INTRINSIC SPECIFICITY OF BINDING AND REGULATORY FUNCTION OF CLASS II BHLH TRANSCRIPTION FACTORS

APPROVED BY SUPERVISORY COMMITTEE

Jane E. Johnson Ph.D.

Helmut Kramer Ph.D.

Genevieve Konopka Ph.D.

Raymond MacDonald Ph.D.

INTRINSIC SPECIFICITY OF BINDING AND REGULATORY FUNCTION OF CLASS II BHLH TRANSCRIPTION FACTORS

by

BRADFORD HARRIS CASEY

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2016

DEDICATION

This work is dedicated to my family, who have taught me pursue truth in all forms.

To my grandparents for inspiring my curiosity,

my parents for teaching me the value of a life in the service of others,

my sisters for reminding me of the importance of patience,

and to Rachel, who is both "the beautiful one", and "the smart one",

and insists that I am clever and beautiful, too.

INTRINSIC SPECIFICITY OF BINDING AND REGULATORY FUNCTION OF CLASS II BHLH TRANSCRIPTION FACTORS

Publication No. _____

Bradford Harris Casey

The University of Texas Southwestern Medical Center at Dallas, 2016

Jane E. Johnson, Ph.D.

PREFACE

Embryonic development begins with a single cell, and gives rise to the many diverse cells which comprise the complex structures of the adult animal. Distinct cell fates require precise regulation to develop and maintain their functional characteristics. Transcription factors provide a mechanism to select tissue-specific programs of gene expression from the shared genome. ASCL1, ASCL2, and MYOD are class II basic Helix-Loop-Helix (bHLH) transcription factors which play crucial roles in lineage specification in the developing embryo. In vivo, these factors bind to distinct genomic sites, and regulate distinct transcriptional programs. The mechanisms by which they select their cognate binding sites

remain poorly defined. Here, we utilize an inducible system to express these master regulatory factors in embryonic stem cells to characterize early events in bHLH factor binding and function in a common cellular context, removed from their role as endogenous master regulators of lineage specification. Using genome-wide sequencing approaches, we demonstrate that these factors maintain distinct binding when ectopically expressed in a common context. We observe that they initiate distinct transcriptional programs, which include key regulators in lineage specification. By comparing chromatin accessibility of bHLH binding sites, we reveal a shared ability for these factors to bind nucleosome-occupied sites, and meet the criteria which define pioneer transcription factors. We further characterize epigenetic features of the empirically observed genome-wide binding sites of these factors, and compare these findings to the conventional understanding of bHLH factor function. This work represents the first comprehensive approach to direct comparison of early events in the binding and transcriptional profiles of ASCL1, ASCL2, and MYOD.

ACKNOWLEDGEMENTS

In preparing this work, I have reflected often on the great many kindnesses which have been visited upon me during my time at UT Southwestern. How can one begin to square the accounts of the generous efforts of so many? Regardless, this is the best opportunity to do so en masse, and I would be remiss to neglect the chance to express my appreciation.

First, I wish to express my most sincere gratitude to my mentor, Dr. Jane E. Johnson. Her hard work, dedication, and collaborative spirit serve as an example for all scientists. Her kindness and generosity are without limit, and her patience, without equal. Many scientists fall prey to the illusion that ours is a field of competitors, rather than colleagues; Jane has always sought to keep friends, rather than make enemies of her peers. She has chosen to maintain her office in the center of a noisy and chaotic lab, rather than a posh and quiet executive suite, and maintains an open door for questions, feedback, approval, and support. She has earned the respect and friendship of everyone who knows her.

I also gratefully acknowledge the generous guidance of my dissertation committee, which has been invaluable in the completion of this work; Dr. Helmut Kramer, who taught me to always try to test a theory, rather than try to confirm it, Dr. Raymond MacDonald, whose hand-written notes were always more valuable than my work deserved, and Dr. Genevieve Konopka, whose thoughtful and considered insight has always encouraged me to push forward, and to find the silver lining. They are all great assets to our institution, and to their respective fields of research. I am forever indebted to each for the guidance and support they have provided me over these long years.

The completion of this work would not have been possible without the support of my many colleagues and friends at UT Southwestern. I have spent many late nights in the lab, and have rarely done so alone. It has been a wild ride, and I am grateful to have shared it with you.

In particular, the members of the Johnson lab have selflessly shared their experience and knowledge to help see this research come to fruition, and have always provided honest feedback. Dr. Helen Lai and Dr. Tou Yia Vue have always challenged me to consider carefully my approach and results, and to find and cultivate the story that it has revealed. My fellow students Dr. Mark Borromeo and Dr. Joshua Chang always shared the wisdom of their experience, and taught me how to survive the gauntlet of the neuroscience graduate program. We oft burned the midnight oil, and our late nights and early mornings were some of the happiest and most memorable times that I spent in lab. They also taught me to play poker, which has ensured that I will never be taken to cards, as I will surely lose. Lauren Tyra, who will certainly have finished her Ph.D. by the time this is printed, has been a constant source of personal support and scientific discussion. She has worked alongside me in various volunteer efforts and leadership roles, and serves as an example of a student who has gone the extra mile to improve opportunities for those less fortunate than ourselves. Erin Kibodeaux, who has left science to follow her dream of caring for others, has been a dear friend, and brought much-needed levity to my time in the lab. Her patients will benefit greatly from her intelligence and skill, and will be comforted by her kind demeanor.

I am especially indebted to Clark Rosensweig, for reasons too numerous to list here. His passion and capability for research is exceptional, his insight, invaluable, and his wit,

peerless. His constant support and encouragement have been immeasurable comfort in hard times. He is indisputably the better scientist, but has steadfastly insisted that we are equal, and that I have something to contribute. The gilded words of true friends are bracing in the face of hardship, even when we know them to be too generous. I am eternally grateful for his kindness and his fraternity over the years.

I also wish to thank Dr. Matthew Goldberg, who took great risk in hiring me, despite my having no significant experience in research. He entrusted me with precious resources, trained me in many techniques, and treated me as a valued member of his laboratory. I never would have imagined that I could be a successful mouse surgeon. He encouraged me to make the most of my time at UT Southwestern, and allowed me to attend lectures and conferences. He supported me when I left to pursue my education, and eagerly wrote letters of introduction and recommendation, without which I certainly would never have had this opportunity.

Often neglected are the many faculty and staff which make up the University of Texas Southwestern Medical Center, and especially the College of Biomedical Sciences Graduate Program. The many people that make up the Office of the Dean spare no effort in keeping this complex operation running smoothly. I am especially grateful for the unique and exceptional opportunities to serve our students and our university, which have been some of the most satisfying experiences I have enjoyed during my tenure. I especially want to recognize the efforts of Wes Norred, who has given myself and others the opportunity to advocate for our students in many venues. He secured an audience with everyone I ever asked, listened, and supported my modest efforts to improve our fair institution. This school

would not be the same without his tireless advocacy for our students. I genuinely cannot fathom why I have been allowed such liberty, and such privilege, but it has been a great honor.

Outside the academe, Chad Lumley, an exceptional mathematician and physicist, has been a great friend, and our many projects, adventures, and discussions have been a welcome and satisfying distraction from my studies. Few friends will let you call them at any hour, even fewer will invite you over for dinner when you wake them up. I thank him especially for teaching me calculus, and for enduring my terrible jokes about maths. I remain certain that my definition of a polar bear is correct.

Miryam Prodanovic has been a constant comfort, and despite great distance has remained my steadfast friend and confidant. Her handwritten letters have lifted my spirits when they were low, and her many drawings and photographs have provided inspiration and opportunity for reflection. To be Miryam's friend is to drink from a hydrant of kindness, support, and creativity. I am forever in arrears for her friendship, and I will forever labor to repay her.

I feel compelled to assert that I could not possibly convey my appreciation for those mentioned above, much less the many who are not mentioned herein. Any future success I may enjoy is predicated on the kindness, generosity, and friendship of these and many others. The gratitude I express here is simply insufficient to even begin to describe my appreciation for all that you have done. I am brought to tears by the magnitude of these debts, and the knowledge of how fortunate I am to have been given these opportunities. I came to science in the hopes of a life in the service of others. I will not rest until I find it.

Such, such were the joys…

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Frank-Cannon, T. C., Tran, T., Ruhn, K. A., Martinez, T. N., Hong, J., Marvin, M., Hartley, M., Trevino, I., O'Brien, D. E., **Casey, B.**, Goldberg, M. S., & Tansey, M. G. (2008). Parkin deficiency increases vulnerability to inflammation-related nigral degeneration. *J Neurosci, 28*(43), 10825-10834. doi:10.1523/JNEUROSCI.3001-08.2008

Hennis, M. R., Seamans, K. W., Marvin, M. A., **Casey, B. H.**, & Goldberg, M. S. (2013). Behavioral and neurotransmitter abnormalities in mice deficient for Parkin, DJ-1 and superoxide dismutase. *PLoS One, 8*(12), e84894. doi:10.1371/journal.pone.0084894

Meredith, D. M., Borromeo, M. D., Deering, T. G., **Casey, B. H.**, Savage, T. K., Mayer, P. R., Hoang, C., Tung, K. C., Kumar, M., Shen, C., Swift, G. H., Macdonald, R. J., & Johnson, J. E. (2013). Program specificity for Ptf1a in pancreas versus neural tube development correlates with distinct collaborating cofactors and chromatin accessibility. *Mol Cell Biol, 33*(16), 3166-3179. doi:10.1128/MCB.00364-13

Nguyen, T. A., Frank-Cannon, T., Martinez, T. N., Ruhn, K. A., Marvin, M., **Casey, B.**, Trevino, I., Hong, J. J., Goldberg, M. S., & Tansey, M. G. (2013). Analysis of inflammation-related nigral degeneration and locomotor function in DJ-1(-/-) mice. *J Neuroinflammation, 10*, 50. doi:10.1186/1742-2094-10-50

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF DEFINITIONS

AP-1 - Activator Protein 1

AS-C - Achaete-scute

ASCL - Achaete-scute-like

ATAC - Assay for Transposase-Accessible Chromatin

ATF - Activating Transcription Factor 2

ATP - Adenosine Triphosphate

BMP - Bone Morphogenic Protein

bHLH - basic Helix-Loop-Helix

BP - Biological Process

cAMP - Cyclic adenosine monophosphate

ChIP - Chromatin Immunoprecipitation

CPB - CREB Binding Protein

CPDB - Consensus Path Database

DEG - Differentially expressed gene

DMEM - Dulbecco's Modified Eagle's minimal Medium

DNA - Deoxyribonucleic acid

DNMT - DNA methyltransferase

EDTA - Ethylenediaminetetraacetic acid

EGTA - Ethylene Glycol Tetraacetic Acid

ES cells - Embryonic Stem Cells

EMSA - Electrophoretic Mobility Shift Assay

ENCODE - Encyclopedia of DNA Elements

FAIRE - Formaldehyde-Assisted Isolation of Regulatory Elements

FDR - False Discovery Rate

GATA - GATA Binding Protein

Gcm1 - Glial-cells-missing

GEO - Gene Expression Omnnibus

GO - Gene Ontology

GREAT - Genomic Regions Enrichment Annotation Tool

HCl - Hydrogen Chloride

HDAC- Histone Deacetylase

H3K4me1 - Monomethylated Histone H3 at Lysine 4

H3K4me3 - Trimethylated Histone H3 at Lysine 4

H3K9ac - Acetylated Histone H3 at Lysine 9

H3K9me3 - Trimethylated Histone H3 at Lysine 9

H3K27ac - Acetylated Histone H3 at Lysine 27

H3K36me3 - Trimethylated Histone H3 at Lysine 36

HEPES - 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HES - Hes Family BHLH Transcription Factor

HOMER - Hypergeometric Optimization of Motif EnRichment

HOXC9 - Homeobox C9

HMM - Hidden Markov Modeling

ICM - Inner cell mass

ID - Inhibitor-of-DNA-binding

iPSC - Induced pluripotent stem cells

IRF1 - Interferon Regulatory Factor 1

KLF - Kruppel-Like Factor

KOH - Potassium Hydroxide

LIF - Leukemia Inhibitory Factor

lncRNAs - Long noncoding RNAs

MAPK - Mitogen-Activated Protein Kinase

MEF2 - Myocyte Enhancer Factor 2

MEFs - Mouse Embryonic Fibroblasts

miRNA - MicroRNA

MNase - Micrococcal Nuclease

mRNA - Messenger RNA

MYOD - Myogenic differentiation?

NANOG - Nanog Homeobox

ncRNA - Non-coding RNA

NEUROD2- Neuronal Differentiation 2

Nonidet - 40 Octylphenoxy poly(ethyleneoxy)ethanol, branched

NPCs - Neural Progenitor Cells

OCT - Octamer-binding transcription factor

PBS - Phosphate buffered saline

PBX - Pre-B-Cell Leukemia Homeobox

PCR - Polymerase chain reaction

PWM - Position Weight Matrix

REST/NRSF - RE1-Silencing Transcription factor/Neuron-Restrictive Silencer Factor

RIN - RNA Integrity Number

RIPA - Radioimmunoprecipitation Assay

RNA - Ribonucleic acid

RPKM - Reads Per Kilobase transcript per Million mapped reads

RT-qPCR - Reverse-transcription quantitative polymerase chain reaction

SDS - Sodium dodecyl sulfate

Shh - Sonic hedgehog

SOX - Sex Determining Region Y-box

SWI/SNF - SWItch/Sucrose Non-Fermentable

TCF4 - Transcription Factor 4

TEAD - TEA Domain Transcription Factor

TF - Transcription factor

TFBS - Transcription Factor Binding Site

TSS - Transcriptional start sites

UBC - Ubiquitin C

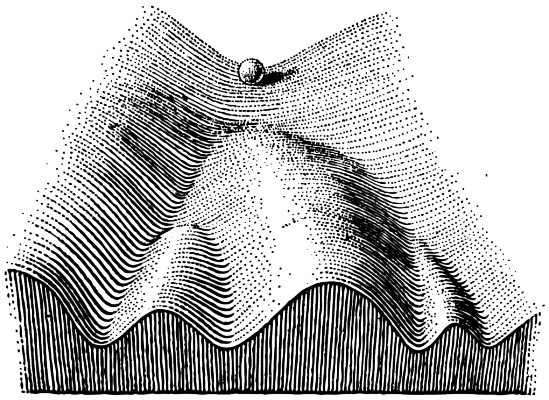
UTR - Untranslated Region

WNT - Wingless-type MMTV integration site family member

ZBTB7b - Zinc Finger And BTB Domain Containing 7B

ZNF - Zinc Finger Protein

Illustration 1 : Waddington's Epigenetic Landscape



Part of an Epigenetic Landscape.

The path followed by the ball, as it rolls down towards the spectator, corresponds to the developmental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

C.H. Waddington, 1957 *The Strategy of the Genes*

CHAPTER ONE

Introduction

THE ROLE OF BHLH FACTORS IN CELL FATE SPECIFICATION

Cell fate specification

All animals begin the course of life in reduced form. Order is gradually bestowed through the process of development, through which specialized cell types are defined, and refined into the myriad tissues of the adult animal. The final forms of cells are commonly referred to as fates, reflecting both the significance, and uncertainty which has defined this field of biology. From the single-celled zygote, organisms grow exclusively through the processes of cell growth and division, initiating the complex chain of events known as embryonic development. From this single cell, every component of the organism must be derived, eventually giving rise to the specified cell fates of the animal. This common progenitor contains the genetic information required for each of these specialized fates, and this process, known as cell fate specification, must be meticulously orchestrated to guide each cell to its appropriate destiny. Thus, the development of the embryo is dependent on a complex series of self-organized cell division, which gradually establishes a hierarchy of distinct populations of cells, known as lineages. Cells within these lineages are themselves gradually specialized, and eventually give rise to the diverse terminal cell fates of the adult organism. This is conventionally illustrated in the form of a ball rolling down a slope, with distinct furrows gradually restricting the ability of these cells to move between fates. This depiction was developed by C.H. Waddington, in concert with a theory about how gene

regulation is shaped by the cellular environment (Waddington, 1947), referred to as the epigenetic landscape (Illustration 1: Waddington's Epigenetic Landscape). In this theory, he posited that each cell's list of potential fates is gradually restricted based on the progressive lineage specification it experiences. As technology and understanding of developmental processes has improved, biologists have gradually uncovered some of the mechanisms by which cell lineage is established and maintained.

Initially, the embryo forms a small number of such lineages; this special subset is referred to as the germ layers: ectoderm, endoderm, and mesoderm. Once established, these layers undergo successive rounds of cell division, and each is itself the apex of a hierarchy of potential cell fates. Each of these higher-order germ layers produces a separate set of tissues, and for this reason, each is indispensable for survival. Crucially, once these lineages are established, cells of a given lineage are generally incapable of achieving cell fates specified from another germ layer, a feature known as *lineage restriction* or *lineage specification*. With each progressive division, the set of possible fates is gradually reduced, and the result of this process is an ever-narrowing set of potential fates, allowing for precise cellular specialization, and providing order from the otherwise chaotic effects of uncontrolled cell division. These limitations of cell fate are a central theme in embryonic development, and precede the formation of the embryo itself; cells derived from different layers of the blastula are restricted into separate embryonic and extraembryonic (trophoectodermal) lineages, and cells from either lineage family cannot functionally replace those of the opponent lineage. Such restriction is relevant at every stage of embryogenesis, and is critical to appropriate developmental progression.

The early embryo is particularly significant, in that it represents a brief window of organized cellular proliferation prior to establishing these lineages. In contrast to the adult organism, cells of the early embryo remain essentially unspecified prior to the self-organized specification of the germ layers. Pluripotency, the capacity to become distinct cell types, is a feature of these early populations in embryonic development; in most populations it is gradually lost as these cells undergo successive rounds of division and specialization. Totipotency, the capacity to become any of the cells of the organism, is a further distinction, and is naturally restricted to the earliest populations of cells generated from the blastocyst, before discrete populations are formed. Cells with this capability are termed stem cells, and represent unique populations in the organism, which can give rise to multiple types of cells. Embryonic Stem Cells (ES cells) are a unique type of cells derived from the inner cell mass (ICM) of the early embryo; clones of single cells can be differentiated into cells from any of the three germ layers (Evans et al., 1981; Martin, 1981). To accomplish this, they express a core pluripotency gene network, which restricts differentiation by maintaining expression of genes which provide this pluripotent capacity, and preventing expression of genes which serve to differentiate cells into discrete developmental lineages. Among these factors are SOX2, NANOG, OCT4, MYC, KLF4 and others. Together, these genes act as a network of master regulators of pluripotency, and have been shown to be highly expressed in ES cells (Liu et al., 2008), where they directly regulate downstream targets associated with maintaining this pluripotent capacity. Through careful handling, these cells can be expanded and maintained in cell culture for extended periods (Evans et al., 1981). Highlighting the transience of this pluripotent state, these cells spontaneously differentiate under standard cell

culture conditions, requiring supportive co-culture on feeder cells (Evans et al., 1981; Martin, 1981), or attentive addition of exogenous growth factors (Smith & Hooper, 1987; Williams et al., 1988), to be maintained in culture indefinitely. This cell population is present only very briefly in the embryo, and quickly gives way once germ layer formation occurs, beginning the process of specification for the many distinct cell types present in the mature organism. Indeed, the considerable majority of cells present in the adult organism are fully specified, and only a few, select populations of multipotent cells persist in adult animals; which are responsible for maintaining specific cell populations throughout life.

The process of lineage specification is not merely a method of establishing the taxonomy of cells, however, but is central to the survival of the organism. Cell fate specification provides additional transcriptional regulatory capacity, and is a central mechanism in preventing uncontrolled gene expression. This is one mechanism that ensures the appropriate numbers and types of cells are generated for various tissues, and that these cells have appropriate properties for their specific function. It has long been suggested that transcriptional programs present in cells are a combination of two categories of genes: a basic set of ubiquitously expressed genes involved in growth, division, and maintenance; and a second set of "luxury" genes, which are necessary for specialized cell function (Weintraub et al., 1972*). Lineage specification is a reflection of this distinction, and while many genes are commonly expressed across all lineages, these "luxury" genes are precisely regulated, and restricted to relevant cell types. The mechanisms regulating lineage-specific gene expression are necessarily complex, as each cell must faithfully execute its specific transcriptional program from the common template of the genome, integrating information from the

environment, and responding to perturbations as they arise. It is this complexity that allows for the development and maintenance of the disparate cell types of all multicellular organisms.

Role of transcriptional regulators in cell fate specification

Early rounds of symmetrical cell division occurring in the blastocyst result in increased cell number, but these cells remain unspecified. Once gastrulation has initiated the formation of germ layers, the embryo demonstrates a remarkable, self-ordered process of asymmetrical cell division, wherein these three layers rapidly establish body planes. These define the basic structure of the embryo, which is maintained throughout life. This, and virtually every subsequent process of cell fate specification, is the result of transcriptional regulators. In vertebrates, one of the earliest structures formed is that of the *neural tube*, which establishes the basis of the nervous system.

This structure provides a compelling example of how cell populations are progressively defined from the germ layers. Initially formed by an invagination of the ectoderm at the neural plate, the neural plate progressively furrows and is closed by convergent extension of cells at the lateral edges of the neural plate to produce a hollow, tubular structure, the neural tube, which extends along one edge of the embryo. This process is referred to as *primary neurulation*. Two signaling proteins, *Sonic hedgehog (Shh)*, expressed at the center of the furrow, and *Bone Morphogenic Protein (BMP)*, expressed at the sealed margin of the neural tube, are expressed on opposing sides of this structure; these establish the ventral, and dorsal aspects of the neural tube, respectively. Together, these

proteins establish opposing gradients across the neural tube, which convey positional information to cells and establish functionally distinct cell populations through the activity of downstream regulatory cascades in a spatiotemporally regulated manner. From these two signaling proteins, the process by which the complex nervous system of the adult animal is set in motion. The effects of neural tube development extend even to non-neural tissues, as mesodermal tissues are themselves partially patterned by their proximity to this structure (Munsterburg & Lassar, 1995; Alveset et al., 2003). This process is one example of the many instances of embryonic patterning, by which complex structures of distinct cells may be generated from largely similar progenitor populations.

While there are many mechanisms by which cells orchestrate the different cell fates within a given lineage, one of the most powerful is the use of a subtype of transcriptional regulators known as transcription factors (TFs), which serve to regulate gene expression. While this broad class of proteins is functionally diverse, and operates through multiple mechanisms, one central aspect is by binding directly to DNA through so-called DNA binding domains. These domains facilitate regulation by identifying specific DNA nucleotide sequences, referred to as *motifs*. In doing so, these factors may act directly or indirectly, recruiting activating or repressive complexes and transcriptional machinery to specifically increase or decrease gene expression. The diversity of DNA-binding TFs is a central mechanism in providing the necessary regulatory complexity to direct appropriate gene expression from the common genetic template of the genome.

One class of transcription factors of particular significance to lineage and cell-fate specification is the basic Helix-Loop-Helix (bHLH) family. These proteins are broadly

conserved, and are present in all eukaryotic organisms from yeast to mammals (Atchley & Fitch, 1997). This family of factors was initially characterized based on diverse functional roles as transcriptional activators in development (Little et al., 1983; Davis et al., 1987; Villares & Cabrera, 1987; Alonso & Cabrera, 1988; Tapscott et al., 1988; Wright et al., 1989; Johnson et al., 1990). Structurally, these factors share a similar domain structure, consisting of two amphipathic α-helices connected by a loop region, which interact in a tertiary structure, and form quaternary complexes with other bHLH proteins, and bind to DNA as a complex (Murre et al., 1989; Murre et al., 1994; Ferre-D'Amare et al., 1993). bHLH proteins are subdivided into seven classes based on their expression pattern, and their functional characterization (Murre et al., 1994; Crews, 1998)[note: classifications presented as reviewed in *Massari & Murre, 2000*, which appears to be the first mention of class VII]. Of these, class I and class II factors are of principal significance in establishing cell fate.

Tissue-specific bHLH transcription factors

Class I bHLH factors, also known as E-proteins, are broadly, if not ubiquitously, expressed, and play crucial roles in development. These factors were among the first proteins associated with a specific DNA binding motif, binding the hexameric E-box, with the sequence *CANNTG*, for which they are named (Murre et al., 1989). These factors form homodimeric and heterodimeric complexes with each other (Murre et al., 1989), and other bHLH factors. Members of this family include E12 and E47 (which are both transcribed from the *E2A* locus (*Tcf3*)), E2-2 (*Tcf4*), HEB (*Tcf12*) in vertebrates, and are homologues to the Drosophila E-protein Daughterless (*da*). E-proteins play central roles in hematopoietic lineages (Murre et al., 1991; Shen et al., 1995), but are also of central importance in

establishing cell types in many different lineages through interaction with tissue-specific bHLH factors, known as class II bHLH factors. Class I factors also interact with other bHLH factors, including tissue-specific class V (*Id)* factors, which do not possess a DNA-binding domain, and functionally repress the activity of both class I and class II factors through this interaction.

Class II bHLH transcription factors are a family of broadly conserved developmental regulatory proteins which play key roles in lineage specification. A number of these factors were discovered as tissue-specific regulators involved in cell fate specification (Lassar et al., 1986; Davis et al., 1987; Villares & Cabrera 1987; Johnson et al., 1990; Guillemot et al., 1992). *Ascl1* and *Ascl2*, also known as *Mash1* and *Mash2*, were discovered as mammalian homologs of the Drosophila *Achaete-scute* (AS-C) genes, and were initially characterized based on the role of AS-C in neural development (Johnson et al., 1990; Johnson et al., 1992). Likewise, *MyoD* was discovered in a screen for gene mediators of muscle cell fate (Davis et al., 1987). As a class, their defining characteristic is the shared presence of the eponymous bHLH domain, and their ability to bind to DNA in a heterodimeric complex with E-proteins (Murre et al., 1989), the class I bHLH factors. While some class II bHLH factors can form homodimers, they have been shown to preferentially act as heterodimers with E-proteins through in vitro reporter assays (Lassar et al., 1991). As with all DNA-binding bHLH factors, class II factors have previously been demonstrated to bind to a degenerate Ebox motif with the nucleic acid sequence *CANNTG* (Ephrussi et al., 1985). Through sequence-specific DNA-binding, and interaction with ubiquitously expressed E-proteins, these factors act as transcriptional activators, and have previously been demonstrated to regulate lineage-specific

gene targets in establishing cell fate (as reviewed in Massari & Murre, 2000). Crucially, the tissue-specific expression of class II factors provides specificity for tissue-specific interaction with the class I E-proteins.

ASCL1, ASCL2, and MYOD are developmentally critical class II bHLH proteins, and play central roles in defining neural, trophectodermal, and muscle lineages in the developing embryo, respectively, where they are expressed in lineage restricted multipotential progenitor populations during development. These proteins act as transcriptional activators, often through their interactions with E-proteins, which feature independent activation domains outside of the bHLH region (Henthorn et al., 1990; Aronheim et al., 1993). The structure of these factors and their E-protein binding partners inform the binding preferences of these complexes, and a number of preferred central dinucleotide motifs have been previously characterized for these factors from *in vivo* and *in vitro* studies (Castro et al., 2011; Cao et al., 2010; Berkes et al., 2004; Jolma et al., 2013; Schuijers et al., 2014), with each binding variations of the aforementioned E-box. A considerable number of other class II bHLH factors exist, and are also involved in tissue-specific gene expression and lineage specification for these and other lineages (as reviewed in Massari & Murre, 2000). In vivo, ASCL1, ASCL2 and MYOD have been shown to have overlapping, but distinct DNA binding patterns in their respective tissues (Cao et al., 2010; Schuijers et al., 2014; Borromeo et al., 2014), and regulate distinct transcriptional programs in these contexts. Based on the known role of these factors in development (Johnson et al., 1990; Guillemot et al., 1993; Guillemot et al., 1994; Davis et al., 1987; Weintraub et al., 1991; Tapscott et al., 1993), these factors have previously been characterized as master

regulatory factors on the basis of their central roles in establishing their respective lineages. Indeed, such is the dramatic effect of these factors as a class that ectopic expression of ASCL1 alone is sufficient to induce cells to differentiate from neural progenitors (Nakada et al., 2004), as well as establish neural lineage cells from P19 cells (Farah et al., 2000), and fibroblasts (Vierbuchen et al., 2010). Similar evidence demonstrates the ability of MYOD to induce myogenic lineages from 10T1/2 embryonic fibroblasts (Lassar et al., 1986; Tapscott & Weintraub, 1988). Importantly, while MYOD expression is sufficient to specify myogenic lineages (Davis et al., 1987), its expression is not strictly necessary for muscle development in a murine model, as *MyoD* null animals still develop normal muscle tissue due to functional redundancy with *Myf5,* another myogenic bHLH factor (Rudnicki et al., 1993). While no precedent literature demonstrates the phenotypic result of ASCL2 overexpression in unrelated cell lineages, ASCL2 overexpression in intestine leads to hyperplasia of crypt cells (van der Flier et al., 2009), and *Ascl2* null animals die at approximately E10.5 due to placental failure (Guillemot et al., 1994). While the trophoectodermal expression and function of *Ascl2* is the best characterized, *Ascl2* is also expressed in adults in a small subset of mesodermally-derived gut tissues (van der Flier et al., 2009; Yan et al., 2015), and in some leukocytes (Liu et al., 2014).

Despite the dramatic functional differences between these proteins, these factors are structurally similar within their bHLH domains, which has previously been shown to define function of these factors (Nakada et al., 2004) (Figure 1-1: Comparison of structure and sequence of ASCL1, ASCL2, and MYOD). While the structure of the family-defining bHLH domain was identified early (Murre et al., 1989), the structure and biophysical basis of their

DNA-binding ability was not fully understood until the bHLH domain of MyoD was solved through x-ray crystallography (Ma et al., 1994). This understanding has been further refined by crystallization of bHLH factors of this and other classes, including the class I E-proteins which serve as the canonical heterodimeric binding partners of class II bHLH factors (Ferre-d'Amare et al., 1993; Ellenberger et al., 1994; Shimizu et al., 1997; Nair et al., 2003; Sauve et al., 2004; Longo et al., 2008; Ahmadpour et al., 2012; El Omari et al., 2013).

While the structural aspects of ASCL1, ASCL2, and MYOD are similar within the bHLH region, important variations exist even within this domain. One difference is observed in the apparent difference in conserved residues of these proteins. Residues which are highly conserved across bHLH family members are found at medially oriented DNA-binding or bH1-H2 interaction interfaces of these proteins, while outward facing residues, which are conserved for specific bHLH factors across species, differ at these sites between family members, thus providing potential sites for bHLH-specific factor-cofactor interactions (Ma et al., 1994; Longo et al., 2008; Nakada et al., 2004). Additionally, the structure of the bHLH domain is itself distinguished between ASCL1/ASCL2 and MYOD, as the length of the basic helix-1 domain of ASCL factors is shortened by one helical turn, based on the position of the helix-terminating arginine residue at the N-terminus. This variation has previously been suggested to provide a mechanistic basis for differences in binding site selection for these factors (Soufi et al., 2014).

Outside the bHLH domain, these proteins show little sequence similarity, and possess unique amino and carboxy terminal sequences. These have previously been demonstrated to be significant in the capacity of bHLH factors as transcriptional activators. MYOD possesses

a 53 residue N-terminal transactivation domain (Davis et al., 1992; Weintraub et al., 1991), which can activate transcription independently of the bHLH domain (Weintraub et al., 1991). As a component of a fusion protein, this peptide can induce expression of nearby genes, and mediates interaction between MYOD and p300/CEBP (Sartorelli et al., 1997), a lysine acetyltransferase complex known to associate with enhancers (Visel et al., 2009). Notably, while ASCL1 lacks this domain, or a known equivalent, it was also found to recruit p300 to a subset of its binding sites *in vivo* (Martynoga et al., 2013; Anderson et al., 2014), suggesting that p300 may partially mediate its transactivational capacity.

Chromatin structure and the epigenetic landscape

Transcription factors are one component of the regulatory mechanism that provides lineage specification. However, their function is itself dependent on DNA binding in their developmental context, which is defined through lineage specification. This has become conventionally known as the *"epigenetic landscape"* (Waddington, 1957), and the study of such mechanisms is referred to as *epigenetics* (Waddington, 1942), referring to aspects of the genome outside the nucleotide sequence. While this term was coined before the mechanisms of epigenetic gene regulation were understood, more recent evidence has implicated specific differences in this epigenetic landscape as constituting the molecular basis of lineage specification in the form of specific chromatin features which confer additional regulatory capacity in the common genome shared by the distinct cells of the organism.

Early evidence came in the observation that gene expression was accompanied by changes in chromosomal structure. To accommodate the large quantity of DNA present in the nucleus of the cell, inactive DNA is maintained in a compact state, precisely wound around

octamers of histone proteins. This highly-ordered structure facilitates cellular processes of transcription and replication, and protects the DNA helix from damage. Using enzymatic DNAse treatment, it was demonstrated that actively transcribed genes were especially sensitive to nuclease activity (Weintraub & Groudine, 1976). It has since been observed that differences in patterns of nucleosome-depleted "open" chromatin are consistently observed between cell lineages (Stergachis et al., 2013). Compellingly, these patterns were found to be bestowed upon daughter cell populations, suggesting that variation in nucleosome occupancy might be central to lineage specification (Groudine & Weintraub, 1982). This variation, termed *chromatin accessibility,* provides a potential regulator of gene expression, and, crucially, a mediator of stable lineage specification. This is especially relevant for transcription factor binding and function, as nucleosomal DNA is theoretically less compatible with TF binding due to reduced exposure to the nuclear solvent. The structure of nucleosomes is highly ordered, and components of the histone octamer directly engage the DNA in both the major and minor grooves (Luger et al., 1997), presumably limiting access by DNA binding proteins. Indeed, this relationship between chromatin accessibility and transcription factor binding has long been observed (As summarized in Gross & Garrard, 1988). While chromatin accessibility was the first specific component of the epigenetic landscape to be characterized, a number of other epigenetic features of DNA have also been identified.

The structure of the nucleosome also serves to highlight another component of the epigenetic landscape; post-translational modifications of the histone subunits which comprise the nucleosomal octamer have been revealed as an additional substrate for gene regulation.

Genome-wide comparisons across disparate tissue types have revealed consistent changes between cell lineages, suggesting that a "histone code" is a second component of the epigenetic landscape which defines cell fate (Wang et al., 2012; Kundaje et al., 2012; *The Encode Project Consortium*, 2012). Both activating and repressive modifications have been identified, primarily in the form of methylation or acetylation, but also through addition of alternative functional groups to these histone proteins. These modifications are deposited through the activity of histone modifying enzymes, and have consistently been shown to play important roles in gene regulation.

While chromatin accessibility is likely to be one mechanism of transcriptional regulation, it is not the case that it necessarily prevents binding. One subclass of transcription factors, referred to as *pioneer factors,* has previously been shown to interact with nucleosomal DNA (Cirillo et al., 2002), suggesting that nucleosome occupancy is not necessarily incompatible with transcription factor function. Pioneer factors are defined by their ability to bind to closed chromatin, direct expression of their gene targets, and displace nucleosomes (as reviewed in Zaret & Carroll, 2011). The first observation of pioneering activity came from *in vivo* footprinting assays performed in the embryonic gut endoderm. This yielded the identification of TF binding at closed sites in the *Alb1* enhancer in liver buds, which were not occupied in other tissue types, suggesting that a mechanism other than DNA sequence was responsible for selection of binding sites within this narrow window (Gualdi et al., 1996). These factors proved to be members of the FoxA family, a winged helix factor family known to bind a forkhead motif (Clark et al., 1993), and the GATA family, which bind variations on the sequence *WGATAR* (Merika & Orkin, 1993). Members of both

families are expressed early in gut endoderm, and their binding at these sites precedes the expression of the target of this enhancer. FoxA proteins are of particular significance, as in addition to binding closed chromatin, they also mimic the structure of linker histones (Clark et al., 1993). These early discoveries have led to the understanding that the ability of transcription factors to bind nucleosomal chromatin is an important mechanism determining their binding and activity. Since then, pioneer activity has been identified for other transcription factor families, including bHLH factors. MYOD has been demonstrated to bind to closed chromatin specifically at the *Myogenin* enhancer (Gerber et al., 1997), and this binding relies on an N-terminal interaction with PBX3 (Berkes et al., 2004). This has been proposed as a central mechanism in the lineage-specifying capacity of MYOD, and distinguishes it from other myogenic factors, such as *Myogenin,* and *Mrf4,* which lack this domain, and cannot functionally replace MYOD and MYF5 in development (Rudnicki et al., 1993; Wang & Jaenisch, 1997). Importantly, pioneering capability is not a feature of all transcription factors, as significant differences in pioneering have been identified, even within a single family of transcription factors, such as bHLH factors (Soufi et al., 2012; Wapinski et al., 2013; Treutlein et al., 2016). The significance and limitations of pioneering capability are represented in the core transcription factors previously demonstrated to reprogram adult cells to so-called induced pluripotent cells (iPSCs). Of the four factors necessary for this conversion (OCT4, SOX2, KLF4, and MYC), only OCT4, SOX2, and KLF4 prove capable of binding to closed chromatin (Soufi et al., 2012), highlighting that the ability, or inability for transcription factors to function is dependent on the chromatin environment in which they function. This has previously been proposed as a mechanism for

progressive or cooperative gene regulation (Soufi et al., 2012; Soufi & Zaret, 2013; Soufi et al., 2015). Thus, differential pioneering ability represents a potential mechanism defining transcription factor function.

Cellular reprogramming

Experimental manipulation of differentiated cell populations has since revealed that lineage specification can be overcome under certain conditions. The first such approach utilized nuclear transfer, and demonstrated that pluripotency could be bestowed on differentiated cells with the addition of a pluripotent nucleus (Gurdon, 1962). The ability to reprogram intact cells was not identified until considerably later, and dramatic reversal of lineage specification to a pluripotent state has since been demonstrated (Takahashi & Yamanaka, 2010). Notably this remarkable conversion was achieved through expression of four transcription factors of notable significance in stem cell biology *Oct4, Sox2, Klf4,* and *Myc*. Thus, at least in response to artificial gene expression, lineage specification is not necessarily permanent. In addition to reversal of lineage specification to pluripotency, conversion between distinct cell lineages, often referred to as direct reprogramming or transdifferentiation, has also been demonstrated (Davis et al., 1987; Feng et al., 2008; Ieda et al., 2010; Vierbuchen et al., 2011). Perhaps unsurprisingly, lineage-specific bHLH transcription factors have been revealed as central mediators of lineage reprogramming. Expression of these master regulatory factors in differentiated cell types is sufficient to confer lineage-alternative cell fate with varying degrees of efficiency (Davis et al., 1987; Tapscott & Weintraub, 1988; Turner & Weintraub, 1994; Farah et al., 2000; Ieda et al., 2010; Vierbuchen et al., 2011; Wapinski et al., 2013).

As many class II bHLH factors are crucial to establishing their respective lineages, and in these lineages they bind to different sites throughout the genome, seemingly without dramatic differences in the primary motif bound, and initiate distinct transcriptional programs, it is tempting to believe that their function is limited primarily by the environment in which they are expressed. However, the observation that ectopic expression of ASCL1 or MYOD is sufficient to functionally reprogram a differentiated cell type such as fibroblasts to neurons or muscle indicates that environment alone is not sufficient to fully restrict these cells to obligate lineages. Indeed, the ability of these factors to enact disparate tissue specific gene expression programs cannot be attributed solely to differences in the chromatin environment, as ectopic expression of these factors in similar cellular contexts results in induction of their respective cell-type specific gene expression programs (Nakada et al., 2004; Nishiyama et al., 2009). Thus, the underlying mechanisms of transcription factor specificity in lineage specification remain unclear.

Introduction to DNA motif discovery

Historically, identification of regulatory regions bound by bHLH and other DNA-binding transcription factors has been accomplished through the use of Electrophoretic Mobility Shift Assays (EMSAs), which detect the presence of DNA binding proteins based on delayed movement through a gel medium (Garner et al., 1981). By pre-treating the DNA fragments to be tested with the protein suspected to have DNA binding capability, the assay can identify differences in mobility by visualizing the different DNA bands on the gel. By careful selection of the DNA regions tested, it is possible to identify sites which contain a

binding element for a given DNA-binding protein. Through iterative study, the minimal site required for a given transcription factor can be identified by making a large number of variant fragments, which progressively move across a region to be tested. While labor intensive, it was through studies such as this that many transcription factors and their cognate binding sites were identified. This approach also allowed for stepwise identification of transcription factor binding complexes, by the addition of multiple proteins in addition to the DNA test templates. These assays were frequently used in concert with in vitro reporter assays, wherein a previously identified binding region is cloned into a reporter line, where it is combined with a reporter to test the ability of a given enhancer region to direct expression of a fluorescent or biochemical reporter. Together, these techniques laid the foundation of functional testing of DNA regulatory regions in vitro and in vivo.

While in vitro binding and reporter assays were crucial to the early understanding of bHLH factors as a family, and to the identification of their preference for an E-box binding motif, they are limited in their ability to test hypotheses regarding the functional capacity of transcriptional regulators in vivo. They require *ab initio* selection of regions to be tested, and thus have an inherent bias in testing only regions previously predicted to bind or not bind, such as previously identified enhancers. They require extensive effort to construct the DNA template regions to be tested, to prepare the protein samples for investigation, and to optimize and complete the assays. Initially, this approach led to the development of Chromatin Immunoprecipitation assays (ChIP), which allows for the purification of specific complexes directly from cells or tissue, allowing for observation of specific protein-DNA binding events without the use of a reduced system (Gilmour & Lis, 1985). However, like

EMSA-based approaches, this strategy is limited by the need to identify specific sites for study. The advent of modern sequencing technologies has supplemented these in vitro techniques by allowing for rapid identification of DNA fragments of unknown composition.

Using an experimental strategy known as ChIP-seq, we can perform similar assays of DNA binding genome-wide in a single experiment, directly comparing the regions bound without the need to limit the experiments to predicted binding regions. This approach allows for direct observation of the revealed DNA sequence binding preferences for a transcription factor, and for bioinformatic inference of potential gene regulatory targets of the factors. In this study, I have used this approach to directly test the binding of three related bHLH factors in an engineered ES cell system, which, as described below, is used here as an environment lacking features of lineage-specific chromatin accessibility or differential expression of co-factors which may influence binding or function of our bHLH factors.

Prior to this study, thousands of bHLH binding sites have been identified for factors ASCL1, ASCL2, and MYOD in vivo using ChIP-seq (Castro et al., 2011; Cao et al., 2010; Borromeo et al., 2014; Schuijers et al., 2014). These sites include a number of previously validated binding sites demonstrated to show significant, focal enrichment for these bHLH factors (Borromeo et al., 2014; Cao et al., 2010), and to transcriptionally activate expression of nearby gene targets (Cao et al., 2006; Castro et al., 2006). From genome-wide binding data sets, it is also possible to computationally determine the revealed preference for DNA binding motifs for each factor, using previously developed analytics packages (Langmead et al., 2009; Bailey et al., 2009; Heinz et al., 2010). These algorithms utilize Markov chain analysis to impute statistical position-weight-matrices of nucleotide positions within the

regions enriched for DNA binding proteins, and can identify the presence of overrepresented binding motifs from these large genomic data sets. This analysis is essential to correctly identify the specific component sequences of binding sites, which are believed to provide the fundamental mechanism by which selective DNA-binding transcription factors select and regulate gene targets throughout the vast stretches of the genome.

Rationale for studies

ASCL1, ASCL2, and MYOD are tissue-specific class II bHLH factors, and are considered to be master regulators of cell fate. Recent genome-wide binding studies for these factors in their respective tissues demonstrate that they bind the same CAGSTG motif, but bind separate subsets of sites (Cao et al., 2006; Borromeo et al., 2014; Schuijers et al., 2014), suggesting that lineage specific differences in chromatin accessibility may play a role in determining where these factors bind, and therefore act. However, the ability of these factors to enact disparate tissue specific gene expression programs cannot be attributed solely to differences in chromatin accessibility, as ectopic expression of these factors in similar cellular contexts results in induction of gene expression programs resembling those of their respective cell-type (Nakada et al., 2004; Nishiyama et al., 2009; Fong et al., 2012*)*.

ES cells, which are among the earliest derived cells of the organism during embryogenesis, represent the last common stage in unspecialized cell division. Uniquely, ES cells demonstrate embryonic totipotency, the capacity to develop into any tissue in the embryo (Martinet al., 1981; Evans et al., 1981). As might be expected based on the ability of bHLH factors to establish their respective lineages, and to reprogram differentiated cell

types, the class II bHLH factors are essentially absent from the transcriptional profile of ES cells. As the factors discussed here function in partially lineage-restricted cell populations, studies performed *in vivo* in their respective cell types are limited to observations of their function in a partially established lineage. In my effort to decipher the fundamental mechanisms by which these factors specifically identify targets and regulate transcription across a complex genome, I utilized modified ES cells, representing a *tabula rasa* on which the activity of these bHLH factors can be compared, to minimize the potential confounding influence of developmental cues present in partially or fully defined lineages. These ES cells were engineered to inducibly express ASCL1, ASCL2 or MYOD (Nishiyama et al., 2009). Here I directly test differences in the genome-wide binding and transcriptional consequences of these master regulatory factors in the unspecified cell environment of embryonic stem cells to gain insight into the specific activity of these master regulators of cell fate.

Research Objective

The objective of these studies is to gain insight into the mechanism or mechanisms underlying the specificity of function of ASCL1, ASCL2, and MYOD in directing lineage-specific gene expression. Through direct manipulation of bHLH factor expression, I interrogate specific candidate mechanisms. These studies specifically focus on early events in bHLH factor function. This project is composed of two specific aims.

Specific Aim 1: Distinguishing mechanisms of binding and specificity for the bHLH factors ASCL1, ASCL2, and MYOD

The functional capacity of DNA-binding class II bHLH factors is dependent on their ability to recognize, bind, and activate transcription of specific, relevant gene targets across the genome. In vitro assays have previously shown that bHLH factors can activate transcription through specific enhancer regions, and *in vivo* binding comparisons have previously demonstrated that despite their defining bHLH domain, they bind to different sites throughout the genome (Borromeo et al., 2014; Cao et al., 2010; Schuijers et al., 2014) thereby activating specific transcriptional targets for expression. The exact mechanism by which they can recognize their cognate binding sites across the genome and specifically give rise to relevant transcriptional programs remains unclear. One possible explanation would be factor-specific preferences in their respective DNA binding motifs.

Class II bHLH factors have previously been demonstrated to selectively bind CANNTG Eboxes in vitro (Johnson et al., 1992; Murre et al., 1996), in reporter assays (Braun et al., 1989; Weintraub et al., 1991; Nakada et al., 2004), and in vivo (Borromeo et al., 2014; Meredith et al., 2014; Cao et al., 2010). From in vitro tiled DNA binding microarray data sets and in vivo ChIP-seq data sets, *de novo* analysis of genome-wide binding reveals distinct preferred binding motifs for a number of these class II bHLH factors. Previous studies of genome-wide binding have demonstrated that at least some of these factors have additional motif specificity either in the central dinucleotide positions of the Ebox (Cao et al., 2010; Borromeo et al., 2014; Meredith et al., 2013; Schuijers et al., 2014), or outside the canonical E-box (Castro et al., 2006; Beres, 2006), with additional identification of stringent spacing requirements for secondary co-factors identified for other class members (Meredith et al., 2013). In vivo ChIP-seq for ASCL1 and MYOD in neural

tissue and myoblasts, respectively, have demonstrated that despite variation in the protein structure of the basic region of the bHLH domain, where proteins of this class directly interact with DNA, they demonstrate a shared binding preference for CAGCTG binding motifs when tested in their respective lineages. However, the basis of this preference has not been fully explored. It may be the case that the preference for a GC-core E-box is partially due to the relative availability, or accessibility, of these binding sites as compared to alternative E-box motifs in these cellular contexts. If so, comparing binding of these factors in a common cell type, presumably with similar availability of binding sites, could reveal additional sequence preference by which these bHLH factors select distinct binding sites throughout the genome, and give rise to distinct transcriptional profiles.

While it is clear that some discrepancies in the amino acid sequence of the family-defining bHLH domain exist, structural studies of class II bHLH factors (Ma et al., 1994; Longo et al., 2008) demonstrate striking similarity in the revealed crystal structure of the DNA-interacting portion of these proteins. The largely unstructured domains of these proteins residing outside the bHLH domain, however, are less clearly characterized, especially for ASCL1 and ASCL2. While a number of interactions have been reported for these factors in their respective lineages, especially MyoD, no specific co-factor has been identified which explains how these factors select their distinct binding sites within a given cell type. Given the ability of ASCL1 to identify specific binding targets in the lineage-inappropriate environment of embryonic fibroblasts (Vierbuchen et al., 2010; Wapinski et al., 2013), it is possible that some previously unidentified co-factor may be partially responsible for directing these bHLH factors to the specific sites necessary for lineage-relevant function

in development, disease, and reprogramming. By investigating the binding of these factors in the common environment of the ES cell, we may be able to uncover additional motif specificity outside the primary binding motif, in the form of significant secondary co-factor binding motifs, supporting a model in which these bHLH factors have a limited ability to recognize and bind to specific sites in the genome, but additional specificity is conferred upon them in the presence of a relevant co-factor.

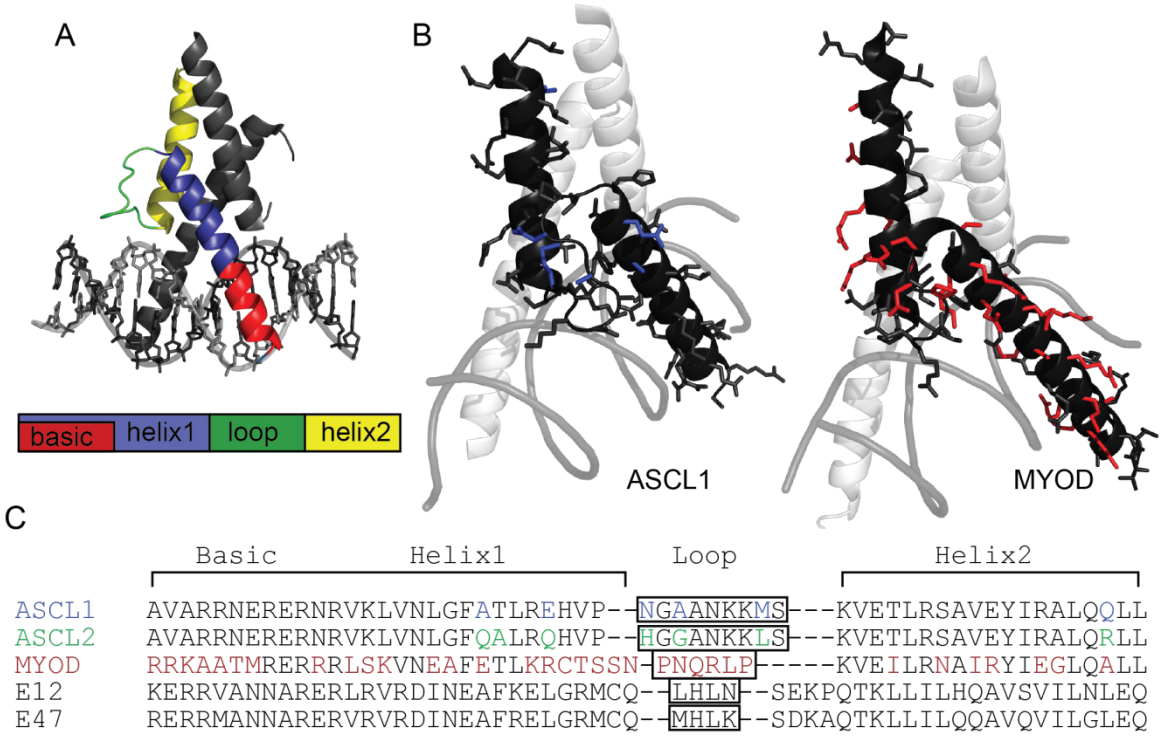
Specific Aim 2: bHLH binding and the chromatin landscape

ASCL1, ASCL2 and MYOD are able to induce expression of distinct transcriptional profiles in ES cells (Nishiyama et al., 2009); thus, these bHLH transcription factors appear capable of directing different transcriptional profiles from a common tissue source, with a common chromatin state. By comparing chromatin accessibility before and after induction of the TFs in inducible ES cell lines with TF-specific ChIP-Seq results for TF binding, I test whether bHLH binding is predicated upon, or informed by, the presence of open chromatin at potential binding sites. Additionally, this allows determination of bHLH-dependent changes at, and beyond, bHLH binding sites, testing whether these factors can induce a comprehensive program in cells without a previously established permissive chromatin state, as recently suggested for ASCL1 (Wapinski et al., 2013), and previously shown for MYOD (Weintraub & Groudine, 1976; Gerber et al., 1997; Bergstrom et al., 2002). By directly observing chromatin accessibility and transcription factor binding in the common context of the ES cells, I directly compare the pioneering ability of these class II bHLH factors.

