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNAse profile magnitude and shape. *Nature Biotechnology*, 32(2):171–8, 2014.

M. Shogren-Knaak, H. Ishii, J. M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson. Histone h4-k16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–7, 2006.

M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, and R. S. Mann. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, 147(6):1270–82, 2011.

S. T. Smale. Pioneer factors in embryonic stem cells and differentiation. *Current Opinion in Genetics and Development*, 20(5):519–26, 2010.

A. G. Smith, J. K. Heath, D. D. Donaldson, G. G. Wong, J. Moreau, M. Stahl, and D. Rogers. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature*, 336(6200):688–90, 1988.

E. Smith and A. Shilatifard. The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Molecular Cell*, 40(5):689–701, 2010.

Z. D. Smith and A. Meissner. Dna methylation: roles in mammalian development. *Nature Reviews: Genetics*, 14(3):204–20, 2013.

Z. D. Smith, H. Gu, C. Bock, A. Gnirke, and A. Meissner. High-throughput bisulfite sequencing in mammalian genomes. *Methods*, 48(3):226–32, 2009.

Z. D. Smith, M. M. Chan, T. S. Mikkelsen, H. Gu, A. Gnirke, A. Regev, and A. Meissner. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, 484(7394):339–44, 2012.

H. Spemann and H. Mangold. Induction of embryonic primordia by implantation of organizers from a different species (reprinted from archiv mikroskopische anatomie entwicklungsmechanik, vol 100, pg 599-638, 1924). *International Journal of Developmental Biology*, 45(1):13–38, 2001.

F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews: Genetics*, 13(9):613–26, 2012.

M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480 (7378):490–5, 2011.

B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–5, 2000.

M. M. Suzuki and A. Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews: Genetics*, 9(6):465–76, 2008.

P. C. Taberlay, T. K. Kelly, C. C. Liu, J. S. You, D. D. De Carvalho, T. B. Miranda, X. J. Zhou, G. Liang, and P. A. Jones. Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell*, 147(6):1283–94, 2011.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76, 2006.

K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–72, 2007.

P. H. Tate and A. P. Bird. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics and Development*, 3(2):226–31, 1993.

R. C. Team. R: A language and environment for statistical computing. 2012.

S. Temple. Stem cell plasticity–building the brain of our dreams. *Nature Reviews: Neuroscience*, 2(7):513–20, 2001.

A. K. Teo, S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang, Z. Chng, E. J. Robertson, N. R. Dunn, and L. Vallier. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes and Development*, 25(3):238–50, 2011.

P. J. Tesar, J. G. Chenoweth, F. A. Brook, T. J. Davies, E. P. Evans, D. L. Mack, R. L. Gardner, and R. D. McKay. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, 448(7150):196–9, 2007.

J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391):1145–7, 1998.

H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings Bioinformatics*, 14(2):178–92, 2013.

R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–11, 2009.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols*, 7(3): 562–78, 2012.

C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature Biotechnology*, 31(1):46–53, 2013.

J. van Arensbergen, J. Garcia-Hurtado, I. Moran, M. A. Maestro, X. Xu, M. Van de Casteele, A. L. Skoudy, M. Palassini, H. Heimberg, and J. Ferrer. Derepression of polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Research*, 20(6):722–32, 2010.

K. E. Varley, J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford, D. M. Absher, B. J. Wold, and R. M. Myers. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research*, 23(3):555–67, 2013.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

T. Vierbuchen, A. Ostermeier, Z. P. Pang, Y. Kokubu, T. C. Sudhof, and M. Wernig. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, 463 (7284):1035–41, 2010.

M. R. Wabl, R. B. Brun, and L. Du Pasquier. Lymphocytes of the toad xenopus laevis have the gene set for promoting tadpole development. *Science*, 190(4221):1310–2, 1975.

