Winter 12-15-2016

Epigenomics of Cell Fate in Development and Disease

Rebecca Faith Lowdon
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics & Genomics

Dissertation Examination Committee:
Ting Wang, Chair
Sarah Elgin
Stephen L. Johnson
Samantha Morris
Nancy Saccone
Tim Schedl

The Epigenomics of Cell Fate in Development and Disease

by

Rebecca Faith Lowdon

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2016
St. Louis, Missouri

Table of Contents

Chapter 5

Evolution of Epigenetic Regulation in Vertebrate Genomes

Chapter 6

Synthesis

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgments

I am grateful to many people for the support and encouragement given me over the past several years. I am thankful for the unwavering support and guidance of my mentor, Ting Wang. He has given me tremendous opportunities, for which I am very thankful. Joining the Wang lab has allowed me to explore science with an exciting new perspective, and much of the credit for that goes to Ting. I am also thankful to my fellow lab members who helped me push my own boundaries as we pursued our research together.

The opportunity to work with various collaborators made this dissertation a very rewarding one. First many thanks are due to Scott Higdon for advice with zebrafish project experimental design and protocols, and Stephen Johnson for his conception of this project. Josh Jang provided invaluable assistance during the final stages of the zebrafish work. The skin project was a collaboration with Jeffrey Chang and Joseph Costello of the University of California – San Francisco. I am grateful to them for collaborating with me early in my graduate career.

I thank my thesis committee members who have provided valuable advice and guidance over the years. I am especially grateful to Tim Schedl, Sally Elgin, and Stephen Johnson for their mentorship over the past four years, and to Nancy Saccone and Samantha Morris for joining my committee.

I am grateful to my many friends in St. Louis and beyond who have supported me over the past five years. They have made a long journey a very enjoyable one.

Above all, I thank my family, whose continuous love and support I treasure immensely.

In addition, I thank the Alvin J. Siteman Cancer Center at Washington University in School of Medicine and Barnes-Jewish Hospital in St. Louis, Mo., for the use of the Siteman Flow Cytometry Core, which provided flow cytometry service as described in Chapter 3. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842.

I thank my funding sources: the NSF Graduate Research Fellowship Program (DGE-1143954) and the Washington University Interface of Psychology, Neuroscience, and Genetics training program (5T32GM081739).

Rebecca Faith Lowdon

Washington University in St. Louis

December 2016

Dedicated to Mom and Dad.

Abstract of the Dissertation

The Epigenomics of Cell Fate in Development and Disease

by

Rebecca Faith Lowdon

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics & Genomics

Washington University in St. Louis, 2016

Professor Ting Wang, Chair

Epigenetic features at regulatory elements provide instructive cues for transcriptional regulation during development. However, the particular epigenetic alterations necessary for proper cell fate acquisition and differentiation are not well understood. This dissertation explores the epigenetic dynamics of regulatory elements during development and uses epigenome annotations to document inappropriate transcriptional regulation in disease. First, I summarize my contributions to developing a new algorithm for detecting differential DNA methylation, M&M. I report the application of the M&M algorithm to identify distinct classes of DNA methylation dynamics in surface ectoderm (SE) progenitor cells and SE-derived lineages: epigenome alterations, and differential DNA methylation in particular, that are present in progenitor cells are transmitted to daughter cells and consequently observed in differentiated cells. I exploit this property of DNA methylation to characterize DNA methylation dynamics in surface ectoderm embryonic tissue and SE-derived cells. Next, I use zebrafish to investigate the biological relevance of the classes of DNA methylation dynamics described in the SE context. In zebrafish, I use the pigment cell development system to understand the contribution of DNA methylation to a particular cell fate

choice: melanocyte or iridophore cell fate. Next, I investigate the consequence of somatic mutations in primary liver cancer by utilizing epigenomic annotations of human tissues to distinguish putatively functional mutations from passenger mutations. Here I present support for the hypothesis that transcriptional regulatory instructions for heterologous cell types are co-opted by cancer cells during malignant tumorigenesis. Finally I present a review of the evolution of epigenetic regulation over regulatory elements. Altogether, this dissertation advances our understanding of epigenetic regulation in cell fate decisions by integrating functional genomics with developmental biology and cancer genetics.

Chapter 1

Cell Fate and Epigenetics

The innovation of cellular differentiation has been key to allowing multicellular organisms to exploit new niches. In the context of multicellular organisms, in order for a division of labor to be efficient and advantageous, each cell in the organism must "know" it's role. The process by which a cell acquires its terminal characteristics or fate is known as differentiation. As development commences with the first cell division of a zygote, each subsequent cell division gives rise to daughter cells with increasingly restricted cell fate choices. Cell fate restriction is necessary in order to create an organism with all required functionalities.

How a differentiation program proceeds to produce a variety of cellular phenotypes from a single genotype is a major question for developmental genetics. Some genetic regulatory networks that mediate specific phenotype specification, commitment, determination, and maintenance have been described [1-3]. However, the epigenome is also critical for restricting cell fate choices during development [4]. For example, imprinting in mammals is regulated by correct placement of 5-methylcytosine residues, and incorrect DNA methylation of imprinted alleles causes congenital developmental syndromes [5]. Thus the molecular epigenome is also a mediator of

cell type-specific gene regulation.

Background for this dissertation will begin with a brief review of epigenetics and the molecular epigenome features that are the focus of this dissertation. Next is a discussion of what is known about the genetic and epigenetic control of cell fate determination, and what happens when these mechanisms go awry. Finally, the outline of this dissertation summarizes the outstanding questions that are the focus of chapters 2-4.

1.1 The Molecular Epigenome

Epigenetics refers to heritable changes in gene expression that are not explained by changes in DNA sequence. Conrad Waddington, who coined the term "epigenetic," provided the first such example in the fly *Drosophila melanogaster*. Waddington used an environmental perturbation, in this case, high temperature incubation of pupae, to create a specific wing phenotype (trait). After breeding treated flies that displayed the wing phenotype for several generations, the trait persisted in new generation without the environmental intervention. Thus the environmentally-induced phenotypic variation was assimilated into the fly genome [6].

Since Waddington's time, the term epigenetics has come to refer to not only the phenomenon of inherited gene expression change, but also the (non-DNA) molecular components that influence gene expression. The complex of DNA and proteins that pack DNA into the cell nucleus is referred to as chromatin. The proteins that package DNA are dynamic and can be chemically modified to affect chromatin behavior (see 1.1.2). Similarly, direct chemical modification of DNA bases can alter local biochemistry. Both DNA-bound protein modifications and modifications of DNA bases are considered molecular epigenetic features.

Molecular genetics has enabled scientists to link features of the molecular epigenome to the modulation of gene expression and inheritance of gene expression patterns. Immediately after fertilization (and again later in primordial germ cell development), the early embryo undergoes epigenome reprogramming [5]. Epigenome reprogramming is the process by which the epigenome features of the parental genomes are erased. This process allows for proper epigenome features to be established in the developing organism, which as we will see, is critical for proper development [4]. Imprinting is the phenomenon of parent-of-origin specific gene expression and is also critical for proper development. Specific application of epigenetic features, in particular DNA methylation (see **1.1.1**), ensures proper imprinting is established in the early embryo [7]. The epigenome also modifies gene expression by position effect variegation – gene silencing due to proximity to condensed chromatin (heterochromatin). For example, the *white* gene is silenced in *Drosophila melanogaster* when a chromosomal rearrangement places *white* next to a region of heterochromatin that then spreads to shut down expression of the *white* gene [8]. Thus there are many aspects of gene regulation during development that require proper regulation by the epigenome.

The primary focus of this dissertation is the regulation of gene expression by epigenetic modification of gene regulatory elements (see 1.2). An example of epi-regulation in mouse is emblematic: the A^{vy} allele contains a retrotransposon containing a cryptic promoter inserted 100Mb upstream of the *agouti* gene. In the unmethylated state, the cryptic promoter is active and drives expression of the *agouti* gene aberrantly, resulting in a distinct phenotype of yellow hair. However when methylated, the *agouti* gene is properly expressed and the mice have a wildtype (brown) coat (Morgan 1999). The *agouti* epi-alleles exemplify the potential of the epigenome to influence gene transcription and phenotype.

1.1.1 DNA Methylation

DNA methylation refers to the addition of a methyl group to the 5' carbon of a cytosine residue, forming 5-methylcytosine [9]. During DNA replication, maintenance DNA methyltransferase and the multifunctional Uhrf1 protein detects the hemimethylated state (single strand methylation) and methylates the opposite strand accordingly, while *de novo* methyltransferases are targeted to genome loci by cofactors [10]. Conversely, DNA methylation is removed by oxidation of 5-methylcytosine followed by base excision repair, or is achieved by the passive loss of 5-methylcytosine during DNA replication (via lack of maintenance DNA methyltransferase) [11].

In vertebrate genomes, DNA methylation is predominately found at CG dinucleotides (CpG residues) [12,13]. Vertebrate genomes are ubiquitously DNA methylated, although methylation level depends on the CpG density of a given DNA fragment [14]. Very CpG dense regions (termed CpG islands) tend to remain unmethylated [10], while DNA methylation is enriched in vertebrate genomes across repetitive DNA fragments and coding exons [13].

The function of DNA methylation is very context-dependent. DNA methylation at promoters tends to repress gene transcription [15-18], and methylation of CpG island promoters accumulates as development progresses, shutting down inappropriate genes [19,20]. Similarly, DNA methylation can enable binding of DNA-binding factors that are sensitive to 5-methylcytosine: MeCP1 is a transcriptional repressor that binds methylated DNA [21]. Conversely, DNA methylation may inhibit binding of transcriptional activators [22]. In intragenic regions, DNA methylation has been shown to regulate alternative promoters [23] as well as alternative splicing [24]. Finally, DNA methylation is enriched over transposable elements across metazoan phyla, repressing transposable element mobilization and mutagenesis

[13]. Intergenic regions with variable DNA methylation are indicative of enhancers [25-27] and will be the focus of much of this dissertation (see **1.2.2**).

1.1.2 Histone Post-translational Modifications

The 6 billion base pairs of DNA in a human cell are carefully packaged in the nucleus in such a manner as to be compact and also make available for transcription the necessary genes for a given cell to carry out its biological functions. DNA is packaged around proteins called nucleosomes, and the DNA-nucleosome complex is referred to as chromatin. Eight histone subunits comprise the nucleosome, which is wrapped by 146 base pairs of DNA. Nucleosomes are subsequently packaged into higher order structures to achieve compaction of DNA.

However, chemical modifications to the unstructured tails of histone proteins can modify the local biochemistry of DNA, modulating the DNA exposure to transcriptional machinery [28]. The combination of histone post-translational modifications (PTMs) may work independently or in concert to change DNA accessibility and therefore transcription [29]. Acetylation of histones along with methylation of specific residues creates "active" chromatin conformations allowing for transcriptional machinery to access DNA. For example, methylation of histone 3 lysine 4 activates promoters and prompts transcriptional elongation [30]. Conversely, compaction of chromatin by Polycomb group complexes [31] or chromatin remodeling enzymes [32] creates "silent" regions that are not amenable to transcription. Deposition of three methyl groups on histone 3 lysine 27 by lysine methyltransferases represses transcription, either by recruiting Polycomb group 1 repressor proteins [33] and/or by spreading of H3K27me3 modification due to lack of transcription [34]. Thus, the specific and combinatorial chemical modifications of histone tails are critical to gene regulation.

1.2 Genetic and Epigenetic Control of Cell Fate

All cells in an organism contain the same genome sequence; however cells in different tissues have very different features and functions. Such cell-specific features are a result of cell-specific gene expression. Gene expression occurs by a process called transcription. ~200 DNA base pairs around the transcription start site of the gene comprise the gene promoter. For a gene to be expressed, binding of DNA-binding proteins termed transcription factors activates the gene promoter. General transcription factors recruit RNA polymerase, which transcribes DNA by synthesizing a messenger RNA molecule as it processes along the length of the gene. One mechanism of cell-specific gene expression is binding of transcription factors that confer specificity to gene expression. Such cell-specific binding may occur at the gene promoter, or at DNA fragments hundreds to millions of base pairs away.

A regulatory element is any region of DNA that modifies expression of a gene. DNA-binding proteins or epigenetic factors, such as those described above, act on regulatory elements to influence gene transcription. Regulatory elements that activate gene expression include gene promoters and distal regulatory elements (enhancers); insulator elements repress gene expression. This dissertation focuses on the epigenetic features and functions at activating regulatory elements (promoters and enhancers), so the subsequent discussion will focus on these elements.

Regulatory elements can be detected by sequence conservation [35-37] or by experimental methods, such as chromatin-immunoprecipitation followed by high-throughput sequencing (ChIP-seq) [38-41]. Of particular relevance to this dissertation are epigenetic features of enhancers, long-range regulatory elements that can be hundreds to millions of base pairs away from a target gene promoter and act in a distance- and orientation-independent manner to

activate gene transcription [42].

1.2.1 Transcription Factors

Transcription factors (TFs) are critical to developmental stage- and position-specific gene expression. Pioneer factors are TFs that bind DNA and modify local chromatin in order that subsequent TFs can bind. By recruiting nucleosome repositioning enzymes or histone modifying complexes, or protecting enhancer DNA from DNA methyltransferases, such early-binding pioneer factors ensure that an enhancer is available to be bound by downstream trans-activators [43]. For example, AP1 recruits chromatin remodeling enzymes that create an open chromatin environment in mouse mammary epithelial cells, priming AP1-bound elements for quick glucocorticoid receptor binding in response to stimulation by hormone [44]. Similarly, enhancers bound by FOXA1 during neuronal differentiation remain hypomethylated and gain histone 3 lysine 4 methylation (an active histone modification) [45]. TFs binding subsequent to pioneer factors may facilitate expression by creating physical contacts with RNA polymerase and other general TFs at the promoter, for example via the MEDIATOR protein [43]. Thus pioneer factors interact with the epigenome to confer cell specific regulation by chromatin remodeling factors and TFs that are constitutively expressed.

1.2.2 Epigenetic Features of Distal Regulatory Elements

Enhancers confer much of the cell- and tissue-specificity of gene regulation. For example, enhancers embedded the mammalian β-globin locus control region will activate specific globin genes in a developmental stage-specific manner [42]. Similarly, specific expression of the morphogen SHH in the posterior limb bud is required for proper digit patterning. An enhancer 1MB away from the *SHH* transcription start site (TSS) is responsible for limb bud expression of *SHH*, and disruption of this long range enhancer results in preaxial polydactyly [46]. Indeed,

disease-associated genetic variants are highly enriched in enhancers [47,48], suggesting that modulation of enhancer activity can have adverse effects on an individual. Thus enhancers are critical to proper gene regulation during development.

Enhancers display distinctive epigenetic features. Cell type-specific lowly methylated DNA fragments that are intergenic or embedded in transposable elements tend to be enhancers [25,26]. Similarly, gain of 5-hydrosymethylcytosine (an oxidized derivative of 5-methylcytosine) is associated with TF binding and increased gene expression during differentiation of adipose and neuronal cells [49], and this is consistent with the finding that DNA-binding proteins are necessary to create lowly methylated regions in mouse embryonic stem cells and neuronal progenitor cells [25].

The DNA fragment that binds transcription factors is often nucleosome-depleted in order that DNA-binding TFs may bind the regulatory element [50]. DNaseI hypersensitivity analysis reveals nucleosome-depleted regions and DNaseI hypersensitive sites (DHSs) are highly correlated with known regulatory elements [42,51]. While the TF-binding fragment is absent of nucleosomes, enhancers flanked by nucleosomes are subject to cell type-specific histone post-translational modifications [50]. The transcriptional activator p300 co-occurs with activating histone modifications H3K4me1 and H3K27ac, as well as DNaseI hypersensitive sites, and are found at evolutionarily conserved sequences; further, the epigenetic component of these enhancer profiles of enhancers are highly cell type-specific [52]. Thus, while evolutionary conservation is an important and useful measure to detect enhancers, it does not provide information regarding the cell type-specificity that is central to enhancer activity. In addition, enhancers are often marked by short, bidirectional transcription, and analysis of short RNA molecules can identify enhancers in a cell-specific manner [53].

As development proceeds cell fate choices become increasingly restricted, and epigenetic mechanisms coordinate to repress genes for other cell fates [4]. Generally, differentiation is correlated with gain of DNA methylation globally, loss of DNA methylation at specific enhancers, and site-specific changes of histone PTMs [54,55]. The histone modifying Polycomb complex works with *de novo* DNA methylation to restrict developmental potential by shutting down inappropriate gene expression [56]. Meanwhile, histone PTMs and DNA methylation are coordinated to drive position-specific gene expression, as exemplified by the precise developmental expression patterns of the *Hox* cluster of genes in the *Drosophlia melanogaster* embryo [57]. Concomitantly, acquisition of DHSs specifically indicates new regulatory elements in emerging cell types during development [58].

Yet, DHSs active in embryonic stem cells persist in derived lineages, an indication of "epigenetic memory." Indeed, elements that are hypomethylated in adult tissues but lack activating histone modifications represent "vestigial" elements that were active at an earlier developmental time point [59,60]. Such elements may be primed to be reactivated inappropriately in cancer [58], which exhibits disruption of gene regulation and a loss of differentiated cell identity. Elucidating the features and dynamics of regulatory elements during development is a primary aim of this dissertation, which is addressed in chapter 2 and 3.

In summary, while the patterns of epigenome change over development have been characterized, what is not understood are the specific epigenome alterations that are critical for a specific developmental outcome. Cell fate acquisition is the developmental context in question here. In this regard, we have a general understanding that regulatory elements for critical master regulators of cell differentiation need to be activated, but what are the specific changes required in a given cell fate decision? Or is the epigenome regulation of cell fate more robust, and it is not

specific epi-alterations that are required, but cell fate decisions are made in a rheostat model, where only a sufficient number of alternations are required to generate a specific fate outcome.

1.2.3 Loss of Cellular Differentiation in Cancer

Stem cells are unique cells that exhibit self-renewal, can contribute to many different cell lineages, and have a high capacity for proliferation. Cells of the early embryo display these stem cell properties and give rise to hundreds of specialized, differentiated cell types over the course of organismal development. Consecutive steps of cell specification, commitment, and differentiation produce a specific cell fate. Yet the properties of embryonic stem cells are shared by so-called tumor stem cells: cells with a high proliferative capacity, that contribute to any part of the tumor, and that can self-renew [61]. Tumor stem cells can arise from progenitor cells that are not terminally differentiated, but lineage-restricted. Conversion of these progenitor cells to tumor stem cells endows them with the ability to produce cell types they would not normally [62,63], thus indicating these tumor cells have lost their original identity as a lineage-progenitor.

Mechanisms for tumor stem cell instigation include classic cancer mutagenesis targets such as chromosomal rearrangements [64] or mutation of known oncogenes [65]. Noncoding mutation may also play a role in tumorigenesis: in melanoma, regulatory elements for melanocytes (from whish melanoma is derived) were depleted for somatic mutation, while enhancers specific to other cell types were enriched for mutations. Depletion of noncoding mutation in melanocyte regulatory elements indicates a regulatory architecture of de-differentiated cells, which may contribute to tumorigenesis [66], and is consistent with the observation that DHS-marked regulatory elements from embryonic stem cells or other lineages are reactivated in cancer [58]. Understanding the mechanism of these co-opted regulatory elements is one aim of this dissertation, addressed in chapter 4.

1.3 Outline

In Chapter 2, I aimed to determine the relative contribution of tissue environment and developmental origin on shaping the epigenome of skin cells. We and other labs have detected robust tissue- and cell type-specific signatures of DNA methylation in various tissues. However, the skin is unique as a tissue because it is composed of a mixture of cell types derived from different embryonic origins and is exposed to the environment. Given these different epigenetic inputs, we aimed to determine if there is a skin-specific epigenetic signature; alternatively, we hypothesized differentiated skin cells will bear an epigenetic signature common to the cell's specific developmental origin. In this aim, I examined epigenome datasets from three skin cell types and similar datasets from other tissues to answer these questions. I used a novel algorithm developed by our lab for detecting differentially methylated regions to test the hypothesis that developmental origin shapes the epigenomes of differentiated cell types. Bioinformatics analysis revealed the function of these different classes of epigenetic elements and elucidated the gene regulatory network these elements contributed to. I found that the dynamics of DNA methylation over regulatory elements was hierarchical, with a small set of early-demethylating regulatory elements and a larger set of late-demethylating elements.

In chapter 3, I characterized DNA methylation dynamics and regulation of pigment cell development. Based on evidence presented in chapter 2, we hypothesized that early-demethylating elements are responsible for cell fate regulation, while late-demethylating elements are responsible for specific terminal phenotypes. Accordingly, I examined if early-demethylating regions are responsible for cell fate choice by characterizing methylation dynamics in cell fate choice. We used zebrafish pigment cell development to test this question. Melanocytes and iridophores are neural crest-derived pigment cells present in the zebrafish

embryo. Pigment cells arise from a pigment cell progenitor population, a subset of the neural crest, that is present by 24 hours post-fertilization (hpf). We isolated a neural crest GFP-tagged population from two early embryo time points, and melanocytes and iridophores from older embryos. We then used molecular biology techniques to generate whole methylome libraries for each sample type and mRNA-seq libraries for the neural crest samples (We used melanocytes and iridophore mRNA-seq published by the Steve Johnson lab [67]). Analysis of these data is ongoing, and preliminary analyses show expected dynamics over known pigment cell genes. Continuing analysis of methylation dynamics in concert with mRNA-seq data are expected to reveal likely regulators of melanocyte/iridophore cell fate choice.

In chapter 4, I used epigenomic annotation to understand the impact of noncoding cancer somatic mutations in primary liver cancer. The majority of somatic mutations in cancer are noncoding [68], yet the functional implication of noncoding somatic mutations remains elusive. Separating likely functional noncoding mutations from silent "passenger" mutations is a critical goal for cancer genomics. Most noncoding mutations of consequence are hypothesized to occur in regulatory elements, and evidence for this is accumulating in the literature [69-74]. Additionally, we predicted that somatic mutation in cell-type regulators contributes to the deterioration of cell identity, as the cell acquires regulatory programs and phenotypes specific to other cell types. I tested this hypothesis by using epigenome data from ~100 different primary human cell types to functionally annotate noncoding somatic point mutations in liver cancer. Specifically, I integrated noncoding somatic variants from COSMIC and TCGA with Roadmap Epigenome Project data. I found noncoding point mutations occur primarily in cell-type regulatory elements, many of which were not liver-cell regulatory elements. Sequence analysis of the mutated sites

predicted gain or loss of TF binding sites, revealing potential downstream gain- or loss-of-TF binding events consequential for gene expression and cell phenotype.

Chapter 5 is a review of the literature of epigenome evolution in the context of transcriptional regulatory elements. The concluding chapter provides a synthesis of the projects described in this dissertation.

Chapter 2

Regulatory Networks Derived from Epigenomes of Surface Ectoderm-Derived Cells

2.1 Preface: M&M Algorithm for Detecting Differentially Methylated Regions

DNA methylation contributes important information to the genome during development and is responsible for important genome biology, including X chromosome inactivation, repression of transposable elements, and modulation of tissue-specific gene expression [4]. Cell- and tissue-specific differentially methylated regions (DMRs) are increasingly associated with cell- or tissue-specific gene regulation [19,75,76]. However identifying DMRs with confidence remains a challenge for several reasons. Technically it is difficult to isolate pure cell populations in many contexts, and DNA methylome data from heterogeneous cell populations can often be difficult to interpret. Furthermore, it is unclear at this point what magnitude of DNA methylation level change is needed for a biologically relevant effect. To answer this second question, it will help the field to have more robust algorithms that can be applied to a wealth of data across biological samples.

The gold-standard method for DNA methylation analysis is whole genome bisulfite sequencing (WGBS). WGBS is still an expensive experiment, requiring many lanes on a sequencing flow cell for a single sample, making WGBS prohibitive for many labs or for many samples. Conversely, our DNA methylation technologies can accommodate 6-8 samples on a single flow cell. Removing the barrier to library sequencing makes obtaining biological replicates much easier. Consequently, increasing replicates can increase confidence for calling DNA methylation levels and subsequently differentially DNA methylated regions.

Our method relies on gathering two data types: methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq), which queries methylated DNA, and methylation-sensitive restriction enzyme digestion and sequencing (MRE-seq), which interrogates unmethylated CpGs at single base pair resolution [23,77].

The data for a given sample are scaled based on the CpG coverage provided by each data type. Read counts are then treated as mutually independent Poisson random variables to modeled the expected read count values. Next, the algorithm will test the significance of read counts in the two samples using a joint distribution of tag counts. The algorithm effectively tests the hypothesis that

$$H_0: \mu_1 = \mu_2 \textit{ versus } H_1: \mu_1 \neq \mu_2 \qquad \text{(Equation 2.1)}$$

where μ_1 is the methylation level for the given window in sample 1 and μ_2 is the methylation level for the same window in sample 2 [76].

Integrating signal for DNA methylation and un-methylation gave our algorithm increased sensitivity and specificity compared to competing methods (**Figure 2.1**). The M&M algorithm

performed well when identifying cell- or tissue-specific DMRs (**Figure 2.2**). Therefore our analysis of skin cell type methylomes during developed relied heavily on the M&M algorithm.

Other cost-saving solutions include only using MeDIP-seq, or using a targeted approach, for example, Reduced Representation Bisulfite Sequencing (RRBS) assays. RRBS suffers from a lack of genome-wide coverage and focus at promoter regions, and it is therefore often unable to detect DNA methylation changes at distal enhancers, which are important site of differential methylation [14,25,78]. In addition, by integrating MRE-seq with MeDIP-seq data, M&M increases sensitivity for monoallelic or intermediately methylated regions [79].

2.1.1 Author Contributions

Full explanation of the M&M algorithm is presented in the manuscript Bo Zhang[1#], Yan Zhou[#2], Nan Lin[3#], Rebecca F. Lowdon[1#], Chibo Hong[4], Raman P. Nagarajan[4], Jeffrey B. Cheng[5], Daofeng Li[1], Michael Stevens[1], Hyung Joo Lee[1], Xiaoyun Xing[1], Jia Zhou[1], Vasavi Sundaram[1], GiNell Elliot[1], Junchen Gu[1], Philippe Gascard[6], Mahvash Sigaroundinia[6], Thea D. Tlsty[6], Theresa Kadlecek[7], Artheur Weiss[7], Henriette O'Geen[8], Peggy J. Farnham[9], Cécile L. Marie[10], Keith L. Ligon[10,11], Pamela A.F. Madden[12], Angela Tam[13], Richard Moore[13], Martin Hirst[13,14], Marco A. Marra[13], Baozue Zhang[2*], Joseph F. Costello[4*], Ting Wang[1*]. "Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm." *Genome Research*. 2013;23(9):1522-1540. [78]

[1] Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, Saint Louis, MO 63108
[2] Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Chanchun, Jilin Province, P.R. China
[3] Department of Mathematics and Division of Biostatistics, Washington University in St. Louis, Saint Louis, MO 63130
[4] Brain Tumor Research Center, Deparment of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California – San Francisco, CA 94143
[5] Department of Dermatology, University of California – San Francisco, CA 94143
[6] Department of Pathology, Center for Translational Research in the Molecular Genetics of Cancer, University of California – San Francisco 94143
[7] Howard Hughes Medical Institude, Division of Rheumatology, University of California – San Francisco, CA 94143
[8] Genome Center, University of California – Davis, Davis, CA 95616
[9] Department of Biochemistry & Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA
[10] Department of Medical Oncology, Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115
[11] Departments of Pathology, Brigham and Women's Hospital, Children's Hospital Boston, and Harvard Medical School, Boston, MA 02115
[12] Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO 63310
[13] BC Cancer Agency, Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada
[14] Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada
[#] These authors contributed equally to this work.
* Corresponding authors

2.2 Author Contributions

This chapter is adapted from the published manuscript: Rebecca F. Lowdon[1], Bo Zhang[1], Misha Bilenky[2], Thea Mauro[3], Daofeng Li[1], Philippe Gascard[4], Mahvash Sigaroudinia[4], Peggy J. Farnham[5], Boris C. Bastian[3], Thea D. Tlsty[4], Marco A. Marra[2,6], Martin Hirst[2,7], Joseph F. Costello[8], Ting Wang[1], Jeffrey B. Cheng[3*]. "Regulatory Networks Derived from Epigenomes of Surface Ectoderm-Derived Cell Types." *Nature Communications*. 2014;5:5442. [60]

R.F.L., J.F.C., T.W., and J.B.C. designed the study; T.M., B.B., and T.D.T. supervised sample collection and processing; P.G., M.S., and J.B.C. processed cell samples; P.J.F., M.A.M, M.H., J.F.C, and J.B.C. designed and supervised library production and sequencing assays; R.F.L., B.Z., and T.W. performed data analysis; M.B. and D.L. contributed computational tools for data processing and quality control; R.F.L., J.F.C., T.W., and J.B.C. wrote the manuscript.

[1] Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, MO 63108
[2] Canada's Michael Smith Genome Science Centre, BC Cancer Agency, Vancouver, British Columbia, Canada V5Z 1L
[3] Department of Dermatology, University of California – San Francisco, CA 94143
[4] Department of Pathology, Center for Translational Research in the Molecular Genetics of Cancer, University of California – San Francisco 94143
[5] Department of Biochemistry & Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA
[6] University of British Columbia, Department of Medical Genetics, Vancouver, British Columbia, Canada V6H 3N1
[7] Centre for High-Throughput Biology, Department of Microbiology & Immunology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4
[8] Department of Neurological Surgery, Helen Diller Family Comprehensive Cancer Center, University of California – San Francisco, CA 94143
* Corresponding Author

2.3 Abstract

Developmental history shapes the epigenome and biological function of differentiated cells. Epigenomic patterns have been broadly attributed to the three embryonic germ layers. Here we investigate how developmental origin influences epigenomes. We compare key epigenomes of cell types derived from surface ectoderm (SE), including keratinocytes and breast luminal and myoepithelial cells, against neural crest-derived melanocytes and mesoderm-derived dermal fibroblasts to identify SE differentially methylated regions (SE-DMRs). DNA methylomes of neonatal keratinocytes share many more DMRs with adult breast luminal and myoepithelial cells than with melanocytes and fibroblasts from the same neonatal skin. This suggests that SE origin contributes to DNA methylation patterning, while the shared skin tissue environment has limited effect. Hypomethylated SE-DMRs are in proximity to genes with SE relevant functions. In addition, they are enriched for enhancer- and promoter-associated histone modifications in SE-derived cells, and for binding motifs of transcription factors important in keratinocyte and mammary gland biology. Thus, epigenomic analysis of cell types with common developmental origin reveals an epigenetic signature that underlies a shared gene regulatory network.

2.4 Introduction

While epigenetic mechanisms are crucial in establishing and maintaining cell identity, the role of developmental origin and tissue microenvironment in shaping the epigenome is just beginning to be unraveled. Marked epigenomic transitions occur upon directed embryonic stem cell differentiation into the three major embryonic lineages [54,55] and over the course of development [80]. Differentiated cells and tissues have specific DNA hypomethylation signatures, particularly at enhancers [19,78]; however, a subset of hypomethylated enhancers are actually dormant in adult tissues and active only in corresponding fetal tissues suggesting that a DNA methylation memory of fetal origin may be retained in adult cells [59]. Similarly, DNase I-hypersensitive patterns in differentiated cells can reflect embryonic lineage and mark a subset of embryonic enhancers [58]. Tissue microenvironment influences cell identity and morphogenesis [81] and consequently, may affect epigenomes. Accordingly, perturbation of tissue microenvironment is associated with epigenomic alteration [82,83]. These studies suggest that embryonic origin and tissue environment may influence normal cellular epigenomic states and that differentiated cell epigenomes can be utilized to infer epigenomic patterns of precursor embryonic cell populations.

To investigate how developmental origin and tissue environment contribute to cell type-specific epigenetic patterns, we utilize skin as a model system. The three most prevalent skin cell types are each derived from a different embryonic origin (keratinocytes from surface ectoderm, fibroblasts from mesoderm, and melanocytes from neural crest [84]), but exist within a shared tissue environment (**Figure 2.3**). We generate DNA methylation and histone modification profiles for these three skin cell types and compared their epigenomes among the skin cell types and against breast, blood, and brain tissue epigenomes. The three skin cell types share few

regions with common DNA methylation and histone modification states, that were not also present in the other tissue samples. Surface-ectoderm derived skin keratinocytes and breast cells however, share many common differentially DNA methylated regions (SE-DMRs). SE-DMRs are enriched for enhancer- and promoter-associated histone modifications in SE-derived cell types and for binding motifs of relevant transcription factors. Reconstruction of the gene regulatory network connecting these transcription factors and putative target genes with nearby SE-DMRs demarcated epigenetic and regulatory events associated with structural components and signaling pathways in SE-derived cell types. Thus, for surface ectoderm-derived cells, their shared developmental origin influences their epigenomes to a greater extent than tissue environment. Furthermore, a shared gene regulatory network emerged from the SE-DMR signature.

2.5 Results

2.5.1 Skin Cell Type-Specific Differentially Methylated Regions

Fibroblasts, melanocytes, and keratinocytes were individually isolated from each of three neonatal human foreskins and expanded as short-term primary cultures. From these samples, we generated nine high-resolution epigenomes encompassing key histone modifications (H3K4me1, H3K4me3, and H3K27ac) and DNA methylation, along with mRNA and miRNA expression profiles (**Appendix 2: Data 1 and 2**). The effects of aging and environmental exposure were minimized by utilizing neonatal samples. Since each set of three cell types shares a common genome, the effect of genetic variation on epigenetic variability was also minimized.

We identified 12,892 regions encompassing 193,202 CpGs with a DNA methylation status unique to one of the three skin cell types and consistent across all three individuals (**Methods, Figure 2.2a, Figures 2.5-2.7, 2.8a, Appendix 1: Notes 1-3, Table 2.1**). The majority of these

skin cell type-specific DMRs were hypomethylated (**Figure 2.4a**), suggesting potential cell type-specific regulatory activity at these regions [25,26,78]. 40-46% of the DMRs were intergenic and 5-9% were associated with RefSeq annotated gene promoters (**Figure 2.9**); non-CGI promoters were enriched among cell type-specific DMRs (**Appendix 1: Note 4; Table 2.2**). 80-91% of hypomethylated cell type-specific DMRs overlapped with regulatory element-associated histone modifications in the same cell type (**Figure 2.4b**). Accordingly, hypomethylation of cell type-specific DMRs at gene promoters correlated with increased gene expression relative to the other two cell types where the DMR was hypermethylated (**Figure 2.4c, Tables 2.3-2.5**). Gene Ontology (GO) analysis using the GREAT [85] tool on hypomethylated cell type-specific DMRs showed strong enrichment for biological processes relevant to each cell type (e.g. extracellular matrix organization for fibroblasts (*P*-value=9.05E-45) and pigmentation for melanocytes (*P*-value=2.43E-06); **Figure 2.4d; Appendix 2: Data 3**). These data suggest skin cell type-specific DMRs occur primarily at distal enhancers and regulate genes relevant to each cell type.

2.5.2 Skin Cell Tissue-Specific Epigenomic Features

We next examined whether the common tissue environment of the three skin cell types would impose an identifiable skin tissue epigenetic signature. For comparison, we generated complete epigenomes and transcriptomes for a panel of non-skin cell types and tissues (including brain tissue and breast and blood cell types) and identified DMRs shared by all three skin cell types relative to other tissues (**Figure 2.8b, Appendix 1: Note 5**). Of the 28,776 total DMRs detected, only 8 regions shared the same methylation status in skin cell types and the opposite methylation status in all other samples (**Figure 2.10a,b**). Hierarchical clustering based on methylation levels at the 28,776 DMRs reveals that while samples of the same cell type cluster together, the three skin cell types do not (**Figure 2.3c**). These results suggested that skin cell type methylomes did

not share many differences compared to breast, brain, and blood cell methylomes and that skin tissue lacks a specific and substantive DNA methylation signature.

To determine whether skin tissue also lacks a shared histone modification signature, we identified cell type-specific chromatin states from H3K4me1, H3K4me3, and H3k27ac ChIP-seq data for each skin cell type, as well as for breast, brain and blood samples. Among the 259,297 enhancer-associated H3K4me1 peaks and 55,859 promoter-associated H3K4me3 peaks identified in the above samples, only 997 H3K4me1 and 57 H3K4me3 peaks are present in all three skin cell types and absent in the other samples (**Figures 2.11, 2.12**). Only 100 of the 997 exclusively skin-specific H3K4me1 peaks overlapped with H3K27ac peaks (a combination which marks active enhancers) in all three skin cell types (**Figure 2.10d**). While GO enrichment analysis for cell type-specific histone modification patterns showed enrichment for relevant terms, analysis for the few exclusively skin tissue shared histone modification peaks did not reveal any relevant enrichment (**Appendix 2: Data 4**). The minimal DNA methylation and histone modification commonalities that separate skin cell types from other tissues and the lack of functional enrichment for these common shared regions strongly suggest that the shared skin tissue environment does not significantly influence its constituent cell type epigenomes at this developmental stage.

2.5.3 Developmental Origin Influences Epigenomes

In the absence of a strong skin tissue-specific epigenetic signature, we hypothesized that developmental origin is a major determinant of skin cell type epigenetic patterns. We explored this hypothesis by focusing on skin keratinocytes and breast epithelial cells, which are both derived from surface ectoderm [86]. Consistent with their shared developmental origin, neonatal skin keratinocytes clustered with adult breast epithelial cell types based on DNA methylation

values at the DMRs previously identified in skin and non-skin cell pairwise comparisons (**Figure 2.10c**). To specifically define the DNA methylation signature of surface ectoderm, we identified DMRs for each of the surface ectodermal cell types in a pairwise manner compared to neonatal skin melanocytes and fibroblasts, which are derived from other embryonic germ layers (**Figure 2.8c**). There were 1,392 DMRs with the same methylation state exclusively in keratinocyte, breast myoepithelial, and breast luminal epithelial cells relative to the two other cell types, which we inferred to be surface ectoderm-specific DMRs (SE-DMRs) (**Methods, Figure 2.13a**). Therefore, common developmental origin influences surface ectoderm-derived cell epigenomes to a greater extent than does the shared skin tissue environment.

We examined whether SE-DMRs, like cell type-specific DMRs, possessed regulatory potential. The majority (97%) of surface ectoderm DMRs (SE-DMRs) were hypomethylated with 12% located in gene promoters and 40% within intergenic regions (**Figure 2.14a**). Hypomethylated SE-DMRs were enriched for promoter- and enhancer-associated histone modifications in both keratinocytes and breast myoepithelial cells, and for DNase I-hypersensitive sites in keratinocytes (**Figure 2.13b, Figure 2.14b**). Hypomethylated SE-DMRs were also enriched for transcription factor binding motifs including TFAP2 and KLF4 (**Figure 2.13c**); transcription factors that bind to these two motifs function in keratinocyte and mammary epithelium development, differentiation, and/or maintenance of cell fate [87-91]. Genes associated with hypomethylated SE-DMRs were enriched for functions relevant to the biology of these cell types, such as “epidermis development” (*P*-value=4.35e-15) and “mammary gland epithelium development” (*P*-value=2.10e-9) (**Figure 2.13d, Appendix 2: Data 5**). DNA hypomethylation status of genes with hypomethylated SE-DMRs in their promoter regions correlated with increased expression in SE-derived cells relative to non-SE cells (**Figure 2.13e, Table 2.6**).

These annotations suggested that the majority of surface ectoderm-DMRs were at distal enhancer or gene promoter elements and regulate genes important for keratinocyte and mammary gland development. More generally, these results offer a new and deeper level of interrogating the origin and function of adult epigenomes, adding significantly to the recent attribution of epigenome signatures to germinal layers[54,55].

2.5.4 Epigenome-Derived Surface Ectoderm Regulatory Network

Given their regulatory element signatures, overlap with DNase I-hypersensitive sites, and enrichment for relevant transcription factor binding site (TFBS) motifs, we hypothesized that hypomethylated SE-DMRs may be regulatory elements that coordinate expression of genes essential for function of surface ectoderm-derived cells. To test this, we sought to connect these putative regulatory elements to genes in a surface ectoderm gene network. We associated DMRs with nearby putative target genes and queried databases of TF-target genes and gene-gene interactions to construct regulatory relationships among these genes (**Methods**). The result is a highly connected network with a statistically significant number of connections (1458 edges, 278 nodes; *P*-value=1.25e-4; **Methods; Table 2.7**), whose distribution follows a power law (R^2=0.89; **Figure 2.15**).

Strikingly, the transcription factors near the top of the inferred SE network were those whose motifs were enriched in the hypomethylated SE-DMRs (**Figure 2.13c**). This observation, along with the network connectivity data, suggested that TFAP2a, TFAP2c, and KLF4 may regulate many of the downstream genes in this network. To identify biological processes associated with each set of hypomethylated DMRs containing either TFAP2 or KLF4 TFBSs, we performed GREAT analysis [85]. The network was characterized by two partially overlapping major branches (summarized data in **Figure 2.16a, Table 2.8**). The first branch included the

transcription factors TFAP2a and TFAP2c and connected to genes associated with surface ectoderm relevant GO terms, e.g. "hemidesmosome assembly" which is a structural complex critical for epithelial cells [92] and Notch signaling which functions in mammary cell fate commitment [93] and keratinocyte homeostasis [94] (**Figure 2.16b**). The second branch was characterized by KLF4 and associated with mammary gland development and Wnt signaling which influences both breast and keratinocyte cell fate decisions [95,96] (**Figure 2.16c**). Thus, we observed a highly structured set of connections between regulatory elements and putative target genes that underlie and integrate signaling pathways vital for both keratinocyte and mammary gland epithelial cell function.

Surface ectoderm hypomethylated DMRs were located near the TSS of six genes that encode hemidesmosome/epidermal basement membrane zone components, five of which contain the TFAP2 TFBS motif (**Figure 2.16e**). These genes were highly expressed in all surface ectodermal cell types (**Figure 2.16d**). Mutations occur in any one of five of these genes in various forms of the inherited epidermolysis bullosa blistering skin diseases [97,98]. These findings suggest SE-DMRs may coordinately regulate a suite of genes that encode for components of a key structural complex in surface ectoderm-derived cells, that when perturbed leads to a clinically relevant phenotype.

Hypomethylated SE-DMRs containing TFAP2 motifs were also identified near the transcription start site of two genes, *IRF6* and *Stratifin*, that are highly expressed in surface ectoderm-derived cells (**Figure 2.17a-d**). *IRF6* is a transcription factor, known to be regulated by TFAP2a[99], that coordinates keratinocyte and breast epithelium proliferation and differentiation [100,101]. *Stratifin* is a member of the 14-3-3 protein family which functions as an adaptor protein and binds to phosphorylated proteins mediating diverse cellular processes, such as cell cycle control,

apoptosis, and keratinocyte differentiation [102]. Stratifin promoter DNA hypermethylation and expression downregulation is found in both breast and skin cancers [103]. Mutations in *IRF6* or *SFN* lead to similar phenotypes with limb and craniofacial developmental abnormalities and an impaired skin barrier due to defective keratinocyte differentiation [104,105].

A KLF4 motif containing hypomethylated SE-DMR was noted near the mir-200c/141 locus. These two microRNAs promote epithelial cell fate and mir-200c/141 expression is often lost in breast cancers [106]. Our findings of mir-200c/141 surface ectoderm-specific expression and DNA hypomethylation (**Figure 2.17e,f**) are consistent with previously demonstrated epigenetic regulation of this locus [107]. Thus, SE-DMRs may modulate key genes that regulate proliferation, differentiation, and epithelial cell fate maintenance in surface ectoderm-derived cells.

2.5.5 Developmental Dynamics of SE Regulatory Elements

To explore the developmental dynamics of DNA methylation at SE-DMRs, we obtained whole genome bisulfite sequencing data for samples representing early stages in surface ectoderm development: H1 embryonic stem cells (ESCs) and ESCs differentiated to represent an early ectoderm developmental stage [54]. A majority of hypomethylated SE-DMRs were methylated in both early developmental stages, but hypomethylated in keratinocytes and mammary gland epithelia (**Methods, Figure 2.18a**). The few exceptions are transcription factors that are upstream in the regulatory hierarchy. For example, the DMR near the *TFAP2a* promoter was demethylated in ES cells, whereas the DMR in *KLF4* was methylated in ES cells but demethylated in early surface ectoderm differentiated cells. Both genes are most highly expressed in keratinocytes (**Figure 2.18b-e**). The remaining hypomethylated SE-DMRs, many of which putatively regulate genes that are TFAP2a, TFAP2c, or KLF4 targets in the network

analysis, were lowly methylated in differentiated cells. Accordingly, expression of these genes was generally increased in keratinocytes relative to H1 ESCs (**Figure 2.18f**). Additionally, hypomethylated SE-DMRs were highly methylated in fetal brain tissue, which is predominantly neuroectoderm-derived, concordant with their specific assignment to surface ectoderm-derived cells rather than embryonic ectoderm as a whole (**Figure 2.19**).

2.6 Discussion

Analysis of an increasingly diverse collection of epigenomes has revealed tissue- and cell type-specific regulatory elements important for cell fate and development [26,41,108-110]. However, the developmental origins of these epigenomic features have been less explored. Studies utilizing in vitro ESC differentiation systems have uncovered early developmental DNA methylation dynamics that are believed to occur with specification of the embryonic germ layers [54,55]. There is a growing realization that this developmental lineage-specific information is maintained in differentiated cells, as DNA methylation and DNase I hypersensitive site profiles of cell types and tissues cluster by their embryonic germ layer of origin [58,59]. The persistence of a subset of DNA hypomethylated enhancers, which are active in early development but quiescent in adulthood, also suggests a developmental memory is encoded in the epigenome of differentiated cells [59].

Here we present our analysis of the epigenomic features of human skin cell types and their origins. In our experimental design, we used three different skin cell types from the same individual, and identified DNA methylation signatures which are consistent for three individuals across each cell type, minimizing variables that confound many other study designs including genetic background, age, and external environmental exposures. Consistent with findings in other cell types, we found many skin cell type-specific DMRs at distal enhancers, enriched for

association with cell type-relevant genes, and correlated with expression at hypomethylated promoters. Thus we demonstrated that, as expected, the cell types within skin tissue possess many regions with cell type-specific epigenomic patterns.

Next we assessed whether the shared environment within skin tissue imparts common epigenomic features upon its constituent cell types to create a skin tissue-specific signature. To investigate this question, we developed an approach to identify "shared differences" between epigenomes. This approach prioritized specificity and minimized the influence of variation between biological replicates. Thus, shared epigenomic signatures should be robust to sources of variation and attributable to the common biological factor of the grouped samples, for example, the shared tissue environment of skin cell types. Utilizing this approach on the skin cell type epigenomes revealed few shared regions compared to epigenomes of other cell types, suggesting that skin tissue environment had little uniform impact on the epigenomes of its constituent cell types.

Since tissue environment had minimal effect on skin cell type epigenomes, we hypothesized that developmental origin may influence differentiated cell epigenomes and confer features specific to their shared origin. We compared the DNA methylomes of surface ectoderm-derived cells, epidermal keratinocytes and breast luminal and myoepithelial cells, to methylomes of non-SE-derived cells to identify "shared differences." We found that SE-derived cell types share many DMRs when compared to non-SE derived cells and that these DMRs possess regulatory potential. This suggests that the common developmental origin of these surface ectoderm-derived cells impacts their epigenomes, and that this influence is greater than that of tissue environment on keratinocyte methylomes.

To gain better insight into the SE-DMR signature, which we defined indirectly through adult cell epigenomes, we identified target genes putatively regulated by SE-DMRs and then connected these genes based on known interactions (Methods). The resulting SE network predicted both upstream regulators and co-regulated suites of genes. Transcription factors predicted to bind to SE-DMRs (**Figure 2.13c**) were encoded by genes with the highest number of connections in the network (**Figure 2.15**). The presence of SE-DMRs containing TFAP2 TFBSs near the transcription start site of hemidesmosome genes suggests their co-regulation by TFAP2. Additionally, TFAP2 TFBS-containing SE-DMRs are found near the TSSs of the cell cycle regulators *IRF6* and *SFN*. Given the genetic interaction of these two genes in epidermal development [101], TFAP2 may coordinately co-regulate their expression in SE-derived cells. These examples of predicted regulatory relationships illustrate the significant value afforded by incorporating epigenetically-defined regulatory elements into gene networks.

A more direct approach to define epigenomic features that arise from a developmental origin would involve isolation and profiling of actual human embryonic tissues and their derivatives at various time points along a single developmental lineage and comparing their epigenomes and transcriptomes. As this type of experiment is not possible for ethical reasons, we selected cell types arising from a major germ layer derivative, surface ectoderm, to infer for the first time a DNA methylation signature derived from this inaccessible human embryonic cell population. Our approach builds upon previous studies that utilized induced differentiation of ESCs to elucidate DNA methylation patterns of the three main embryonic germ layers [54,55]. Our SE-specific signature findings substantially extend the general concept that epigenomes of differentiated cell types cluster by their embryonic origin [59,108]. We demonstrate that a gene network regulating shared biological processes and functional components can be decoded from

DNA methylation profiles of cell types specifically chosen for their common embryonic origin. Thus, analysis of differentiated cell types with shared developmental origin may be widely applicable for inference of regulatory epigenomic states derived from other inaccessible precursor human cell populations.

2.7 Methods

2.7.1 Cell Type and Tissue Isolation

Fibroblasts, keratinocytes, and melanocytes were isolated from neonatal foreskins obtained from circumcision using standard techniques [111]. Briefly, epidermis was mechanically separated from dermis after overnight incubation at 4 degrees Celsius with dispase solution. The epidermal sheet was incubated with trypsin for 15 minutes at 37 degrees Celsius. The disassociated cells were then incubated in selective growth media. Keratinocytes were grown in keratinocyte growth media (Medium 154CF supplemented with 0.07 mM $CaCl_2$ and Human Keratinocyte Growth Supplement (Life Technologies)). Melanocytes were grown in melanocyte growth media ((Medium 254 with Human Melanocyte Growth Supplement (Life Technologies)). Fibroblasts were extracted from the dermis by mincing and digesting with collagenase. The cell suspension was plated in Medium 106 supplemented with Low Serum Growth Supplement (Life Technologies). All skin cell types were harvested after two passages by snap freezing in liquid Nitrogen.

A pure population of keratinocytes was verified by examination of cell morphology and immunofluoresence staining for keratinocyte markers (cytokeratin (acidic), clone AE1, Life Technologies,18-0153) and lack of staining for melanocyte markers (HMB45+Mart-1+Tyrosinase cocktail, Biocare Medical, CM165 or Mel-5, Covance, Sig-38150). A pure population of melanocytes was verified by examination of cell morphology and

immunofluoresence staining for melanocyte markers and lack of staining for keratinocyte markers. A pure population of fibroblasts was verified by examination of cell morphology and positive staining for vimentin (Sigma, V6630) and lack of staining for keratinocyte and melanocyte markers.

Breast, blood, and fetal brain samples were isolated as previously described [78]. Briefly, for blood cell types, peripheral blood mononuclear cells (PBMCs) were isolated from buffy coat using Histopaque 1077 separation medium (Sigma-Aldrich) according to the manufacturer's protocol. CD4 naïve, CD4 memory, and CD8 naïve cells were isolated from PBMCs using the following isolation kits: EasySep Human Naive CD4+ T Cell Enrichment Kit, EasySep Human Memory CD4+ T Cell Enrichment Kit, and Custom Human Naıve CD8+ T Cell Enrichment Kit (Stemcell Technologies). Pure populations of PBMCs and T cell subsets were confirmed by staining with the following antibodies (anti-CD3 TRI-COLOR (Invitrogen), anti-CD4 PE (BD Biosciences), anti-CD8 FITC (BD Biosciences), anti-CD4 TRI-COLOR (Invitrogen), anti-CD45RO PE (Invitrogen), anti-CD45RA FITC (BD Biosciences), and anti-CD8 TRI-COLOR (Invitrogen)) and FACS analysis.

Briefly, for breast cell types, breast tissue from disease-free premenopausal women was obtained from reduction mammoplasty samples under UCSF CHR protocol #10-01563. Tissue was mechanically and enzymatically dissociated with collagenase and hyaluronidase. Cell suspensions were serially filtered through 150-um and 40-um nylon mesh to obtain epithelial cell enriched clusters (breast cell organoids). To obtain single cell suspensions, organoids were further digested with trypsin and dispase and filtered with a 40-um cell strainer followed by incubation for 60-90 minutes in MEGM medium (Lonza). The resulting cells were stained and sorted by FACS to isolated purified breast myoepithelial and luminal epithelial cells. For

positive selection, a PE-Cy7 labelled anti-CD10 antibody (for myoepithelial cells, BD Biosciences, 341092) and a FITC labelled anti-CD227/MUC1 antibody (for luminal epithelial cells, BD Biosciences, 559774) were used. For negative selection of hematopoietic, endothelial, and leukocyte cells, cells were stained with the following antibodies respectively: anti-CD2, -CD3, CD16, CD64 (BD Biosciences, 555325, 555338, 555405, and 555526); CD31 (Invitrogen, MHCD3115); and CD45, CD140b (BioLegend, 304003 and 323604).

Briefly, for fetal brain samples, brain tissue was obtained post-mortem from fetuses whose death was attributed to environmental/placental etiology, under Partner's Healthcare/Brigham and Women's.

2.7.2 Genomic DNA Isolation

Cells were lysed in extraction buffer (50 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0), 0.5 % SDS, and 1 mg/ml proteinase K) at 55 degrees Celsius for 12-16 hours. The lysed cells were incubated with 40 ug/ml of RNase A for 1 hour at 37 degrees Celsius to remove RNA. DNA was purified by two rounds of phenol/chloroform/isoamyl alcohol extractions and then two rounds of chloroform extractions. DNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ethanol, washed in 70% ethanol, and resuspended in TE.

2.7.3 Methylation-sensitive Restriction Enzyme (MRE)-seq

MRE-seq was performed as in Maunakea, et al. [23] with modifications as detailed below. Five parallel restriction enzyme digestions ((HpaII, Bsh1236I, SsiI(AciI) and Hin6I (Fermentas) and HpyCH4IV (NEB)) were performed, each using 1 ug of DNA per digest for each of the skin cell type samples. Five units of enzyme were initially incubated with DNA for 3 hours and then an additional five units of enzyme was added to the digestion for a total of 6 hours of digestion time. DNA was purified by phenol/chloroform/isoamyl alcohol extraction, followed by

chloroform extraction using phase lock gels. Digested DNA from the different reactions was combined and precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ethanol. The purified DNA was size selected and purified (50-300 bp) by gel electrophoresis and Qiagen MinElute extraction. Library construction was performed as per the Illumina Genomic DNA Sample Prep Kit protocol with the following modifications. During the end repair reaction, T4 DNA polymerase and T4 PNK were excluded and 1 uL of 1:5 diluted Klenow DNA polymerase was utilized. For the adapter ligation reaction, 1 uL of 1:10 diluted PE adapter oligo mix was utilized. 10 uL from the 30 uL of purified adapter ligated DNA was utilized for the PCR enrichment reaction with PCR PE Primers 1.0 and 2.0. PCR products were size selected and purified (170-420 bp) by gel electrophoresis and Qiagen Qiaquick extraction. DNA libraries were checked for quality by Nanodrop (Thermo Scientific) and Agilent DNA Bioanalyzer (Agilent).

Reads were aligned to hg19 using BWA, and pre-processed using methylQA (an unpublished C program; available at http://methylqa.sourceforge.net/). MRE reads were normalized to account for differing enzyme efficiencies, and methylation values were determined by counting reads with CpGs at fragment ends [23]. To enable comparison between MRE-seq data from blood, brain, and breast samples which utilized three restriction enzymes and skin cell types which utilized five restriction enzymes, skin cell type MRE reads that resulted from the use of additional restriction enzymes (Bsh1236I and HpyCH4IV) were removed. Detailed library construction protocols for MRE-seq, MeDIP-seq, ChIP-seq, RNA-seq, and miRNA-seq are publicly available at the NIH Roadmap Epigenomics project website http://www.roadmapepigenomics.org/protocols/type/experimental/.