One potential mechanism by which these bHLH factors may direct cell-type specific expression is the recruitment of epigenetic modification mechanisms to bHLH binding sites, thus modifying histone proteins to activate or repress gene targets. By comparing the distribution of histone modifications, we test whether interactions between the different bHLH factors and specific histone modifying enzymes might selectively target specific binding sites for epigenetic modification. Using ChIP-seq for H3K27ac, I test the influence of this marker of active enhancers on bHLH binding, and the effects of these bHLH factors on H3K27ac distribution genome-wide. Additionally, using these and other ChIP-seq data sets for histone modifications, I test whether specific histone signatures inform the binding of these factors in the context of ES cells.

Together, these studies provide insight into a fundamental question about how tissue specific class II bHLH factors function to specify distinct gene expression programs: Are binding differences dependent solely on chromatin state, or do factor-specific differences in interaction with DNA or co-factors enact discrete transcriptional programs? Additionally, these experiments provide insight into whether these bHLH factors are themselves able to induce changes in chromatin accessibility, and whether the capacity to do so differs between these bHLH transcription factors. Together, these experiments provide new insight into the mechanisms by which this essential class of developmental regulators enacts disparate transcriptional programs in cells.

Figure 1-1: Comparison of structure and sequence of ASCL1, ASCL2, and MYOD



ASCL1, ASCL2, and MYOD have similar bHLH domains.
(A) structural rendering of class II bHLH factor ASCL1 modeled in complex with E47 and DNA helix (PDB structure 2ql2 from Longo et al., 2008). Colored regions represent structural domains as indicated, basic domain is included in Helix1.
(B) structural rendering of mouse ASCL1 (left) and MYOD (right)(ASCL1 modeled on NEUROD:E47, PDB 2ql2, Longo et al., 2008; MYOD PDB structure 1mdy from Ma et al., 1994). Colored residues are factor-specific between ASCL1, ASCL2, and MYOD, black residues are shared between these factors.
(C) Aligned protein coding sequences for mouse ASCL1, ASCL2, and MYOD shown. Domains annotated as previously described for these factors (Ma et al., 1994; Nakada et al., 2004; Longo et al., 2008). Colored residues are unique to the factor within this comparison. Loop regions depicted in boxes, residues are contiguious with adjacent helices. bHLH coding sequence of class I E-proteins E12 and E47 also shown for comparison.

CHAPTER TWO

Methods

The Inducible ES Cell System

Overview of ES cells used in experiments

These studies were performed in three inducible murine ES cell (mESC) lines derived from 129S6/SvEvTac (Simpson et al., 1997; Olson et al., 2003), and express *Ascl1*, *Ascl2*, or *MyoD,* under the control of a tetracycline-repressive promoter system (Gossen & Bujard, 1992; Nishiyama et al., 2009). Full-length cDNAs of *Ascl1, Ascl2,* or *MyoD* (Carter et al., 2005; Sharov et al., 2003) were cloned into a transgenic construct which expresses the bHLH factor as a bHLH-His6-FLAG fusion, and an internal ribosome entry site allows for expression of the fluorescent Venus reporter from the same transcript. This construct was targeted to the *ROSA* locus (*R26R)* (Masui et al., 2005)*,* providing a mechanism to regulate expression of the bHLH transgene at the transcriptional level, without the use of viral or chemical vectors. (Figure 2-1: Schematic diagram of transgenic construct used to generate ES cells).

Culture of ES cells

Mouse ES cells were cultured in a variant of Dulbecco's Modified Eagle's minimal Medium (DMEM), termed *ESLX*, which was formulated based on previously described

media conditions (Nishiyama et al., 2009; Coriell Institute for Medical Research, 2005, 2014). Culture media was further optimized with input and gracious assistance from Robin Gilmore and Mylinh Nguyen of the UT Southwestern Transgenic Center. ES culture media was formulated to contain Dulbecco's Modified Eagle's Medium (Millipore, SLM-120-B), 20% v/v Fetal Bovine Serum (Gemini, 100-525), 0.1mM β-mercaptoethanol (Millipore, ES-007-E), 1.93mg/mL L-glutamine (Fisher Scientific, BP379-100), 1% v/v penicillin/streptomycin (Gibco, 15070-063), 100 μM non-essential amino acids (Millipore, TMS-001-C), 1% v/v nucleosides (Millipore, ES-008D), 1mM sodium pyruvate (Sigma, P5280-25G), 1000U/mL Leukemia Inhibitory Factor (Gemini, 400-495), 1 μg/mL puromycin (1μg/mL), and .2μg/mL (Sigma, D9891) (Table 2-2: Table of cell culture medium components). This formulation was found to support robust growth, and induction of expression of the bHLH transgene.

ES cells are low-adherence cell types, but will grow readily in a monolayer on gelatinized culture vessels when co-cultured with adherent murine embryonic fibroblast feeder cells (SNLP). Mitomycin-C treated SNLP cells were plated at ~$1.0*10^6$ cells per 10cm plate, and cultures were allowed to grow for at least 24 hours prior to plating of ES cells. All cell culture was performed in a laminar flow hood using aseptic technique. Cells were maintained in a water-filled 37C incubator with 5% CO_2, and passaged at 48h after plating to new feeder cultures. After initial recovery from frozen stocks, ES cells were plated at a density of ~$1.0x10^7$ cells to 10cm plates in 5-10mL of media, and maintained in culture at below 80% confluence to avoid spontaneous differentiation. ES cell media was changed at least every 24 hours, and cells were passaged every 48 hours. Cell pluripotency was assessed

by staining formaldehyde-fixed cells with alkaline phosphatase as per the provided protocol (Millipore, SCR004). All cell lines tested showed strong positive alkaline phosphatase staining, indicating that these cells are effectively maintained in a pluripotent state at passage numbers beyond those used for experiments.

Preparation of Experimental Samples

Induction of ES cells

ES cells were grown on SNLP feeder cells for expansion prior to experiments. As murine SNLP cells are also present in these cultures, they represent a potential source of experimental bias. To reduce the influence of these cells in our genomic studies, ES cells were passaged at equivalent density to gelatinized plates without feeder cells for the last two passages prior to induction. As mitomycin-C treated cells are non-proliferative, this is sufficient to effectively remove these cells from culture prior to harvest for experiments. Prior to induction, cells were plated to 6-well plates, 10cm plates, or 150cm^2 plates at ~1.5*10^7 cells/10cm plate, or similar density for 6-well and 150cm^2 plates, in ESLX media Immediately before induction, all cells were observed under microscope, and observed to contain phase-bright, rounded colonies, as previously described for proliferative ES cell cultures.

To induce expression of their respective transgenes, doxycycline must be removed from the culture media. This is accomplished by serial washes and replacement of the media with an induction specific media, which is identical except for the absence of doxycycline.

Three rounds of washes, using 37C Ca(-),Mg(-) phosphate buffered saline (PBS), with three hour delays between these washes, proved effective for inducing robust expression of the VENUS reporter. To minimize the potential effect of the additional media changes in induced cells, uninduced control cells were removed from the incubator, and the media is replaced with fresh ESLX media. At the last induction round, induced and uninduced cells were passaged to new plates without feeder cultures. Induction of these cells is performed to maintain similar culture conditions, and reduce the potential for procedural bias (such as media deprivation, or distinct culture timelines).

Harvest of ES cells for RNA purification

RNA was purified from ES cells which were cultured and induced as previously described. RNA preparation was performed in parallel with chromatin preparation, during the incubation period allotted for fixation. Cells were observed for VENUS fluorescence to confirm induction of transgene prior to harvest. Prior to sample collection, laminar flow hood, instruments, benchtop surfaces, and centrifuge interiors were treated with RNAseAway (Fisher Scientific, 10328011) to remove potential contaminants. 10cm plates containing samples of induced and uninduced control ES cells (~$1.0*10^7$ – $2.0*10^7$ cells) were removed from 37C incubator, and washed once in 15mL ice-cold PBS. The PBS was decanted away, and 1 mL of RNA lysis buffer (Zymo, R1054) was immediately added to each plate. Disposable nuclease-free cell lifters were used to detach cells from the plate surface. Lysates were transferred to 1.5mL microfuge tubes via pipette. Samples were stored at -80C as

stabilized lysates, and were purified for RT-qPCR analysis and sequencing. RNA purification was performed using a small volume column elution as per the *Zymo Research* provided protocol, including 15 minute DNAse I treatment to remove residual trace DNA prior to column elution (Zymo Research, R1054). All samples were eluted into nuclease-free water, and quantified using a NanoDrop benchtop spectrophotometer (Thermo Scientific, ND1000 or ND2000).

Based on the result of spectrophotometric analysis, 5ug samples of purified RNA in nuclease-free water were prepared for potential sequencing, and stored at -80C pending analysis of sample quality. Transgene expression was evaluated by reverse-transcription quantitative polymerase chain reaction analysis (RT-qPCR). cDNA preparation was performed from 1µg purified RNA using Invitrogen SuperScript III (Invitrogen, 18080-044), using a Bio-Rad C1000 thermocycler. Expression of *Ascl1, Ascl2* and *MyoD* transcripts was evaluated using primers directed against endogenous and transgenic transcripts for these factors (sequences for these primers can be found in the appendix – Primers). All RT comparisons validated to be RNA-transcript specific by comparison of non-reverse-transcribed control reactions. Quality of RNA samples was assessed by bioanalysis, and all samples used for sequencing demonstrated RNA Integrity Number (RIN) ≥ 9. 5 µg of purified RNA was submitted to the UT Southwestern Microarray Core facility for sequencing library preparation, and single-end 50bp sequencing on an Illumina HiSeq 2500 line. Analysis of the resulting data sets is described in a later section.

Chromatin immunoprecipitation for bHLH proteins

In brief, fixed, frozen whole cells prepared as described were transferred to conical tubes, washed briefly in 5mL ice-cold modified RIPA buffer solution (50mM HEPES-KOH, 140mM NaCl, 1mM EDTA, 10% v/v glycerol, 0.5% v/v Nonidet P-40 substitute IGEPAL 630, 0.25% v/v Triton X-100). Cells were then centrifuged in a refrigerated centrifuge for 5 minutes at 400 x g, and supernatant was discarded. Pellets were then resuspended in 5mL of a slightly basic ice-cold saline solution (1mM Tris pH 8.0, 200mM sodium chloride, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0), and centrifugation repeated. Supernatant was then removed with care, and nuclear pellets suspended in 275 µL of ice-cold lysis buffer (10mM Tris-HCl pH 8.0, 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% w/v sodium deoxycholate, 0.5% w/v N-laurylsarcosine), transferred to siliconized low-adhesion microfuge tubes, and incubated on ice for 15 minutes.

Samples were then sonicated using an ice-chilled Diagenode Bioruptor Standard Sonicator, for a total of 35 minutes, using a 50:50 cycle, in 7 bouts of 5 minutes each. Bath temperature was regulated by regular replacement of ice water. Samples were then diluted in chromatin immunoprecipitation buffer (20mM Tris-HCl pH 8.0, 150mM NaCl, 0.1% Triton X-100, and 2mM EDTA), and centrifuged 30 minutes at max speed (~30,000 x g) to remove cellular debris. The clear supernatant was transferred to new siliconized microfuge tubes, and incubated overnight with the antibodies described below. Each reaction was performed using 5µg of mouse anti-MASH1 (BD Pharmingen 556604) for ASCL1 ChIP or mouse anti-FLAG (Sigma F1804) antibody for ASCL2 and MYOD ChIP. Antibody/lysate were added to new tubes contain 25ug Protein G Dynabeads (Life Technologies 10003D) and incubated for 4-6 hours at 4C on a benchtop rotator to immunoprecipitate bound fragments.

Samples were washed on a benchtop rotator in a series of 4 minute washes in 1 mL volumes, in ice cold solutions. First, washed once using a low-salt buffer (20mM Tris-HCl pH 8.0, 150mM NaCl, 2mM EDTA, 0.1% w/v SDS, 1% v/v Triton X-100), and once in a high-salt buffer (20mM Tris-HCl pH 8.0, 400mM NaCl, 2mM EDTA, 0.1% w/v SDS, 1% v/v Triton X-100). 5 washes were performed in lithium chloride Wash buffer (250mM LiCl, 1% v/v NP-40 substitute, 1% w/v sodium deoxycholate, 1mM EDTA, 10mM Tris-HCl pH 8.0). Beads were washed once in Tris-EDTA to remove trace detergents. Elution was performed at 70C in a heated robotic shaker, using two sequential elutions with heated buffer (10mM Tris-HCl pH 8.0, 1% w/v lithium dodecyl sulfate, and 1mM EDTA). Samples were treated with Proteinase K solution (11uL 5M NaCl, 5uL Proteinase K 10mg/mL) for 4 hours shaking at 55C, and incubated overnight at 65C in a heated robotic shaker to reverse crosslinks. ChIP and input samples were then purified using Qiagen QIAquick miniature affinity purification columns, and stored in the provided elution buffer. Samples were evaluated by qPCR and quantified using the Qubit DNA high sensitivity kit, and duplicates combined prior to library preparation.

One distinction of the approach used here is that chromatin normalization is calculated based on cell number at harvest and fractional input, rather than quantification. This approach was chosen based on the results observed in optimization of the ChIP protocol; comparison of input samples demonstrated that buffer components prevent accurate observation of chromatin concentration by spectrophotometry, as conventionally used to quantify chromatin in ChIP protocols. Due to the cross-linking of protein and DNA, this cannot be readily overcome by affinity column purification or phenol-chloroform extraction

due to the long delay introduced by reversal of cross-linking. The approach used here instead relies on extrapolation of quantified input control, which more precisely reflects the amount of chromatin template present within the sample.

Chromatin immunoprecipitation for acetylated H3K27

ChIP purification for H3K27ac was performed using the same protocol as for bHLH factors. Importantly, unlike ChIP for bHLH factors, ChIP for H3K27ac utilized 0.1% sodium dodecyl sulfate (SDS). ChIP purification was performed on fixed, flash frozen aliquots of whole cells prepared as described for ES cell harvests. ChIPs for histone markers were performed from aliquots of $1.0*10^7$ cells. Each reaction was performed using 5ug of anti-H3K27ac antibody (Abcam, ab4729). ChIP was performed similarly as for bHLH factors; importantly, unlike ChIP for bHLH factors, ChIP for H3K27ac utilized 0.1% sodium dodecyl sulfate (SDS) in the lysis buffer to facilitate fragmentation. 25ug Protein G Dynabeads (Life Technologies, 10003D) were used to immunoprecipitate bound fragments. Elution was performed at 70C in a heated robotic shaker using a solution of lithium dodecyl sulfate. Samples were treated with Proteinase K solution, and incubated overnight at 65C in a heated robotic shaker to reverse crosslinks. ChIP and input samples were then purified using Qiagen QIAquick miniature affinity purification columns. Samples were evaluated, and duplicates combined prior to library preparation.

Preparation of ChIP sequencing libraries

Illumina DNA sequencing library preparation was performed as per NEBNext ChIP-Seq Library Preparation protocol using Illumina-compatible multiplexing primers. Libraries were generated using 2-4ng of ChIP purified chromatin template from 24 hour induced ES cells (*tTA-Ascl1*) and 24 hour induced and uninduced control samples (*tTA-Ascl2,* and *tTA-MyoD*), as well as purified 10ng input controls from the same samples. Library amplification was performed using multiplexing primer pairs (New England Biolabs, E7735S). Size selection was performed using Ampure XP bead purification (Figure 2-3: Example Agilent DNA Bioanalysis result). The resulting libraries were sequenced by the UT Southwestern Microarray facility, using single-end 50bp sequencing, on an Illumina HiSeq 2500. The resulting reads were demultiplexed and aligned to the mouse *mm10* genome (GRCm38), using Bowtie2 (Langmead et al., 2009). Peak calling, intersection, annotation, and motif analysis were performed using HOMER v4.7 (Heinz et al., 2010). Peak to gene calling was performed using HOMER v4.7 (Heinz et al., 2010) and GREAT (McLean et al., 2010). Discussion of these methods is discussed later in this chapter.

Assay for Transposase-Accessible Chromatin (ATAC-seq)

ATAC-seq from ES cells was performed as per the previously published protocol outlined in *Buenrostro et al., 2013:* and *Buenrostro et al., 2015*. Cells were harvested at 24h post-induction by dissociation with warm trypsin-EDTA, quenched by the addition of ice-cold serum-containing media with (control cells), or without doxycycline (induced cells), and diluted in 4C PBS. Cells were then counted by serial dilution by haemocytometer. 50,000 cells from 24 hour induced and uninduced cultures of each cell line were isolated and were

used in the preparation of these libraries. Traditional polymerase chain reaction amplification (PCR) was utilized for amplification of transposed elements in a Bio-Rad C1000 thermocycler, using the programs specified in *Buenrostro et al., 2015.* Multiplexed single-end 50bp sequencing was performed using an Illumina HiSeq 2500, using Nextera-compatible amplification primers. Due to differences in multiplexing primers, backwards-compatible Nextera sequencing primers were used to sequence these samples. Outputs of this sequencing were demultiplexed using sample-specific Nextera primer sequences, and fastQC (Andrews 2010) was used to filter and score the resulting sequencing runs. This sequencing provided high read depth and complexity, as expected for these samples. The sequencing results of each sample were aligned to the mouse mm10 genome (Kent et al., 2002; Kent et al., 2010), and processed for downstream analysis using Bowtie2 (Langmead et al., 2009), and HOMER (Heinz et al., 2010), as described in the accompanying text.

Bioinformatics and computational analysis

The study presented here makes extensive use of a number of previously developed open source computational algorithms and software packages, which are introduced here in brief. The value of these resources in completing this study cannot be overstated. In particular, *Hypergeometric Optimization of Motif EnRichment* (HOMER), a comprehensive genomic analytics package (Heinz et al., 2010) has been extensively utilized.

Data handling and software used for analysis

Genome-wide DNA sequencing data (ChIP-seq and ATAC-seq) from multiple lanes were demultiplexed using sample-specific Illumina primer sequences. The resulting data sets were aligned to the mm10 genome using Bowtie2 v2.2.6 (Langmead et al., 2009). Reads with a Bowtie2 quality score less than 10 were removed using SAMtools v1.3 (Li et al., 2009) with parameters (*-bh -F 0x04 -q 10*). Duplicate reads were removed using picardtools v1.119, and the remaining reads were normalized to 10M reads using HOMER v4.7 (Heinz et al., 2010). All UCSC Genome Browser plots shown (Kent et al., 2002; Kent et al., 2010; Raney et al., 2014; Rosenbloom et al., 2015) reflect these normalized tag counts.

Sequencing of RNA samples was aligned to the mouse mm10 genome using TopHat 2.1.0 (Langmead et al., 2009; Trapnell et al., 2009). Default settings were used, with the exception of *–G*, specifying assembly to the mm10 genome, *--library-type fr -first strand*, and *–no-novel-juncs*, which disregards noncanonical splice junctions when defining alignments. *edgeR* (Robinson et al., 2010; Nikolayeva et al., 2014) was used to incorporate RNA-seq data from three biological replicates for each factor tested, and identify genes which were differentially expressed between samples, using the default parameters. Experimental details of gene expression analysis are further discussed in Chapter 3.

Use of previously published genomic data sets

In addition to the experiments described here, the results of these analyses make use of a number of previously available data sets. These sets were downloaded from the public *Gene Expression Omnibus* (GEO) from the accession numbers provided in the works cited, and processed for analysis using the same approach as our ChIP-seq data sets (as described

below) for unbiased comparison. In each instance, the original source of these data is indicated in the text accompanying its use.

Identification of bHLH binding sites from sequencing data (peak calling)

The sequencing data sets generated from ChIP-seq from the inducible ES cells were used to identify putative binding sites (ChIP-enriched peak regions) genome-wide based on the distribution of the aligned reads. ChIP-seq data sets for transcription factors were normalized to 10M reads. Peaks for each sample were called based on respective input control samples created during immunoprecipitation. This approach addresses potential bias from variation in immunoprecipitated fragment length and sequence bias in sequencing library preparation. Peak calling was performed using a sequencing-depth independent approach to correct for variation between data sets used; the *findPeaks* library of HOMER 4.7 (Heinz et al., 2010) was used to call significantly enriched peak regions from each data set. Significance is evaluated as a FDR ≤0.0010 (0.1%). Parameters used specify for selection of focal peak regions (*-factor*), which uses the autocorrelated predicted fragment length (derived from aligned read distribution) to identify changes, modeled on a Poisson distribution. These peaks are subjected to local filtering of fourfold compared to the surrounding interval (*-L 4*), and discards regions not meeting significance by Poisson p-value threshold of ≤1.00e-04. Sequencing data sets from ASCL2 and MYOD data sets were subjected to additional filtering based on the uninduced control sample from the same experiments. This additional filtering was added to address the presence of a small number of non-specific peak regions (as described in *Chapter 3*) in an unbiased manner. Peak calling

was performed comparing ASCL2 and MYOD 24h induced data sets to their respective input controls, as described for ASCL1. Peak calling was then performed using ASCL2 and MYOD induced samples versus an uninduced control sample at lower stringency (*-L 2*), and the resulting intervals annotated for the presence of a canonical REST/NRSF motif (*JASPAR MA0138.2*) (Mathelier et al., 2016). This was observed in ChIP-seq from both induced, and uninduced ES cells of *tTA-Ascl2-FLAG*, and *tTA-MyoD-FLAG* cell lines, but not in *tTA-Ascl1-FLAG*, and thus nonspecific. Regions which showed the presence of this motif within 100bp of the empirically determined peak center were removed from the ASCL2 and MYOD peak lists.

Identification of potential regulatory targets by peak-to-gene association

To identify potential regulatory targets of these factors, peak-to-gene calling was performed, which observes the location of transcriptional start sites (TSS) present in genomic intervals surrounding the putative binding sites identified from ChIP-seq data. HOMER 4.7 (Heinz et al., 2010) and *Gene Region Enrichment Association Tool* (GREAT) v3.0 (McLean et al., 2010) were used to perform this analysis in two distinct ways, addressing two distinct aims. Peak-to-gene calling HOMER utilizes a straightforward approach to gene calling, comparing the distance from the peak to the RefSeq-curated catalog of transcription start sites (TSS) of nearby genes, and selecting the closest TSS as the gene identified. HOMER also includes specific non-genic features present in the RefSeq catalog, such as miRNAs, ncRNAs, and pseudogenes (Pruitt et al., 2014), and identifies a single feature for each genomic position. This allows for objective comparison of distances to nearby start sites, and

identifies non-genic features which represent potential targets. However, as class II bHLH transcription factors are known to function primarily at distal enhancer regions, this approach is expected to underestimate the number and identity of potential regulatory targets.

To address this, GREAT v3.0 (McLean et al., 2010) was used to perform peak-to-gene calling when surveying potential gene targets. GREAT observes the location of multiple TSS for each putative regulatory region, and reports the genes identified, based on the canonical isoform for each gene, and reports gene ontology associations (GO) from the list of genes identified. For all analysis performed in these studies, the association rules are 5kb 5' of the TSS, 1kb3' of the TSS, and an extension to the next gene of up to 1mb. Thus, this approach can associate a single peak with multiple genes, and a single gene with multiple peaks. In practice, this increases the total number of genes associated with each putative binding site, allowing for association with multiple genes. While neither approach identifies every potential regulatory target of a given binding site, GREAT generally identifies more genes associated with binding sites identified in ChIP, generating larger lists of genes for further comparisons (Figure 2-4: HOMER vs. GREAT genes called). HOMER identifies a single genomic feature, and reports non-genic features as well as RefSeq genes. The algorithm used in each analysis is noted in the accompanying text.

De novo motif discovery

This study makes considerable use of *de novo* motif discovery to observe the revealed preference for specific DNA sequence of bHLH factors; HOMER v4.7 was used for all motif discovery presented here. To compare bHLH factor binding sites, a narrow observational

window of 50bp centered on the peak apex identified by HOMER was used. This narrow window was selected specifically to avoid the influence of adjacent regions when identifying the primary motif bound by these factors. To screen for potential DNA-binding co-factor motifs, a broader window of 150bp centered on the peak apex was used to identify enriched motifs adjacent to the primary binding site. The interval used for each analysis is noted in the accompanying text. Parameters *–S 10 –bits* were used unless otherwise noted. The statistical comparisons shown reflect the significance identified by HOMER for the specific *de novo* motifs identified from ChIP-seq. The *de novo* motifs shown for each analysis represent the most significantly enriched motif identified from the specified set of intervals. In instances where non-specific or low-information motifs are identified, the next best match is shown. Statistics reflect the motif shown in each instance. For genome-wide motif discovery, the binomial distribution is used for statistical comparisons, whereas for promoter motifs, the hypergeometric distribution is used to observe significance as compared to promoter-specific background regions.

Scatterplot, histogram, and heatmap generation from genome-wide sequencing data

To visualize genomic density of sequencing reads, HOMER's *annotatePeaks.pl* library was utilized to generate incidence matrices from the normalized sequencing reads generated for each data set (Heinz et al., 2010). The resulting matrices were sorted using either the Linux command line or *Microsoft Excel*, based on criteria described in the text accompanying each analysis. Heatmap plots were created using *Java TreeView 1.6* (Saldanha et al., 2004), and *MatLab v. R2016b*®. Histogram representations of sequencing reads and

motif distribution at binding sites were generated in HOMER (Heinz et al., 2010), and were plotted using *GraphPad Prism® 7.0*. Scatterplot comparisons were created using HOMER (Heinz et al., 2010), and plotted in *RStudio* (RStudio Team, 2015) using *ggPlot2* (Wickham, 2009; Wickham, 2016).

Figure 2-1: Schematic diagram of transgenic construct used to generate ES cells

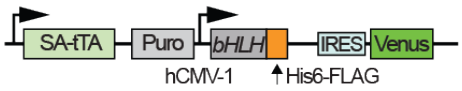


Diagram of transgenic construct used in generation of inducible ES cell lines. Constructs differ only by the bHLH cDNA inserted into the targeting vector, and reflects the bHLH factor specified for each cell line. SA-tTA represents tetracycline repressible promoter system described in *Gossen & Bujard, 1992*. Diagram adapted from *Nishiyama et al., 2009*.

Figure 2-2: Diagram of induction strategy and analysis

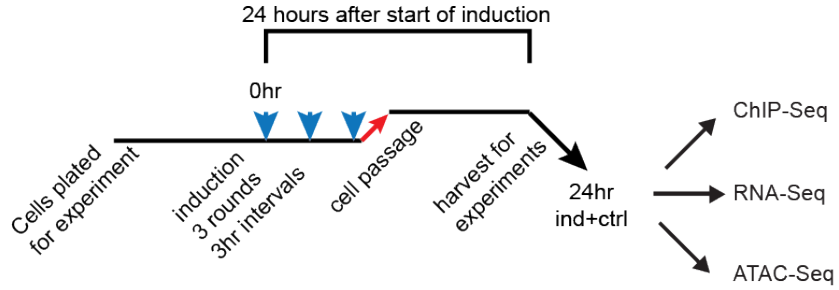
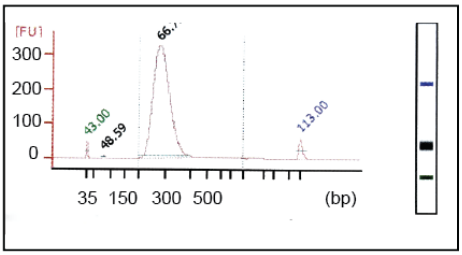


Diagram of induction strategy used to generate experimental sample, as described in text. Blue arrows denote induction rounds. Red arrow denotes passage to new plates. Induction time is considered as time since initiation of the first media change to doxycycline-free media.

Table 2-1: Table of cell culture medium components

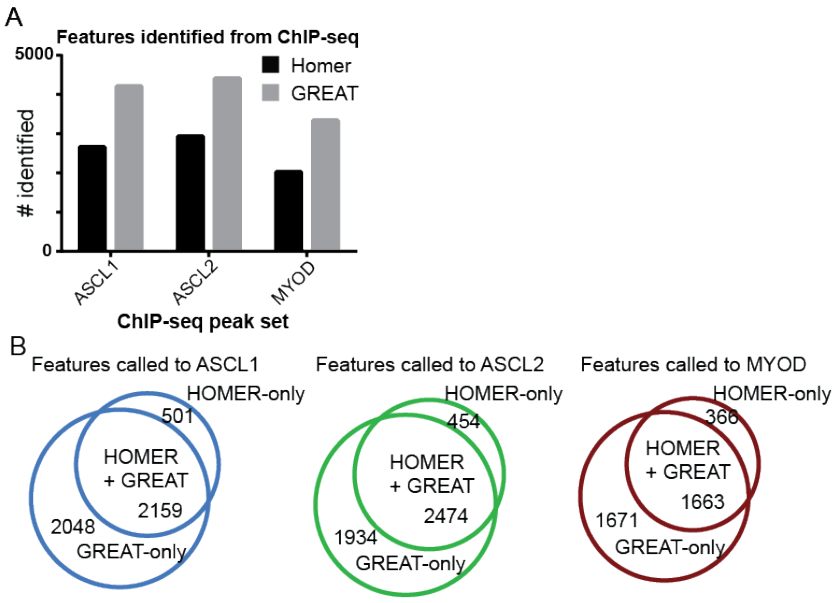
Embryonic Stem Cell Culture Medium (ESLX)		
Reagent stock	Supplier/Catalog	final concentration
Dulbecco's Modified Eagle's Medium	Millipore SLM-120-B	75% v/v
Stem Cell Grade Fetal Bovine Serum	Gemini 100-525	20% v/v
b-Mercaptoethanol	Millipore ES-007-E	.1mM
L-Glutamine 200mM	Fisher Scientific BP379-100	1.93mg/mL
Penicillin/Streptomycin	Gibco 15070-063	1% v/v
Non-Essential Amino Acids	Millipore TMS-001-C	100 uM
Nucleosides	Millipore ES-008-D	1% v/v
Sodium Pyruvate	Sigma P5280-25G	1mM
Leukemia Inhibitory Factor (LIF)	Gemini 400-495	1000U/mL
Puromycin	Sigma P8833	1ug/mL
Doxycycline	Sigma D9891	.2ug/mL

Figure 2-3: Example bioanalyzer result from prepared ChIP-seq library



Sequencing libraries generated from bHLH ChIP purified samples show consistent fragment length and high yield. Agilent high sensitivity bioanalyzer result from ChIP library submitted for sequencing.Plot shows result from ASCL1 ChIP library, and is representative of other ChIP libraries sequenced. 2-4 ng of purified ChIP product was used in creation of each library. X-axis represents size distribution of amplified fragments in sample.

Figure 2-4: HOMER vs GREAT features called



(A) Comparison of numbers of RefSeq features identified from ChIP-seq binding sites by HOMER and GREAT, using default peak association parameters for each algorithm. Graph shows the total number of entries identified by each method from the total set of ChIP-seq peaks identified for each bHLH factor.
(B) Proportional overlap of RefSeq features called by HOMER and GREAT using same parameters as shown in (A). Values reflect the numbers of features present in each subset, from the total set of peaks identified for each bHLH factor, as indicated.

CHAPTER THREE

CLASS II BHLH FACTORS MAINTAIN DISTINCT BINDING GENOME-WIDE WHEN EXPRESSED IN ES CELLS

Introduction

Embryonic development is a complex process, during which the cells of the early organism divide, specialize, mature, and integrate into distinct populations and networks to form a complex organism. This process gives rise to distinct populations of cells, each with dramatically different properties, despite sharing an identical DNA blueprint. This process requires precise regulation of transcriptional programs within the cell, which are partially established by transcription factors. The power of transcription factors to specifically, and selectively regulate gene targets within the complex genome is central to their role in development. Class II bHLH factors ASCL1, ASCL2, and MYOD have specific roles within development, establish discrete embryonic and extraembryonic lineages, and selectively regulate ordered expression of gene targets within their respective lineages. As a class, they form heterodimers with relevant E-proteins, bind to CANNTG Eboxes, and activate expression of specific targets, but the mechanism underlying this specificity remains unclear. One challenge in understanding these mechanisms is the discrete pattern of expression of these factors in the developing embryo. As they primarily function after lineage specification has begun, untangling the influence of the distinct cell lineages of these respective tissues on bHLH function is especially challenging.

Expression of these factors begins relatively early in embryogenesis, with Ascl1 expression noted in neuroectoderm by E8.5 (Guillemot et al., 1993), Ascl2 noted in extraembryonic tissue precursors at E3.5 (Rossant et al., 1998) and transiently in the embryo at E7.5 (Guillemot et al., 1994), and MyoD expression noted by E10.5 in mesoderm (Chen et al., 2002). The mouse embryo is divided into largely discrete tissues and organ structures by E9.5, and these lineages remain distinct throughout the life of the animal. Compared to differentiated tissues, the distinct lineages that these factors function within are generally poorly characterized, in part due to the inherent challenge of studying a transient population *in vivo* during development. Additionally, these lineages are themselves made up of heterologous cell populations, which are successively created in a spatiotemporally restricted manner. Furthermore, while advances have been made, the direct mechanisms regulating the expression of these bHLH factors have not been fully characterized. While the genome itself is theoretically identical in every cell, the function of these bHLH factors may be influenced by other events which establish these populations, thus affecting the mechanism by which these transcription factors select appropriate binding sites and regulate expression of appropriate transcriptional targets.

One such possibility would be differences in binding site accessibility, which might restrict the binding of these factors to specific subsets of Eboxes within the genome in discrete lineages. This could be established by the presence or absence of lineage-specific transcription factors which may directly compete with class II bHLH factors for binding sites, or interact with these bHLH factors to repress their DNA binding capability. Alternately the presence or absence of lineage-specific co-factors may facilitate their binding

to DNA at lineage-appropriate gene targets when expressed in their respective lineages. Another possibility would be factor-specific differences in binding motifs masked by these or other factors present in differentiated lineages. Were these mechanisms responsible for the distinct patterns of binding and expression, we would anticipate changes in the binding profiles of these factors when expressed outside their specific endogenously defined lineages.

Alternately, the apparent functional specificity of these proteins may lie beyond their ability to bind DNA motifs. In such a model, differences in binding observed *in vivo* (Cao et al., 2010; Castro et al., 2011; Borromeo et al., 2014; Schuijers et al., 2015) would not be central to the regulation of direct and downstream gene targets of these factors. Indeed, a number of studies have previously demonstrated that functional specificity is not clearly attributed to the basic DNA-binding interface, but relies on multiple domains, which function in a semi combinatorial fashion to initiate transcriptional activation of gene targets (Chien et al., 1996; Nakada et al., 2004). This might suggest a mechanism in which the majority of binding sites are bound by multiple factors, either transiently or constitutively, in their developmental contexts. In such a model, differences in the transcriptional complement present in these lineages, including lineage-specific transcriptional activator or repressors, might then interact with these factors to directly regulate specific targets, thus allowing for factor-specific gene regulation with diminished dependence on factor-specific differences in binding. Such a model may not be evident in studies performed *in vivo* from partially or fully specified tissues, as co-factor expression may also be context-dependent, and identification of the bound sites would presumably require observation in the specific context in which they interact. Relatively few co-factors (Mao et al., 1996; Black et al., 1998; Mal et al., 2001) for

these class II bHLH factors have been identified outside the class-defining interaction with E-proteins, and expression levels of bHLH factors and relevant co-factors in their endogenous contexts remain largely uncharacterized.

Thus, while it is clear that ASCL1, ASCL2, and MYOD demonstrate functional specificity, the complexity of studying the mechanisms underlying this specificity in different cell contexts has prevented mechanisms involved in the specification to be uncovered. In these studies, I utilized the common cellular environment of ES cells to selectively express each factor, testing potential mechanisms underlying the powerful, lineage-directive function of these master regulators. In this system, which minimizes the potential influence of the disparate cellular environments in which these bHLH factors endogenously function, I test potential models which may underlay the distinct effects of these TFs. I utilize ChIP-seq to directly observe genome-wide binding of ASCL1, ASCL2, and MYOD, and RNA-seq to characterize the transcriptional changes initiated by these bHLH factors in a common biological context. Additionally, these data are compared to identify potential direct regulatory targets of bHLH factors, some of which may represent novel downstream mediators of lineage specification. In doing so, I address fundamental questions about how these bHLH factors give rise to transcriptionally diverse cell lineages, and present evidence that these factors maintain considerable specificity even when ectopically expressed in ES cells.

Results

Induction of ASCL1, ASCL2 or MYOD in designer embryonic stem cells

Inducible transgenic ES cell lines featuring a "dox-off" *tTA-bHLH-FLAG-IRES-VENUS* construct (Nishiyama et al., 2009) were used for these studies. Designer ES cells with inducible ASCL1, ASCL2 or MYOD were obtained and characterized for induction properties. The cells were cultured without SNL-P feeders in LIF-containing ES cell culture medium, and induced by removing doxycycline from the culture media. While the uninduced ES cells have essentially no ASCL1, ASCL2, or MYOD, robust expression of each bHLH factor was detected within 12 hours with mRNA levels approaching maximum by 24 hours (Figure 3-1: Inducible ES cells demonstrate robust expression of bHLH factors within 24 hours). As the goal of this study is to identify early binding events in bHLH function that explain the specificity of these factors, I chose the 24 hour time point for my experiments.

The engineered ES cells were designed to express a bHLH-FLAG fusion to facilitate analysis, immunocytochemistry, western blot analysis, and ChIP. However, although by sequence the ASCL1 cells demonstrate each of the appropriate components of the transgenic construct, the FLAG moiety is undetectable in multiple assays. Thus, for the following ChIP-seq studies, mouse monoclonal *α-Ascl1* (BD Pharmingen, 556604) was used to detect ASCL1, while an *α-FLAG M2* (Sigma-Aldrich F1804) was used to detect ASCL2 and MYOD in these cells.

ASCL1, ASCL2, and MYOD bind largely distinct sites within the mouse ESC genome

To directly compare the binding and function of these three related factors in a common context, I identified the sites bound by the bHLH factors genome-wide by ChIP-seq

in ASCL1, ASCL2, and MYOD-expressing ES cells at 24 hours after the onset of induction. Sequence reads were mapped to the mouse mm10 genome using Bowtie2 (Kent et al., 2002; Kent et al., 2010; Raney et al., 2014; Rosenbloom et al., 2015; Langmead et al., 2009). Peak calling, overlap analysis, and motif analysis were performed using HOMER v4.7 (Heinz et al., 2010). 3188 ASCL1, 3504 ASCL2, and 2385 MYOD peaks were called in their respective cell lines. Comparison of these tracks on the UCSC Genome Browser from the ChIP-seq data to their respective input controls reveals primarily narrow, focal peaks with peak morphology resembling a normal distribution for the read lengths used in this analysis, similar to previously sequenced and validated ChIP-seq data sets.

Confidence in the quality of these data sets is illustrated by observing known developmental targets of ASCL1, ASCL2, and MYOD (Nelson et al., 2009; Castro et al., 2006; Malone et al., 2011). Sharp, highly enriched peaks near *Dll1, Dll3, and Hes6* are seen (Figure 3-2: ASCL1, ASCL2, and MYOD bind at distinct and shared sites near key developmental genes). ASCL1, ASCL2, and MYOD are each enriched at a site approximately 1 kb upstream of the TSS of *Hes6 (Hairy-and-Enhancer-of-Split 6),* a bHLH factor which acts as a transcriptional repressor (Gao et al., 2001), and a known target for class II bHLH regulation (Bae et al., 2000). These factors were also found at two sites proximal to *Dll1 (Delta Drosophila-like 1)*, a canonical NOTCH ligand, and crucial mediator of lateral inhibition in development. All three factors were enriched at a *Dll1* intronic peak, while ASCL1 and ASCL2 but not MYOD was found at the upstream peak of *Dll1*, an initial indication that MYOD may bind to distinct targets when compared to the ASCL factors. Similarly, a site within the 5' promoter of *Dll3 (Delta Drosophila-like 3)* was bound by ASCL1 and ASCL2,

but not by MYOD. Thus, these ChIP-seq data are of good quality, are in agreement with previous identification of functionally validated regulatory sites near a subset of developmental genes, and differences in binding of the different factors are evident.

Primary among the questions addressed in this research is whether the lineage-specific bHLH factors maintain distinct binding when presented with a common chromatin environment. To perform this comparison, *mergePeaks*, a HOMER module (Heinz et al., 2010), was used to identify coincident peaks across the bHLH ChIP-seq data sets for the 3188 ASCL1, 3504 ASCL2, and 2385 MYOD peaks within 150bp. This analysis identified 625 peaks which were shared across all three factors, representing 18-27% of the peaks identified for each bHLH factor (Figure 3-3: Proportional overlap diagram of binding sites identified in ChIP-seq). Importantly, the majority of peaks identified for each bHLH were found to be factor-specific, being identified in only one of the three bHLH ChIP-seq data sets. Thus, specificity in binding of bHLH factors in neural versus muscle lineages (for example) is not simply due to context dependent chromatin accessibility or tissue-specific co-factors.

Comparison of this overlap also demonstrates considerably greater overlap of ASCL1 and ASCL2, as compared to MYOD. ASCL1 and ASCL2 share 40-45% of their sites, whereas MYOD shares only 25% with ASCL1 and ASCL2. As the bHLH domain of ASCL1 and ASCL2 are more structurally similar, especially in the DNA-interaction domain (Figure 1-1: Comparison of structure and sequence of ASCL1, ASCL2, and MYOD), their increased similarity may reflect greater shared specificity for specific DNA features.

The observation that bHLH factors maintain distinct binding in ES cells is crucial to understanding the role of bHLH factors in development and disease, as well as in reprogramming. Had we identified largely overlapping patterns of binding in these experiments, this would suggest that bHLH binding, and presumably function, was primarily a consequence of the chromatin environment in which these bHLH factors act. This might suggest that the primary mechanism by which cells regulate the targets of these bHLH factors would be the establishment of a permissive binding environment, which would then lead to changes in ASCL1 target gene expression. The clear distinction in genome-wide binding seen in the common chromatin environment of the ES cell demonstrates that this is clearly not the case. Rather, each bHLH factor has an intrinsic activity to recognize specific sites within the genome. Whether this is due to specific DNA recognition properties of the bHLH, or recruitment of specific co-factors is addressed below through analysis of the factor specific bound chromatin.

bHLH factors primarily bind distal enhancer regions, with similar preferences for genic features

Another characteristic of the genomic binding properties of a transcription factor is where they bind relative to coding sequences. This approach allows one to test whether transcription factors demonstrate similar or different preferences in binding site proximity, such as a preference for transcription start sites, 5' promoter regions, or distal enhancer regions. Here I utilized two algorithms to identify genes associated with the bHLH binding sites identified by ChIP-seq, HOMER (Heinz et al., 2010) and GREAT (McLean et al.,

2010). The distinction between these approaches is described in Chapter 2: Methods. I first utilized HOMER to characterize bHLH bound sites with respect to genic features, and found that each bHLH tested demonstrates similar preference for distal enhancer regions (Figure 3-4: ASCL1, ASCL2, and MYOD have similar binding distribution relative to gene features). This is evident by the large number of binding sites annotated as intergenic and intronic sites as compared to promoter, UTR, or exon sites, and by the similar distribution of absolute distance to the nearest TSS in the mouse genome. This comparison shows that ASCL1, ASCL2, and MYOD binding sites are generally comparable, binding relatively few sites within ±1kb (2^{10} bp) of the TSS. ASCL2 binding demonstrates moderately increased preference for proximal binding sites, and ASCL1 and MYOD demonstrate virtually identical distribution profiles. However, for all three factors tested, the majority of bound sites are identified as distal enhancer elements. This analysis was also performed using GREAT (McLean et al., 2010), which supported this result. Together, these data show comparable preference for binding distal enhancer regions for each bHLH factor, indicating that specificity in bHLH binding is not dramatically influenced by proximity to genic features.

ASCL1, ASCL2, and MYOD demonstrate largely similar preferences in binding motif

Class II bHLH transcription factors are defined by their shared bHLH domain, and their ability to form heterodimeric complexes with ubiquitously expressed E-proteins, such as E12 or E47, and bind to DNA. They bind to DNA at short sequences known as E-boxes, so

named for their identification by association with E-proteins (Church et al., 1985; Ephrussi et al., 1985; Murre et al., 1989; Johnson et al., 1992). Eboxes are defined by a relatively common (probabilistically 1 occurring every 256 bases) 6-nucleotide CANNTG motif. In vitro, class II bHLH factors and E-proteins will promiscuously bind these sequences, irrespective of the composition of the central dinucleotide. However, using either array based or genome-wide ChIP-seq data, the preference of ASCL1, ASCL2, and MYOD in their respective tissues was revealed through *de novo* motif analysis. When compared in differentiated cell types, these factors have been shown to preferentially bind an E-box with a GC-core dinucleotide (Borromeo et al., 2014; Castro et al., 2011; Liu et al., 2014; Cao et al., 2010) (Figure 3-5: Comparison of *de novo* binding motifs identified for ASCL1, ASCL2, and MYOD in ES cells and differentiated cell types).

To determine if the motif preference is altered when binding of these bHLH factors is observed outside of their normal context, I performed *de novo* motif discovery from the binding regions identified in ChIP-seq for each factor, using a 50bp interval surrounding the center of the binding sites (center ±25bp) (using HOMERs *findMotifsGenome* module). There was a clear, dramatic enrichment for Ebox motifs, present in the majority of peak regions that is well above the expected occurrence throughout the genome (Figures 3-5, 3-6: Primary Ebox motifs identified in bHLH ChIP-seq). Comparison of the preferred *de novo* motifs identified for each of the three factors demonstrates a shared preference for a CASSTG Ebox. While there is some variation in the degeneracy of this motif, especially at the second position of the central dinucleotide, this motif appears to be the most strongly enriched for each of the bHLH factors. In addition, I performed a similar comparison on the

factor-specific and shared binding sites, and identified only modest variability in the degeneracy of the central dinucleotide residues specific for each subset. This modest variability could indicate some factor-specific preference in motif but it is difficult to determine the relevance here.

To further test the conclusion that these motifs represent the primary binding site identified within the ChIP-seq data sets, and not an artifact of the discovery methodology, I also generated position weight matrices for each possible permutation of the CANNTG Ebox, and directly annotated the frequency of these Eboxes at binding sites identified the ChIP-seq data (Figure 3-7: Ebox Distribution at bHLH ChIP-seq peaks). This comparison demonstrated that both GC-dinucleotide, and GG-dinucleotide Eboxes were strongly enriched at the peak center, in roughly equal measure, indicating that these factors are not strongly selective between these CAGSTG Eboxes. Enrichment of CAGATG Eboxes, was also identified, but at dramatically lower frequency than GC and CC motifs. Together, this suggests that these factors share a common preference for CAGSTG Eboxes, but do not clearly discriminate between G/C in this position.

To test whether the specific enrichment for these motifs was primarily a reflection of the relative genomic frequency of the 10 possible Ebox permutations, I also compared the relative incidence of these features across the genome (Figure 3-8: Genome-wide distribution of E-boxes). This comparison revealed that genome-wide, the relative frequency of these Ebox features is variable, and each permutation differs from its expected frequency throughout the genome. Compared to the expected incidence for each permutation, the identification of GC core Eboxes is especially striking. While these Eboxes are similarly

enriched in ChIP-seq peaks, CAGCTG Eboxes represent only 7.05% of the total set of genomic Eboxes, and therefore more significantly enriched versus the genomic background. The ability to bind less common genomic features may be one mechanism by which these factors establish sufficient regulatory complexity underlying their binding specificity.

De novo motif discovery identifies preferences for distinct flanking sequence of primary Eboxes

Indeed, these specific extended Ebox motifs are apparent near sites associated with lineage-specific functional relevance. For example, MYOD localizes within an intron of *Myocyte-specific-enhancer-factor 2-D* (*Mef2d*), a muscle specific target, in a peak containing three separate ACAGSTG sites within 50bp of the peak center (Figure 3-10: MYOD binding sites within *Mef2D* locus). As ASCL1 and ASCL2 are nearly identical in their basic-H1 domain, and have notable differences when compared to MYOD in this region, the shared preference for GCAGSTG by ASCL factors, and ACAGSTG in MYOD suggests that these preferences may be due to specific differences in the structure of the basic-H1 domain junction in these proteins. Based on the previously solved structure of class II bHLH dimeric binding complexes (Ma et al., 1994; Ellenberger et al., 1994; Ferre-D'Amare et al., 1993), the identification of additional preferential binding specificity at the periphery of the Ebox motif is likely due to structural variation occurring N-terminally to the ‘basic’ region. This region is believed responsible for selection of the primary binding site, generally defined as R111 to T115 in MYOD (Weintraub et al; 1991; Brennan et al., 1991). In these crystal structures, the N-terminal end of the basic-Helix1 domain extends laterally along, and nearly central within,

the major groove of the DNA helix, thus 5' expansion in motif preference based on bH1-DNA interactions would likely occur N-terminally to the region responsible for binding the Ebox motif (Ma et al., 1994). Alternatively, DNA interactions by the loop region, or N-cap of the H2 domain might play a role in establishing this preference, as seen for NeuroD1-E47 heterodimers (Longo et al., 2008), or by interactions with relevant co-factors (Figure 3-11 : potential bHLH:DNA interaction sites). While the role of these structural elements is less clear, their proximity to the DNA helix presents a potential secondary interaction site, which might be less stringent than differences in the bH1 binding domain, potentially allowing for tuning of motif preference. As ASCL1 and ASCL2 share considerable protein sequence in the Helix1-loop-Helix2 region, and together are more distinct from MYOD (see Figure 1-1 comparison of bHLH domain structure), it is reasonable to attribute this distinct specificity to this region of the protein. However, it is of critical importance to note that while these motifs represent the most significantly enriched E-boxes for each factor, ASCL1, ASCL2, and MYOD binding sites contain both forms of the variant motifs identified. Thus, while the apparent disparity in preference may inform the binding of these factors, this difference alone is not sufficient to restrict the binding of these factors to the preferred flanking variant. These possibilities are especially intriguing in light of the finding that the variant flanking motif expansion identified confers directionality on the otherwise palindromic Eboxes bound, suggesting that bHLH factors might be able to distinctly regulate expression of nearby targets by differential recruitment of transcriptional complex components to one side of the Ebox.

Secondary motifs identify specific co-factor families that are unlikely to explain the specificity of bHLH binding

Analysis of the primary Ebox binding motifs identified only modest differences in the primary binding motifs identified for each factor, and these differences failed to support a model in which these factors are differentially recruited to their distinct binding sites genome-wide. To test whether additional specificity might be conferred by co-factors, I compared the secondary motifs identified from *de novo* motif analysis. Additional enriched motifs from the bHLH peak regions represent a potential secondary mechanism for differential recruitment of bHLH factors to distinct gene targets. In their normal roles in development, co-factors may help direct and stabilize the bHLH factors at particular binding sites. I repeated *de novo* motif analysis on a 150bp interval surrounding the center of the peak regions identified in ChIP-seq for the three bHLH factors. This broader interval was selected to improve the ability to identify secondary binding motifs, which, being indirectly enriched by their proximity to primary binding motifs, are generally located farther from the center of the enriched regions. This analysis yielded a relatively small number of secondary motifs that were identified as enriched for any given factor.

