Transcriptional Enhancer Activity of Biochemically Marked Genomic Elements

Thesis by
Gilberto DeSalvo

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended June 11th, 2018

Gilberto DeSalvo
ORCID: 0000-0002-8957-1699

ACKNOWLEDGEMENTS

I would like to thank my adviser Barbara Wold, as you spent countless hours shaping me into the young scientist I am today. Without you giving me many opportunities, starting with a SURF and culminating in my PhD, I would not be where I am today. I also thank you for your unending patience while I learned and failed my way towards performing and controlling experiments, analyzing and presenting data indubitably. I look forward to continue learning from you as we work on our manuscript and on future works together.

I would also like to thank the members of my thesis committee: Paul Sternberg, Ellen Rothenberg, Eric Davidson and Ross Hardison. Your advice, feedback and interpretation of my results has been a priceless contribution to shaping my thinking and approach to problems in the world of transcriptional enhancers and beyond. I hope that in my professional career I may be able to count on your feedback. I admire your dedication to training others and sharing your collectively acquired knowledge.

I, like many others at Caltech, have had the pleasure of working with Georgi Marinov and owe him many thanks. I hope that we continue to cross paths and I enjoy the times that we do get to spend together.

I would also like to thank Brian Williams and Sreeram Balasubramanian for your time and dedication to us graduate students who are fortunate to have you as friends, and colleagues in the lab. Explaining your contributions to the Wold lab, succinctly, would require writing an additional document approximately the length of this thesis, but I do hope we continue to work together over the coming years.

I would like to thank Tony Kirilusha. I look forward to being a friend and colleague. I also look forward to working with you over the summer on a joint manuscript. I would like to thank all of the other members of the lab that I have had the fortune of being able to collaborate with including Diane Trout, Henry Amrhein and Sean Upchurch. Finally, within the lab I would like to thank Anna Abelin for her feedback and help and I hope to cross paths with you again!

I would like to thank and acknowledge Jost Vielmetter and the Protein Expression Center for allowing me to use their robotics for my experiments. Jost, your advice has been priceless over the years and I'm glad we have developed a friendship that also spans outside of work. I look forward to keeping in touch with you!

I would also like to thank Sandra Sharp for providing some of the CRMs that I used to tune my assay. Sandy thank you for your feedback and insight over the years!

I would like to thank Christopher Partridge and Richard Myers for sharing their data with me and allowing me the opportunity to analyze it and have their feedback and support on transfection assays and the biochemical data that we used to derive and interpret them.

I would like to thank Nergiz Dogan and Ross Hardison for sharing their transfection and biochemical measurements allowing me to draw comparisons with another system. I was also fortunate to be able to access STARR-seq data produced in the Reddy lab by Christopher Vockley.

I would like to thank Elizabeth Bertani for allowing me to shape an undergraduate class into a unique teaching experience for Matthew Smalley that resulted in a small publication.

I would like to thank Shirley Pepke, Santiago Lombeyda and Camden Jensen for their help on the Self Organizing Maps.

On a personal note, I would also like to thank all of my friends for their unending support but especially Ian, Andy, Chris, Tony, and Darlena. Your friendships are important to me and I cannot begin to explain just how much I have come to value each of you.

Finally I would like to thank my parents for the unwavering support they have given me over the years.

I hope I can one day pay back each and every one of you as without the individual contributions of each of you, and many others, I doubt I would be where I am fortunate to be today.

ABSTRACT

Functional genomics aspires to explain how a transcription factor (TF) and its measured biochemical occupancy relates to the enhancer activity of the underlying sequence elements. Tissue-specific TFs exhibit remarkable selectivity and reproducibility in the available genome-wide sequence motifs accessed. A consistent central conclusion is that, irrespective of the element selection criteria used, ~50% of candidate Enhancers score as transcriptionally active in both mouse and human cell types, while the remaining 50% of similarly biochemically marked regions are unable to activate transcription on their own. This finding is based on an integrated comparison of a group of functionally assayed elements containing TF-occupied elements, evolutionarily conserved elements, and TF agnostic elements with hallmark biochemical signatures of known enhancers. Quantitatively, the level of TF occupancy signal was the best predictor of the proportion of active enhancers detected, but overall (and contrary to expectation) it is a weak predictor of the magnitude of enhancer activity readout. In specific cell types, elements can display all of the hallmark signatures of enhancers, but can remain inactively poised prior to a stimulus that either activates them or releases a repressive factor. Against previous expectations these poised occupancy sites, once released, behave comparatively in magnitude of enhancer activity as their counterparts that are only directly accessed upon stimulation. Based on our findings, the vast majority of active enhancers in the genome, including some of the most individually powerful ones, are expected to display relatively modest biochemical signatures. Finally, the combined set of over a hundred genomic regions that lacked biochemical marks, even while containing the motifs known to be necessary to bind the relevant TFs, did not support significant enhancer function. We also found evidence that both enhancer orientation and combinations of relatively closely spaced candidate Enhancers, can yield

additive functions, with possible fine tuning of the enhancer activity controlled by the type and the distance between individually accessed motifs. In special cases, these elements might cooperate to recruit stable complexes resulting in a synergistic transcriptional activation, suggesting that both local "super-enhancers" and recruited multi-element combinatorics are likely to play an important role *in vivo*. These findings provide an expectation for enhancer function in the comprehensive annotations provided by the new ENCODE encyclopedia and may help guide future efforts to define the mechanisms by which enhancer activity is achieved and conferred selectively to target genes. Surprisingly, elements that deeply sample the biochemical occupancy of complex loci, match a random population of selected elements remarkably well. Our findings also indicate that carefully designed and lower throughput approaches, rather than high numerical assays that focus on the outstanding features, will bring widely applicable answers to the remaining questions of how relative enhancers are tuned and how seemingly identical regions at a biochemical and motif level are selected for or against function.

PUBLISHED CONTENT AND CONTRIBUTIONS

DeSalvo, G, Marinov GK, Partridge P, Vockley CM, and et al. (2018). "Pursuing function in candidate Enhancer signatures". In: *To be submitted*.
GD participated in the conception of the project, developed and produced data on a robotic enhancer assay across four cell types, analyzed the enhancer data in the context of biochemical measurements in my own system and across the multiple systems presented, and participated in writing the draft manuscript in Chapter 3.

MD, Smalley, Marinov GK, Bertani LE, and DeSalvo G. (2015). "Genome Sequence of Magnetospirillum magnetotacticum Strain MS-1. Genome Announcements." In: *Genome Announcements* 3.2, pp. 00233–15. DOI: 10.1128/genomeA.00233-15.
GD conceptualized this project for a student in a lab class and oversaw the project from inception to publication.

SG, Landt, Marinov GK, Kundaje A, DeSalvo G, and et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." In: *Genome Research* 22.9, pp. 1813–1831. DOI: 10.1101/gr.136184.111.
GD created Figure 4B to annotate ChIP-seq regions for previously characterized enhancers in the the context of muscle.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

1.1 Thesis Outline

Chapter 1 provides a review of the key components currently understood to regulate the transcription of genes and sets the context necessary to interpret the findings presented in this thesis.

Chapter 2 outlines the candidate Enhancer selection, development and limits of my enhancer assay.

Chapter 3 presents my draft manuscript outlining the combined findings of our collaboration by contrasting enhancer activity proportion and magnitude from candidate regions selected in each cell types.

Chapter 4 pertains to the relative activities of individual candidate Enhancer elements within complex loci.

Chapter 5 proposes models and key experiments for future developments.

1.2 Gene Regulatory Networks

As developmental biologists, we aspire to obtain a complete understanding of the process that decodes the genomic blueprint using the maternal inputs found in the oocyte, to produce the hundreds of cell types needed in an adult. During this process, each cell has to juggle many concurrent regulatory programs to support the establishment and maintenance of cell type-specific functions (Huang L, 2000; Ma JY, 2013; HoughEvans BR, 1977; Wang DG, 1995) while carrying on the common "housekeeping" (ie. metabolic and homeostatic) functions shared by most cell types. (Eisenberg E, 2003) Both the housekeeping and cell type-specific

functions are regulated by gene regulatory networks (GRNs) orchestrated by the sum of many ubiquitous and selectively expressed regulatory proteins known as Transcription Factors (TFs). (Davidson EH, 2005) These, in turn, occupy cis Regulatory Elements (cREs) — stretches of DNA containing the necessary TF binding sequences — in response to either internal feedback loops or stimuli external to the cell. (Longabaugh WJ, 2005; Yee SP, 1993; Cameron R, 1997) In order to understand the drivers behind this process, many labs, including several here at Caltech, began studying and dissecting these functional elements of the genome. Yet despite the combined knowledge of numerous examples of functional regions, the motifs bound by individual factors, the TFs expressed in each system, and the regions of the genome occupied by each TF, we are still unable to predict the functional logic of individual cis-regulatory elements that are used to guide the developmental process from their DNA sequences.

1.3 DNA sequence mediating TF occupancy

Transcription factors (TFs) are proteins that are able to bind DNA directly in a sequence specific manner to decode to the transcriptomic instructions contained in the genome based on the cellular context. This property distinguishes them from transcriptional co-factor proteins, which define the context, and are recruited to the DNA through protein-protein interactions with the TFs. EMSA (Electrophoretic Mobility Shift Assay) and Selex (multiple rounds of oligonucleotide selection and TF binding) studies allowed for the identification of the DNA sequence bound by individual TFs. For example, MyoD, a TF able to transdifferentiate fibroblasts into muscle can bind to the CAGSTG motif called an Ebox. (Davis RL, 1987; Blackwell TK, 1990; Kabadi AM, 2015; Wright WE, 1991) The mapping of TF binding sites to a sequenced genome might seem like an easy step once the recognition sequences are known. (Kophengnavong T, 2000) Unfortunately, the genome

provides huge numbers of such motifs and identifying active cREs used to control these transcriptomic processes proved to be a challenge. (Cirillo LA, 2002) For example, in the mouse genome, there are over 2.1 million instances of the Ebox motif alone, which is several orders of magnitude more than the number of genes. This numerical imbalance raises questions as to which of the available motifs are biochemically accessed and how much of the genome supports biological regulatory function. (Kirilusha A, 2014 thesis)

1.4 Identifying candidate Regulatory Elements

A goal for the field has thus been to identify, map and characterize all regulatory DNA sequences in the genome, sometimes called the cis-regulome. This effort has used a combination of biochemical correlates for regulatory activities, direct functional assays and mutagenesis perturbations of cis-elements to identify and define regulatory elements. The scale of this effort has increased from discovery and dissection of single genes and their associated regulatory elements to larger locus-level studies and finally to the whole-genome scale.

Mutagenesis screens and random integration events were first used to identify DNA elements that result in phenotypical variations in Drosophila. (OKane CJ, 1987) Reviewed in: (Rossant J, 1992) However these methods are inefficient when applied to large mammalian genomes. (Weihner H, 1984; Queen C, 1984) Long contiguous pieces of DNA (eventually large BACs) from upstream of key genes were tested for their ability to support a given temporal/spatial expression pattern during development. (Mautner J, 1995; Springer PS, 2000; Carvajal JJ, 2001) These assays can also support deletion analysis of smaller target regions within the test DNA to identify local sequence features that support a given expression pattern. For example, the myf5 gene is driven with different temporal dynamics in different spatial domains of the developing mouse embryo by multiple, separable, large elements in

a 100kb upstream segment of DNA within a yeast artificial chromosome (YAC) (Hadchouel J, 2000). The 14kb region immediately spanning the Myf5/MRF4 locus, which controls the early epaxial expression of myf5, can be further subdivided and contains individual enhancers with separable functions (Summerbell D, 2000; Teboul L, 2002). Separability of functional elements is also true for muscle-specific control elements located in a region ~50kb upstream of myf5 that are required for later expression. (Hadchouel J, 2003) However, these designs are also numerically limited by both time and expense. (Carvajal JJ, 2001; Johnson JE, 1989; Jaynes JB, 1988; Hadchouel J, 2003) Deletions in transgenic BACs that result in loss of function also have limited interpretability, because they cannot identify whether a loss of function is due to removal of one or more positive-acting cREs within the deletion. Conversely, if a gain-of-function is detected upon making a deletion, it cannot be known whether the loss of a negative acting element (insulators or silencers) normally contributes to the functional output of the region. Although transgenic BAC methods support the idea that functional elements are separable and can be assayed out of their normal genomic context, they raise the possibility that motifs which function in the normal genomic context may not individually function in an experiment, unless they are provided the context of other control elements. In order to increase the detection of likely candidate enhancers that can support tissue specific expression patterns, targeted mutations or DNAse hypersensitivity mappings in representative cell lines could be used to identify the sequences most likely responsible for the identified function of a larger fragment of DNA.

At the time biochemical measurements over such vast regions as a BAC (and much less genome-wide) were not feasible but the use of evolutionary sequence conservation permitted to identify candidate Regulatory Elements on a local scale by comparing the upstream sequences across species of a few genes identified by genetic screens such as myoglobin. (Blanchetot A, 1986) The eventual sequencing of

multiple genomes allowed for large scale comparisons and the identification of candidate enhancers in a way that facilitated detection of regulatory elements positioned at more distal locations from genes. For example, the olfactory receptor enhancer H region was first identified by a large deletion that was then narrowed using sequence conservation between mouse and human to a smaller distal 2kb region for further study. (Serizawa S, 2003)

These candidate cis-regulatory elements identified using sequence conservation were again characterized for function using the combined power of CAT reporters (function test), (Weintraub H, 1989; Weintraub H, 1990; Weintraub H, 1991) LacZ based embryo reporter assays (function, localization and temporal pattern) and cis-element knock-out experiments (necessity). (Gossett LA, 1989; Fickett JW, 2000; Hardison R, 1997a; Hardison R, 1997b; Brown CT, 2002; Gottgens B, 2000; Clayton CE, 1982; Chandler KJ, 2009; Teboul L, 2002; Rosenthal N, 1990; Hockheimer A, 2003; Levine M, 2003; Morokuma J, 2003) Once identified and tested these enhancers led to the identification of the key characteristics of genomic elements with discernible functions. Enhancers or Silencers were functionally defined genomic elements that over short distances or through physical connections (sometimes guided by insulator/blocker elements) to bridge more distant ones, modulate the transcriptional output provided by the basal promoter. (Long L, 2013; KiefferKwon KR, 2013; Dekker J, 2002; Banerji J, 1981; Butler JE, 2001; Yuh CH, 1996)

1.5 Defining the biological functions of Regulatory Element

The basal promoter contains a local collection of regulatory elements that will only work in one orientation, guided by TF binding to recruit the machinery necessary to enable the production of a transcript from a transcription start site (TSS). (Sartorelli V, 1999; Mikhaylichenko O, 2018; Levine M, 2003; Andres ME, 1999) The TSS proximal region can, but does not always, contain additional regulatory elements

beyond what is minimally necessary to enable the start of transcription. (Goldhamer Dj, 1995) For example, individual promoter regions can contain DNA elements which provide some level of enhancement or repression either interdigitated within the basal promoter sequence itself or immediately proximal. Indeed this region, especially for genes that must work across different cell types, needs to support occupancy of different sequence specific TFs and biological function in a relatively compact region, putting it under an unusual evolutionary constraint when compared to the average gene distal region. Since the TSS of each gene can be identified from the highly conserved exons, the relative functional output of similarly sized promoter regions was compared across hundreds of genes showing that promoters have varied levels of activity conferred by cis-elements localized within the promoter proximal region. (Cooper SJ, 2006) The occupancy of TF binding sites thus acts as a regulatory gate to transcriptional regulatory functions, including fine tuning by the combinations of off-DNA cofactors which are able to recruit other complexes and by the modification of histones or modification of the proteins themselves. (Sartorelli V, 1999) These TF controlled DNA elements can modulate both spatial and temporal gene expression during development of the mouse embryo. (Gutman A, 1994; Hadchouel J, 2003; Carvajal JJ, 2001; Peter IS, 2011) It has since become clear that transcription can be further modulated by the integrated contributions of distal elements, DNA methylation and regulation of transcriptional pausing. (Li B, 2007; Kouzarides T, 2007; Barski A, 2007; Taft RJ, 2009; Core LJ, 2008; Rahl A, 2010; Schwartz YB, 2007; Schor IE, 2017; Cheng TS, 1993) While this definition, which combines positional and experimentally determined criteria seems clear, there turns out to be considerable room for confusion because the underlying biochemical mechanisms and their relationships to DNA sequence features are still not fully understood.

Indeed some “promoter proximal” regulatory elements described above can also

meet the experimental definition of “enhancer”, while others fall outside the operational definition of “enhancer”. Unlike promoters, one of the generally accepted characteristics of an enhancer is that it is able to function in both orientations and with independence of position relative to its TSS. Exceptions to this principle do exist as some intronic enhancers are position-sensitive relative to the promoter in specific cell contexts. (Banerji J, 1981; Tai PWL, 2011; Zomorodipour A, 2017) The logic encoded in DNA and accessed by transcription factors is largely modular and can be transported (at least in some specific cases) from one gene to another or even to an exogenous test system. (Yuh CH, 1996; Kirchhamer CV, 1996) Reviewed in Fickett JW, 2000. Some of these enhancer elements are relatively compact while others can require individual cREs to be integrated over a larger region for the correct interpretation of the functional output. The varying size of these elements can make identifying the complete logic for a gene difficult as external elements might provide additional inputs over great distances. (Long L, 2013; KiefferKwon KR, 2013) In all of these instances to decode these instructions transcription factors act as the master regulators responsible for recruiting co-factors (of both activating and repressive natures), non-coding RNAs (to modulate a variety of functions including recruiting and tethering complexes), enzymes responsible for histone marking and modification of the TF/cofactors themselves. (Katayama S, 2005; He XJ, 2009; Faghihi MA, 2009; Aravin AA, 2008; Bartel DP, 2009) While much of the focus of this thesis is on enhancer elements, in order to maintain functional genomic DNA domains which are accessible in a cell type, insulator/anchor elements mark chromatin boundaries and function as anchors for bridging certain DNA domains in physical proximity to each other. (Barkess G, 2012; Fudenberg G, 2018; West A, 2002) For example these elements are used to either exclude (through looping) or allow for the interaction of an enhancer with a nearby gene. (KiefferKwon KR, 2013) They can also form chromatin boundaries that separate (“insulate”) inaccessible heterochromatin

domains from relatively more accessible regions. (Reviewed in Laat W, 2013) The functional orientation requirement of chromatin insulator elements, which when changed in orientation can result in the rewiring of chromatin-chromatin contacts, raises questions as to whether and how this observation can also apply to select target genes for enhancer and silencer elements. (Ghirlando R, 2016) (Further discussed in section 1.13) In contrast to enhancers, silencer elements bind repressors (ie Oct-1, NRSF) and dampen the transcriptional level of a gene, either by binding at the promoter proximal region directly or also through looping a distal silencer element to a promoter. (Lawinger P, 2000; Hwang SS, 2016; Becker NA, 2013) Silencer elements have been observed to have similar positional dependence to that more recently observed in some enhancer elements; where changing relative position to a promoter affects the biological activity of an element. (Chow KL, 1990; Edwards JG, 1992; MouraNeto V, 1996; Bessis A, 1997; Cheng CK, 2002; Kim MK, 1996)

1.6 Developmental model systems

The differentiation of a precursor into a mature cell type has long been used to characterize the key functions of factors that bind regulatory elements, as it provides a differential where biochemical measurements can be contrasted to a prior state. As an example, during myogenesis, undifferentiated specified myoblast cells become differentiated into myocyte muscle cells. This process is primarily regulated by four key basic Helix Loop Helix (BHLH) TFs known as Myogenic Regulatory Factors (MRFs) which heterodimerize with E-proteins to bind DNA and function along with numerous other TFs and cofactors, such as MEF2, Pbx1 and co-activators (EP300/pCAF). (Taylor SM, 1979; Davis RL, 1987; Hasty P, 1993; Wright WE, 1993; Buckingham M, 2014; Hu P, 2008; Hernandez JM, 2017; Neuhold LA, 1993; Fong AP, 2015; Conerly ML, 2016; Cao Y, 2010; Weintraub H, 1991).

The mouse C2C12 cell line (Yaffe D, 1977) (derived from normal mouse thigh

muscle following a crushing injury) provides the main cell culture model system for studying this transition from myoblast to myocyte, which can be experimentally controlled by the researcher by changing the composition of the culture medium. (Figure 1.1) In C2C12s, the key specification MRF detected at high levels in myoblasts is MyoD while myogenin is the predominant differentiation TF expressed in myocytes. (Figure 1.1) (Davis RL, 1987; Cornelison DD, 2000) The remaining two MRFs are Myf5 (early specification in the embryo) and Myf6 (late differentiation) and are quantitatively very minor in the C2C12 system. (Braun T, 1989; Miner JH, 1990) MyoD was originally isolated for its ability to convert different cell types *in vitro* to the myoblast lineage (Davis RL, 1987) including the 10T1/2 cell line, which normally provides a mesodermal precursor that normally does not express MRFs and does not normally differentiate into muscle. (Mordan LJ, 1982) Since MyoD and myogenin are not expressed in 10T1/2 cells, this is a useful experimental system to test for specificity of enhancer activity where neither of the two key MRFs present in C2C12s are expressed. (Figure 1.1) This cell line also allows to identify cREs which respond to serum starvation and exposure to insulin, both of which are used as a stimulus to induce muscle differentiation of C2C12 cells and are conditions known to induce changes in gene expression. (Keeton AB, 2002; Pirkmajer S, 2011)

During early myogenesis, the expression of myogenin and several other key genes is kept off in myoblasts even though MyoD is present. This has been hypothesized to be controlled by Snai1 which is thought to repress key genes by occupying the same Ebox motifs preferred by myogenin in myoblasts, until Snai1 is later targeted by an inhibitory microRNA, releasing the repression of terminal differentiation genes. (Soleimani VD, 2012) Zeb1 has also been found to bind preferentially to the promoter of terminal differentiation genes, potentially working as a repressor that is released through a yet to be understood mechanism. (Siles L, 2013) Musculin is a third repressor expressed during myogenesis. In muscle, musculin was found to

compete for available E2A proteins (which are a positive regulator of HAT activity in select cell types) indicating that the mechanism through which musculin functions as a sequence specific repressor might be context dependent. (Kumar D, 2007; Hyndman BD, 2012; Wu C, 2017; Yuh L, 2003; Yuh L, 2004)

Using modern RNA-seq measurements, we showed that 8704 genes are detected at a level of 10 FPKMs or more in C2C12s myocytes. (Kirilusha A, 2014 thesis) Of these 8704 genes, as a direct result of myogenin expression and its downstream targets (including Mef2C), a small but important set of 628 genes are transcriptionally up-regulated at least 3-fold compared to the myoblast state. For at least for a subset of these genes the "classic" form of the muscle enhancer is responsible for generating this response, including the enhancer 1kb upstream of the CKM TSS from which the combination of cis-acting sequences that bind MEF2 (CTAAAAATAG) and MRFs (RRCAGSTG – the Ebox) motifs was derived. (Amacher SL, 1993; Andres V, 1995) In contrast, a set of 824 genes are downregulated including several antagonists to the process (IDs, MSC, snail, msx1) by unknown regulatory mechanisms upon the expression of myogenin. (Taylor SM, 1979; Davis RL, 1987; Hasty P, 1993; Wright WE, 1991; Buckingham M, 2014; Hu P, 2008; Hernandez JM, 2017; Neuhold LA, 1993; Fong AP, 2015; Conerly ML, 2016; Cao Y, 2010; Weintraub H, 1991) The ID family of proteins are also involved in repression of the myogenic progression by sequestering the MRFs from their BHLH partner proteins E2 and HEB to which they have the greatest affinity. (Langlands K, 1997; Liu CJ, 2002) The ID family proteins also bind MYOD and Myf5, but not significantly to myogenin, thus providing a way to circumvent this block once myogenin molecules are present in effective numbers. (Langlands K, 1997; Liu CJ, 2002)

Hematopoiesis can be used as a second developmental process involving a distinct set of elements also derived from tissue specific factors to contrast or validate any findings from myogenic ccREs. Mouse G1E cells have served as the key

model system for erythropoiesis for many years. (Weiss MJ, 1997) These cells are derived from *in vitro* differentiated mouse embryonic stem cells in which the endogenous GATA1 gene has been knocked out; a subclone of them, termed G1E-ER4, constitutively expresses a GATA1-ER fusion that can be specifically activated by estradiol exposure, allowing for differentiation to be triggered rapidly and in a controlled manner. (Tsang AP, 1997; Rylski M, 2003) Lineage commitment during the process of erythropoiesis is accomplished through the so called GATA switch. (Kaneko H, 2010) The onset of terminal differentiation is marked by the replacement of the GATA2 transcription factor pioneering for occupancy at thousands of sites occupied by GATA1, which then regulate the expression of the genes involved in red blood cell development and functions, with the bHLH protein SCL/TAL1 being an important cofactor of GATA1, often forming closely spaced heterodimers with it. (Han GC, 2015; Wu W, 2014; Tripic T, 2009; Yu M, 2009) GATA1 binds to a WGATAA consensus recognition motif, while TAL1 is a bHLH factor targeting a CAGMTG Ebox. (Han GC, 2015)

1.7 ChIP-seq maps global occupancy profiles

ChIP-seq data provides high resolution genome-scale occupancy maps of direct DNA interacting factors. (Johnson DS, 2007) With small technical adjustments, this approach can also map general coactivators such as EP300 to identify candidate active enhancers. This quite general biochemical approach is independent of sequence conservation or motif content and identifies candidate enhancers missed by those approaches. (Blow MJ, 2010; Hong JW, 2008; Levine M, 2003) For example, these new measurements effectively identify candidate enhancers in tissues such as heart, where phylogenic conservation failed to be effective, as well as a novel class of candidate enhancer elements found within otherwise highly conserved regions of genes. (Blow MJ, 2010; Ponting CP, 2011; Hong JW, 2008; Berthelot C, 2018;

Wong ES, 2015) Taken together, these non-coding candidate regulatory elements, highlighted by preferential sequence conservation and/or by various biochemical signatures, possess the characteristics correlated with regulatory function. (Sections 1.8 and 1.10) This collection of elements derived from these biochemical measurements across many tissues number in the millions but it remains uncertain how many can, in fact, function as enhancers when tested experimentally. (Yue F, 2014; ENCODE, 2012) In addition, some apparently exceptional systems suggest there are enhancer elements with previously undefined properties. For example, the H enhancer element in the olfactory receptor system, together with a small set of similarly biochemically marked function specific enhancers in the genome, feature highly selective action aimed at just one among hundreds of OR genes, selected on a cell by cell basis, and operating efficiently across chromosomes. (Markenscoff E, 2014) However, only a subset of these identified enhancers are in physical contact with the active gene; raising questions about whether specific enhancers are targeted to pre-established genes. These experiments also suggests that elements can display the biochemical hallmarks of enhancers, function as enhancers in a cell type, but need not display physical connectivity to a TSS in all cases while awaiting heterochromatin formation over the enhancer element. (Lyons DB, 2013; Magklara A, 2011)

In order to help understand part of the control mechanism behind the process of myogenic differentiation in culture, a fellow lab member measured the occupancy for MyoD (specification factor) and myogenin (differentiation factor) in C2C12 cells. (Kirilusha A, 2014 thesis) Of the occupied sites, 40% of elements are proximal to gene models while the remaining 60% need to be physically brought to distant promoters. The MRF ChIP-seq occupied regions can be sub-selected for ones that contain conserved elements, and our lab showed that they contain ~400bp long conserved elements with the Ebox enriched towards the center of each region.