C. H. Waddington. *Organisers and genes*. Cambridge biological studies. The University Press, Cambridge Eng., 1940.

J. A. Wamstad, J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer, and B. G. Bruneau. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1):206–20, 2012.

Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008.

Z. Wang, E. Oron, B. Nelson, S. Razis, and N. Ivanova. Distinct lineage specification roles for nanog, oct4, and sox2 in human embryonic stem cells. *Cell Stem Cell*, 10(4): 440–54, 2012.

O. L. Wapinski, T. Vierbuchen, K. Qu, Q. Y. Lee, S. Chanda, D. R. Fuentes, P. G. Giresi, Y. H. Ng, S. Marro, N. F. Neff, D. Drechsel, B. Martynoga, D. S. Castro, A. E. Webb, T. C. Sudhof, A. Brunet, F. Guillemot, H. Y. Chang, and M. Wernig. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell*, 155 (3):621–35, 2013.

M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schubeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–62, 2005.

M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4):457–66, 2007.

H. Wei, G. Tan, Manasi, S. Qiu, G. Kong, P. Yong, C. Koh, T. H. Ooi, S. Y. Lim, P. Wong, S. U. Gan, and W. Shim. One-step derivation of cardiomyocytes and mesenchymal stem cells from human pluripotent stem cells. *Stem Cell Res*, 9(2):87–100, 2012.

A. Weiner, H. V. Chen, C. L. Liu, A. Rahat, A. Klien, L. Soares, M. Gudipati, J. Pfeffner, A. Regev, S. Buratowski, J. A. Pleiss, N. Friedman, and O. J. Rando. Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS Biology*, 10(7):e1001369, 2012.

H. Weintraub, S. J. Tapscott, R. L. Davis, M. J. Thayer, M. A. Adam, A. B. Lassar, and A. D. Miller. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of myod. *Proceedings of the National Academy of Sciences of the United States of America*, 86(14):5434–8, 1989.

O. Wendling, C. Dennefeld, P. Chambon, and M. Mark. Retinoid signaling is essential for patterning the endoderm of the third and fourth pharyngeal arches. *Development*, 127(8):1553–62, 2000.

W. A. Whyte, S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton, C. T. Foster, S. M. Cowley, and R. A. Young. Enhancer decommissioning by lsd1 during embryonic stem cell differentiation. *Nature*, 482(7384):221–5, 2012.

M. Wiench, S. John, S. Baek, T. A. Johnson, M. H. Sung, T. Escobar, C. A. Simmons, K. H. Pearce, S. C. Biddie, P. J. Sabo, R. E. Thurman, J. A. Stamatoyannopoulos, and G. L. Hager. Dna methylation status predicts cell type-specific enhancer activity. *EMBO Journal*, 30(15):3028–39, 2011.

B. Wilczynski, Y. H. Liu, Z. X. Yeo, and E. E. Furlong. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Computational Biology*, 8(12):e1002798, 2012.

I. Wilmut, A. E. Schnieke, J. McWhir, A. J. Kind, and K. H. Campbell. Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385(6619):810–3, 1997.

J. Wysocka, T. Swigut, H. Xiao, T. A. Milne, S. Y. Kwon, J. Landry, M. Kauer, A. J. Tackett, B. T. Chait, P. Badenhorst, C. Wu, and C. D. Allis. A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442 (7098):86–90, 2006.

Y. Xi and W. Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC Bioinformatics*, 10:232, 2009.

W. Xie, M. D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J. W. Whitaker, S. Tian, R. D. Hawkins, D. Leung, H. Yang, T. Wang, A. Y. Lee, S. A. Swanson, J. Zhang, Y. Zhu, A. Kim, J. R. Nery, M. A. Urich, S. Kuan, C. A. Yen, S. Klugman,

P. Yu, K. Suknuntha, N. E. Propson, H. Chen, L. E. Edsall, U. Wagner, Y. Li, Z. Ye, A. Kulkarni, Z. Xuan, W. Y. Chung, N. C. Chi, J. E. Antosiewicz-Bourget, I. Slukvin, R. Stewart, M. Q. Zhang, W. Wang, J. A. Thomson, J. R. Ecker, and B. Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–48, 2013.