In preliminary analysis of ASCL2 and MYOD-expressing cells, a striking enrichment was noted for a REST/NRSF motif which, as detailed below, turned out to be an artifact of the FLAG antibody used for the ChIP (Figure 3-12: Comparison of REST/NRSF motif). The long and highly stereotyped nature of this motif was of sufficient significance to initially identify REST not simply as a secondary motif, but as the primary motif for the peaks shared ASCL2 and MYOD binding, rather than the expected Ebox motif. Expression data from the

ES cells demonstrated that REST/NRSF is expressed at considerable levels. As REST, along with mSin3a and CoREST, is known to be a crucial element of the Polycomb repressor complex, and has a defined role in maintaining the repression of neural targets in non-neural tissues, the identification of this motif in ASCL2 and MYOD ChIP-seq data presented an exciting candidate for a mechanism by which these transcriptional activators might function to repress neural gene expression in their respective lineages. However, after many experiments to confirm the validity of this finding failed to support an interaction of REST/NRSF with ASCL2 and MYOD, I tested the idea that it was an artifact based on the FLAG antibodies used that were not used in the ASCL1 ChIP-seq. To test this possibility, I sequenced the uninduced control samples from the original ChIP-seq experiment for ASCL2 and MYOD. This identified strong, focal enrichment of a number of REST/NRSF binding sites in uninduced control samples, revealing a highly specific non-specific binding interaction with the FLAG antibody. As a result, all analysis of ASCL2 and MYOD ChIP-seq data utilizes multi-way corrected peak sets, which selectively remove nonspecific REST/NRSF peaks from our data. (See detailed methodology of this correction in Methods)

Comparison of secondary motifs identified in *de novo* motif analysis revealed a number of interesting candidates for potential co-factors (Figure 3-13: Comparison of significant secondary motifs identified in ChIP-seq). In ASCL1-expressing cells, there were a number of secondary motifs which were enriched considerably above background (20-100 fold), with significant p-values (1e-20 to 1e-60), including Sox, Osr, Meis, and a motif identified as ZNF354C. Some, such as Osr and ZNF354, appear to represent considerably longer motifs than the known motif to which it is matched, suggesting that they may

represent binding sites of these or other factors, in complex with a second DNA-binding entity. ASCL2 also demonstrates enrichment of Sox, Osr (identified as ESC-Nanog), and ZNF354C. MYOD identified considerably fewer motifs, generally of lesser statistical significance, including Pbx3, and a rather degenerate Tcf3 (Ebox) motif. While several of the secondary motifs identified have strong PWMs, and represent intriguing targets for potential regulatory interactions, the numbers of sites present containing these motifs make it clear that while potentially meaningful, they do not numerically represent a viable mechanism to explain the dramatic level of factor-specific binding observed. Thus, the mechanism underlying the specificity of binding for the bHLH factors is not evident from secondary motif identification. Nevertheless, Osr in particular remains an intriguing potential co-factor for ASCL1 and ASCL2 (see discussion).

Distinct gene expression programs are induced by ASCL1, ASCL2, and MYOD within 24 hours

In vivo, and in vitro, ASCL1, ASCL2, and MYOD have previously been shown to directly regulate gene expression of downstream targets. Given the dramatic role of ASCL1 and MYOD in reprogramming terminally differentiated cell types to neurons and muscle cells, respectively, it is clear that they demonstrate some functional specificity, and activate distinct targets. Direct comparison of targets of these TFs when tested in a common cellular context has not been performed. Using ChIP-seq and RNA-seq data, I identify potential mediators by which these bHLH factors function to specify their respective lineages, and identify common

or factor-specific mechanisms by which ASCL1, ASCL2, and MYOD may regulate their expression. Analysis of the genome-wide binding profiles of ASCL1, ASCL2, and MYOD clearly demonstrate that these factors have distinct binding properties when ectopically expressed in the environment of the ES cell. To get insight into whether these binding sites are relevant to the lineage-specific functions of these bHLH factors, I used Gene Ontology (GO) analysis to compare the genes associated with the bHLH binding sites. I first performed GO analysis directly from the binding sites identified for each bHLH factor in ChIP-seq. While there are many methods and tools which can be used to perform GO analysis, all rely on statistical comparisons of the list of genes in a given set to a catalogue of defined pathways, biological processes, or molecular functions. The probability of a list of genes containing members of a defined set are used to rank the sets which are significantly enriched, suggesting that the input list may have similar functional relevance. I used GREAT (McLean et al., 2010) to identify peak associated genes, and performed GO analysis directly on these lists. GREAT was selected as it identifies potential regulatory targets based on their distance and orientation to nearby genes, and allows for association with multiple genes meeting these criteria. The full sets of peak regions identified in ChIP-seq from each cell line, as well as the lists of peak regions identified as either shared, or factor-specific, were submitted using the same parameters as in the previous analysis (5kb upstream, 1kb downstream, or 1,000kb beyond these basal regions).

GO analysis of genes associated with each set of bHLH binding sites

GO analysis of the total sets of ASCL1, ASCL2, and MYOD bound sites identifies enrichment for a broad range of biological process categories (Figure 3-14: GREAT Analysis of full bHLH GO categories). Significantly enriched categories include a number of entries with clear developmental relevance, including a number of processes involved in stem cell maintenance, embryonic patterning, or tissue differentiation. This suggests that genes proximal to sites preferentially bound at this early time point are involved both in stem cell processes and in developmental processes, rather than gene targets central to mature cell lineages, or profiles unrelated to development. Some categories identified appear specifically relevant for the lineages which these factors serve to establish, such as compartment pattern specification and hindbrain morphogenesis in ASCL1, and heart morphogenesis in MYOD-expressing cell lines. However, these categories were not exclusive to their respective cell lines, and a considerable number of GO biological process categories are identified for lineage-inappropriate ontological categories.

I also performed analysis on the shared and factor-specific subsets of peaks from ChIP-seq (Figure 3-15: GREAT analysis of shared and factor-specific gene ontology categories). As the bHLH factors under study here are crucial for establishing lineage specific gene expression patterns, factor-specific binding sites are likely the most informative in identifying key mediators of lineage specification downstream of each bHLH factor. However, the overall gene ontology profiles did not demonstrate dramatic, factor or lineage-specific category enrichment for any of these factors. Interestingly, some of the lineage-relevant GO categories identified for the full sets as being factor-specific were not enriched from the factor-specific sets of peaks, which suggests that the gene sets that make up these

lineage-relevant GO categories are split between shared and factor-specific genes. As factor-specific comparisons eliminate these shared binding sites, the components present in the factor-specific binding sites fall below the threshold of significance required to identify enrichment of a given GO term. This serves to illustrate that GO analysis is highly dependent on well-annotated gene networks for the identification of functional categories, and that these functional categories are often made up of a combination of both highly specific functional genes and less specific pathway components. Highlighting this point, perhaps the most intriguing result from this GO analysis is from the shared gene targets. The results included a number of highly lineage-relevant categories, including glial cell development, neural tube formation and closure, placenta development, and heart morphogenesis. As such, while each bHLH factor is known to be crucial for establishing their respective lineages and regulating distinct transcriptional profiles, the most lineage-specific gene ontology categories identified from ChIP-seq binding sites are identified from the 625 shared bHLH binding sites, rather than the 1000+ factor-specific binding sites for our ES cell lines. Thus, while these factors show modest association with lineage-specific genes at 24h post-induction, dramatic, factor-specific enrichment for lineage-specific GO categories is not found.

Sites with central ACAGSTG or GCAGSTG motifs do not enrich for lineage-relevant GO terms

As ASCL1/2 and MYOD appear to demonstrate distinct preferences for an Ebox motif featuring differences in its flanking motif, I also tested whether the sites featuring these variant motifs might be preferentially associated with lineage-specific genes. I utilized the

full, and factor-specific lists of bHLH binding sites for each factor, and annotated peaks containing an ACAGSTG (for MYOD binding sites), or GCAGSTG Ebox (for ASCL1/2 binding sites) within a 50bp window centered on the peak center. The lists of peaks were filtered for each factor containing the variant Ebox motif identified for the given cell line. These peaks were then submitted for GO analysis using GREAT. Neither full nor factor-specific sets of binding sites containing these variant flanking motifs identified increased association with factor-specific targets. Broadly, the regions demonstrated less lineage-relevant association with GO categories than the full, unfiltered lists of peaks discussed previously. As only a few, marginally significant GO categories were identified, the lists of genes associated with these peaks were also submitted to the Mouse Consensus Path Database (CPDB) for a more comprehensive GO screen. Each list provided identified enrichment for a number of GO categories relevant to developmental pathways, but failed to demonstrate specific enrichment for lineage-relevant categories when compared to the other factors tested. From this analysis, it does not appear that the presence of these specific motifs near the peak centers of bHLH binding sites is predictive of lineage-specific gene function.

RNA-seq analysis demonstrates differential expression of distinct genes in response to ASCL1, ASCL2, or MYOD expression within 24 hours.

As binding sites for TFs, particularly the bHLH class studied here, are found at large distances from the genes they regulate, it is often difficult to unambiguously identify the regulated genes. To directly observe changes in the transcriptional profile of these cells,

single-end 50 bp RNA-seq was performed from three separate experiments. Each experiment included cells induced to express ASCL1, ASCL2, or MYOD for 24 hours, with matching uninduced controls. Using edgeR, a statistical analytics package which identifies significant changes from replicate RNA-seq, a few hundred genes were identified in response to each bHLH factor tested. Using a statistical cutoff of FDR≤0.05, edgeR identified 234 genes in response to ASCL1 demonstrating a significant increase or decrease, 326 in ASCL2, and 608 in MYOD. These lists include a number of genes which, while promising, are expressed at very low levels, or demonstrate a consistent, but low fold change upon bHLH induction. For the majority of this analysis, a two-fold increase or decrease, and an average RPKM (*Reads Per Kilobase transcript per Million mapped reads*) of at least 1 in either the induced or control sample was used. To demonstrate the trend seen between samples, RPKM represents the mean RPKM between the three biological replicates, and fold change is calculated from these values. Using these parameters, 116 genes were increased with ASCL1, 170 with ASCL2 and 315 with MYOD. (Figure 3-16: Comparison of genes showing significant differential expression at 24h).

As expected based on RT-qPCR analysis, ASCL1, ASCL2, and MYOD demonstrate high levels of differential expression, showing negligible expression in the uninduced cells, and high levels of expression in 24h induced samples. (Table 1-1: Table overview of *Ascl1, Ascl2, and MyoD* expression as derived from RNA-seq data). The list of positively regulated DEGs includes known targets of these bHLH factors, including genes previously identified in the ChIP-seq data, such as *Dll1, Dll3, Hes6*, *Paprp12*, *Lfng*, and *Fgf5* (Table 1-2: Comparison of expression of known bHLH targets). The majority of DEGs showed increased

expression upon induction, rather than decreased, which supports the characterization of these and other class II bHLH factors as transcriptional activators.

ASCL1 and ASCL2 identify considerably fewer positively differentially expressed genes than MYOD (116 and 170, versus 315, respectively). Furthermore, expression levels of 199 genes are uniquely changed in MYOD-expressing cells, whereas only 18 or 50 genes demonstrate significant factor-specific increase in response to ASCL1 or ASCL2 expression within 24 hours. As ~70 genes demonstrate shared increase of expression in response to all three of these transcription factors, the discrepancy in the number of genes which are differentially expressed suggests that *MyoD* may inherently possess a greater capacity for early transcriptional activation of gene targets, despite being induced to a comparable level of mRNA expression by 24h. This is particularly remarkable, as the MYOD ChIP-seq data identified the fewest number of ChIP-seq binding sites, implying that the nearly doubled number of genes which demonstrate differential expression may be regulated through the activity of MYOD at a considerably smaller number of regulatory enhancers (ChIP-seq peak regions) as compared to ASCL1 and ASCL2.

Together, the RNA-seq data reveals dramatically more genes expressed in response to MYOD expression than ASCL1 or ASCL2 expression at 24h. To test whether this difference in the number of DEGs identified for MYOD-expressing cells was due to the specifics of our analysis, or was suggestive of a more robust ability to activate gene expression, I compared several approaches to defining DEGs. I first utilized different thresholds for statistical, fold change, and RPKM filtering to see whether adjustments to these thresholds might demonstrate more comparable numbers of genes identified as differentially expressed in each

ES cell line (Figure 3-17: Comparisons of DEG Criteria). While dramatic differences in the total numbers of genes identified are apparent in these comparisons, the relative overlap between these ES cell lines remains virtually identical. This suggests that the pattern of gene expression changes seen across the cell lines is not primarily due to differences in the strategy or thresholds used to select the lists of candidates. Thus, it is not the case that these bHLH factors merely implement the same profile of gene expression with subtle distinctions in expression levels, but appears that the distinct patterns of expression noted represent veritable differences in the early transcriptional profile established by ASCL1, ASCL2, and MYOD when expressed in ES cells. As MYOD is known to contain a transactivation domain N-terminally to the bHLH domain (Weintraub et al., 1991), and this domain is known to be sufficient to improve transactivation and reprogramming capability when fused to non-bHLH transcription factors (Hirai et al., 2011), MYOD may simply possess more potent regulatory capacity than ASCL1 and ASCL2.

Taken in aggregate, these data demonstrate that both shared and factor-specific transcriptional changes occur within 24h of induction, and further demonstrate that these related bHLH proteins have intrinsic abilities to induce distinct transcriptional profiles in a common context. The relatively limited number of genes identified supports our prediction that the changes seen at 24h represent early events in the transcriptional programs established by these bHLH factors. As ASCL1 and MYOD have been thoroughly demonstrated to have lineage-reprogramming capability, it is clear that they are able to direct the dramatic changes in gene expression necessary to overcome lineage restriction. As such, I conclude that the

relatively small number of factor-specific changes in expression, and the considerable overlap between bHLH factors, is due to the early time point studied.

Direct comparison of the full lists of DEGs for each bHLH factors reveals a striking number of genes involved in developmental pathways. Of particular significance are genes involved in signaling pathways central to embryonic development, such as the NOTCH, WNT, BMP, and FGF pathways (Figure 3-19: Overview of selected developmental pathways identified from shared targets, Appendix 3-1:Table overview of genes identified in developmental pathways). This suggests that bHLH factors may initially induce expression of a common regulatory core of powerful developmental regulators, rather than primarily directing factor-specific programs with lineage specific activity.

The lists of genes demonstrating factor-specific expression within 24 hours reveal a small number of genes with ASCL1 (18 genes) and ASCL2 (50 genes)-dependent patterns of induction. ASCL1-expressing cells identify intriguing candidates suggestive of neural function, such as *Gcm1* (*Glial-cells-missing*), which has been linked to a neural lineage deficit phenotype, as well as *Itpr3*, and *Homer2*, both of which are identified as having functional roles at glutamatergic synapses. Additionally, induction of ASCL1, and to a lesser degree ASCL2, was also associated with increased expression of *Dll3,* which is found in differentiating cells throughout the neural tube (Henke et al., 2009), whereas MYOD induction led to a subtle decrease in its expression (Figure 3-20: bHLH-dependent activation of *Dll3* in induced ES cells at 24h). *MYOD*-specific genes include many genes with known function in differentiated muscle cell types. *MyoD*-specific DEG included *Actc1,* and *Des,* markers of early skeletal muscle differentiation which are known targets of MYOD (Minty et

al., 1986; Allen et al., 1991) as well as *Tnni2, and Tnnc1*, which are skeletal and cardiac troponin family members. Although MYOD is known to induce muscle specific gene expression, induction of these genes at this early time point demonstrates that MYOD is able to utilize the existing transcriptional machinery in a cell type which is not primed for its activity by the effects of lineage specification. Together, these data reveal significant changes in gene expression by 24 hours post-induction, demonstrating that these bHLH factors are able to enact changes in the transcriptional profile of ES cells within this early period. However, the numbers of genes regulated in response to only one factor are relatively low in ASCL1 and ASCL2, versus MYOD-expressing cell lines, suggesting a discrepancy in the ability of these factors to directly regulate factor-specific targets when ectopically expressed. Furthermore, while some lineage-relevant genes are differentially expressed, no clear path to lineage specification is revealed by these early, factor-specific targets.

Identification of potential direct targets of bHLH factors at 24 hours post-induction does not reveal obvious regulators of lineage specification

To focus our analysis on genes which represent likely candidates in bHLH mediated lineage specification, I identified DEGs from RNA-seq which were associated with bHLH binding sites identified in ChIP-seq. As most binding sites are located in intergenic regions, comparing them based on the nearest single gene does not fully address their potential regulatory targets, which may occur at great distances. Additionally, some transcription factors regulate expression of multiple targets from a single binding site, and intronic binding sites may regulate genes other than the one in which they reside. For these analyses, I utilized

an algorithm known as GREAT (*Genomic Regions Enrichment Annotation Tool*) (McLean et al., 2010). GREAT utilizes a more comprehensive approach to identification of potential regulatory targets, addressing distance, orientation, and gene density to identify probable regulatory targets (genes) for a given set of binding sites, and uses gene ontology to predict functional roles of the targets identified. Specifically, GREAT identifies genes called to binding sites in basal regions within 5kb 5' of the TSS, 1kb 3'of the TSS, or 1,000kb beyond these regions. Additionally, GREAT compares a list of curated regulatory domains with previously identified regulatory interactions outside these parameters. Roughly half of the genes demonstrating differential expression in response to the bHLH factors are associated with at least one ChIP-seq peak region in the same cell line, suggesting that these genes represent potential direct targets of ASCL1, ASCL2, or MYOD. (Figure 3-21: Potential direct targets of bHLH factors identified from ChIP-seq and RNA-seq).

To test whether these lists of potential direct targets were more suggestive of lineage-specific transcriptional roles than DEG overall, GO analysis was performed. The full lists of potential direct target genes identified for *Ascl1, Ascl2, and MyoD* ES cell lines at 24h post-induction were submitted to the Consensus Path Data Base (MCPDB)(Kamburov et al., 2009; Kamburov et al., 2011). Unlike GREAT, MCPDB does not rely on association rules, but identifies enrichment directly from any list of gene symbols, such as those generated by edgeR from RNA-seq data sets, comparing them to defined ontological categories. As was seen in the GO analysis from the ChIP-seq binding sites, these analyses identified a number of significantly enriched category entries. For each bHLH factor, developmental categories (such as cell development, system development, cell fate commitment, cell differentiation, *et*

cetera) were identified among the most significant categorical enrichments. Genes associated with these terms included many of the most significantly differentially expressed genes, such as components of the NOTCH signaling pathway (Dll1, Dll3, Id1, Id2, Id3, Id4), Smads, and other broadly defined components of developmental significance. While these categories demonstrate the crucial relevance of these potential targets to general cellular differentiation, their broad definition, and the fact that many are shared, again suggests that they do not represent the unique components by which these bHLH factors specify their respective lineages.

Other, more specific categories, (such as muscle organ development, neurogenesis, heart morphogenesis, neuron differentiation, and others) were also identified, but as in GO analysis from ChIP-seq, these categories were largely shared, and no clear distinction can be made based on the GO categories identified from the lists of targets for each factor. The primary outlier in this respect was a higher significance for muscle-specific categories identified for MyoD-specific targets, which fits with both GO analysis from the ChIP-seq data, and comparison of the differential expression noted in the RNA-seq data, both of which identify more muscle lineage-associated genes. MYOD-expressing cells appear to induce a more thorough complement of transcriptional changes by 24h, and this is reflected in the number of potential direct targets identified. However, comparing the factor-specific GO categories identified for each ES cell line from the lists of potential direct targets reveals that while MYOD-expressing ES cells identify numerically more GO categories, the factor-specific categories identified are not dramatically muscle-directed. While a number of significant muscle-associated categories are identified as MyoD-specific (such as cardiac

ventricle formation, contractile fibers, sarcomere, muscle cell proliferation, etc.), Neural categories are also enriched in this set, frequently with considerable significance as compared to genomic background (such as axonal projection, glial cell differentiation, neuron projection, neuroblast proliferation, neuron maturation, and others). Comparing the genes contained in the lists of MyoD-specific direct targets, we see that they do not simply reflect the broad ontological association of common developmental pathways or structural genes, but include a number of potential target genes primarily associated with neural specific function, including *EphA4,* and *Sema6c*. This reflects a trend towards the considerable overlap of bHLH-responsive transcriptional profiles identified in these cells at this time point.

To test whether factor-specific direct target genes might be individually significant, the lists of potential direct targets for each ES cell line were compared across factors to identify shared and factor-specific genes (Figure 3-22: Comparison of potential direct targets of ASCL1, ASCL2, and MYOD). While the total list of these potential direct targets is quite small (only 225 genes in total were identified as a potential direct target of any bHLH factor tested), and the lists of factor specific targets even smaller, these genes represent potentially crucial mediators of lineage direction. MCPDB GO analysis from these smaller lists (17, 36, and 104 genes from ASCL1, ASCL2, and MYOD-expressing cell lines, respectively), using the same strategy and statistical cutoffs as in previous analysis, reveals only a few enriched categories. MYOD-specific potential direct targets demonstrated enrichment for categories associated with myogenic lineages. Observation of these factor-specific potential direct regulatory targets is revealing, however, as a few functionally distinct genes are identified.

Genes relevant in neurons such as *Homer2, Itpr2,* which are components of glutamatergic synapses, and *Gcm1*, a gene associated with glial development, are identified only in ASCL1-expressing cells, as was *Ascl1* itself, due to the presence of a binding site 3' of the *Ascl1* locus. *Cdh15, Cxcr4, Mef2D,* and *Smarcd3*, which are associated with muscle-specific ontologies, were identified only in MYOD-expressing cells. From these results, it appears that while the lists of potential direct regulatory targets of ASCL1, ASCL2, and MYOD are not clearly associated with dramatic factor-specific transcriptional profiles reflective of their respective lineages, they include genes potentially relevant to establishing functional transcriptional programs in these cells. (Figures 3-22: Comparison of potential direct targets of ASCL1, ASCL2, and MYOD, 3-23: Potential direct targets with lineage significance)

Binding sites near differentially expressed genes do not show additional motif specification

ASCL1, ASCL2 and MYOD bind to distinct sites, and initiate different transcriptional profiles. One possible explanation for how differential expression may be accomplished would be factor-specific differences in motifs present at regulatory enhancers for these genes. To test whether specific DNA binding motifs can be identified that predict early direct targets of bHLH function, I performed known and *de novo* motif analysis on the peak regions near these genes, using both narrow (50bp) and broad (150bp) size intervals to directly observe differences in the primary motif, and screen for co-factor motifs, respectively. Results of this analysis show strong enrichment for a GC-core Ebox as the primary motif, which was present in the majority of peak regions, suggesting that binding sites associated with differentially expressed genes largely reflect the same central

dinucleotide preference as the total set of binding sites for each bHLH factor (Figure 3-24: Comparison of *de novo* motifs at peaks associated with potential direct targets). While ASCL1 and ASCL2 failed to demonstrate a strong expansion of the 5' guanine flanking motif seen in the previous motif analysis, MYOD did demonstrate modest enrichment of the 5' adenine residues, and was identified as resembling a reference *de novo* MyoG motif. Together, this suggests that while these variant motifs may confer additional specificity for bHLH binding, they are not clearly associated with increased expression of putative targets near these sites.

Comparison of potential co-factor motifs associated with direct targets identified Sox motifs as enriched in each ChIP-seq data set. ASCL1 demonstrated a significantly enriched motif resembling a Sox2 binding site (Chen et al., 2008), present in over 30% of the 150bp peak regions tested. Sox2 is a core component of pluripotency, and has previously been studied in neural stem cells, Ascl1-mediated reprogramming (Colasante et al., 2015), and neural and lung cancers, which are also known to express high levels of ASCL1 (Borromeo et al., 2016). Sox2 is highly expressed in ES cells, including those used here. In contrast, ASCL2 and MYOD showed a marginal enrichment for a motif identified as Sox9, which was identified in 5% and 2.5% of peak regions tested. However, as Sox family binding motifs are quite similar, distinction between Sox factors based on *de novo* motifs is not possible from these data alone. Additional secondary motifs were identified, but all were present only in a minority of peak regions, and were only marginally enriched. ASCL1 identified IRF1 and HOXC9 motifs in approximately 3% of peak regions, which were virtually absent from background regions. HOXC9 is notable for playing a significant role in anterior-posterior

patterning in development. ASCL2 showed marginal enrichment for a ZBTB7B motif. *Zbtb7b* encodes a zinc finger protein known to be critical to T-cell development in the immune system, where ASCL2 is known to play a role (Liu et al., 2014). Interestingly, MYOD was enriched for PBX3 and MEF2A binding motifs, in 7% and 4.2% of peak regions, respectively. These genes are part of critical myogenic lineage pathways, suggesting that while they are present in a minority of the peak regions tested, these peaks may be associated with gene targets co-regulated by multiple factors. In aggregate, while peaks associated with direct regulatory targets of bHLH function reflect bHLH binding preference in ES cells, they are unlikely to be key determinants of early differential activity of these factors.

While bHLH binding sites identified by ChIP-seq are associated with a considerable number of differentially regulated genes, the mechanism by which these regulatory loci lead to specific gene expression changes remains unclear. The large number of binding sites which are not associated with transcriptional changes in nearby gene expression demonstrates the challenge in inferring the functional role of transcription factor binding. One possible mechanism which might explain the differential expression seen would be a factor-specific co-factor functioning to recruit transcriptional machinery to the promoter region of their distinct regulatory targets. Such a model would provide a mechanism by which bHLH factors might themselves function in a transactivating capacity without directly binding to promoter DNA, via chromatin looping or other long distance interactions. To test this possibility, I created a Linux shell script which utilizes HOMERs basic annotation and motif finding libraries to identify *de novo* motifs at upstream promoter regions of specific

sets of genes. In brief, it sequentially queries a list of genes across the mm10 database of transcription start sites, and aggregates these coordinates based on position, strand, and size arguments, and generates a list of coordinates which represent the promoters of every gene present in a given list. I then used this script to query the lists of differentially expressed genes identified as up, or downregulated in response to induction of ASCL1, ASCL2, or MYOD at 24h. I then submitted these coordinates as peak regions, and performed *de* novo motif analysis to test whether specific motifs were identified in the promoters of genes demonstrating differential expression in response to bHLH induction, providing a potential mechanism mediating bHLH function at distal enhancers (Figure 3-25: Comparison of *de novo* motifs identified at promoters of DE genes).

The results of this analysis revealed a number of motifs, such as the Sp1 zinc finger motif, and the poly-G Maz motif, which showed modest enrichment, and are associated with promoter regions in general, suggesting that they are not representative of a bHLH-specific mechanism of gene regulation. This analysis was also performed using a curated list of mouse promoter regions as a promoter-specific background. This approach did not identify these motifs, or any other highly enriched motifs suggestive of a bHLH-dependent regulatory mechanism. Thus, despite the distinct patterns of bHLH-dependent differential gene expression identified at 24h, no compelling factor-specific promoter motifs are identified which are strongly suggestive of a factor-specific transcriptional regulatory mechanism targeting promoters for expression. Furthermore, the relatively low enrichment for E-boxes at promoter regions of genes showing differential expression at this early time point suggests

that promoter-bound sites are not dramatically more capable of directing early transcriptional changes than those mediated by distal enhancer elements.

Summary and Conclusions of bHLH binding and transcriptional analyses

A considerable body of evidence has demonstrated the ability of class II bHLH factors ASCL1, ASCL2, and MYOD to interact with E-proteins, bind to CANNTG Ebox motifs throughout the genome, and enact transcriptional changes in response to their expression. Additionally, genome-wide binding of these factors from their respective tissues established the preference of these factors for a CAGSTG Ebox in these contexts (Borromeo et al., 2014; Cao et al., 2010; Schuijers et al., 2015). Furthermore, the ability of ectopically expressed ASCL1 and MYOD to function as part of a cocktail to reprogram differentiated fibroblasts into their respective lineages (Vierbuchen et al., 2010; Farah et al., 2000; Davis et al., 1987; Weintraub et al., 1989) demonstrates the ability of these master regulatory factors to enact disparate transcriptional effects in a differentiated cell type. Here, I have directly compared the genome-wide binding and transcriptional consequences of expression of ASCL1, ASCL2, or MYOD in undifferentiated embryonic stem cells at 24 hours after expression of these factors, using doxycycline-inducible ES cells. This reductionist approach allows interrogation of the molecular basis for the specificity of these factors. In doing so, I have tested their ability to bind to distinct sets of sites throughout the genome, and to enact distinct transcriptional changes in a common cellular context in the absence of lineage differences present in the tissues in which these factors function in development.

ChIP-seq data revealed that within 24 hours of their expression in the inducible ES cell system, these bHLH factors were able to bind to previously identified enhancer elements, with known functional roles in the regulation of canonical targets of bHLH function. (Figure 3-3 bHLH factors bind near known targets). These data demonstrate that even without lineage-specific cues, central components of the lateral inhibitory pathways used by these bHLH factors in establishing the proper types and numbers of cells *in vivo* remain early and robust targets of bHLH binding. Crucially, these data clearly demonstrate that while some of the sites bound are shared between ASCL1, ASCL2, and MYOD, these factors demonstrate largely distinct patterns of binding when ectopically expressed in ES cells. Thus, they demonstrate intrinsic specificity in binding that goes beyond the environment in which they are expressed.

Potential mechanisms which might explain this specificity were also explored, but no mechanism was identified that could account for the distinct patterns of binding observed. ASCL1, ASCL2, and MYOD do not demonstrate factor-specific preferences with regards to genic features or proximity to gene targets, and each binds primarily to distal enhancers in this context, as previously observed *in vivo*. Nor were differences in primary or secondary binding motifs revealed that would explain the binding specificity of these factors. The distinct flanking sequences observed for these factors may provide additional specificity for these factors, but the physical basis of this expansion and its role in bHLH functional specificity remain unclear. While intriguing secondary motifs are identified at ChIP-seq peak regions, they are numerically insufficient to explain the dramatic differences seen in genome-wide binding of these factors.

I also utilized RNA-seq to identify genes responding to bHLH induction to focus analysis of bound sites to those associated with the differentially expressed genes. This also did not reveal differential binding motifs or identify co-factor motifs uniquely associated with each bHLH factor. Thus, ASCL1, ASCL2, and MYOD maintain intrinsic, factor-specific ability to bind to factor-specific targets throughout the genome of embryonic stem cells, and do so through mechanisms other than clear differences in motif preference.

Differentially expressed gene sets from these ES cells exhibit a strong core of shared transcriptional changes, including enrichment for both neural and muscle categories in response to expression of ASCL1, ASCL2, or MYOD. Many common pathways exist that contain genes necessary for the development of multiple lineages. This may be of particular importance in light of their role, and capacity as mediators of cellular reprogramming. While cellular reprogramming is often described as "direct", in contrast to an induced pluripotent state, the definition of "direct" remains inadequately characterized. From the analysis presented here, it is clear that the transcriptional changes induced in these cells are distinct, but the presence of the considerable number of common transcriptional changes may represent a transcriptionally intermediate state. It may be the case that expression of these or other reprogramming factors may initially give rise to a transient transcriptional profile dissimilar to the lineage specified by these factors, and that this may be similar between bHLH factors. Such a model may rely on factor specific gene targets or co-factor interactions with slower dynamics to tailor the transcriptional consequences of specific bHLH factors towards their respective lineages. Previous studies have concluded that reprogramming is "direct" as cells do not express high levels of markers of stemness or specific progenitor

states (Ieda et al., 2010; Vierbuchen et al., 2010) but did not attempt to characterize an intermediate state, beyond observing an absence of markers of pluripotency. It remains possible that a transient state still occurs, where these factors have repressed mechanisms by which lineage specificity is maintained, but not yet conferred the new cell identity. This is supported by recent work which demonstrates that such a state may be transiently induced, and that some cells in this state may never fully reprogram (Treutlein et al., 2016), providing evidence that this may represent one component of the limitations that lead to very low efficiency in the reprogramming of fibroblasts to neurons, and other cellular reprogramming models. Such a model would suggest that these bHLH factors might first commonly function to establish a permissible state for reprogramming before initiating expression of lineage-specific genes, perhaps in complex with, or mediated by early downstream targets of these bHLH factors.

One sub aim of this analysis is to use the reduced system of the ES cells to identify potential downstream targets and regulatory mechanisms crucial to the functional and reprogramming capabilities of these bHLH factors. While a relatively small number of genes demonstrate factor-specific changes in gene expression within 24h, factor-specific expression at this early stage may not be critical to factor-specific mechanisms of function. Even if commonly induced by bHLH factors, these gene targets may then differentially interact with these master regulators to enact more dramatic changes in transcriptional profile. Highlighting this, while there exists rather dramatic similarity in the bHLH domain of these factors, there are structural differences between these factors, especially in residues which do not directly contact the DNA (Nakada et al., 2004), or in the H1/H2 interaction region

(Figures 1-1, 3-11). These provide potential interaction sites, and may allow for a common transcriptional target to confer factor-specific functional capabilities on one or more bHLH factors. While ASCL1 has been shown to depend heavily on its bHLH domain for functional specificity in the context of the nervous system (Nakada et al., 2004), MyoD has been shown to also require cooperation of its terminal domains for its full complement of myogenic capabilities (Ishibashi et al., 2005), demonstrating that these proteins rely on more than simply their ability to recognize specific Eboxes to give rise to their distinct transcriptional profile in the context of partially defined cell lineages. Even ASCL1 and ASCL2, which are nearly identical within their bHLH domain, have dramatic differences outside this domain. Each of these differences provide opportunity for factor-specific co-factor interactions. While the distinct binding profiles of these factors, and DNA motif analysis of their respective binding sites suggest that factor-specific DNA binding co-factors are not likely to be the crucial determinant of binding specificity, there remain potential roles for factors which do not have direct DNA-binding activity, including a number of transcription factors identified in our study as potential direct targets of ASCL1, ASCL2, and MYOD. For example, Id factors, which possess an HLH domain, are known to repress bHLH function by interacting with class I bHLH factors. However, ID1 and ID2 have also been demonstrated to interact with MYOD and repress its ability to initiate transcription of a myogenic reporter, but not interact with SCL/Tal-1 (Langlands et al., 1997). Factor-specific interactions such as this may provide a mechanism by which commonly expressed target genes might differentially regulate the downstream transcriptional profiles established by these factors without being evident in motif analysis. Similar mechanisms may provide additional factor-specific

functional specificity, allowing for further transcriptional regulation which may not be apparent in motif analysis.

Another potential mechanism for functional specificity is through nongenic RNA transcription. While gene-coding RNA transcription is the oldest and best understood function of DNA-RNA transcription, it is now widely understood that noncoding RNA can lead to dramatic changes in gene expression through a surprisingly diverse range of functional mechanisms. One form of noncoding RNA, microRNA (miRNA), has previously been revealed as crucial posttranscriptional regulators of gene expression, binding to mRNA and modifying their translational properties either by directly repressing translation, or by targeting these mRNAs for modification by other mechanisms within the cell. miRNAs have specifically been implicated in modulating transcriptional response to stochastic gene expression in blood lineages, which are also heavily dependent on bHLH mediated lineage specification, and miRNA dysregulation has been shown to be relevant in leukemias (as reviewed in Musilova & Mraz, 2015), highlighting their central role in the regulation of gene expression. As potent regulators of cell fate, expression of ASCL1, ASCL2, and MYOD, as well as their downstream mediators of cell fate must be tightly controlled to prevent inappropriate lineage specification in vivo, and miRNA-mediated mechanisms may be central to regulating factor-specific gene expression. Indeed, specific miRNAs have been implicated in both maintenance of somatic stem cell populations, and in differentiation into specific lineages (as reviewed in Shenoy & Blelloch, 2014). Expression of miRNAs alone has also been shown to be at least partially sufficient to reprogram fibroblasts to neural cell types (Yoo et al., 2011), suggesting a potential role for miRNA in neural lineage

specification and reprogramming. miRNA function has been demonstrated as central to the stem-like properties of *Ascl2*-expressing tumors in a model of colorectal cancer (Zhu et al., 2012). Furthermore, MYOD has itself been demonstrated to target expression of a number of miRNAs implicated in myogenesis (Rao et al., 2006), and to repress TWIST-1 function via a specific miRNA (Koutalianos et al., 2015). Thus it is clear that miRNAs are already implicated in bHLH function in development and disease, and may mediate some components of factor-specific gene regulation.

Long noncoding RNAs (lncRNAs) are a second, distinct form of noncoding RNAs which were more recently discovered. Like miRNAs, lncRNAs have been demonstrated as a crucial mechanism for posttranscriptional modulation of gene expression, but appear to have more diverse modes of function, and in particular are implicated in spatiotemporal regulation of neural gene expression. They are implicated in activation of specific genes by RNA polymerase II targeting, regulation of pluripotency in cis and in trans (Luo et al., 2016), and have known roles in both activation and repression of gene expression (as reviewed in Geisler & Coller, 2013). Especially intriguing is a potential role for lncRNAs to interact directly with DNA to target proteins, potentially providing a high degree of specificity for DNA sequences which may not be apparent in motif analysis. These and other noncoding RNA may provide additional functional specificity by regulating binding, transcription, and posttranscriptional changes.

Alternately, it may be the case that genes which are expressed at different levels in response to each bHLH factor lead to downstream regulatory complexity. These genes may have different effects at different expression levels. *Hes6*, for example, is identified as a

shared target gene, but demonstrates dramatically higher expression in response to MYOD as compared to ASCL1. HES6 has known roles in both neural and muscle tissues, but may function differently at different levels. In neural development, HES6 is believed to interfere with HES1-mediated repression of ASCL1 activity, and inhibits neural differentiation (Bae et al., 2000), but also promotes skeletal muscle differentiation (Gao et al., 2001). Thus distinct levels of expression of a common target may be sufficient to produce distinct transcriptional effects in response to bHLH factors.

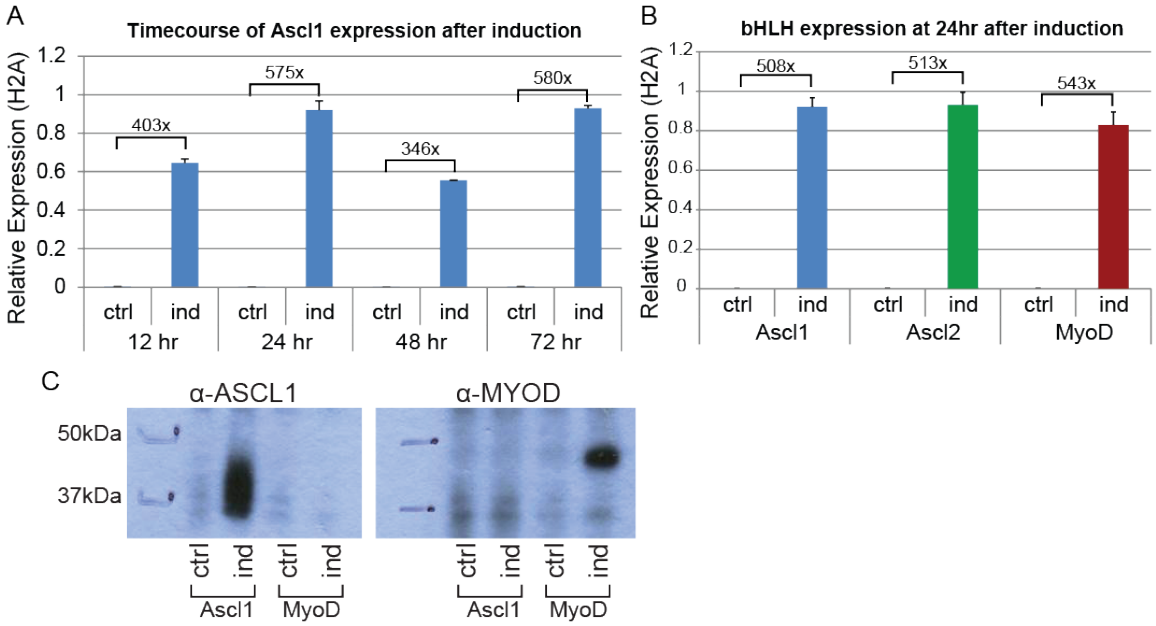
While the ES cell model with inducible TFs allows us to directly regulate expression of the bHLH transgene, our ability to induce expression of these factors is restricted to an "on/off" system. As the mechanisms underlying endogenous expression of these factors remains unclear, it is almost certainly the case that their expression during development is considerably more complex and nuanced than can be recapitulated in the paradigm used here. Indeed, it has previously been demonstrated that ASCL1, ASCL2 and MYOD each demonstrate autoregulatory mechanisms (Meredith & Johnson, 2000; Yan et al., 2015; Thayer et al., 1989; Zingg et al., 1994; Penn et al., 2004) which positively (ASCL2 and MYOD) or negatively (ASCL1) modulate expression, directly, and indirectly, respectively. Additionally, ASCL1 expression has recently been demonstrated to oscillate in neural progenitors *in vivo*, and manipulation of this oscillation is sufficient to direct proliferation and differentiation decisions (Imayoshi et al., 2013). Likewise, MYOD has been demonstrated to be directly regulated by the circadian clock in skeletal muscle in an oscillatory manner, and this oscillation is necessary for function and maintenance of skeletal muscle (Andrews et al., 2010). Graded expression of ASCL1 by retinoic acid treatment gives

rise to alternative neural populations (Jacob et al., 2013), suggesting that even static differences in the level of *Ascl1* expression may direct different transcriptional programs. Thus, some aspects of the binding or transcriptional program induced by these factors may be masked, or lost due to dramatic overexpression of the bHLH. However, it has previously been demonstrated that overexpression of MYOD binds largely the same set of sites as compared to endogenously expressed MYOD, and showed the same E-box preference and co-factor enrichments (Yao et al., 2013), suggesting that expression level is not a primary factor in establishing binding site specificity. Furthermore, the ability of ASCL1 and MYOD to reprogram differentiated cells into alternate lineages suggests that they remain capable of binding appropriate sites and directing the necessary gene expression to accomplish these transitions.

Together, these data demonstrate that ASCL1, ASCL2, and MYOD maintain distinct binding and transcriptional profiles even when expressed outside the cellular contexts in which they function in development. Thus, factors other than cellular environment provide the mechanism by which they bind to distinct sites, and enact their disparate transcriptional programs. One component of this environment believed to differ between cells of different lineages and developmental stages is chromatin accessibility, which certainly plays a role in regulating the binding of these factors in a factor specific-manner. Nevertheless, even in a common context such as the ES cells utilized here, the binding of these factors may reflect factor-specific differences in preference for open or closed chromatin, or other elements of the chromatin landscape, such as histone modifications. Such a model may help explain the differences in binding observed between the bHLH factors. In Chapter 4, we consider these

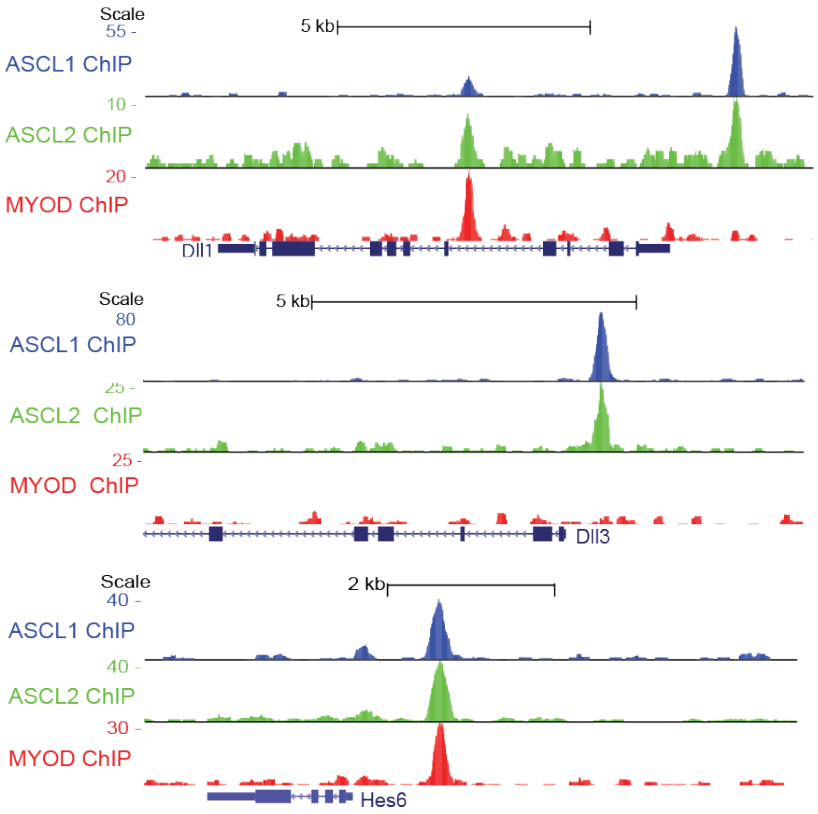
potential mechanisms regulating the binding and function of ASCL1, ASCL2, and MYOD, and examine their interplay with the chromatin landscape.

Figure 3-1: Inducible ES cells demonstrate robust expression of bHLH factors within 24 hours



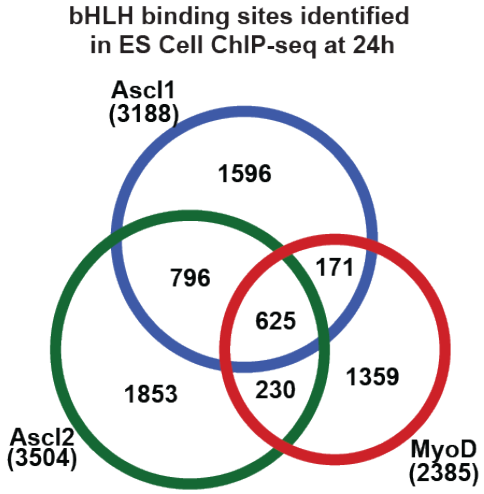
Inducible ES cell lines demonstrate robust expression of bHLH transgenes within 24h post-induction. (A) Timecouse demonstrating *Ascl1* transgene expression after induction. Plot depicts H2A-normalized expression of Ascl1 in *tTA-Ascl1* ES cells, as measured by RT-qPCR for *Ascl1*. (B) Comparison of bHLH transgene induction at 24h post-induction, as measured by RT-qPCR for bHLH transgenes, as indicated. Values represent mean H2A normalized expression of technical replicates taken at each timepoint. Fold change of induced vs. control samples is indicated. Error bars shown represent standard deviation between technical replicates for each sample. (C) western blot showing induction of ASCL1 and MYOD protein expression at 48h post-induction. Protein detection using α-ASCL1 and α-MYOD antibodies. Sizes correspond to previously those previously reported for these proteins. Size indicated by standard ladder.

Figure 3-2: ASCL1, ASCL2, and MYOD bind at distinct sites near key developmental genes



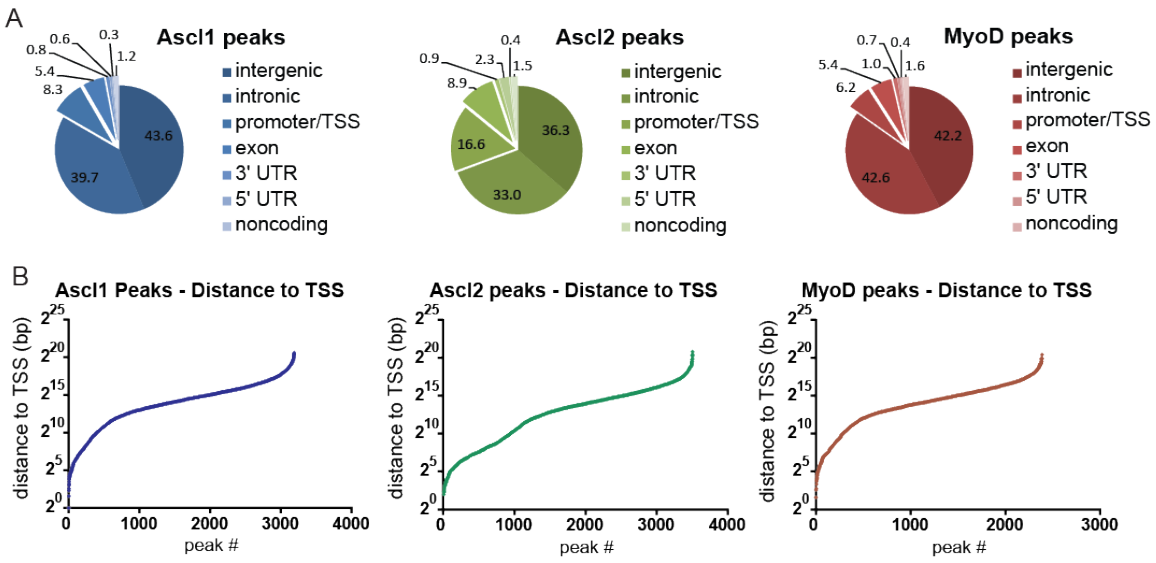
ASCL1, ASCL2, and MYOD bind previously observed targets.
UCSC Genome browser tracks comparing bHLH ChIP-seq enrichment near Notch pathway components *Dll1*, *Dll3*, and *Hes6*. Each track shown represents ChIP-seq data from the indicated bHLH factor, normalized to 10M reads and aligned to the mm10 genome. Track scale indicated represents total track height shown, base of track is zero reads. Distance scale of observed window indicated at the top of each set.

Figure 3-3: Overlap of bHLH binding sites identified in ChIP-seq



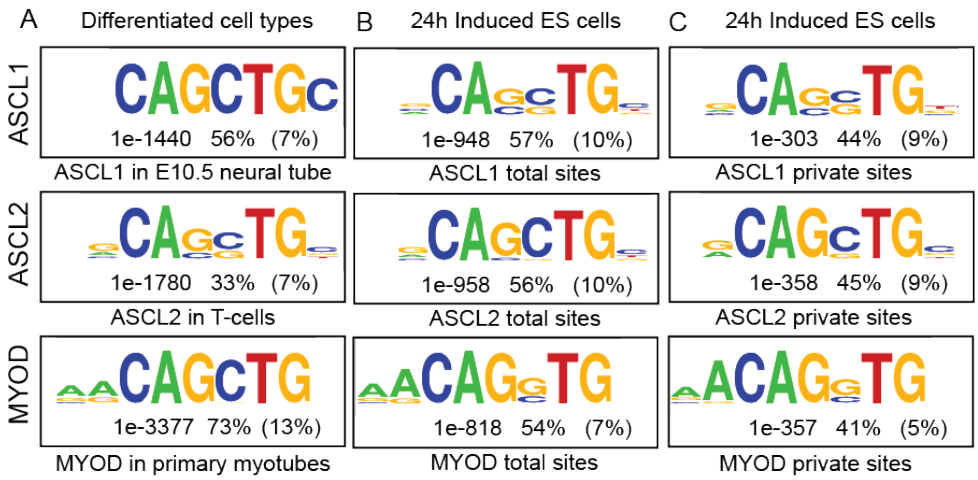
ChIP-seq reveals distinct binding sites for bHLH factors ASCL1, ASCL2, and MYOD in ES cells.
(A) Proportional area diagram of bHLH binding sites identified by bHLH or FLAG ChIP-seq for ASCL1, ASCL2, or MYOD in ES cells at 24h post induction. Numbers represent distinct peaks in each subset, and total number of peaks identified for each factor. Overlapping peaks are defined by peaks identified within 150bp between ChIP-seq data sets for each bHLH factor tested.