(Pepke S, 2009) Overall the expected Ebox motif is found in >90% of MyoD or myogenin occupancy sites. However, the p=0.05 significantly bound 14,088 myogenin sites determined by ChIP-Seq occupy only ~0.5% of muscle class Ebox motifs (CAGSTG) in the mouse genome. (Kirilusha A, 2014 thesis) Both factors occupy a highly overlapped set of sites with the majority containing an extended RRCAGSTG form of the muscle Ebox. Moncaut N, 2013; Tapscott SJ, 2005; Molkentin JD, 1996; Yun K, 1996; Valdez MR, 2000 Many possible combinations of flanking nucleotides to the Ebox are detected within the ChIP-seq measurements and the actual motif preference is likely dictated by the individual combinations of the MRFs and the heterodimeric partners. (Yin Y, 2017; Yun K, 1996; Neuhold LA, 1993; Lu Y, 2002) However, over-expression of MyoD in murine embryonic fibroblasts (MEF) shows ChIPSeq occupancy that highly overlaps the native MyoD occupancy sites indicating the possibility that additional identical motif sites may be kept inaccessible at the chromatin level. (Yao Z, 2013) This selectivity of which motifs are being accessed in the genome also raises questions as to the minimum number of motifs required to support biological function such as an enhancer.

1.8 Mapping ChIP-seq measured occupancy to validated enhancer regions

The classic 2R-TK-CAT experiment in muscle demonstrated that 2 copies of the right Ebox from the CKM -1kb enhancer are sufficient for enhancer function. (Weintraub H, 1990; Weintraub H, 1991) However, 4 copies of the motif were necessary to recapitulate the native enhancer activity, which contains 2 Eboxes and a Mef2 motif. (Weintraub H, 1990; Weintraub H, 1991) Indeed the combination of motif types, numbers and distances are likely the key driver by which the level of enhancer activity is encoded within in the more compact promoter proximal regions to provide a local logic that can be further modified by the contribution of distal elements. (Istrail S, 2005; Vockley CM, 2016) The myogenin ChIP-seq regions contain from

single to multiple Eboxes. (Kirilusha A, 2014 thesis) A significant portion of the myogenin occupancy regions (over 25%) contain two Eboxes and are most similar to the 2R-TK-CAT construct. The ChIP-seq regions with a single Ebox sometimes contain a paired Mef2 and Ebox motifs, suggesting an initial test of the relative enhancer contribution for each class of elements (Ebox/Ebox and Ebox/Mef2) paired at varying distances. (Kirilusha A, in drafting) While many combinations of sites exist within the occupancy these two simple examples might set the ground rules for enhancer function. (Vockley CM, 2016) As previously discussed, ChIP-seq experiments also revealed that many potentially important sites and even entire new classes of sites had not been studied. (Levine M, 2003; Barolo S, 2013; Hong JW, 2008) In the context of muscle these included candidate Enhancers that differentially bind MyoD across differentiation, multiple factor binding sites and non-conserved candidate Enhancers.

However, the biochemical measurements done by ChIP-seq only denote the occupancy by a factor, and can result from the chemical bridge to another protein depending on the amount, temperature and type of fixative agent(s) used. (Vockley CM, 2016) Our current methods for TF occupancy mapping allow for only a limited understanding of the signaling context due to the poor availability of antibodies that reliably detect such modifications. Even so, the highly selective occupancy relative to the number of motifs available in the genome (14K accessed vs 2.1M available) displayed by the sequence specific TF myogenin (Kirilusha A, 2014 thesis) raised key questions as to whether such specificity in the biochemical occupancy always results in biological function (indicating that ChIP-seq is perhaps under sensitive) or that conversely much of ChIP-seq is noise in the lower ranked signals.

Unfortunately, initial functional studies have often focused only on the most outstanding biochemical signatures (Visel A, 2007), thus obtaining deceptively high success rates that some have generalized as expectations for all reproducible sites

of occupancy. For example, a study focused on some of the top signals from among the first available but limited quality EP300 occupancy measurements in limb and brain tissue to find that 80% of the conserved and occupied sites are active enhancers versus 40% by conservation alone. (Visel A, 2007) Several follow-up studies confirmed that absence of conservation within a biochemically occupied region does not imply non-functionality. (Visel A, 2009; Berthelot C, 2018; Wong ES, 2015; Ballester B, 2014) This 80% estimate of elements functionally validated is likely an upper bound due to the quality of ChIP occupancy data and the way candidates were selected. (Landt SG, 2012) Within the context of muscle MyoD is known to interact with EP300 and recruits this cofactor together with Mef2 to cREs. This should allow for potential enhancer sensitivity comparisons, albeit limited to the top signal EP300 sites, between ones selected from within the occupancy an individual tissue specific factor compared to this "TF agnostic" EP300 based selection. (Chapters 2 and 3) (Eckner R, 1996; Yuan W, 1996; Puri PL, 1997; Sartorelli V, 1997)

A collection of 24 previously characterized "gold standard" enhancer regions derived from the literature (Appendix E) contain at least one core recognition motif for MyoD/Myogenin (CAGSTG) and show significant MRF occupancy signals (Figure 1.2), though the magnitude of the occupancy signals varied by more than an order of magnitude (Figure 1.3). The histogram summary of these validated enhancers adapted from Landt SG, 2012 clearly illustrates that these functional elements map throughout the range of ChIP-seq signal (Figure 1.3). Nevertheless, the manner in which these elements were selected and their relatively small number mean that one cannot make a statistical case for the power of ChIP to predict enhancer activity or its magnitude. Because of discovery bias for these long-studied enhancers most are promoter proximal elements associated with genes highly transcribed in our muscle cell model system. As a result, there are multiple reasons why they do not accurately represent the myogenin genomic occupancy repertoire as a whole,

which includes 14088 sites; many of which are found near genes of more diverse behaviors, including ones that are not expressed. (Kirilusha A, 2014 thesis) I also selected another set of 24 enhancers from the literature which are affiliated with genes specifically expressed in T-cell and neuron development. These literature validated "non muscle" enhancer regions were selected to contain muscle MRF-class Ebox motifs which are shared by paralogs of the Tal or NeuroD families responsible for the development of those cell types. In muscle these genes are not transcribed, and the validated enhancer regions show no occupancy in our ChIP-seq measurements for the muscle specific factors or non-tissue specific co-activators such as EP300. (Figure 1.2) In contrast NRSF occupancy is detected at comparable rates in both populations of muscle and non-muscle enhancers (Figure 1.2). When I began this work, it was unclear whether this tissue specific occupancy (and function) would be retained in a transient transfection assay where naked DNA is introduced (and at high copy number) without any prior epigenetic history. Although the specificity of enhancer function was largely maintained in our assay for biochemically marked regions (Chapter 3) sufficient flanking DNA context might be necessary. (further discussed in 1.15)

1.9 Relating ChIP-seq observations to physical occupancy

In vivo molecular footprinting experiments (which quantify the physical occupancy) of the well-characterized CKM -1kb enhancer demonstrate that the reproducibly modest but statistically significant ChIP-seq signal detected in this cRE does not correlate with the physical occupancy (over 90%). (Mueller PR, 1989; Garrity PA, 1992; Buskin JN, 1989; Horlick RA, 1990; Cserjesi P, 1994; Horlick RA, 1989) This disparity between ChIP signal strength, occupancy level and enhancer function level could be explained if the stably bound and active MRF-recruited complex involving multiple proteins provides poorer physical access to the ChIP antibody. These stable

complexes at enhancers might however be necessary to provide the transcriptional bursting frequency observed for strongly transcribed genes. (Fujita K, 2016) The CKM enhancer observation, together with the weak myogenin occupancy signal observed for a small set of literature-validated enhancers (Figure 1.3), also raises the question of whether ChIP-seq, for both a TF such as myogenin or a cofactor such as EP300 are equally sensitive for enhancers across different signal strengths.

1.10 Derivation of large scale biochemical signatures for cREs

The genome-wide mapping of candidate REs has been a major activity of the ENCyclopedia Of DNA Elements (ENCODE) consortium, including our contributions for skeletal muscle and C2C12 cells. (ENCODE, 2011; Yue F, 2014; Moore J, 2018) To this end, measurements of DNAse hypersensitivity, multiple histone modifications, hundreds of sequence-specific transcription factors, and RNA transcripts across many cell types and tissues were performed to provide an increasingly comprehensive biochemical annotation of the human and mouse genomes. (ENCODE, 2011; Yue F, 2014; ENCODE, 2012; Moore J, 2018) Previously characterized cis-regulatory elements (such as the ones in Appendix E) provided a framework that is substantially supported by contemporary genome-wide biochemical data, with many findings being "rediscovered" in the literature over the last decade. Indeed, interpreting the genome-scale biochemical data in light of previously characterized regulatory elements was the first step in creating a dictionary of benchmark biochemical signatures that could be extrapolated to other similarly indexed regions. (Schor IE, 2017; Visel A, 2007)

As discussed in section 1.5; promoters, enhancers, and insulators are sequence specific elements that are each associated with specific molecular functions in the control of transcriptional processes, and when active these elements are each marked by distinctive biochemical signatures. (Riethoven JJ, 2010) For example, active pro-

moters in eukaryotes feature the trimethylation of lysine 4 on histone 3 (H3K4me3) (Vermeulen M, 2010) and the H3K27ac histone mark. These regions often display several ChIPSeq signals for TFs and polymerase. (Thurman RE, 2012; Visel A, 2009) Enhancer elements exhibit their own biochemical signature where sequence specific transcription factors recruit EP300, and as a result of a yet to be defined biological activity also feature the H3K27ac histone mark deposited on nearby histones. (Figure 3.1A) (Creyghton MP, 2010; RadaIglesias A, 2011; Thurman RE, 2012) Transcription factor occupancy also allows for increased accessibility of chromatin organized DNA by DNAses, allowing for the TF agnostic detection of the union set of occupied regions in individual cell types. The correlation of significant DNAse HyperSensitivity (DHS) with TF occupancy and the histone marks that result from this biological function are strong when all data are of high quality. However, the appearance of histone marks can sometimes lag active regulatory changes in which regulatory elements and promoters can become active and DHS sensitive before histone marks appear, or conversely, they can be shut down before the marks disappear or change. While the H3K4me1 mark had been proposed, with initial enthusiasm, as a landmark signature for active enhancers, we and others are finding that it marks only a subset of activating TF occupied elements, and is also found at many promoters. (Cheng J, 2014)

Conversely, the presence of PolII and the basal transcriptional machinery is now also well documented at distal enhancers, and at least some active enhancers produce enhancer RNAs (eRNAs). However, the transcripts produced are non-coding, often short, and apparently short-lived. This phenomenon however blurs a previously clarifying biochemical distinction between promoters and enhancer functions in the face of histone mark and TF occupancy overlaps. (Koch F, 2011) At least in Drosophila, it was shown that the level of these eRNAs correlate positively with enhancer activity, that some intergenic enhancers can work as weak promoters, and that conversely

alternative gene promoters can work as weak developmental enhancers. A corollary is that cell type and temporal transcription of eRNAs is largely specific to active enhancers, though their detection requires the use of methods that do not depend on polyadenylation and provide relatively high sensitivity. (Mikhaylichenko O, 2018) In mammalian cells, however, there are reports of active enhancers that do not produce detectable eRNA, indicating that this property is not obligatory, making eRNA a limited positive predictor of enhancer activity with yet to be determined sensitivity. (Andersson R, 2014) Moreover, promoters can be "stalled" and show very similar characteristics with POLII occupancy, H3K27ac and sometimes detectable aborted transcripts. (Adelman K, 2005; Wu JQ, 2008)

In the context of muscle, MyoD and Myogenin show both H3K27 histone acetylation (H3K27Ac) and EP300 occupancy at 50-60% of their respective occupancy sites for both myoblasts and myocytes. MyoD has been shown to interact with Histone Acetyl Transferases (HATs) EP300 and pCAF/KAT2b (which are thought to have partially redundant functions) as well as a wide variety of other cofactors. (DelaSerna A, 2005; Berkes CA, 2005) There is at least one more significantly expressed HAT in our system (KAT2A) which is expressed in both myoblasts and myocytes at a level comparable to EP300. While we do not find EP300 at all elements demarked by (H3K27ac) and myogenin or MyoD occupancy this could be explained by the presence of these other HATs which we were not able to ChIP due to the lack of a ChIP-grade antibody. Interestingly, E2A (the key MRF heterodimeric partner) has also been suggested as a positive regulator of Histone AcetylTransferase (HAT) activity, providing a possible mechanism through which ID family proteins can dampen the onset of differentiation. (Soleimani VD, 2012; Hyndman BD, 2012)

As expected from previous studies reporting a lagging histone mark pattern, we also find significant H3K27ac present in DHS negative regions, especially in developmental systems. (Zhang JA, 2012) The converse mechanism where an enhancer

might be primed for function at the biochemical level but unable to function until it is activated by occupancy by specific factor is also observed. For example, although MyoD and Myf5, the two specification factors, occupy a largely overlapping set of sites when both are expressed, displaying the same hallmarks of enhancers including occupancy and H3K27ac there is evidence for distinct functions where MyoD is needed to efficiently recruit Pol II and mediate strong activation of early myogenic genes. (Weintraub H, 1991; Conerly ML, 2016) However, in order to fully execute the myogenic differentiation program myogenin appears able to access sites occupied by either myf5 or MyoD on top of ones only occupied upon its expression. (Conerly ML, 2016)

Additional questions about the mechanism of early steps in CRM recognition, occupancy, histone modification and regulatory activity comes from the fact that several different TFs, depending on the cellular context, can bind to identical motifs. However, in different genomic locations the occupancy pattern appears to depend on the pioneered status of the site and/or the presence of specific co-factors (and their post translational modifications (PTMs)). (Casey BH, 2018) While the detailed molecular mechanisms and dynamics are not fully understood, the genome-scale maps show that this machinery coordinates events at tens of thousands of genomic sites, with individual TFs typically occupying in the range of one to fifty thousand sites in each cell type.

1.11 Cross-tissue motif and occupancy comparisons

MyoD was one of the first identified trans-differentiation factors, followed by NeuroD which was able to differentiate cells to neurons. (Davis RL, 1987; Lee TC, 1994) The transcription factors NeuroD and MyoD independently function as heterodimers with an E-protein that specifies occupancy for an Ebox sequence motif (CANNTG). This shared motif is an essential component of the cell-type specific gene regulatory

networks in neurons and muscle respectively. (Yao Z, 2013) ChIP-seq allowed the field of genomics to refine and contrast the consensus motifs of TFs that were previously known to recognize this Ebox. From these measurements of DNA occupancy we learned that these two factors share a common core CAGCTG Ebox, with NeuroD also accessing CAGATG and MyoD occupying CAGGTG motifs. The nucleotides immediately external to the hexamer core motif are thought to be a shared RR-core-G (MyoD) and RR-core-GG (NeuroD2) which are hypothesized to specify binding by heterodimers containing the common E-protein partner. (Yao Z, 2013) Domain-swap experiments have demonstrated that the MyoD BHLH domain is sufficient to drive NeuroD occupancy to the majority of myogenic sites (GC core Eboxes), while PBX1 is responsible for specifying the remaining subset (GG core) of occupancy. (Fong AP, 2015) ChIP-Seq experiments also show that motifs bound by the musculin repressor and MyoD share a common central Ebox motif, but musculin occupancy maps to sites with different flanking nucleotides to the muscle Ebox (CCAGCTGG). (Macquarrie KL, 2013; Yao Z, 2014) A second example of ChIP-seq being used to refine occupancy motifs are PFT1A (pancreatic TF1A) and ASCL1 (also known to bind the CANNTG Ebox) occupancy having been mapped in neural tube tissue where these TFs regulate the expression of several factors important for proper development. (Borromeo MD, 2014) In that study, motif mapping analysis revealed a "muscle" class Ebox (CAGCTG) in 85% of common occupancy sites, and a G-CAGCTG-C Ebox in 73.6% of ASCL1 exclusive occupancy sites (Borromeo MD, 2014), whereas PTF1 specific occupancy is enriched for CAGATG and RBPJ motifs. (Borromeo MD, 2014)

Even though the motifs can be refined, comparing the genomic locations of ASCL1 occupancy with that of the MRF factors measure in our system, I found a relatively low (22%) overlap. I also found minimal (1%) overlap when comparing the MRFs to PTF1 which use different motifs. Comparing the occupancy regions of key TFs

related to the establishment of the muscle and erythroid lineage (Tal1 and MRFs, both of which recognize the Ebox) resulted in a similarly low overlap of ~10%. Based on the low TF occupancy overlap for factors that access identical or very similar motifs in different cell types, especially at distal occupancy regions (Figure 3.1D), it is expected that cross-testing the function of elements in different cell types by transient transfection assay might not be directly informative of native activity for the majority of the candidate cRE population. Because these regions are largely selected for cell-type specificity, the introduction of naked DNA fragments within cells that allow access to identical binding motifs could result in non-native gene expression in non-homologous cell types. (discussed further in section 1.15 and chapter 2) This exclusivity of genomic occupancy regions accessed by factors that recognize identical sequence motifs suggests that set of sites accessed in a given cell-type relies on additional molecular regulation external to the cell type specific TFs. One possibility is the leading effect of pioneering factors that bind at or near the location of final TFs. (Soufi A, 2015) This possibility poses other questions about the initial specificity of the pioneer(s), as suggested by the requirement of Pbx1 for MyoD to access its exclusive set of occupancy sites in domain swap experiments. (Fong AP, 2015) Despite currently unresolved questions about how TF pioneering activity is achieved, the idea of multi-layered controls that may include repressors and heterochromatic states as the starting quiescent substrate that is then penetrated and activated by multiple coordinated and even cooperative steps is emerging as the model for controlling the onset of myogenesis.

A stratified system of control can prevent inappropriate activation by regulating TF activity by the onset of expression of necessary co-factors, or by post-translational modification and subcellular compartmentalization. This complex system of control, along with cell population effects in our current measurements, pioneering of sites in developmental stages, and the redundancy of occupancy by multiple factors renders

difficult the attribution of the effects of transcript expression to any one individual input. Any attempt to correlate observed changes in expression to the biochemical inputs is further complicated by any gene which has multiple cREs, especially distal or previously undetected ones, contributing to its expression level, or where multiple cREs contribute in a partly redundant manner. (Barolo S, 2013; Levine M, 2003; Osterwalder M, 2018; Casey BH, 2018)

1.12 Regulatory functional themes confirmed by a standardized catalog of enhancer marks

While transcriptional regulatory elements can be defined using combinatorial biochemical inputs because there is considerable redundancy between select individual biochemical marks cREs can also be identified by following simpler rules. For example, candidate enhancers occupied by sequence specific TFs, TSS proximal and distal alike, generally exhibit DNAse Hypersensitivity. These TFs appear to cause at least some local histone rearrangement and recruit acetyltransferases (e.g. EP300) which deposit the H3K27ac mark on nearby histones. (Figure 3.1B) (Camerini RD, 1976; Hatzi K, 2013) It is however unclear as to whether activating transcription factors and these matching biochemical patterns (often peak called by different bioinformatic methods) remain similarly predictive across different tissue types or between developmental and established cell types. DNAse Hypersensitivity patterns and H3K27Ac deposition patterns have provided two over-arching biochemical signatures that are now being standardized and used for the compilation of candidate cis-regulatory elements (ccREs). As of the writing of this thesis, DNAse-seq maps, CTCF occupancy, and histone mark profiles across many cell types allowed the ENCODE consortium to annotate 1.31M and 0.53M ccREs in the human and mouse genomes respectively. (ENCODE, 2012; Moore J, 2018) It is important to note that while active regulatory elements often exhibit characteristic biochemical signatures,

the reverse — that characteristic signatures always infer activity in all of the possible spatial and temporal localizations where they are found — does not necessarily hold true. (Kellis M, 2014) Biochemical marks by themselves do not provide a direct understanding of how functional REs are specified, or exercise their function since some may be concurrent while others may be setting up for, or lagging behind, the functional output. Any inference of active biological function is further complicated by the observation that biochemical signatures exist on a quantitative continuum between outstanding features — often the first to be assayed — and what is likely biochemical background noise. (Figure 3.1C) For example, it is far from clear that all transcription factor binding sites of a characterized trans-activator, which can be reproducibly identified using ChIP-seq, are in fact active enhancers. (Fisher WW, 2012) Therefore, individual ccREs in the catalogs compiled by efforts such as the ENCODE and mouseENCODE consortia (ENCODE, 2012; Yue F, 2014) have to be subsequently tested and functionally characterized in detail.

Also, ChIP-seq does not allow us to distinguish between direct binding vs recruitment of TFs, two fundamentally different interactions. In the first case, the TF is interacting physically with the DNA while in the latter case, a protein is recruited to the DNA by a secondary protein-protein interaction. (Johnson DS, 2007) What was perhaps most surprising from genome-wide enhancer mark data is that each gene, especially genes dynamically expressed and involved in developmental contexts, presented a much larger than expected number of candidate enhancers, many of them distally located to their linked promoters. In contrast, the majority of housekeeping genes appear to rely largely on a strong promoter control mechanism, with proximal elements. Another surprising finding is that occupancy sites appear able to promiscuously recruit up to hundreds of different TFs in a broad population of cells. (Partridge C. 2018 in drafting) It is unclear how this overall promiscuous occupancy, which might result in occupancy by sets of different TFs in individual

cells of a population can be beneficial for the controlled expression of nearby genes.

1.13 Enhancer to promoter targeting

The important but still difficult task of assigning distal elements to their correct gene (or genes) and understanding how multiple active distal element inputs are integrated at a given promoter are central problems. While most enhancers are able to function in either orientation relative to their targeted promoter a peculiar insight into enhancer/promoter targeting is found in a specific cRE found in the CKM first intron. This cRE exhibits a cell-type-specific position dependency in the valence it provides to its targeted promoter. It can act as a suppressor of gene transcription when placed upstream of its native promoter in cardiomyocytes, but not in skeletal muscle cells. (Banerji J, 1981; Tai PWL, 2011) These findings suggest specific targeting of enhancers to promoters.

As a second example, at the combined Myf5/Myf6(MRF4) locus, the A17 enhancer selectively targets either gene using anatomical and temporal cues within the embryo. (Chang TH, 2007; Yoon JK, 1997) A promoter deletion of the natural target gene caused the upstream enhancer to retarget the nearby Myf6/MRF4 promoter as well as several cryptic Myf5 promoters. This result resolved long standing confusion about mixed results obtained by several Myf6 gene targeting strategies from different labs which generated different expression patterns across development. (Chang TH, 2004; Chang TH, 2007; Zammit PS, 2004) The originally proposed model for this equilibrium involves nonproductive interactions between the enhancers and multiple promoters. (Carvajal JJ, 2008) Elements that are involved with multiple promoters at the same spatial/temporal locus go against the previous models described for the regulation of the ß-globin locus by the LCR and the Drosophila abd-B gene suggesting that enhancer to promoter targeting may be a somewhat random process that can also bridge multiple promoters or enhancers together, rather than enhancers

always to a single promoter. (Wilderme M, 1995; Zhou J, 1999)

These examples raise the key question of how individual promoters in a region are excluded from the activity of nearby strong cREs driving the expression of genes involved in the differentiation of muscle. The information could be encoded at the promoter or it could be regulated at the chromatin level from the combined sequence near the gene. In addition to binding profiles for the MRFs and several cofactors, we have two CTCF ChIP-Seq data sets produced from C2C12 cells from our lab and the Broad Institute to integrate into our analysis and experimental design. (unpublished data) The CTCF insulator TF has been shown to have two distinct activities. It can act as both a block between enhancer/promoter interactions by creating looped domains as well as provide a barrier between heterochromatin and active transcription domains. This makes it a likely candidate to function as a buffer between the strong CREs and the TSS of nearby genes with moderate or no transcription. (Felsenfeld G, 2004; Oti M, 2016) While CTCF insulator boundaries have provided a way to begin to restrict the chromosomal domains in which a given distal element could access a target gene, we know that only some CTCF sites act as boundary sites. This means that mere motif presence or even ChIP-detectable occupancy by CTCF is not sufficient.

A second possible molecular mechanism for selective promoter targeting could rely on TAF3/TRF3 transcription initiation complex subunits. These subunits are selectively switched during muscle differentiation as indicated by genome-wide occupancy measurements. (Deato MDE, 2007) This paper reported evidence that the composition of the Holo-TFIID was altered across differentiation from myoblast to myocyte by the substitution of a novel subunit. This switching of the core promoter subunits provides a potential molecular mechanism to selectively target promoters for activation and or silencing. In collaborative work with the Tijan laboratory we could not replicate these findings, and subsequent studies suggest that the antibody

for TAF3 was not as specific as originally believed, allowing the possibility that other TAFs normally present in the system were mistakenly detected. A previous study however suggested that a switch of the core machinery might be necessary only for function in the postmitotic state which might not have been reliably detectable in early differentiated C2C12 cells. (Apone LM, 1996)

A third possible mechanism for selective promoter usage might be encoded by the combination of context present and established pioneering in each cell type. This is supported by the observation that cell type specific TFs from different lineages occupy separate regions of the genome whereas TFs within the same context (eg MYF5 and MyoD) occupy an highly overlapped set of sites. This suggests that site selection is either mediated by recruitment of co-expressed co-factors or TFs pioneering the correct sites (or a combination of the two). The combination of access to distal and proximal sites in a cell type, together with the CTCF occupancy, is likely a key driver of at least some of the specificity observed. For example in muscle, PBX1, a protein thought to have a role in recruiting myoD to certain sites in C2C12, is present where MyoD alone occupies the E-box motif but Myogenin does not (Heatmaps in Chapter 3 Supplemental Figure 5). (Yao Z, 2013; Macquarrie KL, 2013; Maves L, 2007; Weintraub H, 1989; Fong AP, 2015) It is not yet clear exactly how PBX1 contributes to pioneering, and it might be necessary mainly to stabilize MyoD binding rather than truly pioneering on its own. (Grebbin BM, 2017) This view has been further supported by evidence that specific histone variants are needed for PBX1 recruitment, and histone substitution is thought to come after pioneering rather than before. (DellOrso S, 2016) Further, even though Myf5 and MyoD share their occupancy, their function is not the same, as MyoD is required for activating transcription; whereas Myf5 is limited to occupancy and recruitment of H3K27Ac at certain genomic locations. The latter observation might appear to be conflicting to current models that suggest that the histone marks are the result of the

biological activity at the occupancy sites. A role for EP300 in modifying the Myf5 protein, similar to that observed for MyoD, might explain this observation. Yet this mechanism provides yet another switch by which two identically marked sites might only be active enhancers when the correct TF is present (such as myogenin) or when known repressors (IDs) are ineffective. (Further discussed in Chapter 5) It also remains unclear whether myogenin can access all of its native sites without this prior pioneering by PBX1 and MyoD, indeed during development of the embryo different compartment show varying orders to the onset of the four MRF factors, which might specify slightly different sets of genes being accessed.

To provide an example of pioneering which might be required for later occupancy and possibly correct promoter targeting, Pbx1/Meis is responsible for recruiting myogenin to the promoter proximal site for the myogenin gene itself. (Yao Z, 2013; Macquarrie KL, 2013) In this specific case, the 1092bp region upstream of the myogenin TSS is able to support a specific activation in the myotome of the embryo (independently replicated in our laboratory in a lentiviral context). The short stretch called the myogHCE (myogenin Highly Conserved Element) is responsible for most of this function based on mutagenesis experiments, and partially overlaps an MSY3 occupancy motif previously characterized in our lab and occupied in myoblasts by MSY3 in ChIP-seq assays. The GZ133 (133bp upstream of TSS) and GZ188 (188bp) and GZ1092 (1092bp) cREs tested from this region show modestly increasing amounts of enhancer activity respectively in an exogenous reporter assay.(Yee SP, 1993) Removing just a few base pairs (Figure 4.9; MYOG125) containing the PBX1 site (125bp) kills all remaining activity. Notably, this mutation also prevents Meis occupancy. (Berkes CA, 2005) This element could therefore function as a bistable switch between MSY3-mediated repression and a Pbx-mediated positive Myod state. (Berghella L, 2008; DeAngelis L, 2015) This bistable quality might also be the mechanism by which MSY3 maintains myogenin repression in differentiated

myotubes after innervation, when MRF4 takes over the MRF role. This is consistent with the measured activity of the 1092-133 myogenin promoter fragment in different cell types when tested with an HSP68 target promoter-reporter, where it loses its muscle specificity and displays strong ectopic expression. (Yee SP, 1993) A coherent interpretation is that both Enhancer and Silencer sub-elements within this CRM are equally important for specifying cell type specific expression and that an element that functions as an enhancer in one tissue might very well be an active silencer in another. (Howard ML, 2004; Istrail S, 2005)

While the mechanism by which promoters are targeted is not yet fully understood, a Pol II ChIA-PET map (Fisher K, Thesis 2016) can reveal which genomic regions are brought into physical proximity and enrich for regions occupied by RNA POLII, a mark largely associated with active transcription. This measurement found that candidate Enhancers (and even characterized enhancers) can skip over several silent genes and loop to their targets which are sometimes located several hundred kilobases away from a reproducible gene target. (Long L, 2013; KiefferKwon KR, 2013) These data however, raise an additional question as to whether the physical connectivity mapped in this way captures mainly functional elements, as the assay was initially designed to do. For example, one might want to ask if the portion of elements selected for myogenin occupancy which are connected, are enriched for activity when compared to the myogenin occupancy elements which are not connected. This part may be difficult to assess reliably because many non-connected elements may be false negatives where the connection was not successfully captured, or may simply be a less “stable” connection. These findings suggest that the targeting is tightly controlled as several genes closest to even very strong enhancers are kept silent, while the correct gene is modulated.