R. A. Young. Control of the embryonic stem cell state. *Cell*, 144(6):940–54, 2011.

X. Yu, T. R. St Amand, S. Wang, G. Li, Y. Zhang, Y. P. Hu, L. Nguyen, M. S. Qiu, and Y. P. Chen. Differential expression and functional analysis of pitx2 isoforms in regulation of heart looping in the chick. *Development*, 128(6):1005–13, 2001.

K. S. Zaret and J. S. Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes and Development*, 25(21):2227–41, 2011.

J. A. Zhang, A. Mortazavi, B. A. Williams, B. J. Wold, and E. V. Rothenberg. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish t cell identity. *Cell*, 149(2):467–82, 2012.

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

X. Y. Zhao, W. Li, Z. Lv, L. Liu, M. Tong, T. Hai, J. Hao, C. L. Guo, Q. W. Ma, L. Wang, F. Zeng, and Q. Zhou. ips cells produce viable mice through tetraploid complementation. *Nature*, 461(7260):86–90, 2009.

V. W. Zhou, A. Goren, and B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews: Genetics*, 12(1): 7–18, 2011.

J. Zhu, M. Adli, J. Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P. L. De Jager, D. A. Bennett, J. A. Houmard, D. M. Muoio, T. T. Onder, R. Camahort, C. A. Cowan, A. Meissner, C. B. Epstein, N. Shoresh, and B. E. Bernstein. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–54, 2013.

L. J. Zhu, C. Gazin, N. D. Lawson, H. Pages, S. M. Lin, D. S. Lapointe, and M. R. Green. Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data. *BMC Bioinformatics*, 11:237, 2010.

M. J. Ziller, F. Muller, J. Liao, Y. Zhang, H. Gu, C. Bock, P. Boyle, C. B. Epstein, B. E. Bernstein, T. Lengauer, A. Gnirke, and A. Meissner. Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. *PLoS Genetics*, 7 (12):e1002389, 2011.

M. J. Ziller, H. Gu, F. Muller, J. Donaghey, L. T. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke, and A. Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463): 477–81, 2013.

M. J. Ziller, R. Edri, Y. Yaffe, C. Gifford, A. Goren, J. Xing, H. Gu, O. Kohlbacher, A. Gnirke, B. Bernstein, Y. Elkabetz, and A. Meissner. Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Currently in revision at Nature*, 2014.

R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, 2009.

Own contributions

The content of Chapters 3 and 4 has previously been published in Ziller et al. 2013 (Ziller et al., 2013). The design of the analysis strategy as well as all analyses were carried out by M.J. Ziller. All wet lab experiments including next-generation sequencing library production were contributed by the co-authors. The manuscript was written by M.J. Ziller.

The content of Chapter 5 has previously been published in Gifford & Ziller et al. 2013 (Gifford et al., 2013). This project was a joint research effort with equal contributions by Casey A. Gifford (leading wet lab scientist) and Michael J. Ziller (leading computational scientist). While most wet lab experiments including cell culture, FACS and next-generation sequencing library production (ChIP-Seq and RNA-Seq) were carried out by C.A. Gifford, computational analysis and design of analysis strategy was contributed by M.J. Ziller. Differential splicing analysis was contributed by C. Trapnell. M.J. Ziller and C.A. Gifford interpreted the data and wrote the paper together.

An extended and updated version of Chapter 6 is part of Ziller & Edri et al. and is currently *in press* at Nature. The paper and analysis strategy was conceived by M. J. Ziller (MJZ). All analysis and computer code development was carried but MJZ as well as the writing of the manuscript. R. Edri performed cell culture and FACS. ChIP-Seq library generation was performed by A. Goren, RNA-Seq library generation was carried out by C. Gifford and WGBS/RRBS library production was done by H. Gu.