2.7.4 Methylated DNA Immunoprecipitation (MeDIP)-seq

MeDIP-seq was performed as in Maunakea et al. [23]. 5 ug of genomic DNA was sonicated to a fragment size of ~100-400 bp using a Bioruptor sonicator (Diagenode). End-repair, addition of 3' A bases, and PE adapter ligation with 2 ug of sonicated DNA was performed as per the Illumina Genomic DNA Sample Prep Kit protocol. Adapter-ligated DNA fragments were size selected to 166-366 bp and purified by gel electrophoresis. DNA was heat denatured and then immunoprecipitated with 5-Methylcytidine antibody (Eurogentec) (1 ug of antibody per 1 ug of DNA) in 500 uL of immunoprecipitation buffer (10 uM sodium phosphate, pH 7.0, 140 mM sodium chloride, and 0.05% Triton X-100) overnight at 4 degrees Celsius. Antibody/DNA complexes were isolated by addition of 1 uL of rabbit anti-mouse IgG secondary antibody (2.4 mg/ml, Jackson Immunoresearch) and 100 uL protein A/G agarose beads (Pierce Biotechnology) for 2 hours at 4 degrees C. Beads were washed six times with immunoprecipitation buffer and then DNA was eluted in TE buffer with 0.25% SDS and 0.25 mg/ml of proteinase K for 2 hours at 50 degrees Celsius. DNA was then purified with the Qiagen Qiaquick kit and eluted in 30 uL EB buffer. 10 ul of DNA was utilized for a PCR enrichment reaction with PCR PE Primers 1.0 and 2.0. PCR products were size selected (220-420 bp) and purified by gel electrophoresis. Methylated DNA enrichment was confirmed by PCR on known methylated (SNRPN and MAGEA1 promoters) and unmethylated (a CpG-less sequence on chromosome 15 and GADPH promoter) sequences. DNA libraries were checked for quality by Nanodrop (Thermo Scientific) and Agilent DNA Bioanalyzer (Agilent). Reads were aligned to hg19 using BWA, and pre-processed using methlyQA.

2.7.5 methylCRF

Genome-wide DNA methylation value predictions were made using a conditional random field model that integrates MRE and MeDIP sequencing data for a given sample. The program was

run using default parameters [112], and can be downloaded from http://methylcrf.wustl.edu/. In **Figure 2.18**, methylCRF predicted values were averaged for each DMR.

2.7.6 Differential DNA Methylation Region Analysis

The M&M statistical model [78] which integrates MeDIP-seq and MRE-seq data to identify differentially methylated regions between two samples was implemented with a window size of 500 bp and a q-value (FDR corrected p-value) cutoff = 1e-5. Scripts utilized for pair-wise comparison are shown in **Appendix 1: Note 2**. Adjacent 500 bp DMRs were merged into a single DMR for further analysis unless otherwise noted. The specific pairwise comparisons performed to generate each DMR set are summarized in Supplementary Fig. 4. Additional details and discussion of the DMR calling strategy and false discovery rate for M&M analyses are in **Appendix 1: Notes 1 and 3**. Comprehensive lists of identified skin cell type specific DMRs are available online (http://epigenome.wustl.edu/SE).

2.7.7 Whole Genome Bisulfite Sequencing

1-5 ug of Qubit quantified genomic DNA was utilized for library construction. Unmethylated Lambda DNA (Promega) was added to genomic DNA for a 0.1% final concentration. DNA was fragmented to ~300 bp using Covaris E series shearing. End-repair, addition of 3' A bases, and adapter ligation was performed as per the Illumina PE Genomic DNA Sample Prep Kit protocol except methylated cytosine PE adapters were used. After each of the previous steps, DNA was purified using Ampure XP beads (Agencourt). Bisulfite conversion of purified adapter ligated DNA was performed using the Epitect bisulfite kit (Qiagen) according to manufacturer's instructions. The DNA was amplified by PCR enrichment using Kapa HiFi Hot Start Uracil+Ready (Kapa Biosystems) for 5 cycles with PCR PE primers 1.0 and 2.0. PCR products were purified with the Qiagen Minelute kit and size selected with PAGE gel purification. DNA

libraries were checked for quantity by Qubit (Life Technologies) and quality by Agilent DNA Bioanalyzer (Agilent). Libraries were sequenced using paired-end 100 nt sequencing chemistry on an Illumina HiSeq2000 following manufacturer's protocols (Illumina).

Raw WGBS sequences were examined for quality, sample swap and reagent contamination using custom in house scripts. Sequence reads were directionally aligned to the human genome (GRCh37-lite) using Bismark [113] v. 0.7.6) running Bowtie [114] (v. 0.12.5) allowing up to two mismatches in the 50 bp seed region (using -n 2 -l 50 parameters). Methylation status for each aligned CpG was calculated using Bismark Methylation Extractor (v. 0.7.10) at a minimum of 5x coverage per site in a strand-specific manner (run-time parameters: -p, no_overlap, --comprehensive, --bedGraph, --counts). Overlapping methylation calls from read_1 and read_2 were scored once.

All WGBS data was processed using custom scripts to obtain CpG methylation values. CpG methylation values were filtered such that only CpGs with 10x coverage were subsequently averaged for each DMR in each sample. Lowly methylated regions were called as DMRs for which the average CpG methylation values were ≤ 0.3. Averaged values were plotted as in **Figure 2.18a** using the R package *pheatmaps*.

2.7.8 ChIP-seq

Standard operating procedures for ChIP-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. ChIP-seq library construction involves the following protocols in order: 1) Crosslinking of frozen cell pellet, 2) DNA sonication using Sonic Dismembrator 550, and 3) SLX-PET protocol for Illumina sample prep. Antibodies used in this study were subjected to rigorous quality assessment to meet Reference Epigenome Mapping Quality Standards

(http://www.roadmapepigenomics.org/protocols) including western blot of whole cell extracts, 384 peptide dot blot (Active Motif MODified Histone Peptide Array) and ChIP-seq using control cell pellets (HL60). Antibody vendor, catalog number and lot are provided along with ChIP-seq library construction details as part of the metadata associated with all ChIP-seq datasets and available through GEO and the NCBI epigenomics portals (e.g. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSM669589). Final library distributions were calculated using an Agilent Bioanalyzer and quantified by fluorometric quantification (Qubit, Life Technologies). Libraries were sequenced using single-end 76 nt sequencing chemistry on an Illumina GAiix or HiSeq2000 following manufacturer's protocols (Illumina) as either single or multiplexed libraries using custom index adapters added during library construction.

Sequencing reads were aligned to NCBI GRCh37-lite reference using BWA 0.6.2-r126 with default parameters. MethylQA (an unpublished C program; available at http://methylqa.sourceforge.net/) was used to directionally extend aligned reads to the average insert size of DNA fragments (150 bp) and to generate a bigWig file for downstream visualization. Reads with BWA mapping quality scores < 10 were discarded and reads that aligned to the same genomic coordinate were counted only once.

2.7.9 Differential ChIP-seq Enrichment Analysis

Mapped read density was generated from aligned sequencing reads using methlyQA. Read density overlapping DMRs and their 5 kb upstream/downstream regions were extracted at 50 bp resolution as RPKM values.

The default parameters were used to apply MACS2 [115] to histone modification ChIP-seq data for the identification of peaks at a 1% false discovery rate. A DMR was defined as enriched for

histone signal when at least 60% of the DMR overlapped with histone peaks. Skin cell type-specific histone peaks were identified using the following two criteria: 1) peaks were identified in at least two of three biological replicates of a skin cell type and 2) peaks were not identified in any of the other two skin cell types or other tissue types (brain, breast, and blood). Skin tissue-specific histone peaks were identified using the following three criteria: 1) peaks were identified in at least two of three biological replicates of a skin cell type, 2) peaks were identified in all three skin cell types, and 3) peaks were not identified in any other tissue type (brain, breast, or blood).

2.7.10 Genomic Features

CpG islands, gene bodies, and RefSeq gene annotations (including 5' and 3' UTRs, exons, and introns) were downloaded from the UCSC Genome Browser. Promoters were defined as the 3.5 kb surrounding the TSS (-3 kb/+500 bp) of all RefSeq genes. Intergenic regions were defined as all regions outside RefSeq gene bodies and promoters.

2.7.11 Gene Ontology Enrichment Analysis

Gene Ontology (GO) analyses for biological processes were performed using the GREAT package [85]. Gene regulatory domains were defined by default as the regions spanning 5 kb upstream and 1 kb downstream of the TSS (regardless of other nearby genes). Gene regulatory domains were extended in both directions to the nearest gene's basal domain but no more than a maximum extension in one direction. Only categories that were below a false discovery rate of 0.05 were reported.

2.7.12 Transcription Factor Binding Site Enrichment

Genome sequences were obtained for hypomethylated SE-DMRs from the hg19 human genome assembly. Motif finding analysis was performed using the FIMO tool from the MEME suite and default vertebrate databases [116,117], with a q-value (FDR-corrected p-value) cutoff of 0.04.

Motif enrichment was calculated as the number of motif instances found in the test data compared to the number found genome wide (for hg19), normalized for length, as in equation 2.2.

$$E(motif) = \frac{\frac{n_{DMRs}}{820000}}{\frac{N_{hg19}}{3200000000}} \qquad \text{(Equation 2.2)}$$

where n_{DMRs} = number of a given motif found in the hypomethylated SE-DMRs and N_{hg19} = number of a given motif found in hg19. 820000 = number of base pairs in hypomethylated SE-DMRs; 3200000000 = number of base pairs in the human genome.

2.7.13 Regulatory Network Construction

Regulatory networks were constructed in the following steps. First, genes (nodes) were identified as putative targets of regulatory (hypomethylated) SE-DMRs either by their association with DMRs that fell in the promoter region of RefSeq genes (-3 kb/+500 bp TSS) or by association as distal intergenic regulatory elements. Intergenic DMRs were associated with genes whose TSS fell in a window of +/- 35 kb (This window size is chosen based on literature assessing the average distance of enhancer-promoter associations [109]). The collection of these promoter- and distal enhancer-associated gene lists were then filtered for a gene expression level ≥ 1 RPKM in any of the surface ectoderm-derived cell types.

To obtain interactions between genes in this list, the gene list was used as nodes in the UCSC Interaction Browser [118]. The Interaction Browser queries known databases for connections (links) between a given set of genes (nodes). Four pathway collections (GEA_CLR TF-targets network; UCSC_Superpathway; UCSC_Superpathway_collapsed; CHEA transcription factors) were used to query for interactions between the given genes. For the SE-DMR network, KLF4 was added to the gene list because its motif was enriched in hypomethylated SE-DMRs (Figure 2c) and because it is known to be important for keratinocyte differentiation [119]. *Klf4* does have two hypomethylated SE-DMRs in its second exon, suggesting it is regulated, but the exonic location of the *Klf4* DMRs excluded it from the stringent method for identifying putatively regulated genes, above. Similarly, TFAP2C was added to the gene list because it is known to be important in keratinocyte differentiation [87] and its motif (shared with TFAP2A) was enriched in our motif analysis (**Figure 2.4c**). For the network overview presented in Figure 3a, the transcription factor p63 was added at the top of the network as it integrates both network branches, is a known regulator of the ZNF750–KLF4 transcriptional cascade [120], and interacts genetically with TFAP2a/c [121,122]; however, p63 and its edges are not included in the data or network structural analysis (**Table 2.7**, **Figure 2.15**).

We applied the same method for generating links between a set of 374 random genes to obtain an expected distribution of links given the number of genes in the test network. This resulted in a distribution as described in **Table 2.7** with a mean of 958 and variance of 136.5. By a *t*-test, the number of links in the SE network is statistically significant (*P*-value=1.245e-4).

To assess the scale-free properties of the SE network, we calculated the number of edges assigned to each node and plotted this distribution in **Figure 2.15**.

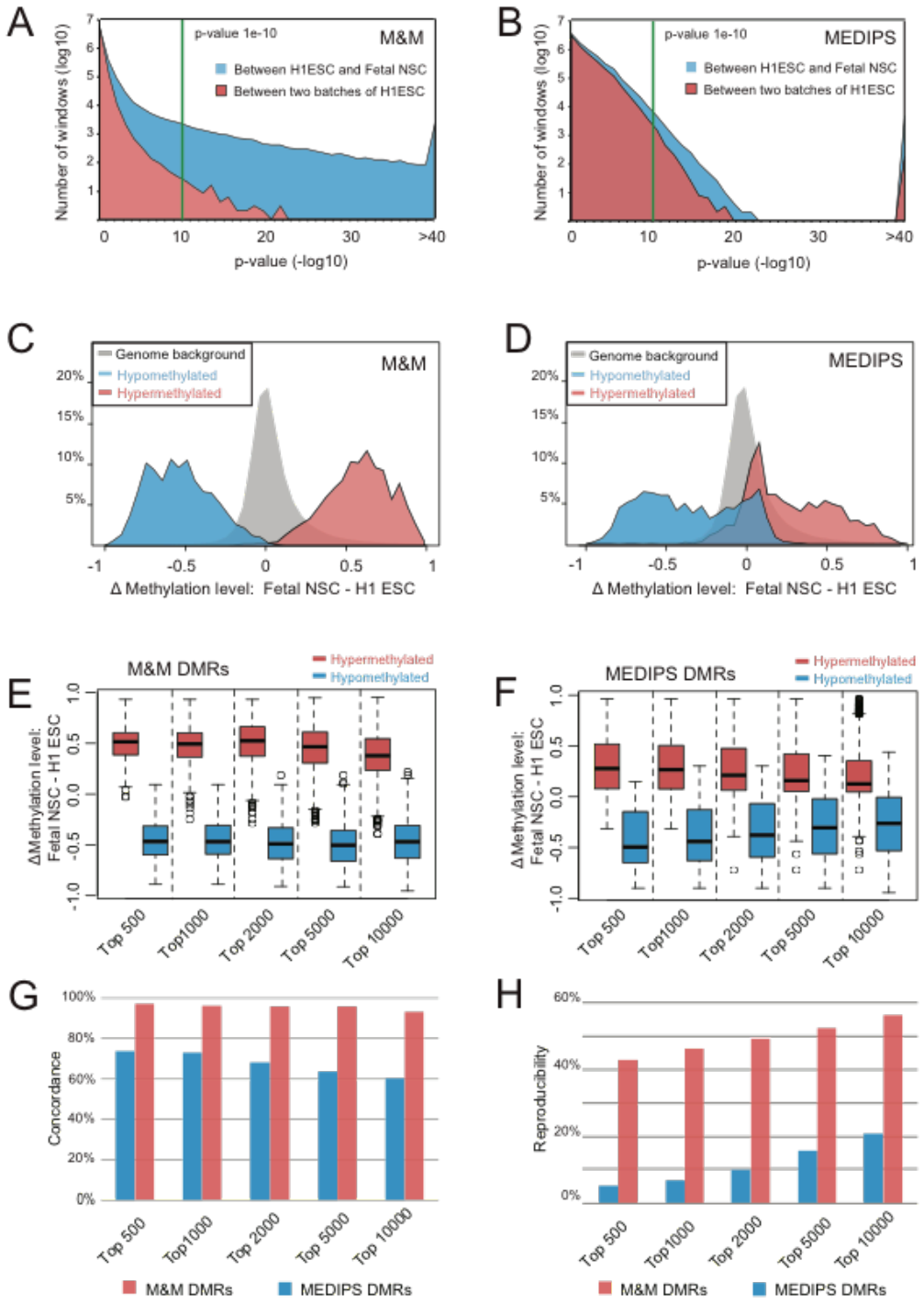


Figure 2.1. Benchmarking the performance of M&M. (A) The distribution of *P*-values generated by M&M when comparing two H1 ESC biological replicates (blue area) and when comparing H1 ESC and fetal neural stem cells (NSC) (red area). At *P*-value cutoff of less than

$1x10^{-10}$ (green line), M&M predicted 70 DMRs between the two H1 samples, and 16,398 DMRs between H1 ESC and fetal NSC. (B) The distribution of *P*-values generated by MEDIPS for the 11.162 DMRs between H1 ESC and fetal NSC. (C) Whole-genome bisulfite sequencing (WGBS) data were used to validate DMRs predicted by M&M between H1 ESC and fetal NSC. DMRs predicted by M&M were ranked according to their *P*-values, then average DNA methylation levels for each of the top 1000 significantly hypermethylated DMRs (red) and the top 1000 significantly hypomethylated DMRs (blue) in fetal NSC were computed using WGBS data from the same two samples. Distribution of the DNA methylation level differences was plotted for hypermethylated DMRs and hypomethylated DMRs separately. The gray area represents the distribution of DNA methylation differences in the whole-genome background, calculated at 500bp window resolution using the same WGBS data sets. (D) Same as (C), except that DMRs were predicted by MEDIPS. (E) DNA methylation differences between H1 ESC and fetal NSC were calculated using WGBS data for individual CpGs within the top 500, 1000, 2000, 5000, and 10,000 hypermethylated and hypomethylated DMRs (predicted by M&M, at varying cutoffs). These values were plotted as a boxplot. (F) Same as (E), except that DMRs were predicted by MEDIPS. (G) Concordance between M&M (red) or MEDIPS (blue) predicted DMRs and differential methylation for these regions calculated from WGBS data. DMRs predicted by M&M and MEDIPS were ranked based on their *P*-values. At different cutoffs, DMRs were determined to be concordant with WGBS data (if differences in WGBS data were greater than 0.1 and were in the correct direction). (H) Reproducibility of DMR predictions in M&M (red) and MEDIPs (blue). DMR discovery was preformed between two cell types from the same individual and repeated in a second individual. DMRs identified in each individual

were ranked according to their *P*-values and intersected between the two individuals. The percentages of overlapping DMRs at different cutoffs were plotted.

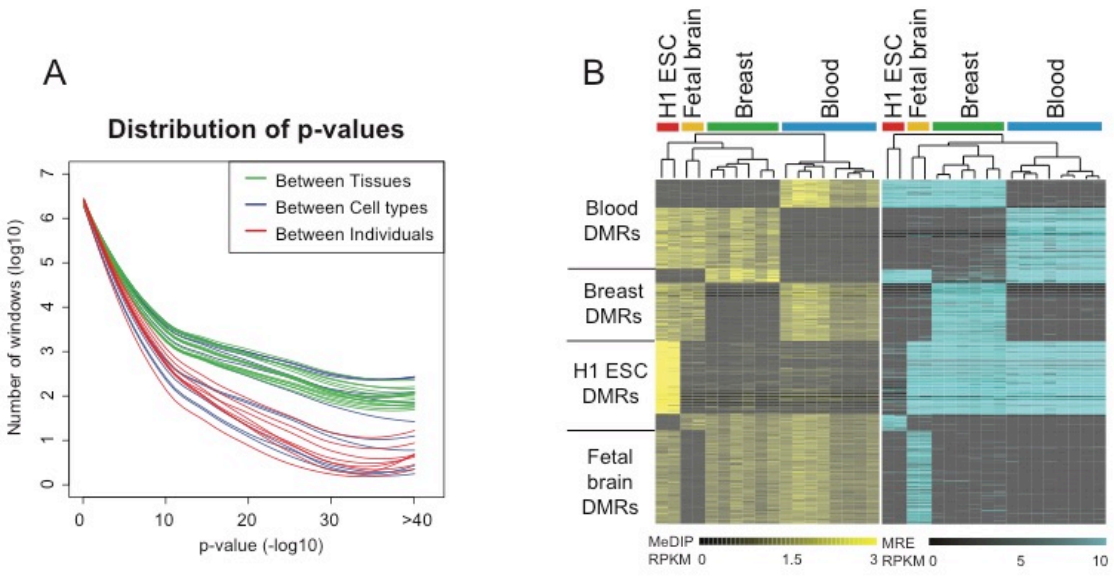


Figure 2.2. M&M analyses of DNA methylation differences across multiple tissue types, cell types, and individuals. (A) *P*-value distributions of M&M predictions between tissue types (green lines), cell types (blue lines), and individuals (red lines). (B) Biclustering analysis of tissue-specific DMRs (*left panel*) based on RPKM values of MeDIP-seq; (*right panel*) based on RPKM values of MRE-seq.

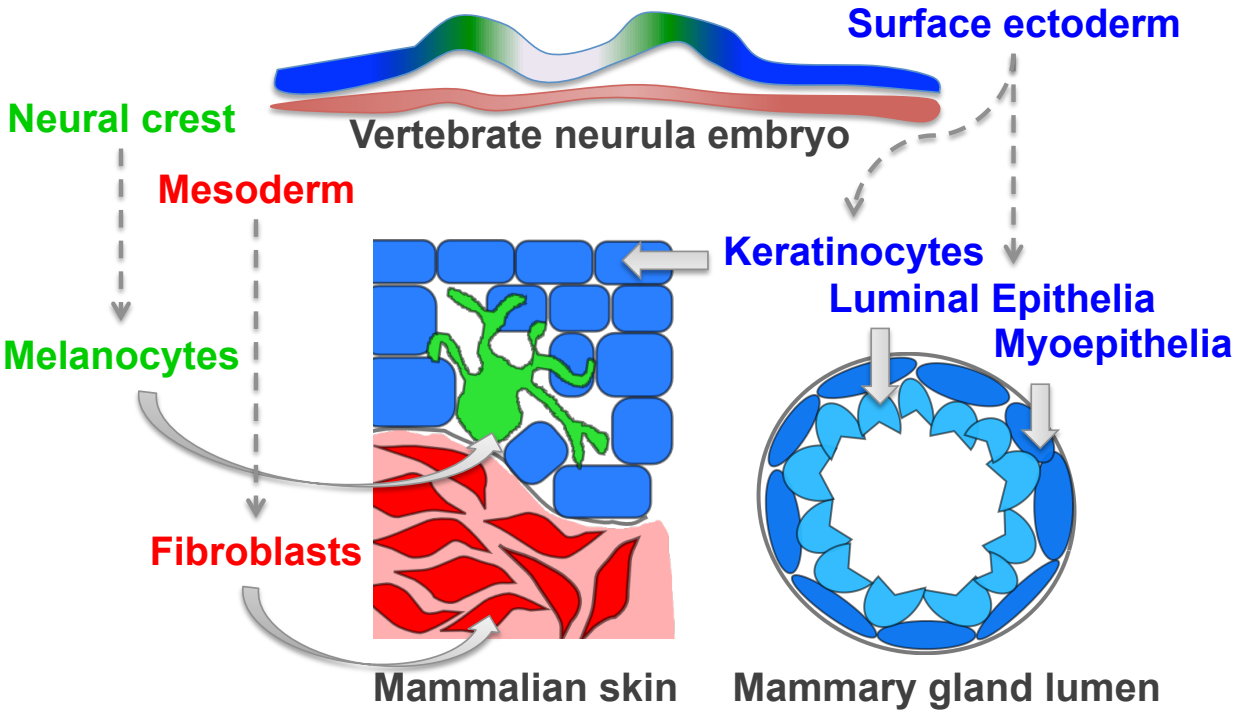


Figure 2.3. Developmental origins of samples. Developmental origins of skin and breast cell types utilized in this study. Embryonic surface ectoderm from the vertebrate neurula stage embryo (blue) gives rise to keratinocytes in the skin and cells of the mammary gland lumen. Embryonic neural crest cells (green) will produce melanocytes that intercalate with epidermal keratinocytes, and skin fibroblasts are derived from embryonic mesoderm (red).

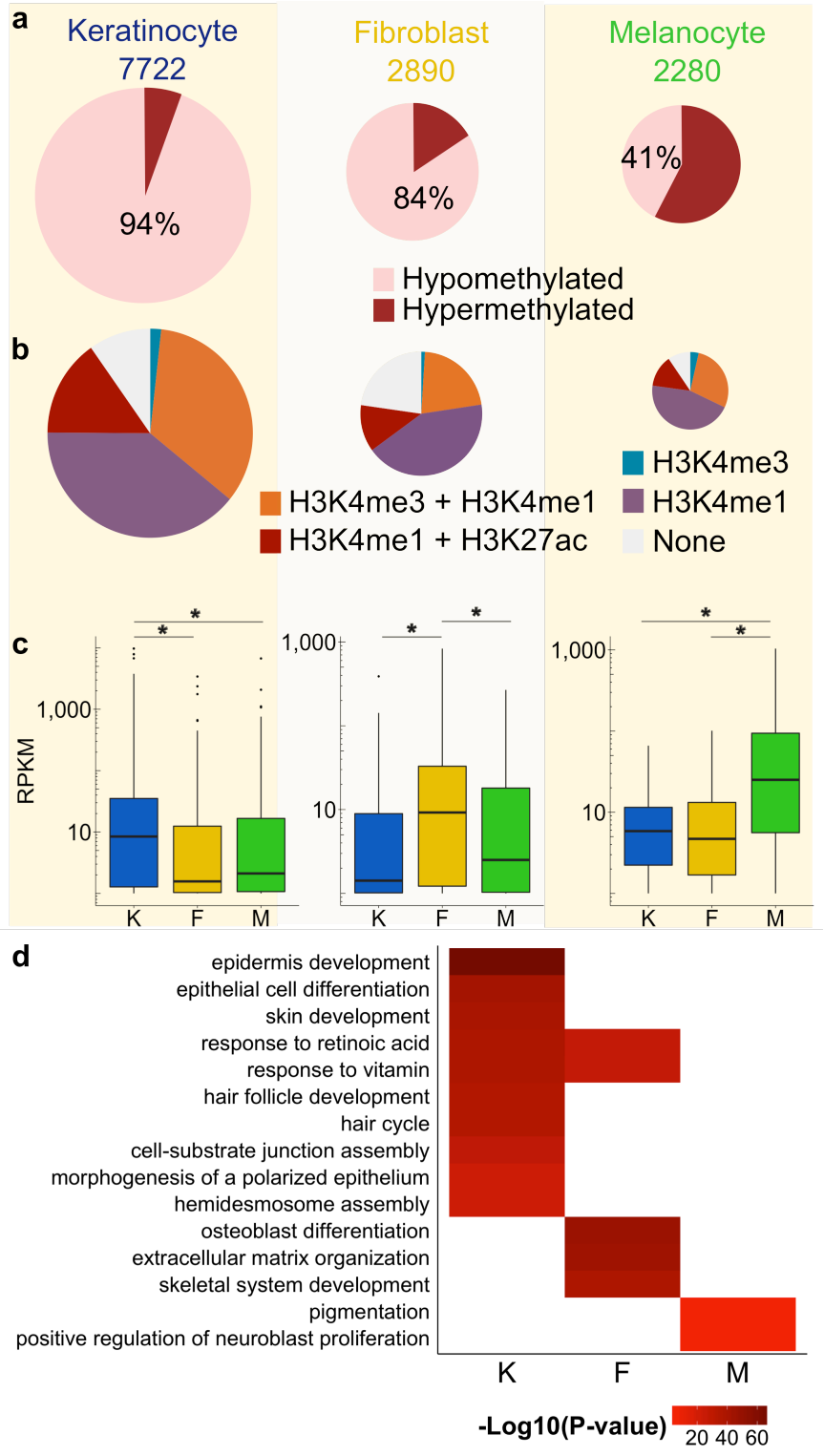


Figure 2.4. Identification and characterization of skin cell type-specific DMRs. (a) Hypomethylation and hypermethylation percentages for each set of skin cell type-specific DMRs

defined by comparison against the other two skin cell types. The total number for each set of cell type-specific DMRs is listed above the pie chart. DMRs are 500bp windows. (b) Histone modification patterns at skin cell type-specific hypomethylated DMRs. (c) Skin cell type RNA expression levels for genes with hypomethylated cell type-specific DMRs in their promoter regions. Each panel depicts expression values for a set of cell type-specific DMR-associated genes. Plotted values are RNA-seq RPKM values over exons, averaged (mean) over three biological replicates. For each boxplot, the middle line indicates the median value, top and bottom box edges are the third and first quartile boundaries respectively. The upper whisker is the highest data value within 1.5 times the interquartile range; the lower whisker indicates the lowest value within 1.5 times the interquartile range. The interquartile range is the distance between the first and third quartiles. Points indicate data beyond whiskers. Logarithmic scale transformations were applied before boxplot statistics were computed. RPKM distributions for a given set of cell type-specific DMR-associated genes in the specified cell type compared to other cell types were statistically significant (Wilcoxon ranked test, paired, * indicates *P*-value < 0.003, Keratinocyte-DMRs $n = 602$, Fibroblast-DMRs $n = 108$, Melanocyte-DMRs $n = 74$; K = keratinocytes, F = fibroblasts, M = melanocytes; Supplementary Tables 3-5). (d) Heat map depicting selected gene ontology terms enriched for keratinocyte, fibroblast, and melanocyte hypomethylated cell type-specific DMRs. K = keratinocytes, F = fibroblasts, M = melanocytes. Color intensity represents the negative log10 transformed p-value of enrichment of a given cell type-specific DMR set for association with the listed gene ontology term. Full datasets are in **Appendix 2: Data 3**.

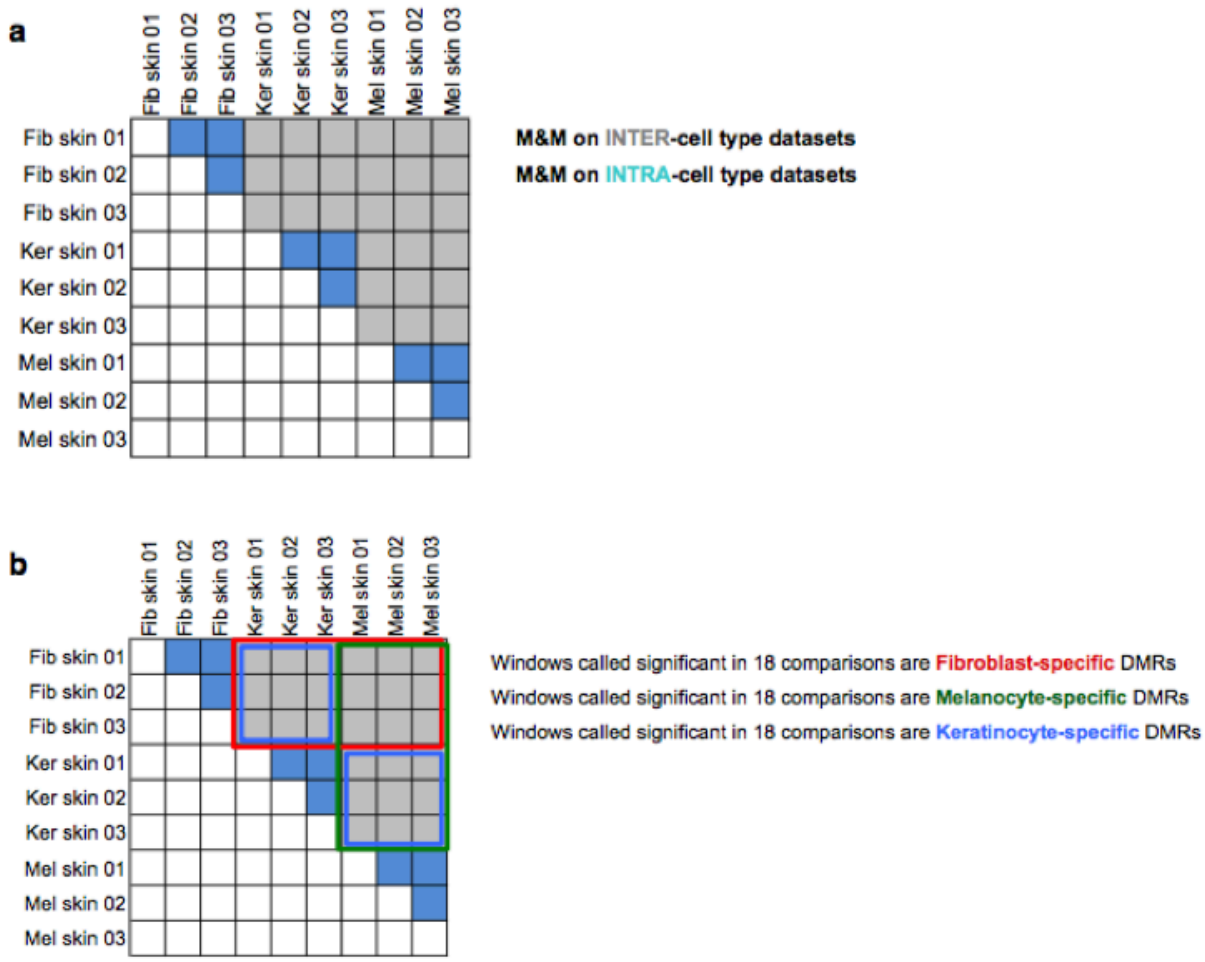


Figure 2.5. Skin cell type-specific DMR calling strategy. (a) Illustration of M&M skin cell type pairwise comparisons. (b) Illustration of intersection strategy for calling skin cell type-specific DMRs. Each gray cell represents one comparison by M&M. DMRs called in the same direction in each of the indicated comparisons (cells within red, green, or blue outlines) were collected as a given cell type-specific DMR set (fibroblast, melanocyte, or keratinocyte, respectively).

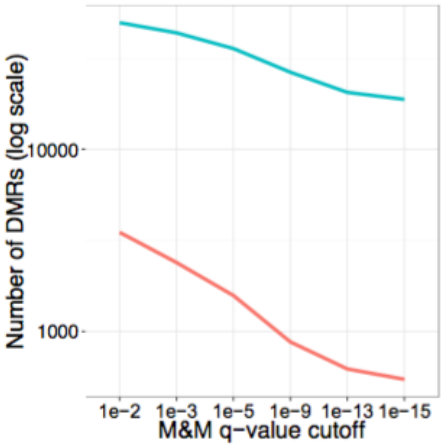


Figure 2.6. Number of DMRs across M&M q-values. Red line = intra-Fibroblast DMRs (Fibroblast 02 vs Fibroblast 03); blue line = inter-cell type DMRs (Fibroblast 03 vs Keratinocyte 03).

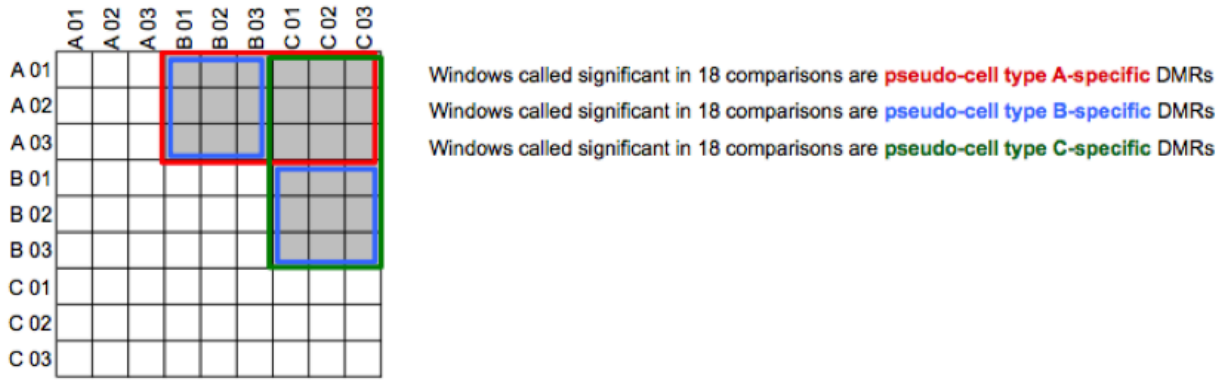


Figure 2.7. Illustration of intersection strategy for identifying pseudo-cell type-specific DMRs. Each gray cell represents one comparison by M&M. DMRs called in the same direction in each of the indicated comparisons (cells within red, blue, or green outlines) were collected as a given pseudo-cell type-specific DMR set (A, B, or C, respectively).

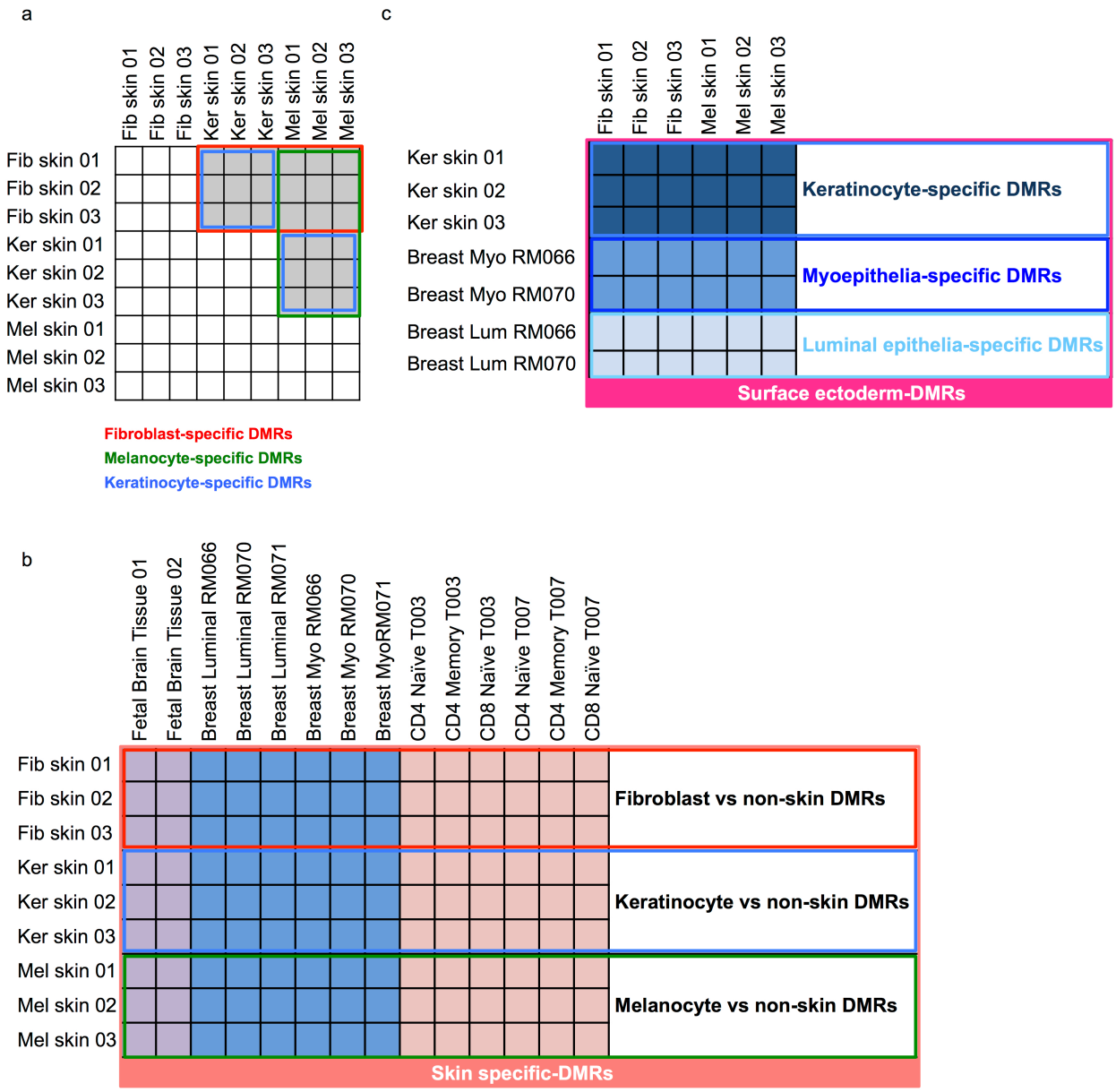


Figure 2.8: Matrices depicting sample comparisons used to identify differentially DNA methylated regions. (a) Matrix depicting pairwise methylome comparisons used to determine skin cell type-specific DMR sets. Each gray cell represents one comparison by M&M (Methods). DMRs called in the same direction in each of the indicated comparisons (cells within red, green, or blue outlines) were collected as a given cell type-specific DMR set (fibroblast, melanocyte, or keratinocyte, respectively). (b) Matrix depicting pairwise methylome comparisons used to

determine skin tissue-specific DMRs. Each cell represents one M&M pairwise comparison. DMRs called in the same direction in all depicted pairwise comparisons (i.e. for each of the 3 skin cell types compared to non-skin cell types) were called “skin tissue-specific DMRs” (of which there were only 8; Figure 2.3a). (c) Matrix depicting pairwise methylome comparisons used to determine surface ectoderm-specific DMRs. Each cell represents one M&M pairwise comparison. DMRs called in the same direction in all depicted pairwise comparisons (i.e. for each of the 3 surface ectoderm cell types) were collected as the surface ectoderm-DMR set.

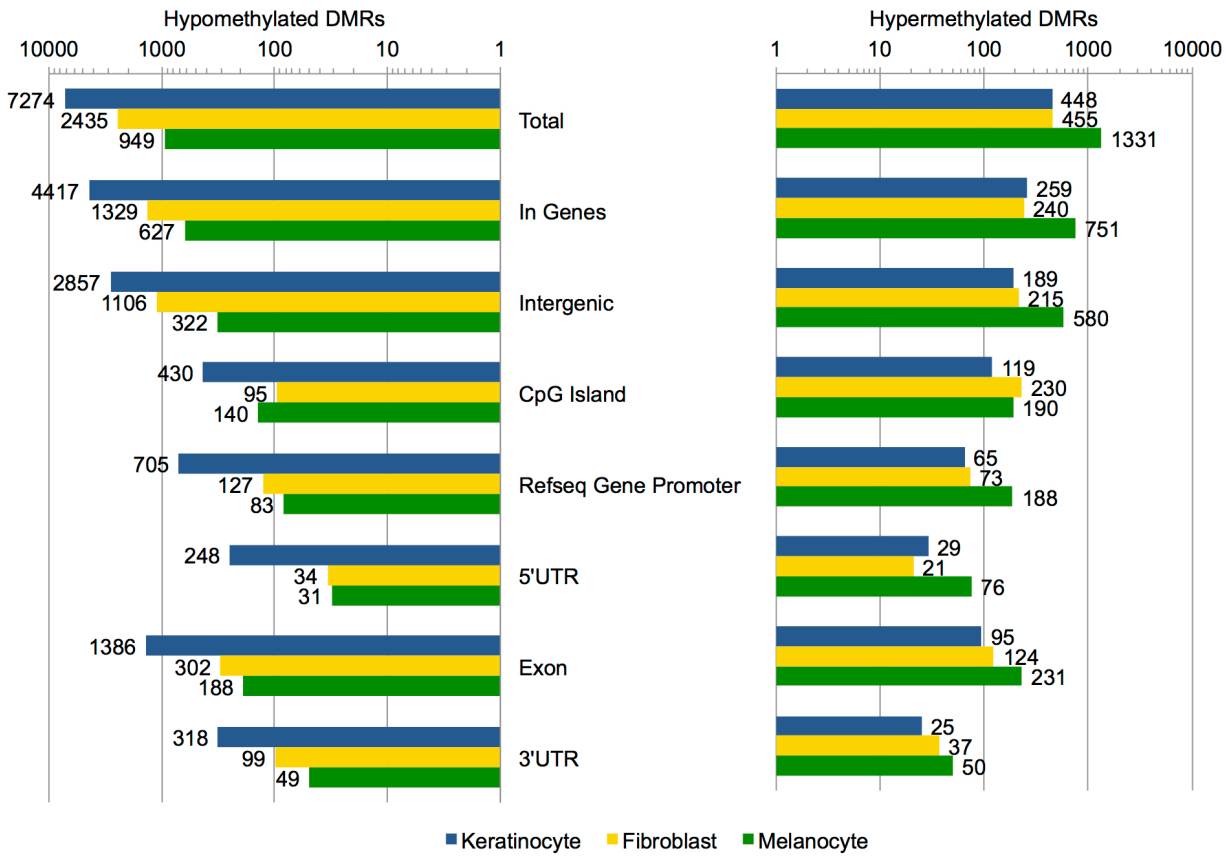


Figure 2.9: Genomic annotation of skin cell type-specific DMRs. Hypomethylated and hypermethylated DMRs plotted independently. DMRs are 500 bp windows. Cell types indicated by bar color. Genomic annotations described in Methods.

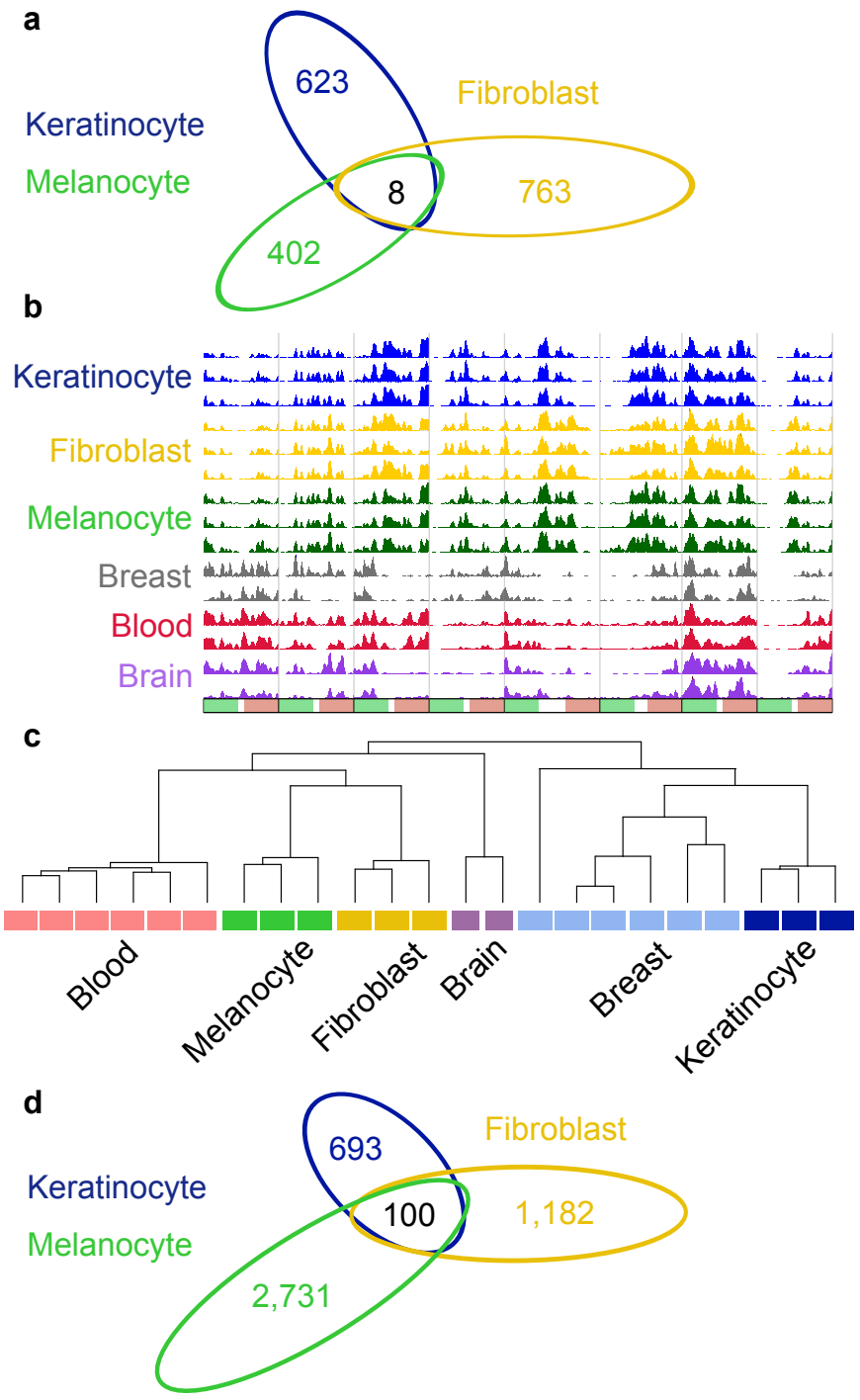


Figure 2.10. Skin-tissue level epigenomic features. (a) Venn diagram showing number of DMRs for each of the skin cell types compared to non-skin samples (brain, breast, and blood). 8 DMRs (overlap region) share the same methylation status in the three skin cell types and have

the opposite methylation status in all non-skin samples. (b) WashU Epigenome Browser screenshot of the 8 DMRs where the three skin cell types share the same methylation status and all non-skin cell types have the opposite methylation status. Each column represents a 500bp window +/- 2.5 kb except for two columns which represent multiple contiguous 500bp windows +/- 2.5 kb. Each row is a MeDIP-seq track for the indicated cell type. Three replicates for each skin cell type and two replicates for each non-skin sample are depicted. (c) Clustering dendrogram based on average DNA methylation levels (predicted by methylCRF [112]) at 39,861 DMRs found between skin and brain tissue, breast, and blood cell types. DMRs are 500bp windows. (d) Venn diagram showing number of H3K4me1 peaks for each skin cell type that are absent in all non-skin samples (brain, breast, and blood), which also have overlapping H3K27ac signal. The intersection represents the 100 overlapping regions where H3K4me1 and H3K27ac peaks are present in all three skin cell types and H3K4me1 peaks are absent in all non-skin samples.

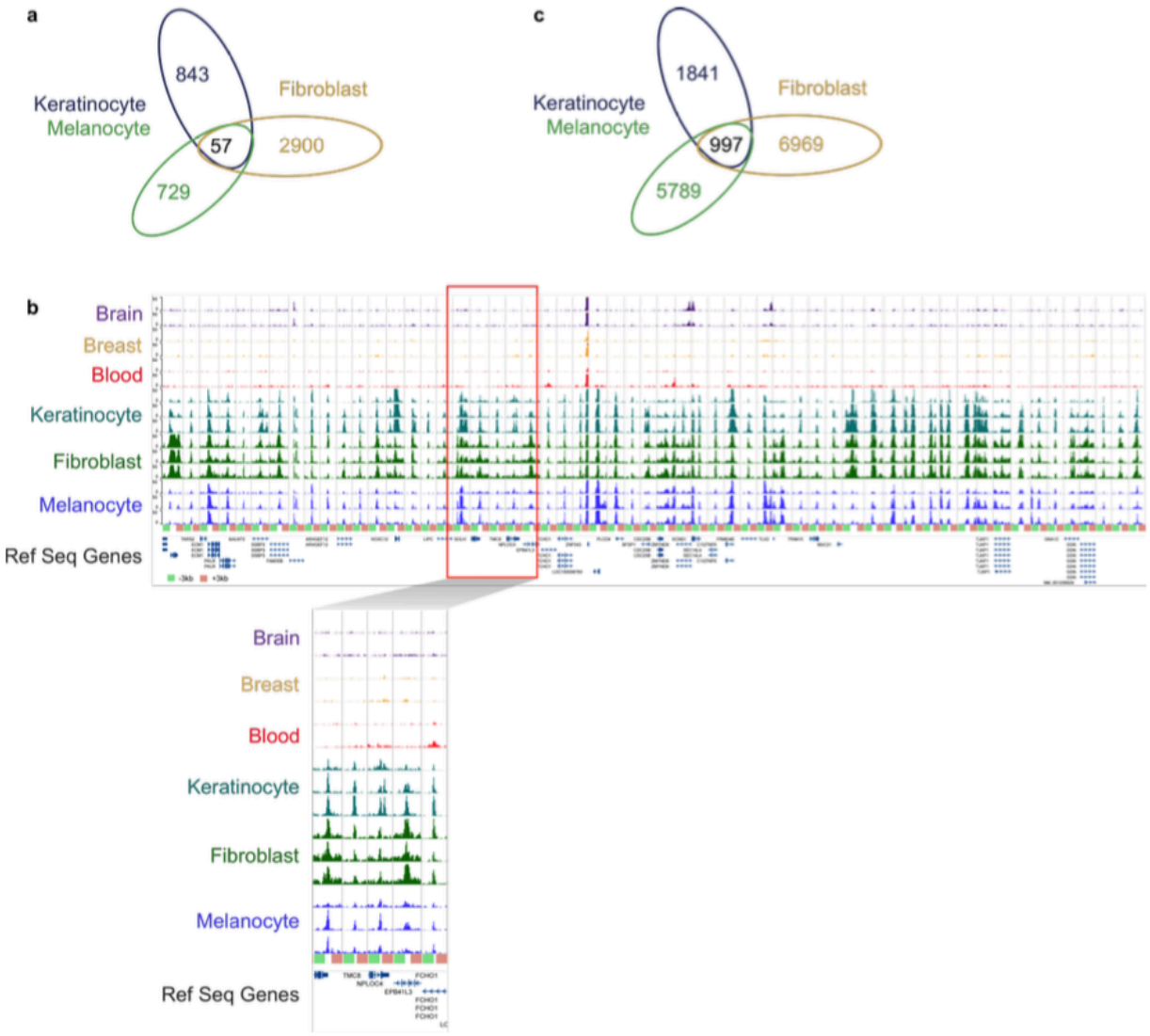


Figure 2.11. Shared histone modification patterns for skin cell types. (a) Venn diagram showing number of H3k4me3 peaks present in each skin cell type that are also absent in all non-skin samples (brain, breast, and blood). There are 57 overlap regions where H3K4me3 peaks are present in all three skin cell types and absent in all non-skin samples. A total of 55,859 H3K4me3 peaks were detected in all samples. (b) WashU Epigenome Browser screenshot of the 57 regions where H3K4me3 is present in all three skin cell types and absent in all non-skin cell types. Each row is a ChIP-seq track for the indicated cell type. Three replicates for each skin cell

type and two replicates for each non-skin sample are depicted. Each column represents one of the 57 different genomic regions +/- 3 kb. Bottom panel is a close-up of the red-boxed region in top panel. (c) Venn diagram showing number of H3k4me1 peaks for each skin cell type that are absent in all non-skin samples (brain, breast, and blood). There are 997 overlap regions where H3K4me1 peaks are present in all three skin cell types and absent in all non-skin samples. A total of 259,297 H3K4me1 peaks were detected in all samples.

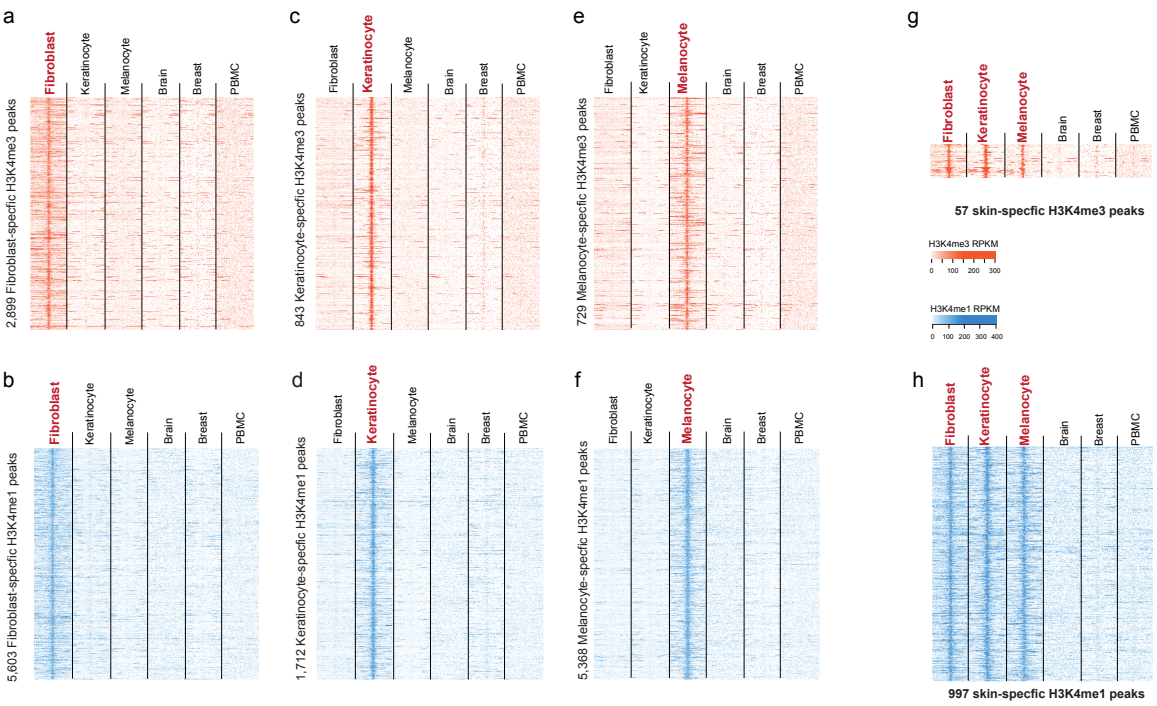


Figure 2.12. Heat maps of ChIP-seq signal around skin cell type-specific and tissue-specific histone modification peaks. (a) ChIP-seq signal for fibroblast-specific H3K4me3 peaks. Each heat map row represents a 10kb region centered on a fibroblast-specific H3K4me3 peak divided into 200 windows, read density (RPKM) was calculated for each window. Each heat map column represents ChIP-seq signal for the labelled cell type. Breast = breast myoepithelial cell, Brain = fetal brain tissue, and PBMC = peripheral blood mononuclear cells.

(b) Similar to (a), but for fibroblast-specific H3K4me1 peaks. (c-d) ChIP-seq signal for keratinocyte-specific H3K4me3 peaks (c) and H3K4me1 peaks (d). (e-d) ChIP-seq signal for melanocyte-specific H3K4me3 peaks (e) and H3K4me1 peaks (f). (g-h) ChIP-seq signal for skin tissue-specific H3K4me3 peaks (g) and H3K4me1 peaks (h).