Functional testing should be performed against a promoter element that is permissive in the cell system and is able to pair with the candidate functional element

tested. However, we do not fully understand the range and mechanism(s) underlying proximal promoter "permissivity". For the vast majority of function studies that, as in this work, use a single, invariant proximal element to allow for direct comparisons of varying distal elements, three main promoter types have been employed.

The most conservative strategy employs a very minimal promoter, including few base-pairs other than a TATA box, which generally produces substantially lower signals for bona fide enhancers. This means that weaker enhancers might be cast aside due to the lower signal to noise ratio, and any other unintended start sites (including vector sequences) might assume a large and potentially confounding role. The use of a minimal promoter that has previously been shown to be permissive for the cell type(s) and enhancer class being surveyed might improve overall sensitivity for enhancer function over a very minimal TATA-like promoter. These historically included viral promoters (HSVtk; CMV; SV40T antigen) or unregulated portions of the HSP or beta globin, among others. Especially relevant to my studies are a set of important prior experiments in C2C12 cells, including ones aimed at exploring the minimal number of muscle class E-boxes which could comprise an active enhancer. They were performed using a Thymidine Kinase (TK) viral promoter. (McKnight SL, 1982; Weintraub H, 1990; Weintraub H, 1991; Molkentin JD, 1996) Figure 1.4 compares the basal level of CytoMegaloVirus (CMV), Thymidine Kinase (TK) and Simian Virus 40 (SV40) promoter driven Renilla co-transfected with several test enhancer constructs in both myoblasts (MB) and myocytes (MC). (Figure 1.4)

A final strategy used in the literature is to select a proximal promoter that is "native" to a locus of special interest or represents activity in a particular cell type or differentiation state that matches a group of enhancers to be tested. The majority of the enhancer control sequences cited in this chapter were derived from historically studied elements that are preferentially located near to their native TSS, and they are also often preferentially conserved. Those tested as part of the study were

not selected for TSS-proximity nor for preferential conservation, but rather for significant occupancy by myogenin (Fig 1.2) At a few key loci, additional elements were tested based on nuclease accessibility in the model cell system (C2C12), regardless of the factor occupancy profile (Chapter 4).

The overall upshot of reporter designs, and of a few more detailed studies of promoter-distal element interactions (Yee SP, 1993; Teboul L, 2002; Summerbell D, 2000, Appendix E) is that practical considerations have led to design compromises for single-promoter reporters that future larger-scale multiplex studies should address via distal element/promoter combinatorics. At present, these considerations play a role in interpreting single-reporter studies. For the muscle assays presented here we selected the same TK promoter because of its historic use and because it offered a lower starting signal to measure fold activation. We note that a key weakness of most current MPRA assays is that they require a relatively strong promoter which allows to sort signal-positive cells prior to sequencing in order to achieve sensitivity against a noisy background of transcripts from the whole genome.

The ultimate functional characterization of candidate enhancers must involve a combination of loss-of-function assays and direct assays for enhancer, silencer or promoter activity without confounding factors, across a very broad range of cellular contexts. Loss of function assays have been (until recently) technically challenging, but are becoming more commonplace with the advent of large-scale CRISPR/Cas9 mediated mutagenesis techniques. (Korkmaz G, 2016; Fulco CP, 2016) In order to approach some of these issues I focused on studying the predictive power for enhancer function of TF occupancy and biochemical marks within the developmental process of myogenesis and contrasted it to that of other systems against a single promoter.

1.14 Massively Parallel Reporter Assays

High-throughput sequencing has enabled the development of assays that go beyond the one-by-one testing of individual cREs by using very large numbers of sequences analyzed in parallel. In these assays the readout is based on sequencing transcribed DNA tags associated with the cRE (cRE-seq) or sequencing the transcribed cRE itself (STARR-seq). These assays are usually referred to collectively as MPRAs (massively parallel reporter assays) (Inoue F, 2015) and a number of variations have been successfully applied to different biological systems. (Patwardhan RP, 2009; Kinney JB, 2010; Kwasnieski JC, 2012; Melnikov A, 2012; Patwardhan RP, 2012; Arnold CD, 2013; Murtha M, 2014) This method has also been used for testing cREs for activity by the ENCODE Consortium. (Kheradpour P, 2013; Kwasnieski JC, 2014; Ernst J, 2016) MPRA designs are attractive because they allow the functional characterization of larger numbers of elements across the genome individually, providing a more comprehensive survey that should be more statistically persuasive for subclasses of infrequent elements. However, several quite substantial issues complicate the interpretation of current MPRA experiments. The nature of MPRA designs, so far, is that the elements tested are short (in the 80-250bp range) compared to the size spectrum of conservation within known enhancer elements, within the occupancy of a TF (~400bp) or compared to the size of thousands of conserved noncoding nucleotide elements in the genome (Chapter 3, Supplemental Figure 1). Assuming that conserved non-coding sequences represent a major subset of functional enhancers, and the regions of conservation are a minimal estimation of a complete cRE (Chapter 3), these assays thus truncate many elements and may generate significant numbers of false negatives. Alternatively, the magnitude of enhancer activity detected might not be complete as other nearby elements are not included within the test element, and these short regions might also lead to a loss of cell-type specificity in some cases. (White MA, 2013) For example, it is not unusual

for CRMs to contain repressor binding sites as well as sites for positive acting factors. Deleting a repressing cis-sequence could leave the remaining element de-repressed in a cellular context where the full element would normally be “off” or “poised”, yet not active. More generally, the high-throughput tests in their current forms do not capture the combinatorial factor array often found to occupy multiple binding sites within a ChIP-Seq region when testing such small fragments for enhancer function. (White MA, 2013; Kwasnieski JC, 2014) Even within genome integrated tests, similarly low activity estimates of upwards of 28% of active elements are observed indicating that the size of elements are key to faithfully replicating the activity observed in the chromosome. (Murtha M, 2014)

Pertinent to both high throughput and more conventional reporter assays, it was recently discovered that the bacterial origin of replication (ORI) can cause interference with the test promoter. This results mainly in constitutive background activity, the magnitude of which appears to depend on the promoter used to drive the reporter. (Muerdter F, 2018) Whether a competing promoter entity is provided by vector sequence or by other unintended elements in a construct, it is expected to increase noise and degrade sensitivity. If different enhancers are differentially “attracted” to the ORI, it is expected to confound quantitative output relationships. (Conte C, 2002)

This issue is also likely to affect the system used in this thesis, which relies on a single plasmid in each cell culture well to measure enhancer activity. However, this problem in our system is not compounded by the presence of multiple enhancers concurrently as in high-throughput MPRA assays. Further, in specific cases, some of the regulatory elements studied in this thesis have been assayed in multiple formats, including integrated transgenics and lentiviral test systems, with similar results for a small set of our muscle elements. The possibility that we may be also underscoring due to a subset of elements favoring interaction with the ORI is however possible

and should be tested.

Other concerns with MPRA interpretation relate to issues that have become better appreciated in individual element assays. It is very difficult to control the number of constructs going into each individual cell in transfection experiments, and MPRA features large numbers of different cREs being tested concurrently. By explicitly titrating mixtures of two well-characterized constructs, initial design work for this thesis documented significant "cross-talk" between two well-defined test reporters. These results led us and our collaborating labs in ENCODE to focus the work on single reporter measurements rather than the superficially elegant, yet ultimately misleading, dual reporter ratio designs.

In the future, these and related concerns can be addressed by using genome-integration (Maricque BB, n.d.; Inoue F, 2016; Muerdter F, 2018), which we hope to combine with longer insert MPRA insert libraries. However, the short length of constructs tested remains a significant issue, and might be responsible for the low positive rates reported by MPRAs. (Vockley CM, 2016)

1.15 Validation of biochemical mark predictions with longer cRE assays

One of the cheapest and most straightforward ways to validate an enhancer for activity has been to test cREs using an exogenous plasmid construct combining a cRE, a promoter and a reporter gene. Numerous enhancers (including developmental ones) have been characterized following this approach. (Appendix E) (Visel A, 2007; Visel A, 2009; May D, 2011; Fisher WW, 2012)

My aim was to develop a robust enhancer test (described in the next chapter) in the muscle system that would allow us to ascertain both function and relative activity of full sized myogenin occupied regions (~1kb) from the full quantitative range of biochemical signal detected by ChIP-seq. The results from this assay are presented in chapter 3 which contains a draft manuscript that reports on the findings from

my assay and contrasts them to other similar assays in other tissue types from our collaborators within ENCODE.

It must be noted that a negative result from a test in a single cell-type context does not predict negative results in all other possible cellular contexts. (Moore J, 2018) It is important to remember that exogenous tests for enhancer activity are done in limited cellular contexts, often against a single promoter element and in cell lines that are established from a single origin. This approach may not reveal the activity level in the native chromosomal context and may miss combinatorial contributions from other nearby elements in either the native locus or in the site of integration. It only measures both the existence and magnitude of effect by single elements. These assays cannot measure the contribution of these enhancers within their native locus which can only be achieved by targeted deletion of the cRE and observing the effects on the nearby genes. (Hnisz D, 2013; Kvon EV, 2016; Dickel DE, 2018) This type of assay does however allow to measure the combined contributions of limited pairs of elements associated with a gene by cloning them into the same test construct.

Transient reporter assays, where the DNA is not integrated into the genome, are criticized as being unable to reflect the native chromatin context; which might be worse for MPRAs due to the diminutive size of the DNA cloned as 250bp is barely sufficient to support one natively positioned nucleosome; compared to <4 in the larger regions tested in these single plasmid assays. Due to this difference it is important to properly control such experiments with large sets of negative control elements which might be expected to be chromatin silent, such as the promoter proximal regions of genes that might result, if expressed, in catastrophic consequences to proper development of a cell type. For example, STARR-seq can now even encompass the entire genome for enhancer tests (albeit at poor fold coverage), but the reliability of such tests in a context that includes regions of less than 1 nucleosome is unclear as enhancer function is detected even in regions

which are not natively accessible. Whether this remains true in a larger test system remains unclear. (Muerdter F, 2018; Liu Y, 2017) Indeed this smaller structure might prevent in some cases the formation or displacement of the histones observed *in vivo* at thousands of occupancy sites across different tissues. (Figure 3.1A)

One of the key benefits of controlled integration is that it minimizes the random effects of integration on the test construct and on the nearby location which might be affected differently on a cell to cell basis where some might be silenced, some might be upregulated and others might disrupt other important genes nearby. The controlled integration site provides a test within a native chromatin approach for a locus that is known to be generally accessible (not heterochromatinized) across most cell types. For example, the mouse genome ROSA 26 locus was identified by gene trapping as a candidate for controlling the locus of integration and low copy number both of which are impossible in transient transfection assays. (Zambetti GP, 1992)

Although tests of all of these variables are possible they are expensive and currently limited to smaller numbers. A first pass screen requires a scalable platform that enables testing several hundred elements at a reasonable cost and in a relatively short timeframe. Medium-throughput luminescence reporter based transfection systems, such as the 96 well based assay that I robotized as part of my thesis work (chapter 2), have proven to be a robust method to assay full size candidate enhancer element activity across cell lines. (Trinklein ND, 2004) With this assay, as part of the ENCODE Project Consortium efforts towards functional validation of cREs, we tested the regulatory activity of hundreds of candidate enhancer elements (cEnhs) using constructs designed to capture full-length cREs of both mouse and human origin. These cEnhs were selected from a wide range of biochemical signature strengths, using both TF-centric selection criteria (identified directly from TFs occupancy) and machine learning "TF-agnostic" approaches (identified from combinatorial biochemical signatures of enhancers) across diverse mammalian cell lines.

Concurrent with our own study, several ongoing studies at LBNL used injection of DNA into developing mouse embryos to temporally and spatially map enhancer function for full size cREs (~1800bp regions - Chan AW, 1999) in randomly integrated genomic locations, in tissues where promoter function is not limiting. (Visel A, 2009; Wang Z, 2005) Scoring the database of the combination of these pronuclear injection studies (n=1993), reveals that 50% of test elements are active enhancers in at least one tissue. (Chan AW, 1999) This result is comparable to our own findings on the proportion of active elements presented in Chapter 3, and to that in the current ENCODE encyclopedia paper indicating that for sufficiently large elements, chromatin integration may not be required for faithful reproduction of the enhancer function for individual elements. (Moore J, 2018) We expect that the findings in this thesis will help guide efforts towards the comprehensive cataloging of functional elements in the human genome, and we discuss the implications of our findings in the context of models of gene regulation mediated by the action of distal enhancers.

Given recent technological developments and some of the initial results that we have today, I would have tested smaller sets of individual enhancers, and more varied combinations of enhancers. I would also have assayed these elements in both orientations, upstream and downstream of native promoter elements affiliated by connectivity. (Garrity PA, 1990) These would be contrasted for activity to generic promoters that contain muscle-type motifs, promoters that do not contain muscle-type motifs and finally TATA vs non-TATA promoters. In this way, the encoded specificity of action for functional regulatory elements might be better understood. We could then begin systematic tests of the molecular logic driving the developmental transition of myogenesis. (Howard ML, 2004; Arnone MI, 1997; Butler JE, 2001)

1.16 Figures (Chapter 1)

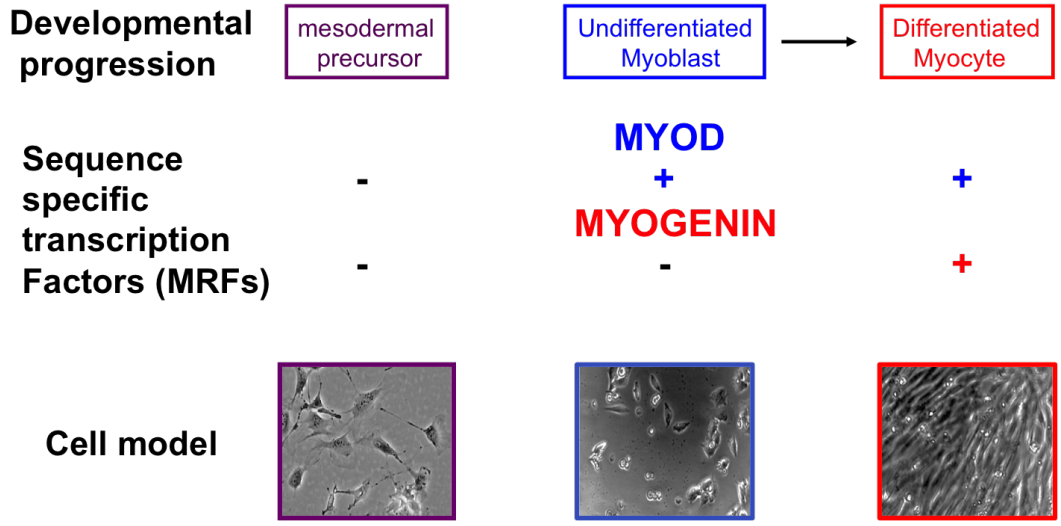


Figure 1.1: Expression pattern of MyoD and myogenin across myocyte differentiation. MyoD is expressed in both C2C12 undifferentiated myoblasts and differentiated myocytes whereas myogenin is expressed solely in differentiated myocytes. In contrast neither is expressed in the mesodermal precursor represented by the 10T1/2 cell line.

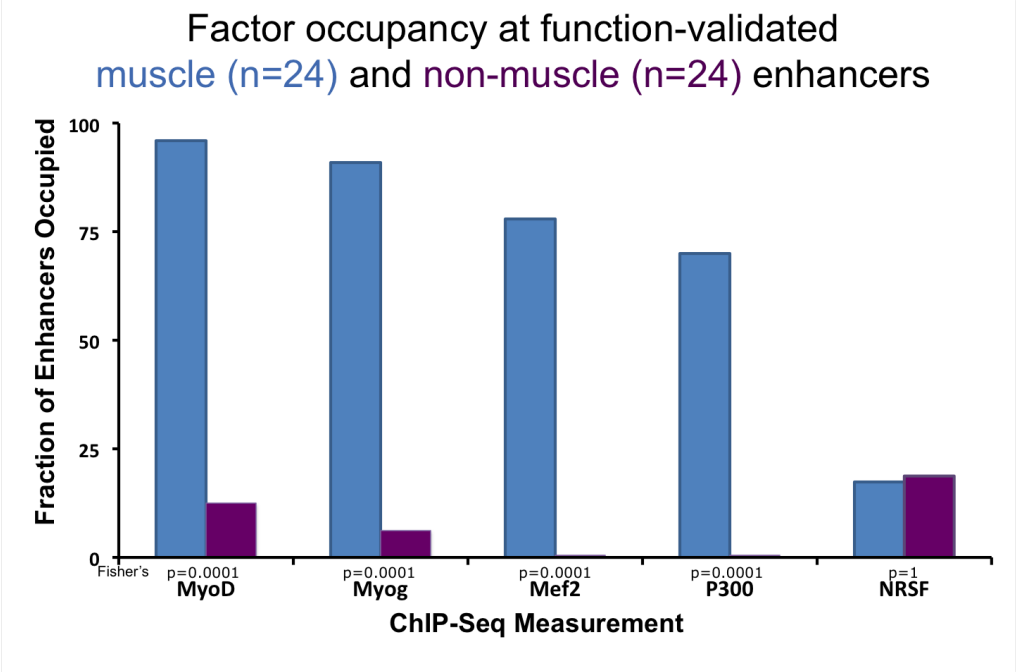


Figure 1.2: Occupancy of two of the myogenic regulatory factors and one non-muscle TFs at function validated Muscle and Non-Muscle Enhancers. 95% of a large set of previously known muscle related Enhancers display significant binding of key myogenic regulatory factors (MRFs) in our global measurement. (blue) In contrast ~10% of non muscle enhancers identified as active in neuronal and t-cell lineages display binding of these key MRFs. (purple) Similarly only ~15% of the muscle related CREs showed binding of NRSF although some of these may be false positives due to background ChIP signals at transcriptional start sites (TSS). Fisher's exact test p values comparing the proportion of TF occupied muscle vs non-muscle enhancers are presented above each TF name.

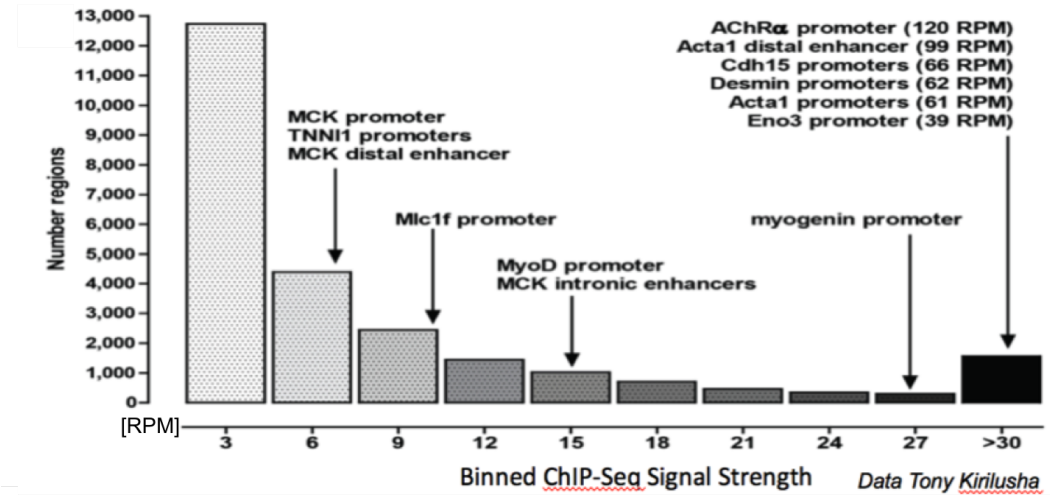


Figure 1.3: Histogram representing myogenin ChIP-Seq signal strength binned by RPM over the medium confidence ERANGE calls. Overlaid are the corresponding ChIP-Seq signature strength for 13 validated enhancer regions selected from the literature (Figure 1.2). Adapted from (Landt SG, 2012).

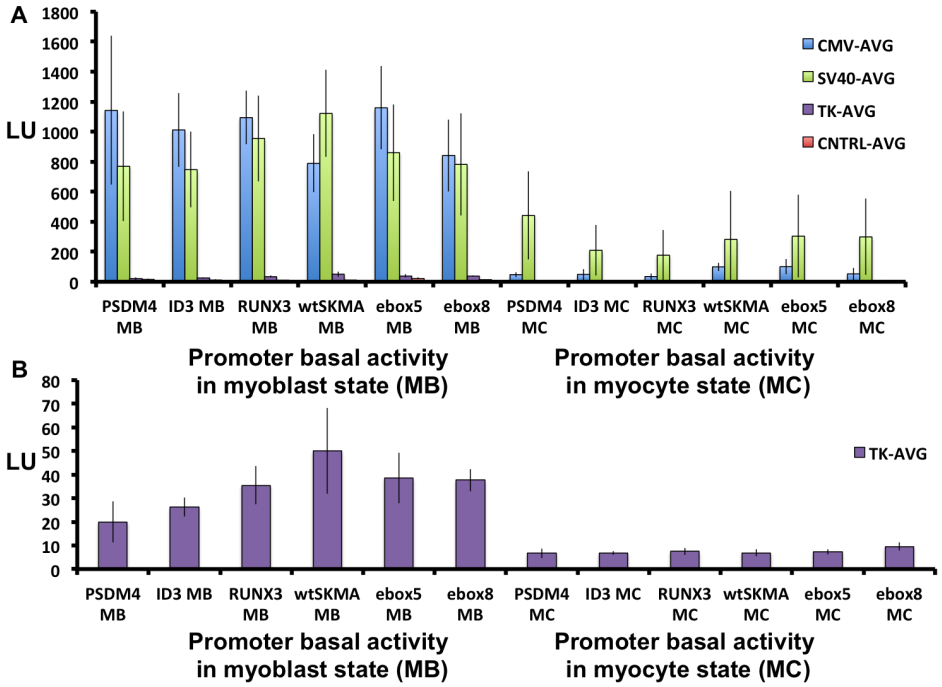


Figure 1.4: Behavior of promoter driven Renilla cotransfectants with several test constructs. Blue = CMV; Green = SV40; Purple = TK and Red = Empty Control promoters in myoblast (MB) and myocyte (MC) C2C12s. (A) Full range of values. (B) Rescaled values for the range of signal pertinent to the selected TK promoter.

References

Adelman K, et al. (2005). "Efficient Release from Promoter-Proximal Stall Sites Requires Transcript Cleavage Factor TFIIS." In: *Mol. Cell* 17, pp. 103–112.

Amacher SL Buskin JN, et al. (1993). "Muscle Creatine Kinase Enhancer Activity in Skeletal and Cardiac Muscle." In: *Science.* 13, pp. 2753–2764.

Andersson R, et al. (2014). "An atlas of active enhancers across human cell types and tissues." In: *Nature.* 507, pp. 455–461.

Andres ME, et al. (1999). "CoREST: a functional corepressor required for regulation of neural-specific gene expression." In: *PNAS* 96, pp. 9873–9878.

Andres V Cervera M, et al. (1995). "Determination of the Consensus Binding Site for MEF2 Expressed in Muscle and Brain Reveals Tissue-specific Sequence Constraints." In: *Journal of Biological Chemistry.* 270, pp. 23246–23249.

Apone LM, et al. (1996). "Yeast TAF(II)90 is required for cell-cycle progression through G2 M but not for general transcription activation." In: *Genes Dev.* 10, pp. 2368–2380.

Aravin AA, et al. (2008). "Small RNA silencing pathways in germ and stem cells." In: *Cold Spring Harb Symp Quant Biol.* 73, pp. 283–290.

Arnold CD Gerlach D, et al. (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." In: *Molecular and Cellular Biology.* 339, pp. 1074–1077.

Arnone MI, et al. (1997). "The hardwiring of development: organization and function of genomic regulatory systems." In: *Development* 124, pp. 1851–1864.

Ballester B, et al. (2014). "Multispecies, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways." In: *eLife* 3, pp. 2626–2627.

Banerji J, et al. (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." In: *Cell.* 27, pp. 299–308.

Barkess G West AG, et al. (2012). "Chromatin insulator elements: establishing barriers to set heterochromatin boundaries." In: *Epigenomics.* 4, pp. 67–80.

Barolo S, et al. (2013). "Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy." In: *Bioessays.* 34, pp. 135–141.

Barski A, et al. (2007). "High resolution profiling of histone methylations in the human genome." In: *Cell.* 129, pp. 823–837.

Bartel DP, et al. (2009). "MicroRNAs a target recognition and regulatory functions." In: *Cell.* 136, pp. 215–233.

Becker NA, et al. (2013). "Mechanism of promoter repression by Lac repressor–DNA loops." In: *Nucleic Acids Res.* 41, pp. 156–166.

Berghella L De Angelis L, et al. (2008). “A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle.” In: *Genes Dev.* 22, pp. 2125–2138.

Berkes CA, et al. (2005). “MyoD and the transcriptional control of myogenesis.” In: *Cell and Developmental Biology.* 16, pp. 585–595.

Berthelot C, et al. (2018). “Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression.” In: *Nat. Ecol. Evol.* 2, pp. 152–163.

Bessis A, et al. (1997). “The neuron-restrictive silencer element: A dual enhancer silencer crucial for patterned expression of a nicotinic receptor gene in the brain.” In: *PNAS* 94, pp. 5906–5911.

Blackwell TK, Weintraub H. (1990). “Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection.” In: *Science.* 250, pp. 1104–1110.

Blanchetot A, et al. (1986). “Enhancer interaction networks as a means for singular olfactory receptor expression.” In: *Eur J Biochem.* 159, pp. 469–474.

Blow MJ, et al. (2010). “ChIP-seq Identification of Weakly Conserved Heart Enhancers.” In: *Nat. Genet.* 42, pp. 806–810.

Borromeo MD Meredith DM, et al. (2014). “A transcription factor network specifying inhibitory versus excitatory neurons in the dorsal spinal cord.” In: *Development.* 141, pp. 2803–2812.

Braun T, et al. (1989). “A novel human muscle factor related to but distinct from MyoD1 induces myogenic conversion in 10T1/2 fibroblasts.” In: *The EMBO Journal.* 8, pp. 701–709.

Brown CT, et al. (2002). “New computational approaches for analysis of cis-regulatory networks.” In: *Dev Biol.* 246, pp. 86–102.

Buckingham M, et al. (2014). “Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis.” In: *Dev. Cell* 28, pp. 225–238.

Buskin JN, et al. (1989). “Identification of a myocyte nuclear factor that binds to the muscle-specific enhancer of the mouse muscle creatine kinase gene.” In: *Mol. Cell Biol.* 9, pp. 2627–40.

Butler JE, et al. (2001). “Enhancer–promoter specificity mediated by DPE or TATA core promoter motifs.” In: *Genes Dev.* 15, pp. 2515–2519.

Camerini RD, et al. (1976). “The organization of histones and DNA in chromatin: evidence for an arginine-rich histone kernel..” In: *Cell* 8, pp. 333–347.

Cameron R, et al. (1997). “LiCl Perturbs Ectodermal Veg1Lineage Allocations in Strongy locentrotus purpuratus Embryos.” In: *Dev. Biol.* 187, pp. 236–239.

