Methods for Comparative Analysis of Chromatin Accessibility and Gene Expression, With Applications to Cellular Reprogramming

by

Dinesh Manandhar

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: ______________________
Approved:

__________________________
Raluca Gordăn, Supervisor

__________________________
Gregory E. Crawford

__________________________
Andrew S. Allen

__________________________
Charles A. Gersbach

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics in the Graduate School of Duke University 2018
Abstract

Methods for Comparative Analysis of Chromatin Accessibility and Gene Expression, With Applications to Cellular Reprogramming

by

Dinesh Manandhar

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: ____________________

Approved:

________________________
Raluca Gordăan, Supervisor

________________________
Gregory E. Crawford

________________________
Andrew S. Allen

________________________
Charles A. Gersbach

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics in the Graduate School of Duke University 2018
Abstract

Cellular reprogramming processes remain poorly characterized at the level of genome-wide chromatin and gene expression changes. Specifically, the extent to which reprogrammed cells differ quantitatively from both the starting cells and the target cells is unknown for most reprogramming systems. In addition, direct comparisons between the genome-wide reprogramming efficiencies in systems driven by the overexpression of endogenous versus exogenous master regulator(s) are rarely performed. This thesis presents methods for comparative analyses of genome-wide gene expression and chromatin accessibility data, applied to myogenic reprogramming systems in order to assess reprogramming efficiency and generate testable hypotheses for improving the reprogramming process. First, gene expression and chromatin accessibility profiles of MyoD-induced transdifferentiated primary human skin fibroblasts are compared to fibroblasts and myoblasts. Second, similar genome-wide changes are assessed for myogenic conversion of iPS cells driven by overexpression of endogenous MyoD versus exogeneous MyoD. Both studies show that (i) while many muscle marker genes are reprogrammed after MyoD overexpression, the genome-wide accessibility and gene expression profiles are still different from those of primary myoblast or myotube cells; (ii) MyoD induces a continuum of changes in chromatin accessibility, with only a fraction of myogenic chromatin sites gaining a completely reprogrammed accessibility status; and (iii) chromatin-remodeling deficiencies are strongly correlated with incomplete gene expression reprogramming. Classification
analyses comparing reprogrammed and non-reprogrammed genes or chromatin sites revealed discriminatory genetic and epigenetic features, suggesting ways to potentially improve the reprogramming efficiency. Genomic analysis of transgene MyoD overexpression in iPS cells, compared to endogenous MyoD activation, also showed that MyoD is more “aggressive” in its chromatin opening behavior, showing a large number of off-target chromatin opening events. To further investigate the effects chromatin remodeling events on gene expression in reprogramming studies, a novel cross-cell type gene expression prediction framework (CPGex) is also developed. By integrating and modeling the non-linear combinatorial effects of chromatin accessibility as well as the expression levels of regulatory TFs, CPGex is able to weigh the importance of regulatory sites or factors for downstream targeted reprogramming of specific gene(s). The methods described in this thesis can be applied to any cellular reprogramming system in order to quantitatively assess the efficiency of reprogramming at the chromatin accessibility and gene expression levels, as well as to generate testable hypothesis for improved genome-wide reprogramming.
To my family.
Contents

Abstract iv
List of Tables xii
List of Figures xiii
List of Abbreviations and Symbols xvi
Acknowledgements xviii

1 Introduction 1

1.1 Cell fates are not determinate 2

1.1.1 A brief history 2

1.2 Reprogramming, transdifferentiation and de-differentiation 6

1.2.1 Significance of reprogramming and transdifferentiation studies 7

1.3 MyoD-induced myogenic reprogramming 9

1.3.1 Significance of MyoD-induced reprogramming studies 9

1.3.2 Normal myogenesis 10

1.3.3 MyoD and other myogenic regulatory factors 12

1.3.4 Structure of MyoD and other MRFs 14

1.3.5 In vivo DNA binding determinants of MyoD and MRFs 15

1.3.6 In vivo binding of MyoD and other factors 17

1.3.7 MyoD-induced gene regulation 18

1.4 Computational tools designed for improving cellular reprogramming 21
3.2.3 List of regulatory TFs and their corresponding target genes (TF-TG list) ............................................. 98
3.2.4 Training artificial neural network models ......................... 99
3.2.5 Availability ........................................... 100
3.2.6 Feature importance scores ................................ 100
3.3 Results ................................................. 102
3.3.1 Models that combine chromatin accessibility and regulatory TF expression levels can accurately predict the expression of target genes ............................................. 102
3.3.2 Our joint DHSs-plus-TFs models compare favorable to JEME 104
3.3.3 Feature importance measurements yield regulatory information 105
3.4 Discussion ............................................. 107

4 Comparative genome-wide myogenic differentiation analyses between CRISPR/Cas9-mediated endogenous and lentiviral transgene MyoD expression ............................................. 110
4.1 Background ............................................. 111
4.2 Methods ................................................. 112
4.2.1 Plasmid Construction .................................. 112
4.2.2 Cell Culture ............................................ 113
4.2.3 Viral Production and Transduction ......................... 114
4.2.4 Quantitative Reverse Transcription PCR ................. 115
4.2.5 Immunofluorescence Staining .......................... 116
4.2.7 Chromatin immunoprecipitation followed by Quantitative Real-Time PCR ......................... 117
4.2.8 DNase-seq experiments ................................ 118
4.2.9 DNase-seq processing .................................. 119
4.2.10 RNA-seq experiments ............................... 121
4.2.11 RNA-seq data processing ............................ 121
4.2.12 Clustering gene expression matrix .......................... 122
4.2.13 MyoD ChIP-seq data in Myotubes .......................... 123
4.2.14 DHS Pattern Enrichment Analysis .......................... 123
4.3 Results .......................................................... 124
   4.3.1 Exogenous and endogenous MyoD-mediated differentiation of human iPS cells generate muscle-like cells ............... 124
   4.3.2 Genome-wide transcriptional profiles in CR-MYOD and TG-MYOD match the muscle marker and fusion phenotypes observed 128
   4.3.3 Transgene MyoD is “aggressive” in chromatin remodeling behavior .................................................. 131
   4.3.4 cis-genic DHS accessibility profiles match the expression pattern of the differentially expressed genes .................. 136
   4.3.5 Gene regulatory network analysis shows CR-MYOD have fibroblast genes activated ............................... 138
4.4 Conclusions ....................................................... 139
5 Epigenetic poising and co-factor binding explain HDACi-induced chromatin opening at PU.1-bound sites 141
   5.1 Background ...................................................... 142
   5.2 Methods ........................................................ 144
      5.2.1 Classification of opened versus stable sites .............. 144
   5.3 Results and conclusionc ........................................ 145
6 Conclusions ......................................................... 149
A Proofs ............................................................ 155
   A.1 Proof that increasing MyoD concentration increases MyoD occupancy at lower affinity sites. ......................... 155
B Installing and running CPGex .................................... 157
Bibliography ......................................................... 161
Biography .......................................................... 180
List of Tables

2.1 Gene expression (TPM) values for some muscle and SAND domain genes ............................................. 60

B.1 CPGex arguments and their descriptions ............................................. 159
List of Figures

1.1 Examples of cellular reprogramming ................................. 4
1.2 Waddington epigenetic landscape .................................... 7
1.3 MyoD heterodimerizes with E-protein to bind to DNA. ............ 15
1.4 Artificial Neural Network architecture ............................... 36
2.1 Induced expression of MyoD in fibroblast cells leads to incomplete genome-wide changes in gene expression and chromatin accessibility. . 56
2.2 Myogenic markers show increased levels in MyoD-induced fibroblast cells compared to uninduced fibroblast cells. ...................... 58
2.3 DNase-seq data around myogenic genes for fibroblasts, fibroblasts expressing MyoD (fibro-MyoD), and primary myoblasts. ............. 59
2.4 In-silico mixing showing gene expression and DNase-seq profiles of cell populations that contain fibroblast and myoblast cells in different ratios. 61
2.5 ChromHMM model output on DNase-signal across the normal fibroblasts, transdifferentiated fibroblasts, and muscle cell lines. ............. 64
2.6 MyoD induction results in continuum of chromatin reprogramming genome-wide. .................................................. 67
2.7 The degree of chromatin reprogramming (or CRL score) is significantly higher for Fibro-MyoD compared to Fibro-control cells. .......... 68
2.8 Pie-chart showing genomic sites bound by MyoD in fibro-MyoD cells, grouped by the binarized DHS patterns corresponding to the sites. . 70
2.9 Comparisons between MyoD binding profiles in fibro-MyoD and myoblasts. ......................................................... 71
2.10 Diagrammatic representation of TF-DNA binding specificity and histone modification signal used as features for classification analyses. . 72
3.3 Our DHSs-plus-TFs models compare favorably to JEME (the best model to date) ......................................................... 105
3.4 Feature Importance analysis for TNNC2 expression .......................................................... 106
4.1 Endogenous MyoD expression shows better fusion but lower level of myogenic gene activation ........................................... 127
4.2 PCA and distance matrix plot for RNA-seq samples across the cell types .......................................................... 129
4.3 Comparative gene expression analyses between CR-MYOD and TG-MYOD samples ...................................................... 130
4.4 Transgene MyoD reprograms sites better but is also “aggressive” in opening chromatin .............................................. 132
4.5 An example of a DHS site more reprogrammed by transgene than endogenous MyoD ...................................................... 134
4.6 Transgene MyoD opens chromatin sites involved in myogenic activities more in number and degree compared to CRISPR/Cas9-based endogenous MyoD activation ...................................................... 135
4.7 DHS-pattern enrichment analysis for gene-clusters with genes differentially expressed between cell types .................. 137
4.8 Tissue/Cell type classification of samples used in MyoD-mediated iPSC reprogramming study obtained with CellNet .......... 138
5.1 Chromatin features distinguish HDACi-opened from stable DHS sites bound by PU.1 ...................................................... 147
5.2 H3K4me1 enrichment and CBX2/8 binding at PU.1-bound DHS sites 148
List of Abbreviations and Symbols

Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Design matrix</td>
</tr>
<tr>
<td>y</td>
<td>Response vector</td>
</tr>
<tr>
<td>m</td>
<td>Total number of examples (or samples) in data</td>
</tr>
<tr>
<td>n</td>
<td>Total number of features in X</td>
</tr>
<tr>
<td>m′</td>
<td>Total number of examples in the training set</td>
</tr>
</tbody>
</table>

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>MRF</td>
<td>Myogenic Regulatory Factor</td>
</tr>
<tr>
<td>DHS</td>
<td>DNaseI hyper-sensitive (site)</td>
</tr>
<tr>
<td>CRL</td>
<td>Chromatin Reprogramming Level (in accessibility)</td>
</tr>
<tr>
<td>GRL</td>
<td>Gene Reprogramming Level (in expression)</td>
</tr>
<tr>
<td>Fibro-control</td>
<td>Primary human skin fibroblasts transduced with lentiviral human MYOD1 gene, but without doxycycline-mediated induction of the transgene</td>
</tr>
<tr>
<td>Fibro-MyoD</td>
<td>Primary human skin fibroblasts reprogrammed using transduction of lentiviral human MYOD1 gene</td>
</tr>
<tr>
<td>iPSC</td>
<td>induced Pluripotent Stem Cell</td>
</tr>
<tr>
<td>PTCI</td>
<td>Pre-treated control iPSC</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>CR-MYOD</td>
<td>PTCI cells reprogrammed using CRISPR/Cas9-mediated endogenous MYOD expression</td>
</tr>
<tr>
<td>TG-MYOD</td>
<td>PTCI cells reprogrammed using transduction of lentiviral MYOD1 transgene</td>
</tr>
<tr>
<td>MB</td>
<td>Myoblasts</td>
</tr>
<tr>
<td>MT</td>
<td>Myotubes</td>
</tr>
<tr>
<td>CPGex</td>
<td>Cross cell-type Prediction of Gene Expression</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads per kilobase length per million reads</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments per kilobase length per million reads</td>
</tr>
<tr>
<td>TPM</td>
<td>Transcripts per million</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>EN</td>
<td>Elastic Net</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
</tbody>
</table>
Acknowledgements

There are many to thank. Special thanks to my advisor, Raluca Gordăn, for her great mentorship and support over the years. I would also like to thank other members of my committee (Gregory Crawford, Charles Gersbach and Andrew Allen) for their guidance and support. Another acknowledgement to my advisor and Greg for helping me with my presentation practices. Special thanks also to Charles Gersbach, Greg Crawford and some members of their laboratories - especially, Ami Kabadi (Gersbach Lab), Lingyun Song (Crawford Lab), Jennifer B. Kwon (Gersbach Lab), and Christopher Frank (Crawford Lab)) for performing the wet-lab experiments and providing me with the data for the projects in the thesis. I would also like to Barbara Engelhardt, Sayan Mukherjee and Florian Wagner for their feedback on the first project, and past and current members of the Gordăn Laboratory for their feedback on all of my projects. Additionally, I would like to thank the Duke CBB (Computational Biology and Bioinformatics) Program for the support. Duke CBB, and Duke Graduate School have covered part of my conference travel expenses, and Gordăn Lab has covered the rest over the years, for which I am also grateful. In addition to the projects in the thesis, I have also worked on another project with Ian McDowell, Barbara Engelhardt, Tim Reddy, and Ami Schmidt and I would like to thank them for helping with that project.
This chapter provides necessary background on cellular reprogramming and transdifferentiation, a central theme throughout the thesis. Cellular reprogramming processes bring genome-wide epigenomic and gene expression changes directing modifications of cell fates. As such, much of the genome-wide computational analyses and predictive models discussed in this thesis, although generally applicable, are motivated by their applications to cellular reprogramming. Therefore, I begin with some history and background on the subject, and then cover myogenic reprogramming study as a model system. After discussing the knowledge and gaps in the field of myogenic reprogramming, I introduce some computational approaches and tools aimed at improving and assessing reprogramming efficiency. I close with a discussion on genomic technologies and machine learning methods that I have used either in my analyses or in order to develop the gene expression prediction framework described in Chapter 3.
1.1 Cell fates are not determinate

1.1.1 A brief history

Embryonic stem cells (ESCs) derived from inner mast cell population in blastocysts are known to be pluripotent, which means that they can derive cell lines from all germ lines - ectoderm, mesoderm and endoderm. The ecto-, meso- and endodermal stem cells themselves are multipotent, and produce definite number of tissues and cell types in response to the developmental cues. During this process of lineage determination, these cells divide to more defined precursor cells and differentiated cellular stages. For instance, ectodermal lineage eventually give rise to neurons and external linings of the body; tissues like muscles, chondrocytes, fibroblasts and adipocytes are generated from mesodermal lineage; while, endoderm gives rise to the intestinal linings, and exocrine and endocrine cells in the pancreas, and epithelial linings of lung.

It was long believed this process of cell differentiation and lineage determination is irrevocably set, and as such differentiated cells cannot change their cell fate (Nicholas and Kriegstein, 2010). However, much like other fields of science that saw revolutionary ideas last century, a number of experiments throughout the last six decades set the stage that contested this line of definitiveness and motivated developmental biologists to study lineage plasticity (Graf, 2011). Thus, before describing in detail the systems and methods used in this thesis, it is worth discussing some history and highlighting major developments that paved the way to our current understanding of cell fate plasticity. (For further details, please refer to (Graf, 2011; Vierbuchen and Wernig, 2011; Xu et al., 2015).)

Essentially, the following cellular conversion successes are to note.

1. **Somatic cell nuclear transfer (SCNT) and animal cloning experiments:**

   Briggs and King developed the SCNT technique in 1952, and showed that blastocyst cell nuclei transplanted into enucleated eggs of Rana pipens generated
swimming tadpoles (Briggs and King, 1952). The authors reported very limited success with gastrula stage nuclei however, and attributed this to progressive specialization of nuclear function during cell differentiation (King and Briggs, 1955). A few years later, Gurdon et al. used the nuclear transplantation technique to generate adult frogs using nuclei obtained from intestinal linings of Xenopus laevis (Gurdon et al., 1958; Gurdon and Uehlinger, 1966). This was the first successful animal cloning experiment performed. These experiments showed early on that the oocyte cytoplasm has factors necessary to reprogram differentiated nuclei of adult vertebrates, the details of which are still not well understood. Animal cloning in mammals did not succeed until 1997, when Dolly, the sheep, was cloned by transplantsing mammary epithelial nucleus into an enucleated oocyte (Wilmut et al., 1997).

2. Transdetermination experiments:

During insect metamorphosis, larval to pupal transition involves the primitive disc-like larval structures called imaginal discs evert and grow into appendages, such as wings, antennae or legs. Between 1962 and 1968, Hadorn et al. showed that if cells from these discs are transplanted to ectopic sites in the larvae, their fate can be transdetermined (Graf, 2011; Hadorn, 1968). This meant that imaginal disc cells otherwise destined to be antenna could be made part of legs or vice-versa. These studies raised the possibility that differentiated cells are pliable in their cell fates.

3. Heterokaryons:

In 1983, Helen Blau fused human amniocytes and mouse muscle cells. The resulting multinucleated cells or heterokaryons had human muscle-specific genes turned on after a 24-hour fusion (Blau et al., 1983). These observations were
repeated by replacing amniocytes with human keratinocytes and hepatocytes (Blau et al., 1985). These studies showed transactivation of silenced human muscle genes in adult cells is possible.

4. Fibroblasts converted to muscles – a story of serendipity:

In the 1970s, Peter Jones was studying the effect of chemotherapeutic drugs on human cells. He noticed that cultured fibroblasts exposed to 5-azacytidine (AzaC) switched their phenotype to largely muscles, and to some extent adipocytes and chondrocytes (Graf, 2011; Taylor and Jones, 1979). Soon after, he and his colleague identified that AzaC achieves this by inhibiting DNA methylation (Jones and Taylor, 1980). The choice of the 10T1/2 fibroblast cell line turned out to be fortuitous because it was later discovered that primary fibroblasts have this gene repressed by a mechanism different than repressive DNA methylation (Tapscott, 2005).

It took about another decade before Weintraub and his colleagues discovered the target gene (MYOD1) responsible for the AzaC-induced myogenic conversion. They used cDNA hybridization subtraction scheme to narrow down
a list of candidate factors, which were later tested individually (Davis et al., 1987). This was the discovery of the first “master regulator transcription factor capable of switching cell fate in non-dividing cells. Since then, a number of studies on cellular conversion and transcription factors directing these cell fate switches have increased, notably in the last few years (Graf, 2011; Xu et al., 2015), some of which are shown in Figure 1.1. For instance, in 2004, Xie and colleagues showed that B-cells can be converted to macrophages by over-expressing macrophage marker protein, C/EBP alpha (Xie et al., 2004), and Kajimura et al., in 2008, showed that skin fibroblasts can be converted to brown fat cells by overexpressing C/EBPbeta together with PRDM16 (Seale et al., 2008). Notably, among these direct lineage conversion studies, conversion across germ lines have also been shown to be possible. While previously, cell types were shown to convert across germ lines with limited efficiency (Weintraub et al., 1989) or under unclear delineations of cell line of origin (Comai and Tajbakhsh, 2014), recently, it was shown that fibroblasts could also be directly converted to neurons by expression of three neuronal TFs (Vierbuchen et al., 2010).

5. Reprogramming to pluripotency:

One of the most seminal reprogramming successes is the 2006 study by Takahashi and Yamanaka study showing that a combination of just four transcription factors (Oct4, Klf4, Sox2, Myc) can de-differentiate adult mouse fibroblast cells to ESC-like cells that they named “induced Pluripotent Stem Cells (iPSCs)” (Takahashi and Yamanaka, 2006). The generation of iPSCs not only showed that complete de-differentiation to pluripotency was achievable (albeit at limited success), but it also solved the ethical issues of using human embryonic stem cells for therapeutic and disease modeling purposes (discussed more
below). The discovery of Yamanaka factors has also helped the field of transdifferentiation (or direct cellular reprogramming, see below), as the interest in identifying similar master TF regulators capable of molding cellular fate has also peaked since 2006 (Graf, 2011).

Interestingly, soon after iPSC discovery, it was also shown that similar to SCNT, fusion of somatic cell with a pluripotent stem cell confers pluripotency properties to the fusion cell (Do and Scholer, 2006; Hall et al., 2006). The fusion cells are however tetraploid and are not clinically relevant as iPSCs or transdifferentiated cells (Nicholas and Kriegstein, 2010).

1.2 Reprogramming, transdifferentiation and de-differentiation

Over the years, reprogramming has developed a different meaning from the two other terms that also represent cellular conversion transdifferentiation and de-differentiation (Jopling et al., 2011). Reprogramming cells refer to converting differentiated cells back to their pluripotency states, i.e. conversion to iPSC state. Transdifferentiation, on the other hand, refers to direct and forced cellular conversion between the differentiated adult cell types, as in the case of AzaC-induced muscle conversion. The third term de-differentiation refers to partial “hike up to some pre-cursor cells, instead of the full reprogramming.

These definitions can be visualized using the Waddington epigenetic landscape (wad, 1939), shown in Figure 1.2. In this landscape, ESC cell at the top of the hill is totipotent. As it rolls down the hill along one of the ridges, its cell fate gets more committed. Finally, the cell becomes one of the many differentiated adult cells at the bottom of the hill. Reprogramming of a differentiated cell represents a reversal to totipotency (path 1 in Figure 1.2), whereas, a conversion of a differentiated cell type to another over a ridge at the bottom of the hill represents the process of
transdifferentiation (path 2, Figure 1.2). De-differentiation on the other hand is incomplete reprogramming up the hill to a precursor cell (Jopling et al., 2011)

**Terminology usage in this thesis:** This thesis is primarily concerned with genomic and epigenomic study during lineage commitment to muscles. As such, for the sake of simplicity, reprogramming is generally used in its literal sense, i.e. re-programming of cell fate. Therefore, in context that the starting (e.g. iPSCs or primary human fibroblasts in Chapters 2 and 3) and target cell types are clear, reprogramming is also sometimes used to mean transdifferentiation or directed differentiations of iPSCs.

### 1.2.1 Significance of reprogramming and transdifferentiation studies

Here, the importance of studying reprogramming and TD studies is broadly discussed.

- **Academic interest.**

  As discussed earlier, since the successes with SCNTs, deep interest in identifying factors responsible have existed. This has led to the discovery of various new master regulators, and also highlighted the novel roles of factors identified.
Reprogramming and TD studies provide a tool to understand how master regulators like MyoD, Gata1, or Yamanaka factors dictate temporal tissue-specific gene regulation, chromatin modification changes and thereby cell fates in various cell types. Developmental biologists can also use these studies to probe lineage commitment and understand how such differentiation processes differ from those observed during normal development (Vierbuchen and Wernig, 2011). A better understanding of transcription factor mediated gene regulation, cell differentiation and lineage commitment is possible with TD and reprogramming studies.

- Therapeutic interest.

The reprogrammed iPSCs and transdifferentiated cells are excellent systems to model diseases in vitro, and test autologous therapeutic applications. As noted above, iPSCs create a way to develop patient-specific stem cells sideling the use of embryonic stem cells. On the other hand, transdifferentiating cell types directly to target cell type possesses advantages of higher cellular conversion efficiencies and shorter reprogramming routes, since cells do not have to undergo a time-consuming and potentially inefficient detour of being de-differentiated before they are directed to target cell fate as in the case of iPSC based cellular therapies (Nicholas and Kriegstein, 2010; Xu et al., 2015).

Regardless, a number of promising applications have been reported. In 2008, Zhou et al. showed that the transdifferentiated insulin-producing pancreatic cells could alleviate hyperglycemia caused by insulin deficiency in a type one diabetic mouse model (Zhou et al., 2008). Similarly, mouse fibroblasts converted to cardiomyocytes with a combination of Gata4, Mef2c and Tbx5, were shown to be functional when transplanted into heart of mice (Ieda et al., 2010). Other experiments showing fibroblasts to muscle conversions also exist and is
detailed in section 1.3 below. In addition to these disease modeling and ingrafting studies, other innovative applications are possible. Recently, it was shown that cyclic expression of Yamanaka factors in vivo could increase the lifespan of mouse models (Mahmoudi and Brunet, 2016).

1.3 MyoD-induced myogenic reprogramming

In this thesis, the reprogramming studies discussed in upcoming chapters pertain to MyoD-induced cellular conversion. TF MyoD is named for being a muscle-specific myogenic determination factor (Davis et al., 1987). MyoD can singularly orchestrate the reprogramming of adult fibroblasts, adipocytes, liver and nerve cells, to cells capable of fusing into myotubes (Weintraub et al., 1989). As such, MyoD-induced cellular reprogramming offers a simple platform to study how a master regulator can direct cellular transformations, compared to systems that require modulated expression of multiple TFs. Below, necessary background concerning MyoD and myogenic cellular differentiation studies is provided.

1.3.1 Significance of MyoD-induced reprogramming studies

Clearly, MyoD-induced myogenic reprogramming also enables a better understanding of how master regulators like MyoD and their effector proteins modulate changes in cell fate. Additionally, MyoD-induced reprogramming presents an opportunity to understand how cellular and cell-type specific chromatin dynamics affect regeneration efficiency. This is because MyoD reprograms some cell types, such as mouse embryonic fibroblasts (MEFs) efficiently, while others such as human HepG2, human HeLa, and mouse P19 are reprogrammed less efficiently or not at all (Albini et al., 2013; Comai and Tajbakhsh, 2014). A few such studies have identified that BAF60C, a BAF chromatin modifier subunit present in mesodermal lineages, together with MyoD can reprogram the refractory HeLa and ESC cells (Albini et al.,
Further studies, especially pertaining to genomic and epigenomic changes, as we do in Chapters 2 and 3, can improve our understanding of the limitations of master regulators in reconfirming cellular identities.

Myogenic reprogramming studies are also great model systems for disease modeling and autologous muscle regeneration. A number of studies have focused to harness the reprogramming ability of MyoD to treating muscle diseases. One such common muscular degeneration diseases is Duchenne muscular dystrophy (DMD), which is caused by the mutation of the dystrophin genes (Bushby et al., 1991; Hoffman et al., 1987; Koenig et al., 1988). En Kimura et al transdifferentiated primary mouse fibroblasts using lentiviral vectors expressing micro-dystrophin/enhanced green fluorescent protein (µDys/eGFP) fused with tamoxifen-inducible MyoD construct (Kimura et al., 2008). The transdifferentiated cells when transplanted back into the muscles of mdx<sup>4cv</sup> mice, a mouse model for DMD, enabled regeneration of myofibers expressing µDys. Others have performed similar experiments with MyoD-induced transdifferentiation of adipose-derived stem cells (Goudenege et al., 2009). As with adipose-derived stem cells, recently, there have been increased interest in generating satellite cells-like self-renewing progenitors to generate myoblast population for regenerative purposes (Takahashi et al., 2007; Yu et al., 2007). The myogenic reprogramming of iPS cells to generate Pax3+/7+ cells provides a feasible avenue for this (Comai and Tajbakhsh, 2014). However, further studies are needed to increase efficiency, and eliminate residual iPSCs (Blum and Benvenisty, 2008; Comai and Tajbakhsh, 2014; Rao et al., 2012).

1.3.2 Normal myogenesis

A complete understanding of the process and limitations of myogenic reprogramming requires an apprehension of myogenesis during normal development. Here, brief background on embryonic and post-natal myogenesis is provided. (Some of the key
myogenic regulatory factors mentioned in this section are further discussed below.)

Early embryonic development involves segmentation of paraxial mesoderm into metameric structures called somites. The somites show a distinct rostral-caudal polarity attributed to gradient concentration of retinoic acid (Duester, 2008), with rostral somites showing myogenic differentiation earlier than caudal somites. The caudal somites are maintained in early undifferentiated mesenchymal states by the morphogen signaling (pertaining to Wnt, BMP and Shh morphogens) originating from the surrounding tissue (Aulehla and Pourquie, 2010). In particular, Wnt1 and Wnt3 from dorsal neural tube; Wnt4, Wnt6, and Wnt7a from surface ectoderm; and Sonic hedgehog (Shh) released from floor plate of neural tube and the notochord provide positive myogenic activation for muscle precursor cells by inducing the expression of Myf5 or MyoD or both (Bentzinger et al., 2012). In contrast, bone morphogen protein (BMP) signaling from lateral-plate mesoderm and Notch signaling controlled by Wnt and Fgf signaling (Hofmann et al., 2004) prevent myogenic differentiation and promote expansion of precursor cells by delaying Myf5 and MyoD induction and fostering Pax3 expression (for BMP: (Pourquie et al., 1995), for NOTCH: (Schuster-Gossler et al., 2007; Vasyutina et al., 2007), (Bentzinger et al., 2012)). A very recent study (Cacchiarelli et al., 2017) has also shown that BMP signaling also marks a differentiation bottleneck towards efficient myogenic transdifferentiation, and inhibiting BMP pathway enhances MyoD-induced transdifferentiation (discussed below).

During embryonic myogenesis, the most dorsal portion of the somite develop to structures called dermomyotome that represent very early myogenic structure, which eventually give rise to body and trunk (but not head) muscles. Dermomyotome show expression of early myogenic regulators like Pax3 and Pax7, and low level of early myogenic regulatory factor (Myf5) discussed below (Goulding et al., 1991; Jostes et al., 1990; Kiefer and Hauschka, 2001). The two lips of the dermomyotome mature into myotome, containing committed myogenic cells expressing MyoD and Myf5
(Cinnamon et al., 2001; Ordahl et al., 2001; Sassoon et al., 1989). It is believed myotome formation involves BMP inhibition through Wnt and Shh-mediated increased levels of Noggin (Hirsinger et al., 1997; Marcelle et al., 1997; Reshef et al., 1998), which helps with localized MyoD upregulation (Bentzinger et al., 2012).

As the embryo develops, central dermomyotome cells intercalate to myotome, a fraction of which gives rise to satellite cells in post-natal skeletal muscles (Gros et al., 2005; Kassar-Duchossoy et al., 2005; Relaix et al., 2005; Schienda et al., 2006). Satellite cells are myogenic stem cells, that reside between the basal lamina and sarcolemma of the developed myofibers (Mauro, 1961), and maintain the normal turnover of terminal muscles as well as muscle regeneration during tissue damage (Hill et al., 2003). Eventually, the epaxial and hypaxial parts of the dermomyotome and myotome form dorsal and lateral trunk-and-limb muscles respectively (Parker et al., 2003). Limb muscles are generated from Pax3+ migratory cells that delaminate from ventrolateral lip of the dermomyotome at the level of limbs (Vasyutina and Birchmeier, 2006).

Myogenesis during peri- and post-natal development is less well understood (Alonso-Martin et al., 2016). During both post-somitogenesis steps discussed above and post-natal development, muscle precursor cells and satellite cells give rise to single celled myoblasts that proliferate and subsequently fuse to contractile, multinucleated myotubes (Bentzinger et al., 2012; Blais, 2015).

1.3.3 MyoD and other myogenic regulatory factors

MyoD is one of the four Muscle Regulatory Factors (MRFs), a subfamily of bHLH (basic helix-loop-helix) family of TFs. The three other MRFs, namely Myf5, Myog and Myf6 (also known as MRF4), were soon discovered after the discovery of MyoD, and were also shown to possess reprogramming ability (for Myf5: (Braun et al., 1989); for Mrf4: (Braun et al., 1990; Miner and Wold, 1990)) and, to a certain extent, to
Myog (Edmondson and Olson, 1989). All four MRFs are also known to be important in myogenic determination and/or differentiation processes. Among these, Myf5 and MyoD, in a redundant fashion, commit cells to the myogenic program, Myog (short for Myogenin) is crucial for the differentiation of these committed muscle cells in vivo, and Myf6 has important roles in both muscle specification and differentiation (Comai and Tajbakhsh, 2014; Kassar-Douchesoy et al., 2005; Tapscott, 2005).

In addition to MRFs, there are important upstream myogenic regulatory factors important for proper spatio-temporal regulation of myogenesis (in concert with MRFs). Some of these include:

- **Six1 and Six4.**

  Six1 and Six4 are homeobox TFs that form complexes with eyes absent (Eya) proteins and stimulate transcription on critical myogenic target genes such as PAX3, MYOD, MRF4 AND MYOG (Grifone et al., 2005). As such Six proteins are known to be important upstream regulators of myogenesis. However, recent studies suggest that Six proteins are also active parallel to the MRFs during terminal differentiation of myoblasts in vitro (Liu et al., 2010). Indeed, one recent study shows that Six1 is critical for terminal myogenic differentiation of muscles (Santolini et al., 2016).