Figure 3-4: ASCL1, ASCL2, and MYOD have similar binding distribution relative to gene features



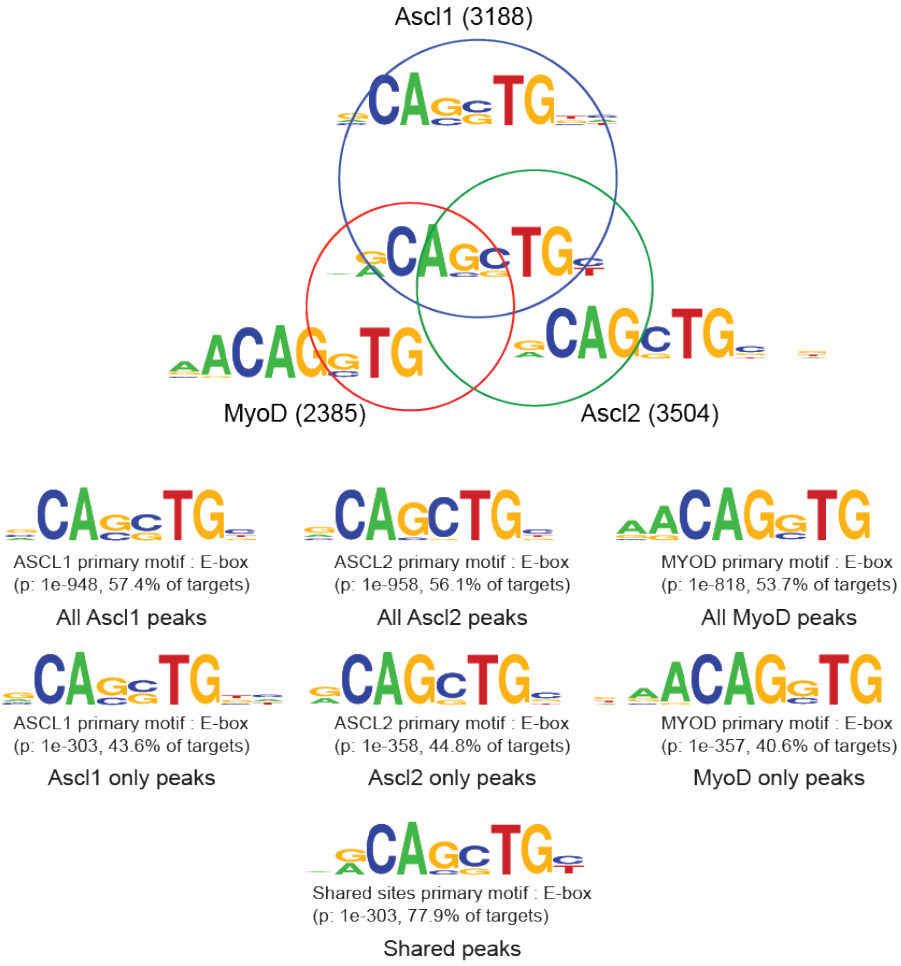
ASCL1, ASCL2, and MYOD binding sites are similarly distributed relative to gene features
A) Genic feature annotation of ChIP-seq peak regions bound by bHLH factors in ES cells at 24h. Numbers represent percent of total peaks for each bHLH bound in each annotation category listed, ordered as in legend. Each peak is assigned to only one category shown. Annotation performed using HOMER.
(B) Comparison showing distribution of distances from bHLH binding sites to TSS. Plot represents absolute distance between peak center and nearest TSS for each peak identified within ChIP-seq data set.

Figure 3-5 - Comparison of *de novo* motifs identified in ES cells and differentiated cell types



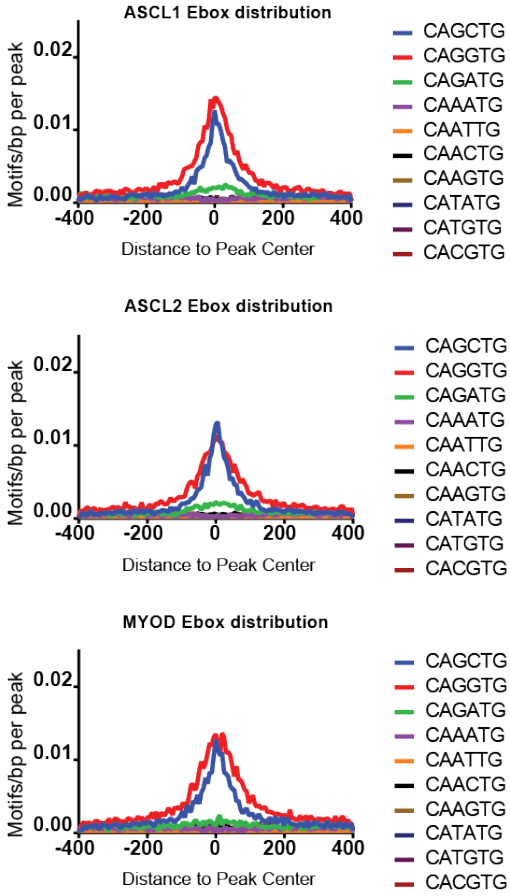
Comparison of primary E-box motifs identified in ChIP-seq for ASCL1, ASCL2, and MYOD.
Motifs shown primary result of *de novo* motif analysis as identified using HOMER. ChIP-seq from differentiated tissues (A), and induced ES cell lines (B,C) using 50bp interval centered on peak apex. Comparative motifs in differentiated cell types were generated from previously published data sets for ASCL1 (Borromeo et al., 2013), ASCL2 (Liu et al., 2014), and MYOD (Cao et al., 2010). Motifs in ES cells are shown as observed for the set of total (B), and private (C) sites for each factor, as indicated. Numbers reflect the P-value significance of the specified motif, the percent incidence of the specific motif shown in ChIP-seq peak regions, and percent incidence in a normalized random background set (indicated in parens), values shown rounded to the nearest integer.

Figure 3-6: Overlap comparison of factor-specific and shared motifs



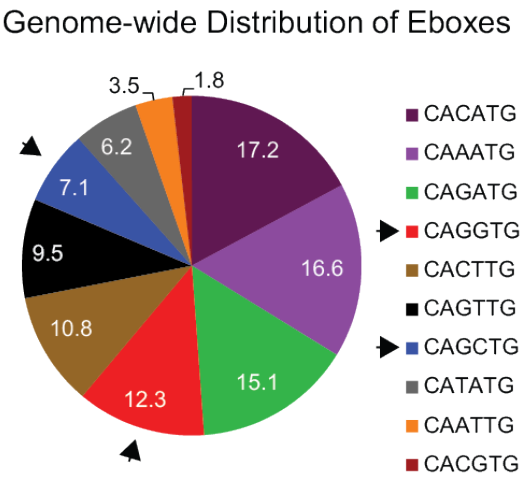
ASCL1, ASCL2, and MYOD preferentially bind a CAGSTG Ebox motif.
(A) Proportional area diagram showing de novo motifs identified from shared and private sites identified for each factor. Motifs shown represent most significant motif identified for each subset, identified from 50bp interval centered on peak identified.
(B) Comparison of de novo motifs identified from total, private, and shared binding sites identified for each factor by ChIP-seq. Motif shown represents most significant de novo motif identified for each set. Numbers reflect the P-value significance of the specified motif, the percent incidence of the specific motif shown in ChIP-seq peak regions, and percent incidence in a normalized random background set, values shown rounded to the nearest integer.

Figure 3-7: Ebox distribution at bHLH ChIP-seq peaks



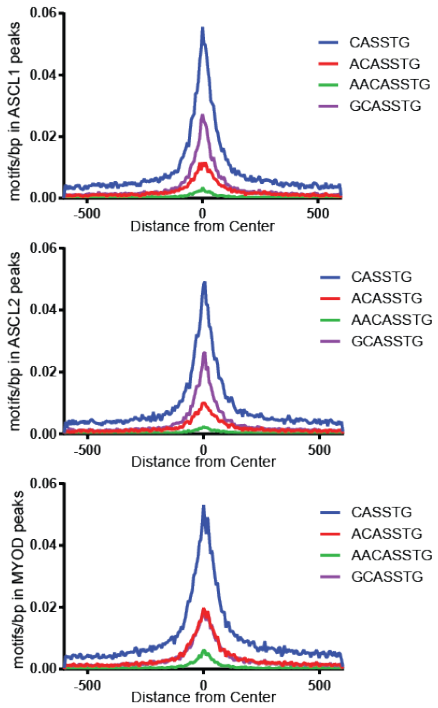
ASCL1, ASCL2, MYOD preferentially bind CAGSTG E-boxes.
Histogram plots compare distribution of all E-box motifs present at binding sites for each factor, as identified in ChIP-seq from ES cells at 24h post-induction. Annotation shows distribution of all E-boxes present within ±400bp of peak center for the total set of binding sites identified for each factor, with all core dinucleotide permutations shown, with equivalent Ebox permutations are collapsed to a single motif for comparison.

Figure 3-8: Genome-wide distribution of Eboxes



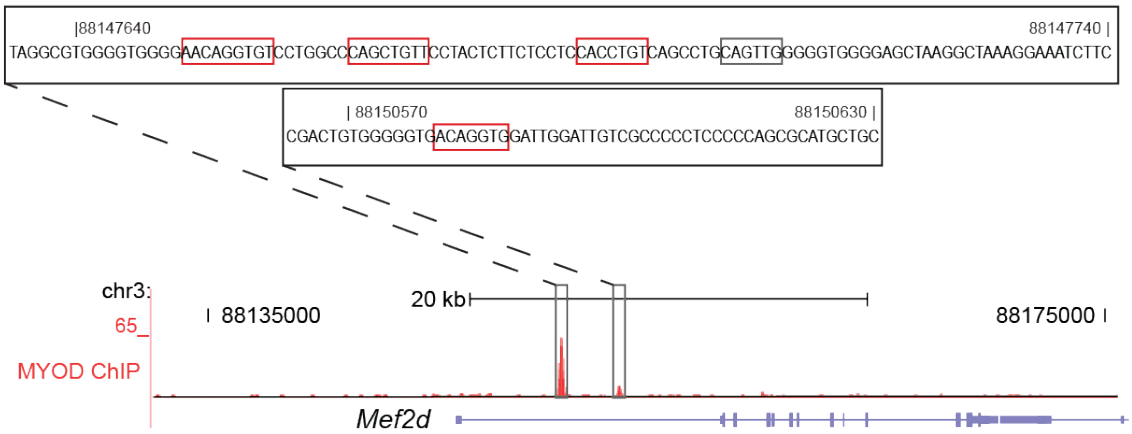
ASCL1, ASCL2, and MYOD Ebox preference is not based on genomic motif prevalence.
Chart shows prevalence of of each core dinucleotide Ebox sequence genome-wide in mm10 genome. Numbers reflect percentage of each Ebox dinucletide as compared to total Ebox sequences identified by genomic annotation. Arrows highlight enriched Ebox sequences in ChIP-seq for bHLH factors.

Figure 3-9: Distribution of flanking variants identified for ASCL1/ASCL2 versus MYOD



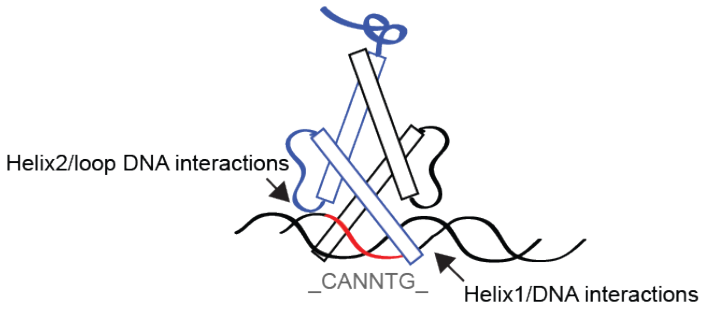
ASCL1/ASCL2 and MYOD are differentially enriched for variant flanking sequences.
Plots depict distribution of variant flanking sequences at ChIP-seq peak regions identified in each inducible ES cell line, within ±600bp of peak centers identified for each bHLH factor as indicated. Each colored trace indicates the distribution of the specific Ebox shown. Expanded E-box motifs identified in bHLH ChIP-seq were directly annotated using HOMER, using the motif sequence indicated.

Figure 3-10: MYOD binding sites within *Mef2D* locus



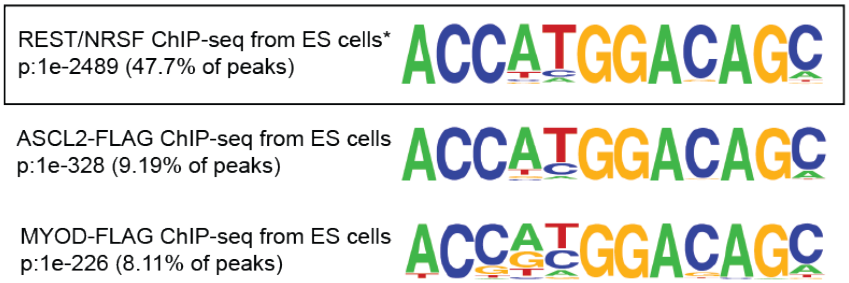
MYOD binding sites identified within *Mef2D* locus feature expanded motif identified by *de novo* motif discovery. Two binding sites are located within intron 1 of *Mef2D*, and demonstrate differential binding by MYOD. Nucleotide sequence for genomic intervals indicated shown as inset with genomic coordinates indicated. Boxed sequences represent all E-boxes identified within these intervals. Red boxes highlight ACANNTG E-boxes within indicated regions.

Figure 3-11: Potential bHLH:DNA interaction sites



DNA adjacent to Ebox motifs are in proximity to specific domains of bHLH factors
Diagram highlighting bHLH dimer components and interactions potentially underlying expanded E-box motif identified by ChIP-seq. Red line and minimal Ebox shown approximate alignment relative to bHLH position as identified in solved crystal structures for class II bHLH factors. Arrows denote bHLH components proximal to Ebox-flanking nucleotide bases.

Figure 3-12: Comparison of REST motif from REST ChIP-seq versus FLAG ChIP-seq



ASCL2 and MYOD FLAG ChIP samples are enriched for selective, non-bHLH specific REST/NRSF motif
Comparison of REST/NRSF motif identified in ChIP-seq for REST/NRSF in ES cells (McGann et al., 2014), and corresponding motif identified in FLAG ChIP from ASCL1-expressing and MYOD-expressing ES cells. Motifs shown represent *de novo* result corresponding to known REST/NRSF motifs. REST/NRSF ChIP-seq motif shown for comparison (*McGann et al., 2014).

Figure 3-13: Comparison of significant secondary motifs identified in ChIP-seq

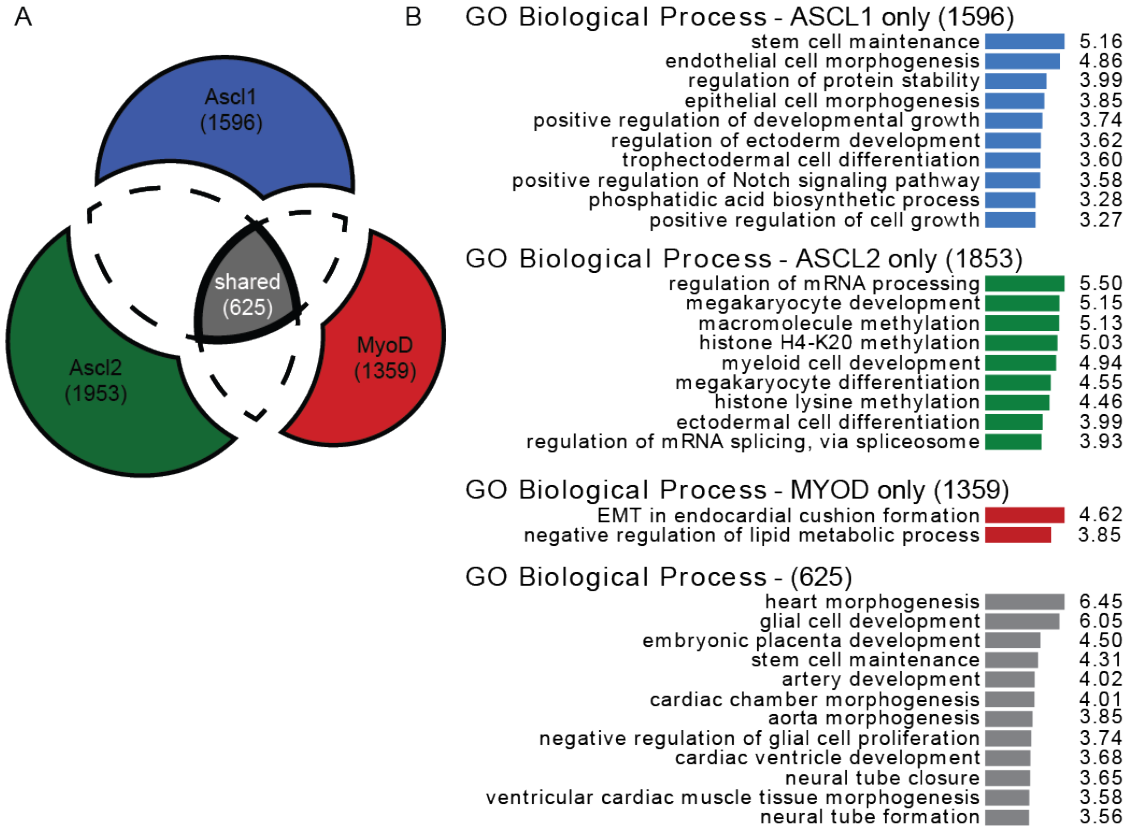
Secondary motifs are present at a minority of binding sites identified in ChIP-seq for bHLH factors. Comparison highlighting most significant *de novo* secondary. motifs identified in ChIP-seq data from each bHLH factor. Motifs shown taken from HOMER analysis of 150bp regions surrounding peak centers. de novo motifs shown were identified in ChIP-seq peaks for the factor listed. Best Match represents most significantly similar motif identified by HOMER. Statistics reflect the binomial significance of the de novo motif identified, as compared to random background, with percentage of binding sites and randomized background regions featuring the motif depicted, as indicated.

Figure 3-14: GREAT analysis of full bHLH GO categories

Genes associated with bHLH binding sites identified by ChIP-seq are enriched for developmental processes Comparison of most significant GO-Biological process categories identified from total sets of bHLH ChIP-seq peaks, as identified by GREAT v3.0. Most significant GO categories for each peak set shown, ranked by association with known GO BP categories. Values and colored bars reflect -log10 binomial P-values as calculated by GREAT. Redundant gene ontology categories identified within the list of a single factor not shown.

ASCL1 20160702-public-3.0.0-XyQJNs
ASCL2 20160702-public-3.0.0-4YT4Ln
MYOD 20160702-public-3.0.0-tQEHib

Figure 3-15: GREAT analysis of shared and factor-specific bHLH GO categories

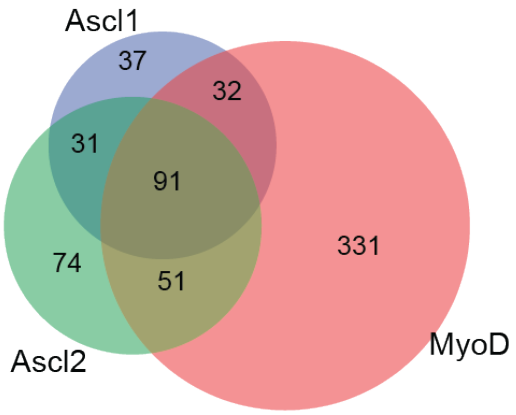


Factor-specific binding sites identified in ChIP-seq are not dramatically enriched for lineage-specific ontologies. (A)Area-proportional diagram showing sets of shared and factor specific binding sites used for gene ontology analysis. Numbers reflect subsets of binding sites used in each set. (B) Comparison of most significant GO-Biological process categories identified from total sets of bHLH ChIP-seq peaks, as identified by GREAT v3.0. Most significant GO categories for each peak set shown, ranked by association with known GO BP categories. Values and colored bars reflect -log10 binomial P-values as calculated by GREAT. Redundant gene ontology categories identified not shown.

ASCL1 -20160702-public-3.0.0-rszwNZ
ASCL2 - 20160702-public-3.0.0-Re7F7s
MYOD - 20160702-public-3.0.0-t9Bb38
Shared - 20160702-public-3.0.0-xVXl4M

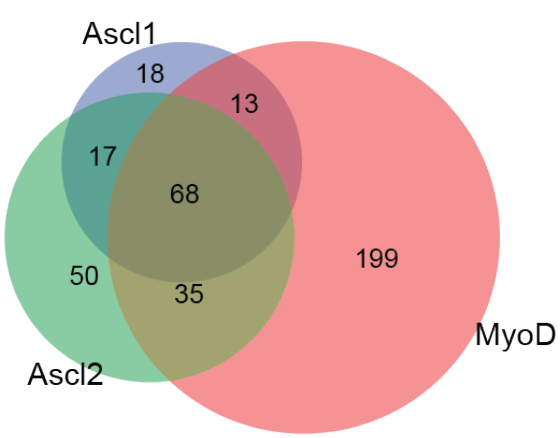
Figure 3-16: Comparison of genes showing significant differential expression at 24h

A

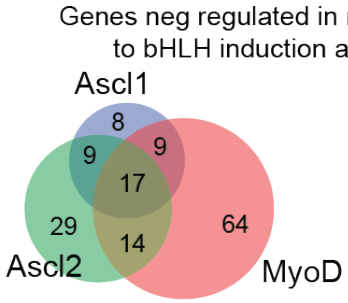


FC≥2, FDR≤.05, no RPKM cutoff

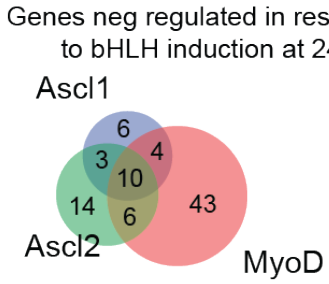
B



FC≥2, FDR≤.05, RPKM(ind)≥1



FC≤0.5, FDR≤.05, no RPKM cutoff



FC≤0.05, FDR≤.05, RPKM(ctrl)≥1

C

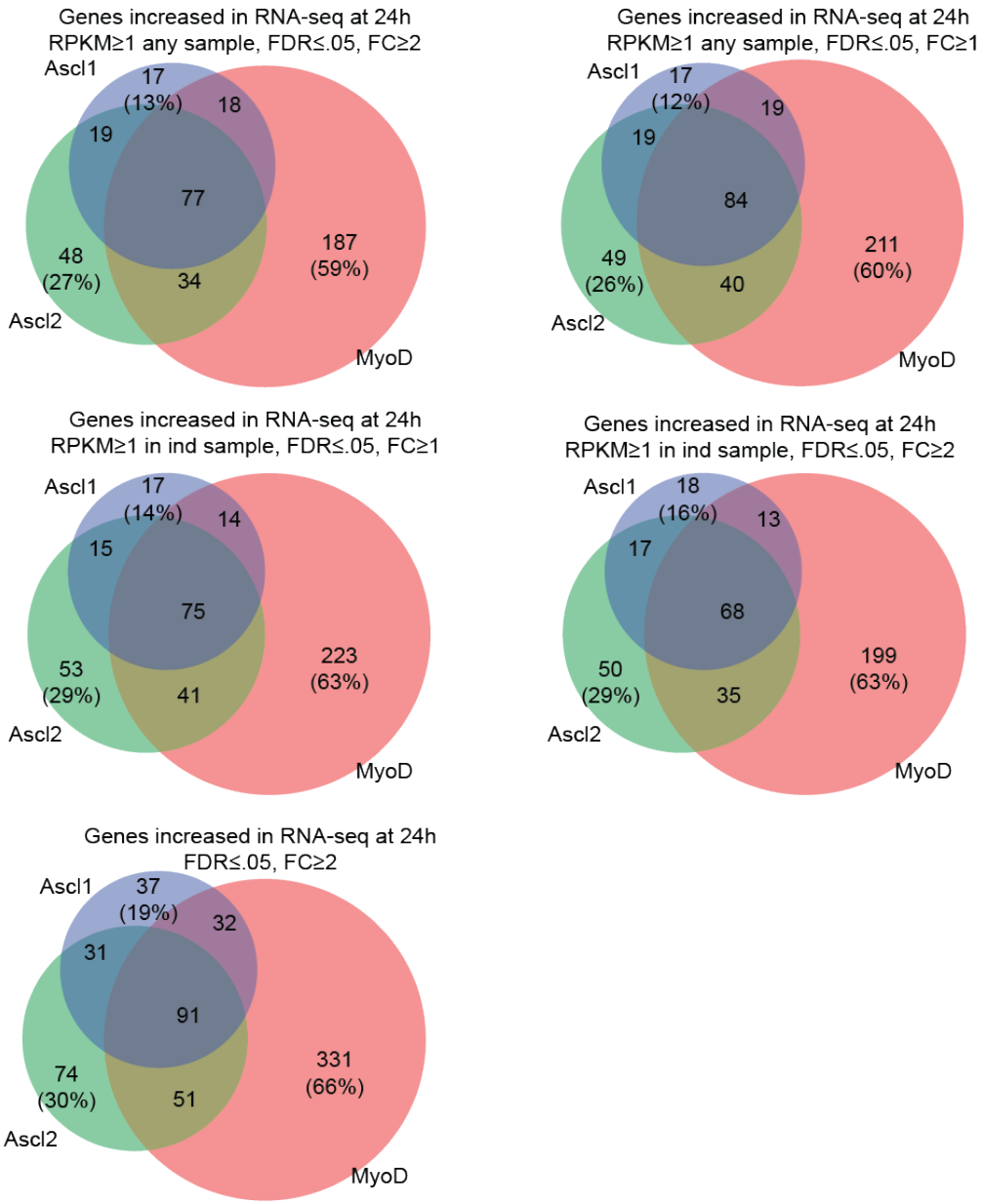
Total genes identified by edgeR analysis (FDR≤0.05)			
bHLH	**Changed**	**Increased**	**Decreased**
Ascl1	234	191	43
Ascl2	326	247	79
MyoD	608	505	103

D

Genes showing expression and change at 24h (ind)			
bHLH	**Changed**	**Increased**	**Decreased**
Ascl1	139	116	23
Ascl2	203	170	33
MyoD	378	315	63

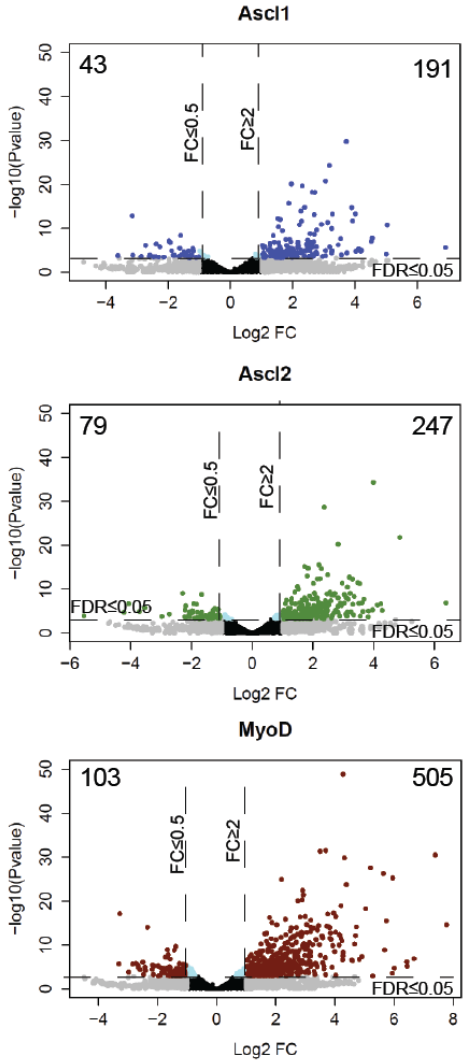
Induction of bHLH factors ASCL1, ASCL2, and MYOD leads to changes in gene expression within 24h
Proportional area diagram comparison of genes showing significant bHLH-dependent changes in expression. Plots show number and concordance of total genes (A) demonstrating significant increase (top) and decrease (bottom), and (B) expressed genes in gene expression at 24 hours post-induction based on fold change for each gene and edgeR significance (FDR≤0.05). Criteria used for each selection are indicated in figure. Tables (C,D) show total numbers of genes identified by each criteria for comparison. Fold change calculated as (mean RPKM ind/mean RPKM ctrl). All data reflect three independent biological replicates for each experimental condition.

Figure 3-17: Comparison of DEG criteria demonstrating largely similar degree of overlap



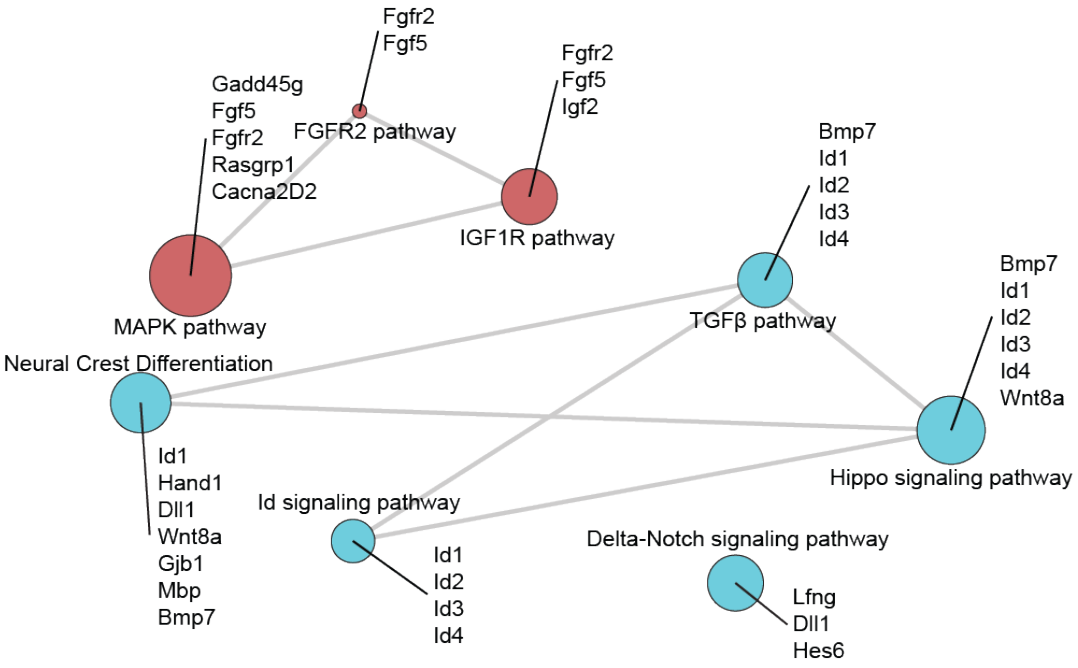
Comparison of overlap between RNA-seq using different parameters for identification of sig upregulated genes. Area-proportional diagrams represent the number of genes in each subset meeting the specified criteria indicated. Numbers of genes in each subset as indicated. Percentage represents fraction of genes showing factor-specific expression, as compared to total DEG for each factor.
Overall conclusion: relative overlap stays almost identical, regardless of the parameters used. Numbers of genes identified vary dramatically based on the thresholds set for FC and whether an RPKM cutoff is used, but conclusion remains the same: A core set of shared, overlapping genes are induced by all three factors, and only a few hundred are significantly induced regardless of parameters. Filtering RPKM had the greatest effect on the number of genes identified.

Figure 3-17b: Comparison of FC and significance of DEGs



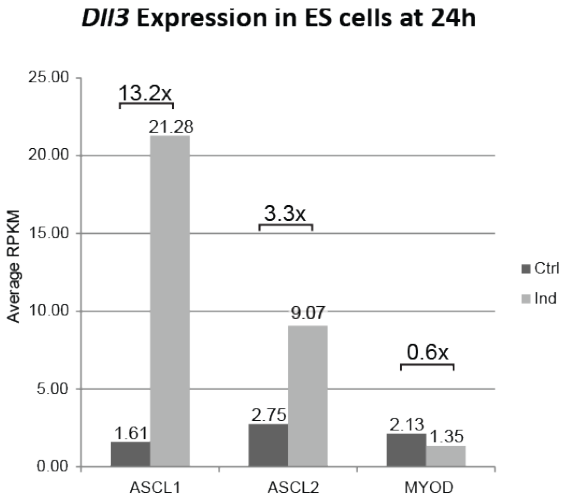
Volcano plot comparison of genes exhibiting significant differential expression in each ES cell line Scatterplots indicate the significance ($-\log_{10}$ p-value of expression change from edgeR results) and fold change ($\log_2$ FC) observed for all genes present in the mouse genome. Dashed lines highlight threshold for cutoffs used in defining differentially expressed genes in our analysis (FC ≥2 or ≤0.5, FDR≤0.05). Cyan points represent genes with a mean fold change below threshold for positive or negative differential expression, but were identified as significant by edgeR across three biological replicates for each cell experimental condition.

Figure 3-19: Overview of selected developmental pathways identified from shared DEGs



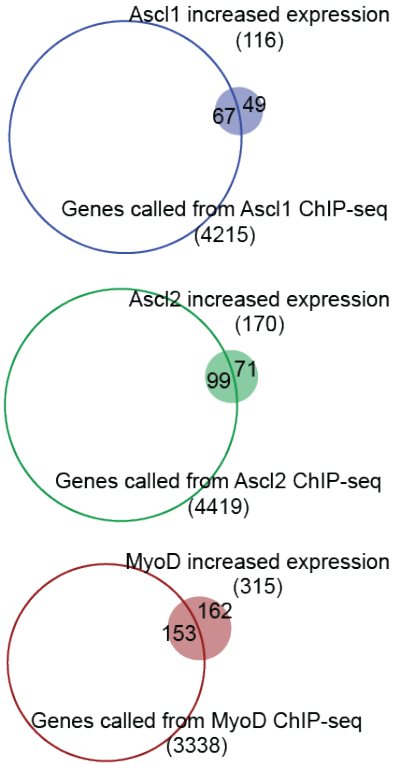
Shared targets of bHLH factors include key components of developmental signaling pathways.
Figure depicts network diagram comparing genes associated with developmental signaling pathways demonstrating shared bHLH-dependent expression in response to Ascl1, Ascl2, and MyoD induction at 24h. Colored circles represent defined pathways based on GO analysis. Lines link pathways with shared components identified in this analysis. Genes listed were identified as DEGs in response to all three factors meeting significance threshold by RPKM≥1, FC≥2, edgeR FDR≤0.05 across replicates.

Figure 3-20: bHLH-dependent activation of Dll3 in induced ES cells



RNA-seq reveals differential activation of Dll3 between bHLH factors ASCL1, ASCL2, and MYOD
Comparison of Dll3 expression from RNA-seq in uninduced control, and 24h post-induction across ES cell lines. Values represent mean RPKM across three biological replicates for each experimental condition. FC (indicated) represents mean RPKM of induced vs uninduced control samples at 24 hours post-induction.

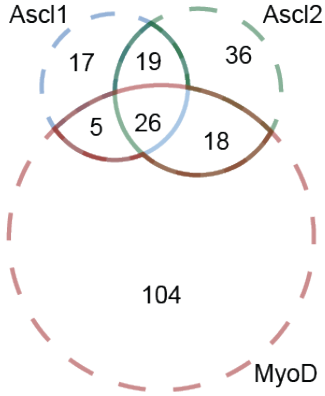
Figure 3-21: Potential direct targets of bHLH factors identified from ChIP-seq and RNA-seq



Genes associated with bHLH binding sites identified in ChIP-seq are over-represented in DEGs.
Area-proportional overlap diagrams comparing overlap between genes identified by peak-to-gene association from ChIP-seq binding sites and DEGs identified in RNA-seq. Numbers represent set of total genes identified by each approach, and the overlapping subset identified as potential direct targets. ChIP-seq represents genes called by GREAT v3.0 using default parameters. RNA-seq from genes with average RPKM≥1 (induced), FC≥2, and FDR≤0.05 in the cell line in which it was identified. .

Figure 3-22: Comparison of potential direct targets of ASCL1, ASCL2, MYOD

A



B

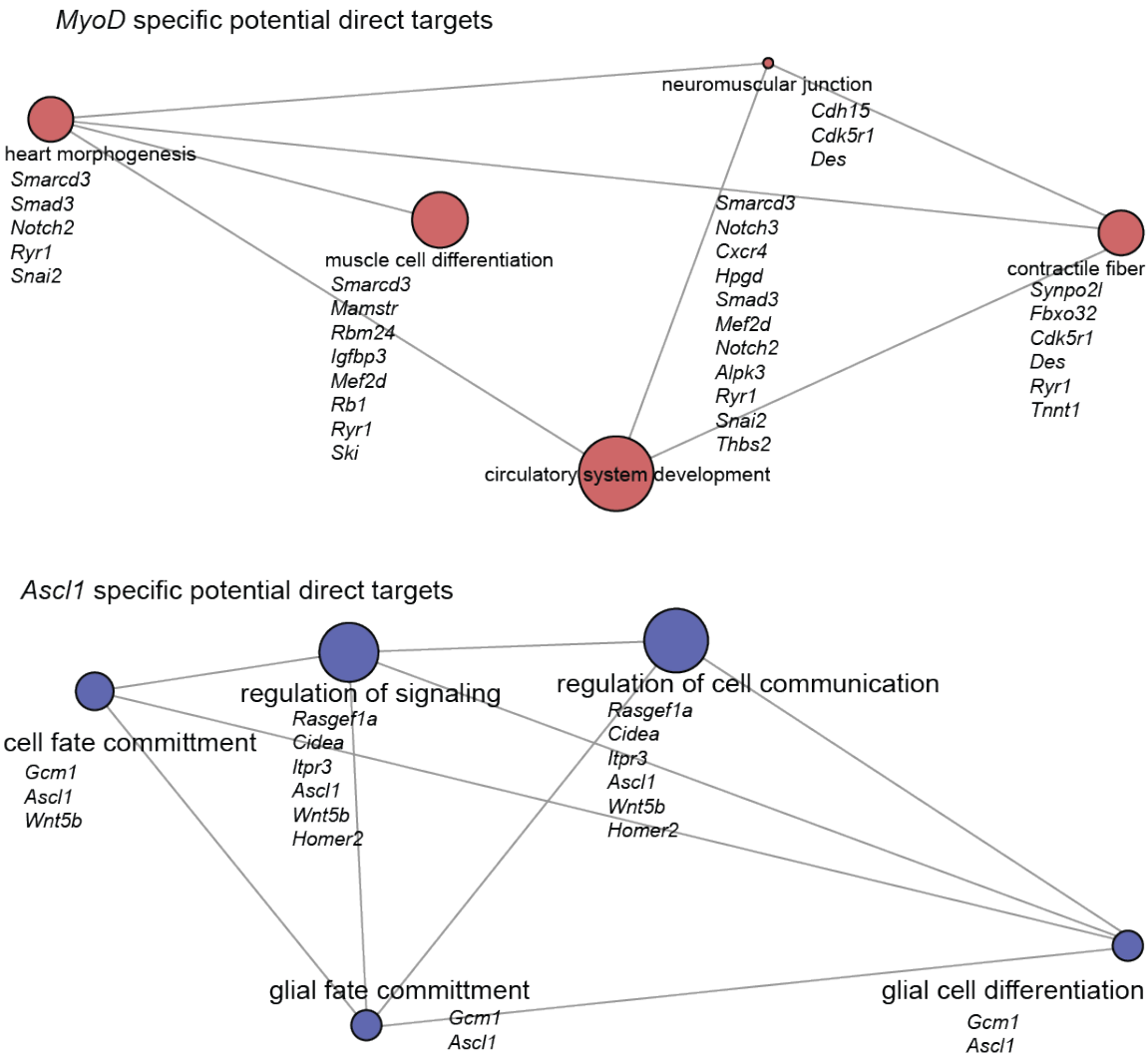
Shared targets:	ASCL1 & ASCL2:	ASCL2 & MYOD:	ASCL1 & MYOD:
26 shared genes	19 shared genes	18 shared genes	5 shared genes
Hes6	Dll3	Rassf4	Nxn
Dll1	Parp12	Igfbp5	Itgam
Krt7	Id3	Ldhb	Cldn5
Lfng	Cdc25b	Lbh	Ninj1
Nsg2	Abcg1	Zbp1	Ap3b2
Kcnh8	Slc27a2	Prox1	
Irf1	Plk3	Tmem88b	
Gadd45g	Tmem145	Snca	
Mbp	Wnt8a	Tubb3	
Id2	Tnnt3	Gprc5b	
Fgfr2	Nbl1	Pak3	
Lzts1	Sel1l3	Id4	
Tbx4	Cdh3	Phf19	
Cdkn1a	Id1	Phc2	
Ppap2b	4632428N05Rik	Dsp	
Syne3	Lgals9	Fam83b	
Lurap1l	Vwf	Nfasc	
Cacna2d2	Tmem184a	Atp2b4	
E2f2	Cbfa2t2		
Zcchc24			
Urgcp			
Rgs16			
Hand1			
Fam19a4			
Btbd17			
Syt13			

C

17 Ascl1 only:	Top 20 Ascl2 only:	Top 20 MyoD only:
Ascl1	Ccdc136	Smarcd3
Tmem154	Myom2	Rbm24
Wnt5b	Fgf5	Tnnt1
Ms4a10	Klhdc7a	Celf2
Klk1	Bmp7	Myadml2
Tspan15	Chrnb1	Cxcr4
Slc43a1	Vwa1	Des
Homer2	Klhl21	Rb1
Gcm1	Foxi3	Sema6c
Sfi1	Pirt	Slc44a2
Plcxd1	Slain1	Rtn2
Gdap1l1	Lrig1	Mamstr
Itpr3	Cgnl1	Pde2a
Rasgef1a	Lypd6b	Cited4
Cidea	Dmrt1	Cdh15
Neurog3	Ldlrad3	Pgap1
Tmem119	Plekhg1	Fitm1
	Aldh1a2	Elfn1
	Lrrc4	Snai2
	Klf13	Aif1l

Potential direct targets identified from ChIP-seq and RNA-seq include shared and factor specific genes
Genes listed represent DEGs identified for each bHLH from RNA-seq, compared to genes identified by peak-to-gene association using GREAT. (A) Proportional diagram of potential direct targets (PDT) identified for each factor. (B) PDTs shared between two or more bHLH factors tested. (C) factor-specific PDTs identified for ASCL1, ASCL2, and MYOD. Lists are sorted based on FDR of expression change in decreasing order of significance, as calculated by EdgeR. Top 20 genes shown for ASCL2 and MYOD.

Figure 3-23: highlights from bHLH-specific potential direct targets

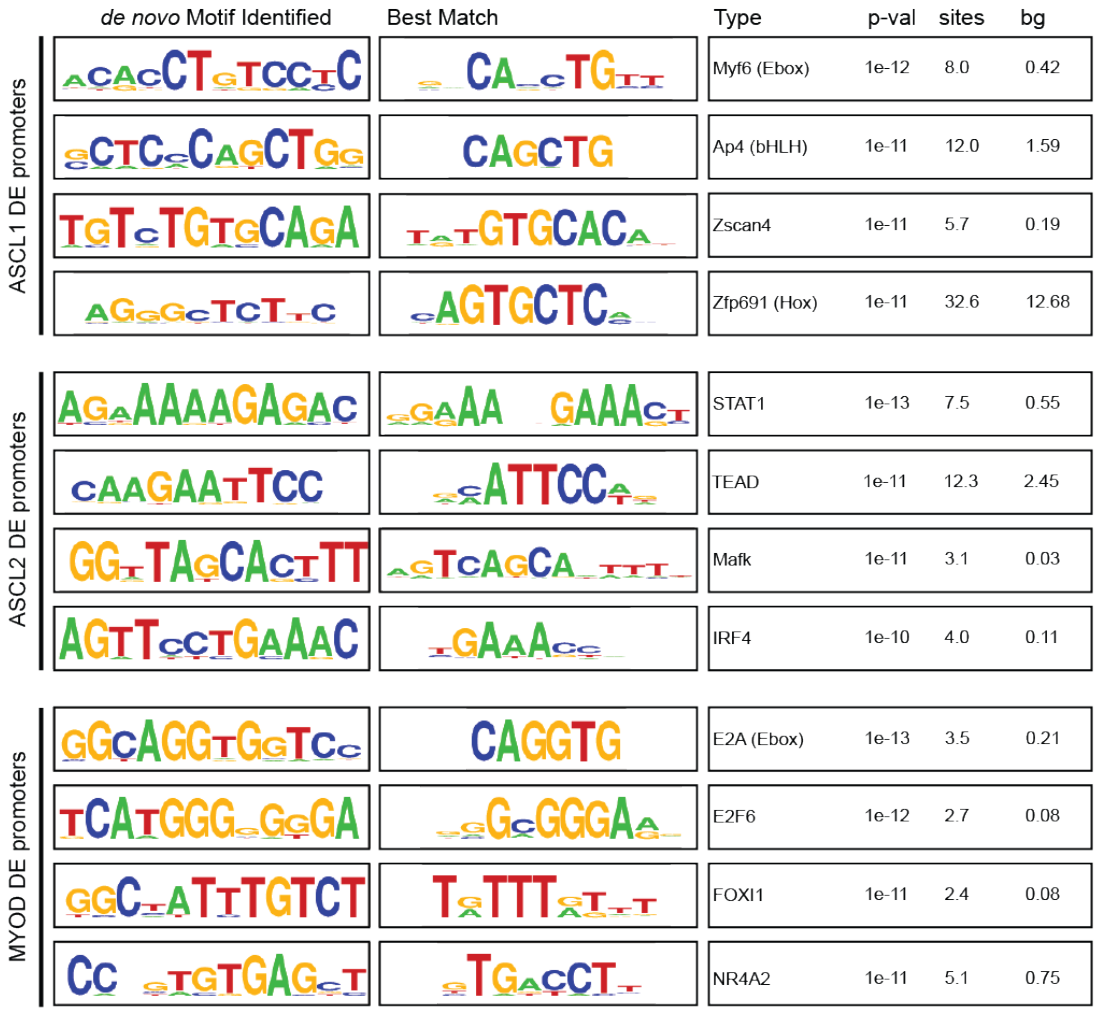


Selected GO categories highlight lineage-relevant genes identified as MyoD, and Ascl1-specific PDTs
Network shows extent of overlap between genes identified in each GO category (distance) and number of genes identified (size) for the total set of factor-specific PDTs identified for MyoD or Ascl1-expressing cells as indicated. Each gene shown meets expression criteria of RPKM≥1, FC≥2, FDR≤0.05, and is identified only in Ascl1 or MyoD expressing cells, respectively. *Ascl1* is identified as a potential target in these analyses, as it was associated with a ChIP-seq binding site 6kb 3' of the *Ascl1* locus.

Figure 3-24: Comparison of significant *de novo* motifs at peaks associated with PDT.

Binding sites associated with potential targets do not demonstrate additional motif specificity
Comparison of *de novo* motifs identified at peak regions associated with differentially expressed genes showing increased expression at 24h post-induction in RNA-seq (from EdgeR analysis). Motif shown represents PWM associated with best match for enriched motif. Best Match represents most significantly similar motif identified by HOMER. Statistics reflect the binomial significance of the *de novo* motif identified, as compared to random background, with percentage of binding sites, and and background regions featuring the motif depicted, as indicated. Motifs generated from 150bp surrounding peak center. (pval < 1e-12 identified as possible false positives by HOMER)

Figure 3-25: Comparison of *de novo* motifs identified at promoters of DE genes.



Comparison of de novo motifs identified at promoters of differentially expressed genes showing increased expression at 24h post-induction in RNA-seq (from EdgeR analysis). Motif shown represents *de novo* PWM associated with best match for enriched motif. Numbers reflect the hypergeometric significance of the de novo motif identified, the percentage of sites featuring the specified motif, and the percentage of promoter-specific background featuring the specified motif.

Table 1-1: Table overview of *Ascl1*, *Ascl2*, and *MyoD* expression as derived from RNA-seq

bHLH tested	ctrl	ind (24h)	FC
Ascl1	0.6	191	337x
Ascl2	2.0	358	181x
MyoD	1.1	250	235x

ES cells show robust induction of bHLH factor transcription in RNA-seq
Table compares bHLH mRNA transcript expression from RNA-seq data from ASCL1, ASCL2, and MYOD-expressing ES cells. in uninduced controls and 24h induced ES cells from each cell line. Values represent mean RPKM of three independent biological replicates for each experiment. Fold Change ([mean ind]/[mean ctrl]) as indicated.

Table 3-2: comparison of known targets of bHLH factors

	Average RPKM						Average FC			FDR (edgeR)		
	ASCL1		ASCL2		MYOD							
Symbol	ctrl	ind	ctrl	ind	ctrl	ind	Ascl1	Ascl2	MyoD	Ascl1	Ascl2	MyoD
Dll3	1.6	21.3	2.8	9.1	2.1	1.4	13.2	3.3	0.6	4.03E-26	5.77E-06	5.36E-01
Hes6	8.3	75.0	20.6	146.5	21.8	280.7	9.1	7.1	12.9	5.70E-21	2.93E-17	3.00E-28
Dll1	1.3	10.0	2.4	11.3	1.9	8.4	8.3	4.7	4.5	1.47E-17	7.91E-10	7.63E-10
Parp12	5.4	21.0	9.6	49.7	9.0	41.3	3.9	5.2	4.6	5.17E-17	1.77E-25	2.96E-22
Lfng	3.1	16.4	3.1	13.8	4.1	31.2	5.2	4.4	7.7	4.26E-12	8.34E-10	8.35E-19
Fgf5	0.3	4.1	0.9	8.8	1.1	17.9	15.3	9.4	16.3	2.46E-09	7.90E-07	6.63E-11

bHLH factor expression in inducible ES cells leads to differential expression of known targets within 24h. Values shown reflect the mean RPKM from RNA-seq across three biological replicates. FC represents the mean RPKM in induced ES cells at 24 hours post-induction versus the mean RPKM in uninduced control cells from same experiment. FDR values shown reflect the false discovery rate for these observations, as reported by edgeR across the three biological replicates for each experimental condition.

CHAPTER FOUR
Results

bHLH factors ASCL1, ASCL2 and MYOD function as pioneering transcription factors

Introduction

Class II bHLH transcription factors ASCL1, ASCL2, and MYOD are key developmental regulators which are crucial to establishing cell lineages in the developing embryo. To do so, these factors function in partially specified cell lineages, influenced by, and interacting with the environment of the cell. While functional gene networks provide the foundation for development and maturation, they fall short in providing a mechanism for some of aspects of cell lineage. As these bHLH factors function in partially differentiated cell lineages, it is possible that their binding and function are informed by features of these lineages, potentially providing a mechanism for the regulatory specificity demonstrated by these factors. Here, we directly compare the binding and functional specificity of ASCL1, ASCL2, and MYOD in when ectopically expressed in ES cells to test whether this specificity may be attributed to specific aspects of the chromatin landscape.

The binding of class II bHLH factors ASCL1 (Castro et al., 2011; Borromeo et al., 2014; Borromeo et al., 2016), ASCL2, (Liu et al., 2014; Schuijers et al., 2015), and MYOD (Cao et al., 2010) has previously been characterized in differentiated cell and tissues. These studies demonstrated unique binding in these cellular contexts, despite the identification of very similar primary Ebox motifs, suggesting that specificity in binding is dependent on

factors other than simple motif recognition. Cells derived from different lineages demonstrate different, lineage-specific patterns of open chromatin (Vierstra et al., 2014), representing one component of the chromatin landscape in which transcription factors must function. One theory about how transcription factors select their specific binding sites posits that binding is defined partially by the motif, and partially by access to this motif, in which they bind a specific DNA binding motif, but their interaction with potential binding sites is restricted by the chromatin environment. However, this cannot be the only mechanism underlying the specificity of bHLH function, as ASCL1 and MYOD, at least, have been previously demonstrated to have a functional capacity for reprogramming (Lassar et al., 1986; Davis et al., 1997; Weintraub et al., 1989; Farah et al., 2000; Vierbuchen et al., 2010; Wapinski et al., 2013).

Chromatin has been revealed to possess multiple mechanisms which promote or restrict both gene expression and the function of regulatory elements. While the genetic (nucleotide) sequence of DNA is the best understood aspect of transcription factor function, especially with regards to transcription factor binding, epigenetic features of the genome are increasingly identified as crucial components of transcriptional regulatory mechanisms. Examples include so-called chromatin modifications, which have been demonstrated to correlate with specific genic features, such as H3K4 monomethylation, identified at enhancer regions, H3K4 trimethylation at promoter regions, and H3K27 acetylation, which is identified in active chromatin. The DNA helix can undergo a specific modification of its nucleotide bases, termed CpG methylation, and this process has previously been demonstrated as necessary for development (Li et al., 1992), and central to a number of

disease states, including developmental pathology, oncogenesis, and neurological disorders (Amir et al., 1999). Finally, the structure of chromatin is itself a barrier to transcription factor function, as the presence of histone proteins occupies potential transcription factor binding sites. This restriction has previously been shown to have the ability to repress transcription factor activity, preventing the binding and function of these factors by regulating chromatin accessibility (Soufi et al., 2012).