Cao Y Yao Z, et al. (2010). “Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming.” In: *Developmental cell.* 18, pp. 662–674.

Carvajal JJ, et al. (2008). “Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5.” In: *Genes and Development.* 22, pp. 265–276.

Carvajal JJ Cox D, et al. (2001). “A BAC transgenic analysis of the Mrf4 Myf5 locus reveals interdigitated elements which control activation and maintenance of gene expression during muscle development.” In: *Development.* 128, pp. 1857–1868.

Casey BH Kollipara RK, et al. (2018). “Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors.” In: *Genome Res.* 28, pp. 484–496.

Chan AW, et al. (1999). “Timing of DNA integration, transgenic mosaicism, and pronuclear microinjection.” In: *Mol. Rep. Dev.* 52, pp. 406–413.

Chandler KJ Chandler RL, et al. (2009). “Identification of an ancient Bmp4 mesoderm enhancer located 46 kb from the promoter.” In: *Developmental Biology.* 327, pp. 590–602.

Chang TH Primig M, et al. (2004). “An enhancer directs differential expression of the linked Mrf4 and Myf5 myogenic regulatory genes in the mouse.” In: *Dev Biol.* 269, pp. 595–608.

Chang TH Vincent SD, et al. (2007). “The A17 enhancer directs expression of Myf5 to muscle satellite cells but Mrf4 to myonuclei.” In: *Dev Dyn.* 236, pp. 3419–3426.

Cheng CK, et al. (2002). “Oct-1 is involved in the transcriptional repression of the gonadotropin-releasing hormone receptor gene.” In: *Endocryn.* 143, pp. 4693–4701.

Cheng J Blum R, et al. (2014). “A role for H3K4 mono-methylation in gene repression and partitioning of chromatin readers.” In: *Molecular cell.* 53, pp. 979–992.

Cheng TS, et al. (1993). “Separable regulatory elements governing myogenin transcription in mouse embryogenesis.” In: *Science* 261, pp. 215–18.

Chow KL, et al. (1990). “A combination of closely associated positive and negative cis-acting promoter elements regulates transcription of the skeletal alpha-actin gene.” In: *Mol. Cell Biol.* 10, pp. 528–38.

Cirillo LA, et al. (2002). “Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4.” In: *Molecular Cell* 9, pp. 279–289.

Clayton CE Murphy D, et al. (1982). “A fragment of the SV40 large T-antigen gene transforms.” In: *Nature.* 299, pp. 69–61.

Conerly ML Yao Z, et al. (2016). "Distinct Activities of Myf5 and MyoD Indicate Separate Roles in Skeletal Muscle Lineage Specification and Differentiation." In: *Dev Cell.* 36, pp. 375–385.

Conte C, et al. (2002). "Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons." In: *Embo J.* 21, pp. 3908–3916.

Cooper SJ, et al. (2006). "Comprehensive analysis of transcriptional promoter structure and function in 1 percent of the human genome." In: *Genome Res.* 16, pp. 1–10.

Core LJ, et al. (2008). "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." In: *Science.* 322, pp. 1845–1848.

Cornelison DD, et al. (2000). "MyoD - satellite cells in single-fiber culture are differentiation defective and MRF4 deficient." In: *Dev Biol.* 224, pp. 122–137.

Creyghton MP Cheng AW, et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." In: *Proc Natl Acad Sci U S A.* 107, pp. 21931–21936.

Cserjesi P, et al. (1994). "Homeodomain protein MHox and MADS protein myocyte enhancer-binding factor-2 converge on a common element in the muscle creatine kinase enhancer." In: *J. Biol. Chem.* 269, pp. 16740–45.

Davidson EH, et al. (2005). "Gene Regulatory Networks." In: *PNAS* 102, pp. 4935–4936.

Davis RL Weintraub H, et al. (1987). "Expression of a single transfected cDNA converts fibroblasts to myoblasts." In: *Cell.* 51, pp. 987–1000.

DeAngelis L, et al. (2015). "Akt mediated phosphorylation controls the activity of the Y box protein MSY3 in skeletal muscle." In: *Skelet. Muscle* 5, pp. 18–19.

Deato MDE Tjian R, et al. (2007). "Switching of the core transcription machinery during myogenesis." In: *Genes and Dev.* 21, pp. 2137–2149.

Dekker J, et al. (2002). "Capturing chromosome conformation." In: *Science* 295, pp. 1306–1311.

DelaSerna A, et al. (2005). "MyoD Targets Chromatin Remodeling Complexes to the Myogenin Locus Prior to Forming a Stable DNA-Bound Complex." In: *Molecular and Cellular Biology.* 25, pp. 3997–4009.

DellOrso S Wang AH, et al. (2016). "The Histone Variant MacroH2A12 is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1." In: *Cell Rep.* 13, pp. 1156–1168.

Dickel DE, et al. (2018). "Ultraconserved Enhancers Are Required for Normal Development." In: *Cell.* 172, pp. 491–499.

Eckner R, et al. (1996). "Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation." In: *Genes Dev.* 10, pp. 2478–2490.

Edwards JG, et al. (1992). "A repressor region in the human b-myosin heavy chain gene that has a partial position dependency." In: *Biochem Phys Res.* 189, pp. 504–11.

Eisenberg E, et al. (2003). "Human housekeeping genes are compact.." In: *Trends Genet.* 19, pp. 362–365.

ENCODE, Project Consortium (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." In: *PLoS Biol.* 9, e1001046.

– (2012). "An integrated encyclopedia of DNA elements in the human genome." In: *Nature.* 489, pp. 57–74.

Ernst J Melnikov A, et al. (2016). "Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions." In: *Genome Res.* 34, pp. 1180–1190.

Faghihi MA, Wahlestedt C. (2009). "Regulatory roles of natural antisense transcripts. Nature reviews Molecular cell biology." In: *Nature reviews Molecular cell biology.* 10, pp. 637–643.

Felsenfeld G, et al. (2004). "Chromatin Boundaries and Chromatin Domains." In: *Skelet Muscle.* 69, pp. 245–250.

Fickett JW, et al. (2000). "Discovery and modeling of transcriptional regulatory regions." In: *Current Opinion in Biotechnology* 11, pp. 19–24.

Fisher WW Li JJ, et al (2012). "DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila." In: *Proc Natl Acad Sci U S A* 109, pp. 21330–21335.

Fong AP Yao Z, et al. (2015). "Conversion of MyoD to a neurogenic factor binding site specificity determines lineage." In: *Cell Rep.* 31, pp. 1937–1946.

Fudenberg G Abdennur N, et al. (2018). "Emerging Evidence of Chromosome Folding by Loop Extrusion." In: *Cold Spring Harb Symp Quant Biol.* 2, pp. 34710–34710.

Fujita K Iwaki M, et al. (2016). "Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA." In: *Nature Communications.* 7, pp. 13788–13789.

Fulco CP Munschauer M, et al. (2016). "Systematic mapping of functional enhancer-promoter connections with CRISPR interference." In: *Science.* 354, pp. 769–773.

Garrity PA, et al. (1990). "Tissue-specific expression from a compound TATA-dependent and TATA-independent promoter." In: *Mol. Cell Biol.* 10, pp. 5646–5654.

Garrity PA Wold BJ, et al. (1992). "Effects of different DNA polymerases in ligation-mediated PCR enhanced genomic sequencing and in vivo footprinting." In: *PNAS.* 89, pp. 1021–1025.

Ghirlando R Felsenfeld G, et al. (2016). “CTCF making the right connections.” In: *Genes Dev.* 30, pp. 881–891.

Goldhamer Dj Brunk BP, et al. (1995). “Embryonic activation of the myoD gene is regulated by a highly conserved distal control element.” In: *Development.* 121, pp. 637–649.

Gossett LA, et al. (1989). “A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes.” In: *Mol. Cell Biol.* 9, pp. 5022–33.

Gottgens B, et al. (2000). “Analysis of vertebrate SCL loci identifies conserved enhancers.” In: *Nature Biotechnology* 18, pp. 181–186.

Grebbin BM Schulte D, et al. (2017). “PBX1 as Pioneer Factor A Case Still Open.” In: *Front Cell Dev Biol.* 5, pp. 9–10.

Gutman A Gilthorpe J, et al. (1994). “Multiple positive and negative regulatory elements in the promoter of the mouse homeobox gene Hoxb-4.” In: *Development.* 14, pp. 8143–8154.

Hadchouel J, et al. (2000). “Modular long-range regulation of Myf5 reveals unexpected heterogeneity between skeletal muscles in the mouse embryo.” In: *Development.* 127, pp. 4455–4467.

Hadchouel J Carvajal JJ, et al. (2003). “Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by elements dispersed throughout the locus.” In: *Development.* 130, pp. 3415–3426.

Han GC Vinayachandran V, et al. (2015). “Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution.” In: *Mol Cell Biol.* 36, pp. 157–172.

Hardison R, et al. (1997a). “Locus control regions of mammalian beta globin gene clusters combining phylogenetic analyses and experimental results to gain functional insights.” In: *Gene.* 205, pp. 73–94.

– (1997b). “Long Human Mouse Sequence Alignments Reveal Novel Regulatory Elements.” In: *Genome Res.* 7, pp. 959–966.

Hasty P Bradley A, et al. (1993). “Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene.” In: *Nature.* 364, pp. 501–506.

Hatzi K, et al. (2013). “A hybrid mechanism of action for BCL6 in B cells defined by formation of functionally distinct complexes at enhancers and promoters..” In: *Cell Rep.* 5, pp. 578–588.

He XJ Hsu YF, et al. (2009). “An effector of RNA directed DNA methylation in Arabidopsis is an ARGONAUTE 4 and RNA binding protein.” In: *Cell.* 137, pp. 498–508.

Hernandez JM, et al. (2017). “The myogenic regulatory factors determinants of muscle development cell identity and regeneration.” In: *Cell Dev. Biol.* 72, pp. 10–18.

Hnisz D, et al. (2013). “Super Enhancers in the Control of Cell Identity and Disease.” In: *Cell.* 155, pp. 934–947.

Hockheimer A Tjian R, et al (2003). “Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression.” In: *Genes and Development.* 17, pp. 1309–1320.

Hong JW Hendrix DA, et al. (2008). “Shadow enhancers as a source of evolutionary novelty.” In: *Science.* 321, p. 1314.

Horlick RA, et al. (1989). “The upstream muscle-specific enhancer of the rat muscle creatine kinase gene is composed of multiple elements.” In: *Mol. Cell Biol.* 9, pp. 2396–2413.

– (1990). “Brain and muscle creatine kinase genes contain common AT-rich recognition protein-binding regulatory elements.” In: *Mol. Cell Biol.* 10, pp. 4826–36.

HoughEvans BR, et al. (1977). “Appearance and persistence of maternal RNA sequences in sea urchin development.” In: *Dev. Biol.* 60, pp. 258–77.

Howard ML, et al. (2004). “cis-Regulatory control circuits in development.” In: *Developmental Biol.* 271, pp. 109–118.

Hu P, et al. (2008). “Codependent activators direct myoblast-specific MyoD transcription.” In: *Dev. Cell* 15, pp. 534–546.

Huang L, et al. (2000). “Involvement of Tcf Lef in establishing cell types along the animal-vegetal axis of sea urchins.” In: *Development genes and Evolution* 210, pp. 73–81.

Hwang SS, et al. (2016). “Role of OCT-1 and partner proteins in T cell differentiation.” In: *Biochim Biophys Acta.* 1859, pp. 825–831.

Hyndman BD, et al. (2012). “E2A proteins enhance the histone acetyltransferase activity of the transcriptional co-activators CBP and p300.” In: *Biochem Biophys Acta* 1819, pp. 446–453.

Inoue F Ahituv N, et al. (2015). “Decoding enhancers using massively parallel reporter assays.” In: *Genomics.* 106, pp. 159–164.

Inoue F Kircher M, et al. (2016). “A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity.” In: *Genome Res.* 27, pp. 38–52.

Istrail S, et al. (2005). “Logic functions of the genomic cis-regulatory code.” In: *PNAS* 102, pp. 4954–4959.

Jaynes JB, et al. (1988). “The muscle creatine kinase gene is regulated by multiple upstream elements, including a muscle-specific enhancer.” In: *Mol. Cell Biol.* 8, pp. 62–70.

Johnson DS, et al. (2007). “Genome-wide mapping of in vivo protein-DNA interactions..” In: *Science* 316, pp. 1497–502.

Johnson JE, et al. (1989). “Muscle creatine kinase sequence elements regulating skeletal and cardiac muscle expression in transgenic mice.” In: *Mol. Cell Biol.* 9, pp. 3393–9.

Kabadi AM, et al. (2015). “EnhancedMyoD-induced transdifferentiation to a myogenic lineage by fusion to a potent transactivation domain.” In: *ACS Synth Biol.* 4, pp. 689–699.

Kaneko H Shimizu R, et al. (2010). “GATA factor switching during erythroid differentiation.” In: *Hematology.* 17, pp. 163–168.

Katayama S, Tomaru Y et al. (2005). “Antisense transcription in the mammalian transcriptome.” In: *Science.* 309, pp. 1564–1566.

Keeton AB Amsler MO, et al. (2002). “Insulin signal transduction pathways and insulin-induced gene expression.” In: *J Biol Chem.* 13, pp. 48565–48573.

Kellis M Hardison RC, et al (2014). “Defining functional DNA elements in the human genome.” In: *Proc Natl Acad Sci U S A* 111, pp. 6131–6138.

Kheradpour P Ernst J, et al. (2013). “Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.” In: *Genome Res.* 23, pp. 800–811.

KiefferKwon KR, et al. (2013). “Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation.” In: *Cell* 155, pp. 1507–1520.

Kim MK, et al. (1996). “A soluble transcription factor Oct-1 is also found in the insoluble nuclear matrix and possesses silencing activity in its alanine-rich domain.” In: *Mol Cell Biol.* 16, pp. 4366–4377.

Kinney JB Murugan A, et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.” In: *Proc Natl Acad Sci U S A.* 107, pp. 9158–9163.

Kirchhamer CV, et al. (1996). “Developmental expression of synthetic cis-regulatory systems composed of spatial control elements from two different genes.” In: *PNAS* 93, pp. 13849–13854.

Koch F Andrau JC, et al. (2011). “Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity.” In: *Transcription.* 2, pp. 263–268.

Kophengnavong T Michnowicz JE, et al. (2000). "Establishment of Distinct MyoD E2A and Twist DNA Binding Specificities by Different Basic Region-DNA Conformations." In: *Mol Cell Biol.* 20, pp. 261–272.

Korkmaz G Lopes R, et al. (2016). "Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9." In: *Nat Biotechnol.* 34, pp. 192–198.

Kouzarides T, et al. (2007). "Chromatin modifications and their function." In: *Cell.* 128, pp. 693–705.

Kumar D, et al. (2007). "Id3 is a direct transcriptional target of Pax7 in quiescent satellite cells." In: *Mol Cell Biol.* 20, pp. 3170–3177.

Kvon EV, et al. (2016). "Progressive Loss of Function in a Limb Enhancer during Snake Evolution." In: *Cell.* 167, pp. 633–642.

Kwasnieski JC Fiore C, et al. (2014). "High-throughput functional testing of ENCODE segmentation predictions." In: *Genome Res.* 24, pp. 1595–1602.

Kwasnieski JC Mogno I, et al. (2012). "Complex effects of nucleotide variants in a mammalian cis-regulatory element." In: *Proc Natl Acad Sci U S A.* 109, pp. 19498–19503.

Laat W, et al. (2013). "CTCF: the protein, the binding partners, the binding sites and their chromatin loops." In: *Philos Trans R Soc Lond B Biol Sci.* 368, pp. 1620–1621.

Landt SG Marinov GK, et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." In: *Genome Res.* 22, pp. 1813–1831.

Langlands K Yin X, et al. (1997). "Differential Interactions of Id Proteins with Basic-Helix-Loop-Helix Transcription Factors." In: *Journal of Biol. Chem.* 272, pp. 19785–19793.

Lawinger P, et al. (2000). "The neuronal repressor REST/NRSF is an essential regulator in medulloblastoma cells." In: *Nat Med.* 6, pp. 826–831.

Lee TC, et al. (1994). "Bifunctional transcriptional properties of YY1 in regulating muscle actin and c-myc gene-expression during myogenesis." In: *Oncogene* 9, pp. 1047–52.

Levine M Tjian R, et al. (2003). "Transcription regulation and animal diversity." In: *Nature.* 424, pp. 147–151.

Li B, et al. (2007). "The role of chromatin during transcription." In: *Cell.* 128, pp. 707–719.

Liu CJ Ding B, et al. (2002). "The MyoD inducible p204 protein overcomes the inhibition of myoblast differentiation by Id proteins." In: *Mol Cell Biol.* 22, pp. 2893–2905.

Liu Y, et al. (2017). "Functional assessment of human enhancer activities using whole-genome STARR-sequencing." In: *Genome Biol.* 18, pp. 1–13.

Long L, et al. (2013). “A far downstream enhancer for murine Bcl11b controls its T-cell specific expression.” In: *Blood* 122, pp. 902–911.

Longabaugh WJ Davidson EH, et al. (2005). “Computational representation of developmental genetic regulatory networks.” In: *Developmental Biology.* 283, pp. 1–16.

Lu Y, et al. (2002). “The Basic Helix–Loop–Helix Domain of the E47 Transcription Factor Requires Other Protein Regions for Full DNA Binding Activity.” In: *Biochem and Biophys Res.* 290, pp. 1521–1528.

Lyons DB, et al. (2013). “An epigenetic trap stabilizes singular olfactory receptor expression.” In: *Cell.* 154, pp. 325–336.

Ma JY, et al. (2013). “Maternal factors required for oocyte developmental competence in mice: transcriptome analysis of non-surrounded nucleolus (NSN) and surrounded nucleolus (SN) oocytes.” In: *Cell Cycle* 12, pp. 1928–38.

Macquarrie KL Yao Z, et al. (2013). “Genome-wide binding of the basic helix-loop-helix myogenic inhibitor musculin has substantial overlap with MyoD implications for buffering activity.” In: *Skelet Muscle.* 1, p. 26.

Magklara A, et al. (2011). “An epigenetic signature for monoallelic olfactory receptor expression.” In: *Cell.* 145, pp. 555–570.

Maricque BB Dougherty JD, et al. “A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells.” In: *Nucleic Acids Res*, pp. 1–2.

Markenscoff E, et al. (2014). “Enhancer interaction networks as a means for singular olfactory receptor expression.” In: *Cell.* 159, pp. 543–557.

Mautner J, et al. (1995). “Identification of two enhancer elements downstream of the human c-myc gene.” In: *Nucleic Acids Res.* 11, pp. 72–80.

Maves L, et al. (2007). “Pbx homeodomain proteins direct Myod activity to promote fast-muscle differentiation..” In: *Development* 134, pp. 3371–3382.

May D Blow MJ, et al. (2011). “Large-scale discovery of enhancers from human heart tissue.” In: *Nat Genet.* 44, pp. 89–93.

McKnight SL Kingsbury R, et al. (1982). “Transcriptional control signals of a eukaryotic protein-coding gene.” In: *Science.* 217, pp. 316–324.

Melnikov A Murugan A, et al. (2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.” In: *Nat Biotechnol.* 30, pp. 271–277.

Mikhaylichenko O Bondarenko V, et al. (2018). “The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription.” In: *Genes and Development.* 32, pp. 42–57.

Miner JH, et al. (1990). “Herculin, a fourth member of the MyoD family of myogenic regulatory genes.” In: *PNAS.* 87, pp. 1089–1093.

Molkentin JD Black BL, et al. (1996). “Mutational analysis of the DNA binding dimerization and transcriptional activation domains of MEF2C.” In: *Molecular and Cellular Biology.* 16, pp. 2627–2636.

Moncaut N Rigby PW, et al. (2013). “Dial M(RF) for myogenesis.” In: *FEBS J.* 280, pp. 3980–3990.

Moore J Purcaro MJ, et al. (2018). “ENCODE Phase III Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse.” In: *Submitted.* Pp. 1–2.

Mordan LJ, et al. (1982). “Isolation of methylcholanthrene-initiated C3H/10T1/2 cells by inhibiting neoplastic progression with retinyl acetate.” In: *Carninogenesis.* 3, pp. 279–285.

Morokuma J, et al. (2003). “Acis-regulatory element within the 5 flanking region of arylsulfatase gene of sea urchin Hemicentrotus pulcherrimus.” In: *Dev. Growth and Diff.* 39, pp. 469–476.

MouraNeto V, et al. (1996). “A 28-bp negative element with multiple factor-binding activity controls expression of the vimentin-encoding gene.” In: *Gene* 168, pp. 261–8.

Mueller PR Wold BJ, et al. (1989). “In vivo footprinting of a muscle specific enhancer by ligation mediated PCR.” In: *Science.* 246, pp. 780–786.

Muerdter F Boryń M, et al. (2018). “Resolving systematic errors in widely used enhancer activity assays in human cells.” In: *Nat Methods.* 15, pp. 141–149.

Murtha M Tokcaer-Keskin Z, et al. (2014). “FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.” In: *Nat Methods.* 11, pp. 559–565.

Neuhold LA, et al. (1993). “HLH forced dimers: tethering MyoD to E47 generates a dominant positive myogenic factor insulated from negative regulation by Id.” In: *Cell* 74, pp. 1033–1042.

OKane CJ, et al. (1987). “Detection in situ of genomic regulatory elements in Drosophila.” In: *PNAS.* 84, pp. 9123–9127.

Osterwalder M, et al. (2018). “Enhancer redundancy provides phenotypic robustness in mammalian development.” In: *Nature* 554, pp. 239–243.

Oti M Falck J, et al. (2016). “CTCF-mediated chromatin loops enclose inducible gene regulatory domains.” In: *BMC Genomics.* 17, p. 252.

Patwardhan RP Hiatt JB, et al. (2012). “Massively parallel functional dissection of mammalian enhancers in vivo.” In: *Nat Biotechnol.* 30, pp. 265–270.

Patwardhan RP Lee C, et al. (2009). “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.” In: *Nature.* 27, pp. 1173–1175.

Pepke S Wold BJ, et al. (2009). “Computation for ChIP-seq and RNA-seq studies.” In: *Nat Methods.* 11, pp. 22–32.

Peter IS Davidson EH, et al. (2011). “Evolution of Gene Regulatory Networks that Control Embryonic Development of the Body Plan.” In: *Cell.* 144, pp. 970–985.

Pirkmajer S Chibalin AV, et al. (2011). “Serum starvation caveat emptor.” In: *Am J Physiol Cell Physiol.* 13, pp. 272–279.

Ponting CP, et al. (2011). “What fraction of the human genome is functional?” In: *Genome Res.* 21, pp. 1769–1776.

Puri PL, et al. (1997). “Differential roles of p300 and PCAF acetyltransferases in muscle differentiation.” In: *Mol. Cell* 1, pp. 35–45.

Queen C, et al. (1984). “Fine mapping of an immunoglobulin gene activator.” In: *Mol Cell Biol.* 4, pp. 1042–1049.

RadaIglesias A, Bajpai R et al. (2011). “A unique chromatin signature uncovers early developmental enhancers in humans.” In: *Nature.* 470, pp. 279–283.

Rahl A, et al. (2010). “Myc Regulates Transcriptional Pause Release.” In: *Cell.* 141, pp. 432–445.

Riethoven JJ, et al. (2010). “Regulatory regions in DNA promoters, enhancers, silencers, and insulators.” In: *Methods Mol Biol.* 674, pp. 33–42.

Rosenthal N Berglund EB, et al. (1990). “A highly conserved enhancer downstream of the human MLC1-3 locus is a target for multiple myogenic determination factors.” In: *Development.* 18, pp. 6239–6246.

Rossant J, et al. (1992). “Of fin and fur: mutational analysis of vertebrate embryonic development.” In: *Genes Dev.* 6, pp. 1–13.

Rylski M Welch JJ, et al. (2003). “GATA-1-mediated proliferation arrest during erythroid maturation.” In: *Mol Cell Biol.* 23, pp. 5031–5042.

Sartorelli V, et al. (1997). “Molecular mechanisms of myogenic coactivation by p300: direct interaction with the activation domain of MyoD and with the MADS box of MEF2C.” In: *Mol. Cell Biol.* 17, pp. 1010–1026.

Sartorelli V Puri PL, et al. (1999). “Acetylation of MyoD directed by PCAF is necessary for the execution of the muscle program.” In: *Mol Cell.* 4, pp. 725–734.

Schor IE, Degner JF et al. (2017). “Promoter shape varies across populations and affects promoter evolution and expression noise.” In: *Nat Genet.* 49, pp. 550–558.

Schwartz YB Pirrotta V, et al. (2007). “Polycomb silencing mechanisms and the management of genomic programmes.” In: *Nat Rev Genet.* 8, pp. 9–22.

Serizawa S, et al. (2003). “Negative Feedback Regulation Ensures the One Receptor-One Olfactory Neuron Rule in Mouse.” In: *Science.* 302, pp. 2088–2094.

Siles L SanchezTillo E, et al. (2013). "ZEB1 imposes a temporary stage-dependent inhibition of muscle gene expression and differentiation via CtBP-mediated transcriptional repression." In: *Mol Cell Biol.* 33, pp. 1368–1382.

Soleimani VD Punch VG, et al. (2012). "Transcriptional dominance of Pax7 in adult myogenesis is due to high-affinity recognition of homeodomain motifs." In: *Dev Cell.* 22, pp. 1208–1220.

Soufi A, et al. (2015). "Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming." In: *Cell* 161, pp. 555–568.

Springer PS, et al. (2000). "Gene Traps." In: *Plant Cell.* 12, pp. 1007–1020.

Summerbell D, et al. (2000). "The expression of Myf5 in the developing mouse embryo is controlled by discrete and dispersed enhancers specific for particular populations of skeletal muscle precursors." In: *Development.* 127, pp. 3745–3757.

Taft RJ, et al. (2009). "Tiny RNAs associated with transcription start sites in animals." In: *Nat Genet.* 41, pp. 572–578.

Tai PWL, et al. (2011). "Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer." In: *Skeletal Muscle.* 1, pp. 1–25.

Tapscott SJ, . (2005). "The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription." In: *Development.* 132, pp. 2685–2695.

Taylor SM Jones PA, et al. (1979). "Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine." In: *Cell.* 17, pp. 771–779.

Teboul L Hadchouel J, et al. (2002). "The early epaxial enhancer is essential for the initial expression of the skeletal muscle determination gene Myf5 but not for subsequent, multiple phases of somitic myogenesis." In: *Development.* 129, pp. 4571–4580.

Thurman RE Rynes E, et al (2012). "The accessible chromatin landscape of the human genome." In: *Nature.* 489, pp. 75–82.

Trinklein ND Aldred SF, et al. (2004). "An Abundance of Bidirectional Promoters in the Human Genome." In: *Genome Research.* 14, pp. 62–66.

Tripic T Deng W, et al. (2009). "SCL and associated proteins distinguish active from repressive GATA transcription factor complexes." In: *Blood.* 113, pp. 2191–2201.

Tsang AP Visvader JE, et al. (1997). "FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation." In: *Cell.* 90, pp. 109–119.

Valdez MR, et al. (2000). "Failure of Myf5 to support myogenic differentiation without myogenin MyoD and MRF4." In: *Dev. Biol.* 219, pp. 287–298.

Vermeulen M, Timmers HT. (2010). "Grasping trimethylation of histone H3 at lysine 4." In: *Epigenomics* 2, pp. 395–406.

Visel A Blow MJ, et al. (2009). “ChIP-seq accurately predicts tissue-specific activity of enhancers.” In: *Nature.* 457, pp. 854–858.

Visel A Minovitsky S, et al. (2007). “VISTA Enhancer Browser-a database of tissue-specific human enhancers.” In: *Nucleic Acids Res* 35, pp. D88–92.

Vockley CM DIppolito AM, et al. (2016). “Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome.” In: *Cell.* 166, pp. 1269–1281.

Wang DG, et al. (1995). “Maternal and embryonic provenance of a sea urchin embryo transcription factor SpZ121.” In: *Mol. Mar. Biol. Biotechn.* 4, pp. 148–153.