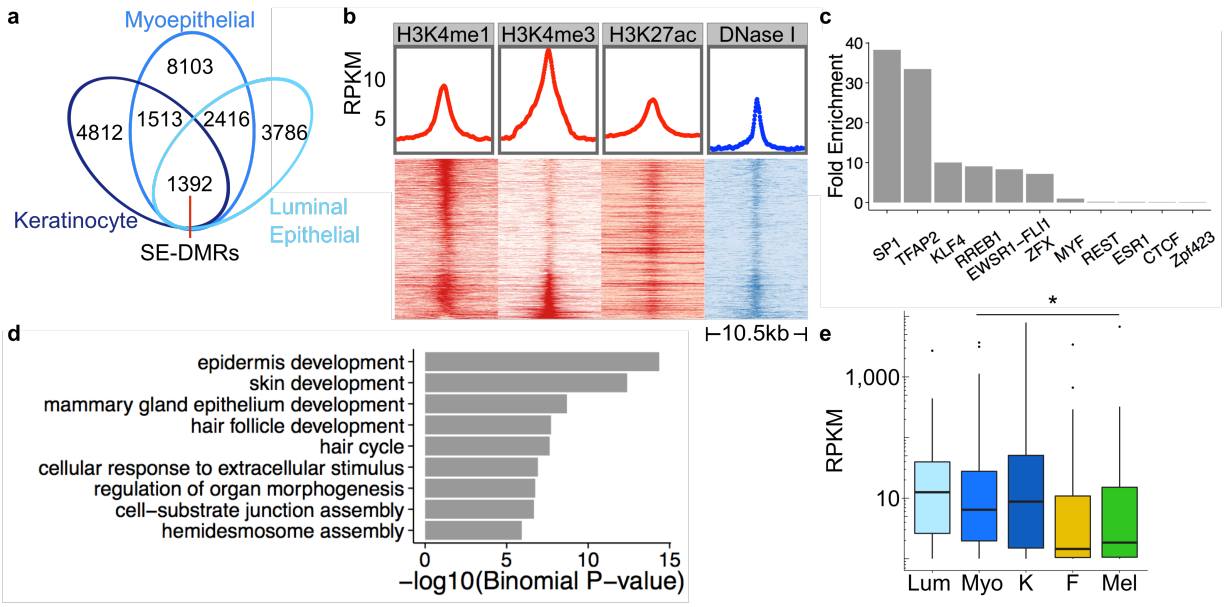


Figure 2.13. Identification and characterization of surface ectoderm-DMRs. (a) Venn diagram showing surface ectoderm-specific DMRs, defined as the overlap of keratinocyte, breast myoepithelial, and luminal epithelial cell DMRs. (b) Enrichment of H3K4me1, H3K4me3, H3K27ac, and DNAse I-hypersensitivity at SE-DMRs. Each heat map column represents histone modification ChIP-seq or DNAse-seq signal at 500bp SE-DMRs +/- 5 kb. Each heat map row represents a single hypomethylated SE-DMR, ordered by decreasing H3K4me1 signal, then increasing H3K4me3 signal. (c) Bar plot of enrichment values for top ten enriched TFBS motifs determined by motif scanning of hypomethylated SE-DMRs using FIMO [116] (**Methods**). Enrichment based on hg19 genome background. (d) Selected gene ontology terms enriched for hypomethylated surface ectoderm-DMRs. *P*-value of enrichment calculated by GREAT [85]. Full list of enriched GO terms is in **Appendix 2: Data 5**. (e) Box plots showing RNA expression levels for genes with hypomethylated SE-DMRs in promoter regions. Skin cell type RNA-seq RPKM values over exons are averages (mean) of three biological replicates; luminal epithelial and myoepithelial values are a single biological replicate. The middle line indicates the median value, top and bottom box edges are the third and first quartile boundaries respectively. The

upper whisker is the highest data value within 1.5 times the interquartile range; the lower whisker indicates the lowest value within 1.5 times the interquartile range. The interquartile range is the distance between the first and third quartiles. Points indicate data beyond whiskers. Logarithmic scale transformation was applied before boxplot statistics were computed. RPKM distributions for SE cell type expression levels vs. non-SE cell type expression levels are statistically significant (Wilcoxon-ranked test, paired, * indicates *P*-value < 0.02; *n* = 150 genes; Lum = breast luminal epithelial cells, Myo = breast myoepithelial cells, K= keratinocytes, F = fibroblasts, M = melanocytes; **Table 2.6**).

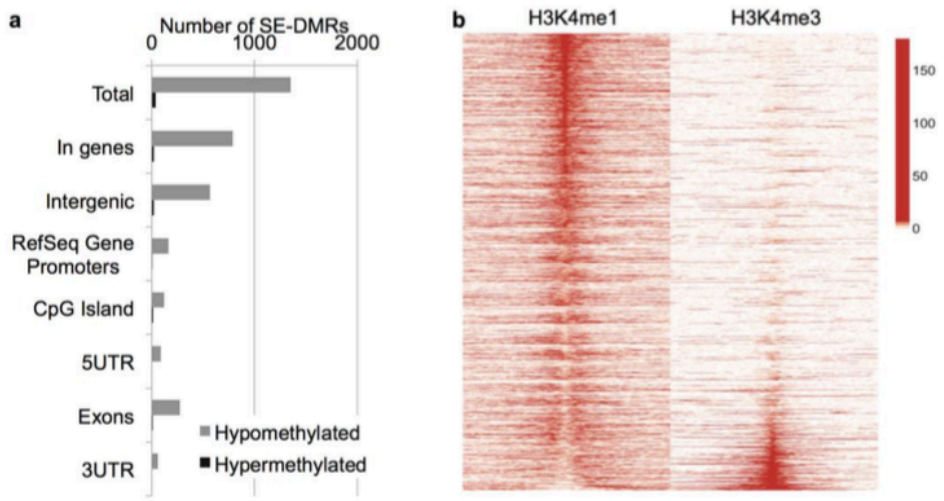


Figure 2.14. Additional SE-DMR characterization. (a) Genomic annotation of SE-DMRs. Hypomethylated and hypermethylated DMRs (1392 total) plotted independently. Genomic annotations described in Methods. (b) Breast myoepithelial cell histone modification ChIP-seq signal at SE-DMRs. Each row represents a 500 bp DMR +/- 5kb (as in Figure 4b). DMRs are sorted in descending order of H3K4me1 signal, then increasing H3K4me3 signal. Values plotted are RPKM normalized to input.

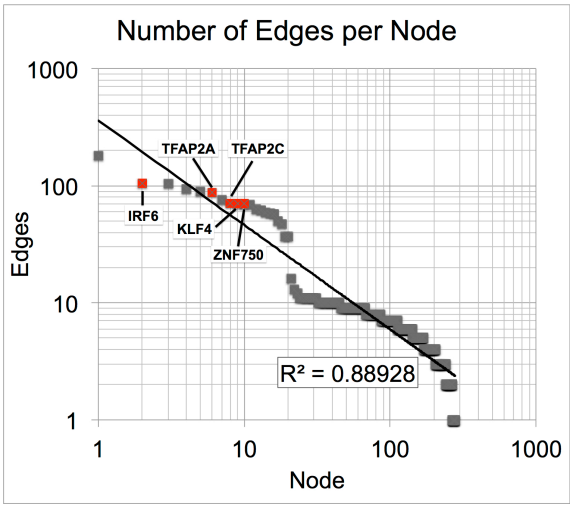


Figure 2.15. Distribution of edges per node in the SE network. Each node is plotted on the x-axis in order of its degree (total number of edges), the number of edges is the y-axis. The distribution fits a power law (black line) with R^2=0.88928. Gray and red boxes are individual nodes (genes). Genes of interest are highlighted in red and labeled. Genes with the highest degree are transcription factors at the top of the SE network (as in **Figure 2.16a**).

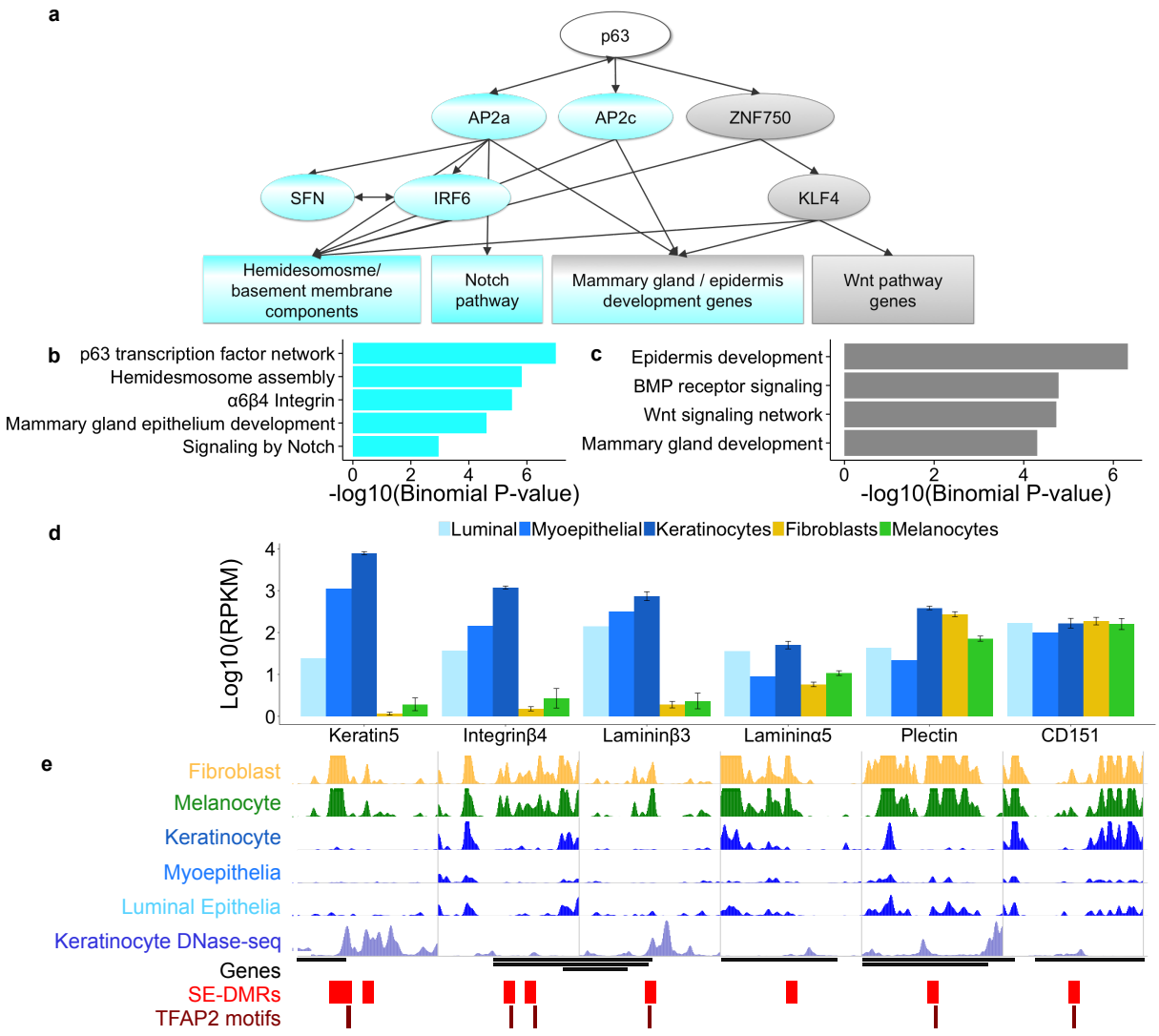


Figure 2.16. Surface ectoderm-DMRs are regulatory elements in a gene network. (a) Summary of the TF-target gene regulatory network derived from SE-DMR analyses. The categories at the bottom of the panel represent enriched biological processes or pathways for genes associated with TFAP2 or KLF4 motifs. TFAP2 associated TFs/pathways highlighted in blue; KLF4 associated pathways in gray. (b) Functional enrichment for TFAP2 motif containing hypomethylated SE-DMRs. (c) Functional enrichment for KLF4 motif containing hypomethylated SE-DMRs. (d) RNA expression values for SE-DMR associated hemidesmosome/basement membrane genes for SE and non-SE cell types. Skin cell type values

are averages (mean) of three biological replicates. Error bars are standard error of the mean (s.e.m.). (e) WashU Epigenome Browser screenshot of hemidesmosome/basement membrane genes. MeDIP-seq tracks depicted in green, yellow, and blue; all track y-axes heights are 60 RPKM. DNase-seq track is shown in light blue. Genes depicted as black lines. SE-DMRs depicted as red boxes and TFAP2 motifs as maroon lines.

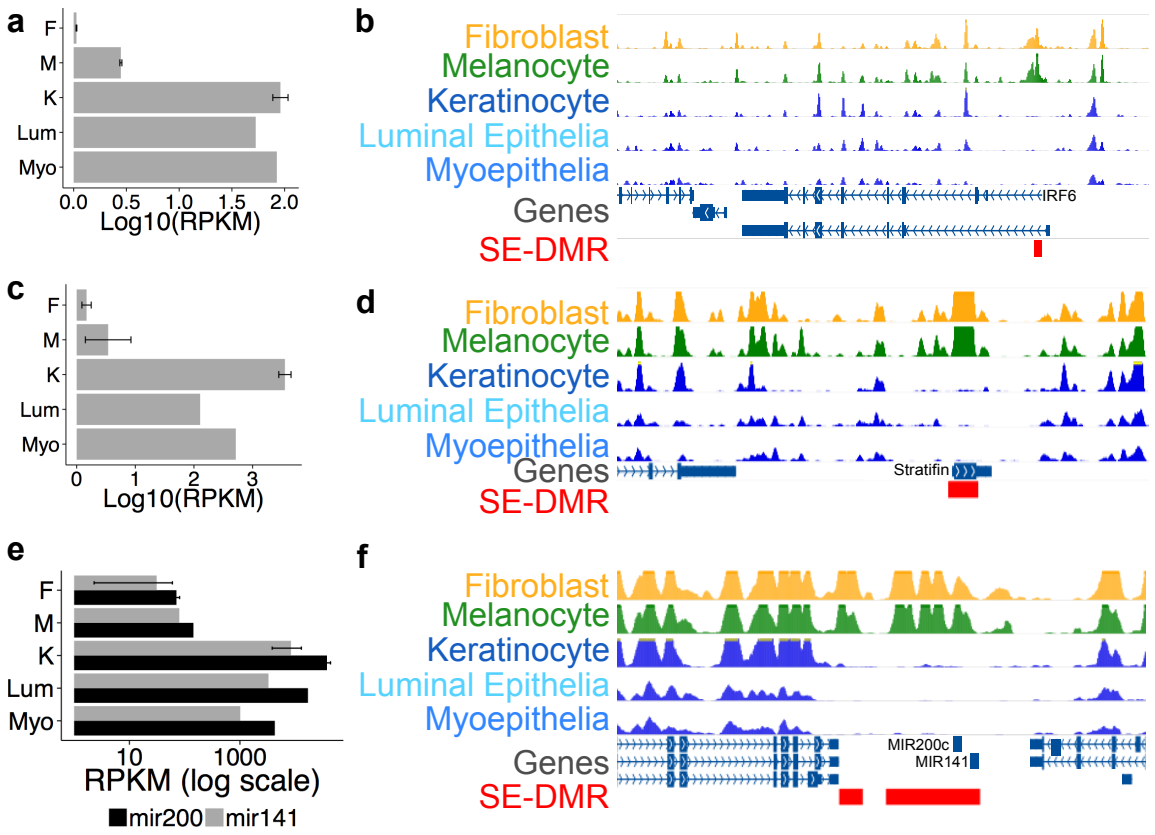


Figure 2.17. RNA expression levels and browser screenshots of selected SE-DMR loci. (a) Expression values for *IRF6* in each cell type as listed on the left. X-axis is expression in RPKM (log10 scale) for each cell type. Skin cell type values are averages (mean) of three biological replicates (error bars are s.e.m.); luminal epithelial and myoepithelial values are a single biological replicate. (b) Browser screenshot of *IRF6* locus and surrounding genomic region. MeDIP-seq tracks are shown for the indicated cell types; all track y-axes heights are 60 RPKM. Red box = hypomethylated SE-DMR near the *IRF6* promoter. (c) Expression for Stratifin (*SFN*) as in (a). (d) Browser screenshot of *SFN* locus. Tracks as in (b). Red box = hypomethylated SE-DMR at *SFN* promoter. (e) Expression values for mir-200c and mir-141 in each cell type as listed to the left. X-axis is RPKM (log scale). Keratinocyte value is the average (mean) of three biological replicates; fibroblast value is the mean of two biological replicates, (error bars are

s.e.m.); melanocyte, luminal epithelial, and myoepithelial values are a single biological replicate. (f) Browser screenshot of mir-200c/mir-141 locus and surrounding genomic region. Tracks as in (b). Red boxes = hypomethylated SE-DMRs including and adjacent to both miRNA loci.

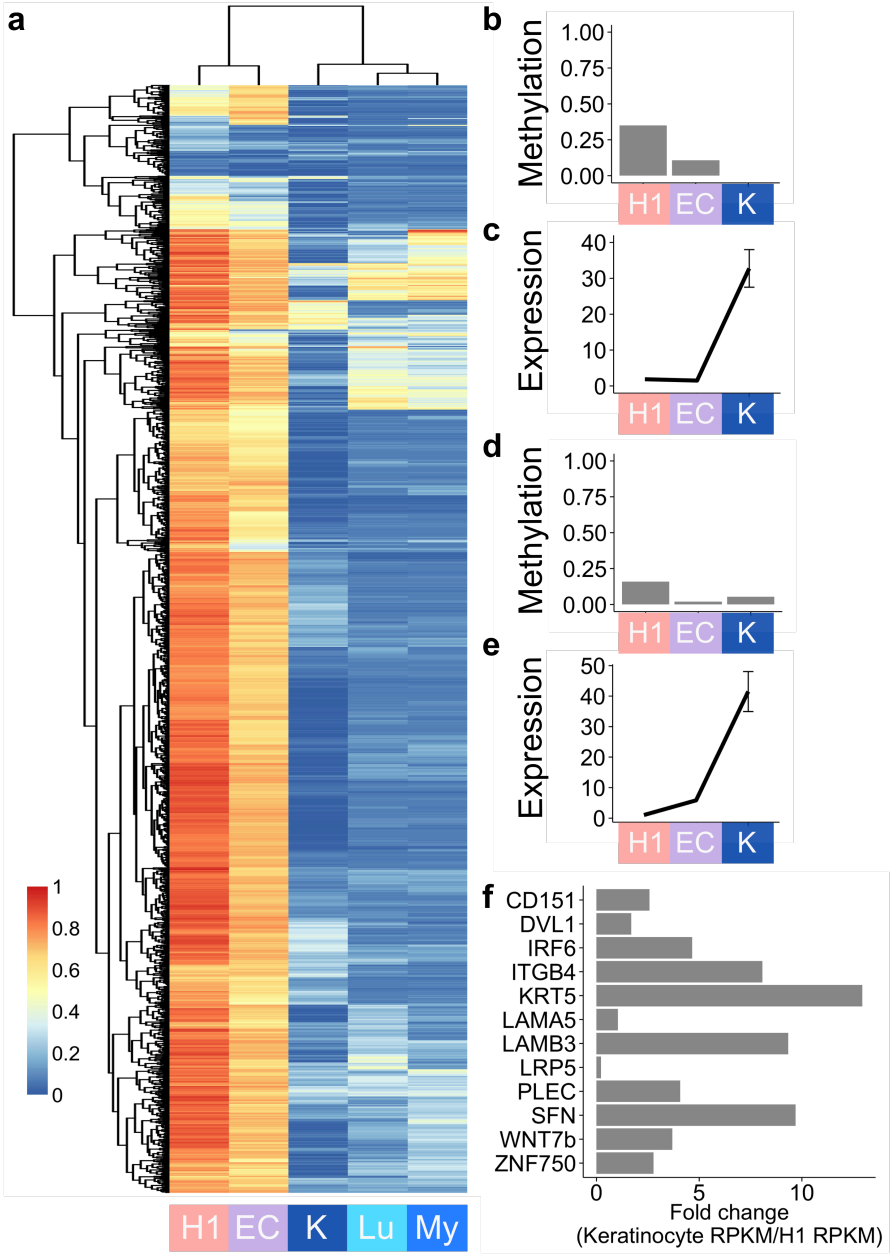


Figure 2.18. DNA methylation dynamics of SE-DMRs across samples from different developmental stages. (a) Heatmap and clustering dendrogram based on average CpG DNA methylation values of hypomethylated SE-DMRs for different developmental samples. Each row represents one of 1307 DMRs for which there are CpGs with ≥ 10x coverage in WGBS data. Methylation values for H1 ESCs, ectoderm differentiated ESCs ("EC"), and keratinocyte ("K") are from WGBS; breast luminal ("Lu") and myoepithelial ("My") values are the average of

single CpG methylCRF predictions in each DMR. MethylCRF predictions are based on MeDIP-seq and MRE-seq data for these samples (Methods). A value of "1" is fully methylated; "0" is completely unmethylated. (b) *KLF4* gene body SE-DMR average CpG DNA methylation levels across developmental stages. (c) *KLF4* RNA expression across developmental stages. Values are RPKM over coding exons; error bars for keratinocytes are s.e.m., $n = 3$. Sample abbreviations as in (a). (d) *TFAP2A* promoter SE-DMR average CpG DNA methylation levels across developmental stages. (e) *TFAP2A* RNA expression across developmental stages. Values are RPKM over coding exons; error bars for keratinocytes are s.e.m., $n = 3$. Sample abbreviations as in (a). (f) RNA expression levels in keratinocytes relative to H1 ESCs for selected genes with hypomethylated SE-DMRs in their promoters. These SE-DMRs, like the majority of hypomethylated SE-DMRs, were methylated in H1 and ectoderm-differentiated ESCs but lowly methylated in differentiated SE cell types. Increased expression relative to an earlier developmental sample suggests these DMRs are transcriptional regulatory regions for their associated genes.

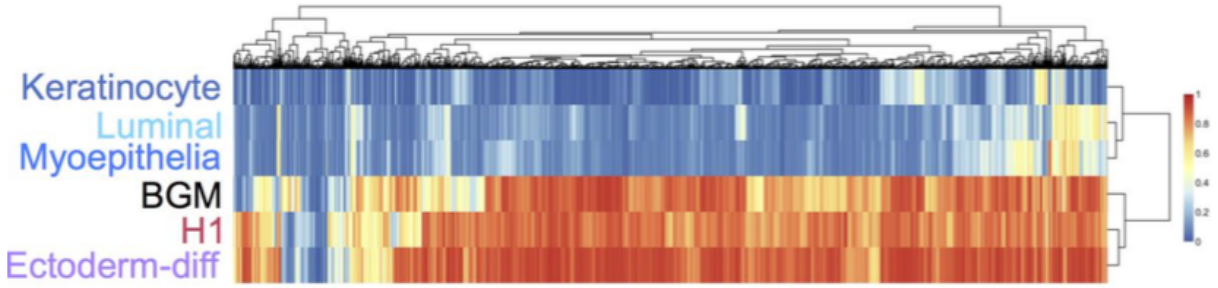


Figure 2.19. Heatmap and clustering dendrogram based on methylCRF CpG methylation values for hypomethylated SE-DMRs. Each column represents one of 1307 DMRs for which there are CpGs with ≥ 10x coverage. Keratinocyte, brain germinal matrix (BGM), H1 ESC, and ectoderm-differentiated ESC values from WGBS; breast luminal and myoepithelial values are the average of single CpG methylCRF predictions in each DMR. MethylCRF predictions are based on MeDIP-seq and MRE-seq data for these samples (Methods). A value of "1" is fully methylated; "0" is completely unmethylated.

Table 2.1. False discovery rate for calling DMRs across M&M q-values.

q-value	1.00E-02	1.00E-03	1.00E-05	1.00E-09	1.00E-13	1.00E-15
FDR	0.071	0.055	0.044	0.033	0.030	0.029

Table 2.2. Numbers of CGI and non-CGI promoters in all skin cell type-specific DMRs. χ^2 test p-value < 2.2e-16.

	# CGI promoters	# non-CGI promoters
Cell type DMRs	267	974
Genome-wide	16638	9691

Table 2.3. Wilcoxon test for keratinocyte-specific expression analysis. Wilcoxon ranked test, paired, *P*-values for RPKM distributions for Ref Seq genes with keratinocyte-specific hypomethylated DMRs at promoters.

	Fibroblast expression (average, *n*=3)	Melanocyte expression (average, n=3)
Keratinocyte expression (average, n=3)	2.20E-16	2.39E-13

Table 2.4. Wilcoxon test for fibroblast-specific expression analysis. Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with fibroblast-specific hypomethylated DMRs at promoters.

	Keratinocyte expression (average, n=3)	Melanocyte expression (average, n=3)
Fibroblast expression (average, n=3)	2.59E-09	3.14E-03

Table 2.5. Wilcoxon test for melanocyte-specific expression analysis. Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with melanocyte-specific hypomethylated DMRs at promoters.

	Keratinocyte expression (average, n=3)	Fibroblast expression (average, n=3)
Melanocyte expression (average, n=3)	3.65E-09	3.53E-09

Table 2.6. Wilcoxon test for surface ectoderm-specific expression analysis. Wilcoxon ranked test, paired, *P*-values for RPKM distributions for RefSeq genes with surface ectoderm-specific hypomethylated DMRs at promoters.

	Melanocyte expression (average, *n*=3)	Fibroblast expression (average, *n*=3)
Keratinocyte expression (average, *n*=3)	1.00E-05	1.31E-09
Luminal epithelia expression	1.07E-04	1.18E-07
Myoepithelia expression	1.51E-02	2.25E-04

Table 2.7. Statistics for network analysis. Control datasets and statistics for SE network.

Random dataset filenames	one	two	three	four	five	six	seven	eight	nine	ten
Number of random genes into the Interaction Browser	374	374	374	374	374	374	374	374	374	374
Number not found in database	13	16	13	11	8	7	8	11	15	15
Number of edges	810	841	996	1068	995	890	1015	1193	1034	738

Mean (# edges in random datasets)	958
Standard Deviation (random datasets)	136.54
Number of edges in SE network	1458
P-value (t-test, upper-tail)	1.25E-04

Table 2.8. TFBS motif-containing DMRs. Number of hypomethylated SE-DMRs (1353 total) which contain TFAP2 and/or KLF4 binding site motifs.

	Number of SE-DMRs
Contains TFAP2 motif only	283
Contains KLF4 motif only	273
Contains both TFAP2 and KLF4 motifs	283
No TFAP2 or KLF4 motifs	514

2.8 Accession Codes

Accession codes for keratinocyte skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the Gene Expression Omnibus (GEO) database under the accession codes GSM707022, GSM707018, GSM751278, GSM817253, GSM669589, GSM669591, and GSM817242 respectively. Accession codes for keratinocyte skin02 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941726, GSM941723, GSM941745, GSM1127113, GSM941735, GSM941736, and GSM941742 respectively. Accession codes for keratinocyte skin03 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, WGBS, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM958180, GSM958169, GSM958177, GSM1127111, GSM1127056/GSM1127058, GSM958155, GSM958161, GSM958156, and GSM958167 respectively. Accession codes for fibroblast skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM707021, GSM707017, GSM751277, GSM817252, GSM817235, GSM817234, and GSM817246 respectively. Accession codes for fibroblast skin02 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941725, GSM941722, GSM941744, GSM941718, GSM941717, and GSM817247 respectively. Accession codes for fibroblast skin03 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under

the accession codes GSM958182, GSM958171, GSM958178, GSM1127116, GSM958158, GSM958164, GSM958163, and GSM958168 respectively. Accession codes for melanocyte skin01 MeDIP-seq, MRE-seq, mRNA-seq, miRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM707020, GSM707016, GSM751276, GSM817251, GSM941719, GSM941728, and GSM941740 respectively. Accession codes for melanocyte skin02 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM941727, GSM941724, GSM941743, GSM941731, GSM941730, and GSM941741 respectively. Accession codes for melanocyte skin03 MeDIP-seq, MRE-seq, mRNA-seq, H3K4me3 ChIP-seq, H3K4me1 ChIP-seq, H3K27ac ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM958181, GSM958170, GSM958174, GSM958151, GSM958152, GSM958157, and GSM958166 respectively. Accession codes for breast luminal epithelia RM071 MeDIP-seq and MRE-seq datasets have been deposited in the GEO database under the accession codes GSM1517154 and GSM613826 respectively. Accession codes for breast luminal epithelia RM080 mRNA-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM669620, GSM669595, and GSM959124 respectively. Accession codes for breast myoepithelia RM071 MeDIP-seq and MRE-seq datasets have been deposited in the GEO database under the accession codes GSM1517153 and GSM613908 respectively. Accession codes for breast myoepithelia RM080 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM693277, GSM613885, and GSM613897 respectively. Accession codes for the Fetal Brain Germinal Matrix HuFGM02 WGBS dataset

have been deposited in the GEO database under the accession code GSM941747. Accession codes for PBMC TC015 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM613811, GSM613814, and GSM613816 respectively. Accession codes for CD8 Naïve TC001 H3K4me3 ChIP-seq, H3k4me1 ChIP-seq, and input ChIP-seq datasets have been deposited in the GEO database under the accession codes GSM1127126, GSM1127143, and GSM1127151 respectively.

2.9 Acknowledgements

We would like to acknowledge support from the NIH Roadmap Epigenomics Program, supported by the National Institute on Drug Abuse (NIDA) and the National Institute of Environmental Health Sciences (NIEHS). J.B.C. is supported by a Career Development Award from the Dermatology Foundation. J.F.C., M.H., and T.W. are supported by NIH grant 5U01ES017154. T.W. is also supported by NIH grants R01HG007354, R01HG007175, and American Cancer Society grant RSG-14-049-01-DMC. T.M is supported by NIH grant R01AG028492, administered by the Northern California Institute for Research and Education, and with resources of the Veterans Affairs Medical Center, San Francisco, California. B.Z. is supported by NIDA's R25 program DA027995. R.F.L. is supported by the NSF Graduate Research Fellowship Program (DGE-1143954) and by the Washington University Interface of Psychology, Neuroscience, and Genetics training program (NIH, grant no. 5T32GM081739).

Chapter 3

DNA Methylation Dynamics in Zebrafish Pigment Cell Development

3.1 Author Contributions

I am a co-contributor on this work in collaboration with Josh Jang, a fellow graduate student in the Ting Wang lab at Washington University in St. Louis. Critical zebrafish resources were provided by Steven Johnson in the Genetics Department at Washington University in St. Louis. Project design and intellectual contributions were by me, Josh Jang, Stephen Johnson, and Ting Wang. Protocols for pigment cell isolation were developed with assistance from Scott Higdon.

3.2 Background

Below is a discussion of the developmental genetic and epigenetic control of neural crest development in the literature to date. A note on nomenclature: when synthesizing the role of a gene across two or more vertebrate systems, the convention for mouse gene nomenclature is used. When discussing a result in the context of a specific vertebrate system, the gene nomenclature appropriate for that organism is used.

3.2.1 Neural Crest Specification

The neural crest is a multipotent, vertebrate-specific tissue with diverse and important biological roles, including peripheral nervous system, cranial neurons and glia, cartilage and bones of the face, connective tissue, cardiac tissue, and pigment cells [123]. The neural crest is remarkable for its generation of a variety of cell types that migrate to all parts of the body.

The neural crest emerges from the border between neural plate and non-neural ectoderm, where instructive cues from the ectoderm and underlying mesoderm signal the neural plate border [124]. Intermediate Bmp signal from the ectoderm is necessary, but not sufficient for neural border specification. Bmp signal is combined with Wnt and Notch/Delta signaling from the ectoderm and Wnt and Fgf signaling from the underlying mesoderm to specify the neural plate boarder zone [125]. Hippo signaling induces *Pax3* in mouse [126], which along with *Zic1* are necessary and sufficient to trigger the neural crest developmental program [127].

Once the neural plate border is established, *Pax3* and *Zic1* drive expression of neural crest specifier genes [128]. These include *Snail1/2*, *SoxE* family transcription factors (notably *Sox10*), and *FoxD3* [125]. As is the case with most developmental transcription factors, transcription factors involved in neural crest cell development and lineage specification are reused in multiple contexts. Besides specifying neural crest, *Snail2* is a central regulator of the epithelial-to-

mesynchymal transition (EMT) [129], a process that is critical to delamination of neural crest cells and their migration throughout the developing organism [130]. Indeed, genes important in cell migration were found to be upregulated in transcriptomes of cranial neural crest cells [131]. In addition, melanocytes and iridophores both need to disperse dorsally and ventrally across the developing embryo [132]. Last, the transcriptional repressor *FoxD3* drives self-renewal and multipotency in pre-migratory NCCs, but later represses ectomesenchymal and melanocyte cell fates while promoting neuronal, glial, and iridophore cell fates [133]. Thus the developmental genetic network of early neural crest cell specification is fairly well understood. However the molecular genetics of NCC-lineage bifurcation is not completely described.

Myc is also and important neural crest gene, although it is not specific to NCCs. However *Myc* activity in NCCs is critical for NCC self-renewal properties, similar to *Myc* function in stem cells and cancer. *Myc* is known to be upregulated in melanoma [134]. Therefore the importance of *Myc* in NCC specification and melanogenesis is an important aspect for future study.

3.2.2 Developmental Genetics of Zebrafish Melanocyte and Iridophore Differentiation

Pigment Cell Ontogeny

The zebrafish *Danio rerio* exhibits three types of pigment cells: melanocytes, iridophores, and xanthophores. Melanocytes are melanin-containing cells orthologous to human melanocytes. Iridophores are guanine-containing cells that produce the iridescent aspect of fish scales. Yellow, pteridine-containing xanthophores are the third pigment cell of the zebrafish. All three are neural crest-derived cells; however melanocytes and iridophores share a common *mitfa*+ precursor cell population [135,136] (**Figure 3.1**). Understanding the melanocyte/iridophore cell fate choice is the focus of the work in the present chapter.

Mitf *and Melanocyte Differentiation*

Melanocytes are a conserved, neural crest-derived cell type responsible for the pigmentation patterns across many vertebrates. Because of their deep conservation, and because disruption of pigment patterns does not confer a detrimental fitness cost in most cases, melanocytes have been the subject of developmental genetic investigation for many species and for many years [137]. While analyzing the promoter of *Tyrosinase*, the gene required for melanin synthesis, *Mitf* was discovered as the master regulator of melanocyte fate [138].

mitfa is the *Danio rerio* ortholog that controls melanocyte development in zebrafish [135]. In zebrafish, melanocytes may be either direct-developing or regenerative [139]. Most of the embryonic melanocyte pigment pattern is derived from direct-developing melanoblasts, although a small fraction of the wildtype pattern is contributed by melanocyte stem cells (MSCs) acting in a regulative manner to complete the pigment pattern [140]. Direct-developing melanoblasts and MSCs share a common *mitfa*+ precursor [136,141]. *mitfa* is not required for melanocyte stem cell establishment, although presumably it is required for MSC-derived melanocyte differentiation [142].

Because the focus of this project is on a specific cell fate decision – that of melanocyte or iridophore cell fate – and because MSCs contribute only a small fraction of mature melanocytes in a wildtype embryo, the focus of this project is on direct-developing melanoblasts and melanocytes. As discussed below, the direct-developing melanoblast is hypothesized to be the melanocyte/iridophore progenitor cell, as shown in **Figure 3.1.**

FoxD3 *and Neural Crest Diversification*

The genetic control of iridophore fate segregation has been generally less well-studied than that of melanocytes, presumably because there is less direct relevance for human biology. However,

from the perspective of fate segregation, iridophore development becomes a very attractive system, as pigmentation can be perturbed without fitness consequences to the organism (in a lab environment). The first comprehensive gene expression analysis of iridophores was recently published [67]; however understanding of iridophore differentiation is still incomplete.

The most important genetic factor known to be involved in iridophore differentiation is *foxd3*. *foxd3* is involved in early NCC specification, where it regulates *snail2* and *sox10* expression in premigratory neural crest cells and causes delayed migration of NCCs [143] (see above). Knock down of *foxd3* caused a reduction of iridophores in morpholino-injected embryos, especially in the trunk and tail regions [133], consistent with the noted migration phenotype. Thus *foxd3* has early roles in NCC establishment but is also required later for pigment cell fate-specification.

foxd3 is key to melanocyte/iridophore fate segregation. Lineage analysis revealed that both pigment cells are derived from a *mitfa*+ precursor [136,141]. Indeed, *foxd3* represses melanocyte cell fate specifically by repressing the *mitfa* promoter [144], and timely down-regulation of *foxd3* is needed for melanocyte differentiation [145]. Similarly, loss of *mitfa* resulted in a decrease in melanocytes and a concomitant increase in iridophores [135]. The reuse of *foxd3* – for early neural crest specification and later for iridophore cell fate segregation – is a classic example of the reuse of developmental genes, particularly transcription factors.

3.2.3 Epigenome in Dynamics in Zebrafish Development

The developmental genetic context for pigment cell differentiation in zebrafish has been the subject of intense study for almost two decades. However, epigenetic regulation of the zebrafish genome has only recently become illuminated with the advent of cost-effective high-throughput sequencing technologies. Below is a brief overview of the extent of the current understanding of

zebrafish epigenetic regulation, with emphasis given to how epigenetic regulation relates to neural crest development in human and other organisms.

DNA Methylation Dynamics and Machinery in Zebrafish

Similar to the human genome, the zebrafish genome is highly methylated (~80% globally), especially over gene bodies and transposable elements [13]. Most 5-methylcytosine nucleotides are found in the CpG context, with very low levels found at CHG and CHH contexts [13], consistent with DNA methylation patterns in mammalian genomes.

The core DNA methylation machinery is conserved between mammals and zebrafish. The zebrafish genome contains homologs for both maintenance and *de novo* DNA methyltransferase enzymes [146] and the hemimethylation-binding factor *Uhrf1*, which is required for maintenance DNA methylation [147]. Notably, zebrafish lack a *Dnmt3L* homolog. *Dnmt3L* is critical for monoallelic methylation of loci at imprinted genes, as well as repression of transposable elements (TEs). While zebrafish do not exhibit imprinting, the genome is moderately enriched for methylation over transposable elements [13], suggesting an alternate mechanism for DNA methylation-mediated repression of TEs in the zebrafish genome. The zebrafish genome also contains homologs for other important methyl-binding proteins including *Mecp2* and *Mdb2* homologs. Finally, regulation of demethylation occurs in part by a suite of active DNA demethylation enzymes [148].

Recently, the application of whole genome bisulfite sequencing technologies to zebrafish gametes and early embryos has revealed the dynamics of DNA methylation in the few hours post-fertilization. Zebrafish oocyte genomes are markedly hypomethylated compared to sperm and somatic tissues [149,150]. Zebrafish undergo zygotic genome activation at the mid-blastula transition, and by this developmental stage the maternal genome DNA methylation pattern has

increased to mirror that of the paternal genome [149,150].

Epigenetic regulation of developmentally important regulatory elements shares many aspects of mammalian transcriptional regulation. As embryogenesis progresses, DNA methylation decreases specifically at promoters and distal regulatory elements, which bear marks of enhancer elements including H3K4me1 and H3K27ac [27]. Also similar to mammals, zebrafish promoters are enriched for H3K4me1/3 and this enrichment is correlated with increased gene expression, and H3K4me1 was found to mark enhancers that can drive tissue-specific expression [151].

Epigenome Regulation in Neural Crest Cell Differentiation

Modulation of histone post-translational modifications has been moderately explored in the context of neural crest development across vertebrate systems. In human and *Xenopus*, the chromatin remodeler CHD7 is required for neural crest specification, specifically *Sox9, Twist,* and *Slug* expression (all neural crest specifier genes) and subsequent neural crest cell delamination and migration [152]. Subsequently, Snail2 recruits Polycomb remodeling complex 2 (PRC2) to regulate neural crest development in *Xenopus*. Knockdown of Ezh2 (the catalytic subunit of PRC2) resulted in neural crest migration defects, as Snail2 interacts with PRC2 to down-regulate E-cadherin, and is required for the epithelial-to-mesenchymal transition and neural crest migration [153].

Histone deacetylases 1 and 2 are required for *Pax3* and *Sox10* expression in mouse, the latter of which are master regulators of neural crest [154]. Concomitantly, the histone demethylase *JUMONJID2A* is required for neural crest specifier gene expression, specifically for *SOX10* derepression in chick embryos [155]. Finally, in human, neural crest-specific enhancers are marked by active histone marks and the transcription factors TFAP2A and NR2F1/2, and these regulatory elements drive expression in the absence of sequence conservation [156].

During pigment cell development, loss of function of *hdac1* in zebrafish caused delayed neural crest migration and differentiation – similar to the phenotypes of *foxd3* loss (reviewed above). In addition, *hdac1* loss resulted in prolonged expression of *foxd3*, reduced *mitfa* expression, and resulted in a reduction of melanocytes, suggesting that *hdac1* normally acts to repress *foxd3* in pigment progenitor cells that acquire a melanocyte fate [157].

3.3 Rationale and Hypothesis

Two decades of work on neural crest biology has yielded a solid understanding of early neural crest gene regulatory networks. Recent application of epigenome analysis to model organisms has begun to extend our knowledge of how the epigenome regulates gene expression in model organisms and of the general principles of the zebrafish epigenome with respect to mammalian systems. Evidence that epigenome regulation is crucial to cell fate decisions is evident, especially considering the instructive cues of histone demethylases in neural crest specification and melanocyte fate, as described above. However, a comprehensive understanding of how epigenetic regulation contributes to a specific cell fate decision is still lacking.

To examine the role of the epigenome in a specific cell fate decision, we examined zebrafish pigment cell differentiation. Neural crest cells generate a variety of cell types that contribute to ectomesenchymal structures, peripheral nerves, and pigment cells. In zebrafish, the neural crest cell population includes a set of melanocyte/iridophore progenitor cells. These progenitor cells can give rise to either a melanocyte or iridophore (**Figure 3.1**). Some genetic regulators of pigment cell fates are known (see section **3.2**), but how the epigenome acts to specify either fate has not been comprehensively examined. We hypothesized that DNA methylation dynamics at regulatory elements drive pigment cell fate specification by modulating transcription factor binding affinity of key regulatory elements. To investigate this hypothesis we asked three questions. First, how do DNA methylation dynamics change over pigment cell development? Next, how do DNA methylation dynamics at transcriptional regulatory elements contribute to pigment cell fate determination? Finally, how do TF-epigenome interactions drive enhancer activity? I addressed these questions by generating genome-wide DNA methylation and transcriptome datasets and analyzing the genetic signatures correlated with cell fate-associated

DNA methylation changes in the context of pigment cell differentiation.

3.4 Experimental Design

To profile the methylome and transcriptome during pigment cell development, we isolated several stages of neural crest cells during the pigment cell differentiation (**Figure 3.2**). To isolate neural crest cell populations, we used the *crestinA>GFP* transgenic line, which expresses *GFP* in neural crest cells [158] (**Methods**). A time course experiment determined that the earliest GFP expression in this line was observed at approximately the 14-somite stage; therefore, this was the earliest time point neural crest cells could be isolated. By 24hpf (prim-5 stage), neural crest cells that will give rise to pigment cells are starting to commit to pigment cell fate [159] as they migrate dorsally and caudally across the embryo. Thus 24hpf was chosen as our second neural crest/pigment cell progenitor sample. By 5dpf, the pigment cell pattern of zebrafish larvae is established, so we collected melanocytes and iridophores at this time point.

After single cell dissociation, target cell types were isolated using fluorescence-activated cell sorting (FACS; **Methods**). For the two early neural crest populations, we collected the GFP- population, which represents the non-neural crest cell population. The GFP- control samples were used to verify the specificity of our neural crest cell (GFP+) during data analysis.

All samples were then processed for whole-genome DNA methylation and gene expression analysis. We chose to use whole genome bisulfite sequencing (WGBS) for this project for several reasons: (1) utilizing low-input protocols in our lab allowed us to gather WGBS data for samples that were previously difficult to process; (2) early MeDIP-seq experiments on these samples had to be thrown out when a change of antibody vendor introduced dramatic and unexpected technical variation into the data; (3) we reasoned that our study would gain more credibility by using WGBS, especially as model organism studies often face a high need for

relevance to human biology, and WGBS data are more accepted in the community than MeDIP-seq and MRE-seq technologies.

For gene expression analysis, we leveraged pigment cell expression data previously published by Higdon, 2013 [67]. To complement these data, we generated gene expression data using mRNA-seq on the neural crest cell samples and their corresponding GFP- controls. We used the TruSeq Illumina library kit for mRNA-seq library construction. We note that the pigment cell expression data was generated using a different kit (ScriptSeqV2) and that this may introduce technical variation into the data. However, preliminary analysis showed differential expression between neural crest pigment progenitors and the pigment cell populations that were biologically meaningful (see section **3.5**). Therefore we conclude that despite technical differences in library preparation, our experimental design can recover biologically relevant results.

3.5 Preliminary Data Analysis

3.5.1 Whole Genome Bisulfite Preliminary Analysis

WGBS Quality Control

Whole genome bisulfite sequencing (WGBS) libraries were constructed from two biological replicates for each sample collected (**Methods**). Adapter sequences were trimmed using cutadapt and reads were aligned using Bismark. CpG read coverage is an important indicator of confidence when calling DNA methylation levels. Greater than 10x coverage is considered a good standard for calling CpG methylation. Our WGBS libraries had 2-10 million CpGs with ≥10x coverage. (**Figure 3.3)** More CpGs (an additional 5.6 – 10 million) had 5-9x read coverage.

Principal component analysis (PCA) was used to assess the variance among the datasets. The methylation values of CpGs covered in all datasets were used for PCA. The first two components showed separation of methylomes from differentiated pigment cells from most of the embryonic stages (**Figure 3.4a**), and the first two components explained 81.6% and 4.2% of the variance respectively (**Figure 3.4c**). PC2 and PC3 also stratify samples with respect to developmental stages, however the outlier dataset, 24hpf GFP+ replicate 2, became more apparent (**Figure 3.4b**).

Overall, the above quality metrics suggested either deeper sequencing or more complex libraries are needed before exhaustive analysis can be done; ideally an average CpG coverage of 10x can be obtained for all datasets.

Differentially Methylated Regions Analysis

We proceeded to analyze the data on hand to find if the samples show expected DNA methylation differences at key loci. We identified differentially methylated regions using DSS [160], a software specific for calling differentially methylated regions (DMRs) from WGBS data.

To capture high-confidence DMRs, we required a DNA methylation difference of ≥25%. Strikingly, most DMRs were hypomethylated in the sample from a "later" developmental time point (**Figure 3.5a**). For example, when comparing 14-somite GFP+ early neural crest samples to 24hpf GFP+ pigment progenitor cells, only 68 DMRs were hypomethylated in the 14-somite stage compared to 24hpf, but 869 DMRs were hypomethylated in the 24hpf GFP+ methylomes compared to 14-somite GFP+ methylomes. Similarly, when comparing the 24hpf GFP+ pigment progenitor samples to differentiated melanocytes or iridophores, we found over 10 times as many hypomethylated DMRs in the pigment cell than in the progenitor. This trend is in line with the observed loci-specific loss of DNA methylation that occurs during embryogenesis [27].

We examined select loci to confirm that the expected DNA methylation dynamics were captured in our libraries. Two DMRs over the *mitfa* promoter revealed progressive DNA demethylation from neural crest and pigment progenitor stages until the locus is demethylated in melanocytes and iridophores. *mitfa* is not expressed in iridophores – in fact is repressed in order for iridophore fate to be established [144]. However, since melanocytes and iridophores share a very recent common precursor, and as DNA methylation often bears imprints of the developmental history of a cell [58,60], finding that the *mitfa* promoter is still demethylated in iridophores is not so surprising. Instead, it would be interesting to investigate if DNA methylation changes at distal enhancers or repressors are responsible for *mitfa* repression in iridophores, or if another epigenetic mechanism is actively repressing *mitfa* in iridophores, for example, Polycomb-mediated repression.

We note that we do observe moderate *mitfa* expression in iridophores in our data (see **Figure 3.9i**). We reason that even a slight melanocyte contamination in our iridophore sample would result in very high *mitfa* expression, given the very high expression of *mitfa* in melanocytes (note

the log scale in **Figure 3.9i**). More comprehensive expression analysis is needed to verify these early results.

Next we examined the *pnp4a* locus, an iridophore-specific marker [136]. We find several DMRs over a 10kb region centered on the *pnp4a* promoter. We observe dynamic demethylation of regions upstream of the promoter and in *pnp4a* introns 3 and 4 (**Figure 3.5c**). In this case, *pnp4a* region DNA methylation is consistent with *pnp4a* expression patterns we observe (see **Figure 3.9k**).

3.5.2 mRNA-seq Preliminary Analysis

mRNA-seq Quality Control

mRNA-seq libraries were generated for the early embryo samples (**Methods**) and pigment cell mRNA-seq fastq files were downloaded from GEO [67]. All libraries were aligned to the danRer10 transcriptome assembly with STAR [161]. Uniquely mapped read alignment rates ranged from 75% to 92% across all libraries (**Figure 3.6**). While the libraries for early embryo stages were sequenced much more deeply than the pigment cell mRNA-seq libraries, we reasoned that the multiple replicates for melanocytes and iridophores will give us power to detect differentially expressed genes. Indeed, replicates are important for determining gene expression variance in a sample. Therefore, biological replicates were kept separate for statistical analysis.

Basic quality control analyses include pairwise correlation metrics between biological replicates. Expression levels by transcript were determined using the htseq-count Python utility [162], then normalized using edgeR [163]. Transcript expression levels were plotted as scatterplots for each pairwise comparison within the early embryo stages (**Figure 3.7**) and pigment cells (**Figure 3.8**). Across the early embryo stages, the Spearman correlation ranged from 0.69 – 0.98. The highest

Spearman correlations were between biological replicates (orange boxes), indicating there were minimal batch effects during sample preparation and library construction.

The moderately high correlation between non-biological replicates is not surprising, as most genes are expected to be similarly expressed between samples, as core cell biological pathways are conserved across cell types. Instead, our analysis aims to find the specifically differentially expressed genes between biologically meaningful comparisons (see **Figure 3.9**), so the results in these plots indicate high-quality data.

Multidimensional scaling (MDS) analysis allows for visualization of the similarity (or dissimilarity) of several datasets at once. A quantitative metric is used to determine (dis)similarity between datasets. For example, in **Figure 3.9a**, points represent the Euclidian distance separating samples based on the gene expression levels of the top 500 genes in each sample. Points that are closer together represent more similar datasets, while those farther apart are more dissimilar. The MDS plot in **Figure 3.9a** shows biological replicates clustering together. Further, the x-axis is strongly correlated with the developmental trajectory of the samples studied: the 14-somite stages are on the far left, followed by 24hpf samples, and the pigment cell samples are to the far right, separated by the y-axis.

Notably, the 14-somite GFP+ and GFP- samples are clustered together; we find other indicators that these mRNA-seq libraries are very similar (see **Figure 3.9b**). Therefore we are considering regenerating these data (see section **3.6**).

Differentially Expressed Genes Analysis

Next, we called differentially expressed genes (DEGs) using the edgeR package [163]. Because we wanted to enrich for high-confidence differentially expressed genes, DEGs were called with

an FDR-corrected p-value of ≤ 0.001. **Figure 3.9b** plots the numbers of DEGs in each pairwise comparison. Yellow bars indicate genes that were upregulated in the first sample compared to the second: for example, there were 28 genes more highly expressed in the 14-somite GFP+ samples than the 24hpf GFP+ samples. Blue bars indicate the number of genes upregulated in the second sample: 3 genes were more highly expressed in 24hpf GFP+ samples than 14-somite GFP+ samples.

We found few DEGs between the early samples. Only the two neural crest cell samples had differential expression (14-somite GFP+ vs 24hpf GFP+; **Figure 3.9c**). It is interesting that the GFP- controls did not have differentially expressed genes with their corresponding GFP+ samples at the p-value used here. One possible reason for this is if our reporter does not mark 100% of neural crest cells, some neural crest cells may infiltrate the GFP- population collected during FACS. However a more likely explanation for lack of gene expression difference is that many developmental genes are reused in different contexts, and it is likely that many of the important neural crest genes are also expressed in other parts of the embryo, all of which comprise the GFP- control populations. For example, *tfap2a* is important for establishing the neural plate boarder and premigratory neural crest [125], but is also involved in epidermal development and required for the establishment of a subset of sensory neurons [164]. Therefore *tfap2a* will not be detected as a DEG. In fact, normalized expression for *tfap2a* in 14-somite GFP- was ~500 reads per million (RPM) in both samples, and ~750 RPM in both GFP+ samples (data not shown). Thus, it is reasonable to expect that many early developmental genes will not be DEGs. However, we note that lowering the stringency of DEG calling (when FDR-corrected p-value is increased to 0.05) does generate DEGs for these samples.

We found many DEGs between the melanocyte/iridophore progenitor samples (24hpf GFP+ samples) and differentiated pigment cells. We found a total of 2753 DEGs between the progenitor population and melanocytes (**Figure 3.9d)**; 2175 DEGs between the progenitor population and iridophores (**Figure 3.9e)**; and 1254 DEGs between melanocytes and iridophores. The breakdown of up- or down-regulated DEGs is depicted in **Figure 3.9b**.

Examination of specific genes involved in neural crest specification and melanogenesis confirms that our mRNA-seq libraries captured biologically meaningful expression patterns. *Sox10* is required for neural crest specification [125] and was expressed in the 14-somite GFP+ samples, which represent the premigratory neural crest, as well as in the 24hpf GFP+ samples, representing the migratory neural crest and pigment cell progenitors (**Figure 3.9f**). *ErbB3* is required for the establishment of direct-developing melanocytes [140], and we observed expression in the 24hpf GFP+ pigment cell progenitor population (**Figure 3.9g**). In zebrafish, *kit* is necessary for melanoblast migration and survival, but not melanocyte differentiation [165]. Appropriately, we find elevated *kit* expression specifically in the 24hpf GFP+ pigment cell progenitor population, and both the melanocyte and iridophore samples (**Figure 3.9h**). *mitf* is required for melanocyte differentiation and maintenance [135]; we observed increased *mitfa* expression in the progenitor pigment cells and in differentiated melanocytes (**Figure 3.9i**). Tyrosinase is the enzyme responsible for melanin formation in melanocytes, and we observed elevated *tyr* expression in pigment progenitor samples as well as melanocytes (**Figure 3.9j**). Lastly, *pnp4a* is a marker of iridophore development [136] and was specifically elevated in iridophore samples (**Figure 3.9k**).

3.6 Future Directions

3.6.1 Preliminary Conclusions

Based on the data generated and analyzed so far, we can start to answer some of the questions posed above in section **3.3**. First, how do DNA methylation dynamics change over pigment cell development? First, we observed no dramatic change in global levels of DNA methylation. Instead we find locus-specific changes in DNA methylation across the developmental time course. These dynamics resulted in a net decrease in loci-specific DNA methylation in each of the differentiated pigment cells. To explore this result further, it will be interesting to determine how many of these changes are localized to specific classes of regulatory elements, such as promoters, enhancers, or insulators.

Further, it will be important to incorporate other epigenome data into the global and loci-specific analyses. ATAC-seq data on these samples is still in preparation, and we anticipate that incorporating information about nucleosome dynamics over developmental time will add value to the DNA methylation dynamics documented here. For example, we found that the *mitfa* promoter is unmethylated in iridophores (**Figure 3.5b**) but *mitfa* is not expressed (**Figure 3.9i**). One explanation is that the *mitfa* promoter is repressed by an alternate epigenetic mechanism in iridophores. If we observe gain of nucleosome occupancy over the *mitfa* promoter in iridophores (e.g. loss of ATAC-seq signal) this would support our hypothesis that nucleosomes at the *mitfa* promoter are rearranged in the absence of promoter-repressive DNA methylation.

Another global analysis that will integrate WGBS and mRNA-seq data is to examine the DNA methylation status and expression of *crestin* elements. The *crestin* long terminal repeat (LTR) is ~1500bp and occurs at ~570 loci in the zebrafish genome. *crestin* is a marker of neural crest cells [158], and about half of annotated *crestin* elements have evidence of transcription (S. Higdon,

unpublished). However the function of *crestin* in the neural crest is unknown. Our unique dataset provides an opportunity to examine the regulation of *crestin* by DNA methylation. As one of the key roles of DNA methylation in vertebrate genomes is repression of transposable elements [13], understanding how and why *crestin* escapes DNA methylation may illuminate the function of this element and provide a more comprehensive picture of the neural crest cell epigenome.

Second, we asked how DNA methylation dynamics contribute to pigment cell fate acquisition. Again, we cannot fully answer this question yet. However preliminary analysis yielded some encouraging results. Even though, as noted above, pigment cells have a net gain of loci-specific DNA methylation, we observed loci exhibiting progressive loss of DNA methylation at promoters of key pigment cell genes. *mitfa* promoter exhibited loss of DNA methylation from neural crest cells and pigment progenitors to differentiated melanocytes (**Figure 3.6b**, **Table 3.1**). Similarly, the promoter of *purine nucleoside phosphorylase 4a* (*pnp4a*), a marker of iridoblasts and iridophores [166], was demethylated specifically in iridophores. It will be important to validate our WGBS datasets with more examples like these positive controls. In addition, uncovering DNA methylation events that are required for melanocyte or iridophore fate specification will be a key analysis to add to this project.

Last, how does DNA methylation interact with transcription factors to regulate pigment cell fate? To answer this question we will first annotate regions of dynamic DNA methylation to identify candidate regulatory elements. Loci that exhibit cell type- or tissue-specific demethylation are strongly enriched for regulatory elements [25,27,60,78]. Therefore, future analyses should include finding more cell-specific DNA methylation patterns, like those in **Figure 3.5b-c**. Such candidate regulatory elements will be analyzed *in silico* for transcription factor binding site motifs. A candidate-approach will prioritize motifs for TFs known to be important in pigment

cell development (mitf, foxd3, myc). An unbiased approach to find novel, important factors for pigment cell development will query motifs for factors that are expressed in the pigment progenitor cell and differentiated pigment cells, but as yet not known to be related to pigment cell development.