- **Pax3, Pax7 and Pitx2.**

  Pax3 and Pax7 are members of paired-homeobox family TFs, and are other pairs of critical upstream myogenic regulators. During embryonic myogenesis, Pax3 is expressed in early precursor cells that migrate from myotome (see above) to form limbs, and Pax3 null mice are either devoid of or have impaired limbs (Bentzinger et al., 2012). However, Pax3 is absent in craniofacial myogenesis, and its absence is rescued by another paired-related homeodomain
TF Pitx2 (Sambasivan et al., 2011). The third protein, Pax7 represses terminal differentiation, and is important to maintain stemness in satellite cells since their formation in mid-embryogenesis through to adulthood (Relaix et al., 2006). Genome-wide ChIP-seq analysis of Pax3 and Pax7 binding in primary adult mouse myoblasts indicates Pax7 binds to much larger number of loci than Pax3 (52000 versus 4600) possibly because of higher affinity of Pax7 for homeodomain binding sequence motif TAAT (Soleimani et al., 2012).

1.3.4 Structure of MyoD and other MRFs

All MRFs, being bHLH TF members, bind to DNA using their basic region, which confers MRFs their myogenic specificity in part (discussed below). The MRFs dimerize with their binding partners, typically E-proteins (Tcf3, Tcf4 and Tcf12) using the HLH domains (Lassar et al., 1991a; Murre et al., 1989). Besides the bHLH structures, MyoD and Myf5 contain an amino terminal acidic-activation domain (Tapscott, 2005). Towards the amino-end, MyoD and Myf5 have the histidine- and cysteine-rich domain (H/C) towards the amino-end. This H/C domain contains a tryptophan residue needed for MyoD to interact with Pbx/Meis complex, which has been shown to recruit MyoD in closed chromatin near myogenic genes such as MYOG (Fong and Tapscott, 2013; Fong et al., 2012; Knoepfler et al., 1999a; Sagerstrom, 2004). MyoD also possesses a third helix structure (closer to the carboxyl terminal tail) necessary for recruiting chromatin modifiers (Tapscott, 2005).

In particular, MyoD uses this third helix structure to bind to p300, a histone acetyl-transferase (HAT) protein, which acetylates histones, setting up the chromatin for downstream TFs to bind (Tapscott, 2005). HAT p300 in turn also recruits another HAT, the p300/CBP-associated factor (PCAF). PCAF acetylates MyoD at lysine residues proximal to its DNA binding domain, which is important for full transcriptional ability of MyoD (Dilworth et al., 2004). In addition to the HAT
recruitment, MyoD also recruits SWI/SNF chromatin remodelers (Simone et al., 2004). Hence, the structure of MyoD explains its ability to be recruited by Pbx1 at closed chromatin and subsequent ensure modification of chromatin for downstream effector TFs. Because MyoD is able to open chromatin and ensue downstream transcriptional program, it is also known as a pioneer factor (Zaret and Carroll, 2011). Hence, MyoD provides a platform to study details of chromatin remodeling changes during myogenic differentiation, a topic discussed in Chapters 2 and 3. Figure 1.3 shows these domains of MyoD, while in-bound with one of its dimerization partner, E-protein (Tapscott, 2005).

1.3.5  *In vivo DNA binding determinants of MyoD and MRFs*

MyoD typically binds with its heterodimer partner, E-protein and binds to the canonical E-box motifs (CAAnnTG) in the genome. Given the presence of large number of E-boxes and the ability of other bHLH factors (like NeuroD2 involved in neuronal differentiation) (Fong et al., 2012), what determines the myogenic specificity of MyoD
and MRFs \textit{in vivo}? The following factors in particular are attributable:

1. Myogenic code.

The three residues at positions 114, 115 and 124 (namely, Alanine, Threonine and Lysine) in the basic region is conserved across all MRFs (Tapscott, 2005), and serves as the myogenic code. It has been associated with DNA binding specificity and serves as the recognition code for binding at myogenic regulatory E-boxes (CAGGTG or CAGCTG sites) and thereby results in activation of muscle-specific genes. In fact, the replacement of the basic region in E-proteins by the myogenic code establishes myogenic activity in them (Fong and Tapscott, 2013).

2. Sequence specificity.

Difference in sequence specificities between MyoD and other bHLH proteins partly explain their in vivo binding differences. While MyoD and NeuroD2, neurogenic differentiation regulator, both bind to a shared CAGCTG E-box, MyoD has its own “specific” E-box sites (CAGGTG), whereas a neurogenic determination factor, NeuroD2, prefers to bind to CAGATG E-box sites (Fong et al., 2012). Fong and colleague also report that presence of two E-boxes, as opposed to just one, enables stable binding of MyoD-E-protein complex (Fong et al., 2012; Tapscott, 2005). Additionally, E-box flanking sequences can affect the binding preference of MyoD (Blackwell and Weintraub, 1990).

3. Determination based on binding of other factors.

Besides interaction with heterodimer partner, E-proteins, and recruitment placeholder protein, Pbx1, MyoD is known to interact with other TFs such as Mef2, Jdp2, Meis, c-Jun and Runx1 (Bengal et al., 1992; Knoepfler et al., 1999a; Ostrovsky et al., 2002; Wang et al., 2005). These factors have been
shown to be regulatory in myogenesis, with their absence marking reduced RNA pol II recruitment at MyoD enhancers (Blum and Dynlacht, 2013).

4. Epigenetic/chromatin context.

It is plausible that cell types refractory to myogenic reprogramming have their E-boxes relevant for myogenic differentiation “hidden” in heterochromatin (Tapscott, 2005). The details of how MyoD binding is epigenetically regulated remain largely unknown. One hypothesis is that factors such as Polycomb group (PcG) proteins, that trimethylates H3K27 histone, control developmental fates through repression of specific genes (Fong and Tapscott, 2013).

1.3.6 In vivo binding of MyoD and other factors

Tapscott and colleagues have studied the MyoD binding using chromatin precipitation followed by sequencing (ChIP-seq, see below) in C2C12 myoblasts and differentiated myotubes (Cao et al., 2010). In this study, the authors quantified > 30,000 MyoD bound sites in C2C12 myoblasts and differentiated myotubes, with MyoD occupancy changing in only a fraction (around 5k) of regions during between the cell types. Sites where MyoD binding increases are closer to muscle-related and other genes upregulating in terminal differentiation, while those that show decreased binding are closer to cell proliferation and communication related genes (Biggin, 2010).

While a large number of MyoD binding is thermodynamically plausible given the protein concentration, comparative analysis between normal MEFs to MyoD-induced transdifferentiated MEFs showed that a large fraction of MyoD-bound sites show increased H4 acetylation (Cao et al., 2010), suggesting MyoD at least induces epigenomic reprogramming and potentially sets up for transcriptional activity, through chromatin looping for example, in some fraction of those sites. Other studies looking into enhancer-specific epigenomic modifications (i.e. H3K4me1, H3K27ac and
HAT p300 binding) along with MyoD binding in C2C12 myoblasts and myotubes have identified thousands of MyoD-dependent enhancer sites (Asp et al., 2011; Blum et al., 2012; Vethantham et al., 2012).

Tapscott and colleague also showed that the overexpressed MyoD in MEFs largely binds to the same genomic sites as endogenous MyoD in primary and C2C12 myotubes (Yao et al., 2013). The authors report a roughly 70% concordance in binding for the top 35k peaks. Additionally, both endogenous and overexpressed MyoD have similar E-box preference. These findings explain why MyoD is able to induce cellular reprogramming in MEFs.

Interestingly, but not surprisingly, similar ChIP-seq studies have identified that proper binding of MyoD co-factors is also important for effective terminal myogenesis. While MyoD binding is similar between normal muscle cells and cells in Rhabdomyosarcoma (RD), a pediatric skeletal muscle tumor, differences at sites enriched for MyoD co-factor motifs (pertaining to RUNX1, MEF2C, JDP2 and NFIC) is observed (MacQuarrie et al., 2013). The authors showed that expression of these factors can rescue myogenesis in RD cells. All these studies add to the growing collection of ChIP-seq experiments done to understand myogenic differentiation (Blais, 2015). However, not many studies have looked into the nature of MyoD binding during myogenic transdifferentiation in primary human cells. MyoD binding analysis in Chapter 2 is one such study.

1.3.7 MyoD-induced gene regulation

Studies aimed at understanding transcriptomic dynamic during myogenesis date back to the early 2000. Two early microarray studies explored the gene expression dynamic during C2C12 myoblast to myotube differentiation (Moran et al., 2002; Shen et al., 2003), while another looked into the time-course expression profiles of genes during MyoD-induced myogenesis in mouse embryo fibroblasts (MEFs) (Bergstrom
et al., 2002). These experiments, while not completely genome-wide, identified large number of muscle, signaling and cell cycle related genes showing dynamic profile during the process, and showed that large clusters of genes coordinate the complex myogenic differentiation process (Blais, 2015).

A working regulatory model from these transcriptome studies is that during myogenesis, MyoD increases its own expression through a self-feed-forward loop, and the expression of downstream genes in a feed-forward mechanism, where the proteins from early myogenic genes are used to regulate the expression of terminal genes (Tapscott, 2005). For instance, MyoD helps in expressing Mef2d and p38 kinase; Mef2d, when activated by p38 kinase, works with MyoD in expressing a late myogenic gene, MYH3. MyoD, thereby, activates genes in a temporal fashion. Specifically, time-course studies have shown that the expression of genes affecting cell migration, adhesion/matrix is followed by activation of a set of TFs and then by myogenic proteins (Bergstrom et al., 2002; Tapscott, 2005). Additionally, the expression of myogenic as well as non-myogenic genes takes place at various kinetics (Bergstrom et al., 2002).

More recently RNA-seq profiling (see below) has given additional insights into the myogenesis process. Tapscott et al. also performed time-course transcriptome profiling of the C2C12 myoblast differentiation process using RNA-seq protocol (Trapnell et al., 2010a). The authors identified a large number (3,724) of previously unannotated transcripts, and a few hundred genes that show splice isoforms. An example of such genes, MEF2D, gives rise to two isoforms, one of which escapes encoding for amino acids involved in binding with corepressors, and is able to target terminal myogenic genes (Sebastian et al., 2013). RNA-seq has also been used to establish that non-coding enhancer RNAs are transcribed at thousands of MyoD-bound sites, including those at two well-known MyoD enhancers - core enhancer region (CE) and distal regulatory region (DRR), and have functional role in myogenesis (Mousavi et al., 2013). Knockdown experiments show that CE-eRNA is essential for Pol II
occupancy at MYOD1 gene, while DRR-eRNA activated MYOG expression, highlight their regulatory roles in myogenesis (Mousavi et al., 2013). Besides eRNAs, miRNAs and long-noncoding RNAs have also been shown to be important for myogenic regulation (Ballarino et al., 2015; Sohi and Dilworth, 2015).

Most recent transcriptome studies are using the single cell based sequencing technologies, which provides a more granular, single-cell based perspective in reprogramming dynamics. Trapnell and colleagues have used single-cell next generation sequencing technologies to study the myogenic differentiation processes during normal myoblast differentiation and forced conversion with MyoD (Cacchiarelli et al., 2017). In , the authors applied pseudo-temporal ordering of single cell RNA-seq samples during terminal differentiation of myoblasts to myotubes, and identified switch-like patterns between Id1, which recruits E-proteins to limit transcriptional activity of MyoD in proliferative myoblasts, and Myog, that is expressed during differentiation of myoblasts.

In a latest (yet non-peer reviewed) study (Cacchiarelli et al., 2017), the authors studied the MyoD-induced transdifferentiation of fibroblasts using single-cell RNA-seq (scRNA) technique. The authors identified that despite an alignment of core differentiation trajectory with normal myoblast differentiation, two key branch nodes exist in the middle and late differentiation trajectories that mark as reprogramming nodal points dictating the fate of a large number of reprogramming cells. Through differential branch-point analyses, the authors hypothesized and later showed that inhibiting BMP signaling and upregulating insulin signaling in the transdifferentiation media could increase the efficiency by fivefold. Additionally, incubating transdifferentiating cells with LSD1 inhibitor, to inhibit its histone demethylase activity, can also increase the reprogramming efficiency in the presence of insulin.
1.4 Computational tools designed for improving cellular reprogramming

In this section, some novel computational tools developed to study and improve reprogramming efficiencies are discussed.

1. Mogrify

Mogrify is a prediction framework developed by Rackham et al. to identify one or more TFs that can potentially reprogram starting cell types to target cell types via transdifferentiation (Rackham et al., 2016). It uses differentially expressed TFs and genes between the starting and target samples, and identifies smallest number of TFs that most maximizes the regulatory coverage in the target samples. Some important details include weighing TFs that covers a larger number differentially expressed genes more, and measuring the influence of a TF in the protein-protein interaction network based on the distance from the TF and out-degree of parent nodes. Mogrify has recapitulated important subsets of TFs used in previous transdifferentiation studies, and is a powerful starting tool for novel reprogramming studies or improving transdifferentiation efficiencies. Regarding skin fibroblast to skeletal muscle cell conversion in particular, the TF shortlist generated by Mogrify are Myog, MyoD, Pitx3, Six2, Hoxa7, Sox8, Foxd1 and Junb, many of which are known to have myogenic regulatory activity.

2. CellNet

Patrick Cahan and colleagues developed a network biology platform, CellNet, that assesses the efficiency of cellular reprogramming based on the fidelity with target tissue-specific gene regulatory network (GRN) (Cahan et al., 2014; Radley et al., 2017). For each tissue, its tissue-specific GRN is generated by
(i) pooling multitude of transcriptome perturbation assays (such as with gene knockout, different media treatment etc.) performed in other studies into a single data set, and (ii) measuring context-specific correlation between any TF and target gene (TG) pair to infer their direct regulatory relationship. CellNet also generates a ranked list of TFs that can potentially improve the reprogramming efficiency based on their measure to influence target genes in the network. The authors of the paper have used CellNet to show that among the reprogramming studies analyzed, reprogramming to pluripotency restores the target GRN profile better than both differentiations of iPSCs and adult tissues to target cell types, with the last showing least efficiency in capturing target GRN profile.

In addition to Mogrify and CellNet, there are a number of other computational resources developed to yield a ranked list of TFs for cellular reprogramming or assess the similarity of reprogrammed cells to target cell type; see (Bian and Cahan, 2016) for a full list. For instance, PluriTest uses a curated database of transcriptome profile across hundreds of human ESC and human iPSC lines as well as non-pluripotent cell types to yield a pluripotency probability score with high degree of sensitivity and specificity (see below) (Muller et al., 2011).

1.5 Limitations in knowledge in myogenic reprogramming

1. Analysis of genome-wide reprogramming efficiency based on chromatin and transcriptome-level changes is lacking.

"Non-permissive nuclear landscape has been hypothesized to be driving cellular resistance towards efficient reprogramming (Comai and Tajbakhsh, 2014; Fong and Tapscott, 2013)." As such, importance of studies on genome-wide chromatin remodeling changes has been highlighted recently (Blais, 2015; Comai
Additionally, most of previous myogenic differentiation studies posit that reprogramming is “efficient based on phenotypic expression of just a few key muscle marker genes such as DESMIN and Myosin heavy chain (MYH3). It is possible that such cells are only partially transdifferentiated (Comai and Tajbakhsh, 2014).

Therefore, a quantitative and integrative genome-wide chromatin- and transcriptome-level reprogramming analysis is lacking in a majority of myogenic reprogramming studies. In particular, (i) the efficiency of MyoD in remodeling chromatin genome-wide by itself or in combination with additional MRF (such as Myf5), chromatin modifiers (such as BAF60C), or reprogramming inducer (such as BMP4 inhibitor) needs to be studied; (ii) association of chromatin remodeling events or deficiencies with transcriptional changes or lack thereof; and (iii) comparative genome-wide myogenic conversion studies between reprogramming efficiency of iPS cells against that of fibroblasts need further examination. Chapters 2 and 3 in this thesis are steps in this direction. To our understanding, our study in Chapter 2 is the first study looking into genome-wide chromatin dynamics during MyoD-induced myogenic conversion. Recently, another study has also looked into the chromatin accessibility dynamic during terminal differentiation of myoblasts using single-cell analysis (Pliner et al., 2017). However, such single-cell chromatin analysis studying MyoD-induced cellular conversion is missing.

2. Inefficiency in reprogramming observed in primary human cells is not well understood.

Many of the previous studies have used cell line, as opposed to primary cells. Cell-line maintenance in culture could result in changes, such as epithelial-
to-mesenchymal transition (EMT), not observed in primary cells (Fong and Tapscott, 2013). In line with this, human primary cells are known to be harder to reprogram than their mouse counterparts (Lattanzi et al., 1998; Vogel, 2008; Weintraub et al., 1989). Not surprisingly, only a small number of transdifferentiation studies with genome-wide analyses in primary human fibroblasts have been reported (Blais, 2015; Cacchiarelli et al., 2017; Kabadi et al., 2015; MacQuarrie et al., 2013), besides our studies in Chapter 2 and 3. While these studies have shown that reprogramming efficiency can be increased using engineered fusion MyoD protein or inhibiting BMP4 repressive signaling, further studies are important to better understand bottlenecks observed in reprogramming primary human cells compared to non-human and non-primary counterparts.

3. Genome-wide changes in lentiviral MyoD-induced reprogramming is not compared against endogenous upregulation of MyoD.

This has been done in the context of MyoD binding, but further studies are needed to understand the differences in chromatin and gene expression between the two systems. Here, chapter 3 aims to do so by comparing systems with lentiviral MyoD over-expression against CRISPR-induced upregulation of endogenous MyoD.

4. Computational models that tie chromatin remodeling events with gene regulation do not exist in the context of cellular reprogramming.

As discussed above, models developed so far use transcriptome network-based framework. However, it is well known that chromatin remodeling deficiencies could be a major determinant of reprogramming inefficiencies, as we also show in chapter 2. Models that integrate chromatin reprogramming profile along with the gene expression data could be a powerful tool for integrative regulatory
1.6 Tools and technologies used

1.6.1 DNase-seq

Deoxyribonuclease I (or DNase I) is an endonuclease that non-specifically cuts the DNA at chromatin unbound by nucleosome or TFs. Regions of DNase I hypersensitive (DHS) sites in the genome have been shown to represent regulatory elements such as promoters, enhancers, silencers, insulators and locus control regions (Gaszner and Felsenfeld, 2006; Gross and Garrard, 1988; Li et al., 1999; Song and Crawford, 2010a; Thurman et al., 2012). The DNase I digestion followed by sequencing (DNase-seq) is used to globally map open chromatin regions in the genome. The basic steps of DNase-seq involve isolation and DNase I digestion of nuclei, adding linker to bind the DNA fragment to DynaI beads at one end, MmeI digestion to shorten the DNA fragment for sequencing, and adding another linker and PCR amplification of the resulting sequence (Song and Crawford, 2010a).

DNase-seq reads are mapped onto the genome using alignment tools like bowtie2 (Langmead and Salzberg, 2012), and further processed to identify regions of DNase hypersensitivity, called DHS peaks. For the latter, peak-calling softwares such as MACS2 (Zhang et al., 2008) are used. For conditions with multiple DNase-seq replicates, Irreproducible Discovery Rate (IDR) algorithm (Li et al., 2011) can be used to identify DHS peaks that are confidently called (or reproducible) in replicate experiments.

1.6.2 ChIP-seq

Chromatin Immuno-precipitation followed by sequencing (ChIP-seq) is used to immune-precipitate and later sequence DNA fragments bound by DNA binding proteins or histone modification marks using protein-specific antibody. The basic steps involve
formaldehyde cross-linking of the protein-DNA complex, nuclear isolation and sonication to shear chromatin to a target size of 100-300bp (Iyer et al., 2001; Ren et al., 2000), immunoprecipitation with the antibody to enrich DNA fragments bound to the protein, reverse cross-linking followed by library preparation and sequencing (Landt et al., 2012a). Whenever possible, comparison of ChIP-seq sample(s) are done against control sample(s) prepared similarly using a non-specific antibody or input chromatin.

ChIP-seq reads are mapped onto the genome using genome alignment softwares (such as bowtie2 (Langmead and Salzberg, 2012)), and genomic regions with enrichment of sequencing reads compared to the control sample (or the local background distribution of reads if control sample is not available) are identified as ChIP-seq “peaks by peak-calling softwares such as MACS2 (Zhang et al., 2008), as with DNase-seq data. Similarly, IDR can be used to identify reproducible set of ChIP-seq peaks across replicate samples (Li et al., 2011).

1.6.3 RNA-seq

RNA-sequencing technology yields more sensitive quantification of mRNA content in the cells than microarray based technologies (Hou et al., 2015). The basic process of mRNA library preparation involves: i) selecting polyA-tailed mRNA using oligo-beads with polyT-oligos, ii) shearing and reverse transcribing the pulled mRNA sequences to double stranded DNAs, iii) adapter tagging and PCR amplifying the reverse transcribed sequences. The sequences are mapped back to the genome using splicing-aware softwares like Tophat (Trapnell et al., 2009) or HISAT (Kim et al., 2015). The reads mapping to a gene are then quantified in fragments per kilobase per million (FPKM) or transcripts per million (TPM) measures.
1.6.4 PBM

Protein Binding Microarrays (PBMs) are high-throughput in vitro assays to quantify DNA binding specificities of DNA binding proteins (Berger and Bulyk, 2009). Universal PBM (or uPBM) can contain multiple copies of all possible 8-mers in the array and therefore can yield quantitative specificity of the protein for each 8-mers. This measure is quantified as enrichment or E-score, which measures the preference of the protein in binding 8-mer containing foreground sequences against those that do not (i.e. the background sequences). The E-scores range from -0.5 for complete non-specificity to +0.5 for highly specific binding. In the subsequent chapters using DNA binding analyses of TFs, I have used the data from the UniPROBE PBM database (Hume et al., 2015).

1.7 Classification and regression models used in this thesis

In my classification or regression analyses in later chapters, I have used one or more of the three models - Random Forest, Elastic Net, and Artificial Neural Network. These are explained in detail below. The reason any one of these modeling framework was used is because they satisfied a majority or all of the following desirable properties:

- Have high prediction performance. Model performances are assessed separately for classifier and regressor models. In the case of classification, first, true positives ($TP$), true negatives ($TN$), false positives ($FP$) and false negatives ($FN$) are identified from the test data. Then, the following performance measures
are computed.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]  \hspace{1cm} (1.1)

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

In the case of univariate binary classification, the area under the receiving operating curve (AUROC) is also calculated. For this, true positive rate, TPR = \(\frac{TP}{TP + FP}\) or sensitivity is computed for all possible values of false positive rate, FPR = \(\frac{FP}{FP + TN}\) or 1 – specificity, and the area under the resulting curve is computed. AUROC values of 0.5 corresponds to a random classifier, and that of 1 corresponds to a perfect classifier.

In the case of regression, the model performance is assessed as the root mean squared error or the pearson correlation coefficient (PCC) between real and predicted response values.

- **Select predictive features and assess their effect sizes.** One of the main goals in the classification and regression analyses in the thesis is to identify features that are most important in the model.

- **Handle correlated features.** Features such as TF binding specificities, and DNase hypersensitivity can be correlated inputs for a model. A classification or regression model that can highlight or ‘group’ important, potentially correlated features is preferred.

- **Is robust to the training data used.** Both the performance as well as the effect size (or importance scores) of the selected features should be stable (i.e. have small variance) under various training sets generated by uniform subsampling.
of the total input data.

Before the models are described, first, the input data is discussed below. Let,

\[ D = \{(x_1, y_1), \ldots, (x_m, y_m)\} \]

be the total collection of \( m \) independent and identically distributed (i.i.d.) samples (or examples) used in model training and testing, where \( x_s = [x_{s,1}, \ldots, x_{s,n}]^T \), \( 1 \leq s \leq m \) is the input vector with a feature measure \( x_{s,p}, 1 \leq p \leq n \) corresponding to feature \( p \), and \( y_s \) is the response class (‘+1’ or ‘−1’ for classification) or continuous real value (for regression) in sample \( s \). Let’s also define \( X = (x_1| \ldots |x_n) \) and \( y = (y_1, \ldots, y_m)^T \) to be the design matrix and the response vector respectively.

1.7.1 Random Forest

Random Forest (RF), developed by Leo Breiman in 2001 (Breiman, 2001), is an ensemble-based learning algorithm that uses binary decision trees as individual learners. To make each decision tree, a subset of training examples are selected at random with replacement (a process also known as “bootstrap-aggregating” or “bagging”); training data that are not used in training a decision tree are called “out-of-bag” (or OOB) examples. Similarly, each decision tree model is made with only a random subset of features (discussed below). This ensures that the prediction correlation between the models is minimal (because the individual learners are “less alike” from each other) (Genuer et al., 2010), which in turn helps reduce the generalization or test error.

Making decision trees

During the construction of the binary decision tree, each node splits examples into two groups to maximize the change in impurity (or cost) measure resulting from
the split. Specifically, for a node $b$ and its left and right children nodes, $b_l$ and $b_r$ respectively, the change in impurity is computed as:

$$\Delta \text{impurity} = \text{impurity}(b) - \left( \frac{|b_l|}{|b|} \text{impurity}(b_l) + \frac{|b_r|}{|b|} \text{impurity}(b_r) \right)$$

(1.2)

All possible feature values for the set of features randomly selected for the tree are considered for the split. The feature and its value that maximizes this $\Delta \text{impurity}$ is used to split the node. The impurity measure of a node in equation 1.2 can be one of the following depending on the regression or classification setting. For regression, typically the mean squared error function $\sum_{i \in b} (y_i - \bar{y})^2$ is considered, where $\bar{y}$ is the mean response value for the elements in node $b$. For classification, the impurity measure at node $b$ is typically computed using the “Gini Index (GI)” score:

$$GI(b) = \sum_{c=1}^{C} \pi_c (1 - \pi_c),$$

(1.3)

where $\pi_c$ is the probability of an example belonging to class $c$, i.e.

$$\pi_c = \sum_{i=1}^{|b|} \mathbb{I}(\text{class of } i = c).$$

Hence, gini index is a measure of expected error rate in classification. There are other impurity metrics like misclassification error and entropy that could also be used (Murphy, 2012).

**Variable selection**

RF can output relative importance of features in the model by two means.

- Feature Permutation.
For any particular feature, its values in the OOB samples are permuted. Its importance score is then computed as the average change in prediction error for the OOB sample across all the trees.

- **Gini-based Importance Score.**

For any particular feature, its aggregated gini index based “splitting prowess” across all nodes in the tree is also its measure of importance. Specifically, for any feature \( p \), its Gini importance score is computed as:

\[
\text{Importance}(p) = \frac{1}{T} \sum_{t=1}^{T} \sum_{b \in \omega_t} \mathbb{I}(p_b = p) \left[ \frac{|b|}{|\omega_t|} \times \Delta \text{impurity} \right],
\]

where \( \omega_t, 1 \leq t \leq T \) is a tree, \( p_b \) represents the value of the feature \( p \) at node \( b \), and impurity is measured using the Gini Index metric in equation 1.2. It has been reported that the Gini Index based importance measure is not suitable for categorical features (Strobl et al., 2007). However, this is not an issue for the analyses in this thesis as all the features used are continuous variables.

**Hyperparameters**

For all RF models, unless otherwise stated, the hyperparameters pertaining to the number of features to be selected at random (\( m_{\text{try}} \)) and the number of trees to build for the forest (\( \text{ntrees} \)) is set to default, i.e. 10 and 500 respectively.

1.7.2 **Elastic Net**

EN is a regularized linear regression method developed by Zou and Hastie in 2005 (Zou and Hastie, 2005). A linear regression function is used to predict the response vector \( \mathbf{y} \) by

\[
\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1 \hat{\beta}_1 + \cdots + \mathbf{x}_m \hat{\beta}_m
\]

(1.4)
For simplicity, we can center the response and standardize the feature vectors such that their mean is 0 and variance is 1 (i.e. \( \frac{1}{m} \sum_{s=1}^{m} y_s = \bar{y} = 0 \), \( \frac{1}{m} \sum_{s=1}^{m} x_{s,p} = \bar{x}_p = 0 \) and \( \sum_{s=1}^{m} x_{s,p}^2 = 1, p = 1, \ldots, n \)).

After the data normalization, let \( m' \) be the number of training examples in training set (Note: \( m \) is the number of examples in the full input data \( D \) above). The model fitting process then produces the coefficient vector \( \hat{\beta} = (\beta_1, \ldots, \beta_{m'}) \), which describes the effect size of each variable in calculating the response, and can be used to predict the response for a new test example \( x^* \). The EN solves the following optimization problem:

\[
\hat{\beta}_{en} = \arg\min_{(\beta_0, \beta)} \left\{ \sum_{s=1}^{m'} (y_s - \beta_0 - x_s^T \beta)^2 + \lambda J(\beta) \right\},
\]

where the elastic net penalty penalizes over the \( l_1 \) and \( l_2 \) norms of the coefficient vector \( \beta \) as follows:

\[
J(\beta) = \lambda_1 |\beta|_{l_1} + \lambda_2 |\beta|_{l_2}^2 = (1 - \alpha)|\beta|_{l_1} + \alpha|\beta|_{l_2}^2, \quad \text{for } \alpha \in [0, 1]
\]

Since \( |\beta|_{l_1} = \sum_{j=1}^{p} |\beta_j| \) and \( |\beta^2|_{l_2} = \sum_{j=1}^{p} \beta_j^2 \), the elastic net penalty can be written as:

\[
J(\beta) = \sum_{j=1}^{p} \left[ (1 - \alpha)|\beta_j| + \alpha \beta_j^2 \right].
\]

Hence, EN optimizes the function:

\[
\hat{\beta}_{en} = \arg\min_{(\beta_0, \beta)} \left\{ \sum_{s=1}^{m'} (y_s - \beta_0 - x_s^T \beta)^2 + \lambda \left( \sum_{j=1}^{p} \left[ (1 - \alpha)|\beta_j| + \alpha \beta_j^2 \right] \right) \right\}.
\]

Since \( J(\beta) \) in equation 1.6 is a convex combination of the lasso (with \( \alpha = 0 \)) and ridge (with \( \alpha = 1 \)) penalties, EN represents a compromise between the two regression
methods – lasso and ridge regression. Note that equation (1.7) has two hyperparameters (λ and α) that will be selected using cross-validation (CV) analysis discussed below.

**Shrinkage and Feature Selection in EN**

The lasso penalty (α = 0) tends to shrink coefficients to 0, but is indifferent to correlated variables and picks one among the highly correlated variables (or features) and ignores the rest. Since (for the analyses in the thesis) we want to select all the correlated (and discriminatory) features for feature interpretation, we do not search for \( \hat{\beta} \) coefficients at exactly α = 0. On the other extreme, the ridge penalty (α = 1) shrinks the correlated features towards each other, but does not tend to zero out the coefficients. EN, being a compromise of the two, performs much like lasso – generating sparser models as α is decreased to be close to 0, but it also shrinks correlated features towards each other like ridge regression (Zou and Hastie, 2005).

Zou and Hastie also show that the distance between two correlated covariates \( i \) and \( j \) in their \( \lambda \)-versus-\( \beta \) coefficient path is bounded as follows:

\[
d_{\lambda_1, \lambda_2}(i, j) = \frac{|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|}{|y|_1} \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)}
\]

where \( \rho = x_i^T x_j \) is the correlation between covariates \( i \) and \( j \) (Zou and Hastie, 2005). This shows the desired “grouping effect” of EN in that the coefficients of correlated variables are going to be close to each other. (This can be observed for correlated features in Figure 2.15 in Chapter 2.)