Wang Z Zhu T, et al. (2005). “Adeno-associated virus serotype 8 efficiently delivers genes to muscle and heart.” In: *Nat Biotechnol.* 23, pp. 321–328.

Weihner H, et al. (1984). “An enhancer sequence from bovine papilloma virus DNA consists of two essential regions.” In: *Nucleic Acids Res.* 12, pp. 2901–2916.

Weintraub H, et al. (1989). “Activation of muscle-specific genes in pigment nerve fat liver and fibroblast cell lines by forced expression of MyoD..” In: *PNAS* 86, pp. 5434–5438.

– (1991). “Muscle-specific transcriptional activation by MyoD.” In: *Genes Dev.* 5, pp. 1377–1386.

Weintraub H Davis R, et al. (1990). “MyoD binds cooperatively to two sites in a target enhancer sequence: occupancy of two sites is required for activation.” In: *PNAS* 87, pp. 5623–5627.

Weiss MJ Yu C, et al. (1997). “Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line.” In: *Mol Cell Biol.* 17, pp. 1642–1651.

West A, et al. (2002). “Insulators: many functions many mechanisms.” In: *Genes Dev.* 16, pp. 271–288.

White MA Myers CA, et al. (2013). “Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks.” In: *PNAS.* 16, pp. 11952–11957.

Wilderme M, et al. (1995). “Transcription complex stability and chromatin dynamics in vivo.” In: *Nature.* 377, pp. 209–213.

Wong ES, et al. (2015). “Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals.” In: *Genome Res.* 25, pp. 167–178.

Wright WE, et al. (1991). “Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site.” In: *Mol Cell Biol.* 11, pp. 4104–4110.

Wright WE Sassoon DA, et al. (1993). “Myogenin, a factor regulating myogenesis, has a domain homologous to MyoD.” In: *Cell.* 56, pp. 607–617.

Wu C, et al. (2017). “The transcription factor musculin promotes the unidirectional development of peripheral Tregcells by suppressing the TH2 transcriptional program.” In: *Nature Immunology* 18, pp. 344–353.

Wu JQ, et al. (2008). “RNA polymerase II stalling: loading at the start prepares genes for a sprint.” In: *Genome Biol.* 9, pp. 220–221.

Wu W Morrissey CS, et al. (2014). “Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis.” In: *Genome Res.* 24, pp. 1945–1962.

Yaffe D, Saxel O. (1977). “Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle.” In: *Nature.* 270, pp. 725–727.

Yao Z Fong AP, et al. (2013). “Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites.” In: *Skeletal Muscle.* 3, pp. 1–8.

Yao Z Macquarrie KL, et al. (2014). “Discriminative motif analysis of high-throughput dataset.” In: *Bioinformatics.* 15, pp. 775–783.

Yee SP Rigby PW, et al. (1993). “The regulation of myogenin gene expression during the embryonic development of the mouse.” In: *Genes Dev.* 7, pp. 1277–1289.

Yin Y, et al. (2017). “Impact of cytosine methylation on DNA binding specificities of human transcription factors.” In: *Science.* 356, pp. 2239–2240.

Yoon JK, et al. (1997). “Different MRF4 knockout alleles differentially disrupt Myf5 expression: cisregulatory interactions at the MRF4 Myf5 locus.” In: *Dev. Biol.* 188, pp. 349–362.

Yu M Riva L, et al. (2009). “Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis.” In: *Mol Cell.* 36, pp. 682–695.

Yuan W, et al. (1996). “Human p300 protein is a coactivator for the transcription factor MyoD.” In: *J. Biol. Chem.* 271, pp. 9009–9013.

Yue F Cheng Y, et al. (2014). “A comparative encyclopedia of DNA elements in the mouse genome.” In: *Nature.* 515, pp. 355–364.

Yuh CH, et al. (1996). “Modular cis-regulatory organization of Endo16 a gut-specific gene of the sea urchin embryo.” In: *Development* 122, pp. 1069–1082.

Yuh L, et al. (2003). “MyoR is expressed in nonmyogenic cells and can inhibit their differentiation.” In: *Exp Cell Res.* 289, pp. 162–173.

– (2004). “The bHLH protein MyoR inhibits the differentiation of early embryonic endoderm.” In: *Differentiation* 72, pp. 341–347.

Yun K, et al. (1996). “Skeletal muscle determination and differentiation: story of a core regulatory network and its context..” In: *Curr Opin. Cell Biol.* 8, pp. 877–898.

Zambetti GP, et al. (1992). “Wild-type p53 mediates positive regulation of gene expression through a specific DNA sequence element.” In: *Genes Dev* 6, pp. 1143–52.

Zammit PS Carvajal JJ, et al. (2004). “Myf5 expression in satellite cells and spindles in adult muscle is controlled by separate genetic elements.” In: *Dev Biol.* 15, pp. 454–465.

Zhang JA Mortazavi A, et al. (2012). “Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity.” In: *Cell.* 149, pp. 467–482.

Zhou J Levine M, et al. (1999). “A novel cis-regulatory element, the PTS, mediates an anti-insulator activity in the Drosophila embryo.” In: *Cell.* 99, pp. 567–575.

Zomorodipour A, et al. (2017). “Position dependence of an enhancer activity of the human betaglobin intron ii within a heterologous gene.” In: *J of Mol Med.* 1, pp. 1–12.

Chapter 2

TRANSFECTION ASSAY DEVELOPMENT

2.1 Introduction

Skeletal Muscle Actin (Acta1) is one of the most prominently detected transcripts (2855 FPKM) expressed selectively, like myogenin (819 KPKM) after the onset of differentiation. (Figure 1.1) My earliest enhancer assay tests used several mutants produced by the Sharp Lab at CSU-LA for the promoter regions of Acta1 and ID2 (78 FPKM, myoblast selective gene) region driving a luciferase reporter. The mutants from the Sharp Lab, done in a conventional luciferase assay, show that much of the enhancer activity for this fragment of DNA depends on the binding of myogenic factors at the Eboxes 5 and 6 (Figure 2.1a and c).

These cell type selective promoter proximal regions, and relative scale of activity of these Acta1 mutant constructs (Figure 2.1a and b) allowed me to refine an assay that was both technically reproducible, sensitive over a wide range and able to function in both the myoblast and myocyte cell types. After seemingly unending testing of different conditions, I was able to replicate their assay successfully in a 96 well format compatible for testing in the order of hundreds of constructs. (Figure 2.1b).

In addition, to validate the assay, I tested 39 human promoter constructs selected from the available switchgear collection that showed myogenin and MyoD binding in the corresponding mouse regions. (ENCODE, 2007) Only 12 (30%) of the human gene promoters reported as significantly active, and only 5 matched the expected activity pattern based on their associated gene transcript behavior. (Figure 2.1d) Since conservation of the occupancy motif was not required for selection, actual expectations for activity remain uncertain.

The different behavior of these promoter elements underscores the importance of comparing the relative activity of candidate enhancers against a single promoter element. (Section 1.24) Indeed, testing each enhancer against their native promoter, while ideal, would result in data that is not directly comparable as select promoters might be tightly gated at one of the two differential stages; resulting in a negative readout, even when the candidate enhancer itself is functional.

2.2 Structure of a ChIP-seq measurement, enhancer assay development and candidate selection

Our early knowledge of the muscle transcription factor network mapped to about 15,000 occupied sites out of 2.2X10E6 CAGSTG genome wide. Even though less than ~1% of all available Eboxes were found occupied and even after the motif was refined to an RRCAGSTG (and occupancy at 1.5% of such Eboxes in the genome (Kirilusha A, 2014 thesis)) We thought it possible that several elements identified by ChIP are either inert or simply "parking" sites for excess factors in the nucleus. (Casey BH, 2018) This seemingly simple task of assigning enhancer function to candidate signatures remains difficult today, even with many biochemical maps that now span CTCF, several cofactors (E proteins, Mef2 and PBX1 in our case); EP300 and histone marks.

The transient transfection assay I developed provides a baseline comparative measurement of relative enhancer activity over a wide range of activity (Figure 2.1), however it also allows for simple classification of several selections of candidate enhancers in a binary score as functional or not. These additional ChIP-seq measurements can be used individually or in combination with chromatin signatures to determine their predictive qualities of enhancer function to provide a first tier functional expectation.

In order accurately assess the predictive nature for enhancer activity of TF occu-

pancy, we had to randomly select elements from the genome which covered both low, medium and high signal from the myogenin ChIP-Seq measurement. (Figure 2.2) This set of candidate elements allowed us to answer outstanding questions on the functional contribution of each class of ChIP-seq signal in an unbiased way, instead of relying on the 80% functional top signals for drawing expectations. (Visel A, 2009; Visel A, 2007)

We also selected elements from several genes key to myogenesis to provide a bridge to the well characterized literature on muscle enhancers (Reviewed and contrasted in Chapters 1 and 5). These loci are clearly special compared to the "run of the mill" loci in the genome because of the number of elements they contain and connect to but not in terms of biochemical strength (which ranges similarly to the genome-wide distribution). (Fisher K, 2017 thesis) We sampled a small set of additional genes from complex loci with matching RNA expression trajectories to our MRF genes, in the hopes to gain the numbers necessary to contrast these elements to their randomly selected counterparts in the genome. This locus based selection is separate from the candidate Enhancers used in this chapter and are discussed in chapter 4.

All the candidate elements from both selections were cloned by Switchgear Genomics upstream of a commonly used basal TK promoter driving a high turnover luciferase reporter gene (h-PEST). (McKnight SL, 1982) Figure 2.3 is a conceptual representation of the selection process of candidate Enhancer elements from the genome at relatively low, medium and high signal myogenin occupancy sites. The actual experimental sampling in each signal class is illustrated in Figure 2.2 (right panel) together with the genome wide distribution (left panel).

I transfected each element in four cell types representing a mesodermal precursor (to provide a system without either MyoD or myogenin), mock differentiated mesodermal precursor (to control for serum and insulin sensitive enhancers), undif-

ferentiated myoblast and a set of myoblasts that undergo a differentiation stimulus bringing them to an early myocyte state. (Figure 2.4)

My assay partially draws strength from this differentiable multi-cell type design to isolate distinct expression patterns of genes and study the modulatory function of specific elements under these separate conditions. This assay design also allows us to set functional expectations for the large numbers of myoblast MyoD occupied elements, many of which go on to be co-occupied at the myocyte state by both MyoD and myogenin compared to the counterparts that are selectively occupied only in the myocyte. (Chapter 5; Figure 5.1) This design also allowed us to draw a contrast and set functional expectations for the relatively smaller, but potentially functionally important set of elements which are highly occupied in the myoblast by MyoD but are excluded from occupancy in the myocyte. (discussed later in this Chapter) Unfortunately, a design spanning so many questions quickly led to a design with several hundred elements to be tested across 4 cell types.

This type of now medium-throughput enhancer assays generally relies on a liposomal reagents (electroporation was expensive and not available in 96 well format when I stated these experiments) which require the cells to be dividing in order for the DNA to be delivered to the nucleus. This created a logistical issue because myocytes, where myogenin is expressed, do not divide and are fusing together to form the multiple nucleus containing myocytes, meaning that even if electroporation was possible each cell (assuming similar external surface area) might have different proportions of transfected DNA going to each nucleus (depending on n of nuclei).

I opted to transfect the DNA at the myoblast state at a relatively low density and assay at a later time-point for the myocyte to allow the cells to differentiate. This also should have allowed me the benefit of matching the established “seed and transfect” single step protocol suggested by Switchgear and successfully used by

our collaborators at HudsonAlpha in K562 cells.

Unfortunately when I tested this process on C2C12 myoblasts it resulted in ball-like structures rather than our cells settling in a mono-layer on the bottom of the well. This created an issue because the onset of differentiation occurs spontaneously in C2C12s myoblasts when they reach confluence (such as in a ball-like structure) making a controlled progression to the onset of differentiation that matched our biochemical measurements impossible.

These transient assays are also known to begin loosing their reproducible signal after approximately 60 hours post-transfection (which I confirmed in a test using a small set of test elements) creating a narrow window of opportunity.

This combination of these technical limitations created a non-trivial challenge as the differentiation stimulus requires up to 16 hours after the cells become confluent to reach significant levels of myogenin transcript meaning one does not want to assay prior to the 24 hour time-point. (Unpublished Wold Lab data)

In order to work around these constraints I had to plate cells at a set density, let them adhere, transfect and (where applicable) wait for them to achieve confluence, then differentiate. I also had to adjust serum conditions to be higher (20% FBS - something also known to affect the efficiency of these reagents) so that the cells grow faster and allow me to create an assay that barely fit within these parameters.

Unfortunately working by hand also is a slow process, especially to test what when counting technical and biological replicates across cell lines is in the order of 50,000 wells being plated, transfected and media changed to induce differentiation. These timing differences which span in the order of many hours alone created a logistical nightmare.

In order to reproducibly test this many enhancers I devised a robotic transfection assay, which closely mimics the manual assay I developed in 96 well plates. (Figure

2.5) The use of robotics resulted in reproducible timing, fast handling of plates (mere minutes instead of hours), and lower loss of cells when changing media for the differentiation, as the robot can remove and dispense reagents to plates in a way and speed that a human simply cannot. These many advantages; along with allowing for a large set of negative controls resulted in a well controlled, reliable assay with low technical error. 2.6

2.3 Enhancer Thresholding

High confidence (IDR=0.02) myogenin data which has been processed by the modern Encode 3 pipeline yields a distribution of signal sites (Figure 2.2 Left Panel) across the genome where the majority of signal is contained in the medium to low signal ranges similar to our original ChIP measurement for this factor. (Figure 1.3) We had used this original ChIP measurement to select our candidates, and the IDR filtering (and 1PCR libraries) excluded a surprisingly large set of elements. (discussed later in Chapter) Even with these exclusions, our sampling was still deep enough to allow us to predict activity for the full signal range of myogenin occupancy. (Figure 2.2 Right Panel)

In order to ensure specificity of activity I had to devise a set of controls. I used the set of literature derived t-cell and neural enhancers as a specificity control. (Figure 1.2 - Purple) These regions provide an identical occupancy motif, and are centered on the CAGSTG motif(s) contained, testing whether assay function is specific for occupied regions. In addition I selected a small set of unoccupied muscle class Eboxes (RRCAGSTG) from each key muscle loci to further establish a reliable baseline of activity the assay.

2.4 Additional Results

The results of the transfection assay for the myocyte state are summarized in Figure 2.6 (Top Panel). The negative control elements are in gray on the left, showing no

function across the entire set of tested elements, whether they were selected from characterized enhancers or non occupied sites in key loci (where the chromatin should be accessible). This indicates that our assay is specific for function in elements that show occupancy by TFs. It is important to note that two of the originally selected negative control elements were actually functional, and stood out from the distribution of the negatives. These were later found to have significant DHS sites which indicate the presence of another TF that was not yet measured in our cells driving their function. As a result these were excluded from the negative control set.

The experimental elements are ranked based on their corresponding ChIP-Seq signal within the region. Enhancer function is present throughout the range of signal and that while strong enhancers are enriched in the top 200 as expected, that the ratio of activity is stable across the higher 2000 sites, all the way down into the low signal range of ChIP-Seq. (Figure 2.6 Bottom Panel) Overall 40% of sites are found to be active in the myocyte. This can be extrapolated to predict that ~5500 of the 15000 myogenin occupied sites are expected to be active genome-wide.

Overall this data tracks well with the activity found in the comparable portions of signal used in other studies since published in the literature. The original EP300 study found ~80% enhancer function among the top ChIP-Seq peaks, whereas we found ~75% function. (Visel A, 2007; Visel A, 2009) In the top 2000 signal occupancy sites we found a comparable function at ~53% to our companion study which derived candidate Enhancers from Tal1 and Gata1 co-occupancy (54% active). (Dogan N, 2015) While we found many inactive elements in high and medium signal elements these may function as poised enhancer sites, or simply sites that function when brought into physical proximity with other sites in the genome. (Further discussed in Chapter 5) Overall we find 40% of candidate Enhancers to be functional compared to <25% as measured by Starr-Seq, which may be suffering

from sensitivity or cross-talk issues. (Arnold CD, 2013; Muerdter F, 2018)

Figure 2.7 presents the overall contribution to the number of genome wide enhancer by signal class in a ChIP-Seq experiment. This means that while the strength and proportion of enhancers in the lowest signal class may be lower, this class of elements far outnumber their counterparts in the higher signal classes across the genome. Even just the lowest signal class accounts for approximately approximately half of active myogenin occupancy sites in the genome!

2.5 Activity of MyoD sites occupied exclusively undifferentiated C2C12

MyoD is a transcription factor that is expressed in both myoblast and in myocyte cells. (Figure 1.1) I selected a set of sites which are exclusively occupied by MyoD in the myoblast state in the hopes that these sites are a class of myoblast enhancers that are then actively repressed to prevent their function in the myocyte (n=16). The results of the myoblast assay in Figure 2.8 (blue) are unambiguous, most of these do not function in the myoblast. One enhancer did function when cloned upstream of a TK promoter but only significantly in the differentiated myocyte state (red). These results, along with those of the entire population of myoblast MyoD occupied elements indicates that little function if present at these sites, with the exception of a few validated enhancers from known loci. (further discussed in Chapter 5)

2.6 Enhancer function analyzed against EP300 signal

The Histone Acetyl Transferase (HAT) EP300 has been associated with functional elements in several enhancer studies. (Visel A, 2007; Vockley CM, 2016; Dogan N, 2015) Ranking our elements by EP300 signal classes results in a similar fraction of activity as reported by Visel A, 2009 where an identical 80% of active enhancers are found in the top200 sites. (Figure 2.9)

In myogenesis the study of the relationship between EP300 occupancy and enhancer

function is complicated as other HATs are present and expressed at levels comparable or greater than EP300. (Further discussed in Chapter 5) (Albini S, 2010; He J, 2011) This means that while no EP300 signal is found within several active enhancers assayed in this study (Figure 2.9), these sites may simply pair to with different HAT other than EP300 (eg CBP/pcaf). Unfortunately we do not have a significant number of these active but non EP300 marked elements, which makes their analysis at the motif level impossible without greater numbers. Attempts to ChIP these other HATs unfortunately failed repeatedly.

2.7 Enhancer function analyzed ChIA-PET connectivity

Our randomly selected elements that have been tested can be separated by their ChIA-PET connectivity status. (Fisher K, 2017 Thesis, Lajoie RR, 2015) Figure 2.10 shows the lack statistical significance at the p=0.05 level for difference of mean activity between elements which are connected vs not connected. (t-test p=0.16) This is contrasted to the expectations of many that connected elements might be enriched for function. These sites may only work as an enhancer when brought into physical contact with their cross-connected locations or may be working as poised enhancer where the 3D chromatin structure is laid out prior to enhancer activation. (Vockley CM, 2016) (Further discussed in Chapter 5)

2.8 Comparing IDR vs Pre-IDR peak calls; 1PCR vs 2PCR libraries

The original selection of elements was done using data processed with the ERANGE peak caller from a 2PCR library. Of the randomly selected 131 elements 31 of the tested candidate elements were discarded by the modern processing by the ENCODE3 pipeline using SPP for peak-calling and processed for IDR=0.2 on biological replicate 1PCR myogenin ChIP-Seq libraries. (Mortazavi A, 2008; Li Q, 2011; Moore J, 2018; ENCODE, 2012) Of these 31 sites, only 2 marginal enhancers are detected in the MRF negative elements. (Figure 2.11)

This indicates that either the 2PCR amplifications of the libraries included several artifactual sites that looked like real signal when processed in a heuristic model based peak caller such as ERANGE. (Mortazavi A, 2008) Unfortunately we do not have two matching 2PCR ChIP-Seq for myogenin that we could run through IDR, to determine if IDR is responsible for the exclusion of this class of sites from that data. Knowing for certain whether these irreproducible sites are largely void of enhancer activity would be helpful in guiding future efforts.

There is however a significant contribution by the combination of 1PCR libraries and IDR to remove a large set of unstable sites across the two biological replicates, which results in enrichment in significant function overall in the population of ChIP-Seq sites (depending on where the IDR cutoff is set).

2.9 Conclusions

1) The transfection assay that I developed is sensitive for function within sites that occupied by myogenin. This is proven by a large set of negative control elements which do not function even though they provide comparable motifs. (Figure 2.6)

2) Elements in all signal classes (Figure 2.2), including top signals, did not function as enhancers. (Figure 2.6) These may be poised enhancers or simply require physical connectivity to other elements (or their native promoter in the correct orientation to function. Overall 40% of myogenin occupancy sites are expected to function, indicating that myogenin is a permissive condition for function but not a determinant. (Figure 2.6)

3) The vast majority of these enhancers are specific for function in the myocyte state. (Figure 5.1) Only a few elements are marginally active in the myoblast state, the mesodermal precursor cells (control for specific TF function) nor result of ectopic activation from the differentiation stimulus (serum change and insulin exposure) indicating that these Enhancers depend on TFs expressed or signaling only present

in myocytes for their function. (Further discussed in Chapter 5)

4) MyoD occupancy which is exclusive in the myoblast failed to capture active enhancers (Figure 2.8), although there is indication the MyoD may function at several important sites as an enhancer when recruited by Pbx1/meis. (Further discussed in Chapter 5)

5) While EP300 has been noted as a requirement for enhancer function, we find several elements that function in our assay without the or with weak presence of EP300, indicating that other HATs may perform redundant functions (Figure 2.9).

6) Physical connectivity, as measured by ChIA-PET, fails to capture exclusively active enhancers. Several of these connected elements are not functional in my assay suggesting a possible requirement for several of these sites to come together physically be biologically active. (Figure 2.10)

7) Modern data processing using IDR on 1PCR biologically replicated ChIP-Seq experiments appears to selectively remove candidate Enhancers that are largely nonfunctional. (Figure 2.11)

2.10 Figures (Chapter 2)

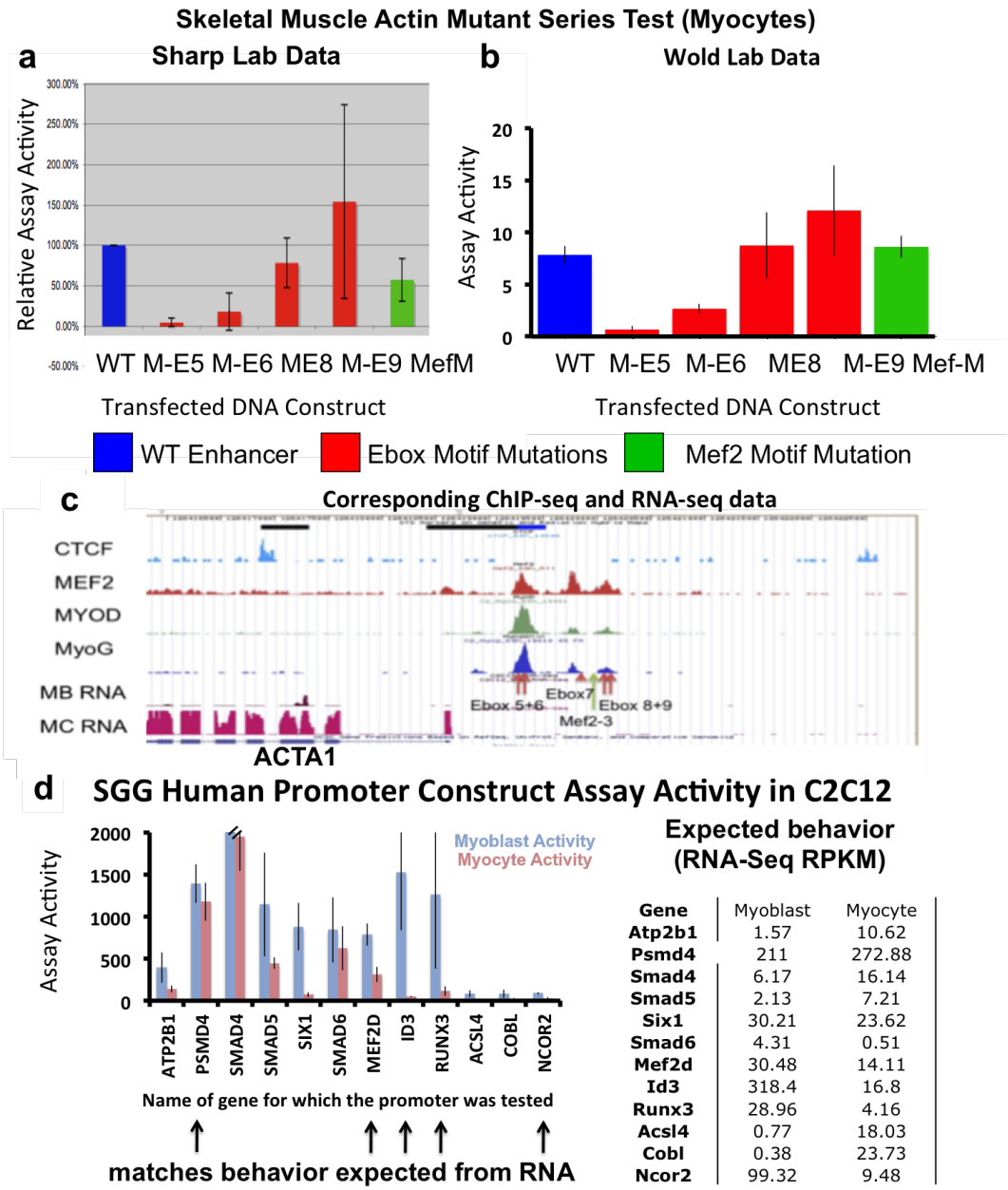


Figure 2.1: (Caption on next page.)

Figure 2.1: 4A Data from Sharp Lab at CSU-LA for a skeletal muscle actin construct (blue) and a series of mutant elements (red-Ebox mutant, green – Mef2 mutant). 4B data produced in one of our initial transfection tests using the same constructs as panel 4A. 4C 2PCR occupancy of MyoD, Myogenin (MyoG), Mef2A and CTCT in the region over and upstream of the Acta1 gene. Red arrows indicate the Eboxes that were mutated, while green arrows point to the mutated Mef2 motif. 4D Initial transfection test using Switchgear Genomics human promoter constructs and comparison with the RNA trajectory of the associated gene.

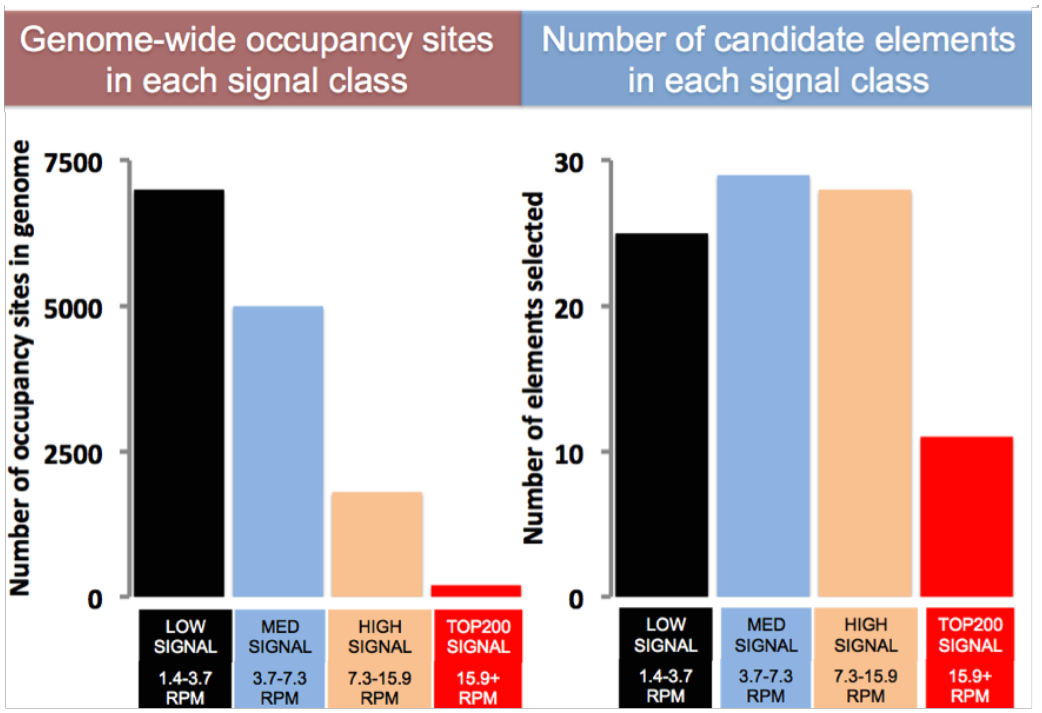


Figure 2.2: LEFT genome-wide distribution of myogenin occupancy sites in each signal class (SPP-IDR p=0.02 peak calls). RIGHT number of candidate Enhancers that were randomly selected in each signal class.