3.6.2 Future Data Generation

In addition to continuing to analyze our first round of data for this project, the quality of some datasets may necessitate regeneration or additional biological replicates. Our 24hpf GFP+ samples showed high global correlation (**Figure 3.7**). However, we found ~850 DMRs between biological replicates. Closer examination showed that ~75% of these DMRs are in repetitive regions. In addition, we found only ~2 million CpGs with ≥ 10x coverage in the 24hpf GFP+ replicate 2. These results likely explain why we found relatively few DMRs between the 24hpf GFP+ and the 14-somite or 24hpf GFP- methylomes (**Figure 3.9a**). (For reasonably pure cell populations we generally expect thousands of DMRs between samples, rather than hundreds.) Therefore it seems worthwhile to generate an additional biological replicate for this sample.

We also found fewer DMRs between the 14-somite GFP+ and 14-somite GFP- methylomes than we would expect for distinct populations (**Figure 3.9a**). We also found no differentially expressed genes between the 14-somite populations. One explanation for this lack of differentiation might be the poor fidelity of our *crestinA>GFP* reporter. During sample collection, we noticed over time (years) that *GFP* expression at the 14-somite stage became increasingly diffuse as we continued to breed the same fish to collect embryos for FACS. Loss of specificity due to epigenetic changes is known to occur with transgenes over time. A potential solution is to use an additional reporter construct to capture early neural crest samples. Our collaborator, Charles Kaufman, has generously offered a zebrafish line carrying

crestin>mCherry, which we will use to regenerate the 14-somite data after validating proper reporter gene expression pattern.

In addition, we note that while we have many biological replicates of pigment cell expression data, the reads counts are very low (**Figure 3.6**). Accordingly, it will be worthwhile to explore other analysis options, including combining the five and 11 iridophore and melanocyte replicates *in silico* to generate two synthetic "biological" replicates with increased read depth.

Finally, as mentioned above, we are working to generate ATAC-seq data for samples in this developmental time course. Josh Jang, a key collaborator on this project, has optimized the ATAC-seq protocol for zebrafish FACS-collected cells. ATAC sequencing data is anticipated in the coming months.

3.7 Methods

3.7.1 Zebrafish strains

Embryos used for neural crest cell isolation were transgenic for a construct driving expression of *EGFP* under control of a *crestin* fragment termed *crestinA*. *Crestin* is a retroelement in the zebrafish genome expressed specifically in the premigratory and migratory neural crest [158].

In the Stephen Johnson lab, approximately 1200 base pairs of a *crestin* element was cloned upstream of *EGFP* in a *Tol2* vector. The transgenic fish was created using *Tol2* transgenesis [167]. Resulting *crestinA* embryos drive *EGFP* expression starting at approximately the14 somite stage and throughout embryogenesis until the onset of melanogenesis.

5dpf *mlpha* larvae were used for pigment cell isolation. *mlpha* is a melanosome dispersion mutant that carries a loss-of-function mutation in the *melanophilin* gene, causing melanocytes to have a reduced dispersion of melanosomes [168]. We used the *mlpha* strain for pigment cell isolation for two reasons; (1) the GFP-fluorescence from the reporter line would interfere with pigment cell isolation (as described in 3.5.3), and (2) many *mlpha* fish were available for mating.

3.7.2 Neural Crest Cell Isolation

The following pigment cell isolation protocol and flow cytometry strategy was adapted from Higdon, 2013 [67].

14-somite or prim-25 stage embryos were dechorionated with 20mg/mL Pronase (Sigma) at room temperature, then rinsed with egg water and incubated on ice.

For 14-somite embryos, egg water was decanted and replaced with 100µL deyolking buffer (55mM NaCl, 1.8mM KCl, 1.25mM NaHCO3) for every 100 embryos. The egg water/deyolking buffer/embryos were spun down gently and the solution aspirated to remove residual egg water.

1mL deyolking buffer was added followed by gentle pipetting to generate a single cell suspension.

For prim-25 embryos, egg water was decanted and replaced with 1mL TrypLE Express (Gibco TrypLE Exrpress Enzyme (1X) Catalog no. 12604013) enzyme solution and incubated in a 37C heat block for 8-10 minutes in a 1.6mL eppendorf tube and inverted 2-3 times to encourage cell separation. After 8-10 minutes, the solution was triturated with a micropipette 10-20 times to mechanically dissociation remaining intact tissue.

Once embryos are brought to a single cell suspension, cells were pelleted in a tabletop fixed angle centrifuge at 300rcf for 8 minutes at 4C. The pellet was fully resuspended in 800µL cold 1XPBS + 2% fetal calf serum. The suspension was filtered through a 100µm mesh (Partec CellTrics® filters Order No. 04-004-2328) into a 15mL conical, adding buffer to flush all cells through the mesh as needed. Cells were pelleted in a fixed angle tabletop centrifuge at 300rcf for 8 minutes at 4C. Cells were resuspended in 1mL cold 1XPBS + 2% fetal calf serum and kept on ice until further processing. If using for FACS, 10µL 7-AAD was added at least 10 minutes before flow cytometry.

Embryonic cell suspensions were separated for GFP+ and GFP- populations using Fluorescence-Activated Cell Sorting (FACS) (**Figure 3.10**). The solution was resuspended and filtered one more time using a 100µm filter (Partec CellTrics®) immediately prior to sorting on the MoFlo cytometer using a 100µm nozzle (with the assistance of the Siteman Cancer Center Flow Cytometry Core Facility). Scatter properties were used to remove debris and doublets from the population. Cells were excited with a 488nm laser using filters for GFP (FL1) and Phycoerythrin (PE; FL2). Cells that were positive for GFP but negative for PE were collected as the GFP+

neural crest cell population. For increased stringency and specificity, the top 2% of the GFP+ population was collected as putative neural crest cells. The GFP/PE double negative population was collected as a GFP- control. Cells were kept on ice until further processing.

3.7.3 Pigment Cell Isolation

The following pigment cell isolation protocol and flow cytometry strategy was adapted from Higdon, 2013 [67].

5dpf embryos were anesthetized with Tricane and subsequently collected into 50mL conical tubes, ~500 embryos per 50mL egg water and stored on ice. Egg water was decanted out of each 50mL tube and refilled with TrypLE Express. 50mL conicals were then pooled into a 1L Erlenmeyer flask containing 500mL TrypLE Express and incubated in a 37C shaking incubator for 30 minutes (until larvae eyes could be seen floating in suspension). The larvae suspension was poured through a 120μm mesh. The supernatant was kept on ice while dividing into 50mL conicals. Cells were pelleted in a swinging bucket rotor (Eppendorf 5810 R) at 500rcf for 10 minutes at 4C. Pellets were resuspended in isotonic Percoll (1 part 10X PBS : 9 parts 100 Percoll (Sigma, Catalog No. P4937), iteratively transferring the resuspended pellet and Percoll to the next pellet for resuspsension; after reuspension, ~1mL isotonic Percoll contained all pigment cells in a single cell suspension.

Several 5mL columns of Percoll were pipetted into 15mL conicals. Then, the single cell suspension was overlayed on 5mL columns of isotonic Percoll, ~1mL of suspension per 5mL column, pipetting carefully. The overlay was spun at 1000rcf for 8 minutes at 4C in a swinging bucket rotor to separate cells by density. Pigment cells were contained in the pellet. If there was much debris in the pellet, the isotonic Percoll column separation step was repeated a second time.

The pigment cell pellet was resuspended in 300µL cold 1XPBS + 2% fetal calf serum and filtered through a 50µm mesh filter (Partec CellTrics® filters Order No. 04-004-2327) into a clean conical containing 500µL 1XPBS + 2% fetal calf serum and kept on ice until further processing.

The single cell suspension of enriched pigment cells was resuspended immediately prior to flow-cytometry. The MoFlo cytometer (with the assistance of the Siteman Cancer Center Flow Cytometry Core Facility) was used for separation of the pigment cell population based on the natural iridescent properties of the cells: iridophores reflect all light and autofluoresced when subjected to irradiation with the 488nm laser using filters for GFP (FL1) and Phycoerythrin (PE; FL2) (**Figure 3.11)**. Therefore the FL1/FL2 double positive population was collected as iridophores. Conversely, melanocytes absorb all wavelengths so were located at the bottom left corner of the FL1/FL2 scatter plots (for example, see **Figure 3.11b**). The FL1/FL2 double negative population was collected as a melanocyte-enriched population. After collection, cells were kept on ice in 1XPBS + 2% fetal calf serum until further processing.

3.7.4 Genomic DNA Isolation and Whole Genome Bisulfite Sequencing

Genomic DNA Isolation

After FACS, cells were pelleted using 500g for 10 minutes and the supernatant was discarded. Cells were then resuspended in 300µL of extraction buffer (50mM Tris (pH 8.0), 1mM 0. EDTA (pH 8.0), 0.5% SDS, 1mg/mL Proteinase K) was added to cell suspension and immediately pipetted with a 1000µL micropipette tip to disrupt cell membranes. The solution was incubated 12-16 hours in a 55C heat block, then centrifuged in a tabletop fixed angle centrifuge at max speed for 10 minutes at 4C. The supernatant was transferred to a 1.5mL Phase Lock Gel tube (5PRIME catalog no. 2302800) followed by one phenyl/chloroform:isoamyl alcohol (PCI)

extraction. The supernatant was removed and incubated with RNase for 30 minutes at 37C. Another PCI extraction was performed to remove RNase, followed by one chloroform-only extraction. The top phase was transferred to a fresh eppendorf tube for ethanol precipitation (Add 1/10 volume sodium acetate, 2.5x volume 100% ethanol, and 1µL of glycogen) overnight at -20C. The precipitation was spun in a fixed angle tabletop centrifuge at 16000g at 4C for 15 minutes. The pellet was washed in 70% cold ethanol, then centrifuged again at 16000g for 5 minutes at 4C. The pellet was air-dried and resuspended in 20-50µL molecular grade water.

WGBS Library Construction

WGBS libraries were constructed using the EpiGenome Methyl-Seq Kit (Epicentre; now Illumina TruSeq DNA Methylation, Catalog ID EGMK81312).

3.7.5 mRNA Extraction, cDNA Synthesis, and mRNA-seq Library Preparation

Total RNA Isolation

Total RNA was extracted using Trizol reagent. 1mL Trizol was added to cell pellet and immediately gently pipetted to homogenize. Solution was incubated 5-10 minutes at room temperature then spun down at 12000g for 10 minutes at 4C in a tabletop microcentrifuge. Supernatant was aspirated to remove debris. 0.2mL chloroform was added and the solution shaken for 15 seconds to mix, then incubated 2-3 minutes at room temperature. The suspension was spun at 12000g for 15 minutes at 4C. The upper aqueous phase was transferred to a fresh eppendorf tube. The 0.2mL chloroform extraction was repeated one more time.

RNA was precipitated by adding 0.5mL isopropanol to the aqueous phase. The solution was mixed by shaking and incubated at -20C for at least one hour. Suspension was spun at 12000g for 10 minutes at 4C and the supernatant aspirated. The pellet was washed with 1mL of 75% ethanol (made with RNase-free water) and spun at 7500g for 5 minutes at 4C. The wash step was

repeated up to twice as need to clean the RNA pellet. All ethanol was aspirated and the RNA pellet allowed to air dry at room temperature.

Cleaned total RNA was treated with TURBO DNase (Ambion, Catalog No. AM1907), according to manufacturer instructions. The final DNase-treated sample was resuspended in 30µL RNase-free water and kept on ice or stored at -20C.

mRNA-seq Library Construction
Total RNA was processed with TruSeq RNA library kit (Illumina, Catalog No. RS-122-2001) for preparation of mRNA-seq libraries according to manufacturer kit instructions.

3.7.6 WGBS Analysis

WGBS libraries were sequenced using NextSeq sequencing machines at the Washington University Center for Genome Sciences and Systems Biology Sequencing Center. Adapter sequences were trimmed using cutadapt [169], and reads were aligned against danRer10 using the Bismark aligner [113] and the following options: `-N 1 -L 28 --score_min L,0,-0.2`.

Differentially methylated CpGs and differentially methylated regions were called using the DSS package [160] in R 3.3.0. Differentially methylated CpGs (differentially methylated loci or DMLs) were called with smoothing, followed by calling DMRs with a dynamic window size and requiring a delta value of 0.25 and a p-value threshold of 0.01.

3.7.7 mRNA-seq Analysis

mRNA-seq libraries were sequenced using NextSeq sequencing machines at the Washington University Center for Genome Sciences and Systems Biology Sequencing Center. mRNA-seq reads were aligned against the Ensembl 80 *Danio rerio* transcriptome build using STAR [161].

Differentially expressed genes were identified using the edgeR package [163] in R 3.3.0. Genes with normalized counts ≥10000 or < 2 were removed to eliminate the effect of outliers and reduce noise. Pairwise differential expression was determined by using a generalized linear model to model gene expression and a quasi-likelihood F-test to determine significant expression. Only transcripts with a log-fold-change of ≥ 2 and FDR-corrected p-value of ≤ 0.001 (by Benjamini-Hochberg) were retained.

3.8 Data Access

mRNA-seq data for melanocytes and iridophores from Higdon, 2013 [67] were downloaded from GEO accession GSE46387 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46387).

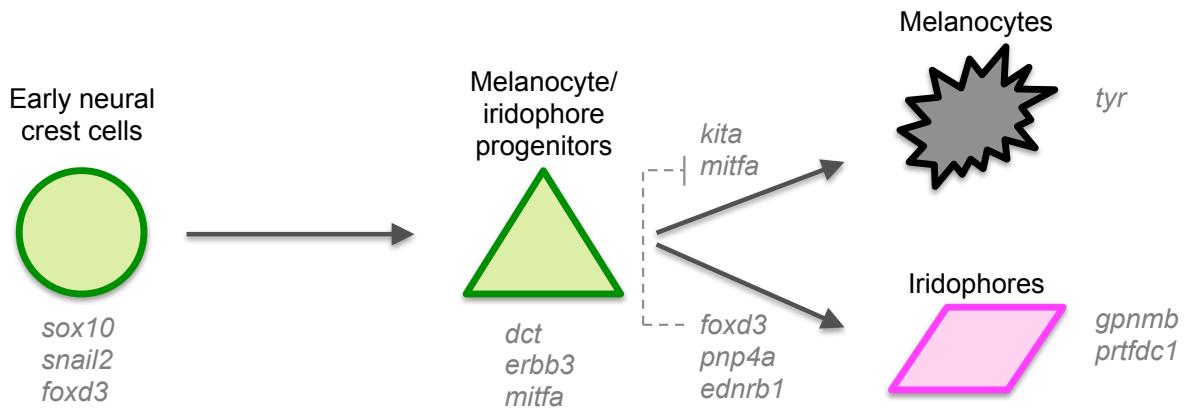


Figure 3.1. Pigment cell ontogeny. Early neural crest cells are represented by the green circle; melanocyte/iridophore progenitor neural crest cells are represented by the green triangle. The black star represents melanocytes and the magenta trapezoid represents iridophores. In our experimental design, the early neural crest cell stages were collected as 14-somite GFP+ cells; the 24hpf GFP+ population represents a melanocyte/iridophore progenitor-enriched population; melanocytes and iridophores were isolated from 5dpf larvae (see **Figure 3.2**). Arrows depict the developmental progression of pigment cells in the zebrafish embryo. In light gray are listed key genes expressed at specific stages. *foxd3* repression of *mitfa* is illustrated by the dotted gray line.

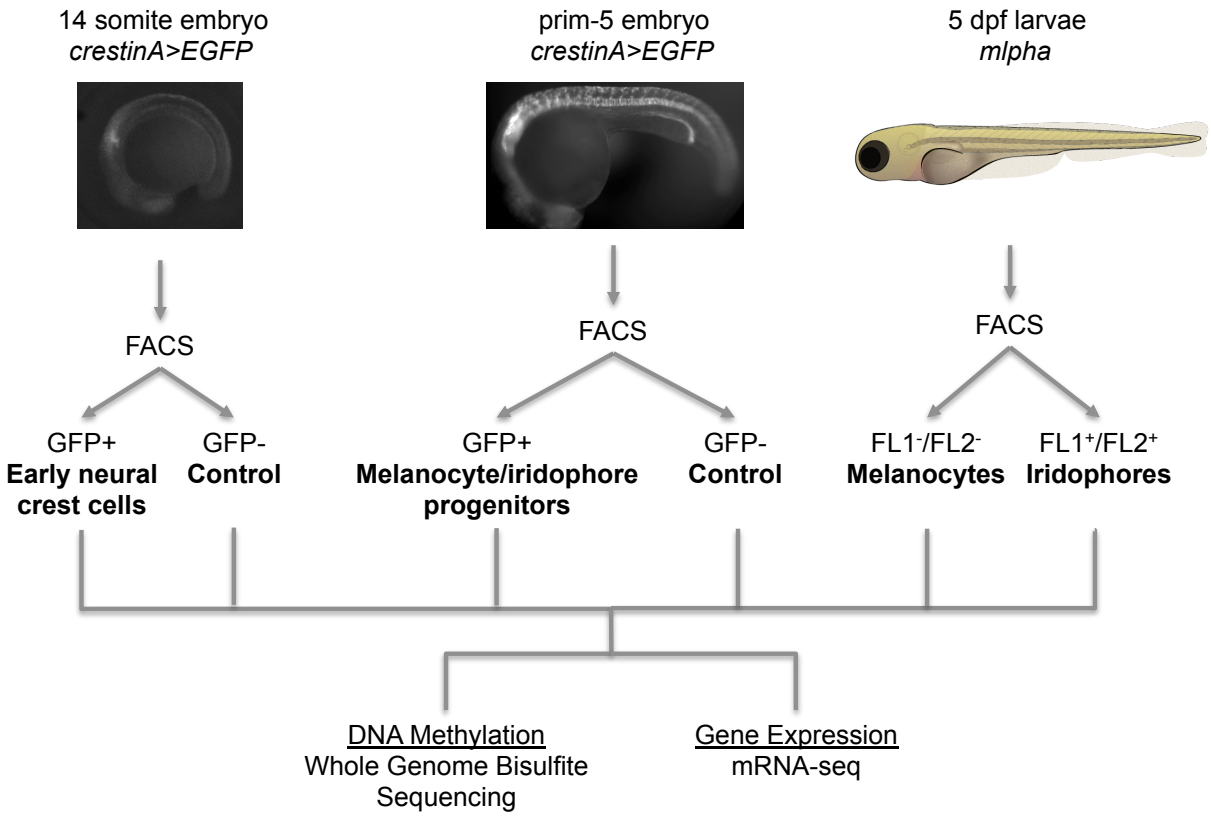


Figure 3.2. Experimental design. Embryos at the specified time point were collected, dissociated and sorted using fluorescence-activated cell sorting (FACS). For neural crest cell stages, GFP+ and GFP- control populations were collected. For the 5dpf embryo dissociated samples, FL1-/FL2- (FITC/PE channels; see Methods) samples were collected as melanocytes; FL1+/FL2+ samples were collected as iridophores (see **Methods**). Samples were processed for whole methylome or transcriptome analysis. The 14 somite and prim-5 embryo stages are depicted by an image of *crestinA>GFP* embryos at the specified stage. The 5dpf larvae image is credited to Lizzy Griffiths (http://zebrafishart.blogspot.com).

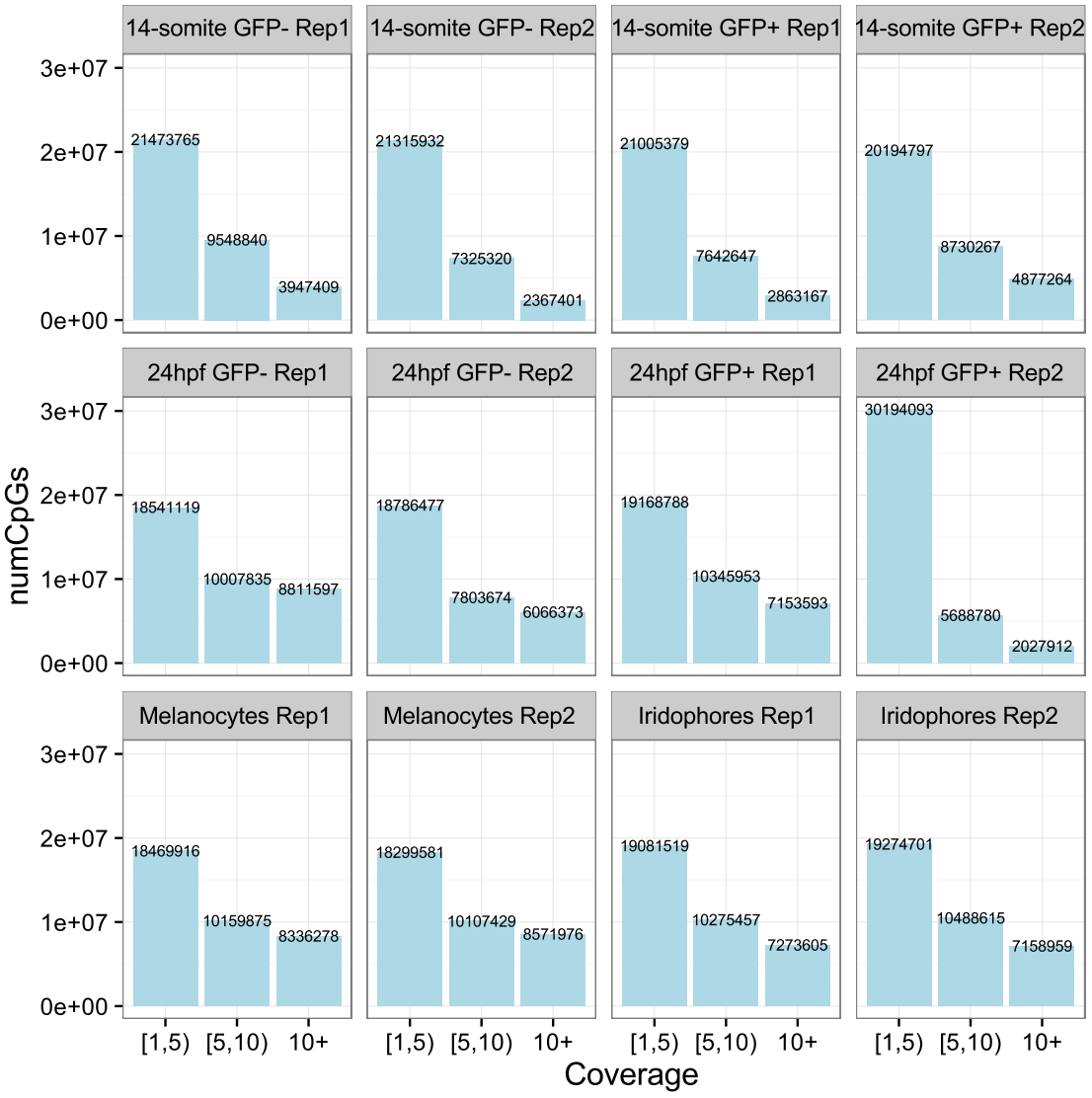


Figure 3.3. WGBS per CpG library coverage. Depth of coverage per CpG by library. Text labels state the number of CpGs with the given interval of coverage.

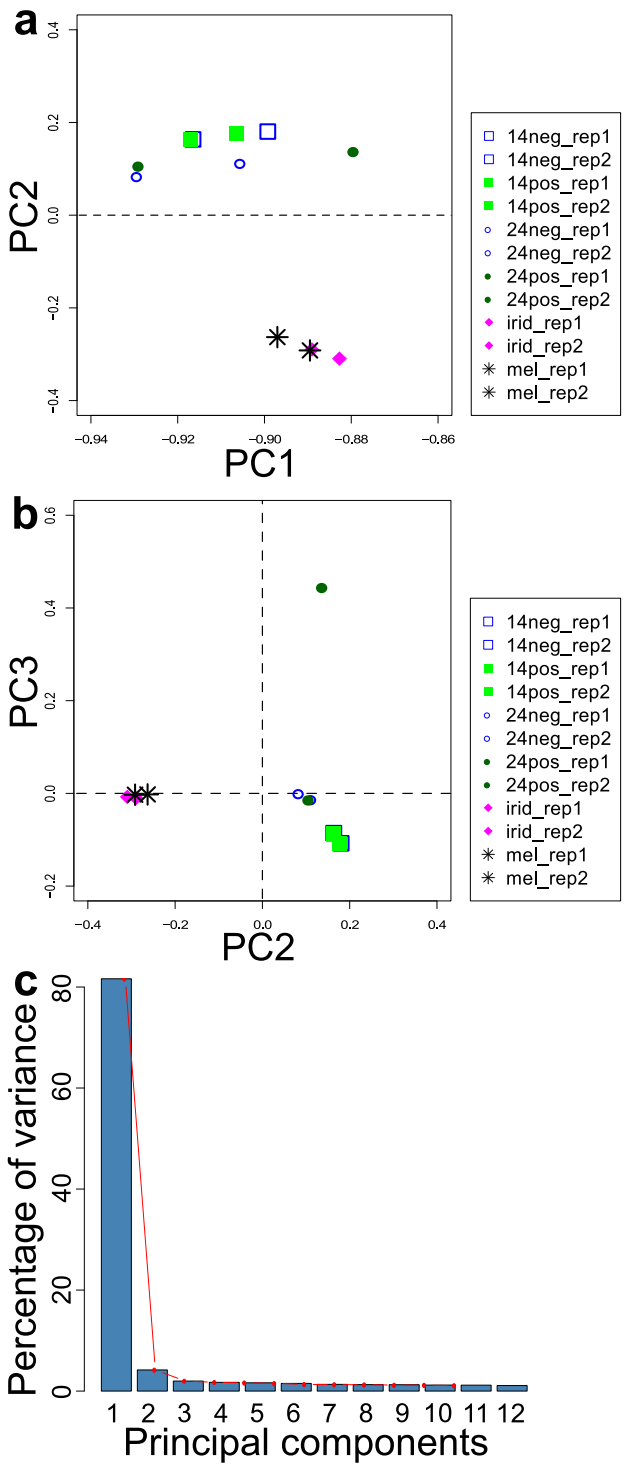


Figure 3.4. WGBS quality control. (a) PCA plot with principal components 1 and 2. (b) PCA plot with principal components 2 and 3. (c) Percentage of variance explained by each principal component.

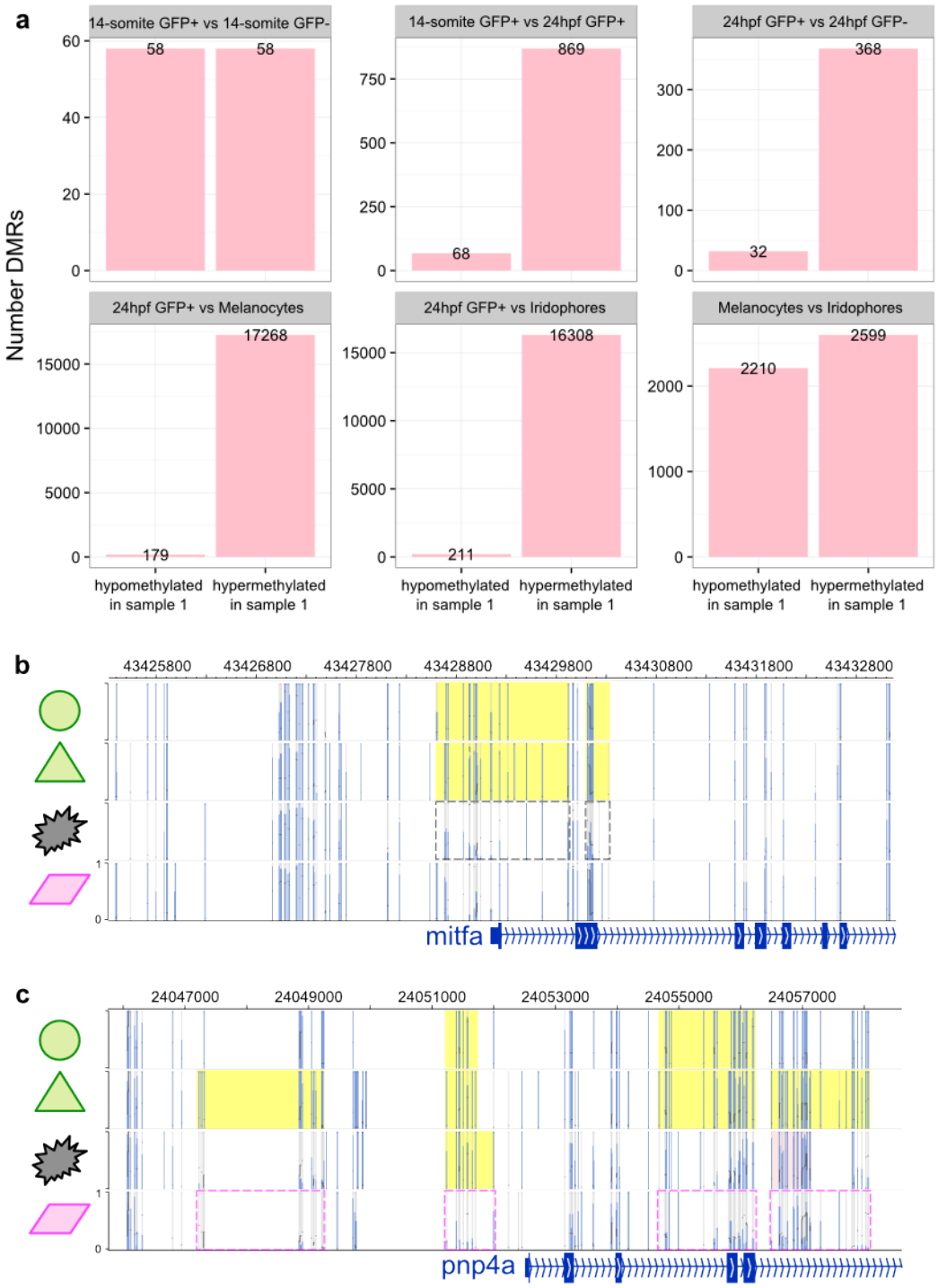


Figure 3.5. WGBS preliminary analysis results. (a) Bar graph of numbers of DMRs between pairwise comparisons that are biologically meaningful. The x-axis label refers to the direction of methylation change with respect to the first sample listed. (b) DMRs at the *mitfa* promoter show

demethylation dynamics between neural crest stages and pigment cells. Blue bars represent the DNA methylation levels at individual CpGs. DMRs that are hypermethylated compared to melanocytes are highlighted by yellow backgrounds in the precursor (hypermethylated) samples and gray dotted rectangles in the melanocyte sample. Tracks correspond to samples according to the cartoon on the left of each track, as depicted in **Figure 3.1**. DMR loci are listed in **Table 3.1**. (c) The *pnp4a* locus contains several DMRs in a 10kb region centered on the *pnp4a* promoter. Blue bars represent the DNA methylation levels at individual CpGs. DMRs that are hypermethylated compared to iridophores are highlighted by yellow backgrounds in the hypermethylated samples and magenta dotted rectangles in the iridophore sample. Tracks are as in (b). DNA methylation dynamics reflect the iridophore-specific nature of *pnp4a* expression (see **Figure 3.9k**). The DMR loci displayed here are listed in **Table 3.2**.

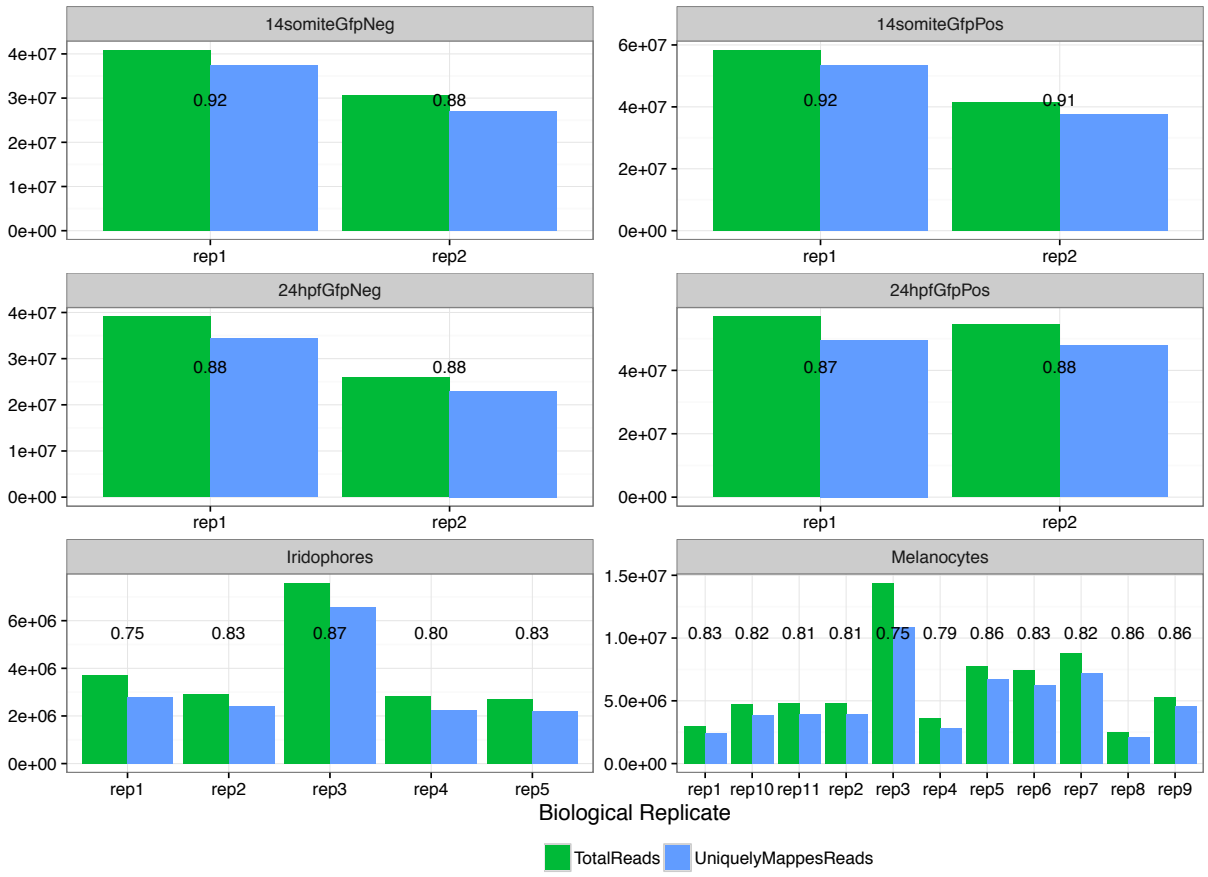


Figure 3.6. mRNA-seq mapping statistics. Total and uniquely mapped reads plotted as bar graphs, by sample type, by replicate. Text labels are the uniquely mapping reads rates.

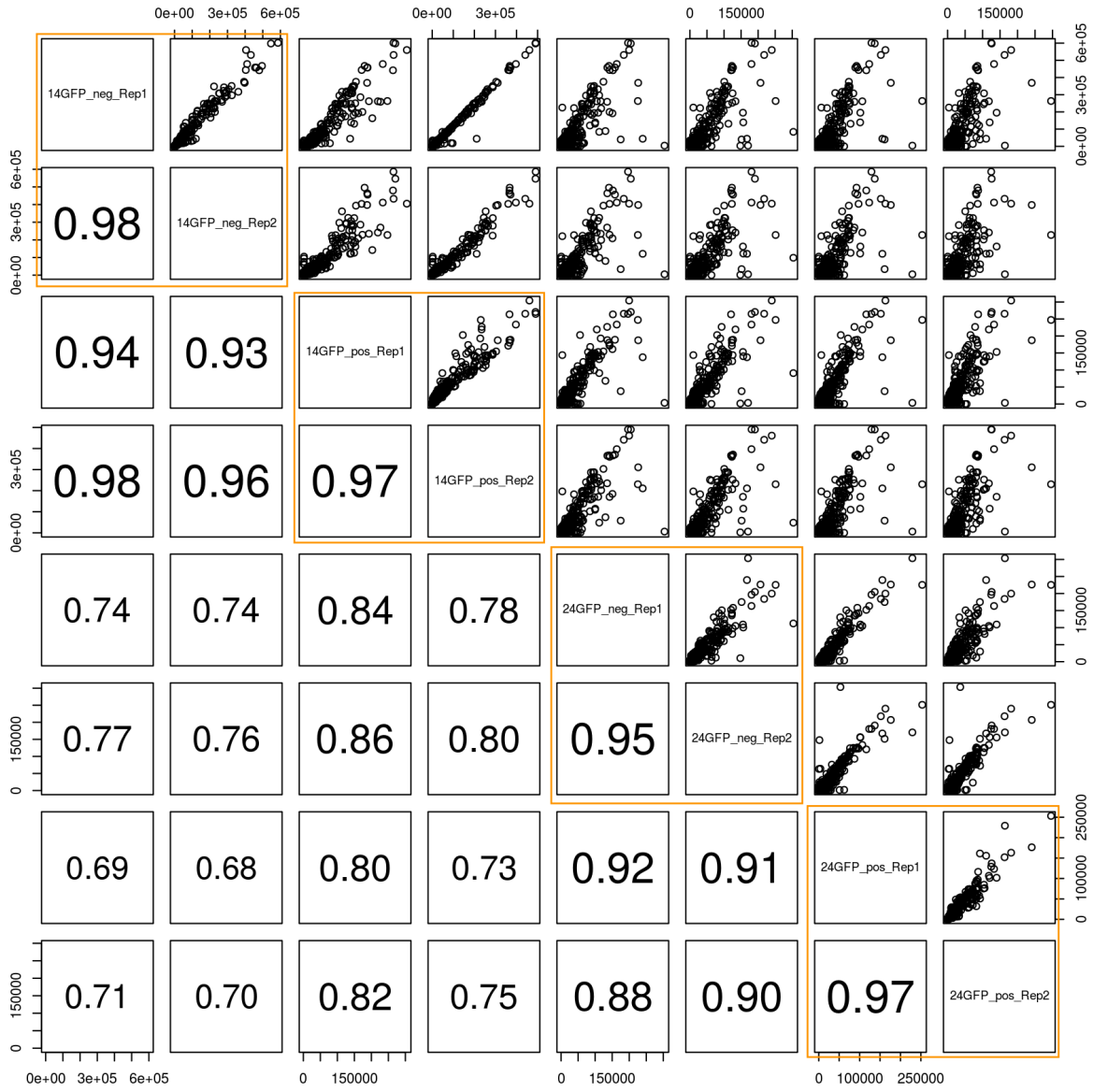


Figure 3.7. Gene expression levels pairs plots for early embryo stages. Pairs plot for gene expression level of processed mRNAseq libraries for 14-somite and prim-25 stages (labeled as 24hpf on plot). Filtering out genes with very high (≥10000 normalized read counts in at least one sample) or low (< 2 in at least one sample) was applied before plotting and subsequent analysis. In the upper panel scatterplots, each point in the scatterplots represents the expression levels of

one transcript in the two samples being compared. X- and y-axes are normalized read counts. In the lower panels, the decimal is the Spearman correlation between the two samples. Biological replicates are highlighted in orange boxes.

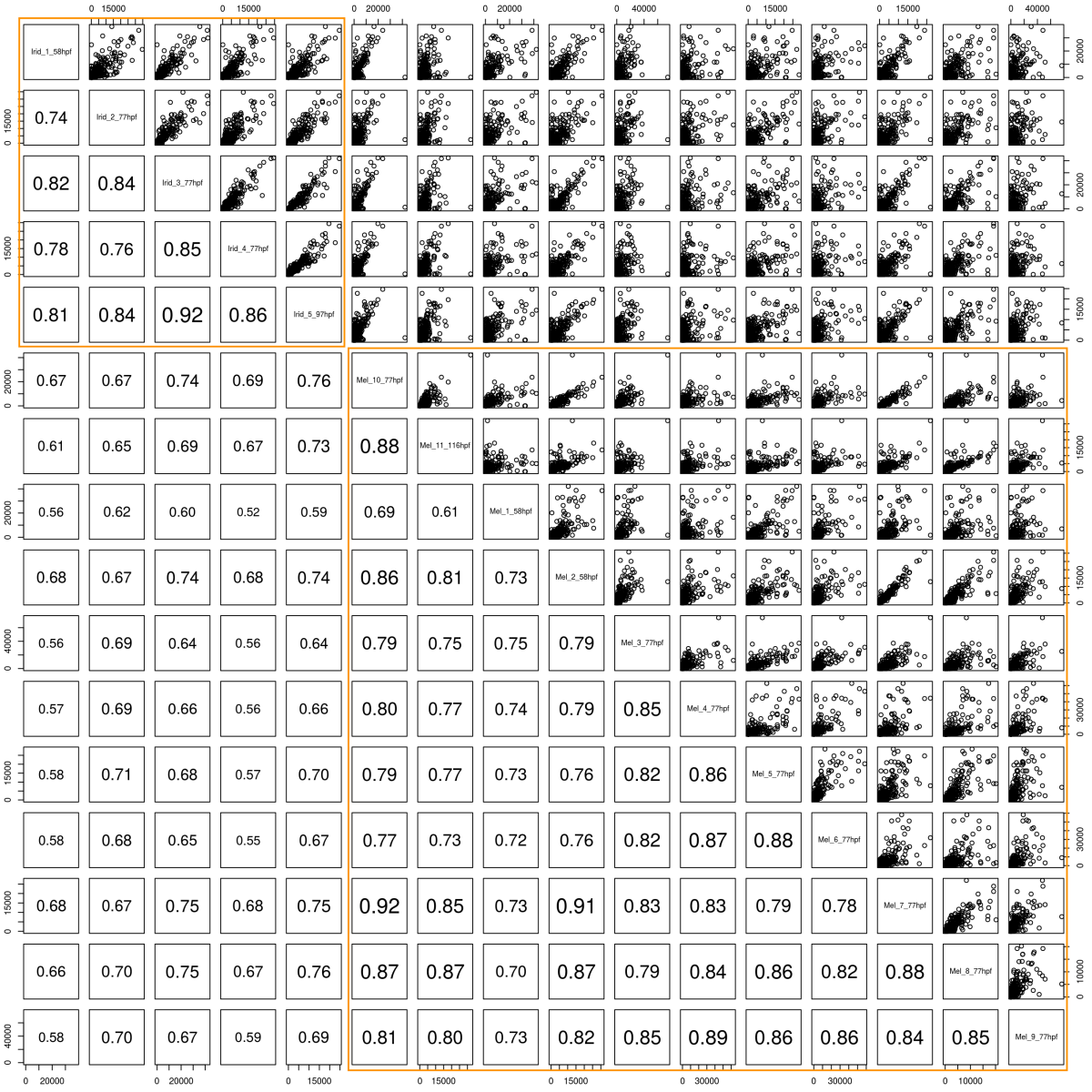


Figure 3.8. Gene expression levels pairs plots for pigment cells. Pairs plot for gene expression level of processed mRNAseq libraries for melanocytes and iridophores. Data processing, panels and axes are as in **Figure 3.7**. Biological replicates are grouped in orange boxes.

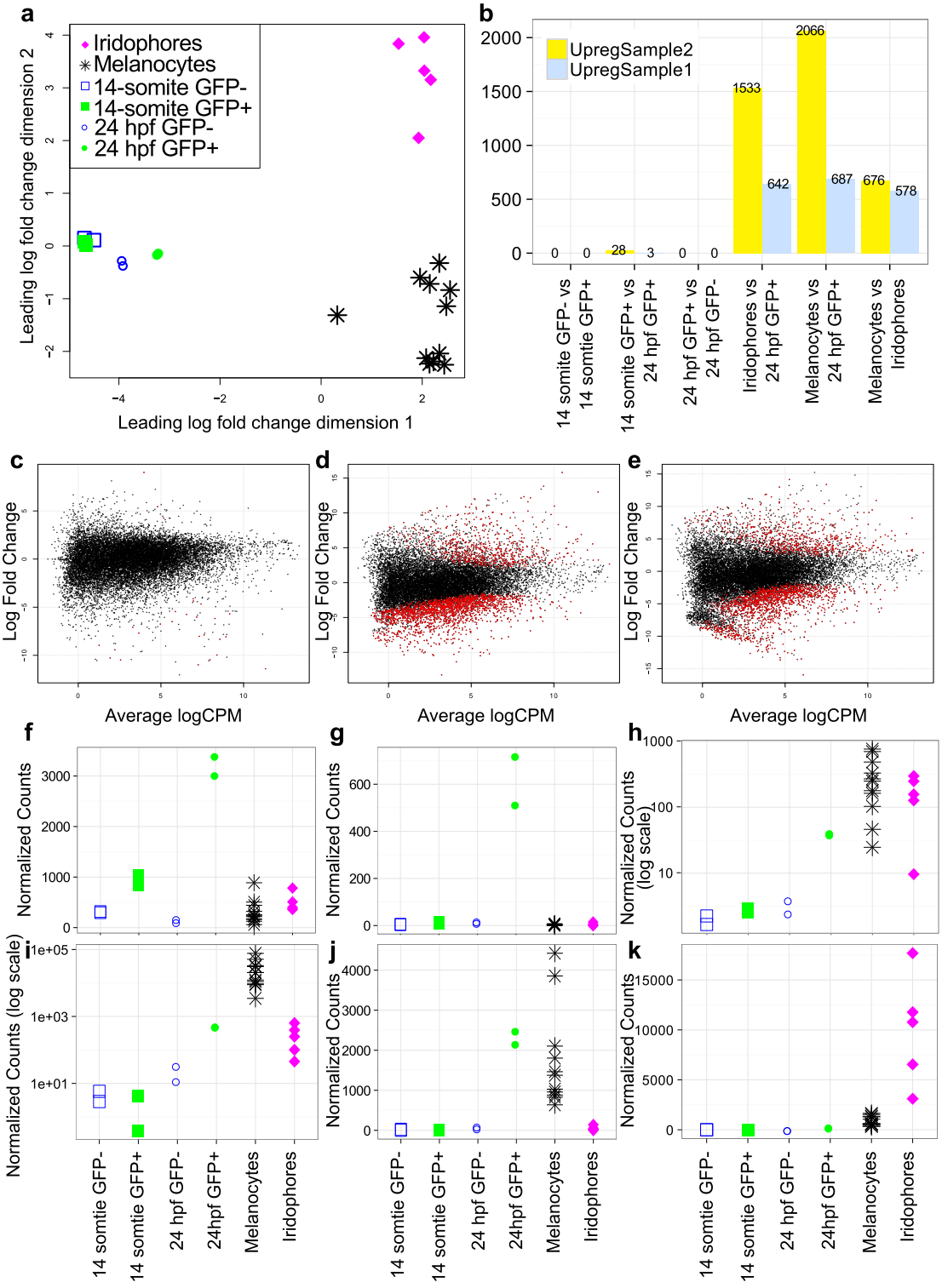


Figure 3.9. mRNAseq analysis summary. (a) MDS plot of mRNA-seq samples. (b) Number of differentially expressed genes between biologically meaningful sample comparisons. Yellow

bars represent the number of genes up-regulated in the first listed sample; blue bars are genes up-regulated in the second sample in the comparison. (c) Smear plot for differentially expressed genes between 14-somite GFP+ vs. 24hpf GFP+. Red dots are differentially expressed genes at a p-value cutoff of ≤ 0.001. (d) Smear plot for melanocytes vs. 24hpf GFP+. Red dots are differentially expressed genes. (e) Smear plot for Iridophores vs. 24hpf GFP+. Red dots are differentially expressed genes. (f-k) Expression of key genes in pigment cell differentiation. Legend is as in (a); y-axes are normalized read counts, log scale where indicated. (f) Expression of *sox10* in each mRNA-seq sample. (g) Expression of *erbb3*. (h) Expression of *kita*. (i) Expression of *mitfa*. (j) Expression of *tyr*. (k) Expression of *pnp4a*.

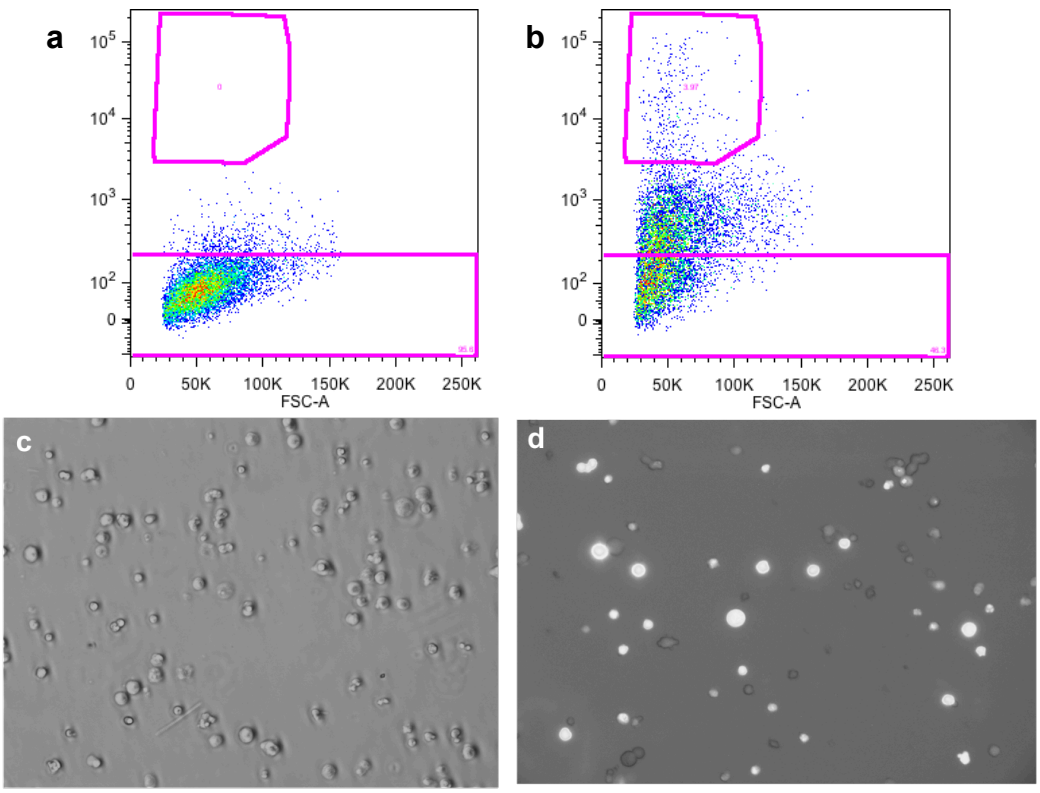


Figure 3.10. FACS separation of embryonic neural crest cells. (a) Representative FACS plot of wildtype control dissociated embryonic cells. X-axis is forward scatter, y-axis is FITC (GFP) fluoresence. (b) Final FACS plot of *crestinA>GFP* dissociated embryonic cells. The bottom gate is the GFP- control collected cells; the upper gate is the GFP+ collected cells. Axes are as in (a). (c) Merged brightfield and FITC fluorescence channels for GFP- cell sample (e.g. the bottom gate in (b)). (d) Merged brightfield and FITC fluorescence channels for GFP+ cell sample (e.g. top gate in (b)). White signal indicates GFP fluorescence.

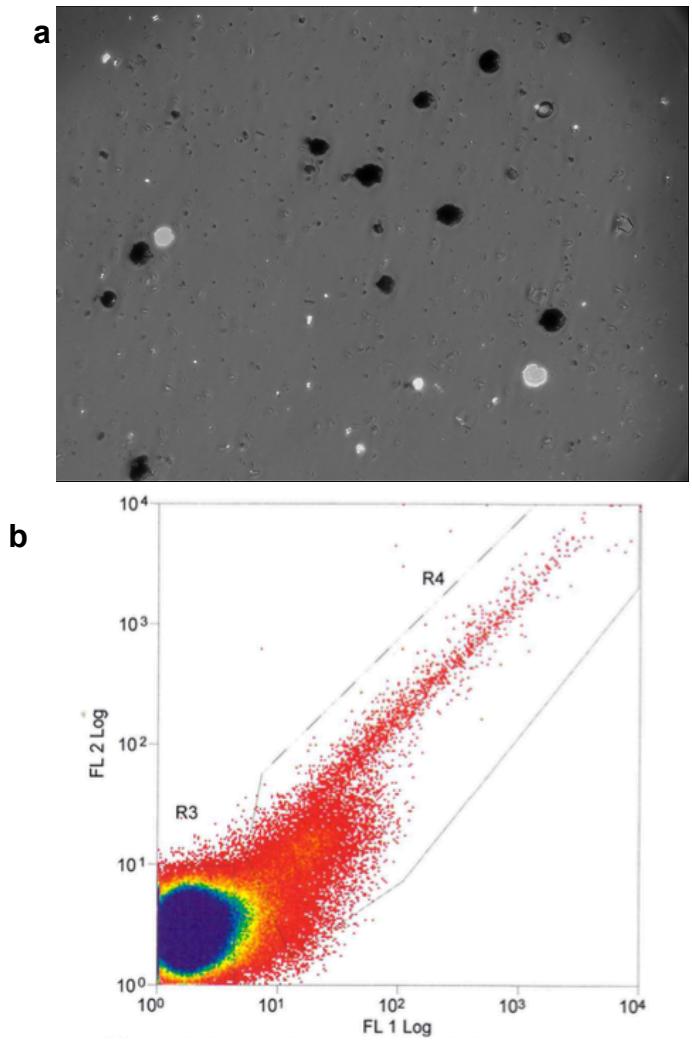


Figure 3.11. FACS separation of pigment cells. (a) Merged FITC and brightfield images of pigment cell isolation, before FACS. Melanocytes are the black cells; iridophores are seen as the reflective white cells in the FITC channel (white). (b) FACS plot of pigment cell suspension. FL1 is fluorescence in the GFP channel; FL2 is fluorescence in the Phycoerythrin (PE) channel. Melanocytes absorb all light and are FL1-/FL2- (R3 on plot). Iridophores reflect all light and are collected as the FL1+/FL2+ population (R4).

Table 3.1. DMRs at *mitfa* locus. DMR results using DSS software to identify DMRs at a dynamic window size. Regions are called DMRs between both replicates in each sample. In parentheses is the average DNA methylation level for both replicates. Num. CpGs = number of CpGs contained the DMR. Delta = the difference between the average DNA methylation levels at the given DMR in the samples being compared (sample 1 – sample 2). These loci are depicted in **Figure 3.9b**.

Locus	**Num. CpGs**	**Sample 1**	**Sample 2**	**Delta (S1 –S2)**
chr6:43428599-43429908	18	14-somite GFP+ (0.938)	Melanocytes (0.360)	0.578
chr6:43428599-43429908	18	24hpf GFP+ (0.851)	Melanocytes (0.361)	0.491
chr6:43430101-43430245	9	14-somite GFP+ (0.951)	Melanocytes (0.527)	0.424
chr6:43430101-43430245	9	24hpf GFP+ (0.965)	Melanocytes (0.527)	0.437

Table 3.2. DMRs at *pnp4a* locus. DMR results using DSS software to identify DMRs at a dynamic window size. Regions are called DMRs between both replicates in each sample. In parentheses is the average DNA methylation level for both replicates. Num. CpGs = number of CpGs contained the DMR. Delta = the difference between the average DNA methylation levels at the given DMR in the samples being compared (sample 1 – sample 2). These loci are depicted in **Figure 3.9c**.

Locus	**Num. CpGs**	**Sample 1**	**Sample 2**	**Delta (S1 –S2)**
chr11:24047238-24049228	16	24hpf GFP+ (0.923)	Iridophores (0.154)	0.769
chr11:24051398-24051658	6	14-somite GFP+ (0.960)	Iridophores (0.194)	0.767
chr11:24051398-24051658	6	24hpf GFP+ (0.40)	Iridophores (0.184)	0.746
chr11:24051398-24051716	7	Melanocytes (0.872)	Iridophores (0.194)	0.677
chr11:24054844-24056209	20	14-somite GFP+ (0.952)	Iridophores (0.286)	0.666
chr11: 24054844-24056209	20	24hpf GFP+ (0.907)	Iridophores (0.286)	0.621
chr11:24056862-24059392	35	24hpf GFP+ (0.943)	Iridophores (0.383)	0.560
chr11: 24056862-24057138	14	Melanocytes (0.913)	Iridophores (0.482)	0.431

Chapter 4

Epigenomic Annotation of Noncoding Mutations Identifies Mutated Pathways in Primary Liver Cancer

4.1 Author Contributions

This chapter is adapted from the manuscript submitted for publication to *PLoS Computational Biology*: Rebecca F. Lowdon[1], Ting Wang[1]. "Epigenomic annotation of noncoding mutations identifies mutated pathways in primary liver cancer."

R.F.L. conducted all data analysis, figure production, and manuscript writing; contributed to experimental design and manuscript editing. T.W. conceived the project design, provided intellectual contributions to the work; contributed to experimental design and manuscript editing.

