Transcription factor occupancy in differentiating skeletal muscle

Thesis by

Anthony Kirilusha

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2014

Defended May 27th, 2014

Acknowledgements

First and foremost, I would like to thank my adviser Barbara Wold, as she has been the single most influential person during my time as a graduate student and a young scientist-in-training. Without her support and tremendous patience while I learned what it takes to pursue a scientific career, I would never have had the opportunity to complete this thesis or to receive a doctoral degree. Along with imparting great enthusiasm and respect for the work, Barbara taught me the most important skills that need to be acquired in graduate school - how to formulate a project, take ownership of it, see the research through, and report it with thoroughness and awareness of the context in which it belongs. I am not sure that I can adequately express my gratitude in writing, but I will say that if for the next 50 years I can pursue my work with as much passion, skill, intellectual integrity and foresight as I have witnessed from Dr. Wold, I will consider it an unequivocal success.

I would also like to thank the other members of my thesis committee, consisting of Dr. Angelike Stathopoulos, Dr. Paul Sternberg, Dr. Ellen Rothenberg, and Dr. Eric Mjolsness. I consider myself very fortunate to have found a group of professors that so enthusiastically supported my work, challenged my interpretations, and contributed different perspectives to problem formulation and analysis. I wish that my future interactions and discussions can be as productive and smooth as those I've had with my committee members. Additionally, while Dr. Erik Winfree was not on my thesis committee, he helped me navigate the Computer Science department during the early parts of my graduate career, and was very gracious in his support for my option change to Biology. Without Dr. Winfree's help, I would not have had the opportunity to be at Caltech, nor would I have successfully completed a Master's degree in Computer Science.

All the members of the Wold group, both past and present, whom I've had the pleasure of working with over the years deserve a collective "Thank you." In particular, I would like to thank Chris Hart and Titus Brown for their help, both scientific and personal, as I was first getting started. I also would like to thank Brian Williams for his exceptional willingness to teach both experimental and analysis techniques - his presence has been a gift to both myself and all my fellow graduate students. Special thanks goes to Ali Mortazavi, Katherine Fisher, Gilberto DeSalvo and Georgi Marinov, who in addition to being fantastic labmates have contributed significant amounts of data, software and discussion that went into my thesis. Finally, given the importance of computation in processing and analyzing genomic data, none of this would have been possible without our excellent team of system administrators. Diane Trout, Henry Amrhein, Sean Upchurch and Brandon King have done everything from building and supporting the computer cluster to writing and debugging code used for data analysis, and, quite remarkably, I do not recall a single time where my work was slowed down or halted because of computer-related issues in the last 8 years.

Some people who are not at Caltech deserve recognition and thanks for their contributions to my research efforts. Dr. Cornelis Murre kindly donated the E47 and E12 plasmid templates that I used for the *in vitro* binding assays, and his experiments measuring E47 occupancy in B-cells were integral to the analysis discussed in chapter 4. Dr. Stephen Tapscott has provided discussion, access to data, and general inspiration for my work, especially as it related to the function and occupancy of MyoD. Dr. Sandra Sharp helped me formulate the initial set of questions as I was setting out to study transcription factor occupancy. Dr. Anne Reifel-Miller supervised my internship project in human genetics at Eli Lilly and Co, and enabled me to make substantial progress on the problem despite the relatively short time-frame, as well as to make a key observation that would prove influential in my own research of skeletal muscle differentiation. She is a very gifted scientist, an excellent leader, and an exceptionally warm and kind person - I could not have wished for a better supervisor and mentor. Dr. Ross Hardison has given

me access to the Tal1 and Gata1 occupancy maps generated in his lab, and while I do not discuss them in the thesis, they were very helpful to the overall understanding of the role and function of bHLH transcription factors in various developmental lineages. Dr. Fancis Collins has given me a superb opportunity to pursue a postdoctoral project in the genomics of type 2 diabetes, and showed a lot of patience while I was completing my defense - both the enthusiasm for the upcoming project and the knowledge that I have a place to go to once I am done at Caltech have been invaluable as I worked to bring my thesis work to a close.

Finally, I would like to thank my friends and loved ones who over the years displayed great love and personal loyalty, as well as providing me with a sense of connectedness and the support structure I needed to persevere (and maintain my sanity). Especially I would like to thank my parents, for their unconditional love. They went above and beyond in their efforts to help me succeed, both materially and intellectually, and while at times it was difficult to accept their admonishments, I think I am a better person for it. Once again, words don't seem enough to express the true measure of gratitude I feel for my parents, so I'll try to be succinct: "Thank you mom. Thank you dad. I love you."

Abstract

With recent advances in high-throughput sequencing, mapping of genome-wide transcription factor occupancy has become feasible. To advance the understanding of skeletal muscle differentiation specifically and transcriptional regulation in general, I determined the genome-wide occupancy map for myogenin in differentiating C2C12 myocyte cells. I then analyzed the myogenin map for underlying sequence content and the association between occupied elements and expression trajectories of adjacent genes. Having determined that myogenin primarily associates with expressed genes, I performed a similar analysis on occupancy maps of other transcription factors active during skeletal muscle differentiation, including an extensive analysis of co-occupancy. This analysis provided strong motif evidence for protein-protein interactions as the primary driving force in the formation of Myogenin / Mef2 and MyoD / AP-1 complexes at jointly-occupied sites. Finally, factor occupancy analysis was extended to include bHLH transcription factors in tissues other than skeletal muscle. The cross-tissue analysis led to the emergence of a motif structure used by bHLH TFs to encode either tissue-specific or "general" (public) access in a variety of lineages.

Table of Contents

Chapter 1: Introduction

1.1 Overview

The core of this thesis will focus on the study of the transcriptional regulation of muscle differentiation. In mammals, this is a stepwise process in which multipotential mesodermal precursors first give rise to unipotential myoblasts, characterized by their ability to proliferate and migrate. The myoblasts in turn differentiate into post-mitotic myocytes, at which point most of the molecular markers associated with muscle are expressed. Myocytes further undergo membrane fusion and form multinucleated mytubes, which mature into functional myofibers, completing the process. To facilitate the genomic study of myogenesis, several myoblast cell lines have been cultured over time, capable of recapitulating the transition with high efficiency. One such cell line - murine C2C12s (utilized very commonly in the field) - was used as a model for my experimental work. A key characteristic of the C2C12 cells is that they are capable of proliferation, but already express MyoD - hence they are at the myoblast stage, having passed beyond the unipotential threshold. They do not begin differentiating into myocytes until proper growth and feeding conditions are met. Therefore, it would be more precise to say that this thesis focuses on the genetic network responsible for the myoblast to myocyte transition, and then on the maturation of the myocyte into a myotube. For a review of the precursor network, see Buckingham and Rigby 2014.

1.2 Primary transcriptional regulators of myogenesis

Since a key feature of C2C12s is the expression of MyoD, we should begin by examining what that means. The "birth" of the myogenic network can be traced back to an observation that took place over 30 years ago - that 10T1/2 cells treated with 5-azacytidine can turn into adipocytes, myoctes or chondrocytes (Taylor and Jones, 1979), with myogenic conversion occurring 25 - 50% of the time. This is consistent with 10T1/2s having the developmental characteristics of multipotential somitic cells. Since 5-azacytidine essentially releases methylation restrictions during the synthesis of the daughter genome, the working hypothesis first proposed by Hal Weintraub and colleagues was that there must be a single dominant-acting factor that can commit cells to a myogenic fate, with the downstream program executed upon meeting the required growth conditions. This would account for the high rate and relative ease of the conversion to myoblast, and indeed led to the cloning of the MyoD cDNA by Davis et al. (1987). Upon transfection under a constitutively active promoter, MyoD initiated a stable myogenic conversion of 10T1/2s and other fibroblast lines. It turned out that MyoD was one of four closely related mammalian genes, all possessing the dominant myogenic property - together, they are commonly referred to as MRFs (muscle regulatory factors). The multiple MRFs have a single ortholog in most invertebrate genomes, including hlh1 in *C. elegans* and nautilus in *D. Melanogaster*. All MRFs are sequence-specific DNA-binding proteins of the bHLH family, and function as heterodimers with E-proteins (also bHLH transcription factors). The dimerization and its implications for DNA binding will be discussed in more detail in the next section of the introduction.

Of the four MRFs, MyoD and Myf5 (Braun et al. 1989) are the most similar functionally and structurally. They are preferentially expressed in myoblasts, and a joint $MyoD^{-/-}$ $Myf5^{-/-}$ knockout results in a complete failure to develop skeletal muscle (Rudnicki et al. 1993), with animals dying immediately upon birth. Single knockouts compensate each other and result in an apparently normal muscle phenotype

(Rudnicki et al. 1992; Braun et al. 1992), although the MyoD$^{-/-}$ genotype was subsequently associated with several subtler abnormalities. Specifically, such mice were viable, but had an impaired capacity to regenerate skeletal muscle following injury (Megeney et al. 1996), while satellite cells carrying the MyoD$^{-/-}$ knockout form unusual aggregate structures, fail to fuse efficiently, show a severe reduction in differentiation efficiency, and are Mrf4 deficient (Cornelison et al. 2000). The initial experiment involving a Myf5$^{-/-}$ knockout led to an unexpected phenotype with severe abnormalities in rib development, resulting in immediate postnatal lethality due to suffocation (Braun et al. 1992). Despite that, the morphology of skeletal muscle in those animals remained unaffected. After some complex issues involving neomorphic effects of early knockout constructs were resolved, the phenotype of Myf5$^{-/-}$ was essentially the same as that of MyoD$^{-/-}$, further reinforcing the idea that MyoD and Myf5 are able to compensate for one another, at least insofar as the generation of viable skeletal muscle is concerned. While these findings may appear surprising due to the perceived importance of MyoD to muscle specification and the initiation of the differentiation cascade, they are not altogether unexpected given the extensive sequence homology shared by the two proteins and their contemporaneous expression. Because of the extensive work surrounding MyoD, and its importance to the system, a portion of the work presented in chapter 3 will focus on analyzing its occupancy in both cycling and differentiating C2C12s. Conversely, we will not focus much attention on Myf5, primarily due to its much lower mRNA levels compared to MyoD (Table 3.7).

Myogenin (Myog), identified and isolated by Wright et al. (1989), is functionally unique among MRFs and therefore central to this thesis. A Myog$^{-/-}$ knockout leads to a severe malformation of all skeletal muscle, and is also lethal at birth, largely due to the absence of a functioning diaphragm and the resulting asphyxiation (Hasty et al. 1993). This makes myogenin indispensable for proper myotube formation, and its expression pattern reflects that property. While MyoD and Myf5 are present in cycling and undifferentiated myoblasts, myogenin is generally not expressed until entry into terminal

differentiation, at which point its drastically up-regulated *in vivo* and *in vitro*. This puts myogenin downstream of MyoD and Myf5, both temporally and in the regulatory cascade - myogenin is a direct target of MyoD and Myf5. What makes myogenin unique is that its the only member of the MRF family whose sole absence leads to a failure to form mature muscle - there appears to be no compensatory mechanism to account for a lack of myog. This served as the impetus for making myogenin the primary focus of chapter 2.

Finally, Mrf4 (Rhodes and Konieczny 1989), independently discovered as herculin (Miner and Wold 1990) and Myf6 (Braun et al. 1990), was isolated and characterized based on its extensive sequence similarity to the other 3 members of the MRF family. Mrf4 shares the ability to initiate fibroblast conversion into myocytes, and is the only one to be prominently expressed in mature muscle (both in mice and humans). The other three MRFs (MyoD, Myf5, Myogenin) are generally absent in healthy mature muscle, although they are up-regulated during muscle repair and regeneration due to the activation of satellite cells. An Mrf4 knockout results in a corresponding reduction of Myf5, and as such is essentially an $Mrf4^{-/-}$ $Myf5^{-/-}$ double knockout, with skeletal muscle morphology unaffected (Braun and Arnold 1995). In aggregate, these data suggest that either Myf5 or MyoD must be present for skeletal muscle specification, while myogenin is crucial to the proper completion of the differentiation process. The exact function of Mrf4 is still not well understood, although it is located in very close genomic proximity to the Myf5 gene, which explains the initial difficulties in the generation of knockout animals.

1.3 MRFs and DNA binding

The 4 members of the MRF family are all bHLH (basic helix-loop-helix) transcription factors (Weintraub et al. 1991), and bind DNA as dimers, with the helix-loop-helix motif responsible for dimerization and the basic domain responsible for DNA binding. Both MyoD and myogenin heterodimerize with another member of the bHLH family - E47 - in order to bind DNA (Murre et al. 1989; Chakraborty et al. 1991; Lassar et al. 1991), effectively making E47 a crucial 5th member of the group. However, because the presence of E47 is not sufficient for myogenesis, nor is it a marker unique to skeletal muscle, it is neither classified nor treated as an MRF. In fact, E47 is itself a primary regulator of B-cell differentiation (Murre 1991; Bain et al. 1997; Lin et al. 2010), and this will be used to improve our understanding of the subtleties of sequence-specific targeting.

MyoD:E47 and Myog:E47 heterodimers have for a long time been reported to bind the motif CANNTG (known as an e-box), but recent findings have substantially refined the definition of the recognition sequence for MyoD:E47 (Kophengnavong et al. 2000; Cao et al. 2010; Fong et al. 2012). Specifically, the central nucleotides have a significant impact on binding affinity, with GS being most optimal, and WW least so. Additionally, flanking nucleotides play an important role, with RR being the preferred prefix (although RY and YR are also viable, but not YY). As such, the binding site is really better characterized as RRCAGSTG - part of the work presented in chapter 2 contributed significantly to this refinement. Even so, there are over a million RRCAGSTG motifs in the mouse genome, requiring additional information to identify those directly involved in muscle differentiation.

Two additional points deserve mention. First, E47 is a product of the E2A gene, but it is not the only one. E12 - another splice isoform of E2A - is also a bHLH protein (Murre et al. 1989a), and it is co-expressed with E47 in both muscle and B-cells. The transcript for E47 is more prevalent, by an approximately 2:1 ratio in differentiating muscle. While E47 can homodimerize efficiently and binds

DNA as either an MRF:E47 heterodimer or an E47:E47 homodimer, E12 homodimers lack the ability to bind DNA (Shirakata and Patterson 1995). MRFs can heterodimerize with E12, and show similar preferences when partnered with either E47 or E12. Second, HEB (also known as Tcf12) is another class I bHLH protein expressed in skeletal muscle. MyoD:HEB heterodimers can bind DNA (Hu et al. 1992) and are thought to be active in the later stages of muscle differentiation (Parker et al. 2006; Davie and Londhe 2011). They appear to have the same sequence preference repertoire as their MyoD:E47 counterparts.

1.4 Secondary transcriptional regulators of myogenesis

Several other transcription factors have over time been strongly linked to the process of muscle differentiation. Myocyte Enhancing Factor 2 (Mef2) was first postulated by Gossett et al. (1989) for its ability to interact with enhancers of muscle creatine kinase (ckm) and myosin light-chain 1/3 (mlc-1). It was then shown that myogenin can induce Mef2 (Cserjesi and Oslon 1991) and that Mef2 in turn interacts with a recognition site in the myogenin promoter (Edmondson et al. 1992). This formed the basis for an auto-regulation loop involving myogenin and Mef2, and coupled with the ability of Mef2 to activate highly muscle specific genes led to the conclusion that it must be crucial to myogenesis. While technically correct, the initial results were an oversimplification due to being based upon the detection of a specific protein-DNA complex. We now know that Mef2 is a MADS-homeobox transcription factor, and that in mouse it is not a single gene, but instead a family of four similar yet distinct genes designated a through d (Martin et al. 1993). The four Mef2 genes are highly homologous in the 56 amino-acid MADS domain responsible for DNA binding and dimerization, although they are divergent at their carboxyl termini (Edmondson et al. 1994). They bind DNA in the form of dimers, and with their relative levels changing over the course of differentiation, so changes the exact nature and concentration of the available species. For instance, Mef2a:Mef2a and Mef2a:Mef2d are both able to recognize the cannonical binding site CAT$(W)_4$TAG (Nurrish and Treisman 1995) but show up as distinct bands on a mobility shift assay and are likely present in different amounts in myoblasts versus myocytes (thesis results, chapter 4). They can also be recruited to DNA through protein-protein interactions, because myogenin:E12 heterodimers and Mef2c were jointly co-precipitated as a protein-DNA complex in the presence of an e-box motif but not the CAT$(W)_4$TAG site (Molkentin et al. 1995). Knockout mice that are Mef2a$^{-/-}$ were born alive, but most died within 10 days from cardiac failure likely caused by severe ventricular chamber dilation (Naya et al. 2002). Some Mef2a$^{-/-}$ mice survived to adulthood, and showed a deficiency in cardiac mitochondria, but without obvious structural defects of the heart.

Neither population exhibited any apparent skeletal muscle abnormalities. While this knockout result strongly implies that Mef2a is important to the proper function and development of myocardial muscle, its implications with regard to skeletal muscle differentiation are less clear. In addition to being active in cardiac muscle (Edmondson et al. 1994), members of the Mef2 family also function in neurons (Lin et al. 1996), and remain a studied regulator of myogenesis (Snyder et al. 2013; Liu et al. 2014) due to their known ability to regulate the expression of several muscle-specific genes. Chapter 3 will test some of the expectations about the genomics of Mef2 in differentiating C2C12s and examine its occupancy.

RP58 (also known as ZFP238 and ZNF238) is a POZ zinc-finger repressor first identified by Aoki et al. (1998), and shown to bind the CAGATGT motif. It was also reported that DNA methyltransferase Dnmt3a functions as a co-repressor with RP58 in a manner that does not require its methyltransferase activity (Fuks et al. 2001). The e-box-containing recognition sequence is very similar to the RRCAGMTG site proposed for NeuroD2 (Fong et al. 2012), and not surprisingly RP58 was initially detected in developing neurons and linked to the regulation of neurogenesis (Ohtaka-Maruyama et al. 2007; Okado et al. 2009). However, RP58 is also expressed, albeit to a lesser degree, in skeletal muscle, and its mRNA levels increase substantially upon entry into terminal differentiation (Chapter 3, Table 3.7). Furthermore, RP58 has flexibility in its binding site, and can bind CAGCTGT motifs, which often overlap recognition sequences for myogenin (I observed this behavior in 2008 while working on regulation of human GPR41, albeit without linking it to myogenesis at the time). In 2009 Yokoyama et al. demonstrated that levels of Id2 and Id3 (both historically important repressors of MyoD that will be discussed shortly) increase when RP58 is knocked down. Simultaneously, an RP58 knockout study (Okado et al. 2009) revealed that $RP58^{-/-}$ mice exhibit defects in muscle development, with a significant reduction in the number of myofibers found in the hind-limb and a severe impairment of diaphragm development, resulting in immediate postnatal lethality. This work was followed up establishing that RP58 can interact directly with sequences in promoters of Id1-4, and that its absence leads to increased

expression of reporter constructs (Hirai et al. 2012). While the latter was done in astrocytes, it stands to reason that the same mechanism is applicable in skeletal muscle. The implications of RP58 binding CAGCTGT, which often overlaps MRF:E targets (RRCAGSTG), and suggests a fluid, competitive regulatory apparatus, will be discussed in Chapter 4.

It would be remiss not to acknowledge the Id family of repressors, despite the fact that they are not transcription factors, and therefore fall outside the scope of this investigation. Inhibitor of differentiation (Id) was first isolated by Benezra et al. (1990) during a search for proteins with homology to the HLH domain of MyoD and c-myc (another member of the bHLH family, most famous for its role as an oncogene). While Id does have an HLH domain, which allows it to heterodimerize with members of the bHLH family, it lacks the basic domain, in turn making the resulting heterodimers unable to bind DNA. Id was shown to heterodimerize with E47, E12, and MyoD (Benezra et al. 1990), which drastically reduced the ability of those proteins to interact with known targets, such as the ckm enhancer. The "titrating" nature of Id repression was confirmed through the use of tethered dimers, where MyoD and E47 monomers were joined by a short peptide bridge, virtually ensuring complete dimerization. The resulting MyoD~E47 species was able to efficiently bind target sequences *in vitro* despite a high concentration of Id, and served as a potent initiator of myogenesis in 10T1/2 and NIH3T3 cells - neither line is inherently myogenic (Neuhold and Wold 1993). In mammals, the Id family consists of four homologous members, Id1-4, the first three of which are heavily expressed in myoblasts along with MyoD and Myf5. All three are also significantly downregulated (over 90% reduction in transcript abundance - thesis data) upon cell cycle exit and entry into terminal differentiation. The Id genes competitively mitigate the effects of various tissue-specific bHLH transcription factors in a variety of setting, including hematopoietic (Deed et al. 1998), neuronal (Cai et al. 2000), muscle (Benezra et al. 1990) and adipose (Moldes et al. 1999). While undoubtedly important to curating the function of bHLH TFs, the exact physiological role of Ids in myogenesis remains unclear. In the C2C12 system, the

myoblast stage shows high levels of transcripts for MyoD, Id1 and Id3; with lower levels for Myf5, Id2, Tcfe2a (E12/E47) and HEB (RNASeq data). The initial Benezra et al. (1990) result was extended by Langlands et al. (1997) to show that Id1-3 have a high dimerization affinity for all class I bHLH proteins present in differentiating muscle (E12/E47, HEB, and also E2-2), although their dimerization affinities for members of the MRF family vary by species of Id and MRF. Since MyoD requires either HEB or E47 to bind DNA, their availability appears to be the limiting factor for the formation of the MyoD:E complex. Furthermore, despite high levels of Ids, MyoD is clearly able to perform its role as a positive regulator of transcription, and drive the transition to the myocyte stage (where, as noted above, transcript levels for Id1-3 fall dramatically). Given this, one reasonable interpretation is that the primary role of Ids is to shift the equilibrium away from the formation of E47:E47 homodimers, which are "inappropriate" for myogenic differentiation, with perhaps a secondary role of titrating the concentration of active MyoD. The latter would then help account for the temporal patterns inherent in the network.

ZEB (zinc finger e-box binding protein) is a repressor that, based on its name alone, one might expect to play a role in the regulation of myogenesis. Much like RP58, it was first described in a non-myogenic context - while studying the immunoglobulin heavy-chain (IgH) enhancer, Genetta et al. (1994) cloned and characterized a zinc-finger protein (ZEB) that was able to bind the e-box sequence CAGGTG (an important element of said enhancer). They then demonstrated that ZEB and E:E homodimers bind that e-box *in vitro* in a competitive, concentration dependent manner by abolishing the E:E~CAGGTG complex in the presence of a large molar excess of ZEB, and vice versa. It was then pointed out that ZEB is a mammalian homolog of the *Drosophila* gene Zfh1, which is expressed in muscle precursors and is crucial to the proper development of muscle (Postigo and Dean 1997). The same study indicated that although ZEB is able to act as a transcriptional repressor, the repression is lost upon expression of MyoD, without decrease in the level of ZEB (in fact, the abundance of ZEB1 transcript increases significantly after cell cycle exit, in a manner similar to that of RP58). More recent efforts in identifying

the role of ZEB in myogenesis involved a targeted knockdown of ZEB1 in differentiating C2C12s (Siles et al. 2013), leading to an earlier than expected expression of molecular markers associated with myogenesis and accelerated myotube formation. By examining the occupancy of CAGGTG elements in the promoters of troponin and MyH4, the authors also concluded that ZEB is present in myoblasts but not myocytes, with the opposite true of MyoD occupancy. These data formed the impetus for considering the importance of ZEB in myogenesis vs. B-cell differentiation (discussed in chapter 4), providing further evidence for a fluid attenuation model of transcriptional regulation. The physiological function of ZEB in myogenesis appears to be more about imposing temporal control on MRF-based activation of transcription rather than about direct repression of MRF targets, although instances of the latter cannot be entirely ruled out.

1.5 Aims of the thesis

At the time I began this work, myogenin was a relatively less-studied transcription factor, with the majority of attention in the field focused on understanding the targets and kinetics of MyoD. High-throughput sequencing was just becoming widely available, and the immediate first step was to utilize it in combination with chromatin immunoprecipitation to assess the binding properties of myogenin *in vivo*. The genome-wide map of myogenin occupancy served as the launching pad for the investigation of myogenesis, and is discusses at length in chapter 2. Some of the issues addressed are the number of occupied sites, their distribution relative to annotated TSSes and gene models, motif content, and association with expression patterns of nearby genes.

Since myogenin does not operate in a vacuum, a logical next step was to consider occupancy for other transcription factors that are either directly involved in myogenesis, or are active in C2C12s. Among them are MyoD, Mef2, E47, Fosl1 (one of the components of the AP-1 complex), CTCF, Klf4, and others. Using data either generated in our laboratory or published in the literature, I conducted a comparative analysis of occupancy profiles, their changes during state transition (myoblast to myocyte), and the underlying motif content. These results are presented in chapter 3.

Finally, because E47 is an important regulator of B-cell differentiation, and at the same time a partner crucial to the ability of MRFs to bind DNA and perform their roles, E47 occupancy in B-cells was compared with MRF occupancy in muscle cells. Using underlying sequence biases, a model was formulated to partially account for the enforcement of context specificity, despite the use of TFs that recognize virtually identical binding sites. These findings are presented in chapter 4, with the overall goal of better understanding transcription factor targeting. Chapter 4 also presents a synthesized view of how different transcription factors interact in a physiological setting, with emphasis on dynamic,

competitive interactions, where affinity (and therefore activity) is influenced by both the subtleties of DNA sequence, as well as the availability of active species able to recognize it.

References (chapter 1)

Aoki, K., Meng, G., Suzuki, K., Takashi, T., Kameoka, Y., Nakahara, K., Ishida, R., and Kasai, M. (1998). RP58 associates with condensed chromatin and mediates a sequence-specific transcriptional repression. J Biol Chem *273*, 26698-26704.

Bain, G., Robanus Maandag, E.C., te Riele, H.P., Feeney, A.J., Sheehy, A., Schlissel, M., Shinton, S.A., Hardy, R.R., and Murre, C. (1997). Both E12 and E47 allow commitment to the B cell lineage. Immunity *6*, 145-154.

Benezra, R., Davis, R.L., Lockshon, D., Turner, D.L., and Weintraub, H. (1990). The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell *61*, 49-59.

Braun, T., and Arnold, H.H. (1995). Inactivation of Myf-6 and Myf-5 genes in mice leads to alterations in skeletal muscle development. EMBO J *14*, 1176-1186.

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

Braun, T., Rudnicki, M.A., Arnold, H.H., and Jaenisch, R. (1992). Targeted inactivation of the muscle regulatory gene Myf-5 results in abnormal rib development and perinatal death. Cell *71*, 369-382.

Braun, T., Winter, B., Bober, E., and Arnold, H.H. (1990). Transcriptional activation domain of the muscle-specific gene-regulatory protein myf5. Nature *346*, 663-665.

Buckingham, M., and Rigby, P.W. (2014). Gene regulatory networks and transcriptional mechanisms that control myogenesis. Dev Cell *28*, 225-238.

Cai, L., Morrow, E.M., and Cepko, C.L. (2000). Misexpression of basic helix-loop-helix genes in the murine cerebral cortex affects cell fate choices and neuronal survival. Development *127*, 3021-3030.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L*., et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell *18*, 662-674.

Chakraborty, T., Brennan, T.J., Li, L., Edmondson, D., and Olson, E.N. (1991). Inefficient homooligomerization contributes to the dependence of myogenin on E2A products for efficient DNA binding. Mol Cell Biol *11*, 3633-3641.

Cornelison, D.D., Olwin, B.B., Rudnicki, M.A., and Wold, B.J. (2000). MyoD(-/-) satellite cells in single-fiber culture are differentiation defective and MRF4 deficient. Dev Biol *224*, 122-137.

Cserjesi, P., and Olson, E.N. (1991). Myogenin induces the myocyte-specific enhancer binding factor MEF-2 independently of other muscle-specific gene products. Mol Cell Biol *11*, 4854-4862.

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

Deed, R.W., Jasiok, M., and Norton, J.D. (1998). Lymphoid-specific expression of the Id3 gene in hematopoietic cells. Selective antagonism of E2A basic helix-loop-helix protein associated with Id3-induced differentiation of erythroleukemia cells. J Biol Chem *273*, 8278-8286.

Edmondson, D.G., Cheng, T.C., Cserjesi, P., Chakraborty, T., and Olson, E.N. (1992). Analysis of the myogenin promoter reveals an indirect pathway for positive autoregulation mediated by the muscle-specific enhancer factor MEF-2. Mol Cell Biol *12*, 3665-3677.

Edmondson, D.G., Lyons, G.E., Martin, J.F., and Olson, E.N. (1994). Mef2 gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. Development *120*, 1251-1263.

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

Fuks, F., Burgers, W.A., Godin, N., Kasai, M., and Kouzarides, T. (2001). Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. EMBO J *20*, 2536-2544.

Genetta, T., Ruezinsky, D., and Kadesch, T. (1994). Displacement of an E-box-binding repressor by basic helix-loop-helix proteins: implications for B-cell specificity of the immunoglobulin heavy-chain enhancer. Mol Cell Biol *14*, 6153-6163.

Gossett, L.A., Kelvin, D.J., Sternberg, E.A., and Olson, E.N. (1989). A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes. Mol Cell Biol *9*, 5022-5033.

Hirai, S., Miwa, A., Ohtaka-Maruyama, C., Kasai, M., Okabe, S., Hata, Y., and Okado, H. (2012). RP58 controls neuron and astrocyte differentiation by downregulating the expression of Id1-4 genes in the developing cortex. EMBO J *31*, 1190-1202.

Hu, J.S., Olson, E.N., and Kingston, R.E. (1992). HEB, a helix-loop-helix protein related to E2A and ITF2 that can modulate the DNA-binding ability of myogenic regulatory factors. Mol Cell Biol *12*, 1031-1042.

Kophengnavong, T., Michnowicz, J.E., and Blackwell, T.K. (2000). Establishment of distinct MyoD, E2A, and twist DNA binding specificities by different basic region-DNA conformations. Mol Cell Biol *20*, 261-272.

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

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

Li, H., and Capetanaki, Y. (1993). Regulation of the mouse desmin gene: transactivated by MyoD, myogenin, MRF4 and Myf5. Nucleic Acids Res *21*, 335-343.

Lin, X., Shah, S., and Bulleit, R.F. (1996). The expression of MEF2 genes is implicated in CNS neuronal differentiation. Brain Res Mol Brain Res *42*, 307-316.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J.*, et al.* (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol *11*, 635-643.

Liu, N., Nelson, B.R., Bezprozvannaya, S., Shelton, J.M., Richardson, J.A., Bassel-Duby, R., and Olson, E.N. (2014). Requirement of MEF2A, C, and D for skeletal muscle regeneration. Proc Natl Acad Sci U S A *111*, 4109-4114.

Londhe, P., and Davie, J.K. (2011). Sequential association of myogenic regulatory factors and E proteins at muscle-specific genes. Skelet Muscle *1*, 14.

Martin, J.F., Schwarz, J.J., and Olson, E.N. (1993). Myocyte enhancer factor (MEF) 2C: a tissue-restricted member of the MEF-2 family of transcription factors. Proc Natl Acad Sci U S A *90*, 5282-5286.

Megeney, L.A., Kablar, B., Garrett, K., Anderson, J.E., and Rudnicki, M.A. (1996). MyoD is required for myogenic stem cell function in adult skeletal muscle. Genes Dev *10*, 1173-1183.

Miner, J.H., and Wold, B. (1990). Herculin, a fourth member of the MyoD family of myogenic regulatory genes. Proc Natl Acad Sci U S A *87*, 1089-1093.

Moldes, M., Boizard, M., Liepvre, X.L., Feve, B., Dugail, I., and Pairault, J. (1999). Functional antagonism between inhibitor of DNA binding (Id) and adipocyte determination and differentiation factor 1/sterol regulatory element-binding protein-1c (ADD1/SREBP-1c) trans-factors for the regulation of fatty acid synthase promoter in adipocytes. Biochem J *344 Pt 3*, 873-880.

Molkentin, J.D., Black, B.L., Martin, J.F., and Olson, E.N. (1995). Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. Cell *83*, 1125-1136.

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

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

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

Naya, F.J., Black, B.L., Wu, H., Bassel-Duby, R., Richardson, J.A., Hill, J.A., and Olson, E.N. (2002). Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. Nat Med *8*, 1303-1309.

Neuhold, L.A., and Wold, B. (1993). HLH forced dimers: tethering MyoD to E47 generates a dominant positive myogenic factor insulated from negative regulation by Id. Cell *74*, 1033-1042.

Nurrish, S.J., and Treisman, R. (1995). DNA binding specificity determinants in MADS-box transcription factors. Mol Cell Biol *15*, 4076-4085.

Ohtaka-Maruyama, C., Miwa, A., Kawano, H., Kasai, M., and Okado, H. (2007). Spatial and temporal expression of RP58, a novel zinc finger transcriptional repressor, in mouse brain. J Comp Neurol *502*, 1098-1108.