**Model training and hyperparameter estimation**

We first randomly select 75% of the input set as the training set and the rest is used as the test set to independently assess the performance of our model. In order to
generate the best trained model, the vector $\beta$ in equation (1.7) needs to be optimized over the values of the regularization coefficient ($\lambda$) and the lasso-ridge tuning parameter ($\alpha$), within the training set. For this, we performed a two-dimensional grid search over the space of $\lambda$ and $\alpha$. Because we are interested in integrated both properties of sparsity (induced by lasso penalty) and grouping effect (induced by the ridge penalty), we limit ourselves to the choices of $\alpha \in (0,1)$. The $(\lambda, \alpha)$ pair that minimized the mean prediction error in a 10-fold CV test was selected to train the final model using the full training set. (The 10-fold cross-validation is performed by training a model on the 9/10th of the training data, and testing on the remaining 1/10th, and repeating this 9 more times for all 10 separate cross-validation test sets. The average cross-validated prediction error measures the performance of the $(\lambda, \alpha)$ pair within the training set.)

The R package `cv.glmnet` was used for these model training and CV analyses. During the grid search, at each of 10 choices of $\alpha \in (0,1)$, we estimate the coefficient vector $\beta$ for various choices of $\lambda$ (or $\log(\lambda)$), and we generate coefficient paths, as shown in Figure 2.15 in Chapter 2. We adopt the `cv.glmnet` default option of selecting the $\beta$ coefficients at $+1$ standard deviation of the best $\lambda$ value identified, so that a sparser and more interpretable model can be generated without much increase in the error.

1.7.3 Feedforward Neural Network

Feedforward Artificial Neural Network (hereafter abbreviated as ANN) is a non-linear predictive model where feature variables are represented as input units in the “input layer”, the response variable(s) are represented as the “output layer” unit(s), and one or more “hidden layer” of units are used to capture the non-linear combinatorial effect on the response variable. At its core, is a stack of logistic regression models that are jointly yield the response vector (Murphy, 2012).
Model training

As discussed earlier, $m'$ and $n$ be the number of training samples, and the number of nodes (or features) in the input layer respectively. Also let $\mathbf{y}'$ represent the training response vector of dimension $m' \times 1$. Let $W_{l_1,l_2}$ be the initial weight matrix between layers $l_1$ (which can be the input layer) and the subsequent layer $l_2$ (which can also be the output layer). The model aims to minimize the following cost function:

$$ J = \sqrt{\frac{\|\mathbf{y}' - \hat{\mathbf{y}}\|^2}{m'}} + \lambda \|W_{l_1,l_2}\|^2_2 $$  \hspace{1cm} (1.8)

where $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{m'})$ is the predicted expression vector, $\|\cdot\|^2_2$ is the square of the Frobenius norm of the matrix, and $\lambda$ is the regularization parameter that needs to be tuned to minimize the generalization error. The response vector $\hat{\mathbf{y}}$ is obtained by non-linear weighted transformation of the input variables. If we assume there is just one hidden layer present, then $\hat{\mathbf{y}}$ is predicted as $\sigma(A_1W_{2,3})$, where $\sigma$ is a non-linear transformation (such as sigmoid or RELU operations), $W_{2,3}$ is the weight matrix corresponding the second and third layers (i.e. hidden and output layer), and $A_1 = \sigma(z(\cdot))$, where $z$ is the batch-normalization operation that normalizes a tensor by its mean and variance (as discussed in (Ioffe, 2015)).

Hyperparameter optimization

Figure 1.4 show the ANN architecture with just one hidden layer. However, we let our model have one or two hidden layers ($h_1$ and $h_2$). Hence, in addition to the regularization parameter $\lambda$ and the choice of non-linear operation (sigmoid versus RELU), the number of hidden layers as well as the number of units in the hidden layer make up the list of hyperparameters for the model. In order to limit the number of hyperparameters to optimize, we set the non-linear operation to RELU for hidden layers and sigmoid for the output layer based on their performances in preliminary
testing. We additionally implement a dropout of 70%, meaning that a random set of 70% of input and hidden layer nodes are used during stochastic gradient descent optimization of the weight parameters. Both dropout and regularization helps reduce the generalization error, and they are important given that the number of training samples is small.

For the hyperparameter optimization, we use the Tree-structured Parzen Estimators (TPE) implementation in hyperopt (Bergstra, 2015, 2013). TPE is an adaptive hyper-parameter search algorithm that iteratively models the likelihood of best performing subsets of hyper-parameters against the rest to sample a hyperparameter set that most maximizes this likelihood ratio. The choices of hyperparameters used for optimization are defined in Chapter 3, where ANN is used to model cross-cell type specific gene expression.
Incomplete MyoD-induced transdifferentiation is associated with chromatin remodeling deficiencies

In this chapter, MyoD-induced myogenic conversion of primary human fibroblasts is discussed. This is a collaborative study, and has been published recently under the same title as the title of this chapter in Nucleic Acids Research 2017, Vol. 45, No. 20 (doi: 10.1093/nar/gkx773; https://doi.org/10.1093/nar/gkx773) (Manandhar et al., 2017). Most of the contents in this chapter are also discussed in the publication. However, some additional analyses not mentioned in the publication are also shown.

The authors in this joint project include members from Gersbach laboratory (involved in designing and performing the reprogramming experiments), Crawford laboratory (involved in performing high throughput next-generation sequencing analyses), Gordän laboratory (involved in performing computational analyses), and Ehrlich laboratory (that provided the muscle samples for comparison). The contributions from individual members of these laboratories are highlighted in the Methods section below.
2.1 Background

In chapter 1, I have discussed the critical role of MyoD in normal muscle development as well as direct myogenic differentiation of adult cell types. Being the first reprogramming factor discovered, MyoD is one of the well-studied transcription factors, whereas myogenic differentiation remains one of the well-characterized lineage determination systems (Comai and Tajbakhsh, 2014; Cacchiarelli et al., 2017) making MyoD-mediated reprogramming studies a great model study for cellular transdifferentiation studies. However, despite decades of study, there remain major gaps in the knowledge of the factors involved in determining the efficiency of myogenic transdifferentiation systems, as discussed in Chapter 1, section 1.5.

In particular, before the publication of our study (Manandhar et al., 2017), gene expression profile of MyoD-induced transdifferentiation system had only been well-profiled in a microarray experiments (Bergstrom et al., 2002; Ishibashi et al., 2005) limiting their scope for a complete genome-wide study. One study that assayed genome-wide gene expression profile focused on comparing gene expression changes induced by wild-type versus engineered MyoD variants (Kabadi et al., 2015). To our knowledge, a direct genome-wide comparison of the gene expression profiles of MyoD-transdifferentiated cells against primary human muscle cells had not been performed – a comparison that is critical to characterize genome-wide efficiency of reprogramming in addition to the fact that human primary cells have been largely under-studied (see Chapter 1). Additionally, chromatin-based genome-wide analyses had not been performed. This is noteworthy because “nuclear/epigenomic landscape” has been purported to be primarily driving variations in MyoD-mediated reprogramming efficiencies across the cell types, as discussed in Chapter 1. Not surprisingly, comprehensive genome-wide analyses incorporating chromatin accessibility, MyoD binding and global gene expression profile has been recognized to be critical to understand
(i) the extent of, and (ii) any deficiencies in, MyoD-mediated reprogramming of cell types (Comai and Tajbakhsh, 2014). In this chapter, genome-wide chromatin and gene expression level changes in MyoD-induced transdifferentiated primary human fibroblasts is discussed with the focus to characterize reprogramming deficiencies at the chromatin and gene expression level. Additionally, this study aimed to generate testable hypotheses that could potentially improve MyoD-mediated transdifferentiation process.

In this study, I have assessed the genome-wide efficiency of the transdifferentiation process by analyzing global gene expression (RNA-seq), chromatin accessibility (DNase-seq), and MyoD binding (ChIP-seq) profiles of transdifferentiated primary skin fibroblasts in relation to those in the starting fibroblasts and target primary human myoblasts. Transdifferentiation was performed by over-expressing tet-inducible MyoD transgene that is transduced using lentiviral delivery system. This study revealed that MyoD overexpression leads to a continuum of DNaseI hypersensitive (DHS) site changes, with only a fraction of DHS sites getting completely reprogrammed, i.e. the degree of chromatin accessibility at those sites in transdifferentiated cells closely resembles that observed in primary myoblasts. I implemented a classification approach to analyze reprogrammed and non-reprogrammed DHS regions, and identify potential explanations for the incomplete reprogramming at the chromatin level. This comparative analysis also resulted in possible avenues for improving chromatin modification, involving additional regulatory factors and induced epigenetic changes. At the end of the chapter, I also present evidence of a strong correlation between chromatin reprogramming deficiencies and lack of complete reprogramming of gene expression. Given the causal role of chromatin accessibility in regulating gene expression, these analyses suggest that incomplete MyoD transdifferentiation is due, at least in part, to deficiencies in remodeling the chromatin landscape.
2.2 Methods

2.2.1 Cell culture

The cell culture were performed by Ami Kabadi and Jennifer B. Kwon in Gersbach Laboratory for normal and differentiated primary human fibroblasts, and Koji Tsumagari in Ehrlich Laboratory for primary human myoblasts. Primary human dermal fibroblasts (Catalog ID: GM03348) were obtained from Coriell Institute (Camden, New Jersey) and were maintained in DMEM supplemented with 10% FBS and 1% penicillin streptomycin. All cells were cultured at 37°C with 5% CO₂. Three biological replicates of both the parental fibroblast cells and transduced fibroblast cells were harvested on different days. Four replicates of myoblasts were originally sampled from biopsies on quadriceps and grown as described and published previously (Chandra et al., 2015; Tsumagari et al., 2013, 2011). As described previously, myoblasts cells were harvested at ~ 70% confluence, and all myoblast preparations had < 1% multinucleated cells. Importantly, immunostaining of the myoblast samples showed 90 – 98% desmin-positive cells (Tsumagari et al., 2011), indicative of highly pure populations of myoblasts.

2.2.2 Viral production and transduction

Viral production and transduction experiments were performed by Ami Kabadi and Jennifer B. Kwon in Gersbach Laboratory. All lentiviral vectors used in this study were produced from second-generation plasmids using standard viral production methods previously described (Kabadi et al., 2015). Briefly, 3.5 million HEK293T cells were plated per 10cm dish. The following day, cells were transfected with 20µg of transfer vector, 6µg of pMD2G, and 10µg psPAX2 using a calcium phosphate transfection. The media was changed 12 – 14 hours post transfection. The viral supernatant was collected 24 and 48 hours after this media change and pooled. For
transduction, the cell medium was replaced with viral supernatant supplemented with 4µg/mL polybrene. The viral supernatant was changed 24 – 48 hours later.

2.2.3 *MyoD*-directed genetic reprogramming

*MyoD*-directed reprogramming experiments were performed by performed by Ami Kabadi and Jennifer B. Kwon in Gersbach Laboratory. Human dermal fibroblasts were transduced with a Tet-ON lentivirus that expresses a 3xFlag-tagged full-length *MYOD1* cDNA (Kabadi et al., 2015). In this vector, a 3xFlag-tagged full-length human *MYOD1* cDNA-T2A-dsRed-Express2 cassette is expressed from the Tetracycline Responsive Element (TRE) promoter. T2A is a peptide that facilitates ribosomal skipping as the mRNA transcript is being translated into protein. The resulting products are two separate peptides that are expressed in a similar ratio. The vector constitutively expresses the Reverse Tetracycline Transactivator (rtTA2s-M2) and the Puromycin resistance gene from the human phosphoglycerate kinase (hPGK) promoter. The rtTA and PuroR are co-expressed from the same mRNA via an Internal Ribosomal Entry Site (IRES). The rtTA is only able to bind to the TRE and activate expression of the downstream genes in the presence of doxycycline.

Transduced cells were selected in 1µg/ml puromycin for 4 days until 100% cell death was observed in untransduced cells, in order to obtain a pure population. Cells were expanded in standard growth medium. Selected cells were seeded in appropriate plates and grown to confluence. 3×Flag-tagged MyoD transgene expression was induced by supplementing the medium with 3µg/mL doxycycline. Cells were given fresh media supplemented with doxycycline every two days. All differentiation studies were conducted in standard growth medium (DMEM supplemented with 10% FBS and 1% penicillin streptomycin). Cells were harvested after 10 days of doxycycline treatment. These cells are referred to subsequently as MyoD-induced. Given that cells were not transferred to differentiation media to further induced terminal
differentiation, in all our computational analyses we compare the transdifferentiated cells against myoblasts, which represent the early myogenic determination stage.

2.2.4 Quantitative reverse transcription PCR

Quantitative reverse transcription PCR experiments were performed by Ami Kabadi, Gersbach Laboratory. RNA was isolated using the RNeasy Plus RNA isolation kit (Qiagen). cDNA synthesis was performed using the SuperScript VILO cDNA Synthesis Kit (Invitrogen). Real-time PCR using PerfeCTa SYBR Green FastMix (Quanta Biosciences) was performed with the CFX96 Real-Time PCR Detection System (Bio-Rad). Oligonucleotide primers and PCR conditions are reported in Table S1. Primer specificity was confirmed by agarose gel electrophoresis and melting curve analysis. Reaction efficiencies over the appropriate dynamic range were calculated to ensure linearity of the standard curve (Kabadi et al., 2015). The results are expressed as fold-increase mRNA expression of the gene of interest normalized to Beta Actin expression using the Ct method. Reported values are the mean and S.E.M. from two independent experiments ($n = 2$) where technical replicates were averaged for each experiment. Effects were evaluated with multivariate ANOVA and Dunnetts post hoc test using JMP 10 Pro.

2.2.5 Western blot

The western blot experiments were done by performed by Ami Kabadi in Gersbach Laboratory. Cells were lysed in RIPA Buffer (Sigma) supplemented with protease inhibitor cocktail (Sigma). Protein concentration was measured using BCA protein assay reagent (Thermo Scientific) and BioTek Synergy 2 Multi-Mode Microplate Reader. Lysates were mixed with loading buffer and incubated at 70°C for 5 min; equal amounts of protein were run in NuPage 10% Bis-Tris Gel polyacrylamide gels (Bio-Rad) and transferred to nitrocellulose membranes. Nonspecific antibody bind-
ing was blocked with TBST (50 mM Tris, 150 mM NaCl and 0.1% Tween-20) with 5% nonfat milk for 1 hour at room temperature. The membranes were incubated with primary antibodies (anti-MyoD (Santa Cruz, Sc-32758) in 5% BSA in TBST, diluted 1:250, overnight at 4°C; anti-Myogenin (Santa Cruz, Sc-12732) in 5% BSA, diluted 1:250, overnight at 4°C; anti-Beta Actin (Sigma, A2066) in 5% milk in TBST, diluted 1:5,000, for 30 min at room temperature, and the membranes were washed with TBST for 15 min. Membranes were incubated with anti-rabbit HRP-conjugated antibody (Sigma, A 6154) or anti-mouse HRP-conjugated antibody (Santa Cruz, SC-2005) diluted 1:5,000 for 30 min and washed with TBST for 15 min. Membranes were visualized using the ImmunStar WesternC Chemiluminescence Kit (Bio-Rad) and images were captured using a ChemiDoc XRS+ System and processed using ImageLab software (Bio-Rad).

2.2.6 Immunofluorescence staining

The immunofluorescence staining experiments were performed by Ami Kabadi and Jennifer B. Kwon in Gersbach Laboratory. Fibroblasts transduced with Tet-ON LV co-expressing 3xFlag human MyoD and dsRed Express2 were plated on autoclaved glass coverslips (1mm, Thermo Scientific). Following 10 days of transgene expression, cells were fixed in 4% PFA and prepared for immunofluorescence staining. Samples were permeabilized in blocking buffer (PBS supplemented with 5% BSA and 0.2% Triton X-100) for one hour at room temperature. Samples were incubated with primary antibodies MF20 (Hybridoma Bank) diluted 1:200 and MyoD (Santa Cruz, Sc-32758) diluted 1:250 in blocking buffer overnight at 4°C, and rinsed for 15 min in PBS. Samples were incubated with secondary antibodies anti-mouse IgG2b Alexa Fluor 488 (Invitrogen, A-21141) and anti-mouse IgG1 Alex Fluor 647 (Invitrogen, A-21240) both diluted 1:500 for one hour at room temperature. Cells were incubated with DAPI diluted 1:5000 in PBS for 5 min and washed with PBS for 15 min.
Coverslips were mounted with ProLong Gold Antifade Reagent (Invitrogen) and imaged using a Leica SP5 inverted confocal microscope.

2.2.7 Flow cytometry analysis

Flow cytometry analysis was performed by Jennifer B. Kwon in Gersbach Laboratory. Following 8 days of transgene expression, cells were analyzed for dsRed expression by flow cytometry. Untransduced fibroblasts and fibroblasts transduced with Tet-ON LV co-expressing 3xFlag human MyoD and dsRed Express2 were harvested, washed once with PBS, and resuspended in 3% FBS in PBS. All cells were analyzed using the SH800 Cell Sorter (Sony Biotechnology).

2.2.8 DNase-seq

The DNase-seq experiments were performed by Lingyun Song in Crawford Laboratory. DNase-seq was performed as previously described (Song and Crawford, 2010b), with one modification: oligo 1b was synthesized with a 5’ phosphate to increase the efficiency of ligation. 5 – 20 million cells were used for each biological replicate. Libraries were sequenced on Illumina GAII or Hi-Seq 2000 sequencing platform in Duke Sequencing and Analysis Core Resource. Raw reads were trimmed to 20bp from 5’ and aligned to hg19 reference genome by using bowtie-0.12.9, with up to two mismatches and four mapping sites allowed. Blacklisted genomic regions (https://sites.google.com/site/anshulkundaje/projects/blacklists,(Consortium, 2012)) and PCR artifacts were filtered out and peaks were called by using MACS2 (version 2.1.0) with parameter --shift -100 --ext 200 at significance threshold of FDR 0.01. Data quality was evaluated using standard quality control (QC) metrics: (i) number of uniquely mapped reads, (ii) PCR Bottleneck Coefficient (PBC), (iii) Normalized Strand Cross-correlation coefficient (NSC), and (iv) Relative Strand Cross-correlation coefficient (RSC) (Landt
et al., 2012b). QC scores were comparable to those for available ENCODE DNase-seq data (Table S2).

2.2.9 Defining fibroblast- or myoblast-specific DHS sites

Differential DNase-seq peaks were determined using DESeq (Anders and Huber, 2010). The top 100k highest confidence DNase-seq peaks from fibroblasts and myoblasts (called by MACS2 as described above) were merged using bedtools merge with the default -d 0 option. Next, DESeq (Anders and Huber, 2010) was called on the merged set of DNase I hypersensitive (DHS) regions from the autosomal chromosomes ($N = 128,080$) to identify differentially accessible chromatin sites between the untransfected fibroblasts and the myoblast cells. DNase-seq read counts were computed for each DHS region for each of the three fibroblast and four myoblast replicates available. All DHS regions with at least two-fold differential enrichment and adjusted p-value $\leq 0.01$ were selected as differentially accessible sites between fibroblasts and myoblasts. Among these sites, DHS sites with significantly higher signal in fibroblasts were defined as “fibroblast-specific. Similarly, DHS sites with significantly higher signal in myoblasts were defined as “myoblast-specific”.

2.2.10 RNA-seq

The RNA-seq experiments were performed by Lingyun Song, Crawford Laboratory. RNA was extracted and purified by the methods in RNeasy Mini Handbook from Qiagen. RNA-seq libraries were made by standard TruSeq library preparation with PolyA selection, and processed 50bp paired-end sequencing on Illumina Hi-Seq 2000 platform in the Duke Sequencing and Analysis Core Resource. After trimming bases with low quality score, reads were aligned to the UCSC Genes hg19 reference transcriptome and filtering sequences containing no insertion, using Tophat with options x 4 and n 2 (Trapnell et al., 2009). Data quality was assessed using RNASeQC
(DeLuca et al., 2012). All samples showed high mapping rate, low number of alternative aligned reads, and low rate of mismatched bases (Table S2). Fragments Per Kilobase of exon per Million fragments mapped (FPKM) was estimated by Cufflinks and differential expression between normalized gene read counts in FPKM with significance threshold of FDR 0.05 by using Cuffdiff (Trapnell et al., 2010b).

2.2.11 Defining fibroblast- or myoblast-specific genes

Given our focus on fibroblast-to-myoblast conversion, we define “fibroblast-specific genes as genes with significantly higher expression levels in fibroblasts compared to myoblasts, as determined from the RNA-seq data (see above). Similarly, we define “myoblast-specific genes as genes with significantly higher expression levels in myoblasts compared to fibroblasts (this set includes, but it is not limited to, genes characterized in the literature as being muscle-specific). Using the definitions above, we identified 220 fibroblast-specific genes and 268 myoblast-specific genes.

For downstream analysis comparing the degree of chromatin versus gene expression reprogramming for these fibroblast- or myoblast-specific genes (Figure 2.22), transcript per million (TPM) expression values were computed jointly for these and all other known UCSC genes in the hg19 assembly.

2.2.12 ChIP-seq of MyoD-reprogrammed fibroblasts cells

The MyoD ChIP-seq experiments in MyoD-reprogrammed fibroblasts cells were performed by Lingyun Song in Crawford Laboratory. Chromatin immunoprecipitation was processed with two biological replicates of fibroblasts cells following induction of MyoD expression; each replicate contains 20 million cells. One of the biological replicates had two technical replicates. Briefly, cells were fixed with 1% formaldehyde (wt/vol) for 15 min, washed with 1× PBS and lysed in buffer with 50 mM Tris(pH8), 1% SDS (wt/vol), and 10 mM EDTA. Lysed cells were sonicated with 30 s on/off
cycles at high intensity by using a bioruptor (Diagenode). Sonicated supernatants were diluted with buffer containing 16.7 mM Tris-HCl(pH8), 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA and 167 mM NaCl. 10µl of FLAG antibody (Monoclonal anti-FLAG M1 antibody produced in mouse, F3040, Sigma-Aldrich) was added into the diluted supernatants, and 60µl of Dynabeads protein A beads were added and incubated for 3 hours at 4°C. ChIP-seq libraries were made by using NEBNext Ultra DNA Library prep kit for Illumina, and 50 bp single-end sequencing was processed on Hi-Seq 2000 platform in the Duke Sequencing and Analysis Core Resource. Sequences were aligned to hg19 reference genome by using bowtie-0.12.9. Uniquely aligned reads were used for peak calling with MACS2 (version 2.1.0) (see below). Data quality was evaluated using standard quality control (QC) metrics: (i) number of uniquely mapped reads, (ii) PCR Bottleneck Coefficient (PBC), (iii) Normalized Strand Cross-correlation coefficient (NSC), and (iv) Relative Strand Cross-correlation coefficient (RSC) (Landt et al., 2012b). QC scores were comparable to those for available ENCODE ChIP-seq data (Table S2).

2.2.13 Calling peaks from MyoD ChIP-seq data

MACS2 v2.10 (Zhang et al., 2008) was used to call peaks for each replicate, and the fragment length input was estimated based on the strand cross-correlation plot generated using SPP (Kharchenko et al., 2008). For the MyoD ChIP-seq data on fibro-MyoD cells, the reproducible peaks were identified using the ENCODE IDR (Irreproducible Discovery Rate) pipeline (Li et al., 2011). Specifically, all MyoD peaks with IDR ≤ 0.01 on at least one pair of replicates were selected for downstream analyses. For the MyoD ChIP-seq data on human myoblast cells, MACS2 was run without a control data set, as no control data was available from the study that reported the myoblast ChIP-seq data (MacQuarrie et al., 2013). Given the lack of a control data set, and the fact that only one replicate was available for the myoblast...
ChIP-seq data, the final set of MyoD peaks in myoblasts were called at stringent FDR threshold of $10^{-10}$.

### 2.2.14 Assessing significance of chromatin reprogramming

Given the large number of myoblast-specific DHS sites with Chromatin Reprogramming Level (CRL) scores around 0 (see Equation 1, Results section), we asked whether MyoD induction is in fact leading to significant chromatin remodeling. For this purpose, we used fibro-control DNase-seq data to compute a null distribution of CRL scores. Briefly, for each differentially accessible DHS site $s$, we computed the score

$$CRL_{fibro\text{-}control}(s) = \frac{DNase_{fibro\text{-}control}(s) - DNase_{fibroblast}(s)}{DNase_{myoblast}(s) - DNase_{fibroblast}(s)},$$

which reflects the level of chromatin accessibility changing due simply to transduction of the human $MYOD1$ gene, prior to its induction. As expected based on the fact fibro-control cells are very similar to fibroblasts, we found that the control distributions of values are tightly centered around zero (Figure 2.7). In addition, the distributions of CRL scores are significantly different between fibro-MyoD and fibro-control (Wilcoxon signed-rank test p-values $< 10^{-293}$, Figure 2.7), indicating that MyoD overexpression does lead to a highly significant level of genome-wide chromatin reprogramming. Nevertheless, many DHS sites failed to be reprogrammed upon MyoD induction.

### 2.2.15 Gene ontology enrichment analyses for genes and DHS sites

For the GO enrichment analyses of the fibroblast- or myoblast-specific reprogrammed or non-reprogrammed genes (Table S3), the DAVID tool (Huang da et al., 2009a,b) was used with the whole genome as background. For the fibroblast-specific or myoblast-specific DHS sites, the GO enrichment analyses were performed using
GREAT (McLean et al., 2010) (Table S4) using the whole genome as background and the default basal plus extension association rule.

2.2.16 Clustering TFs based on their DNA-binding specificities

TF-DNA binding specificity features were derived from protein-binding microarray (PBM) data (Berger et al., 2006). All PBM data sets for mammalian TFs were downloaded from UniPROBE (Hume et al., 2015). Additionally, DREAM challenge PBM data sets (Weirauch et al., 2013) of good quality (defined as data sets for which the derived DNA motifs have at least three consecutive positions with an Information Content $\geq 0.3$) were also included. The PBM data sets were used to generate a Pearson correlation coefficient (PCC) matrix. For each pair of data sets, we computed the PCC using only 8-mers with PBM E-score $\geq 0.35$ according to at least one of the data sets (i.e. 8-mers specifically bound by the TFs (Berger et al., 2006)) (Figure 2.11). We then searched for a threshold value $t$ such that 90% of all pairs of replicate TF data sets have PCC > $t$. The dendrogram generated from the PCC matrix was then cut at distance corresponding to the PCC threshold, i.e. at a distance $d = (1 - t)/2$, where $t = -0.0259$. For every cluster, the TF with the smallest average intra-cluster distance was chosen to represent the cluster. The distance between any two TFs was computed as $d = (1 - \text{PCC})/2$, where PCC is the Pearson correlation coefficient computed as discussed above. If there were only two elements in the cluster, one was chosen at random. This resulted in 140 clusters, with an average cluster size of 4. A full list of the 140 TF cluster representatives with their corresponding TF cluster members is available in Table S5.

2.2.17 Classification analyses of reprogrammed versus non-reprogrammed DHS sites

We performed classification analyses for the top 1000 reprogrammed versus non-reprogrammed DHS sites, ignoring outliers with very low or very high CRL scores.
Specifically, sites with CRL scores lower than −0.1 or higher than 1.1 were considered outliers. Within the [−0.1, +1.1] interval for myoblast-specific sites, the top 1000 sites (i.e. the sites in the “reprogrammed set) had CRL scores between +0.55 and +1.10, while the bottom 1000 sites (i.e. the sites in the “non-reprogrammed set) had CRL scores between −0.10 and +0.01. Within the [−0.1, +1.1] interval for fibroblast-specific sites, the top 1000 sites (i.e. the sites in the “reprogrammed set) had CRL scores between +0.67 and +1.10, while the bottom 1000 sites (i.e. the sites in the “non-reprogrammed set) had CRL scores between −0.10 and +0.31. We note that the results of our classification analyses are robust to variations in the precise CRL score intervals used to select reprogrammed and non-reprogrammed sites. Before performing the classification, we observed that the GC content was significantly different between the two sets (Figure 2.12), which could result in biased selection of TFs with GC-rich motifs as highly predictive features. In order to correct for the GC bias, we selected subsets of reprogrammed and non-reprogrammed DHS sites that are matched in their GC content, computed in the ±150bp regions around DHS centers. For each of the 1000 reprogrammed DHS sites, we search for a ‘matching non-reprogrammed site with similar GC content (i.e. GC content different by at most 1%). If a matching site was found, then both the reprogrammed DHS site and its matching non-reprogrammed site were selected. Otherwise, the reprogrammed DHS site was filtered out. Non-reprogrammed sites were selected without replacement. This resulted in a total of 847 reprogrammed and 847 non-reprogrammed myoblast-specific DHS sites. Similarly, a total of 531 reprogrammed and 531 non-reprogrammed fibroblast-specific DHS sites were selected for downstream classification analyses.

The following features were used in classification analyses: (1) 140 TF binding specificity features, defined for each TF cluster representative as the maximum 8-mer PBM enrichment score (E-score) in ±150bp genomic regions centered at DHS
sites, (2) 11 histone modification ChIP-seq signals from the Normal Human Dermal Fibroblasts (NHDF) cell line, computed as read pileups in $\pm 0.7$kb region around the DHS centers, (3) maximum CTCF and EZH2 ChIP-seq pileup signals in $\pm 0.7$kb region around DHS centers, obtained by scanning 200bp windows that overlap by 100bp, and (4) normalized fibroblast DNase-seq reads (as computed by DESeq) in the $\pm 150$bp region around DHS centers.

R packages `randomForest` and `glmnet` were used to run the Random Forest (RF) and Elastic Net (EN) analyses, respectively. RF and EN were repeated 10 times to assess the stability of the classification accuracy and of the feature importance scores, using 75% of the data (randomly selected) for training and the remaining 25% for testing. For RF analyses, the number of trees (`ntrees`) and number of predictor variables for splitting a node (`mtry`) were set to their default values. For EN analyses, the best parameter was selected from the training data using cross-validation by the `cv.glmnet()` function.

### 2.2.18 DHS pattern enrichment analysis

The top 100,000 DNase-seq peaks from each of the three cell lines (fibroblast, fibro-MyoD and myoblast) were merged using the `bedtools merge -d 0` command, resulting in a total of 133,714 autosomal merged DHS sites, which were used in subsequent analyses. For every merged DHS site, a binary tag was assigned to indicate whether the DHS site was present (‘1’) or absent (‘0’) in each of the three cell lines. Hence, every merged DHS site was annotated with one of seven binary triplets (‘001, ‘010, ‘011, ‘100, ‘101, ‘110, ‘111’), where the binary tags are for fibroblast, fibro-MyoD and myoblast, in this order. Next, we checked whether these patterns are significantly enriched or depleted around the genes corresponding to fibroblast-specific or myoblast-specific reprogrammed or non-reprogrammed gene sets, hereby referred to as treatment gene-sets (TS), when compared against a corresponding background set.
We analyzed four treatment sets: 1) myoblast-specific reprogrammed genes, 2) myoblast-specific non-reprogrammed genes, 3) fibroblast-specific reprogrammed genes, and 4) fibroblast-specific non-reprogrammed genes. For the myoblast-specific treatment sets, the background set was defined as all genes expressed in myoblasts at FPKM $\geq 5$. Similarly, for the fibroblast-specific treatment sets, the background set was defined as all genes expressed in fibroblasts at FPKM $\geq 5$.

The GSEA algorithm (Subramanian et al., 2005) was implemented to perform enrichment analyses for the DHS patterns. For any DHS pattern $p$ and gene set TS (short for “treatment-set”), all the genes in the corresponding background set BG were sorted in decreasing order of the number of occurrences of pattern $p$ in $\pm 100$kb of the TSS. Then, a running sum statistic, or enrichment score (ES) was computed at every position in the ranked list of genes, as follows:

$$ES_i = \begin{cases} \frac{1}{N_S} & \text{if gene } i \in \text{TS} \\ \frac{1}{N-N_S} & \text{otherwise} \end{cases}$$

where $N$ represents the size of BG and $N_S$ represents the size of TS. The maximum running sum statistic (MES = maximum enrichment score) was recorded as the enrichment for pattern $p$ in TS compared to BG. To assess the statistical significance of MES scores, null distributions were generated by randomly permuting the genes in BG 2000 times. Empirical p-values were computed from the null distributions, and corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) as implemented in the `statsmodels.sandbox.stats` python module.