[1] Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, MO 63108

4.2 Abstract

Evidence that noncoding mutations can result in cancer driver events is mounting. However, it is more difficult to assign molecular biological consequences to noncoding mutations than to coding mutations, and a typical cancer genome contains many more noncoding mutations than protein-coding mutations. Accordingly, parsing functional noncoding mutation signals from noise remains an important challenge. Here we use an empirical approach to identify putatively functional noncoding somatic single nucleotide variants (SNVs) from liver cancer genomes. Annotation of candidate variants using publically available epigenome datasets finds that 40.5% of SNVs fall in regulatory elements. When assigned to specific regulatory elements, we find that the distribution of regulatory element mutation mirrors that of non-synonymous coding mutation, where few regulatory elements are recurrently mutated in a patient population but many are singly mutated. We find potential gain-of-binding site events among candidate SNVs, suggesting a mechanism of action for these variants. When aggregating noncoding somatic mutation in promoters, we find that genes in the ERBB signaling and MAPK signaling pathways are significantly enriched for promoter mutations. Altogether, our results suggest that functional somatic SNVs in cancer are sporadic, but occasionally occur in regulatory elements and may affect phenotype by creating binding sites for transcriptional regulators. Accordingly, we propose that noncoding mutations should be formally accounted for when determining gene- and pathway-mutation burden in cancer.

4.3 Author Summary

Cancer develops after multiple mutations to a cell's genome confer a growth advantage on that cell, enabling it to outcompete other cells. Only about 1.5% of the DNA in our genome codes for proteins, the molecules that carry out biological functions in the cell. 50 years of genetics research has focused on understanding the molecular nature of this 1.5% of our genome, and the consequences that occur when protein-coding DNA fragments are mutated. The remaining ~98% of the genome is the "noncoding" genome, and our understanding of mutations in noncoding DNA is less mature. Yet anecdotal reports suggest that noncoding mutations can initiate a variety of cancer diseases, including melanoma, glioma, and liver cancer. Here we identify regions of noncoding mutation that may be important in primary liver cancer by utilizing recent advances that map the functions of the noncoding genome. Notably, we find that noncoding mutations accumulate at promoters (the "on" switch) of genes known to be involved in liver cancer. Our results suggest that mutations in the noncoding genome provide a more complete picture of altered biology in cancer and accordingly should be accounted for in patient diagnosis and therapy.

4.4 Introduction

Cancer genomics suffers from a dramatic signal to noise problem, where the majority of somatic mutations are not expected to cause cancer phenotypes, but to be passenger mutations that do not contribute to selective growth advantage [170-172]. The challenge of identifying mutations that change cancer phenotype is especially difficult in the noncoding genome: whereas over 50 years of molecular genetics research has given cancer investigators a toolkit for understanding the deleteriousness of coding mutation, the same code book does not exist for noncoding mutations. Instead, anecdotal instances of oncogenic noncoding mutations in the cancer literature include a variety of mechanisms, including transcription factor binding site creation (or deletion) by point mutation [69-71,173,174], enhancer hijacking by structural rearrangements [72,175], or abrogation of chromatin neighborhoods by disruption of cohesion binding sites [176]. Considering this mechanistic diversity, we aim to increase our sensitivity for recovering functional noncoding mutations by focusing our analyses on point mutations that may appropriate regulatory elements from heterologous cell types by creation of transcription factor binding sites.

As the importance of regulatory variation has become illuminated [41,177] several tools for detecting deleterious noncoding mutation have been developed in recent years. These tools implement empirical scoring algorithms and machine learning approaches to determining functional noncoding variants. These algorithms use a combination of negative selection [178,179], mutation recurrence [178], and/or functional element annotation data [179-182] (e.g. from the ENCODE Project [41]) to predict noncoding variant significance [183]. In the study presented here, we expand noncoding variant annotation to include the wealth of epigenomic data, now publically available, by resources such as the Roadmap Epigenomics Project

[177,184]. Epigenome annotation data allow us to investigate the hypothesis that somatic mutations might activate transcriptional regulatory programs not native to the tumor cell type of origin.

One model of regulatory element-mediated oncogenesis in the literature is the cancer enhancer model (Figure 4.1). In the cancer enhancer model, coding mutations can have oncogenic effects by mis-regulation of the epigenome. For example, mutation of chromatin modifier genes (for example, *mixed lineage leukemia* (*MLL*) family genes) may adjust the affinity of transcriptional activators for cognate enhancers, driving over-expression of a proto-oncogene [185]. Alternately, in a tumor suppressor context, mutated chromatin modifiers may reduce affinity of trans-activators for enhancers, in either case leading to tumor progression [185].

Analogously, we propose the *cis*-cancer enhancer model, whereby somatic mutation of regulatory elements changes their regulatory potential (**Figure 4.1**). The *cis*-cancer enhancer model predicts that functional noncoding mutation may activate transcriptional regulatory programs intrinsic to heterologous cells. In our model, noncoding somatic mutation might change the regulatory potential of an element by creating a binding site for a DNA-binding protein, subsequently allowing the protein to bind DNA and recruit other chromatin modifiers. Such activity is reminiscent of pioneer factor action, as has been demonstrated to occur in the context of breast cancer mutations that modulate FOXA1 binding [173].

Accordingly, here we use epigenomic annotation from diverse cells and tissues to test the hypothesis that noncoding mutation activates regulatory elements used in heterologous cells. We find that after filtering, approximately 40.5% of noncoding variants fall in transcriptional regulatory elements. Subsequently, we find widespread potential gain or loss of transcription

factor binding sites, suggesting specific mechanisms by which noncoding mutation may influence cancer phenotype and progression. Lastly, we find that noncoding regulatory mutations in primary liver cancer (PLC) accumulate in promoters for genes involved in transcriptional misregulation in cancer, ERBB signaling, and MAPK signaling pathways.

Genome-wide studies of regulatory mutation in cancer have analyzed noncoding mutation from a pan-cancer perspective [186-188]. These studies have found repeatedly a limited set of candidate noncoding variants that are responsible for phenotype in the pan-cancer context. Fewer have queried the effect of noncoding mutation in cancer on a single disease basis [68,73,189-193]. In the present study, we aimed to increase our specificity by focusing on a single disease. We chose to study PLC for two reasons: first, normal liver tissue is relatively homogeneous, making determination of regulatory elements easier. Second, there are many publically available liver cancer samples, and a large sample size is necessary in order to detect rare events.

4.5 Results

4.5.1 Isolating Putatively Functional Noncoding SNVs

The Catalog of Somatic Mutations in Cancer (COSMIC) project houses publically available cancer genetics data [194]. The repository includes data from a variety of diseases and various assay types (e.g. whole genome resequencing, ExomeSeq). For the present work, we used the noncoding variants dataset from the COSMIC Genomes project.

To isolate putatively functional noncoding SNVs in the COSMIC dataset, we took a stringent filtering approach (**Figure 4.2.a; Methods**). After isolating noncoding SNVs from primary liver cancer (PLC) samples, we removed variants at positions of known population variants and kept only variants that were confirmed somatic (e.g. not observed in the matched normal genome) and that were discovered from whole genome resequencing (WGS) (**Methods**). We focused our analysis on WGS-derived variants because we wanted an unbiased interrogation of somatically mutated genome-wide regulatory elements.

Next, we determined the distribution of noncoding SNVs per sample ID in COSMIC. Hypermutator phenotypes occur when DNA repair genes have been inactivated and DNA mutation occurs unchecked [195]. To remove noise due to hypermutation, variants from samples with the top 2.5% of SNVs/sample were removed (7 samples with 79817 total SNVs; **Figure 4.3a**; **Methods**). This noncoding SNV filtering strategy resulted in 7893 noncoding SNVs from 235 unique liver cancer samples in the COSMIC database.

The same strategy applied to ExomeSeq noncoding variants returned 1,477,249 noncoding SNVs from 789 unique liver cancer samples (**Figure 4.3b-c**).

All analyses were run on filtered WGS and ExomeSeq SNV sets separately; however all results are for WGS-SNVs unless otherwise noted.

4.5.2 Genome Feature Annotation of Noncoding SNVs in Liver Cancer

Annotating noncoding SNVs by the UCSC known genes annotation set revealed that noncoding somatic mutations were markedly enriched in UTRs and promoters (**Figure. 4.2b**). Promoters and UTRs are sites with a high density of regulatory elements. Thus, noncoding SNVs that passed our filtering strategy were likely enriched in genome regions that host regulatory features.

4.5.3 Epigenomic Annotation of Noncoding SNVs in Liver Cancer

The Roadmap Epigenome Project generated reference epigenomic datasets for 111 primary human cell types and tissues [177]. Among the data generated were chromatin immunoprecipitation-sequencing (ChIP-seq) for various histone modifications. Histone ChIP-seq data for each tissue were then synthesized by the ChromHMM algorithm to produce a genome-wide annotation of epigenomic status [177]. Other experiments included DNaseI-hypersensitivity sequencing and were conducted on a subset of tissues.

DNaseI hypersensitive regions are enriched for transcriptional regulatory elements such as enhancers and promoters [51]. To validate that noncoding SNVs delivered by our algorithm were likely to be regulatory, we analyzed the SNV locations in the context of the Roadmap DNaseI hypersensitivity site (DHS) data. The catalog of DHS regions was collected from the 39 Roadmap Epigenomes for which data were available, and the ChromHMM promoter or enhancer status of these DHS positions was queried in all 111 Roadmap Epigenome primary cell types. Notably, the single primary liver sample in the Roadmap Project did not have DNaseI hypersensitivity in the pan-Roadmap DHS site catalog. However, we wanted to determine if non-liver regulatory element accumulated PLC noncoding mutations. Therefore, we partitioned the

DHSs into cell type-shared or cell type-restricted regions, as determined by the Roadmap Project analysis of DHS data (**Methods**). Then we assigned each SNV location to a DHS if it fell in a DHS peak as called by the Roadmap Project (**Methods**).

Noncoding somatic PLC SNVs that passed filtering were found in DHSs were annotated as promoters more often than random expectation (**Figure 4.2c**). Both cell type-shared and cell type-restricted DNaseI-promoters were somatically mutated more than expected (2.06- and 1.88-fold over expectation based on background, respectively). The enrichment for SNVs in cell type-restricted DNaseI-promoters indicates that promoters not specific to liver sustain regulatory mutations in PLC. Enrichment of cell type-shared promoters reflects mutation of promoters for genes that are constitutively expressed. On the other hand, both cell type shared and cell type restricted DNaseI-enhancers were slightly depleted for somatic mutations (0.62-fold and 0.84-fold compared to background expectation respectively). It is likely that the low fold enrichment for DNaseI-enhancers was due to the large expected value, as DNaseI-annotated enhancers accounted for a large percentage of genome base pairs.

4.5.4 PLC SNVs are Enriched in Bivalent Chromatin Features

We suspected analyzing enhancer chromatin states in finer detail would provide a more nuanced picture of the patterns of somatic regulatory mutation. Thus, we analyzed the filtered noncoding PLC SNVs in the context of the ChromHMM-18 state model for Roadmap Epigenome Project primary tissues. We tabulated the occurrence of liver cancer SNVs in each ChromHMM-18 state in each of the 78 cells and tissues for which data were available and compared this value to the expected number of SNVs, assuming a random mutation distribution (**Figure 4.2d**; **Methods**). Strikingly, we found elevated observed/expected values across most tissues analyzed in regulatory ChromHMM states, including active promoters (1_TssA), flanking promoter regions

(2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD), genic enhancers (7_EnhG1, 8_EnhG2), and bivalent states (14_TssBiv, 15_EnhBiv), which have regulatory potential. Surprisingly, active enhancer states (9_EnhA1, 10_EnhA2) did not have elevated observed/expected values across most cell types. Again, this was likely because these enhancer states occupied a large fraction of the genome (34% of merged epigenome base pairs were annotated as potential enhancer state (active, weak, genic, and bivalent enhancer states) versus 8.4% annotated as potential promoter (active, flanking, and bivalent) (**Methods**).

Specifically in liver annotations, we found elevated observed/expected values in active and flanking promoters states, genic enhancers, and bivalent states. The strongest enrichment was for the bivalent transcription start site (TSS) and bivalent enhancer states. Bivalent chromatin is best understood in the embryonic stem cell context, where simultaneous modification of nucleosomes by activating (H3K4me3) and Polycomb-repressive (H3K27me3) histone modifications is thought to keep promoters in a "poised" state until the cell further differentiates [196]. The function of bivalent domains in differentiated cells is less understood, but may enable the cell to respond quickly to environmental stimuli [197,198].

Finding elevated SNVs at bivalent enhancers and promoters prompts the hypothesis that these liable regulatory sites may be central to transcriptional mis-regulation in PLC. For example, dysregulation of bivalent promoters has been shown to lead to oncogene activation in colorectal tumors [199]. Indeed, dysregulation of bivalent domains is a reported phenomenon in cancer genomes [200]. In a process called "epigenome switching," the Polycomb-deposited repressive histone modification (histone 4 lysine 27 trimethylation) is aberrantly replaced by DNA methylation, which is relatively more stable [201]. It would be interesting to explore if the

accumulation of SNVs in bivalent domains is mechanistically linked to recruitment of DNA methyltransferases to these regions in cancer.

Altogether, we find that 40.5% (3200/7893) of SNVs were found in regulatory elements from 78 cell types and tissues genome-wide. Thus, analysis of candidate somatic noncoding mutations in epigenetically defined regulatory elements supports our hypothesis that noncoding somatic mutation may influence cancer phenotype by modulating regulatory elements.

4.5.5 Patterns of Noncoding Somatic Mutation in Regulatory Elements Mirrors that of Coding Mutations in Genes

Coding mutations in cancer display a stereotypic distribution across genes, where a few genes are recurrently mutated across patients, while a long tail of genes is rarely mutated [171]. This is true for most cancer types, even though the identity of the highly or lowly-mutated genes varies depending on the disease [186,202]. We hypothesized that the distribution of putatively functional regulatory element mutations might mirror the pattern of coding mutation. Indeed, plotting the number of candidate somatic mutations from the COSMIC PLC samples for each regulatory element mapped revealed a striking distribution: one regulatory element is mutated in 16 patients, two regulatory elements are mutated in 7 patients each, and a long tail of individual elements are mutated in 1, 2, or 3 patients (**Table 4.1**). The most-mutated regulatory element is the *TERT* promoter, which was mutated 16 times at the same position in the ETS binding site, as has been previously reported in the literature [203].

We sought to connect the candidate noncoding liver mutations to putative target genes. First we assigned the candidate SNVs to regulatory elements, epigenetically defined by the Roadmap Project (**Methods**). Next we assigned each SNV-containing regulatory element to putative target gene promoters (using a +/-35kb window [109]; **Methods**). Based on these target gene

assignments, we asked if some target genes have an elevated rate of mutated regulatory elements. We queried the collection of target gene regulatory elements -- their promoters and putative distal enhancers -- and tabulated the number of somatically mutated regulatory elements associated with each gene (**Table 4.2**). The distribution is qualitatively similar to that of coding mutations in cancer, where in a patient population, a few genes have several noncoding somatic mutations in their regulatory elements, while a long tail of genes have only one mutated regulatory element.

Three genes had three putative regulatory elements with noncoding somatic mutations. One of these was *C1orf61* (**Figure 4.2e**), which has been characterized as a tumor activator in hepatocellular carcinoma [204]. *C1orf61* is located on 1q22, which experiences copy number amplifications in several cancers including hepatocellular carcinoma [204]. Investigation of the effect of upregulation of *C1orf61* revealed that it was correlated with liver disease and HCC progression, and ectopic expression of C1ORF61 promoted cell proliferation, metastasis, and EMT [204].

In our analysis, each of the three somatically mutated *C1orf61* regulatory elements was found in three unique samples. Importantly, these samples were not recorded with 1q22 amplifications in the COSMIC database, indicating that noncoding regulatory mutation may upregulate *C1orf61* in hepatocellular carcinoma in a similar tumorigenic manner as copy number amplification. We examined The Cancer Genome Atlas expression data for PLC samples and matched normal tissue [188] and found that *C1orf61* expression was elevated in a subset of tumors (**Figure 4.2f**).

Epithelial splicing regulatory protein 1 (*ESRP1*) also had three SNV-containing putative regulatory elements (**Figure 4.2g**). *ESRP1* can promote the epithelial-to-mesenchymal transition

(EMT) by regulating alternative splicing of *CD44* [205]. Knockdown of *ESRP1* activity in breast cancer cells restored the non-EMT-inducing isoform of *CD44* and suppressed metastasis [206], evidence that *ESRP1* acts as an oncogene. *ESRP1* acts as a master regulator of EMT in melanoma [207] and somatotroph adenomas [208]. However, upregulation of *ESRP1* is correlated with fewer metastasis and better prognosis in pancreatic ductal adenocarcinoma [209], and acts a tumor suppressor in colorectal cancer [210], reflecting the cell type-specific nature of cancer genes [202].

The filtered PLC SNVs contained three mutations in regulatory elements whose putative target was *ESRP1*. TCGA expression data from PLC and matched normal samples showed that 27% of tumors had elevated expression of *ESRP1* (**Figure 4.2h**).

The gene with the most somatically mutated regulatory elements was *MAP2K1*, part of the mitogen-activated signaling pathway, which is a central regulator of cell growth. The five *MAP2K1* regulatory elements found mutated in our data set contained seven unique mutated positions in seven samples. At time of writing, *MAP2K1* has not been directly implicated in liver cancer; however the MAPK signaling pathway has been identified as important for PLC [211,212]. *MAP2K1* has been identified as an occasional driver in non-small cell lung cancer [213], and sustained gain-of-function mutations in melanoma [214]. Variation among genes in the MAPK pathway predisposes to colon and rectal cancer, including susceptibility variants in *MAP2K1* [215].

4.5.6 Regulatory Element-Annotated SNVs Cause Gain-of-Binding Site Events Upstream of Known Oncogenes

Since our hypothesis was that noncoding somatic mutations might activate transcriptional regulatory programs from heterologous cell types, we predicted that functional noncoding

mutations in regulatory elements should result in gain-of-function genetic events. Such events may be evident as gain-of-binding site motifs for transcriptional trans-activators.

To test this prediction, we conducted a systematic analysis of somatic SNVs in regulatory elements to look for gain-of-binding site events. First, we queried the COSMIC Cancer Gene Census for transcription factors (termed CGC-TFs), of which there were 93. For these factors, we searched the JASPAR and TRANSFAC motif databases for motifs that are bound by the cognate CGC-TFs; 106 motif position weight matrices (PWMs) were found, including motifs for heterodimers. Finally, for each of the 106 motif PWMs we constructed a position-specific scoring matrix (PSSM) and determined the threshold PSSM value for a false-positive rate of 0.001 (**Figure 4.4a; Methods**).

We then analyzed each SNV from the filtered COSMIC noncoding variant set that occurred in a regulatory element for its ability to modulate the motif PSSM score. For each SNV, we generated *in silico* wildtype and mutant alleles, using hg19 as the reference (wildtype) allele. Each pair of alleles was scored against each CGC-TF PSSM to obtain a log-odds ratio score compared to a background of genomic nucleotide frequencies; only scores passing the CGC-TF-specific threshold were retained.

We determined the delta value for each pair of PSSM scores by subtracting the mutant allele score from the wildtype score (**Figure 4.5a,b**). To enrich our dataset for events with high effect size, we kept only pairs of CGC-TF motif scores where at least one score (wildtype or mutant) was log odds score over background ≥ 2. The resulting distribution reveals that 1234 pairs of wildtype-mutant alleles from whole genome-resequenced samples create potential gain-of-binding site events, in which the mutant allele score is higher than the wildtype allele score for a

particular CGC-TF (**Figure 4.4b**; **Figure 4.5c**). 1393 allele pairs represent potential loss-of-binding events, where the wildtype allele score was greater than the mutant allele score. Allele pairs residing in promoter regions from ExomeSeq samples resulted in 25600 and 29410 gain and loss of binding sites, respectively. Thus we find a substantial number of potential gain-of-binding site events from candidate noncoding somatic mutations.

We examined the gain-of-binding site candidates for evidence of oncogenic events. The mutation event with the highest effect size in our dataset was a noncoding mutation in the last intron of *ZFAS1* lncRNA (**Figure 4.6a**). The *ZFAS1* mutated position is annotated as a genic enhancer in human Mammary epithelial cells (vHMEC) cells by ChromHMM. This T>G mutation creates a strong JUND binding site where the reference sequence is less likely than background to bind JUND (wildtype allele = -0.12; mutant allele = 14.75). Importantly, *ZFAS1* is known to promote metastasis in hepatocellular carcinoma [216,217]. *ZFAS1* is a regulator of normal mammary gland development, where it inhibits miR-150, which in turn inhibits *ZEB1* [216], a regulator EMT [218]. When *ZFAS1* is upregulated in HCC, it acts as a sponge to decrease the concentration of miR-150, thereby upregulating *ZEB1*, which induces tumor cell invasion and metastasis in *in vitro* and animal models [217].

Since many SNVs from whole genome-resequenced PLC samples did fall in promoter regions, and promoters are often captured in ExomeSeq data, we expanded the motif mutation analysis to promoter ExomeSeq variants from COSMIC PLC samples. Among the ExomeSeq SNVs, we find a COSMIC patient sample with an A>T mutation in the *FGF5* promoter that creates a MYC binding site (**Figure 4.6b**). The somatic mutation creates a binding site where the reference sequence is slightly less likely than background to bind MYC (wildtype allele = -0.2; mutant

allele = 12.9). *FGF5* is a known oncogene in glioblastoma where it promotes proliferation and inhibits apoptosis [219].

Thus, at least two known oncogenes were recovered in our gain-of-binding candidate somatic mutations. These events suggest that noncoding mutation may mimic oncogenic coding mutations by upr-egulating proto-oncogenes. Importantly, such gain-of-function mutations may occur at regulatory elements not annotated in the cancer tissue-of-origin (in this case liver) but in regulatory elements active in other cell types (for example, *ZFAS1* in breast tissue).

4.5.7 Noncoding Mutations Add to Pathway Level Mutation Burden

An important aspect of cancer genomics is that deleterious mutations can inactivate a pathway at several points [202]. For example, in colorectal cancer, *BRAF* mutations are mutually exclusive with mutations in *KRAS* [220], indicating that a single alteration of the activity of a pathway member is sufficient to induce misregulation of that pathway. We suggest that the positions of deleterious somatic mutation can be used to probe pathways affected by somatic mutation. When considering the noncoding genome, we hypothesized that accumulation of noncoding somatic mutation in the transcriptional regulatory regions of genes belonging to a single pathway may indicate pathways with a significant noncoding mutation load in a population of liver cancer patients.

To identify pathways with significant noncoding mutation burden, we first obtained cancer-related pathways as reported in the pan-cancer literature [202] and in liver cancer-specific reports [211,221]. For each pathway, gene lists were collected from publically available databases [222-224]. We then used SNVs assigned to promoters to tabulate the genes hit by somatic regulatory mutation in liver cancer, and identified pathways with a significant noncoding regulatory mutation load in the population of samples tested (**Figure 4.7**; **Methods**).

In the ExomeSeq data, the most significantly hit pathway was "Transcriptional misregulation in cancer" (KEGG; p-value = 2.67e-11) (**Figure 4.7**, blue box), a positive result. The next most significant pathway hit was MAPK signaling (p-value = 3.81e-6) (**Figure 4.7**, purple box; **Figure 4.8**). This result was consistent with the finding that five *MAP2K1* regulatory elements were mutated (see above). Additionally, the MAPK pathway is a central regulator of cell growth, so mis-regulation of the MAPK pathway in cancer is not surprising: our data suggest that noncoding mutation may impact MAPK pathway function. Last, the ERBB signaling pathway was significantly mutated (p-value = 1.14e-4) (**Figure 4.7**, green box; **Figure 4.9**).

SNVs from whole genome resequenced PLC samples had fewer pathways significant hit, as the sample size was much smaller. However pathway hits were consistent with the larger, ExomeSeq SNV set. The MTOR signaling pathway was the most significant pathway mutated in this sample set (p = 8.10e-4) (**Figure 4.7**, gold box). This pathway shares several gene members with the ERBB pathway. Additionally, the ERBB signaling pathway was just under the threshold for significance for the WGS SNV set, after correcting for multiple-testing. We anticipate that more samples would replicate the ERBB enrichment result seen for the ExomeSeq SNV set.

4.6 Discussion

Cancer is initiated by sequential somatic mutation until a cell acquires a selective growth advantage and becomes malignant [170,172,225]. Most characterized somatic mutation is to coding genes, either activating proto-oncogenes or inactivating tumor suppressor genes, and is readily identified by sequence-based methods that detect changes to open reading frames. However, the majority of somatic mutation occurs in noncoding regions [68,171]. Identifying the small fraction of noncoding somatic mutation that has a phenotypic effect remains a challenge, as changes to noncoding regulatory DNA are less straightforward to interpret.

While difficult to detect, mounting evidence suggests that noncoding somatic mutations can act as cancer drivers. Amplification of a locus hosting a proto-oncogene is a common oncogenic mechanism: the *ERBB2* locus is amplified in breast cancer [226] and *EGFR* in glioma multiforme [227,228]. Similarly, amplification of a super-enhancer drives overexpression of oncogenes such as *MYC* and *KLF4* in epithelial cancers [229]. Other structural rearrangements place an enhancer near novel oncogenes, such as *GFI1* and *GFI1b* in subtypes of medulloblastoma [72]. Point mutations can also be detrimental, especially in solid tumors [202]. Point mutations that abrogate cohesion binding sites disrupted chromatin neighborhoods, resulting in mis-regulation of proto-oncogenes by enhancers in neighboring chromatin neighborhoods in T-ALL [176]. In addition, point mutations may create transcription factor binding sites near oncogenes, as has been well-documented at the *TERT* promoter in melanoma, breast cancer, liver cancer, and other diseases [69,71,230-232].

Here we describe an algorithm for filtering noncoding somatic mutation data to arrive at potentially functional SNVs. Our algorithm relies on an empirical measure of hypermutation to remove extremely noisy cancer genomes. Subsequently, epigenomic annotation of variants

informed which variants had the potential to modulate transcriptional regulatory states: we found 40.5% of filtered variants occurred in regulatory states in one of the 78 Roadmap Project primary cell and tissue types analyzed. SNVs in liver cancer kept from our filtering method were enriched in regulatory states, especially active promoter states, genic enhancers, and bivalent enhancers and promoters.

The distribution of functional coding mutations per gene in a population tend to be highest in a few, specific genes that vary by disease, while many genes will be infrequently mutated in a population [186,233]. Genes highly recurrently mutated in a disease population are expected to be potent cancer drivers. Alternately, low-frequency recurrently mutated genes are thought to drive cancer by mitigating specific pathways; that is, a single pathway may be mutated in several different ways (by mutation of different genes) across individual patients [170,234]. We hypothesized that noncoding mutation may follow a similar pattern.

We were not surprised that the *TERT* promoter mutation remained the strongest signal in terms of mutation recurrence. However, by continuing to probe the publically PLC samples, we were able to find new, moderately strong signals, including recurrent regulatory mutations for *C1orf61*, *ESRP1*, and *MAP2K1*. By assigning SNV-containing regulatory elements to putative target genes, we showed that the distribution of noncoding mutations in regulatory elements for specific genes qualitatively mirrors that of coding mutations.

Pathway level analysis is increasingly an important way to interpret cancer mutations [171,189,202,235]. Genes with a low frequency of coding mutations in a population can still have a functional effect in an individual, and aggregating these low-frequency mutated genes has been used to identify pathways deregulated in hepatocellular carcinoma [212,221,236-238]. To

ask if noncoding mutations accumulated across samples at regulatory elements for genes of specific pathways, we examined somatically mutated promoters in the context of cancer-involved biological pathways. We found significant involvement of mutated promoters for MAPK signaling, ERBB signaling, MTOR signaling, and transcriptional mis-regulation in cancer pathways.

The result of our pathway analysis is consistent with literature that reports MTOR and MAPK pathway activation in HCC [237]. Hepatocyte proliferation is spurred in cirrhotic liver cells by activation of the MAPK pathway via transforming growth factor-α or insulin-like growth factor-2 [239]. ExomeSeq studies of HCC samples have also identified the mTOR and MAPK pathways as significantly enriched for coding mutations [212,221,236]. Indeed, both the mTOR and MAPK pathways are well known to be involved in several cancers via coding mutation [171,202].

Our analysis suggests noncoding mutations might burden the same pathways as coding mutations. In the future, it will be important to explore new, unanticipated pathways that have a high somatic noncoding mutation load. Additionally, including distal enhancers in this analysis can increase the sensitivity and specificity of analyzing regulatory element mutation burden effects at a pathway level; however more robust and reliable distal regulatory element to target promoter assignment is needed for the analysis to have a reasonable signal to noise.

One way noncoding mutation can influence phenotype is by altering transcriptional regulation, for example, by modulating transcription factor binding site affinities. We found that 15.6% of whole genome resequenced candidate SNVs created putative gain-of-binding site events while 17.6% resulting in potential loss-of-binding site events. Thus, once a stringent algorithm was

applied to noncoding mutation, a significant amount of noncoding mutation had a potential effect on transcriptional regulation. Our method recovered transcriptional regulatory alterations at known oncogenes (*FGF5*) and at cell biological pathway genes that are important for tumor cell biology (*ZFAS1* and tumor cell invasion).

As we gain a better understanding of how noncoding somatic mutation alters transcriptional regulation, it will be important to incorporate noncoding somatic mutation information into algorithms that predict network-level mutation burden [240]. Eventually, such information might better inform differential diagnosis and therapeutic recommendations.

4.7 Methods

4.7.1 Filtering COSMIC Noncoding Variants

COSMIC v77 noncoding variants file <CosmicWGS_NCV.tsv.gz> and the sample metadata file <CosmicWGS_SamplesExport.tsv.gz> were downloaded from the COSMIC database (http://cancer.sanger.ac.uk/cosmic) (13 July 2015) [194]. Noncoding SNVs were then parsed as follows (see also **Figure 4.1a**):

1. Using custom python code, filter variants for:
    1.1. Variant's sample ID had primary site metadata for as "liver"
    1.2. Variant not annotated as known variant position in (e.g. in dbSNP or 1000 Genomes; see ref. [194])
    1.3. Variant is a confirmed somatic mutation (e.g. was not observed in matched normal sample)
    1.4. Variant is from a whole genome resequenced sample
2. Then find the distribution of variants per sample. Based on the distribution:
    2.1. Define hypermutated samples as those above the percentile on the ordered set of SNVs / sample where the rate of change between percentiles is the greatest (0.5% resolution). This was the top 2.5% samples.
    2.2. Remove variants from hypermutated samples.

A similar strategy was used for filtering ExomeSeq derived variants by modifying step 1.4 above (**Fig. 3b-c**).

4.7.2 ChromHMM-18 Enrichment

ChromHMM-18 Segmentation Data

ChromHMM-18 segmentations by the Roadmap Project on hg19 were downloaded from the Roadmap Project data repository (http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state; mnemonics bedfiles archive). Each of the 78 (non-ENCODE cell lines) mnemonics bed files were parsed using custom python code for each EID and each ChromHMM-18 state.

Calculating observed, expected values of filtered noncoding SNVs in ChromHMM-18 segmentations

For the set of 78 EIDs' ChromHMM-18 bedfiles, bedops annotateBed function was used to determine the overlap of filtered noncoding SNVs with each ChromHMM-18 state. The total expected SNVs in state *m* in cell type (EID) *n* was calculated using custom R code as in Equation 4.1.

$$E_{m,n} = total\ SNVs \times \frac{total\ ChromHMM\ annotated\ bp\ for\ state\ m\ in\ cell\ type\ n}{total\ ChromHMM\ annotated\ bp\ in\ cell\ type\ n} \quad \text{(Equation 4.1)}$$

Then the total observed SNVs in state *m* in cell type *n* was tabulated and compared to expectation to create plot in **Figure 4.1d**.

4.7.3 DNaseI Shared Versus Restricted Regulatory Elements

Delineation of DNaseI-accessible regulatory regions" data were downloaded from the Roadmap Epigenome Project data repository (http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation; RData files (hg19 coordinates)). Shared or restricted determination for each DNaseI region was made using the k-centroid clustering algorithm results provided by Roadmap (text files for order of modules at the same URL). Overlap of filtered COSMIC noncoding SNVs with regions in each DNaseI cluster was done in R using GRanges package and custom R code.

4.7.4 Regulatory Element Annotation

Cosmic noncoding SNVs kept after filtering were annotated using the UCSC Known Genes track and the GenomicFeatures R package functions and custom R code.

4.7.5 Assigning Noncoding Regulatory SNVs to Target Gene Promoters

SNV-to-regulatory element assignment

First we constructed a merged regulatory epigenome: The merge of all 78 ChromHMM-18 states was compiled (for autosomes only). For each 200bp window in the ChromHMM-18 annotations, a regulatory classification of enhancer, promoter, transcribed, or inert was given based on observations in the 78 ChroMHMM-18 annotations. Priority was as follows: assignment to enhancer states (states 7,8,9,10,11, and 15); promoter state (states 1,2,3,4, and 14); transcribed states (states 5 and 6); inert states (states 12,13,16,17, and 18). Filtered SNVs were assigned to overlapping ChromHMM-18 state regulatory element annotations (enhancer and promoter state regions only) using bedtools. Adjacent regulatory elements were merged and the total number of Cosmic noncoding SNVs / full-length element counted using a custom python script. Regulatory elements multiply mutated in the same sample ID were counted as mutated twice, except in the case of adjacent SNVs, which were counted a single nucleotide mutation.

Regulatory element-to-target gene assignment

The TxDb.Hsapiens.UCSC.hg19.knownGenes R package was used to construct a transcript database (TxDb) of UCSC known genes. Filtered PLC SNVs were assigned regulatory elements using custom python code. SNV-regulatory elements assignments were read into into R as GRanges object. The start and end of the regulatory elements' intervals were extended by +/- 35kb [109] and overlap with UCSC known promoters was found using the GenomicRanges package mergeByOverlaps function.

4.7.6 Motif Mutation Analysis

Identifying cancer-related TFs and their motifs

Searched PUBMED for transcription factors using the search terms "("transcriptional activator" OR "transcriptional repressor") AND ("transcription factor") AND ("DNA-binding") AND "Homo sapiens"[porgn:__txid9606]". The resulting list of transcription factor genes was cross-listed the PUBMED-TF set with Cancer Gene Census list [194]. The resulting CGC-TFs list was queried to against JASPAR [241] and TRANSFAC [242] databases to find any motif that is bound by CGC-TFs (106 motifs including heterodimers). For each CGC-TF motif, the position-specific scoring matrix (PSSM) was determined using Biopython tools [243], and threshold PSSM was determined at FPR = 0.001.

Motif scanning on wildtype and mutated allele sequences

Sequences were generated for wildtype (hg19 reference) and tumor alleles using custom python code and Biopython modules. For each allele, and for each CGC-TF motif, the log-odds PSSM score that the allele creates the given motif site compared to background nucleotide frequencies was determined using Biopython tools and custom python code. Only PSSM scores above the CGC-TF-specific FPR threshold were kept.

Data were then curated to keep only predicted motif-altering instances with a reasonable effect size: pairs of alleles must have had a PSSM log-odds score >= 2 in at least one allele. The delta value was computed for each pair of wildtype-mutant alleles where delta = mutant allele score – wildtype allele score.

4.7.7 Pathway Analysis

For each set of SNVs (WG resequenced or ExomeSeq derived), SNVs were filtered to retain only those in UCSC Known Gene promoter regions (-2000bp, +500bp from TSS). Gene names of these promoters were retained. List of pathway gene members was downloaded from the

Molecular Signatures database (MsigDB) [224] (v5.1); pathways selected were from the KEGG [222] or Amigo [223] databases. The retained genes list was intersected with each pathway gene list, and the number of overlapping genes were counted as "hits".

4.7.8 Binomal Test

A one-sided binomial test was conducted using R for each pathway overlap hits count, where k = number of overlapping genes, n = number of promoters hit by SNV set, p = corrected length of pathway gene list / promoters for UCSC Known Genes ("corrected" as some of the gene symbols in the downloaded pathway gene lists were not present in the UCSC Known Genes track). Bonferroni-correction was used to determine significant p-values.

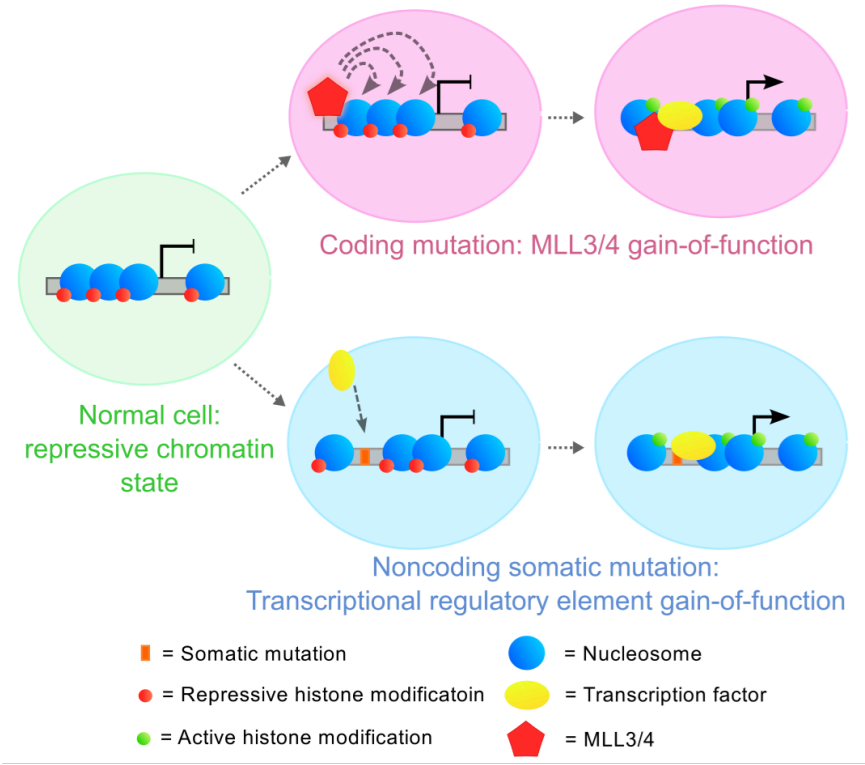


Figure 4.1. Models for regulatory element involvement in cancer. In the *trans*-model of cancer enhancers, somatic mutation to a chromatin modifier gene, here *MLL3/4* (red pentagon), results in that chromatin modifier binding more tightly to a DNA-bound transcription factor (yellow oval) and aberrantly creates a persistently open chromatin state, up-regulating the target gene. In the *cis*-model of cancer enhancers, a somatic mutation to a noncoding regulatory element (orange bar) creates the same open chromatin state, perhaps by creating a binding site for a transcription factor that is recruited to the locus and facilitates opening local chromatin.

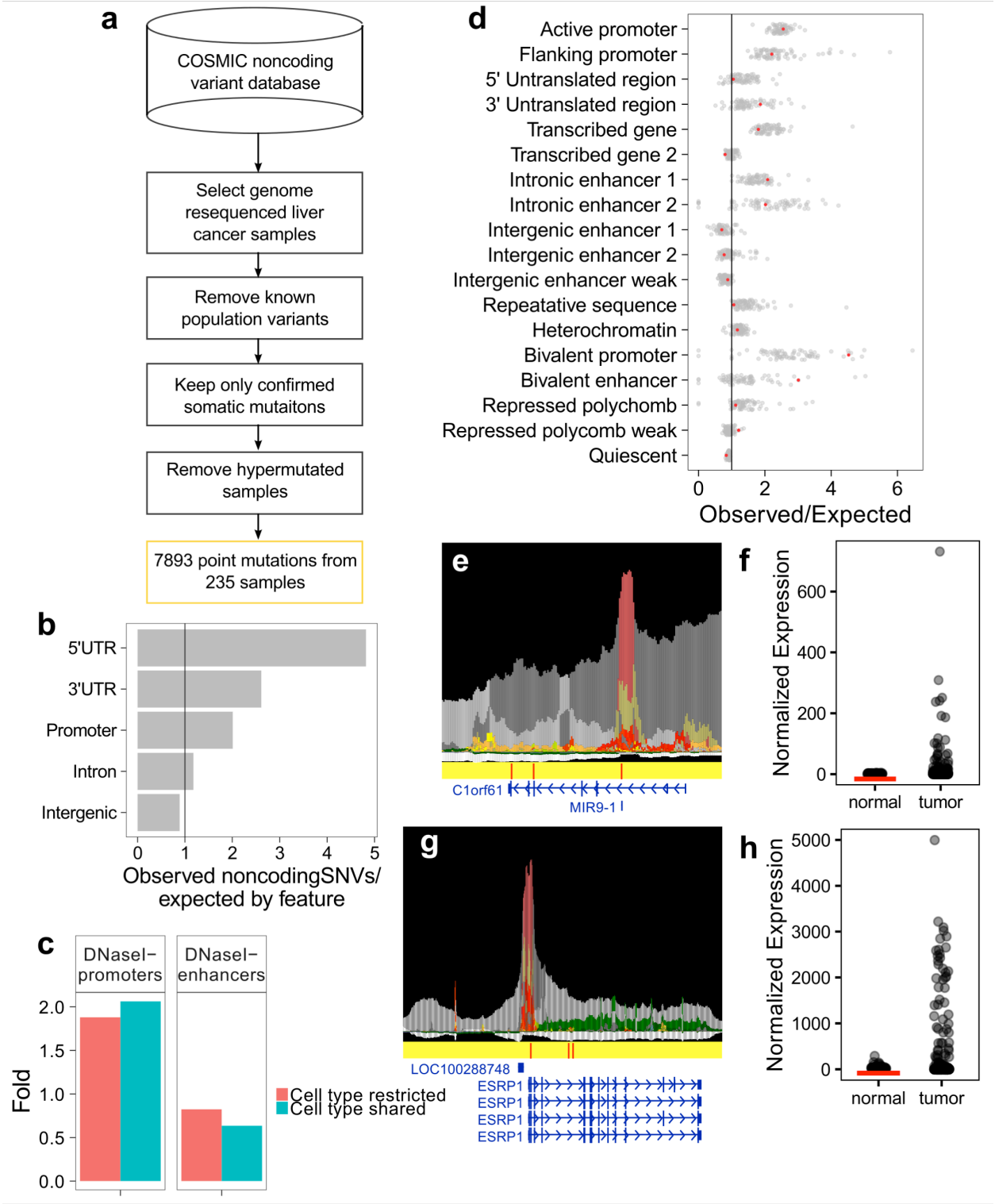


Figure 4.2. PLC SNVs occur more often than expected in heterologous cell type-specific regulatory elements. (a) Filtering strategy for SNVs from whole genome resequenced samples

in COSMIC. (b) Annotation of filtered SNVs by UCSC known genes. (c) SNVs in cell type restricted or shared DNaseI promoters or enhancers. Y-axis is fold observed over expected, based on background distribution of cell type restricted or shared DHSs. (d) Observed versus expected SNVs in each ChromHMM-18 state in each of the 78 Roadmap cells and tissues with available data. Orange dot is primary liver sample (Roadmap E066); gray dots are the other 77 Roadmap samples; black line is 1. (e) Browser view of *C1orf61* locus and three regulatory elements mutated in three unique samples. The top track is the Epilogos track (http://compbio.mit.edu/epilogos/), which provides a visualization of the chromatin state models for several cell types at once. The presented track depicts the ChromHMM-18 state model 127 Roadmap cell types (primary and cell lines) at a 200bp resolution. Red and orange colors represent active promoter annotations; light green and yellow colors represent genic enhancers and enhancers, respectively; pink and beige are bivalent states; grays are repressed Polycomb states. Middle track: Positions of PLC WGS SNVs (red lines) on a yellow background. Bottom track: RefSeq genes track. (f) Expression from TCGA PLC tumor and matched normal samples for *C1orf61*. Red line = median expression for normal samples. (g) Browser view for *ESRP1* and three regulatory elements mutated in three unique samples. Tracks are as in (e). (h) Expression as in (f) for *ESRP1*.

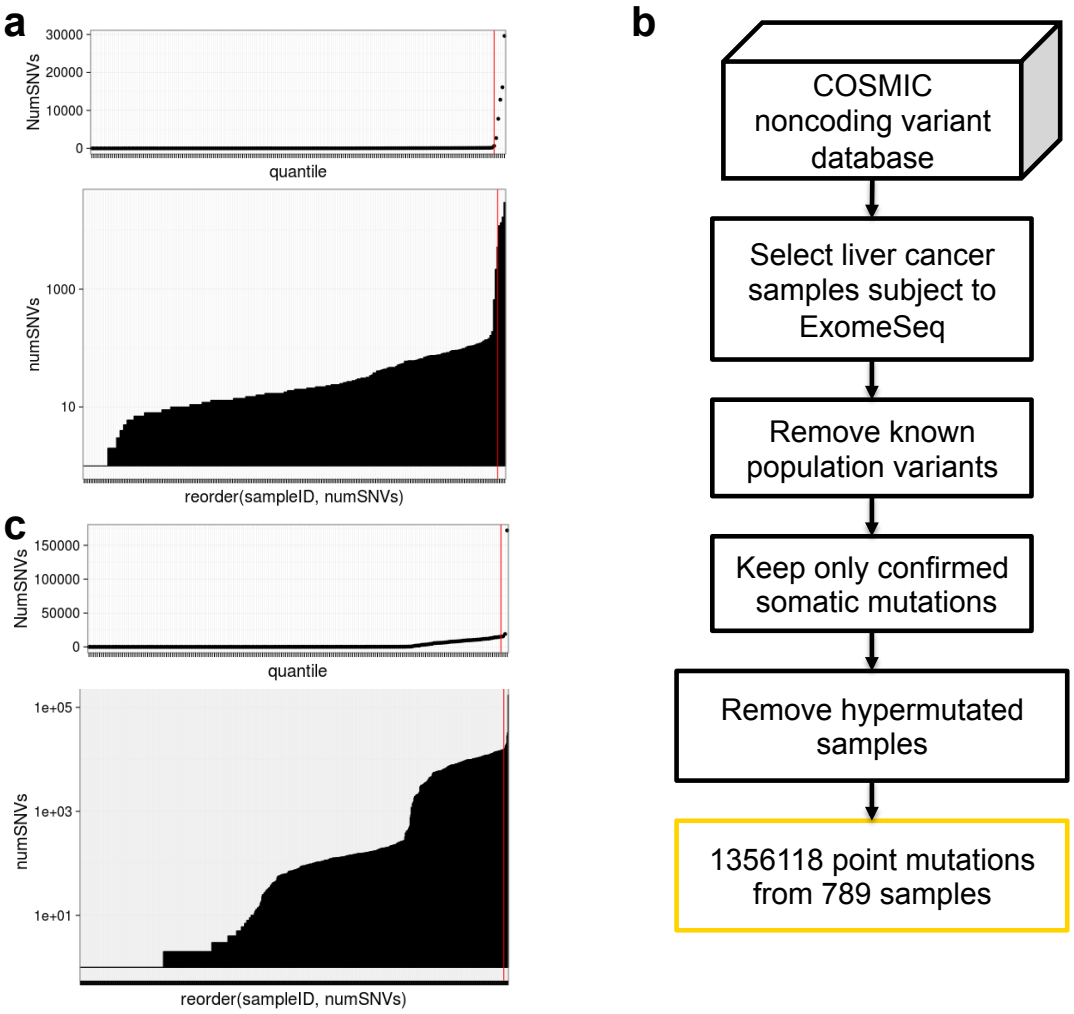


Figure 4.3. Data filtering strategy. (a) Top: For COSMIC PLC samples with whole genome resequenced data, each percentile (x-axis) was plotted against the number of SNVs (y-axis). Bottom: Samples ordered from fewest to largest number of SNVs. Red line = cutoff at the greatest rate of change between percentiles. (b) Filtering strategy for COSMIC PLC samples with ExomeSeq data. (c) Same as (a) but for SNVs from PLC samples with ExomeSeq-derived SNVs.

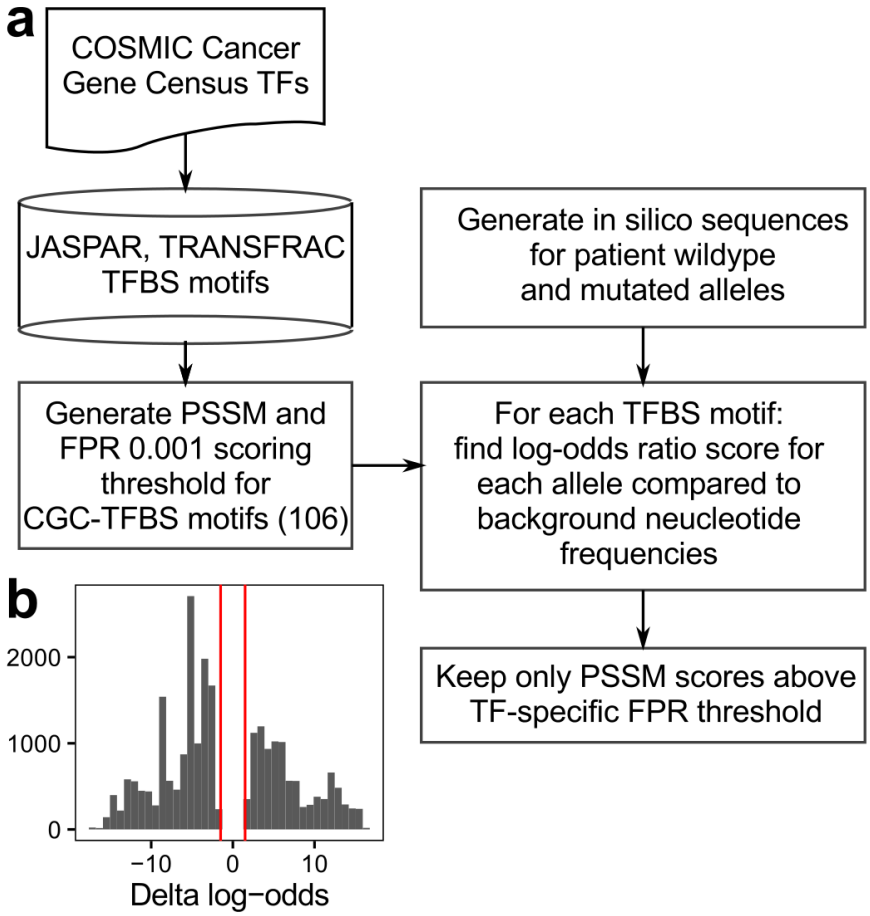


Figure 4.4. Systematic motif detection identifies oncogenic TFBS gain-of-binding events. (a) Analysis pipeline for detecting motifs from wildtype and mutant allele sequences. (b) Histogram of delta values for WGS SNV allele pairs after filtering to keep only allele pairs with at least one motif score of absolute value ≥ 2.

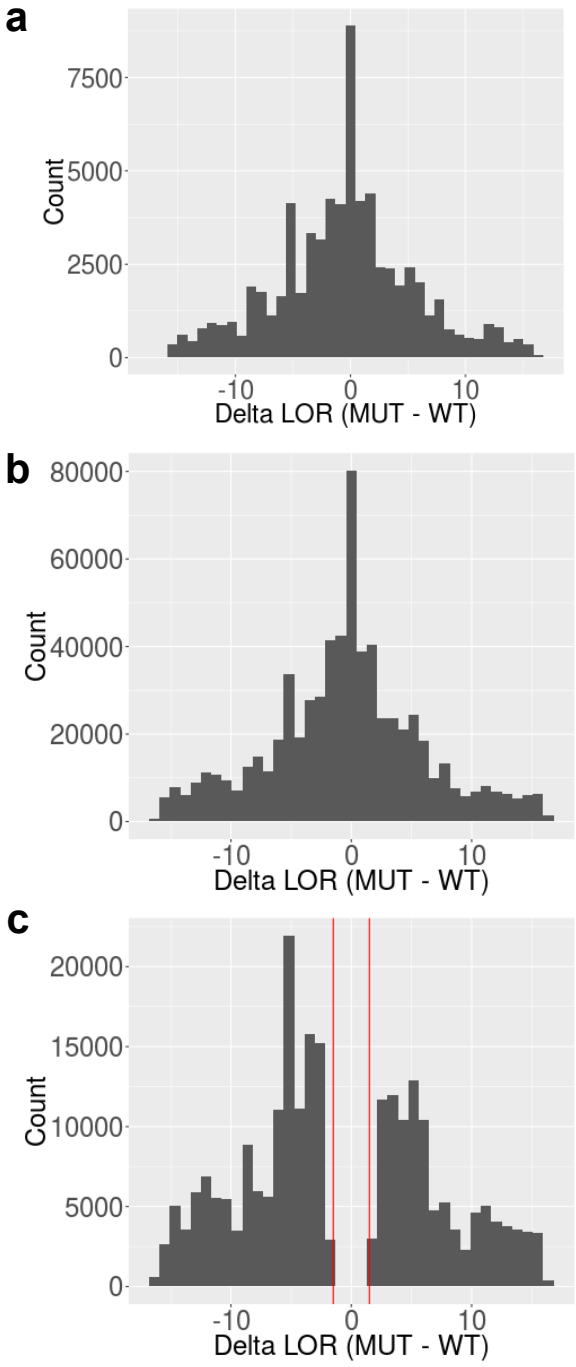


Figure 4.5. Delta values from systematic motif detection. (a) Delta values (mutant allele log-odds score – wildtype allele log-odds score) for WGS SNVs before applying threshold criteria.

(b) Same as (a) but for ExomeSeq SNVs. (c) ExomeSeq SNVs after applying threshold criteria (at least one score ≥ 2 log-odds over background).

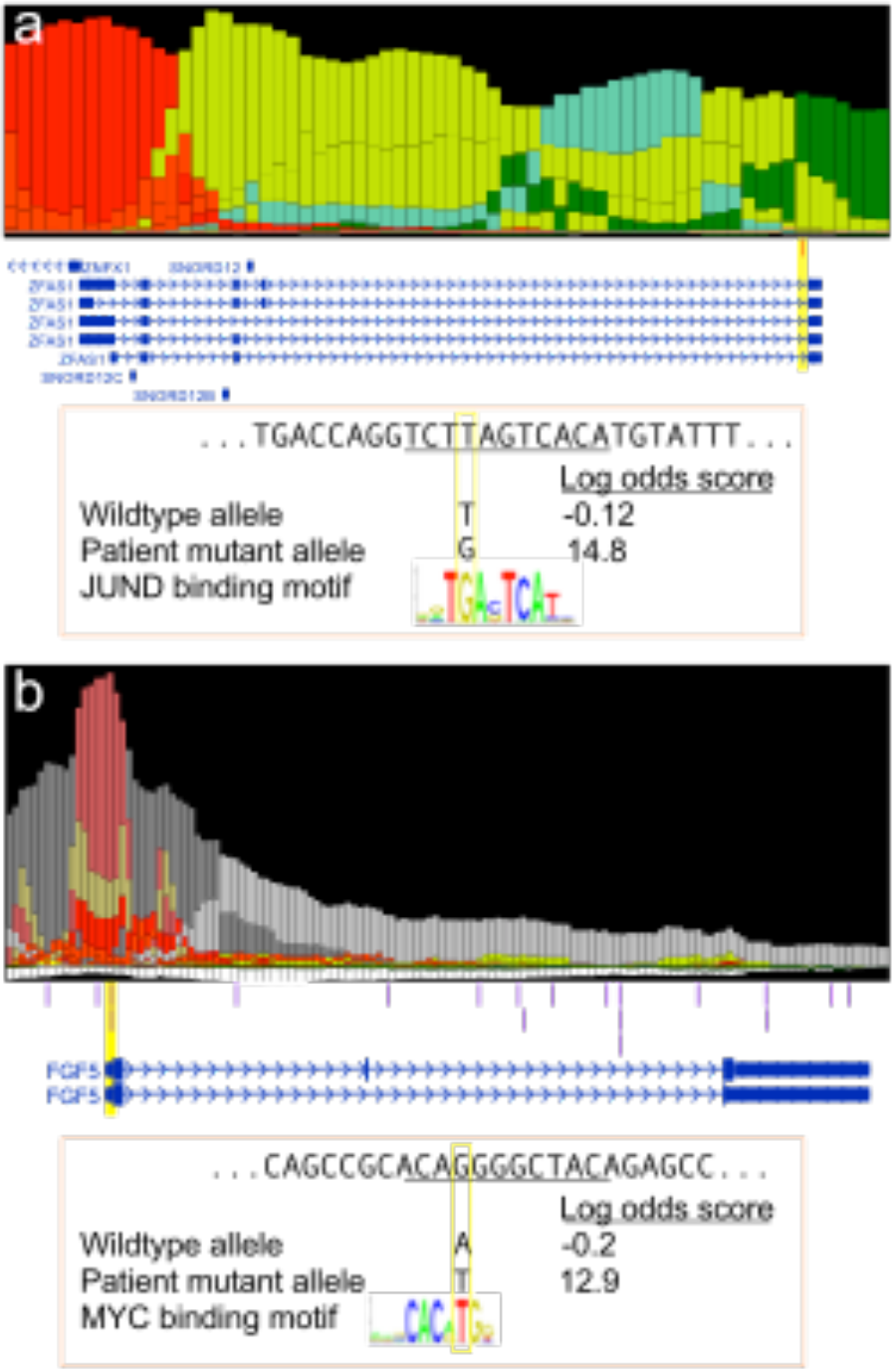


Figure 4.6. Gain-of-binding site events at known oncogenes. (a) *ZFAS1* locus. SNV occurs in the last intron creating a JUND binding site. (b) *FGF5* locus. SNV in the promoter creates a MYC binding site.