Okado, H., Ohtaka-Maruyama, C., Sugitani, Y., Fukuda, Y., Ishida, R., Hirai, S., Miwa, A., Takahashi, A., Aoki, K., Mochida, K.*, et al.* (2009). The transcriptional repressor RP58 is crucial for cell-division patterning and neuronal survival in the developing cortex. Dev Biol *331*, 140-151.

Parker, M.H., Perry, R.L., Fauteux, M.C., Berkes, C.A., and Rudnicki, M.A. (2006). MyoD synergizes with the E-protein HEB beta to induce myogenic differentiation. Mol Cell Biol *26*, 5771-5783.

Postigo, A.A., and Dean, D.C. (1997). ZEB, a vertebrate homolog of Drosophila Zfh-1, is a negative regulator of muscle differentiation. EMBO J *16*, 3935-3943.

Rhodes, S.J., and Konieczny, S.F. (1989). Identification of MRF4: a new member of the muscle regulatory factor gene family. Genes Dev *3*, 2050-2061.

Rudnicki, M.A., Braun, T., Hinuma, S., and Jaenisch, R. (1992). Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. Cell *71*, 383-390.

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

Shirakata, M., and Paterson, B.M. (1995). The E12 inhibitory domain prevents homodimer formation and facilitates selective heterodimerization with the MyoD family of gene regulatory factors. EMBO J *14*, 1766-1772.

Siles, L., Sanchez-Tillo, E., Lim, J.W., Darling, D.S., Kroll, K.L., and Postigo, A. (2013). ZEB1 imposes a temporary stage-dependent inhibition of muscle gene expression and differentiation via CtBP-mediated transcriptional repression. Mol Cell Biol *33*, 1368-1382.

Snyder, C.M., Rice, A.L., Estrella, N.L., Held, A., Kandarian, S.C., and Naya, F.J. (2013). MEF2A regulates the Gtl2-Dio3 microRNA mega-cluster to modulate WNT signaling in skeletal muscle regeneration. Development *140*, 31-42.

Taylor, S.M., and Jones, P.A. (1979). Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine. Cell *17*, 771-779.

Weintraub, H., Davis, R., Tapscott, S., Thayer, M., Krause, M., Benezra, R., Blackwell, T.K., Turner, D., Rupp, R., Hollenberg, S.*, et al.* (1991). The myoD gene family: nodal point during specification of the muscle cell lineage. Science *251*, 761-766.

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

Yokoyama, S., Ito, Y., Ueno-Kudoh, H., Shimizu, H., Uchibe, K., Albini, S., Mitsuoka, K., Miyaki, S., Kiso, M., Nagai, A*., et al.* (2009). A systems approach reveals that the myogenesis genome network is regulated by the transcriptional repressor RP58. Dev Cell *17*, 836-848.

Chapter 2: Genome-wide analysis of myogenin occupancy in differentiating skeletal myocytes.

2.1 Introduction: genome-wide occupancy mapping for a better tomorrow

At the inception of this project, relatively little was known about the number and nature of direct myogenin targets, although an attempt to broadly map them using ChIP-chip (Blais et al. 2005) yielded a list of 198 genes thought to be directly regulated by either MyoD, myogenin or Mef2. Work was also done studying DNA binding affinities, although focusing on that of MyoD, and the recognition site narrowed from the total e-box motif CANNTG to the "myogenic" e-box CAGSTG (CAGCTG and CAGGTG) (Blackwell and Weintraub 1990; Huang et al. 1996; Kophengnavong et al. 2000). While CAGCTG can be recognized and bound by multiple bHLH proteins with varying lineage specificities (more on that in Chapter 4), for the purposes of defining sites favorable to MRF occupancy CAGSTG was indeed a correct and useful refinement. This reduced the total pool of potential occupancy sites in the mouse genome from 14.2 million down to 2.7 million, but the remaining number of targets was intractably large for individual investigation. Fortunately, a technique for mapping whole-genome transcription factor occupancy by coupling chromatin immunoprecipitation and high-throughput sequencing (ChIPSeq) had just been developed by a collaboration that involved a colleague (Mortazavi et al. 2007). A high quality genome-wide occupancy map would answer the question of which CAGSTG e-boxes are occupied almost at a glance, as well as aid in a more thorough identification of target genes.

Beyond the search for target genes and assessing the number of occupied sites, several other questions could be addressed by a myogenin occupancy determination. Sequence content analysis of occupied regions can be used to refine, or verify, the primary motif, and to help in the search for likely collaborating transcription factors. Conservation at and around occupied sites can be compared to conservation at large. Occupied sites can also be classified based on their association with genes

belonging to a particular expression category, with the goal of identifying common properties within a class and distinguishing properties between classes. For example, every gene that had been studied as a direct target of myogenin was one highly specific to muscle, and whose expression increased dramatically in differentiating myocytes - would that pattern persist in the global map, or would there be a substantial number of associations with genes following different expression trajectories?

To answer these and other questions, and to gain a better understanding of myogenin's role in skeletal muscle differentiation, a genome-wide occupancy map was determined using cultured C2C12 cells, harvested 60 hours after withdrawal of serum, which triggers terminal differentiation. Immunoprecipitation was performed using a monoclonal antibody against myogenin (F5D) (Wright et al. 1996) and following the protocol from Mortazavi et al. (2007). Sequenced fragments (Illumina) were mapped to the mm9 mouse genome assembly using Eland and consolidated into regions of predicted occupancy using ERANGE (Mortazavi et al. 2007). Two stringency thresholds were used to survey the data, producing 14786 and 27765 candidate regions for, respectively, the high confidence (HC) and medium confidence (MC) settings (see Methods). For simplicity, the bulk of the discussion will focus on the HC set of regions, although comparable results from the MC set will be mentioned when necessary. Typical ChIPSeq region length was between 400 and 600 nucleotides, reaching four kilobases at the upper extreme (Table 2.1). Inspection of the small fraction of very long regions (> 800 nt) found that they have multiple motif instances. The resulting overlap of sequence reads in such regions was interpreted by the peak calling algorithm as a single, continuous region.

2.2 Associating myogenin occupancy with gene expression during differentiation

Regions occupied by myogenin were associated with genes by proximity to the nearest transcription start site (TSS), regardless of directionality (see Methods). In the absence of a direct measure of physical connectivity, such as ChIP-PET (which has only recently become possible) proximity on the chromosome was an unbiased and rational way for associating occupancy events with their candidate target gene. The distribution of distances between region peak and nearest TSS generally followed an inverse log distribution (Figure 2.2), with 14% of sites (2034) located within ± 1000 bp of an annotated TSS. Of these, three quarters (1597, 10.8%) were actually within ± 500 bp. Of the 31680 RefSeq gene models used in measuring gene expression, approximately 23% (7240) had at least one associated myogenin region. To better understand the type of genes likely to have a proximal myogenin site, the entire set was classified into groups based on differential RNASeq levels at 4 time points - undifferentiated (cycling), differentiating (60 hours), differentiating (5 days) and differentiating (7 days). Genes that were significantly (5x) up-regulated during myogenesis were classified as "myocyte" genes, while those significantly down-regulated (5x) were classified into the "myoblast" category. Genes showing a stable level of expression (no more than 2x variation either up or down) were called "flat", and genes showing no expression beyond the margin of error at any of the 4 time points were classified as "unexpressed". This method of classification resulted in a substantial number of genes that didn't fit into either of the four categories, and collectively they comprise a 5th group - "undetermined" (sometimes referred to as "wobbly").

While 68% of myocyte-specific genes had an associated myogenin region - by far the highest fraction out of the five categories, they account for only 5% of the total number of associated genes (Figure 2.3b, 2.3c). Perhaps surprisingly, at least at the time this was first observed, 57% of genes with an associated myogenin occupancy event fell into either "flat" or "undetermined" categories, showing a large amount

of promiscuity with regard to target gene behavior (Figure 2.3b). Similarly, of the 14786 myogenin regions, only 6.7% were associated with muscle-up genes, while 69% were associated with either flat or wobbly genes (Figure 2.3d). All associations by expression group were statistically different from each other ($p < 0.0002$), with order of likelihood of having a myogenin region following the coverage table in figure 2.3C (muscle up > flat > muscle down > undetermined > unexpressed). Muscle up genes were also likely to have a higher number of myogenin regions associated with them than members of the other expression groups (Figure 2.4). Absolute transcript abundance of the gene did not have a significant impact on the average number of occupancy events per gene - only heavy up-regulation (five-fold or more) mattered.

While these findings support the importance of myogenin to the regulation of muscle-specific genes, the majority of observed myogenin occupancy events (93%) did not associate with genes belonging to the muscle up category (Figure 2.3d). We now know that this is consistent with the behavior of other transcription factors from the bHLH family, including MyoD, NeuroD2, Tal1, and E47 (Cao et al. 2010; Fong et al. 2012; Kassouf et al. 2010; Lin et al. 2010); and suggests that myogenin fulfills a significant "housekeeping" role in addition to acting as a master regulator of myogenesis. In fact, the most statistically significant difference observed in this analysis was also perhaps the most stunningly obvious one - genes that have a myogenin association are more likely to be expressed than those that do not. One way to think about this is that as long as a region of chromatin is accessible and has a binding site not being actively blocked by a competitor, myogenin will occupy it some fraction of the time.

2.3 Sequence content of myogenin-occupied regions.

2.3.1 Refining myogenin primary motif based on the *in vivio* occupancy repertoire

Many introduction sections of papers published on the subject of myogenesis to this day cite the generic CANNTG e-box motif as the recognition site for MRFs specifically and for bHLH transcription factors in general. This is perhaps true if literally taken as the sum of all binding preferences of all members of the bHLH family, as diverse as they are: MyoD, myogenin, NeuroD, Tal-1, c/l-myc, Twist (Ozdemir et al. 2011) to name a few of the more 600 surveyed in a recent phylogenetic study (Stevens et al. 2008; Skinner et al. 2010). However, it is clearly a misrepresentation of the binding preferences of the MRFs themselves. For that matter, when they are considered individually, its not representative of any of the other transcription factors named above. Different bHLH proteins have different preferences for the two central nucleotides, and often exhibit biases for flanking base pairs as well. Understanding this specificity is crucial to understanding the regulatory networks governed by these transcription factors, and the possible interactions they may have with one another.

At the time this measurement was done, most evidence pointed to CAGSTG as the likely target site for MyoD:E (as discussed in section 2.1), and functional studies of elements regulated by myogenin suggested the same would be true of Myog:E (Brennan and Olson 1990; Prody and Merlie 1992; Catala et al. 1995). A genome-wide occupancy map was the perfect opportunity to assess the accuracy of this prediction, so both matrix mapping and de-novo motif discovery were used to analyze myogenin-occupied regions. To make the data more uniform and easier to interpret and process, regions were restricted to ± 250 bp from the computational peak. This was done in part to remain consistent with the average region length, and in part to allow detection of potential co-factor binding sites that might be located 100 - 200 bp away from the primary myogenin target. It was additionally informed by an analysis of conservation done by a colleague in the lab, which found that preferential conservation

around these ChIPSeq peaks covered ~ 400 bp (Pepke, unpublished). Using longer regions was not desirable because doing so would dilute enrichment of sites and place more computational strain on de novo motif finders.

The hexamer CAGSTG is indeed present in a vast majority of myogenin regions (87.0% HC, 80.6% MC) (Table 2.5), and its frequency is significantly ($p < 0.01$) enriched over genomic background (Figure 2.6a). It was also derived *de novo* by MEME from virtually any subset of myogenin occupancy data consisting of at least 100 HC regions, including the set of 100 weakest regions ranked by the strength of ChIPSeq signal they presented. One noteworthy observation from the *de novo* searches is that the matrices reported by meme consistently included the octamer RRCAGSTG, which served as the impetus for measurements that I will discuss below. Both components of the CAGSTG motif (CAGCTG and CAGGTG) were centrally concentrated in myogenin regions (Figure 2.6c), with density dropping sharply when looking more than 100 nucleotides away from the computationally defined peak. A majority of myogenin HC regions (8237, 55.7%) contained more than 1 CAGSTG, which can be interpreted biologically as support for the notion of collaborative binding (Weintraub et al. 1990; Neuhold and Wold 1993), and technically as an indication that regions with multiple CAGSTG motifs are more likely to produce a strong ChIPSeq signal. The latter will be discussed in section 2.4. The number of CAGSTG motifs present in the region did not correlate with the expression profile or mRNA levels of the associated gene.

The hexamer motif was then expanded to include the RR prefix suggested by meme searches and hinted at by Blackwell and Weintraub (1990). The same bias was independently reported for MyoD by Fong et al. (2012) and can be observed in the myogenin occupancy data from Mousavi et al. (2013), although the latter did not report on the sequence content of their occupancy determination. The RRCAGSTG octamer covered 76% of myogenin regions (compared to 87% for the hexamer), with strongest

enrichment within a 100 nucleotide radius around the computational peak (Figure 6c). It turned out that 75% of all CAGSTG e-boxes contained in the myogenin ChIPSeq regions were of the RRCAGSTG form - a startling enrichment ($p < 0.0001$) compared to the genome at large (27%). Genome-wide RRCAGGTG and RRCAGCTG occur at a similar frequency - 495K vs. 560K occurrences per genome, respectively, but the GC version dominates in myogenin regions, by a ratio of 2.7:1, which is a highly significant bias ($p < 0.0001$). Why would GC be concentrated nearly threefold with respect to GG? This will be examined in more detail in chapter 4, but its interesting to note that many of the early myogenin targets described functionally by enhancer mutagenesis were of the GG variety, so this observation was not expected. These data reinforced the idea of an octameric binding site, rather than the classically thought of hexamer. There was no correlation between the central nucleotides of the e-box (GC vs. GG) or its prefix (RR vs. non-RR) and the expression pattern of the associated gene. There does appear to be a stronger pressure on CAGGTG e-boxes to have the RR prefix than on their more numerous CAGCTG counterparts, and a potential reason for why that is will also be proposed in chapter 4.

2.3.2 Candidate motifs for collaborating and modulating functions

Based on detailed studies of canonical muscle differentiation enhancers, MRFs often act in conjunction with other transcription factors in CRMs (cis-regulatory modules) to exert the regulatory effect. Partnership with Mef2 at the ckm promoter has already been discussed in the introduction. It was also shown that MyoD and Sp1 jointly act to induce the cardiac alpha-actin promoter in skeletal muscle (Biesiada et al. 1999), and MyoD has been reported to collaborate with Pbx/Meis (Maves et al. 2007) to regulate myogenin. A variety of CRMs with differing additional factor sites are likely to exist. On the simple end of the spectrum, modules might consist of two or three sites for different TFs (all of which are expressed in a given lineage) in close proximity to one another, where joint occupancy is needed for activation of the target gene. On the other extreme, complex modules containing permissive sequences

recognized by a wider variety of transcription factors present in different developmental lineages or cell types, where subtleties in affinity coupled with additional tissue-specific elements ultimately modulate transcriptional output differently across a number of cell types. A genome-wide occupancy map provides an excellent avenue to search for such regulatory elements and for evidence of transcription factor collaboration via joint presence of recognition sites.

In addition to *de novo* searches using MEME, an extensive motif mapping effort was undertaken. It included a variety of e-box sequences, recognition sites for transcription factors previously reported to collaborate with MRFs, and a number of motifs taken from JASPAR (only motifs with reasonably high information content were taken). While a filter could have been applied to only search for binding sites of TFs actually expressed in cycling or differentiating C2C12s, such a winnowing could have predisposed the analysis towards missing "negative interactions" - motifs depleted in myogenin-occupied regions due to functional pressure for keeping certain networks separate. In addition to the octamer RRCAGSTG, MEME was able to derive versions of the AP-1 and Runx motifs from several subsets of myogenin regions - these matrices were added to the mapping database. Motif maps were then used to determine relative frequencies with which various sites were encountered (see Methods). There was an over-representation ($p < 0.01$) of binding sites for AP-1, Meis, Runx, Sp-1, c-Myc, Usf1, CTCF, and Klf4 (Figure 2.6a). The first three were reported to be associated with MyoD occupancy in myotubes, and we know that 82.3% of myogenin-occupied regions in myotubes are also occupied by MyoD (Cao et al. 2010, Chapter 3). Contrary to current expectation, there was a significant depletion ($p < 0.01$) of the Mef2 half-site AAATAG said by Cao et al. (2010) to be co-enriched with MyoD. A behavior more in line with previously known biology, where myogenin and Mef2 collaborate positively at a few well studied myocyte enhancers, was observed when using the canonical Mef2 site (CTAWWWWTAG). The latter occurs at background frequency in the myogenin ChIPSeq dataset as a whole, but I found it to be enriched in the subset of regions associated with muscle-up-regulated genes (Figure 2.6a). Despite this

confirmation of enrichment for myocyte-up-regulated genes, the total number of Mef2 sites falling within myogenin regions is relatively small - 229 total sites, 29 of them in regions associated with muscle-up genes, especially in light of its commonly accepted role as a crucial regulator of myogenesis. The observation of infrequent Mef2 sites agrees with the results of a Mef2 ChIPSeq determination on myocytes in the same stage of differentiation (Fisher-Aylor unpublished; Chapter 3). One explanation for this is that cases like the ckm enhancer, where Mef2 and MyoD/Myog have binding sites in relatively close proximity to one another, and have been proven by mutagenesis to be necessary for full enhancing activity of the corresponding CRM, are the exception rather than the rule. It is also possible that Mef2 operates at other sites without its primary recognition motif, relying instead on protein-protein interactions with MyoD or myogenin. This question will be addressed further in Chapter 3, where results of a Mef2a ChIPSeq and their correlation with the myogenin data will be analyzed.

Three promoter-associated motifs stood out in the analysis, but for different reasons. There was a very strong depletion ($p < 0.01$) of TATA-like motifs in all myogenin-occupied regions, most especially those that are promoter proximal. The TATA motif was also depleted in the dataset as a whole, being true for all stratifications by distance from the nearest TSS. The depletion of TATA motifs in regions far from any known TSS is, unlike promoters, consistent with its known function. Conversely, CGCGCG, a reported target for the for the Cfp1 subunit of the Set1 H3K4 methyltransferase complex (Clouaire et al. 2012), and the Sp1 motif CCGCCC, both promoter-associated (Figure 2.9a), were significantly ($p < 0.01$) enriched. The enrichment was very prominent in myogenin regions centered within 2500 bp from the nearest TSS, but persisted in other groups and in the dataset as a whole (Figure 2.6b). Because the G/C content of myogenin regions was not significantly different from the genome at large these differences likely suggest function. The Sp1 result is consistent with previous work, and points to a number of promoters where MRFs and Sp1 joint occupancy leads to transcriptional activation. Taken together, TATA-box and CGCGCG data suggest that the myogenic network may rely less on TATA-linked promoters

and more on those controlled by initiator regions where CGCGCG tends to be abundant. Alternatively, the presence of CGCGCG could be indicative of DNA methylation/demethylation playing an active role in controlling the activity of that subset of promoters, at least in the C2C12 cell line - a question that could benefit from further investigation focused on DNA methylation patterns in this system.

2.3.3 Repressor motifs in myogenin regions

Myogenin-occupied regions contain a higher than expected number of recognition sites for the repressive regulator Klf4, which is thought to play a role in regulating the cell cycle and can function as either an oncogene or a tumor suppressor in a context-dependent manner (Rowland and Peeper 2006). It is also implicated in triggering membrane fusion as multinucleated myotubes form (Sunadome et al. 2011). The G-rich Klf4 site is most heavily present in promoter-proximal myogenin regions (Figure 2.6b), but is enriched throughout the entire dataset. Its frequency is not influenced by the expression behavior of the associated gene (Figure 6a), and its central tendency is much lower than that of other secondary motifs. An alternative explanation is that Klf4 sites are preferentially promoter-proximal, leading to a higher rate of coincidental overlap (Figures 2.9a and b).

The consensus binding site for bHLHb2/Dec1 - an orange-class transcriptional repressor (Sun et al. 2007) up-regulated during myogenesis - is globally enriched in myogenin regions. Further analysis shows myogenin regions associated with muscle up genes do not have an enrichment for this repressor site, while regions associated with all other gene categories do (Figure 2.6a). It is important to note that this version of the Dec1 motif, as well as the Usf1 site, contain the myc-class e-box CACGTG, suggesting a need for a finer parsing of binding preferences of these transcription factors. Unfortunately, multiple attempts to generate a Dec1 occupancy map in C2C12s via ChIPSeq failed consistently, and as a result a direct derivation of the consensus site via sequence analysis has not been done.

CTCF is an 11-Zinc finger transcription factor that is highly conserved in most vertebrates (Filippova et al. 1996), and has been associated with transcriptional repression (Baniahmad et al. 1990), activation (Klenova et al. 1993) and most prominently insulation (Bell et al. 1999). It was first described as a regulator of c-myc in gallus gallus (Lobanenkov et al. 1990), but has since been ascribed a variety of functions having to do with chromatin structure and remodeling - such as facilitation of chromatin looping (Splinter et al. 2006). It is also theorized that CTCF uses different combinations of zinc fingers in different situations (Filippova et al. 1996), which makes defining a fully informative ubiquitous recognition motif more challenging. A version of the CTCF site derived from its occupancy map in differentiating myocytes (Mikkelsen unpublished) was over-represented in myogenin-occupied regions, with heaviest enrichment in promoter-proximal regions - this is consistent with the overall enrichment for CTCF sites observed in TSS-proximal regions (± 250 bp from TSS) (Figure 2.9b). However, the CTCF motifs encountered in myogenin-occupied regions had a central positional tendency that is unlikely to arise purely from accidental overlap. There was a slight bias towards regions associated with genes down-regulated during myogenesis, although that tendency was not as pronounced as the one dictated by distance to the nearest TSS.

2.3.4 Non-CAGSTG e-box motifs in myogenin regions

To further address the question of how CANNTG motifs correlate with myogenin occupancy, all possible e-box sequences (accounting for the reverse complement) were mapped and their densities calculated. There was a heavy depletion of the CAWWTG class e-boxes associated with Twist (Kophengnavong et al. 2000). The depletion was more prominent in promoter-proximal myogenin regions, but turned out to be a feature of the aggregate sample as well (Figure 2.6d). Two explanations arise for this observation. First, there is evidence of functional impetus for keeping the two networks separate, as expression of Twist and Myf5 are mutually exclusive (Hebrok et al. 1994), resulting in a depletion of Twist-like sites in

areas of genome active during myogenesis. Recent findings in drosophila (Ozdemir et al. 2011) suggest Twist frequently uses the CACATG e-box motif, and that CATATG sites may require additional inputs to be properly activated by Twist, if these biases extend to mammals. This deserves further investigation, as it could be evidence of fine-tuned Twist targeting, or it could be evidence of difference between drosophila Twist and mouse Twist. The CACATG motif is also slightly but significantly ($p < 0.01$) depleted in myogenin-occupied regions. Second, there is a general depletion for these motifs in close proximity to annotated TSSes (Figures 2.9a and 2.b), likely to limit the occurrence of spurious translation start sites (ATG). When taken together, these two phenomena help account for both the general lack of CAWWTG in myogenin-occupied regions, and for the even heavier depletion observed specifically in the TSS-proximal subset. There was also a twofold enrichment ($p < 0.01$) for CACGTG myc-class e-boxes. Otherwise, there was a lack of appreciable enrichment or depletion for any e-box that didn't belong to the SS or the WW classes (Figure 2.6d). While all density differences discussed were above the threshold for statistical significance ($p < 0.01$), the absolute rate of incidence for non-SS, non-WW e-boxes was always in the range of 0.9 - 1.2x whole genome rate, even for the CACATG Twist site, which correlates with results obtained from analyzing randomly selected regions. Beyond biological implications of enrichment and depletion above, these data should definitively put to rest the notion of CANNTG being the reference binding site for myogenin or MRFs as a group.

2.3.5 Conservation of RRCAGSTG motifs in myogenin regions

Conservation has been a traditional, if not always reliable, measure of functional importance of a given DNA element. High conservation across multiple species, and especially across phyla, often highlights sequence areas that are either involved in multiple pathways or are essential to the survival of the organism, and as a result are under selective pressure against change. While a conserved element is likely to have some functional importance, the reverse is not necessarily true. Non-conserved elements

can also be functional if they arose recently or belong to a system that is in some way unique to the organism being investigated. Finally, the definition of the word "functional" can itself be debated, so perhaps a better way to phrase this is to say that conserved elements have a higher probability of being "necessary" - their destruction or significant alteration is more likely to have deleterious consequences. That having been said, conservation of sites occupied by myogenin was natural to consider. Of the slightly over 1 million RRCAGSTG motifs present in the mouse genome, only ~23,500 are occupied by myogenin. Are those 23,500 motifs more likely to be conserved than the other ~1,000,000? And if so, what fraction are conserved? Will conservation serve as a good predictor of occupancy? Note that these motif totals are somewhat different from those used for enrichment calculations, where simple repeats were filtered out prior to computing density. A repeat could be conserved, hence the entire motif population was used for this analysis.

The results matched expectation - RRCAGSTG motifs in regions occupied by myogenin are more likely to be conserved than those in the genome at large. The exact fraction varies depending on the stringency of conservation required and the set of phastCons scores used, so after extensive testing the placental mammals set of phastCons scores was used with 0.7 entropic conservation threshold (see Methods). Under those parameters, 11.7% of all RRCAGSTG motifs are conserved in the whole genome, whereas 26.6% of RRCAGSTG motifs are conserved in regions of myogenin occupancy. This represents a significant ($p < 0.0001$) increase in the frequency with which conserved motifs are encountered, although it is interesting that it does not account for the majority of observed occupancy. If only conserved motifs are considered, 5.1% of available RRCAGSTG sequences are occupied by myogenin in differentiating myocytes (versus 2.2% if conservation is not factored in). While not a good predictor of occupancy, conservation can be useful for selecting candidate regions to undergo functional testing.

2.4 Correlation between signal intensity and motif content

Another question pertinent to the evaluation of the myogenin occupancy data was about the meaning of observed signal strength - a normalized measure of the total number of sequenced reads mapping within the genomic boundaries of the identified region. Does signal intensity correlate in any meaningful way with motif content or, perhaps more interestingly, with the behavior of the associated gene? It quickly became clear that there is some positive correlation between the observed number of reads and the number of RRCAGSTG motifs present in the region (Figure 2.7a). For comparison, the same analysis was performed using the AP-1 recognition motif CTAGTCA, and showed no correlation between the number of AP-1 motifs and signal intensity (Figure 2.7b). While longer regions are more likely to contain more copies of the RRCAGSTG octamer, the correlation remains true even when regions are normalized for length.

The relationship between signal intensity and associated gene expression is weak, if present at all. Several analyses designed to detect a positive quantitative relationship, differing in input and method, all showed remarkably little effect. A representative example is shown in Figure 2.8, which takes each region-gene association and compares ChIPSeq signal intensity and transcript levels at 60 hr after withdrawal of serum. The red category represents only the top 10% strongest myogenin regions (by ChIPSeq signal intensity), while the blue category represents all HC regions. There is no significant correlation in either case: $R = 0.01$ for the strongest 10%, $R = 0.05$ for all regions combined. The same conclusion holds true if myogenin ChIPSeq signal strengths from all sites associated with a gene are summed together first. Similarly, there is no correlation between ChIPSeq signal strength and the category of the gene expression profile, using the expression categorization method described in section 2.2. In fact, the most statistically significant correlation (by a χ^2 test, $p < 0.0001$) is the same one discussed in section 2.2 - genes with an associated myogenin region are more likely to be expressed than

those without one. This conclusion, while in many ways obvious, raises a "chicken or egg" question - does myogenin occupancy lead to the expression of neighboring genes, or are the genes being expressed by definition in areas of open chromatin, which allows myogenin occupancy, but without noticeable effect on the levels of transcription. The answer, undoubtedly, is a combination of the two (we know myogenin is crucial to the expression of many genes, both *in vivo* and *in vitro*), but the relative contribution of each phenomenon remains unknown. Functional assays can shed light on the likelihood of a given element's ability to influence transcriptional output, though even that is not necessarily a full reflection of its physiological function when it has one (or more).

References (chapter 2)

Baniahmad, A., Steiner, C., Kohne, A.C., and Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. Cell *61*, 505-514.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387-396.

Biesiada, E., Hamamori, Y., Kedes, L., and Sartorelli, V. (1999). Myogenic basic helix-loop-helix proteins and Sp1 interact as components of a multiprotein transcriptional complex required for activity of the human cardiac alpha-actin promoter. Mol Cell Biol *19*, 2577-2584.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104-1110.

Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B.D. (2005). An initial blueprint for myogenic differentiation. Genes Dev *19*, 553-569.

Brennan, T.J., and Olson, E.N. (1990). Myogenin resides in the nucleus and acquires high affinity for a conserved enhancer element on heterodimerization. Genes Dev *4*, 582-595.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L*., et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell *18*, 662-674.

Catala, F., Wanner, R., Barton, P., Cohen, A., Wright, W., and Buckingham, M. (1995). A skeletal muscle-specific enhancer regulated by factors binding to E and CArG boxes is present in the promoter of the mouse myosin light-chain 1A gene. Mol Cell Biol *15*, 4585-4596.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev *26*, 1714-1728.

Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol Cell Biol *16*, 2802-2813.

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

Hebrok, M., Wertz, K., and Fuchtbauer, E.M. (1994). M-twist is an inhibitor of muscle differentiation. Dev Biol *165*, 537-544.

Huang, J., Blackwell, T.K., Kedes, L., and Weintraub, H. (1996). Differences between MyoD DNA binding and activation site requirements revealed by functional random sequence selection. Mol Cell Biol *16*, 3893-3900.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenkov, V.V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol Cell Biol *13*, 7612-7624.

Kophengnavong, T., Michnowicz, J.E., and Blackwell, T.K. (2000). Establishment of distinct MyoD, E2A, and twist DNA binding specificities by different basic region-DNA conformations. Mol Cell Biol *20*, 261-272.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J.*, et al.* (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol *11*, 635-643.

Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743-1753.

Maves, L., Waskiewicz, A.J., Paul, B., Cao, Y., Tyler, A., Moens, C.B., and Tapscott, S.J. (2007). Pbx homeodomain proteins direct Myod activity to promote fast-muscle differentiation. Development *134*, 3371-3382.

Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Mol Cell *51*, 606-617.

Neuhold, L.A., and Wold, B. (1993). HLH forced dimers: tethering MyoD to E47 generates a dominant positive myogenic factor insulated from negative regulation by Id. Cell *74*, 1033-1042.

Ozdemir, A., Fisher-Aylor, K.I., Pepke, S., Samanta, M., Dunipace, L., McCue, K., Zeng, L., Ogawa, N., Wold, B.J., and Stathopoulos, A. (2011). High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. Genome Res *21*, 566-577.

Prody, C.A., and Merlie, J.P. (1992). The 5'-flanking region of the mouse muscle nicotinic acetylcholine receptor beta subunit gene promotes expression in cultured muscle cells and is activated by MRF4, myogenin and myoD. Nucleic Acids Res *20*, 2367-2372.

Rowland, B.D., and Peeper, D.S. (2006). KLF4, p21 and context-dependent opposing forces in cancer. Nat Rev Cancer *6*, 11-23.

Skinner, M.K., Rawls, A., Wilson-Rawls, J., and Roalson, E.H. (2010). Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. Differentiation *80*, 1-8.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev *20*, 2349-2354.

Stevens, J.D., Roalson, E.H., and Skinner, M.K. (2008). Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation. Differentiation *76*, 1006-1022.

Sun, H., Ghaffari, S., and Taneja, R. (2007). bHLH-Orange Transcription Factors in Development and Cancer. Transl Oncogenomics *2*, 107-120.

Sunadome, K., Yamamoto, T., Ebisuya, M., Kondoh, K., Sehara-Fujisawa, A., and Nishida, E. (2011). ERK5 regulates muscle cell fusion through Klf transcription factors. Dev Cell *20*, 192-205.

Weintraub, H., Davis, R., Lockshon, D., and Lassar, A. (1990). MyoD binds cooperatively to two sites in a target enhancer sequence: occupancy of two sites is required for activation. Proc Natl Acad Sci U S A *87*, 5623-5627.

Wright, W.E., Dac-Korytko, I., and Farmer, K. (1996). Monoclonal antimyogenin antibodies define epitopes outside the bHLH domain where binding interferes with protein-protein and protein-DNA interactions. Dev Genet *19*, 131-138.

Figures and Tables (chapter 2)

	Myogenin HC	Myogenin MC
Number of regions	14,798	27,793
Nucleotide coverage	7,923,964	9,742,958
Read % in regions	22.37%	25.98%
Average region length	534	351
Median region length	481	311
Length standard deviation	279	151
Minimum length	70	52
Maximum length	3,793	2,071

Table 2.1. Primary characteristics of the myogenin occupancy measurement and comparison between HC and MC sets of regions. Nucleotide coverage refers to the total number of nucleotides considered "occupied", as a union of all called regions. Read % generally correlates with quality of determination - higher is better, although the exact fraction depends in large part on the stringency of peak calling.