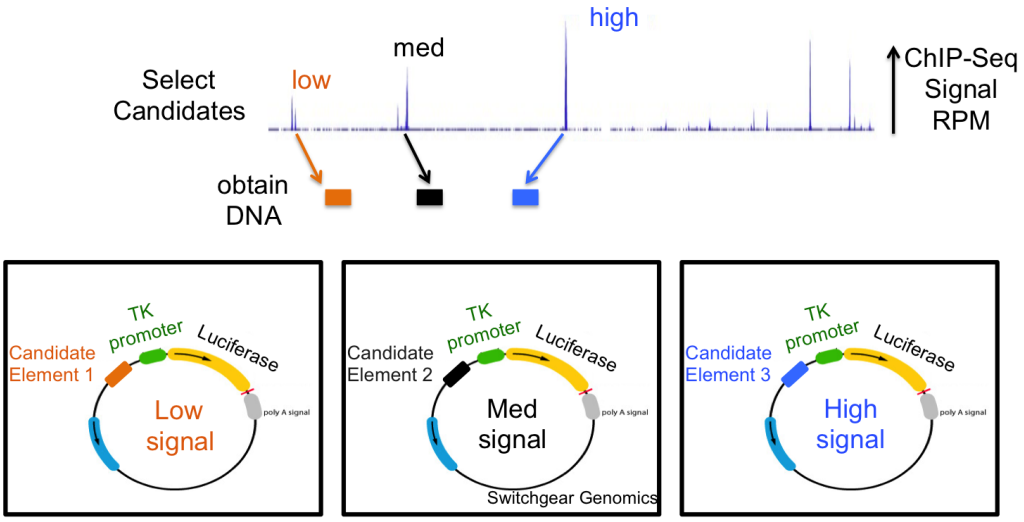


Figure 2.3: This figure shows the sampling and cloning of a representative low, medium and high signal element into the same basal TK promoter construct driving a high-turnover (h-PEST) Luciferase reporter gene. This process done for comparable numbers in each class. (Figure 2.2)

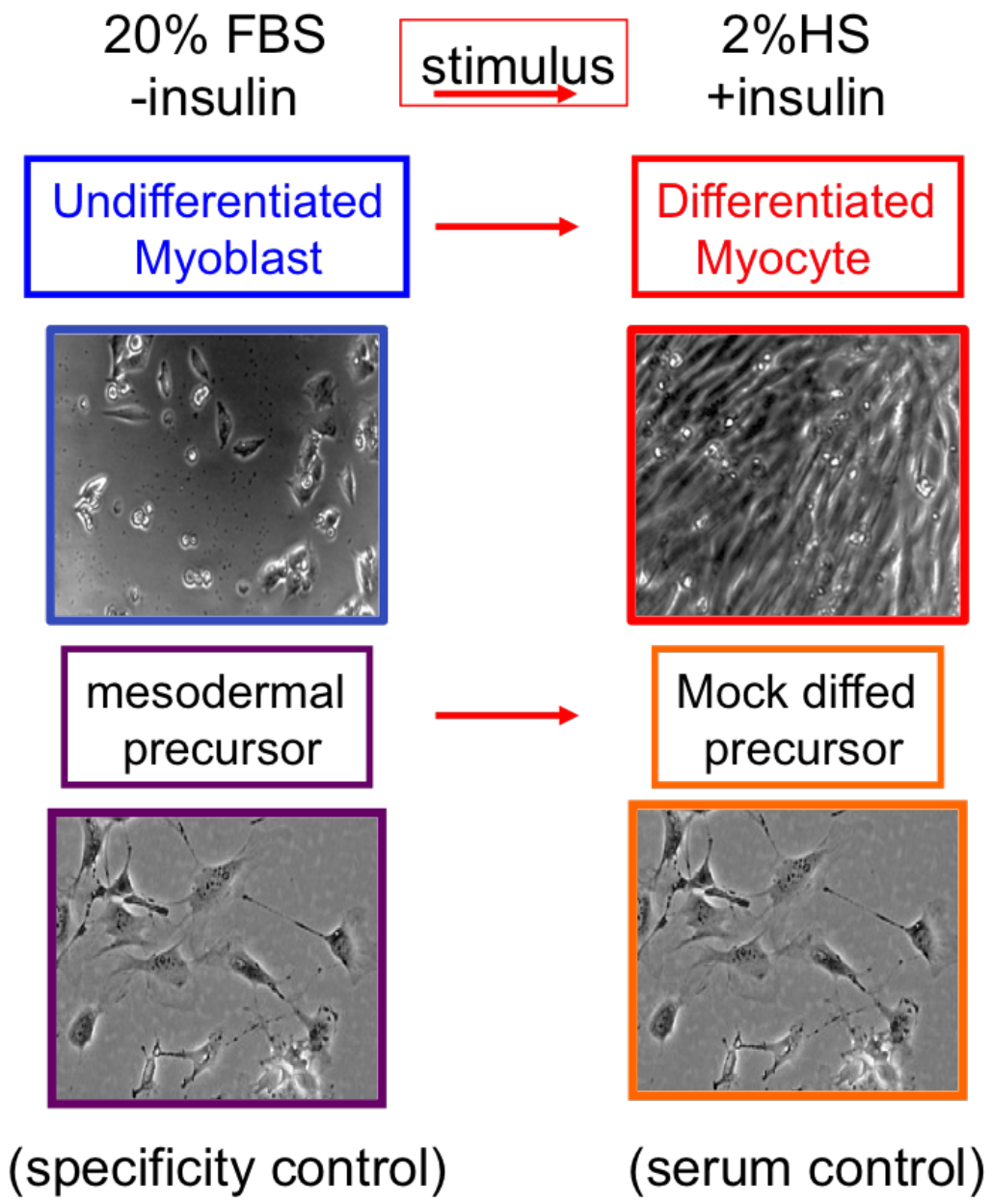


Figure 2.4: Top Left Panel: C2C12 undifferentiated myoblast expressing MyoD. Top Right Panel: C2C12 differentiated myocyte expressing both MyoD and myogenin. Bottom left Panel: 10T1/2 mesodermal precursor which acts as a specificity control since neither MyoD or myogenin are expressed. Bottom Right Panel: 10T1/2 mock differentiated mesodermal precursor cells which act as a control for serum and insulin used to differentiated the C2C12 cells. MyoD and myogenin are also not expressed in mock differentiated 10T1/2 cells.

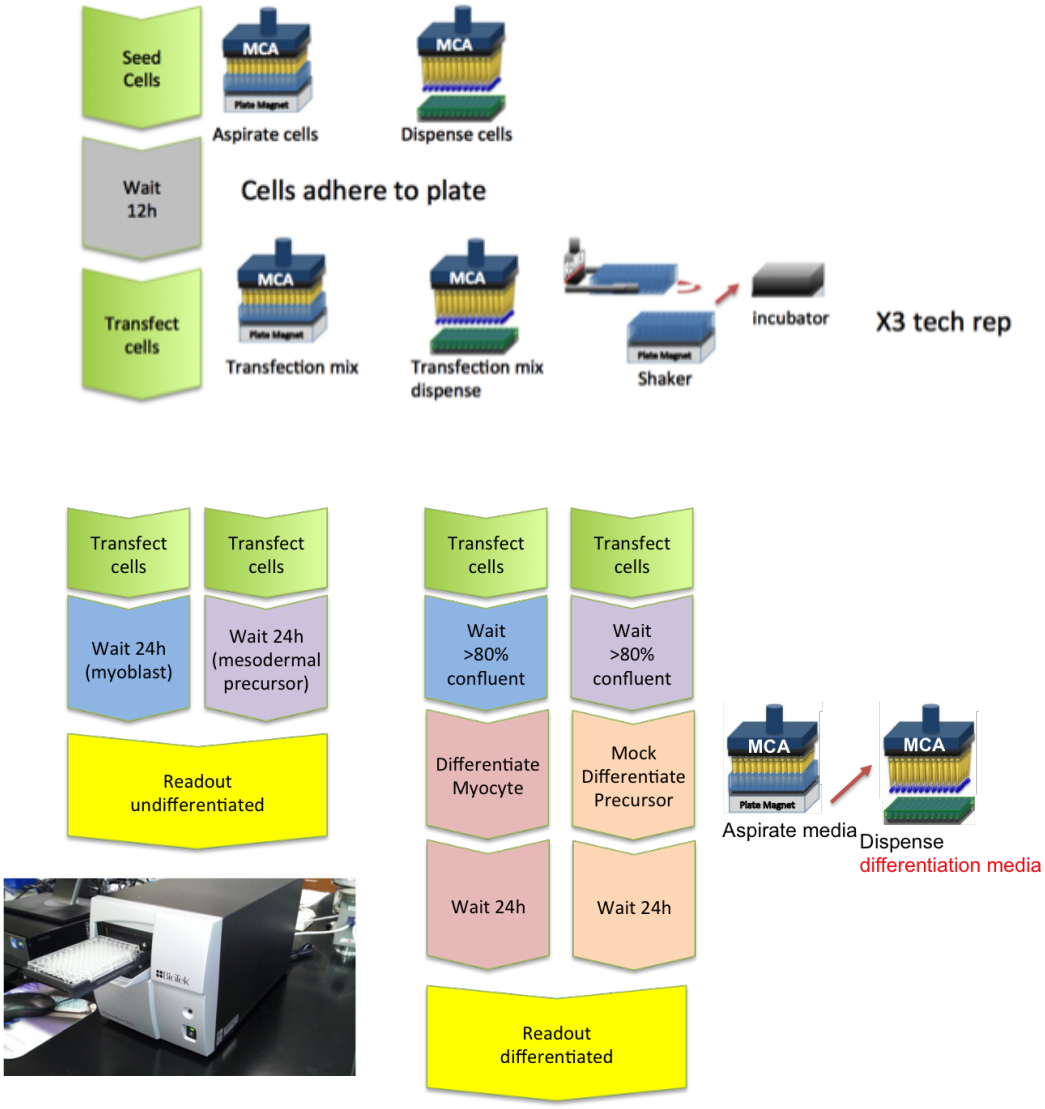


Figure 2.5: Flowchart for the transfection assay. Each cell line is plated independently in triplicate technical plates. For the undifferentiated myoblast and mesodermal precursor a simple 24 hour wait post transfection is followed by a readout on a plate luminometer. For the differentiated C2C12 myocyte and mock differentiated 10T1/2 mesodermal precursor there is a wait for the cell density to increase to ~80% confluence, the differentiation stimulus is then added, and the luminescent is read after 24 hours post transfection.

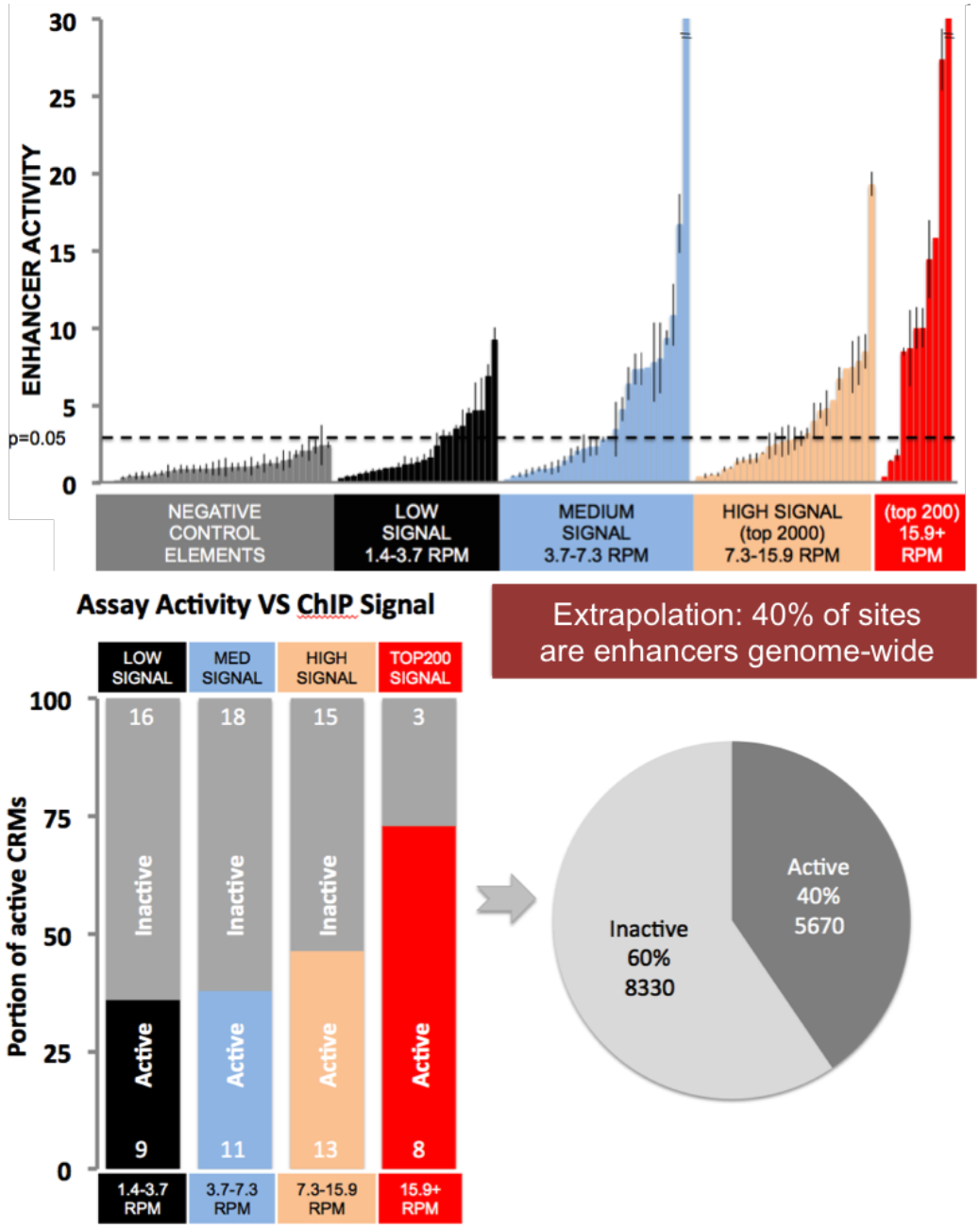


Figure 2.6: TOP Randomly selected elements sorted by their underlying myogenin ChIP-seq signal (SPP-IDR peak calls). Negative control elements are represented in gray and are used to set the activity threshold. BOTTOM LEFT Proportion of active enhancers in each signal class, active enhancers represented in color, inactive portion represented in gray. The number of elements tested that fall active and inactive in each bin is represented at the top and bottom of each bin. BOTTOM RIGHT Pie chart showing the overall expected number of enhancers using the transfection assay as a base for genome-wide extrapolation. Overall, 40% of ChIP-Seq sites for myogenin are expected to be active enhancers.

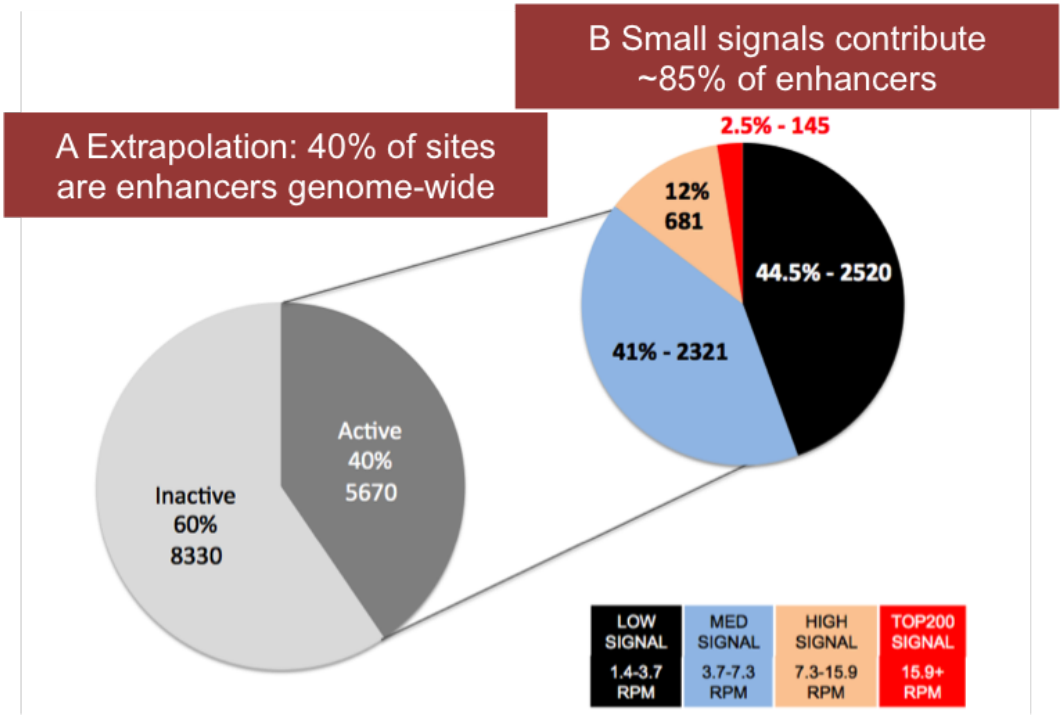


Figure 2.7: A Expected number of genome wide enhancers using the transfection assay as a base for prediction. 40% of ChIP-Seq sites for myogenin (SPP-IDR peak calls) are expected to be active enhancers totaling over 5500 sites. B Shows the contribution of predicted enhancers from each signal class, where low signal sites contribute almost 45% of active enhancers genome-wide.

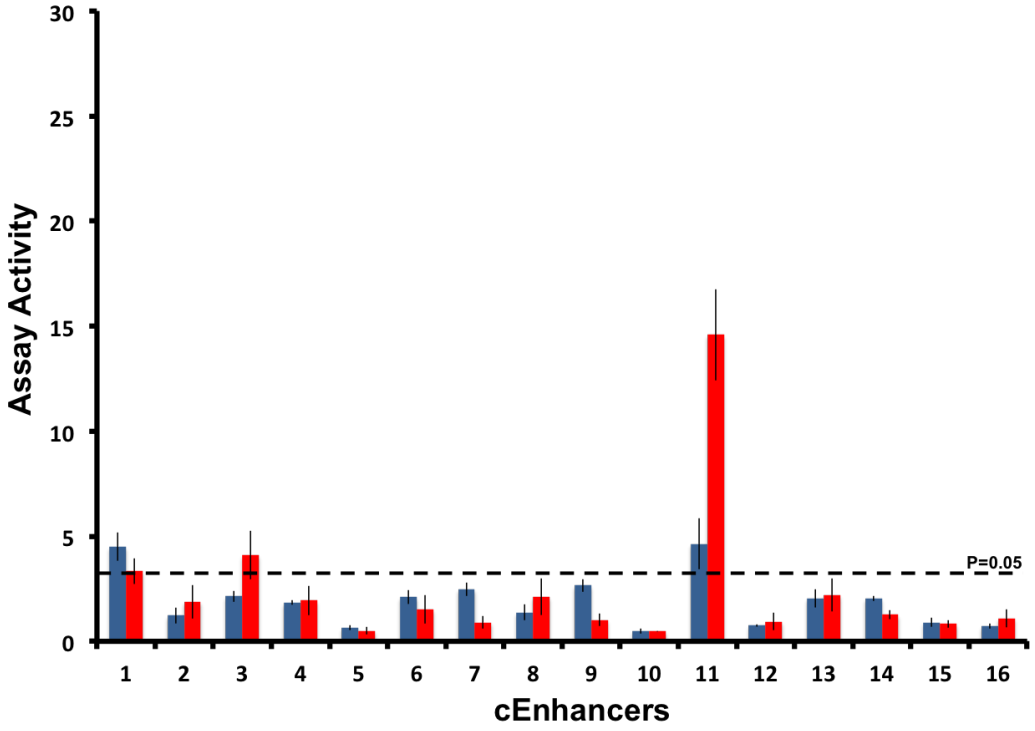


Figure 2.8: Transfection assay activity for elements that are exclusively occupied in the myoblast state by MyoD (no MRF MC signal). Myoblast (blue) and myocyte (red) transfection assay activity for elements ranked low to high by myoblast MyoD signal. These elements turned out generally void of EP300 occupancy, but element 11 is an exception which shows high EP300 signal present within the tested region in myocyte indicating that it likely is a non myogenic enhancer.

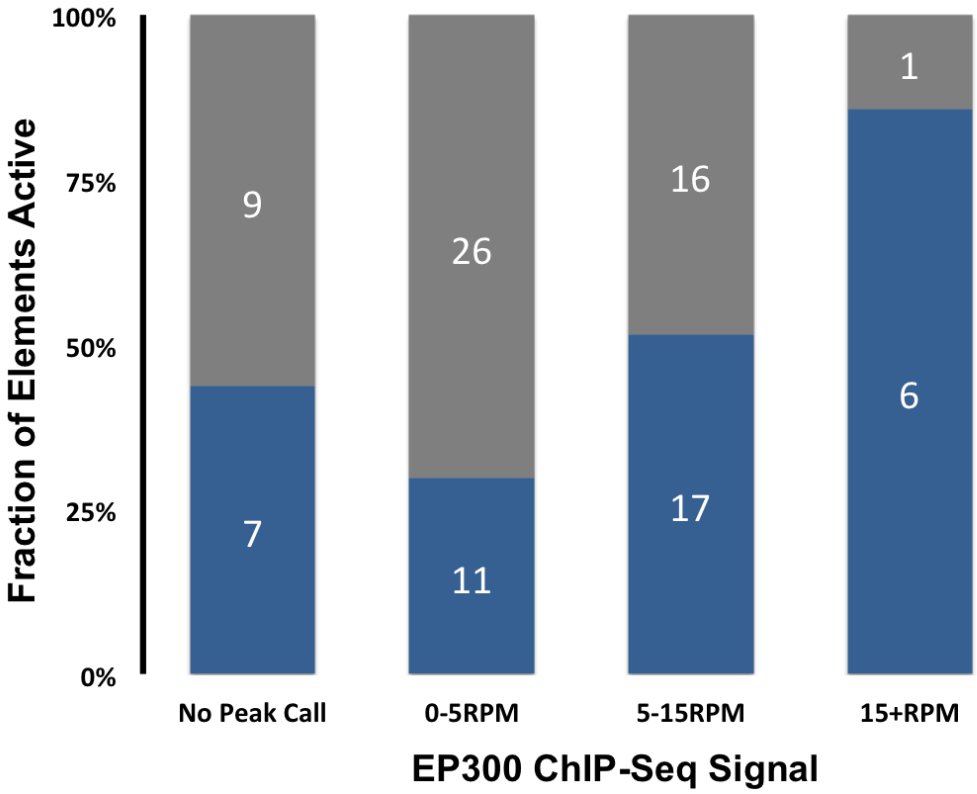


Figure 2.9: Proportion of active enhancer elements (blue) vs inactive (gray) among myogenic candidate Enhancers scored by signal class of EP300 occupancy. Number of cEnhancers active/inactive in each class are in white.

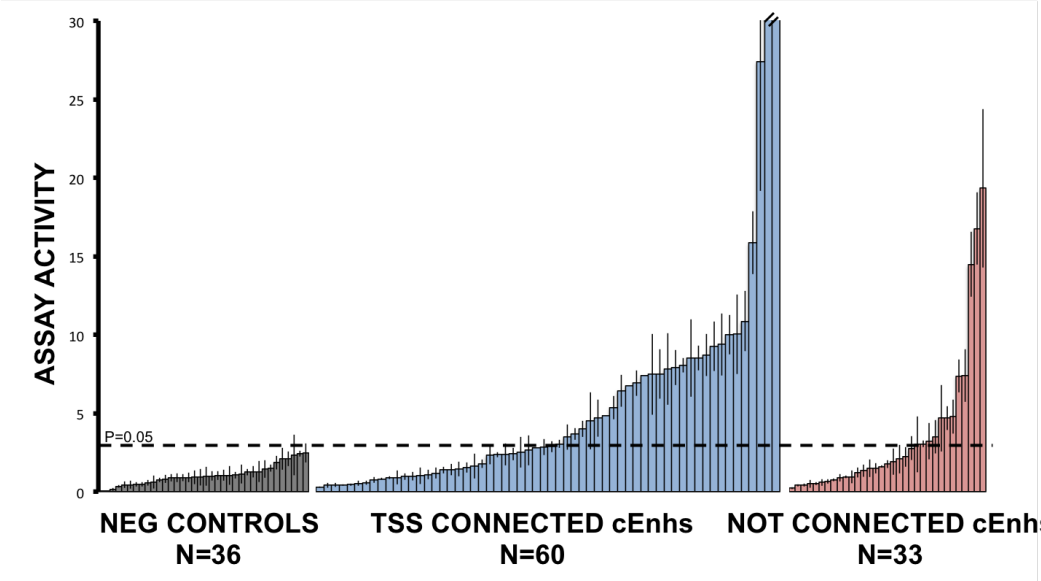


Figure 2.10: Myocyte transfection activity for randomly selected elements are not statistically different at the p=0.05 level for proportion, mean, and distribution of activity between populations that are ChiA-pet connected (Blue) vs not connected (red). T test p=0.16, Fisher's Exact p=0.28, KS test p=0.15. Negative control elements are shown in gray on the left.

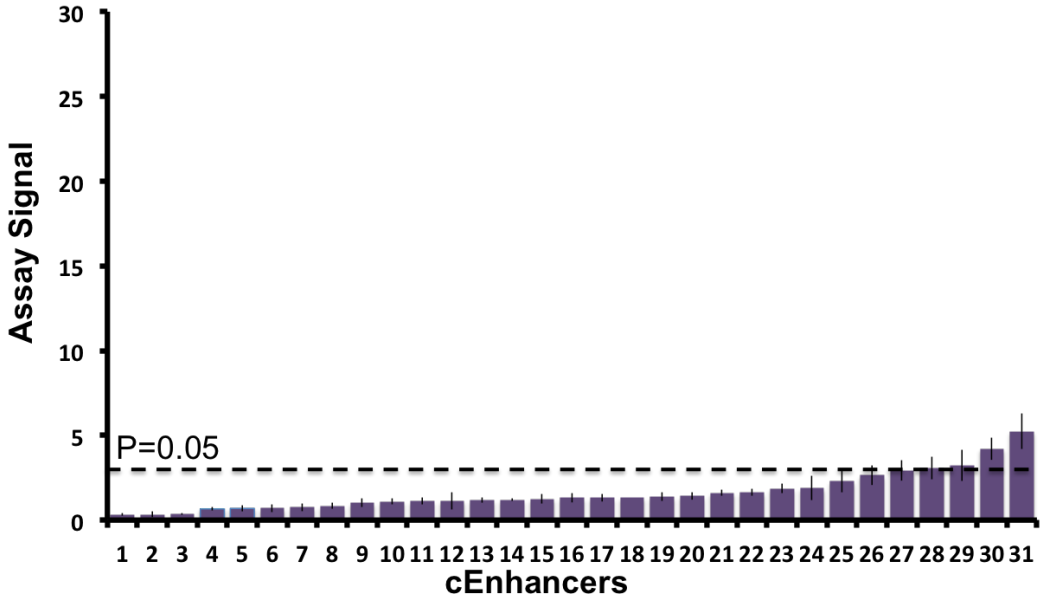


Figure 2.11: 2PCR ERANGE elements that failed to reproduce in 1PCR IDR called replicate experiments. Additional elements that met a previous myogenin occupancy criteria (based on a 2PCR ERANGE ChIP-seq run), but failed to meet the modern selection standards (IDR on true biological replicate data sets and 1PCR libraries) used elsewhere in this chapter. The removal of the data not conforming to IDR based on 1PCR biological replicate libraries appears to discard mainly inactive cEnhancers.