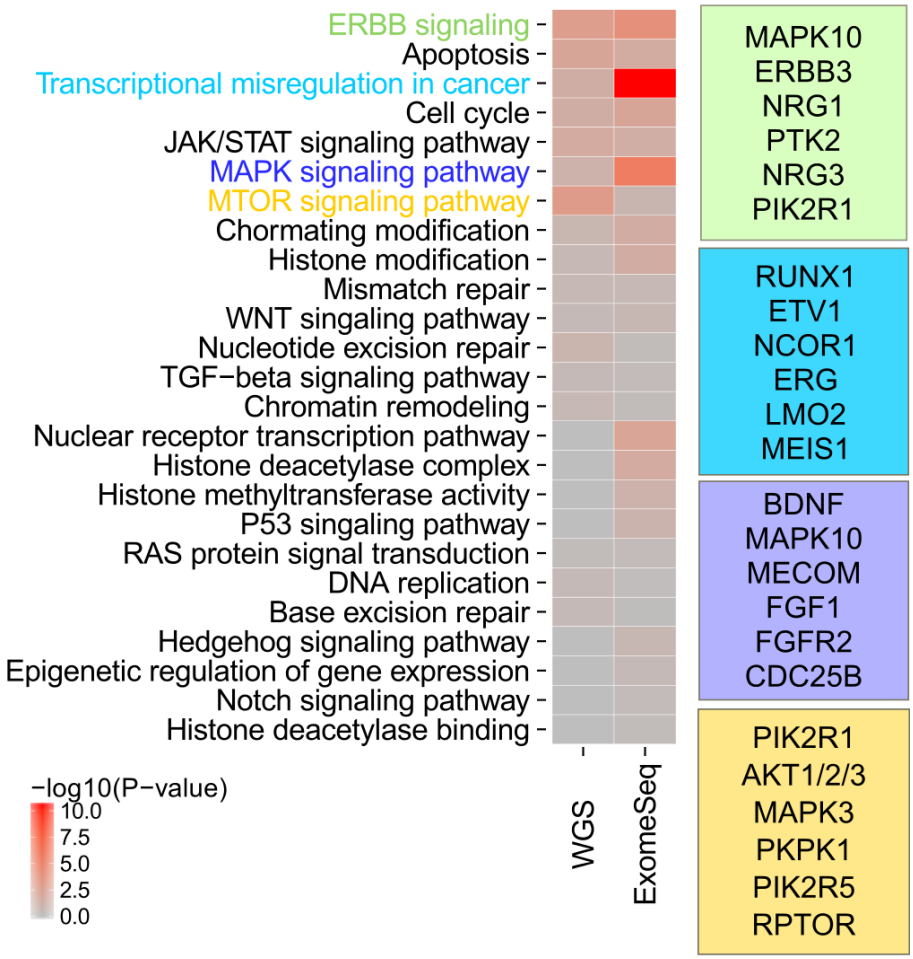


Figure 4.7. Liver cancer SNV pathway enrichment. Right: Heat map of 25 pathways tested. Color intensity represents the significance of enrichment (–log10(P-value)) for PLC SNVs in promoters that are found in genes for each pathway. WGS = whole genome resequencing-derived PLC SNVs; ExomeSeq = ExomeSeq-derived PLC SNVs. Left: Colored boxes depict a sample of top hits from significantly enriched pathways. Genes listed have the most recurrently hit promoters for the given pathway. Green box = ERBB signaling pathway; blue box = transcriptional misregulation in cancer; purple box = MAPK signaling pathway; gold box = MTOR signaling pathway.

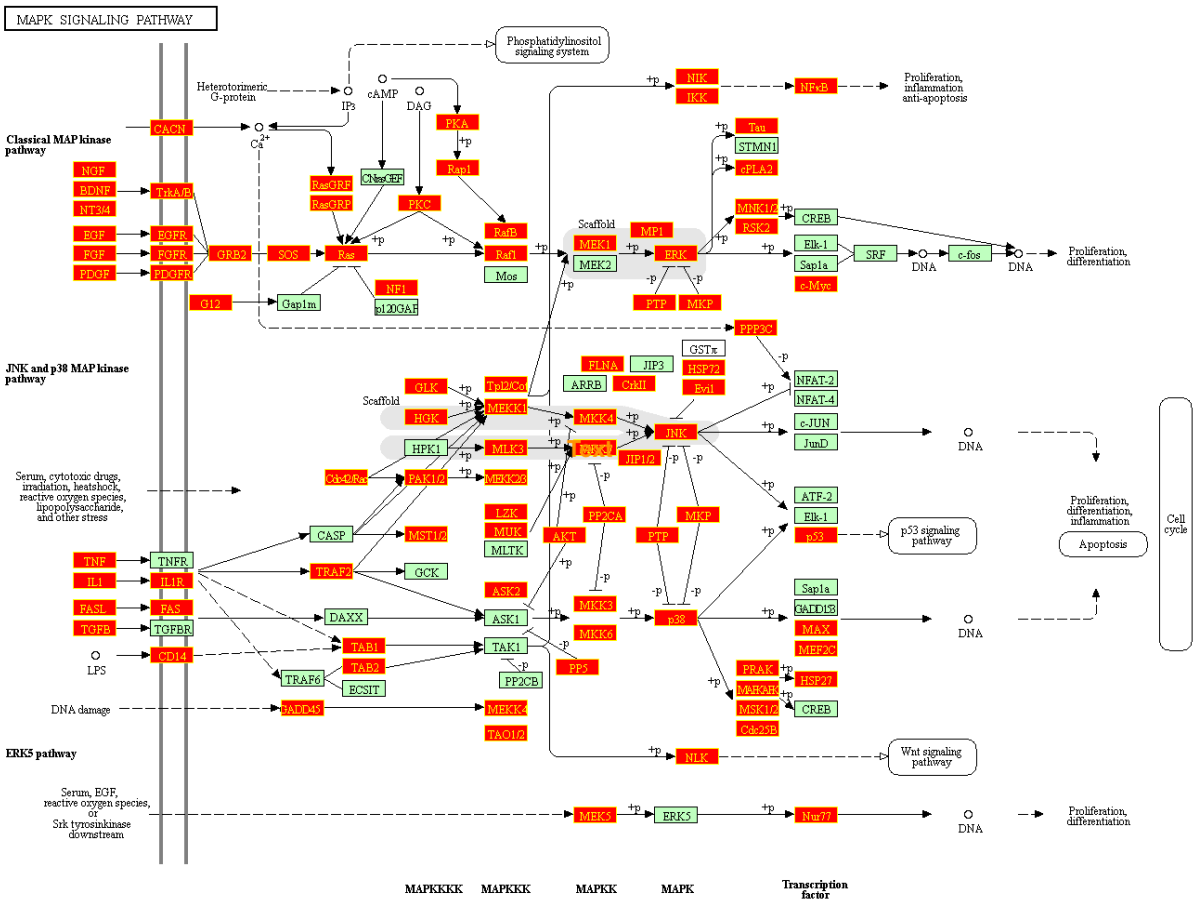


Figure 4.8. KEGG pathway map for MAPK signaling pathway (hsa04010). Red boxes are genes that have SNV promoter mutations in PLC data. Constructed using Pathway Painter [244].

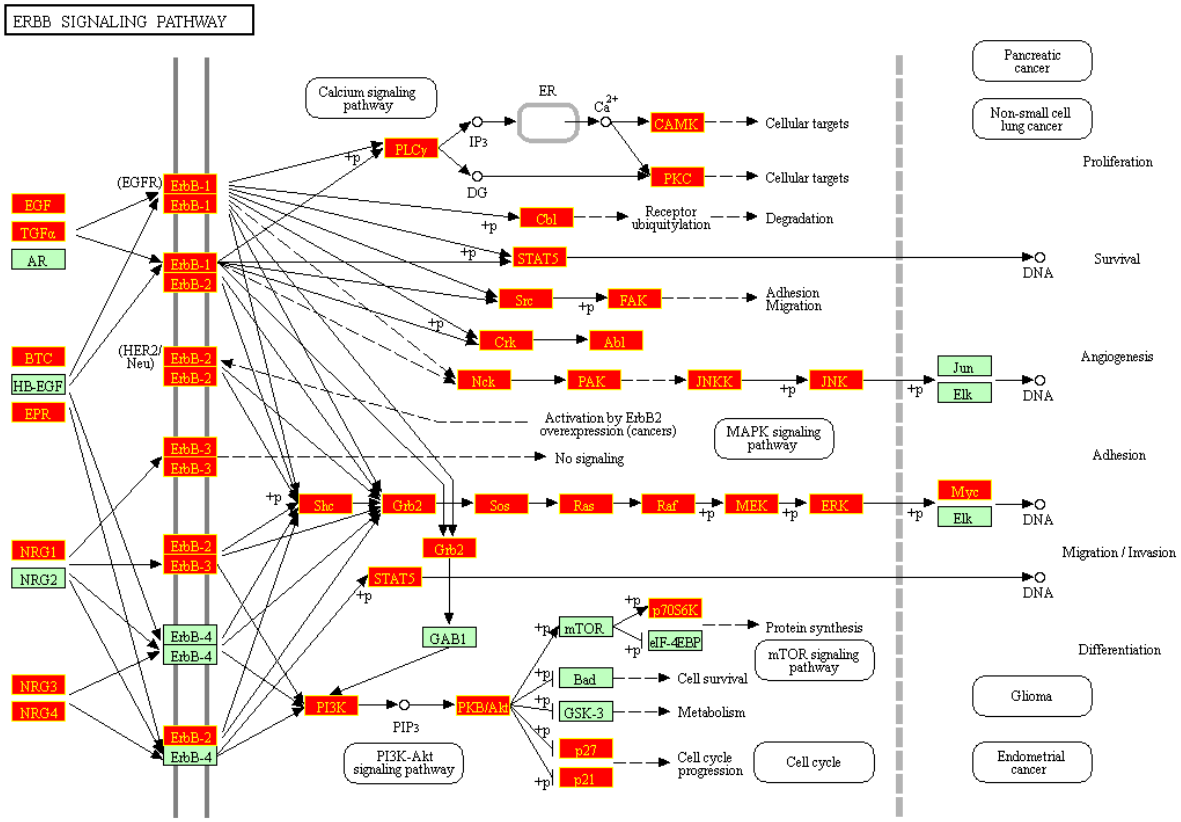


Figure 4.9. KEGG pathway map for ERBB signaling pathway (hsa04012). Red boxes are genes that have SNV promoter mutations in PLC data. Constructed using Pathway Painter [244].

Table 4.1. Number of SNVs per regulatory element.

Number SNVs per element	Number of regulatory elements
1	3035
2	43
3	6
4	2
5	1
7	2
16	1

Table 4.2. Number of genes with SNV-containing putative regulatory elements.

Number SNV-containing regulatory element per gene	Number of genes
1	1031
2	52
3	3
5	1

4.8 Datasets and URLs

Dataset	Description	Filename	URL	Version	Download Date	Ref.
COSMIC Whole Genomes	Noncoding variants	CosmicWGS_NCV.tsv.gz	http://cancer.sanger.ac.uk/cosmic	v77	13-Jun-16	[194]
COSMIC Whole Genomes	Sample metadata	CosmicWGS_SamplesExport.tsv.gz	http://cancer.sanger.ac.uk/cosmic	v77	14-Jun-16	[194]
ChromHMM-18 segmentation	Mnemonics bed files	mnemonics.bed	http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state	ChromHMM-18	10-Mar-16	[177]
ChromHMM-18 segmentation	State by line files	all.statesByLine.tgz	http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state	ChromHMM-18	10-Mar-16	[177]
DNaseI delineation	Promoter state calls	state_calls_prom.RData	http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation		4-Apr-16	[177]
DNaseI delineation	Enhancer state calls	state_calls_enh.RData	http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation		4-Apr-16	[177]
Cancer Gene Census	Cancer-related genes	CosmicCGC.csv	http://cancer.sanger.ac.uk/census	v77	13-Apr-16	[194]
JASPAR	Nonredundant vertebate JASPAR CORE motifs	jaspar_vertebrates_nonredudntant.pfm	http://jaspar.genereg.net	JASAPR 2016	21-Mar-16	[241]
TRANSFAC	Transcription factor motifs	TRANSFAC.tar	http://www.gene-regulation.com/pub/databases.html		11-Oct-06	[242]
MSigDB Collections	Selected KEGG and Amigo pathways	*various*	http://software.broadinstitute.org/gsea/msigdb/collections.jsp	v5.1	8-Jul-16	[222-224]
TCGA Gene expression	Gene expression quantification for Liver hepatocellular carcinoma samples and matched normal	c035a280-e2be-4844-8ef8-6340746a8c91.tar	https://tcga-data.nci.nih.gov/docs/publications/tcga/		28-Apr-16	[188]
Human Genome Reference Consoritum	Assembly statistics for hGRC37	n/a	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/	hGRC37.p13	n/a	

Chapter 5

Evolution of Epigenetic Regulation in Vertebrate Genomes

5.1 Author Contributions

This chapter is adapted from the published manuscript Rebecca F. Lowdon[1], Hyo Sik Jang[1], Ting Wang[1*]. "Evolution of epigenetic regulation in vertebrates." *Trends in Genetics*. 2016;32(5):269-283. [245]

R.F.L. and H.S.J. conducted all background research. R.F.L., H.S.J., and T.W. wrote and edited the manuscript.

[1] Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, MO 63108
* Corresponding author

5.2 Abstract

Empirical models of sequence evolution have spurred progress in the field of evolutionary genetics for decades. We are now realizing the importance and complexity of the eukaryotic epigenome. While epigenome analysis has been applied to genomes from single cell eukaryotes to human, comparative analyses are still relatively few, and computational algorithms to quantify epigenome evolution remain scarce. Accordingly, a quantitative model of epigenome evolution remains to be established. Here we review the comparative epigenomics literature and synthesize its overarching themes. We also suggest one mechanism, transcription factor binding site turnover, which relates sequence evolution to epigenetic conservation or divergence. Lastly, we propose a framework for how the field can move forward to build a coherent quantitative model of epigenome evolution.

5.3 Comparative Epigenomics as a Tool to Explore Epigenome Evolution

The epigenome is an integral part of genome biology, comprised of DNA modifications, most notably 5-methylcytosine (DNA methylation), histone post-translational modifications, and nucleosome positioning (Figure 1). The epigenome is crucial for proper gene regulation [246], genome integrity [247], dosage compensation [248,249], and proper development [4] across eukaryotic phyla. Nevertheless, an empirical model of epigenome evolution has yet to be established. Decades of interrogating the chromatin remodeling of specific loci over development and across species provide early examples of comparative epigenomics, defined here as the comparison of epigenetic status between syntenic regions.

Comparative epigenomics is based on determining epigenetic conservation: two homologous sequences that host similar epigenetic modifications in homologous cell types (**Figure 5.2**). The homologous loci may be orthologous in distantly related species or paralogs in the same genome. It follows that epigenome comparison requires determination of sequence homology, epigenetic status, and biological homology between two species [250].

This review focuses on what comparative epigenomics has taught us about vertebrate epigenome evolution, although comparisons with invertebrate and plant epigenomes have been invaluable to build a full picture of epigenetic regulation [251-253]. Additionally, the focus is confined to the use of comparative epigenetics, which can reveal epigenetic regulatory features by identifying regions of conserved and divergent epigenetic status across phyla, to understand gene regulation.

Lastly, the scope of this review is constrained by the scope of comparative epigenomics studies in existing literature. **Figure 5.1** outlines common epigenetic marks and related assays that are

covered in this review, along with a representative example of the data and the interpretation of the epigenetic situation in a cartoon.

The arrival of high-throughput sequencing (HTS) technologies and genome-wide biochemistry experiments has moved the study of the epigenome into the 'omics' era. With HTS tools and databases of thousands of epigenome mapping experiments across thousands of eukaryotes [254], the field can begin to create models of epigenome conservation and divergence and interpret the biological meaning behind these signals.

5.4 Epigenome Evolution at Orthologs

Rooted in a strong theoretical foundation [255], comparative genomics enables the identification of conserved sequences, elucidating functional genomic elements [35,256]. However, not all functional genome regions are conserved [36,257] suggesting other genomic features are responsible for adaptive gene regulation [258,259]. Two possible explanations for non-conserved functional elements are the limitation of sequence alignment algorithms [260] or that these non-conserved regions can serve as genuine species- or lineage-specific (a genetic or epigenetic feature specific to an evolutionary lineage) regulatory elements [1,258,261]. Accordingly, experimental approaches have shown many non-conserved sequence elements are gene regulatory [2,39,262].

Pioneering work in comparative epigenetics detail the structure and function of chromatin and epigenetic modifications at orthologous loci across model organisms. Well-studied developmental loci including the insulin-like growth factor 2 receptor locus, macrophage colony-stimulating factor, and the beta-globin locus, exhibit conserved epigenetic status [42,263,264], transcription factor regulation [42,264-266], and function [42,266,267] across species. Taken together, analysis of the sequence and epigenetic conservation at these loci suggests that epigenome comparison is a viable method for identifying elements modulating gene regulation.

From the above observations, it can be postulated that epigenetic features are correlated with underlying sequence features (**Figure 5.2**). This review presents evidence both for and against this hypothesis in an effort to establish a framework for epigenome evolutionary studies.

5.4.1 Vignette: Locus-Specific Example of Epigenome Evolution: the c-FMS Locus

Macrophage colony stimulating factor receptor (c-FMS) expression marks hematopoietic commitment to the myeloid fate. Alternate first exons accompany transcripts in placental trophoblasts, and c-FMS is also an oncogene. Accordingly, c-FMS is subject to specific transcriptional regulation. The c-FMS locus in human and mouse have high sequence identity, especially at each alternative promoter and two intronic enhancers [264,268]. Despite high sequence conservation, c-FMS regulatory regions are bound by the same ensemble of TFs, but in different arrangements, along with some species-specific TFs [264]. However, the TF ensembles recruit the same chromatin remodeling factors in each species (Brg-1, HDAC) and drive the same transcriptional output in a cell-type and developmental- specific manner [264]. This example supports the hypothesis that evolutionarily conserved regulation may be driven by evolutionarily conserved regulatory element sequence and transcriptional programs, although there exists inter-species variability in the execution of such a program.

5.4.2 Relative DNA Methylation Conservation Across Sequence Contexts

Analysis of epigenetic marks at paralogs allows for studying epigenetic evolution without the confounding environmental variability that exists in inter-species comparisons [269]. In the human genome, 78% of paralogous CpGs had an absolute DNA methylation difference of 20% or less [269]. Thus duplicons tend to retain their DNA methylation signature, supporting the hypothesis that epigenetic features are correlated with underlying sequence (Figure 2).

When comparing genome wide DNA methylation levels between species, 70-74% and 80-82% similarity was found in peripheral blood and prefrontal cortex, respectively in the great ape somatic tissues [270,271]. Correlation coefficients from inter-species pairwise comparisons of whole genome bisulfite sequencing (WGBS) data from primate blood samples show agreement

with species phylogeny [271], suggesting DNA methylation variation is related to sequence variation.

However, pairwise correlations of DNA methylation levels between species showed only moderate concordance at individual CpGs. For example, examination of primate peripheral blood samples using the Illumina Methylation450 array showed 22% of probes covering orthologous CpGs were not significantly different among human, chimp, bonobo, gorilla, and orangutan (mean beta-value difference of < 0.1) [270]. What accounts for individual CpG methylation level variance between species?

To test how DNA methylation status varies with sequence, regions of incomplete lineage sorting (ILS), where the sequence genealogy is different from the known species phylogeny were used [271]. An example of an ILS would be an orthologous region in humans that is more similar in sequence to gorilla than chimp. The authors isolated 360,000 CpGs in ILS regions from human, chimp, gorilla, and orangutan. Strikingly, DNA methylation patterns over ILS regions followed the sequence relationships, suggesting a physical dependence of DNA methylation status at ILS regions on sequence variation [271]. Additionally, most of the 570 human-specifically methylated regions were distal to transcription start sites (TSS) and showed accumulation of nucleotide substitutions [271], suggesting that methylome evolution may be coupled to sequence evolution at regulatory elements.

5.4.3 Relationships between Histone Post-Translational Modification Conservation and Sequence Conservation

Comparative studies of histone posttranslational modifications (PTM) show that not all orthologs have conserved epi-mark status. Comparative analysis of H3K4me2 chromatin immunopreciptation (ChIP) followed by microarray (ChIP-chip) over two orthologous loci in

mouse and human lung fibroblasts showed that functional conservation of histone methylation did not correlate with elevated sequence conservation [31]. Unlike in species-specific DNA methylated regions, human cortex-specific gains of H3K27ac or H3K4me2 did not have concomitant accelerated sequence evolution compared to rhesus macaque and mouse brain cortex samples [272].

The first study to calculate epigenetic conservation explicitly in the context of sequence evolution analyzed a panel of epigenomic features in human, mouse, and pig embryonic stem cells (ESCs) [273]. The authors quantified epigenetic conservation between two species for a given epigenetic mark as the ratio of observed orthologous nucleotides conserved for that epi-mark over expectation. They then calculated the epigenetic conservation score for each modification over a range of binned PhyloP scores, a proxy for nucleotide substitution rate [274]. It was found that regardless of species being compared, epigenetic score profiles fell into three distinct patterns. First, subsets of epi-marks (Polycomb deposited H3K27me3 and promoter-associated H3K4me3) were conserved more often than expected at orthologous regions with low substitution rates, agreeing with the hypothesis that epigenetic conservation is correlated with genetic conservation. Second, three marks (DNA methylation, gene body-associated H3K36me3, and enhancer-associated H3K27ac) were enriched for conserved nucleotides over orthologous sequences with reduced substitution rate *and* sequences with an accelerated substitution rate in the human genome. The remainder of interrogated epi-marks (H3K9me3, H3K4me1, H3K4me2, and H2A.Z) had a uniform level of epigenetic conservation regardless of sequence evolution. Accordingly, fast-diverging orthologous sequences are more conserved than expected for DNA methylation, H3K36me3, and H3K27ac [273] (**Figure 5.2**), another example that sequence conservation is not required for conserved epi-mark status. The authors suggest that epi-

conserved but genetically non-conserved regions may buffer against genetic mutations and provide functional stability to fast-evolving genome regions [273].

Sequence conservation is not always required for epi-conservation [275]. Instead, some epi-marks (DNA methylation, H3K36me3, and H3K27ac) may be conserved over orthologs whose sequence is under purifying selection. Understanding what features cause fast-evolving DNA fragments to undergo divergent epigenetic evolution is an important area of future study (**Figure 5.4**).

5.5 Epi-mark Influence on Conserved or Divergent Gene Regulation

The promoter is the regulatory DNA sequences surrounding a gene TSS and is responsible for transcription initiation. Epigenetic modifications at vertebrate promoters are well-studied, and epi-marks common to active promoters are depicted in **Figure 5.1**.

5.5.1 Epigenetic Conservation at Promoters

DNA Methylation Conservation Status at Promoters

CpG islands (CGI) have long been recognized as non-methylated regions associated with protein-coding gene promoters [276]. Isolation of non-methylated DNA fragments is achieved via affinity purification with biotinylated CxxC (Bio-CAP), which preferentially binds non-methylated DNA (**Figure 5.1**). Genome-wide Bio-CAP experiments followed by massively parallel sequencing in seven vertebrate genomes revealed that non-methylated islands (NMI) are a conserved feature of orthologous promoters, as well as distal regulatory elements [277]**.**

Methylation of CpGs in promoters is negatively correlated with gene expression [278,279], and some vertebrate CGIs are tissue-specifically methylated [19,277,280]. Array-based analysis of DNA methylation at ~27k CpG loci at proximal promoters in heart, liver, and kidney samples from human and chimp found that 18-26% of tissue-specific differentially methylated regions (tsDMRs) were conserved between human and chimp (varying by tissue). Conserved tsDMRs were enriched for negative correlations between methylation level and the associated gene's expression level (72% negative correlation values, regardless of species) [280]. Additionally, promoters with conserved tsDMRs were enriched for genes annotated as developmental process genes. Thus epigenetically-mediated tissue-specific regulation over core developmental genes tends to be conserved between human and chimp [280].

A similar study queried the DNA methylation status of ~326k probes shared between human, chimpanzees, bonobos, gorillas, and orangutan peripheral blood. Inter-species differentially methylated CpGs were depleted from proximal promoters and CpG islands (CGIs), suggesting broad conservation of promoter methylation status between primates [270]. However, there are promoters with inter-species differential DNA methylation, which explained 12-18% of gene expression level differences between primates [280]. Thus species-differential DNA methylation at promoters can mediate species-differential gene expression. What sequence features underlie conserved or divergently methylated promoters?

CpG methylation status varies by CpG density, where denser CpG regions, such as CGIs, tend to be lowly-methylated, while regions of sparse or intermediate CpG density are variably methylated [10,18,281]. However, comparative analysis suggests that CpG density does not fully explain DNA methylation status at promoters [282] or predict DNA methylation divergence between paralogs [269,283]. Indeed, sequence analysis of experimentally-determined NMIs revealed that the ratio of observed over expected CpGs and GC content of NMIs varies in a species-dependent manner [277]. Therefore, while NMIs are a common feature of vertebrate genomes and central to promoter regulation, the specific underlying sequence characteristics driving this conserved epigenetic feature may vary between species. Instead, transcription factor binding site (TFBS) motifs at methylation determining regions [282] were found to be sufficient for proper methylation status at promoters [282-286]. In summary, while sequence features cannot fully predict DNA methylation status, transcription factors can alter DNA methylation patterns at promoters and drive promoter function. Comparative epigenome analysis incorporating multiple species is needed to understand how sequence evolution and motif turnover drive DNA methylation status at promoters (**Figure 5.3**).

Histone Post-translational Modification Conservation

H3K4me3 defines active gene promoters [49]. Comparative analysis of ChIP for H3K4me3 and H3K27ac followed by sequencing (ChIP-seq) in liver samples from 20 mammals demonstrated that the basic regulatory landscape was very similar between species, including an average of ~12500 active promoter elements [287]. Similarly, H3K4me3 enrichment was conserved in a cell type-specific manner between mouse and human, where ~80% of queried promoters were conserved in four homologous cell types [51]. Genome wide, the magnitude of H3K4me3 promoter conservation increases in more closely related phylogenies: 16% of human liver promoters are conserved for H3K4me3 ChIP-seq signal across 20 mammals [287], while ~36% of orthologous promoters were conserved for H3K4me3 ChIP-seq between human, chimp, and rhesus macaque lymphoblastoid cell lines [52].

Species-differential promoter histone modification can also indicate species-specific gene expression [275]: up to 7% of differentially expressed genes between human and chimpanzee were explained by H3K4me3 distinctions [52]. These studies reveal that H3K4me3 and H3K27ac are features of highly expressed genes across mammals, and quantify how homologous cell types utilize orthologous genes for shared or species-specific functions. What sequence features drive conserved histone promoter marks, or mediate their turnover, remain to be investigated.

5.5.2 Gene Body Epi-mark Conservation

The “gene body” is the collection of introns and exons in the open reading frame of a gene. An archetypical vertebrate gene body epi-modifications is depicted in **Figure 5.1**.

Differential DNA Methylation over the Intron-Exon Junction

Seminal surveys of eukaryotic methylomes determined that CpG methylation over gene bodies is a conserved feature of eukaryotic genomes [13,288]. Internal exons typically display 6-20%

elevated CG methylation compared to flanking introns [13], and gene body CpG methylation is conserved at 70-76% in human and chimp prefrontal cortex samples [58]. In both human and mouse genomes, recently duplicated genes retain conserved methylation patterns at gene body regions [283,289], suggesting that there is an overarching epigenetic mechanism that can identify duplicated fragments and properly methylate them. In most somatic tissues, high gene body methylation correlates with intermediate expression levels [288], with the exception of primate brain samples where gene body CpG methylation decreased linearly with increasing levels of gene expression [290].

Mammalian placenta are remarkable for their conserved global hypomethylation compared to somatic tissues (less than 66% genome-wide methylation level by MethylC-seq in human, rhesus macaque, squirrel monkey, mouse, dog, horse, and cow placentas, as well as opossum extraembryonic membrane [291]). One exception to placental hypomethylation was that gene bodies displayed elevated methylation across all species. As in somatic tissues, high methylation over gene bodies in the placenta correlated with intermediate gene expression level, while genes with low gene body methylation were less likely to be expressed [291]. Additionally, high placental gene body methylation was conserved across species over genes with similar gene ontologies, including genes involved in cell cycle, protein localization, and chromatin modification [291]. Gene body methylation is a conserved feature of eukaryotic genomes, and methylation level has a parabolic relationship to gene expression level in most eukaryotic somatic tissues and placenta, although the mechanistic links between genic DNA methylation and expression level are still unclear.

Histone Post-translational Modifications over Exons

Exon-specific H3K36me3 modifications are conserved across eukaryotes [292,293] and are associated with exon inclusion [292]. Indeed, in exons, ChIP-seq signals of H3K36me3, and to a lesser extent, H3K79me1, H4K20me1, and H2BK5me1 were found to increase as gene expression level increased [293]. However, histone modification enrichments could be a downstream result of differential nucleosome occupancy over exons and introns, where introns tend to be nucleosome-depleted. Emerging evidence for the role of H3K36me3 in gene splicing [294] or mismatch repair [295] may help resolve the functional importance of H3K36me3 modification over exons.

Furthermore, co-localization of H3K27me3 and H3K36me3 over gene exons is associated with monoallelic gene expression, and this signal was conserved in human and mouse [296]. This signature is conserved over genes important for embryonic development and cell surface protein genes. Notably, monoallelic expression and the genes' corresponding epigenetic signature were lineage-specific and maintained in differentiated tissues [296], suggesting that H3K36me3/H3K27me3 modifications may play a role in maintaining monoallelic expression and expression regulation in general [296].

5.5.3 Evolution of Epigenetic Regulation at Vertebrate Enhancers

Epigenetic modifications are often have complimentary functions and are studied together to explore specific genetic elements. Enhancers display a characteristic histone modification profile of H3K4me1 and H3K27ac, usually in association with p300 [52,297], are usually hypomethylated [25], and are responsible for regulating cell type-appropriate gene expression [298]. While characterization of novel enhancers are improving through advances in profiling techniques and computational models [299], our understanding of enhancer evolution is still unfolding.

Comparative studies analyzing histone PTM ChIP-seq signals between human and mouse developing heart [300], limb [275], and adipogenesis [301], and human and chimp cranial neural crest [302], and between developmental stages in distantly related zebrafish and medaka [303] reveal both shared and lineage-specific epi-marks. One study found that human neural crest cell (NCC) enhancers, defined by co-localization of H3K27ac and H3K4me3, showed strong enrichment of H3K27ac in the chicken orthologs and conserved TFAP2A binding in both species [156], suggesting these are conserved NCC enhancers. This study validated several of these predicted enhancers to have gene regulatory capabilities by reporter assays in zebrafish, and demonstrated that TFAP2A binding was necessary for specific enhancer activity [156]. Such multi-species analysis of cell type enhancer elements shows histone modification can be a strong indicator of conserved, functional enhancers.

Leveraging epigenomic data from multiple organisms can identify species-specific enhancer elements [275,300,304,305]. DNaseI footprints are identified from DNaseI-seq datasets where the cleavage pattern of DNaseI digested fragments is abrogated by a DNA binding protein, such as a transcription factor, occupying the DNA [306]. DNaseI footprinting analyses in human and mouse revealed that while 65% of DNaseI footprints in mouse have an orthologous sequence in human, only 22% of those orthologs also showed DNaseI footprinting signal, suggesting that there has been a large scale turnover of TF binding since the human-rodent split [304,307-310]. An assessment of DNaseI hypersensitive sites (DHSs) in human, chimp, and macaque skin fibroblasts and lymphoblastoid cell lines showed that most DHSs are conserved across species, as pairwise comparison of genome-wide DHS signals were highly correlated [311]. However, several hundred DHSs were gained or lost in each of the human and chimp lineages, particularly at distal enhancers and introns [311].

Conversely, regulatory element conservation decreases when increasing the number of epigenomes compared. ChIP-seq in liver samples from 20 mammals showed that while enhancers were more common than promoters, only 1% of human liver enhancers had conserved H3K27ac signal at the orthologous sequence in at least 10 other mammalian genomes [287].

Species-specific enhancers are clearly common, but are they functional? Human DHS gains in skin and lymphoblastoid cells significantly overlapped with ChIP-seq signals for enhancer-associated chromatin marks H3K4me1 (~80% overlap), H3K4me2 (~80%), and H3K27ac (~70%) [311]. Lineage-specific epi-marks were enriched near tissue-relevant genes and genes associated with lineage-specific marks were discordantly expressed between species [275,287,301,302]. While more conserved than genomic background, human limb-specific enhancers were found to have less sequence conservation than limb enhancers shared with rhesus macaque and mouse [275]. In addition, human liver DHS gains/losses had stronger signal of positive selection on the human lineage, suggesting these regions are likely functional and may contribute to specific-specific gene regulation [311]. The high rate of lineage-specific enhancer turnover may be driven by transcription factor binding site (TFBS) turnover (**Figure 5.3**; See TFBS turnover as a mechanism for epigenome evolution section) [300,301,311].

5.6 Transcription Factor Occupancy at Orthologs

With the publication of the Mouse ENCODE Project, genome wide large-scale comparative analysis between mouse and human transcription factor (TF) ChIP-seq data is now available [76]. Since a comprehensive review of the genetics of TF occupancy across species has been published elsewhere [77], in this review, we highlight what has been learned about the epigenetic context of TF binding to human-mouse orthologous regulatory regions from the Mouse ENCODE Project.

Cis-regulatory sequences in mouse are enriched for conserved sequences: ~67% of both DNaseI hypersensitivity sites and TF ChIP-seq peaks had homologs in human, while ~79% of both chromatin-based promoter and enhancer predictions had homologs in human [312]. However, a smaller fraction of the human orthologs of predicted regulatory sequences in mouse were also predicted to be promoters (44%) or enhancers (40%) in human [312], suggesting that regulatory element conservation does not always track with sequence similarity [305,313]. How did species-specific regulatory elements evolve? The authors found that 89% of histone-defined mouse-specific promoters and 85% of mouse-specific enhancers overlap transposable elements (TEs) or mobile elements and were enriched for specific TE classes [312], suggesting that DNA derived from transposable elements may be responsible for a large fraction of species-specific gene regulation [308].

In order to examine TF ChIP-seq binding peaks in a cell type specific manner, the binding profiles of 32 TFs in two human and mouse homologous cell types: erythroid progenitors (mouse MEL; human K562) and lymphoblastoid cells (mouse CH12; human GM12878) were examined [307]. The TFs queried included Pol2, CTCF, and other general and cell type-specific transcription activators. Conservation of TF occupancy for an orthologous site varied in a TF-

specific manner, and conservation was highest at proximal promoter regions (even after controlling for elevated sequence conservation at promoters), with the exception of CTCF [307]. Epigenetic modification mimicked TF binding occupancy across species: DNA methylation levels were low in both species over orthologs with occupancy-conserved binding, but DNA methylation increased over unbound orthologs. In aggregate, bound fragments show elevated evolutionary constraint, but ~50% of bound regions in one species were not alignable in the other, representing species-specific binding events that may be mediated in some instances by transposable elements [307].

In the vertebrate genome, repetitive sequences and transposable elements (TEs) contribute ~6-60% of total sequence content [314]. However, TE sequences are thought to be silenced through epigenetic defense mechanisms since transposition events can be deleterious [315]. Interestingly, the dynamic epigenetic silencing of transposon elements during development is conserved among vertebrates although numerous TEs are known to be species-specific. Furthermore, TEs are hypothesized to shape gene regulatory networks through exaptation [316]. Exaptation describes the process where the TEs evolved to acquire new function in the genome, such as novel TFBS, that provided some fitness benefits in the host [316]. There are two models of TE exaptation: 1) surplus of TE insertions in the genome provided raw sequence material that can be mutated into novel TFBS and 2) TEs with functional TFBS transposed throughout the genome until a functional gain that led to a fitness benefit and fixation [317].

Recent work revealed that transposable elements contributed 2-40% of TF binding events in human or mouse, depending on the TF and cell type [308]. Yet only 2% of human TE-derived TF binding sites and 1% of mouse TE-derived sites were occupied by the *same* TF at a syntenic site in the opposite genome. Furthermore, 99% of human and 98% of mouse TE-derived binding

sites were species-specific, suggesting either the host TE amplified after the primate-rodent split, or that the TE ancestor accumulated too many mutations to be recognized as a TE sequence in the other genome [308] (**Figure 5.2**). Regardless, this comparative analysis supports the hypothesis that TEs may rewire gene regulatory networks in a species-specific manner [1,261].

5.7 TFBS Turnover as a Mechanism for Epigenome Evolution

TFBS turnover can explain species-specific TF binding events [307,308,310,318,319]. For example, across 4000 orthologous promoters between mouse and human, 41-89% of liver transcription factor binding locations were species-specific [309], suggesting a high amount of TFBS turnover since the last mouse-human common ancestor [304,320].

Formation of novel TFBS can disrupt and shift methylation patterns in the promoter region [282,284,321]. For example, one study described methylation determining regions that direct the DNA methylation status of promoter proximal CpGs during differentiation [282]. TFBS sites, including SP1, CTCF, and Rfx, were required for proper methylation [282]. Moreover, RE1-Silencing transcription factor (REST)-bound lowly methylated regions (LMRs) showed increased DNA methylation in *Rest* knock-out ESCs, indicating that REST is required for proper demethylation of LMRs [285]. Thus, evidence is mounting the DNA sequence polymorphisms in TFBSs may modulate DNA methylation status.

Accordingly, TFBS turnover events have been found to explain paralogs- or lineage-specific differential DNA methylation [269,307,308,322,323], DNase hypersensitivity [311], histone post-translational modifications [300,302], and TF binding events [2,304,307,308,324]. Comparison of orthologous CpGs in the primate lineage revealed that human-specific DMRs genome-wide were enriched for nucleotide substitutions in TFBSs, suggesting a close relationship between TFBS and DNA methylation patterns during human evolution [271]. Similarly, motifs for chromatin regulators or TFs associated with a particular chromatin state (such as SP1 and CTCF respectively) were enriched in epigenetically divergent paralogs [269]. Analysis of binding sites of pluripotency transcription factors in human, mouse, and pig ESCs

demonstrated that inter-species epigenetic differences explain species-differential binding and expression better than sequence differences [273]. Overall, examination of binding site turnover events and their epigenetic context supports the hypothesis that DNA sequence changes in the form of TFBS turnover events drive epigenetic variation that may regulate gene expression (**Figure 5.3**).

Specific TFBS motifs may also mediate epigenome conservation [324]. Preservation of DNA methylation patterns over duplicated genes was associated with the SP1 motif at paralogous promoters [283]. TFBS also mark regulatory elements with DNA methylation (Zhou, unpublished) or histone modification conservation between species [287] (**Figure 5.3B**). In each case, TFBS turnover might be a mechanism for canalization, codifying epigenetic modifications into genetic knowledge and ensuring robustness of a phenotype [323,325] (**Figure 5.4**).

5.8 Concluding Remarks

5.8.1 Challenges and Limitations for Comparative Epigenomics

Because epigenetic status varies with cell type, matching homologous tissues or cell types between species is required for rigorous epigenome comparison. However, matching homologous cell types between species is a non-trivial task [250,326], especially when developmental stage and environment may also need to be matched. One complication is how to determine identity by descent when differentiated cell types must be specified every generation [327]. This might be achieved by comparative molecular cell biology, as has been shown in the case of retinal cell evolution [326]. The comparative molecular biology approach analyzes expression of orthologous genes to identify homologus cell types. However the issue is made more complex by the realization that homologous genes may not direct development of homologous structures [327]. Instead, it has been proposed that conserved gene regulatory networks (GRNs) may control the development of homologous structures, and these GRNs may be comprised of non-orthologous genes [327]. By taking into account the nuances involved in determining biological homology when designing experiments [250], comparative epigenomics may determine how conserved (or species-specific) epigenomic features of homologous cell types regulate GRNs.

Between distantly related species, establishing homology becomes increasingly challenging. Thus the evolutionary history we can recover using comparative epigenomics is limited. Other technical challenges remain, which include variable genome build qualities and the accuracy of multiple genome sequence alignments [260]. As all Next-Generation Sequencing (NGS)-based assays are subject to batch effects, systematic biases in experimental design should be limited or corrected for as much as possible, as the combination with the technical limitations of inter-

species analysis could hinder the interpretability of results. Lastly, given the challenges for multi-species sequence alignment, how do we go about aligning the epigenome? Nucleotide bases are the units of the genome sequence; what are the units of the epigenome?

5.8.2 Future Directions for Epigenome Evolution Research

To efficiently benefit from comparative epigenome research, a model of epigenome evolution needs to be established. Epigenome analysis is a quickly maturing field, and the combination of epigenomics and computational modeling with classic technologies can make inroads on previously intractable questions of epigenetic gene regulation.

To build a model of epigenome evolution, the field must first answer basic questions about how epigenetic conservation relates to sequence evolution (**Figure 5.4**). For example, what kinds of epi-modifications occur at slowly evolving DNA sequences compared to fast evolving sequences? At slowly evolving sequences, we suggest the null hypothesis is that orthologs should display conserved epigenetic modifications; the alternate hypothesis is DNA-conserved orthologs have epi-divergent status. In this case support for the alternate hypothesis is evidence of regulatory innovation driven by epigenetic novelty. At quickly evolving DNA loci the hypotheses are flipped: the null hypothesis is that epi-marks will be different, but the alternate hypothesis is that they would be conserved. Evidence of the alternate hypothesis in this case would be a signature of epigenetic buffering [273,323].

Few studies have quantified the magnitude of epigenetic divergence or conservation with respect to sequence evolution. As a result, it is not yet clear how sequence evolution influences epigenome evolution (**Figure 5.2**). To ameliorate this knowledge gap, the table in the center of the proposed framework (**Figure 5.4**) should be populated with comparative epigenome studies that take into account sequence evolution as well as inter-species epigenetic differences. First,

conservation of three epi-marks (DNA methylation, H3K36me3, and H3K27ac) in mammal ESCs is independent of DNA sequence evolution [273], so regulatory elements identified in ref 34 fulfill both the null hypothesis (H_0) for slowly evolving sequence and the alternate hypothesis (H_{alt}) for fast evolving sequence. On the other hand, ample evidence exists that gene body epi-marks are conserved, and since most gene bodies are slowly evolving [328], the conservation of gene body epigenetic regulation is evidence for the null hypothesis of epigenome evolution over slowly evolving sequences (**Figure 5.4**).

Beyond evidence for each hypothesis, this framework prompts other questions such as what features distinguish divergent vs. conserved epi-marks at fast evolving DNA? How and at what rate can a regulatory element transition from a slow DNA-evolving/epigenetically-divergent status to a fast DNA-evolving/epigenetically-conserved status? What mechanisms mediate genetic assimilation of stable epigenetic regulatory architectures to be genetically encoded?

We view genetic evolution (the evolution of networks and pathways) as subsequent to epigenetic evolution. We note that comparative analysis of genetic networks has found that conservation of regulatory motifs and network topologies between mouse and human is much higher than conservation of individual TF binding sites [58]. Thus we suggest that evolution on the epigenome level is what mediates the flexibility of TF binding sites and thus innovation or stability of gene regulatory networks.

In the Outstanding Questions section, key areas of research are proposed that will help to fill in the proposed framework with empirical data. Lastly, rigorous mathematical models must accompany a mature model of epigenome evolution that quantitatively assess epigenome

conservation or divergence based on data, similar to the Jukes-Cantor model for sequence evolution.

5.8.3 Outstanding Questions

Hybrid cell culture systems can be used to study lineage-specific epigenome features. Cells harboring chromosomes from different species may help disentangle the environmental, trans-, or cis- effects on epigenome status.

Genomic editing technologies make genome editing feasible in any species, and can identify drivers of conserved and divergent epigenetic programs. Targets in homologous cell models may be altered in order to understand how the cis-landscape impacts epigenome state in different species. Similarly, analyzing epigenome status after knockout of tissue-specific transcription factors or chromatin modifiers may point to epigenetic effects that are conserved or species-specific, revealing mechanisms of the evolution of complex tissues and organs.

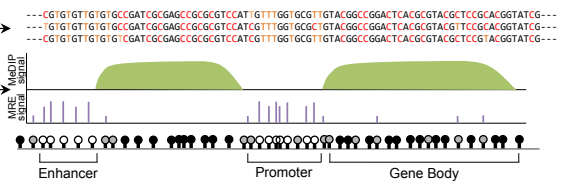
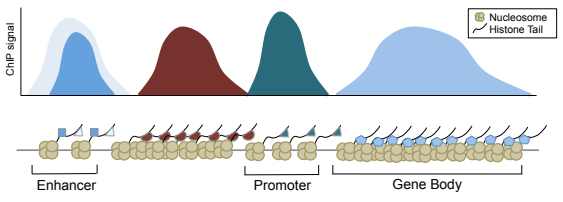
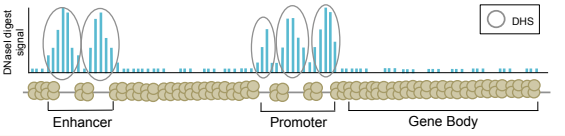
Applying single-cell technologies to comparative epigenomics can identify cell-type specific and temporal differences at higher resolution, elucidating how epi-modifications depend on sequence in an allele-specific manner.

While the determinants of nucleosome positioning in eukaryotic cells are similar, quantitative, comparative studies of nucleosome positioning is best explored in yeast. Some comparative analyses between phyla exist, but comparative nucleosome positioning in vertebrate genomes is a promising an area for future research.

Hi-C technology permits comparative analysis of chromatin architecture and nuclear territories. 3D fluorescent *in situ* hybridization experiments demonstrated that chromosome territories in the nucleus are largely conserved across primates. Comparative analyses of chromosome

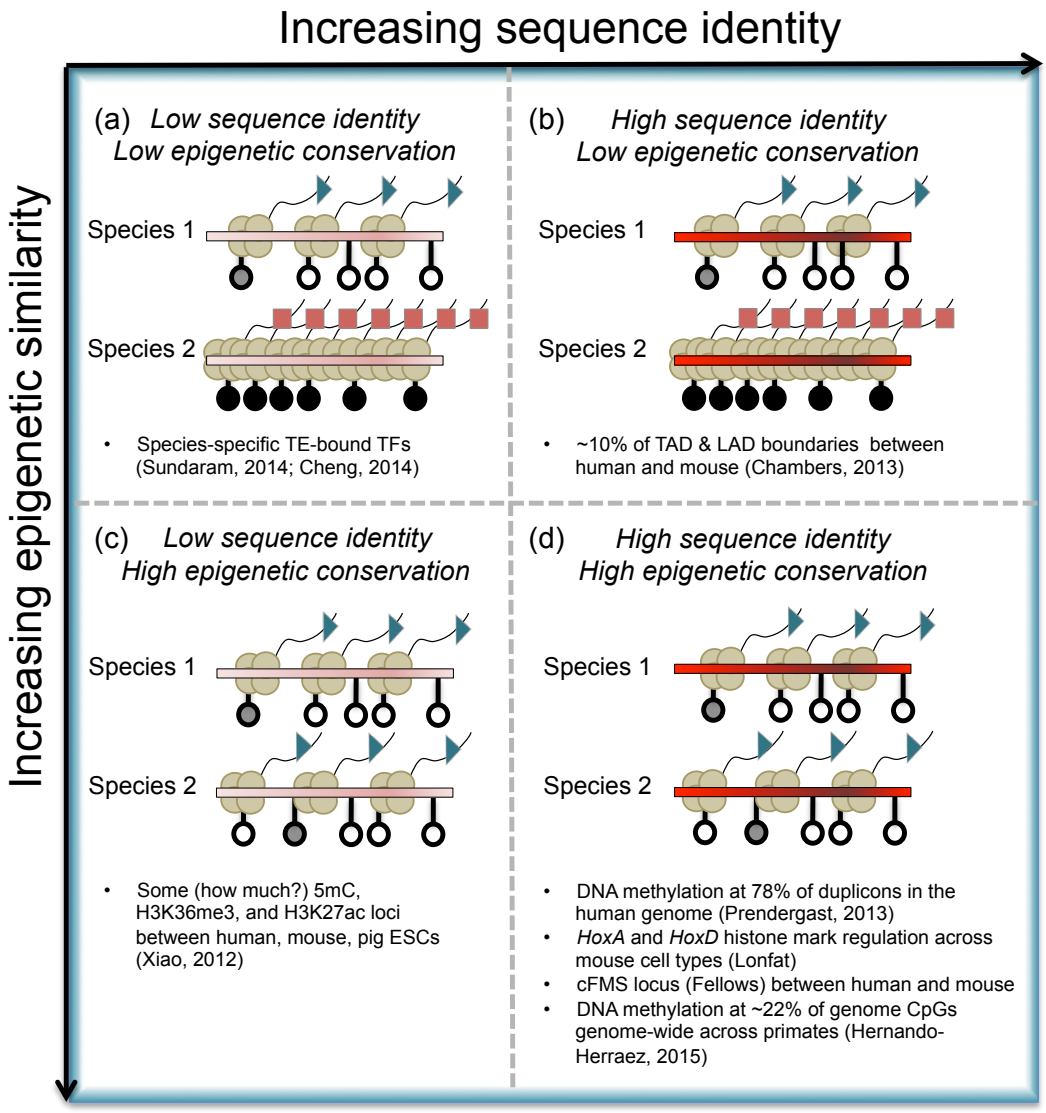
conformation can reveal how genome organization evolved and how nuclear architecture contributes to gene regulation.

Both tumor cells and induced pluripotent stem cells (iPSC) undergo epigenetic reprogramming. Oncogenic transformation may model epigenome evolution on an accelerated time scale. iPSC-cancer cell comparative epigenomics can expose how epigenetic mis-regulation contributes to oncogenesis.

Epigenetic Mechanisms	Modification Types*	Assay Examples	Simplified Diagrams
DNA Methylation	Low Intermediate High	Chemical-based Bisulfite-treatment# [329] RRBS[330] Enrichment-based MeDIP# [23] Bio-CAP# [331] Enzyme digest-based MRE# [23]	
Histone PTM	H3K4me1 H3K4me2 H3K4me3 H3K9me3 H3K27ac H3K27me3 H3K36me3	Enrichment-based Histone-specific ChIP# [332]	
Nucleosome Occupancy	DNase I Hypersensitive sites (DHS)	Enzyme digest-based DNase I# [51]	

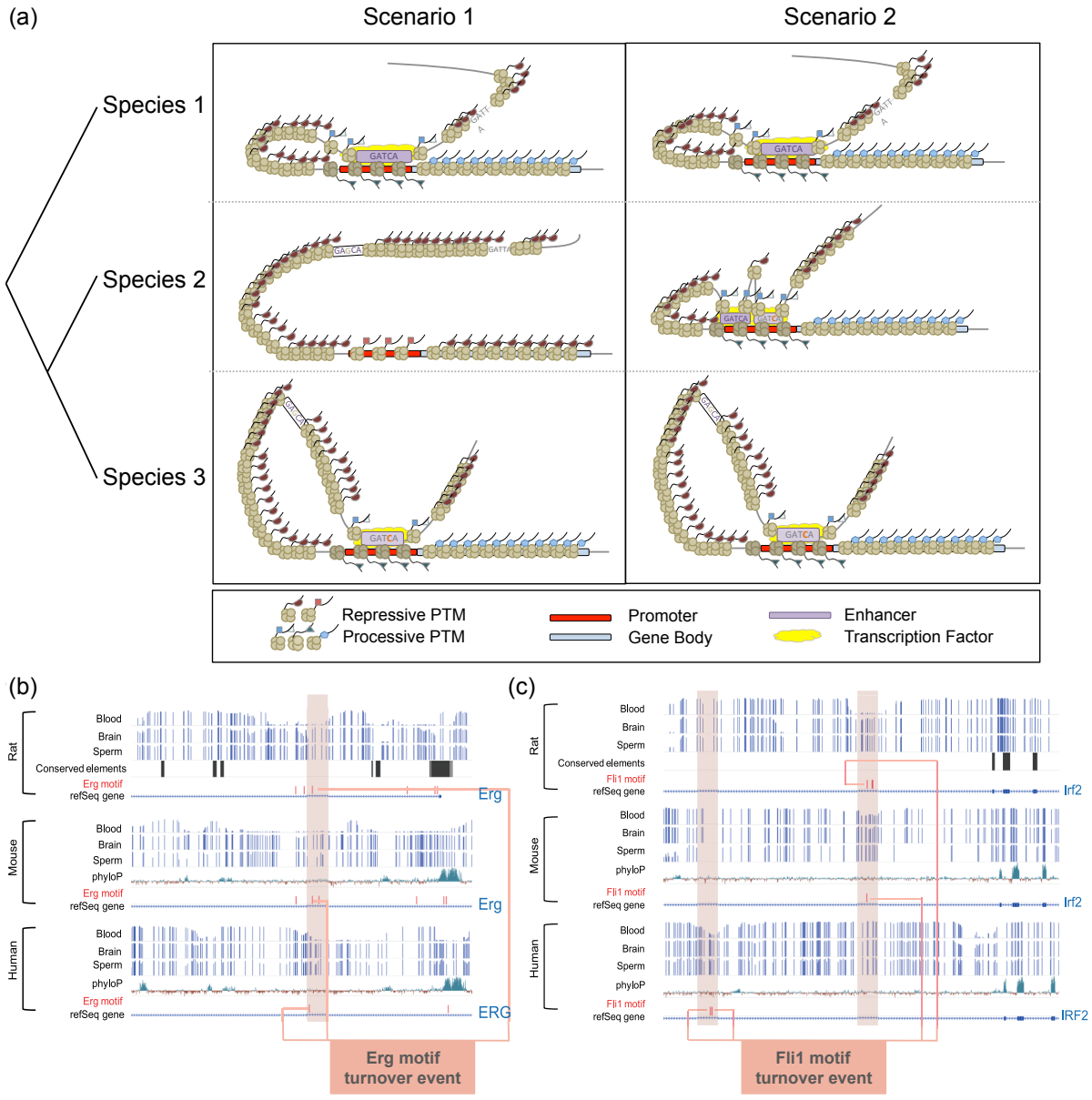
* = Red and blue terms correspond to repressive and processive state, respectively.
# = These assays can be quantified by numerous techniques including, but not limited to, gel-imaging, targeted sequencing, RT-qPCR, microarrays and high-throughput sequencing.
[99] = Reference

5.1. Dynamic epigenetic interactions. Innovation in sequence resolution and identification of novel DNA-modifying mechanisms have provided novel opportunities to develop intricate techniques to explore epigenetic interactions. This review focuses on four unique, but usually complementary, epigenetic modifications that are universally shared across vertebrates. Different types of modifications can have processive function, allowing expression of genes, or recessive function, hindering gene expression, or an intermediate poised state that has potential to go either direction. The biological function of individual epigenetic marks has been widely studied but the combinatorial interactions across epigenetic modifications have still yet to be fully defined and understood. Here, we diagram simplified models to illustrate how results of epigenetics assays can be interpreted in the resolution of DNA-context and chromatin-context [23,329-332].