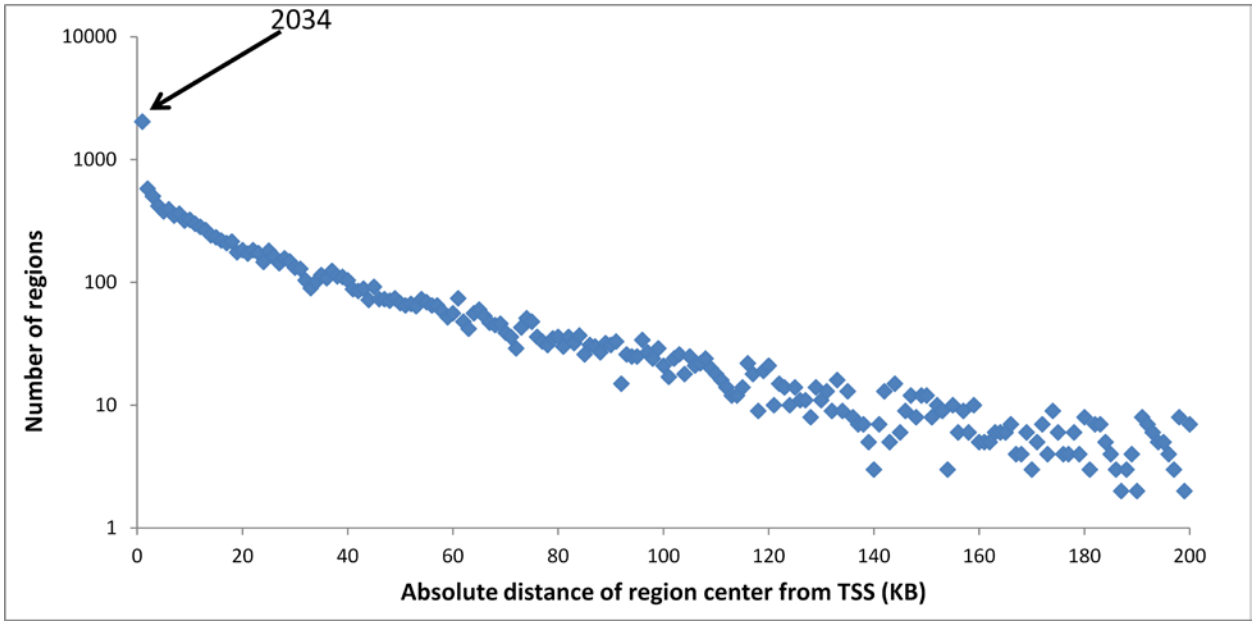


Figure 2.2. Distribution of myogenin-occupied regions relative to nearest annotated TSS. Distance was measured by taking the absolute difference between the coordinate of the region peak (as defined by ERANGE) and the coordinate of the TSS.

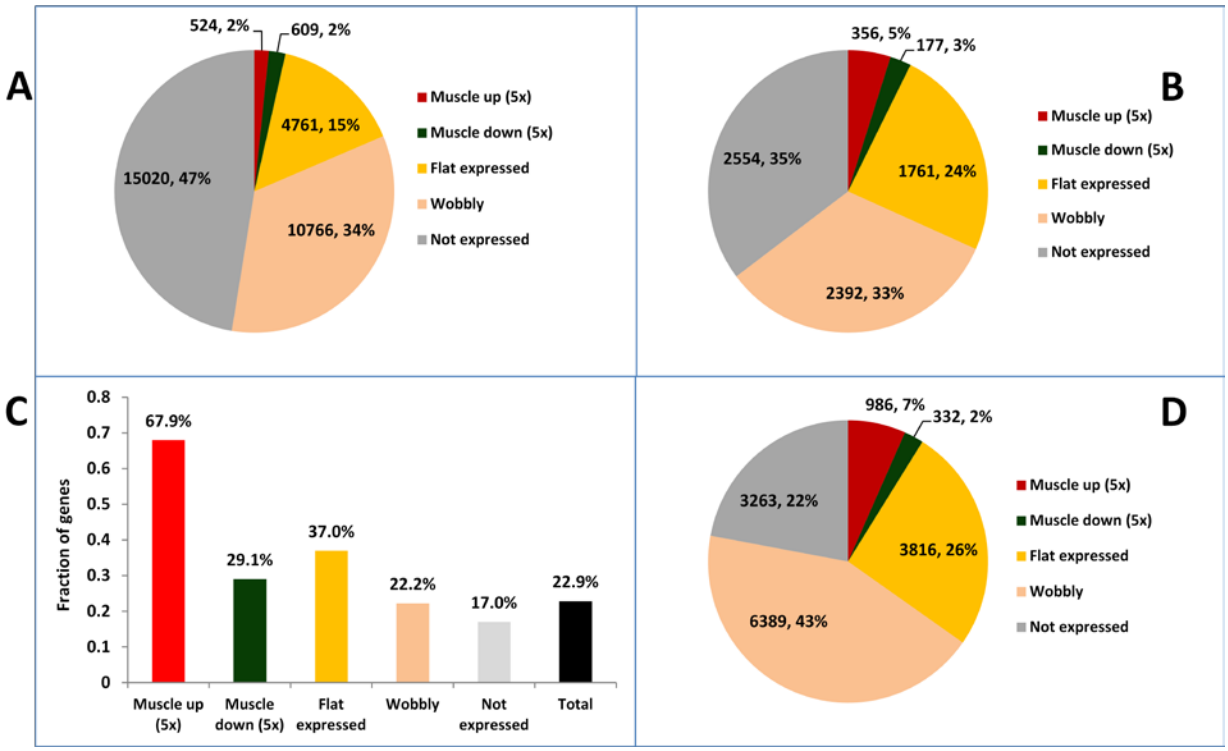


Figure 2.3. A) Classification of gene models by expression profile in C2C12s. All 31,680 RefSeq gene models used for RNASeq data processing were considered. The five classes are defined as follows: **Muscle up (5x)**: genes for which (expression myocytes) / (expression myoblasts) ≥ 5 and (expression myocytes) ≥ 1 rpkm; **Muscle down (5x)**: genes for which (expression myocytes) / (expression myoblasts) ≤ 0.2 and (expression myoblasts) ≥ 1 rpkm; **Flat expressed**: genes for which 2 ≤ (expression myocytes) / (expression myoblasts) ≤ 0.5 and (expression myocytes) ≥ 1 rpkm and (expression myoblasts) ≥ 1 rpkm; **Not expressed**: genes for which (expression myoblasts) ≤ 0.5 rpkm and (expression myocytes) ≤ 0.5 rpkm; **Wobbly (undetermined)**: genes that do not belong to any of the first 4 categories.

B) Expression profiles of genes with an associated myogenin-occupied region (N = 7240).

C) Incidence of myogenin occupancy as a function of expression profile Over two-thirds of muscle up genes have at least one associated myogenin region, compared to only one in six unexpressed genes.

D) Region-centric association of myogenin occupancy and gene expression profiles.

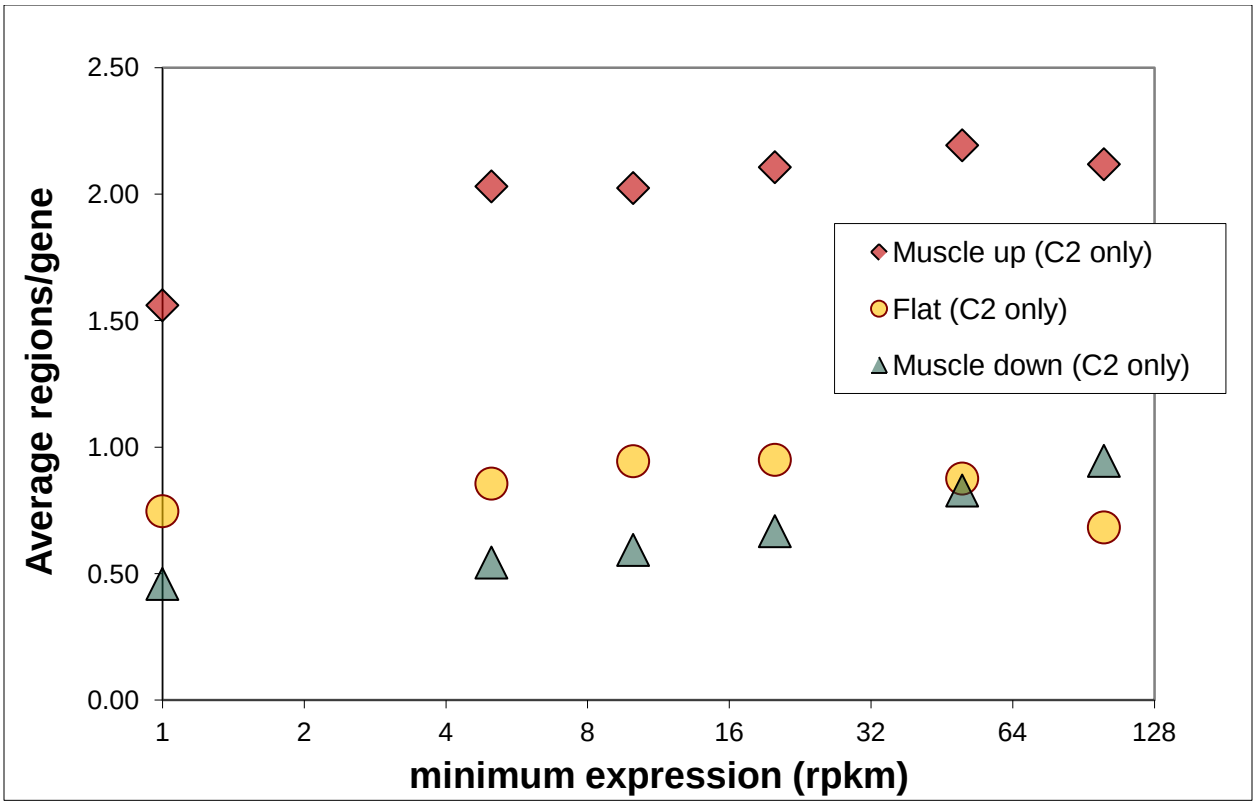


Figure 2.4. Genes whose RNA levels are up-regulated five-fold or more during differentiation have more regions of myogenin occupancy associated with them (by near-neighbor proximity) than do either down-regulated or constantly expressed genes ($p < 0.0001$). This positive correlation hinges on the criterion of substantial upregulation upon differentiation, not on overall transcript abundance, which is not significant.

Motif	Myogenin	Randomized
CAGSTG	87.0%	42.6%
CAGCTG	75.4%	21.8%
CAGGTG	45.0%	28.0%
RRCAGSTG	76.0%	20.0%
RRCAGCTG	63.9%	11.9%
RRCAGGTG	27.6%	9.5%
CACGTG	9.2%	5.2%
AP-1	15.1%	9.0%
Mef2 (half-site)	16.9%	33.0%
Mef2 (lit)	1.5%	1.3%
Klf4	29.1%	10.4%
CTCF	5.7%	1.5%
Usf1	1.7%	0.7%
CGCGCG	5.9%	0.7%
TATA-box	10.2%	32.8%
Sp-1	23.9%	6.4%

Table 2.5. Coverage of myogenin ChIPSeq regions (C2C12, 60 hr after differentiation) by select motifs. Coverage is defined as the fraction of regions containing one or more copies of the motif. 100% coverage means every region has at least one copy of the motif within its boundaries, 0% means the union of all regions in the set is completely devoid of the motif. For comparison, a set of approximately 101,000 regions of the same length (501 bp) was selected at random from the genome, with provisions to avoid poorly sequenced areas, telomere/centromere repeats, and areas of highly repetitive sequence (see Methods). Coverage of these randomized regions was computed and is provided for comparison.

Motif	All	Muscle-up	Muscle-dn	Flat expr.	Unexpr.	Random
CAGCTG	2.55	2.49	2.44	2.47	2.57	0.29
CACCTG	1.08	1.20	1.15	1.03	1.21	0.23
CAGSTG	1.86	1.87	1.82	1.80	1.93	0.26
RRCAGSTG	2.78	2.80	2.66	2.66	2.79	0.27
RRCAGCTG	3.16	3.13	3.03	3.04	3.14	0.29
RRCAGGTG	1.97	2.13	1.88	1.87	2.08	0.24
CACGTG	1.02	0.72	1.11	1.17	1.03	0.15
Meis	0.77	0.82	0.68	0.64	0.73	0.20
AP-1	0.97	0.70	0.46	0.55	0.88	0.11
Mef2 (half)	-1.20	-1.16	-1.76	-1.56	-1.32	-0.07
Mef2 (hybrid)	0.19	0.67	0.15	0.67	0.55	0.13
Runx	1.06	0.97	0.97	0.87	1.02	0.21
RP58	-0.22	-0.54	-0.25	-0.44	-0.31	0.13
CTCF	2.42	2.76	3.54	3.05	2.89	0.46
Mef2 (lit)	0.22	0.99	-0.30	-0.12	0.08	-0.01
Klf4	1.89	2.18	2.29	2.26	2.11	0.20
Usf1	1.56	1.05	1.52	2.05	2.04	0.12
CGCGCG	3.37	3.08	4.45	4.35	3.79	-0.04
TATA-box	-2.36	-2.57	-2.75	-2.75	-2.30	-0.26
Sp-1	2.80	2.88	3.72	3.54	3.08	0.42
Dec1	1.03	0.43	1.68	1.40	1.11	0.13

Figure 2.6a. Comparative sequence content analysis of all myogenin-occupied regions, and groups of regions associated with genes of a particular expression category (and within ± 20KB of nearest TSS). **All** = all myogenin regions (n = 14786); **Muscle-up** = regions associated with muscle-up genes (n = 1097); **Muscle-dn** = regions associated with muscle-down regions (n = 370); **Flat expr.** = regions associated with flat expressed genes (n = 3746); **Unexpr.** = regions associated with unexpressed genes (n = 2702); **Random** = randomized regions (n ≈ 101000). Enrichment is given as $\log_2$ of (observed density) / (expected density), values that are not significantly different from background ($p > 0.01$) are displayed in grey, negative values represent depletion of the associated motif.

Mef2 (lit) = canonical Mef2 motif CTAWWWWTAG.

Mef2 (half) = half site motif AAATAG used by Cao et al. (2010).

Mef2 (hybrid) = Mef2-associated motif GGRANHYGTAGT derived from a subset of Mef2-occupied regions (see chapter 3).

Motif	all	0 - 500	501 - 2.5K	2.5K - 20K	20K+	random
CAGCTG	2.55	2.11	2.59	2.56	2.59	0.29
CACCTG	1.08	0.95	1.04	1.13	1.06	0.23
CAGSTG	1.86	1.54	1.88	1.89	1.88	0.26
RRCAGSTG	2.78	2.36	2.72	2.78	2.81	0.27
RRCAGCTG	3.16	2.67	3.13	3.14	3.21	0.29
RRCAGGTG	1.97	1.76	1.82	2.03	1.98	0.24
CACGTG	1.02	1.62	1.23	0.86	0.90	0.15
Meis	0.77	0.29	0.68	0.80	0.84	0.20
AP-1	0.97	-0.40	0.77	0.93	1.20	0.11
Mef2 (half)	-1.20	-1.97	-1.58	-1.23	-0.99	-0.07
Mef2 (hybrid)	0.19	1.48	0.12	0.06	-0.25	0.13
Runx	1.06	0.38	0.79	1.12	1.18	0.21
RP58	-0.22	-1.70	-0.14	-0.18	-0.05	0.13
CTCF	2.42	4.14	3.00	2.09	1.41	0.46
Mef2 (lit)	0.22	-0.32	0.16	0.30	0.27	-0.01
Klf4	1.89	2.91	2.30	1.81	1.41	0.20
Usf1	1.56	3.08	1.02	1.42	1.01	0.12
CGCGCG	3.37	5.79	4.20	1.91	1.39	-0.04
TATA-box	-2.36	-3.55	-2.53	-2.36	-2.15	-0.26
Sp-1	2.80	4.83	3.35	2.03	1.67	0.42
Dec1	1.03	2.16	0.77	0.73	0.85	0.13

Figure 2.6b. Comparative sequence content analysis of all myogenin-occupied regions, and groups of regions classified by distance from nearest TSS. **All** = all myogenin regions (n = 14786); **0 - 500** = regions with a TSS 0 - 500 bp away from peak (n = 1597); **501 - 2.5K** = regions with a TSS 501 - 2500 bp away from peak (n = 1273); **2.5K - 20K** = regions with a TSS 2501 - 20000 bp away from peak (n = 5118); **20K+** = regions with a TSS 20001 bp or more away from peak (n = 6798); **Random** = randomized regions (n ≈ 101000). Enrichment is given as $\log_2$ of (observed density) / (expected density), values that are not significantly different from background ($p > 0.01$) are displayed in grey, negative values represent depletion of the associated motif.

	-250				0					250
RRCAGSTG	0.57	0.88	1.27	2.24	4.50	4.56	2.19	1.26	0.92	0.61
RRCAGCTG	0.55	0.94	1.43	2.55	4.94	5.02	2.51	1.44	1.08	0.66
RRCAGGTG	0.59	0.80	1.03	1.63	3.49	3.51	1.54	0.95	0.66	0.53
CAGCTG	0.50	0.89	1.18	2.10	4.22	4.28	2.05	1.17	0.95	0.61
CAGGTG	0.28	0.45	0.54	0.87	2.20	2.24	0.75	0.51	0.46	0.40

Figure 2.6c. Both CAGSTG and RRCAGSTG motifs show a central enrichment tendency in the ± 50 bp radius from the called peak of a myogenin-occupied region. Enrichment is given as $\log_2$ of (observed density) / (expected density), all values are significantly different from background ($p < 0.01$).

Motif	All	Muscle-up	Muscle-dwn	Flat expr.	Unexpr.	Random
CAGCTG	2.55	2.49	2.44	2.47	2.57	0.29
CAGGTG	1.08	1.20	1.15	1.03	1.21	0.23
CACGTG	1.02	0.72	1.11	1.17	1.03	0.15
CATATG	-1.34	-1.63	-1.93	-1.81	-1.58	-0.12
CAAATG	-0.67	-0.93	-1.19	-1.07	-0.88	0.05
CAATTG	-1.16	-0.98	-1.44	-1.40	-1.45	0.01
CAGATG	0.26	0.17	0.09	0.02	0.19	0.18
CAAGTG	0.11	0.13	-0.16	-0.08	0.04	0.11
CACATG	-0.15	-0.29	-0.53	-0.43	-0.35	-0.02
CAACTG	0.29	0.28	0.29	0.07	0.19	0.12

Motif	All	0 - 500	501 - 2.5K	2.5K - 20K	20K+	Random
CAGCTG	2.55	2.11	2.59	2.56	2.59	0.29
CAGGTG	1.08	0.95	1.04	1.13	1.06	0.23
CACGTG	1.02	1.62	1.23	0.86	0.90	0.15
CATATG	-1.34	-2.94	-1.93	-1.34	-1.04	-0.12
CAAATG	-0.67	-1.92	-0.91	-0.74	-0.41	0.05
CAATTG	-1.16	-1.61	-1.45	-1.32	-0.93	0.01
CAGATG	0.26	-0.90	0.13	0.30	0.43	0.18
CAAGTG	0.11	-0.67	-0.04	0.12	0.26	0.11
CACATG	-0.15	-1.85	-0.45	-0.13	0.10	-0.02
CAACTG	0.29	-0.37	0.26	0.31	0.41	0.12

Figure 2.6d. E-box motifs in myogenin-occupied regions. Top panel same as 2.6a, bottom panel same as 2.6b, but only for the 10 possible e-box motifs. Enrichment is given as $\log_2$ of (observed density) / (expected density), values that are not significantly different from background ($p > 0.01$) are displayed in grey, negative values represent depletion of the associated motif.

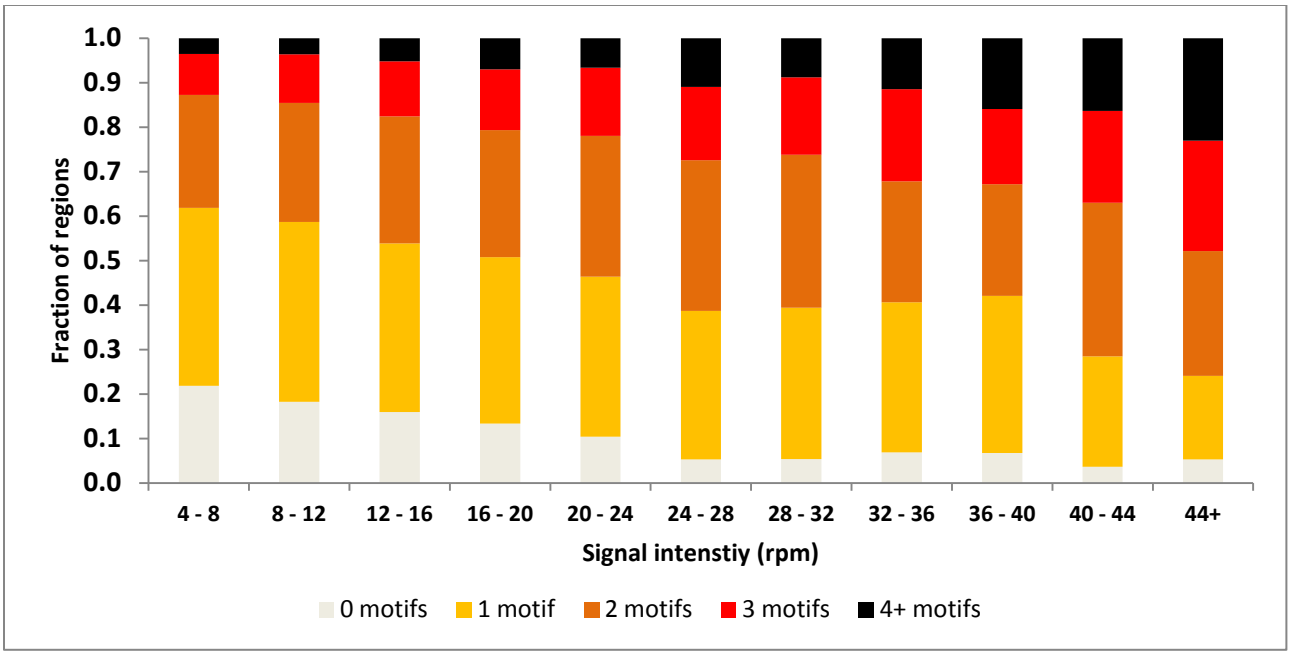


Figure 2.7a. Fraction of regions with multiple RRCAGSTG motifs in myogenin-occupied regions grouped by ChIPSeq signal strength.

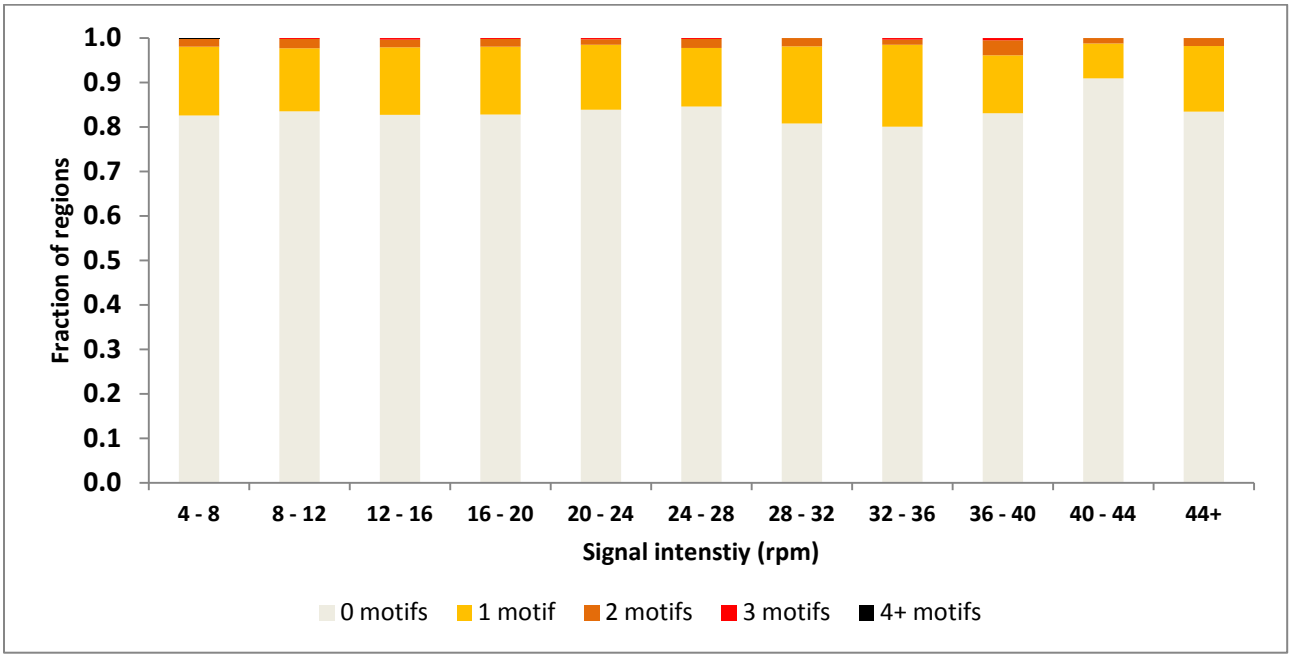


Figure 2.7b. Fraction of regions with multiple AP-1 motifs (TGAGTCA) in myogenin-occupied regions grouped by ChIPSeq signal strength.

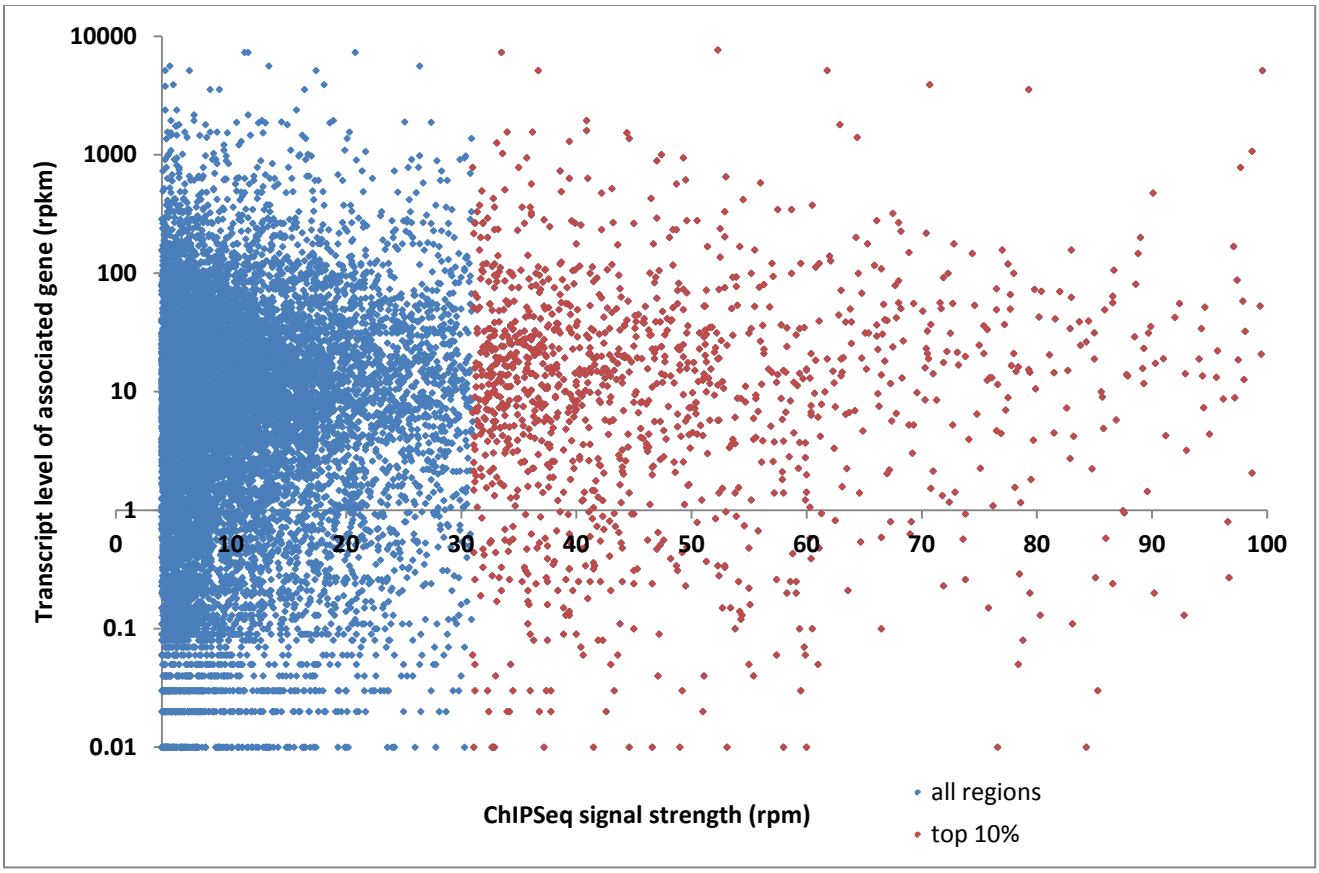


Figure 2.8. ChIPSeq signal strength for each myogenin-occupied region was correlated with transcript abundance of the associated gene. There is no correlation for either the set of all HC regions (blue, R=0.052) or the set of regions with ChIPSeq signal strength above 90th percentile (red, R=0.014).

Motif	-2500					0				2500
CAGCTG	0.14	0.18	0.17	0.16	0.32	0.75	0.40	0.30	0.27	0.25
CACCTG	0.15	0.19	0.17	0.15	0.20	0.48	0.36	0.24	0.26	0.25
CAGSTG	0.15	0.18	0.17	0.15	0.25	0.59	0.37	0.26	0.27	0.25
RRCAGSTG	0.18	0.26	0.25	0.22	0.28	0.74	0.49	0.36	0.32	0.32
RRCAGCTG	0.12	0.22	0.19	0.17	0.33	0.89	0.50	0.36	0.31	0.30
RRCAGGTG	0.24	0.30	0.31	0.28	0.22	0.54	0.47	0.36	0.33	0.35
CACGTG	0.11	0.15	0.26	0.40	0.95	0.50	0.45	0.28	0.22	0.17
CATATG	-0.20	-0.15	-0.20	-0.32	-0.73	-1.22	-0.59	-0.36	-0.25	-0.28
CAAATG	-0.17	-0.17	-0.17	-0.23	-0.54	-0.85	-0.40	-0.24	-0.18	-0.18
CAATTG	-0.19	-0.12	-0.22	-0.20	-0.31	-0.93	-0.41	-0.20	-0.22	-0.34
CAGATG	0.01	0.01	-0.03	-0.15	-0.31	-0.12	-0.08	-0.06	-0.05	-0.01
CAAGTG	0.16	0.18	0.12	0.10	-0.07	-0.20	0.00	0.13	0.14	0.19
CACATG	0.02	0.00	-0.02	-0.20	-0.56	-0.99	-0.46	-0.21	-0.09	-0.08
CAACTG	0.06	0.03	0.03	-0.01	-0.12	-0.14	0.02	0.03	0.05	0.04
Meis	0.03	0.05	0.02	0.01	0.08	0.33	0.20	0.12	0.11	0.14
AP-1	0.10	0.10	0.10	0.06	0.01	-0.46	-0.21	0.00	0.09	0.12
Mef2 (half)	-0.19	-0.14	-0.12	-0.14	-0.42	-1.05	-0.43	-0.23	-0.19	-0.17
Mef2 (hybrid)	-0.05	0.07	0.28	0.29	1.67	0.40	0.11	0.23	0.22	0.01
Runx	0.05	0.07	0.05	-0.01	0.08	0.52	0.38	0.25	0.18	0.20
RP58	-0.09	-0.07	-0.12	-0.21	-0.48	-0.38	-0.24	-0.18	-0.12	-0.09
CTCF	0.69	0.66	0.83	1.27	2.92	2.76	1.62	0.99	0.66	0.68
CACACA	0.08	0.07	0.08	0.01	-0.53	-1.07	-0.39	-0.05	0.01	-0.02
Mef2 (lit)	-0.19	-0.16	-0.08	-0.16	-0.28	-1.59	-0.53	-0.12	-0.21	-0.25
Klf4	0.43	0.48	0.60	0.93	2.28	1.42	1.03	0.71	0.63	0.54
Usf1	0.10	0.21	0.21	0.44	2.55	0.97	0.54	0.36	0.18	0.24
CGCGCG	0.31	0.52	1.09	2.00	4.17	4.31	2.45	1.38	0.72	0.49
TATA-box	-0.23	-0.28	-0.24	-0.20	-0.80	-1.98	-0.78	-0.33	-0.37	-0.25

Figure 2.9a. Mapping motifs in a 2500 bp radius around TSSes of all 31680 gene models used for the RNASeq determination in C2C12s. Mapping was adjusted for directionality of the associated ORF - notice the sharp enrichment gradients around the TSS (position 0) for some of the motifs. Each cell represents a 500 bp "step". Enrichment values computed that same way as above.