2.2.19 Computing correlations between chromatin and gene expression reprogramming levels (CRLs vs. GRLs)

For fibroblast- and myoblast-specific genes, we asked whether the extent to which gene expression is reprogrammed correlates with the chromatin reprogramming level
(CRL) around the gene. To answer this question, we focused on DHS sites in either 5 kb or 2 kb regions centered at the TSSs. Fibroblast-specific DHS sites were considered in analyses of fibroblast-specific genes, and myoblast-specific DHS sites were considered in analyses of myoblast-specific genes. The CRLs and the gene reprogramming levels (GRLs) were computed in log-scale as follows:

\[
\text{CRL}(s) \text{ or } \text{GRL}(g) = \frac{\log(Tb) - \log(Fb)}{\log(Mb) - \log(Fb)} \times 100,
\]

where \( s \) and \( g \) represent the DHS site and gene under consideration, respectively; \( Fb, Tb \) and \( Mb \) represent normalized accessibility or transcript per million (TPM) scores for the fibroblasts, transdifferentiated fibroblasts (i.e. fibro-Myod) or myoblasts, respectively. We chose to use log scale in this analysis in order to better observe the entire range of gene expression changes, including small and moderate changes.

2.3 Results

2.3.1 MyoD induces incomplete genome-wide changes in chromatin accessibility and gene expression

(Note: My contributions in this section pertain to Figures 2.1D,F,G; Figure 2.4; Figure 2.5; and partly Table 2.1).

In order to study the extent to which MyoD can transform non-muscle cells into muscle cells, we characterized the genome-wide chromatin accessibility and gene expression profiles of primary human skin fibroblasts, primary human myoblasts, and human fibroblasts that were transdifferentiated by induction of MyoD from a tet-inducible lentiviral vector for 10 days followed by puromycin selection for transduced cells. These cells are henceforth referred to as ‘fibro-MyoD’ (Figure 2.1A). The induction of expression of the \( MYOD1 \) transgene by doxycycline was confirmed by qRT-PCR (Figure 2.1B), western blot (Figure 2.1C), and immunofluorescence using an antibody to MyoD (Figure 2.2A). As expected, no MyoD protein was detected
in untransfected fibroblast cells, and minimal MyoD expression was observed in control fibroblast cells transduced with the lentiviral MYOD1-containing vector but not induced with doxycycline (cells henceforth referred to as ‘fibro-control’). In fibro-MyoD samples, ~75% of cells showed MyoD-positive nuclei by immunofluorescence staining (Figure 2.2A). In addition, to confirm induction of the MYOD1 cassette, flow cytometry was used to quantify expression of DsRed, which is co-expressed with MyoD (see Methods) (Figure 2.2B). Transdifferentiation was also confirmed by western blot of myogenin (Figure 2.1C), immunofluorescence staining for skeletal myosin heavy chain (Figure 2.2A), qRT-PCR for additional myogenic markers (Figure 2.2C), and RNA-seq for myogenic markers (Figure 2.1D), similar to previous studies (Choi et al., 1990; Davis et al., 1987; Kabadi et al., 2015).

To characterize the genome-wide chromatin changes induced by MyoD, we used DNase-seq to measure chromatin accessibility in fibroblast, fibro-control, fibro-MyoD, and primary myoblast cells (see Methods). We found that for many muscle-specific genes, such as ITGA7 (integrin, alpha 7), MYOG (myogenin) and DES (desmin), chromatin accessibility changes in and around these genes are readily detected as a result of MyoD induction, and these changes largely mirror the chromatin accessibility profiles for primary myoblasts (Figure 2.1E). This indicates that the over-expression of MyoD induces chromatin structure changes in a non-random manner that is highly similar to primary muscle cells. We also found that not all myoblast-specific DHS sites (defined here as DHS sites with significantly higher DNase-seq signal in myoblasts compared with fibroblasts; see Methods) open up in fibro-MyoD, with some opening up only partially (Figure 2.1E, Figure 2.32). This suggests that induction of MyoD alone cannot reorganize chromatin at all potentially myogenic DHS sites in transduced primary human fibroblasts. Furthermore, we also detected fibroblast-specific DHS sites that are not lost after induction of MyoD, indicating that some fibroblast-related regulatory element memory is maintained during the
A
Primary skin fibroblasts → Viral MyoD → Fibro-control → +Dox (10 days) → Fibro-MyoD ↔ Primary myoblasts

B
Relative RNA

C
DHS sites: Myoblast-specific Fibroblast-specific

D
RNA-seq (log2 (FPKM+1))

E
ITGA7
chr12:56,038,064-56,113,386

F
MYOG
chr1:203,015,163-203,089,513

G
DES
chr2:220,258,303-220,323,526

55
Figure 2.1: Induced expression of MyoD in fibroblast cells leads to incomplete genome-wide changes in gene expression and chromatin accessibility. (A) Schematic representation of the MyoD-induced transdifferentiation experiment performed. (B) qRT-PCR and (C) western blot showing an increased level of MyoD after induction by addition of doxycycline (dox) for 10 days. Neither MyoD nor another myogenic regulatory factor, Myogenin, is detected in either untransfected fibroblasts or in fibroblasts that contain the MyoD vector but are not treated with doxycycline. (D) RNA-seq data (logarithm of fragments per kilobase per million base shifted by 1, or log($FPKM + 1$)) showing that myogenic genes are significantly increased in fibro-MyoD cells (** = Welch’s t-test p-value < 0.01, * = Welch’s t-test p-value < 0.05). Error bars show the upper bound of the 95% confidence interval, as computed by Cufflinks. Note that early myogenic genes (ITGA7, CDH15) are reprogrammed, while middle (MYOG) and late (TNNT1, DES, TNNC1, MYBPH, MYLPF, CKM, LDB3, TNNC2) muscle markers are generally partially reprogrammed. (E) Representative DNase-seq signal around myogenic genes ITGA7 (early myogenic gene), MYOG (intermediate myogenic gene), and DES (late myogenic gene). Red boxes represent myoblast-specific DHS sites that are partially, completely or not reprogrammed in fibro-MyoD cells. Blue boxes represent fibroblast-specific DHS sites that are maintained in fibro-MyoD cells. Heatmaps of (F) RNA-seq and (G) DNase-seq for each replicate (B1-B4) of fibroblast, fibro-control, fibro-MyoD, and primary myoblasts are shown for differentially expressed genes and differentially accessible DHS sites, respectively. The data indicate that fibro-Myod cells display partial characteristics of both fibroblasts and myoblasts.

This incomplete reprogramming at the chromatin level is reflected in our RNA-seq data, which shows that although the levels of key myogenic marker genes are significantly higher in fibro-MyoD compared to fibroblasts (consistent with our qRT-PCR results), several of these genes are not expressed at the same levels as observed in primary myoblasts (Figure 2.1D). Some of the common myogenic marker genes have been previously designated as important in early (ITGA7, CDH15), intermediate (MYOG), or late (TNNT1, DES, TNNC1, MYBPH, MYLPF, CKM, LDB3, TNNC2) myogenic differentiation (Bergstrom et al., 2002; Cao et al., 2006; Ishibashi et al., 2005). Our RNA-seq data indicate that ‘early’ genes display a large degree of reprogramming (Figure 2.1D). This result, combined with the majority of trans...
A

<table>
<thead>
<tr>
<th></th>
<th>MyoD positive nuclei</th>
<th>MHC positive nuclei</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibroblasts</td>
<td>0 +/- 0 %</td>
<td>0 +/- 0 %</td>
</tr>
<tr>
<td>Fibro-MyoD</td>
<td>74.6 +/- 5.5 %</td>
<td>17.8 +/- 6.2 % (of MyoD positive nuclei)</td>
</tr>
<tr>
<td>Myoblasts</td>
<td>MyoD and MHC staining not available. However, 90-98% of cells are Desmin-positive, and &lt;1% are multinucleated, as shown previously [5, 6].</td>
<td></td>
</tr>
</tbody>
</table>

B

C

Myogenin

Troponin T

Desmin

Creatine Kinase
differentiated cell nuclei being MyoD-positive (Figure 2.2A), indicates that MyoD induction is indeed differentiating cells toward a myogenic fate. In contrast to ‘early’ genes, ‘late’ marker genes show a wide range of reprogramming levels, with a majority being moderately upregulated and a few showing significant (\textit{TNNC1}) or no (\textit{LBD3} and \textit{TNNC2}) change in expression (Figure 2.1D, Table 2.1). It is possible that ‘late’ myogenic genes were not expressed at very high levels in fibro-MyoD cells because transdifferentiation was performed in growth media, without inducing terminal differentiation (see Methods). Nevertheless, according to our direct comparison to myoblasts (which are also not terminally differentiated), some ‘late’ marker genes are, at best, only partially reprogrammed. Importantly, muscle marker genes previously reported to be upregulated during induction of MyoD expression in transfected mouse embryonic fibroblasts incubated in growth medium are also upregulated in our study (Ishibashi et al., 2005).

In addition to incomplete reprogramming at individually assayed myogenic genes, we also observed incomplete reprogramming at the genome-wide level for both gene expression (Figure 2.1F) and chromatin accessibility (Figure 2.1G). This indicates
Figure 2.3: DNase-seq data around myogenic genes for fibroblasts, fibroblasts expressing MyoD (fibro-MyoD), and primary myoblasts. The red boxes highlight myoblast-specific DHS sites that are partially or completely reprogrammed in fibro-MyoD cells. The blue boxes highlight fibroblast-specific DHS sites that are partially or completely reprogrammed in fibro-MyoD cells.
Table 2.1: Gene expression TPM values for some muscle genes, SAND domain genes and a chromatin modifying enzyme, UHRF1 across normal fibroblast (“F”), fibro-MyoD (“Fd”) and myoblast (“Mb”) samples. The TPM values are rounded to nearest integers. Values for fibro-control samples are close to normal fibroblast and are not shown. (Adapted from (Manandhar et al., 2017) Table S7)

<table>
<thead>
<tr>
<th>GeneId</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>Fd1</th>
<th>Fd2</th>
<th>Fd3</th>
<th>Mb1</th>
<th>Mb2</th>
<th>Mb3</th>
<th>Mb4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYOD1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>109</td>
<td>117</td>
<td>77</td>
<td>172</td>
<td>171</td>
<td>27</td>
<td>113</td>
</tr>
<tr>
<td>ITGA7</td>
<td>35</td>
<td>9</td>
<td>5</td>
<td>120</td>
<td>181</td>
<td>87</td>
<td>348</td>
<td>213</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>CDH15</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>63</td>
<td>88</td>
<td>49</td>
<td>83</td>
<td>210</td>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>MYOG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>8</td>
<td>43</td>
<td>6</td>
<td>669</td>
<td>136</td>
<td>1</td>
</tr>
<tr>
<td>TNNT1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>104</td>
<td>103</td>
<td>136</td>
<td>60</td>
<td>673</td>
<td>315</td>
<td>32</td>
</tr>
<tr>
<td>DES</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>24</td>
<td>13</td>
<td>56</td>
<td>928</td>
<td>2734</td>
<td>1414</td>
<td>563</td>
</tr>
<tr>
<td>TNNC1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>15</td>
<td>119</td>
<td>63</td>
<td>3156</td>
<td>623</td>
<td>8</td>
</tr>
<tr>
<td>MYBPH</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>10</td>
<td>42</td>
<td>38</td>
<td>832</td>
<td>304</td>
<td>4</td>
</tr>
<tr>
<td>MYLPF</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>54</td>
<td>25</td>
<td>2999</td>
<td>804</td>
<td>3</td>
</tr>
<tr>
<td>CKM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>108</td>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>LDB3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>34</td>
<td>16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TNNC2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>309</td>
<td>112</td>
<td>1</td>
</tr>
<tr>
<td>SIX1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>62</td>
<td>42</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>SIX4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>GMEB1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>SP100</td>
<td>19</td>
<td>15</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>14</td>
<td>10</td>
<td>9</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>SKI</td>
<td>16</td>
<td>21</td>
<td>16</td>
<td>18</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>11</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>AIRE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DEAF1</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>7</td>
<td>13</td>
<td>16</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>UHRF1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>49</td>
<td>29</td>
<td>13</td>
<td>60</td>
</tr>
</tbody>
</table>

that fibro-MyoD cells exhibit some of the characteristics of both fibroblasts and myoblasts simultaneously. To verify that the fibro-MyoD cell population is not simply a mixture of undifferentiated (fibroblast) and transdifferentiated (myoblast) cells, we performed in silico mixing experiments at different ratios of fibroblast to myoblast cells. As expected, we found that the gene expression and chromatin accessibility profiles of mixed populations (Figure 2.4) are substantially different from the profiles we observed for fibro-MyoD cells (Figure 2.1F,G).
Figure 2.4: In-silico mixing showing (A) gene expression and (B) DNase-seq profiles of cell populations that contain fibroblast and myoblast cells in different ratios.

Figure 2.1F shows the expression levels of genes significantly differentially expressed between the starting (fibroblast) and target (myoblast) cell types. We refer to these genes as “fibroblast-specific” or “myoblast-specific”, depending on whether the genes are significantly more expressed in fibroblasts or myoblasts, respectively (see Methods). As expected, all genes have similar expression levels between fibroblasts and fibro-control cells. After MyoD induction, some genes are up- or down-regulated and, as a result, their expression levels become similar to those in myoblasts i.e., the
genes have their expression reprogrammed. However, a large fraction of fibroblast-specific or myoblast-specific genes also remain non-reprogrammed. We identified these reprogrammed and non-reprogrammed genes through pairwise comparisons of the expression levels in fibroblasts, fibro-MyoD and myoblasts, as computed by Cuffdiff (see Methods). In order to assess the functional significance of these reprogrammed or non-reprogrammed genes, we performed gene ontology (GO) analysis using DAVID (Huang da et al., 2009a,b). The myoblast-specific non-reprogrammed genes \( N = 100 \), as well as the reprogrammed genes \( N = 168 \), are highly enriched for muscle-specific GO categories such as “sarcomere and “striated muscle contraction (Table S3). This indicates that non-reprogrammed genes are also relevant to muscle biology. Similarly, both reprogrammed \( N = 34 \) and non-reprogrammed \( N = 186 \) fibroblast-specific genes are enriched for fibroblast-associated terms such as “extracellular matrix structural constituent (Table S3). These observations suggest that, based on its gene expression profile, the fibro-MyoD cell line represents an intermediate between fibroblast and myogenic cells.

Similar to the gene expression analysis, global chromatin accessibility (DNase-seq) data revealed the fibro-MyoD cell line represents an intermediate between fibroblast and myogenic cells (Figure 2.1G). While many DHS sites are well reprogrammed (i.e. their accessibility level is similar to that in myoblasts, see Methods), a large number of sites that are different between fibroblasts and myoblasts do not change their accessibility profile after MyoD induction. To assess the potential regulatory activities of these sites, we performed GO enrichment analyses using GREAT (McLean et al., 2010) (Table S4; see Methods). As expected, myoblast-specific DHS sites that open up in fibro-MyoD (i.e. are reprogrammed) are highly enriched for GO annotations pertaining to muscle and satellite cell differentiation. In addition, myoblast-specific DHS sites that remain closed in fibro-MyoD (i.e. are non-reprogrammed) also show enrichment for muscle and regeneration ontologies, albeit to a lower extent than
for the reprogrammed DHS sites. Similarly, fibroblast-specific DHS sites that lose accessibility (i.e. are reprogrammed) are associated with fibroblast-specific genes that have been previously reported to be down-regulated in response to induction of MyoD expression in transduced 3T3 mouse fibroblasts (de la Serna et al., 2005) (Table S4). In addition, fibroblast-specific DHS sites that remain open during transdifferentiation (i.e. are not reprogrammed) are involved in epithelial-mesenchymal transition, cell motility and interstitial matrix, which are fibroblast-associated annotations. The fact that we observe such non-reprogrammed fibroblast-specific DHS sites is consistent with previous reprogramming studies that also found evidence of an “epigenetic memory” from the cells of origin (Feng et al., 2008; Kim et al., 2010; Marro et al., 2011).

**ChromHMM genome-wide chromatin accessibility profile**

Most of the analyses in this study focus on genome-wide comparative study of transdifferentiated cells in relation to the starting fibroblasts and target myoblasts. However, as part of the preliminary analysis, we also profiled genome-wide chromatin accessibility dynamic across the cell lines used in this study together with Myotubes (as terminal muscle cell lines, as is used in Chapter 4). For this, we ran ChromHMM model (Ernst and Kellis, 2012) on the DNase-seq data across the cell lines (Figure 2.5). ChromHMM is a multivariate hidden Markov model that can model the presence (“1”) or absence (“0”) of chromatin accessibility across the cell lines as a product of independent Bernoulli random variables, and identify binarized-chromatin states that are enriched in the genome ((Ernst and Kellis, 2012)). As input for ChromHMM, the genome is binned into 200bp non-overlapping windows with each bin marked as accessible (“1”) or (“0”) for all cell lines as identified by MACS2 peak caller at p-value cutoff of 0.01 and FDR cutoff of 1.5 (Zhang et al., 2008). For every bin, the bin is designated accessible if and only if at least half of a MACS2-called peak overlaps
its domain. The binarized chromatin accessibility profiles for the cell lines are passed into ChromHMM to learn a 12-state model using default options.

**ChromHMM Model**

Figure 2.5: ChromHMM model output on DNase-signal across the normal fibroblasts, transdifferentiated fibroblasts, and muscle cell lines.

Figure 2.5 shows the emission and transition probabilities learned from the 12-state chromHMM model. The model identifies 12 distinct chromatin states, with probability of observing an accessible chromatin at any particular cell line indicated by the shade of the corresponding cell. For instance, state 1 is representative of chromatin states that is extremely likely to be accessible in transdifferentiated fibroblasts, minimally likely to be accessible in normal fibroblasts, and extremely unlikely to be accessible in other cell types. As such, this state could be designated as “Improperly reprogrammed chromatin” as the chromatin is by and large only accessible in
the transdifferentiated fibroblasts. These “improperly reprogrammed chromatin” are discussed briefly in the Discussion section later. The genome coverage plot shows that roughly 0.3% of the genome has this chromatin state. Similarly, states representative of “non-reprogrammed fibroblast-specific chromatin” (state 2, ~ 0.5% genome coverage), “reprogrammed fibroblast-specific chromatin” (state 3, ~ 0.5% genome coverage) and “reprogrammed muscle-specific chromatin” (state 6, ~ 0.35% genome coverage) also exist. State 6 also jointly represents “non-reprogrammed muscle-specific sites” since the probability of observing accessible chromatin in transdifferentiated fibroblasts is roughly around 0.5. Hence, state 5 and 6 jointly represent the collection of sites that are “myoblast-specific” and “non-reprogrammed” (which would cover < 0.5% of the genome) in this study. Among all chromatin states, while state 9, representative of ubiquitously inaccessible chromatin, is the most abundant (~ 96% genome coverage), the next most abundant state of chromatin is ubiquitously accessible chromatin (~ 1.3% genome coverage).

The transition probabilities between states shows that all 200 bp chromatin states transition into itself or closed chromatin (state 9) with next highest probability. Not surprisingly, among states with at least one cell line accessible, it can also be observed that states that are closer to each other (such as states 4 and 7, states 10 and 7) see higher transition probabilities. It should be noted however, transition probability to ubiquitously accessible chromatin (state 7) is not similar to transition probability out of the state.

As discussed, chromHMM succinctly captures the global chromatin accessibility profile. Nevertheless, a couple of disadvantages do exist. First, the chromatin states need to be binarized and therefore, cannot capture the degree that chromatin accessibility changes between cell lines. This is one of the primary interest in this transdifferentiation study to quantitatively characterize genome-wide accessibility change. The next sections are motivated by this goal. Second, as mentioned above
with state 6, some chromatin state representation is not best. State 6 would ideally be split to two states, one representative of “non-reprogrammed” and another representative of “reprogrammed”, muscle-states.

2.3.2 MyoD induction results in a continuum of reprogrammed DHS sites

In order to quantify reprogramming efficiency at DHS sites differentially accessible between fibroblasts and myoblasts, we introduce the ‘chromatin reprogramming level (CRL) score. For each differential DHS site \( s \), the CRL score is defined as the ratio of the change in chromatin accessibility due to MyoD induction in fibroblasts relative to the difference in accessibility between fibroblasts and myoblasts:

\[
CRL_{\text{fibro-MyoD}}(s) = \frac{DNase_{\text{fibro-MyoD}}(s) - DNase_{\text{fibroblast}}(s)}{DNase_{\text{myoblast}}(s) - DNase_{\text{fibroblast}}(s)},
\]

where \( DNase_c(s) \) represents the normalized DNase-seq read signal at DHS site \( s \) in cell type \( c \) (see Methods). For both fibroblast- and myoblast-specific DHS sites, a CRL score close to 0 indicates that the site \( s \) is not reprogrammed, i.e. its chromatin accessibility in fibro-MyoD cells is similar to that in fibroblasts, while a CRL score close to 1 indicates that the site \( s \) is reprogrammed, i.e. its chromatin accessibility in fibro-MyoD cells changed significantly and is similar to the level of chromatin accessibility observed in myoblasts (Figure 2.6A,B).

We detect a continuum of CRL scores for myoblast-specific (Figure 2.6A) and fibroblast-specific (Figure B) DHS sites that span mostly between 0 (non-reprogrammed sites) and 1 (reprogrammed sites). Interestingly, while CRL distribution for myoblast-specific DHS sites is positively skewed (median of only 0.16), the median CRL for fibroblast-specific DHS sites is 0.49. This suggests that reprogramming efficiency is generally higher for fibroblast-specific compared to myoblast-specific sites. We confirmed that these positively skewed CRL scores are significant by comparing them to CRL scores computed using fibro-control DNase-seq data.
Figure 2.6: MyoD induction results in continuum of chromatin reprogramming genome-wide. (A) Distribution of chromatin reprogramming level (CRL) scores for myoblast-specific DHS sites. For classification analyses we selected 1000 non-reprogrammed sites and 1000 reprogrammed sites from the ends of the distribution, ignoring potential outliers with extreme CRL scores. Schematics of sites that are not reprogrammed (pink) or reprogrammed (dark red) are shown. (B) Analogous to panel A but for fibroblast-specific DHS sites. (C) Scatterplot showing positive correlation between CRL scores and MyoD ChIP-seq signal for myoblast-specific DHS in fibro-MyoD cells (Spearman correlation coefficient: 0.66). The colors for the top 1000 reprogrammed or non-reprogrammed sites (dark red and pink, respectively) correspond to panel A. (D) ROC curves for classification of reprogrammed versus non-reprogrammed DHS sites that are myoblast-specific (red) or fibroblast-specific (blue), obtained using MyoD ChIP-seq signal from either fibro-MyoD cells (solid lines) or from myoblasts (dotted lines). (E) Analogous to panel C, but for fibroblast-specific DHS sites (Spearman correlation coefficient: −0.17).
Figure 2.7: Chromatin Reprogramming Level (CRL) distributions for fibro-MyoD cells compared against fibro-control cells. (A) Distributions of chromatin reprogramming level (CRL) scores for myoblast-specific DHS sites in fibro-MyoD cells (solid red) and in fibro-control cells (gray curve). The two distributions are significantly different (Wilcoxon signed-rank test p-value < 10e-293). (B) Distributions of chromatin reprogramming level (CRL) scores for fibroblast-specific DHS sites in fibro-MyoD cells (solid blue) and in fibro-control cells (gray curve). The two distributions are significantly different (Wilcoxon signed-rank test p-value < 10e-293). We note that the variance of $CRL_{fibro\rightarrow control}$ scores is larger for fibroblast-specific DHS sites (panel B) than myoblast-specific DHS sites (panel A). This is because the differences in DNase-seq signal between myoblast and fibroblast cells (i.e. the changes required for complete reprogramming) are significantly smaller for fibroblast-specific compare to myoblast-specific DHS sites, i.e. the denominator for is smaller for fibroblast-specific compared to myoblast-specific DHS sites.

In order to understand why many DHS sites fail to open up during transdifferentiation, we first compared two sets of myoblast-specific sites: a set containing 1000 non-reprogrammed sites with the smallest CRL scores, and another set containing 1000 reprogrammed sites with the largest CRL scores, ignoring potential outliers (Figure 2.6A, see Methods). Similarly, in order to understand why some DHS sites fail to close down during transdifferentiation, we compared two sets of 1000 fibroblast-specific DHS sites with CRL values indicative of reprogrammed versus non-reprogrammed status (Figure 2.6B, see Methods). Using the control distribution of CRL scores derived from fibro-control DNase-seq data (Figure 2.7), we estimate the false discovery rate (FDR) for the 1000 reprogrammed DHS sites to be 0.01 and
0.005 for myoblast-specific and fibroblast-specific sites, respectively (see Methods). Due to a low level of leaky expression of MyoD in fibro-control cells (Figure 2.1B), these FDR values are likely conservative. Importantly, the results presented below are robust to variations in the precise ranges of CRL scores used to call the 1000 reprogrammed and 1000 non-reprogrammed sites.

Previous small-scale studies reported that MyoD can open chromatin and attract chromatin-remodeling enzymes to myogenic enhancers (de la Serna et al., 2005; Gerber et al., 1997; Tapsecott, 2005). Our ChIP-seq data for MyoD in fibro-MyoD cells shows that about half of the binding sites occurred in previously open chromatin regions, while the other half occur at sites that were closed in fibroblast cells but opened up in response to MyoD overexpression. We also found that the majority of these MyoD binding sites in fibro-MyoD are also accessible in myoblasts (Figure 2.8). Therefore, we asked whether the lack of chromatin reprogramming at myoblast-specific DHS sites might be correlated with lack of MyoD binding. Indeed, we observed a positive correlation between the MyoD ChIP-seq signal in fibro-myoD cells (quantified at ±150bp of DHS center) and the corresponding CRL values (Spearman correlation coefficient: 0.66, Figure 2.6C). Moreover, we also found that the MyoD ChIP-seq signal from fibro-MyoD cells can almost perfectly separate reprogrammed from non-reprogrammed myoblast-specific DHS sites (area under the receiver-operating characteristic curve (AUC): 0.96, Figure 2.6D, solid red curve). Broadly, these results indicate that MyoD binding in transdifferentiated cells is strongly associated with chromatin opening, while non-reprogrammed myoblast-specific sites lack MyoD binding. We performed a similar classification analysis using the MyoD ChIP-seq signal from primary myoblasts (MacQuarrie et al., 2013), and found that binding of MyoD in myoblasts is less accurate in distinguishing reprogrammed from non-reprogrammed DHS sites (AUC: 0.76, Figure 2.6D, dotted red curve). This drop in AUC is likely due to the fact that ~33% of myoblast-specific
DHS sites that are not reprogrammed and not bound by MyoD in transdifferen-
tiated cells show MyoD ChIP-seq peaks in myoblasts (at MACS2 FDR cutoff $10^{-10}$, see Methods). Overall, these results show a large but not complete overlap between MyoD-bound sites in fibro-MyoD and myoblasts (Figure 2.9).

![Figure 2.8](image)

**Figure 2.8**: Genomic sites bound by MyoD in fibro-MyoD cells, grouped by the DHS pattern of the site. The DHS pattern is a binary triplet representing the presence (“1”) or absence (“0”) of a DNase-seq peak in fibroblasts, fibro-MyoD cells, and myoblasts, in this order. Note that close to half of MyoD-bound sites in fibro-MyoD cells were closed in fibroblast cells.

In contrast to MyoD being closely associated with the opening of chromatin at myoblast-specific DHS sites, MyoD binding in fibro-MyoD cells is slightly inversely correlated with chromatin closing at fibroblast-specific sites (Spearman correlation coefficient: $-0.17$, Figure 2.6E). Hence, MyoD binding is a weak predictor of reprogramming status of fibroblast-specific sites (AUC ~ 0.38, Figure 2.6D). This suggests that binding of MyoD at some already open chromatin sites could potentially prevent closing down of these sites, but this would only explain a small fraction of non-reprogrammed sites. Indeed, only 8 of the 1000 non-reprogrammed fibroblast-specific sites show MyoD binding signal in fibro-MyoD (ChIP-seq peaks called by MACS2 and filtered at IDR 0.01, see Methods).
**Figure 2.9:** Comparisons between MyoD ChIP-seq data in fibro-MyoD and myoblasts. (A) Concordance plot of MyoD binding between fibro-MyoD and primary human myoblasts, plotted as in (Cao et al., 2010). Briefly, in order to get each cell in the matrix, the corresponding overlapping number of peaks between fibro-MyoD and myoblasts is divided by the minimum between the number of peaks in fibro-MyoD and the number of peaks in myoblasts, and then multiplied by 100. The plot shows high concordance between the two sets of peaks. For example, > 85% of the top 2.5k peaks in fibro-MyoD also overlap with MyoD peaks in myoblasts; similarly, > 65% of the top 2.5k peaks in myoblasts are present in our set of reproducible MyoD peaks in fibro-MyoD. (B, C) MyoD binding events in fibro-MyoD (B) and Myoblasts (C) relative to TSS (left panels), as obtained using GREAT (McLean et al., 2010), and in genic regions (right panels), as obtained using ChiPeakAnno (Zhu et al., 2010).
Figure 2.10: Diagrammatic representation of TF-DNA binding specificity and histone modification signal used as features for classification analyses. (A) The DNA-binding preference of a TF at a DHS site is captured by its binding specificity score (or E-score) for the highest affinity 8-mer (shown in black) in s. (B) For the feature capturing the in vivo enrichment of HM m in the starting cell line, we use the integrated ChIP-seq pileup for m in the fibroblast cells, over a region covering ±0.7kb relative to the DHS-center. (Note that for CTCF and Ezh2, the features are similarly computed, but using overlapping 200bp windows (see text).

2.3.3 TF motifs distinguish reprogrammed from non-reprogrammed myoblast-specific DHS sites

To understand why the non-reprogrammed DHS sites are not bound by MyoD in fibro-MyoD cells, and what other factors might contribute to their lack of MyoD binding and/or their lack of reprogramming, we used a classification approach to distinguish the top 1000 reprogrammed sites from the top 1000 non-reprogrammed sites based on genomic and epigenomic features. The features used for this analysis were: 1) how open the site was in the starting cell type (i.e. its DNase-seq signal in fibroblast cells); 2) the DNA binding specificities of various TFs for the DHS site, as determined using protein-binding microarray (PBM) data (see 2.10A and Methods) (Berger et al., 2006); and 3) the pre-existing chromatin environment at the DHS site, as reflected by ChIP-seq signal for 11 histone marks (2.10B), CTCF, and EZH2 proteins in normal human dermal fibroblast (NHDF) cells (Consortium, 2012). To derive TF-DNA binding specificity features, we used PBM data for human and mouse TFs (Hume et al., 2015; Weirauch et al., 2013). In addition, since PBM data are not
available for human MyoD, we used data for the MyoD ortholog in *C. elegans*, HLH-1, henceforth referred to as cMyoD. Since many TFs in our data set have similar DNA binding specificities, which could lead to high collinearity among features used in our classifier, we clustered the PBM data into 140 clusters (Methods). For each cluster, we selected a representative TF that minimizes the intra-cluster distance (Table S5; Methods), and we used as a feature the maximum DNA binding specificity score (i.e. the maximum 8-mer PBM E-score) in the ±150bp genomic region around the center of the DHS site (see Methods). The PBM E-score is a modified form of the Wilcoxon-Mann Whitney statistic and ranges from −0.5 (least favored sequence) to +0.5 (most favored sequence), with values above 0.35 corresponding, in general, to sequence-specific TF-DNA binding (Berger et al., 2006) (Figure 2.11).

To ensure that the two sets of sequences used in our classification have similar GC content, we randomly sampled reprogrammed and non-reprogrammed sites so that the GC-content distributions for the two sets of sites were matched (Figure 2.12). This reduced the number of sites in each class from 1000 to 847. This step was necessary to avoid the potential bias of selecting certain TFs as predictive features simply because of the GC-content of their motifs. Next, to perform the classification, we randomly selected 75% of the input sequences to train a Random Forest or an Elastic Net model, and used the remaining 25% for testing. We repeated this training-testing step 10 times in order to get an assessment of the stability of the classification accuracy, as well as the stability of the feature ranking. The results of our Random Forest classification analysis (Figure 2.13) are similar to the results obtained using Elastic Net (Table S6) and also similar to the results obtained for sites not matched by GC content (Figure 2.14).