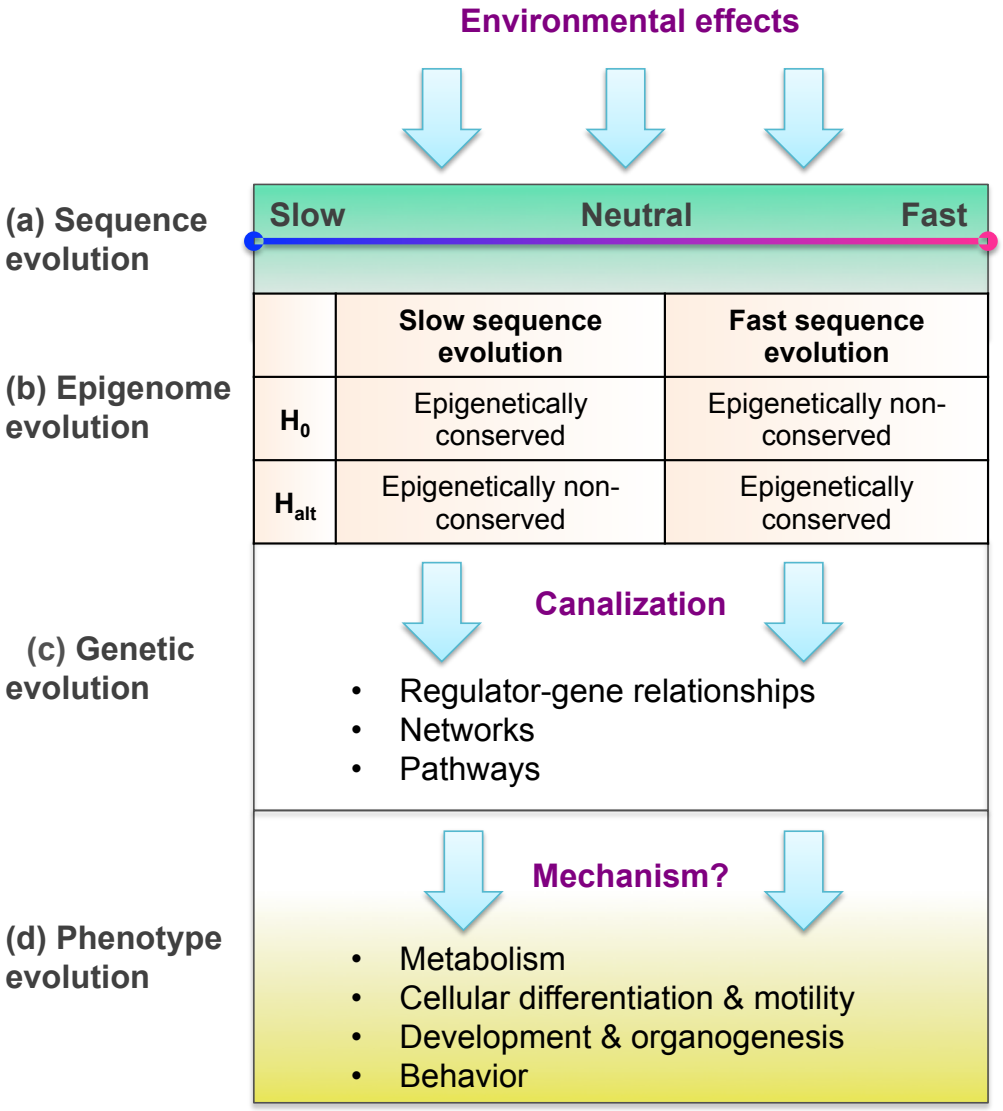
5.2. Genetic and epigenetic conservation correlation. The degree of sequence or epigenetic similarity between syntenic loci each run on a continuum. Here we define epigenetic similarity as having the same epigenetic signal at orthologous loci between the two species. While there are many degrees of variation within each of these four possibilities, we offer this general framework and examples from the literature for each combination of extremes. (a) Loci with low sequence identity and low epigenetic similarity may represent lineage-specific loci and include

non-orthologous regions. (b) A minority of orthologs demonstrate faster epi-mark divergence than sequence divergence. (c) Orthologs where the genome sequence is diverging faster than the epigenetic state represent loci that experienced enhancer turnover. Additionally, some marks are found in both fast and slowing evolving sequences, suggesting a mechanism for buffering genetic variation. (d) The majority of examples we can categorize exhibit both sequence conservation and epigenetic conservation. This is the status for most orthologs (inter- and intra-species) and represents the null hypothesis. Legend as in **Figure 5.1**; the intensity of shading of DNA strands represents the degree of sequence conservation.



5.3. TFBS turnover models and examples. Understanding TFBS turnover during evolution has been a non-trivial challenge. Since TFBS turnover is coupled with epigenetic changes, a null hypothesis that TFBS turnover is also associated with epigenetic evolution across species can be proposed. Although numerous examples of TFBS turnover has been documented, we propose two simple but powerful scenarios that can capture the process of TFBS turnover by comparing

epigenetic signal across orthologous regions across species. The diagram illustrates an orthologous gene region across three species. This can be interpreted differently by varying the window size or synteny. First scenario represents loss-gain TFBS turnover where species 1 had a TFBS but species 2 lost the TFBS by a single mutation in the binding site. However, in species 3, another genomic region was mutated to recover the lost TFBS and become the new enhancer. The second scenario is a competitive model where species 2 gained a mutation that generated another TFBS that competed with the species 1 enhancer. After selection or mutation, the species 1 enhancer is lost and the novel enhancer becomes the sole cis-regulator for the gene in species 3. This mechanism may mediate lineage-specific epigenetic marks [2,304,307,308,310] or conserve epigenetic features as in (b) and (c) [283]. (b) and (c) are representative examples of TFBS turnover events mediating a conserved tissue-specific DNA hypomethylated regions (pink shaded boxes) between rat, mouse, and human (adapted from Zhou, unpublished). Tracks in blue are single-CpG DNA methylation levels from the given tissue in each species. In (b) the Erg motif is found at the same position in rat and mouse, but shifted by 84bp in human. The Erg motif conserved the blood-specific DMR at this locus and is an example of senario 1 depicted in (a). In (c), The Fli1 motif is in a slightly different position in the conserved blood-specific DMR in rat and mouse, but absent in human. Instead the Fli1 motif is found in a nearby blood-specific DMR in human. Importantly, the conserved DMRs in these examples show low sequence conservation, evidence that conserved DNA hypomethylated regions do not depend on sequence conservation.



5.4. Model for building a theory of epigenome evolution. (a) Determining the rate of sequence evolution is now a straightforward process. (b) The expectation for epigenome evolution is different depending on the sequence evolution context. Completing this contingency table with specific examples is a challenge for the field. (c) Epigenetic gene regulation that is adaptive is genetically assimilated into the genome, codifying gene regulation and driving genetic evolution.

(d) Genetic networks drive phenotypic evolution, all of which is motivated by environmental inputs.

Chapter 6

Synthesis

6.1 Detecting Differential DNA Methylation During Development

Somatic genomes are highly methylated, but where they vary DNA methylation differences are biologically important [19,25,75,333,334]. Specifically, DNA methylation differences can direct differential gene regulation with important consequences. Epigenome and gene expression studies of twins discordant for disease provide valuable evidence in this regard [335-337]. For example, analysis of white blood cell DNA methylation in twins discordant for systemic lupus erythematosus (SLE) displayed differential methylation at 49 genes when compared to non-SLE co-twins and healthy controls, including SLE markers including interferon gamma receptor 2 (*IFGNR2*) [337]. Thus differential DNA methylation can have important consequences for human disease.

DNA methylation differences tend to occur at regulated promoters with low to intermediate CpG density and distal, cell-specific enhancers [14,25,54,55,108]. Importantly, active regulatory elements tend to be demethylated [27,78], so detecting signal of lack of methylation is an important factor when modeling the effect of DNA methylation on gene expression.

Our algorithm excels where others cannot provide data, namely, by integrating DNA methylation signal with DNA un-methylation signal. Our algorithm combines two data types, methylated DNA immuneprecipitation and sequencing (MeDIP-seq) and methylation-sensitive restriction enzyme digest and sequencing (MRE-seq), which query DNA methylation and single nucleotide unmethylated CpGs respectively [23,77].

The M&M algorithm dynamics scales sequencing data based on the number of covered CpGs in the MRE and MeDIP datasets. The algorithm then treats read counts for a specified non-overlapping window length as mutually independent Poisson random variables. M&M then models the expected values using a joint distribution of tag counts to test the hypothesis that

$$H_0: \mu_1 = \mu_2 \textit{ versus } H_1: \mu_1 \neq \mu_2 \qquad \text{(Equation 6.1)}$$

where μ_1 is the methylation level for the given window in sample 1 and μ_2 is the methylation level for the same window in sample 2 (Zhang, 2013).

Our algorithm was subject to rigorous testing and in several biological contexts. We compared biological replicates of H1 ESC datasets and H1 ESC to fetal neural stem cell data (Zhang, 2013). We found that M&M performs with higher sensitivity and specificity than other published methods and is better able to discriminate cell- and tissue-specific DMRs.

Because the M&M method was robust for identifying cell- and tissue type-specific DMRs, we employed MeDIP-seq & MRE-seq and M&M to the analysis of skin cell type epigenome. We compared skin cell type-specific methylomes to find skin cell type-specific and tissue-level DMRs [60]. Differential DNA methylation modifications are passed on to daughter cells and observed in differentiated tissues. Thus our particular samples allowed for a novel analysis

strategy: identifying DMRs specific to a developmental tissue, in this case, surface ectoderm. We achieved this by analyzing the cell type methylomes of skin and other surface ectoderm-derived cells to find the DMRs shared only by surface ectoderm-derived cell types.

Finally, we applied the M&M algorithm to DNA methylomes for several developmental times points in zebrafish embryogenesis. Zebrafish is an important developmental system as the embryo can be manipulated much more readily than human tissues at orthologous stages. We found thousands of DMRs between early developmental stages, many of which were validated in zebrafish reporter assays [27]. The findings in Lee, H.J., *et al.* validate that the M&M algorithm works appropriately on the zebrafish methylome. More importantly, this paper reiterated the literature on human data that describes dynamic DNA methylation over cell- and tissue-enhancers. Thus, we were encouraged to pursue experiments in zebrafish on specific subpopulations to understand dynamic DNA methylation in cell fate determination.

6.2 Validation of Developmental DMR Classes found in Human Skin Epigenome Analysis with Zebrafish Neural Crest Cell Experiments

In human, distinct genetic and epigenetic differences between developmental-stage regulatory elements have been observed. Genes involved early in lineage-specification, tend to be CG-dense and primarily regulated by H3K27me3 Polycomb deposited PTM, rather than DNA methylation [14,55,56].

Late-stage lineage-specific promoters tend to be less CG-dense and primarily repressed by DNA methylation [14,55]. However, distal enhancers undergo a transition from high DNA methylation to H3K27me3 in a lineage-specific fashion [54,55,338]. For example, increase of H3K27me3 coincided with FOXA2 pioneer factor binding in the ESC-differentiated endoderm sample [54].

A similar pattern was observed in cardiomyocyte differentiation in a zebrafish model [338].

The developmental dynamics analysis presented in chapter 2 agreed with the above characterization of chromatin and DNA methylation dynamics at enhancers. I found that DNA methylation dynamics over transcription factor binding sites in hypomethylated DMRs shared among surface ectoderm-derived cell types (SE-DMRs) separated into two classes. Class II contained most of the SE-DMRs, which were methylated in a model of early ectoderm-differentiated cells (dEC; hESCs differentiated into early ectoderm cells, Gifford 2013), but demethylated in the surface ectoderm-derived differentiated cell types, epidermal keratinocytes and mammary gland epithelial cells (called late-demethylated). Class I DMRs contained SE-DMRs that were lowly methylated in the dEC cells (called early-demethylated).

My analysis took the observation of individual enhancer epigenetic dynamics one step further. By assigning hypomethylated SE-DMRs to putative target genes, network analysis revealed that the early-demethylated class of SE-DMRs were predicted to regulate the genes of the cognate transcriptions factors for the motifs contained in the late-demethylated class of SE-DMRs. These regions were demethylated in a progenitor cell type represented by the dEC model (**Figure 2.18a**).

The pigment cell development project discussed in chapter 3 allowed an opportunity to validate the findings from chapter 2 in an independent system. Further, the zebrafish has the advantage that we can directly access progenitor cell populations, unlike in human, where instead we rely on cell-differentiation models for cell populations in early development. The zebrafish *crestinA>GFP* transgenic allowed us to isolate neural crest cells specifically at various developmental stages. The 24hpf GFP+ cell population represented a melanocyte/iridophore

progenitor-enriched population of neural crest cells. Thus in this project we can directly determine the DNA methylation dynamics of a progenitor cell population, rather than the indirect determination of surface ectoderm-differentially methylated regions described in chapter 2.

Our preliminary findings presented in chapter 3 are encouraging that we will observe similar progenitor-specific differentially methylated regions at sites of important regulatory elements in the 24hpf GFP+ cell population. One validation of our surface ectoderm work will be validating the class I and class II DMRs (as described in **6.1** above). As validation of class I DMRs, we anticipate finding a few regions that are DNA de-methylated in the 24hpf GFP+ population compared to earlier stages (14-somites) and non-neural crest cells (24hpf GFP-), but that remain lowly methylated in differentiated pigment cells. Class II DMRs will be observed as regions that are specifically hypomethlyated in melanocytes or iridophores compared to the 24hpf GFP+ progenitor population. Further, we hypothesize that class I DMRs occur at regulatory elements that control expression of key pigment cell genes, for example, *mitfa*, a marker of melanoblast and iridoblasts [136] (**Figure 3.9b**). Class II DMRs will be those that control melanocyte- or idirophore-specific genes, for example, the DMR just upstream of the *pnp4a* promoter (**Figure 3.9c**). Lastly, we will examine these class I and class II DMRs for enrichment of transcription factor binding site motifs. As demonstrated in chapter 2, sequence analysis of DMRs can reveal important transcriptional regulators in development. Thus, epigenomic analysis of pigment cell differentiation in zebrafish should validate and illuminate DNA methylation dynamics that we observed in human development.

6.3 Enhancer Dysregulation in Cancer

Chapter 4 takes a different approach to understanding the role of the epigenome in development and disease. While chapters 2 and 3 focus on DNA methylation (and to a lesser extend histone

modification) dynamics as instructive cues for cell fate decisions, in chapter 4 we utilize epigenomic data to annotate genetic variation in a specific disease. We chose to focus on primary liver cancer (PLC) for two reasons: (1) there is an abundance of somatic mutation data available for PLC, and (2) the relatively homogenous nature of normal liver tissue makes identifying epigenetically-defined regulatory elements in liver more straightforward and specific. In this project we hypothesized that annotating mutated fragments with epigenomic data could reveal the function of somatic mutations in PLC. Further, we are particularly interested in the control of cell fate and how cell identity deteriorates in cancer. It has long been noted that loss of differentiation occurs during malignant tumorigenesis [339]. We had previously observed that epigenomic alterations in cancer samples had the effect of down-regulating genes expressed in the normal tissue, but up-regulating genes for heterologous cell types [340]. Thus in chapter 6 we proposed that loss of differentiation might coincide with appropriation of transcriptional regulatory regimes from heterologous cells.

To investigate our hypothesis, we examined PLC somatic mutations in the context of normal epigenome annotations for normal liver and 77 heterologous cell or tissue types. We found elevated somatic mutation rates in DNaseI-defined promoters as well as enhancer and promoter elements as defined by histone modification integration and segmentation algorithms (**Figure 4.2c-d**). We further characterize the potential mechanism of regulatory element somatic mutations by examining gain or loss of transcription factor binding sites. We identified widespread gain-of-binding site events, some of which occurred in promoters of proto-oncogenes, activating regulatory elements native to other cell types, but not normal liver (**Figure 4.4**). Thus we consider the evidence presented in chapter 4 as support for our hypothesis that

transcriptional regulatory instructions for heterologous cell types are co-opted during malignant tumorigenesis.

The analysis presented in chapter 5 is related to the previous chapter insofar as all the projects in this dissertation are concerned with the establishment and activity of regulatory elements in development. Chapters 2 and 3 examine the normal establishment of regulatory elements and their epigenomic dynamics in two distinct developmental contexts. Chapter 4 examines the consequences of disrupting regulatory element function, in this case by somatic mutation. Other work that I contributed to during the course of my dissertation examines the consequences for cell fate and cancer when regulatory elements are disrupted by epigenomic alteration [340]. Thus altogether, this dissertation has pushed the field of functional genomics and developmental biology in several ways: (1) by establishing novel algorithms for detecting differentially methylated regions [78]; (2) by developing new analysis approaches to detecting DNA methylation dynamics for an early embryonic tissue [60]; (3) by pioneering application of DNA methylome technologies to specific embryonic cell populations; (4) by presenting evidence that transcriptional regulatory instructions are re-directed in malignant cancer cells such that cancer cells acquire regulatory cues native to heterologous cells.

References

1. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969 Jul 25;165(3891):349–57.

2. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature Genetics. 2010 Jul;42(7):631–4.

3. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nature Genetics. 2011 Nov;43(11):1154–9.

4. Leeb M, Wutz A. Establishment of epigenetic patterns in development. Chromosoma. 2012 Mar 17;121(3):251–62.

5. Feng S, Jacobsen SE, Reik W. Epigenetic reprogramming in plant and animal development. Science. 2010 Oct 29;330(6004):622–7.

6. Waddington CH. Genetic assimilation of an acquired character. Evolution. 1953;7(2):118.

7. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. 2011 Jul 18;12(8):565–75.

8. Wallrath LL, Elgin SC. Position effect variegation in Drosophila is associated with an altered chromatin structure. Genes & Development. Cold Spring Harbor Lab; 1995;9(10):1263–77.

9. Wyatt GR. The purine and pyrimidine composition of deoxypentose nucleic acids. Biochemical Journal. Portland Press Ltd; 1951 May 1;48(5):584–1495.

10. Bird A. DNA methylation patterns and epigenetic memory. Genes & Development. 2002 Jan 1;16(1):6–21.

11. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013 Oct 23;502(7472):472–9.

12. Sinsheimer RL. The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. Journal of Biological Chemistry. 1955.

13. Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. PNAS. National Acad Sciences; 2010 May 11;107(19):8689–94.

14. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. Nature Publishing Group; 2008 Jul 6;454(7205):766–70.

15. Chomet PS, Wessler S, Dellaporta SL. Inactivation of the maize transposable element Activator (Ac) is associated with its DNA modification. EMBO J. 1987 Feb;6(2):295–302.

16. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature Genetics. 1998 Oct;20(2):116–7.

17. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature. 2001 May 10;411(6834):212–4.

18. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nature Publishing Group; 2008 Jun;9(6):465–76.

19. Illingworth R, Kerr A, Desousa D, Jørgensen H, Ellis P, Stalker J, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol. Public Library of Science; 2008 Jan;6(1):e22.

20. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011 May 16;25(10):1010–22.

21. Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell. 1991 Mar;64(6):1123–34.

22. Watt F, Molloy PL. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes & Development. Cold Spring Harbor Lab; 1988 Sep 1;2(9):1136–43.

23. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010 Jul 8;466(7303):253–7.

24. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. 2013 Nov;23(11):1256–69.

25. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011 Dec 22;480(7378):490–5.

26. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics. 2013 May 26;45(7):836–41.

27. Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, et al. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. Nat Comms. 2015;6:6315.

28. Wang Y, Wysocka J, Perlin JR, Leonelli L, Allis CD, Coonrod SA. Linking Covalent Histone Modifications to Epigenetics: The Rigidity and Plasticity of the Marks. Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press; 2004 Jan 1;69(0):161–70.

29. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. Nature Publishing Group; 2000 Jan 6;403(6765):41–5.

30. Milne TA, Dou Y, Martin ME, Brock HW, Roeder RG, Hess JL. MLL associates specifically with a subset of transcriptionally active target genes. Proceedings of the National Academy of Sciences. National Acad Sciences; 2005 Oct 11;102(41):14765–70.

31. Francis NJ, Kingston RE, Woodcock CL. Chromatin Compaction by a Polycomb Group Protein Complex. Science. American Association for the Advancement of Science; 2004 Nov 26;306(5701):1574–7.

32. Ho L, Crabtree GR. Chromatin remodelling during development. Nature. Nature Publishing Group; 2010 Jan 28;463(7280):474–84.

33. Simon JA, Kingston RE. Mechanisms of Polycomb gene silencing: knowns and unknowns. Nature Reviews Molecular Cell Biology. Nature Publishing Group; 2009 Oct 1;10(10):697–708.

34. Hosogane M, Funayama R, Shirota M, Nakayama K. Lack of Transcription Triggers H3K27me3 Accumulation in the Gene Body. Cell Reports. 2016 Jul;16(3):696–706.

35. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. Nature Publishing Group; 2006 Nov 5;444(7118):499–502.

36. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research. Cold Spring Harbor Lab; 2007 Jun 1;17(6):760–74.

37. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011 Oct 27;478(7370):476–82.

38. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009 Sep 10;461(7261):199–205.

39. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nature Genetics. Nature Publishing Group; 2010 Aug 22;42(9):806–10.

40. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. Nature Publishing Group; 2007 Jun 14;447(7146):799–816.

41. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

42. Hardison R, Slightom JL, Gumucio DL, Goodman M, Stojanovic N, Miller W. Locus control regions of mammalian β-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. Gene. 1997 Dec;205(1-2):73–94.

43. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nature Publishing Group; 2012 Sep 1;13(9):613–26.

44. Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, et al. Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding. Mol Cell. 2011 Jul;43(1):145–55.

45. Sérandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C, et al. Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome Research. Cold Spring Harbor Lab; 2011 Apr;21(4):555–65.

46. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics. Oxford University Press; 2003 Jul 15;12(14):1725–35.

47. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nature Genetics. 2015 Oct 26.

48. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013 Nov;155(4):934–47.

49. Serandour AA, Avner S, Oger F, Bizot M, Percevault F, Lucchetti-Miganeh C, et al. Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. Nucleic Acids Research. 2012 Sep 26;40(17):8255–65.

50. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, et al. Nucleosome dynamics define transcriptional enhancers. Nature Genetics. Nature Research; 2010 Apr

1;42(4):343–7.

51. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012 Sep 6;489(7414):75–82.

52. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009 May 7;459(7243):108–12.

53. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. Nature Publishing Group; 2014 Mar 27;507(7493):455–61.

54. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, et al. Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. Cell. 2013 May.

55. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell. 2013 May.

56. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, et al. Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. Mol Cell. 2008 Jun;30(6):755–66.

57. Vasanthi D, Mishra RK. Epigenetic regulation of genes during development: a conserved theme from flies to mammals. J Genet Genomics. 2008 Jul;35(7):413–29.

58. Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, et al. Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. Cell. Elsevier; 2013 Aug;154(4):888–903.

59. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nature Genetics. Nature Publishing Group; 2013 Sep 1;45(10):1198–206.

60. Lowdon RF, Zhang B, Bilenky M, Mauro T, Li D, Gascard P, et al. Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. Nat Comms. 2014;5:5442.

61. Jordan CT, Guzman ML, Noble M. Cancer Stem Cells. N Engl J Med. 2006 Sep 21;355(12):1253–61.

62. Cozzio A, Passegué E, Ayton PM, Karsunky H, Cleary ML, Weissman IL. Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. Genes & Development. Cold Spring Harbor Lab; 2003 Dec 15;17(24):3029–35.

63. Krivtsov AV, Twomey D, Feng Z, Stubbs MC, Wang Y, Faber J, et al. Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. Nature. 2006 Jul 16;442(7104):818–22.

64. Huntly BJP, Shigematsu H, Deguchi K, Lee BH, Mizuno S, Duclos N, et al. MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. Cancer Cell. 2004 Dec;6(6):587–96.

65. Barnett SC, Robertson L, Graham D, Allan D, Rampling R. Oligodendrocyte-type-2 astrocyte (O-2A) progenitor cells transformed with c-myc and H-ras form high-grade glioma after stereotactic injection into the rat brain. Carcinogenesis. 1998 Sep;19(9):1529–37.

66. Parker SCJ, Gartner J, Cardenas-Navia I, Wei X, Ozel Abaan H, Ajay SS, et al. Mutational Signatures of De-Differentiation in Functional Non-Coding Regions of Melanoma Genomes. Horwitz MS, editor. PLoS Genet. 2012 Aug 9;8(8):e1002871.

67. Higdon CW, Mitra RD, Johnson SL. Gene Expression Analysis of Zebrafish Melanocytes, Iridophores, and Retinal Pigmented Epithelium Reveals Indicators of Biological Function and Developmental Origin. PLoS ONE. Public Library of Science; 2013 Jul 9;8(7):e67801.

68. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2009 Dec 16;463(7278):191–6.

69. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, et al. TERT promoter mutations in familial and sporadic melanoma. Science. American Association for the Advancement of Science; 2013 Feb 22;339(6122):959–61.

70. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science. 2013 Feb 22;339(6122):957–9.

71. Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA, et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. PNAS. National Acad Sciences; 2013 Apr 9;110(15):6021–6.

72. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. Nature Publishing Group; 2014 Jul 24;511(7510):428–34.

73. Drier Y, Cotton MJ, Williamson KE, Gillespie SM, Ryan RJH, Kluk MJ, et al. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. Nature Genetics. Nature Publishing Group; 2016 Mar 1;48(3):265–72.

74. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of

proto-oncogenes by disruption of chromosome neighborhoods. Science. American Association for the Advancement of Science; 2016 Mar 3;351(6280):aad9024–1458.

75. Kitamura E, Igarashi J, Morohashi A, Hida N, Oinuma T, Nemoto N, et al. Analysis of tissue-specific differentially methylated regions (TDMs) in humans. Genomics. 2007 Mar;89(3):326–37.

76. Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Research. 2013 Sep;23(9):1522–40.

77. Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods. 2015 Jan 15;72:29–40.

78. Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Research. 2013 Sep 1;23(9):1522–40.

79. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature Research; 2010 Oct 1;28(10):1097–105. Available from: http://www.nature.com/doifinder/10.1038/nbt.1682

80. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013 Aug 9;341(6146):1237905–5.

81. Nelson CM, Bissell MJ. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. Annu Rev Cell Dev Biol. 2006;22(1):287–309.

82. Dumont N, Wilson MB, Crawford YG, Reynolds PA, Sigaroudinia M, Tlsty TD. Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers. PNAS. 2008 Sep 30;105(39):14867–72.

83. DesRochers TM, Shamis Y, Alt-Holland A, Kudo Y, Takata T, Wang G, et al. The 3D tissue microenvironment modulates DNA methylation and E-cadherin expression in squamous cell carcinoma. Epigenetics. 2012 Jan 1;7(1):34–46.

84. James WD, Elston DM, Berger TG, Andrews GC. Andrews' Diseases of the Skin : Clinical Dermatology. London : Saunders Elsevier; 2011.

85. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. Nature Research; 2010 May 1;28(5):495–501.

86. Sell S, editor. Stem cells handbook. Second. New York: Springer; 2004.

87. Wang X, Pasolli HA, Williams T, Fuchs E. AP-2 factors act in concert with Notch to

orchestrate terminal differentiation in skin epidermis.

88. Dai X, Segre JA. Transcriptional control of epidermal specification and differentiation. Current Opinion in Genetics & Development. 2004 Oct;14(5):485–91.

89. Yori JL, Johnson E, Zhou G, Jain MK, Keri RA. Kruppel-like factor 4 inhibits epithelial-to-mesenchymal transition through regulation of E-cadherin gene expression. Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology; 2010 May 28;285(22):16854–63.

90. Zhang J, Brewer S, Huang J, Williams T. Overexpression of transcription factor AP-2α suppresses mammary gland growth and morphogenesis. Dev Biol. 2003 Apr;256(1):128–46.

91. Jäger R, Werling U, Rimpf S, Jacob A, Schorle H. Transcription Factor AP-2γ Stimulates Proliferation and Apoptosis and Impairs Differentiation in a Transgenic Model. Mol Cancer Res. Molecular Cancer Research; 2003 Oct 1;1(12):921–9.

92. Green KJ, Jones JC. Desmosomes and hemidesmosomes: structure and function of molecular components. The FASEB Journal. 1996 Jun;10(8):871–81.

93. Bouras T, Pal B, Vaillant F, Harburg G, Asselin-Labat M-L, Oakes SR, et al. Notch Signaling Regulates Mammary Stem Cell Function and Luminal Cell-Fate Commitment. Cell Stem Cell. 2008 Oct 9;3(4):429–41.

94. Okuyama R, Tagami H, Aiba S. Notch signaling: its role in epidermal homeostasis and in the pathogenesis of skin diseases. J Dermatol Sci. 2008 Mar;49(3):187–94.

95. Slavik MA, Allen-Hoffmann BL, Liu BY, Alexander CM. Wnt signaling induces differentiation of progenitor cells in organotypic keratinocyte cultures. BMC Dev Biol. 2007;7(1):9.

96. Alexander CM, Goel S, Fakhraldeen SA, Kim S. Wnt signaling in mammary glands: plastic cell fates and combinatorial signaling. Cold Spring Harb Perspect Biol. 2012;4(10):a008037–7.

97. Fine J-D, Eady RAJ, Bauer EA, Bauer JW, Bruckner-Tuderman L, Heagerty A, et al. The classification of inherited epidermolysis bullosa (EB): Report of the Third International Consensus Meeting on Diagnosis and Classification of EB. Journal of the American Academy of Dermatology. 2008 Jun;58(6):931–50.

98. Crew VK, Burton N, Kagan A, Green CA, Levene C, Flinter F, et al. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. Blood. American Society of Hematology; 2004 Oct 15;104(8):2217–23.

99. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, et al. Disruption of an AP-2α binding site in an IRF6 enhancer is associated with cleft lip.

Nature Genetics. Nature Publishing Group; 2008 Oct 5;40(11):1341–7.

100. Bailey CM, Hendrix MJC. IRF6 in development and disease: A mediator of quiescence and differentiation. cc. 2008 Jul 1;7(13):1925–30.

101. Richardson RJ, Dixon J, Malhotra S, Hardman MJ, Knowles L, Boot-Handford RP, et al. Irf6 is a key determinant of the keratinocyte proliferation-differentiation switch. Nature Genetics. Nature Publishing Group; 2006 Oct 15;38(11):1329–34.

102. Medina A, Ghaffari A, Kilani RT, Ghahary A. The role of stratifin in fibroblast–keratinocyte interaction. Mol Cell Biochem. Springer US; 2007;305(1-2):255–64.

103. Lodygin D, Hermeking H. Epigenetic silencing of 14-3-3sigma in cancer. Seminars in Cancer Biology. 2006 Jun 1;16(3):214–24.

104. Herron BJ, Liddell RA, Parker A, Grant S, Kinne J, Fisher JK, et al. A mutation in stratifin is responsible for the repeated epilation (Er) phenotype in mice. Nature Genetics. 2005 Nov;37(11):1210–2.

105. Ingraham CR, Kinoshita A, Kondo S, Yang B, Sajan S, Trout KJ, et al. Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (Irf6). Nature Genetics. 2006 Oct 15;38(11):1335–40.

106. Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008 May;10(5):593–601.

107. Neves R, Scheel C, Weinhold S, Honisch E, Iwaniuk KM, Trompeter H-I, et al. Role of DNA methylation in miR-200c/141 cluster silencing in invasive breast cancer cells. BMC Research Notes 2010 3:1. BioMed Central; 2010 Aug 3;3(1):1.

108. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013 Aug 22;500(7463):477–81.

109. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012 Jul 1;488(7409):116–20.

110. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Research. Cold Spring Harbor Lab; 2013 Mar;23(3):555–67.

111. Normand J, Karasek MA. A method for the isolation and serial propagation of keratinocytes, endothelial cells, and fibroblasts from a single punch biopsy of human skin. In Vitro Cell Dev Biol Anim. 1995 Jun;31(6):447–55.

112. Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction

enzyme sequencing methods. 2013.

113. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. Oxford University Press; 2011 Jun 1;27(11):1571–2.

114. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

115. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. BioMed Central; 2008 Sep 17;9(9):1.

116. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. Oxford University Press; 2011 Apr 1;27(7):1017–8.

117. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Research. Oxford University Press; 2013 Nov 4;42(D1):gkt997–D147.

118. Wong CK, Vaske CJ, Ng S, Sanborn JZ, Benz SC, Haussler D, et al. The UCSC Interaction Browser: multidimensional data views in pathway context. Nucleic Acids Research. 2013 Jun 22;41(W1):W218–24.

119. Koster MI, Roop DR. Mechanisms Regulating Epithelial Stratification. Annu Rev Cell Dev Biol. Annual Reviews; 2007 Nov;23(1):93–113.

120. Sen GL, Boxer LD, Webster DE, Bussat RT, Qu K, Zarnegar BJ, et al. ZNF750 is a p63 target gene that induces KLF4 to drive terminal epidermal differentiation. Dev Cell. 2012 Mar 13;22(3):669–77.

121. Koster MI, Kim S, Huang J, Williams T, Roop DR. TAp63α induces AP-2γ as an early event in epidermal morphogenesis. Dev Biol. 2006 Jan;289(1):253–61.

122. Antonini D, Rossi B, Han R, Minichiello A, Di Palma T, Corrado M, et al. An autoregulatory loop directs the tissue-specific expression of p63 through a long-range evolutionarily conserved enhancer. Molecular and Cellular Biology. 2006 Apr;26(8):3308–18.

123. Bronner-Fraser M. Origins and Developmental Potential of the Neural Crest. Experimental Cell Research. 1995 Jun;218(2):405–17.

124. LaBonne C, Bronner-Fraser M. Neural crest induction in Xenopus: evidence for a two-signal model. Development. 1998 Jul;125(13):2403–14.

125. Prasad MS, Sauka-Spengler T, LaBonne C. Induction of the neural crest state: Control of stem cell attributes by gene regulatory, post-transcriptional and epigenetic interactions. Dev Biol. 2012 Jun;366(1):10–21.

126. Manderfield LJ, Engleka KA, Aghajanian H, Gupta M, Yang S, Li L, et al. Pax3 and Hippo Signaling Coordinate Melanocyte Gene Expression in Neural Crest. Cell Reports. 2014 Dec;9(5):1885–95.

127. Milet C, Monsoro-Burq AH. Neural crest induction at the neural plate border in vertebrates. Dev Biol. 2012 Jun 1;366(1):22–33.

128. Plouhinec J-L, Roche DD, Pegoraro C, Figueiredo AL, Maczkowiak F, Brunet LJ, et al. Pax3 and Zic1 trigger the early neural crest gene regulatory network by the direct activation of multiple key neural crest specifiers. Dev Biol. 2014 Feb;386(2):461–72.

129. Taneyhill LA, Coles EG, Bronner-Fraser M. Snail2 directly represses cadherin6B during epithelial-to-mesenchymal transitions of the neural crest. Development. 2007 Apr;134(8):1481–90.

130. Kerosuo L, Bronner-Fraser M. What is bad in cancer is good in the embryo: importance of EMT in neural crest development. Semin Cell Dev Biol. 2012 May;23(3):320–32.

131. Simões-Costa M, Tan-Cabugao J, Antoshechkin I, Sauka-Spengler T, Bronner ME. Transcriptome analysis reveals novel players in the cranial neural crest gene regulatory network. Genome Research. 2014 Feb;24(2):281–90.

132. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Dev Dyn. 1995 Jul;203(3):253–310.

133. Lister JA, Cooper C, Nguyen K, Modrell M, Grant K, Raible DW. Zebrafish Foxd3 is required for development of a subset of neural crest derivatives. Dev Biol. 2006 Feb;290(1):92–104.

134. Schlagbauer-Wadl H, Griffioen M, van Elsas A, Schrier PI, Pustelnik T, Eichler HG, et al. Influence of increased c-Myc expression on the growth characteristics of human melanoma. J Investig Dermatol. 1999 Mar;112(3):332–6.

135. Lister JA, Robertson CP, Lepage T, Johnson SL, Raible DW. nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate. Development. 1999 Sep;126(17):3757–67.

136. Curran K, Lister JA, Kunkel GR, Prendergast A, Parichy DM, Raible DW. Interplay between Foxd3 and Mitf regulates cell fate plasticity in the zebrafish neural crest. Dev Biol. 2010 Aug 1;344(1):107–18.

137. Rawls JF, Mellgren EM, Johnson SL. How the Zebrafish Gets Its Stripes. Dev Biol. 2001 Dec;240(2):301–14.

138. Goding CR. Mitf from neural crest to melanoma: signal transduction and transcription in the melanocyte lineage. Genes & Development. 2000.

139. Hultman KA, Budi EH, Teasley DC, Gottlieb AY, Parichy DM, Johnson SL. Defects in

ErbB-Dependent Establishment of Adult Melanocyte Stem Cells Reveal Independent Origins for Embryonic and Regeneration Melanocytes. Kelsh RN, editor. PLoS Genet. Public Library of Science; 2009 Jul 3;5(7):e1000544.

140. Hultman KA, Johnson SL. Differential contribution of direct-developing and stem cell-derived melanocytes to the zebrafish larval pigment pattern. Dev Biol. 2010 Jan 15;337(2):425–31.

141. Tryon RC, Higdon CW, Johnson SL. Lineage relationship of direct-developing melanocytes and melanocyte stem cells in the zebrafish. PLoS ONE. 2011;6(6):e21010.

142. Johnson SL, Nguyen AN, Lister JA. mitfa is required at multiple stages of melanocyte differentiation but not to establish the melanocyte stem cell. Dev Biol. Elsevier; 2011;350(2):405–13.

143. Stewart RA, Arduini BL, Berghmans S, George RE, Kanki JP, Henion PD, et al. Zebrafish foxd3 is selectively required for neural crest specification, migration and survival. Dev Biol. 2006 Apr;292(1):174–88.

144. Curran K, Raible DW, Lister JA. Foxd3 controls melanophore specification in the zebrafish neural crest by regulation of Mitf. Dev Biol. 2009 Aug;332(2):408–17.

145. Nitzan E, Krispin S, Pfaltzgraff ER, Klar A, Labosky PA, Kalcheim C. A dynamic code of dorsal neural tube genes regulates the segregation between neurogenic and melanogenic neural crest cells. Development. Oxford University Press for The Company of Biologists Limited; 2013 Jun;140(11):2269–79.

146. Anderson RM, Bosch JA, Goll MG, Hesselson D, Dong PDS, Shin D, et al. Loss of Dnmt1 catalytic activity reveals multiple roles for DNA methylation during pancreas development and regeneration. Dev Biol. 2009 Oct 1;334(1):213–23.

147. Tittle RK, Sze R, Ng A, Nuckels RJ, Swartz ME, Anderson RM, et al. Uhrf1 and Dnmt1 are required for development and maintenance of the zebrafish lens. Dev Biol. 2011 Feb 1;350(1):50–63.

148. Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR. DNA Demethylation in Zebrafish Involves the Coupling of a Deaminase, a Glycosylase, and Gadd45. Cell. 2008 Dec;135(7):1201–12.

149. Jiang L, Zhang J, Wang J-J, Wang L, Zhang L, Li G, et al. Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. Cell. 2013 May;153(4):773–84.

150. Potok ME, Nix DA, Parnell TJ, Cairns BR. Reprogramming the Maternal Zebrafish Genome after Fertilization to Match the Paternal Methylation Pattern. Cell. 2013 May;153(4):759–72.

151. Aday AW, Zhu LJ, Lakshmanan A, Wang J, Lawson ND. Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1

and H3K4me3 binding sites. Dev Biol. 2011 Sep 15;357(2):450–62.

152. Bajpai R, Chen DA, Rada-Iglesias A, Zhang J, Xiong Y, Helms J, et al. CHD7 cooperates with PBAF to control multipotent neural crest formation. Nature. 2010 Feb 18;463(7283):958–62.

153. Tien CL, Jones A, Wang H, Gerigk M, Nozell S, Chang C. Snail2/Slug cooperates with Polycomb repressive complex 2 (PRC2) to regulate neural crest development. Development. 2015 Feb 10;142(4):722–31.

154. Jacob C, Lotscher P, Engler S, Baggiolini A, Varum Tavares S, Brugger V, et al. HDAC1 and HDAC2 Control the Specification of Neural Crest Cells into Peripheral Glia. Journal of Neuroscience. 2014 Apr 23;34(17):6112–22.

155. Strobl-Mazzulla PH, Bronner ME. A PHD12-Snail2 repressive complex epigenetically mediates neural crest epithelial-to-mesenchymal transition. J Cell Biol. 2012 Sep 17;198(6):999–1010.

156. Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. Epigenomic Annotation of Enhancers Predicts Transcriptional Regulators of Human Neural Crest. Cell Stem Cell. 2012 Nov;11(5):633–48.

157. Ignatius MS, Moose HE, El-Hodiri HM, Henion PD. colgate/hdac1 repression of foxd3 expression is required to permit mitfa-dependent melanogenesis. Dev Biol. 2008 Jan;313(2):568–83.

158. Rubinstein AL, Lee D, Luo R, Henion PD, Halpern ME. Genes dependent on zebrafish cyclops function identified by AFLP differential gene expression screen. Genesis. 2000 Jan;26(1):86–97.

159. Kelsh RN, Brand M, Jiang YJ, Heisenberg CP, Lin S, Haffter P, et al. Zebrafish pigmentation mutations and the processes of neural crest development. Development. 1996 Dec;123:369–89.

160. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. Oxford University Press; 2016 May 15;32(10):1446–53.

161. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

162. Anders S, Pyl PT, Huber W. HTSeq – A Python framework to work with high-throughput sequencing data. Bioinformatics. Oxford University Press; 2014 Sep 25;31(2):btu638–169.

163. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research. 2014 Jun;42(11):e91–1.

164. Li W, Cornell RA. Redundant activities of Tfap2a and Tfap2c are required for neural crest induction and development of other non-neural ectoderm derivatives in zebrafish embryos. Dev Biol. 2007 Apr 1;304(1):338–54.

165. Parichy DM, Rawls JF, Pratt SJ, Whitfield TT, Johnson SL. Zebrafish sparse corresponds to an orthologue of c-kit and is required for the morphogenesis of a subpopulation of melanocytes, but is not essential for hematopoiesis or primordial germ cell development. Development. The Company of Biologists Ltd; 1999 Aug 1;126(15):3425–36.

166. Lang MR, Patterson LB, Gordon TN, Johnson SL, Parichy DM. Basonuclin-2 requirements for zebrafish adult pigment pattern development and female fertility. Barsh GS, editor. PLoS Genet. 2009 Nov;5(11):e1000744.

167. Kawakami K. Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. Methods Cell Biol. 2004.

168. Sheets L, Ransom DG, Mellgren EM, Johnson SL, Schnapp BJ. Zebrafish melanophilin facilitates melanosome dispersion by regulating dynein. Curr Biol. 2007 Oct 23;17(20):1721–34.

169. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011 May 2;17(1):pp.10–2.

170. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007 Mar 8;446(7132):153–8.

171. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. Nature Publishing Group; 2009 Apr 9;458(7239):719–24.

172. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 14;500(7463):415–21.

173. Cowper-Sal lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, et al. Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature Genetics. 2012 Sep 23;44(11):1191–8.

174. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science. American Association for the Advancement of Science; 2013 Oct 4;342(6154):1235587–7.

175. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature. 2007 Aug 2;448(7153):595–9.

176. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of

proto-oncogenes by disruption of chromosome neighborhoods. Science. American Association for the Advancement of Science; 2016 Mar 3;64(2):aad9024–248.

177. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317–30.

178. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. BioMed Central; 2014 Oct 2;15(10):1.

179. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. Nature Publishing Group; 2014 Mar 1;46(3):310–5.

180. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research. Cold Spring Harbor Lab; 2012 Sep;22(9):1790–7.

181. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. Oxford University Press; 2015 May 15;31(10):1536–43.

182. Svetlichnyy D, Imrichova H, Fiers M, Kalender Atak Z, Aerts S. Identification of High-Impact cis-Regulatory Mutations Using Transcription Factor Specific Random Forest Models. Tanay A, editor. PLoS Comput Biol. 2015 Nov;11(11):e1004590.

183. Li J, Drubay D, Michiels S, Gautheret D. Mining the coding and non-coding genome for cancer drivers. Cancer Letters. 2015 Dec;369(2):307–15.

184. Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, et al. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. Nat Biotechnol. 2015 Apr;33(4):345–6.

185. Herz H-M, Hu D, Shilatifard A. Enhancer Malfunction in Cancer. Mol Cell. 2014 Mar;53(6):859–66.

186. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013 Oct 16;502(7471):333–9.

187. Araya CL, Cenik C, Reuter JA, Kiss G, Pande VS, Snyder MP, et al. Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nature Genetics. Nature Publishing Group; 2015 Dec 21;48(2):117–25.

188. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013 Oct;45(10):1113–20.

189. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. Nature Publishing Group; 2008 Oct 23;455(7216):1061–8.

190. Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, et al. Putative cis-regulatory drivers in colorectal cancer. Nature. Nature Publishing Group; 2014 Aug 7;512(7512):87–90.

191. Castro MAA, de Santiago I, Campbell TM, Vaughn C, Hickey TE, Ross E, et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nature Genetics. 2015 Nov 30;48(1):12–21.

192. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016 May 2;534(7605):47–54.

193. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nature Genetics. 2012 May 27;44(7):760–4.

194. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research. 2015 Jan;43(Database issue):D805–11.

195. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer. 2014 Nov 24;14(12):786–800.

196. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell. 2006 Apr;125(2):315–26.

197. Bapat SA, Jin V, Berry N, Balch C, Sharma N, Kurrey N, et al. Multivalent epigenetic marks confer microenvironment-responsive epigenetic plasticity to ovarian cancer cells. Epigenetics. 2010 Nov;5(8):716–29.

198. Voigt P, Tee W-W, Reinberg D. A double take on bivalent promoters. Genes & Development. Cold Spring Harbor Lab; 2013 Jun 15;27(12):1318–38.

199. Hahn MA, Li AX, Wu X, Yang R, Drew DA, Rosenberg DW, et al. Loss of the Polycomb Mark from Bivalent Promoters Leads to Activation of Cancer-Promoting Genes in Colorectal Tumors. Cancer Res. American Association for Cancer Research; 2014 Jul 1;74(13):3617–29.

200. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. Nat Rev Cancer. 2011 Sep 23;11(10):726–34.

201. Gal-Yam EN, Egger G, Iniguez L, Holster H, Einarsson S, Zhang X, et al. Frequent

switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. PNAS. National Acad Sciences; 2008 Sep 2;105(35):12979–84.

202. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546–58.

203. Nault JC, Mallet M, Pilati C, Calderaro J, Bioulac-Sage P, Laurent C, et al. High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. Nat Comms. 2013;4:2218.

204. Hu HM, Chen Y, Liu L, Zhang CG, Wang W, Gong K, et al. C1orf61 acts as a tumor activator in human hepatocellular carcinoma and is associated with tumorigenesis and metastasis. The FASEB Journal. 2013 Jan 2;27(1):163–73.

205. Reinke LM, Xu Y, Cheng C. Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition. Journal of Biological Chemistry. 2012 Oct 19;287(43):36435–42.

206. Yae T, Tsuchihashi K, Ishimoto T, Motohara T, Yoshikawa M, Yoshida GJ, et al. Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. Nat Comms. 2012;3:883.

207. Yao J, Caballero OL, Huang Y, Lin C, Rimoldi D, Behren A, et al. Altered Expression and Splicing of ESRP1 in Malignant Melanoma Correlates with Epithelial-Mesenchymal Status and Tumor-Associated Immune Cytolytic Activity. Cancer Immunol Res. 2016 Jun;4(6):552–61.

208. Lekva T, Berg JP, Fougner SL, Olstad OK, Ueland T, Bollerslev J. Gene expression profiling identifies ESRP1 as a potential regulator of epithelial mesenchymal transition in somatotroph adenomas from a large cohort of patients with acromegaly. J Clin Endocrinol Metab. 2012 Aug;97(8):E1506–14.

209. Ueda J, Matsuda Y, Yamahatsu K, Uchida E, Naito Z, Korc M, et al. Epithelial splicing regulatory protein 1 is a favorable prognostic factor in pancreatic cancer that attenuates pancreatic metastases. Oncogene. 2014 Sep 4;33(36):4485–95.

210. Leontieva OV, Ionov Y. RNA-binding motif protein 35A is a novel tumor suppressor for colorectal cancer. cc. 2009 Feb 1;8(3):490–7.

211. Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nature Genetics. 2015 Mar 30;47(5):505–11.

212. Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, Savich GL, et al. Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology. 2013 Nov 1;58(5):1693–702.

213. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. The Lancet

Oncology. Elsevier; 2011 Feb 1;12(2):175–80.

214. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nature Genetics. 2012 Feb;44(2):133–9.

215. Slattery ML, Lundgreen A, Wolff RK. MAP kinase genes and colon and rectal cancer. Carcinogenesis. Oxford University Press; 2012 Dec;33(12):2398–408.

216. Li T, Xie J, Shen C, Cheng D, Shi Y, Wu Z, et al. Amplification of Long Noncoding RNA ZFAS1 Promotes Metastasis in Hepatocellular Carcinoma. Cancer Res. American Association for Cancer Research; 2015 Aug 1;75(15):3181–91.

217. Wang W, Xing C. Upregulation of long noncoding RNA ZFAS1 predicts poor prognosis and prompts invasion and metastasis in colorectal cancer. Pathol Res Pract. 2016 Aug;212(8):690–5.

218. Wellner U, Schubert J, Burk UC, Schmalhofer O, Zhu F, Sonntag A, et al. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. Nat Cell Biol. 2009 Nov 22;11(12):1487–95.

219. Allerstorfer S, Sonvilla G, Fischer H, Spiegl-Kreinecker S, Gauglhofer C, Setinek U, et al. FGF5 as an oncogenic factor in human glioblastoma multiforme: autocrine and paracrine activities. Oncogene. Nature Publishing Group; 2008 Jul 10;27(30):4180–90.

220. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature. Nature Publishing Group; 2002 Aug 29;418(6901):934–4.

221. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nature Genetics. 2014 Nov 2;46(12):1267–73.

222. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. Oxford University Press; 2000 Jan 1;28(1):27–30.

223. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. Oxford University Press; 2009 Jan 15;25(2):288–9.

224. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. National Acad Sciences; 2005 Oct 25;102(43):15545–50.

225. Croce CM. Oncogenes and Cancer. N Engl J Med. 2008 Jan 31;358(5):502–11.

226. Kuwahara Y, Tanabe C, Ikeuchi T, Aoyagi K, Nishigaki M, Sakamoto H, et al.

Alternative mechanisms of gene amplification in human cancers. Genes Chromosom Cancer. 2004;41(2):125–32.

227. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. PNAS. National Acad Sciences; 2007 Dec 11;104(50):20007–12.

228. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan;17(1):98–110.

229. Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nature Genetics. Nature Publishing Group; 2016 Feb 1;48(2):176–82.

230. Huang P, Xiao A, Zhou M, Zhu Z, Lin S, Zhang B. Heritable gene targeting in zebrafish using customized TALENs. Nat Biotechnol. 2011 Aug;29(8):699–700.

231. Vinagre J, Almeida A, Pópulo H, Batista R, Lyra J, Pinto V, et al. Frequency of TERT promoter mutations in human cancers. Nat Comms. 2013 Jul 26;4.

232. Heidenreich B, Rachakonda PS, Hemminki K, Kumar R. TERT promoter mutations in cancer development. Current Opinion in Genetics & Development. 2014 Feb;24:30–7.

233. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific Reports. Nature Publishing Group; 2013 Oct 2;3:2650.

234. Guda K, Veigl ML, Varadan V, Nosrati A, Ravi L, Lutterbaugh J, et al. Novel recurrently mutated genes in African American colon cancers. PNAS. National Acad Sciences; 2015 Jan 27;112(4):1149–54.

235. Yeang C-H, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. The FASEB Journal. Federation of American Societies for Experimental Biology; 2008 Aug 1;22(8):2605–22.

236. Hasse A, Schulz WA. Enhancement of reporter gene de novo methylation by DNA fragments from the alpha-fetoprotein control region. Journal of Biological Chemistry. ASBMB; 1994;269(3):1821–6.

237. Zucman-Rossi J, Villanueva A, Nault JC, Llovet JM. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. Gastroenterology. 2015 Oct;149(5):1226–1239.e4.

238. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics. Nature Research; 2015 Feb

1;47(2):106–14.

239. Thorgeirsson SS, Grisham JW. Molecular pathogenesis of human hepatocellular carcinoma. Nature Genetics. 2002 Aug;31(4):339–46.

240. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research. Cold Spring Harbor Lab; 2012 Feb 1;22(2):398–406.

241. Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research. 2008 Jan;36(Database issue):D102–6.

242. Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research. Oxford University Press; 2003 Jan 1;31(1):374–8.

243. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. Oxford University Press; 2009 Jun 1;25(11):1422–3.

244. Manyam G, Birerdinc A, Baranova A. KPP: KEGG Pathway Painter. BMC Systems Biology. BioMed Central; 2015 Apr 15;9(2):1.

245. Lowdon RF, Jang HS, Wang T. Evolution of Epigenetic Regulation in Vertebrate Genomes. Trends Genet. 2016 May;32(5):269–83.

246. Montavon T, Duboule D. Chromatin organization and global regulation of Hox gene clusters. Philosophical Transactions of the Royal Society of London B: Biological Sciences. The Royal Society; 2013 Jun 19;368(1620):20120367–7.

247. Lukas J, Lukas C, Bartek J. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. Nat Cell Biol. 2011 Oct 3;13(10):1161–9.

248. Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, Daniel N, et al. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. Nature. 2011 Apr 6;472(7343):370–4.

249. Chow J, Heard E. X inactivation and the complexities of silencing a sex chromosome. Current Opinion in Cell Biology. 2009 Jun;21(3):359–66.

250. Wagner GP. The biological homology concept. Annual Review of Ecology and Systematics. 1989;20:51–69.

251. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010 Dec 24;330(6012):1775–87.

252. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010 Dec 24;330(6012):1787–97.

253. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. Nature Publishing Group; 2015 Feb 19;518(7539):317–30.

254. Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. 2012 Aug 13;13(8):418.

255. Kimura M. Evolutionary rate at the molecular level. Nature. 1968 Feb 17;217(5129):624–6.

256. Clarke SL, VanderMeer JE, Wenger AM, Schaar BT, Ahituv N, Bejerano G. Human Developmental Enhancers Conserved between Deuterostomes and Protostomes. Brosius J, editor. PLoS Genet. Public Library of Science; 2012 Aug 2;8(8):e1002852.

257. Cooper GM, Brown CD. Qualifying the relationship between sequence conservation and molecular function. Genome Research. Cold Spring Harbor Lab; 2008 Feb 1;18(2):201–5.

258. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975 Apr 11;188(4184):107–16.

259. Wray GA. The evolutionary significance of cis-regulatory mutations. Nature Publishing Group; 2007 Mar;8(3):206–16.

260. Chen X, Tompa M. Comparative assessment of methods for aligning multiple genome sequences. Nat Biotechnol. 2010 Jun;28(6):567–72.

261. McClintock B. Controlling Elements and the Gene. Cold Spring Harbor Symposia on Quantitative Biology. 1956 Jan 1;21(0):197–216.

262. Dogan N, Wu W, Morrissey CS, Chen K-B, Stonestrom A, Long M, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. Epigenetics & Chromatin. 2015;8:16.

263. Koide T, Ainscough J, Wijgerde M, Surani MA. Comparative analysis of Igf-2/H19 imprinted domain: identification of a highly conserved intergenic DNase I hypersensitive region. Genomics. 1994 Nov 1;24(1):1–8.

264. Follows GA, Tagoh H, Lefevre P, Morgan GJ, Bonifer C. Differential transcription factor occupancy but evolutionarily conserved chromatin features at the human and mouse M-CSF (CSF-1) receptor loci. Nucleic Acids Research. Oxford University Press; 2003 Oct 15;31(20):5805–16.

265. Arney KL, Bae E, Olsen C, Drewell RA. The human and mouse H19 imprinting control regions harbor an evolutionarily conserved silencer element that functions on transgenes in Drosophila. Dev Genes Evol. Springer-Verlag; 2006 Dec;216(12):811–9.

266. Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, et al. Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. Nature Genetics. Nature Publishing Group; 2008 Aug 1;40(8):971–6.

267. Killian JK, Byrd JC, Jirtle JV, Munday BL, Stoskopf MK, MacDonald RG, et al. M6P/IGF2R Imprinting Evolution in Mammals. Mol Cell. 2000 Apr;5(4):707–16.

268. Tagoh H, Himes R, Clarke D, Leenen PJM, Riggs AD, Hume D, et al. Transcription factor complex formation and chromatin fine structure alterations at the murine c-fms (CSF-1 receptor) locus during maturation of myeloid precursor cells. Genes & Development. Cold Spring Harbor Lab; 2002 Jul 1;16(13):1721–37.

269. Prendergast JGD, Chambers EV, Semple CAM. Sequence-level mechanisms of human epigenome evolution. Genome Biology and Evolution. Oxford University Press; 2014 Jul;6(7):1758–71.

270. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, Hvilsom C, et al. Dynamics of DNA Methylation in Recent Human and Great Ape Evolution. Gilad Y, editor. PLoS Genet. 2013 Sep 5;9(9):e1003763–12.

271. Hernando-Herraez I, Heyn H, Fernandez-Callejo M, Vidal E, Fernandez-Bellon H, Prado-Martinez J, et al. The interplay between DNA methylation and sequence divergence in recent human evolution. Nucleic Acids Research. Oxford University Press; 2015 Sep 30;43(17):8204–14.

272. Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, et al. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science. American Association for the Advancement of Science; 2015 Mar 6;347(6226):1155–9.

273. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, et al. Comparative Epigenomic Annotation of Regulatory DNA. Cell. 2012 Jun;149(6):1381–92.

274. Siepel A, Pollard KS, Haussler D. New Methods for Detecting Lineage-Specific Selection. In: Research in Computational Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. pp. 190–205. (Lecture Notes in Computer Science; vol. 3909).

275. Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013 Jul 3;154(1):185–96.

276. Aïssani B, Bernardi G. CpG islands: features and distribution in the genomes of

vertebrates. Gene. 1991 Oct 15;106(2):173–83.

277. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Ferguson-Smith A, editor. eLife. eLife Sciences Publications Limited; 2013;2:e00348.