References

Albini S, Puri PL. (2010). "SWI SNF complexes chromatin remodeling and skeletal myogenesis: It s time to exchange!" In: *Exp Cell Res.* 316, pp. 3073–3080.

Arnold CD Gerlach D, et al. (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." In: *Molecular and Cellular Biology.* 339, pp. 1074–1077.

Casey BH Kollipara RK, et al. (2018). "Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors." In: *Genome Res.* 28, pp. 484–496.

Dogan N Wu W, et al. (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." In: *Epigenetics Chromatin.* 8, p. 16.

ENCODE, Project Consortium (2007). "Identification and analysis of functional elements in 1 percent of the human genome by the ENCODE pilot project." In: *Nature.* 447, pp. 799–816.

– (2012). "An integrated encyclopedia of DNA elements in the human genome." In: *Nature.* 489, pp. 57–74.

He J Ye J, et al. (2011). "Structure of p300 bound to MEF2 on DNA reveals a mechanism of enhanceosome assembly." In: *Nucleic Acids Res* 39, pp. 4464–4474.

Lajoie RR Dekker J, et al. (2015). "The Hitchhiker's Guide to Hi-C Analysis Practical guidelines." In: *Methods.* 72, pp. 65–75.

Li Q Brown J, et al. (2011). "Measuring reproducibility of high-throughput experiments." In: *Ann Appl Stat.* 5, pp. 1752–1779.

McKnight SL Kingsbury R, et al. (1982). "Transcriptional control signals of a eukaryotic protein-coding gene." In: *Science.* 217, pp. 316–324.

Moore J Purcaro MJ, et al. (2018). "ENCODE Phase III Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse." In: *Submitted.* Pp. 1–2.

Mortazavi A Williams BA, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nat Methods.* 5, pp. 621–628.

Muerdter F Boryń M, et al. (2018). "Resolving systematic errors in widely used enhancer activity assays in human cells." In: *Nat Methods.* 15, pp. 141–149.

Visel A Blow MJ, et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." In: *Nature.* 457, pp. 854–858.

Visel A Minovitsky S, et al. (2007). "VISTA Enhancer Browser-a database of tissue-specific human enhancers." In: *Nucleic Acids Res* 35, pp. D88–92.

Vockley CM DIppolito AM, et al. (2016). “Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome.” In: *Cell.* 166, pp. 1269–1281.

Chapter 3

DRAFT MANUSCRIPT: PURSUING FUNCTION IN CANDIDATE ENHANCER SIGNATURES

The introductory section to the draft manuscript was merged within Chapter 1 (introduction). I append an abstract of the results presented in this chapter.

Genome-wide DNAse hypersensitivity, histone modifications and transcription factor occupancy measurements from the ENCODE and NIH Roadmap consortiae have been used to identify candidate regulatory elements (cREs) and profile them as candidate enhancers, promoters and insulators based on their characteristic biochemical signatures. However, accurately predicting the regulatory activity of cREs from the biochemical maps alone has turned out to be difficult, and it is not yet clear to what extent regulatory elements active *in vivo* can be identified from these biochemical signatures. To assay the predictivity of the biochemical signatures characteristic of enhancers we carried out tests for transcriptional enhancer activity for hundreds of candidate elements from five human or mouse cell types, including immortalized cells and cell lines modeling the early developmental differentiation to muscle or red blood cells. Our candidate Enhancer (cEnh) collections were selected using biochemical signature criteria ranging from from individual transcription factor (TF) occupancy to tissue differential integrative machine learning models.

We find that in general ~50% of both TF-selected and TF-agnostic cEnhs demonstrated significant enhancer activity in transfection assays, observing similar proportions across all cell lines and conditions examined. We observe that the presence or number of TF recognition motifs in cEnhs displays no correlation with enhancer activity. Most of the active enhancers detected displayed relatively modest biochemical signatures in their genomic context, some of which scored among the

most powerful on our assay. However, the relative activities of the cEnhs measured reveal that the majority likely only exhibit modest regulatory activity on their own. Conversely genomic regions that lacked the biochemical marks, even while containing the motifs necessary for the relevant TFs, generally did not sustain enhancer function.

We demonstrated that as expected from the redundancy of biochemical measurements that DHS and H3K27ac are equally predictive of enhancer activity compared to TF occupancy alone. We also observe a positive but weak correlation between biochemical signal strength and predictivity of enhancer activity, with even the most biochemically outstanding group displaying a significant fraction of cEnhs displays no discernible enhancer activity.

We corroborated these findings by using the STARR-seq MPRA to assay the activity of thousands of genomic regions occupied by the glucocorticoid receptor (GR) in stimulated A549 cells.

Finally, we discuss these results in the context of current models of the regulatory effect of enhancers on their cognate genes. We expect our findings to help guide future efforts towards cataloging the functional repertoire of mammalian genomes.

3.1 Results: Large-scale activity test of full-length cEnhancers

While powerful approaches now enable measuring the regulatory activity of thousands of elements simultaneously, the diminutive size of the elements assayed is a common concession. (Patwardhan RP, 2009; Kinney JB, 2010; Kwasnieski JC, 2012; Melnikov A, 2012; Patwardhan RP, 2012; Arnold CD, 2013; Murtha M, 2014; Kheradpour P, 2013; Kwasnieski JC, 2014; Ernst J, 2016). If the fragments tested (typically 80-250bp) are significantly shorter than the size of many full-length

candidate functional regulatory elements found in mammalian genomes they may lead to significant numbers of false negatives being scored.

In order to assess the ability of MPRAs to assay the activity for complete, full-length cREs one requires a large collection of likely candidate regulatory elements to estimate their representative size, preferably across many tissues.

Unfortunately the boundaries of distal elements defined by the current biochemical measurements are often unclear, especially when regions are merged from several partially redundant measurements. This creates boundaries that will extend much further than any likely regulatory element. However the distribution of the lengths of conserved noncoding segments (e.g. excluding the sequences overlapping with or in the vicinity of annotated exons) in the human genome (Supplementary Figure 3.1) provides a representative library of candidate functional elements. While this collection excludes the representation of tissues known to depend heavily on recently evolved enhancers (such has heart), it provides a core library of elements with well defined boundaries that stood the test of time.

From this library, we found that tens to hundreds of thousands (depending on the definition) of such blocks fall outside the range of testing supported by MPRAs. The additional concern with MPRAs is the possibility of cross-talk between different REs, the effects of which may vary and self-compound depending on the nature of the input library.

To address these issues, and accurately asses the functional predictivity of biochemical marks, we carried out our tests of cEnhs identified on the basis of biochemical signatures in a variety of mammalian systems using single plasmid transient luciferase reported enhancer activity assays, which allow for much larger segments of DNA to be tested for enhancer activity (Supplemental Figure 3.2).

While the full catalog of REs in the genome includes promoters, insulators, en-

hancers, silencers, and elements with other function; for the purposes of this study we focused primarily on distal candidate transcriptional enhancers (cEnhs).

A major reason for this choice is that distal cEnhancers can readily be identified using TF occupancy, by the overlap of DNAse hypersensitive sites (DHS) and regions marked by H3K27ac, or by all three at genomic locations that do not exhibit the biochemical signatures characteristic of promoter function or a GENCODE TSS annotation.

These historically underrepresented distal cEnhs also constitute the bulk of cREs distinguishing different cell types from each other, in contrast to TSS proximal enhancers the majority of which are biochemically occupied in multiple tissues that make predictions for cell type specific active enhancers difficult because of the often overlapping promoter function. (Figure 3.1D and supplemental figure 3.4).

Because these distal candidate enhancers are largely tissue specific we focused on testing their ability to individually support active biological function in the tissue where they are annotated. We also selected regions that contain DNA motifs relevant to each individual cell context

Immortalized cell lines were targeted for our enhancer assay tests as they have been extensively studied by the ENCODE consortium(ENCODE, 2011; ENCODE, 2012) and they are the source of a significant portion of ENCODE data. We aimed at comparing the predictivity of multiple approaches for identifying functional cEnhs, and at incorporating a diversity of biological systems in our analysis (Figure 1E). We targeted model systems for major developmental transitions (myogenesis and erythropoiesis), immortalized cell lines (the human K562 and HepG2 cell lines), and a model system for cellular response to exogenous signaling stimuli (activation of the glucocorticoid receptor in the human alveolar epithelium cell line A549). We also aimed to represent the full spectrum of cEnh biochemical signatures (Figure 3.1D

and Supplementary Figure 3.3), as multiple studies have shown that the landscape of transcription factor occupancy, DNAse hypersensitivity and histone modification maps includes many more weaker sites than very strong peaks (Landt SG, 2012; Kellis M, 2014).

3.2 Results: Tissue specific enhancers from muscle and red blood cell differentiation

In the context of muscle differentiation, we selected TSS distal cEnh regions based on myogenin ChIP-seq data in differentiated C2C12 myocytes. We randomly selected a set of regions (n = 89) spanning the full range of myogenin occupancy levels, most of which contain a CAGSTG (n = 88) and the extended RRCASGTG (n = 84) E-box. We also used myogenin to select additional cEnh elements (n = 88) associated with genes playing a well characterized role in muscle development; to test whether they exhibit higher levels of functionality than the genome-wide average.

We selected a group of negative controls to test the specificity of biochemical signatures for prediction of functional activity in a given cell type. To this end, we selected elements (n = 23) from a set of well characterized enhancer elements that are active in T cells and in neurons but that are not occupied by myogenin in C2C12 cells; 21 of these elements contain the CAGSTG motifs. We specifically selected the majority of the regions to contain and be centered on a muscle class E-box as we wanted to allow for the potential binding of myogenin in an exogenous plasmid system. A second set of negative control elements were selected in order to assess the baseline functional activity of biochemically and functionally neutral regions. We selected candidate regions (n = 11), 6 of which contain E-box motifs from areas void of enhancer signatures. No significant difference in activity was found between the two sets which were merged for the purposes of this paper.

Erythroid cEnhancers were selected to provide a cross tissue comparison to our

mouse elements from GATA1 or Tal1 ChIP occupied regions (n = 113) of the mouse genome and tested in the easily transfectable human K562 cells. GATA1 ChIP-chip data in the cell line G1E-ER4 was used to select GATA1 cREs. Elements were subjected to independent validation by ChIP-qPCR, with validated comprising candidate Enhancers (n = 53). TAL1 ChIP-seq data was used to select ChIP-seq positive regions (n = 60) that were tested for enhancer activity after transient transfection in K562 cells. Genomic elements that were not significantly occupied by GATA1 were selected as occupancy negative controls for the assay (n = 74)Taylor J, 2006; Cheng Y, 2008. Of these, GATA1 ChIP-chip that failed ChIP-qpcr validation comprised 63 of the negative control elements. Another set of 11 DNA segments were not called as peaks in the ChIP-chip analysis, these are labeled GHN for GATA1 hit negative. Although they are not biochemically occupied, 45 of the control elements contain a GATA1 motif (WGATAA); while 40 contain a Tal1 occupancy motif (CAGMTG) providing a baseline functional activity.

Figure 3.2 shows the measured enhancer activities for TF-centric cEnh selections and negative controls in myogenesis (Figure 3.2A) and erythropoiesis (Figure 3.2B). We provide a summary of the enhancer assay results from each of these two systems where we selected candidates based on TF occupancy in Figure 3.2; together with the biochemical signal measured and statistically significant peak calls (IDR$\geq$ 0.05) for the key TF; DNAse-seq and H3K27ac ChIP-Seq. We also provide the number of DNA motifs present within each cEnhancer tested. We found 39 out of 89 or 44% of muscle cEnhs to pass the threshold of activity in contrast to 3 out of 34, or 9% of muscle negative controls. Similarly, 50 out of 112, or 45% erythropoetic cEnhs and 2 out of 73, or 3% of negative controls, were found to be active. We note that even though in both cases the occupancy is highly selective for a tiny fraction of the available motifs genome-wide, only $\sim$ 50% of candidate Enhancers are active in each system and that both the strength or presence of the biochemical signals or the

number of motifs appears to be non-deterministic of enhancer function. Many of the active elements are contributed by fairly modest signals, with related biochemical measurements such as DNAse and TF occupancy often differing widely in signal strength compared within each individual element.

3.3 Results: The compendium of elements within the closest CTCF boundaries near genes mirrors that behavior of the genome wide population.

We also used myogenin to select additional cEnh elements (n=88) associated with genes playing a well characterized role in muscle development; to test whether they exhibit higher levels of functionality than the genome-wide average. We found 49 out of 88 or 56% of muscle cEnhs to pass the threshold of activity in contrast to 2 out of 34, or 5% of muscle negative controls. Although the locus selected elements appear slightly more predictive of enhancer activity, we found no measurable difference in muscle development between the distribution of active elements (KS p=0.167) selected for TF occupancy between elements randomly selected (figure 3.2) and the specific locus affiliated cEnhancers (supplementary figure 3.7). As such these elements were merged for the further analysis done in this manuscript.

3.4 Results: The occupancy of multiple TFs are equally predictive of enhancer activity as their individual tissue specific counterparts.

While C2C12s and G1E-ER4s provided a low hanging fruit for selecting tissue specific candidate enhancers; K562 and HepG2 comprised the primary immortalized cell lines selected by the ENCODE consortium for extensive annotationENCODE, 2011; ENCODE, 2012.

We aimed at comparing the predictivy of cEnhancers selected from a cell-type specific factors (myogenin and GATA1/Tal1) to that of a collection of overlapping occupancy peak calls by non cell-type specific positive acting factors (such as JUND, MAX, TEAD4 and USF1) mapped by ENCODE in K562s. One might expect that

by requiring the occupancy of multiple factors, one would improve the predictivity of active enhancer function.

To this end we used the vast trove of biochemical annotations in K562 cells to generate cEnhancers ($n = 28$) occupied by several generic activating TFs, to contrast the predictivity of cell type specific TFs used to select cEnhs in our differentiation models. A set of regions similarly TF occupied exclusively in HepG2 ($n = 32$) were used as negative controls. In K562 cells, we found 14 out of 28, or 50%,TF-centric cEnhs to be active, in contrast to 2 out of 32, or 6%) in the corresponding sets of negative controls. The individual results for K562 cEnhancers, including the number of TFs occupying each element, are available in Supplementary Figure 3.14.

We find no significant difference in the predictivity of cEnhancers selected by using cell-lineage specific contrasted with selection of cEnhancers based on generic trans-activator TFs. The results of these enhancer assays based on TF-centric cEnhancers and occupancy negative control elements are summarized in Figure 3.3A.

3.5 Results: Biochemical marks and TF occupancy are similarly predictive of active enhancers in multiple cell types

Next, we aimed at comparing the predictivity of multiple approaches for identifying functional cEnhs, and at incorporating in our analysis a diversity of biological systems (Figure 3.1E). As such, we applied several different strategies for compiling lists of cEnhs to be tested, broadly contrasting the TF-centric with TF-agnostic picks based primarily on chromatin state signatures and evolutionary conservation in order to compare their predictivity for active enhancers.

We relied on multiple computational approaches for integrating high-dimensional collections of functional genomic datasets into a small set of chromatin states that have been devised over the last few years and applied to the problem in ENCODE

cell lines, including the Hidden Markov Model-based SegwayHoffman MM, 2012 and chromHMM(Ernst J, 2012), as well as Self-Organizing Maps(Mortazavi A, 2013) (SOMs). We selected cEnhs in K562 cells based on Segway and chromHMM chromatin state assignments and the presence of DNAse and H3K27ac (n = 30), with elements lacking both marks used as negative controls (n = 21). We also selected cEnhs based on SOMs trained DNAse and histone mark ChIP-seq data over multiple ENCODE cell types; these cEnhs were picked so that they were specifically in an open chromatin state and marked by histone modifications associated with enhancer activity in HepG2 cells (n = 32). The negative control elements were also derived from other SOM regions that lacked both marks in HepG2 (n = 18). The individual enhancer assay results are summarized in the supplemental figure 3.15 with the correlations to the biochemical data (supplemental figure 3.16) being comparable to those found in C2C12 and G1E-ER4.

In the context of hematopoiesis, a TF-agnostic cEnhancer selection was based on evolutionary conservation. A total of 46 cREs were selected from regions conserved in alignments of multiple mammalian species characteristic of regulatory regions but containing a GATA1 motif. Six additional DNA segments highly conserved to outside of mammals that also contain a GATA1 binding motif were also included. These elements were annotated for DHS and H3K27ac in G1E-ER4 cells; with 28 elements scored positive for both biochemical marks and were used as "cEnhancers" while the 24 elements lacked both of the modern marks were used to score as a negative control set.(Wang H, 2006; Taylor J, 2006) The individual results for cEnhancers are presented in supplemental figure 3.11.

The same analysis for TF-agnostic cEnh selections showed that 18 out of 32 (56%) of SOM picks were active in HepG2 cells, 14 out of 42 (33%) of chromHMM/Segway picks were active in K562 cells (Supplemental Figure 15), and 13 out of 28 (46%) of GATA1 conservation selections exhibited significant activity contrasted to 2 out

of 24 negative controls (8%) supplemental figure 3.11.

However, as expected from previous studies that sampled candidate elements solely on conservation Visel A, 2007, the GATA1 motif containing and evolutionarily conserved cEnhs scored blind to the biochemistry yield a reduced fraction of functionally active elements.

3.6 Results:DNAse and H3K27ac predictivity.

In order to contrast and extrapolate to genome-wide "gold standard" cEhancers we applied the current criteria proposed by ENCODE. We required the elements for this comparison to be DNAse hypersensitive and ChIP-seq H3K27ac occupied. This enabled us to compare these regions to their TF agnostic counterparts that were selected almost exclusively from regions that contain these two biochemical annotations. Because no measurable differences were found between the two sets (random and locus) in muscle; both were merged and used. The results of these enhancer assays and occupancy negative control elements are summarized in Figure 3.3B. We extrapolate genome-wide predictions for each system based on the proportion of active enhancers. Overall, we find no difference in the predictivity of enhancer activity for either TF-centric or TF-agnostic selections across multiple tissues, and estimate that approximately half of the population of biochemically DNAse/H3K27ac co-marked cEnhancers are biologically active.

However, as expected from previous studies that sampled candidate elements solely on conservation Visel A, 2007, the GATA1 motif containing and evolutionarily conserved cEnhs scored blind to the biochemistry yield a reduced fraction of functionally active elements.

We find no significant difference in the predictivity of enhancer function for candidates selected by using cell-lineage specific TFs contrasted with selection of cEnhancers based on generic trans-activator TFs. The results of these enhancer

assays based on TF-centric cEnhancers and occupancy negative control elements are summarized in Figure 3.3A. We extrapolate genome-wide predictions for each system based on the proportion of active enhancers. We note that while G1E-ER4 appears to have significantly less enhancers, this is simply attributable to a weaker DNAse hypersensitivity measurement.

We then examined the subsets of cEnhs in each system bearing the simultaneous biochemical signature of TF ChIP-seq occupancy, H3K27ac demarcation and DNAse hypersensitivity, which would represent the most likely to be functional cEnhs. We observed 55.6% active such cEnhs in myogenic cEnhs, 50% among erythroid ones, and 49.5% in the set of cEnhs selected from human immortalized cell lines (Figure 3.3A). The same analysis for TF-agnostic cEnh selections showed that 18 out of 32 (56.2%) of SOM picks were active in HepG2 cells, 14 out of 30 (46.6%) of chromHMM/Segway picks were active in K562 cells (Supplemental Figure 15), and 12 out of 26 (46.2%) of GATA1 conservation selections exhibited significant activity contrasted to 2 out of 24 negative controls (8%) (Supplemental Figure 3.11). We note that HepG2 SOM selections were biased towards more strongly H3K27ac/DNAse positive regions (Supplementary Figure 3.3), which might explain the higher levels of observed activity within that set.

Overall we find similar levels of activity in both the TF-centric and TF-agnostic sets of cEnhs predictions, around 50%. Using the total number of cEnhs in each cell type and the observed proportions of cEnhs active in functional assays, we estimate that there are 8421 active myogenin+/H3K27ac+/DNAse+ enhancers in C2C12 cells, 3789 GATA1 + /H3K27ac + /DNAse + ones in erythroid cells, 13255 TF + /H3K27ac + /DNAse + ones in K562 cells. Self Organizing Maps combined with functional testing predict 13464 active enhancers in K562 cells while chromHMM/Segway predict 12460 ones. The combination of GATA1 conservation and the H3K27ac/DNAse biochemical signature predicts 5253 active enhancers in

erythroid cells. We note that the low prediction in G1E-ER4 cells is likely due to weaker measurement which thresholds the available number of sites to be lower in the genome.

We also note that in all functional tests we carried out using luciferase assays, we find a skewed distribution of activity, similar what is observed for biochemical signal strength in ChIP-seq and other functional genomic experiments. A small number of cEnhs appear to be highly active, while the majority of even active cEnhs exhibits only modest activity (Figure 3.2, Supplementary Figure 3.3 and Supplementary Figure 3.15).

3.7 Results: The bulk of enhancers in a given cell type are marked by modest biochemical signatures

Although we presented evidence that the overall predictivity for biochemical signatures are equivalent across multiple systems, the biochemical signals measured by a ChIP-seq experiment vary widely within statistically significant regions (Figure 3.1D).

As mentioned, we found no significant difference in the predictivity or relative activity between the subpopulation of cEnhancers culled from a random selection of TF occupied sites compared to ones affiliated with transcriptionally outstanding genes. The genomic region surrounding the BTG2-MYBPH-myogenin genes (figure 3.4A) shows that, like functional activity, the general range of myogenin ChIP-Seq signals is well represented even near outstandingly transcribed genes. However, overall smaller biochemical signals make up the vast majority of the catalog of myogenin occupancy based cEnhancers in the genome (Figure 3.4B). One key unresolved question is whether biological activity is significantly present within this lower spectrum of the biochemical measurement.

In order to asses the relationship between occupancy strength and functional activity,

we split the cEnhs we tested in C2C12s into four bins (“low”, “medium”, “high”, and “top”) according to the level of myogenin ChIP-seq signal observed (Figure 3.4C). We find that the fraction of active cEnhs increases steadily with the strength of myogenin signal, with only ~20% of low-myogenin cEnhs exhibiting significant activity in contrast to ~65% of the most strongly occupied ones. These observations superficially imply a close relationship between biochemical signal strength and functional activity, however, we note that even in the latter group a large fraction (nearly half) of cEnhs is inactive when directly functionally tested.

We then asked how many active enhancers genome-wide are likely to be found among each portion of the signal strength distribution. Figure 3.4D shows the extrapolated numbers of active enhancers in each bin of myogenin occupancy strength. However, this function is specific for the biochemically marked regions as demonstrated by the negative control elements. While the strongest myogenin sites are mostly likely to be functionally active, the much greater numbers of weaker sites mean that the 75% of active enhancers in muscle cells are expected to be found among the sites belonging to the “low” and “medium” bins.

Similar observations apply to the combinations of biochemical marks used by the ENCODE encyclopedia within myogenin occupied sites. (Supplementary Figure 3.17). It should be noted that while the proportion of active enhancers remains the same, when using multiple overlapping measurements the predicted numbers of enhancers is affected; especially in the range of medium to low signal. While these elements that “drop out” are within a lower biochemical marks range, their relative activity on the assay is sometimes significant. This indicates that while utilizing multiple overlapping marks one will sometimes miss weaker marked sites that can nonetheless be functionally important. This results in thousands less sites being predicted; which make up a significant portion of the genome. It is important to note that we only tested elements as individuals and many more of these weakly

marked sites may worked when paired with other elements included in the region proximal to their natural promoter target.

Finally, we sought to contrast the contribution of enhancers from regions biochemically marked and occupied by a TF necessary for differentiation with that of regions occupied as the result of an external signaling stimulation. We chose these two main types of dynamic transition of cellular states associated with regulatory alterations of chromatin states to represent both the slower and typically irreversible differentiation of one cell type into another, and the much faster and reversible cellular response to signaling molecules.

The transcriptional response to glucocorticoids, on the other hand, is characterized by a more rapid kinetics of gene expression activation, and by its general reversibility. The A549s are a human lung epithelial cell line which has been used to study the genomic response to the stimulation of the glucocorticoid receptor (GR), normally involved in the suppression of inflammatory responses. We used the response of A549 cells to activation of the GR transcription factor by the cortisol analog dexamethasone (Dex) as a model system for our study. Upon activation by Dex GR rapidly associates with thousands of sites along the genome, both directly through its cognate motif and indirectly through association with cofactors such as AP-1 (Reddy TE, 2009; So AY, 2007; Gertz J, 2013), leading to changes in the expression of hundreds of genes. The epigenomic landscape of A549 cells during GR activation is illustrated for reference in Supplementary Figure 3.18. Using an IP for GR against the DNA of A549 cells induced by exposure to dexamethasone; we created a library to be used in a self-reporting enhancer assay (STARR-seq) (supplemental figure 3.19) (Vockley CM, 2016).

As such, we corroborated these finding by examining our A549 GR ChIP-STARR-seq data. Comparing STARR-seq reads to their input DNA libraries, and only

including cEnhs with sufficiently deep representation in sequencing libraries (see the Methods section for details), we identified ~5% of GR cEnhs to be significantly active in Dex-stimulated A549 cells and ~10% to be active in untreated cells (Supplementary Figure 3.12A). This fraction is considerably lower than what is observed for cEnhs with luciferase assays, an observation that is explained by a combination of the generally lower sensitivity of MPRAs and the shorter fragments being represented in STARR-seq libraries, which likely do not capture complete regulatory elements (although we note that we also carried out an activity analysis at the level of individual DNA fragments and did not observe longer DNA fragments to be preferentially active compared to shorter ones; Supplementary Figure 3.12C). We also note that, similar to luciferase functional assays, the majority of active enhancers in ChIP-STARR-seq datasets exhibit moderate levels of activity, with only a small minority of very highly active functional enhancers (Supplementary Figure 3.14). The distribution of GR ChIP-seq signal strength in cEnhs tested by ChIP-STARR-seq is not as skewed in favor of low-occupancy sites as it is in other contexts (Figure 3.4E), which is due to the fact that representation in the ChIP-STARR-seq input libraries is biased towards stronger sites and that we excluded cEnhs with insufficient number of reads to evaluate their activity. Nevertheless, we do see more weaker sites than stronger ones, and we also observe much higher levels of activity (~40%) in the "top" bin (Figure 3.4F) than within the "low" and "medium" ones (~10%). Thus we can conclude that even though the most visible biochemical signatures are most likely to correspond to active regulatory elements, in most biological contexts most functional enhancers in fact reside among the population of cEnhs characterized by only modest biochemical signatures, underscoring the complexity of the task of identifying active cEnhs from biochemical measurements alone.

Unlike our previous selections, which were heavily biased for TSS distal regions, this provided an assay that tested both TSS proximal and distal cEnhancers in reporter

vector that positioned the regions 3' relative to the contained promoter element.

We selected the genomic regions which were significantly represented in the library and mapped the corresponding GR ChIP-seq signal (figure 3.4E). While the assay biased for the overall presence of relatively strong biochemical signals the overall proportion of enhancers found active in each signal class remains remarkably similar figure 3.4F.

3.8 Results: Biochemical signature signal is not correlated with measured enhancer strength

A most prominent misconception is that the top biochemical signals measured will invariably contain the strongest enhancers in the genome.

The signature combination of H3K27ac and DNAse hypersensitivity outside of promoter regions has been used by ENCODE to characterize cREs with the potential for enhancer function.

The overlap of H3K27ac and DNAse hypersensitivity signal is significant for either C2C12 and G1E-ER4 in terms of numbers, but the two measurements are only marginally correlated even where they overlap (supplementary figures 3.6 and 3.10).

The human HepG2 and K562 display a similar relationship between DNAse hypersensitivity and H3K27 acetylation to that found in mouse C2C12 cells with comparable numbers of candidate enhancers found in both TSS proximal and distal portions(supplemental figure 3.13).