Motif	-250					0				250
CAGCTG	0.07	0.35	0.46	0.43	0.81	1.12	0.93	0.76	0.71	0.72
CACCTG	0.08	0.22	0.16	0.23	0.11	0.33	0.52	0.52	0.58	0.41
CAGSTG	0.08	0.27	0.27	0.31	0.40	0.67	0.68	0.61	0.63	0.53
RRCAGSTG	0.08	0.39	0.30	0.36	0.61	0.81	0.87	0.80	0.81	0.71
RRCAGCTG	0.03	0.42	0.34	0.45	0.96	1.17	1.08	0.96	0.94	0.88
RRCAGGTG	0.15	0.34	0.26	0.25	0.08	0.25	0.57	0.58	0.65	0.49
CACGTG	0.60	1.00	1.15	1.44	1.60	0.78	0.42	0.31	0.38	0.50
CATATG	-0.74	-0.77	-1.12	-1.16	-1.46	-1.72	-1.96	-1.74	-1.40	-1.33
CAAATG	-0.52	-0.63	-0.57	-0.86	-1.06	-1.18	-1.19	-1.23	-0.92	-0.91
CAATTG	-0.36	-0.42	-0.46	-0.25	-0.96	-1.17	-1.10	-1.17	-0.96	-0.88
CAGATG	-0.40	-0.41	-0.32	-0.50	-0.49	-0.38	-0.21	-0.29	-0.24	-0.16
CAAGTG	-0.15	-0.11	-0.18	-0.30	-0.39	-0.46	-0.34	-0.39	-0.29	-0.20
CACATG	-0.60	-0.62	-0.66	-0.70	-1.09	-1.14	-1.26	-1.12	-1.17	-1.02
CAACTG	-0.20	-0.15	-0.22	-0.06	-0.15	0.15	-0.18	-0.27	-0.35	-0.31
Meis	0.06	0.03	0.10	0.14	0.18	0.48	0.34	0.28	0.18	0.31
AP-1	-0.20	-0.18	0.34	-0.06	-0.17	-0.52	-0.63	-0.77	-0.47	-0.76
Mef2 (half)	-0.51	-0.54	-0.59	-0.65	-0.69	-1.87	-1.32	-1.34	-1.05	-1.31
Mef2 (hybrid)	1.19	1.93	2.20	2.77	1.86	0.46	0.13	0.22	0.35	0.70
Runx	0.06	0.20	0.07	0.18	0.06	0.82	0.67	0.61	0.49	0.47
RP58	-0.67	-0.71	-0.63	-0.69	-0.75	-0.79	-0.73	-0.90	-0.66	-0.23
CTCF	2.56	3.06	3.55	3.97	3.43	2.88	2.74	2.77	2.90	2.51
CACACA	-0.68	-0.60	-0.93	-0.96	-1.38	-1.65	-1.83	-1.39	-1.18	-1.26
Mef2 (reference)	-0.86	-0.25	-0.41	0.06	-0.16	-2.79	-2.11	-1.11	-1.94	-2.79
Klf4	1.84	2.25	2.73	3.32	3.05	1.07	1.21	1.36	1.52	1.63
Usf1	1.78	2.30	2.94	3.64	4.00	2.00	1.17	0.75	0.17	1.21
CGCGCG	3.56	4.09	4.51	4.97	5.11	4.78	4.36	4.29	4.23	4.32
TATA-box	-0.89	-0.98	-1.11	-1.23	-0.97	-3.03	-2.85	-2.45	-2.34	-1.98

Figure 2.9b. Same as 2.9a, except using a 250 bp radius; each cell represents a 50 bp "step".

Chapter 3: Comparative occupancy analysis of transcription factors active during myogenesis.

3.1 Introduction: joint occupancy versus exclusive occupancy

Having examined the myogenin occupancy map in differentiating C2C12s and its properties, a natural next step was to compare it to other transcription factors important to muscle differentiation. A number of questions, both empirical and conceptual, could be addressed by a cross-comparison of occupancy data, several of which will be discussed in this chapter. First, an analysis of transcriptional regulation of myogenesis would not be complete without consideration for MyoD, and section 3.2 was dedicated to the comparison between myogenin and MyoD occupancy, as well as differential occupancy by MyoD in cycling myoblasts. The latter provided motif-level evidence for collaboration between MyoD and AP-1 (also noted by Cao et al. 2010), which was further studied in section 3.3 by determining and analyzing a Fosl1 occupancy map in cycling C2C12 myoblasts.

Several hypotheses existed regarding the spatial relationship between Mef2 and MRF occupancy (based on functional studies of active promoter/enhancer elements), and about Mef2 targeting mechanisms (Molkentin el al. 1995). With Mef2 being the primary non-bHLH transcription factor associated with regulation of myogenesis, testing them was a high priority. To that effect, Mef2 occupancy in differentiating myocytes was determined, analyzed for sequence content, and compared to myogenin - these results are presented in section 3.4.

Finally, I will consider two transcription factors that are not considered to have a direct role in skeletal muscle specification, but that are both present and active in C2C12 cells. One is CTCF, whose recognition motif was enriched in several MRF measurements. CTCF itself is an important regulatory factor that fulfills a variety of roles, hence its occupancy in relation to myogenin and MyoD will be

evaluated (section 3.5). The other is Usf1 - a c-myc-like bHLH transcription factor associated with a set of Mef2-occupied regions. While Usf1 is believed to be a "housekeeping" transcription factor, its association with Mef2 prompted a closer look (section 3.6).

The analysis focused heavily on sequence content of occupied regions. It also focused on identifying joint and differential occupancy by the various TFs, and on characterizing its properties. Region-gene associations and distance distribution from nearest annotated TSS were also measured to gain a sense of the overall network architecture.

3.2 Myogenin and MyoD occupancy in differentiating myocytes is highly concordant

Both MyoD and myogenin mRNAs are highly abundant in differentiating C2C12 myocytes (Table 3.7), with myogenin transcript approaching structural protein levels 60 hours after withdrawal of serum. The two factors are closely related by sequence homology, yet function in different stages of the differentiation process, and there is no readily available compensator for the absence of myogenin, which leads to a failure of muscle fiber maturation (see Introduction). An important question for understanding the transcriptional network of skeletal muscle differentiation was to quantify the extent to which MyoD and myogenin occupy the same DNA elements, as well as to identify and examine any apparent differences.

MyoD occupancy was measured in C2C12s using a monoclonal antibody and the same protocol as that for myogenin (Kwan, unpublished). Unlike myogenin, MyoD is present both in the myoblast and the myocyte stages, hence occupancy was measured in cycling myoblasts and in differentiating myocytes at 24 and 60 hours after withdrawal of serum, with primary region calling results summarized in Table 3.1a. The occupancy maps were compared to identify preferential state-specific occupancy, if any. The 60 hr measurement was particularly successful, and almost fully encompassed the 24 hr determination. Of the 9922 HC MyoD 24h regions, 8828 (89.0%) overlapped a MyoD 60h HC region, with the fraction increasing to 93.4% (9269) when compared to MyoD 60h MC regions instead (Table 3.3a). In light of this, I will at times refer to the 60 hr set of MyoD regions as representative of occupancy in differentiating cells. The 24 hr dataset proved useful for computing differential occupancy, due to the smaller number of sites identified in cycling cells.

To answer the question of shared occupancy, maps for MyoD and myogenin at the 60 hr timepoint were compared. There were 14798 HC myogenin regions and 16460 HC MyoD regions at 60 hr after differentiation, with 12904 (87.2%) myogenin HC regions overlapping MyoD MC (Table 3.3a). Only 1257

myogenin regions (8.5%) were completely disjoint from a MyoD-occupied site (Table 3.3b). The same is true in reverse - 82.3% of MyoD 60 hr regions are shared with myogenin, and only 2088 out of 16460 (12.7%) are completely disjoint. Using the MyoD 24 hr dataset yielded a similarly high degree of overlap (Table 3.3). Essentially, MyoD and myogenin occupy the same loci in differentiating myocytes - at least four out of every five elements displaying a MyoD signature also display one for myogenin, and vice versa. This degree of overlap is remarkable because it is comparable to biological replicates, even though the antibodies used for MyoD and myogenin have been tested extensively, and they do not cross-react.

The situation changed somewhat when MyoD occupancy in cycling myoblasts was examined. Of the 6651 HC MyoD cycling regions, 3661 (55.0%) overlapped myogenin HC, and 4709 (70.1%) overlapped myogenin MC regions (Table 3.3). While the similarity is still substantial, these results point to changes in the occupancy profile of MyoD upon cell cycle exit and initiation of terminal differentiation. To further define differential occupancy by MyoD, three sub-groups of regions were selected: those only occupied in cycling progenitor cells ("early" sites), those occupied only in differentiating myocytes ("late" sites), and those occupied in both states ("continuous" sites). The set of "late" sites consisted of 2716 regions, of which 2348 (86.5%) overlapped myogenin MC regions. The remaining 13.5% failed to manifest any readily observable properties that would make them stand out for further study. The "late" MyoD regions also had the highest coverage by the RRCAGSTG motif, at 84.8% (Table 3.8c). There were, 3515 "continuous" regions of MyoD occupancy, defined as sites that had a HC signal in both the cycling and the 60 hr differentiated cells. Of them, 2689 (76.5%) overlapped myogenin MC regions. The 653 that did not were examined separately as "continuous exclusive" sites, and will be referred to when sequence content is discussed. Finally, there were 951 "early" MyoD sites - those occupied in cycling myoblasts but not in differentiating cells. Only 5 "early" sites overlapped myogenin HC regions, with 29 further included when using myogenin MC instead; 903 (95.0%) "early" sites were completely disjoint

from myogenin. This led to the conclusion that the primary difference between MyoD and myogenin lies in the sites that MyoD occupies exclusively in cycling cells - once terminal differentiation begins, the two factors occupy virtually identical sets of elements, although subtle differences do exist.

Sequence content analysis was performed on the four base ChIPSeq determinations (MyoD cycling, MyoD 24h, MyoD 60h, myogenin), the four sub-groups of differentially occupied MyoD regions discussed above, and two additional sub-groups - sites unique to either MyoD or myogenin at the 60 hr timepoint. It consisted of *de novo* motif discovery using MEME and extensive mapping of known matrices, with the *de novo* analysis deriving an extended version of the AP-1 recognition site (TGASTCACW) when presented with "early" MyoD regions. The results of sequence content analysis are summarized in Figure 3.4a. The motif RRCAGSTG showed a strong central enrichment tendency in all 4 primary datasets (Figure 3.4b) and also covered a majority of the regions (Table 3.8a). The "early" MyoD regions clearly stood out from the other differentially occupied sub-groups, showing a significant difference in motif content ($p < 0.01$) for a variety of sites, including the primary recognition motif RRCAGSTG, AP-1, RP58 and Klf4 (Figure 3.5). Over half of the "early" regions contain at least one AP-1 motif (Table 3.8c). The frequency of RRCAGSTG was greatly diminished, with only 41% coverage, compared to 84.8% for the "late" set (Table 3.8c). Furthermore, 60% of the regions with an RRCAGSTG motif also contained the AP-1 site. This is in sharp contrast with the "late" MyoD regions, where the AP-1 motif is significantly ($p < 0.01$) depleted (Figure 3.5). The other MyoD region groupings were well covered by RRCAGSTG (Table 3.8c). Notably, MyoD-exclusive regions had an almost even split between RRCAGCTG and RRCAGGTG motifs, while myogenin-exclusive regions heavily favored the RRCAGCTG version, suggesting subtle targeting differences between the two MRFs.

To identify potential differences in regulatory targets of MyoD and myogenin, region-gene association profiles were generated for the 3 MyoD datasets, and compared to each other and to myogenin. The

normalized results (due to differing numbers of regions in each dataset) are summarized in Figure 3.2. The overall profiles are very similar, and some tendencies become apparent. The vast majority of regions (60-70%, depending on the set) are associated with genes that are expressed in both myoblasts and myocytes, and whose mRNA levels do not change drastically (flat and undetermined categories). Together, regions associated with myoblast and myocyte genes comprise less than 15% of the total number of observed occupancy events. After the onset of differentiation there is an increase in the frequency of "muscle up" (myocyte) associations, and myocyte associations outnumber their myoblast counterparts by approximately 3:1. MyoD regions in cycling myoblasts are almost equally likely to be associated with either a "muscle-up" or a "muscle-down" gene, although the probability of either association is less than 15%. These results are not unexpected given the high degree of similarity between MyoD and myogenin occupancy.

Overall, I found that MyoD and myogenin occupy a largely shared set of sites in differentiating myocytes. Both factors recognize the motif RRCAGSTG, but subtle differences exist in the distribution of the GC vs. GG e-box "cores". Myogenin-occupied regions heavily favor GC over GG, while in MyoD-occupied regions the distribution is more even (but still slanted in favor of GC). Regions preferentially occupied by MyoD in cycling myoblasts are distinct from the rest, both in terms of sequence content and their genomic distribution relative to nearest TSS. Their high coverage by the AP-1 motif TGAGTCA suggests a collaboration between MyoD and AP-1, which is investigated in section 3.3.

3.3 MyoD collaboration with AP-1 is limited to cycling myoblasts

AP-1 (jun/fos heterodimer) was first linked to myogenesis when it was shown that Mef2a and Mef2d can regulate the promoter of c-Jun (Han et al. 1992; Han and Prywes 1995). It was later proposed that AP-1 is involved in a feed-forward loop that leads to the activation of the MRF transcription factors (Andreucci et al. 2002), and its recognition sequence was independently reported enriched in MyoD-occupied regions during muscle differentiation (Cao et al. 2010). Data discussed in the previous section show a strong presence of the AP-1 site in a subset of regions occupied by MyoD in cycling myoblasts ("early" MyoD regions). This association abates completely, on the sequence level, in "late" MyoD regions (those preferentially occupied in differentiating myocytes), where the AP-1 motif is depleted (Figure 3.5). To better understand the nature of AP-1 action in myoblasts and its collaboration with MyoD, Fosl1 (also known as Fra1) occupancy was measured in cycling C2C12s (Marinov, unpublished). The ChIPSeq measurement gave 2261 HC regions (Table 3.1b), which I analyzed for motif content, overlap with MyoD, proximity to nearest TSS, and expression patterns of nearest gene.

Motif mapping showed a strong central enrichment for the cited motif TGAGTCA, as well as for the alternative matrix TGASTCACW derived from early MyoD regions by MEME (Figure 3.11a). Given their similar enrichment patterns and the fact that TGAGTCA covered twice as many regions as TGASTCACW (70% and 35%, respectively) (Table 3.11b), the former was used as the representative Fosl1 site for the purposes of analysis. When compared side-by-side to MRF-occupied regions, differences in sequence profiles were readily apparent (Figure 3.12). The RRCAGSTG motif occurred at background levels, with no statistically significant difference from genomic density, although this was the result of a minor enrichment of the RRCAGCTG component and a minor depletion of RRCAGGTG (both statistically significant, $p < 0.01$). This is consistent with lack of binding affinity between Fosl1 and RRCAGSTG. There was also no enrichment of CTCF or Klf4 sites, both of which occurred at background levels, and

there was a significant ($p < 0.01$) and substantial (almost threefold) depletion of the CGCGCG Set1 target site. Curiously, the bias against TATA-boxes and CAWWTG class e-boxes persisted, with all species depleted ($p < 0.01$) compared to expectation (Figure 3.12). The motif content differences became more pronounced when occupancy maps for Fosl1 and MyoD in cycling myoblasts were compared for overlap. Regions specific to Fosl1 (no overlap with MyoD MC regions) showed a significant ($p < 0.01$) two-fourfold depletion for all e-boxes of the form CAGSTG, and were depleted of every other e-box motif except CACGTG and CAGATG, which occurred at background levels (Figure 3.12). Taken together with a similar depletion of AP-1 sites in MyoD "late" regions, these data point to a strong functional impetus for keeping parts of the two networks separate.

On the other end of the spectrum, there were 215 regions of joint occupancy between the "early" MyoD subset and Fosl1. The AP-1 motif covers 68.8% of the shared regions, whereas RRCAGSTG only covers 24.2%. Furthermore, of the 52 regions with an RRCAGSTG motif, 40 also contained an AP-1 site, leaving only 12 out of 215 (5.6%) that have an MRF recognition sequence without an accompanying AP-1 motif. When both motifs were present, DNAse hypersensitivity footprints were more proximal to the AP-1 motif than to the octameric e-boxes. While there was no way to distinguish between joint occupancy and sample heterogeneity effects, sequence content strongly suggests that MyoD is recruited to these sites primarily by means of protein-protein interactions. There was no evidence to support this recruitment the other way, as only 12 jointly occupied regions that contain RRCAGSTG but not TGAGTCA. In comparison, "late" MyoD regions had zero overlap with Fosl1 occupancy, even when compared to MC-threshold regions - Fosl1 occupancy and MyoD "late" occupancy are completely disjoint, without even a partial overlap. Such a strong exclusion is rather remarkable, even if consistent with motif content analysis - "late" MyoD sites are significantly ($p < 0.01$) depleted of AP-1 motifs. The interaction between MyoD and Fosl1 is therefore limited to cycling myoblasts, and indeed the levels of Fosl1 transcript diminish by almost 90% in differentiating myocytes (Table 3.7). This does not preclude a

potential interaction with Fosl2 (also known as Fra2), which is expressed in differentiating C2C12, but makes it less probable due to the depletion of supporting motifs. If such an interaction were to occur, it would have to rely on MyoD:E binding DNA, then bringing in Fosl2 through secondary means.

Gene association analysis of Fosl1 occupancy gave results similar to those obtained from the MRFs (Figure 3.6a). There was a higher fraction of regions associated with myoblast genes (5.2%), however it was small compared to the total size of the dataset (Table 3.6b). Fosl1 occupancy is likely ($p < 0.01$) to be associated with genes that are expressed, and 49.2% of Fosl1 occupied regions are associated with either flat or undetermined genes. While a lower fraction than was observed for the MRFs, it still encompasses half of all detected Fosl1 occupancy events.

3.4 Understanding the role of Mef2 in the regulation of skeletal muscle differentiation

Mef2 has been shown to directly regulate several genes specific to developing skeletal muscle (see Introduction), where functional studies suggest a positioning pattern. Typically, Mef2-responsive regulatory modules contained a Mef2 A/T rich recognition site and an MRF-class e-box (CAGSTG) within 100 - 150 nucleotides of one another. This suggests a collaborative relationship, where the two factors occupy proximal sequence elements and jointly regulate the expression of the target gene. Further studies showed that Mef2 is able interact *in vitro* with Myog:E12 heterodimers, but not myogenin or E12 on their own (Molkentin et al. 1995); and that Myog:E12 complexes can be recruited to a sequence element by Mef2 even in the absence of an CAGSTG e-box, provided a Mef2 recognition site was present. Consequently, Mef2 can regulate transcription of its target genes both through direct protein-DNA binding, and through secondary recruitment by an MRF:E heterodimer. To evaluate the relative prevalence of these interactions, I used a Mef2 occupancy determination in differentiating C2C12s (Fisher-Aylor, unpublished), which was analyzed for sequence content, gene associations, and overlap with the myogenin occupancy map.

Initially, several chromatin immunoprecipitation experiments using the protocol from Mortazavi et al (2007) and targeting Mef2 were unsuccessful, necessitating the use of a stronger fixing agent (gluteraldehyde) before usable data could be collected. Because three members of the Mef2 family are present in C2C12s at various stages of development (Table 3.7), a pan-Mef2 antibody was used. ERANGE identified 3072 HC regions occupied by Mef2 in differentiating myocytes (60 hours after withdrawal of serum), which were analyzed for sequence content through a combination of *de novo* motif discovery and pre-defined motif mapping, yielding some unanticipated results. Based on the prior observations from myogenin, MyoD and Fosl1, I expected to a strong coverage of the dataset by the canonical Mef2 motif CTAWWWWTAG. In fact, only 3.8% of Mef2 regions (Table 3.8b) contained the

literature site - this represents a three-fold enrichment ($p < 0.01$) over genomic background (Figure 3.9a), but by itself fails to account for the vast majority of occupied regions. Another form of the motif - AAATAG, used by Cao et al. (2010) in their investigation of differential MyoD occupancy, showed a significant ($p < 0.01$) depletion relative to genomic background, and consequently its coverage of the Mef2 occupancy map was lower than that of randomly generated regions (Table 3.8b). *De novo* motif finding derived a version of the canonical site when presented with the list of 500 strongest occupancy events (organized by ChIPSeq signal associated with the event), but it was superseded in frequency and abundance by three other motifs. One was the primary MRF motif RRCAGSTG, observed in 52.5% of Mef2-occupied regions. Another was CACGTGAC, which based on a literature search and existing motif databases was attributed to Usf1 - a c-myc related member of the bHLH family that will be discussed in section 3.6. The Usf1 motif was highly enriched compared to the whole genome background (twentyfold enrichment, $p < 0.01$) but only covered 9.3% of the Mef2 dataset. The third non-canonical motif derived by MEME was GGGANWTGWAGT, which resembles a combination of the canonical Mef2 and the canonical SRF motifs. This site will be referred to as Mef2-hybrid in the analysis, and covered 17.1% of the Mef2 regions (Table 3.10).

The high coverage by RRCAGSTG and general lack of canonical sites in Mef2 occupancy events implied that most Mef2-MRF collaboration events in muscle cells are the result of protein-protein recruitment. An occupancy map comparison between Mef2 and myogenin produced 1834 regions with some direct overlap (referred to as Mef2+Myog), and 1238 regions that had Mef2 occupancy and were completely disjoint from myogenin regions (referred to as Mef2-Myog). Mef2-Myog regions were highly promoter proximal - 840 (67.9%) were centered between -1000 and +250 from an annotated TSS; and 1034 (83.5%) were associated with either flat (640) or undetermined (394) genes. They showed no enrichment for RRCAGSTG (Figure 3.9a). A substantial minority were covered by the Mef2-hybrid motif (34.5%, Table 3.10), with the central enrichment tendency typical of the primary sequence responsible

for the occupancy event (Figure 3.9b). This led me to theorize that Mef2-hybrid is an alternative recognition sequence for Mef2, and to test it using EMSA. There was no detectable *in vitro* binding between either Mef2a or Mef2d and the Mef2-hybrid motif, although both Mef2a:Mef2a homodimers and Mef2a:Mef2d heterodimers were able to bind the canonical motif CTAWWWWTAG (Figure 3.13). Mef2-Myog regions additionally had a thirtyfold enrichment of the Usf1 motif ($p < 0.01$), with the same central tendency as Mef2-hybrid (Figure 3.9e). Its relatively low coverage (16.3%) and known association made it an unlikely candidate for direct Mef2 binding, but in turn made Usf1 a likely collaborating factor.

Mef2+Myog regions closely resemble myogenin regions in terms of their motif content (Figure 3.9a). They have the highest incidence of the Mef2 canonical motif CTAWWWWTAG, but it only covers 4.3% of Mef2+Myog regions (Table 3.10). The majority of Mef2+Myog regions are covered by RRCAGSTG (75.6%), which is consistent with rate of coverage for the myogenin set as a whole. The Mef2-hybrid motif was also enriched ($p < 0.01$), but only covered 5.3% (Table 3.10). The RRCAGSTG motifs in Mef2+Myog, as well as both Usf1 and Mef2-hybrid motifs in Mef2-Myog show a strong central tendency (Figures 3.9b-e). The canonical motif CTAWWWWTAG first found in a number of muscle-specific enhancer and promoter elements does associate preferentially with genes that are heavily upregulated during terminal differentiation (twofold motif enrichment, $p < 0.01$). This association is, however, rare, and a vast majority of RRCAGSTG boxes showing myogenin occupancy (>99%) do not contain such a motif within a 250 bp radius.

The occupancy map of Mef2 consists of groups of regions, each featuring a different motif with the central tendency characteristics of a primary target, but for which Mef2 has no direct binding affinity. Mef2a and Mef2d do not bind either RRCAGSTG, CACGTGAC, or the "hybrid" site GGGANWTGWAGT *in vitro*. The one motif Mef2 does bind - CTAWWWWTAG - covers only 5.3% of detected occupancy

events. MRFs and Mef2 do share occupancy - 60% of detected Mef2 regions have a corresponding myogenin signature, and, conversely, 12% of myogenin HC regions have an overlapping Mef2 signal. Finally, gluteraldehyde fixation used for the Mef2 determination increases the odds of detecting secondary interactions when compared to paraformaldehyde. Put together, these results show that Mef2 is primarily directed to its target sites through protein-protein interactions. Direct binding by Mef2 is not necessary for recruitment or exertion of regulatory activity, although functional studies demonstrate that it leads to an increase in transcriptional output when present. In skeletal muscle, the vast majority of interactions are initiated through MRF:E binding DNA, then bringing in Mef2. There were no instances where a myogenin signal originated over a Mef2 canonical site in the absence of an accompanying RRCAGSTG, compared to 1323 instances where a Mef2 signal originating over a myogenin-occupied RRCAGSTG without an accompanying CTAWWWWTAG.

Overall, Mef2 acts as a transcriptional regulator in differentiating skeletal muscle, but its targeting mechanism differs from that of the MRFs. It is more limited in the number of sites it occupies and correspondingly in the number of genes it likely affects. A large portion of Mef2 occupancy is directed at proximal promoters of genes that maintain a stable transcript level in both progenitor myoblasts and terminally differentiated myocytes. This can be interpreted in two ways - Mef2 could have a strong "housekeeping" role, or it could be recruited to active promoters because it is available. Both are plausible, and the role of Mef2 merits further investigation. It can be further studied by mapping of occupancy for Mef2a, Mef2c and Mef2d explicitly, as all three are present and up-regulated in differentiating skeletal muscle (Mef2d is also present in cycling myoblasts), but could have different binding preferences or sets of targets. Clarifying the binding preferences for various Mef2 dimers and functional studies that focus on the effects of secondary Mef2 presence at regulatory elements would also provide valuable insights.

3.5 CTCF-occupied sites act as insulators in differentiating skeletal muscle

CTCF is an 11-Zinc finger factor that is highly conserved in most vertebrates (Filippova et al. 1996), and has been associated with transcriptional repression (Baniahmad et al. 1990), activation (Klenova et al. 1993) and most prominently insulation (Bell et al. 1999). It was first described as a regulator of c-myc in *gallus gallus* (Lobanenkov et al. 1990), but has since been ascribed a variety of functions related to chromatin remodeling - such as facilitation and maintenance of looping (Splinter et al. 2006). CTCF message is present in both myoblasts and myocytes, although its amount diminishes over time (Table 3.7). A motif associated with CTCF occupancy is enriched ($p < 0.01$) in regions occupied by a number of assayed transcription factors, and although the absolute number of instances varies, the motif is complex enough so that even a relatively small number of occurrences can represent a statistical enrichment. The consistency with which this enrichment occurs suggests an overlap between occupancy maps of CTCF and studied TFs. Additionally, a CTCF occupancy map provides an immediate avenue to test its function as an insulator in differentiating skeletal muscle. I used data from Mikkelsen (unpublished) to analyze CTCF occupancy in cycling myoblasts and in differentiating myocytes.

The first, and primary, observation to come out of the CTCF analysis was about insulation. When comparing CTCF and myogenin occupancy at 60 hrs after differentiation, CTCF regions were much more likely to occur between a myogenin region and a TSS of down-regulated gene than any other type of TSS. The percentage of "open" connections (no CTCF occupancy between myogenin region and TSS) was highest for myocyte genes, somewhat lower for flat and unexpressed genes, and lowest for myoblast genes. This frequency was statistically indistinguishable when comparing the flat and unexpressed sets. Statistical difference was highest between myocyte and myoblast sets ($p < 1x10^{-6}$), although myocyte connections were also somewhat more likely to be open than either flat or unexpressed ones ($p < 0.01$) (Table 3.15). This behavior is consistent with insulation of active myogenin sites from "undesirable"

genes, e.g., those that are down-regulated upon entry into terminal differentiation. A similar observation was used by Cao et al. (2010) when associating regions of MyoD occupancy events with their target genes, leading support to the idea that CTCF acts as an insulator in differentiating myocytes.

The CTCF occupancy determination itself was subjected to the same sequence content, gene association and overlap analyses as the other ChIPSeq datasets discussed herein. Overview of the primary results for both sets of regions are summarized in Table 3.1b. CTCF-occupied regions are among the most likely to be associated with unexpressed genes - nearly 40% of occupancy events in cycling myoblasts were associated with genes lacking measurable transcript levels (Figure 3.6a), which is consistent with CTCF's ability to act as both a repressor and an insulator, and leads us to expect a higher likelihood of CTCF occupancy associating with unexpressed or down-regulated genes. The latter expectation, however, does not pan out - the fraction of associations with down-regulated genes is comparable to that for myogenin and other positive-acting factors (MyoD and Usf1), and did not increase between cycling and differentiated states. In conjunction with the results from Table 3.15, this implies that CTCF more likely acts as an insulator than a direct repressor during skeletal muscle differentiation.

3.6 Usf1 occupancy in C2C12s suggests minimal involvement in the muscle differentiation network

Usf1 is a bHLH transcription factor similar to c-myc by virtue of having the leucine-rich domain (Gregor et al. 1990), and was first isolated from HeLa cells due to its ability to activate the adenoviral major late promoter (MLP) (Sawadogo et al. 1988). Like all bHLH proteins, it binds DNA in the form of a dimer, and can both homodimerize and heterodimerize with Usf2 (Sirito et al. 1992). The reported recognition motif is the octamer CACGTGAC (Rada-Iglesias et al. 2008), although the strictness of the flanking nucleotides, as well as potential promiscuity in the central nucleotides, deserves further investigation. Usf1 originally garnered attention during the analysis of Mef2 occupancy data, where a high density of CACGTG hexamers and CACGTGAC octamers was discovered in the set of regions that had no overlap with myogenin (Figure 9a). The latter were predominantly associated with "flat" genes (those that remain stably expressed both in cycling and differentiating C2C12s) , and the mRNA levels for both Usf1 and Usf2 remain virtually constant throughout the differentiation timeline, with less than a 10% difference between cycling myoblasts and myocytes 7 days after onset of differentiation (Table 3.7). Based on the Mef2 analysis and published data, the expectation was for Usf1 to preferentially occupy promoters of genes whose expression remains stable throughout the differentiation process. To put it another way, it has the markings of a "housekeeping" transcription factor. To test this, as well as to investigate any potential role that Usf1 plays in skeletal muscle differentiation, genome-wide occupancy mapping was performed (Marinov, unpublished).

Peak calling generated 3375 and 4516 regions at HC stringency for the 60 hr and cycling timepoints, respectively. The two sets of regions showed remarkable concordance - of the 3375 regions occupied at 60 hrs, 2875 were a 90%+ match to those occupied in the cycling state. When the criteria for a match were relaxed to include the MC set of cycling regions, 3096/3375 were a 90%+ match by nucleotide overlap (see Methods) - a similarity this high was often observed in technical replicates of the same

experiment. Conversely, 3856/4516 cycling HC regions matched 60h MC regions at the same 90%+ threshold, suggesting that the ultimate difference between the two measurements is likely due to experimental variation rather than a physiological change in occupancy profile. The sites occupied by Usf1 remain virtually unchanged between myoblast and myocyte states.

There is relatively little overlap between Usf1 and myogenin occupancy in differentiating myocytes (60 hrs after withdrawal of serum). Out of 3375 Usf1 HC regions, only 321 overlapped myogenin HC regions with 80%+ nucleotide identity, while 2845 (84.3%) showed no overlap at all. When using the broader MC myogenin region set, 2319 Usf1 sites (68.8%) remained completely disjoint, while the number of overlapping regions increasing to 596. Of those, 224 (37.6%) were within 1KB of a TSS, with the rest distributed over a range of distances. Furthermore, 448 out of 596 (75.1%) such regions were associated with genes that fell into either the flat or the undetermined categories, as opposed to 24 (4.0%) that were associated with differentially expressed genes (muscle-up or muscle-down). This pattern was consistent with expectation.