278. Keshet I, Yisraeli J, Cedar H. Effect of regional DNA methylation on gene expression. Proceedings of the National Academy of Sciences. 1985 May;82(9):2560–4.

279. Feenastr A, Fewell J, Lueders K, Kuff E. In vitromethylation inhibits the promotor activity of a cloned intracisternal A-particle LTR. Nucleic Acids Research. 1986;14(10):4343–52.

280. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. Gibson G, editor. PLoS Genet. Public Library of Science; 2011 Feb 24;7(2):e1001316.

281. Elliott G, Hong C, Xing X, Zhou X, Li D, Coarfa C, et al. Intermediate DNA methylation is a conserved signature of genome regulation. Nat Comms. 2015;6:6363.

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

283. Keller TE, Yi SV. DNA methylation and evolution of duplicate genes. PNAS. National Acad Sciences; 2014 Apr 22;111(16):5932–7.

284. Macleod D, Charlton J, Mullins J, Bird AP. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes & Development. Cold Spring Harbor Lab; 1994 Oct 1;8(19):2282–92.

285. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D. Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions. Barsh GS, editor. PLoS Genet. Public Library of Science; 2013 Dec 19;9(12):e1003994.

286. Schübeler D. Function and information content of DNA methylation. Nature. Nature Publishing Group; 2015 Jan 15;517(7534):321–6.

287. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015 Jan 29;160(3):554–66.

288. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. American Association for the Advancement of Science; 2010 May 14;328(5980):916–9.

289. Chang AY-F, Liao B-Y. DNA methylation rebalances gene dosage after mammalian gene duplications. Mol Biol Evol. Oxford University Press; 2012 Jan;29(1):133–44.

290. Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am J Hum Genet. 2012 Sep 7;91(3):455–65.

291. Schroeder DI, Jayashankar K, Douglas KC, Thirkill TL, York D, Dickinson PJ, et al. Early Developmental and Evolutionary Origins of Gene Body DNA Methylation Patterns in Mammalian Placentas. Kelsey G, editor. PLoS Genet. Public Library of Science; 2015 Aug;11(8):e1005442.

292. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nature Genetics. 2009 Feb 1;41(3):376–81.

293. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009 Sep;16(9):990–5.

294. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of Alternative Splicing by Histone Modifications. Science. American Association for the Advancement of Science; 2010 Feb 19;327(5968):996–1000.

295. Li F, Ortega J, Gu L, Li G-M. Regulation of mismatch repair by histone code and posttranslational modifications in eukaryotic cells. DNA Repair. 2016 Feb;38:68–74.

296. Nag A, Vigneau S, Savova V, Zwemer LM, Gimelbrant AA. Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types. G3 (Bethesda). Genetics Society of America; 2015;5(8):1713–20.

297. Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. Predicting enhancer transcription and activity from chromatin modifications. Nucleic Acids Research. 2013 Sep 12.

298. Rubinstein M, de Souza FSJ. Evolution of transcriptional enhancers and animal diversity. Philosophical Transactions of the Royal Society of London B: Biological Sciences. The Royal Society; 2013 Dec 19;368(1632):20130017–7.

299. Wilczynski B, Liu Y-H, Yeo ZX, Furlong EEM. Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State. Segal E, editor. PLoS Comput Biol. Public Library of Science; 2012 Dec 1;8(12):e1002798.

300. Hsu C-H, Ovcharenko I. Effects of gene regulatory reprogramming on gene expression in human and mouse developing hearts. Philosophical Transactions of the Royal Society of London B: Biological Sciences. The Royal Society; 2013 Jun 19;368(1620):20120366–6.

301. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, et al. Comparative Epigenomic Analysis of Murine and Human Adipogenesis. Cell. 2010 Oct;143(1):156–69.

302. Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, et al. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell. 2015 Sep;163(1):68–83.

303. Tena JJ, González-Aguilera C, Fernández-Miñán A, Vázquez-Marín J, Parra-Acero H, Cross JW, et al. Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. Genome Research. Cold Spring Harbor Lab; 2014 Jul;24(7):1075–85.

304. Yokoyama KD, Zhang Y, Ma J. Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework. Zhong S, editor. PLoS Comput Biol. Public Library of Science; 2014 Aug 21;10(8):e1003771.

305. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, et al. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell. 2013 Aug;154(3):530–40.

306. Stamatoyannopoulos JA, Goodwin A, Joyce T, Lowrey CH. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. EMBO J. European Molecular Biology Organization; 1995 Jan 3;14(1):106–16.

307. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, et al. Principles of regulatory information conservation between mouse and human. Nature. 2014 Nov 19;515(7527):371–5.

308. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Research. Cold Spring Harbor Lab; 2014 Dec;24(12):1963–76.

309. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature Genetics. 2007 May 21;39(6):730–2.

310. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science. American Association for the Advancement of Science; 2010 May 20;328(5981):1036–40.

311. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. Akey JM, editor. PLoS Genet. 2012 Jun 28;8(6):e1002789.

312. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. Nature Publishing Group; 2014 Nov 20;515(7527):355–64.

313. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, et al. Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. Cell. 2005 Jan;120(2):169–81.

314. Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biology and Evolution. 2015 Feb;7(2):567–80.

315. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006 May 4;441(7089):87–90.

316. de Souza FSJ, Franchini LF, Rubinstein M. Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? Mol Biol Evol. 2013 May 9;30(6):1239–51.

317. Feschotte C. Transposable elements and the evolution of regulatory networks. 2008 May;9(5):397–405.

318. Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. Oxford University Press; 2002 Jul;19(7):1114–21.

319. Frith MC. Evolutionary turnover of mammalian transcription start sites. Genome Research. 2006 Jun 1;16(6):713–22.

320. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014 Nov 20;515(7527):365–70.

321. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Names A, et al. Spl elements protect a CpG island from de novo methylation. , Published online: 29 September 1994; | doi:101038/371435a0. Nature Publishing Group; 1994 Sep 29;371(6496):435–8.

322. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics. Nature Publishing Group; 2009 Jan 18;41(2):178–86.

323. Feinberg AP, Irizarry RA. Colloquium Paper: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. PNAS. 2010 Jan 26;107(suppl_1):1757–64.

324. Göke J, Jung M, Behrens S, Chavez L, O'Keeffe S, Timmermann B, et al. Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. Teichmann SA, editor. PLoS Comput Biol. Public Library of Science; 2011 Dec;7(12):e1002304.

325. Waddington CH. Canalization of development and the inheritance of acquired characters. Nature. 1942 Nov 14;150(3811):563–5.

326. Arendt D. Evolution of eyes and photoreceptor cell types. International Journal of Developmental Biology. 2003.

327. Wagner GP. The developmental genetics of homology. 2007 May 8;8(6):473–9.

328. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

329. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences. National Acad Sciences; 1992 Mar 1;89(5):1827–31.

330. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Research. 2005;33(18):5868–77.

331. Blackledge NP, Long HK, Zhou JC, Kriaucionis S, Patient R, Klose RJ. Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. Nucleic Acids Research. 2012 Feb;40(4):e32–2.

332. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007 May 18;129(4):823–37.

333. Gautsch JW, Wilson MC. Delayed de novo methylation in teratocarcinoma suggests additional tissue-specific mechanisms for controlling gene expression. Nature. 1983;301(5895):32.

334. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nature Genetics. 1999 Oct 20;23(3):314–8.

335. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences. National Acad Sciences; 2005 Jul 26;102(30):10604–9.

336. Kaminsky ZA, Tang T, Wang S-C, Ptak C, Oh GHT, Wong AHC, et al. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics. 2009 Jan 18;41(2):240–5.

337. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Research. Cold Spring Harbor Lab; 2010 Feb;20(2):170–9.

338. Paige SL, Thomas S, Stoick-Cooper CL, Wang H, Maves L, Sandstrom R, et al. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell. 2012 Sep 28;151(1):221–32.

339. The Origin of Malignant Tumors. Baltimore: The Williams & Wilkins Company; 1929.

340. Zhang B, Xing X, Li J, Lowdon RF, Zhou Y, Lin N, et al. Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers. BMC Genomics. 2014;15(1):868.

Appendix 1

Notes for Chapter 2

Note 1. Skin Cell Type-Specific DMR Calling Strategy

The specific skin cell type pairwise comparisons processed by M&M are as shown in **Figure 2.S1a**. Each of the 3 skin cell type datasets from 3 different individuals is compared against every other skin cell type dataset, for a total of 36 pairwise comparisons. Pairwise comparisons between two different cell types are inter-cell type comparisons (27 total, gray boxes), while M&M comparisons between two of the same cell type datasets are intra-cell type comparisons (9 comparisons, 3 per cell type, blue boxes).

To maximize the specificity of our DMR prediction, we took advantage of the presence of three biological replicates for each cell type, and required that the same DMR call was reproduced in all analogous pair-wise comparisons. Therefore, to call cell type-specific DMRs, we took the intersection of all comparisons involving the three replicates of a given cell type, and required that a 500bp window be called significantly differentially methylated (in the same direction) by our M&M statistic in each of 18 pairwise comparisons. Our intersection strategy is illustrated in **Figure 2.S1b**.

Note 2. M&M Command Line and Output Description

R scripts used to generate pairwise comparisons using the methylMnM R package (http://epigenome.wustl.edu/MnM/). The **compare.pv.R** script contains the functions to perform the actual pairwise comparison that generates p-values for each 500bp window (**MnM.test()** function). The **qv.DMR.R** script calculates q-values for every window (**MnM.qvalue()**) and selects significant windows based on a user-given q-value threshold (**MnM.selectDMR()**).

compare.pv.R

```
library(methylMnM)

cpgbin <- 'num500_cpgbin.bed'

mrecpgbin <- 'num500_Five_mre_cpg.bed'

medip.list = read.table('skin_medip.list')

mre.list = read.table('skin_mre.list')

c_s <- NULL


for  (i in 1:length(medip.list[,1])) {

        medipfile1 <- paste('num500_',medip.list[i,1],sep='')

        mrefile1 <- paste('num500_',mre.list[i,1],sep='')

        name <- paste(medip.list[i,1])

        first <-
paste(strsplit(name,"_")[[1]][1],strsplit(name,"_")[[1]][2],sep="")


        for (j in (i+1):length(medip.list[,1])) {

                medipfile2 <- paste('num500_',medip.list[j,1],sep='')

                mrefile2 <- paste('num500_',mre.list[j,1],sep='')

                name <- paste(medip.list[j,1])

                second <-
paste(strsplit(name,"_")[[1]][1],strsplit(name,"_")[[1]][2],sep="")

                dataset <- c(medipfile1, medipfile2, mrefile1, mrefile2)
```

```
                w_f <- paste('pv_',first,"_",second,".bed",sep="")

                r_f <- paste('pv_',first,"_",second,".report",sep="")

                MnM.test(file.dataset=dataset, chrstring=c_s,
file.cpgbin=cpgbin, file.mrecpgbin=mrecpgbin, writefile=w_f,
reportfile=r_f, mreratio=3/7, method='XXYY', psd=2, mkadded=1, a=1e-
20, cut=100, top=500)

        }

}
```

qv.DMR.R

```
library(methylMnM)

qv.list = read.table('skin_qv_files.list')


for  (i in 1:length(qv.list[,1])) {

        name <- paste(qv.list[i,1])

        qval_f <- paste('qv_',name,sep='')

        r_f <- paste('qv_',strsplit(name,".bed")[[1]][1],".report",sep="")

        MnM.qvalue(pval_f, writefile=qval_f, reportfile=r_f)

        frames <- read.table(qval_f, header=TRUE, sep="\t", as.is=TRUE)

        DMR <- MnM.selectDMR(frames=frames, up=2, down=1/2,
q.value=1e-5, cutoff="q-value", quant=0.9)

        fname <- paste(qval_f,sep="")

        sname <-strsplit(fname,"pv")[[1]][2]

        writeDMRfile <- paste('DMR_q1e-5',sname,sep="")

        write.table(DMR, writeDMRfile, sep="\t", quote=FALSE,
row.names=FALSE)

}
```

The output of M&M pairwise comparisons were p-value and q-value measurements for the likelihood that the methylation levels of the two samples were different for each 500bp window

across the genome. Note that q-value is the false discovery rate analogue of the p-value. The genome-wide false discovery rate (FDR) was controlled using the previously described Group Benjamini-Hochberg method. We then chose a q-value cutoff to call differentially methylated regions. All of our analyses used a q-value cutoff of 1e-5.

Note 3. Estimation of M&M and Cell Type-Specific DMR FDR

To estimate the false discovery rate of DMRs called by M&M, we chose a pairwise comparison as a test case: Fibroblast skin 03 vs Keratinocyte skin 03. These results were compared to the M&M results from a within-cell type comparison: Fibroblast skin 02 vs Fibroblast skin 03, which are biological replicates (i.e. the same cell type from two different newborn males). For these pairwise comparisons, we examined the number of DMRs called at varying q-value cutoffs. As seen in **Figure 2.S2**, the number of DMRs in both cases decreased with decreasing q-value cutoff. As expected, the numbers of DMRs found between biological replicates is very small. Thus, our pairwise DMR false discovery rate is very low (**Table 2.S1**). We used M&M q-values of 1e-5 throughout, which by this analysis had a FDR of 0.044. FDR calculations using other within-cell type comparisons yielded similar results.

To assay the false discovery rate of our skin cell type-specific DMR calling strategy, we performed a permutation experiment to empirically estimate this value. In this experiment, we randomly shuffled our datasets by labeling them as three "pseudo" cell types (A, B, and C) with three replicates each (01, 02, and 03). Because we have already performed all possible pair-wise comparisons using the M&M algorithm, we called pseudo-cell type specific-DMRs by the same criteria as in **Note 1.1** (above), i.e. that a window must be called differentially methylated 18/18 times to be a DMR in any pseudo-cell type at a q-value cutoff of 1e-5. The strategy is illustrated in **Figure 2.S3**. We repeated this process of shuffling, assigning pseudo-cell type names, and finding DMRs 10 times. Each time the analysis of the pseudo-cell types returned zero windows called as pseudo-cell type-specific DMRs. Thus, our cell type-specific DMRs are very far from the random expectation for these data.

Note 4. Analysis of CpG Islands in Cell Type-Specific DMRs

It is known that approximately 70% of all gene promoters are associated with a CpG island (CGI). We defined a CGI promoter as any promoter that has ≥ 0.05% of a given CGI contained in it and found 16638 RefSeq gene promoters (or 63.2%) were CGI promoters. Then we counted the numbers of CGI promoters and non-CGI promoters in each DMR class and tested the null hypothesis that the percentage of promoters that contain CGIs for each DMR class is similar to the CGI promoter distribution found across the genome. We found that across our DMR sets, the numbers of CGI promoters in DMRs are significantly depleted relative to their genome-wide distribution, while non-CGI promoters are significantly enriched (**Table 2.S2**). In general, the majority of DMRs at promoters were within non-CGI promoters, which is consistent with the concept that non-CGI promoters are involved in tissue and cell type specificity.

Note 5. Skin Tissue-Specific DMR Calling Strategy

We sought to identify the unique DNA methylation signature that the skin environment might contribute to its resident cell types. Therefore, we asked what shared regions of the skin fibroblast, keratinocyte, and melanocyte methylomes were differentially methylated compared to cell types of other tissues. To do this, we compared skin cell type methylomes to those of non-skin cell types and tissues (including brain tissue and breast and blood cell types) to identify DMRs in a pairwise manner. 28,776 total DMRs were identified in these pair-wise comparisons. Compared to the non-skin samples, keratinocytes, fibroblasts, and melanocytes each possessed 623, 763, and 402 consensus DMRs respectively. We then took the intersection of these three DMR sets to identify the shared differences between skin cell types and cell types residing in a different tissue environment (i.e. the same methylation status in all skin cell types and the opposite methylation status in all non-skin cell types). The result was, surprisingly, a very small set of only 8 regions. To be clear, we do expect much of the methylome for the three skin cell types is similar, but the shared methylome signature that is unique to the skin is very small.

Identification of skin tissue-specific DMRs follows the exact same logic as that of cell type-specific DMRs (for which FDR and reproducibility are documented above in **Note 1.3**). Both M&M and our DMR identification strategy are designed to optimize specificity. We use the same M&M q-value threshold for our tissue-specific analysis as for the cell type-specific analysis.

Note 6. Supplementary Methods for Chapter 2

RNA isolation

Total RNA was extracted from cells using Trizol reagent (Life Technologies) following the manufacturer's instructions.

RNA-seq

Standard operating procedures for RNA-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. RNA-seq library construction involves the following protocols in order: 1) Purification of polyA+ mRNA and mRNA(-) Flow-Through Total RNA using MultiMACS 96 separation unit, 2) Strand specific 96 Well cDNA Synthesis, and 3) Strand specific 96-well library construction for Illumina sequencing. Briefly, polyA+ RNA was purified using the MACS mRNA isolation kit (Miltenyi Biotec) from 2-10 ug of total RNA with a RIN>=7 (Agilent Bioanalyzer) as per the manufacturer's instructions. The process included on-column DNase I treatment (Invitrogen). Double stranded cDNA was synthesized from the purified polyA+ RNA using the Superscript II Double-Stranded cDNA Synthesis kit (Invitrogen) and 200 ng of random hexamers. After first strand synthesis, dNTPs were removed using 2 volumes of AMPure XP beads (Beckman Genomics). GeneAmp 12.5mM dNTPs blend (Invitrogen) was used in the second strand synthesis mixture in the presence of 2 ug of Actinomycin D. Double stranded cDNA was purified using 2 volumes of Ampure XP beads, fragmented using Covaris E series shearing (20% duty cycle, Intensity 5, 55 seconds), and used for paired-end sequencing library preparation (Illumina). Prior to library amplification, uridine digestion was performed at 37 degrees Celsius for 30 minutes, followed by a 10 minute incubation at 95 degrees Celsius in Qiagen Elution buffer (10mM Tris-Cl, pH 8.5) with 5 units of Uracil-N-Glycosylase (UNG: AmpErase). The

resulting single stranded sequencing library was amplified by PCR (10-13 cycles) to add Illumina P5 and P7 sequences for cluster generation. PCR products were purified on Qiaquick MinElute columns (Qiagen) and assessed and quantified using an Agilent DNA 1000 series II assay and Qubit fluorometer (Invitrogen) respectively. Libraries were sequenced using paired-end 76 nt sequencing chemistry on a cBot and Illumina GAiix or HiSeq2000 following manufacturer's protocols (Illumina).

RNA-seq pair-end reads were aligned to a transcriptome reference consisting of the reference genome extended by the annotated exon-exon junctions17. To generate a transcriptome reference, we used the JAGuaR v 1.7.6 pipeline (http://www.bcgsc.ca/platform/bioinfo/software/jaguar) which is specifically developed to allow for a single read to span multiple exons. Reads aligned to a custom transcriptome reference (build from NCBI GRCh37-lite reference and Ensembl v65 (GenCode v10) annotations) are then "repositioned" onto genomic coordinates, transforming reads spanning exon-exon junctions into large-gapped alignment. Using repositioned reads, we generated genome wide coverage profiles (wiggled files) using BMA2WIG java program for further analysis and visualization in genome browsers. To generate profiles we included pairs that are marked as duplicated as well as pairs mapped in multiple genomic locations.

A custom RNA-seq QC and analysis pipeline was applied to the generated profiles and a number of QC metrics were calculated to assess the quality of RNA-seq libraries such as intron-exon ratio, intergenic reads fraction, strand specificity, 3'-5' bias, GC bias, and RPKM discovery rate. To quantify exon and gene expression we calculated modified RPKM metrics[4]. For the normalization factor in RPKM calculations, we used the total number of reads aligned into coding exons and excluded reads from the mitochondrial genome, that fall within genes encoding

ribosomal proteins, or that fall into the top 0.5% expressed exons. RPKM for a gene was calculated using total number of reads aligned into all its merged exons normalized by total exonic length. All mRNA-seq analyses used a pseudocount of 1.

miRNA-seq

Standard operating procedures for miRNA-seq library construction are available at http://www.roadmapepigenomics.org/protocols/type/experimental/. miRNA-seq library construction involves the following protocols in order: 1) purification of polyA+ mRNA and mRNA(-) Flow-Through Total RNA using MultiMACS 96 separation unit, 2) strand specific 96 Well cDNA Synthesis, and 3) strand specific 96-well library construction for Illumina sequencing. A more detailed description of miRNA-seq library construction and data processing in Gascard, P. et al. (submitted REMC companion paper).

Appendix 2

Supplementary Data for Chapter 2

Data 1. Samples and Datasets.

GEO accessions for data generated in this project.

Samples and Datasets											
								Histone Modification ChIP-seq			
Sample	**Donor Id**	**MeDIP-seq**	**MRE-seq**	**mRNA-seq**	**miRNA-seq**	**WGBS**	**Dnase I-seq**	**H3K4me3**	**H3K4me1**	**H3K27ac**	**Input**
Keratinocyte	Skin 01	GSM707022	GSM707018	GSM751278	GSM817253			GSM669589	GSM669591		GSM817242
Keratinocyte	Skin 02	GSM941726	GSM941723	GSM941745	GSM1127113		GSM817196	GSM941735	GSM941736		GSM941742
Keratinocyte	Skin 03	GSM958180	GSM958169	GSM958177	GSM1127111	GSM1127056, GSM1127058		GSM958155	GSM958161	GSM958156	GSM958167
Fibroblast	Skin 01	GSM707021	GSM707017	GSM751277	GSM817252			GSM817235	GSM817234		GSM817246
Fibroblast	Skin 02	GSM941725	GSM941722	GSM941744				GSM941718	GSM941717		GSM817247
Fibroblast	Skin 03	GSM958182	GSM958171	GSM958178	GSM1127116			GSM958158	GSM958164	GSM958163	GSM958168
Melanocyte	Skin 01	GSM707020	GSM707016	GSM751276	GSM817251			GSM941719	GSM941728		GSM941740
Melanocyte	Skin 02	GSM941727	GSM941724	GSM941743				GSM941731	GSM941730		GSM941741
Melanocyte	Skin 03	GSM958181	GSM958170	GSM958174				GSM958151	GSM958152	GSM958157	GSM958166
Breast Lumimal epithelia	RM066	GSM613856	GSM613833								
Breast Lumimal epithelia	RM070	GSM613843	GSM613818								
Breast Lumimal epithelia	RM071	GSM1517154	GSM613826								
Breast Lumimal epithelia	RM080			GSM669620					GSM669595		GSM959124
Breast Myoepithelia	RM066	GSM613857	GSM613834					GSM613869	GSM613870		GSM613891
Breast Myoepithelia	RM070	GSM613846	GSM613821								
Breast Myoepithelia	RM071	GSM1517153	GSM613908								

Breast Myoepithelia	RM080			GSM669621				GSM693277	GSM613885		GSM613897
Fetal Brain Tissue	HuFNSC01	GSM669614	GSM669604					GSM806943	GSM806942		GSM806948
Fetal Brain Tissue	HuFNSC02	GSM669615	GSM669605					GSM806935	GSM806934		GSM817243
Fetal Brain Germinal Matrix	HuFGM02					GSM941747					
CD4 Naïve	TC003	GSM543025	GSM543011								
CD4 Naïve	TC007	GSM613913	GSM613901								
CD4 Memory	TC003	GSM613862	GSM613842								
CD4 Memory	TC007	GSM613914	GSM613903								
CD8 Naïve	TC003	GSM543027	GSM543013								
CD8 Naïve	TC007	GSM613917	GSM613905								
CD8 Naïve	TC001							GSM613811	GSM613814		GSM613816
PBMC	TC015							GSM1127126	GSM1127143		GSM1127151
H1 ESC				GSM438361		GSM432686, GSM432685					
Ectoderm-differentiated ESC				GSM1112844		GSM1112820, GSM1112821, GSM1112849					

Data 2. Library Statistics.

Library statistics for MeDIP-seq, MRE-seq, and ChIP-seq data generated in this study.

Data 2.1 MeDIP-seq library statistics.

Sample	Donor	GEO Accession	Mapped Reads	High quality, unique reads	Used reads rate
Fibroblast	Skin01	GSM707021	103695039	60652204	58.49%
Keratinocyte	Skin01	GSM707022	123896883	69591154	56.17%
Melanocyte	Skin01	GSM707020	98450409	56665014	57.56%
Fibroblast	Skin02	GSM941725	239981572	123090682	51.29%
Keratinocyte	Skin02	GSM941726	236055826	114217107	48.39%
Melanocyte	Skin02	GSM941727	224042936	116271310	51.90%
Fibroblast	Skin03	GSM958182	220904929	125936493	57.01%
Keratinocyte	Skin03	GSM958180	238516753	128371876	53.82%
Melanocyte	Skin03	GSM958181	202789509	115465059	56.94%

Data 2.2 MRE-seq library statistics.

Sample	Donor	GEO Accession	Mapped Reads	High quality, MRE filtered reads	Used reads rate	Sampled CpG sites
Fibroblast	Skin01	GSM707017	67639307	44762753	66.18%	2185568
Keratinocyte	Skin01	GSM707018	70299709	48172559	68.52%	1878071
Melanocyte	Skin01	GSM707016	65093003	43554308	66.91%	1972108
Fibroblast	Skin02	GSM941722	200635751	107286642	53.47%	2580687
Keratinocyte	Skin02	GSM941723	199621155	126235449	63.24%	2188999
Melanocyte	Skin02	GSM941724	260071399	107200110	41.22%	1382622
Fibroblast	Skin03	GSM958171	66391182	27560887	41.51%	1660293
Keratinocyte	Skin03	GSM958169	67559889	20181816	29.87%	1269026
Melanocyte	Skin03	GSM958170	64504135	24268667	37.62%	1506660

Data 2.3 ChIP-seq library statistics.

H3K4me1

Sample	Donor	GEO accession	mapped reads	unique reads
Keratinocyte	Skin01	GSM669591	40128416	27543090
Keratinocyte	Skin02	GSM941736	44725534	34462102
Keratinocyte	Skin03	GSM958161	106681834	83793791
Fibroblast	Skin01	GSM817234	52729214	32248772
Fibroblast	Skin02	GSM941717	103855713	78219051
Fibroblast	Skin03	GSM958164	102832408	74623773
Melanocyte	Skin01	GSM941728	41428106	23826872
Melanocyte	Skin02	GSM941730	54054254	45881462
Melanocyte	Skin03	GSM958152	72311326	58158457
Breast Lumimal epithelia	RM080	GSM669595	36920574	7217943
Breast Myoepithelia	RM066	GSM613870	27419931	19460712
Breast Myoepithelia	RM080	GSM613885	39884011	24760406
Fetal Brain Tissue	HuFNSC01	GSM806942	33409083	22225887
Fetal Brain Tissue	HuFNSC02	GSM806934	34528219	29546991
CD8 Naïve	TC001	GSM613814	32253632	21499085
PBMC	TC015	GSM1127143	27420488	20697125

H3K4me3

Sample	Donor	GEO accession	mapped reads	unique reads
Keratinocyte	Skin01	GSM669589	26301506	17040581
Keratinocyte	Skin02	GSM941735	111420479	26005875
Keratinocyte	Skin03	GSM958155	89072108	37175759
Fibroblast	Skin01	GSM817235	35872336	29602160
Fibroblast	Skin02	GSM941718	81594474	54134847
Fibroblast	Skin03	GSM958158	121459804	68806151
Melanocyte	Skin01	GSM941719	50258524	23705505
Melanocyte	Skin02	GSM941731	61379885	29826357
Melanocyte	Skin03	GSM958151	85101010	35039322
Breast Myoepithelia	RM066	GSM613869	30075469	7313255
Breast Myoepithelia	RM080	GSM693277	36627300	9743144
Fetal Brain Tissue	HuFNSC01	GSM806943	34776884	22443075
Fetal Brain Tissue	HuFNSC02	GSM806935	33348651	25099253
CD8 Naïve	TC001	GSM613811	30715940	12682909
PBMC	TC015	GSM1127126	32660152	18423440

H3K27ac

Sample	Donor	GEO accession	mapped reads	unique reads
Keratinocyte	Skin03	GSM958156	90088926	73780386
Fibroblast	Skin03	GSM958163	85996616	70420493
Melanocyte	Skin03	GSM958157	60175730	41969129

Input

Sample	Donor	GEO accession	mapped reads	unique reads
Keratinocyte	Skin01	GSM817242	43100255	24274329
Keratinocyte	Skin02	GSM941742	23721661	17575551
Keratinocyte	Skin03	GSM958167	63756790	44519558
Fibroblast	Skin01	GSM817246	72737986	50480503
Fibroblast	Skin02	GSM817247	70434054	51294727
Fibroblast	Skin03	GSM958168	46473822	31420011
Melanocyte	Skin01	GSM941740	30804003	20723138
Melanocyte	Skin02	GSM941741	35981661	26012467
Melanocyte	Skin03	GSM958166	72601148	60068291
Breast Lumimal epithelia	RM080	GSM959124	14417625	8605267
Breast Myoepithelia	RM066	GSM613891	27014513	20218185
Breast Myoepithelia	RM080	GSM613897	32030813	22049938
Fetal Brain Tissue	HuFNSC01	GSM806948	33227925	21625397
Fetal Brain Tissue	HuFNSC02	GSM817243	16558529	10508213
CD8 Naïve	TC001	GSM613816	36016608	25937704
PBMC	TC015	GSM1127151	29092457	24043241

Data 3. Gene Ontology Enrichment Results I

GO term enrichment from GREAT for skin cell type-specific hypomethylated DMRs.

Data 3.1 Fibroblast hypomethylated DMR GREAT enrichment.

Fibroblast hypomethylated DMRs	
Term Name	**Binomial P-Value**
negative regulation of glycolysis	3.20E-56
osteoblast differentiation	2.67E-49
negative regulation of osteoblast differentiation	1.39E-48
extracellular matrix organization	9.05E-45
ossification	6.73E-44
chordate embryonic development	2.87E-42
negative regulation of cellular carbohydrate metabolic process	4.02E-42
embryonic cranial skeleton morphogenesis	2.03E-41
embryonic skeletal system development	2.25E-41
embryo development ending in birth or egg hatching	1.91E-40
skeletal system development	2.30E-37
osteoblast development	1.25E-35
skeletal system morphogenesis	2.67E-35
negative regulation of transcription from RNA polymerase II promoter	6.28E-35
regulation of generation of precursor metabolites and energy	3.27E-32
tooth mineralization	3.31E-32
response to retinoic acid	7.28E-29
intramembranous ossification	4.54E-28
response to vitamin	8.26E-27
regulation of osteoblast differentiation	3.48E-26
embryonic organ development	6.36E-26
cellular response to external stimulus	9.91E-26
negative regulation of muscle cell differentiation	2.15E-25
cellular response to nutrient levels	3.55E-25
response to vitamin A	6.63E-25
cellular response to extracellular stimulus	9.98E-25
negative regulation of cell differentiation	1.41E-24
cellular response to retinoic acid	2.19E-24
endochondral ossification	2.34E-24
cellular response to vitamin A	3.06E-24
regulation of ossification	4.64E-24
regulation of cardiac muscle contraction	7.43E-23
embryonic organ morphogenesis	8.51E-23

positive regulation of cell morphogenesis involved in differentiation	2.32E-22
response to extracellular stimulus	3.32E-22
response to nutrient levels	6.87E-22
regulation of striated muscle contraction	8.04E-22
cellular response to vitamin	1.04E-21
regulation of glucose metabolic process	2.62E-21
response to nutrient	3.49E-21

Data 3.2 Keratinocyte hypomethylated DMR GREAT enrichment.

Keratinocyte hypomethylated DMRs	
Term Name	**Binomial P-Value**
epidermis development	6.32E-70
epithelial cell differentiation	1.15E-43
skin development	9.32E-42
response to retinoic acid	1.03E-39
response to vitamin	4.29E-39
induction of apoptosis by extracellular signals	4.47E-37
hair follicle development	8.76E-37
hair cycle	2.40E-36
mammary gland epithelium development	3.16E-35
response to vitamin A	1.76E-32
lens fiber cell differentiation	2.22E-32
cell-substrate junction assembly	8.68E-30
inner ear development	1.38E-29
negative regulation of sequence-specific DNA binding transcription factor activity	9.75E-29
stem cell development	2.11E-28
negative regulation of neurogenesis	1.75E-26
placenta development	2.11E-26
stem cell differentiation	6.28E-26
stem cell maintenance	5.33E-25
positive regulation of Rho GTPase activity	7.06E-25
regulation of lipid biosynthetic process	2.04E-24
negative regulation of cell development	2.74E-24
morphogenesis of a polarized epithelium	3.73E-24
cellular response to extracellular stimulus	1.59E-23
hemidesmosome assembly	2.09E-23
negative regulation of osteoblast differentiation	1.16E-20
neuronal stem cell maintenance	2.46E-20
negative regulation of phosphate metabolic process	9.66E-20

regulation of protein kinase B signaling cascade	2.06E-19
regulation of morphogenesis of a branching structure	4.61E-19
negative regulation of glial cell proliferation	3.52E-18
somatic stem cell maintenance	6.19E-18
osteoblast development	1.24E-17
response to vitamin D	1.33E-17
negative regulation of gliogenesis	1.67E-17
digestive tract morphogenesis	1.18E-16
exocrine system development	1.34E-16
negative regulation of phosphorylation	4.21E-16
mammary gland duct morphogenesis	4.99E-16
establishment of tissue polarity	3.97E-15

Data 3.3 Melanocyte hypomethylated DMR GREAT enrichment.

Melanocyte hypomethylated DMRs	
Term Name	**Binomial P-Value**
regulation of cardioblast proliferation	1.73E-16
negative regulation of muscle cell differentiation	7.41E-15
negative regulation of developmental process	1.30E-14
negative regulation of cell differentiation	2.01E-13
regionalization	1.99E-12
regulation of muscle cell differentiation	7.17E-09
segmentation	8.16E-09
negative regulation of osteoblast differentiation	2.03E-08
mammary gland epithelium development	3.31E-08
in utero embryonic development	1.23E-07
positive regulation of Ras GTPase activity	1.89E-07
stem cell maintenance	5.92E-07
stem cell differentiation	1.09E-06
stem cell development	2.39E-06
pigmentation	2.43E-06
cell proliferation in forebrain	2.53E-06
positive regulation of neuroblast proliferation	4.20E-06
anterior/posterior pattern specification	4.34E-06
regulation of neural precursor cell proliferation	6.48E-06
positive regulation of neural precursor cell proliferation	1.13E-05
mesoderm morphogenesis	1.14E-05
lymph vessel development	1.33E-05
myelination	4.42E-05

embryonic limb morphogenesis	4.79E-05
cellular response to lipid	4.97E-05
cardiac chamber development	5.80E-05
axon ensheathment	6.04E-05
negative regulation of cell development	9.52E-05
cardiac ventricle development	9.98E-05
cardiac chamber morphogenesis	1.26E-04
regulation of muscle organ development	1.77E-04
regulation of action potential in neuron	1.79E-04
negative regulation of gliogenesis	2.01E-04
negative regulation of neurogenesis	2.28E-04
organ growth	2.48E-04
somatic stem cell maintenance	2.54E-04
hindbrain development	2.67E-04
regulation of astrocyte differentiation	3.69E-04
proximal/distal pattern formation	4.78E-04
amine transport	5.22E-04

Data 4. Gene Ontology Enrichment Results II

GO term enrichment from GREAT for skin cell type- or tissue-specific histone modification ChIP-seq peaks.

Data 4.1 Fibroblast H3K4me1 + H3K27ac peaks GREAT enrichment.

GO Biological Process	Binomial P-value
platelet-derived growth factor binding	5.83E-22
SMAD binding	6.34E-12
collagen binding	1.79E-08
negative regulation of transforming growth factor beta receptor signaling pathway by extracellular sequestering of TGFbeta	2.94E-20
extracellular matrix organization	1.29E-19
response to oxygen levels	9.80E-18
protein heterotrimerization	1.93E-17
extracellular matrix part	2.56E-43
proteinaceous extracellular matrix	7.69E-38
extracellular matrix	1.35E-37
collagen	7.39E-30
extracellular region part	2.36E-29
fibrillar collagen	4.39E-21
basement membrane	5.19E-17
collagen type VI	5.78E-17

Data 4.2 Keratinocyte H3K4me1 + H3K27ac peaks GREAT enrichment

GO Biological Process	Binomial P-value
keratinocyte differentiation	2.41E-13
keratinization	5.18E-13
epidermis development	1.45E-12
epidermal cell differentiation	2.80E-12
epithelial cell differentiation	7.82E-10
positive regulation of MAPKKK cascade	3.45E-07
branch elongation of an epithelium	7.40E-07
axis elongation	2.02E-06
ear development	4.18E-06
inner ear development	4.98E-06
ear morphogenesis	6.88E-06

limb morphogenesis	8.78E-06
uterus development	1.14E-05
metanephric mesenchyme morphogenesis	2.23E-05
embryonic limb morphogenesis	2.28E-05
response to transforming growth factor beta stimulus	2.62E-05
limb development	4.39E-05
cellular response to transforming growth factor beta stimulus	1.14E-04
regulation of fibroblast growth factor receptor signaling pathway	2.57E-04
embryonic forelimb morphogenesis	3.85E-04
cornified envelope	2.11E-08
desmosome	9.63E-05

Data 4.3 Melanocyte H3K4me1 + H3K27ac peaks GREAT enrichment

GO Biological Process	Binomial P-value
developmental pigmentation	1.20E-27
Ocular albinism	7.77E-29
Reduced iris pigmentation	2.50E-24
Abnormality of the iris	1.39E-15

Data 4.4 Fibroblast H3K4me3 peaks GREAT enrichment

GO Molecular Function	Binomial P-value
platelet-derived growth factor binding	1.05E-51
growth factor binding	5.68E-26
integrin binding	4.80E-08
SMAD binding	8.85E-08
insulin receptor binding	2.84E-07
collagen binding	4.98E-07

GO Biological Process	
extracellular matrix organization	1.91E-41
collagen biosynthetic process	3.41E-27
collagen fibril organization	1.37E-24
cellular response to acid	2.03E-24
collagen metabolic process	1.76E-22
multicellular organismal macromolecule metabolic process	2.47E-22

Human Phenotype	
Joint laxity	4.22E-42
Soft skin	1.22E-34
Joint hypermobility	1.48E-34
Blue sclerae	2.27E-29
Abnormality of the sclera	1.47E-26
Molluscoid pseudotumors	1.49E-26
Mitral valve prolapse	2.51E-26

GO Cellular Component	
extracellular matrix part	1.29E-30
fibrillar collagen	5.23E-27
collagen	6.84E-27
actin cytoskeleton	3.14E-20
basement membrane	1.16E-19
actomyosin	1.35E-16
focal adhesion	6.39E-12
actin filament bundle	2.80E-11
stress fiber	2.98E-11
cell-substrate adherens junction	5.65E-11
cell-substrate junction	5.99E-10
extrinsic to internal side of plasma membrane	2.19E-05

Data 4.5 Keratinocyte H3K4me3 peaks GREAT enrichment

GO Molecular Function	Binomial P-value
Ras guanyl-nucleotide exchange factor activity	1.70E-06

GO Biological Process	
epidermis development	3.53E-13
hair follicle development	2.12E-08
hair cycle	2.48E-08
epithelial cell differentiation	7.18E-08
skin development	8.21E-08
establishment of planar polarity	9.02E-08
establishment of tissue polarity	1.17E-07

morphogenesis of a polarized epithelium	7.42E-07
lateral sprouting from an epithelium	9.10E-07
negative regulation of epidermis development	1.41E-06
response to ionizing radiation	2.73E-06
prostate glandular acinus development	4.18E-06
regulation of Rho protein signal transduction	4.89E-06
response to radiation	4.90E-06
ectoderm development	1.50E-05
response to light stimulus	5.49E-05
positive regulation of DNA replication	9.38E-05
digestive tract morphogenesis	1.20E-04
positive regulation of DNA metabolic process	1.63E-04
prostate gland epithelium morphogenesis	1.96E-04

GO Cellular Component	
cell-cell junction	4.23E-10
desmosome	4.59E-07
anchoring junction	5.43E-06
gap junction	8.24E-05
cell-cell adherens junction	2.58E-04

Data 4.6 Melanocyte H3K4me3 peaks GREAT enrichment

GO Biological Process	**Binomial P-value**
developmental pigmentation	1.19E-09
melanin biosynthetic process	1.93E-09
melanin metabolic process	7.53E-09
regulation of Rap GTPase activity	7.79E-06
cell proliferation in forebrain	5.45E-05
protein autophosphorylation	4.06E-04

GO Cellular Component	
melanosome	1.12E-18
melanosome membrane	4.88E-09

Human Phenotype	
Abnormality of hair pigmentation	4.45E-14

Hypopigmentation of the skin	5.86E-14
Reduced iris pigmentation	1.04E-13
Hypopigmentation of hair	1.91E-13
Abnormality of the iris	1.05E-11
Abnormality of the uvea	2.82E-11
Albinism	4.32E-10
Abnormality of the musculature of the limbs	6.67E-10
Ocular albinism	1.37E-09
Generalized hypopigmentation	3.64E-09

Data 4.7 Skin cell type shared H3K4me1 peaks GREAT enrichment

GO Biological Process	**Binomial P-value**
regulation of cell adhesion	3.83E-09
positive regulation of cell adhesion	2.33E-08
cytoplasmic mRNA processing body assembly	5.37E-05
positive regulation of branching involved in ureteric bud morphogenesis	1.07E-04
positive regulation of cell-substrate adhesion	1.15E-04
metanephric renal vesicle morphogenesis	2.16E-04
renal vesicle development	2.81E-04
renal vesicle morphogenesis	3.18E-04
metanephric nephron morphogenesis	3.25E-04
organ induction	3.48E-04
kidney epithelium development	3.72E-04
metanephros morphogenesis	5.87E-04
regulation of interleukin-17 production	8.45E-04
regulation of embryonic development	1.04E-03
metanephric nephron development	1.33E-03

Data 5. Gene Ontology Enrichment Results III

GO term enrichment from GREAT for hypomethylated surface ectoderm-DMRs.

Term Name	Binomial P-Value
epidermis development	4.35E-15
skin development	4.03E-13
response to extracellular stimulus	8.72E-12
induction of apoptosis	2.73E-10
response to vitamin	4.38E-10
transforming growth factor beta receptor signaling pathway	1.30E-09
mammary gland epithelium development	2.01E-09
response to nutrient	2.24E-09
signal transduction by p53 class mediator resulting in induction of apoptosis	3.03E-09
positive regulation of protein serine/threonine kinase activity	3.28E-09
response to retinoic acid	3.66E-09
regulation of fibroblast proliferation	8.51E-09
hair follicle development	1.90E-08
hair cycle	2.31E-08
cellular response to external stimulus	2.51E-08
positive regulation of fibroblast proliferation	3.42E-08
regulation of MAP kinase activity	4.92E-08
induction of apoptosis by extracellular signals	7.78E-08
sterol metabolic process	9.52E-08
stem cell development	1.00E-07
cellular response to extracellular stimulus	1.21E-07
stem cell maintenance	1.39E-07
induction of apoptosis by intracellular signals	1.72E-07
regulation of organ morphogenesis	1.76E-07
negative regulation of osteoblast differentiation	1.91E-07
cell-substrate junction assembly	2.12E-07
regulation of morphogenesis of a branching structure	2.50E-07
cholesterol metabolic process	3.60E-07
neuronal stem cell maintenance	6.68E-07
hemidesmosome assembly	1.18E-06
lung-associated mesenchyme development	2.02E-06
Ras protein signal transduction	2.47E-06
regulation of leukocyte degranulation	3.66E-06
regulation of osteoblast differentiation	3.95E-06
negative regulation of glial cell proliferation	3.96E-06
regulation of myeloid leukocyte mediated immunity	4.22E-06

stem cell differentiation	1.06E-05
positive regulation of mesenchymal cell proliferation	1.07E-05
regulation of mesenchymal cell proliferation	1.23E-05
cell volume homeostasis	1.87E-05

Rebecca Lowdon

4515 McKinley Ave. Rm 5121 Saint Louis, MO, 63108
Phone: (540)-580-8816 E-Mail: rebecca.lowdon@gmail.com

Education

Ph.D. Candidate, Molecular Genetics & Genomics **2016**

Division of Biology & Biomedical Sciences, Washington University in St. Louis, St. Louis, MO

Bachelor of Science, *cum laude*; Biology, Chemistry (minor) **2009**

College of William & Mary, Williamsburg, VA

Research Experience

Graduate Research, Dr. Ting Wang **2012 – Present**

Genetics Dept., Washington University in St. Louis, St. Louis, MO

- Developing algorithms to analyze hundreds of genomic datasets to identify functional noncoding somatic mutations in cancer and assign putative functional impact using epigenomic and sequence analysis.
- Pioneered protocol to isolate specific cell populations from transgenic zebrafish embryos by fluorescence-activated cell sorting (FACS). Isolated populations were then prepped for epigenomic analysis.
- Led a collaboration to computationally distinguish epigenetically regulated regions unique to human surface ectoderm-derived cells. Resulted in first author publication Lowdon, et al., *Nature Communications* (2014).

Undergraduate Research, Dr. Margaret Saha **2006 – 2009**

Biology Dept., College of William & Mary, Williamsburg, VA

- Conducted independent research investigating the role of transcription factors and calcium signaling on neurotransmitter phenotype determination in the *Xenopus laevis* retina.
- Optimized conditions for culturing optic placode microdissections and subsequent imaging analysis.

Publications

1. *In review*: **Lowdon RF**, Wang T. Epigenomic annotation of noncoding mutation

identifies mutated pathways in cancer.

2. **Lowdon RF**, Jang HS, Wang T. Evolution of epigenome regulation in vertebrates. *Trends in Genetics.* 32, 269-283 (2016).
3. Zhou X, Li D, Zhang B, **Lowdon RF**, Rockweiler NB, *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature Biotechnol.* 33, 345-346 (2015).
4. Roadmap Epigenomics Consortium, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*. 518, 317-330 (2015).
5. Lee HJ, **Lowdon RF**, Maricque B, Zhang B, Stevens M, Li D, Johson SL, Wang T. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nature Comms.* 6, 6315 (2015).
6. Yue, F., *et al.*, A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355-364 (2014).
7. Zhou X, Li D, **Lowdon RF**, Costello JF, Wang T. methylC Track: Visual integration of single-base resolution DNA methylation data on the WashU Epigenome Brwoser. *Bioinformatics* 30, 2206-2207 (2014).
8. Zhang B, Xing X, Li J, **Lowdon RF**, *et al.* Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers. *BMC Genomics* 15, 868 (2014).
9. **Lowdon RF**, *et al.* Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. *Nature Comms.* 5, 5442 (2014).
10. Zhang B, Zhou Y, Lin N, **Lowdon RF**, *et al.* Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Research* 23, 1522-1540 (2013).
11. Xie M, Hong C, Zhou X, Li D, Lee HJ, **Lowdon RF**, *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genetics* 45, 836-841 (2013).
12. Zhou, X., **Lowdon RF**, *et al.*, Exploring long-range genome interactions using the WashU Epigenome Browser. *Nature Methods* 10, 375-376 (2013).
13. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012).
14. Stamatoyannopoulos, J.A., *et al*. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*. 13, 428 (2012).
15. Roy, S. *et al.* Identification of functional elements and regulatory circuits by Drosophila

modENCODE. *Science* 330 1787-1797 (2010).

16. Gerstein, M.B., *et al*. Integrative analysis of the Caenorhabditis elegans gneome by the modENCODE project. *Science* 330 1775-1787 (2010).

Selected Presentations

1. DNA methylation dynamics in pigment cell development. Poster presentation, Biology of Genomes, Cold Spring Harbor, NY. May 5-9, 2015.
2. Normal cell epigenomes underlie functional and regulatory relationships between disease genes. Poster presentation, Keystone Epigenomics meeting, Keystone, CO. March 29-April 3, 2015.
3. WashU Epigenome Browser Workshop. Organizer and workshop presenter, American Society of Human Genetics, San Diego, CA. Oct. 19-23, 2014.
4. Regulatory network construction from the epigenome of normal cells reveals functional connections between disease genes. Poster presentation, American Society of Human Genetics, San Diego, CA. Oct. 19-23, 2014.
5. Epigenetic regulation of pigment cell fate. Selected speaker, WU DBBS Molecular Genetics & Genomics Program Annual Retreat. St. Louis, MO, Sept. 19-20, 2014.
6. Deciphering lineage-specific regulatory circuitry from differentiated cell epigenomes." Selected speaker, Roadmap Epigenomics meeting, Boston, MA. Oct. 20-21, 2013.
7. Differential DNA methylation in surface ectodermal cells marks regulatory elements for epithelial cell identity. Selected speaker, WU DBBS Molecular Genetics & Genomics Program Annual Retreat, St. Louis, MO. Sept. 20-21, 2013.
8. Finding lineage-specific epigenetic signatures from genome-wide data using a novel statistical algorithm. Poster presentation, Biology of Genomes, Cold Spring Harbor Laboratory, May 7-11, 2013.
9. Mechanisms of tissue-specific gene regulation driven by transposable elements. Poster presentation, WU DBBS Molecular Genetics & Genomics Program Annual Retreat, St. Louis, MO. Sept. 28-29, 2012. Honorable mention.
10. Deciphering the human genome with ENCODE: the Encyclopedia of DNA Elements. Poster presentation, National Human Genome Research Institute Division of Intramural Research Annual Retreat, Bethesda, MD. Nov. 15-16, 2010.
11. Role of calcium activity in neurotransmitter phenotype determination in the *Xenopus* retina. Poster presentation, Morphogenesis & Regenerative Medicine Symposium, Charlottesville, VA. May 27-28, 2009.

12. Transcriptional control of *GAD66* in the *Xenopus* retina. Poster presentation, Society for Developmental Biology Annual Meeting, Philadelphia, PA. July 26-30, 2008.

Awards and Honors

NSF Data Science Workshop **2015**

White paper on genomic data visualization selected.

Clinton Global Initiative University (CGI U) Attendee **2013, 2014**

Twice selected to the CGI U for original proposals to expose barriers to innovation in technology and biomedicine.

National Science Foundation Graduate Research Fellowship Program Award **2013**

HHMI Undergraduate Science Education Award **2007 – 2008**

College of William & Mary Program Grant; five awards totaling $5300.

Work Experience

National Human Genome Research Institute (NHGRI), NIH **2009 – 2011**

Scientific Program Analyst (Contractor, Kelly Services)

- Worked closely with 20 Primary Investigators who were grantees of the ENCODE and modENCODE projects to monitor data production and organize analysis across each Consortium.
- Developed controlled vocabulary to record technical and biological metadata for genomic datasets.
- Managed projects and tasks for a five-person NHGRI mod/ENCODE scientific management team.
- Planned two annual mod/ENCODE Consortia meetings of ~200 people.

Leadership Experience

Co-Founder, WU Graduate Students Promoting Science Policy, Education, and Research **2012**

- Conceived and planned career development events, including the Early Career Transitions Symposium (June 2015), which attracted >60 graduate students to network with 15 local scientists.

- Organized a team to pioneer signature event series, *"Where's My Jetpack?"* which brings together scientists and the public to discuss barriers to innovation in technology and biomedicine. Three speaker seminars and one podcast series distributed over the past 3 years (https://thought.artsci.wustl.edu/wheres-my-jetpack).
- One of the first University science policy groups in the country.

Central Region Co-Director, National Science Policy Group **2015 – 2016**

- Mentor graduate student leaders as they establish science policy groups on campus in the Midwest US.
- Initiating an NSPG-wide survey of member groups to gather data on our members.

Science Communication Accomplishments

Editorial experience **2012 – Present**

- Four years experience editing Wang lab manuscripts (9) and grant applications (3) for language, clarity.

***The Biochemist*, Feature Article** **2015**

"From DNA to a human" (http://www.biochemist.org/bio/03705/0024/037050024.pdf)

- Invited to write feature story for *The Biochemist* issue themed "What makes us human?"

American Society of Human Genetics, WashU EpiGenome Browser Workshop, Presenter **2014**

San Diego, CA

- Organized team of three, including a fellow post-doc and my PI, to develop original content for Workshop.
- Published the EpiGenome Browser Handbook.
- Presented to an audience of >250 attendees.

Advanced Genetics, Guest Lecturer **2013**

Washington University in St. Louis

- Planned and delivered 3 lectures on genetics principles, yeast two-hybrid assays, and epigenomics as part of Teaching Assistant assignment.

Roadmap Epigenomics Meeting, Selected Speaker **2013**

Boston, MA

- Selected to present original research on using developmental biology to interpret cell type epigenomes to an audience of >200.

Research presentations **2006 – 2015**

- Poster presentation at DNA Methylation meeting, Keystone, CO (2015).
- Poster presentations at annual Biology of Genomes meetings, Cold Spring Harbor, NY (2013, 2015).
- Poster presentation at American Society of Human Genetics annual meeting (2015).
- Oral presentations at Genetics Department seminars (2013, 2014).
- Poster presentations an annual Cell & Molecular Biology Symposium, Washington University in St. Louis, (2013, 2014).
- Selected speaker (2013) and poster presentations annual Genetics Department retreats, (2012, 2014).
- Poster presentations at Student Research Symposia at The College of William & Mary, (2007-2009).

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics & Genomics

Dissertation Examination Committee:
Ting Wang, Chair
Sarah Elgin
Stephen L. Johnson
Samantha Morris
Nancy Saccone
Tim Schedl

The Epigenomics of Cell Fate in Development and Disease

by

Rebecca Faith Lowdon

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2016
St. Louis, Missouri

Table of Contents

Chapter 5

Evolution of Epigenetic Regulation in Vertebrate Genomes

Chapter 6

Synthesis

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgments

I am grateful to many people for the support and encouragement given me over the past several years. I am thankful for the unwavering support and guidance of my mentor, Ting Wang. He has given me tremendous opportunities, for which I am very thankful. Joining the Wang lab has allowed me to explore science with an exciting new perspective, and much of the credit for that goes to Ting. I am also thankful to my fellow lab members who helped me push my own boundaries as we pursued our research together.

The opportunity to work with various collaborators made this dissertation a very rewarding one. First many thanks are due to Scott Higdon for advice with zebrafish project experimental design and protocols, and Stephen Johnson for his conception of this project. Josh Jang provided invaluable assistance during the final stages of the zebrafish work. The skin project was a collaboration with Jeffrey Chang and Joseph Costello of the University of California – San Francisco. I am grateful to them for collaborating with me early in my graduate career.

I thank my thesis committee members who have provided valuable advice and guidance over the years. I am especially grateful to Tim Schedl, Sally Elgin, and Stephen Johnson for their mentorship over the past four years, and to Nancy Saccone and Samantha Morris for joining my committee.

I am grateful to my many friends in St. Louis and beyond who have supported me over the past five years. They have made a long journey a very enjoyable one.

Above all, I thank my family, whose continuous love and support I treasure immensely.

In addition, I thank the Alvin J. Siteman Cancer Center at Washington University in School of Medicine and Barnes-Jewish Hospital in St. Louis, Mo., for the use of the Siteman Flow Cytometry Core, which provided flow cytometry service as described in Chapter 3. The Siteman Cancer Center is supported in part by an NCI Cancer Center Support Grant #P30 CA91842.

I thank my funding sources: the NSF Graduate Research Fellowship Program (DGE-1143954) and the Washington University Interface of Psychology, Neuroscience, and Genetics training program (5T32GM081739).

Rebecca Faith Lowdon

Washington University in St. Louis

December 2016

Dedicated to Mom and Dad.

Abstract of the Dissertation

The Epigenomics of Cell Fate in Development and Disease

by

Rebecca Faith Lowdon

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics & Genomics

Washington University in St. Louis, 2016

Professor Ting Wang, Chair

Epigenetic features at regulatory elements provide instructive cues for transcriptional regulation during development. However, the particular epigenetic alterations necessary for proper cell fate acquisition and differentiation are not well understood. This dissertation explores the epigenetic dynamics of regulatory elements during development and uses epigenome annotations to document inappropriate transcriptional regulation in disease. First, I summarize my contributions to developing a new algorithm for detecting differential DNA methylation, M&M. I report the application of the M&M algorithm to identify distinct classes of DNA methylation dynamics in surface ectoderm (SE) progenitor cells and SE-derived lineages: epigenome alterations, and differential DNA methylation in particular, that are present in progenitor cells are transmitted to daughter cells and consequently observed in differentiated cells. I exploit this property of DNA methylation to characterize DNA methylation dynamics in surface ectoderm embryonic tissue and SE-derived cells. Next, I use zebrafish to investigate the biological relevance of the classes of DNA methylation dynamics described in the SE context. In zebrafish, I use the pigment cell development system to understand the contribution of DNA methylation to a particular cell fate

choice: melanocyte or iridophore cell fate. Next, I investigate the consequence of somatic mutations in primary liver cancer by utilizing epigenomic annotations of human tissues to distinguish putatively functional mutations from passenger mutations. Here I present support for the hypothesis that transcriptional regulatory instructions for heterologous cell types are co-opted by cancer cells during malignant tumorigenesis. Finally I present a review of the evolution of epigenetic regulation over regulatory elements. Altogether, this dissertation advances our understanding of epigenetic regulation in cell fate decisions by integrating functional genomics with developmental biology and cancer genetics.