An exception to these limitations is provided by a class of transcription factors known as "pioneering" factors. Pioneering factors, as conventionally defined (reviewed in Zaret and Carroll, 2011), possess the ability to bind to closed chromatin, initiate transcriptional changes, and modify the chromatin landscape. The canonical example of such factors is the FoxA family, which, along with GATA were identified through DNA footprinting as being able to bind to inaccessible enhancers, and initiate transcription (Gualdi et al., 1996). These factors demonstrate that unfavorable chromatin is not necessarily sufficient to restrict transcription factor binding. However, it has also been demonstrated that not all transcription factors possess pioneering ability (Cirillo et al., 2002). Thus, differential pioneering ability represents a potential mechanism defining transcription factor function.

One possible explanation for the ability of ASCL1 and MYOD factors to function in differentiated cell lineages and enact distinct transcriptional profiles is that these factors may function as pioneering factors. If this is the case, impediments to binding and function presented by the chromatin landscape present in partially or fully differentiated cell types may not be sufficient to restrict the binding and function of these factors. As it is known that distinct lineages demonstrate identifiable differences in chromatin accessibility through the

presence or absence of nucleosomes (Vierstra et al., 2014), the presence or absence of this open chromatin represents a likely candidate for differential regulation of transcription factor activity.

ASCL1, ASCL2, and MYOD bind to both open and closed chromatin when ectopically expressed in ES cells

To test whether binding of these bHLH factors is restricted based on chromatin accessibility, I performed ATAC-seq (Buenrostro et al., 2013; Buenrostro et al., 2015) on chromatin from the ES cells engineered to inducibly express *Ascl1, Ascl2, or MyoD.* ATAC-seq is a recent innovation which provides a readout defining open and closed chromatin genome-wide, and is based on differential integration of Tn5 transposase, a topoisomerase which targets open chromatin, and inserts a novel DNA fragment into its insertion locus. Using standard and quantitative PCR amplification, and multiplexed primer sets, ATAC-seq selectively replicates the DNA fragments between these Tn5 insertion sites, allowing for selective amplification of the open chromatin in a cell. ATAC-seq requires low cell numbers, and provides greater signal to noise ratios as compared to previous approaches to assaying open chromatin, such as MNase or DNAse hypersensitivity assays, or Formaldehyde-Assisted-Isolation-of-Regulatory-Elements (FAIRE-seq)(Simon et al., 2012).

Cells were collected as uninduced or induced by removal of doxycycline as in the ChIP-seq and RNA-seq studies in Chapter 3, and were prepared at 24 hours post induction. 50,000 cells were collected for ATAC-seq analysis. The data from these samples provided high read depth and complexity. The sequencing results of each sample were aligned to the

mouse mm10 genome (Kent et al., 2002; Kent et al., 2010), and processed for downstream analysis using Bowtie2 (Langmead et al., 2009), and HOMER (Heinz et al., 2010). Visualization of resulting ATAC-seq tracks on the mm10 UCSC genome browser (Kent et al., 2002; Kent et al., 2010; Raney et al., 2012; Rosenbloom et al., 2015) demonstrates that regions of the genome previously identified as open (Buenrostro et al., 2013) by conventional strategies, such as FAIRE-seq, are were also highly enriched for open chromatin in our data, suggesting that these data reflect expected enrichment of open chromatin (Figure 4-1: ATAC-seq enrichment in uninduced ES cells). This provides confidence in the quality of these data, and its usefulness in identifying the chromatin state present in the ES cells used in these experiments.

To test whether binding of ASCL1, ASCL2 or MYOD is restricted to regions of open chromatin, I examined the ATAC-seq signal in uninduced ES cells at the ASCL1, ASCL2, and MYOD bound sites identified in the induced cells in the ChIP-seq experiments (*Chap. 3*). I utilized HOMERs *annotatePeaks* module to specifically survey the normalized tag counts of ATAC-seq reads in 10bp intervals centered on bHLH binding sites identified in each induced ES line. The resulting data were then used to prepare heatmaps (Figure 4-2: Heatmap comparison of bHLH ChIP-seq and uninduced ATAC-seq). Much as the UCSC genome browser depicts the ATAC-signal in two dimensions, genomic position (x) and signal height (y), the heatmap depicts the specified subset of genomic positions at a defined interval around a given ChIP-seq peak, and the signal at each position as intensity, thus collapsing these data to facilitate observation of trends at many genomic intervals. Figure 4-2 shows the ATAC-seq signal at each bHLH binding site. The narrow, focal band representing

ChIP enrichment for each bHLH factor illustrates the called peaks in the relevant induced cells. ATAC-seq data from the uninduced cells over the same intervals reveals a central region of high signal representing enrichment for open chromatin. However, a significant fraction of these intervals do not demonstrate central enrichment, representing closed chromatin. This suggests that ASCL1, ASCL2, and MYOD bind both open and closed chromatin when ectopically expressed in ES cells, thus meeting one of the criteria defining them as pioneer transcription factors (Cirillo et al., 2002).

That ASCL1 and MYOD can bind at closed sites is crucial to understanding their capacity as reprogramming factors. It is generally believed that the process of lineage specification is partially dependent on repression of lineage-appropriate gene expression through chromatin modification. This provides insight into the mechanism by which reprogramming factors, including the bHLH factors studied here, are able to direct expression of lineage-inappropriate gene targets when ectopically expressed in differentiated tissues. If bHLH binding was restricted to open chromatin, this would suggest that the subset of bHLH binding sites crucial for their role in reprogramming must be open in source cells for binding and bHLH mediated reprogramming to occur. Our results show that these factors are able to bind to sites which are not previously marked by the presence of open chromatin; even regions which appear to have very high nucleosomal occupancy (closed sites) feature significant ChIP-seq peaks. The identification of a significant number of such sites suggests that even the presumably unfavorable binding environment presented by such closed sites does not appear sufficient to restrict the binding of these factors in this paradigm.

bHLH binding sites show distinct motif preference and distribution in open versus closed chromatin

Class II bHLH factors are known to function by binding to Ebox motifs throughout the genome. As described in *Chapter 3*, we previously identified a preference for a central GC core dinucleotide by *de novo* motif analysis, and confirmed this finding by detailed annotation and comparisons of Ebox distribution at ASCL1, ASCL2, and MYOD binding sites. This preference for GC-core Ebox motifs was shared across bHLH factors, and demonstrated dramatic central enrichment at bHLH binding sites, suggesting that this is the preferred binding motif for these factors when ectopically expressed in this paradigm. However, we found that even in this system, which minimizes potential confounds presented *in vivo* by lineage specification, these bHLH factors maintained distinct binding genome-wide, suggesting intrinsic binding specificity beyond the minimal Ebox motif, or the revealed preference for a specific motif.

Recently, *Soufi et al.* proposed that bHLH factors, including, but not limited to class II bHLH factors, might exhibit distinct motif preferences in open versus closed chromatin (Soufi et al., 2015). They reveal distinct *de novo* motif preferences for MYC in nucleosome-bound versus nucleosome depleted chromatin regions, when assayed by ChIP-seq and MNase-seq, an alternative measure of chromatin accessibility which reflects nucleosome occupancy, from embryonic fibroblasts (MEFs). Specifically, they find that binding sites present in nucleosome-depleted (open) chromatin demonstrated reduced degeneracy, suggesting that MYC's preference for DNA binding motifs is shaped in part by chromatin accessibility at binding sites in MEFs. They interpret this as evidence that the long α-helix of

the bH1 domain of MYC stabilizes binding at these sites, and is incompatible with the presence of nucleosomes. From this, they also compare previous structural and *de novo* binding motif comparisons of ASCL1 and MYOD (among others), and propose that differences in the bH1 domain alone are sufficient to direct differential preferences to nucleosomal binding, based on their protein structures (Figure 4-3: Proposed mechanism for distinction in pioneering capacity of bHLH factors ASCL1 and MYOD). Specifically, extrapolating from previous X-ray crystallography of NEUROD1:E47:DNA heterodimer, MYOD:MYOD:DNA homodimer, and (TAL) SCL:E47:DNA heteromer, they observe that the N-terminus of the bH1 domain is physical longer in MYOD than in ASCL1, and interpret this as a factor in increased motif specificity (decreased dinucleotide degeneracy) at MYOD binding sites, due to steric hindrance imposed by this extension. Additionally, based on modeling of bHLH:DNA complexes, they predict that the ability of bHLH factors to bind to nucleosome-occupied sites is limited by the increased length of the bH1 domain. Such a model predicts that MYOD is less capable of binding to closed chromatin as compared to ASCL1, which features a shorter bH1 domain. Furthermore, they suggest that the structure of ASCL1 predicts it to have differential pioneering capacity as compared to other bHLH factors, including MYOD, consistent with a recent report of ASCL1 functioning in this capacity in a fibroblast-to-neuron reprogramming paradigm (Wapinski et al., 2013). This supposed difference in the ability to bind to closed chromatin supports a model wherein differential preferences in chromatin accessibility may lead to differential binding of these factors, based on the structural differences present in their bH1 domains.

However, our study demonstrates that ASCL1 and MYOD both demonstrate the ability to bind closed chromatin in ES cells, and does not reveal clear differences in motif degeneracy at closed sites bound in this paradigm (see below for more detail on motif analysis at open versus closed chromatin). Additionally, ASCL2 appears to bind to fewer sites in closed chromatin, suggesting that either it is less capable of binding to such sites, or that its cognate sites are simply more open in ES cells, and that this specificity exists despite the dramatic similarity to the bH1 domain of ASCL1. While ChIP-seq identifies a higher number of ASCL1 binding sites (3188 ASCL1-bound vs. 2385 MYOD-bound), the level of chromatin accessibility, as measured by ATAC-seq, is comparable between ASCL1 and MYOD-bound sites. Furthermore, both "open" sites and "closed" sites are bound in considerable numbers. Together, these data suggest that the degree of chromatin accessibility does not appear to be a clear determinant in defining differential binding in ES cells.

To test whether ASCL1, ASCL2, and MYOD demonstrate differential motif preference in open versus closed chromatin, I used two different methods to characterize the motifs present in these subsets of sites. I first compared *de novo* motifs identified from the most, and least accessible chromatin regions, utilizing ChIP-seq and ATAC-seq data sets from each ES cell line at 24h post-induction (ChIP-seq) and uninduced samples (ATAC-seq). I generated matrices of ATAC-seq data on 6kb intervals, centered on the bHLH binding sites identified from ChIP-seq. An in-house PERL script was used to sort these regions based on their central 50bp, in decreasing order. We then informatically split these data sets to identify the highest and lowest ranked bHLH binding sites based on open chromatin, and performed *de novo* motif analysis on each set, using HOMER's *findMotifsGenome* module, and utilized

a length-specific seed to aid in identification of the primary Eboxes at these binding sites (-*len 8),* despite the lower number of sites in each set. This allows comparison of the primary Ebox motif from these subsets, and comparison to the primary binding motif identified from the total set of binding sites identified for each factor. I then utilized HOMER's *annotatePeaks* feature to specifically annotate the distribution of Eboxes within the intervals identified as peaks, as described above using parameters *-size 800 -hist 6 -rmrevopp 1*, which corrects for the palindromic CAGCTG motif.

Comparison of the distribution histogram plots at accessible and inaccessible sites shows a clear distinction in the spatial distribution of these motifs (Figure 4-4: ATAC-ranked bHLH Ebox Comparison). At the subset of "open" sites, CAGCTG motifs are slightly enriched as compared to CAGGTG/CACCTG motifs, similar to what was observed for the total set of binding sites identified for each factor. When compared to the subset of closed sites identified for each factor, a clear shift towards a GG/CC dinucleotide core is observed. Further, the total Ebox enrichment over these sites increased as compared to the subset of open sites. This suggests that numerically, more total Eboxes are associated with this set of closed sites. As we observe that the majority of bHLH binding sites identified by ChIP-seq possess at least one CAGSTG Ebox motif, peaks associated with closed chromatin appear to have a higher density of Eboxes.

Comparison of the motifs identified from these data sets reveals that each bHLH factor identifies a CAGSTG Ebox motif between open and closed sites. The subset of sites representing bHLH binding sites with the highest enrichment for open chromatin demonstrate modest preference for a GC-biased central dinucleotide motif. Compared to the

de novo motif identified from "closed sites", a shift in the second central dinucleotide position is noted. Whereas open sites exhibit a modest preference for GC-core dinucleotides, closed sets exhibit preference for a GG core dinucleotide, indicating a change in binding preference conferred by the presence of open or closed chromatin at these sites. Intriguingly, both open and closed sites bound by each bHLH factor showed enrichment for the disparate flanking motifs identified from the total set of sites bound by each factor (*i.e.* a preference for GCAGSTG in ASCL1 and ASCL2, and a preference for ACAGSTG in MYOD-bound sites), indicating that the preference for variant flanking motifs is not strongly influenced by the presence or absence of open chromatin at these binding sites, and that the presence of these variants is not restricted to open or closed chromatin. While there were subtle differences in the flanking motifs identified from the subsets of open and closed sites as compared to the total set of binding sites identified for each factor, these differences were relatively minor, and I interpret them as an artifact of the dramatic reduction in the number of templates used to perform this analysis. As one reduces the number of bHLH bound sites utilized in these comparisons, the effects of outliers on the resulting PWM are dramatically increased, and this is reflected in *de novo* motif analysis from small sets.

These results demonstrate that bHLH factors maintain the ability to bind both GC and GG/CC core Ebox motifs in open and closed chromatin, and do not support a model in which their binding in either the open or closed chromatin environment is restricted by dramatic differences in the preferred motif, as identified for MYC, and proposed for class II bHLH factors ASCL1 and MYOD (Soufi et al., 2015). However, the finding that ASCL1, ASCL2, and MYOD demonstrate a consistent shift in preference from a GC-core Ebox at open

binding sites to a GG-core Ebox at closed binding sites suggests that binding may be partially informed by distinct motifs between open and closed chromatin, which does support a model in which binding site sequence and chromatin accessibility each influence the binding of these factors. Indeed, the ChIP-seq and ATAC-seq data do consistently identify a trend towards local open chromatin at the binding sites of ASCL1, ASCL2, and MYOD binding sites as compared to adjacent regions (Figure 4-2: heatmap comparison of bHLH ChIP-seq and uninduced ATAC-seq).

Together, these data demonstrate that when expressed in ES cells, the binding of ASCL1, ASCL2, and MYOD is not differentially predicated on the presence of open or closed chromatin at their binding sites. Furthermore, they suggest that the structural differences between these factors do not appear to dramatically restrict the binding of any of these factors to either open or closed chromatin. This suggests that a model in which class II bHLH factors have dramatic differences in their ability to bind to closed chromatin is too simplistic to address the distinct binding seen between these factors, and that the intrinsic binding specificity observed for these factors is defined elsewhere.

Chromatin accessibility at bHLH binding sites is not predictive of lineage-specific gene ontology

bHLH factors are capable of directing expression of lineage-specific transcriptional profiles when ectopically expressed outside their normal developmental contexts (Lassar et al., 1986;

Weintraub et al., 1991; Farah et al., 2000; Vierbuchen et al., 2010; Wapinski et al., 2013). To test whether binding sites associated with open or closed chromatin might be associated with specific gene functions, I also compared the sets of genes located near these regions. I performed peak-to-gene calling on coordinates identified from the bHLH factor ChIP-seq at 24 hours post-induction, and ranked them by the presence of open or closed chromatin near the bHLH binding site, as described above for *de novo* motif analysis. I then utilized GREAT v3.0 (McLean et al., 2010), to identify the genes associated with binding sites in open or closed chromatin. As gene ontology analysis is highly dependent on statistical comparisons across data sets tested, I utilized the 1000 binding sites for each factor demonstrating the highest, and lowest enrichment for open chromatin, as measured by ATAC-seq. These were submitted for comparison based on the mm10 genome, using the default gene calling parameters, as previously described in Chapter 3.

The GO Biological Process (BP) categories identified for the subset of bHLH binding sites identified as open prior to binding varied between factors (Figure 4-5: Overview of GO categories enriched in open and closed bHLH binding sites). ASCL1 and MYOD open sites showed significant association with developmental GO categories, including stem cell maintenance (*Cdx2, Dll1, Esrrb, Hes1*, and others), presumably reflecting the environment in which these experiments were performed. Interestingly, only ASCL1 and MYOD bound sites which were open prior to binding demonstrated significant enrichment for any GO BP category, with ASCL2 failing to identify any significantly enriched (binomial qFDR <=.05) BP categories. A number of developmentally relevant genes that are increased in response to ASCL1, ASCL2, and MYOD, such as *Dll1, Hes6, Lfng, Fgfr2,* and *Hand1*, are associated

with open chromatin at the bHLH bound sites. This suggests that these developmental genes may be primed for rapid expression upon induction of the bHLH factors.

The most intriguing finding from GO analysis of the bHLH binding sites associated with closed chromatin came from ASCL1-bound sites (Figure 4-5: Overview of GO categories enriched in open and closed binding sites). This subset of sites identified significantly enriched ontological categories associated with lineage-relevant neural targets: Hindbrain morphogenesis and cerebellar cortex morphogenesis. The genes associated with these ontologies included *Agtpbp1, Cacna1a, Dab1, Dlc1, Gli2, Herc1, Hes1, Kndc1, Mtpn, Pcnt, Prox1, Rfx4,* and *Skor2*. This shows that some lineage-specific targets are associated with closed regulatory elements, and that ASCL1, at least, is able to bind to these closed sites, potentially regulating expression of these genes.

While comparison of the genes associated with binding sites in open chromatin suggests that many of the differentially expressed genes (DEGs) identified for these bHLH factors are associated with open chromatin, it is not the case that the presence of open chromatin at bHLH sites is required for transcriptional activity. One example is presented by bHLH binding sites for all three factors ~15kb 5' of the *Hes1* locus, which demonstrates no visible enrichment for open chromatin in any sample tested. *Hes1* is constitutively expressed in uninduced ES cells at an average RPKM of ~12, and its expression is increased 50-90% at 24 hours post induction of the bHLH factors. Additionally, a site ~1kb upstream of the *Hes1* TSS is enriched for open chromatin in uninduced ES cells, but is only significantly bound by ASCL1 with ASCL2 showing limited binding, and MYOD not binding this site at all. These findings highlight that the presence of open chromatin is not a primary determinant for bHLH

binding, and each factor retains specificity for binding, even against the backdrop of other potential sites nearby.

Together, these results demonstrate that bHLH factors are able to bind to both nucleosome-depleted "open" chromatin regions and nucleosome-occupied "closed" chromatin regions. The observation that highly enriched bHLH peaks are present at both ends of this spectrum suggests that they are relatively agnostic in binding with respect to the presence or absence of open chromatin regions, and that this characteristic appears to be largely shared between factors. Furthermore, we identify a shift in the preferred binding site from the GC-core dinucleotide preference revealed in ChIP-seq and identified at sites which appear to be open prior to binding, to an increase in the number of GG/CC-core dinucleotide sites identified in the subset of sites associated with binding to closed chromatin. It is possible that this shift is due to steric hindrance of bHLH/DNA binding between these two chromatin states, which in altering the binding geometry of the interface region, may lead to changes in the favorability of binding specific E-box motifs at these sites. Alternately, it may be the case that both sets of sites demonstrate a similar preference for the GC core Ebox revealed in our *de novo* motif analysis, and in agreement with conventionally identified motif preferences for each of these factors (Cao et al., 2010; Castro et al., 2011; Borromeo et al., 2014; Schuijers et al., 2015). At nucleosome-depleted sites, which are more accessible, even a single preferred motif may be sufficient for binding and transcriptional activity of these sites, whereas binding of less accessible sites may be improved by the existence of multiple adjacent Eboxes, perhaps of differing sequences, which together improve recruitment and retention of these, or other bHLH factors and co-factors, stabilizing bHLH/DNA binding.

Regardless, the identification of relatively similar preferences by the factors tested here suggests that this is not a primary mechanism in the binding specificity observed for these factors.

ASCL1, ASCL2 and MYOD increase chromatin accessibility at binding sites identified in ES cells

The results of bHLH ChIP-seq and ATAC-seq demonstrate that bHLH binding is not predicated on the presence of open chromatin in the context of ES cells. In addition to testing hypotheses regarding the influence of chromatin state on bHLH binding, we tested whether ASCL1, ASCL2, or MYOD might have the ability to themselves alter the chromatin landscape. To test this possibility, we compared ATAC-seq data from uninduced and bHLH induced ES cells to test their ability to modify the chromatin landscape.

To accommodate variation in the number of sequencing reads present in these samples, and allow for direct comparison, I first randomly subsampled 11.4M reads from each data set, to match the sample which had the lowest number of uniquely aligned mapped reads (24 h induced *MyoD* ES cells). I then used the HOMER findMotifsGenome module to identify differential peaks, comparing induced versus uninduced, and vice versa, using a relatively strict filtering cutoff (cumulative Poisson p-value 1e-06). This identified between 807 and 1734 differentially called peaks which met criteria for local and genomic filtering. Visualization of identified peaks on the UCSC genome browser showed visible changes at the intervals identified, demonstrating that the majority of these peaks appear to reflect objective differences between induced and uninduced ATAC-seq data sets for these samples

(Figure 4-6: Regions identified by differential peak calling from ATAC-seq). While some sites demonstrated modest changes, or changes which appeared to correspond to regions with general enrichment rather than focal enrichment, the sites called generally appear to reflect narrow, site-specific changes in open chromatin.

As cursory inspection of sites demonstrating bHLH-dependent changes in open chromatin revealed that some of these changes appear to occur directly at bHLH binding sites identified in ChIP-seq, I tested to what extent ATAC-seq changes overlapped with bHLH ChIP-seq binding sites identified in ES cells. To make this comparison, I utilized mergePeaks (Heinz et al., 2010), which observes peak regions identified from aligned fragments, and generates lists of shared and unique coordinates based on their overlap, as defined by a distance parameter. Surprisingly, even using a maximal overlap of 300bp, bHLH binding sites demonstrated relatively low overlap with the sites identified as changed in ATAC-seq (Figure 4-7: Overlap comparison of ChIP-seq peaks vs. local increases in ATAC-seq). ~100 overlapping sites in each cell line were identified, representing less than 15% of the total sites showing an increase in the bHLH induced samples. Of these overlapping sites, the majority were identified in only one cell line, suggesting that these changes may represent early effects of bHLH binding at these sites, or that differences in the sensitivity of these assays are masking the extent of overlap.

The identification of sites demonstrating local increases in ATAC-seq signal which do not correspond with bHLH binding sites led me to test what might mediate changes at these sites. Because ASCL1, ASCL2 and MYOD function as lineage-specific master regulators of transcription, it is reasonable to suspect that some of the changes in open

chromatin might occur as a result of transcriptional cascades and not be due to direct binding of the bHLH. To test whether these sites might reflect the effects of identifiable downstream regulators, I again performed *de novo* motif analysis on the sets of sites demonstrating bHLH-dependent increases or decreases in open chromatin. As ATAC-seq peaks represent nucleosome positions rather than specific transcription factor binding sites, they often appear broader, and clustered together. This complicates conventional approaches to peak calling and motif analysis, which rely on algorithms designed to identify the rare, focal enrichments associated with transcription factor binding. To accommodate this, I utilized a genomic interval of 200bp, centered on the region changed in ATAC-seq, with the goal of identifying significant primary and secondary motifs.

An Ebox was identified as the primary motif identified from each set of sites demonstrating a bHLH-dependent increase in open chromatin. As with the *de novo* motif identified from ChIP-seq data, each was significantly (p-value of 1e-116 to 1e-207) enriched for a CAGSTG Ebox (Figures 4-8 & 4-9: comparison of primary Ebox motifs identified at ATAC-seq changes). These motifs were present in 31-55% of target sites, representing a ~3-7 fold increase over background. These numbers would be considered relatively low for ChIP-seq data for a factor with a well-defined binding motif, but as they reflect ATAC-seq data, this enrichment suggests that these sites contain a considerable number of motifs for these or other Ebox binding transcription factors. Remarkably, these sites not only identify an Ebox as the primary motif, but identify Eboxes which strongly resemble those bound by ASCL1, ASCL2 and MYOD. Nevertheless, the factor specific ChIP-seq data show these

factors bind relatively few of the sites identified as changed in ATAC-seq between uninduced and induced conditions.

To identify potential mechanisms which might explain the changes in open chromatin identified, I compared the lists of secondary motifs associated with these changes. In addition to the primary Ebox motifs, a number of significant secondary motifs were identified in each cell line tested. In ASCL1-expressing cells, an E2F motif resembling E2F6 was identified as significantly enriched. E2F6 appears to have a role as a component of transcriptional repressor complexes, and may interact with chromatin remodeling enzymes (Leung et al., 2012)). While the overall enrichment for this motif is perhaps underwhelming, it stands out due to this potential functional role. Motifs resembling Fox, Runx, and Sox motifs were also identified, but demonstrate marginal enrichment in this set of sites. In ASCL2, modest enrichment is noted for a Sox motif, but shows considerable degeneracy. Given the relatively short size, and high degree of stereotypy, at least for the core Sry-box which Sox factors bind, this may represent a false positive. However, it does resemble known Sox motif examples, and considering the significance of Sox factors in pluripotency, stem cell biology, and development, its presence is worth noting. FoxO is also identified in ~25% of peaks vs. 14% of background, showing a strong matrix. Pou2f2 (*Oct2*) is also present, but only at 4% of sites, and with a significance of p 1e-16, which we consider marginal for motif analysis, especially for short motifs. MYOD identifies Pbx3 and Oct11/Pou as significantly enriched, with modest enrichment for MEF2A/MEF2D (which share a motif). Given the significance of MEF and Pbx factors in myogenic lineages, and their selective increase in expression in

the RNA-seq data from the factor induced ES cells, this may represent a meaningful finding despite the relatively low number (29) of sites identified with this motif.

Comparing the motifs identified at regions which showed diminished ATAC-seq signal upon induction (potentially bHLH dependent chromatin closing/repressive events), we find that roughly 900 sites are identified with such a reduction in ASCL1 and ASCL2, and ~1500 sites in MYOD in cells at 24h post induction. Strikingly, in contrast with sites demonstrating induction-dependent increases in ATAC-seq signal, E-boxes were not identified at sites showing diminished ATAC-seq signal. As with the ATAC-increased regions, Sox motifs are identified, but the significance and percentage of targets were low (10%-20%, with roughly two-fold enrichment over background). *Oct*/Pou motifs were identified in relatively few instances as well. While the presence of these motifs is not in itself meaningful, they may represent changes relevant in ES cell differentiation and maintenance of pluripotency (given the roles that Oct/Sox factors play in these processes). Considerable enrichment for a CAGTCA motif is noted in all three cell lines, called as AP-1 in ASCL1 and ASCL2 induced cells, and Atf3 in MYOD induced cells. While differently identified, they appear to represent the same motif, and are called in ~10% of the regions surveyed. AP-1/c-Jun is a cell cycle regulator and proto-oncogene that interacts with Fos, ATF-2, MAPK8, MAPK9, and UBC, and is an immediate-early transcription factor, which responds to cAMP signaling. ATF3 is also a response TF involved in cAMP signaling, with many isoforms noted, which appears to repress transcription by stabilizing binding of inhibitory co-factors p65 and HDAC1 to promoter elements (Kwon et al., 2015). Thus, it appears that the most significantly enriched motifs represent potential binding sites for early

response genes involved in cellular response signaling and transcriptional regulation. While the motifs identified do not provide clear evidence of potential bHLH-dependent function at these sites, a striking similarity is apparent. Only TEAD, identified in ASCL2 induced cells, and GATA in MYOD induced cells are uniquely identified. The remaining motifs present analogous entries in each data set, suggesting that changes at these sites may be regulated at least in part through a common mechanism. Thus, the most notable finding from this comparison is the discrepancy in E-box enrichment, which suggests that while some induction-dependent decreases in open chromatin are identified, they are not directly associated with canonical bHLH binding.

Together, the results of *de novo* motif analysis demonstrate that the regions demonstrating bHLH dependent changes in open chromatin at 24 hours post-induction are poorly enriched for co-factor motifs. Thus, no specific DNA-binding co-factors for the bHLH factors at this early stage of induction have been identified from this analysis. However, some of the moderately enriched secondary motifs, such as E2F6 in ASCL1-expressing cells, and Pbx3 and MEF motifs in MYOD-expressing cells suggest potential mechanisms for bHLH-dependent changes in chromatin remodeling and gene expression. Furthermore, several of the motifs associated with regions demonstrating bHLH-dependent decreases in open chromatin, including Pou5f1 (*Oct4*), Sox/Sry motifs, and Klf motifs are compelling, as they are also components of the reprogramming cocktail which confers pluripotency onto differentiated cell types in iPS reprogramming (Takahashi and Yamanaka, 2006). Their identification here may reflect early changes in bHLH-mediated suppression of the transcriptional profile of ES cells, although how this occurs remains unclear.

bHLH direct and indirect mechanisms direct chromatin accessibility changes upon bHLH expression

The relatively low apparent overlap between bHLH factor binding and changes in ATAC-seq are especially perplexing in light of the observation of Ebox motifs as the most significantly enriched motif at sites demonstrating a local increase in open chromatin. One possibility is that these changes reflect the indirect effects through early targets of these factors, or general response to their induction. Alternatively, these changes may reflect direct effects of bHLH binding which escape identification in our ChIP-seq, such as transient or weak binding events. If so, this might suggest that our ChIP-seq peak calling was overly stringent. Alternately, the changes seen may reflect the consequences of binding of other bHLH factors to the Eboxes identified.

To test whether the sites identified as changed in ATAC-seq were indeed occurring at sites other than those bound by these bHLH factors, heatmaps which compare bHLH ChIP-seq enrichment to the ATAC-seq enrichment present, specifically at genomic sites demonstrating local changes in open chromatin were generated (Figure 4-10: Heatmap comparison of bHLH ChIP-seq and ATAC-seq at sites demonstrating changes in open chromatin). These plots illustrate the intervals identified by differential peak calling in the ATAC-seq in uninduced versus induced cells show dramatic gain, or loss of signal at the center of the regions identified. bHLH ChIP-seq signal at these intervals recapitulates the relatively low overlap with ATAC-seq changes identified by informatic comparisons of the peak regions identified in these data sets (see Figs. 4-6, 4-7). This supports our finding that

many of the changes in open chromatin are not directly associated with observable binding of ASCL1, ASCL2, or MYOD in their respective cell lines. Furthermore, these data demonstrate an obvious difference in the bHLH ChIP-seq signal associated with regions showing increases versus decreases in open chromatin. This is in agreement with the finding from *de novo* motif analysis that regions showing bHLH-dependent increases in open chromatin are enriched for Ebox motifs, whereas regions showing decreases in open chromatin are not. Thus, many of the changes in chromatin accessibility appear to be indirect effects of bHLH induction.

Another distinguishing feature of pioneering factors is that by binding to closed chromatin, they are able to displace nucleosomes directly at their cognate binding sites, therefore increasing local chromatin accessibility at these sites. To quantify the aggregate changes in open chromatin identified at bHLH binding sites, I directly compared the ATAC-seq signal at these sites between the induced and control samples, using HOMERs *annotatePeaks* module to calculate the mean ATAC-seq tag height at each position across a 2kb region, centered on the bHLH binding site. I then plotted the mean ChIP-seq tag count along with these results to compare the mean ATAC-seq signal between the uninduced control, and the 24 hour induced sample, at every bHLH binding site identified (Figure 4-11: Histogram comparison of ATAC-seq signal at bHLH binding sites). In each ES cell line, an increase in the mean ATAC-seq signal is detected, which directly corresponds to the peak center identified by ChIP-seq. This is compatible with a model in which ASCL1, ASCL2, and MYOD directly displace nucleosomes from their respective sites, as defined in the model for pioneering factors (Zaret & Carroll, 2011). However, the observation that each of these

factors binds to both open and closed is in contrast to the differential pioneering model of bHLH binding specificity proposed in *Soufi et al., 2015* (summarized in Fig. 4-3). Instead, these data reveal that ASCL1 and MYOD are equally capable of binding to nucleosomal chromatin, and that each is capable of directing increases in chromatin accessibility.

Genes identified as differentially expressed in response to bHLH induction do not show clear changes in open chromatin

Ectopic expression of ASCL1, ASCL2, and MYOD is sufficient to induce changes in gene expression within 24 hours. These genes included known targets of these bHLH factors, but did not demonstrate dramatic lineage-specific or lineage-directive profiles of gene expression. As *de novo* motif analysis performed on sets of regions demonstrating bHLH-dependent changes in open chromatin primarily identified enrichment for Ebox motifs, one compelling possibility was that these changes might reflect vivication of inactive enhancer regions in ES cells. Increases in open chromatin may potentially provide clues as to the location of such enhancers, and insight regarding potential regulatory targets. Such a model would allow for discrete regulation of progressive sets of gene targets, providing additional direct regulatory capacity for these bHLH factors.

To test whether the apparent change in open chromatin was reflective of changes in gene expression which might not be readily associated with ChIP-seq binding sites, I first compared the genes associated with changes in open chromatin to the lists of differentially expressed genes (DEGs) identified in RNA-seq. Interestingly, this comparison demonstrated that relatively few genes of the genes associated with ATAC-seq changes were associated

with differential expression at 24h post-induction by RNA-seq (Figure 4-12: Comparison of genes associated with increased in open chromatin). This suggests that changes in open chromatin are not broadly associated with regulation of early targets of bHLH factor activity. Pathway and GO analysis on the lists of genes meeting these criteria for each cell line was performed using CPDB (Kamburov et al., 2009), which identified a number of significantly enriched ontological categories including neurogenesis and neural development in ASCL1-expressing cells, and muscle cell differentiation and muscle cell development in MYOD-expressing cells. Few of these genes demonstrate factor-specific expression, and in general did not demonstrate strong lineage-associated function. Notable exceptions included *Gcm1* in ASCL1-expressing cells, as well as *Smarcd3*, *Mamstr*, *Mef2d*, *Rb1*, and *Ski*, which have roles in myogenic lineages, and were associated with increases in open chromatin only in response to MYOD induction, and demonstrated factor-specific expression in MYOD-expressing cells. Together, the regions demonstrating changes in open chromatin, and genes identified through this approach included intriguing candidates for potential downstream analysis as components of the transcriptional program of these bHLH factors.

Another possibility is that loci of differentially expressed genes might themselves be associated with local changes in open chromatin; bHLH factors may initiate recruitment of transcriptional or regulatory complexes to promoter regions of transcriptional targets. To test whether differentially expressed genes demonstrated factor-specific changes in open chromatin, I utilized an in-house Linux shell script to compare the distribution of open chromatin at the loci of DEGs in response to each bHLH factor tested, in 24 hour induced and uninduced conditions (Figure 4-13: Distribution of open chromatin at differentially

expressed genes). The results of this comparison show that the TSS of both positively and negatively regulated genes are enriched for open chromatin, as expected based on previous characterizations of open chromatin through FAIRE-seq (Song et al., 2011) and DNAse assays (Guenther et al., 2007). No clear change in open chromatin was observed at the loci of DEGs, suggesting that any transcription-dependent increase in open chromatin at these sites is limited. Overall, enrichment for open chromatin at the TSS of positively and negatively regulated genes appears similar at 24h post-induction, and this pattern does not differ between bHLH factors tested. Together, these results suggest that loci of differentially expressed genes are not associated with early changes in open chromatin, and that this does not represent a clear mechanism for factor-specific regulation of transcription.

bHLH factor binding is informed by the presence of H3K27ac at potential binding sites

One of the most significant histone modifications with regards to enhancer function is H3K27ac, which has been consistently identified at active enhancers (Wang et al., 2008), and is believed to differentiate active versus poised chromatin (Creyghton et al., 2010). One possible mechanism by which bHLH factors select their complement of binding sites would be differential preference for previously established active or poised enhancer regions. I performed ChIP-seq for H3K27Ac from the ES cells before and after 24 hours of bHLH induction. To test whether bHLH binding sites are enriched for H3K27ac, I utilized the same approach used to compare bHLH ChIP-seq and ATAC-seq data sets. Using HOMERs *annotatePeaks* module, I generated heatmaps of the bHLH and H3K27ac ChIP-seq data sets, comparing the signal intensity on 6kb intervals surrounding the peak centers of every bHLH

ChIP-seq peak identified within a given cell line. Using an in-house script, these intervals were ranked based on their central enrichment, and plotted for comparison to bHLH ChIP-seq to observe the pattern of enrichment surrounding bHLH binding sites identified from their respective cell lines. These results reveal the presence of H3K27ac enrichment at a subset of intervals surrounding the peak centers of bHLH binding sites identified by ChIP-seq (Figure 4-14: heatmap comparison of bHLH and H3K27ac ChIP-seq at bHLH binding sites). The distribution of H3K27ac enrichment reveals a bimodal distribution with central depletion, centered on the peak sites identified by ChIP-seq for the bHLH TF. This is representative of previously published data sets revealing such a distribution for a number of common histone marks (Nie et al., 2013), and is believed to reflect the location of nucleosomes adjacent to transcription factor binding sites (TFBSs). This pattern is found in H3K27ac ChIP-seq in each of the cell lines tested, but is noted to be less distinct at the binding sites identified for MYOD. Thus, while some bHLH-bound sites are identified which lack enrichment for H3K27ac, a considerable number of sites are enriched for this mark of active chromatin, and show a distribution suggestive of nucleosomes adjacent to the binding sites identified

The presence of H3K27ac does not predict bHLH binding or transcriptional changes of nearby genes

To test whether the sites enriched for H3K27ac were of specific functional or mechanistic relevance with regards to bHLH factor function, I first characterized the degree to which bHLH factor binding overlapped with H3K27ac-enriched regions identified in

ChIP-seq. Peak calling on the uninduced H3K27ac ChIP-seq data sets using HOMER was performed. To accommodate the bimodal distribution expected for this histone mark, I utilized the nucleosome-free region parameter [-*nfr*] to preferentially center these peaks on regions showing this morphology. This approach identified ~25,000 regions of H3K27ac enrichment genome-wide. These intervals were compared to the bHLH binding sites identified by ChIP-seq. Using a relatively broad overlap interval of 500 base pairs to better reflect the broad regions identified in ChIP-seq for histone markers, a few hundred overlapping sites were identified in each ES cell line, which mirrors the distribution observed in the heatmap comparison (Figure 4-15: Proportion of bHLH sites associated with H3K27ac enrichment). To test whether the subset of bHLH binding sites enriched for H3K27ac were informative of bHLH factor function, I then compared the genes associated with the H3K27Ac enriched bHLH TF bound sites with the relevant RNA-seq data using GREAT. The majority of the genes associated with H3K27ac enrichment were expressed at appreciable levels prior to induction. Having observed the presence of H3K27ac near a number of genes identified as potential direct targets, including *Dll1, Dll3, and Hes6*, I then tested whether the bHLH-dependent DEGs might be associated with these overlapping sites. This comparison demonstrated that while several hundred genes are associated with these sites, relatively few of the bHLH-dependent DEGs were associated with H3K27ac in uninduced ES cells, suggesting that while some targets may be regulated by H3K27ac-enriched enhancer regions, the presence of H3K27ac at these sites this is not the central mechanism in defining transcriptional targets of bHLH factors. Together, these results demonstrate that bHLH factor binding is accompanied by H3K27ac enrichment at a subset of

binding sites in ES cells, but fail to demonstrate a dependence on the presence of this mark for binding, suggesting that neither binding, nor transcription is predicated on its enrichment at the sites.

bHLH factor induction leads to changes in H3K27ac at binding sites within 24h

ASCL1, ASCL2, and MYOD are known transcriptional activators (Davis et al., 1987; Johnson et al., 1992; Guillemot et al., 1994), and ASCL1 and MYOD have been reported to bind p300, a histone acetyl-transferase responsible for acetylating H3K27 (Sartorelli et al., 1997; Vojtek et al., 2003). This predicts that H3K27Ac will be increased at the bHLH bound sites upon induction, relative to the control ES cells. To test this prediction, we compared H3K27ac enrichment at the bHLH bound sites in control and bHLH-induced ES cells (see Figure 14: Heatmap comparison of bHLH and H3K27ac ChIP-seq from uninduced and 24h induced ES cells). Heatmap comparisons illustrate that the H3K27ac signal in induced cells is increased surrounding bHLH ChIP-seq peaks relative to that in the uninduced cells. This comparison reveals that bHLH binding both increases existing H3K27ac, and leads to enrichment of this mark at sites not enriched for H3K27ac prior to bHLH expression. To explore this further, I used differential peak calling to identify intervals showing significant increases in H3K27ac enrichment genome-wide, intersected the resulting intervals with the bHLH binding sites, and compared the change in H3K27ac enrichment at these sites (Figure 4-16: Histogram comparison of bHLH and H3K27ac enrichment at overlapping sites). There is a clear increase in H3K27Ac at the bHLH bound sites. More bHLH bound sites show increases versus decreases in H3K27ac (Figure 4-17: overlap comparison of bHLH binding

sites and H3K27ac changes identified in ES cells). This comparison also revealed that a relatively small fraction of sites featuring changes in H3K27ac enrichment were associated with bHLH binding sites identified in ChIP-seq. This is unexpected, based on the increase in H3K27ac signal noted when compared specifically at these intervals (Figure 4-16), and suggests that while H3K27ac is clearly increased at bHLH binding sites, this enrichment is modest compared to other sites at 24h post-induction. This may reflect early observation in progressive deposition of this mark. As observed at bHLH-dependent increases in ATAC-seq, *de novo* motif analysis identifies E-box enrichment at sites with increased H3K27ac (Figure 4-18: Comparison of de novo motifs identified at sites demonstrating increased H3K27ac), further implicating these, or other bHLH factors, in these increases. Broadly, our results demonstrate that bHLH factors can direct increases in the active enhancer mark H3K27Ac at a subset of binding sites within 24h after induction.

bHLH factor binding in ES cells does not appear to be associated with a trivalent chromatin signature, in contrast to results observed in fibroblasts

I demonstrated that ASCL1, ASCL2, and MYOD are each able to bind to largely distinct sets of sites, and that this binding is not restricted to open sites. While this helps to address how these master regulatory factors are able to effect the dramatic changes needed to reprogram differentiated cell types, it does not resolve the intrinsic specificity which underlies their distinct binding, nor does it explain the ability of these factors to identify the minor subset of potential Ebox binding sites genome-wide. Recently, *Wapinski et al.*

characterized the binding of ASCL1 in cultured mouse embryonic fibroblasts (MEFs) and found that binding in this context largely resembled that of ASCL1 in neural progenitor cells (NPCs). Using FAIRE-seq, they determined that ASCL1 binding sites were locally enriched for nucleosomal occupancy, suggesting that ASCL1 was preferentially binding closed sites despite apparently unfavorable binding conditions. Using previously published ChIP-seq data comparing histone modifications in MEFS, they identified a trivalent signature of H3K4me1, H3K27ac, and H3K9me3 enrichment at sites bound by ASCL1 in NPCs, and suggested that this combination may be predictive of ASCL1 binding in MEFs. Such a finding is unexpected, as H3K9me3 is widely associated with repressive chromatin, and specifically associated with repression of lineage-specific genes through gene silencing. To test whether this, or alternative signatures might be predictive of bHLH binding in ES cells, we used an unbiased approach to model the chromatin landscape at bHLH binding sites.

The complexity of the chromatin landscape precludes direct observation of all possible combinations of such signatures, and rational combinatorial models may not identify seemingly unlikely combinations. To address this challenge, a methodology known as Hidden Markov Modeling (HMM) is utilized, which performs sequential comparisons of potential combinations, and attempts to identify subsets of events which characterize the combinations identified based on their empirically revealed probabilities (Rabiner & Juang, 1986; Rabiner, 1989). This approach allows for the unbiased identification of potentially relevant combinations without relying on *a priori* expectations based on previous characterization of histone marks. To test whether bHLH binding sites identified by ChIP-seq in ES cells were associated with an identifiable signature, ChromHMM (Ernst & Kellis,

2012) was used to apply this technique to our study. To build a model of potential chromatin states, we utilized ATAC-seq and H3K27ac ChIP-seq data sets from uninduced ES cells, and combined them with previously published ChIP-seq data sets for H3K4me1, H3K4me3, H3K9me3, H3K9ac, and H3K36me3, which were generated by the ENCODE consortium as part of the mouse ENCODE database (The Mouse ENCODE Consortium, 2012). These data sets were selected as they met ENCODE quality standards, and were generated from murine ES-E14 cells, which, like the inducible ES cells used in this study were derived from *Mus musculus* strain 129 animals, thus representing the most comparable cell type represented in the ENCODE database. These data sets were downloaded directly from the ENCODE repository and processed based on the same in-house pipeline used in the analysis of my genomic data sets for direct comparison to the ChIP-seq data sets from inducible ES cells. To build the Markov model, replicate data sets were paired based on the specific histone mark used for ChIP, and undirected learning was performed based on an 11 state model, which was selected based on observations from multiple state trials. The 11-state model was selected as it appeared to allow sufficient complexity to observe apparent differences in histone marker distribution, and additional states did not demonstrate additional informative complexity. This model describes the chromatin states identified from these histone data sets genome-wide, and allows for observation of these states at genomic intervals of interest, such as bHLH binding sites.

The states identified in these models are best compared by observing the emission diagram (Figure 4-19: comparison of emission states identified in Markov model). This diagram compares the degree to which the states identified are associated with each of the

marks utilized in the model learning algorithm. This diagram reveals the presence of a number of relevant chromatin states of interest. State 1, which is almost entirely exclusive from the other states identified, is characterized by its association with the repressive mark H3K9me3. State 3 shows strong association with open chromatin, as measured by ATAC-seq, and also modest association with the enhancer-associated H3K4me1, thus representing both open chromatin and the presence of this enhancer-specifying modification. States 4 and 5 both demonstrate association with several markers of active chromatin, as well as high chromatin accessibility, and differ primarily in the presence of H3K27ac, suggesting that these may represent poised (state 4) and active (state 5) chromatin intervals. State 6 is defined by H3K27ac and H3K4me1, and shows considerable association with open chromatin, but lacks the promoter mark H3K4me3, suggesting that this state represents active enhancers in uninduced ES cells. State 11 appears to represent a "Pan-Active" profile, and is associated with signal enrichment in all activity-correlated data sets. Together, the 11 states present in this model represent the sum of all chromatin regions in the genome.

Comparing sites identified in bHLH ChIP-seq, these data reveal a trend on the interval surrounding the peak centers (Figure 4-19: Comparison of HMM states at bHLH binding sites). ASCL1, ASCL2, and MYOD each reveal modest central enrichment for states 3 and 4, and considerably higher enrichment for state 5. Thus, binding of these factors is associated with local enrichment of open chromatin, as well as states suggestive of active promoter and enhancer regions. State 6 is also centrally enriched, particularly at ASCL1 and MYOD-bound sites, indicating that these sites represent a greater association with the histone signature classically attributed to active enhancer regions. States 7, 9, and 11 are also slightly

enriched on the 2kb interval surrounding these binding sites, implying that these states are broadly present at these sites, but not restricted to the region immediately adjacent to the binding site. In addition to comparing the full sets of binding sites for these factors, I compared the states identified at factor-specific and shared binding sites to test whether these states might inform differential binding, potentially implicating histone signatures for the distinct patterns of genome-wide binding identified for these TFs. Broadly, the state profiles at these binding sites were largely reflective of the total sets of sites bound by each factor. Factor-specific binding sites identified for ASCL1 and MYOD showed a modest decrease in enrichment for state 6. Together, these results demonstrate that bHLH binding sites in ES cells are associated with markers of active chromatin, but that these associations appear insufficient to explain the distinct patterns of genome-wide binding seen for these factors in this context.

Notably, neither ASCL1, nor the other factors tested were found to be strongly associated with state 1 chromatin, which is defined by the presence of H3K9me3, and, minimally, H3K27ac. Recently, *Wapinski et al.* observed ASCL1 binding in the context of a fibroblast-to-neuron reprogramming paradigm. In this study, they found that a transcriptionally conflicted signature of active marks H3K4me1, H3K27ac, and the repressive mark H3K9me3 was predictive of ASCL1 binding (Wapinski et al., 2013). The presence of H3K9me3 is compelling, as it suggests that ASCL1 binds to sites marked for repression. ASCL1 ChIP-seq from ES cells does not readily identify a state with strong enrichment for H3K9me3, but does reveal the presence of modest enrichment for H3K9me3 in states 5 and 6, suggesting that ASCL1 binding in ES cells is less clearly associated with

this component of the trivalent signature identified in MEFs. One possible explanation for this is that the chromatin environment present in the ES cells is generally depleted of this, and other repressive marks. As this signature was initially identified in MEF cells, this distinction may primarily reflect the distinct chromatin landscapes between these cell types. While proliferative, MEFs represent a developmentally differentiated cell type, and may be relatively enriched for repressed chromatin domains as compared to ES cells.

To test whether the absence of this predictive signature might be due to differences in the chromatin environment, I directly compared the data sets on which this discovery was based. To address the possibility that ES cells may have lower levels of H3K9 methylation genome-wide, I compared the chromatin states in ES cells surrounding ASCL1 binding sites identified from ChIP-seq in the inducible ES cells, and MEF cells (Wapinski et al., 2013). As with other data sets utilized for HMM model learning, these were aligned and processed to the mm10 genome using our in-house pipeline. Peak calling was performed from ASCL1 ChIP-seq data generated from MEF cultures 48 hours after transfection with an rtTA-*Ascl1* viral vector. The 11-state Markov model was then used to compare the relative state enrichments between ASCL1 binding sites identified in ES cells and ASCL1 binding sites identified from MEFs (Figure 4-20: Comparison of HMM states for ASCL1 binding sites identified in ES and MEF cells). Comparison of these data demonstrates that these two data sets show largely similar state enrichments. However, observable distinctions are evident. ES binding sites are considerably more enriched for state 6 chromatin, which is centrally enriched at ASCL1 binding sites in ES cells, but only broadly enriched at ASCL1 binding sites identified in MEFs. Conversely, states 3,4, and 11 demonstrate visibly increased central

enrichment at ASCL1 bound sites identified in MEFs. Importantly, state 1 enrichment remained conspicuously weak in both data sets, although central depletion of this state is noted at ASCL1 binding sites in ES cells, likely due to the presence of H3K27ac, and other active marks at these sites. Thus, no state featuring the repressive mark H3K9me3 appears predictive of ASCL1 binding in ES cells, and ASCL1 binding sites do not appear to be significantly associated with this mark in the context of naïve ES cells. This demonstrates that the trivalent signature identified in MEF cells is not clearly predictive of binding in ES cells, and neither ASCL1 binding sites identified in the contexts of ES cells or MEFs are associated with the repressive mark H3K9me3 when observed in the chromatin environment of ES cells.

Another possible explanation is that due to differences in experimental strategy, ChIP-seq simply identifies more binding sites in the MEF reprogramming paradigm, and that this signature is primarily predictive of a subset of these additional sites specific to the MEF environment, rather than ASCL1 binding as a whole. As one goal of our approach is the identification of early binding sites, it is possible that the sites identified in 24-hour induced ES cells are subject to different selective pressures relevant to establishing early transcriptional events in the process of lineage specification. To test whether these sites might be associated with a distinct chromatin state in MEFs, I performed the same comparison using the data sets utilized in the original study. Using ChIP-seq data sets for H3K4me1, H3K4me3, H3K9me3, and H3K27ac, I performed undirected Markov model learning to identify chromatin states in MEFs, and characterized the presence of these states at ASCL1 binding sites identified in MEFs and ES cells, using the same approach as

previously described. Due to the lower number of histone markers characterized, an eight state model was sufficient to recapitulate the states identified.