The canonical primary motif CACGTGAC showed a strong and central tendency in the 60 hr Usf1 dataset, however, only 1088 (32.2%) regions had such a motif (Table 3.8b). The coverage increased substantially when the motif was relaxed to the core hexamer form (CACGTG), adding a further 798 regions, for a total coverage of 1886/3375 (55.9%). Still, this is below the primary motif coverage observed for MyoD, myogenin or Fosl1. Most other motifs enriched in regions occupied by Usf1 in myocytes - Mef2 (hybrid), CTCF, Klf4, CGCGCG, Sp-1, were those generally found in promoters (Figure 2.9a). Enrichment of CAGSTG and RRCAGSTG motifs was abolished by removing the ~600 regions jointly occupied by Usf1 and myogenin (Figure 3.14). In aggregate, these data support the hypothesis that Usf1 is an active bHLH-class transcription factor in the C2C12 system, and that it plays a role in the maintenance of overall cell function, but bears little direct impact on the developmental network responsible for skeletal muscle

differentiation. Instead, Usf1 activates genes the need to be transcribed in both the cycling progenitors and the terminally differentiated myocytes.

A couple questions merit further investigation. First, based on the behavior of other bHLH transcription factors, most of which allow some variation in their recognition sites, the possibility that Usf1 has flexibility at both flanking nucleotide positions and the primary e-box core must be considered. The obvious approach of taking Usf1 regions lacking CACGTG and running them through *de novo* motif finding failed to yield conclusive results, however, a deeper investigation of this issue could prove fruitful. I theorize that some amalgam of CASSTGNN sites are actually acceptable, with affinity varying based on exact sequence. This problem was partly addressed during the analysis of Usf1 and Usf2 occupancy in humans (Rada-Iglesias et al. 2008), however the latter only confirmed CACGTGAC as the mostly likely optimal occupancy site, without ruling out occupancy of other, less optimal configurations. A more thorough statistical analysis of occupied motifs or *in vitro* protein-DNA binding experiments are logical next steps in determining the range of binding preferences for Usf1/2. Second, because Usf1 associates with active promoters, likely to harbor large regulatory complexes, the prevalence of Usf1 occupancy from secondary interactions should also be considered. Distinguishing between primary and secondary Usf1 sites could lead to a finer understanding of its function, its partners, and its targeting.

References (chapter 3)

Andreucci, J.J., Grant, D., Cox, D.M., Tomc, L.K., Prywes, R., Goldhamer, D.J., Rodrigues, N., Bedard, P.A., and McDermott, J.C. (2002). Composition and function of AP-1 transcription complexes during muscle cell differentiation. J Biol Chem *277*, 16426-16432.

Baniahmad, A., Steiner, C., Kohne, A.C., and Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. Cell *61*, 505-514.
Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387-396.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L.*, et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell *18*, 662-674.

Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol Cell Biol *16*, 2802-2813.

Gregor, P.D., Sawadogo, M., and Roeder, R.G. (1990). The adenovirus major late transcription factor USF is a member of the helix-loop-helix group of regulatory proteins and binds to DNA as a dimer. Genes Dev *4*, 1730-1740.

Han, T.H., Lamph, W.W., and Prywes, R. (1992). Mapping of epidermal growth factor-, serum-, and phorbol ester-responsive sequence elements in the c-jun promoter. Mol Cell Biol *12*, 4472-4477.
Han, T.H., and Prywes, R. (1995). Regulatory role of MEF2D in serum induction of the c-jun promoter. Mol Cell Biol *15*, 2907-2915.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenkov, V.V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol Cell Biol *13*, 7612-7624.

Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743-1753.

Molkentin, J.D., Black, B.L., Martin, J.F., and Olson, E.N. (1995). Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. Cell *83*, 1125-1136.

Rada-Iglesias, A., Ameur, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T.R., and Wadelius, C. (2008). Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res *18*, 380-392.

Sawadogo, M., Van Dyke, M.W., Gregor, P.D., and Roeder, R.G. (1988). Multiple forms of the human gene-specific transcription factor USF. I. Complete purification and identification of USF from HeLa cell nuclei. J Biol Chem *263*, 11985-11993.

Sirito, M., Walker, S., Lin, Q., Kozlowski, M.T., Klein, W.H., and Sawadogo, M. (1992). Members of the USF family of helix-loop-helix proteins bind DNA as homo- as well as heterodimers. Gene Expr *2*, 231-240.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev *20*, 2349-2354.

Figures and Tables (chapter 3)

	MyoD cycling	MyoD 24h diff	MyoD 60h diff
Number of regions (HC)	6,651	9,922	16,437
Number of regions (MC)	11,790	18,844	30,287
Nucleotide coverage (HC)	4,092,025	3,399,686	5,473,031
Read % in regions (HC)	7.74%	10.72%	22.45%
Average region length (HC)	616	343	333
Median region length (HC)	561	320	300
Length standard deviation (HC)	247	111	139
Minimum length (HC)	164	110	103
Maximum length (HC)	2,786	1,403	1,687

Table 3.1a. Primary statistics for MyoD occupancy measurements. HC = high confidence, MC = medium confidence. The stringency settings for HC vs. MC were consistent across all data sets, and as a result most exhibited the property of having approximately 2x more MC regions. To limit false positives, primary sequence and gene association analysis was focused on HC regions. However, when evaluating differential occupancy, HC regions were filtered using the MC set from the other time point, and only those without a match were declared differentially occupied. Nucleotide coverage refers to total number of nucleotides covered by regions as called, giving a rough estimate of the portion of the genome associated with factor occupancy. Read % in regions is a quality control metric, and refers to the total number of sequenced reads that fall within enriched areas. Higher ratios are desirable, although usable data was obtained from experiments with as few as 1% of reads falling into enriched areas.

	Fosl1	Mef2	CTCF c	CTCF m	Usf1 c	Usf1 m
Number of regions (HC)	2,261	3,072	21,236	14,351	4,516	3,373
Number of regions (MC)	6,269	11,179	29,036	22,381	9,002	7,149
Nucleotide coverage (HC)	851,347	1,247,555	7,338,588	4,448,579	2,295,114	1,501,411
Read % in regions (HC)	2.67%	3.10%	20.88%	12.23%	11.26%	9.19%
Average region length (HC)	377	406	346	310	508	445
Median region length (HC)	366	368	329	288	468	397
Length std. deviation (HC)	94	171	111	107	257	242
Minimum length (HC)	78	60	89	102	104	103
Maximum length (HC)	892	1,486	1,970	1,854	10,628	4,641

Table 3.1b. Primary statistics for ChIPSeq occupancy measurements for Fosl1, Mef2, CTCF and Usf1. HC = high confidence, MC = medium confidence. CTCF c and Usf1 c were done in cycling myoblasts. CTCF m and Usf1 m were done in myocytes - 60 hours after withdrawal of serum for Usf1 and 7 days after withdrawal of serum for CTCF. Nucleotide coverage refers to total number of nucleotides covered by regions as called, giving a rough estimate of the portion of the genome associated with factor occupancy. Read % in regions is a quality control metric, and refers to the total number of sequenced reads that fall within enriched areas. Higher ratios are desirable, although usable data was obtained from experiments with as few as 1% of reads falling into enriched areas.

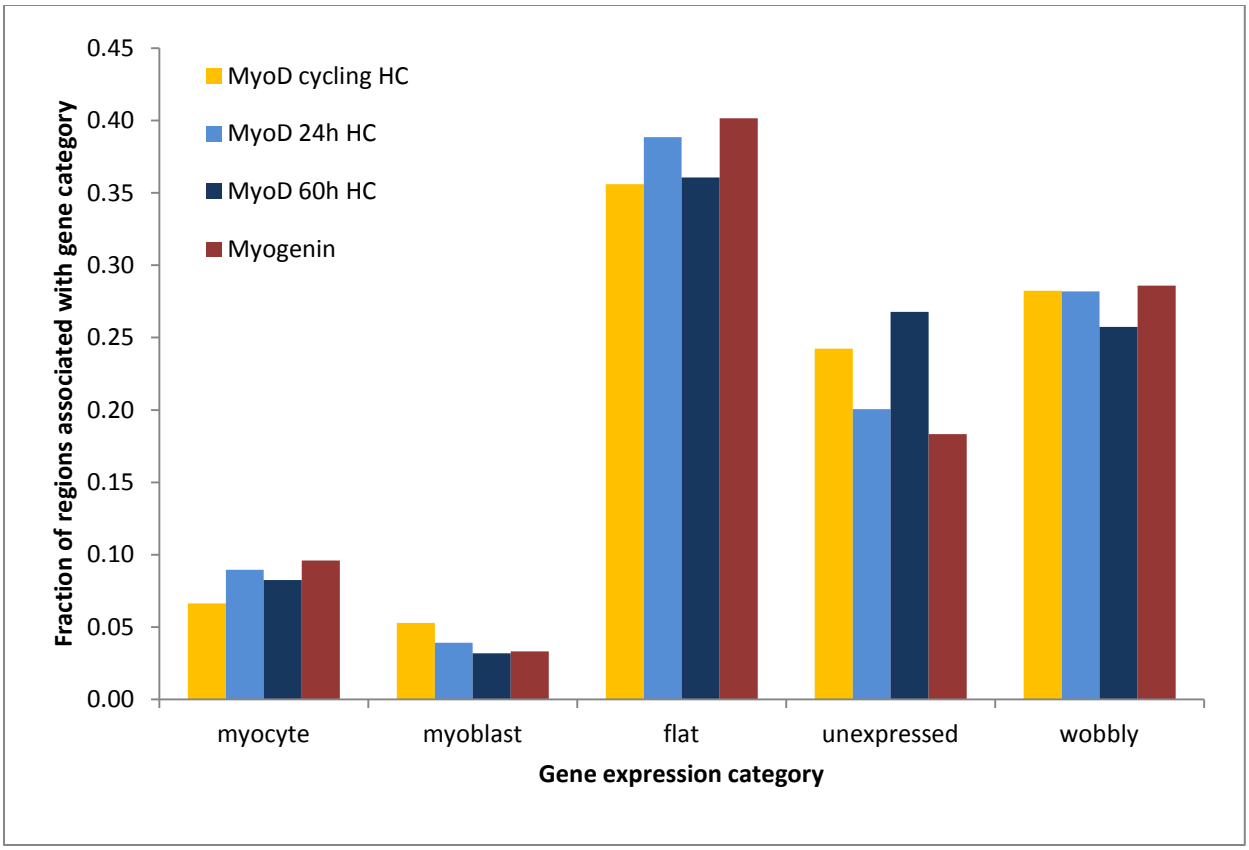


Figure 3.2. Association between MyoD-occupied regions and their nearest gene, organized by expression category of the gene. Fraction of regions associated with myocyte genes increases upon differentiation, while the opposite is true of the associations with myoblast genes. The majority of regions in each dataset (60-70%) are associated with expressed genes whose mRNA levels do not change drastically (flat and wobbly categories). The total number of regions associated with radically changing genes is <15%.

	MyoD cycling	MyoD 24h	MyoD 60h	Myogenin
MyoD cycling	-	70.8%	71.7%	52.5%
MyoD 24h	52.2%	-	93.4%	80.9%
MyoD 60h	*35.5%*	78.5%	-	82.3%
Myogenin	*29.1%*	71.5%	87.2%	-

Table 3.3a. Region overlap between respective datasets. Regions were length-normalized to ±250 bp around the summit, comparison was done by requiring an 80%+ coordinate overlap between a region in the source HC set (horizontal) compared to one from the target MC set (vertical). HC to MC comparison was performed to limit the number of false negatives. Note that some negatives will inevitably arise if the source set is larger than the target set, as the excess regions will have no match by definition. Such comparisons are italicized, and it is a known limitation of ChIPSeq data.

	MyoD cycling	MyoD 24h	MyoD 60h	Myogenin
MyoD cycling	-	1,760	1,630	2,812
MyoD 24h	4,474	-	364	1,413
MyoD 60h	*10,197*	3,085	-	2,088
Myogenin	*10,067*	3,682	1,257	-

Table 3.3b. Absolute number of HC regions without a match to the target MC set. Comparisons where the source set is larger than the target set are italicized.

Motif	Myog	MyoD cyc	MyoD 24h	MyoD 60h
RRCAGSTG	2.78	2.54	2.96	2.74
RRCAGCTG	3.16	2.95	3.34	3.04
RRCAGGTG	1.97	1.67	2.19	2.18
CACGTG	1.02	0.51	0.73	1.01
CATATG	-1.34	-0.80	-1.08	-1.28
CAAATG	-0.67	-0.42	-0.56	-0.72
CAATTG	-1.16	-0.69	-0.96	-1.12
CAGATG	0.26	0.42	0.42	0.25
CAAGTG	0.11	0.01	0.17	0.04
CACATG	-0.15	-0.16	-0.09	-0.19
CAACTG	0.29	0.20	0.29	0.19
Meis	0.77	0.65	0.84	0.76
AP-1	0.97	2.34	1.14	1.03
Mef2 (half)	-1.20	-0.91	-1.08	-1.19
Runx	1.06	1.02	0.98	0.93
RP58	-0.22	0.40	0.08	-0.19
CTCF	2.42	0.78	1.74	2.65
Mef2 (lit)	0.22	-0.36	0.07	0.10
Klf4 - GGGYGKGG	1.89	0.99	1.49	1.85
CGCGCG	3.37	0.04	2.05	3.71
TATA-box	-2.36	-1.71	-2.06	-2.22
Sp-1	2.80	1.21	2.09	2.95
Dec1 - KCACGTGM	1.03	0.29	0.51	1.14

Figure 3.4a. Comparative motif enrichment analysis for MyoD and myogenin regions. Enrichments are computed over genomic background, and normalized using $\log_2$. Ratios that do not pass a statistical significance test ($p > 0.01$) are displayed in grey.

	-250				0					250
MyoD cycing	0.49	0.64	1.03	1.52	4.35	4.43	1.55	1.04	0.88	0.40
MyoD 24h	0.68	1.04	1.62	2.17	4.74	4.79	2.08	1.56	1.13	0.74
MyoD 60h	0.54	0.73	1.40	1.90	4.53	4.56	1.84	1.31	0.74	0.59
Myogenin	0.57	0.88	1.27	2.24	4.50	4.56	2.19	1.26	0.92	0.61

Figure 3.4b. The motif RRCAGSTG is highly centrally enriched in all MyoD and myogenin occupied regions. Enrichment values are normalized using $\log_2$, every ratio in this example is significantly different from genomic background ($p < 0.01$).

Motif	early	cont.	late	cont. ex.	MyoD ex.	Myog ex.
RRCAGSTG	1.31	2.81	3.16	2.35	2.31	2.41
RRCAGCTG	1.50	3.24	3.48	2.67	2.35	2.87
RRCAGGTG	1.00	1.85	2.53	1.71	2.25	1.32
CACGTG	0.07	0.68	0.56	0.12	0.65	0.78
CATATG	-0.48	-0.94	-1.02	-0.72	-1.06	-1.52
CAAATG	-0.39	-0.46	-0.40	-0.30	-0.60	-0.67
CAATTG	-0.61	-0.85	-0.97	-0.87	-0.91	-1.54
CAGATG	0.41	0.42	0.47	0.51	0.29	0.24
CAAGTG	-0.01	0.11	0.27	-0.02	-0.10	0.08
CACATG	-0.16	-0.11	0.09	-0.29	-0.25	-0.38
CAACTG	0.17	0.25	0.62	0.11	0.07	0.37
Meis	0.22	0.78	0.93	0.66	0.67	0.61
AP-1	2.99	1.94	-0.78	1.82	1.12	0.65
Mef2 (half)	-0.74	-1.00	-1.00	-0.62	-0.81	-1.26
Runx	0.73	1.11	0.84	0.60	0.50	1.00
RP58	0.79	0.27	-0.20	0.66	-0.04	-0.03
CTCF	-1.50	1.24	0.69	0.20	2.12	2.68
Mef2 (lit)	-0.49	-0.30	0.28	-1.56	0.06	0.26
Klf4	0.20	1.36	1.18	0.58	1.15	1.95
CGCGCG	-3.29	0.69	0.01	-2.76	3.14	3.45
TATA-box	-1.36	-1.98	-1.91	-1.41	-1.55	-2.60
Sp-1	-0.71	1.55	1.13	0.52	2.27	2.82
Dec1 - KCACGTGM	0.19	0.46	0.31	0.06	0.60	0.79

Figure 3.5. Sequence content analysis of differentially occupied regions. Enrichment values are normalized using $\log_2$. **cont.** = regions occupied by MyoD in both myoblasts and myocytes. **cont. ex.** = same as cont., but with no overlap to myogenin occupancy. **MyoD ex.** = regions occupied by MyoD but not by myogenin at 60h after withdrawal of serum. **Myog ex.** = same as MyoD ex., but in reverse. Ratios that do not pass a statistical significance test ($p > 0.01$) are displayed in grey.

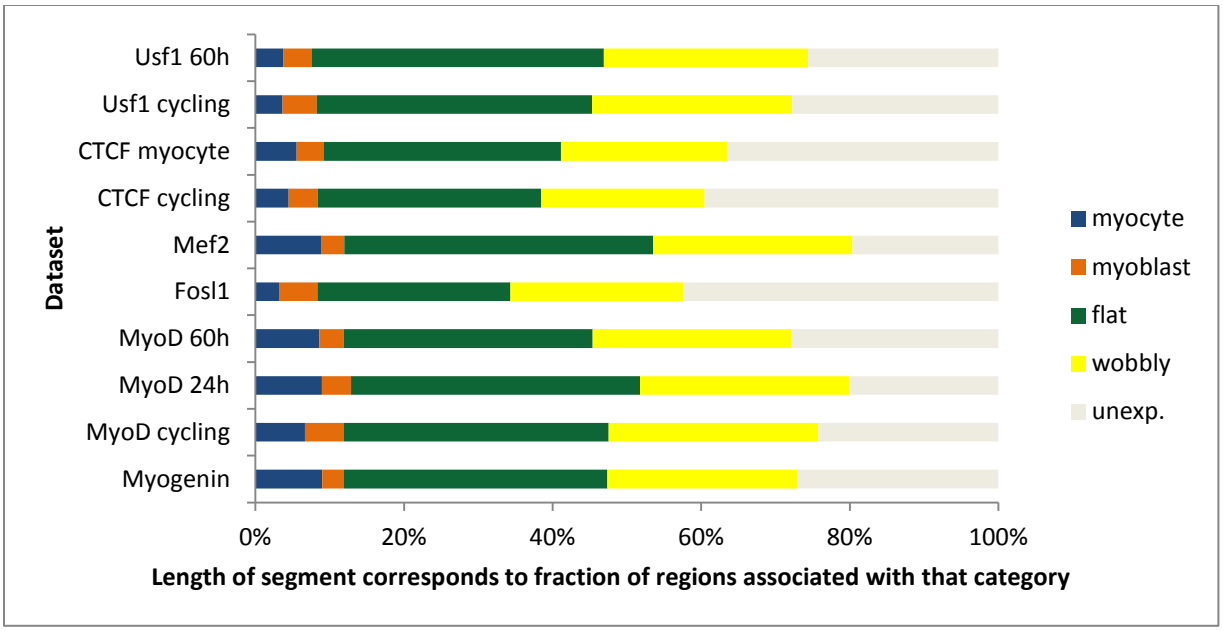


Figure 3.6a. Region-gene associations, ordered by expression category of the gene. This analysis counts the number of regions associated with each category of genes, so occasionally the same gene contributes more than one region to the count. The region-centric approach attempts to detect biases towards a particular expression profile, if any.

	myocyte	myoblast	flat	wobbly	unexp.
Myogenin	1,331	430	5,242	3,788	3,995
MyoD cycling	441	351	2,364	1,876	1,609
MyoD 24h	888	388	3,850	2,793	1,988
MyoD 60h	1,357	522	5,297	4,231	4,400
Fosl1	72	117	587	526	959
Mef2	273	94	1,277	825	603
CTCF cycling	949	826	6,381	4,667	8,380
CTCF myocyte	795	519	4,584	3,216	5,237
Usf1 cycling	164	209	1,673	1,218	1,252
Usf1 60h	127	130	1,324	928	864

Table 3.6b. Absolute numbers of regions associated with genes in each expression category. These numbers were used to generate Figure 3.6a.

Transcription factor	cycling	diff - 60h	diff - 5d	diff - 7d
MyoD	167.4	200.6	131.8	84.5
Myogenin	15.4	911.5	545.1	346.8
Mef2a	7.9	33.1	27.1	25.8
Mef2b	0.0	0.0	0.0	1.4
Mef2c	1.0	20.6	27.6	31.3
Mef2d	22.3	58.2	70.7	98.8
Myf5	6.9	5.5	6.1	6.1
Mrf4	0.1	2.4	9.3	24.3
Fosl1	94.2	9.2	9.3	7.1
Fosl2	11.3	19.4	23.6	35.6
Junb	82.1	61.5	33.2	58.6
Jund	69.8	115.1	72.6	129.7
CTCF	17.2	10.1	6.3	4.6
Dec1	12.0	73.0	70.7	64.1
E47/E12	64.4	33.2	27.4	35.8
Tcf12 (HEB)	22.5	25.0	18.0	14.0
Klf4	6.9	42.0	18.8	18.1
Sp1	8.9	7.2	5.6	5.5
RP58	2.2	14.2	10.9	9.0
Zeb1	6.2	20.7	19.0	16.8

Table 3.7. mRNA levels for various relevant transcription factors. Levels are given in RPKM, based on an RNASeq measurements taken at four time points. The time points are: cycling myoblasts (cycling), differentiating myocytes 60 hours after withdrawal of serum (diff - 60h), differentiating myocytes 5 days after withdrawal of serum (diff - 5d), and maturing myocytes 7 days after withdrawal of serum (diff - 7d).

Motif	MyoD cyc	MyoD 24h	MyoD 60h	Myogenin	Randomized
CAGSTG	81.4%	89.6%	86.0%	87.0%	42.6%
CAGCTG	66.0%	77.4%	69.5%	75.4%	21.8%
CAGGTG	38.8%	47.5%	48.3%	45.0%	28.0%
RRCAGSTG	69.9%	82.9%	74.7%	76.0%	20.0%
RRCAGCTG	56.0%	68.3%	58.6%	63.9%	11.9%
RRCAGGTG	23.9%	31.9%	31.1%	27.6%	9.5%
CACGTG	6.7%	7.7%	9.0%	9.2%	5.2%
AP-1	36.1%	16.8%	15.7%	15.1%	9.0%
Mef2 (Tapscott)	20.2%	18.3%	17.0%	16.9%	33.0%
Mef2 (lit)	1.0%	1.4%	1.4%	1.5%	1.3%
Klf4	17.3%	23.3%	28.0%	29.1%	10.4%
CTCF	1.9%	3.8%	6.6%	5.7%	1.5%
Usf1	1.2%	1.2%	1.8%	1.7%	0.7%
CGCGCG	0.7%	2.6%	7.2%	5.9%	0.7%
TATA-box	14.9%	12.2%	11.0%	10.2%	32.8%
Sp-1	10.6%	17.1%	24.8%	23.9%	6.4%

Table 3.8a. Motif coverage of MRF-occupied regions. Coverage is defined as the percentage of the total number of regions that contain at least one copy of the motif. Randomized regions are included as a representation of expected coverage - the set consists of approximately 101000 randomly selected regions, each 501 nucleotides long (matching the standardized length of ChIP regions). They were filtered to eliminate repeats and unsequenced nucleotides, as well as telomeric/centromeric sequence. This table list coverages for all occupied regions obtained from the dataset denoted at the top of the column.

Motif	Fosl1	Mef2	CTCF cyc	CTCF myoc	Usf1 cyc	Usf1 60h	Randomized
CAGSTG	35.0%	66.9%	43.6%	47.5%	42.4%	44.7%	42.6%
CAGCTG	22.0%	52.7%	21.2%	23.4%	26.5%	28.9%	21.8%
CAGGTG	18.3%	38.5%	30.1%	32.8%	23.2%	23.8%	28.0%
RRCAGSTG	17.0%	52.5%	21.3%	23.8%	21.5%	23.6%	20.0%
RRCAGCTG	12.5%	42.4%	11.6%	13.1%	15.4%	17.5%	11.9%
RRCAGGTG	5.7%	23.4%	11.5%	12.8%	7.7%	7.9%	9.5%
CACGTG	5.1%	18.1%	5.6%	6.1%	53.8%	55.9%	5.2%
AP-1	70.5%	14.7%	7.3%	7.5%	24.0%	20.1%	9.0%
Mef2 (Tap.)	24.6%	22.5%	21.4%	20.7%	16.6%	15.8%	33.0%
Mef2 (lit)	1.1%	3.8%	0.8%	0.7%	1.2%	1.1%	1.3%
Klf4	10.1%	32.4%	14.3%	14.7%	26.9%	29.5%	10.4%
CTCF	0.6%	7.0%	47.0%	50.2%	6.1%	7.0%	1.5%
Usf1	0.8%	9.3%	0.8%	0.9%	30.3%	32.2%	0.7%
CGCGCG	0.2%	14.2%	2.6%	2.3%	9.2%	10.8%	0.7%
TATA-box	21.5%	12.1%	17.5%	15.7%	13.2%	12.4%	32.8%
Sp-1	5.2%	37.0%	12.8%	13.3%	28.3%	32.4%	6.4%

Table 3.8b. Motif coverage of ChIPSeq defined regions of occupancy for non-MRF transcription factors. Coverage is defined as the percentage of the total number of regions that contain at least one copy of the motif. Randomized regions are included as a representation of expected coverage - the set consists of approximately 101 thousand randomly selected regions, each 501 nucleotides long (matching the standardized length of ChIP regions). They were filtered to eliminate repeats and unsequenced nucleotides, as well as telomeric/centromeric sequence. This table list coverages for all occupied regions obtained from the dataset denoted at the top of the column. **CTCF cyc** and **Usf1 cyc** = occupancy in cycling myoblasts. **CTCF myoc** = CTCF-occupied regions in myocytes. **Usf1 60h** = Usf1-occupied regions 60 hours after withdrawal of serum.

Motif	early	late	cont.	cont. ex.	MyoD ex.	Myog ex.	Randomized
CAGSTG	59.8%	89.9%	88.5%	79.3%	76.2%	78.1%	42.6%
CAGCTG	38.8%	80.6%	75.1%	58.8%	50.9%	67.7%	21.8%
CAGGTG	31.7%	55.4%	41.9%	39.3%	46.6%	35.7%	28.0%
RRCAGSTG	40.9%	84.8%	80.1%	66.6%	63.3%	62.6%	20.0%
RRCAGCTG	27.4%	72.8%	66.3%	48.6%	40.4%	54.6%	11.9%
RRCAGGTG	16.7%	38.4%	26.9%	24.8%	31.7%	18.1%	9.5%
CACGTG	5.2%	7.0%	7.4%	4.5%	7.0%	7.9%	5.2%
AP-1	52.3%	5.0%	28.7%	27.3%	16.7%	12.4%	9.0%
Mef2 (half)	22.2%	19.1%	19.0%	23.3%	20.8%	15.8%	33.0%
Mef2 (lit)	0.9%	1.6%	1.1%	0.5%	1.3%	1.6%	1.3%
Mef2 (hybrid)	2.0%	1.7%	1.9%	1.4%	1.6%	2.0%	2.1%
Klf4	10.3%	19.2%	21.7%	13.3%	18.1%	29.4%	10.4%
CTCF	0.4%	1.9%	2.7%	1.4%	4.8%	6.3%	1.5%
Usf1	0.7%	0.9%	1.2%	0.6%	1.1%	1.3%	0.7%
CGCGCG	0.1%	0.6%	1.0%	0.2%	4.8%	6.4%	0.7%
TATA-box	20.2%	13.1%	12.9%	19.0%	15.8%	9.2%	32.8%
Sp-1	3.6%	11.5%	13.4%	6.3%	15.6%	26.0%	6.4%

Table 3.8c. Coverage of differential MyoD regions by select motifs. Mef2 (lit) corresponds to the canonical Mef2 motif CTAWWWWTAG. **Early** = sites occupied by MyoD in cycling myoblasts but not in differentiating myocytes; **Late** = opposite of early; **Cont.** (continuous) = sites occupied by MyoD in both cycling and differentiating C2C12s; **Cont. ex.** (continuous exclusive) = same as continuous, but without corresponding myogenin occupancy; **MyoD ex.** = regions occupied by MyoD in differentiating myocytes that do not have corresponding myogenin occupancy; **Myog ex.** = same as MyoD ex., but for myogenin; **Randomized** = set of ~101,000 randomly generated regions.

Motif	Mef2	1	2	Myog
CAGCTG	2.14	0.42	2.70	2.55
CACCTG	0.84	-0.07	1.23	1.08
CAGSTG	1.52	0.16	2.02	1.86
RRCAGSTG	2.37	0.25	2.97	2.78
RRCAGCTG	2.68	0.25	3.31	3.16
RRCAGGTG	1.76	0.24	2.28	1.97
CACGTG	2.22	2.81	1.62	1.02
CATATG	-1.37	-1.57	-1.25	-1.34
CAAATG	-0.82	-1.42	-0.53	-0.67
CAATTG	-1.10	-1.15	-1.07	-1.16
CAGATG	0.01	-0.60	0.31	0.26
CAAGTG	-0.14	-0.64	0.12	0.11
CACATG	-0.22	-0.82	0.08	-0.15
CAACTG	0.13	-0.49	0.44	0.29
Meis	0.53	-0.10	0.84	0.77
AP-1	0.93	0.15	1.29	0.97
Mef2 (half)	-0.72	-0.93	-0.59	-1.20
Mef2 (hybrid)	3.76	4.81	1.86	0.19
Runx	0.72	0.03	1.04	1.06
RP58	-0.42	-1.06	-0.11	-0.22
CTCF	2.70	3.15	2.28	2.42
Mef2 (lit)	1.57	1.24	1.75	0.22
Klf4	2.07	2.29	1.89	1.89
Usf1	4.33	5.17	3.19	1.56
CGCGCG	4.49	5.18	3.68	3.37
TATA-box	-2.17	-2.37	-2.06	-2.36
Sp-1	3.58	4.24	2.83	2.80
Dec1 - KCACGTGM	3.12	3.95	2.04	1.03

Figure 3.9a. Sequence content analysis of Mef2-occupied regions. **1** = regions occupied by Mef2 that do not have corresponding myogenin occupancy, **2** = regions occupied jointly by Mef2 and myogenin, **Mef2** = all Mef2 regions, **Myog** = all myogenin regions.

	-250					0				250
Mef2	1.74	1.81	2.00	3.53	5.44	5.39	3.74	2.42	2.06	1.66
Mef2-Myog	2.90	2.63	2.73	4.66	6.51	6.47	4.88	3.43	2.81	2.43
Mef2+Myog	-0.84	0.75	1.16	1.16	3.45	3.41	1.33	0.75	1.16	0.75
Myog	0.15	-0.10	0.19	0.58	-0.27	-0.15	0.19	0.24	0.70	0.10

Figure 3.9b. Positional distribution of Mef2-hybrid site

	-250					0				250
Mef2	0.55	0.55	0.55	2.14	2.49	3.18	1.87	0.29	-0.03	-2.03
Mef2-Myog	0.86	0.86	0.28	1.60	1.60	2.74	1.60	0.86	-0.72	-0.72
Mef2+Myog	0.30	0.30	0.71	2.41	2.88	3.41	2.03	-0.29	0.30	-6.64
Myog	-1.30	0.22	0.28	0.74	0.40	0.70	0.65	0.56	-0.30	-1.49

Figure 3.9c. Positional distribution of Mef2 canonical site CTAWWWWTAG

	-250					0				250
Mef2	0.88	1.10	1.64	2.37	3.77	3.84	2.36	1.37	0.83	0.76
Mef2-Myog	0.37	-0.08	0.19	0.19	0.45	0.71	0.14	0.19	0.28	-0.13
Mef2+Myog	1.14	1.56	2.16	2.98	4.45	4.52	2.98	1.83	1.11	1.16
Myog	0.57	0.88	1.27	2.24	4.50	4.56	2.19	1.26	0.92	0.61

Figure 3.9d. Positional distribution of RRCAGSTG

	-250					0				250
Mef2	1.79	1.79	2.44	3.94	6.37	6.07	3.07	1.79	2.15	2.15
Mef2-Myog	2.29	2.29	1.88	4.88	7.30	7.00	3.75	2.61	2.29	1.88
Mef2+Myog	1.31	1.31	2.72	2.53	5.01	4.68	2.31	0.72	2.05	2.31
Myog	0.71	1.17	1.41	1.76	1.92	2.62	1.76	1.30	0.41	1.17

Figure 3.9e. Positional distribution of the Usf1 motif CACGTGAC

Motif	Mef2	1	2	Myog	Randomized
Mef2 (Tapscott)	22.5%	18.4%	25.2%	16.9%	33.0%
Mef2 (lit)	3.8%	3.2%	4.3%	1.5%	1.3%
Mef2 (hybrid)	17.1%	34.5%	5.4%	2.1%	2.1%
RRCAGSTG	52.5%	18.4%	75.6%	76.0%	20.0%
RRCAGCTG	42.4%	10.8%	63.7%	63.9%	11.9%
RRCAGGTG	23.4%	8.9%	33.2%	27.6%	9.5%
CACGTG	18.1%	25.5%	13.1%	9.2%	5.2%
Usf1	9.3%	16.3%	4.5%	1.7%	0.7%

Table 3.10. Coverage of Mef2 data by selected motifs. **1** = Mef2-Myog regions, **2** = Mef2+Myog regions, **Mef2** = all Mef2 regions, **Myog** = all myogenin regions.