Using a Random Forest classifier on reprogrammed versus non-reprogrammed sites, we obtained a high classification accuracy of 80.78 ± 2.35% (Figure 2.13A). The most predictive feature for distinguishing reprogrammed from non-reprogrammed
Figure 2.11: Correlation between PBM data sets. (A) Distributions of the Pearson correlation coefficient (PCC) between all (black) and replicate (red) PBM data sets. PCCs were computed using only 8-mers that are likely to be bound specifically by the TFs, i.e. 8-mers with PBM E-score ≥ 0.35 (Berger et al., 2006) for one or both of the TFs in each comparison. At this E-score cutoff, replicate data sets typically have positive Pearson correlation coefficients (PCCs), with a median of about 0.5. Vertical dotted line shows the PCC threshold $t$ used in our clustering analysis (see Methods). The cutoff was selected such that 90% of pairs of replicate PBM data sets have PCC $> t$. (B-D) Scatterplots comparing two PBM data sets. Red: 8-mers with PBM E-score ≥ 0.35. Black: 8-mers with PBM E-score < 0.35. Panel (B) shows that unrelated TFs have a strong negative correlation for 8-mers with E ≥ 0.35. Panel (C) shows that good replicate data sets can have low correlation over all 8-mers, but still correlate well for 8-mers with E ≥ 0.35. Panel (D) shows that an example of replicate data sets with high correlation over all 8-mers, as well as over 8-mers with E-score ≥ 0.35. We note that over all 8-mers, the correlation is lower for Bhlhb2 (panel D) than for Arid5a (panel C). However, this correlation is largely driven by non-specific 8-mers (i.e. those with E-score < 0.35). When focusing on specifically-bound 8-mers, the correlation is much higher for Bhlhb2, which is indicative of better replicate data.
Figure 2.12: GC content distributions for the ±150bp regions around the centers of myoblast- or fibroblast-specific DHS sites categorized as reprogrammed or not. The distributions were computed (A) before and (B) after matching the reprogrammed and non-reprogrammed sets according to GC content (see Methods). The GC content was significantly different between reprogrammed and non-reprogrammed myoblast-specific (KS test p-value: 2.18e−11) and fibroblast-specific (KS test p-value: 1.31e−101) DHS sites before matching the sets according to their GC content.

sites is the DNA-binding specificity of cMyoD and Myf6, which are myogenic regulatory factors (MRFs) with similar DNA binding specificity (Braun and Arnold, 1991). This is consistent with our results in Figure 2.6D. The greater binding specificity of cMyoD/Myf6 for DNA sequences present in reprogrammed vs. non-reprogrammed DHS sites indicates that the lack of MyoD binding at non-reprogrammed DHS sites can be largely attributed to low MyoD-DNA binding affinity. The second most predictive feature is the DNA binding specificity of Ascl2, which is a basic helix-loop-helix factor involved in neuronal differentiation that has a DNA binding specificity similar to MyoD (Johnson et al., 1992; Li et al., 2006; Meierhan et al., 1995). The
Figure 2.13: Genetic and epigenetic features that distinguish reprogrammed from non-reprogrammed DHS sites. (A) For myoblast-specific sites, the reprogramming status is largely predicted by DNA binding specificities of MyoD, MyoD cofactors (Tcfe2a, Meis1/Pknox1), and SAND-domain factors (Gmeb1/2, Sp100). (B) For fibroblast-specific sites, the most predictive features include several active epigenetic marks (H3K4me2, H3K4me3, etc.) that were present before MyoD induction.

Top panels show the classification accuracy and the top 15 most predictive features obtained from Random Forest classifiers. For each feature, the color of the bar corresponds to the class most correlated with the feature (reprogrammed = dark color; non-reprogrammed = light color). Bottom panels show, for selected features, the heatmaps corresponding to either the DNA binding specificities of the TF cluster representatives (in panel A), or the ChIP-seq signal for histone marks and CTCF in a dermal fibroblast cell line, NHDF (in panel B). The summary plots above each heatmap show the average signal for either of the two classes.
Figure 2.14: Classification results showing genetic and epigenetic features that distinguish reprogrammed from non-reprogrammed DHS sites, using sets of sites that were not match according to GC content (see Methods). This figure is analogous to Figure 2.13, which showed the classification result and heatmaps for sites after matching according to GC content. Note that the set of top features largely overlap with those in Figure 2.13.

The third most predictive feature is the DNA-binding specificity of Tcfe2a (or E2A), a mouse E-protein that is homologous to human E12/E47, known heterodimer partners of MyoD (Lassar et al., 1991b). Our classification models also selected other known cofactors of MyoD, namely Pknox1/Meis1, as features predictive of chromatin opening (Heidt et al., 2007; Knoepfler et al., 1999b, 1997; Sagerstrom, 2004). Thus, reprogrammed myoblast-specific DHS sites are largely characterized by a high
Figure 2.15: Coefficient paths for feature selection in classifying reprogrammed versus non-reprogrammed “myoblast-specific” DHS sites. The coefficient paths are shown for varying values of log($\lambda$) (at $\alpha = 0.12$ selected by cross-validation). The paths were obtained using the cv.glmnet package in R. The coefficient paths of TFs Myf6, Ascl2, Sp100 and Gmeb1 are highlighted. Note that Myf6 and Ascl2 have higher affinity sites in dDHSs+ (i.e. the ‘+1’ class) and thus have positive coefficients, while Sp100 and Gmeb1 have higher affinity sites in the ‘-1’ class, and thus have negative coefficients. The vertical lines at about $-3.5$ represent the log($\lambda$) value selected via cross-validation (at $\alpha = 0.12$) to minimize the error. The numbers at the top represent the number of the variables selected at specified log($\lambda$) values.

affinity for MyoD and its co-factors. In contrast, the non-reprogrammed myoblast-specific DHS sites have lower affinity for MyoD and its cofactors, but higher affinity for SAND-domain factors such as Gmeb1, Gmeb2, and Sp100. SAND-domain proteins, which are a class of DNA-binding proteins involved in chromatin-associated transcriptional regulation, were only minimally expressed in fibro-MyoD (Table 2.1).

Compared to the TF features, epigenetic features based on fibroblast histone modification data played only a minor role in the classification of myoblast-specific sites. The most predictive epigenetic mark, H3K27me3, was moderately enriched at reprogrammed sites compared to non-reprogrammed sites (Figure 2.17). This pre-existing relative abundance of H3K27me3 at MyoD-induced reprogrammed sites is in agreement with similar observations at genomic sites targeted by other pioneering factors,
Figure 2.16: Mean effect sizes of the features used in classification of sites that open up properly compared to the sites that remain closed during MyoD-induced transdifferentiation of fibroblasts. The means are taken over $N = 10$ runs using random subsamples of the data. The error bars capture 1 standard deviation. Variables with negative effect sizes are associated with the “negative” class (i.e. non-reprogrammed myoblast-specific DHS sites) and are shown in pink, whereas features that have positive weights, shown in dark red, correspond more to sites that are correctly reprogrammed (i.e. reprogrammed myoblast-specific DHS sites). Note that histone modification-specific features are obtained from multiple fibroblast cell lines.

such as Neurod1 during neuronal differentiation (Pataskar et al., 2016) and PU.1 during remodeling induced by histone deacetylase inhibitors (HDACi) (Frank et al., 2016). It is known that H3K27me3, when concomitantly observed with H3K4me1 or H3K4me2, marks poised enhancers (Heinz et al., 2015a). Because we also see pre-existing H3K4me1 and H3K4me2 enrichment at myoblast-specific reprogrammed chromatin sites, albeit non-differential from non-reprogrammed sites (Figure 2.17), this suggests that MyoD is potentially targeting poised enhancers during reprogramming.

2.3.4 Histone modification marks distinguish reprogrammed from non-reprogrammed fibroblast-specific DHS sites

We performed similar classification analyses for reprogrammed versus non-reprogrammed fibroblast-specific DHS sites (Figure 2.13B), using the same set of features described above. As for myoblast-specific sites, the GC-content was matched between reprogrammed and non-reprogrammed fibroblast-specific DHS sites (Figure
2.12), resulting in 531 sequences in each class. The classification accuracy achieved by the Random Forest classifier on fibroblast-specific sites was $68.87 \pm 3.19\%$ (Figure 2.13B). Similar to myoblast-specific sites, the classification accuracies and predictive features selected by Elastic Net models largely match those of Random Forest models (Table S6).

![Figure 2.17: Summary plots showing average ChIP-seq signal of some predictive histone modification marks obtained from the classification analyses between reprogrammed (dark red) versus non-reprogrammed (pink) myoblast-specific DHS sites, as well as reprogrammed (dark blue) versus non-reprogrammed (light blue) fibroblast-specific DHS sites. (A) Signal in NHDF cells (normal human diploid fibroblasts). (B) Signal in HSMM cells (human skeletal muscle myoblasts).](image)

Unlike myoblast-specific sites, the majority of the top predictors for fibroblast-specific site classification are histone marks. H3K4me2, H3K4me3, H3K27ac, H2az, and H3K9ac are enriched at non-reprogrammed vs. reprogrammed fibroblast-specific DHS sites in fibroblasts, suggesting that sites with these active marks are refractory to direct or indirect heterochromatization by MyoD. Importantly, this is true despite the fact that the pre-existing chromatin accessibility is not discriminatory between the sets of sites (Table S6, Figure 2.18). In addition, we found that high-affinity
sites for the activator proteins E2f2/E2f3 and in vivo binding of CTCF in fibroblast cells, were enriched in the non-reprogrammed compared to the reprogrammed sites, suggesting that binding of these factors could be involved in maintaining chromatin accessibility at some sites.

![Fibroblast-specific sites](image)

**Figure 2.18:** The DNase-seq signal in fibroblast cells is not significantly different between the reprogrammed and non-reprogrammed fibroblast-specific DHS sites. The sites included in this analysis are the ones selected after matching the GC content distributions of reprogrammed and non-reprogrammed sites.

2.3.5 Chromatin remodeling deficiencies are associated with nearby non-reprogrammed genes

As discussed in Chapter 1, a recent study suggested that the Six1 and Six4 transcription factors are critical for MyoD-induced reprogramming of mouse embryonic fibroblast cells (MEFs), since MyoD is completely unable to reprogram SIX1/4 mutant MEFs (Santolini et al., 2016). In this study of primary human cells, we find
that SIX1 is one of the myoblast-specific non-reprogrammed genes, while SIX4 is minimally expressed in all samples (Table 2.1). Looking more closely at Six1, our classification analysis did not find the specificity of this factor to be an important feature for distinguishing between reprogrammed and non-reprogrammed DHS sites (Table S6, Figure 2.19A), so it is unclear whether up-regulating SIX1 would lead to better chromatin reprogramming in our system. However, given the synergy between MyoD and Six1 at myogenic regulatory regions (Santolini et al., 2016), reprogramming of SIX1 could potentially improve reprogramming of other myogenic genes. Interestingly, our data also reveals two non-reprogrammed DHS sites in the SIX1 gene region (Figure 2.19B), which could contribute to the lack of reprogramming of this gene. Efficient chromatin reprogramming surrounding genes that encode myogenic regulatory factor (such as Six1) could improve the overall transdifferentiation efficiency.

![Figure 2.19](image.png)

**Figure 2.19:** (A) Average Six1-DNA binding specificity profiles are similar for reprogrammed and non-reprogrammed myoblast-specific DHS sites. (B) DNase-seq signal in the neighborhood of the SIX1 gene (80kb, coordinates chr14 : 61,075,000 – 61,155,000). Red rectangles highlight the non-reprogrammed myoblast-specific DHS sites in this region, located within the SIX1 gene (coordinates chr14 : 61113286 – 61113548) and about 6.6 kb upstream of the gene’s TSS (coordinates chr14 : 61122413 – 61123202). CRL = Chromatin Reprogramming Level score.

As was observed for SIX1, in order to determine whether global chromatin-remodeling deficiencies can explain the incomplete reprogramming observed at the gene level, we tested whether different types of chromatin remodeling events are en-
DHS sites from fibroblasts, fibro-MyoD and myoblasts were binarized with respect to state of chromatin accessibility (1 = open, 0 = closed). For example, the ‘011’ DHS pattern refers to sites that are closed in fibroblasts (0), open in fibro-MyoD (1), and open in myoblasts (1). Next, for each possible pattern we asked whether it is significantly enriched in the neighbourhood of genes in test sets compared to background sets. We analyzed four test sets: 1) myoblast-specific reprogrammed genes (Figure 2.20A), 2) myoblast-specific non-reprogrammed genes (Figure 2.20D), 3) fibroblast-specific reprogrammed genes (Figure 2.21A), and 4) fibroblast-specific non-reprogrammed genes (Figure 2.21C). For the myoblast-specific treatment sets, the background set was defined as all genes expressed in myoblasts at FPKM ≥ 5. Similarly, for the fibroblast-specific treatment sets, the background set was defined as all genes expressed in fibroblasts at FPKM ≥ 5. To perform enrichment analyses for DHS patterns we implemented the GSEA algorithm (Subramanian et al., 2005), sorting genes by the number of occurrences of a pattern in the genomic regions ±100kb of gene TSSs (see Methods). We used the Benjamini-Hochberg procedure to correct for multiple hypothesis testing, and characterized DHS patterns enriched at a false discovery rate (q-value) < 0.05.

Reprogrammed (‘011’) DHS sites are enriched around reprogrammed genes (Figure 2.20B, q-value: 0.004), which indicates that these sites may be bound by regulatory factors including MyoD (Figure 2.20C) required for turning on myoblast-specific genes. We also found that ‘011’ DHS sites are modestly enriched around non-reprogrammed genes (q-value: 0.047, Figure 2.20E), suggesting that some myogenic regulatory regions that open up in response to MyoD induction could still be missing critical trans factors necessary to activate gene expression. Classification of ‘011’ DHS sites nearby reprogrammed versus non-reprogrammed genes, using the same genetic and epigenetic features mentioned above, was not better than random,
Figure 2.20: Reprogrammed genes are associated with reprogrammed chromatin profiles. (A) Normalized gene expression levels for reprogrammed myoblast-specific genes. Inset shows a schematic representation of the gene expression level in the three cell lines. (B) GSEA analysis (Subramanian et al., 2005) showing reprogrammed ‘011’ DHS sites are significantly enriched in the ±100kb regions around TSSs of reprogrammed genes. Heatmap shows all genes expressed in myoblasts (FPKM ≥ 5) sorted by the number of ‘011’ DHS sites, with reprogrammed genes shown in red and all other genes shown in white. Based on the ordered list of genes, we computed an enrichment score (ES) at every position in the list, and used the maximum enrichment score (MES) as our test statistic. We assessed the significance of the test statistic by comparing it to a null distribution (grey histogram) computed from 1,000 random gene sets (see Methods). (C) MyoD binding is observed at reprogrammed ‘011’ DHS sites in both fibro-MyoD and myoblast cells.

(continued)
(D-F) Similar to A-C, but for non-programmed myoblast-specific genes. (G) Reprogrammed genes have significantly more ‘011’ DHS sites compared to non-reprogrammed genes (one-sided Mann-Whitney U test p-value: 6.52e-4). (H) Chromatin reprogramming level (CRL) is higher for DHS sites around reprogrammed compared to non-reprogrammed genes (one-sided Mann-Whitney U test p-value: 5.78e-5). (I) Reprogrammed genes have ‘011’ DHS sites closer to the TSS compared to non-reprogrammed genes (one-sided Mann Whitney U test p-value: 2e-3). For each gene, we considered the ‘011’ DHS site closest to its TSS, if present within 100kb of the TSS.

possibly also indicating missing non-DNA binding co-factors which our classification cannot identify. In addition, we found that ‘011’ sites around reprogrammed versus non-reprogrammed genes have similar levels of MyoD binding in fibro-MyoD (Figures 2.20C and 2.20F, left panels) and myoblast cells (Figures 2.20C and 2.20F, right panels).

Although both reprogrammed and non-reprogrammed gene sets show over-representation of ‘011’ DHS sites, there exist significant differences between the two sets themselves. First, reprogrammed vs. non-reprogrammed genes have significantly more ‘011’ DHS sites in their neighborhood (within 100kb; p-value: 6.52 × 10^-4; Figure 4G), which suggests that although non-reprogrammed genes have chromatin opening more than one would expect by chance (i.e. more than control gene sets), they may still lack sufficient chromatin reprogramming. Second, compared to non-reprogrammed genes, the reprogrammed genes displayed a significantly higher degree of chromatin opening (p-value: 5.78 × 10^-5, Figure 4H). Third, chromatin reprogramming occurred closer to the TSS for reprogrammed myoblast-specific genes (p-value: 2 × 10^-3, Figure 2.20I).

We observed similar results for fibroblast-specific genes (Figure 2.21). Both the reprogrammed and non-reprogrammed gene sets are enriched for reprogrammed (‘100’) and non-reprogrammed (‘110’) DHS sites (q-values < 0.006; Methods). The abundance of reprogrammed ‘100’ sites is not different in the two sets of genes (Figure
Figure 2.21: Chromatin remodeling profile around fibroblast-specific reprogrammed or non-reprogrammed genes. (A) Normalized gene expression levels for reprogrammed fibroblast-specific genes. Inset shows a schematic representation of the gene expression level in the three cell lines. (B) GSEA analysis showing that reprogrammed ‘100’ DHS sites are significantly enriched in the ±100kb regions around TSSs of reprogrammed genes. Heatmap shows all genes expressed in fibroblasts (FPKM $\geq 5$) sorted by the number of ‘100’ DHS sites, with reprogrammed genes shown in blue and all other genes shown in white. Based on the ordered list of genes, we computed an enrichment score (ES) at every position in the list, and used the maximum enrichment score (MES) as our test statistic. We assessed the significance of the test statistic by comparing it to a null distribution (gray histogram) computed from 1,000 random gene sets (see Methods). (C,D) Similar to A-B, but for non-programmed fibroblast-specific genes. (E) Reprogrammed and non-reprogrammed genes have similar numbers of ‘100’ DHS sites (Mann-Whitney U test p-value: 0.47). (F) Chromatin reprogramming level (CRL) is higher for DHS sites around reprogrammed compared to non-reprogrammed genes (one-sided Mann-Whitney U test p-value: 5.6e-5). (G) Non-reprogrammed genes have non-reprogrammed ‘110’ DHS sites closer to their TSS compared to the reprogrammed genes (one-sided Mann-Whitney U test p-value: 6e-3).
2.21E), but the CRL scores (i.e. levels of chromatin reprogramming) are significantly higher around reprogrammed genes (one-sided Mann-Whitney U test p-value: 0.005, Figure 2.21F). Interestingly, non-reprogrammed (‘110’) chromatin sites are significantly closer to the TSSs of non-reprogrammed versus reprogrammed fibroblast-specific genes (one-sided Mann-Whitney U test p-value: 0.006, Figure 2.21G), indicative of an association between gene expression and promoter-region accessibility.

We also assessed whether promoter-proximal chromatin remodeling, or lack thereof, explains the variation in gene expression reprogramming (Figure 2.22). For this analysis, we focused on promoter-proximal (TSS ±5kb or TSS ±2kb) myoblast-specific and fibroblast-specific DHS sites (Methods). The extent of gene expression reprogramming level (GRL) was computed similarly to CRL (Methods). Log-transformed values were used to compute both CRLs and GRLs, in order to capture even small and moderate changes in expression level and accessibility. We observed strong correlation between the degree of chromatin reprogramming and the degree to which gene-expression is reprogrammed (PCC: 0.48, p=2.7e-16 for myoblast-specific genes; PCC: 0.41, p = 0.0044 for fibroblast-specific genes at TSS ±5kb, Figure 2.22A). The observed correlation was stronger for TSS ±2kb regions (PCC: 0.6, p = 6.4e-18 for myoblast-specific genes; PCC:0.48, p = 0.0079 for fibroblast-specific genes, Figure 2.22B), suggesting that chromatin accessibility remodeling in promoter regions partially explains the variation in reprogramming of gene expression during MyoD-induced transdifferentiation. To our knowledge, this is the first genome-wide study comparing gene expression and chromatin accessibility profiles of myoblasts versus MyoD-transdifferentiated cells.

2.4 Discussion

The MyoD-induced transdifferentiation model system can be used to understand how master regulatory TFs can transform cell fates by inducing global changes in chro-
Figure 2.22: Correlation between Chromatin Reprogramming Level (CRL) of promoter-proximal myoblast-specific (left) or fibroblast-specific (right) DHS sites and the Gene Reprogramming Level (GRL) of the corresponding genes. Like CRL, the GRL quantifies the extent that the gene is reprogrammed in its expression level. Both CRL and GRL scores are computed after the normalized DNase-seq read counts, or the expression values (TPMs), respectively, are log-transformed in order to capture moderate changes in accessibility or expression dynamics (Methods). For these plots, only the myoblast-specific genes (left panels, in red) or fibroblast-specific genes (right panels, in blue) are considered. Also, for every myoblast- or fibroblast-specific gene, only myoblast- or fibroblast-specific DHS sites in TSS ±5kb (A) or TSS ±2kb (B) are considered, respectively.
matin and gene expression profiles. However, it is unknown at a genome-wide scale how much transdifferentiated cells quantitatively differ from both the starting cells and the target cells. Here, we use MyoD-induced transdifferentiation of primary human skin fibroblasts to the myogenic cell fate as a model system to develop a general approach for investigating the extent of chromatin- and gene-level reprogramming induced by forced overexpression of TF master regulators. In our system, we find that while many of the early muscle marker genes are reprogrammed, global gene expression and accessibility changes are still incomplete when compared to myoblasts, the early myogenic determination stage. Our findings suggest how incomplete transdifferentiation can be quantified, characterized, and potentially improved.

This study is the first to quantify the effects of MyoD on chromatin accessibility in a transdifferentiation system. We are using primary human cells, which are known to be more challenging to reprogram with MyoD compared to murine and/or immortalized cell lines (Cao et al., 2006, 2010; Chaouch et al., 2009; Comai and Tajbakhsh, 2014; Davis et al., 1987; Fong and Tapscott, 2013; Lattanzi et al., 1998; Vogel, 2008; Yutzey et al., 1990). In this system, we found that MyoD targets both closed and accessible chromatin, indicating its mixed role as a pioneer factor that opens chromatin, and a TF that binds already accessible DNA. This result is similar to those described for other pioneer factors, such as neurogenic factor Ascl1 (Wapinski et al., 2013) or pluripotency factors Oct4, Sox2, Klf4. Similar to previous reports (Cao et al., 2010; Yao et al., 2013), we observed a large overlap between the MyoD-bound sites in fibro-MyoD and myoblasts (Figure 2.9). However, MyoD appears to be limited in its capacity to bind to all of its potential targets in fibro-MyoD cells, and thereby binds and opens only a fraction of myoblast-specific sites, consistent with previous studies suggesting that MyoD's role as a pioneer factor is limited (Comai and Tajbakhsh, 2014; Fong and Tapscott, 2013). This may partially explain the incomplete transdifferentiation in our system. Importantly, the MyoD
binding signal in myoblasts at myoblast-specific chromatin sites that remained closed in fibro-MyoD cells, compared to the sites that did become accessible, MyoD was bound more weakly or not at all, even in myoblasts.

In addition to the reduced MyoD binding at non-reprogrammed chromatin sites in fibro-MyoD cells, we found that these sites displayed an enrichment of binding motifs for SAND-domain factors, a family that includes Aire, Gmeb1 and Ski (Figure 2.13A). Since SAND-domain factors were minimally expressed in fibro-MyoD cells, this finding points to an attractive possibility for improving reprogramming efficiency. Aire is known to bind unmethylated H3K4 molecules, and it has been hypothesized to play a role in transcriptional regulation via chromatin remodeling, although the precise molecular mechanisms are not well understood. Gmeb1 is known to interact with CBP (CREB Binding Protein), which can open closed chromatin sites (Chen et al., 2002). Ski can convert quail non-muscle cells into muscle cells (Colmenares and Stavnezer, 1989; Nyman et al., 2010). Therefore, we hypothesize that increasing expression of these SAND-domain proteins could help open a subset of myoblast-specific sites not reprogrammed by MyoD alone. Our quantitative approach for analyzing chromatin changes during transdifferentiation, based on classifying genomic regions with different chromatin reprogramming levels (CRLs) also suggests that increasing the nuclear concentration of MyoD or its co-factors could lead to higher CRL scores for myoblast-specific DHS sites. This is consistent with the finding that increasing nuclear localization of MyoD improves reprogramming efficiency (Jeong et al., 2011) (also see Appendix A). However, increasing MyoD to levels that are not physiologically normal could also lead to off-target binding events and improper reprogramming, as discussed below.

Chromatin-remodeling factors are another class of proteins that may improve cellular reprogramming. Our classification analyses indicate that the pre-existing repressive mark H3K27me3 is moderately predictive of chromatin opening at myoblast-
specific DHS sites, while pre-existing active marks (H3K4me1, H3K4me2) are similarly enriched at both reprogrammed and non-reprogrammed sites (Figure 2.17). This suggests that inducing a “poised” chromatin state (Heinz et al., 2015a) at myoblast-specific DHS sites prior to MyoD overexpression, by targeted deposition of repressive mark H3K27me3, could lead to increased MyoD binding and improved reprogramming efficiency. Our observations are in agreement with the pre-existing enrichment of H3K27me3 at sites bound by NeuroD1 during induced neuronal differentiation (Pataskar et al., 2016), and by pioneer factor PU.1 during HDACi-induced remodeling (Frank et al., 2016). Similar to H3K27me3, another repressive mark, H3K9me3, was also found to be enriched along with active marks H3K27ac and H3K4me1, at sites bound by pioneer factor Ascl1 during transdifferentiation of mouse embryonic fibroblasts to neuronal cells (Wapinski et al., 2013). Interestingly, during initial stages of pluripotency reprogramming of fibroblasts, large H3K9me3-marked chromatin domains were reported to prevent binding of pioneer factors Oct4, Sox2, and Klf4, diminishing overall conversion efficiency. Our study did not find any discriminatory capacity of H3K9me3.

We also observed enrichment of active chromatin marks at fibroblast-specific sites that do not close down in fibro-MyoD cells. One of the reasons could be that repressive chromatin modifiers such as UHRF1 (Bostick et al., 2007) are significantly underexpressed in normal and transdifferentiated fibroblasts compared to myoblasts (Table 2.1). If this relationship is causal, then erasing active marks in some non-reprogrammed fibroblast-specific DHS sites or depositing repressive DNA methylation marks could facilitate closing down of regulatory sites to shut off fibroblast-specific genes. Our findings highlight the need for genome-wide reprogramming studies to explore the concerted effects of other TF master regulators and chromatin remodelers in transdifferentiation. One such study observed that expression of BAF60C, a component of the SWI/SNF chromatin remodeling complex, induces
myogenic priming in MyoD-mediated reprogramming of human ES cells (Albini et al., 2013). We note that the enrichment of active chromatin marks at sites that fail to close down during transdifferentiation could also be due to the fact that fibroblast-specific TFs, which might target these regions, are encoded by some of the genes not appropriately repressed. While possible, this hypothesis is not supported by our classification analyses (Figure 2.13B, Figure 2.14B).

Our DHS pattern analyses also revealed chromatin opening at off-target sites, which resulted in a total of 6,792 fibro-MyoD-specific ‘010’ DHS sites. Only about 6% of these sites show MyoD binding in fibro-MyoD, suggesting that off-target chromatin remodeling is mostly indirect. These off-target DHS sites are scattered randomly throughout the genome. In addition, they do not appear to have a significant effect on gene expression, as we only found seven genes that are ‘improperly reprogrammed’ in the sense that they are significantly upregulated in fibro-MyoD compared to both fibroblast and myoblast cells. These misprogrammed DHS sites might be attributable to supraphysiological levels of MyoD.

Our genome-wide analyses show that although many fibroblast-specific DHS sites are successfully closed, only a small fraction of fibroblast-specific genes are repressed. It is not entirely clear why this is the case, but we offer a few possibilities. First, we note that, despite this apparent discrepancy, there is a good correlation between the overall chromatin reprogramming level in promoters of fibroblast-specific genes and the extent to which these genes are reprogrammed in terms of expression (R > 0.4; Figure 2.22). Thus, the additional decommissioned DHS sites beyond these promoters may not have strong effects on gene expression. Second, the number of fibroblast-specific DHS sites is about 10 times higher than the number of fibroblast-specific genes. If multiple DHS sites are used to redundantly upregulate a gene, then closing only a fraction of these sites may not be sufficient to reverse the upregulation. To more fully investigate this possibility, we would need precise information regard-
ing which fibroblast-specific DHS sites are used to regulate each fibroblast-specific gene, taking into account both proximal and distal DHS sites. Such chromatin conformation information (e.g., Hi-C, etc.), unfortunately, is not currently available and could be the focus of future studies. Third, it is possible that some fibroblast-specific genes are regulated not through fibroblast-specific DHS sites, but through ubiquitous DHS sites bound by fibroblast-specific TFs, which are still be produced. Again, in order to investigate this possibility, we would need information regarding which DHS regions (proximal or distal) are used to regulate each fibroblast-specific gene.

In summary, our study revealed a continuum of chromatin remodeling changes genome-wide and showed that chromatin remodeling deficiencies are associated with global transcriptional reprogramming bottlenecks during MyoD-induced transdifferentiation. We identified potential explanations for the incomplete reprogramming at the chromatin level, and suggest mechanisms to improve the process. Our approach for genome-wide analysis of the efficiency of cellular transdifferentiation driven by a TF master regulator can be applied to any transdifferentiation system, and will likely be particularly useful for characterizing transdifferentiation systems with relatively low epigenetic conversion efficiency.
Cross cell-type predictive modeling of gene expression using chromatin accessibility and transcription factor expression profiles

3.1 Background

Eukaryotic gene expression is dictated by a combination of regulatory transcription factors (TFs) and epigenetic chromatin elements (Vaquerizas et al., 2009). This is true not only during normal cell growth and differentiation, and also during cellular reprogramming. As shown in Chapter 2, during transdifferentiation of fibroblasts to the myogenic lineage, the transcriptional and epigenetic profiles of the reprogrammed/“converted” cells can be very different from those of the target cells. Our study showed that chromatin remodeling deficiencies, including those involving genes that encode key regulatory factors (e.g. Six1), represented an obstacle to full conversion efficiency. Thus, as a result of the observed chromatin remodeling deficiencies, critical regulatory factors remained under-expressed, and so did their target genes. Motivated by these findings, I aimed to understand whether the transcriptional profiles of the converted cells could be improved by targeted reprogramming of specific
regulatory TFs and/or specific chromatin sites. As such, under the supervision of my advisor, I developed a computational model that uses the expression levels of regulatory TFs and the chromatin accessibility profile of the cells to predict target gene expression. The main goal of our model is to allow users to generate hypotheses regarding which regulatory TFs expression level and/or which chromatin sites should be manipulated experimentally in order to improve reprogramming efficiency.

Models to predict gene expression levels have been developed previously. However, such models are not generally applicable to reprogramming studies. Many gene expression prediction models make use of large collections of cell/tissue-specific data sets, such as ChIP-seq data for transcription factors (Ouyang et al., 2009) or histone modifications marks (Cao et al., 2017; Cheng et al., 2011; Singh et al., 2016; Wang et al., 2008), or a combination of these and other data types (Althammer et al., 2012; Angelini and Costa, 2014; McLeay et al., 2012; Park and Nakai, 2011), which are not generally available in cellular reprogramming studies. In addition, many of these models are specific to a singular cell line or tissue type (Althammer et al., 2012; Cheng et al., 2011; Singh et al., 2016; Wang et al., 2008), and are therefore not generalizable across cell types. Such models are not applicable to reprogramming studies, where the cells might be in intermediary stages between the starting and the target cells. Cross-cell type models, on the other hand, can leverage regulatory information across cell types (Cao et al., 2017; Marstrand and Storey, 2014; Natarajan et al., 2012), and as such, can highlight the role of new or absent features (like cis-regulatory elements or TFs) during reprogramming or disease modeling studies.