Overall, the Markov model derived from MEFs demonstrated similar enrichment for active chromatin marks H3K4me1, H3K4me3, and H3K27ac at ASCL1 binding sites identified in either cellular context. Both sets of binding sites demonstrated central enrichments of states associated with active chromatin, including states 4 and 6, resembling active enhancer, and promoter regions, respectively. State 4 of the MEF-directed HMM, characterized by strong association with H3K4me1 and H3K27ac, demonstrates modest enrichment for H3K9me3, and most closely resembles the state identity identified in *Wapinski et al.* This state showed strong central enrichment at ASCL1 sites identified in MEFs, suggesting that this signature is enriched at ASCL1 binding sites in the context in which it was previously described, but not at the same set of sites in ES cells. As this state largely resembles state 6 of the MEF-directed HMM, which shows a comparable pattern of enrichment on intervals surrounding ASCL1 binding sites in MEFs, I interpret these associations as evidence that H3K27ac, which is strongly associated in both states, may be predictive of ASCL1 binding in this environment. As ASCL1 binding sites identified in ES cells show a similar enrichment for a largely equivalent state (6) in the Markov model derived from ES cells, this suggests that the presence of H3K27ac in combination with H3K4 monomethylation or trimethylation may partially inform ASCL1 binding in either cell environment. Intriguingly, ASCL1 binding sites identified in ES cells demonstrate further distinction in the MEF derived in their association with MEF chromatin states 7 and 8. These states are defined primarily by association with H3K4me3 and H3K4me1, but are not

strongly associated with H3K27ac, suggesting that these may represent inactive enhancer and promoter regions. As H3K27ac is believed to distinguish active from poised enhancers, this distinction suggests that a subset of these sites are active in ES cells, but inactive in MEFs. As binding sites identified in this cellular context are less enriched for these states, this provides further evidence that H3K27ac may be partially predictive of ASCL1 binding. Together, these results show that of the histone marks compared H3K27ac is the most predictive component of ASCL1 binding in both ES cells and in MEFs, providing further support that ASCL1 binding may be informed by the presence of this hallmark of active chromatin.

Summary and Conclusions from chromatin landscape studies

The progressively changing landscape established through lineage specification represents an appealing framework for regulation of transcription factor binding and function within lineages. As bHLH factors bind different sites, and initiate lineage-directive gene expression against the backdrop of this landscape in vivo, it is reasonable that this binding may be partially reflective of the changing environment in which they function. The effects of this landscape on transcription factor binding remain largely uncharacterized. It has previously been suggested that differential preferences in transcription factor binding may be one mechanism by which these factors identify appropriate binding sites against the many potential sites present in the genome. Here, I utilized an embryonic stem cell model with inducible expression of TFs to compare the binding of ASCL1, ASCL2, and MYOD against

epigenetic features present in undifferentiated ES cells. In doing so, we interrogate the influence of specific features of the chromatin landscape on bHLH factor binding and function, and the effect of expression of these factors on this landscape.

It has recently been postulated that bHLH factors such as ASCL1, ASCL2 and MYOD may identify their cognate binding sites through differential pioneering capability (Soufi et al., 2015). Using ATAC-seq, I characterized the chromatin environment at the empirically defined binding sites of these factors in ES cells. In contrast to the model put forward in *Soufi et al., 2015*, my data demonstrate that each of the three class II bHLH factors compared here is able to bind to both open and closed sites. As such, binding site specificity of these factors is not due to obvious differences in chromatin accessibility, and the presence of closed chromatin is not sufficient to restrict binding of these factors.

This finding is crucial to understanding the previously demonstrated role of ASCL1 and MYOD in cell reprogramming paradigms. The ability of these factors to bind both open and closed chromatin demonstrates that lineage-specific chromatin accessibility is not the primary determinant in defining the binding of these factors. However, it does not explain how these factors initiate the specific gene regulatory networks necessary for reprogramming when expressed outside their normal developmental context. Additionally, it does not provide a mechanistic explanation for the context-specific binding of these factors, which we have shown to occur without a clear distinction in binding motif.

As the binding of these factors is not restricted to a subset of open sites, the capacity of these factors to identify lineage-appropriate targets lies elsewhere. While differential preference for primary or co-factor motifs are appealing explanations for where this binding specificity

resides, our results demonstrate that the motifs bound are not dramatically different between binding sites with high, or low chromatin accessibility. The modest differences in motif identified are suggestive of a shift in sequence of the E-box, wherein a GC dinucleotide core is slightly more enriched in nucleosome-depleted chromatin, versus GG/CC in nucleosome-occupied chromatin. However, both E-box sequences are well represented in both open and closed chromatin, and this shift is consistent for all three bHLH factors tested; thus, chromatin-dependent E-box preference is not the primary mechanism for binding site selection, or distinct binding in this context. The distinct difference in distribution of these motifs is suggestive of potentially distinct mechanisms of transcription factor binding in open versus closed chromatin, but the significance of this distinction presently remains unclear. Similarly, while secondary motifs were tested to identify potential co-factor influence in binding to open and closed chromatin, they did not prove a reasonable candidate for distinct binding. Secondary motifs identified in open chromatin were largely similar to those identified for these factors overall, but were generally associated with peaks identified in open chromatin. As with secondary motifs identified for the total sets of bHLH binding sites, the low enrichment for these motifs at bHLH binding sites identified implies that they are not present in sufficient numbers to mediate the distinct binding observed for these factors. Together, these data suggest that factor-cofactor interactions are not central mechanisms in binding site selection in this context, and that the pioneering ability of ASCL1, ASCL2, and MYOD is intrinsic to these bHLH factors, rather than conferred upon them by specific DNA-binding co-factors. However, this is not necessarily representative of bHLH binding in their respective developmental contexts. The binding sites identified in the context of ES cells

represent a subset of the total set of binding sites for these factors, and it may be the case that this subset is less dependent on factor-cofactor interactions. As downstream transcriptional targets of these bHLH factors are likely not present at this early stage, sites at which bHLH binding is mediated by factor-cofactor interactions would necessarily be absent from this subset. In their normal developmental context, such co-factors may provide additional preference for genomic features, including chromatin accessibility.

Additionally, I demonstrate that ectopic expression of ASCL1, ASCL2, and MYOD lead to increases in chromatin accessibility at their respective binding sites within 24 hours. This finding, combined with the ability to direct gene expression, and access closed chromatin, meet the formal criteria set for pioneer factors (as defined in Zaret & Carroll, 2011). These changes are supportive of a potential role for bHLH-dependent chromatin remodeling, and suggest that this may be one mechanism mediating transcriptional activation by these factors. This finding is especially compelling, as these bHLH factors are considered master regulators in cell fate and lineage specification. In addition to the known role of these factors in initiating gene regulatory networks in development, bHLH-dependent chromatin remodeling may allow for additional lineage-specific regulation. We observe a general increase in chromatin accessibility at bHLH binding sites in this context, which, in addition to recruitment of transcriptional machinery, may reflect recruitment of chromatin remodeling complexes. As the process of lineage-specification is thought to involve durable changes in the chromatin landscape, these factors may be one mechanism by which these factors establish long term changes in the regulatory networks of their respective lineages.

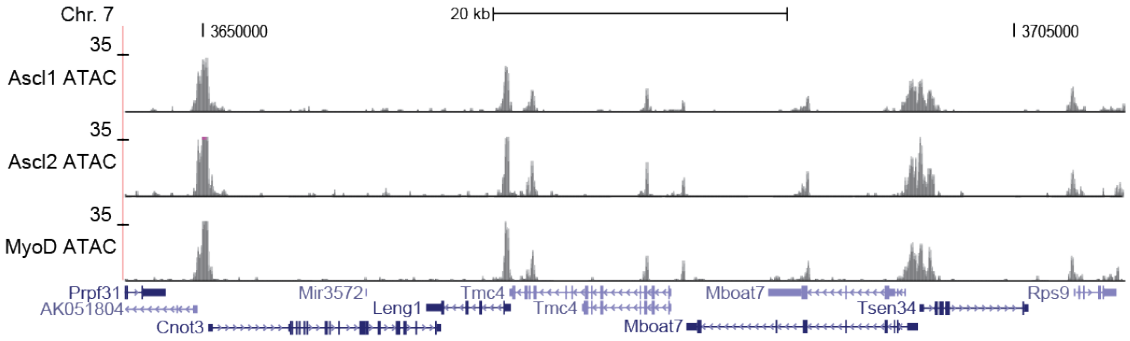
In characterizing the motifs identified at sites showing bHLH-dependent increases in chromatin accessibility, we find that the most significant motif is an E-box resembling that identified for these factors from ChIP-seq, further supporting a role for these factors in defining chromatin accessibility. In contrast, we find that sites demonstrating a reduction in chromatin accessibility are not appreciably enriched for an E-box motif, suggesting that these factors are likely not responsible for these decreases *in cis,* concordant with their previous characterization as transcriptional activators. This suggests that these factors may have both a direct role in increasing chromatin accessibility, and an indirect role in decreasing chromatin accessibility. While the core components of the *SWI/SNF* ATP-dependent chromatin remodeling mechanism are not themselves direct targets of these factors, these genes are expressed in ES cells, and are crucial to maintenance of pluripotency (Kidder et al., 2009). It may be that bHLH factors activate expression of other genes which mediate decreases in open chromatin, thus allowing for both bHLH dependent increases *in cis*, and bHLH-dependent decreases *in trans*.

Finally, we further characterize the chromatin landscape at bHLH binding sites using a hidden Markov modeling approach. This comparison demonstrated that ASCL1, ASCL2, and MYOD binding sites are enriched for markers of active enhancers, including H3K4me1, H3K4me3, and H3K27ac. However, these factors shared this enrichment, and did not reveal factor-specific preferences for the histone marks tested. In addition, ASCL1 binding was compared to the data from a previous study, which identified the presence of a trivalent signature which was predictive of ASCL1 binding. This comparison revealed that ASCL1 binding sites identified in MEF cells were similarly enriched for markers of active enhancers,

and that that these patterns of enrichment were context dependent. However, this does not preclude the possibility that other histone modifications, alone or in combination with the marks tested, might play a role in the binding specificity observed for these factors.

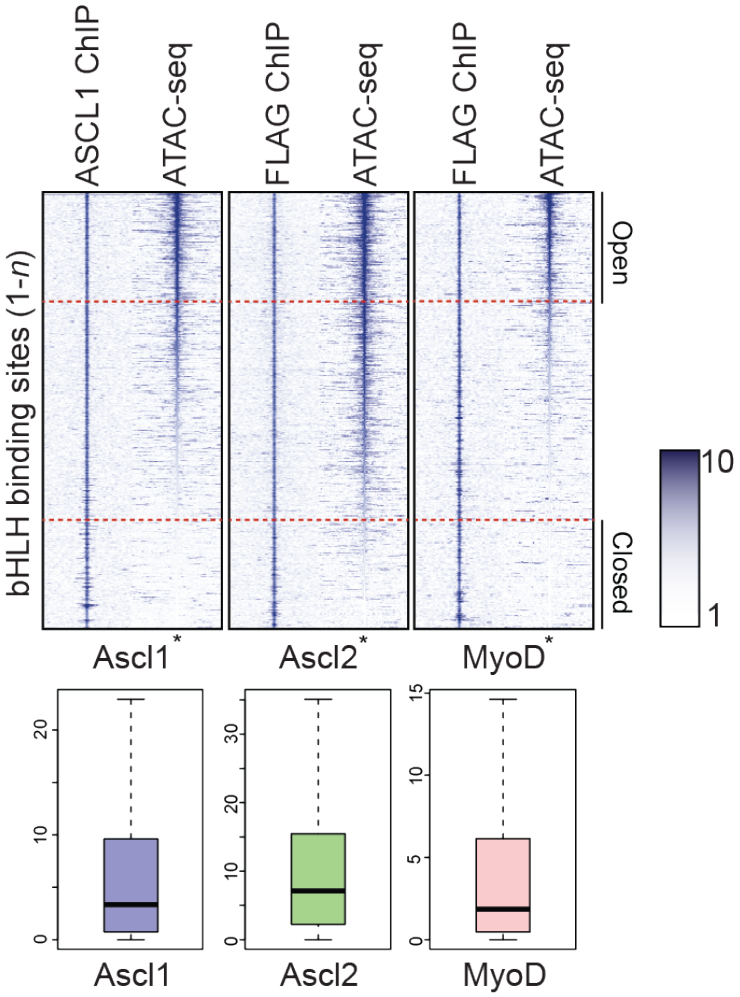
Thirty years ago, *MyoD* was discovered as the first master regulatory factor, and its remarkable ability to revise the identity of an established cell fate fundamentally changed the field of developmental biology. For many years, these class II bHLH factors have been characterized primarily based on their class-defining interaction with E-proteins, and their shared preference for an E-box binding motif. In their respective developmental contexts, and when ectopically expressed, they are able to engage relevant transcriptional networks from an invariant genomic template and define their respective cell populations. Together, the results of these experiments demonstrate that these master regulators of cell fate possess the intrinsic capacity to access specific binding sites, even when ectopically expressed in ES cells. We further demonstrate that this specificity is not dependent on differential preference or ability to bind closed chromatin, and show that each is able to induce changes in the chromatin landscape. The mechanism by which these factors identify specific binding sites from the many possible sites genome-wise remains a central question in defining the activity of these crucial regulators of cell fate.

Figure 4-1: ATAC-seq enrichment in uninduced ES cells near *Mboat7* locus



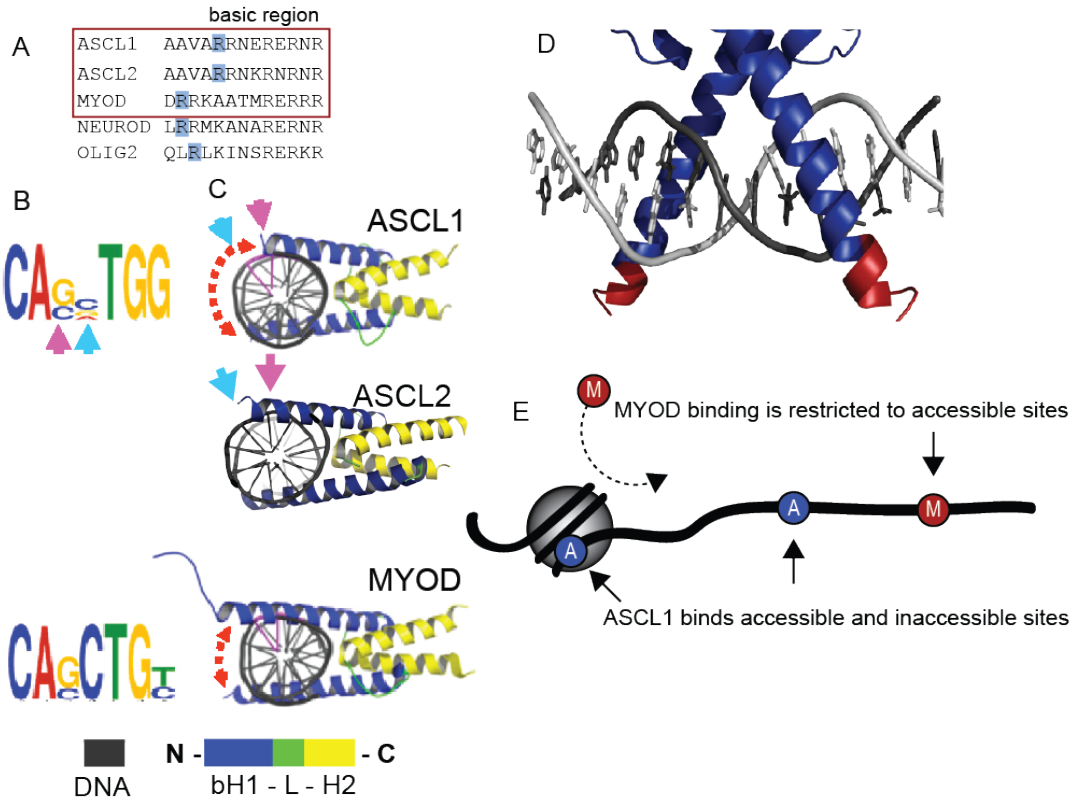
ATAC-seq reveals enrichment for open chromatin in ES cells.
Shown is UCSC mm10 genome browser plot of region ~100kb surrounding Mboat7 locus, previously shown to contain several regions of open chromatin (Simon et al., 2012). Tracks represent ATAC-seq signal in uninduced cells from *Ascl1, Ascl2*, and *MyoD* ES cell lines, normalized to 10M reads. Refseq transcripts shown reflect positions of nearby features. Signal scale and mm10 genomic coordinates shown.

Figure 4-2: Heatmap comparison of bHLH ChIP-seq and uninduced ATAC



ASCL1, ASCL2, and MYOD bind to both open and closed chromatin in ES cells
Heatmap comparison of ChIP-seq and ATAC-seq signal at bHLH binding sites identified by ChIP-seq. Columns represent (A)ChIP-seq from 24h induced ES cells and ATAC-seq from uninduced ES cells. Each histogram represents the set of sites identified in ChIP-seq from respective cell lines, using 6kb interval centered on ChIP-seq peak center. Dashed lines reflect quartile cutoffs for “open” vs. “closed” sites based on mean ATAC-seq signal from 100bp interval surrounding bHLH ChIP peak apex. Asterisk denotes data set used for ranking. (B) Box-whisker plot of ATAC signal at bHLH binding sites identified by ChIP. Plot depicts distribution of ATAC-seq signal at ChIP-seq peaks, using mean ATAC-seq signal from 100bp interval surrounding bHLH binding sites. Scale is indicated for each plot.

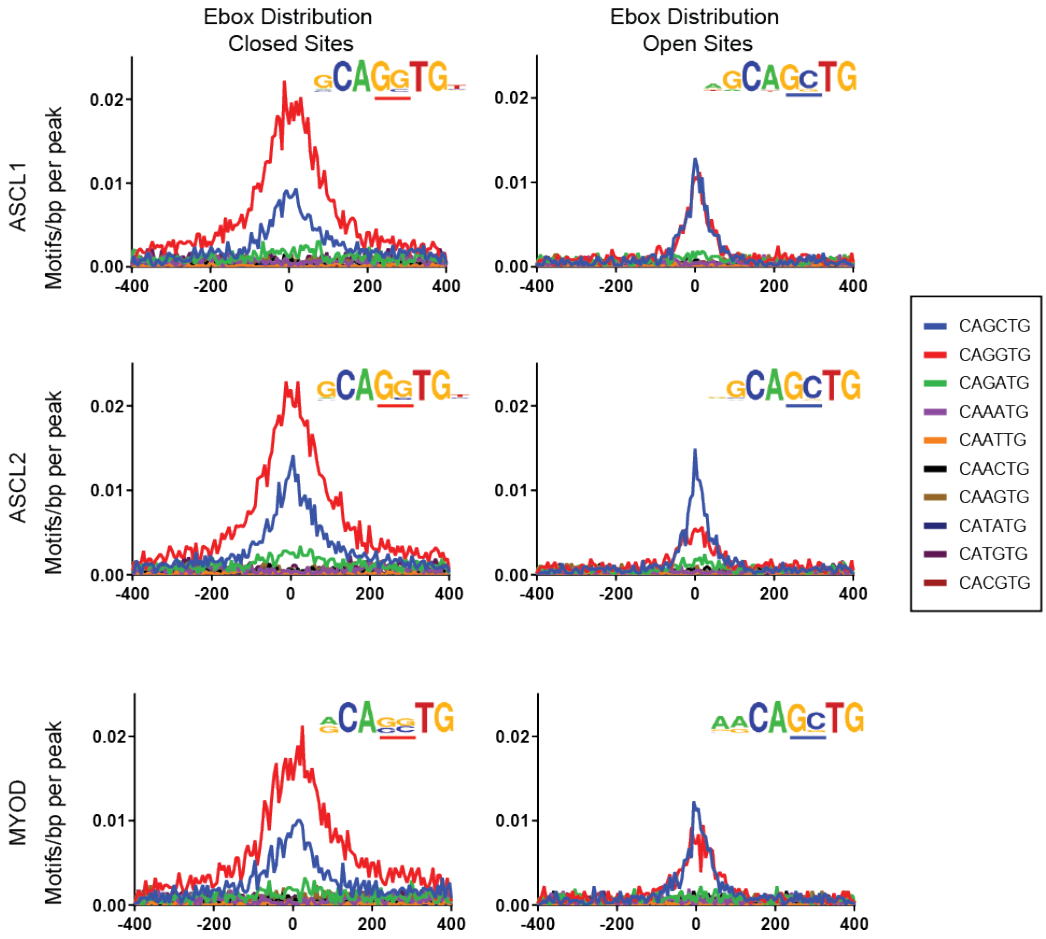
Figure 4-3: Proposed mechanism for distinction in pioneering capacity of bHLH factors



Suggested mechanism describing predicted differential pioneering capacity of class II bHLH factors (Adapted from *Soufi et al., 2015*)
A) Comparison of protein structure of bH1 domain of bHLH factors. Blue box indicates last basic residue of bH1 alpha helices for each factor compared.
B) Comparison of de novo motifs identified from ChIP-seq data for ASCL1 and MYOD (as reported in *Soufi et al., 2015*). Arrows depict relative positions of central dinucleotides for comparison to structural models.
C) Inferred structural models depicting bHLH:DNA complexes. Colors reflect map of helical domains shown below. Structural prediction demonstrates difference in bH1 domain. ASCL1, MYOD reproduced from Soufi et al., 2015, ASCL2 modeled using same approach for comparison. Colored arrows highlight position of central dinucleotide relative to bHLH.
D) Lateral view of MYOD in complex with Ebox. Blue residues indicate predicted length of helix for ASCL1/ASCL2, red residues indicate positions predicted for additional helix in MYOD. Structure modeled on solved structure for MyoD homodimer bound to DNA (Ma et al., 1994, PDB structure accession *1mdy*).
E) Inferred model of differential pioneering ability, Colored circles represent ASCL1 (blue) and MYOD (red)

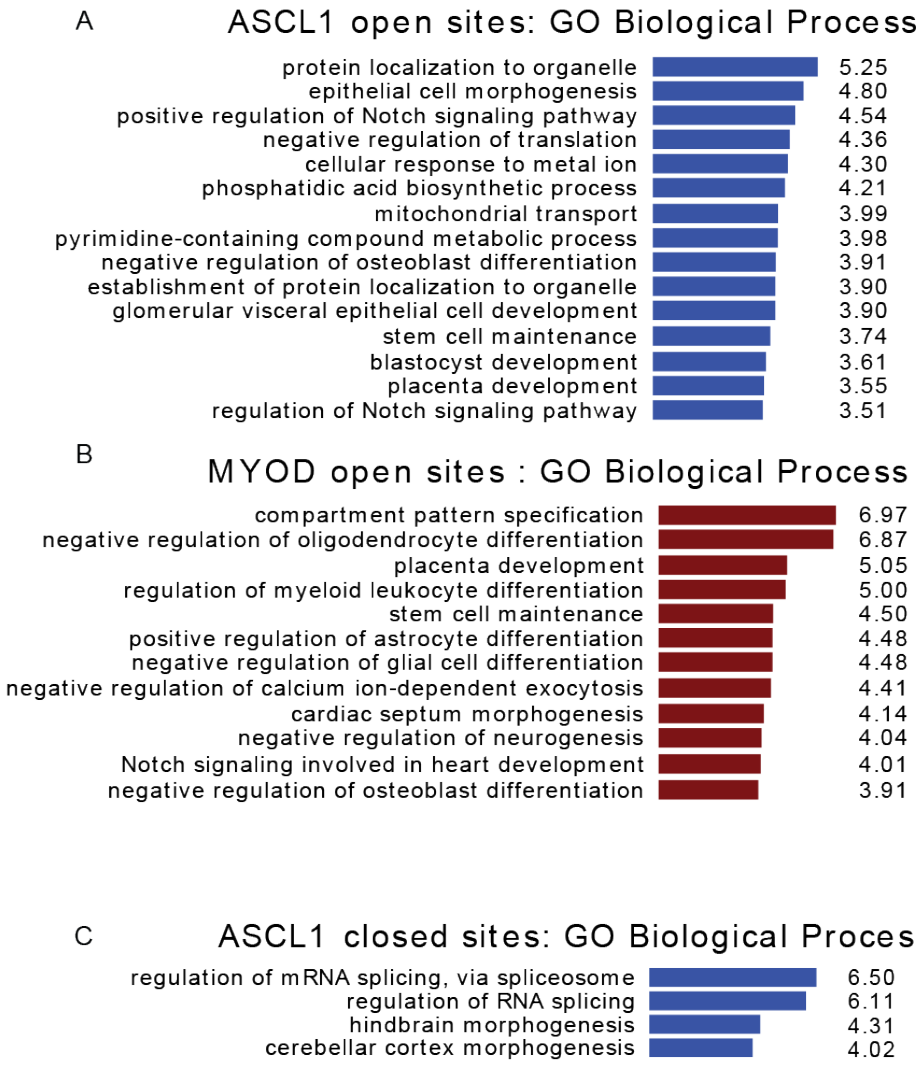
Adapted from Soufi et al, Cell (2015)

Figure 4-4: ATAC-ranked bHLH Ebox Comparison



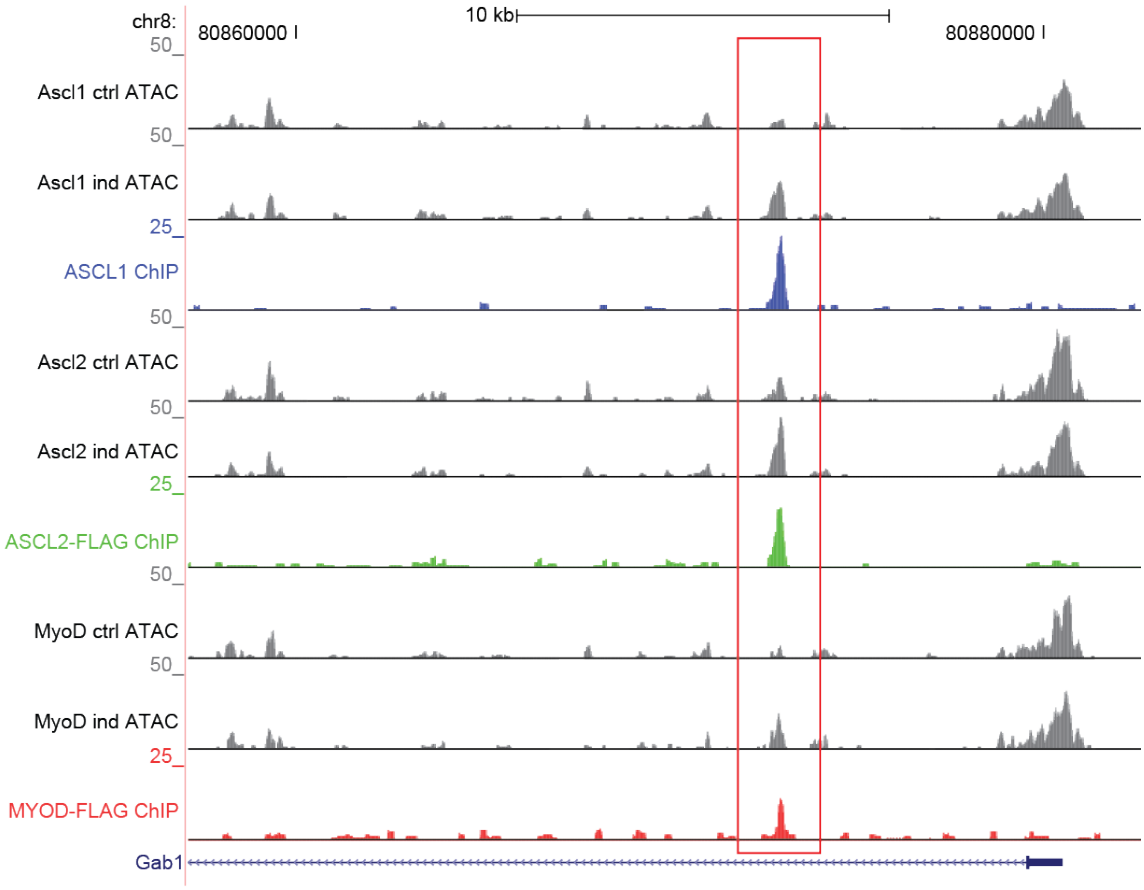
Binding sites present in open versus closed chromatin show distinct patterns of Ebox distribution.
Inset for each plot shows the most significantly enriched motif identified from each subset of peaks shown, as identified by HOMER, using parameters *-S 10 -len 8 - bits*. Plots show spatial distribution for each Ebox permutation within a 800bp window, centered on the peak center identified in ChIP-seq. Shown are plots for each bHLH factor tested (rows) in the highest, or lowest quartile of peaks (columns), as ranked by mean ATAC-seq signal from the 50bp interval surrounding bHLH peak center.

Figure 4-5: Overview of GO categories enriched in open and closed binding sites



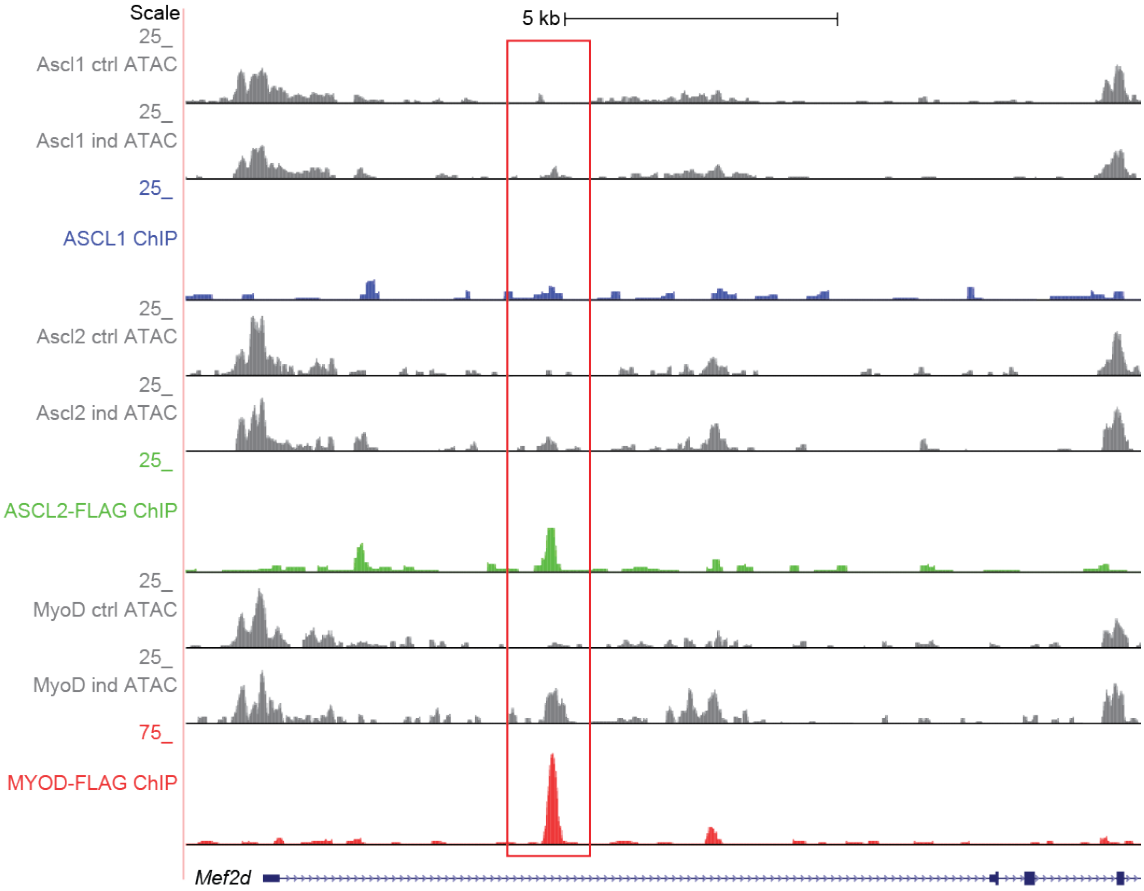
Overview of GO analysis from bHLH binding sites demonstrating highest enrichment for open (ASCL1, MYOD), and closed (ASCL1) chromatin prior to induction.
Results from *GREAT* v3.0 GO BP analysis of top 1000 sites based on ATAC-seq signal. Length of bars and values shown represent -log10 binomial P-value of category shown. Analysis performed on ASCL1, ASCL2, MYOD binding sites at open and closed chromatin. Sets not demonstrating significant (FDR≤0.05 by hypergeo-metric comparison of gene regions and binomial genomic comparison) enrichment for any GO category not shown (ASCL2 open and closed sites, MYOD closed sites identify no significantly enriched categories).

Figure 4-6A: Regions identified by differential peak calling from ATAC-seq



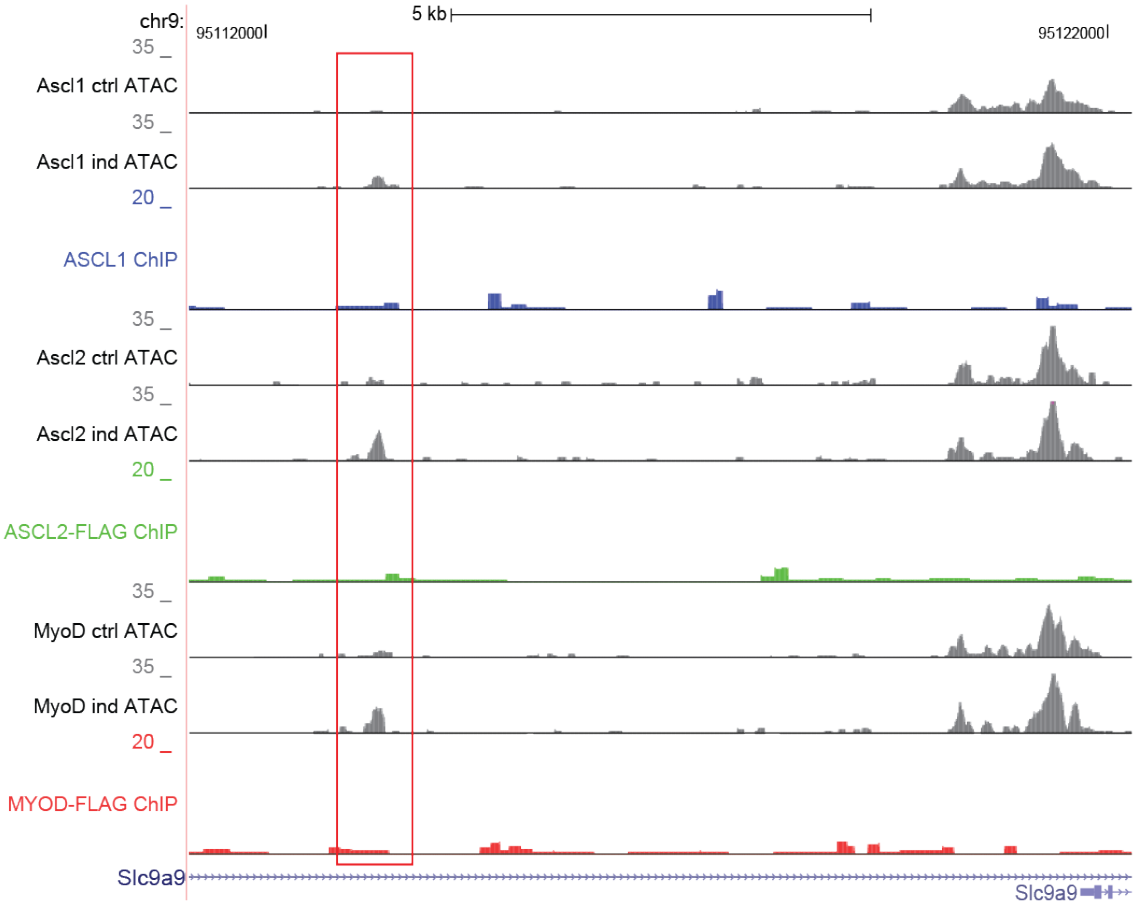
ATAC-seq identifies bHLH-dependent changes in open chromatin in ES cells.
Shown is UCSC mm10 genome browser plot of region within first intron of *Gab1* locus, demonstrating bHLH-dependent change in ATAC-seq signal at bHLH binding site identified by ChIP-seq. Tracks represent ChIP-seq (24h post-induction) and ATAC-seq signal (uninduced) normalized to 10M reads from *Ascl1, Ascl2*, and *MyoD* ES cell lines. Red box indicates area of peak and corresponding change in ATAC-seq.Refseq transcript shown indicates position within intron of *Gab1*. Scale and mm10 genomic coordinates shown.

Figure 4-6B: *Mef2d* locus is differentially bound and selectively opened by MYOD



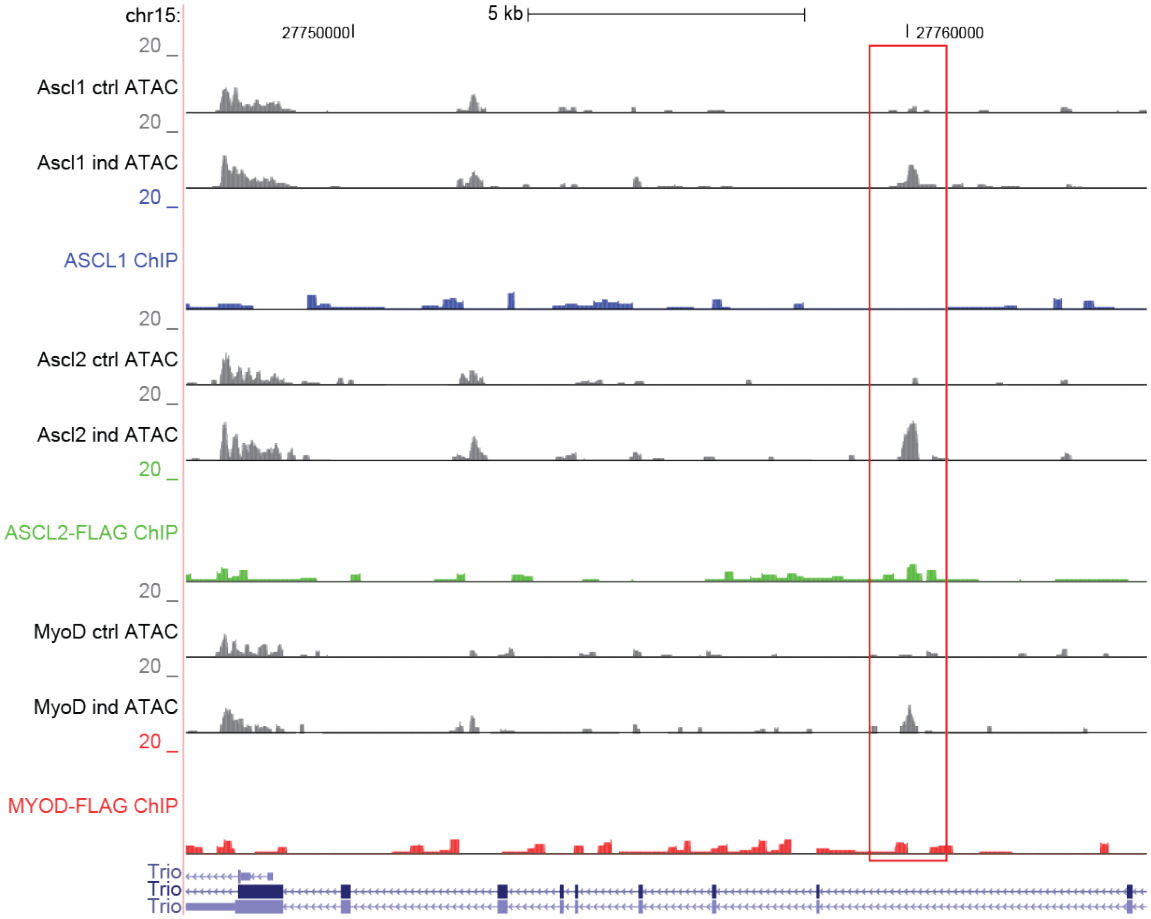
MYOD binds closed chromatin and leads to increased accessiblity within 24h of MYOD expression. Shown is UCSC mm10 genome browser plot of region near *Mef2d* locus. Tracks represent ChIP-seq (24h post-induction) and ATAC-seq signal (uninduced) normalized to 10M reads from *Ascl1, Ascl2*, and *MyoD* ES cell lines. ASCL1 and ASCL2 ChIP-seq tracks plotted at increased scale to show difference in signal between factors tested. Red box indicates area of peak and corresponding change in MYOD ATAC.

Figure 4-6C: Regions identified by differential peak calling from ATAC-seq



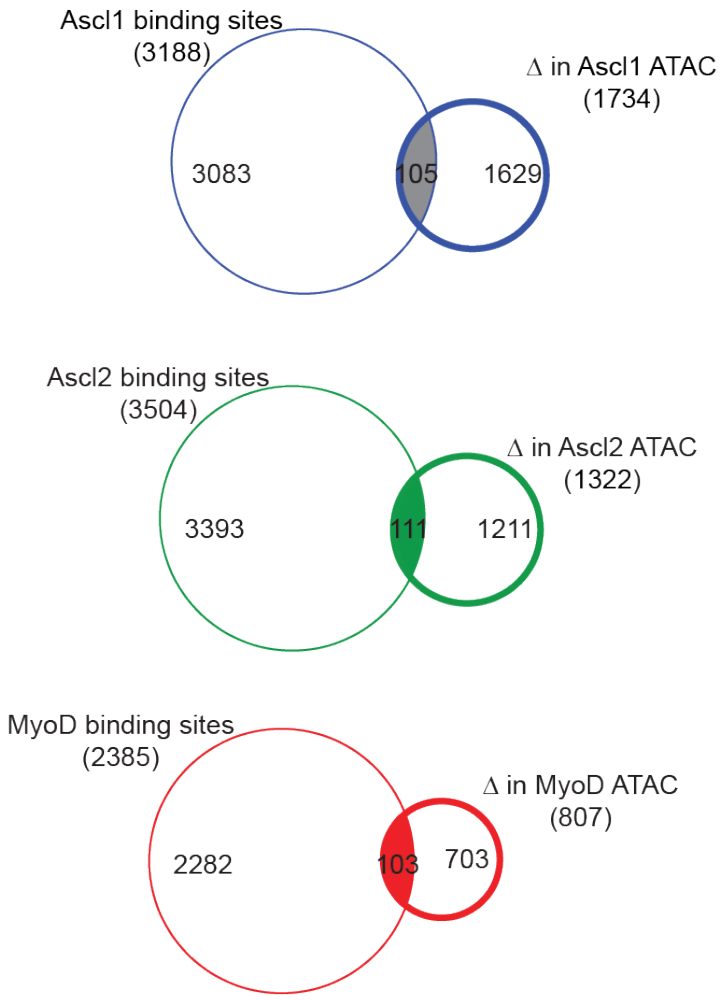
Changes in open chromatin are not restricted to bHLH binding sites in ES cells.
Shown is UCSC mm10 genome browser plot of region within first intron of *Slc9a9* locus, demonstrating bHLH-dependent change in ATAC-seq signal in the absence of a bHLH binding site. Tracks represent ChIP-seq, and ATAC-seq signal normalized to 10M reads in uninduced cells from *Ascl1*, *Ascl2*, and *MyoD* ES cell lines. Refseq transcript shown indicates position within intron of *Slc9a9*. Scale and mm10 genomic coordinates shown.

Figure 4-6D: Overview of GO categories enriched in open and closed binding sites



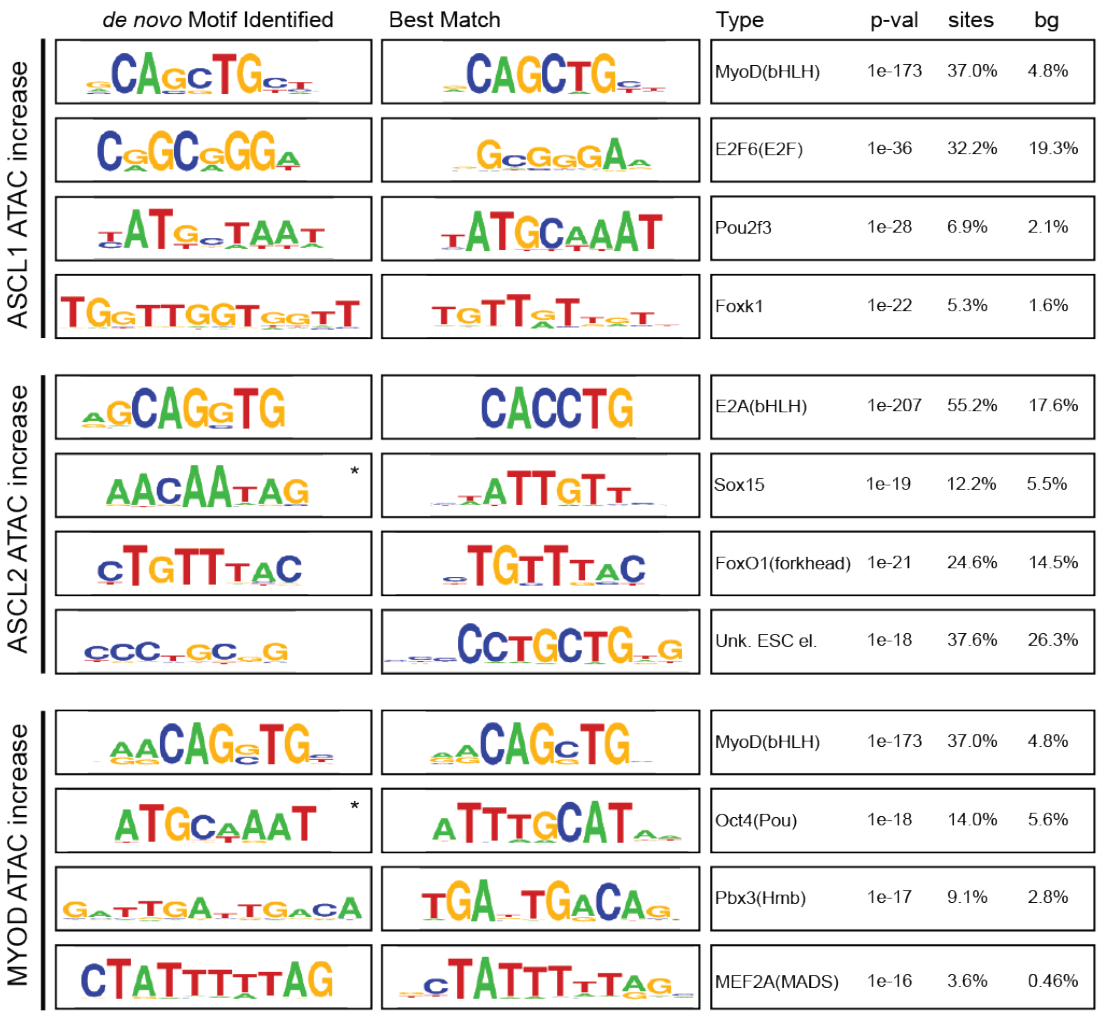
Changes in open chromatin are not restricted to bHLH binding sites in ES cells.
Shown is UCSC mm10 genome browser plot of region within first intron of *Trio* locus, demonstrating bHLH-dependent change in ATAC-seq signal in the absence of a bHLH binding site. Tracks represent ChIP-seq, and ATAC-seq signal normalized to 10M reads in uninduced cells from *Asc*Tracks represent ChIP-seq (24h post-induction) and ATAC-seq signal (uninduced) normalized to 10M reads from *Ascl1, Ascl2*, and *MyoD* ES cell lines. Refseq transcript shown indicates position within intron of *Trio*. Scale and mm10 genomic coordinates shown.

Figure 4-7: Overlap comparison of ChIP-seq peaks vs local increases in ATAC-seq



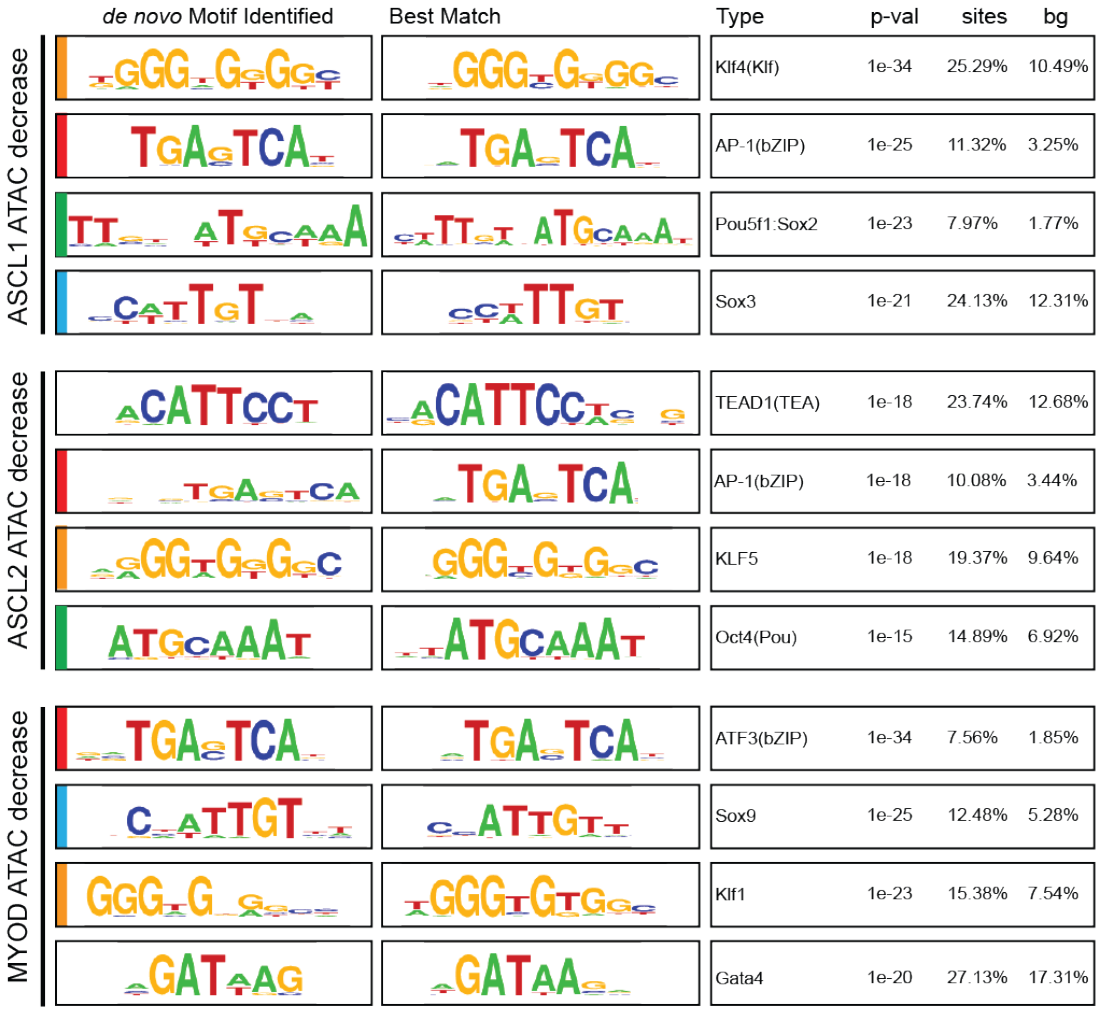
Comparison of bHLH-dependent increases in open chromatin.
Area proportional diagram comparison of bHLH binding sites and regions showing focal increase in ATAC-seq. ATAC-seq peaks identified from induced ES cells at 24h post-induction. Numbers reflect total number of sites in each component of set. Shaded areas reflect sites overlapped.