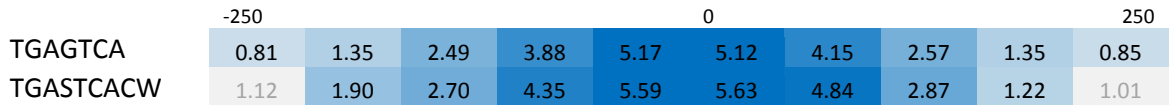


Figure 3.11a. Positional distribution of two alternative AP-1/Fosl1 recognition motifs.

Motif	Fosl1	Randomized		# of occ.
TGAGTCA	70.54%	8.97%		2457
TGASTCACW	35.47%	2.84%		1096

Table 3.11b. Coverage and absolute incidence of two potential AP-1/Fosl1 recognition motifs. Motif TGAGTCA was cited as a consensus binding cited by Cao 2010 and others. Motif TGASTCACW was derived via MEME from regions preferentially occupied by MyoD in cycling myoblasts, but not differentiating myocytes. **# of occ.** = number of instances of the motif present in the Fosl1 HC set of occupancy signals. **Fosl1** = coverage of Fosl1 HC regions by the motif. **Randomized** = coverage of the randomized set of regions by the motif.

Motif	MyoD cyc	Fosl1	Fosl1-MyoD	MyoD-Fosl1	Fosl1+MyoD	Random
CAGCTG	2.34	0.32	-0.96	2.46	1.48	0.29
CACCTG	0.72	-0.52	-1.27	0.78	0.34	0.23
CAGSTG	1.61	-0.11	-1.13	1.71	0.92	0.26
RRCAGSTG	2.54	0.01	-1.84	2.68	1.39	0.27
RRCAGCTG	2.95	0.30	-1.77	3.09	1.80	0.29
RRCAGGTG	1.67	-0.55	-1.94	1.80	0.50	0.24
CACGTG	0.51	0.14	-0.03	0.52	-0.10	0.15
CATATG	-0.80	-0.57	-0.56	-0.80	-0.44	-0.12
CAAATG	-0.42	-0.26	-0.28	-0.44	-0.24	0.05
CAATTG	-0.69	-0.72	-0.74	-0.67	-0.73	0.01
CAGATG	0.42	0.20	-0.15	0.37	0.68	0.18
CAAGTG	0.01	-0.13	-0.28	0.00	0.01	0.11
CACATG	-0.16	-0.25	-0.40	-0.18	0.09	-0.02
CAACTG	0.20	-0.14	-0.41	0.20	0.15	0.12
Meis	0.65	-0.07	-0.30	0.71	0.32	0.20
AP-1	2.34	3.64	3.64	1.84	3.59	0.11
Mef2 (half)	-0.91	-0.63	-0.52	-0.92	-0.75	-0.07
Mef2 (hybrid)	0.03	0.00	-0.03	0.04	0.45	0.13
Runx	1.02	0.59	0.29	1.02	0.84	0.21
RP58	0.40	0.20	-0.19	0.35	0.59	0.13
CTCF	0.78	-0.94	-0.79	0.96	-1.42	0.46
Mef2 (lit)	-0.36	-0.21	-0.96	-0.32	-0.27	-0.01
Klf4	0.99	0.07	-0.24	1.11	0.27	0.20
CGCGCG	0.04	-1.54	-1.16	0.34	-1.21	-0.04
TATA-box	-1.71	-1.13	-0.97	-1.72	-1.56	-0.26
Sp-1	1.21	-0.18	-0.26	1.37	-0.35	0.42
Dec1 - KCACGTGM	0.29	0.13	-0.41	0.34	-0.19	0.13

Figure 3.12. Motif content analysis of Fosl1-occupied regions, and for regions of joint and separate occupancy by Fosl1 and MyoD in cycling C2C12s.

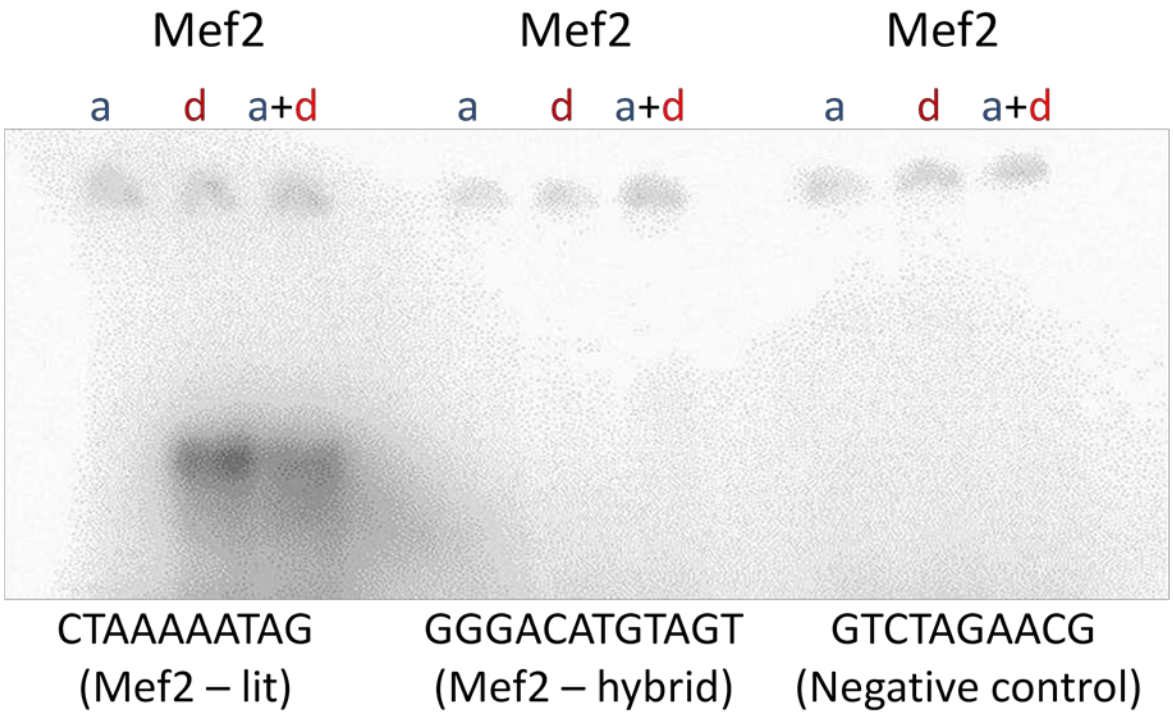


Figure 3.13. Binding affinity assay between Mef2a, Mef2d and the hybrid site derived from motif analysis of Mef2a occupancy. Mef2a:Mef2a homodimers do not bind either the canonical or the hybrid motifs. Mef2a:Mef2d heterodimers and Mef2d:Mef2d homodimers are able to bind the canonical but not the hybrid motifs. (Gel shift #64)

Motif	Usf1	Usf1-Myog	Mef2-Myog	Mef2	Myog
CAGCTG	0.79	0.11	0.42	2.14	2.55
CACCTG	-0.11	-0.32	-0.07	0.84	1.08
CAGSTG	0.33	-0.12	0.16	1.52	1.86
RRCAGSTG	0.58	-0.24	0.25	2.37	2.78
RRCAGCTG	0.91	-0.10	0.25	2.68	3.16
RRCAGGTG	-0.07	-0.45	0.24	1.76	1.97
CACGTG	3.74	3.57	2.81	2.22	1.02
CATATG	-1.61	-1.51	-1.57	-1.37	-1.34
CAAATG	-1.02	-0.95	-1.42	-0.82	-0.67
CAATTG	-0.90	-0.85	-1.15	-1.10	-1.16
CAGATG	-0.35	-0.43	-0.60	0.01	0.26
CAAGTG	-0.39	-0.40	-0.64	-0.14	0.11
CACATG	0.30	0.33	-0.82	-0.22	-0.15
CAACTG	-0.43	-0.53	-0.49	0.13	0.29
Meis	0.08	-0.07	-0.10	0.53	0.77
AP-1	1.44	1.49	0.15	0.93	0.97
Mef2 (half)	-1.28	-1.24	-0.93	-0.72	-1.20
Mef2 (hybrid)	1.35	1.42	4.81	3.76	0.19
Runx	0.28	0.09	0.03	0.72	1.06
RP58	-0.54	-0.49	-1.06	-0.42	-0.22
CTCF	2.63	2.35	3.15	2.70	2.42
Mef2 (lit)	-0.32	-0.68	1.24	1.57	0.22
Klf4	1.97	1.56	2.29	2.07	1.89
Usf1	5.98	5.77	5.17	4.33	1.56
CGCGCG	4.02	3.49	5.18	4.49	3.37
TATA-box	-2.03	-1.86	-2.37	-2.17	-2.36
Sp-1	3.44	3.05	4.24	3.58	2.80
Dec1	4.55	4.32	3.95	3.12	1.03

Figure 3.14. Sequence content analysis of the Usf1 occupancy map in differentiating myocytes (60h after withdrawal of serum). **Usf1-Myog** = regions occupied by Usf1 but not by myogenin in differentiating myocytes (60 hr); **Mef2-Myog** = regions occupied by Mef2 but not by myogenin in differentiating myocytes (60 hr); **Usf1, Mef2, Myog** = all HC regions of occupancy for Usf1, Mef2 and myogenin, respectively, in differentiating myocytes (60 hr). Enrichment is given as $\log_2$ of (observed density) / (expected density), values that are not significantly different from background ($p > 0.01$) are displayed in grey, negative values represent depletion of the associated motif.

Expression Category	# Open connections	# Blocked connections	% Total connections that are open
Muscle down	502	433	**53.7%**
Muscle up	2401	898	**72.8%**
Flat expressed	3740	1765	**67.9%**
Unexpressed	4964	2286	**68.5%**

Table 3.15. CTCF occupancy is more likely to occur between myogenin regions and TSSes of down-regulated genes in differentiating myocytes. A blocked connection is defined as one of the form myogenin -> CTCF -> TSS for forward-oriented reading frames or TSS <- CTCF <- myogenin for reverse-oriented reading frames. An open connection is defined as one where the no CTCF occupancy event exists between a myogenin region and its nearest gene. There is no statistical difference in the likelihood of connections being open for flat expressed and unexpressed genes. Muscle-up genes have the highest probability of having open connections ($p < 0.01$), muscle-down genes have the lowest probability of having open connections ($p < 0.0001$).

Chapter 4: Competitive transcriptional regulation and its role in defining lineage-specific cis-regulatory activity.

4.1 Introduction: bHLH diversity and the problem of lineage specificity

The family of CANNTG-binding bHLH transcription factors is large, with over 600 proteins catalogued in a recent phylogenetic survey (Stevens et al. 2008), and its members perform a wide range of functions. Leucine-zipper bHLHs, named for the presence of a leucine-rich protein-protein interaction domain, serve a number of roles in cycling cells, with myc garnering significant attention due to its oncogenic properties (Bretones et al. 2014). Usf1, also a bHLH-lz, is a "housekeeping" transcription factor active in both cycling and terminally differentiated cells - it maintains expression of genes that are necessary for cell function but not differentially transcribed in a lineage-specific manner (discussed in section 3.6). Members of the bHLH-orange family, such as Dec1 (also known Stra13 and bHLHb2) and Dec2 (also known as Sharp1) are transcriptional repressors (Shen et al. 2002; Honma et al. 2002), and are up-regulated in multiple cell types during terminal differentiation. They likely serve in an anti-proliferative capacity in a broad range of lineages, and together with bHLH-lzs form the basis for the model I will discuss in section 4.5. Other bHLH proteins are crucial to cell fate determination - MRFs (MyoD, Myf5, Mrf4, myogenin) control skeletal myogenesis; NeuroD and neurogenin regulate neuronal development; E2A gene products - the two splice isoforms E12 and E47 have different bHLH domains - are important for B and T cell differentiation as homodimers, in addition to being crucial heterodimerization partners for many tissue-specific bHLHs; Tal1 (also known as Scl) directs terminal erythroid cell formation. Curiously, all of the lineage-directing factors just mentioned recognize and bind the same hexamer motif CAGCTG (Cao et al. 2010; Fong et al. 2012; Lin et al. 2010; Kassouf et al. 2010), and the E2A proteins are expressed alongside them. How, then, is lineage specificity maintained in the presence of E2A, which

recognizes the same CAGSTG sites as the MRFs it heterodimerizes with, and can contribute critically towards inducing its own differentiation program?

The answer, as it does for many things in mammalian biology, rests in a combination of mechanism. The analysis and data in this chapter point to several ways in which superficially similar transcription factors establish and maintain lineage specificity. First, and perhaps simplest, binding affinities for specific subsets of likely recognition motifs can help us to understand the targeting of bHLH dimers, and in section 4.2 I examine *in vitro* affinities of MRFs and E proteins for variations of the e-box sequence at both central and flanking nucleotides. The resulting sequence preferences are not sufficient to fully account for lineage specificity once I compare the occupancy map for E47 in differentiating B-cells (Lin et al. 2010) with my occupancy maps of MyoD and myogenin in differentiating myocytes (section 4.3). Many sites are occupied in both systems, and binding affinity alone neither accounts for exclusion of E:E homodimers from occupying CAGSTG motifs accessible in differentiating muscle, nor precludes MRFs from occupying some of the E:E specific sites. Instead, the comparison of underlying sequence structure points to a cohort of co-expressed transcription factors that are all able to bind identical or very similar recognition sites, suggesting a competitive regulatory system. I capture these interactions in a model that helps further account for lineage-specific occupancy, although it too does not fully explain it (section 4.4). To address the likely prevalence and physiological importance of competitive systems, I introduce a second set of factors that, based on sequence analysis and current knowledge, are likely to behave in a similar fashion, with multiple co-expressed TFs competing for the same sets of target DNA elements. Finally, I identify some specific examples in the genome worthy of functional testing to verify parts of the model (section 4.5). This analysis and model are discussed in the context of transcriptional regulation, which is only one of several known ways of achieving lineage specificity. Others include epigenetic silencing via chromatin compaction and DNA methylation, post-transcriptional regulation via non-coding RNAs, or post-translational factor modifications, to name a few. It is the totality of these

inputs that drives developmental pathways. In the scope of direct transcriptional regulation, competitive factor binding plays an important role in both cell fate determination and in the execution of the differentiation program.

4.2 Binding affinities of myogenin and E47.

Sequence content analysis of myogenin-occupied regions showed that 75% of CAGSTG motifs contained therein had an RR prefix - a much higher than expected rate ($p < 0.01$). While this enrichment was substantial and significant, it does not by itself preclude other prefixes from forming viable binding sites, although they are likely to be of lower affinity. Additionally, when the RRCAGSTG motif components were analyzed separately based on central nucleotides, RRCA**GC**TG was observed ~3 times more frequently than RRCA**GG**TG in regions of myogenin occupancy. Interestingly, the GC:GG ratio was lower in the subset of MyoD regions that were continuously occupied in both myoblasts and myocytes, and did not have an overlapping myogenin signal, although it remained > 1 in every subset of MyoD regions. Fong et al. (2012) reported that MyoD has a high affinity for the CAGGTG motif. One hypothesis was that MyoD:E47 favors CAGGTG, while Myogenin:E47 favors CAGCTG. An alternative hypothesis is that E47:E47 homodimers might have higher affinity for CAGCTG sequences due to their palindromic nature and the symmetric nature of homodimer. To quantify the influence of central and flanking nucleotides on binding affinity of bHLH dimers, a series of binding assays were performed using ^{32}P-labeled double-stranded oligonucleotide probes and transcription factors synthesized *in vitro* using a coupled rabbit reticulocyte lysate expression system (see Methods).

E47:E47 homodimers bind both RRCAGCTG and RRCAGGTG, with a slight preference for CAGGTG, and showed no detectable affinity for the scrambled e-box sequence used as a negative control. Unexpectedly, they also bound the myc-class e-boxes RRCACGTG (Figure 4.1). Further examination using a panel of ^{32}P labeled oligonucleotides standardized for specific activity revealed a preference for CAGGTG over CAGCTG, regardless of prefix, although both showed detectable binding. Some interaction was observed with every probe except CTCAGCTG (Figure 4.2). For E47:E47 homodimers binding affinity is influenced primarily by central nucleotides, with the prefix playing a lesser role.

Additionally, E47:E47 binds the myc-class e-box CACGTG *in vitro*, expanding its consensus to CASSTG. No interaction was observed between E47:E47 homodimers and CATATG e-boxes associated with Twist, nor any e-box of the type CAWWTG (data not shown - Gel shift #37). Additionally, E47:E47 has a higher affinity for GG centered e-boxes rather than the GC-centered ones. While this is counter-intuitive given the initial assumption about symmetry of homodimers and palindromic nature of CAGCTG, the same result was observed by Blackwell and Weintraub (1990) in their SELEX/SAAB experiments to determine MyoD and E2A binding preferences. Their result for E47 remained somewhat speculative due to technical issues with the experiment, but my data confirm it. A later effort to address the targeting and binding preferences of MyoD:E and Twist:E heterodimers (Kophengnavong et al. 2000) used only E12 as the heterodimerization partner, and since E12 does not homodimerize (Shirakata and Paterson 1995), no data for E:E homodimers was generated.

Myogenin and MyoD were unable to bind MRF-class e-box sequences in the absence of E47 (Figure 4.3). Myog:E47 heterodimers efficiently interact with GACAGCTG and GACAGGTG, and, as expected, show very little affinity for GACACGTG myc-class e-boxes (Figure 4.1). Lane 7 shows an E47:E47 complex interacting with GACA**CG**TG, accompanied by an almost complete lack of a heterodimeric band. An expanded competition panel was used to assess the effects of prefix and central nucleotides on E47:Myogenin affinity, which, in agreement with the *in vivo* site derivation, is highest for RRCAGSTG motifs, dropping off noticeably for RYCAGSTG and YRCAGSTG motifs, and similar to that of negative control for YYCAGSTG motifs (Figure 4.4). These data are consistent with sequence content analysis, which confirms the strong preference for an RR prefix. They do not, however, explain the heavy numeric bias towards RRCAGCTG motifs observed in myogenin-occupied regions *in vivo*, as the *in vitro* affinity indicates equal preference for GC and GG centered e-boxes. Both Myog:E12 and MyoD:E12 heterodimers can also bind MRF-class e-boxes, but with lower efficiency than corresponding

heterodimers involving E47 (Figure 4.5). MyoD:E47 heterodimers show a slight preference for GACA**GG**TG over GACA**GC**TG (Figure 4.5), which is also consistent with expectation.

The binding experiments show that simple site preference can explain some of the biased found through motif content analysis of MRF-occupied regions. Myog:E47 and MyoD:E47 preferentially bind GS-centered e-boxes with an RR prefix. MyoD:E47 shows a slight preference towards the GG-centered motifs compared to GC-centered ones, while myogenin:E47 binds both equally. This can be seen in two subsets of MyoD-occupied regions - those present only in cycling myoblasts ("early" MyoD) and those present in both myoblasts and myocytes, but without an overlapping myogenin occupancy event ("continuous exclusive" regions) - both sets show a higher fraction of occupied RRCA**GG**TG motifs, although in both cases the ratio of GC:GG is still > 1. Both of these groups of MyoD regions lack overlapping myogenin occupancy, and thus serve to illustrate the *in vivo* sequence bias of MyoD. E47:E47 homodimers showed three primary differences from the MRF:E hybrid species. First, they do not have the RR prefix bias, accepting a much broader range of flanking nucleotides. Second, they have a higher affinity for CAGGTG e-boxes compared to CAGCTG ones. Finally, E47:E47 homodimers are able to bind to CACGTG e-boxes with comparable efficiency to that of CAGCTG. In aggregate, these results do not explain the heavy bias towards GC-centered motifs observed in myogenin-occupied regions, but they do provide baseline expectations for an occupancy map comparison of Myog:E47 in differentiating muscle versus E47:E47 in differentiating B-cells.

4.3 Comparison of bHLH occupancy in differentiating skeletal muscle and B-cells.

In addition to being a crucial partner for MyoD and myogenin's ability to bind DNA, the transcription factor E2A plays a key role in B-cell differentiation (Murre et al. 1991; Lin et al. 2010). To better understand mechanisms used to enforce system-specific occupancy, data for E47:E47 homodimers in differentiating B-cells (Lin et al. 2010) were compared to MRF:E47 data in differentiating myocytes (C2C12). To maintain consistency, raw reads were mapped to the mm9 genome assembly, then used as ERANGE input to identify 8103 HC regions occupied by E47. They were compared to the 14786 HC regions occupied by myogenin in differentiating C2C12s 60h after withdrawal of serum, and to the 6641 HC regions occupied by MyoD in cycling C2C12s. MyoD 24 and 60 hr sets were also used in the comparison, but due to their high overlap discussed in section 3.2, their contribution to identifying system-specific occupancy proved minimal.

The analysis focused on three sets of regions - those occupied in B-cells but not in muscle (4726), those occupied in muscle but not in B-cells (11212), and those occupied in both states (1842) (Figure 4.6a). I should note that this analysis is likely to be more sensitive towards regions occupied in a B-cell specific manner, due to the almost twofold excess in the number of myogenin regions compared to the number of E47 regions. It is difficult to say precisely how many sites "should" be occupied in a given setting, which makes attempts at normalization not well justified. This, in turn, means that the stronger (or, more accurately, the larger) of the two determinations will contain a set of regions that, from a differential occupancy point of view, could all be false positives. This problem is inherent to comparative ChIPSeq analysis, and is the limitation of current data. The three sets of regions were analyzed for sequence content using the same combination of motif mapping and *de novo* searches as before.

Regions specific to B-cells exhibited noticeable differences in their sequence content. The relative enrichment of RRCAGCTG and RRCAGGTG motifs compared to genomic background was virtually identical (Figure 4.6b), as opposed to the substantial preference for RRCAGCTG observed in muscle-specific regions. Furthermore, consistent with the *in vitro* results reported in 4.2, there was no bias in favor of RR-prefix bias, nor was there any bias against it (Figure 4.6c and d) - the enrichment of the octamers is solely the result of increased density of the core hexamers. For muscle-specific occupancy by Myog:E, the non-RR versions of the CAGGTG motif occur at background levels, whereas the non-RR variants of CAGCTG were enriched regardless of prefix (as expected, those preceded by RR were the most heavily enriched) (Figure 4.6c). This suggests that the RR prefix is of special importance to CAGGTG motifs in muscle-specific regions. Part of the reason for this bias stems from the fact that E47:E47 has a higher affinity for CAGGTG than for CAGCTG, thus placing a higher premium on the use of the RR prefix to achieve muscle-specific occupancy. Another likely reason for this bias is Zeb1, which will be discussed below.

The motif for zinc finger repressor RP58 - CAGATGT (see Introduction) is significantly enriched in B-cell-specific regions ($p < 0.01$) compared to genomic background, but depleted in their muscle counterparts ($p < 0.01$) (Figure 4.6B). The canonical Mef2 motif and CGCGCG exhibited the opposite behavior (Figure 4.6b), being significantly enriched in muscle-specific regions and depleted in B-cell-specific regions.

The set of regions occupied in both B-cells and muscle had a motif content profile that is very similar to that of the muscle-specific regions, with the same preference for RRCAGSTG and a high ratio of GC:GG cores (central e-box nucleotides). Their content for AP-1, Mef2 and CGCGCG was essentially the same as that of muscle-specific regions. Taken at face value, this is consistent with the idea that modules shared between muscle and B-cells exhibit muscle-like characteristics indicative of MRF:E occupancy, in part due to the RR prefix requirement associated with the latter.

A blueprint emerges for the separation between muscle and B-cell networks. On the sequence level, MRF-specific sites are distinguished by the presence of an RR prefix, and more loosely by the requirement that at least one of the two flanking nucleotides be an A or a G - Myog:E and MyoD:E dimers show little to no affinity for motifs of the form YYCAGSTG. Conversely, E:E homodimers bind any sequence of the form CASSTG, although the CACGTG component is unique due to the presence of a CpG dinucleotide that serves as a methylation target. CACGTG sequences occur less frequently in the mouse genome when compared to either CAGCTG or CAGGTG, with only ~250,000 instances, compared to ~1 million for CACGTG and ~1.7 million for CAGGTG, before accounting for simple repeats (repeat masking reduces these numbers to 149K, 638K and 909K, respectively). In addition to being a methylation target, CACGTG and its derivates also serve as binding sites for myc-family bHLHs, and consequently play an important role in cell cycle regulation. In the B-cell-specific regions occupied by E47, CACGTG motifs occur twice as often as expected ($p < 0.01$), but their coverage is low (9.3%) compared to that of CAGSTG (89.4%), suggesting that E47:E47 occupancy is primarily achieved at the "myogenic" motifs CAGCTG and CAGGTG. While CAGGTG is preferred *in vitro*, *in vivo* CAGCTG and CAGGTG are enriched at virtually the same level (Figure 4.6b) and both cover a majority of regions - 65.1% for CAGCTG, 71.8% for CAGGTG. This behavior that suggests that both can, and do, serve as viable occupancy centers for E47 homodimers.

Sequence content alone does not fully capture lineage specificity. Of special interest are motifs of the form RRCAGGTG, as both E47:E47 and MRF:E47 show a high affinity for them *in vitro*, and both MRFs and E2A proteins (E12/E47) are present in cycling and differentiating skeletal muscle cells. A similar relationship was described between MyoD:E and NeuroD2:E (Fong et al. 2012), where both MyoD and NeuroD bind the motif RRCAGCTG, but NeuroD additionally binds RRACGATG, whereas MyoD additionally binds RRCAGGTG. This led to dubbing GC-e-boxes "public" and GA/GG-e-boxes "private" in relation to the transcription factors studied (Figure 4.12). However, in a physiological setting MyoD and

NeuroD are not expressed at the same time, and E:E homodimers show little *in vitro* affinity for CAGATG. In differentiating skeletal muscle, however, both MRFs and E are present, and both have a high affinity for RRCAGGTG. Some additional mechanisms are employed to enforce lineage specificity. One is chromatin silencing, which can be assessed by measuring the overlap between B-cell specific regions and repressive chromatin marks (such as H3K27me3) in differentiating muscle cells. Another rests on relative availability of bHLH proteins present and their dimerization affinity. Based on RNASeq data (Table 3.7), the myogenin message is present in almost thirtyfold excess compared to the E2A message in differentiating myocytes (at the 60 hr timepoint), although this ratio drops to ~10:1 by day 7. In cycling myoblasts Id1 and Id3 combined also exceed E2A by over tenfold at the RNA level. It is likely that part of the reason for such high transcript abundance is to prevent formation of E:E homodimers through saturation - an effect primarily accomplished by Ids in cycling progenitors and by myogenin in terminally differentiating myocytes. While the latter mechanism deserves further investigation, it is made all the more likely by the *in vitro* binding data, where a threefold molar excess of myogenin greatly diminished the band associated with E:E homodimers (Figure 4.4). A further method for maintaining appropriate lineage-specific occupancy is proposed in sections 4.4 and 4.5.

4.4 Model: RP58 and Zeb1 as attenuators in muscle CRMs.

Recent work in astrocytes established that RP58 directly interacts with elements regulating the expression of Id1-4 (Hirai et al. 2012). While the canonical binding site for RP58 is CAGATGT, it tolerates single nucleotide mismatches in certain positions, including the central one, and I noticed that several of the elements shown to interact with RP58 by Hirai et al. were of the form CAGCTGT. This is supported by preliminary *in vitro* binding data, in which I observed that RP58 formed a band when presented with a ^{32}P-labeled GACAGCTGTC oligonucleotide. Because CAGCTG is palindromic, all ACAGCTG sequences must be accompanied by CAGCTGT on the opposite strand, making fully half of all RRCAGSTG motifs recognizable by both MRF:E47 and RP58. The relationship is less straightforward for CAGGTG, since its not symmetric. It is likely the this sequence is also acceptable for RP58, although further testing needs to be done. If true, it would make all RRCAGGTGT also recognizable by MRF:E and RP58. Furthermore, the primary RP58 motif - CAGATGT - is enriched in B-cell-specific occupancy regions ($p < 0.01$) and depleted in muscle-specific occupancy regions ($p < 0.01$) (Figure 4.6b). Jointly, this suggests two functions for RP58 in myogenesis. One is the direct repression of regions used in non-myogenic networks, the B-cell one amongst them. While this is most likely accomplished through the optimal site CAGATGT, some qualifying CAGSTG sites could also be involved. This repression most likely occurs in a "classic" manner - repressor binding to a recognition motif facilitates recruitment of co-repressors and/or histone deacetylases, leading to silencing of the target element. The other is an attenuating effect exerted at accessible CAGSTG motifs. Given the relative physiological concentrations of MRFs and RP58 present in differentiating myocytes, it is unlikely that RP58 can prevent occupancy of optimal MRF octamers (RRCAGSTG). However, it should be able to delay occupancy and exert a titrating effect on sites recognized by both it and MRF:E heterodimers. I expect the titration effect to be more pronounced at sites that are suboptimal for MRF:E, with the strongest attenuation achieved at sequences of the form YYCAGSTGT. In fact, it is highly plausible that such sequences serve as repressive rather than

activating sites in a myogenic context, which is in sharp contrast to the "expected" role of a CAGGTG motif. At the moment, both pathways represent working hypotheses.

Consistent with the hypothesis, albeit not decisive, is the fact that 23% of B-cell specific E47 elements are covered by the RP58 motif (CAGATGT), although how often RP58 actually participates in the repression of those elements in skeletal muscle is unclear. A direct measurement of RP58 occupancy by ChIPSeq would be highly desirable, but requires the availability of a strong ChIP-competent antibody, which is currently lacking. Additionally, it is unclear how quickly RP58-meidated repression takes place - cases of efficient chromatin silencing followed by decoupling may prove challenging to capture via traditional occupancy measurements. There is also evidence that a large subset of MRF motifs RRCAGSTG can be bound by RP58, leading to the attenuation hypothesis - though some of them could be completely repressed instead (which one can consider as the most severe form of attenuation). A knockout of RP58 leads to deficiencies in hind limb myofiber development and immediate postnatal death due to absence of diaphragm function (Okado et al. 2009). The underlying molecular basis for disruption in muscle development caused by the knockout merits a detailed investigation, with focus on possible de-repression and the set of genes that would be affected by it. The impact of RP58 deficiency on genes expressed in B-cells could also be tested in a knockout.

There is at least one other characterized zinc finger repressor that likely carries out an attenuating function in differentiating skeletal muscle, and could be involved in preventing MRF:E occupancy at B-cell-specific sites. Zeb1 binds CAGGTG e-boxes, and in C2C12 shows a similar expression pattern to RP58, although its transcript is present at a somewhat higher level (Table 3.7). Recent results show that in the absence of Zeb1, expression of muscle-related genes proceeds at a greatly accelerated pace (Siles et al. 2013). Zeb1 exerts an additional layer of control over the RRCAGGTG motif and its derivatives, for which both E:E and MRF:E have strong affinity. Taken together, RP58 and Zeb1 help explain some of the

sequence content results described earlier. In muscle-specific regions occupied by MRFs, there is a strong preference for the RR prefix - based on motif density analysis, I expect that virtually all functional motifs based around CAGGTG are of the RRCAGGTG form. This takes advantage of the extra specificity of the MRF:E heterodimers, which require the prefix to bind efficiently, but are present in overwhelming abundance. Motifs lacking the RR prefix are good targets for E:E homodimers, but the latter are unlikely to be present in significant physiological concentrations, due to large excess of Ids and/or MRFs (depending on the cell state). Furthermore, such motifs are optimal targets for Zeb1, and any CAGGTG motif with a T suffix is also a target for RP58. This mechanism, if true, helps account for the tight control over the availability of CAGGTG motifs on a sequence level. The most "myogenic" motif - RRCAGGTG - is also the most tightly regulated one. The expression patterns of both Zeb1 and RP58 (Table 3.7) also support them functioning in primarily attenuating roles during skeletal muscle differentiation, as both are up-regulated upon cell cycle exit, and neither transcript is particularly abundant in cycling myoblasts. Nevertheless, this does not preclude the possibility that both are important to myogenic fate determination through repression/attenuation of a few key sites early in the specification process. Interestingly, the RP58 promoter is strongly occupied by myogenin at 60 hr after differentiation. The associated occupancy region contains two RRCAGSTG motifs, both highly conserved, and both viable RP58 targets. While not conclusive, this raises the possibility of a feedback loop involving MRFs and RP58 that ultimately controls the expression levels of RP58. And could the same mechanism be used in other tissue types where RP58 is expressed and CAGSTG-binding bHLH factors abound? A further interesting problem to address would be to quantify the effects of over-expression of RP58 and Zeb1 in cycling myoblasts or in early progenitors prior to the expression of MyoD.