JEME, a recently developed enhancer-target prediction model (Cao et al., 2017), appears to the most predictive cross-cell type gene expression prediction model. However, this model also requires a large collection of enhancer-related epigenetic modifications (such as H3K4me1, H3K27ac, H3K27me3) to be used as model features, which limits its application in reprogramming studies for which such data is
not available. Enhancer-specific features can indeed have high predictive power, as they are known to correlate with gene expression levels (although these correlations do not always indicate causality). However, chromatin features do not incorporate information about the regulatory TFs and their expression levels, which can be very informative for accurate gene expression prediction models, as they capture our understanding of how genes are regulated. Therefore, we reasoned that a simple yet powerful cross-cell type model for predicting gene expression should be able to learn regulatory information from chromatin accessibility and the expression levels of regulatory TFs across the cell types.

Here, a cross-cell type gene expression prediction framework, named CPGex, that leverages chromatin accessibility and the expression levels of predicted regulatory factors to model the expression of target genes is presented. Specifically, simple, target gene-specific artificial neural networks are trained to model non-linear combinatorial effects of chromatin accessibility and regulatory TFs towards the expression of the gene(s) they regulate. We show that despite limited use of input data, CPGex’s performance is at least as good as that of JEME (Cao et al., 2017). In addition, based on the importance scores of the features using in our trained models, CPGex can be used to generate testable hypotheses on whether specific changes in the accessibility of chromatin sites or the expression level of specific regulatory TFs could lead to improved reprogramming of gene expression in cellular transdifferentiation studies.

3.2 Data and methods

3.2.1 Collection and processing of DNase-seq and RNase-seq data

We used DNase-seq and RNA-seq data for 127 human cell types, tissue types, and cell lines (111 from the Roadmap Epigenomics Project and 16 from the ENCODE project), downloaded from the Roadmap Epigenomics website (Roadmap Epigenomics et al., 2015). We chose these data sets because 1) they were uniformly
processed, and 2) they were also used in (Cao et al., 2017), which allowed us to perform a direct comparison between our models and JEME (Cao et al., 2017). As in (Cao et al., 2017), we used all 127 samples in our study, including those for which the DNase-seq or RNA-seq data were imputed from other data types (Ernst and Kellis, 2015). Given that we use the imputed data, we performed a quick test to assess the quality of the RNA-seq data by analyzing the expression profiles of key tissue-specific marker genes (such as \textit{SOX2}, \textit{NANOG} required for stem-cell maintenance, \textit{MYOD1} and \textit{MYOG} involved in myogenic differentiation, and \textit{ASCL1} and \textit{NEUROD2} key to neuronal commitment), and we observed that these genes are indeed most highly expressed in their corresponding tissue types (not shown).

ENCODE enhancers (from states 6, 7 and 12 in the core 15-state ChromHMM model) were obtained from (Cao et al., 2017) for each sample. These enhancer regions were merged across the samples using the default \texttt{mergeBedTools d 0} option, yielding a total of 249,079 regions with an average length of 1,048bp and a median length of 1,000bp. The DNase-seq signal at these sites (henceforth also referred to as DNase I hypersensitive sites, or DHS sites) was considered in our analyses.

In general, all DHS sites within \(\pm 150\) kb of a transcription start site (TSS) were considered. If this region overlapped a TAD domain boundary, then only the region up to the TAD boundary was considered in order to select the DHS sites to be used as features in the model (see below). In order to restrict the number of input features, only those DHS sites most correlated with the gene expression profiles were used. This ranged in number from just top 6 DHS sites for the joint DHS-TF model (see below) to top 10 for the DHS-only model.

For processing of the RNA-seq data, the RefSeq gtf gene annotation file (hg19 version) was downloaded from the https://genome.ucsc.edu/cgi-bin/hgTables. For each gene, only reads mapping to the \(\pm 500\)bp region around each TSS was used as a
proxy for the expression level, as done in (Cao et al., 2017). For genes with multiple TSSs, one TSS was selected at random. This resulted in a total of 27,544 genes considered in our analyses.

Similar to the Roadmap DNA-seq data, RNA-seq signal data for all 127 samples were downloaded from the Roadmap Epigenomic website. The RPKM bigwig signal file for each sample was converted to bedgraph file format, and mapped onto each TSS ±500bp region as a proxy of raw expression value of the gene.

### 3.2.2 TAD boundaries

TAD boundaries for 21 cell types were obtained from ((Schmitt et al., 2016), Table S3). All TAD boundaries are merged using the `mergeBedTools` (using default -d 0) option. Only those merged TAD boundaries present in at least 20% of the cell types (i.e. 4 or more out of 21 cell types) were considered as confident TAD domain boundaries consistent across the cell types. The center of each TAD boundary marks the edge of corresponding TAD domain. From the resulting set of TAD domains, all TADs containing the centromere and telomere regions were filtered out. This resulted in a total of 3,081 TAD domains genome-wide. TAD domain boundaries were used to mark the boundaries in the cis-TSS regions. Specifically, we only considered DHS sites overlapping the range: \([\text{max}(\text{TAD lower boundary, TSS} - 150\text{kb}), \text{min}(\text{TAD upper boundary, TSS} + 150\text{kb})])\].

### 3.2.3 List of regulatory TFs and their corresponding target genes (TF-TG list)

The human “Transcription factor Target gene” (TF-TG) gene regulatory network was obtained from previous studies (Cahan et al., 2014; Radley et al., 2017). All predicted regulatory TFs with TF-TG attributed confidence scores (z-scores) of 6 or higher were considered in our study. This list was further filtered to only select those TFs that were also at least minimally correlated (Pearson Correlation Coefficient...
3.2.4 Training artificial neural network models

We used the DNA accessibility information and the expression levels of regulatory TFs to derive features that were used in the input layer of gene-specific neural network models. For each target gene, we build three different models using either: (1) DHS sites only, i.e. only the normalized chromatin accessibility signal in cis-TSS chromatin sites (e.g. TSS ± 150kb) was used to train and predict the expression of the target gene; (2) regulatory TFs only, i.e. only the normalized expression of predicted regulatory TFs for the target gene were used in the model; and (3) DHS sites with TFs, i.e. both normalized chromatin accessibility and TF expression features were used in the model. These models are henceforth referred to as “DHS-sites-only”, “TFs-only” and “DHSs-plus-TFs” models respectively. For all these types of models, both chromatin accessibility and gene expression profiles were log-transformed using the functions before training the model. Figure 3.1 shows a schematic representation of the processing steps, and Figure 1.4 in Chapter 1 shows the modeling framework for each of the three types of models mentioned above.

Similarly to the strategy used in (Cao et al., 2017) to train JEME models, we implemented a leave-one-group-out procedure to train and test the model performance. The 127 Roadmap Epigenomics samples has been grouped into 19 categories, such as “ES-derived”, “Blood and T-cell”, “Muscle”, based on their lineages by Roadmap Consortium (see this Roadmap metadata spreadsheet for details on the samples and their groups). We used one of these categories/groups, termed “ENCODE2012”, as a validation group in order to select best choices of hyperparameters. This choice was motivated by the fact that the ENCODE2012 group represents a collection of cell types from various lineages. Out of the remaining 18 groups, we used one group for testing and the other 17 groups to train the model, thus performing 18 rounds of
training (which includes hyperparameter optimization on the validation group) and testing.

For each training-validation-test split, the training design matrix and corresponding response (or “expression”) vector were normalized to the [0, 1] scale by dividing by the maximum values in the respective arrays. The same normalizing factors were used to normalize the corresponding test design matrix and response vector. The validation set was used to select best set of hyperparameters for the training set (see below). In our implementation, the numbers of units for the hidden layers were selected from the range [$\frac{1}{3}$ size of preceding layer, $\frac{1}{3}$ size of preceding layer], while the parameters were selected from the set $5 \times 10^{-i}, i \in \{1, \ldots, 7\}$. Once the best hyperparameters were obtained, the training and validation sets were used together, with the selected hyperparameter values, before testing on the held-out test set.

For the comparison against the JEME models, the generalization error was computed as absolute median percentage error in the test (Cao et al., 2017), i.e. model performance was assessed as $\text{median}\{\frac{|y_s - \hat{y}_s|}{y_s}\}$, where $s$ is a test cell type/sample.

### 3.2.5 Availability

We have made our model available to download in the Github repository - https://github.com/dineshmdh/predicting_gene_expression. Appendix B has some basic outline and examples showing how the software can be downloaded and run.

### 3.2.6 Feature importance scores

The importance scores corresponding to a particular DHS site or TF is measured as the absolute change in error obtained when the specific feature attains a background mutated level of measure in the model. Since the data is normalized in [0 – 1] scale, for every feature, we set its background mutated level to an array of 0 across the design matrix. Importance measures thus obtained are pooled across the test groups.
Figure 3.1: Data processing schematic for predicting gene expression using artificial neural network. (Details in Methods.)

in Roadmap samples to get a distribution of feature importance measures as shown in Figure 3.4.
3.3 Results

3.3.1 Models that combine chromatin accessibility and regulatory TF expression levels can accurately predict the expression of target genes

We assessed the performance of our gene-specific neural network models using a collection of 261 genes that were either reprogrammed (161 genes) or non-reprogrammed (100 genes) during MyoD-induced cellular reprogramming of primary human fibroblasts, according to our previous work (Manandhar et al., 2017). For each gene, we trained neural network models (Figure 1.4 in Chapter 1) using the three modes of training described in Data and Methods (Figure 3.1): DHS-sites-only, TFs-only or the joint DHS-sites-plus-TFs. For each training mode, we compared the performance of models that use “true” features versus models that uses the same number of features but randomly sampled from the pool of (i) DHS sites genome-wide, or (ii) TFs, depending on the training mode. The “true” features represent the collection of either (i) ENCODE-regulatory sites in the TSS ±150kb region that are most correlated with the expression of the target gene (see Data and Methods section 3.2), or (ii) TFs that are most likely to be regulating the target gene, based on their expression profile and their TF-TG (transcription factor target gene) association score obtained from (Cahan et al., 2014; Radley et al., 2017) (see Data and Methods section 3.2).

In order to constrain the model complexity, we limited the number of features to the top six regulatory sites or ten predicted regulatory TF for all genes and modes of training.

For each target gene and each model, we assessed model performance using the Pearson Correlation Coefficient (PCC) between the true and the predicted expression levels of the target genes over the held-out test samples (see Data and Methods section 3.2). As shown in Figure 3.2, we found that the performance of the joint model (DHSs-plus-TFs) is significantly better than the performance of models using either
Figure 3.2: Models that combine chromatin accessibility and regulatory TF expression levels can accurately predict the expression of target genes. Boxplots show the Pearson Correlation Coefficients (PCCs) between the true and the predicted gene expression levels for three different training modes, across 261 genes (see text for details). The PCC values are highest for models that combine DNA accessibility and TF expression level information (i.e. the joint models). For all three modes of binding, the PCC values are significantly higher for models based on the “true” features, i.e. DHS and regulatory TFs relevant to the target gene (orange boxes) compared to models that use random DHS sites or TFs (green boxes). One-side Mann Whitney U-test p-value for all pairwise comparisons between the distribution of PCCs is $< e^{-60}$.

Only DNA accessibility information (one-sided Mann Whitney U-test p-value: 0.0) or only regulatory TF information (p-value 9.27e-52), suggesting that chromatin sites and regulatory TFs jointly confer larger effects on the gene expression than either of these features by themselves. In addition, we found that our models perform significantly better when using the true input features as opposed to randomly samples features (one-sided Mann Whitney U-test p-value: 3.44e-151 for DHSs-only model, 2.66e-105 for TFs-only model, and 6.16e-117 for the joint model). This indicates that our models are able to distinguish real from random biological input signal.
3.3.2 Our joint DHSs-plus-TFs models compare favorable to JEME

We also compared our model performance against that of the JEME model (Cao et al., 2017). To ensure a fair and direct comparison, we processed our data similarly to the JEME study (see Data and Methods section 3.2). This includes incorporating a leave-one-group-out learning framework as in (Cao et al., 2017), in order to minimize the overlap between training and test samples. In our design, for any held-out-test group, we used “ENCODE2012” as the validation group, and we trained on the remaining 17 groups of samples. Our choice of validation set was motivated by the fact that the “ENCODE2012” group comprises of samples from diverse lineages, such as lung, blood, muscle, and skin cell types, and therefore, is an effective validation group.

Figure 3.3 shows the distribution of medium percentage errors for each of the held-out test groups across the models learnt for the target genes. Specifically, for any target gene, if a held-out test group has cell types $c_i, i \in \{1, \ldots, k\}$, then median of $\frac{|y_i - \hat{y}_i|}{y_i}$, for all $c_i$, is reported for the test group. The same procedure was used to assess models performance in (Cao et al., 2017). The mean of the medium test percentage errors across all samples and genes reported in (Cao et al., 2017) is approximately (Figure 3.3, dashed line). Interestingly, our model performance is not significantly different from that of the JEME model (which to our knowledge is the best cross-cell type model to date). This is significant considering that JEME requires a collection of H3K4me1, H3K27ac, and H3K27me3 histone modification data sets across all the samples, in addition to the chromatin accessibility and gene expression data used in our models.
Figure 3.3: Our DHSs-plus-TFs models compare favorably to JEME, the best model to date (Cao et al., 2017). Results are shown for each of the 18 test categories, where each category comprises of samples with similar gene expression profiles (see Data and Methods section 3.2, and (Cao et al., 2017)). For each category, DHSs-plus-TFs models were trained for all 261 genes in our study (i.e. the genes that were either reprogrammed or non-reprogrammed during myogenic transdifferentiation according to our previous work 2; see Data and Methods section 3.2). In order to compare our models versus JEME, for each gene we computed the absolute median percentage error over the samples in each held-out test set, as in (Cao et al., 2017). Briefly, if a held-out test group has cell types \(c_i, i \in \{1, \ldots, k\}\), then the median of \(\left|\frac{y_i - \hat{y}_i}{y_i}\right|\) for all \(c_i\), is reported. The density plots in the figure show the models performance (i.e. the absolute median percentage error) over all 261 test genes. The dashed line represents the mean of group-specific median test percentage errors, as obtained by the JEME model for the same data (see (Cao et al., 2017), Supplemental Figure 6).

3.3.3 Feature importance measurements yield regulatory information

Using trained models, relative contributions of the DHS sites or TFs in the model can be computed by measuring the effect that feature mutation has on the overall error for the model (see Methods). Figure 3.4 shows the importance measures computed for features corresponding to a gene randomly selected from our previous study 2,
**Figure 3.4:** Importance scores computed for the features used in predicting \textit{TNNC2} under joint (or DHSs-plus-TFs) model. Importance scores were obtained by imputing feature values and computing change in error (see Methods). Heatmap shows the log transformed DNase-seq or RNA-seq signal across all Roadmap samples for the corresponding feature. For each feature, the suffix shows the Pearson correlation coefficient between the feature signal and \textit{TNNC2} gene expression (shown on the right).

\textit{TNNC2.} \textit{TNNC2} codes for a terminal muscle-specific protein 2, and is expressed high in muscle and heart tissue samples in the Roadmap samples as well. We trained a “DHSs-plus-TFs” model across the held-out Roadmap test groups, and measured the feature importance for each test group. Figure 3.4 shows the distribution of importance scores thus generated. As shown, TFs contribute more than DHS sites
in the overall model performance. More interestingly, the top three TFs (based on the median values), namely *MYF6*, *SIX1* and *MYOG*, are all known to be important myogenic regulatory factors (Manandhar et al., 2017; Santolini et al., 2016). Because *SIX1*, like *TNNC2*, remains under-expressed during MyoD-induced transdifferentiation of fibroblasts (Manandhar et al., 2017), it is possible that *SIX1* expression is critical for *TNNC2* upregulation.

### 3.4 Discussion

In this study, we present a new gene expression prediction framework, CPGeX based on artificial neural networks. Importantly, the goal of our framework is not to predict gene expression data de novo in cell types without any available gene expression (as our models use the expression levels of regulatory TFs as input features). Instead, we aim to use our models in cellular reprogramming studies to identify chromatin sites and/or TFs that are important for predicting the expression of target genes, with the goal of generating testable hypotheses for improving reprogramming efficiency. To our knowledge, our DHSs-plus-TFs models are the first cross-cell type models that incorporates both chromatin accessibility and the expression levels of regulatory TFs in order to predict expression of target genes. Our models are based on the hypothesis that joint regulation of relevant chromatin sites and regulatory factors affects the expression of a gene. In addition, the fact that our DHSs-plus-TFs models compare favorable to JEME models (Cao et al., 2017) shows that non-linear combinatorial interactions of features pertaining to just DNA accessibility and TF expression levels can predict the expression of target genes as well as models that use a combination of histone modification and chromatin accessibility data sets. As such, our model represents a novel computational framework for gene expression studies, allowing users to assess, in silico, how the expression of a target gene of interest would change if we manipulate the accessibility of specific chromatin sites around the gene, and/or the
expression level of specific regulatory TFs. This offers an ability to generate testable hypotheses in cellular reprogramming and disease modeling studies, where it is of interest to manipulate the expression of certain genes is predictable ways. Our model is available to download at https://github.com/dineshmdh/predicting_gene_expression. Instructions for installation and running the software is included in Appendix B.

There exist models like CellNet (Cahan et al., 2014; Radley et al., 2017) and Morgridy (Rackham et al., 2016) that can also provide information on the overall significance of TFs in converting cells types. However, our model is different from previous work in a couple of ways: First, previous models are based on global transcriptional network analyses and, hence, they do not model the interactions between TFs and regulatory chromatin sites. Second, while previous models can yield a global ranked list of TFs for improved cellular conversion efficiency, our DHSs-plus-TFs model is gene-centric and thus provides a “zoomed-in” capacity to manipulate, in silico, the regulatory factors and/or chromatin sites for targeted regulation of a single gene or a subset of genes.

Despite the advantages mentioned above, our model has its limitations. It is known that regulatory elements control the expression of their target genes by long-range interactions that may be specific to particular cell types. Because of the lack of 3-D chromatin interaction data across the samples, we do not consider cell-type specific regulation of chromatin sites. Another limitation is that the TF-TG interaction list obtained from CellNet (Cahan et al., 2014; Radley et al., 2017), which we use to select regulatory factors TFs, is obtained computationally. The sensitivity of the algorithm implemented to generate the TF-TG network in (Radley et al., 2017) is contingent on the availability of a multitude of transcriptomic perturbation assays (Faith et al., 2007), which is a limitation for human samples. Therefore, many potentially regulatory factors do not end up in the model because the sensitivity of TF-TG predictions is not high. Additionally, not all predicted TF-TG interactions
are real. Given all this, it was not surprising to us that even sets of randomly chosen TFs can, in some cases, perform well (albeit still worse than when using the “real” regulatory TFs based on the CellNet predictions) (Figure 3.2). Clearly, we expect our model performance to improve with better TF-TG interaction data. A third limitation is that, as with long-range 3D chromatin interactions, we also ignore modes of gene regulation involving micro-RNAs and long non-coding RNAs. While including these interactions could enhance our performance, this is beyond the scope of the current study, considering that we are primarily interested in modeling gene expression using chromatin accessibility and TFs. This is notwithstanding the fact that large collection of such potentially regulatory data sets or histone modification data are not readily available for gene expression studies on novel studies, where our model can be useful.
Comparative genome-wide myogenic differentiation analyses between CRISPR/Cas9-mediated endogenous and lentiviral transgene MyoD expression

In this chapter, genome-wide chromatin and gene expression changes during myogenic reprogramming is studied in two systems: (i) CRISPR/Cas9-mediated endogenous MyoD expression, and (ii) traditional lentiviral MyoD construct mediated overexpression. Both systems use human iPS cells (hiPSCs) as the starting cell line, providing a platform for comparative analyses between the two. This work is not yet published.

Similar to Chapter 2, this is a collaborative project between the members of Gersbach Laboratory, Crawford Laboratory, Gordon Laboratory and Reddy Laboratory, (Duke University). Ami Kabadi from Gersbach Laboratory is the main contributor in designing and performing the reprogramming assays, and is the lead author in the manuscript-in-preparation. Lingyun Song in Crawford laboratory has generated next generation sequencing (NGS) data on the cell lines. I have performed all com-
putational analyses discussed in this chapter using the NGS data similar to Chapter 2, and am currently second author in the manuscript-in-preparation. The details of all author contributions to the project and their respective laboratory affiliations are presented in the Methods section below.

4.1 Background

In Chapter 2, we looked into genome-wide changes induced by lentiviral-mediated overexpression of exogenous MyoD. Lentiviral transgene expression is an effective means of inducing cellular reprogramming because: (i) the method is effective for both dividing as well as non-dividing cells, and (ii) the transgene expression is not transient since the transgene randomly integrates into the host genome ((Sakuma et al., 2012)). However, as opposed to transgene MyoD, endogenous MyoD expression provides an alternate and potentially advantageous myogenic reprogramming system, where MyoD expression is induced in its local chromosomal context. It has been shown that endogenous gene expression could facilitate better chromatin modeling ((Gao et al., 2013)) and ensure regulatory process similar to normal differentiation ((Rebar et al., 2002)).

As discussed in Chapter 1, endogenous versus exogenous MyoD expression systems have been compared but only in relation to the binding consistency of MyoD from the perspective of genome-wide analysis. (Yao et al., 2013) showed that exogenous MyoD in transdifferentiated mouse embryonic fibroblasts (MEFs) show highly concordant MyoD binding between in vivo binding of endogenous MyoD in mouse myoblasts and myotubes; specifically, the authors report close to 70% overlap of top 35k MyoD ChIP-seq peaks in the two sets. However, studies focused on understanding chromatin remodeling differences between the two systems are lacking. Additionally, chromatin or gene expression-level reprogramming efficiency or differences have not been directly compared between transgene MyoD-induced and en-
dogenous MyoD-induced myogenic reprogramming of the same cell/tissue-system. Understanding chromatin and gene expression differences between the two systems can offer insights into how chromatin stability (or instability) affects genome-wide transcriptional and subsequently phenotypic reprogramming efficiency.

In this chapter, I discuss the comparative genome-wide analyses performed between two myogenic reprogramming systems - one where MyoD is overexpressed using traditional lentiviral vector medium, and another where endogenous MyoD is expressed by targeting CRISPR/Cas9-mediated transcriptional machinery at the endogenous MyoD promoter. For both systems, the same starting cell line (human iPS cells) is used. Briefly, we found that transgene MyoD is more “aggressive” in the degree and number of chromatin opening events identified compared to endogenous MyoD. In line with this, we observed that transgene MyoD also targets more myogenic genes. However, endogenous MyoD targets reprograms cellular fusion related genes better, which aligns with better cellular fusion phenotype observed for this system. We also observe that the endogenous, but not exogenous, MyoD system shows activation of some fibroblast-specific genes.

4.2 Methods

4.2.1 Plasmid construction

The plasmid construction was done by Ami Kabadi in Gersbach Laboratory. Individual gRNAs were tested for transcriptional activation using the pLV-hU6-gRNA-hUbC-VP64dCas9VP64-T2A-GFP destination vector. This plasmid contains convenient BsmBI restriction sites that allow for direct ligation of the desired protospacer sequence using previously described methods (Step 1 in (Kabadi et al., 2014)). The CRISPR/Cas9 transcriptional activators and four gRNAs were lentivirally delivered using our previously published platform (Kabadi et al., 2014). This method allows for assembly of up to four gRNAs along with Cas9 into a single vector in under
5 days. The destination vector pLV-hUBC-VP64dCas9VP64T2A-GFP was used to assemble four gRNA cassettes expressing different gRNAs targeted to the MyoD proximal regulatory region. Human WTMyoD was delivered using LV-hUbC-MyoD-IRES-dsRed. The MyoD gene was PCR amplified from cDNA extracted from human cells and ligated into the FUGW backbone. This study also made use of previously published LV-CRE and LV-Floxed Luc.

4.2.2 Cell culture

Cell culture was done by Ami Kabadi in Gersbach Laboratory. HEK293T cells were obtained from the American Tissue Collection Center (ATCC, Manassas, VA, USA) through the Duke University Cell Culture Facility and were maintained in Dulbeccos modified Eagles medium (DMEM) supplemented with 10% FBS and 1% penicillin/streptomycin. Primary human dermal fibroblasts (Catalog ID: GM03348) were obtained from Coriell Institute (Camden, NJ, USA) and were maintained in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin. mND2-0 iP-SCs were obtained from the WiCell Stem Cell Bank and maintained in mTESR (Stem Cell Technologies) and plated on tissue culture treated plates coated with ES-qualified matrigel (Corning). Prior to all experiments, the iPSCs were pre-treated with skeletal muscle media (PromoCell, Heidelberg, Germany) supplemented with 20% fetal bovine calf serum (Sigma, St. Louis, MO), 50 µg/ml fetuin (Sigma), 10 ng/ml human epidermal growth factor (Sigma), 1 ng/ml human basic fibroblast growth factor (Sigma), 10 µg/ml human insulin (Sigma), 1% GlutaMAX (Invitrogen), 1% penicillin/streptomycin (Invitrogen), and 10 µM Rock Inhibitor Y27632 (Stemgent) for 5 days to push the cells towards the myogenic lineage. Due to the short half-life, the Rock Inhibitor was added fresh to the media just prior to putting on the cells. To help induce differentiation, iPSCs were grown in differentiation media following transduction (DMEM supplemented with 2% horse serum and 1%
penicillin/streptomycin).

Primary human myoblast cell strains were generated from muscle needle biopsies of three healthy individuals using a collagenase/dispase dissociation step (Pavlath and Gussoni, 2005) and preplating to remove contaminating fibroblasts. Duly signed consent forms were obtained that had been approved by the Institutional Review Boards of Tulane Health Science Center and the University of Rochester School of Medicine. Myoblasts were propagated as previously described by Tsumagari et al. (Tsumagari et al., 2011) and checked by immunocytochemistry (Thermo, RB-9014-P) for desmin, a marker for muscle cells. Fibroblasts immunostained using the same antibody give negligible signal. An aliquot of the same batch of cells used for harvesting myoblasts was differentiated to myotubes to generate the differentiated myotube samples. Differentiation was obtained by incubation in limiting serum (1.5% horse serum for 1 day and 15% horse serum for 4–6 days). The batches of myoblasts used for these experiments contained >85% of nuclei in desmin-positive cells. The extent of myotube formation was determined by desmin-immunostaining as well as immunostaining with a monoclonal myosin heavy chain antibody (MF20). More than 70% of the nuclei in myotubes used for these experiments were in cells positive for MF20 and desmin. All cells were cultured at 37°C with 5% CO2.

4.2.3 Viral production and transduction

The viral production and transduction experiments were performed by Ami Kabadi in Gersbach Laboratory. All lentiviral vectors used in this study are second generation and were produced using standard viral production methods that have been previously described. Briefly, 3.5 million HEK293T cells were plated per 10 cm dish. The following day, cells were transfected by the calcium phosphate transfection method with 20 µg of transfer vector, 6 µg of pMD2G and 10 µg psPAX2. The media was changed 12 – 14 h post-transfection. The viral supernatant was
collected 24 and 48 h after this media change for a total of 20 mL of virus, and passed through a 0.45 um filter. The viral supernatant was concentrated to 20x using Lenti-XTM concentrator (CloneTech) prior to being snap frozen. For human dermal fibroblast transductions, cells were resuspended and plated into viral supernatant supplemented with 4 ug/ml polybrene. The viral supernatant was exchanged for fresh medium 12 – 24 h later. For skeletal muscle pretreated iPSCs, cells were transduced one day post plating by adding the desired concentration of virus to the media along with 4 ug/ml polybrene. The viral supernatant was exchanged for fresh medium 1224 h later. LV-MyoD, LV-CRE, and LV-Floxed Luciferase treated cells were transduced using a final concentration of 1× virus while CRISPR/Cas9 transduced cells were given 3× virus due to the lower viral titer of this large lentiviral vector (Kabadi et al., 2014).

4.2.4 Quantitative reverse transcription PCR

The Quantitative Reverse Transcription PCR experiments were performed by Ami Kabadi in Gersbach Laboratory. RNA was isolated using the RNeasy Plus RNA isolation kit (Qiagen). cDNA synthesis was performed using the SuperScript VILO cDNA Synthesis Kit (Invitrogen). Real-time PCR using PerfeCTa SYBR Green FastMix (Quanta Biosciences) was performed with the CFX96 Real-Time PCR Detection System (Bio-Rad). Primer specificity was confirmed by agarose gel electrophoresis and melting curve analysis. Reaction efficiencies over the appropriate dynamic range were calculated to ensure linearity of the standard curve. The results are expressed as fold-increase expression of the gene of interest normalized to β-actin expression using the ∆∆Ct method (Livak and Schmittgen, 2001).
4.2.5 Immunofluorescence staining

Immunofluorescence staining experiments were performed by Ami Kabadi in Gersbach Laboratory. iPSCs pretreated with skeletal muscle media were plated on autoclaved glass coverslips (1 mm, Thermo Scientific) coated with ES-qualified matrigel (Corning). The following day, lentivirus delivering either a CRISPR/dCas9 transcriptional activator or MyoD was added to the media. One day post-transduction, media was changed to differentiation media (DMEM supplemented with 2% horse serum and 1% penicillin/streptomycin). Seven days post-transduction, cells were fixed in 4% PFA and prepared for immunofluorescence staining. Samples were permeabilized in blocking buffer (PBS supplemented with 5% BSA, 0.2% Triton X-100, and 2% goat serum) for 1 h at room temperature. Samples were incubated with MF20 supernatant primary antibody (1:200 dilution, Hybridoma Bank) or MyoD 5.8A (1:200 dilution, Santa Cruz) in blocking buffer overnight at 4

Quantitative Cell Fusion Assay experiments were performed by Ami Kabadi in Gersbach Laboratory. To conduct the fusion assay, iPSCs pretreated with skeletal muscle media were transduced with either LV-Cre or LV-Floxed Luc. Following transduction, these cell populations were mixed and plated in a 1:1 ratio of LV-Cre transduced cells to LV- Floxed Luc cells in 48 well plates. The next day, cells were transduced with CRISPR/dCas9 transcriptional activator in a modified LV backbone without LoxP sites or MyoD expressing lentivirus. Cells were harvested 3, 5, 7, or 10 days post-transduction and assayed for luciferase expression. Cells were pelleted and washed with PBS. Pellets were resuspended in 100 μL of lysis buffer (100 mM KH2PO4 + 0.2% Triton-X, pH 7.8) and incubated at room temperature for 10 min. The cell debris was pelleted and 30 μL of the supernatant from each sample was transferred to an opaque 96-well plate. Each sample was mixed with 30 μL of Bright-Glo reagent (Bright-Glo Luciferase Assay System, Promega). Lumines-
cence was measured by a BioTek Synergy 2 Multi-Mode Microplate Reader with 1-s scan time. Each time point had matched control samples that were not transduced with a transcription factor. All luciferase data is presented as a fold increase over background from the matched control.

4.2.7 Chromatin immunoprecipitation followed by quantitative real-time PCR

Chromatin Immunoprecipitation followed by Quantitative Real-Time PCR (ChIP qRT-PCR) experiments were performed by Ami Kabadi, Gersbach Laboratory. The protocol for this was adapted from the Myers lab protocol. For each sample, ten million cells were transduced and harvested seven days post transduction. Adherent cells were crosslinked in the plate with a final concentration of 1% formaldehyde. Cells were re-suspended in 300 µL RIPA buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS in 1% PBS). To fragment the chromatin, samples were sonicated for a total time of 45 min (30 seconds on, 30 seconds off) using a Diagenode XL Bioruptor at 4°C. Cell debris was removed by centrifugation at 14,000 rmp for 15 minutes at 4°C. 5 µg of anti-H3K4me3 antibody (Abcam, ab8580) or anti-H3K27ac antibody (Abcam, ab4729) was conjugated to 200 µL of sheep anti rabbit IgG beads (Invitrogen. 11204D) by rocking overnight at 4°C. The sheared chromatin in RIPA was then incubated with the conjugated beads on a rotator overnight at 4°C. Beads were washed five times with LiCl IP wash buffer (100mM Tris pH 7.5, 500mM LiCl, 1% NP-40, 1% sodium deoxycholate) and one time with TE buffer (10mM Tris-HCl pH 7.5, 0.1 mM Na2EDTA) at 4°C. Chromatin was eluted from the beads by incubating in IP elution buffer (1% SDS, 0.1% NaHCO3) for 1 hour at 65°C. Beads were pelleted by centrifugation at 14,000 rpm for 3 min at room temperature. The chromatin supernatant was incubated overnight at 65°C to complete the crosslinking reversal. DNA was purified using the Qiaquick PCR Purification Kit (Qiagen). Real-time PCR using PerfeCTa SYBR Green FastMix (Quanta Biosciences) was
performed with the CFX96 Real-Time PCR Detection System (Bio-Rad) using 1 ng of precipitated chromatin as template. Primer specificity was confirmed by agarose gel electrophoresis and melting curve analysis. Reaction efficiencies over the appropriate dynamic range were calculated to ensure linearity of the standard curve. The results are expressed as fold-enrichment of the target genomic region normalized to a β-actin (H3H4me3) or GAPDH (H3K27ac) promoter region that remains constant using the ∆∆Ct method (Livak and Schmittgen, 2001).