Figure 4-8: Comparison of *de novo* motifs identified at regions increased in ATAC-seq



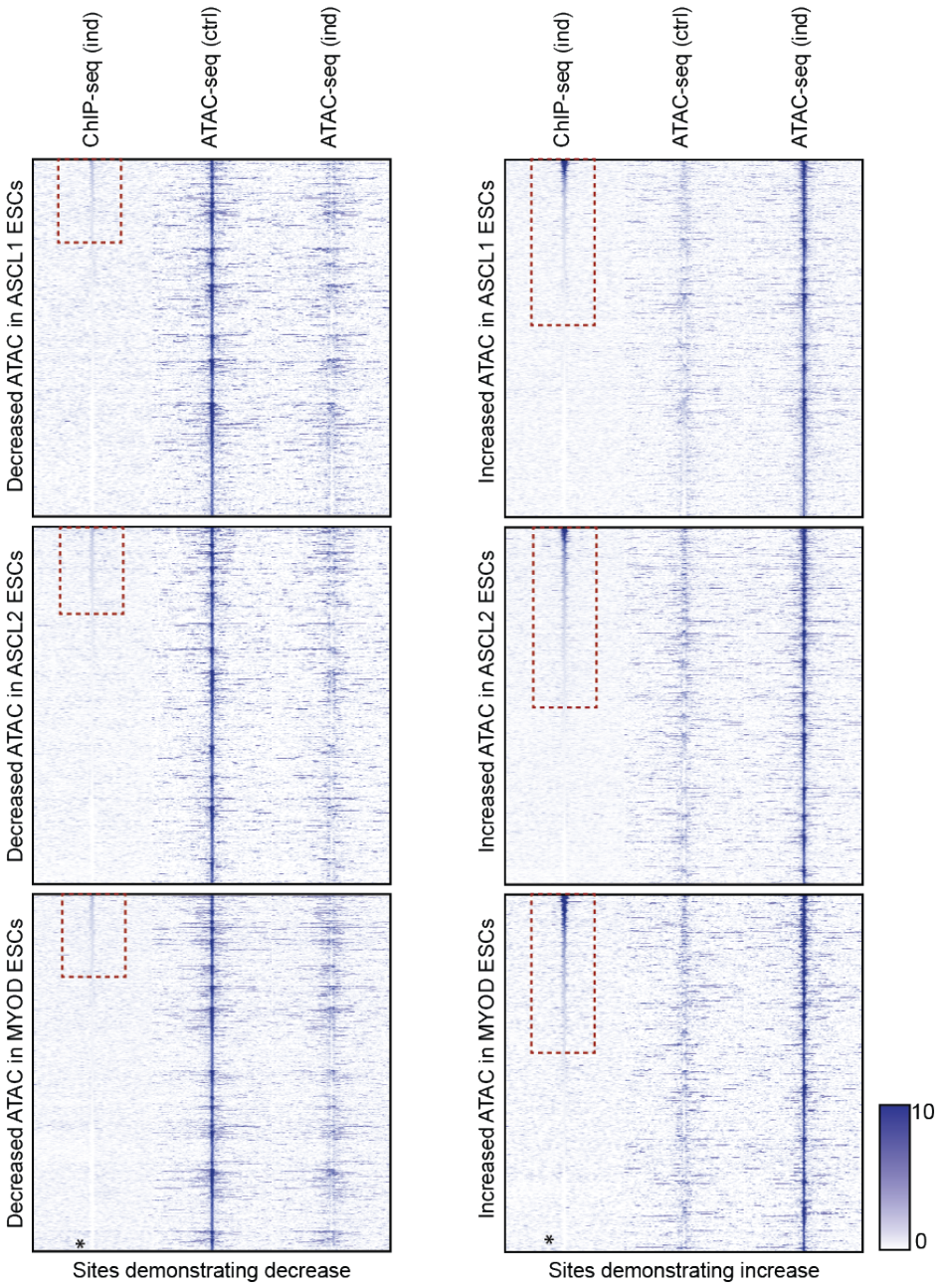
Regions demonstrating local increase in open chromatin are enriched for Ebox motifs
Comparison of *de novo* motifs identified at regions demonstrating local increase in open chromatin as assayed by ATAC-seq. Differential ATAC changes identified by HOMER *findPeaksGenome* using induced/uninduced ATAC-seq data sets, with *-p 1e06* parameter. *de novo* motif discovery performed using 200bp interval centered on ATAC-seq change used for analysis, HOMER parameters used: *-size 200 -bits*. Asterisk denotes motifs identified as low stringency, and replaced by higher strength PWM. Numbers reflect the binomial significance of the de novo motif identified, the percentage of sites featuring the specified motif, and the percentage of normalized random background featuring the specified motif.

Figure 4-9: Comparison of *de novo* motifs identified at regions decreased in ATAC-seq



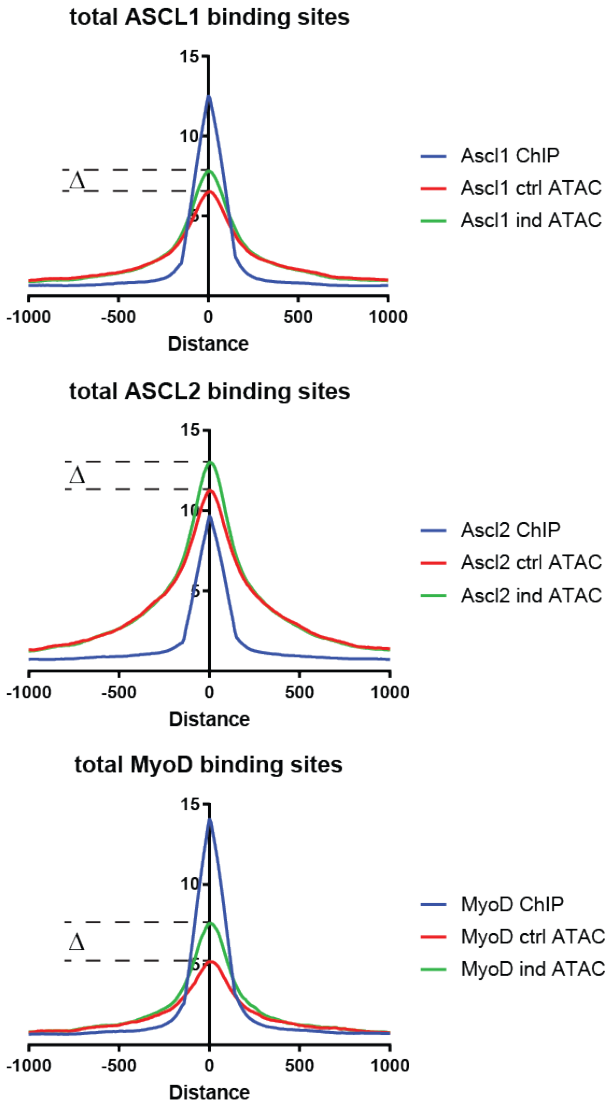
Regions demonstrating local increase in open chromatin are enriched for Ebox motifs
Comparison of *de novo* motifs identified at regions demonstrating local decrease in open chromatin as assayed by ATAC-seq. Colored bars represent motifs apparent in multiple data sets; similar motifs between factors are indicated by same color. Differential ATAC changes identified by HOMER *findPeaksGenome* using uninduced/induced ATAC-seq data sets, with *-p 1e06* parameter. *de novo* motif discovery performed using 200bp interval centered on ATAC-seq change used for analysis, HOMER parameters used: *-size 200 -bits*. Numbers reflect the binomial significance of the de novo motif identified, the percentage of sites featuring the specified motif, and the percentage of normalized random background featuring the specified motif.

Figure 4-10: Heatmap comparison of ChIP-seq and ATAC-seq at sites demonstrating changes in open chromatin



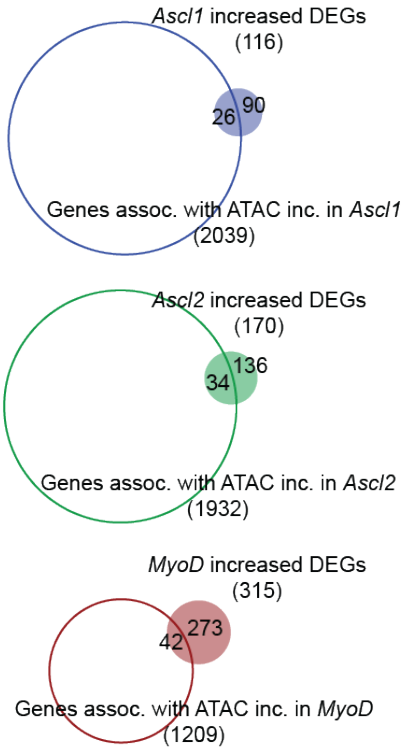
Heatmaps comparing ATAC-seq and ChIP-seq enrichment at sites demonstrating significant changes in open chromatin. Sites demonstrating decreases in open chromatin (left column) and increases (right column), were used to define intervals demonstrating local changes. Heatmap plots represent 6kb regions centered on the sites identified as changed in ATAC-seq in ASCL1, ASCL2, or MYOD-expressing ES cells at 24h. Each plot shows bHLH ChIP-seq signal, uninduced ATAC-seq signal, and induced ATAC-seq signal. Red dashed boxes highlight distinction in bHLH ChIP-seq signal. All plots sorted based on peak apex of bHLH binding site (denoted by asterisk).

Figure 4-11 : Histograms comparing ChIP and ATAC-seq signal at bHLH binding sites



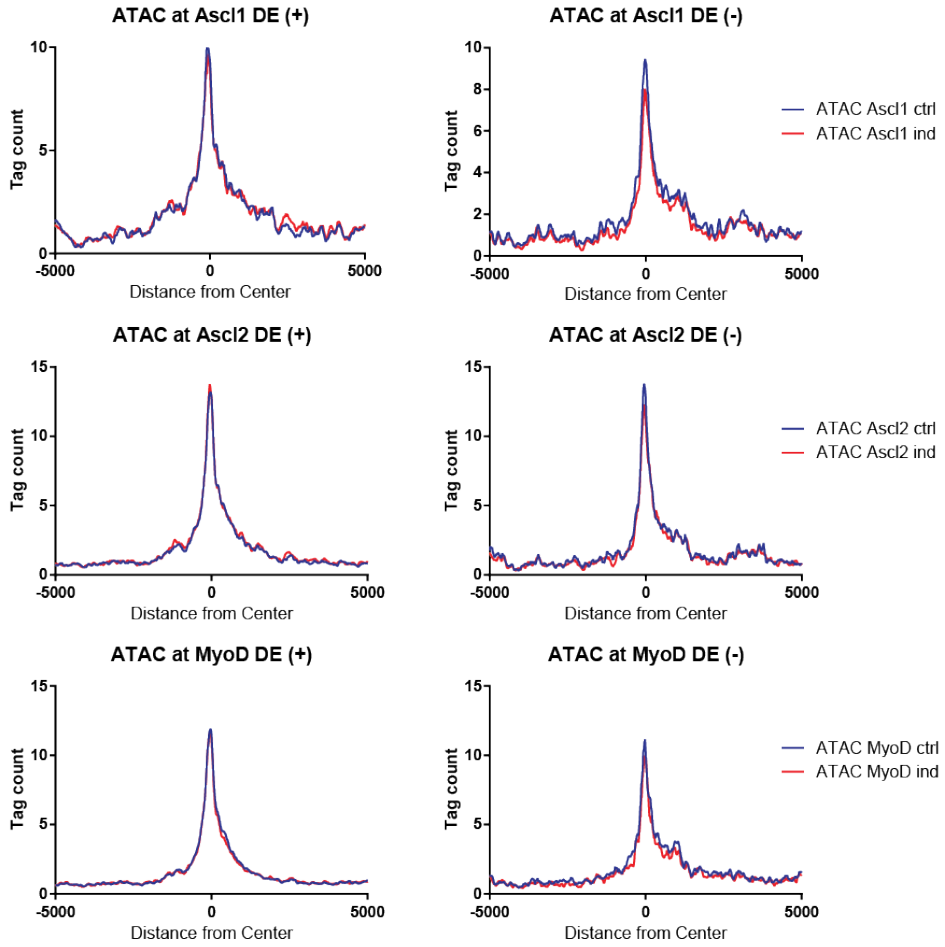
Histograms comparing bHLH ChIP-seq and ATAC-seq data at bHLH binding sites identified in ES cells. Plots depict the mean tag count at each position on 2kb intervals centered on peak apex of bHLH binding sites identified by ChIP-seq for the data sets shown. Each plot represents the aggregate result from the total set of bHLH binding sites identified for each factor Dashed lines indicate change in mean ATAC-seq signal.

Figure 4-12: Comparison of differentially expressed genes associated with increases in open chromatin



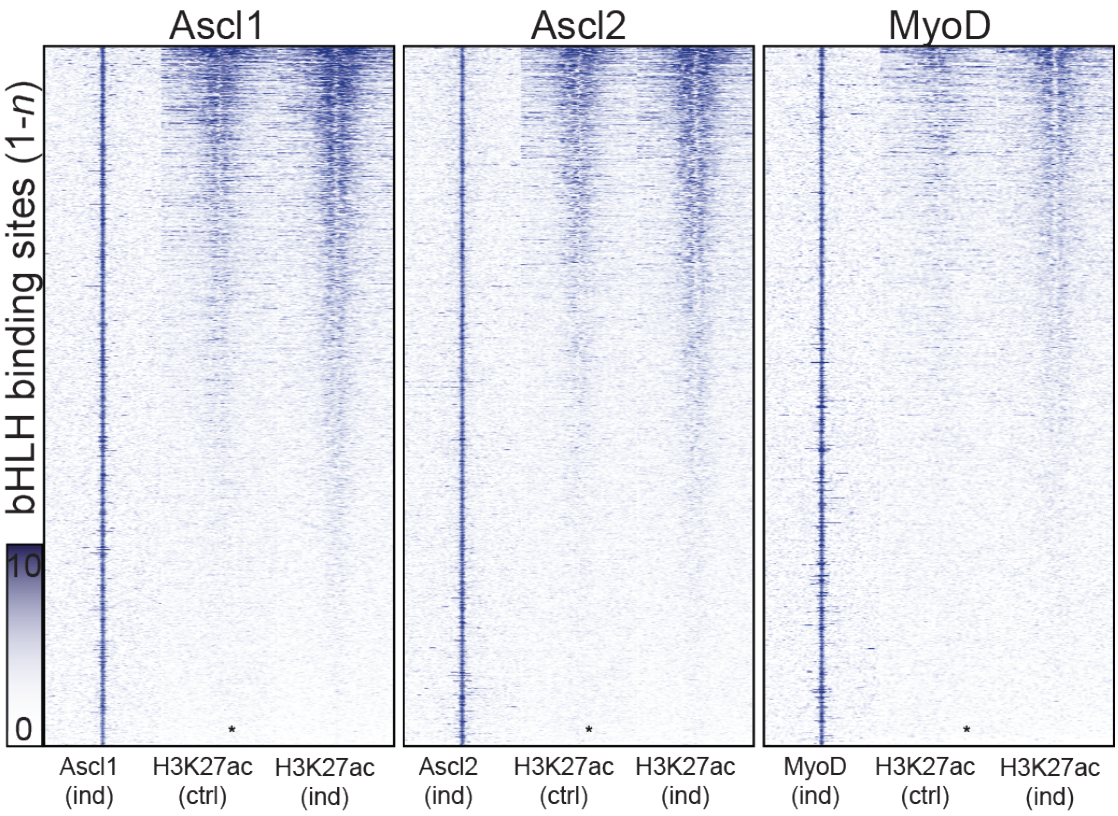
Comparison of differentially expressed genes associated with increases in open chromatin. Proportional venn diagrams comparing overlap between changes in open chromatin identified by ATAC-seq and RNA-seq. ATAC-seq region-to-gene calling performed using GREAT v3.0 using default association parameters. ATAC-seq changes identified by HOMER v4.7 using poisson distribution parameter -p 1e06. RNA-seq from genes with average RPKM≥1 (induced), FC≥2, and FDR≤.05 in a given cell line.

Figure 4-13: Distribution of open chromatin at DE genes



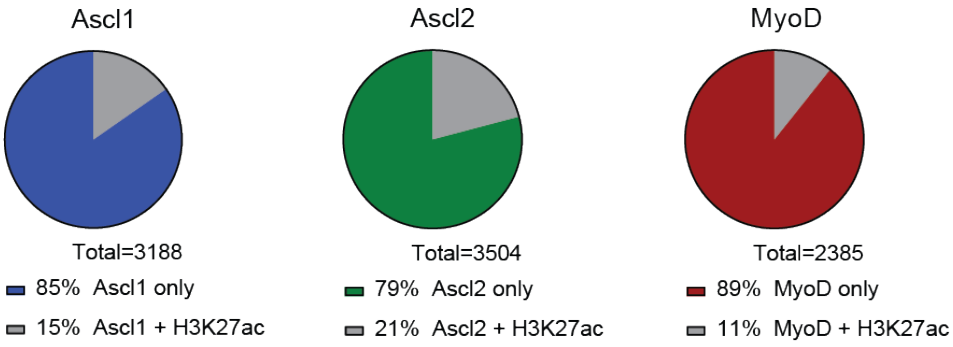
Differentially expressed genes do not demonstrate bHLH-dependent changes in open chromatin. Comparison of aggregate ATAC-seq signal at TSS of positively or negatively differentially expressed genes identified in edgeR analysis of RNA-seq data from 24h, with calculated FDR of ≤0.05. ATAC tag counts from 10kb region oriented to, and centered on TSS, shown for induced and uninduced control samples across cell lines. Each interval oriented as 5'(left) to 3'(right). Both uninduced control (blue) and induced (red) are plotted for each cell line to demonstrate lack of change.

Figure 4-14: Heatmap comparison of bHLH and H3K27ac ChIP-seq at bHLH binding sites



Heatmaps comparing bHLH and H3K27ac ChIP-seq enrichment at bHLH binding sites identified in ChIP-seq. ChIP-seq enrichment plotted on 6kb interval surrounding ChIP-seq peak center. Each plot shows bHLH/FLAG ChIP-seq signal from 24h induced ES cells, and H3K27ac ChIP-seq signal from uninduced, and 24h induced cells of respective cell line. All plots sorted in descending order based on mean H3K27ac signal across central 100bp (denoted by asterisk).

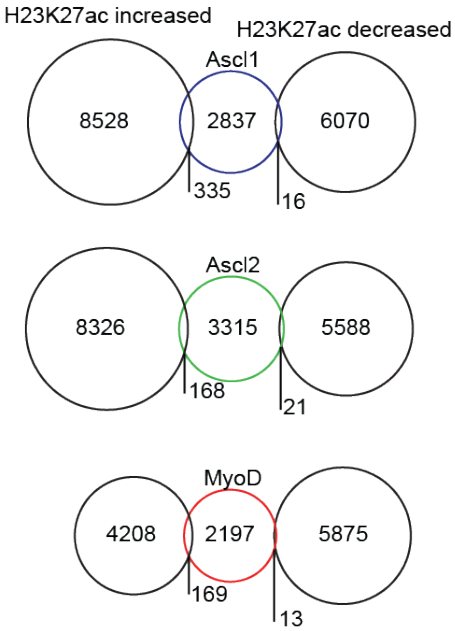
Figure 4- 15: Proportion of bHLH binding sites associated with H3K27ac enrichment



Percentage of bHLH binding sites identified in ChIP also significantly enriched for H3K27ac. Plots depict numbers of bHLH binding sites identified in ES cells associated with H3K27ac enrichment in uninduced ES cells. Numbers represent total number of bHLH binding sites identified in each bHLH ChIP-seq data set. Percentages reflect fraction of total bHLH sites identified for each factor which were associated with significant H3K27ac enrichment. Peak overlap defined as apex distances of ≤500 bp.

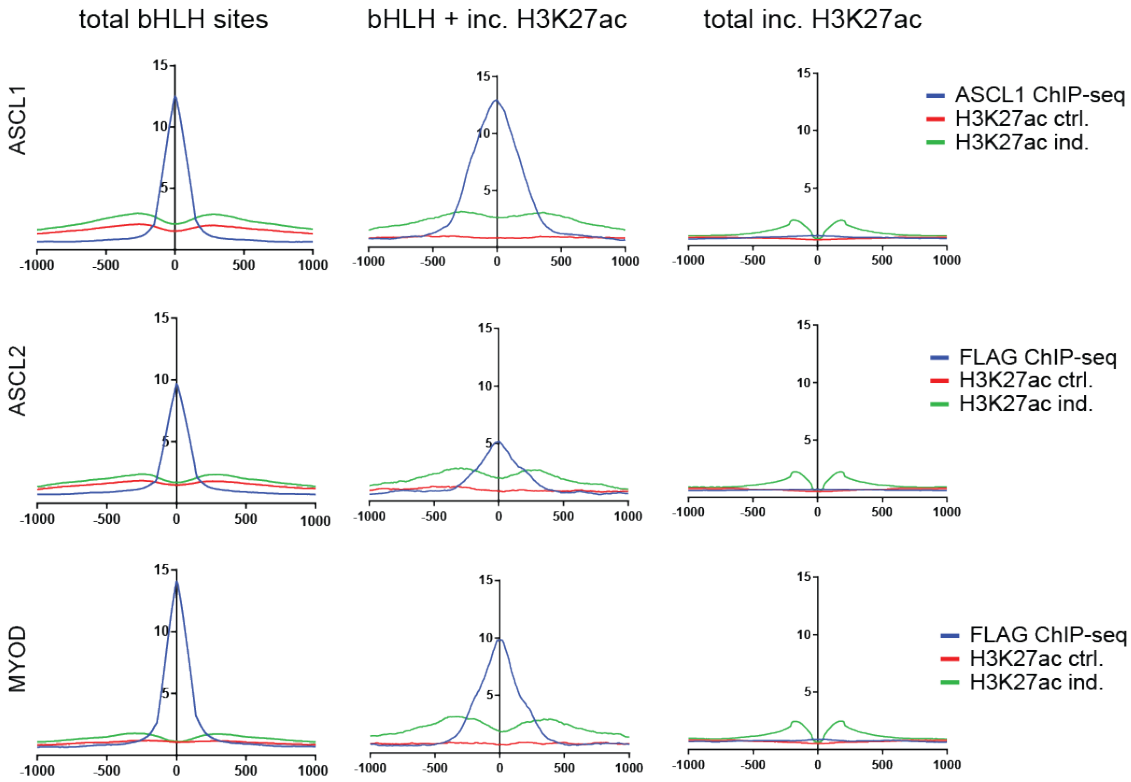
(H3K27ac analysis performed using -nfr parameter, default parameter showed comparable levels of overlap.)

Figure 4-17: Overlap comparison of bHLH binding sites and H3K27ac changes in ES cells



Sites of H3K27ac changes show limited coincidence with bHLH binding sites.
Proportional Venn diagram comparing overlapping intervals identified in ChIP-seq as bHLH binding sites or significantly changed in H3K27ac. Numbers reflect subsets of intervals in each component. H3K27ac changes identified through differential peak calling between control and induced samples from each ES cell line at 24h post induction, using HOMER's *findPeaksGenome* module with parameters *-histone* and *-nfr*.

Figure 4-16: Histogram comparison of bHLH and H3K27ac at total and overlapping sites



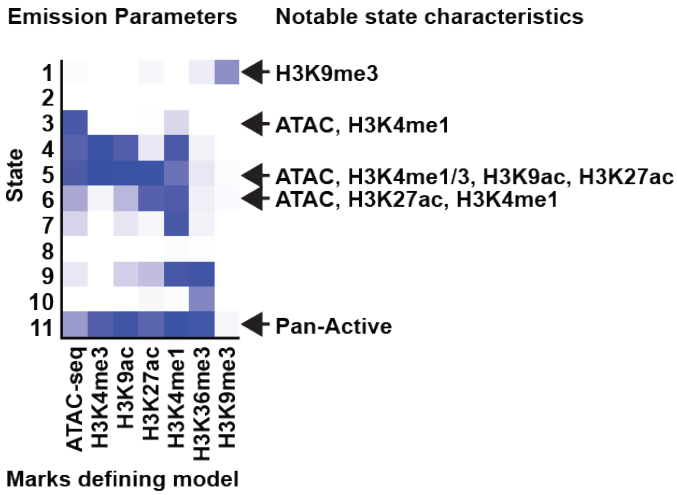
bHLH binding sites show increased H3K27ac enrichment within 24h

Histogram comparison of mean tag count at sites identified in each cell line. Histogram data generated based on total bHLH sites identified in ChIP-seq (left), sites identified as showing significant increase in H3K27ac (right), or subset of bHLH binding sites showing a significant increase in H3K27ac (center). H3K27ac changes were identified using *-histone* and *-nfr* parameters to identify peak regions with expected bimodal distribution. All data shown reflects comparisons within a given inducible ES cell line.

Figure 4-18: Comparison of *de novo* motifs identified at H3K27ac increases

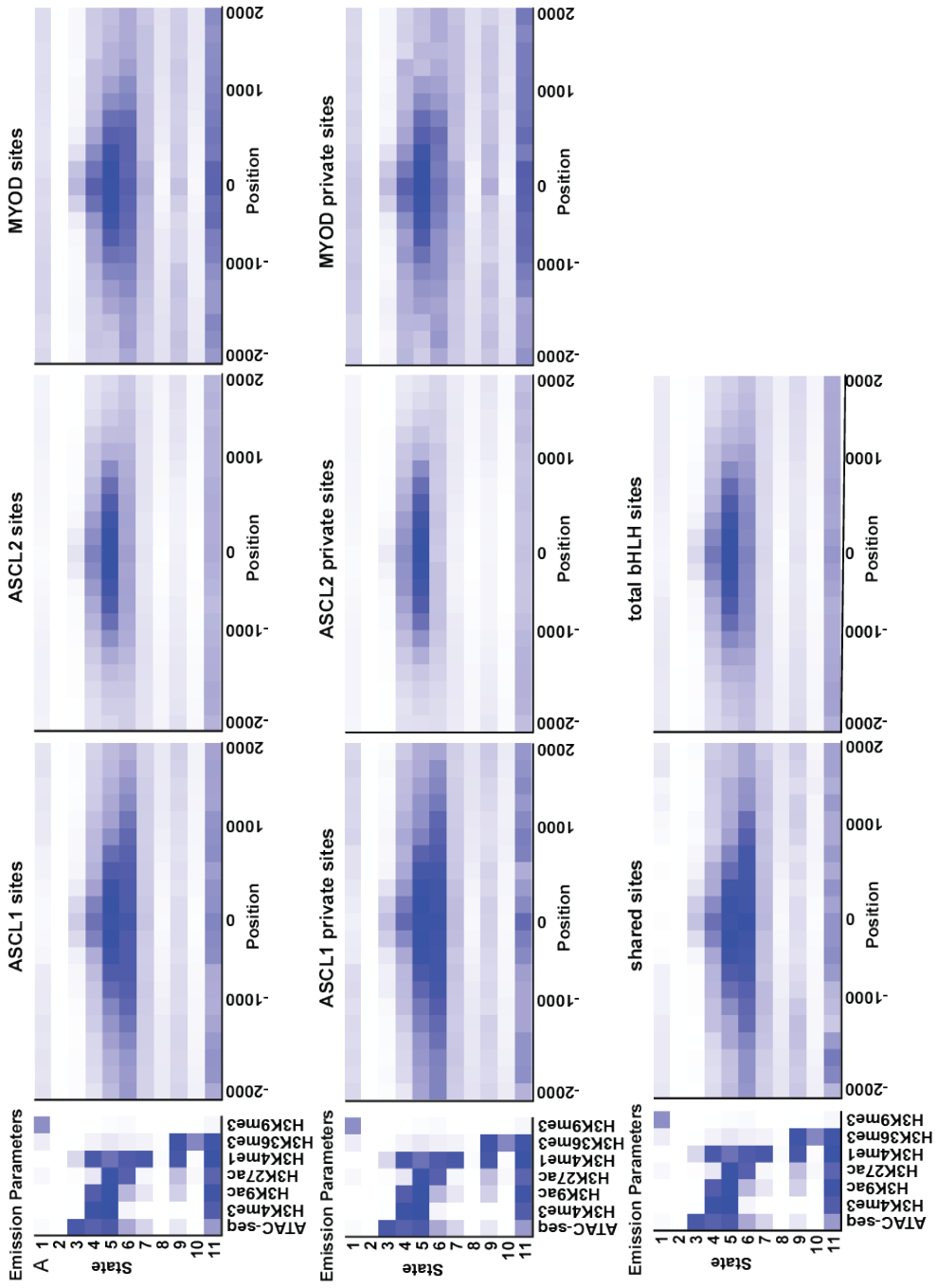
Comparison of *de novo* motifs identified at regions demonstrating increase in H3K27ac. Differential enrichment identified by HOMER *findPeaksGenome* using uninduced/induced H3K27ac ChIP-seq data sets. Differential peak calling performed with nucleosome free peak centering (parameter-*nfr)*. *de novo* motif analysis performed using 200bp interval centered on ATAC-seq change used for analysis. Numbers reflect the binomial significance of the *de novo motif* identified, the percentage of sites featuring the specified motif, and the percentage of normalized random background featuring the specified motif.

Figure 4-19: Emission states identified from Markov model



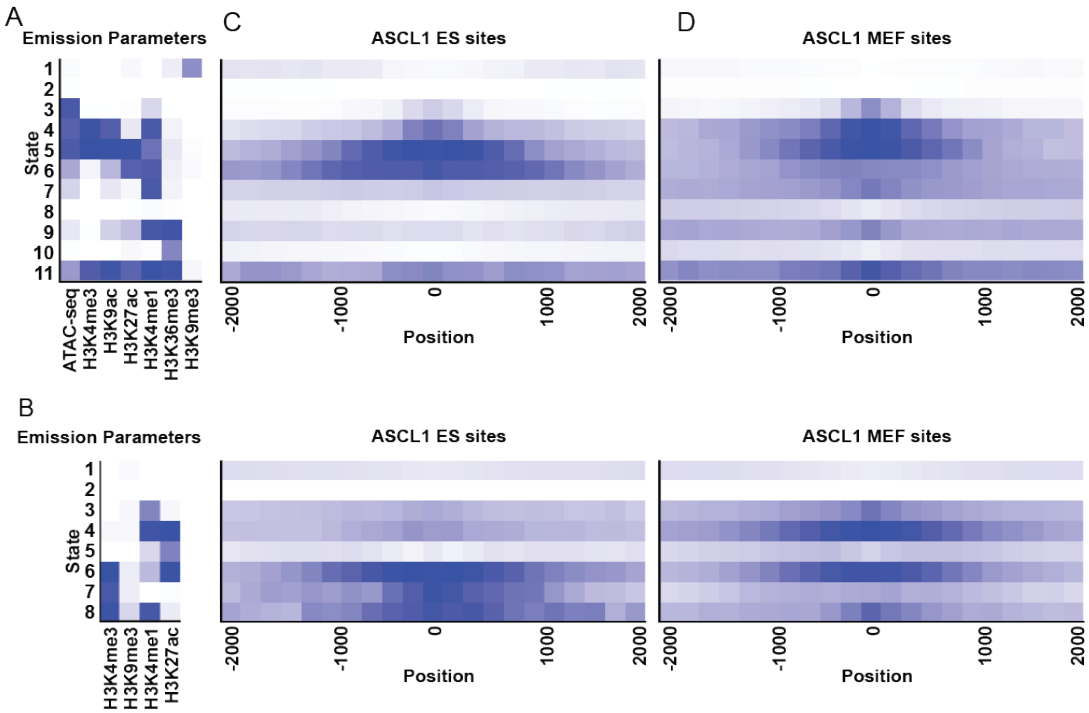
Comparison of chromatin states identified in 11 state model. Heatmap plot represents Markov model derived from ChIP-seq and ATAC-seq data sets from inducible ES cells (H3K27ac, ATAC), and ENCODE (H3K4me1, H3K4me3, H3K9me3, H3K36me3) from ES-E14 cells. Increasing association for each component of each state corresponds to darker shading. Arrows highlight specific combinations of features noted in text.

Figure 4-20: Comparison of HMM states at bHLH binding sites



Comparison of state enrichment for total, shared, and factor-specific subsets of bHLH binding sites identified in ChIP-seq in ES cells. All comparisons utilize 11-state Markov model (A) derived from chromatin data sets generated in uninduced (ATAC-seq, H3K27ac), or WT ES cells (The ENCODE consortium, 2013). State distribution plots show enrichment for these states at bHLH binding sites as indicated above.

Figure 4-21: Comparison of HMM states at ASCL1 binding sites in ES and MEF cells



Hidden Markov model comparison of ASCL1 binding sites identified in ES and MEF cells. Emission plots from (A) ES cells (11 state) and (B) MEF cells (8 state) shown with histone marks defining these states. State enrichment diagrams for ASCL1 binding site intervals identified in inducible ES cells at 24 hours post-induction (C), or *rtTE-Ascl1* transfected MEFs at 48 hours post transfection (D). Plots represent enrichment for Markov states, as identified from ChIP-seq for histone modifications in respective cell types.

CHAPTER FIVE
FUTURE DIRECTIONS

PROPOSED APPROACHES TO IDENTIFYING MECHANISMS UNDERLYING BHLH FACTOR SPECIFICITY

Introduction

In these studies, I directly compared the binding and transcriptional consequences of ASCL1, ASCL2, and MYOD when ectopically expressed ES cells. I found that these master regulators of cell fate each show a remarkable capacity to bind closed chromatin, and direct distinct transcriptional profiles. I characterized the chromatin environment of these cells through a number of measures, and observe the presence of markers of active chromatin near the bHLH binding sites. Despite these efforts, the fundamental mechanism by which these factors identify their specific complement of binding sites from the many potential sites within the genome, and engage transcription of relevant gene targets remains unexplained.

Comparison of bHLH factor binding and function at additional time points

The research presented here was performed exclusively at 24 hours post-induction of the bHLH factors tested. This time point was selected specifically to identify early events in bHLH function in a naïve cellular context. Analysis at this time point allowed different mechanisms for specificity of bHLH function to be tested. Another approach to identifying specificity might be the observation of additional time points, using a similar approach to that described here. As these class II bHLH factors function in partially differentiated cell

lineages, and are critical to lineage-appropriate cell fate specification, it is likely the case that they command key components of transcriptional cascades, and establish gene regulatory networks that provide further specification. Our ChIP-seq data captures early binding events, but provides only a single set of such sites to query for each factor. Observation of binding at later time points may reveal progressive changes in the binding of these factors. If this is the case, comparison of early-bound and late-bound sites may allow inference of novel mechanisms which specify temporal changes in TF binding. These rules of engagement may reveal how a single transcription factor can function to progressively define cell fate and establish multiple sub-lineages. This would allow for observation of the interaction between these factors and their direct and indirect targets, such as DNA-binding co-factors or chromatin remodeling enzymes, as well as the epigenetic and transcriptional consequences of sustained expression of these factors, but more importantly, may also allow for more specific comparison of the early sites identified.

One important finding from these studies is the observation that these class II bHLH factors each exhibit pioneering capacity. In addition to identifying binding to both open and closed chromatin for each factor tested, we also identify significant numbers of sites which exhibit bHLH-dependent changes in open chromatin, which are not limited to bHLH binding sites identified by ChIP-seq. However, these data are captured simultaneously, and bHLH-dependent changes are likely masked by this simultaneous observation. As with bHLH factor binding at later time points, further comparison of chromatin accessibility by ATAC-seq or other means may allow for identification of broader reorganization in response to bHLH expression.

Characterization of binding and activity of bHLH dimeric complexes

Class II bHLH factors interact with class I factors, known as E-proteins, to form a heterodimeric DNA binding complex. In vivo, and in vitro, these complexes have been shown to preferentially bind to CAGSTG E-boxes. However, binding and function of these complexes is not limited to the preferred E-box binding motif, and a number of distinct E-boxes have been functionally tested in reporter assays. Additionally, it has previously been demonstrated that in vitro, these factors can complex with different E-proteins. The structure of the bHLH domains of a number of class II factors has been fully solved through high-resolution X-ray crystallography, and these structures have conclusively demonstrated the basis of the heterodimer interaction in both heterodimeric, and homodimeric complexes. However, these structures were tested in a reduced system, and it is not known whether these factors exhibit E-protein preference, or what role the E-protein plays in defining the binding motifs of these factors. Furthermore, while MYOD has been crystallized in a homodimeric complex, it is unknown whether this complex is formed outside of a reconstituted system. Here I show that both CAGCTG and CAGGTG/CACCTG E-boxes are consistently the most significantly enriched E-boxes identified by ChIP-seq. It may be the case that these similar, but distinct motifs are the result of heterodimeric vs. homodimeric binding complexes. Alternately, it may be the case that in dimeric binding complexes, only one of the bHLH proteins present contributes to core dinucleotide preference, and its binding partner interacts only with the CA/TG terminal residues and/or the phosphate backbone, and thus does not

affect the motif preference observed. However, this does not necessarily imply that E-protein selection, or homodimeric/heterodimeric complex partners are not biologically meaningful. It may be the case that dimeric partner selection allows for additional functional specificity independent of motif selection. As previously described, bHLH factors are largely similar in their inward-facing, DNA-binding and HLH-formation surfaces, and differ primarily in their outward-facing residues, suggesting that they may exhibit differential co-factor binding capability. As such, selection of dimeric partners may allow for distinct co-factor interactions, based either on partner preference, or context specific differences in bHLH factor expression. Furthermore, class I bHLH proteins are further distinguished by their ability to interact with other bHLH families, notably including the class V *(Inhibitor-of-DNA-binding, ID)* factors, which function as repressors of bHLH function, and are believed to do so through competitive interactions with class I E-proteins. Here, we identify *Id* factors as early targets for bHLH-dependent changes in gene expression. This finding suggests that these factors may play a role in bHLH factor function, either by negatively regulating spurious activity of bHLH factors in precursor populations, or by selectively repressing the activity of these or other bHLH factors.

Two distinct approaches can be taken to address the possible role of class I factors in defining specificity of the class II factors tested here. One approach is to characterize binding of endogenous E-proteins to the genome. As with class II factors, this can be accomplished by performing ChIP-seq in the same inducible ES cells, using antibodies directed against specific E-proteins. E2A antibodies are widely used, and have been proven effective for ChIP-seq (Lin et al., 2010). However, this antibody cannot distinguish between E12 and E47,

two distinct isoforms transcribed from the *E2A* locus. Isoform-specific human E12 and E47 antibodies exist, but have not been reported for ChIP-seq (Active Motif 39016, Active Motif 39017). TCF4/E2-2 and TCF12/HEB antibodies are also available. Theoretically, by performing ChIP-seq from the ES cells with these antibodies, it should be possible to characterize the binding of these class I bHLH factors. By comparing the sites bound against the class II factors tested, it should be possible to identify homodimeric, and heterodimeric binding class II complexes through reductive analysis. By comparing the sites identified for these factors it may be possible to determine whether specific complexes have additional binding specificity, or functional significance.

One challenge of such an approach is that endogenously expressed factors are not expressed at equivalent levels. Additionally differences in antibody specificity or sensitivity add experimental variability to these assays. One way of testing the role of different bHLH dimers is to force these factors to interact with a specific binding partner by adapting the inducible ES cell system to express both factors as a fusion protein. The inducible factors are engineered to express a single bHLH with a carboxy-terminal FLAG tag. If the coding sequence for a specific binding partner, along with an unstructured linker region, were cloned between these components, a single protein could contain both halves of a homodimeric, or heterodimeric binding complex. Like the inducible ES cell model, expression of these fusion complexes could be controlled by the manipulation of doxycycline in culture media. This model would allow for serial testing of multiple binding partners. One of the strengths of this model is that the single transcript expressing the two component bHLH factors would serve to provide equivalent expression levels for the two components, thus circumventing the

problem of differences in expression levels between binding partners. This approach has previously been used to characterize homodimeric binding in vitro from bacterially expressed MASH1/ASCL1, and reportedly bind with higher affinity (Sieber et al., 1998). This would allow for characterization of complex binding, including complexes which may be biologically unlikely in ES cells, or alternative contexts, due to stoichiometric differences in available partners. As in the studies reported here, the genome-wide binding of these factors could be characterized using ChIP-seq. This may allow for identification of potential differences in the binding preferences of these bHLH dimeric complexes. If differences were identified, it may be the case that E-protein interaction is one mechanism by which these factors select their specific complement of binding sites. One possibility is that certain E-proteins impose greater E-box specificity on these complexes. If so, *de novo* motif analysis from ChIP-seq data sets from different complexes should reveal these differences. Alternatively, it may be the case that these heterodimeric complexes bind to the same E-box sequence, but interact with specific co-factors. As E12 and E47 are functionally active as homodimers, they may also provide class II complex partners with additional transactivational capacity. While we have performed similar analysis here, our model does not directly regulate E-protein selection, and identification of different binding sites for the different bHLH:E-protein complexes would allow for more specific comparisons, both for a given factor, and between factors.

Characterization of synthetic bHLH hybrids

Class II bHLH factors are defined by their shared bHLH DNA binding motif, and are known to bind Eboxes through interactions with class I bHLH factors. My results show that despite this similar structure, they bind similar E-boxes sequences, at distinct sites in the genome. Functionally, they specify different cell types, as demonstrated by their role in reprogramming of fibroblasts to muscle and neural cell types. The nature of this specificity remains elusive.

Previously, it has been shown that mutation of only a few key residues within the bHLH domain is sufficient to confer functional specificity onto the ubiquitous class I bHLH E12 (Weintraub et al., 1991; Davis et al., 1992). Functionally, chimeric hybrids of ASCL1 and MYOD were tested in chicken neural tube using *in ovo* electroporation, and demonstrated that the bH1 domain of ASCL1 was sufficient to recapitulate a neural differentiation phenotype onto MYOD (Nakada et al., 2004). Binding sites of these proteins were not characterized in either study. However, the preferred E-box sequence of these master regulators is essentially the same, both in differentiated cell types (Cao et al., 2010; Borromeo et al., 2014) and in when ectopically expressed in ES cells, as described here. These data suggest that features other than sequence specificity dictate which specific E-boxes are bound from the large number present within the genome, and that the source of this specificity lies in the bHLH domain. However, while these studies clearly demonstrate a factor-specific role for the bH1 domain, it does not appear to be the case that this domain is solely responsible for binding specificity. A subset of MYOD binding sites, including an important site regulating the myogenic factor *Myogenin,* are associated with a specific interaction between MYOD and PBX, and both the amino and carboxy terminal regions are

necessary for this interaction, and stable binding of MYOD (Berkes et al., 2004). More recently, MYOD has been shown to demonstrate altered binding to NEUROD2 sites by replacement of the bHLH domain with that of NEUROD2, and additional mutations outside the bHLH (Fong et al., 2015), suggesting that this approach can be used to identify the components of these factors responsible for binding and transcriptional specificity.

Structurally, there are two possibilities which might explain this specificity. One possibility is that the distinct differences present within the bHLH domain direct binding to distinct sites through interactions with co-factors, chromatin, or the DNA itself. If this is the case, binding specificity should be primarily a product of the bHLH domain of the factors tested. Alternately, it may be the case that regions outside the bHLH are central to site selection, potentially through co-factor interactions. Both can be directly tested, accordingly.

The inducible ES cell model utilized here provides a potential strategy to identify specific features of these bHLH factors which are responsible for their distinct binding and function. By generating ES cells which express chimeric hybrids of the bHLH factors tested here, it is possible to identify specific features of these TFs which provide for factor-specific binding. Using a directed design approach, such as that used by *Nakada et al.,* specific components of the bHLH domains of these factors can be combined to create chimeric bHLH proteins, and the binding of these factors can be compared as in my study. By comparing this binding to that of unaltered bHLHs, this may reveal the source of the distinct binding observed. If specific components of the bHLH domain of these proteins are the source of this specificity, these residues can be mutated, and their effect on binding and transcriptional influence can be observed either genome-wide, or in a site-specific manner. While previous

comparisons have suggested that functional specificity resides within the bHLH domain, binding specificity may require additional interactions which have not been observed.

Alternatively, it may be the case that, while the bHLH domain is critical for DNA binding at the preferred E-box motif, it contributes little to the distinct patterns of genome-wide binding observed for these factors. Crystal structures have not identified specific interactions with DNA outside the bHLH domain, however, these were observed in truncated form to allow for crystallization (Ma et al., 1994; Ellenberger et al., 1994; Longo et al., 2004), suggesting that other parts of the protein may play a role in selection of cognate binding sites. If this is the case, selection of appropriate binding sites from the many instances of the preferred E-box may be partially mediated by regions outside the bHLH itself. Previously, it has been demonstrated that regions outside the bHLH domain are involved in directing gene expression of myogenic targets (Gerber et al., 1997; Berkes et al., 2004), and that their addition can functionally confer transactivating potential on a chimeric bHLH in which the basic DNA-binding domain of MyoD replaces that of E12 (Weintraub et al., 1991; Blackwell et al., 1997). By interchanging these domains between factors, or by deleting them entirely, we can test whether they are central components of binding selection. Together, these component-based approaches will provide further insight into how these factors function, and may identify the precise components of these bHLH factors which underlie their specific binding and function.

Characterization of pioneering ability of additional bHLH transcription factors

Here, I have directly compared binding of ASCL1, ASCL2, and MYOD within a common cellular context, demonstrating that each demonstrates pioneering capacity evidenced by binding to closed chromatin. This may be a key capability for the role of these bHLH factors. However, it remains unknown whether this is a feature of all bHLH factors, or whether it is specific to a subset of class II bHLH factors. It has previously been shown that pioneering capability is not a feature of all transcription factors (Soufi et al., 2012; Wapinski et al., 2013). It has specifically been suggested that MYC, a class IV bHLH factor, has dramatically reduced pioneering capacity as compared to OCT4, SOX2, and KLF4, which represent the other members of the pluripotent cocktail of "Yamanaka factors" (Soufi et al., 2012; Soufi et al., 2015). Class I bHLH factors, which function both as homodimeric complexes, and heterodimerize with class II and class V factors, are a particularly compelling candidate for study in this system. While we observe the binding of the bHLH factors tested here to closed chromatin, we have not tested whether specific dimeric partners are present at different sets of binding sites. It may be the case that some aspects of the pioneering ability of these class II factors is dependent on specific class I partners. One way of testing the limitations, and relative abilities of these factors to bind to closed chromatin would be to characterize their binding in ES cell by ChIP-seq, in the presence and absence of these or other class II bHLH factors, allowing for direct comparison of the chromatin accessibility of observed binding sites in a common cellular context. *Myc*, *Mycn*, *Tcf3* (E12/E47), and *Tcf4* (E2-2) cell lines have been generated in the same manner as the ES cells utilized here (Nishiyama et al., 2009; Correa-Cerro et al., 2011), which allow the same ChIP-seq approach using a FLAG antibody. While pioneering ability could be tested in other contexts, the use of

a common system will allow for direct comparison between factors, and the use of this ES cell model should not require dramatic optimization. This may allow for identification of what specific feature of these factors establishes pioneering ability, and what extent pioneering ability is a feature of bHLH factors as a family.

Expanded comparison of histone modifications

It has previously been reported that ASCL1 preferentially binds to a conflicted trivalent signature of H3K4me1, H3K27ac, and H3K9me3 when observed in a fibroblast-to-neuron reprogramming paradigm. Here, I have characterized the presence of these marks in the context of ES cells, and find that the sites bound by class II bHLH factors were similarly enriched for H3K4me1, H3K4me3, and H3K27ac, with very modest enrichment for H3K9me3. This suggests that differential binding by these factors is not likely due to differences in preferences for this signature. However, these modifications represent only a small subset of the total set of histone modifications observed. Additionally, these marks have been characterized based on their presence at large numbers of active or repressive sites, suggesting that these are representative of relatively common changes throughout the genome. Thus, their pattern of enrichment may not provide sufficient complexity for the distinct patterns of binding observed. It may be that differential binding is based on a signature of less prevalent histone modifications, and that this signature has not been observed due to lack of direct comparison of these marks.

Using ChIP-seq data, more complex Markov models can be built based on additional data sets, including histone markers. Public data sets, such as the mouse ENCODE database, provide access to a number of these data sets, including those derived from ES cells. Using these additional data may allow further refinement of the binding sites identified. In addition to genome-wide binding data, the chromatin remodeling pathways responsible for regulating many of these histone modifications are known. If a signature is identified, these represent candidates for experimental validation of specific components of the signature identified. Furthermore, this undirected modeling approach can theoretically be used to incorporate disparate types of sequencing data, such as the ATAC-seq data used in our analysis. In addition to histone marks, the genome-wide distribution of many transcription factors has been determined in ES cells. While no specific co-factor binding motifs were suggestive of bHLH:co-factor complex binding, other TFs may serve to influence bHLH binding. Together with the bHLH binding data presented here, this may allow for better characterization of the binding sites selected from a complex chromatin landscape.

Characterization of bHLH binding through inducible differentiated cell types

One particularly compelling feature of ASCL1 and MYOD is their remarkable ability to trans-fate cells from one lineage to another. Recently, both have been studied in the context of fibroblast reprogramming strategies, alone, and with other transcription factors. However, this activity has not been directly compared for these factors in this context. While the inducible ES cells used here are particularly useful as a reductive model, it is likely the

case that the binding and activity of these bHLH factors is influenced by this context. It has long been observed that ES cells spontaneously differentiate into fibroblasts, and this must be overcome through constitutive inclusion of LIF and fetal serum to maintain proliferative, pluripotent cell growth (Evans et al., 1981; Martin et al., 1981). However, this property can be also be utilized as a distinct experimental strategy to observe the changes in binding in distinct cellular contexts. By altering culture conditions, inducible ES cultures can be readily differentiated, generating inducible fibroblast cultures which would be otherwise genetically identical to the inducible ES cells utilized here. These cultures can then be induced to express the bHLH factors compared here, and their binding can be observed in fibroblasts for comparison between these contexts. While the binding of these factors in fibroblasts has previously been observed through enhancer screens (Johnson et al., 1992; Davis et al., 1992), and ChIP-seq (Cao et al., 2013; Fong et al., 2012; Wapinski et al., 2013; Treutlein et al., 2016), the use of a common inducible expression system may allow for direct comparison of these factors across cellular contexts, as the cellular environment and induction dynamics would presumably be more comparable than models relying on primary fibroblast cultures.

While the specific mechanism which provides for distinct genome-wide binding of these factors has not been revealed here, their direct comparison alone, or in a cocktail of reprogramming factors may reveal specific differences which are key to their role in lineage reprogramming. This approach can further be expanded through the use of specific protocols which allow for specified differentiation into distinct specified lineages. This would allow for observation of the binding and activity of these factors in factor-appropriate and alternative cell lineages, using similar approaches as described here. A number of such directed

differentiation approaches have been demonstrated, and it may be possible to directly compare bHLH binding and activity in specific cell types which would be otherwise technically challenging, or impossible, such as expression of ASCL1 in differentiated muscle lineages or MYOD in neural lineage cells. This may provide insight as to the roles of established chromatin features or co-factors in shaping the function of the bHLH factors tested here. These approaches would allow for observation of how these factors function in partially, or fully specified cell fates, and may reveal new insight into the specific complement of features which provide specificity of binding and function for these factors.

Testing the role of phosphorylation on bHLH binding by ASCL1 and MYOD

It has previously been demonstrated that phosphorylation of ASCL1 affects its ability to function in *Xenopus laevis,* and that hypophosphorylation mutations lead to improved neuronal maturity in fibroblast to neuron paradigms (Ali et al., 2014). Similar results have been identified for *Neurogenin2*, (Ali et al., 2011), a related neurogenic class II bHLH factor which shares a number of similarities with ASCL1, including interactions with HES proteins and lateral inhibition via NOTCH pathway involvement. Intriguingly, through in vitro stability assays and ChIP-qPCR, this was revealed to rely at least partially on interaction with E-proteins, and phosphomutant NEUROG2 was shown to have enhanced stability with E12, which was further demonstrated to increase binding to *Neurod1* and *Dll1* promoters, which are known targets of NEUROG2 (Ali et al., 2011). They further demonstrated that increased phosphorylation led to progressively diminished binding, suggesting a progressive negative regulatory mechanism for NEUROG2 binding and activity. Compellingly, they suggest that

this may provide a mechanism for the previously noted distinction in binding activity between these two promoters, of which *Neurod1* has been suggested to require extensive chromatin modification for binding and activity (Koyano-Nakagawa et al., 1999; *as interpreted in* Ali et al., 2011). This suggests that phosphorylation of a related bHLH factor may affect its ability to bind to a subset of its cognate binding sites which are associated with repressive chromatin. If ASCL1 binding were similarly regulated, it would indicate that specific kinases might play a role in the distinct binding identified for this factor. Phosphorylation-null forms of proneural bHLH proteins show enhanced function in fibroblast-to-neuron reprogramming paradigms (Ali et al., 2011; Ali et al., 2014), and this change is not due to enhanced or diminished stability of these mutant forms (Ali et al., 2014). This suggests that phosphorylation negatively regulates the function of proneural bHLH factors through a separate mechanism.