The attenuation model is an example of "competitive regulation" - where an activating and a repressing factors both recognize the same DNA sequence, and are both present in the system (Figure 4.8). This creates a site-specific equilibrium, with relative availability of competitors and their affinity for the

sequence in question determining the ultimate outcome. It implies the ability to fine-tune the output of an active site through subtle sequence variation, where instead of destroying affinity for one of the regulators, relative affinities for a number of regulators are affected instead. This allows for more complex and dynamic interactions than the "classic" model of transcription factors binding side-by-side, although the latter is clearly an important part of transcriptional regulation. I believe that exploring and understanding competitive regulation will be vital to a refined understanding of not only myogenesis, but of every developmental pathway. In addition to the MRFs - E2A - RP58 - Zeb1 system just discussed, in the next section I will provide another example of a potential competitive system at work in C2C12s.

4.5 Of mice and mycs: An active *in vivo* network presents a model for studying competitive transcriptional regulation

In the C2C12 muscle system, transcripts for Myc and its primary dimerization partner Max (Kato et al. 1992) are present in both cycling and differentiating states (Table 4.7). The former is expected for rapidly dividing cells, the latter can be explained in part by the role Myc plays in regulating mitochondrial biogenesis (Li et al. 2005). Myc:Max heterodimers recognize and bind the e-box CACGTG (Krepelova et al. 2014), which in our discussion came up as an acceptable *in vitro* binding site for E47:E47 homodimers. The latter are unlikely to be present in a sufficient physiological concentration to compete with Myc:Max (due to saturation by MRFs and Ids), but other bHLH transcription factors with known or possible motif overlaps are available. Mad, Hes6, Dec1, Dec2, Usf1 and Usf2 all show substantial transcript levels throughout the differentiation process (Table 4.7), and are not reported to dimerize with either E47 of Ids. They belong to two families of bHLH TFs - bHLH-lz (Myc, Max, Mad, Usf1, Usf2) and bHLH-orange (Hes6, Dec1, Dec2) (Dawson et al. 1995; Sun et al. 2007), with dimerization generally permissible among family members. Their roles vary according to the species of dimer, some acting as transcriptional activators and others as transcriptional repressors. All factors listed above bind motifs containing the core CACGTG (Figure 4.9), hence all compete for sections of the available population of CACGTG motifs. While E47:E47 occupancy of CACGTG in muscle is unlikely to occur for other reasons (mainly lack of E47:E47), in B-cells a similar system could serve the additional role of controlling interaction between E:E and CACGTG. Sequence content analysis of E47 regions of occupancy in B-cells supports this theory - while the *in vitro* affinity of E47 for CAGCTG and CACGTG is similar, *in vivo* occupancy favors CAGCTG, and the number of occupied CACGTG motifs is relatively small.

The above analysis of motif preference and of the implied factor competition, synergy, or silencing at specific motif families provides a richer and more comprehensive starting framework for understanding *in vivo* transcription factor occupancy. It is, however, only a starting point, since additional mechanisms,

both cis and trans, are likely at work. On the sequence level, our analysis and other data show that there is some acceptable variation in most core e-box recognition sites for bHLH factors. My analysis also shows that the influence of flanking nucleotides on binding affinity is important to understanding E:E vs. MRF:E occupancy in muscle and B-cells. At the protein level, post-translational modifications (such as phosphorylation) and dimerization affinities play a significant role in the availability and activity of any particular species. Nor is the provided list of TFs potentially interacting with CACGTG-based motifs exhaustive. Nevertheless, this model serves as a starting point for representing a complex regulatory system, where multiple transcription factors could all potentially occupy the same DNA element. The nature and strength of occupancy will be dictated by the overall availability of TFs themselves, their relative affinities for the target site in question, and cooperative effects (either direct or indirect through scaffolding proteins). If properly understood, such a system allows for a highly fine-tuned control of transcriptional output, directed both through subtle alterations of the base motif (often at the flanking rather than the core nucleotides) and variations in the amounts of competing TFs. It also allows for a given CRM to function with a dynamic range of activity, instead of the simple on/off switch behavior usually associated with cis-regulatory elements. More specifically, the bHLH-lz and bHLH-o system presented herein is important both to the proper maintenance of the cell cycle, and to the exit from it that accompanies terminal differentiation. It is therefore likely to be involved in other developmental networks beyond myogenesis, and while not central to cell fate determination, it is important to the successful execution of the differentiation program.

4.6 Testing cis-repression in differentiating muscle

Repressor elements have received less attention than positive acting elements in functional studies of muscle CRMs. I would like to conclude by discussing a likely regulatory module that features several of the motifs highlighted during sequence content analysis of transcription factor occupancy in myogenesis. The region in question is the promoter region of Atoh8, encompassing nucleotides between -600 bp and +0 bp from the RefSeq TSS (chr6:72,185,570 in the mm9 assembly). Atoh8 (atonal homologue 8) is itself bHLH transcription factor that has been implicated in the development of brain, pancreas and kidneys (Chen et al. 2011); it is expressed in cycling myoblasts, but the transcript levels drop twentyfold upon terminal myogenic differentiation (23.1 / 1.4 / 0.8 / 1.1 at cycling / 60hr / 5d / 7d, respectively). Interestingly, its strongly up-regulated upon denervation (Berghella, unpublished). ChIPSeq data for MyoD and myogenin indicate that the neighborhood around Atoh8 contains several regions of high signal in both cycling and differentiating C2C12s (Figure 4.10). The three sites outlined by blue boxes - one upstream, one in the first intron, and one downstream of the gene body - are only occupied by MyoD and myogenin in differentiating myocytes, when Atoh8 expression levels are minimal. The site highlighted in yellow shows evidence of occupancy in both states.

Several explanations for the apparent disparity between factor occupancy and transcriptional output are possible. The first option, which I currently favor, arises from my analysis of motif content of the promoter adjacent region of Atoh8 (Figure 4.11) and the candidate distal elements. Within the promoter region, I found and highlighted four motifs, three of which are associated with repressors present in the system - Dec1, KLf4 and RP58. Furthermore, all three are up-regulated in differentiating myocytes and may explain the Atoh8 down-regulation in the face of MRF occupancy. The 4th motif - an MRF-class octamer AACAGCTG - is also a viable target for RP58, and has no associated MRF signal. I propose that one or more of Dec1, Klf4 or RP58 act to reduce the activity of the Atoh8 promoter, leading

to down-regulation of transcriptional output even in the presence of nearby MRF occupancy. Two alternative possibilities should be mentioned for the sake of completeness. One is that MRF-occupied sites proximal to Atoh8 do not have any interaction with the Atoh8 promoter, and instead either target other genes or accomplish nothing - if an area of chromatin is accessible and contains the correct binding site, MRF presence is likely even when it serves no regulatory function. Another is that a myocyte-specific repressor site(s) exists that was not detected because the motif for the repressor in question was not included in sequence analysis, and because occupancy data for it are not available. This repressor would then act independently of the proposed promoter elements to reduce transcriptional output of Atoh8.

The idea of repression mediated by the sites found in the promoter proximal region of Atoh8 is testable, and I've designed and put in motion a series of experiments to do so. The first stage is to verify that a large construct containing the distal elements and native promoter recapitulate the pattern observed for Atoh8. Stage two is an element-by-element dissection of the locus, measuring activity of each distal element with a simple, standard heterologous basal test promoter (TK minimal). The goal of stage two is to learn which elements have activity, and to measure it in both the myoblast and the myocyte settings. Simultaneously, the activity of the Atoh8 proximal promoter region will be tested on its own, and with fusion to cis-elements with known functional activity, including a myoblast-specific enhancer, a myocyte-specific enhancer, and a constitutive enhancer. Presently, DNA constructs are designed, but the experiments themselves have not been carried out. While speculative, this region serves as a good candidate for functional testing and study of repressive effects.

References (chapter 4)

Bretones, G., Delgado, M.D., and Leon, J. (2014). Myc and cell cycle control. Biochim Biophys Acta.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L*., et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell *18*, 662-674.

Chen, J., Dai, F., Balakrishnan-Renuka, A., Leese, F., Schempp, W., Schaller, F., Hoffmann, M.M., Morosan-Puopolo, G., Yusuf, F., Bisschoff, I.J*., et al.* (2011). Diversification and molecular evolution of ATOH8, a gene encoding a bHLH transcription factor. PLoS One *6*, e23005.

Dawson, S.R., Turner, D.L., Weintraub, H., and Parkhurst, S.M. (1995). Specificity for the hairy/enhancer of split basic helix-loop-helix (bHLH) proteins maps outside the bHLH domain and suggests two separable modes of transcriptional repression. Mol Cell Biol *15*, 6923-6931.

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

Hirai, S., Miwa, A., Ohtaka-Maruyama, C., Kasai, M., Okabe, S., Hata, Y., and Okado, H. (2012). RP58 controls neuron and astrocyte differentiation by downregulating the expression of Id1-4 genes in the developing cortex. EMBO J *31*, 1190-1202.

Honma, S., Kawamoto, T., Takagi, Y., Fujimoto, K., Sato, F., Noshiro, M., Kato, Y., and Honma, K. (2002). Dec1 and Dec2 are regulators of the mammalian molecular clock. Nature *419*, 841-844.

Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.

Kato, G.J., Lee, W.M., Chen, L.L., and Dang, C.V. (1992). Max: functional domains and interaction with c-Myc. Genes Dev *6*, 81-92.

Krepelova, A., Neri, F., Maldotti, M., Rapelli, S., and Oliviero, S. (2014). Myc and max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. PLoS One *9*, e88933.

Li, F., Wang, Y., Zeller, K.I., Potter, J.J., Wonsey, D.R., O'Donnell, K.A., Kim, J.W., Yustein, J.T., Lee, L.A., and Dang, C.V. (2005). Myc stimulates nuclearly encoded mitochondrial genes and mitochondrial biogenesis. Mol Cell Biol *25*, 6225-6234.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J*., et al.* (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol *11*, 635-643.

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

Okado, H., Ohtaka-Maruyama, C., Sugitani, Y., Fukuda, Y., Ishida, R., Hirai, S., Miwa, A., Takahashi, A., Aoki, K., Mochida, K.*, et al.* (2009). The transcriptional repressor RP58 is crucial for cell-division patterning and neuronal survival in the developing cortex. Dev Biol *331*, 140-151.

Shen, M., Yoshida, E., Yan, W., Kawamoto, T., Suardita, K., Koyano, Y., Fujimoto, K., Noshiro, M., and Kato, Y. (2002). Basic helix-loop-helix protein DEC1 promotes chondrocyte differentiation at the early and terminal stages. J Biol Chem *277*, 50112-50120.

Shirakata, M., and Paterson, B.M. (1995). The E12 inhibitory domain prevents homodimer formation and facilitates selective heterodimerization with the MyoD family of gene regulatory factors. EMBO J *14*, 1766-1772.

Siles, L., Sanchez-Tillo, E., Lim, J.W., Darling, D.S., Kroll, K.L., and Postigo, A. (2013). ZEB1 imposes a temporary stage-dependent inhibition of muscle gene expression and differentiation via CtBP-mediated transcriptional repression. Mol Cell Biol *33*, 1368-1382.

Stevens, J.D., Roalson, E.H., and Skinner, M.K. (2008). Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation. Differentiation *76*, 1006-1022.

Sun, H., Ghaffari, S., and Taneja, R. (2007). bHLH-Orange Transcription Factors in Development and Cancer. Transl Oncogenomics *2*, 107-120.

Figures and Tables (chapter 4)

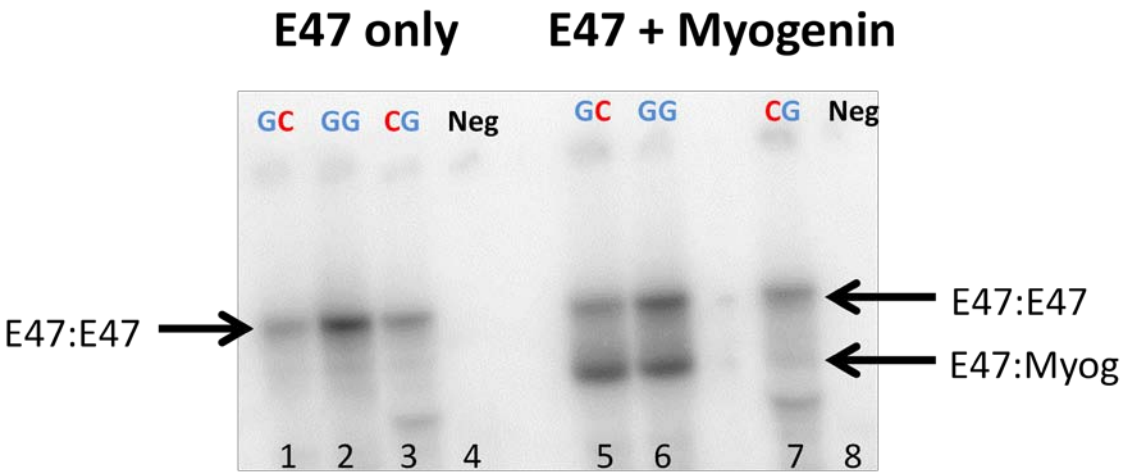


Figure 4.1. Binding affinity assay using E47, Myogenin and ^{32}P-labeled double-stranded oligonucleotide probe. Both proteins were synthesized *in vitro* using a rabbit reticulocyte lystae expression system. Lanes 1-4 contain E47 only, lanes 5-8 contain E47 and myogenin in equimolar ratios. Probes are: GACAGCTG (lanes 1 and 5), GACAGGTG (lanes 2 and 6), GACACGTG (lanes 3 and 7), GTCTAGAACG (lanes 4 and 8). (Gel shift #72)

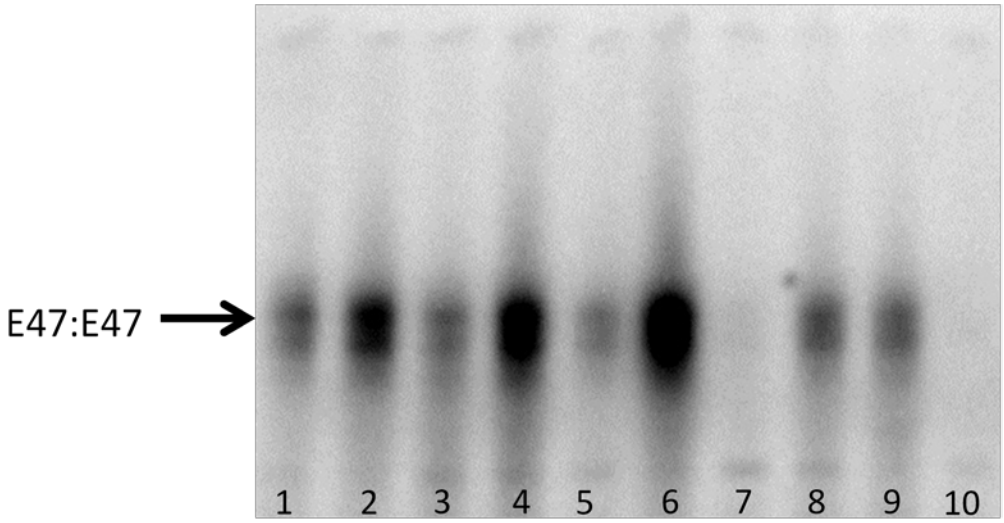


Figure 4.2. Interaction between E47:E47 homodimers (all lanes) and a panel ^{32}P labeled e-box sequences. Probe intensity was standardized at 500K cpm per reaction. Each lane contains E47 synthesized *in vitro*. Probes are: GACAGCTG (1), GACAGGTG (2), GCCAGCTG (3), GCCAGGTG (4), CGCAGCTG (5), CGCAGGTG (6), CTCAGCTG (7), CTCAGGTG (8), GACACGTG (9), GTCTAGAA (10). (Gel shift #93)

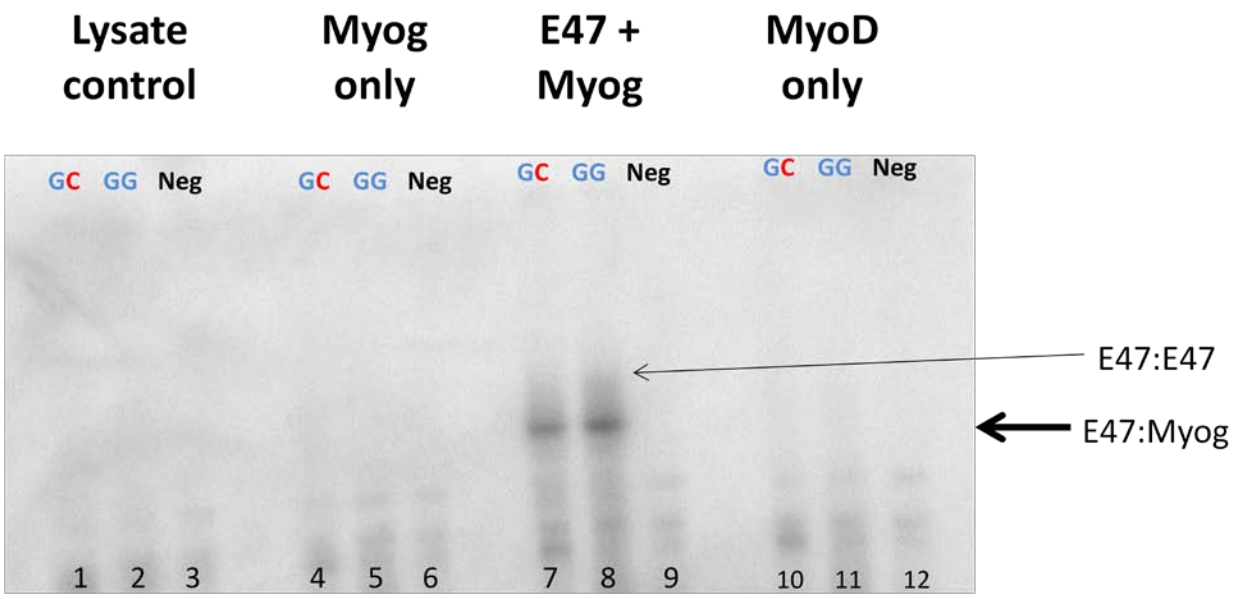


Figure 4.3. Myogenin and MyoD do not bind RRCAGSTG in the absence of E47. Proteins were: reticulocyte mixture incubated without plasmid template (lanes 1-3), myogenin (lanes 4-6), myogenin and E47 in equimolar concentrations (lanes 7-9), MyoD (lanes 10-12). Oligonucleotide probes were: GACAGCTG (lanes 1, 4, 7, 10), GACAGGTG (lanes 2, 5, 8, 11), GTCTAGAACG (lanes 3, 6, 9, 12). (Gel shift #89).

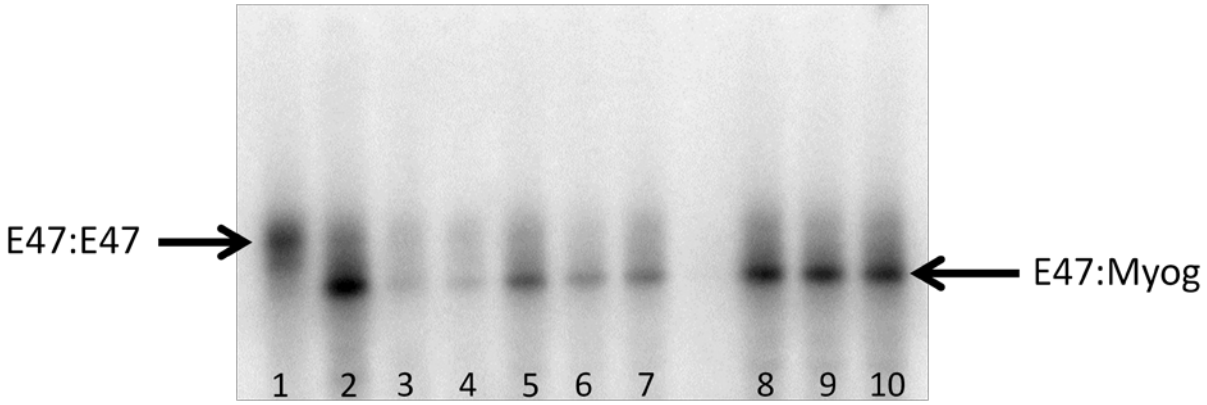


Figure 4.4. Competitive binding assay using E47 only (lane 1) and E47:Myog mixture at a 1:3 molar ratio (lanes 2-10). ^{32}P-labeled probe GACAGGTG at a 0.34 µM concentration was used in all lanes (1-10). Unlabeled competitor probes were added to lanes 3-10 at a 3.4 µM concentration (10x molar excess). Competitors are: GACAGCTG (3), GACAGGTG (4), GCCAGCTG (5), GCCAGGTG (6), CGCAGCTG (7), CTCAGCTG (8), CTCAGGTG (9), GTCTAGAA (10). (Gel shift #103)

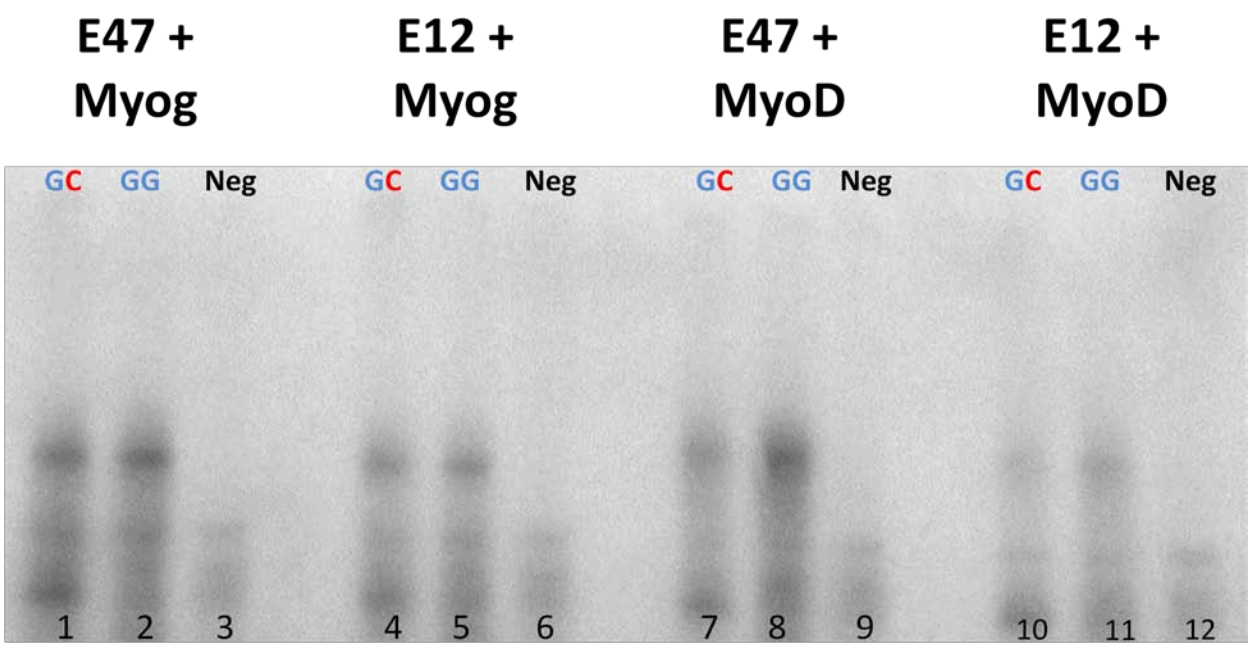
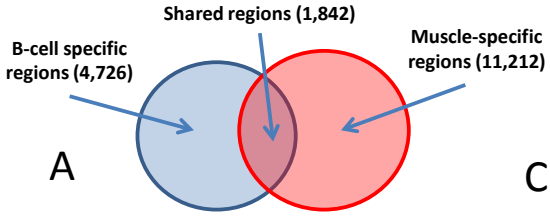


Figure 4.5. Both MyoD:E12 and Myogenin:E12 formed *in vitro* from a linked transcription system bind MRF-class e-boxes in a manner similar to MyoD:E47 and Myogenin:E47. Oligonucleotide probes were: GACAGCTG (lanes 1, 4, 7, 10), GACAGGTG (lanes 2, 5, 8, 11), GTCTAGAACG (lanes 3, 6, 9, 12). (Gel shift #90)

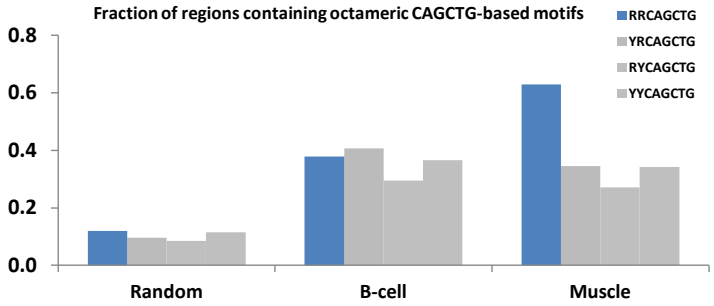


A

B

Motif	muscle	B-cell
CAGCTG	2.42	2.33
CACCTG	0.82	2.17
CAGSTG	1.69	2.24
RRCAGSTG	2.67	2.13
RRCAGCTG	3.08	2.11
RRCAGGTG	1.77	2.16
CACGTG	0.96	0.97
AP-1	1.00	0.11
RP58	-0.34	1.34
Mef2 (reference)	0.22	-1.12
Usf1	1.41	0.68
RYCAGCTG	2.09	2.23
RYCAGGTG	0.38	2.11
YRCAGCTG	2.22	2.73
YRCAGGTG	0.17	2.48
YYCAGCTG	1.79	2.21
YYCAGGTG	-0.11	1.92
CGCGCG	3.23	1.37
TATA-box	-2.27	-2.84
Sp-1	2.61	1.56
Dec1	0.89	0.69

C



D

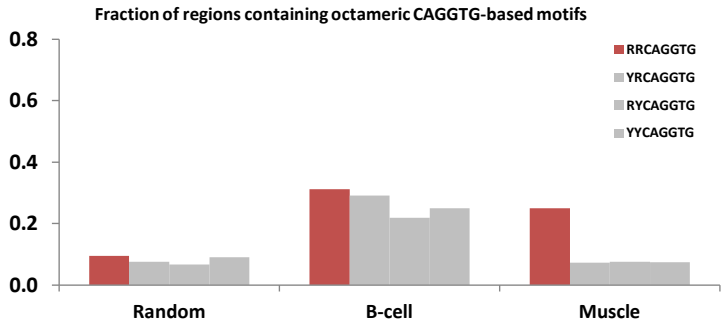


Figure 4.6.

A) Overlap between regions occupied by E47:E47 homodimers in differentiating B-cells (blue) and Myog:E47 heterodimers in differentiating muscle cells (red).

B) Motif content analysis of B-cell-specific (**B-cell**) and muscle-specific (**muscle**) regions of occupancy.

C) CA**GC**TG motifs preferentially contain the RR prefix in muscle-specific regions but not in B-cell specific regions.

D) CA**GG**TG motifs preferentially contain the RR prefix in muscle-specific regions but not in B-cell specific regions. Note that CAGGTG motifs without the RR-prefix occur at background levels in muscle-specific regions

Transcription factor	cycling	diff - 60h	diff - 5d	diff - 7d
Myc	25.3	20.5	12.4	13.0
Mycl1	7.9	5.7	4.3	3.9
Max	30.1	16.9	12.0	12.5
Mad	33.9	21.0	17.5	13.1
Hes6	55.4	134.4	107.3	68.3
Dec1	12.0	73.0	70.7	64.1
Dec2	4.3	14.4	12.8	16.7
Usf1	26.7	28.9	26.3	28.4
Usf2	40.0	39.3	32.7	39.8

Table 4.7. mRNA levels of several CA**CG**TG-binding bHLH transcription factors in cycling and differentiating C2C12s, based on RNASeq measurements. Units of expression are FPKM (fragments over kilobase of RNA per million reads)

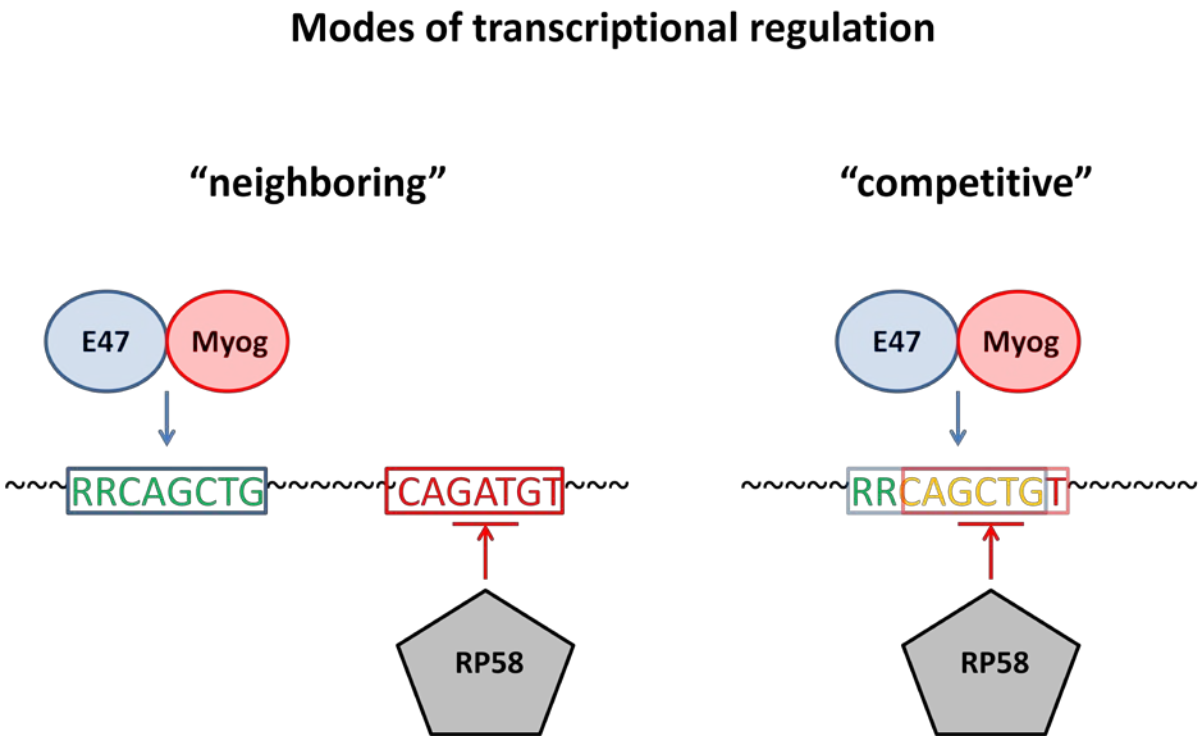


Figure 4.8. Comparison between two modes of transcriptional regulation involving the same set of TFs. In the **"neighboring"** (or classic) case, each TF has its own binding site within the module, and joint occupancy leads to the desired output (be it repressive, activating, or insulating). In the **"competitive"** case, both factors compete for the same sequence, and the output of the module depends on the current equilibrium conditions achieved through physiological availability of competing TFs and their respective affinities for the target site (as well as any secondary interactions that might be involved).