4.2.8 DNase-seq experiments

The DNase-seq experiments were performed by Lingyun Song in Crawford Laboratory. All experiments were performed using three to four biological replicates starting from independent cell transductions for the iPSCs or independent donors for the myoblast and myotube samples. Library preparation and analysis were performed as previously described with the one exception of adding a 5’ phosphate group to oligo 1b to increase ligation efficiency (Song and Crawford, 2010b; Song et al., 2011). iPSC replicates were harvested seven days after transduction while myoblasts and myotube samples were harvested 4 – 5 days after plating. Approximately 2.5 × 10^7 nuclei were extracted and then digested with different amounts of DNase I for 16 min at 37°C. Reactions were terminated by the addition of 50 mM EDTA. Libraries were constructed from pooled digests as described and sequenced on the Illumina HiSeq2000 or Illumina GAIIx platform with 50-bp single-end reads at the Duke Genome Sequencing Shared Resource. Resulting reads were filtered for delivered vector sequences and aligned by Bowtie (Langmead et al., 2009).

The myoblast and myotube data was re-analyzed from a previous study, (Chandra et al., 2015), along with one additional myoblast and myotube replicate.
4.2.9 DNase-seq processing

First, for simplicity, the cell types used in this study are abbreviated as follows:

1. iPSC: Induced pluripotent stem cells

2. PTCI: Pre-treated control iPSC cells

3. CR-MYOD: PTCI cells reprogrammed using CRISPR/Cas9-mediated endogenous MYOD1 overexpression

4. TG-MYOD: PTCI cells reprogrammed using transduction of lentiviral MYOD1 transgene

5. MB: Primary Human Myoblasts

6. MT: Primary Human Myotubes

DNase hypersensitive (DHS) sites were called using MACS2 v2.1.10 (Zhang et al., 2008) at a loose cutoff. The cell line specific peaks from multiple replicates were merged using mergeBed with `-c 4,5,6,7,8,9,10 -o count,mean,first,mean,min,min,first` options (Quinlan and Hall, 2010). Afterwards, IDR (Li et al., 2011) was then run on all replicates using the cell line-specific merged peaks as the oracle peak list.

The cell line specific DHS peaks that had a global IDR score of ≤ 0.1 on at least half, or ≤ 0.05 on at least one-third of all pairwise replicates were picked for downstream analyses. This resulted in a total of 80,159 peaks in iPSC, 82,774 peaks in PTCI, 92,103 peaks in CR-MYOD, 127,760 peaks in TG-MYOD, 79,150 peaks in MB, and 120,657 peaks in MT. These peaks were then merged using mergeBed (with default options). Peaks in chromosomes X, Y and M, as well as those overlapping UCSC unmappable regions were removed, resulting in a total of 211,864 DHS sites.
For these merged DHS sites, the mean peak length is 577 bp and the median peak length is 486 bp.

DESeq2 was used to find DHS sites differentially accessible between PTCI and MT at adjusted p-value of $\leq 0.01$ and $|\log_{2}FC| \geq 2$ (Figure 4.4D). For this, DHS signal at DHS-center $\pm$ 150 bp regions were used to generate the input count matrix.

For some DHS patterns, GREAT analysis was run using default options to identify potential functional relevance of the sites (McLean et al., 2010).

For chromatin sites that are differentially accessible between the starting cell line, PTCI, and the target cell line, MT, the degree of change in chromatin accessibility observed after reprogramming is computed as “Chromatin Reprogramming Level” (CRL) measure, similar to equation 2.3 in Chapter 2. Specifically, for a DHS site $s$ in cell line $c \in \{\text{TG-MYOD, CR-MYOD}\}$:

$$CRL_c(s) = \frac{DNase_c(s) - DNase_{PTCI}(s)}{DNase_{MT}(s) - DNase_{PTCI}(s)},$$

where $DNase_x(s)$ represents the number of normalized DNase-seq reads mapping to site $s$ in cell line $x$.

Hence, a DHS site differentially more open or closed in PTCI when compared against MT will have a $CRL_c(s)$ score $\sim 1$ if its state of accessibility in cell line $c \in \{\text{TG-MYOD, CR-MYOD}\}$ is similar to that in MT, or a $CRL_c(s)$ score of $\sim 0$ if it similar to PTCI instead. Also, a CRL score of $> 1$, as is observed for TG-MYOD in Figure 4.4B and 4.6B, suggests that the state of chromatin accessibility particular to the DHS site $s$ in cell line $c \in \{\text{TG-MYOD, CR-MYOD}\}$ is more than what is observed for MT.
4.2.10 RNA-seq experiments

The RNA-seq experiments were performed by Lingyun Song in Crawford Laboratory. All experiments were performed using three to four biological replicates starting from independent cell transductions for the iPSCs or independent donors for the myoblast and myotube samples. Myoblast and myotube RNAs were isolated from donor different than those used for DNAse-seq. RNA-seq libraries were constructed as previously described (Gertz et al., 2012). Briefly, first-strand cDNA was synthesized from oligo(dT) Dynabead-captured mRNA using SuperScript VILO cDNA Synthesis Kit (Invitrogen). Second-strand cDNA was synthesized using DNA polymerase I (New England Biolabs). cDNA was purified using Agencourt AMPure XP beads (Beckman Coulter). Nextera transposase (Illumina; 5 min at 55°C) was used to simultaneously fragment and insert sequencing primers into the double-stranded cDNA. Transposition reactions were halted using QG buffer (Qiagen) and fragmented cDNA was purified on AMPure XP beads. Indexed sequencing libraries were generated by six cycles of PCR. Libraries were sequenced using 50-bp paired-end reads on two lanes of an Illumina HiSeq 2500 instrument at the Duke Genome Sequencing and Analysis Core Resource.

For myoblast and myotube cell types, RNA-seq data was re-analyzed from Chandra et al. (Chandra et al., 2015).

4.2.11 RNA-seq data processing

The raw fastq files were processed with sortmerna (Kopylova et al., 2012) and trimmomatic (Bolger et al., 2014) to filter out the rRNA and adapter sequences respectively. The output reads were then aligned to the h19 genome assembly using HISAT2 (Kim et al., 2015). Afterwards, the read counts were obtained for each gene using HTSeq (Anders et al., 2015). The GTF file used for HTSeq was obtained from the PrimerSeq library ((Tokheim et al., 2014), http://primerseq.
As with DNase seq analysis, DESeq2 was used to obtain the significantly differentially expressed genes using cutoffs of 0.01 and 1.5 for adjusted p-value and log2-fold-change respectively (Anders and Huber, 2010). Also, SVA was used to control for the batch effect (Leek et al., 2012). The normalized read counts were then converted to TPM scores (see 1.6.3 in Chapter 1, for which Picard (https://broadinstitute.github.io/picard/) was used to compute the mean fragment length of each sample.

4.2.12 Clustering gene expression matrix

All genes that are significantly differentially expressed between any two cell lines are selected for clustering. The gene expression matrix in Figures 4.3 and 4.7 was clustered using DPGP (short for Dirichlet Process, Gaussian Process) clustering algorithm (McDowell et al., 2018). It is a non-parametric clustering approach that yields the best number of cluster appropriate for the data. DPGP was run with n 300 unscaled do-not-mean-center options, on the log transformed and standard normalized gene expression matrix. Specifically, for each gene \( g \), let

\[
g(s_i) = \log_2(\text{TPM of gene } g \text{ in sample } s_i)
\]

be the \( \log_2 \)-transformed TPM value of \( g \) in sample \( s_i \). Let \( g(s) \) be the corresponding expression vector across all samples \( s_i \in \{ipsc_{rep1}, \ldots, mt_{rep5}\} \). Then, the z-scores of the expression values were computed for each sample as

\[
z(g(s_i)) = \frac{g(s_i) - \text{mean}(g(s))}{\text{standard deviation}(g(s))}.
\]

The median \( z \)-scores expression values for each cell lines were used as input for DPGP algorithm.
4.2.13 MyoD ChIP-seq data in myotubes

The MyoD ChIP-seq data on primary myotubes (treatment sample: GSM1218850, and control sample: GSM1218848) were obtained from (MacQuarrie et al., 2013) and is available in GEO (accession code: GSE50415). Over-represented sequences form the fastq files were removed using cut-adapt software. The processed fastq reads were aligned to hg19 using bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009). MACS2 was then run to identify peaks at FDR 0.001 threshold, which yielded a total of 56,547 peaks (Zhang et al., 2008).

4.2.14 DHS pattern enrichment analysis

For some gene-clusters identified in Figure 4.3, any DHS pattern significantly enriched in the cis-genic regions of the cluster-genes compared against a background set of genes were identified. For this, the genes in the cluster of interest is treated as the treatment genes. These were filtered in a couple of ways:

1. They are filtered to have their expression profiles significantly differential between the pairs of cell lines to match the profiles observed in Figure 4.3. For example, treatment genes in cluster 2 (second from the top in Figure 4.3) were filtered so that these were significantly differentially upregulated in MB and MT when compared against PTCI, CR-MYOD and TG-MYOD cell lines. (More cluster-specific detail are highlighted at the end of this sub-section below.)

2. Additionally, the treatment genes are required to have their expression values > 2 TPM in the cell lines showing the highest expression profile. For example, for cluster 2, only genes with TPM value > 2 in MB and MT were selected.

The control gene sets were constructed to match the expression values of the treatment genes in Myotubes. Specifically, if a gene in the treatment set has its
expression in MT to be \(x\) TPM, all genes with expression in myotubes in the range \((x - 2^{0.2} \times x, x + 2^{0.2} \times m)\) are selected without duplication to be in the control gene set.

Once the treatment and control gene sets are finalized, the number of DHS patterns overlapping each of the genes in TSS \(\pm\) 200kb regions are compared using Mann-Whitney U-Test. The one-sided p-values recorded were corrected for multiple hypothesis testing using Benjamini-Hochberg FDR correction with \texttt{statsmodels.stats.multitest} python package.

**More notes on cluster-specific treatment gene selection:**

1. For cluster 3 (third from the top in Figure 4.3), treatment genes are required to have their expression in TG-MYOD significantly higher than in PTCI, MB and MT. Also, their expression values in TG-MYOD cell line are required to be \(> 2\)TPM.

2. For cluster 4, treatment genes are required to have their expression in MT to be \(> 2\)TPM and significantly higher than in PTCI, CR-MYOD and TG-MYOD.

3. For cluster 16 (second to last cluster from the top in Figure 4.3), treatment genes are required to have their expression in PTCI, CR-MYOD and TG-MYOD to be all \(> 2\) TPM and significantly higher than in both MB and MT.

### 4.3 Results

#### 4.3.1 Exogenous and endogenous MyoD-mediated differentiation of human iPS cells generate muscle-like cells

The experiments discussed in this section are performed by Ami Kabadi, Gersbach Laboratory. Human iPS cells (iPSC) were converted to myocyte-like cells by either endogenous \textit{MYOD1} overexpression or lentiviral \textit{MYOD1} transgene expression as
shown in the schematic in Figure 4.1. Briefly, iPSC were treated to skeletal muscle media for 5 days to get pre-treated control iPS cells (henceforth called PTCI). This protocol has been shown to improve iPSC engraftment efficiency in an in vivo mouse model of muscle regeneration (Goudenege et al., 2012). Transduction and activation of lentiviral human MYOD construct in PTCI, as done previously in (Manandhar et al., 2017) and (Kabadi et al., 2015), yielded cells, TG-MYOD (short for “Trangene-MyoD reprogrammed cells”) after 7 days. As an alternative reprogramming system, lentiviral vector co-expressing combinations of guide RNAs (gRNAs) and transcriptional activator VP64-dCas9-VP64 were transduced in PTCI to target MYOD1 promoter-proximal regulatory region (PRR) in vivo. This combination of gRNAs, total four in number, were identified from a list of thirty gRNAs tested as follow: (i) First each gRNA was individually assessed in its ability to express endogenous MYOD and MYOG genes; and (ii) Random groups of gRNAs that showed highest efficiency in step (i) were tested similarly to get the group of gRNAs that maximizes the expression of endogenous MYOD1 and MYOG (not shown). These experiments were performed in primary human fibroblasts initially. However, endogenous MYOD1 expression did not yield quantifiable phentoypic myogenic conversion (not shown), possibly in part because the PRR region is not accessible in primary fibroblasts (Fong and Tapscott, 2013) (not shown). However, the same combination of gRNAs was able to generate myocyte and some myotube-like cells for iPSC cells (Figure 4.1). For simplicity, these CRISPR/dCas9-mediated reprogrammed PTCI cells are called “CR-MYOD” cells. Both CR-MYOD and TG-MYOD cells were compared against primary myoblasts (MB) and myotubes (MT) in the genome-wide chromatin accessibility and gene expression profile changes in the following sections.

Figure 4.1B shows comparative time-course expression profiles of key muscle marker genes, and total as well as MYOD expression. Both MyoD expression systems resulted in robust expression of the probed genes. Comparatively, while MyoD
Skeletal muscle media
5 days
7 days
+Dox
compared against iPSCs PTCI
CR-MYOD
TG-MYOD
Primary myoblasts (MB)
and myotubes (MT)

A

B

Total MyoD
Endogenous MyoD
Myogenin
Desmin
Creatine Kinase
Troponin T

Log(10) fold change (relative mRNA)

Time (Days)

Control
VP64-dCas9-VP64 + gRNAs
MyoD

C

D

Percent MyoD Positive Nuclei

Control
VP64-dCas9-VP64 MyoD

E

Cell Fusion

F

DAP1
MyoD

DAPI/Myosin Heavy Chain

126
**Figure 4.1:** Endogenous MyoD expression shows better fusion but lower level of myogenic gene activation. (A) Schematic representation of the MyoD-induced reprogramming experiments of human iPSC cells. iPSCs were pre-treated with skeletal muscle media for 5 days to get PTCI (pre-treated control iPSC) cells. PTCI were either subjected to (i) endogenous expression of MYOD1 gene by targeting dCas9 with two VP64 transactivation domain epitopes to MYOD1 promoter using a combination of CRISPR/Cas9 guide RNAs, or (ii) lentiviral transduction of MYOD1 transgene. After 7 days of MyoD expression on both systems, the resulting cell lines (CR-MYOD and TG-MYOD) were compared against primary myoblasts and myotubes for downstream analyses. B Time course mRNA expression profiles of total MyoD, endogenous MyoD, and some muscle marker genes for the same three systems - PTCI control, VP64-dCas9-VP64 and transgene-MyoD systems. (C,E) DAPI (in C,E), MyoD (in C) or Myosin Heavy Chain (in E) immunostaining performed on PTCI control, VP64-dCas9-VP64 and transgene-MyoD systems. (D) Percentage MyoD-positive nuclei across the three systems, (F) Cell Fusion Efficiency quantified for the three systems.

transgene yielded total MyoD mRNA, CR-MYOD contained the largest values of endogenous mRNA. Expression of the muscle marker genes, however, is induced higher in TG-MYOD compared to CR-MYOD indicating that overall myogenic conversion efficiency is likely higher for lentiviral MyoD overexpression system. Immunostaining of MyoD positive nuclei also show higher fraction of TG-MYOD nuclei (~ 85%) is MyoD positive than CR-MYOD nuclei (~ 77%) (Figure 4.1C,D). Interestingly, however, CR-MYOD cells show better cellular fusion than TG-MYOD cells, both visually (as shown by myosin heavy chain immunostaining in Figure 4.1E), and quantitatively (4.1F, see Methods).

It should be noted that difference in the total expression of MyoD or MyoD-positive nuclei between the two systems can bias the downstream comparative analyses and conclusions. In other words, some of the differences observed could be a direct result of the differences in MyoD concentration between the two systems. This is taken into consideration in downstream analyses and are discussed in the following sections. Nevertheless, this difference in MyoD concentration does not bias (i) a
non-comparative system-specific analysis of these myogenic reprogramming systems, or (ii) a joint analysis looking into the relationship between chromatin accessibility remodeling and change in gene expression (discussed in sub-section 4.3.4).

4.3.2 Genome-wide transcriptional profiles in CR-MYOD and TG-MYOD match the muscle marker and fusion phenotypes observed

RNA-seq data were collected in iPSC, PTCI, CR-MYOD and TG-MYOD samples, while MB and MT RNA-seq samples were reanalyzed from a previous study (see Methods). Principal Component Analysis (PCA) on the RNA-seq samples shows that samples in specific cell types, by and large, cluster together except for MB samples; previous studies have attributed heterogeneity in myoblasts to variations in cellular origin, location and possibility of non-myoblast contamination (Zeng et al., 2016; Cusella-De Angelis et al., 1994) which could be the case for these myoblasts (see Methods). The PCA plot also shows gradual progression of cell types from iPSC states at one end towards muscle-like states at the other along the first principal component (PC1), which explains a large fraction (59%) of variance in the data. Samples along PC1 show that TG-MYOD is closer to muscle states than CR-MYOD. However, on the second principal component (PC2), which explains 20% of variance, TG-MYOD is also slightly further than CR-MYOD from the muscle samples, indicating that the transgene MyoD also induces non-myogenic changes than the endogenous MyoD does. Similar to the PCA plot, the distance matrix plot between the samples show that CR-MYOD and TG-MYOD samples are closer to myogenic lineage than iPSCs, but at the same time, are distinctly different from muscle-cells (Figure 4.2).

A number of key marker genes, some including those in Figure 4.1, showed high upregulation in both CR-MYOD and TG-MYOD cells (Figure 4.3A). Differential expression analysis using DESeq2 (Anders and Huber, 2010), showed that TG-MYOD
has more than twice as many significantly differentially expressed genes (n=3967) compared to CR-MYOD (n=1879) (p-value and FDR thresholds used: 0.01 and 1.5 respectively, see Methods). While this could be at least in part because TG-MYOD has overall total MyoD expressed than CR-MYOD as discussed in section 4.3.1, not all genes in PTCI that are differentially expressed in TG-MYOD, but not in CG-MYOD, necessarily conform to myogenic conversion as indicated by the PCA plot (Figure 4.2A). We also looked into a joint analysis of all differentially expressed genes across the cell types (Figure 4.3C). Any gene differentially expressed in at least one pair of cell types shown were included resulting in a total number of 10,949 genes differentially expressed across the cell types. These genes were then clustered using the non-parametric DPGP clustering algorithm ((McDowell et al., 2018), see Methods). Gene ontology enrichment analysis performed using DAVID ((Huang da et al., 2009b)) showed that genes in clusters enriched for terms related to muscle proteins and functions were generally visually more expressed in TG-MYOD than CR-MYOD. Again, this could be a direct reflection of the fact that MyoD concentration or MyoD positive nuclei is significantly higher in TG-MYOD than CR-MYOD (see section 4.3.1, Figure 4.1). Interestingly, two gene clusters with higher gene expression in

\textbf{Figure 4.2:} (A) PCA and (B) distance matrix plot for RNA-seq samples across the cell types.
Figure 4.3: Comparative gene expression analyses between CR-MYOD and TG-MYOD samples. (A) $\text{Log}_2$(TPM) measures of some muscle specific genes across the cell types. (B) MA plot showing genes differentially expressed between PTCI and CR-MYOD (left panel), and PTCI and TG-MYOD (right panel). (C) Heatmap showing standard-normalized TPM values and adjoining median cell type-specific profiles for the genes significantly differentially expressed between any two cell types. These genes were clustered using the DPGP clustering algorithm (McDowell et al., 2018). For some of these gene-clusters, Gene Ontologies enriched were identified using DAVID (Huang da et al., 2009b).
CR-MYOD than TG-MYOD were enriched for “cell junction”, “cell adhesion” or “transmembrane” related terms, which fits the observation that CR-MYOD fuse better than TG-MYOD. Targeted reprogramming of these “fusion” genes could help TG-MYOD cells fuse better.

### 4.3.3 Transgene MyoD is “aggressive” in its chromatin remodeling behavior

DNase-seq data were also generated for iPSC, PTCI, CR-MYOD and TG-MYOD samples. For MB and MT samples, DNase-seq data previously generated were used (see Methods). DHS peaks for each cell line were called by filtering loosely called MACS2 DHS peaks to get reproducible peaks using IDR (Li et al., 2011) (see Methods). The range of the peaks called in each cell line ranged from about 80k for MB to 128k for TG-MYOD (see Methods). These peaks were then merged and filtered to remove peaks overlapping UCSC unmappable regions or in chromosomes X,Y and M (see Methods) to get a total of 211,864 DHS sites, with mean and median peak lengths of 577 bp and 486 bp respectively.

For each of the processed DHS peaks, a binary DHS-pattern profile was also generated by assessing whether the peak is present (“1”) or absent (“0”) across the cell lines iPSC, PTCI, CR-MYOD, TG-MYOD, MB, and MT in order. Hence, a DHS site with pattern “000100” would refer to it being accessible only in TG-MYOD cells. Figure 4.4A shows the frequency of such DHS patterns genome-wide. Figure 4.4A shows the frequency of each such binarized DHS pattern throughout the whole genome. Surprisingly, the TG-MYOD-specific (“000100”) DHS sites is the most frequent (n ~ 23k) DHS pattern observed. Comparatively, the number of CR-MYOD-specific (“001000”) DHS sites is only ~ 3k. This indicates that the transgene MyoD opens chromatin in a non-specific manner with much higher frequency compared to CR-MYOD. Similar to TG-MYOD, the iPSC-, PTCI- and MT-specific DHS sites are also highly frequent suggesting that the chromatin architecture of these cell
Figure 4.4: (A) Frequency of DHS sites with accessibility patterns across the cell lines as shown. For each accessibility pattern, dark square box represents a DHS site and white box represents closed chromatin in the corresponding cell line. (B) Distributions of Chromatin Reprogramming Level (CRL) corresponding to CR-MYOD ($CRL_{CR-MYOD}$, red curves) and TG-MYOD ($CRL_{TG-MYOD}$, blue curves) obtained by considering PTCI as the starting cell line and MT as the target cell line are shown. The top panel corresponds to sites that are significantly more accessible in MT than in PTCI ($n = 17577$), and the bottom panel corresponds to sites that show the opposite accessibility profile ($n = 20267$), as obtained from the MA plot shown in the center.

(continued)
(C) Distributions of DNA-binding specificity of hlh-1, a MyoD ortholog in *C. elegans*, shown for sites that are bound by MyoD in MT (blue curves) or not (green curves). As in B, the top panel corresponds to sites that are significantly more accessible in MT than PTCI, and the bottom panel corresponds to sites showing opposite DNA accessibility profile. Note that in C, lower panel, only about 1% of the total sites make up the green distribution (i.e. sites that are significantly less accessible in MT than PTCI and also bound by MyoD in MT).

lines are notably distinct, similar to the RNA-seq profiles (Figure 4.2). Nonetheless, large number of chromatin sites open up to be “reprogrammed” in both systems (i.e. “011111”, “001111” DHS sites) suggesting both endogenous and trangene MyoD expression open a large number of common DHS sites.

In order to assess the degree of chromatin reprogramming observed with MyoD overexpression, DESeq2 (Anders and Huber, 2010) was used to find DHS sites differentially accessible between PTCI and MT at adjusted p-value of \( \leq 0.01 \) and \( |\log_{2} FC| \geq 2 \). For this, DHS signal at DHS-center ± 150bp regions were used to generate the input count matrix. Figure 4.4D shows that \( n = 17577 \) sites are significantly more accessible in MT than PTCI, and \( n = 20267 \) sites are more accessible in PTCI compared to MT. These together constitutes roughly 20% of total DHS sites across the cell lines in the genome.

For the differentially accessible sites, the degree of chromatin reprogramming observed were computed for CR-MYOD and TG-MYOD as Chromatin Reprogramming Level (CRL) measures, (see subsection 4.2.9 in Methods). As discussed previously, CRL score of \( \sim 1 \) corresponds to accessibility profile matching target (in this case, MT) cells, and a score close to 0 corresponds to the accessibility profile matching the starting (in this case, PTCI) cells. Figure 4.4B shows the CRL distributions for CR-MYOD (red curves) and TG-MYOD (blue curves). These profiles are similarly skewed as \( CRL_{fibro-MyoD} \) distribution in Chapter 1. For both sets of sites that are significantly more accessible in MT than PTCI (Figure 4.4B, upper panel) or
Figure 4.5: An example of a DHS site more reprogrammed by transgene ($CRL_{TG-MYOD} > 1.5$) than endogenous ($CRL_{TG-MYOD} < 0.6$) MyoD. significantly more accessible in PTCI than MT (Figure 4.4B, lower panel), transgene MYOD reprograms the chromatin significantly more than endogenous MyoD (Wilcoxin signed-rank test p-values: < 10-e10 for upper panel, and < 10e-273 for lower panel). A large fraction of chromatin sites that open up and “reprogram” are show higher accessibility achieved than is observed for MT (i.e. $CRL_{TG-MYOD} > 1$). Figure 4.5 shows an example of one such chromatin site that opens up more in TG-MYOD ($CRL_{TG-MYOD} = 1.6$) than CR-MYOD ($CRL_{CR-MYOD} = 0.6$). Figure 4.4B shows that transgene MyoD not only opens chromatin significantly more than endogenous MyoD, but it directly or indirectly also closes chromatin more efficiently.

Figure 4.4C shows the DNA binding specificity of hlh-1, a *C. elegans* ortholog of MyoD (Murre et al., 1989), (henceforth, also called as MyoD in the context of DNA-binding specificity) in the differentially accessible sites highlighted in Figure 4.4D. For each DHS site, the mean of top three MyoD 8-mer e-scores were assigned as the DNA binding specificity measure of MyoD corresponding to the site (see subsection 1.6.4...
Figure 4.6: Transgene MyoD opens chromatin sites involved in myogenic activities more in number and degree compared to CRISPR/Cas9-based endogenous MyoD activation. (A) Heatmaps of chromatin accessibility profiles of sites significantly more accessible in MT than PTCI across all cell types are shown. These differentially accessible sites were secondarily filtered to have only those sites that are called by MACS2 peak-caller to have a peak in MT but not in PTCI. The left heatmap shows the normalized raw signal, while the right shows the binarized calls of the same sites as DHS sites or not, as identified by MACS2. (B) For the patterns of DHS sites shown, distributions of Chromatin Reprogramming Level (CRL) scores are also shown for CR-MYOD CRL_{CR-MYOD} (blue distributions) and TG-MYOD CRL_{TG-MYOD} (orange distributions). These are obtained by taking PTCI to be the starting and MT to be the target cell type. (C) For some DHS sites following specific DHS-patterns shown, Gene Ontologies enriched are also listed as identified by GREAT.

in Chapter 1). Figure 4.4C (upper panel) show that the DNA binding specificity of MyoD is high (i.e. E-score > 0.35 for a majority of sites) and roughly 55% of these are bound by MyoD in MT. Hence, Figure 4.4B (upper panel) and Figure 4.4C (upper panel) together show that endogenous MyoD, more so than transgene MyoD, does not bring about chromatin accessibility change in a large fraction of regions.
containing its potential binding sites. This could be either because the chromatin epigenome provides a hindrance to MyoD-mediated chromatin remodeling at many of these sites (with \( CRL \sim 0 \)) or MyoD itself is not sufficient to bring about the change in accessibility. Figure 4.4C (lower panel) shows that MyoD is bound only in \sim 1\% of sites that are significantly inaccessible in MT than PTCI, and this \sim 1\% of sites have significantly higher binding specificity for MyoD than the rest in the set.

The DHS sites significantly more accessible in MT than PTCI were further analyzed in their chromatin reprogramming efficiency and enrichment of muscle-related Gene Ontology (GO) terms (Figure 4.6). For this, first the DHS sites were further filtered so that they are called as peaks in MT but not in PTCI (see Methods), which resulted in a total of 16191 DHS sites shown (Figure 4.6A). The binarized representation of “DHS versus not a DHS site” is shown in the adjoining heatmap. In Figure 4.6B, the CRL distributions are shown for few binarized patterns of DHS sits. These heatmaps and CRL distributions show that MyoD overexpression opens roughly half of “MT-accessible-PTCI-inaccessible” DHS sites, for which transgene MyoD generally shows significantly higher chromatin accessibility achieved than the endogenous MyoD. As mentioned previously, this could be partly because CR-MYOD has lower total MyoD concentration than TG-MYOD. GO enrichment analysis performed on the DHS sites showed that the non-reprogrammed sites are also likely involved in regulating myogenic phenotype.

4.3.4 Cis-genic DHS accessibility profiles match the expression pattern of the differentially expressed genes

In order to understand the relationship between chromatin remodeling events and gene expression change, DHS-pattern enrichment analysis was performed for some gene sets identified in Figure 4.3 (see Methods for details), similar to our study in Chapter 2. Specifically, for some gene-clusters identified in Figure 4.3, any DHS
Figure 4.7: For a few gene-clusters identified in 4.3, binarized patterns of DHS sites across the cell types significantly more enriched around the genes in the cluster (also referred to as “treatment genes”) compared to control genes are shown. The treatment genes are filtered to match their expression profiles observed in the heatmap (see Methods, section 4.2.14). The control genes are all the genes that have expression in MT comparable to the treatment genes (see Methods, section 4.2.14). For a couple of DHS-patterns enriched in the treatment genes shown, distributions of the specific DHS-patterns enriched are also shown for the corresponding treatment and control genes.

pattern significantly enriched in the cis-genic regions of the cluster-genes compared against a background set of genes were identified. Briefly, the genes were first filtered to match their expression profile in the standard-normalized heatmap. For instance, cluster 2 (second from the top in Figure 4.3) were filtered so that these were significantly differentially upregulated in MB and MT when compared against PTCI, CR-MYOD and TG-MYOD cell lines. These genes were compared against a background set of control genes that have similar levels of expression in MT (see
Methods) to identify any binarized DHS patterns significantly enriched in TSS ± 200kb regions of the genes in the cluster than would be expected by chance. As shown, the third cluster has two patterns of DHS patterns (“00100” and “01100” - note iPSCs are not shown) enriched, which match the expression profile in the heatmap. This observation also follows for few other gene clusters and DHS patterns highlighted, but is interestingly, not universally true for all gene clusters.

4.3.5 Gene regulatory network analysis shows CR-MYOD have fibroblast genes activated

In addition to the chromatin and gene-expression based analyses, we also looked into the gene regulatory network (GRN) profiles of the cell lines in this study using CelNet
(Cahan et al., 2014; Radley et al., 2017) (Figure 4.8. CellNet classifies cell/tissue-
samples to various tissue types based on the expression profile of genes in the target
tissue or cell type-specific GRN. From Figure 4.8, it is clear that iPSC samples
closely resemble ESC-specific GRN, while MB and MT samples closely resemble the
GRN profiles of skeletal muscle. Also, TG-MYOD samples have higher similarity to
skeletal muscles compared to CR-MYOD, in line with the previous observations based
on chromatin and gene-expression profiles. Interestingly, we also observed that PTCI
samples and somewhat in CR-MYOD samples, genes in fibroblast-specific GRNs are
also activated, which is not observed for TG-MYOD samples. This could mean
that transgene MyoD is able to deactivate transitory expression of fibroblast genes,
unlike CRISPR/dCas9-mediated expression of endogenous MyoD. Further analyses
is required to assess whether proper deactivation of these fibroblast genes in CR-
MYOD could enhancer their overall similarity to skeletal muscles.