Interestingly, phosphorylation of MYOD by kinases of the *Cdk* family has also been observed, but these modifications have been characterized as regulators of protein degradation (Song et al., 1998; Kitzmann et al., 1999). However, perhaps the most intriguing result comes not from MYOD studies, but those of its avian homolog *CMD1*; DNA binding activity of phosphorylated CMD1 homodimers is lost, but binding of CMD1:E12 is not affected (Mitsui et al., 1993). Thus, phosphorylation of this protein can alter dimer formation and binding, perhaps providing for distinct ability to bind specific sites, such as those in closed or repressive chromatin (as described above in regards to dimerization). While uncharacterized, this may also be the case for ASCL1, or other class II bHLH factors.

While studies regarding the role of phosphorylation could be performed in other models, its observation and manipulation in this system would allow for comparison of binding of ASCL1 and MYOD in a common cellular context, providing for direct comparisons between these factors. One approach to this analysis would be to generate ES cells expressing a phosphomutant (Ali et al., 2014) in place of the wild type ASCL1 transgene. These could then be induced, and their ability to dimerize could be assayed using Co-IP. This would allow for direct testing of whether phosphorylation of ASCL1 is subject to differential dimerization based on its phosphorylation state.

Using ChIP-IP, the binding of these specific complexes to specific enhancer regions could be compared, capitalizing on the characterization of sites identified in this study. If phosphorylation of ASCL1 affects binding to specific sites, such as those in open or closed chromatin, or those associated with a specific chromatin signature, this may explain some of the distinct binding observed. ASCL1 contains a conserved serine residue present in the loop region, adjacent to the amino terminus of helix 2, and phosphorylation of this site may affect either H1-H2 interaction, or bHLH interaction with the DNA itself. This may potentially provide a structural mechanism for the distinct flanking sequence identified for ASCL1 and ASCL2 (as discussed in Chapter 3). Based on alignment to the similar class II bHLH NeuroD, this residue is the physically closest residue to this position in the motif, and thus represents a potential mechanism in the additional selectivity for this position. Furthermore, if ASCL1 phosphorylation is a significant aspect of its activity, it may be possible to disrupt this activity using selective protein kinase inhibitors. As ASCL1 is thought to be a factor in neural and lung cancers (Borromeo et al., 2016), the pathway modulating its phosphorylation

represents a potential drug target. A MYOD phosphomutant also exists, and if introduced into the ES cell system, could then be compared to test whether phosphorylation may play a role in the distinct binding observed for these factors.

Investigating potential co-factors identified from motif analysis

ASCL1 and MYOD are central regulators of cell fate, and have now been thoroughly demonstrated as central components of cellular reprograming cocktails (Davis et al., 1987; Dekel et al., 1992; Farah et al., 2000; Vierbuchen et al., 2010), however, mechanisms by which this is accomplished, and the limitations to which it can be performed in different lineages are poorly characterized. In these studies, I identified a number of genes which are expressed in response to one or more of the bHLH factors tested. While some genes with clear potential (such as *Gcm1* in ASCL1, and *Mef2d* and *Smarcd3* in MYOD are intriguing based on their previous characterization, broadly, the transcriptional programs established by these factors in the ES cell system are not dramatically enriched for obvious mediators of lineage specification. However, these or other targets of ASCL1 and MYOD likely include downstream effectors of lineage specification, and these are likely not characterized in this role. As such, these early targets represent candidates for future analysis in reprogramming via fibroblast to neuron or fibroblast to muscle paradigms. By expressing these candidates along with ASCL1 or MYOD, and observing the reprogramming efficiency, a synergistic role for these genes as potential facilitators of lineage reprogramming can be tested. Finally, it may be that these genes are necessary downstream components for the reprogramming

roles of these factors. By performing knockdown or knockout of these genes, and comparing the activity of ASCL1 or MYOD in reprogramming assays, can be directly tested. Such an approach may allow for identification of genes which are also significant in developmental lineage specification, but whose function remains undiscovered due to their regulation by these master regulators.

Role of DNA Methylation in class II bHLH binding

In this research, I characterized the binding and function of three class II bHLH factors within a common cellular context, and have partially characterized the chromatin landscape that defines this context. Vexingly, this has failed to identify a clear determinant of how these factors distinctly select their binding sites to drive lineage-specific gene transcription. There remain many aspects of this landscape, however, which we have not tested. One particularly intriguing possibility is that presented by DNA methylation, which occurs not in the histone proteins associated with DNA, but within the nucleotides of the double helix. Cytosine can be methylated into multiple forms through the activity of a class of regulators known as DNA methyltransferases, including the *DNMT* family. This form of methylation occurs specifically at CpG positions. CpG features are relatively uncommon genome-wide, and methylation of these sites has long been implicated in modulation of gene expression. Methylation over gene bodies has specifically been implicated as a repressive regulatory mechanism. In the blastocyst, CpG methylation levels are relatively low, and these features are increasingly methylated throughout development, and in adult animals over 90%

of CpG islands are methylated. The central importance of this process is highlighted by the observation that *DNMT1* mutant ES cells are stable, but *DNMT1*$^{-/-}$ embryos die at midgestation (Li et al., 1992). This suggests that methylation of these positions occurs simultaneously with the process of lineage specification.

I determined that in ES cells, ASCL1, ASCL2, and MYOD preferentially bind to CAGSTG E-box features genome-wide. Within the binding sites, both GC and GG/CC dinucleotides are well represented. These sites are distinguished by the differential capacity for methylation. One possibility is that these bHLH factors may preferentially bind to a specific methylated or unmethylated form of this site. Such preference is not without precedent, as MECP2 is characterized largely based on its ability to preferentially bind to methylated CpG islands. MYC, a class IV bHLH, has been shown to preferentially bind to unmethylated CpG sites, and this preference results in observable changes in gene expression (Perini et al., 2005). It may be that the class II bHLH factors tested here may also preferentially bind to an E-box with a specific methylation state. This may lead to reduced degeneracy in the central dinucleotide of the bound E-boxes in cells with a greater degree of DNA methylation, and provide a mechanism for progressive changes in bHLH binding. ASCL1, ASCL2, and MYOD exhibit a preference for CAGCTG E-boxes in differentiated cell types (Borromeo et al., 2014; Liu et al., 2014; Cao et al., 2010). Comparatively, the *de novo* motifs identified from ChIP-seq in the ES cells reveal greater degeneracy than reported in the differentiated cell types. This is supportive of a model in which progressive methylation of E-boxes leads to progressive changes in the binding of these factors. This is

also intriguing in light of the observation that DNA methylation plays a role in chromatin structure. (Keshet et al., 1986)

To test this possibility, it is necessary to determine the distribution of methylated CpG sites. Using bisulfite sequencing, it is possible to characterize DNA methylation genome-wide. Comparison with E-box features and with empirically defined bHLH binding sites would allow for observation of the methylation state of the specific binding sites identified in the ES cell model. A preliminary screen of the methylation state at observed bHLH binding sites may be possible through the use of publically available data sets which characterize DNA methylation in different cell types, including murine ES cells (Zhao et al., 2014). Comparison of binding sites identified in ES cells may allow preliminary characterization of the distribution of these features. Should significant overlap suggestive of methylation state be identified for one, or more of these factors, genome-wide sequencing could be used to directly characterize the distribution of methylated cytosine in the context of the inducible ES cells used here. Further characterization could include manipulation of the cellular machinery responsible for regulating DNA methylation, potentially observing the effects of knockdown or overexpression of these components on both DNA methylation at observed binding sites and bHLH binding through further ChIP-seq, or ChIP-qPCR at specific sites. This approach could further be used in combination with differentiation and reprogramming experiments, as previously discussed.

Postscript

Historically, identification of the mechanisms by which lineage-specifying transcription factors has been limited to the observation of small subsets of their genome-wide binding, and complement of transcriptional activity. While the dedicated efforts of countless researchers have dramatically revised and refined our understanding of transcription factor function through decades of careful work, the identification of the fundamental mechanisms underlying their regulation and activity is still incomplete. Recent advances in experimental technology, including high-throughput sequencing of DNA and RNA, and the ability to readily generate mutant animal lines is facilitating a dramatic expansion of our knowledge of these and other transcription factors. These approaches allow for observation of transcription factors in greater depth and detail than any conventional methodology. Here, I have proposed a set of experiments that leverage these new technologies to approach fundamental questions in how this familial class of transcription factors engages the shared genome to create the myriad disparate cell fates of the organism.

APPENDIX 1
PCR AND RT-QPCR PRIMERS

Primers used to evaluate bHLH expression in inducible ES cells		
Gene Symbol	**Forward Sequence**	**Reverse Sequence**
Gene specific qRT-PCR primers for total transcript containing ORF		
Venus	CAACAGCCACAACGTCTATATCACCG	CTTTACTTGTACAGCTCGTCCATGCC
Ascl1	CACCATCTCCCCCAACTACTCCAAC	GAACCAGTTGGTAAAGTCCAGCAGC
Ascl2	ATGGAAGCACACCTTGACTGGTACG	TTTGCACCTTCACGGGCCTC
MyoD1	GTGGCGACTCAGATGCATCCAG	GTCGTAGCCATTCTGCCGCC
Gene specific qRT-PCR primers for endogenous transcript		
Ascl1	TTAGCCCAGAGGAACAAGAGCTGC	TGCTTCCAAAGTCCATTCCCAGG
Ascl2	AGGAGCTGCTTGACTTTTCCAGTTG	TTGGGCTAGAAGCAGGTAGGTCCAC
Myod1	ATCCAGCCCCAAAGAAAGGACATAG	TGGCCACTCAAGGATCAGCTCTG
H2A specific qRT-PCR primer used in normalization		
H2A (H2afz)	TTGCAGCTTGCTATACGTGGAGATG	TGTTGTCCTTTCTTCCCGATCAGC
ES cell line genotyping		
ROSA26 wild type	AAAGTCGCTCTGAGTTGTTAT	GGAGCGGGAGAAATGGATATG
ROSA26 knock-in	AAAGTCGCTCTGAGTTGTTAT	ACCCTGGGGTTCGTGTCC

APPENDIX 2
Expression of genes associated with developmental signaling pathways identified from RNA-seq from each ES cell line

		Average RPKM						Fold Change (Avg ind. Vs. Avg ctrl.)			Significance of Change		
	Gene Symbol	Ascl1_ctrl	Ascl1_ind	Ascl2_ctrl	Ascl2_ind	MyoD_ctrl	MyoD_ind	FC_Ascl1	FC_Ascl2	FC_MyoD	Ascl1 FDR	Ascl2 FDR	MyoD FDR
ESC Pluripotency													
	Fgfr2	3.06	7.92	3.31	9.49	3.63	9.30	2.59	2.87	2.56	3.39E-06	9.55E-08	1.26E-06
	Fgf5	0.27	4.14	0.94	8.81	1.10	17.87	15.31	9.39	16.27	2.46E-09	7.90E-07	6.63E-11
	Fgf14	0.05	0.25	0.20	0.31	0.25	0.21	4.73	1.52	0.84	2.48E-03	1.00E+00	1.00E+00
	Fgf15	5.28	11.23	7.91	9.42	8.25	11.90	2.13	1.19	1.44	1.41E-02	1.00E+00	6.96E-01
	Fgf18	1.71	3.75	1.22	4.90	1.43	0.70	2.19	4.03	0.49	6.44E-01	5.93E-03	5.68E-01
	Sepp1	2.78	8.13	3.34	10.56	3.50	10.09	2.93	3.16	2.89	1.88E-03	3.56E-04	7.37E-04
	Wnt5b	0.57	2.12	0.36	0.99	0.43	0.56	3.74	2.74	1.29	2.34E-05	8.43E-03	1.00E+00
	Wnt10a	0.04	0.29	0.04	0.55	0.10	0.29	8.27	13.97	2.81	3.03E-03	1.00E-05	2.21E-01
	Wnt16	0.02	0.16	0.06	0.25	0.06	0.02	6.80	4.56	0.30	3.30E-02	1.98E-02	9.72E-01
IRF activation													
	Dtx4	2.05	3.86	3.06	5.47	3.10	8.37	1.88	1.79	2.70	5.26E-01	5.76E-01	6.68E-03
	Nlrp4a	0.32	0.91	0.22	0.76	0.18	0.52	2.86	3.45	2.87	2.46E-02	2.87E-03	2.00E-02
	Nlrp4c	0.22	0.71	0.33	0.50	0.30	0.60	3.26	1.51	1.97	1.65E-02	1.00E+00	4.03E-01
	Zbp1	0.51	3.22	0.87	6.43	0.91	4.32	6.29	7.42	4.77	6.33E-04	6.24E-05	2.31E-03
Hippo Signaling													
	Id1	42.04	160.87	63.42	232.64	66.33	222.88	3.83	3.67	3.36	2.87E-03	3.14E-03	4.33E-03
	Id2	11.94	70.36	18.07	166.10	19.83	216.68	5.89	9.19	10.93	1.33E-06	3.78E-10	3.29E-12
	Id3	33.43	219.36	28.65	455.96	30.85	229.40	6.56	15.91	7.44	2.05E-15	6.34E-31	9.91E-18
	Id4	1.11	3.65	1.57	8.29	1.47	11.37	3.29	5.29	7.71	1.49E-02	2.10E-05	1.51E-08
	Ccnd3	80.80	82.99	64.89	84.67	62.72	166.10	1.03	1.30	2.65	1.00E+00	1.00E+00	2.33E-03
	Smad3	2.26	4.55	2.52	6.26	2.38	6.46	2.02	2.48	2.71	1.26E-01	6.60E-03	8.39E-04
	Snai2	0.42	0.67	0.35	1.26	0.27	1.74	1.60	3.58	6.49	1.00E+00	7.09E-04	8.16E-09
	Bmp7	0.65	2.50	0.97	4.03	0.95	4.16	3.84	4.14	4.37	7.29E-05	9.05E-06	1.04E-06
	Wnt5b	0.57	2.12	0.36	0.99	0.43	0.56	3.74	2.74	1.29	2.34E-05	8.43E-03	1.00E+00
	Wnt7a	0.14	0.05	0.10	0.03	0.22	0.03	0.36	0.32	0.14	4.40E-01	5.37E-01	1.98E-04
	Wnt10a	0.04	0.29	0.04	0.55	0.10	0.29	8.27	13.97	2.81	3.03E-03	1.00E-05	2.21E-01
	Wnt10b	0.01	0.13	0.05	0.17	0.09	0.09	11.03	3.77	1.05	7.82E-03	1.13E-01	1.00E+00
	Wnt16	0.02	0.16	0.06	0.25	0.06	0.02	6.80	4.56	0.30	3.30E-02	1.98E-02	9.72E-01
Notch Signaling													
	Notch2	3.33	3.94	3.31	3.94	2.94	6.63	1.18	1.19	2.26	1.00E+00	1.00E+00	4.38E-02
	Notch3	5.23	7.20	6.88	8.07	6.29	19.68	1.38	1.17	3.13	1.00E+00	1.00E+00	6.96E-03
	Lfng	3.13	16.37	3.10	13.75	4.06	31.16	5.23	4.44	7.66	4.26E-12	8.34E-10	8.35E-19
	Dll1	1.25	10.38	2.40	11.28	1.87	8.43	8.30	4.70	4.52	1.47E-17	7.91E-10	7.63E-10
	Dll3	1.61	21.28	2.75	9.07	2.13	1.35	13.19	3.30	0.64	4.03E-26	5.77E-06	5.36E-01
	Dll4	0.04	0.23	0.06	0.20	0.06	0.06	5.49	3.40	0.99	3.54E-03	1.05E-01	1.00E+00
	Dtx4	2.05	3.86	3.06	5.47	3.10	8.37	1.88	1.79	2.70	5.26E-01	5.76E-01	6.68E-03
	Nfkbia	20.34	31.74	16.37	33.05	15.54	29.05	1.56	2.02	1.87	4.09E-01	5.78E-03	1.42E-02
	Smad3	2.26	4.55	2.52	6.26	2.38	6.46	2.02	2.48	2.71	1.26E-01	6.60E-03	8.39E-04

Table shows selected genes from developmental signaling pathways identified as significantly upregulated in response to bHLH induction. Genes shown were identified as components of the associated pathways by CPDB analysis. Values reflect average RPKM across three biological replicates, with fold change indicated for each factor. Significance represents the edgeR calculated significance of observed change across three biological replicates for each condition. Coloration indicates genes which were identified as significant for each bHLH factor tested.

REFERENCES

Ali, F., Hindley, C., McDowell, G., Deibler, R., Jones, A., Kirschner, M., Guillemot, F., & Philpott, A. (2011). Cell cycle-regulated multi-site phosphorylation of Neurogenin 2 coordinates cell cycling with differentiation during neurogenesis. *Development, 138*(19), 4267-4277. doi:10.1242/dev.067900

Ali, F. R., Cheng, K., Kirwan, P., Metcalfe, S., Livesey, F. J., Barker, R. A., & Philpott, A. (2014). The phosphorylation status of Ascl1 is a key determinant of neuronal differentiation and maturation in vivo and in vitro. *Development, 141*(11), 2216-2224. doi:10.1242/dev.106377

Allen, R. E., Rankin, L. L., Greene, E. A., Boxhorn, L. K., Johnson, S. E., Taylor, R. G., & Pierce, P. R. (1991). Desmin is present in proliferating rat muscle satellite cells but not in bovine muscle satellite cells. *J Cell Physiol, 149*(3), 525-535. doi:10.1002/jcp.1041490323

Alonso, M. C., & Cabrera, C. V. (1988). The achaete-scute gene complex of Drosophila melanogaster comprises four homologous genes. *EMBO J, 7*(8), 2585-2591.

Alt, F. W., DePinho, R., Zimmerman, K., Legouy, E., Hatton, K., Ferrier, P., Tesfaye, A., Yancopoulos, G., & Nisen, P. (1986). The human myc gene family. *Cold Spring Harb Symp Quant Biol, 51 Pt 2*, 931-941.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet, 23*(2), 185-188. doi:10.1038/13810

Andrews, J. L., Zhang, X., McCarthy, J. J., McDearmon, E. L., Hornberger, T. A., Russell, B., Campbell, K. S., Arbogast, S., Reid, M. B., Walker, J. R., Hogenesch, J. B., Takahashi, J. S., & Esser, K. A. (2010). CLOCK and BMAL1 regulate MyoD and are necessary for maintenance of skeletal muscle phenotype and function. *Proc Natl Acad Sci U S A, 107*(44), 19090-19095. doi:10.1073/pnas.1014523107

Aronheim, A., Shiran, R., Rosen, A., & Walker, M. D. (1993). The E2A gene product contains two separable and functionally distinct transcription activation domains. *Proc Natl Acad Sci U S A, 90*(17), 8063-8067.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res, 37*(Web Server issue), W202-208. doi:10.1093/nar/gkp335

Beres, T. M., Masui, T., Swift, G. H., Shi, L., Henke, R. M., & MacDonald, R. J. (2006). PTF1 is an organ-specific and Notch-independent basic helix-loop-helix complex containing the mammalian Suppressor of Hairless (RBP-J) or its paralogue, RBP-L. *Mol Cell Biol, 26*(1), 117-130. doi:10.1128/MCB.26.1.117-130.2006

Berkes, C. A., Bergstrom, D. A., Penn, B. H., Seaver, K. J., Knoepfler, P. S., & Tapscott, S. J. (2004). Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential.
. *Mol Cell, 14*(4), 465-477.

Borromeo, M. D., Meredith, D. M., Castro, D. S., Chang, J. C., Tung, K. C., Guillemot, F., & Johnson, J. E. (2014). A transcription factor network specifying inhibitory versus excitatory neurons in the dorsal spinal cord. *Development, 141*(14), 2803-2812. doi:10.1242/dev.105866

Borromeo, M. D., Savage, T. K., Kollipara, R. K., He, M., Augustyn, A., Osborne, J. K., Girard, L., Minna, J. D., Gazdar, A. F., Cobb, M. H., & Johnson, J. E. (2016). ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. *Cell Rep, 16*(5), 1259-1272. doi:10.1016/j.celrep.2016.06.081

Braun, T., Buschhausen-Denker, G., Bober, E., Tannich, E., & Arnold, H. H. (1989). A novel human muscle factor related to but distinct from MyoD1 induces myogenic conversion in 10T1/2 fibroblasts. *EMBO J, 8*(3), 701-709.

Brennan, T. J., Chakraborty, T., & Olson, E. N. (1991). Mutagenesis of the myogenin basic region identifies an ancient protein motif critical for activation of myogenesis. *Proc Natl Acad Sci U S A, 88*(13), 5675-5679.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi:10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol, 109*, 21 29 21-29. doi:10.1002/0471142727.mb2129s109

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature, 523*(7561), 486-490. doi:10.1038/nature14590

Cao, Y. (2012). MyoD - Revised Chromatin IP Protocol in obscene detail for beginners. *Tapscott Lab correspondence*.

Cao, Y., Kumar, R. M., Penn, B. H., Berkes, C. A., Kooperberg, C., Boyer, L. A., Young, R. A., & Tapscott, S. J. (2006). Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters. *EMBO J, 25*(3), 502-511. doi:10.1038/sj.emboj.7600958

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., MacQuarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., & Tapscott, S. J. (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell, 18*(4), 662-674. doi:10.1016/j.devcel.2010.02.014

Carter, M. G., Sharov, A. A., VanBuren, V., Dudekula, D. B., Carmack, C. E., Nelson, C., & Ko, M. S. (2005). Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biol, 6*(7), R61. doi:10.1186/gb-2005-6-7-r61

Castro, D. S., Martynoga, B., Parras, C., Ramesh, V., Pacary, E., Johnston, C., Drechsel, D., Lebel-Potter, M., Garcia, L. G., Hunt, C., Dolle, D., Bithell, A., Ettwiller, L., Buckley, N., & Guillemot, F. (2011). A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. *Genes Dev, 25*(9), 930-945. doi:10.1101/gad.627811

Castro, D. S., Skowronska-Krawczyk, D., Armant, O., Donaldson, I. J., Parras, C., Hunt, C., Critchley, J. A., Nguyen, L., Gossler, A., Gottgens, B., Matter, J. M., & Guillemot, F. (2006). Proneural bHLH and Brn proteins coregulate a neurogenic program through cooperative binding to a conserved DNA motif. *Dev Cell, 11*(6), 831-844. doi:10.1016/j.devcel.2006.10.006

Chen, J. C., Ramachandran, R., & Goldhamer, D. J. (2002). Essential and redundant functions of the MyoD distal regulatory region revealed by targeted mutagenesis. *Dev Biol, 245*(1), 213-223. doi:10.1006/dbio.2002.0638

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., & Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell, 133*(6), 1106-1117. doi:10.1016/j.cell.2008.04.043

Chien, C. T., Hsiao, C. D., Jan, L. Y., & Jan, Y. N. (1996). Neuronal type information encoded in the basic-helix-loop-helix domain of proneural genes. *Proc Natl Acad*

Sci U S A, 93(23), 13239-13244.

Church, G. M., Ephrussi, A., Gilbert, W., & Tonegawa, S. (1985). Cell-type-specific contacts to immunoglobulin enhancers in nuclei. *Nature, 313*(6005), 798-801.

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., & Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell, 9*(2), 279-289.

Clark, K. L., Halay, E. D., Lai, E., & Burley, S. K. (1993). Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature, 364*(6436), 412-420. doi:10.1038/364412a0

Colasante, G., Lignani, G., Rubio, A., Medrihan, L., Yekhlef, L., Sessa, A., Massimino, L., Giannelli, S. G., Sacchetti, S., Caiazzo, M., Leo, D., Alexopoulou, D., Dell'Anno, M. T., Ciabatti, E., Orlando, M., Studer, M., Dahl, A., Gainetdinov, R. R., Taverna, S., Benfenati, F., & Broccoli, V. (2015). Rapid Conversion of Fibroblasts into Functional Forebrain GABAergic Interneurons by Direct Genetic Reprogramming. *Cell Stem Cell, 17*(6), 719-734. doi:10.1016/j.stem.2015.09.002

Correa-Cerro, L. S., Piao, Y., Sharov, A. A., Nishiyama, A., Cadet, J. S., Yu, H., Sharova, L. V., Xin, L., Hoang, H. G., Thomas, M., Qian, Y., Dudekula, D. B., Meyers, E., Binder, B. Y., Mowrer, G., Bassey, U., Longo, D. L., Schlessinger, D., & Ko, M. S. (2011). Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Sci Rep, 1*, 167. doi:10.1038/srep00167

Crews, S. T. (1998). Control of cell lineage-specific development and transcription by bHLH-PAS proteins. *Genes Dev, 12*(5), 607-620.

Davis, R. L., & Weintraub, H. (1992). Acquisition of myogenic specificity by replacement of three amino acid residues from MyoD into E12. *Science, 256*(5059), 1027-1030.

Davis, R. L., Weintraub, H., & Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell, 51*(6), 987-1000.

Dyce, J., George, M., Goodall, H., & Fleming, T. P. (1987). Do trophectoderm and inner cell mass cells in the mouse blastocyst maintain discrete lineages? *Development, 100*(4), 685-698.

Ellenberger, T., Fass, D., Arnaud, M., & Harrison, S. C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev, 8*(8), 970-980.

Ephrussi, A., Church, G. M., Tonegawa, S., & Gilbert, W. (1985). B lineage--specific interactions of an immunoglobulin enhancer with cellular factors in vivo. *Science, 227*(4683), 134-140.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods, 9*(3), 215-216. doi:10.1038/nmeth.1906

Evans, M. J., & Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature, 292*(5819), 154-156.

Farah, M. H., Olson, J. M., Sucic, H. B., Hume, R. I., Tapscott, S. J., & Turner, D. L. (2000). Generation of neurons by transient expression of neural bHLH proteins in mammalian cells. *Development, 127*(4), 693-702.

Feng, R., Desbordes, S. C., Xie, H., Tillo, E. S., Pixley, F., Stanley, E. R., & Graf, T. (2008). PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proc Natl Acad Sci U S A, 105*(16), 6057-6062. doi:10.1073/pnas.0711961105

Ferre-D'Amare, A. R., Prendergast, G. C., Ziff, E. B., & Burley, S. K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature, 363*(6424), 38-45. doi:10.1038/363038a0

Fong, A. P., Yao, Z., Zhong, J. W., Cao, Y., Ruzzo, W. L., Gentleman, R. C., & Tapscott, S. J. (2012). Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev Cell, 22*(4), 721-735. doi:10.1016/j.devcel.2012.01.015

Fong, A. P., Yao, Z., Zhong, J. W., Johnson, N. M., Farr, G. H., 3rd, Maves, L., & Tapscott, S. J. (2015). Conversion of MyoD to a neurogenic factor: binding site specificity determines lineage. *Cell Rep, 10*(12), 1937-1946. doi:10.1016/j.celrep.2015.02.055

Gao, X., Chandra, T., Gratton, M. O., Quelo, I., Prud'homme, J., Stifani, S., & St-Arnaud, R. (2001). HES6 acts as a transcriptional repressor in myoblasts and can induce the myogenic differentiation program. *J Cell Biol, 154*(6), 1161-1171. doi:10.1083/jcb.200104058

Garner, M. M., & Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res, 9*(13), 3047-3060.

Geisler, S., & Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol, 14*(11), 699-712. doi:10.1038/nrm3679

Gerber, A. N., Klesert, T. R., Bergstrom, D. A., & Tapscott, S. J. (1997). Two domains of MyoD mediate transcriptional activation of genes in repressive chromatin: a mechanism for lineage determination in myogenesis. *Genes Dev, 11*(4), 436-450.

Gilmour, D. S., & Lis, J. T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Mol Cell Biol, 5*(8), 2009-2018.

Gossen, M., & Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc Natl Acad Sci U S A, 89*(12), 5547-5551.

Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J. R., & Zaret, K. S. (1996). Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes Dev, 10*(13), 1670-1682.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell, 130*(1), 77-88. doi:10.1016/j.cell.2007.05.042

Guillemot, F., Lo, L. C., Johnson, J. E., Auerbach, A., Anderson, D. J., & Joyner, A. L. (1993). Mammalian achaete-scute homolog 1 is required for the early development of olfactory and autonomic neurons. *Cell, 75*(3), 463-476.

Guillemot, F., Nagy, A., Auerbach, A., Rossant, J., & Joyner, A. L. (1994). Essential role of Mash-2 in extraembryonic development. *Nature, 371*(6495), 333-336. doi:10.1038/371333a0

Gurdon, J. B. (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *J Embryol Exp Morphol, 10*, 622-640.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell, 38*(4), 576-589. doi:10.1016/j.molcel.2010.05.004

Henke, R. M., Meredith, D. M., Borromeo, M. D., Savage, T. K., & Johnson, J. E. (2009). Ascl1 and Neurog2 form novel complexes and regulate Delta-like3 (Dll3) expression in the neural tube. *Dev Biol, 328*(2), 529-540. doi:10.1016/j.ydbio.2009.01.007

Henthorn, P., Kiledjian, M., & Kadesch, T. (1990). Two distinct transcription factors that bind the immunoglobulin enhancer microE5/kappa 2 motif. *Science, 247*(4941), 467-470.

Hori, K., Cholewa-Waclaw, J., Nakada, Y., Glasgow, S. M., Masui, T., Henke, R. M.,

Wildner, H., Martarelli, B., Beres, T. M., Epstein, J. A., Magnuson, M. A., Macdonald, R. J., Birchmeier, C., & Johnson, J. E. (2008). A nonclassical bHLH Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling. *Genes Dev, 22*(2), 166-178. doi:10.1101/gad.1628008

Ieda, M., Fu, J. D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B. G., & Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell, 142*(3), 375-386. doi:10.1016/j.cell.2010.07.002

Imayoshi, I., Isomura, A., Harima, Y., Kawaguchi, K., Kori, H., Miyachi, H., Fujiwara, T., Ishidate, F., & Kageyama, R. (2013). Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. *Science, 342*(6163), 1203-1208. doi:10.1126/science.1242366

Jacob, J., Kong, J., Moore, S., Milton, C., Sasai, N., Gonzalez-Quevedo, R., Terriente, J., Imayoshi, I., Kageyama, R., Wilkinson, D. G., Novitch, B. G., & Briscoe, J. (2013). Retinoid acid specifies neuronal identity through graded expression of Ascl1. *Curr Biol, 23*(5), 412-418. doi:10.1016/j.cub.2013.01.046

Johnson, J. E., Birren, S. J., & Anderson, D. J. (1990). Two rat homologues of Drosophila achaete-scute specifically expressed in neuronal precursors. *Nature, 346*(6287), 858-861. doi:10.1038/346858a0

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., & Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell, 152*(1-2), 327-339. doi:10.1016/j.cell.2012.12.009

Juang, R. a. (1986). A tutorial on Hidden Markov Models. *IEEE ASSP*(Jan), 0740-0767.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res, 39*(Database issue), D712-717. doi:10.1093/nar/gkq1156

Kamburov, A., Wierling, C., Lehrach, H., & Herwig, R. (2009). ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res, 37*(Database issue), D623-628. doi:10.1093/nar/gkn698

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics, 26*(17), 2204-2207. doi:10.1093/bioinformatics/btq351

Keshet, I., Lieman-Hurwitz, J., & Cedar, H. (1986). DNA methylation affects the formation of active chromatin. *Cell, 44*(4), 535-543.

Kidder, B. L., Palmer, S., & Knott, J. G. (2009). SWI/SNF-Brg1 regulates self-renewal and occupies core pluripotency-related genes in embryonic stem cells. *Stem Cells, 27*(2), 317-328. doi:10.1634/stemcells.2008-0710

Koutalianos, D., Koutsoulidou, A., Mastroyiannopoulos, N. P., Furling, D., & Phylactou, L. A. (2015). MyoD transcription factor induces myogenesis by inhibiting Twist-1 through miR-206. *J Cell Sci, 128*(19), 3631-3645. doi:10.1242/jcs.172288

Koyano-Nakagawa, N., Wettstein, D., & Kintner, C. (1999). Activation of Xenopus genes required for lateral inhibition and neuronal differentiation during primary neurogenesis. *Mol Cell Neurosci, 14*(4-5), 327-339. doi:10.1006/mcne.1999.0783

Kwon, J. W., Kwon, H. K., Shin, H. J., Choi, Y. M., Anwar, M. A., & Choi, S. (2015). Activating transcription factor 3 represses inflammatory responses by binding to the p65 subunit of NF-kappaB. *Sci Rep, 5*, 14470. doi:10.1038/srep14470

Langlands, K., Yin, X., Anand, G., & Prochownik, E. V. (1997). Differential interactions of Id proteins with basic-helix-loop-helix transcription factors. *J Biol Chem, 272*(32), 19785-19793.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Lassar, A. B., Davis, R. L., Wright, W. E., Kadesch, T., Murre, C., Voronova, A., Baltimore, D., & Weintraub, H. (1991). Functional activity of myogenic HLH proteins requires hetero-oligomerization with E12/E47-like proteins in vivo. *Cell, 66*(2), 305-315.

Lassar, A. B., Paterson, B. M., & Weintraub, H. (1986). Transfection of a DNA locus that mediates the conversion of 10T1/2 fibroblasts to myoblasts. *Cell, 47*(5), 649-656.

Leung, J. Y., & Nevins, J. R. (2012). E2F6 associates with BRG1 in transcriptional regulation. *PLoS One, 7*(10), e47967. doi:10.1371/journal.pone.0047967

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA

methyltransferase gene results in embryonic lethality. *Cell, 69*(6), 915-926.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Lin, Y. C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C. A., Dutkowski, J., Ideker, T., Glass, C. K., & Murre, C. (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol, 11*(7), 635-643. doi:10.1038/ni.1891

Little, C. D., Nau, M. M., Carney, D. N., Gazdar, A. F., & Minna, J. D. (1983). Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *Nature, 306*(5939), 194-196.

Liu, X., Chen, X., Zhong, B., Wang, A., Wang, X., Chu, F., Nurieva, R. I., Yan, X., Chen, P., van der Flier, L. G., Nakatsukasa, H., Neelapu, S. S., Chen, W., Clevers, H., Tian, Q., Qi, H., Wei, L., & Dong, C. (2014). Transcription factor achaete-scute homologue 2 initiates follicular T-helper-cell development. *Nature, 507*(7493), 513-518. doi:10.1038/nature12910

Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G., & Kang, J. (2008). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Res, 18*(12), 1177-1189. doi:10.1038/cr.2008.309

Longo, A., Guanga, G. P., & Rose, R. B. (2008). Crystal structure of E47-NeuroD1/beta2 bHLH domain-DNA complex: heterodimer selectivity and DNA recognition. *Biochemistry, 47*(1), 218-229. doi:10.1021/bi701527r

LR, R. (1989). A tutorial on Hidden Markov Models and Selected Applications in Selected Speech Recognition. *IEEE ASSP, 77*(February), 257-286.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature, 389*(6648), 251-260. doi:10.1038/38444

Luo, S., Lu, J. Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., Shao, W., Lu, Z., Li, H., Na, J., Tang, F., Wang, J., Zhang, Y. E., & Shen, X. (2016). Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell, 18*(5), 637-652. doi:10.1016/j.stem.2016.01.024

Ma, P. C., Rould, M. A., Weintraub, H., & Pabo, C. O. (1994). Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell, 77*(3), 451-459.

Mal, A., Sturniolo, M., Schiltz, R. L., Ghosh, M. K., & Harter, M. L. (2001). A role for histone deacetylase HDAC1 in modulating the transcriptional activity of MyoD: inhibition of the myogenic program. *EMBO J, 20*(7), 1739-1753. doi:10.1093/emboj/20.7.1739

Maleki, S. J., Royer, C. A., & Hurlburt, B. K. (1997). MyoD-E12 heterodimers and MyoD-MyoD homodimers are equally stable. *Biochemistry, 36*(22), 6762-6767. doi:10.1021/bi970262m

Maleki, S. J., Royer, C. A., & Hurlburt, B. K. (2002). Analysis of the DNA-binding properties of MyoD, myogenin, and E12 by fluorescence anisotropy. *Biochemistry, 41*(35), 10888-10894.

Malone, C. M., Domaschenz, R., Amagase, Y., Dunham, I., Murai, K., & Jones, P. H. (2011). Hes6 is required for actin cytoskeletal organization in differentiating C2C12 myoblasts. *Exp Cell Res, 317*(11), 1590-1602. doi:10.1016/j.yexcr.2011.03.023

Mao, Z., & Nadal-Ginard, B. (1996). Functional and physical interactions between mammalian achaete-scute homolog 1 and myocyte enhancer factor 2A. *J Biol Chem, 271*(24), 14371-14375.

Massari, M. E., Jennings, P. A., & Murre, C. (1996). The AD1 transactivation domain of E2A contains a highly conserved helix which is required for its activity in both Saccharomyces cerevisiae and mammalian cells. *Mol Cell Biol, 16*(1), 121-129.

Masui, S., Shimosato, D., Toyooka, Y., Yagi, R., Takahashi, K., & Niwa, H. (2005). An efficient system to establish multiple embryonic stem cell lines carrying an inducible expression unit. *Nucleic Acids Res, 33*(4), e43. doi:10.1093/nar/gni043

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., & Wasserman, W. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res, 44*(D1), D110-115. doi:10.1093/nar/gkv1176

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol, 28*(5), 495-501. doi:10.1038/nbt.1630

Meredith, A., & Johnson, J. E. (2000). Negative autoregulation of Mash1 expression in CNS development. *Dev Biol, 222*(2), 336-346. doi:10.1006/dbio.2000.9697

Meredith, D. M., Borromeo, M. D., Deering, T. G., Casey, B. H., Savage, T. K., Mayer, P. R., Hoang, C., Tung, K. C., Kumar, M., Shen, C., Swift, G. H., Macdonald, R. J., & Johnson, J. E. (2013). Program specificity for Ptf1a in pancreas versus neural tube development correlates with distinct collaborating cofactors and chromatin accessibility. *Mol Cell Biol, 33*(16), 3166-3179. doi:10.1128/MCB.00364-13

Merika, M., & Orkin, S. H. (1993). DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol, 13*(7), 3999-4010.

Minty, A., & Kedes, L. (1986). Upstream regions of the human cardiac actin gene that modulate its transcription in muscle cells: presence of an evolutionarily conserved repeated motif. *Mol Cell Biol, 6*(6), 2125-2136.

Mitsui, K., Shirakata, M., & Paterson, B. M. (1993). Phosphorylation inhibits the DNA-binding activity of MyoD homodimers but not MyoD-E12 heterodimers. *J Biol Chem, 268*(32), 24415-24420.

Munsterberg, A. E., & Lassar, A. B. (1995). Combinatorial signals from the neural tube, floor plate and notochord induce myogenic bHLH gene expression in the somite. *Development, 121*(3), 651-660.

Murre, C., Bain, G., van Dijk, M. A., Engel, I., Furnari, B. A., Massari, M. E., Matthews, J. R., Quong, M. W., Rivera, R. R., & Stuiver, M. H. (1994). Structure and function of helix-loop-helix proteins. *Biochim Biophys Acta, 1218*(2), 129-135.

Murre, C., McCaw, P. S., & Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell, 56*(5), 777-783.

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. V., Buskin, J. N., Hauschka, S. D., Lassar, A. B., & et al. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell, 58*(3), 537-544.

Murre, C., Voronova, A., & Baltimore, D. (1991). B-cell- and myocyte-specific E2-box-binding factors contain E12/E47-like subunits. *Mol Cell Biol, 11*(2), 1156-1160.

Musilova, K., & Mraz, M. (2015). MicroRNAs in B-cell lymphomas: how a complex biology gets more complex. *Leukemia, 29*(5), 1004-1017.

doi:10.1038/leu.2014.351

Nair, S. K., & Burley, S. K. (2003). X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell, 112*(2), 193-205.

Nakada, Y., Hunsaker, T. L., Henke, R. M., & Johnson, J. E. (2004). Distinct domains within Mash1 and Math1 are required for function in neuronal differentiation versus neuronal cell-type specification. *Development, 131*(6), 1319-1330. doi:10.1242/dev.01008

Nelson, B. R., Hartman, B. H., Ray, C. A., Hayashi, T., Bermingham-McDonogh, O., & Reh, T. A. (2009). Acheate-scute like 1 (Ascl1) is required for normal delta-like (Dll) gene expression and notch signaling during retinal development. *Dev Dyn, 238*(9), 2163-2178. doi:10.1002/dvdy.21848

Nie, Y., Liu, H., & Sun, X. (2013). The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines. *PLoS One, 8*(3), e60002. doi:10.1371/journal.pone.0060002

Nikolayeva, O., & Robinson, M. D. (2014). edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods Mol Biol, 1150*, 45-79. doi:10.1007/978-1-4939-0512-6_3

Nishiyama, A., Xin, L., Sharov, A. A., Thomas, M., Mowrer, G., Meyers, E., Piao, Y., Mehta, S., Yee, S., Nakatake, Y., Stagg, C., Sharova, L., Correa-Cerro, L. S., Bassey, U., Hoang, H., Kim, E., Tapnio, R., Qian, Y., Dudekula, D., Zalzman, M., Li, M., Falco, G., Yang, H. T., Lee, S. L., Monti, M., Stanghellini, I., Islam, M. N., Nagaraja, R., Goldberg, I., Wang, W., Longo, D. L., Schlessinger, D., & Ko, M. S. (2009). Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell, 5*(4), 420-433. doi:10.1016/j.stem.2009.07.012

Olson, L. E., Bedja, D., Alvey, S. J., Cardounel, A. J., Gabrielson, K. L., & Reeves, R. H. (2003). Protection from doxorubicin-induced cardiac toxicity in mice with a null allele of carbonyl reductase 1. *Cancer Res, 63*(20), 6602-6606.

Penn, B. H., Bergstrom, D. A., Dilworth, F. J., Bengal, E., & Tapscott, S. J. (2004). A MyoD-generated feed-forward circuit temporally patterns gene expression during skeletal muscle differentiation. *Genes Dev, 18*(19), 2348-2353. doi:10.1101/gad.1234304

Perini, G., Diolaiti, D., Porro, A., & Della Valle, G. (2005). In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation. *Proc Natl*

Acad Sci U S A, 102(34), 12117-12122. doi:10.1073/pnas.0409097102

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., & Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res, 42*(Database issue), D756-763. doi:10.1093/nar/gkt1114

Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., Nguyen, N., Paten, B., Zweig, A. S., Karolchik, D., & Kent, W. J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics, 30*(7), 1003-1005. doi:10.1093/bioinformatics/btt637

Rao, P. K., Kumar, R. M., Farkhondeh, M., Baskerville, S., & Lodish, H. F. (2006). Myogenic factors that regulate expression of muscle-specific microRNAs. *Proc Natl Acad Sci U S A, 103*(23), 8721-8726. doi:10.1073/pnas.0602831103

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., & Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res, 43*(Database issue), D670-681. doi:10.1093/nar/gku1177

Rossant, J., Guillemot, F., Tanaka, M., Latham, K., Gertenstein, M., & Nagy, A. (1998). Mash2 is expressed in oogenesis and preimplantation development but is not required for blastocyst formation. *Mech Dev, 73*(2), 183-191.

Rudnicki, M. A., Schnegelsberg, P. N., Stead, R. H., Braun, T., Arnold, H. H., & Jaenisch, R. (1993). MyoD or Myf-5 is required for the formation of skeletal muscle. *Cell, 75*(7), 1351-1359.

Saldanha, A. J. (2004). Java Treeview--extensible visualization of microarray data. *Bioinformatics, 20*(17), 3246-3248. doi:10.1093/bioinformatics/bth349

Sartorelli, V., Huang, J., Hamamori, Y., & Kedes, L. (1997). Molecular mechanisms of

myogenic coactivation by p300: direct interaction with the activation domain of MyoD and with the MADS box of MEF2C. *Mol Cell Biol, 17*(2), 1010-1026.

Schuijers, J., Junker, J. P., Mokry, M., Hatzis, P., Koo, B. K., Sasselli, V., van der Flier, L. G., Cuppen, E., van Oudenaarden, A., & Clevers, H. (2015). Ascl2 acts as an R-spondin/Wnt-responsive switch to control stemness in intestinal crypts. *Cell Stem Cell, 16*(2), 158-170. doi:10.1016/j.stem.2014.12.006

Sharov, A. A., Piao, Y., Matoba, R., Dudekula, D. B., Qian, Y., VanBuren, V., Falco, G., Martin, P. R., Stagg, C. A., Bassey, U. C., Wang, Y., Carter, M. G., Hamatani, T., Aiba, K., Akutsu, H., Sharova, L., Tanaka, T. S., Kimber, W. L., Yoshikawa, T., Jaradat, S. A., Pantano, S., Nagaraja, R., Boheler, K. R., Taub, D., Hodes, R. J., Longo, D. L., Schlessinger, D., Keller, J., Klotz, E., Kelsoe, G., Umezawa, A., Vescovi, A. L., Rossant, J., Kunath, T., Hogan, B. L., Curci, A., D'Urso, M., Kelso, J., Hide, W., & Ko, M. S. (2003). Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol, 1*(3), E74. doi:10.1371/journal.pbio.0000074

Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y., & Hakoshima, T. (1997). Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J, 16*(15), 4689-4697. doi:10.1093/emboj/16.15.4689

Simon, J. M., Giresi, P. G., Davis, I. J., & Lieb, J. D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc, 7*(2), 256-267. doi:10.1038/nprot.2011.444

Simpson, E. M., Linder, C. C., Sargent, E. E., Davisson, M. T., Mobraaten, L. E., & Sharp, J. J. (1997). Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nat Genet, 16*(1), 19-27. doi:10.1038/ng0597-19

Smith, A. G., & Hooper, M. L. (1987). Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. *Dev Biol, 121*(1), 1-9.

Song, J., Ugai, H., Nakata-Tsutsui, H., Kishikawa, S., Suzuki, E., Murata, T., & Yokoyama, K. K. (2003). Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves interactions with the same cis-elements. *Int J Mol Med, 11*(5), 547-553.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., Sheffield, N. C., Graf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N. D., Birney, E., Iyer, V. R., Crawford, G. E., Lieb, J. D., & Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory

elements that shape cell-type identity. *Genome Res, 21*(10), 1757-1767. doi:10.1101/gr.121541.111

Soufi, A., Donahue, G., & Zaret, K. S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell, 151*(5), 994-1004. doi:10.1016/j.cell.2012.09.045

Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., & Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell, 161*(3), 555-568. doi:10.1016/j.cell.2015.03.017

Soufi, A., & Zaret, K. S. (2013). Understanding impediments to cellular conversion to pluripotency by assessing the earliest events in ectopic transcription factor binding to the genome. *Cell Cycle, 12*(10), 1487-1491. doi:10.4161/cc.24663

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell, 126*(4), 663-676. doi:10.1016/j.cell.2006.07.024

Thayer, M. J., Tapscott, S. J., Davis, R. L., Wright, W. E., Lassar, A. B., & Weintraub, H. (1989). Positive autoregulation of the myogenic determination gene MyoD1. *Cell, 58*(2), 241-248.

Treutlein, B., Lee, Q. Y., Camp, J. G., Mall, M., Koh, W., Shariati, S. A., Sim, S., Neff, N. F., Skotheim, J. M., Wernig, M., & Quake, S. R. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature*. doi:10.1038/nature18323

van der Flier, L. G., van Gijn, M. E., Hatzis, P., Kujala, P., Haegebarth, A., Stange, D. E., Begthel, H., van den Born, M., Guryev, V., Oving, I., van Es, J. H., Barker, N., Peters, P. J., van de Wetering, M., & Clevers, H. (2009). Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell, 136*(5), 903-912. doi:10.1016/j.cell.2009.01.031

Vierbuchen, T., Ostermeier, A., Pang, Z. P., Kokubu, Y., Sudhof, T. C., & Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. *Nature, 463*(7284), 1035-1041. doi:10.1038/nature08797

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., Thurman, R. E., Johnson, A. K., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Giste, E., Haugen, E., Dunn, D., Wilken, M. S., Josefowicz, S., Samstein, R., Chang, K. H., Eichler, E. E., De Bruijn, M., Reh, T. A., Skoultchi, A., Rudensky, A., Orkin, S. H., Papayannopoulou, T., Treuting, P. M., Selleri, L., Kaul, R., Groudine, M.,

Bender, M. A., & Stamatoyannopoulos, J. A. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science, 346*(6212), 1007-1012. doi:10.1126/science.1246426

Villares, R., & Cabrera, C. V. (1987). The achaete-scute gene complex of D. melanogaster: conserved domains in a subset of genes required for neurogenesis and their homology to myc. *Cell, 50*(3), 415-424.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., & Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature, 457*(7231), 854-858. doi:10.1038/nature07730

Vojtek, A. B., Taylor, J., DeRuiter, S. L., Yu, J. Y., Figueroa, C., Kwok, R. P., & Turner, D. L. (2003). Akt regulates basic helix-loop-helix transcription factor-coactivator complex formation and activity during neuronal differentiation. *Mol Cell Biol, 23*(13), 4417-4427.

Waddington, C. H. (1942). The epigenotype. 1942. *Endeavor, 1*(1), 18–20. doi:10.1093/ije/dyr184

Waddington, C. H. (1952). *The epigenetics of birds*. Cambridge Eng.: University Press.

Waddington, C. H. (1957). *The strategy of the genes; a discussion of some aspects of theoretical biology*. London,: Allen & Unwin.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., & Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res, 22*(9), 1798-1812. doi:10.1101/gr.139105.112

Wang, Y., & Jaenisch, R. (1997). Myogenin can substitute for Myf5 in promoting myogenesis but less efficiently. *Development, 124*(13), 2507-2513.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., & Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet, 40*(7), 897-903. doi:10.1038/ng.154

Wapinski, O. L., Vierbuchen, T., Qu, K., Lee, Q. Y., Chanda, S., Fuentes, D. R., Giresi, P. G., Ng, Y. H., Marro, S., Neff, N. F., Drechsel, D., Martynoga, B., Castro, D. S., Webb, A. E., Sudhof, T. C., Brunet, A., Guillemot, F., Chang, H. Y., & Wernig, M. (2013). Hierarchical mechanisms for direct reprogramming of

fibroblasts to neurons. *Cell, 155*(3), 621-635. doi:10.1016/j.cell.2013.09.028

Weintraub, H. (1972). A possible role for histone in the synthesis of DNA. *Nature, 240*(5382), 449-453.

Weintraub, H., & Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. *Science, 193*(4256), 848-856.

Wickham, H. (2009). *Ggplot2 : elegant graphics for data analysis*. New York: Springer.

Wickham, H. (2016). *Ggplot2*. New York, NY: Springer Science+Business Media, LLC.

Wright, W. E., Sassoon, D. A., & Lin, V. K. (1989). Myogenin, a factor regulating myogenesis, has a domain homologous to MyoD. *Cell, 56*(4), 607-617.

Yan, K. S., & Kuo, C. J. (2015). Ascl2 reinforces intestinal stem cell identity. *Cell Stem Cell, 16*(2), 105-106. doi:10.1016/j.stem.2015.01.014

Yao, Z., Fong, A. P., Cao, Y., Ruzzo, W. L., Gentleman, R. C., & Tapscott, S. J. (2013). Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites. *Skelet Muscle, 3*(1), 8. doi:10.1186/2044-5040-3-8

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes Dev, 25*(21), 2227-2241. doi:10.1101/gad.176826.111

Zhao, L., Sun, M. A., Li, Z., Bai, X., Yu, M., Wang, M., Liang, L., Shao, X., Arnovitz, S., Wang, Q., He, C., Lu, X., Chen, J., & Xie, H. (2014). The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res, 24*(8), 1296-1307. doi:10.1101/gr.163147.113

Zingg, J. M., Pedraza-Alva, G., & Jost, J. P. (1994). MyoD1 promoter autoregulation is mediated by two proximal E-boxes. *Nucleic Acids Res, 22*(12), 2234-2241.