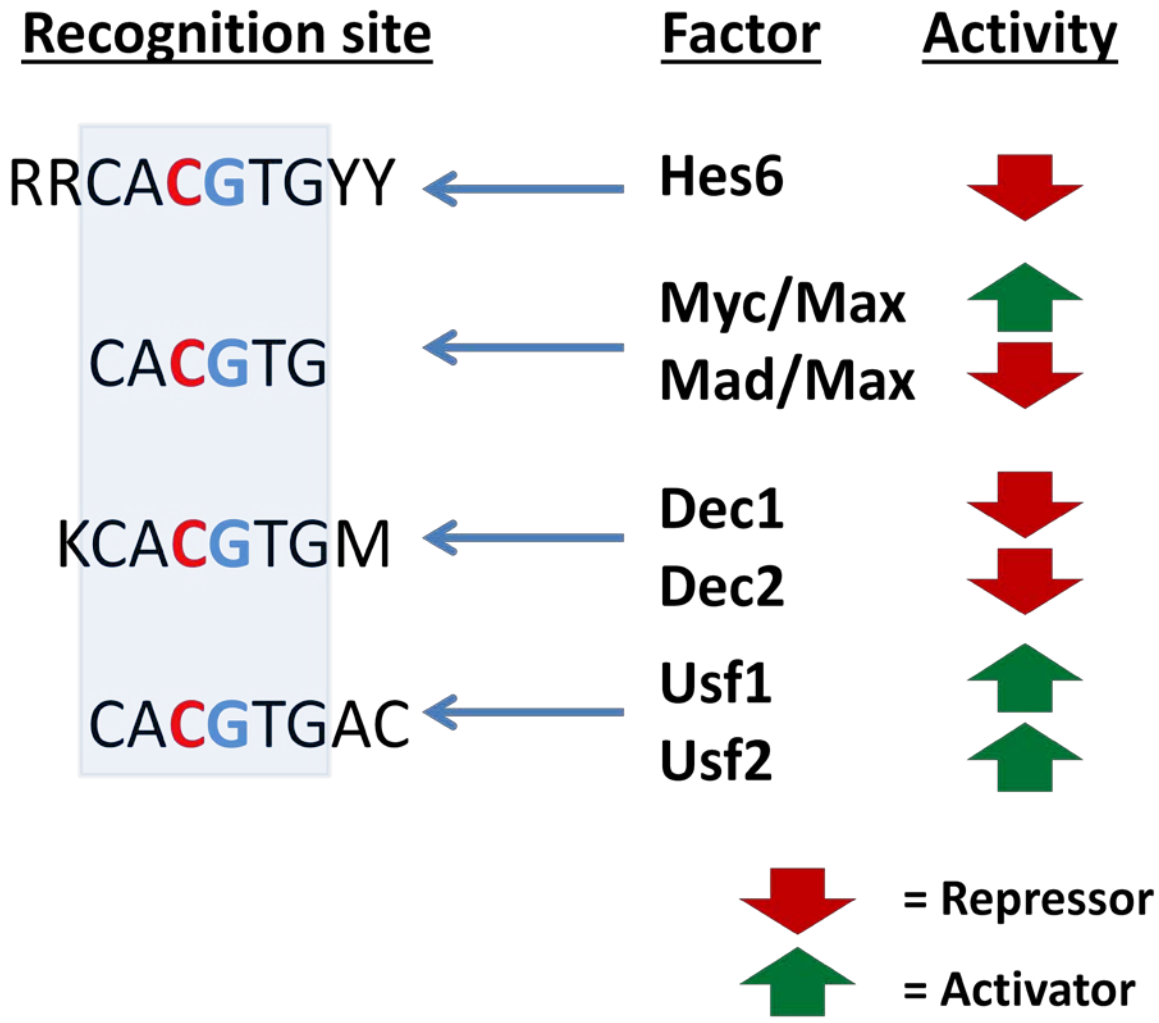


Figure 4.9 Various bHLH-lz and bHLH-o TFs present in C2C12s recognize common sets of DNA elements, but exert contradictory influences on transcriptional output from the linked TSS.

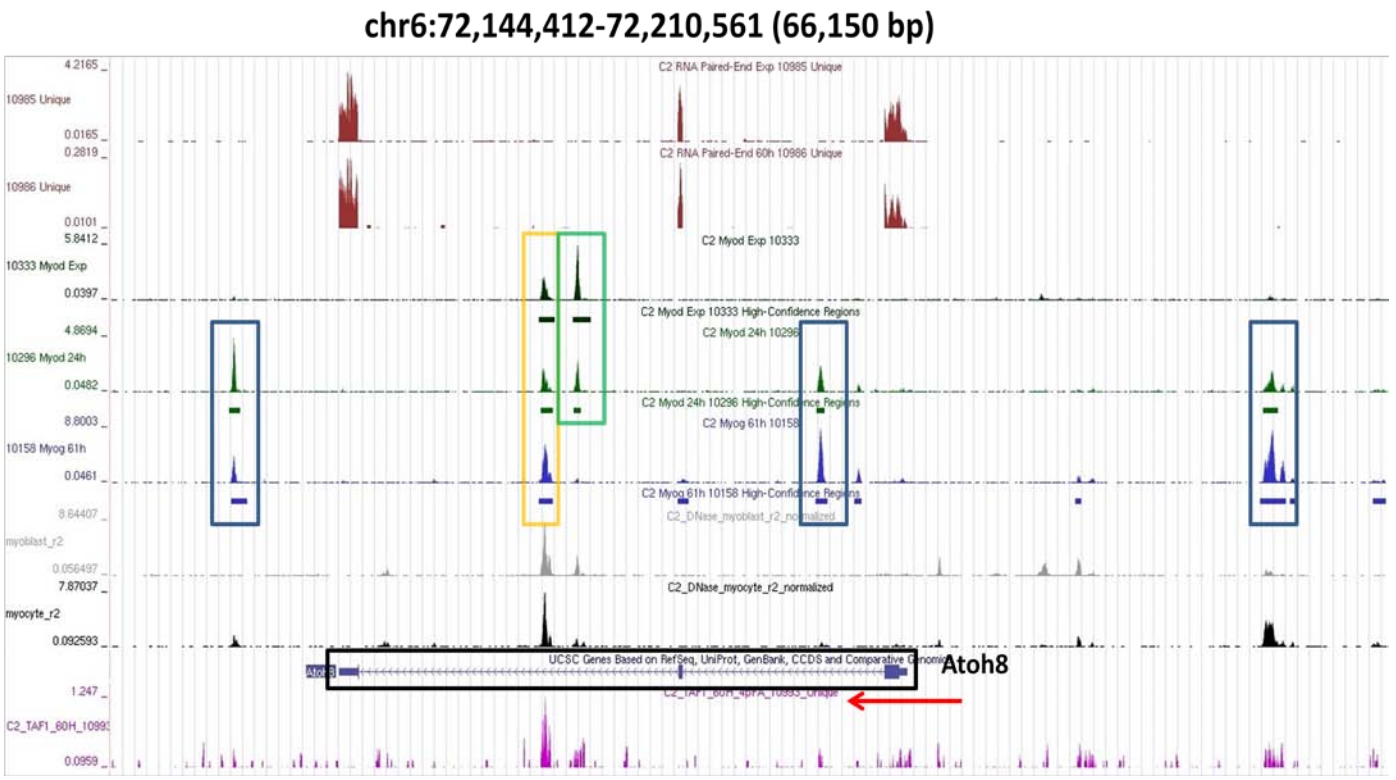


Figure 4.10. MRF-occupied elements around the Atoh8 gene. Elements highlighted in blue are only occupied by MyoD and myogenin in differentiating myocytes. The element in gold is continuously occupied, showing a MyoD signature in cycling myoblasts and both MyoD and myogenin signatures in differentiating myocytes. The element in green is preferentially occupied in cycling myoblasts - while it registers a MyoD signal at 24 hours after differentiation, it lacks any myogenin occupancy at 60 hrs. The red arrow denotes the direction of the Atoh8 gene model relative to the diagram.

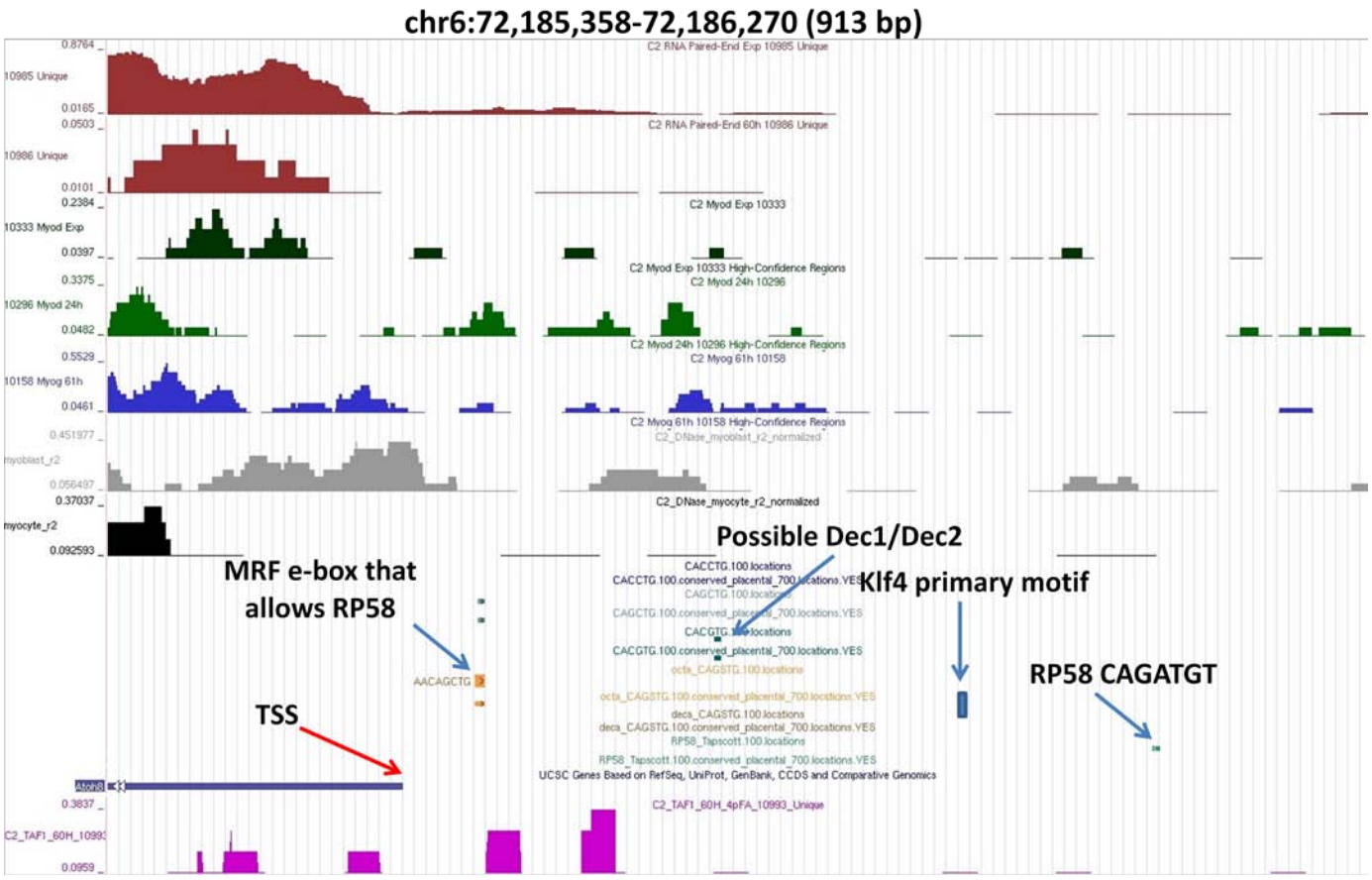


Figure 4.11. Potential regulatory motifs in the Atoh8 promoter. The e-box motif immediately upstream of the TSS is of the form AACAGCTG, with the reverse compliment CAGCTGTT. This makes it both an MRF:E and an RP58 binding site. The CACGTG e-box to the right of it is a possible Dec1/Dec2 interaction site, although this interaction is purely hypothetical. Also present are recognition sites for Klf4 and RP58 (the latter being a cited optimal site CAGATGT). Both are repressors, and both are up-regulated in differentiating myocytes.

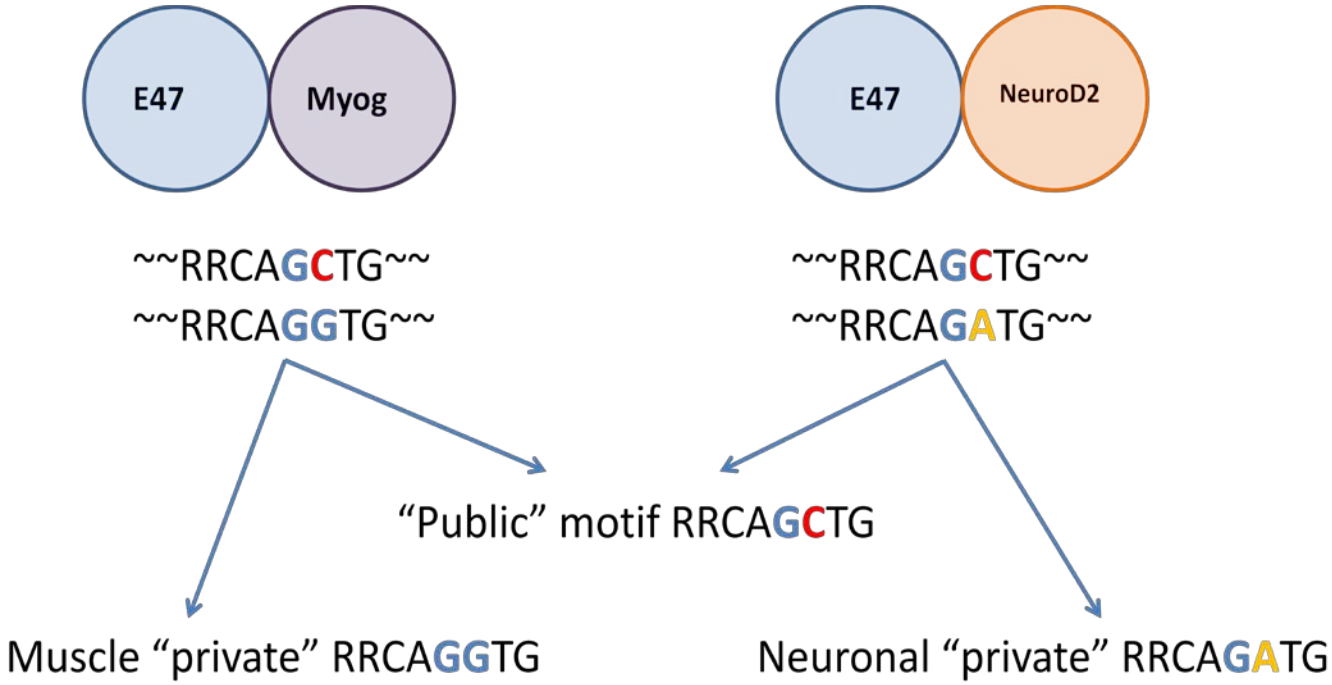


Figure 4.12. Public and private e-boxes in myogenic vs. neurogenic lineages (based on Fong et al. 2012)

Chapter 5: Conclusions

I would like to conclude the presentation of my thesis results with a few summary remarks, and discuss the major question raised by my work, potential approaches the answering them, and their implication to the study of transcriptional regulation in skeletal muscle specifically and in development generally. To begin with, an interesting result arose from the analysis of the genome-wide occupancy map of myogenin - a highly differentiation-specific transcription factor was found associating with genes whose main distinguishing characteristic is that they are expressed, without regard for the actual expression trajectory. It immediately leads to two questions that merit further consideration. First - is there a need for a better method of associating TF-occupied elements with candidate regulatory targets; and second - is myogenin occupancy causal of transcriptional output, or merely coincidental with it?

The first question is, in a sense, rhetorical - association based on chromosomal proximity is unbiased, but represents a "best guess" scenario. Ideally, a method utilizing evidence of physical interaction between a regulatory element and its target promoter(s) should be used to make region-gene assignments. Obtaining such a measurement is by no means trivial, and method for doing so has only recently become available (Fisher-Aylor, unpublished). It provides a clear path for moving forward, with an opportunity to refine target associations, and data gathering/analysis are currently underway to address the larger question of 3D interactions between cis-regulatory elements and actively transcribed genes. Such a map will increase the veracity of the list of genes affected by myogenin, and will lead to a better understanding of muscle-specific transcriptional activation. However, it is unlikely to change the overarching conclusion that myogenin-occupied regions preferentially associate with genes expressed in differentiating skeletal muscle, with myocyte-specific genes making up an important but relatively minor (no more than 10%) fraction thereof.

The second question poses an interpretation dilemma. The preponderance of myogenin occupancy associating proximally with expressed genes can be explained in two ways. One is that myogenin, as a positive-acting transcription factor, occupies cis-elements around genes that need to be expressed in differentiating muscle (including those that are not differentiation-specific), and contributes to their expression. The somewhat more pessimistic explanation is that chromatin in the area of an expressed gene is more likely to be accessible, in turn increasing the likelihood of making a myogenin binding site available. Under this model myogenin will occupy any binding site that is not being obstructed by a competitor or actively repressed, but will have little to no influence on nearby gene expression, except at a relatively small (less than 10% of total occupied elements, limited to differentiation-specific genes) fraction of sites. In truth, both explanations are almost certainly valid. Based on its expression pattern, loss of function phenotype, and extensive mutagenesis analysis of select CRMs, it is clear that myogenin is crucial to the expression of muscle-specific genes and proper progress of terminal differentiation. It is also very likely that myogenin, especially due to its high abundance, can occupy most elements in the genome that are presented to it unobstructed and meet the criteria for occupancy, such as having an RRCAGSTG recognition site, or perhaps more sophisticated combinations of targeting motifs. The fundamental question, therefore, is how much does myogenin really contribute to the expression of genes that are not differentiation-specific?

A way to begin evaluating this is through functional testing, where select elements are attached to a reporter construct, transfected into a skeletal muscle system, and construct activity measured under conditions of differentiation. A number of such experiments were performed by a colleague - Gilberto DeSalvo, and one of them I will mention specifically. Elements from two groups were selected and evaluated based on their ability to activate a reporter construct in differentiating C2C12 myocytes. Group 1 consisted of elements associated with genes expressed in myocytes, regardless of their expression trajectory over the course of differentiation or absolute mRNA abundance (so long as the

latter was statistically significant based on the RNASeq measurement). Group 2 consisted of elements associated with genes lacking detectable (by RNASeq) transcript levels in myocytes. Elements in group 1 were more likely to activate a reporter construct, with $p < 0.01$. While on its own not definitive, this strongly suggests that at least a number of non differentiation-specific genes are being regulated by myogenin in differentiating skeletal muscle.

This raises a larger evolutionary question that extends beyond the muscle system. It is reasonable and likely to assume that in other tissue types the predominant bHLH factor, if there is one, will have a large repertoire of accessible binding sites, and therefore exert influence over a large fraction of expressed genes. In the case of skeletal muscle, nearly one third of the ~ 15,000 expressed genes have an associated myogenin-occupied CRM. NeuroD2 and E47 occupancy studies in neurons and B-cells, respectively, tell a similar tale, albeit with varying percentages. But in all three cases the number of "differentiation-indifferent" targets greatly outweighs the number of differentiation-specific ones. What, then, imparts tissue specificity? Regulatory elements containing the motif RRCAGCTG are "suitable" to a number of bHLH TFs, and the presence of such an element would lead to the expression of the target gene in a variety of tissues. Perhaps such generic CRMs form the core of the overall regulatory network, where more precise control over expression levels has evolved over time through a number of mechanisms, such as targeted epigenetic modifications, tissue-specific repression (perhaps at a CRM altogether), or competitive binding, where multiple species of TFs can recognize and bind the same motif. It implies that a fairly large number of genes would have SOME level of expression in a variety of tissues - a hypothesis is at least superficially supported by the currently available RNASeq data. A more in-depth analysis of the prevalence and use of such cross-tissue elements through a combination of ChIPSeq, RNASeq, and DNAse-hypersensitivity measurements would be invaluable to the overall understanding of transcriptional regulation in mammalian development.

Two similar results emerged from the cross-factor occupancy comparisons, both providing clues about a different aspect of transcriptional regulation. It has long been known that at a number of well-studied elements several TFs contribute to the overall transcriptional output, such as the example of the "classic" muscle CRM involving joint occupancy/binding by Mef2 and MRF:E. It is therefore believed that CRM diversity is at least partially achieved through the combinatoric use of recognition motifs, allowing for an extremely diverse range of "recipes" to be implemented through joint occupancy and/or repression. While this model certainly remains true, both the Mef2 and early MyoD occupancy data point to an alternative mechanism that is also prevalent, where recruitment of a co-factor is accomplished primarily through protein-protein interactions, rather than joint DNA binding. The immediate question - what makes those elements "special" - does not have a simple answer, at least on the level of underlying sequence. Detailed analysis of 1323 regions jointly occupied by myogenin and Mef2 and containing only an RRCAGSTG motif without an accompanying CTAWWWWTAG failed to reveal a secondary motif that would help account for Mef2 occupancy. While ChIPSeq data point to the presence of Mef2 at these elements, from a motif content standpoint there is no basis for it being there. Once again, a number of explanations are plausible. The most straightforward one is that Mef2 is being recruited to these sites, either by interacting directly with myogenin:E (as suggested by Molkentin et al. 1995), or by means of another part of the complex, such as p300/pCaf. But the question remains as to what distinguishes these ~ 1,300 sites from the remaining ~ 11,000 sites that have myogenin occupancy and an underlying RRCAGSTG motif, but no joint Mef2 presence. A more involved explanation is that Mef2 is recruited via protein-protein association, but only at elements where 3D interactions with other CRMs permit it. Under this model looking for distinguishing sequence characteristics that separate co-occupied elements would be futile, as there is no reason for them to be present in the first place. Instead, groups of connected elements would have to be considered as a whole, with specificity encoded either in the joint sequence content or the spatial arrangement of the resulting super-complex.

A priori it is difficult, if not impossible, to predict how such CRMs would be grouped - only through physical observation and study of long-range interactions can this model be evaluated. It further emphasizes the need for refining techniques such as ChIA-PET, and is an exciting problem to study not just in relation to myogenin and Mef2, but as general question in the control of gene expression. While transcriptional activity is regulated by complexes of collaborating TFs and co-factors, not all TFs with the ability to bind DNA that are present in a complex need to be encoded for in any given CRM.

Finally, I would like to emphasize the importance of understanding binding affinities and their implications for transcriptional regulation and competitive occupancy. Cohorts of factors that recognize variants of the same core site are present at various levels in both myoblasts and myocytes. Undoubtedly, similar groups exist in other cell types. The bHLH family is large and diverse, and although various members have diverged to recognize different versions of the e-box motif (both with regard to the two interior nucleotides and the flanking sequence), much overlap still exists. The information encoded in the motif provides a gradient of affinities for a variety of DNA-binding factors, which are present at different concentrations and in different combinations depending on the cell type or state. Additionally, non bHLH TFs and zinc-finger repressors can often recognize essentially the same sites. Understanding affinities and relative concentrations of potential binders can help us better evaluate the likely effect of a given CRM on gene expression. For example, not long ago we might have considered the motif CTCAGGTGT to be a likely activating element in the skeletal muscle system - superficially it matches the CAGSTG primary binding site associated with MRF:E heterodimers. However, I have now shown that such a motif is most likely to have a repressive function in skeletal muscle, if it has one at all. The YY "prefix" makes it an unfavorable target for MRF:E heterodimers, while the GG center and T suffix make it an optimal recognition site for Zeb1 and a likely recognition site for RP58 - both repressors that are up-regulated in differentiating myocytes. Understanding competitive binding systems is an

important, and as yet relatively little-studied question in the context of understanding dynamic regulation of transcriptional output.

Appendix: Materials and Methods

Read mapping and peak calling

Sequenced reads were mapped to the mm9 genome assembly using Eland originally and Bowtie after 2009. Initial Eland mapping were re-done using Bowtie to maintain consistency, which did not have a noticeable impact on the overall data analysis. Peaks were mapped using the ERANGE algorithm (Mortazavi et al. 2007), which looks for areas with an over-representation of reads from the ChIP sample compared to the control sample (carried out with no antibody). Absolute read numbers were normalized to reads per million (rpm) to allow for comparison of datasets with differing levels of sequencing depth. For medium confidence (MC) region definition, a minimum read level of 3 rpm (reads per million) and 2x ratio of experimental/control reads was required. For high confidence (HC) regions, a minimum read level of 5 rpm and a 4x ratio of experimental/control reads was required.

Script library

Extensive use was made of a collection of python scripts that I wrote. They will be made available publically at http://woldlab.caltech.edu/~akirilus/ChIPSeq_utilities/ . A user manual will be supplied for some of the more commonly used scripts.

Association between regions of occupancy and genes

In the absence of a direct measure of physical connectivity (such as ChIA-PET), proximity on the chromosome was deemed a rational and unbiased of associating occupancy events with candidate target genes. A list of 31680 RefSeq models used for mapping of RNASeq data was used in conjunction, with a python script designed to perform a "closest neighbor" search between the list of TSSes and the list of occupancy peaks (either as determined by ERANGE, or taken as the midpoint of the occupied region for some of the very early data). Initially association were restricted to a maximum distance of either 20K or 50K nucleotides, but analyses presented in this thesis all use unlimited range associations unless explicitly stated. This assignment is performed by the script tag_regions_by_TSS_distance.py .

Region length normalization

Due to some variability in the individual lengths of called regions of occupancy, efforts were taken to normalize their length. An unbiased method was to take either the computationally defined occupancy "peak" (nucleotide position corresponding to the highest ChIPSeq signal), or the midpoint of the region for some very early data that did not contain peak predictions, and extend it by a certain number of nucleotides on either side, referred to as "region radius". Three sets of radii were considered extensively - 50 bp, 100 bp and 250 bp. Most of the primary motif information could be captured using the 50 bp radius. The 250 bp radius provided the greatest flexibility for locating collaborating motifs, but resulted in an overall dilution of the density of primary motifs (when taken over the whole of the region). This was later corrected in motif density mapping, where each region would be split into 10 "steps" of equal length (each 1/10th of the total length), and motif densities would be calculated per

cell. The latter was made possible by the assumption that the computationally predicted peak correlates reasonably well with an actual occupancy event (which turns out to be true for most factors that were assayed). Ultimately, the 250 bp radius was used for the analysis, meaning all standardized regions were of the same length - 501 nucleotides long (central nucleotide and 250 on either side). A peak caller output file generated by ERANGE can be automatically length normalized to any desired radius using the utility convert_hts_to_radius-hts.py .

Mapping of motifs in the mm9 genome

To rapidly create motif location libraries, a C++ program was used that compare a supplied PWM to all motifs of the appropriate length, and reported all matches above a certain similarity threshold. As inputs, the program required a PWM, a minimum similarity score required to generate a match, a list of all chromosome names in the genome to be considered, and appropriately named FASTA files for each chromosome (the latter contain the actual sequence). The chromosome list allows for mapping to only select chromosomes, if desired. Similarity score was computed through summation of fitness scores, where a fitness score corresponded to the likelihood of seeing a particular nucleotide at that position, as defined by the PWM. The total score was then divided by the maximal score attainable for the PWM - if the ratio exceeded the pre-defined threshold, the location of the motif was reported, otherwise it was disregarded. In general, for shorter sequences (8 nucleotides or fewer), a 100% match was required, while for longer and more degenerate sequences (such as the CTCF and NRSF motifs), this was relaxed to 80-85%. Generation of libraries that included all locations for a given motif at a given stringency was desirable, as it allowed for rapid mapping of motifs to regions of ChIPSeq occupancy, and for additional filtration of motifs based on such criteria as conservation or overlap with simple repeats catalogued in the repeat masker database.

Motif density calculations

To compute motif density and the resulting enrichment/depletion relative to the background density in the genome at large, a stepwise process was used. First, motifs were mapped to ChIPSeq regions, with the distance between center of the motif and center of the region used to identify its relative location (this number could be either positive or negative). All mapped motifs were then "stacked" on top of each other, and each region split into 10 "steps" of equal length. Note that the number of steps is flexible, and reducing it to 1 will simply return the overall density/enrichment measurement for regions as a whole. Next, all motifs in a given step were summed across all regions, and that number divided by (size of step * number of regions in the list) to create a density measurement in motifs/nucleotide. The resulting density was then divided by the total genome density of the motif, defined as (# of motifs in the genome / # of nucleotides in the genome). This provided the relative enrichment or depletion value. Statistical significance was evaluated using a χ^2 test for each "step", with a requirement of $p < 0.01$ in order for an enrichment or depletion to be considered significantly different from the background. Using too many steps in conjunction with a small enough input set of regions could render the output meaningless due to lack of statistical significance - this is reported through the associated χ^2 calculation.

Motif mapping, motif density calculations, and $\chi 2$ values were computed in an automated pipeline available in the script library (batch_map_motifs.py and batch_analyze_regions.py). As inputs these programs require a list of motifs to be mapped, a list of regions to which they will be mapped, and a pre-generated locations library for each motif. While mapping will be performed accurately on regions of any length, density calculations and associated significance tests can only be performed on regions of standardized length, due to the "stacking" requirement.

Region overlap detection

A script was written to detect overlap between regions in two occupancy maps (matching_regions3.py). It takes as input two occupancy maps and a threshold of similarity. The regions in each determination need to be sorted in ascending coordinate order - this is done automatically if ERANGE output is being used. Overlap between two regions is defined as (# of nucleotides in common / length of the smaller of the two regions). Any pair of regions for which overlap ≥ threshold is reported. Several ways of reporting overlapping regions exist, and 4 of them are available with this script (see manual). In the analysis reported as a part of this thesis, 80% similarity thresholds were required to declare two occupancy events overlapping, unless otherwise stated. Because region lengths were standardized, this means that they would have to share at least 401 out of 501 contiguous nucleotides.

Repeat masking

To reduce potential motif density biases introduced by simple repeats, total motif libraries were filtered prior to being mapped onto ChIPSeq regions. The filter used as input a total motif library and a list of simple repeats (available from UCSC or repeatmasker), and returned a list of motifs that did not overlap any of the coordinates included in the repeats database. The script is available as motif_repeat_masker_filter.py . Note that doing so reduces both the total pool of motifs and the effective size of the genome - both were taken into account when computing relative and background densitites.

Conservation of motifs

In order to evaluate the relative conservation of a motif, placental mammals PhastCons data were used. Each nucleotide is given a conservation score by PhastCons, ranging from 0 (not conserved) to 1 (completely conserved). Several ways exist of computing a total conservation score for a motif, with the simplest being to sum up all scores for all base positions, divide it by length, and compare it to the desired threshold. I chose to adopt a different method (dubbed "entropic" for the purpose of this discussion), which aimed to take advantage of binding affinities encoded by a PWM. Instead of looking at the sum of conservation scores, each nucleotide is considered individually. The threshold value (between 0 and 1) is multiplied by the normalized information content of that base as defined by the PWM (1 for a perfectly defined base, 0 for an N) to compute the adjusted threshold. If the PhastCons score listed is ≥ adjusted threshold, the nucleotide passes, otherwise it fails. If every nucleotide passes the threshold test, the motif is considered entropically conserved at that level. Extensive testing pointed to 0.7 as a reasonable entropic conservation threshold to be used in the analysis.

Randomized region generation

To provide an measure of how likely a motif is to occur in a population of 501-nucleotide long regions selected at large, a list of randomized regions was created. Initially, the criteria was only set that the regions be 501 nucleotides long. The resulting 500,000 regions were filtered for overlap with simple repeats (via repeatmasker) and for overlap with centromere/telomere regions of the chromosome to which they were assigned. Any regions with overlap to either were discarded. The winnowed set was then further filtered for sequencing quality, and any region containing unsequenced nucleotides (Ns) was also discarded. The resulting set of ~101,300 regions was treated as a control set of ChIPSeq regions for evaluating central tendencies or enrichments of motifs (it was not used to compute density enrichment - total genomic density was used, as described previously). No effort was made to remove coding sequence from the set of control regions because ChIPSeq regions at times overlap exons.

Gel shift assays

Mobility shift assays were performed in 20 µl volume using a final concentration of 25 µM HEPES, 1 µM DTT, 100 mM NaCl, 5 µM $MgCl_2$, 5 µM EDTA, 5% glycerol by volume. 400 ng poly dI*dC was used in each reaction to inhibit non-specific binding. Labeled probe, cold competitors and protein products were added as appropriate, with 0.15 - 0.4 pM of labeled probe used per reaction (based on emission standardization or the requirements of the reaction). Emission intensity standardization was done using dry scintillation counting. Reactions were incubated at room temperature for 2 - 3 hours, then run on 8% Tris-Glycine non-denaturing gels.

Oligonucleotide labeling and protein synthesis

Oligonucleotide probes were labeled ordered as single-stranded complementary pairs. The forward strand oligo was radioactively labeled using ^{32}P γ-ATP from Perkin Elmer and a 5'-DNA labeling kit (polynucleotide kinase) from Promega. After the labeling reaction, forward and reverse strand oligos were annealed by melting (95^{o}C for 5 mins) and allowing to cool to room temperature. NEB2 buffer was used as annealing buffer. Transcription factor synthesis was performed *in vitro* using a rabbit reticulocyte coupled expression kit (Promega) with Sp6. Non-linearized templates (as per the protocol provided with the kit) were used, with a total amount of input DNA in the range of 400-500 ng / reaction. Protein synthesis was assessed by ^{35}S labeling (Perkin Elmer) and running an aliquot of the reaction product on a 10% Bis-Tris denaturing gel. Although a radioactive protein can be used in conjunction with cold (unlabeled) oligonucleotides to generate a "reverse labeled complex" (this was done at one point to correct technical issues), cold proteins were used in conjunction with labeled probes to generate the actual mobility shift data presented in chapter 4.

ChIPSeq

The protocol used for ChIPSeq, associated library building, and sequencing, is essentially the same as the one used by Mortazavi et al. (2007) and currently utilized by ENCODE.