4.4 Conclusions

This study has shown CRISPR/dCas9-mediated overexpression of endogenous
MYOD1 gene as an alternative approach of myogenic reprogramming to traditional
lentiviral MyoD transgene overexpression. While both systems can yield cell pop-
ulations with over 70% MyoD positive nuclei, total MyoD expression in our study
is higher for TG-MYOD than CR-MYOD cells, which can partly explain the higher
efficiency of transgene MyoD in reprogramming myogenic genes or potentially muscle-
related sites compared to endogenous MyoD. We acknowledge that a future study
with similar expression level of total MyoD would present an unbiased approach to
comparative analyses introduced in this chapter. Nonetheless, despite this potential
source of variance, transgene MyoD not only opens large number of TG-MYOD-
specific DHS sites, but also opens chromatin more so than is observed in MT for “re-
programmed” sites as shown by our CRL analysis. Our CRL analysis also highlights
that endogenous MyoD largely targets the same sets of “open-in-MT” DHS sites as transgene MyoD, but the degree of chromatin accessibility change is significantly lower. This could partly be a result of the difference in total MYOD expression observed between the two systems, as discussed above. However, it should also be noted that both systems remodel but a fraction of all the chromatin and expression profiles potentially relevant myogenic differentiation and activities. These findings suggest that the transcriptional and epigenetic status of the reprogrammed cells are distinctly somewhere in-between the starting iPSC and target MT.

An interesting difference between CR-MYOD and TG-MYOD observed was that CR-MYOD had some fibroblast-specific genes activated as shown by the CellNet gene regulatory network analysis (GRN) (Figure 4.8). It would be interesting to understand how TG-MYOD is able to suppress fibroblast GRN activation that the skeletal muscle media treatment started in PTCI, whereas CR-MYOD is not. Further analysis on this might be insights into fibroblast-specific genes that are refractory to myogenic reprogramming.

One of the possible direction for future studies to assess whether overexpressing MYOD concomitantly with BAF60C improves the efficiency of myogenic conversion in one or both of the myogenic reprogramming systems discussed in this study. Baf60c, a member of SWI/SNF chromatin remodeling complex, has been shown to convert mouse embryonic stem cells (ESCs) and HeLa cells significantly better than by MyoD alone (Albini et al., 2013; Forcales, 2012), as discussed in Chapter 1. Introducing BAF60C could potentially help MyoD reprogram a larger fraction of sites currently missed by both systems.
Epigenetic poising and co-factor binding explain HDACi-induced chromatin opening at PU.1-bound sites

This chapter described computational analyses I performed as part of a collaborative project with Christopher L. Frank, currently an alumnus of Crawford Laboratory. The original article has been published by the title “HDAC inhibitors cause site-specific chromatin remodeling at PU.1-bound enhancers in K562 cells”, by Christopher L. Frank, Dinesh Manandhar, Raluca Gordan and Gregory E. Crawford in Epigenetic & Chromatin 2016, 9:15 (doi:10.1186/s13072-016-0065-5; https://doi.org/10.1186/s13072-016-0065-5) (Frank et al., 2016). The original article is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/2.0/). The analysis in this chapter is similar to the classification analyses in Chapter 2, and as such is another application of that approach. As highlighted below, the findings helped establish epigenomic and in vivo TF-binding differences between a class of enhancers that bind with PU.1 and open up significantly versus a class containing enhancers that bind with PU.1 but maintain their accessibility upon histone deacety-
lase inhibitor (HDACi) drug treatment in a leukemia cell line, K562. Details leading up to this analysis and the findings are discussed below.

5.1 Background

Epigenetic regulators are frequent targets of oncogenic mutation (Cancer Genome Atlas Research, 2014). As such, molecules that can inhibit or reverse aberrant epigenetic states show potential anti-cancer treatment potential (Yoo and Jones, 2006). HDACi (short for Histone deacetylase inhibitor) represents a class of such anti-cancer molecule that inhibits the histone deacetylation in histone lysines leading to greater chromatin accessibility for TFs to bind genome-wide (West and Johnstone, 2014; Xu et al., 2007). HDACi treatment has shown to induce erythrocytic lineage differentiation in myelogenous leukemia cell line, K562 (Andersson et al., 1979; Kruh, 1982). Nevertheless, HDACi-mediated details of anti-cancer differentiation is largely unknown. Previous studies have reported that only \( \sim 10\% \) of genes change their expression in response to HDACi treatment, which indicates that these compounds are more specific in transcriptional response they induce than a global histone hyperacetylation suggests. Additionally, HDACi, despite its potential, has severe dose-limiting toxicities (Guha, 2015). This highlight a need for better understanding of genomic and epigenomic changes that result from HDACi treatment. This study is aimed towards this goal. Below, I summarize the experiments and major findings leading up to my computational analysis that is then discussed in the subsequent sections below.

In order to understand the dynamics of chromatin and transcriptional response to HDACi, C. Frank, the first author in the paper, measured genome-wide chromatin accessibility (DNase-seq) and gene expression (RNA-seq) changes induced by sub-lethal HDACi treatment for 72 hours in the myelogenous K562 leukemia cell line. As the cell proliferation slowed, several thousand DHS sites were identified to signif-
icantly change their accessibility profile. This was in alignment with the direction of gene expression change observed for the genes nearest to these differential DHS sites. Two HDACi molecules, NaBut and SAHA, tested showed overlapping but dosage-dependent response in chromatin remodeling.

C. Frank performed motif enrichment analysis and identified that ETS family motif, closely matching that of its member PU.1, was unique to HDACi-induced DHS (hereafter called “HDACi-opened”) DHS sites. Interestingly, PU.1 was the only ETS factor significantly upregulated following HDACi treatment. PU.1, also known as SPI1, is a pioneer factor in macrophage and B lymphocyte differentiation, and has also been shown to induce myeloid lineage commitment in hematopoetic pro-genitor cells (Nerlov and Graf, 1998), and even fibroblasts when expressed together with C/EBPα/β (Feng et al., 2008). Therefore, PU.1 was likely involved in inducing chromatin opening at HDACi-opened sites in K562 cells. To test this, C. Frank performed ChIP-seq experiments to identify PU.1-binding sites in SAHA-treated K562. As predicted, HDACi-opened DHS sites were highly enriched with PU.1 binding, relative to closed or all DHS sites in K562. However, not all PU.1-bound sites showed increase in DNA accessibility. HDACi-opened PU.1-bound DHS sites opened more than HDACi-opened PU.1-unbound or PU.1-bound but “HDACi-unresponsive” DHS sites. This indicated that HDACi opens only a subset of PU.1-bound sites.

In order to test whether PU.1 is in fact necessary to open HDACi-opened PU.1-bound DHS sites, Frank performed DNase-seq and RNA-seq experiments on shRNA-mediated PU.1 knockout K562 cells with HDACi treatment. Surprisingly, depletion - albeit not complete absence - of PU.1 produced no significant difference in chromatin accessibility from SAHA-treated K562 after 72 hours (at adjusted p-value cutoff of 0.10), and only produced expression change in a minimal number of genes showing HDACi-induced differential expression. These results therefore suggested that while PU.1 binding correlates with HDACi-induced chromatin and gene expression change,
PU.1 is not required for a vast majority of HDACi-induced changes observed in K562.

In order to test whether PU.1 co-factors or epigenetic modifications are likely predictive of HDACi-induced genomic response, I then performed the predictive computational analysis discussed in the rest of the chapter. In particular, the goal of the analysis involves identifying factors that can distinguish HDACi-opened PU.1-bound DHS sites from PU.1-bound but HDACi-unresponsive DHS sites. Using all available TF and histone modification ChIP-seq data for K562 cell line in Encyclopedia of DNA Elements (ENCODE) project, the random forest classifier identified that HDACi-opened PU.1-bound DHS sites are epigenetically poised (marked by H3K27me3 and H3K4me1 histone marks) and bound by GATA1 and TAL1, which show increased expression following HDACi treatment, as is discussed below. Further experiments are needed to validate whether these TFs necessarily act as regulators of HDACi-induced enhancers.

5.2 Methods

5.2.1 Classification of opened versus stable sites

In order to identify transcription factors and epigenetic marks associated with increased accessibility of DHS sites bound by PU.1, we implemented a random forest classifier to distinguish sites that open from those that do not significantly change accessibility following HDACi treatment. We used all ChIP-seq data available through ENCODE (genome assembly hg19) for the untreated K562 cell line (“treatment = None”) with the exception of experiments labeled with “revoked” status. A total of 112 transcription factor, 15 histone modification, and 18 chromatin-modifying factor ChIP-seq alignments were downloaded and processed with MACS2 (v2.10) (Zhang et al., 2008) to generate fragment pileup scores (bedGraph format) over input control. Before classification, we removed all DHS sites localizing to known gene promoters or exons (based on UCSC hg19 known genes). Next, to control for the
DNase-seq and PU.1-binding signal in the sets of opened versus stable sites used for classification, for each opened DHS site we randomly selected a stable DHS site with close to equal baseline DNase-seq and PU.1 ChIP-seq signal in untreated K562 cells. DHS center ± 200 bp was used for DNase signal, and DHS center ± 150 bp was used for PU.1 ChIP signal matching. This resulted in ~ 930 opened and stable sites used for classification.

For each selected DHS site, transcription factor features were computed as the maximum ChIP-seq pileup signal over 200-bp windows with 100-bp overlap in the DHS site (defined as DHS site center ± 300 bp). For chromatin-modifying factors, we used a larger region of DHS center ± 700 bp. For histone modifications, we summed the total ChIP-seq signal in the DHS center ± 700 bp. The random forest classifier was run through R package randomForest with mtry set to 10 and ntree set to 500. 75% of the input data were used to train the model and 25% reserved for testing. The random splitting of training and testing samples followed by classification was repeated 10 times to assess the stability of the top features and classification accuracy. The importance score of each feature was computed as the Gini index value and the mean decrease in accuracy when its class labels are randomly permuted. Heatmap and summary plots were generated using deepTools software (Ramirez et al., 2014).

5.3 Results and conclusion

As discussed in the Background section above, PU.1 binding was not necessary to mediate HDACi-induced genome-wide changes in chromatin modification and gene expression in K562 cells. In order to identify additional transcription factor or any histone modifications that might be required for HDACi-induced chromatin opening and potential regulatory activation at PU.1 bound sites, ChIP-seq data for all available TFs (n = 112), histone modification marks (n = 15) and chromatin modifying
factors \((n = 18)\) in K562 cells were obtained from ENCODE repository (Thurman et al., 2012). These were used in a random forest (RF) classifier (see Chapter 1) to identify most informative discriminatory features between between PU.1-bound DHS sites that open with SAHA treatment and those that remain stable in accessibility. Using random 75% of the input DHS sites to train and rest to test the model on ten separate classification runs, the RF classifier consistently performed better than chance (mean accuracy=72.5%, Figure 5.1a).

The analysis identified H3K27me3 as the top predictors for HDACi-opened DHS sites (Figure 5.1b-e). Chromobox proteins, Cbx2 and Cbx8, which are components of Polycomb repressive complexes involved in maintaining H3K27me3 are also enriched in these HDACi responsive sites (Figure 5.2b,c). While not predictive in distinguishing between opened and unchanged DHS sites, H3K4me1 signal also appears enriched in all PU.1-bound DHS sites ahead of HDACi treatment (Figure 5.2a). Joint enrichment of H3K27me3 and H3K4me1 at HDACi-opened PU.1-bound DHS sites indicates that these sites are epigenetically poised (Heinz et al., 2015b). In contrast, PU.1-bound, HDACi-unresponsive sites appear to be activated enhancers prior to HDACi treatment, as indicated by comparative enrichment of H3K36me3, H3K4me2, RNA polymerase II (Pol2), H3K27ac (Figure 5.1b) and H3K4me1 (Figure 5.2a) (Heinz et al., 2015b). In addition to H3K27me3, GATA1 and TAL1 ChIP-seq signal are also enriched in HDACi-opened sites (Figure 5.1b,d,e). In line with this, TAL1 and GATA1 also significantly increase their expression in response to HDACi treatment. It has been shown that TAL1 is necessary for chromatin looping between the \(\beta\)-globin locus control region and the \(\gamma\)-globin gene, which upregulates its expression in K562 cells (Yun et al., 2014). GATA1, on the other hand, has been shown to be antagonistic to PU.1, but is also known to induce erythroid differentiation (Nerlov et al., 2000; Zhang et al., 2000). Currently, the mechanism by which either of these factors together interact to control the enhancer activity of HDACi-opened,
Figure 5.1: Chromatin features distinguish HDACi-opened from stable DHS sites bound by PU.1. (a) Performance of random forest classifier in distinguishing PU.1-bound DHS sites that remain stable in accessibility from those that significantly open. Plot shows accuracy, specificity, and sensitivity for 10 separate runs with 75% of input data used for training and the remaining 25% used for testing. (b) Ranking of top 15 features by Gini Index. The same factor may appear more than once if data are available from multiple ENCODE centers. Broad Broad Institute, HAIB Hudson Alpha Institute for Biotechnology, UW University of Washington, SYDH Stanford, Yale, Davis, Harvard. Full results can be found in the supplemental file corresponding to this chapter. (ce) Top positive predictors of PU.1-bound DHS site opening are shown as heatmaps of ChIP-seq signal present at each DHS site (DHS center ± 5 kb for H3K27me3 and ± 1.5 kb for TFs). Mean ChIP-seq signal summary plots are shown underneath the heatmaps for the same genomic intervals. (This figure corresponds to Figure 6 in the original paper (?).)

PU.1 bound sites is unclear. Further experiments targeting these factors in HDACi treatment is needed to understand the role of these regulatory factors in anti-cancer erythroid differentiation of the leukemia K562 cell line.
Figure 5.2: H3K4me1 enrichment and CBX2/8 binding at PU.1-bound DHS sites. Heatmaps of ChIP-seq signal present at each DHS site (DHS center ± 5 kb) for the H3K4me1 mark (a), and chromobox proteins CBX2 (b) and CBX8 (c). Plots underneath the heatmaps show the corresponding mean ChIP-seq signal. Note that H3K4me1 is similarly enriched in both opened and stable PU.1-bound DHS sites, and was not picked as a predictive factor by the classifier. On the other hand, CBX2 and CBX8 display stronger signal in a subset of opened DHS sites and are therefore informative features for the random forest analysis presented in Figure 5.1. (This figure corresponds to Supplemental Figure 7 in the original paper (Frank et al., 2016).)
Cellular reprogramming processes are biological marvels. Over-expression of one or a few key TFs “trickles down” the cellular system to bring phenotypic as well as functional reflection of a differentiated cell type. Many early reprogramming studies focused on the expression of key target cell type marker genes to assess or quantify the level of the target phenotype observed in this process. Now, with the technological advancements that the modern biology is seeing, especially in next generation sequencing (NGS), we have a better understanding that such direct (as in transdifferentiation studies) or directed (as in conversions of ESC or iPSC) reprogramming processes involve genome-wide expression change of thousands of genes and epigenetic profile of thousands of chromatin sites. Clearly, a better understanding of these genome-wide chromatin and gene-expression dynamics is essential to understand the extent, efficiency and limitations of reprogramming processes in relation to the natural differentiation processes and the target cell types. This requires, on the one hand, thorough analyses of the NGS data in relation to the starting and target cell types, and, on the other hand, the development of novel methods that give additional perspectives to quantifying and understanding the reprogramming efficiency. This
thesis is a step forward in these directions.

In this thesis, myogenic cellular conversion is used as a model reprogramming system. The myogenic regulator, MyoD, being the first reprogramming factor discovered, has been studied in a number of MyoD-induced myogenic conversion studies, both in culture and in vivo mouse models. As such, MyoD-induced cellular reprogramming is a well-characterized model system. Nevertheless, as we discussed in Chapter 1, a number of unanswered questions remain, especially in light of genome-wide chromatin and gene expression changes that MyoD induces or fails to induce. Therefore, MyoD-mediated reprogramming offered a good platform for the computational analyses employed or developed in this thesis.

One of the methods developed is the Chromatin Reprogramming Level (CRL) measurement, which quantifies the degree that chromatin accessibility changes during differentiation process in relation to the starting and target cell type. Indeed, the CRL measurement is easily applicable to other NGS data types, such as histone modification, TF binding and even gene expression data. We showed in Figure 2.22, Chapter 2 that the degree that a gene “gets reprogrammed” can be captured using a similar measure: GRL, which is short for “Gene Reprogramming Level”. We also showed that for the case of MyoD-induced myogenic conversion of fibroblasts, the correlation between GRL and promoter-specific CRL is high, which shows that the level to which genes get upregulated are partly explained by the extent to which promoter-specific chromatin is made accessible. Including distal chromatin sites also show that genes that are upregulated and reprogrammed in the process have significantly higher level of chromatin accessibility attained at the cis-genic reprogrammed sites compared to genes that remain down-regulated and therefore, non-reprogrammed. This finding would have been missed by simply analyzing binarized chromatin accessibility profiles (e.g. using chromatin models like chromHMM (Ernst and Kellis, 2012)) without downstream quantitative analyses.
One thing to note in relation to the CRL score is that while this method has been developed to analyze a single reprogrammed or transdifferentiated state in relation to starting and target cell types, CRL scores can also be applied to time-course experiments. In this scenario, CRL scores can be measured for every time point in relation to the same starting and target cell types, and the resulting time-course CRL profile can then be analyzed to identify, for instance, chromatin sites that increasingly get reprogrammed, or not, over time. Advantages of using time-course CRL over the usual time-course DNase (or other NGS data) are that: (i) the time-course CRL measurements are already normalized in context of starting and target cell types, and hence, need not include these for downstream clustering or other analyses, and (ii) the measurements can be easily interpreted as the level of reprogramming observed for the site or gene under consideration. Nonetheless, one drawback exists. This approach to time-course analysis would only be feasible for genomic regions or genes that show significantly differential signal between starting and target cell type, as CRL scores are not applicable when the starting cells are not different from target cells for specific sites or genes.

Another analysis approach that this thesis highlights is the comparative classification between different states of chromatin accessibility and gene expression states. Both Random Forest (RF) and Elastic Net (EN) implementations have shown that discriminatory genomic and epigenomic factors can explain differences in reprogrammed versus non-reprogrammed chromatin states. These factors can give biological insights, as in the cases of (i) MyoD-induced reprogramming of fibroblasts (discussed in Chapter 2), where DNA-binding specificity of myogenic regulatory factors (MRFs) and known cofactors can explain why MyoD binding and chromatin reprogramming is different for reprogrammed and non-reprogrammed sites, and (ii) HDACi-treatment in K562 cells (discussed in Chapter 5), where DHS sites that show hyper-accessibility after the drug treatment were identified to be potentially epige-
netically poised. In addition, these RF and EN classifications can also help generate testable hypotheses for converting the “negative” (i.e. non-reprogrammed) class elements to the “positive” elements. We identified that during MyoD-induced transdifferentiation of fibroblasts, DHS sites that are non-reprogrammed show higher binding specificity for SAND-domain factors compared to sites that are reprogrammed. In light of the facts that SAND-domain factors are not highly expressed in fibroblasts and that SAND domain containing factors (like SKI) have been shown to induce myogenic conversion previously (Colmenares and Stavnezer, 1989), one future downstream analysis resulting from the classification analysis is to assess joint reprogramming effects of MyoD and Ski, for instance. In addition to these model-based comparative approaches, a direct comparison of class-specific distributions could also clearly be implemented. Our comparison of reprogrammed and non-reprogrammed genes in Chapter 2 showed that chromatin remodeling deficiencies around the genes are correlated with the state, and even level, of expression for the genes.

Reprogramming studies highlight the need for an integrative analytical approach that can assess or help improve reprogramming efficiency. In Chapter 3, a new model is developed that can predict gene expression using the local chromatin accessibility profile and expression levels of regulatory factors in a cross-cell type framework. As discussed in Chapter 3, this is the first cross-cell type model that integrates the expression levels of TFs with chromatin accessibility data. As such, a non-linear combinatorial effect of regulatory TFs and chromatin sites on the target gene expression can be assessed. The model developed, \texttt{CPGex}, is aimed to predict the effect of atypical or missed chromatin remodeling events or regulatory TF expression that is common for non-reprogrammed genes, as evidenced in these chapters. However, \texttt{CPGex} can also establish a relative ranking of sites or TFs that could have the most prominent effect in reprogramming the expression of the gene. These regulatory players could then be targeted for efficient gene-specific reprogramming.
CPGex can be further improved. The model performance can be improved with a larger number of cell lines used as training data, as well as ensuring that the list of regulatory TFs predicted for the target gene (TG) is accurate. Using a gene regulatory network (GRN) prediction model with higher precision and recall rates than what is used currently (CellNet) will increase the predictive power of CPGex. Nevertheless, the GRN-based analysis can also yield interesting insights during the reprogramming events. As discussed in Chapter 4, skeletal muscle media treatment of iPSC activates fibroblast-specific genes in the control iPSC (or PTCI) cells, which endogenous MyoD overexpression (in CR-MYOD cells) does not fully deactivate, as opposed to transgene MyoD overexpression (in TG-MYOD cells). This could mean that endogenous overexpression-mediated reprogramming activates a different differentiation route than transgene overexpression. Further experimental and computational analyses are needed to confirm this.

In conclusion, how can myogenic reprogramming be made more efficient? As discussed above, our study presents SAND-domain factors as candidate factors for MyoD-induced fibroblast reprogramming. Further experiments are needed to validate their significance. However, other studies have identified some ways that could be implemented for enhanced myogenic conversion. As discussed in Chapter 1, in the case of MyoD-induced reprogramming of fibroblasts, introducing BMP inhibitor (BMPi) has been shown to differentiate MyoD overexpressed cells away from proliferative myoblast state towards terminal differentiation with high efficiency (Cacciarelli et al., 2017). Therefore, downstream computational analyses as in Chapter 2 following BMPi exposure to transdifferentiating fibroblasts would give insights into how BMPi remodels the chromatin and gene expression genome-wide. For the case of MyoD-directed reprogramming of iPS cells in Chapter 4, overexpressing BAF60C, which is part of the chromatin modifying SWI/SNF complex (Albini et al., 2013), can help improve the MyoD-mediated reprogramming observed. In addition, it has
also been shown that C2C12 myoblast cells show significantly improved terminal differentiation with bioflavinoid (Hesperedin) exposure (Jeong et al., 2011), suggesting that this small molecule could also improve myogenic conversion in primary human iPSC or differentiated cell types.

Chromatin modifying enzymes present an interesting class of factors that need further study in MyoD-mediated or other cellular reprogramming processes. The analyses in Chapter 2 have shown that the pre-existing histone modifications associated with active chromatin are significantly more enriched in fibroblast-specific DHS sites that fail to close down compared to the sites that do. In addition, our preliminary analysis showed that DNA-methylation, which is known to be a repressive epigenetic mark, is also differentially enriched between the classes. Specifically, myoblast-specific sites that open up have significantly less pre-existing DNA methylation in fibroblasts than myoblast-specific sites that do not open up. These observations could be a reflection of epigenetic hurdle that explains the reason behind “memory-like” states in chromatin and gene expression observed after reprogramming. Epigenetic bottleneck is also largely thought to be the reason behind varied reprogramming efficiency of MyoD across cell types. As with the case of BAF60C in ES cells, targeted upregulation of key epigenetic regulator could potentially improve the overall reprogramming efficiency. One such epigenetic regulator of interest for fibroblast reprogramming is Uhrf1, which is known to recruit DNA methyl transferase 1 (DNMT1) enzyme, as well as participate in chromatin modification, and negatively regulate transcription (Bostick et al., 2007). *UHRF1* was found to be significantly under-expressed in normal as well as transdifferentiated fibroblasts compared to myoblasts.
Appendix A

Proofs

A.1 Proof that increasing MyoD concentration increases MyoD occupancy at lower affinity sites.

Let $S$ represent a potential binding site for MyoD. The binding event of MyoD at $S$ that forms the complex $MyoD \cdot S$ can be represented as follows:

$$ MyoD + S \rightleftharpoons MyoD \cdot S \text{ complex.} $$

When the rate of formation of $MyoD \cdot S$ complex is the same as rate of its dissociation, i.e. when the system is in equilibrium, the dissociation constant ($K_d$) of TF MyoD for site $S$, which is inversely proportional to its affinity for site $S$, is given by:

$$ K_d = \frac{[MyoD][S]}{[MyoD \cdot S]} . \quad (A.1) $$

Using equation (A.1), we have the probability that site $S$ is bound by MyoD as:

$$ P(S \text{ bound}) = \frac{\text{Concentration of } S \text{ bound by MyoD}}{\text{Total concentration of } S} $$

$$ = \frac{[MyoD \cdot S]}{[MyoD \cdot S] + [S]} $$

$$ = \frac{[MyoD]}{[MyoD] + K_d} \quad (A.2) $$
which implies that increasing MyoD concentration will increase MyoD occupancy at any site S. Here, we are interested in the magnitude of the increase in binding site occupancy (i.e. the increase in $P(S \text{ bound})$), at sites S of various affinity, as the protein concentration increases. From equation (A.2), we have the rate of change of $P(S \text{ bound})$ with respect to $[MyoD]$ as

$$P'(S \text{ bound}) = \frac{K_d}{([MyoD] + K_d)^2}.$$ 

For simplicity, let us consider only two sequences $S_1$ and $S_2$ with dissociation constants $K_{d1}$ and $K_{d2}$ respectively. Also, let $K_{d1} < K_{d2}$, meaning that sequence $S_1$ has higher affinity for MyoD than $S_2$. Then,

$$P'(S_1 \text{ bound}) - P'(S_2 \text{ bound}) > 0$$

$$\Rightarrow \frac{K_{d1}}{([MyoD] + K_{d1})^2} - \frac{K_{d2}}{([MyoD] + K_{d2})^2} > 0 \quad (A.3)$$

$$\Rightarrow 0 \leq [MyoD] < \sqrt{K_{d1} \times K_{d2}}.$$ 

And, similarly,

$$P'(S_1 \text{ bound}) - P'(S_2 \text{ bound}) < 0 \Rightarrow 0 \leq \sqrt{K_{d1} \times K_{d2}} < [MyoD]. \quad (A.4)$$

From equations (A.3) and (A.4), we conclude that increasing MyoD concentration will initially lead to a higher increase in MyoD occupancy at the high affinity sites ($S_1$) compared to the low affinity sites ($S_2$), until a critical concentration is reached ($[MyoD] = \sqrt{K_{d1} \times K_{d2}}$). A further increase in MyoD concentration will then lead to a higher increase in MyoD occupancy at the lower affinity sites ($S_2$).
Appendix B

Installing and running CPGex

CPGex, short for Cross cell-type Prediction of Gene EXpression, is the software implementation of the model described in Chapter 3. This appendix describes the details of CPGex installation and running, which are also explained in detail in the Github repository: https://github.com/dineshmdh/predicting_gene_expression.

CPGex uses cross cell/tissue-type cis-TSS chromatin accessibility and known TF-TG (i.e. transcription factor to target gene) interactions for a given gene of interest to predict its expression in a held-out cell line of interest. Currently, this program has been tested for Roadmap Epigenomics Data, that has 127 matching cell/tissue-samples. (Training and testing on ENCODE samples is in progress.)

Dependencies and installation

The program has been developed and tested in Python v2.7 that comes with Anaconda v4.4.7. The Python packages required to run CPGex are:
numpy, pandas, hyperopt, matplotlib, seaborn, logging, functools, pybedtools, re

While a few of these are already pre-installed in Anaconda, which is recommended
over other Python installations, the rest can be installed as follows:

1. conda install -c jaikumarm hyperopt
2. conda install -c anaconda seaborn
3. conda install -c hargup/label/pypi logging
4. conda install -c travis functools
5. conda install -c bioconda pybedtools

In order to run the program,

1. Make sure these python modules are installed.
2. Download and unzip source code.
3. Download and unzip the input files from https://www.dropbox.com/sh/6u9dth281x6q5as/AACpwMTNjGuctIYN8fA-BOzRa?dl=0, which contains processed Roadmap and CellNet files, and are large to share on the Github repository.
4. Move these input files in a directory named Input_files in the main directory for the package.
5. Go to the /directory-to-package/Functions/model-scripts, and call the python script below.

**Usage**

CPGex can be run using the following command line options.

```bash
```

The details of these arguments are listed below.

Table B.1: CPGex arguments and their descriptions.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-gene</td>
<td>Gene of interest</td>
</tr>
<tr>
<td>-d</td>
<td>Distance from TSS (in kb) to cover as region of interest (Default: 150)</td>
</tr>
<tr>
<td>-u</td>
<td>Use TAD boundaries to demarcate the boundaries for the region of interest. (Default: True)</td>
</tr>
<tr>
<td>-Fd</td>
<td>Take this many DHS sites. If this is set to ‘-1’, then all the known DHS sites in the region TSS ± -d or regulatory TFs is used. Note that if random set of features are to be used, namely by setting ‘-rd’ option, then the same number of DHS sites are considered as in the non-random (i.e. original) set. (See details on ‘-rd’ below.) (Default: 6)</td>
</tr>
<tr>
<td>-dl</td>
<td>Lower limit of the absolute pearson correlation between the DHS site accessibility and expression of this gene to be used in filtering the top dhs sites. All DHS sites with pearson correlation score below this threshold are ignored. (Note, if more than -F remain, only top -F are selected based on the pearson correlation scores. (Default: 0.2)</td>
</tr>
<tr>
<td>-rd</td>
<td>If set, a set of -Fd number of DHS sites are randomly selected from the genome. The DHS sites selected could be from a different TAD domain or chromosome. (Default: False)</td>
</tr>
<tr>
<td>-tff</td>
<td>For the TF-TG association, filter the predicted list of regulatory TFs for the given gene using one of two measures: 1) Pearson Correlation Coefficient between the expression of TF and the target gene TG, or 2) Z-score indicating the significance of one TF-TG association given perturbation measurements of the expression of the TF and the TG across various experimental or biological conditions (see CellNet paper and CLR algorithm). (Default: ‘zscores’)</td>
</tr>
<tr>
<td>-Ft</td>
<td>Take this many TFs. If this is set to ‘-1’, then all the known TFs that are predicted to be regulatory for the gene are used. Note that if random set of features are to be used, namely by setting ‘-rt’ option, then the same number of TFs are considered as in the non-random (i.e. original) set. (See details on ‘-rt’ below.) (Default: 8)</td>
</tr>
</tbody>
</table>
Lower limit of the measure \(-\text{filter-tfs-by}\) in absolute value. The value should be \(>0\) for \(-tff\) and \(\geq 4.0\) for \(-tff\) zscores. Note that the respective upper limits are 1.0 and infinity respectively, and therefore need not be declared. (Default: 5.0 for the default \(-tff\) zscores.)

Relates to the initial wts set between the nodes. Can be one of ‘random’ or ‘corr’. If ‘random’, random initial wts are set between any two nodes; if ‘corr’, initial wts between input and hidden nodes are set to the correlation values between the node feature and the expression of the gene, and the initial weights between hidden layers or the hidden layer and output is set to 0.5 (Default: ‘corr’)

Maximum number of interactions for neural net optimization (Default: 300)

Output directory. A directory for this gene of interest and set of parameters used is created at this location. (Default is ‘../Output’)

Run id for multiple parallel runs. This is useful in slurm. (Default: -1)

If set, all supplemental plots are also generated in addition to the scatterplots showing the performances after hyperparameter optimization and re-training (i.e. training the full training and validation set using the optimized hyperparameters). (Default: Not set)

Some of the examples are shown below.

- `python main.py SIX1` # runs with default parameters

- `python main.py SIX1 -d 200 -Fd 10` # considers top 10 DHS sites in TSS \(\pm 200\)kb region that most correlate with the expression of SIX1 across the samples

- `python main.py SIX1 -Ft 20 -m 200 -o /output/dir` # considers top 6 DHS sites (by default) and top 20 TFs that are predicted to regulate the target gene - `SIX1`; runs 500 iterations (instead of default, 300) for training the neural net; send all output to the designated directory.


Edmondson, D. G. and Olson, E. N. (1989), “A gene with homology to the myc similarity region of MyoD1 is expressed during myogenesis and is sufficient to activate the muscle differentiation program,” Genes Dev, 3, 628–40.


Biography

Name: Dinesh Manandhar

Born: March 2, 1988 in Banepa, Nepal

Past Degree: Bachelor of Arts in Mathematics, Gettysburg College, Pennsylvania

Papers published:

1. Manandhar et al. (2017)
2. Frank et al. (2016)
4. Shats et al. (2017)

Future prospect: Currently, looking for Computational Biologist positions in Research Institutions and Industries

Github repo: https://github.com/dineshmdh