Even the TSS distal genome-wide TF occupancy and DNAse hypersensitivity, which have almost identical genome-wide profiles (Figure 3.1B) have a signal strength correlation that is mediocre at best.

This lack of correlation between even the partially redundant biochemical measurements made it unlikely to find any correlation to activity on the assay.

Finally, we examined the quantitative correlation between biochemical signal and functional activity. Figure 3.5 shows the distribution of active and inactive muscle (Figure 3.5A) and erythroid (Figure 3.5B) cEnhs relative to the spectrum of DNAse and H3K27ac signal genome-wide.

While highly active enhancers are more often found among the most strongly H3K27ac+/DNAse+ regions in muscle cells, overall there is only modest correlation between biochemical marking and functional activity, and it is even less apparent in the erythroid context. We calculated the correlation between TF occupancy and biochemical marks on one hand, and functional activity on the other and found only a small positive quantitative correlation (Pearson r $2 \leq 0.10$; Spearman rank correlation $r \leq 0.40$) between each of these signatures and enhancer activity (Figure 3.5C), observations that also hold in the other contexts we examined (Supplementary Figures 3.5A-B, 3.8A-B, 3.12A-B, and 3.16B). We also evaluated the predictivity of biochemical signatures using a receiver operating characteristic curve analysis (Supplementary Figures 3.4C-F, 3.7C-D, and 3.10C-F). With the exception of erythroid cEnhs, where the combination of GATA1 and TAL1 was most predictive of functional activity, we find that DNAse and H3K27ac are most often the best predictors of functional enhancers. Their predictivity, however, is not incredibly strong, with AUROC values only exceeding 0.8 in K562 and HepG2 cells. Overall, the combination of H3K27ac and DNAse hypersensitivity appears to be as reliable a predictor of functional activity as any other biochemical signature, however, even it is in no way absolutely predictive of function, with only approximately half of H3K27ac+/DNAse+ actually exhibiting significant enhancer activity.

3.9 Discussion

In this study we examined the degree to which the biochemical signatures of a candidate enhancer infer active biological function.

Even though candidate enhancer signatures are selective for a remarkably small fraction of the genome, we found that only ~50% are able to individually activate transcription of a reporter gene. We found that predictivity remained constant across cell types and different signal-unbiased methods for candidate selection centered on DNase hypersensitivity sites flanked by H3K27ac modified histones.

We showed that the vast majority of distal candidate enhancer regions are biochemically marked in a tissue specific manner. However, many contain DNA motifs recognized by TFs specific to other tissue types; potentially allowing them to operate in a transient plasmid based transfection assay where the DNA likely remains outside of its native chromatin context. To address this criticism we demonstrated that the presence of multiple motifs in a region that is biochemically void are not sufficient to drive function in our assay, even when the region is an enhancer in other tissue types.

Finally, by studying cEnhs sampled across the full spectrum of ChIP occupancy for multiple transcription factors, we have demonstrated that the most strongly biochemically marked cEnhs are highly enriched for functionality. However, active elements, including some of the most powerful, are detected throughout the whole biochemical signal spectrum; with the bulk genome-wide residing in the vast population of modest biochemical signatures.

A possible explanation for this result can be gleaned from an enhancer located proximal to the muscle creatine kinase gene. This enhancer, studied over the last three decades for its outstanding function, displays only a modest myogenin ChIP signal that does not correlate with the $\geq$ 95% physical occupancy measured *in vivo* on the chromosome (Mueller PR, 1989; Garrity PA, 1992).

We did not observe a particularly strong correlation between the magnitude of enhancer activity in functional assays and strength of biochemical marks, as expected

if the decoupling of the biochemical measurements from the true physical occupancy is a genome-wide phenomenon.

These findings are in contrast to earlier studies, which reported over 80-90% activity for cEnhs defined using, for example EP300 ChIP-seq(Visel A, 2009). This is most likely due to the fact the these studies only focused on elements selected among the most strongly enriched and likely to be functional cEnhs rather than the full spectrum of ChIP-seq signal.

We find a smaller fraction (15-25%) of active cEnhs using a high-throughput ChIP-STARR-seq MPRA, but similar qualitative patterns across the spectrum of biochemical signatures defining cEnhs. The reasons for the lower activity rates returned by MPRAs are manifold, and include (but are likely not limited to) the fact that the DNA fragments used as input to the MPRA are shorter than the length of fully functional regions, and that ChIP-STARR-seq libraries do not provide deep and complex representation of the original pools of ChIP-seq fragments, leaving many modestly active enhancers with insufficiently many reads to cross the thresholds of statistical significance; both of these factors are expected to lead to high false negative rate.

We observe that biochemical marks can be decoupled from each other temporally, which can impact cEnh predictions based on their co-occurrence. For example, during muscle differentiation thousands of TF occupancy sites also exhibit DNAse hypersensitivity and EP300 localization but are not yet robustly acetylated at H3K27, and conversely, in differentiated cells H3K27ac can remain for some time associated with sites previously TF-occupied DNAse hypersensitive sites even though they are no longer open.

A possible mechanism of action for these orphaned H3K27Ac sites is demonstrated by GR stimulated A549 cells; where a significant subset of biologically active enhancers are from GR occupancy at these formerly orphaned H3K27ac locations from

unstimulated A549 cells. These orphaned sites may thus function as a guide for different responses to external stimuli; or lack thereof in different tissues (Supplemental figure 3.20).

The ENCODE consortium recently compiled a comprehensive list of candidate regulatory elements (cREs) based on DNAse-seq maps, CTCF occupancy and histone marks profiles resulting with 1.31M and 0.53M cREs in humans and mouse respectively. (ENCODE, 2012; Moore J, 2018) Elements within the registry were classed as "Promoter Like State" (PLS), "Enhancer Like State" (ELS) and CTCF.

In order to survey the coverage of the new encyclopedia we annotated our collection of tested candidate Enhancers which spans several tissues represented in the encyclopedia either directly, or indirectly (through similar tissues types). Figure 6A provides an example of our C2C12 tested elements in a 300kb region together with any ENCODE cRE annotation type (Figure 6B) found within our tested regions. The full collection of ENCODE Element Registry within this genomic region can be observed in Figure 6C (Agnostic) together with the biochemical annotations for two tissues. Embryonic limb (E14.5) was used as the closest match to muscle development, whereas liver from the same stage of development was included to provide an exogenous comparison.

The overall encyclopedia cRE coverage of our combined tested constructs is almost total (Figure 7A), including a remarkable 83% coverage for Mouse C2C12s (which had no direct representative). In the negative controls similar coverage was observed for elements that are expected to work in other tissues based on biochemical occupancy, whereas control elements picked from random regions which were negative for ChIP-seq signals generally did not register on the encyclopedia.

We also scored all of our elements again split into ones that score as enhances and ones that do not for cEnhancer (ELS) and cPromoter (CLS). (Figure 7B) We find

that the majority of elements that we tested score as expected in a cEnhancer (ELS) state. However we found a surprising number of PLS candidates within our tested elements, which were preferentially picked to be promoter distal (>2kb). These PLS elements are often found in intronic regions of genes (for example Figure 6 – R3) and appear to be an artifact of the high levels of H3K4me3 often found in gene bodies.

It is unclear if the many elements found inactive in this study, including some which have been shown to be physically connected to a gene, may not function as individuals, or they may require a specific enhancer-enhancer or promoter-enhancer pairing indicated by the physical interaction maps (Zabidi MA, 2015). Indeed local examples of elements found to function more strongly when paired with a neighboring cEnhancer can be observed in R2, R3, R4 and R5 of Figure 6A. Interestingly, R1 shows the opposite behavior indicating that cryptic repressor sites may be involved in either tuning or suppressing the function of some of these individual enhancers.

Some of these elements found to be inactive may only able to function in specific cell contexts. For example, the formation of muscle across different compartments of the embryo commonly uses myogenin, but involves both different temporal domains and the exposure to different cofactors. This would impart for an additional level of control to activate only a subset of the programming.

We expect these findings from the ENCODE functional tests, paired with the collection of higher throughput enhancer assays, to help guide future efforts towards annotating and further testing the other possible functions of elements in mammalian genomes. Taken as a combination these elements provide a high quality repertoire of tested elements and indicate that further testing will likely be needed to fully catalog the genome in future efforts that expand beyond the high quality DNAse and

H3K27ac occupancy.

3.10 Methods

Except where otherwise stated, all analyses were performed using custom-written python scripts. The GENCODE

Cell culture

C2C12 cells

C2C12 myoblasts were maintained and seeded for transfection in 20% FBS supplemented DMEM medium. Upon reaching >80% confluence, the cells were were differentiated using 2% horse serum and 1 μM insulin in DMEM medium.

G1E cells

G1E cells were grown according to previously published protocolsWang H, 2006; Cheng Y, 2008; Dogan N, 2015.

K562, HepG2 and A549 cells

K562; HepG2 and A549 cells were grown according to the approved ENCODE cell culture protocols publicly available through the ENCODE portal (https://www.encodeproject.org/).

Functional assays

Cloning and DNA purification

The specifics of each selection set, the promoters used for each cell line, and other details are publicly available through the ENCODE portal (https://www.encodeproject.org/).

Functional assay testing of cEnhs in C2C12 cells

Candidate REs and negative control regions were either PCR-amplified from female BALB/C purified mouse genomic DNA (Switchgear Genomics) or synthesized *de*

novo (Genscript). The resulting DNA was cloned into a reporter vector 5′ of a custom TK promoter (SwitchGear Genomics) driving a high-turnover sequence-optimized luciferase reporter gene. Plasmids were purified using Miniprep kits (Qiagen) and standardized to 30 ng/μL using the Qubit® dsDNA HS (High Sensitivity) Assay Kit.

For the purpose of testing elements in the myoblast state, undifferentiated C2C12 cells were seeded in 96-well delta surface plates (NUNC) quadruplicates 12 hours before transfection at a concentration of 2500 cells/well. For the purpose of testing elements in the myocyte state, undifferentiated C2C12 cells were seeded at a density of 3500 cells/well. Transfections were carried out with 50 ng of DNA per construct in each replicate using Lipofectamine LTX, after a 5 minute incubation with a 1:16 dilution with the PLUS reagent (Thermo Fisher). Myoblast plates were lysed using a Steady-Glo® kit, and luminescence was measured on a plate luminometer 24 hours post-transfection. Myocyte plates had their media exchanged with differentiation 12-16 hours post transfection and measured following the same procedure 24 hours later.

Aside from the plate reading step, the entirety of the transfection process was automated and carried out on a Tecan Freedom EVO 200 robot.

Functional assay testing of cEnhs in K562 and HepG2 cells

The set of K562 and HepG2 cEnh regions was PCR-amplified and cloned 5′ of the promoter of enhancer assay plasmids containing luciferase and Renilla reporter genes; cloning was performed by SwitchGear Genomics. Each construct was quantified (using Qubit) and standardized to 30ng/μL before use in transfection assays. **Chris Partridge please review this**

Functional assay testing of erythropoetic cEnhs

G1E candidate enhancer regions were tested in K562 cells according to protocols publicly available through the ENCODE portal (https://www.encodeproject.org/.

Functional Assay Data processing

For each cEnh or negative control measurement, the ratio between its value and the corresponding basal promoter vector (relative assay activity) was calculated. Active cREs were discriminated from inactive using a z-score analysis, comparing the population of test element technical replicate values to the set of negative controls.

ChIP-seq experiments

Chromatin immunoprecipitation in A549 cells was performed as previously described (Reddy et al. 2009) using 2×10^7 A549 cells per replicate. Cells were sonicated using a Bioruptor XL (Diagenode) on the high setting until the resulting chromatin was fragmented to a median fragment size of 250 nt as assayed by agarose gel electrophoresis. GR ChIP was performed using 5 μg of a rabbit polyclonal α-GR antibody (Santa Cruz Biotechnology sc-1003), and 200 μl of magnetic sheep anti-rabbit beads (Life Technologies M-280). H3K27ac ChIP was performed using **XXX Ab source XXX**. After reversal of formaldehyde crosslinks at 65 °C overnight, DNA was purified using MinElute DNA purification columns (QIAGEN). Illumina sequencing libraries were then generated using the Apollo 324 liquid handling platform according to manufacturer's specifications (Wafergen).

ChIP-seq in C2C12 cells was performed using chromatin from 2×10^7 nuclei, which was fragmented using a Misonix probe tip sonicator and subjected to immunoprecipitation using a robotic ChIP pipeline described before. The resulting purified DNA was then converted into sequencing libraries and sequenced on an HiSeq 2500 (Illumina) as described previouslyGasper WC, 2014. The following antibodies were

used: α-myogenin (Santa Cruz Biotechnology SC-12732, lot K2311), α-MyoD (Santa Cruz Biotechnology SC-32758, lot J3115), α-MEF2 (Santa Cruz Biotechnology SC-17785, lot H1913), α-p300 (Santa Cruz Biotechnology SC-585, lot H3115), α-E2A (Santa Cruz Biotechnology SC-349X, lot B1207), α-H2B (Santa Cruz Biotechnology SC-357, F2305), and α-H3K27ac (Active Motif 39133, lot 34849).

In addition, publicly availableDellOrso S, 2016 Pbx1 ChIP-seq and Control datasets were downloaded from GEO accession GSE76010.

For G1E, K562 and HepG2 cells, previously publicly availableENCODE, 2011; ENCODE, 2012; Yue F, 2014 ChIP-seq datasets were downloaded from the ENCODE portal (https://www.encodeproject.org/).

DNAse-seq experiments

In C2C12 cells, DNAse-seq was carried out as follows: **XXXXX DETAILS XXX**

In A549 cells, DNAse-seq was carried out as follows: **XXXXX DETAILS XXX.**

For G1E, K562 and HepG2 cells, previously publicly availableENCODE, 2011; ENCODE, 2012; Yue F, 2014 DNAse-seq datasets were downloaded from the ENCODE portal (https://www.encodeproject.org/).

STARR-seq experiments

The STARR-seq experiments previously published by Vockley et al.Vockley CM, 2016 were used in this study.

Genomic coordinate conversion

The regions to be tested using functional assays were designed based on the `mm8` and `mm9` verions of the mouse genome and the `hg19` version of the human genomes. Conversion of the original coordinates to `mm10` and `hg20` coordinates was performed

using the `liftOver` tool from the UCSC Genome Browser UtilitiesTyner C, 2017.

Conservation analysis

Sequence conservation analysis were carried out using the `phastCons60way` and `phastCons100way` conservation tracks, which were downloaded from the UCSC Genome BrowserTyner C, 2017.

ChIP-seq data processing and analysis

ChIP-seq reads were trimmed down to 36 bp in length and mapped against the `hg20` (for human samples; the male or female version depending on the sex of the cell line the sample originated from) and `mm10` (for mouse samples) using Bowtie Langmead B, 2009 (version 1.0.1) with the following settings: `-v 2 -k 2 -m 1 --best --strata`. DNAse-seq reads were processed similarly except that they were trimmed down to 20bp for A549 samples and 36bp for C2C12 cells (due to differences in the experimental protocol used to generate the data).

Peak calling was carried out as follows. For DNAse and H3K27ac datasets, MACS2 Feng J, 2012 (version 2.1.0) was run on individual replicates and on pseudoreplicates (generated by randomly splitting the pooled set of reads for both replicates into two) with relaxed settings (`--to-large -p 1e-1`). For H3K27ac control datasets were subjected to the same treatment (no background/control is available for DNAse data) The top 100,000 peaks from each replicate or pseudoreplicate (ranked by q-value) were then used as input into IDR Li Q, 2011. The number of peaks above a given IDR threshold called as reproducible between true replicates (N_t) and between pseudoreplicates (N_p) were recorded. Peak calling was then carried out on the pooled set of reads and the top $max(N_t, N_p)$ peaks were chosen as the final set of reproducible peaks. For point-source Pepke S, 2009 datasets (transcription factors), peak calling was carried out following the same procedure but using SPP Kharchenko PV, 2008 (version 1.10.1), using the top 300,000 peaks as input to IDR.

The pooled sets of reads were also used to calculate RPM (reads per million) enrichment values over elements tested in functional assays.

STARR-seq data processing and analysis

STARR-seq and STARR-seq control/input reads (2×25mers) were mapped as paired ends to the `hg20` version of the human genome using Bowtie with the same settings as described above. Post-IDR peaks obtained from GR ChIP-seq were used as the list of candidate cEnhs to be scored using the STARR-seq data. For each STARR-seq and STARR-seq control/input replicate, raw fragment counts were obtained from every GR ChIP-seq peak; in addition, the rest of the genome (i.e. the regions that fall between the post-IDR GR ChIP-seq peaks) was split into bins of at most 50 kb length, and read counts were calculated for all such regions. The fragment counts for GR ChIP-seq peaks and for the intervening regions were combined together and used as input to DESeq2Love MI, 2014 for estimating differentially represented regions between STARR-seq and control/input libraries (at FDR-adjusted $p \leq 0.05$). The lowest average fragment counts value which was scored as significantly significant by DESeq2 was identified for each comparison, and all GR ChIP-seq regions with average fragment counts lower than this value were excluded from subsequent analysis, as such regions were not sufficiently represented in the available sequencing data to be reliably scored as active or inactive. We also carried out a fragment-level analysis, in which read counts were calculated for each individual sequencing fragment (defined as the pair of positions $\{i, j\}$, where i and j are respectively the $5'$ and $3'$ ends of the first and the second sequencing reads in a pair), using the same DESeq2 framework.

3.11 Acknowledgments

Robotic ChIP-Seq (Gasper WC, 2014) and transfections for C2C12 cells were performed in collaboration with Jost Vielmetter at the California Institute of Technology

Protein Expression Center.

Library generation and high-throughput sequencing for C2C12 ChIP-Seq samples was performed by Igor Antoshechkin at the California Institute of Technology Millard and Muriel Jacobs Genetics and Genomics Laboratory.

The authors would also like to thank Diane Trout and Henry Amrhein for technical assistance with maintaining the computational infrastructure used to carry out this study.

This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433.

3.12 Figures (Chapter 3)

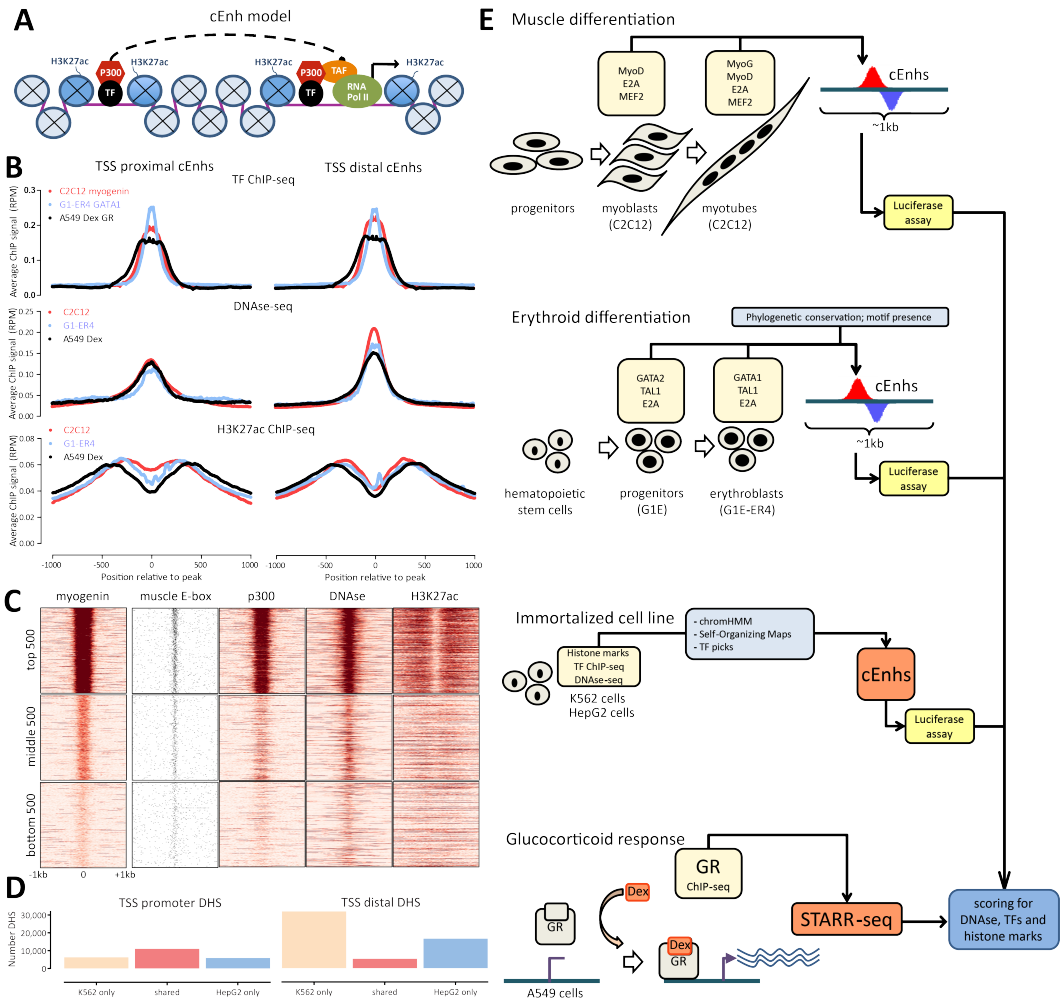


Figure 3.1: (Caption on next page.)

Figure 3.1: Biochemical signatures and functional testing of candidate enhancer elements (cEnhs) in mammalian genomes. (A) Biochemical signatures of cEnhs and promoters. Active enhancers are characterized by DNAse hypersensitivity due to nucleosome depletion, by p300 occupancy and by H3K27ac, as well H3K4me1 (not shown). Promoter elements share some of these features, but also associate with components of the transcription and transcription initiation machineries, and are marked by H3K4me3 (not shown); (B) Genome-wide commonalities and differences between the biochemical signatures of enhancers and promoters. Shown is the average signal profile around TSS distal (right; defined as regions more than 1kb away from an annotated TSS) and TSS proximal (left) cEnhs (defined as statistically significant peaks in the respective datasets; see the Methods section for further detail) in mouse and human cells for TFs (myogenin in differentiating mouse muscle cell, GATA1 in erythroid mouse cells, and the glucocorticoid receptor upon Dexamethasone stimulation of human A549 cells), DNAse hypersensitivity and H3K27ac; (C) The distribution of biochemical signal strength varies over a large continuum. Shown are the signal distribution for myogenin, p300, DNAse-seq, and H3K27ac relative to the summits of the top 500, middle 500 and bottom 500 reproducible myogenin ChIP-seq sites (total $n = 32,278$) in differentiated C2C12 muscle cells, as well as the distribution of the cognate myogenin TF binding motif. (D) Different cell types share a small fraction of their distal cEnh elements, in contrast to promoter elements. Shown are the common and cell-type specific TSS proximal (within 1kb of an annotated TSS) and TSS distal DHSs between the human erythroid K562 and hepatocyte HepG2 immortalized cell lines; E) Outline of cENH selection approaches, biological systems, experimental design and functional assays used in this study. Sets of cEnhs for functional testing were compiled based on: TF ChIP-Seq occupancy measurements (of the master regulators of muscle differentiation, MyoD and myogenin) in differentiating mouse C2C12 cells; phylogenetic conservation patterns and TF occupancy measurements (of the regulators of erythropoiesis GATA1 and TAL1) in differentiating mouse G1E-ER4 cells; TF occupancy (multiple TFs) in immortalized K562 cells; machine learning methods (Self-Organizing Maps, chromHMM and Segway) defining integrated chromatin states over multiple histone modification, DNAse and TF occupancy measurements. These cEnhs were tested using luciferase assays. In addition, DNA fragments from GR ChIP-seq experiments in Dex-stimulated A549 cells were cloned and assayed for activity using the STARR-seq assay. Active elements identified using these methods were then evaluated for the presence and distribution of various biochemical signatures.

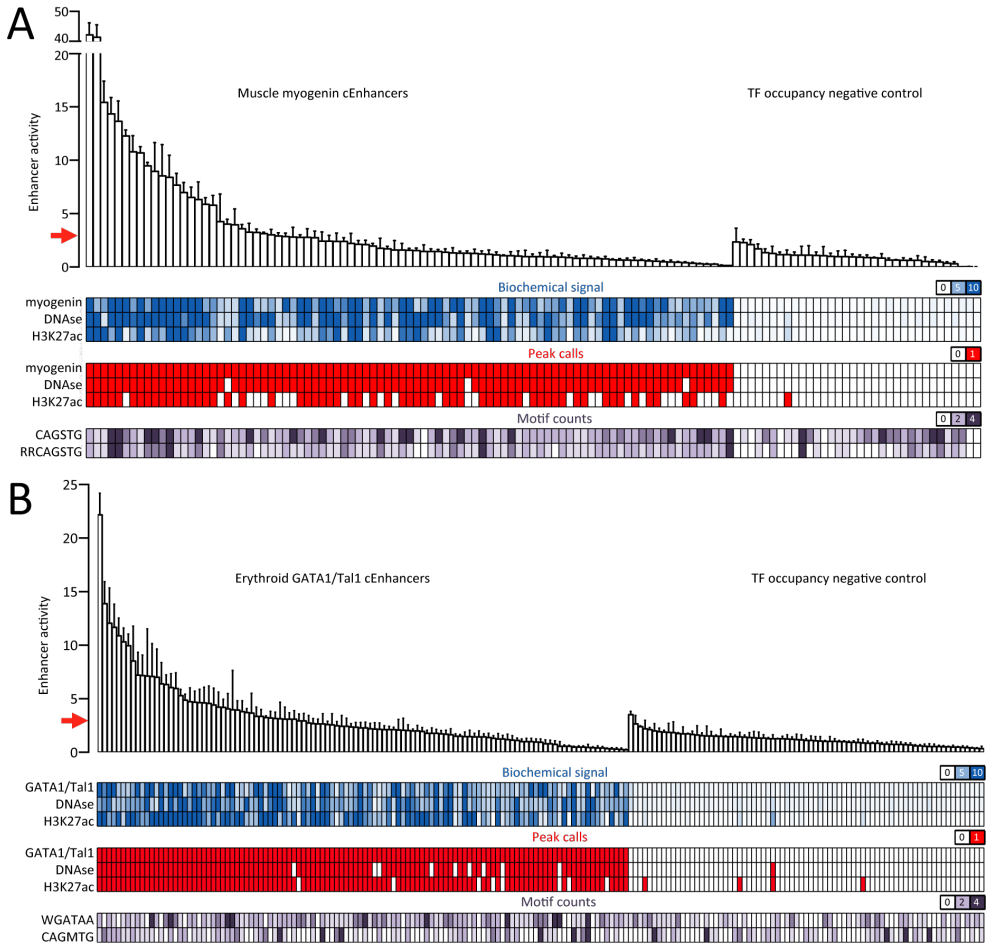


Figure 3.2: (Caption on next page.)

Figure 3.2: Functional testing of cEnh regulatory activity in mammalian cells. (A) Functional assay testing of cEnh regulatory activity in the context of muscle differentiation. Shown is luciferase assay fold activity in differentiated C2C12 myocytes across technical replicates ($n = 4$). The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, for each cEnh DNAse hypersensitivity, H3K27ac status, and myogenin occupancy are shown, both as RPM (Read Per Million) signal intensity values and as binary peak calls, as well as the number of myogenin motif (RRCAGSTG, derived from myogenin ChIP-seq data) occurrences. Tested cEnhs are sorted by mean fold activity. (B) Functional assay testing of cEnh regulatory activity in the context of erythropoiesis. Shown is luciferase assay fold activity in K562 cells across biological ($n \in [1 : 9]$) and technical replicates ($n = 4$ for each biological replicate). The red arrow corresponds to the mean fold activity threshold above which elements are considered active. In addition, for each cEnh DNAse hypersensitivity, H3K27ac status, and GATA1/TAL1 occupancy are shown, both as RPM (Read Per Million) signal intensity values and as binary peak calls, as well as the number of TAL1 (CAGMTG) and GATA1 (WGATAA) motif occurrences. Tested cEnhs are sorted by mean fold activity.

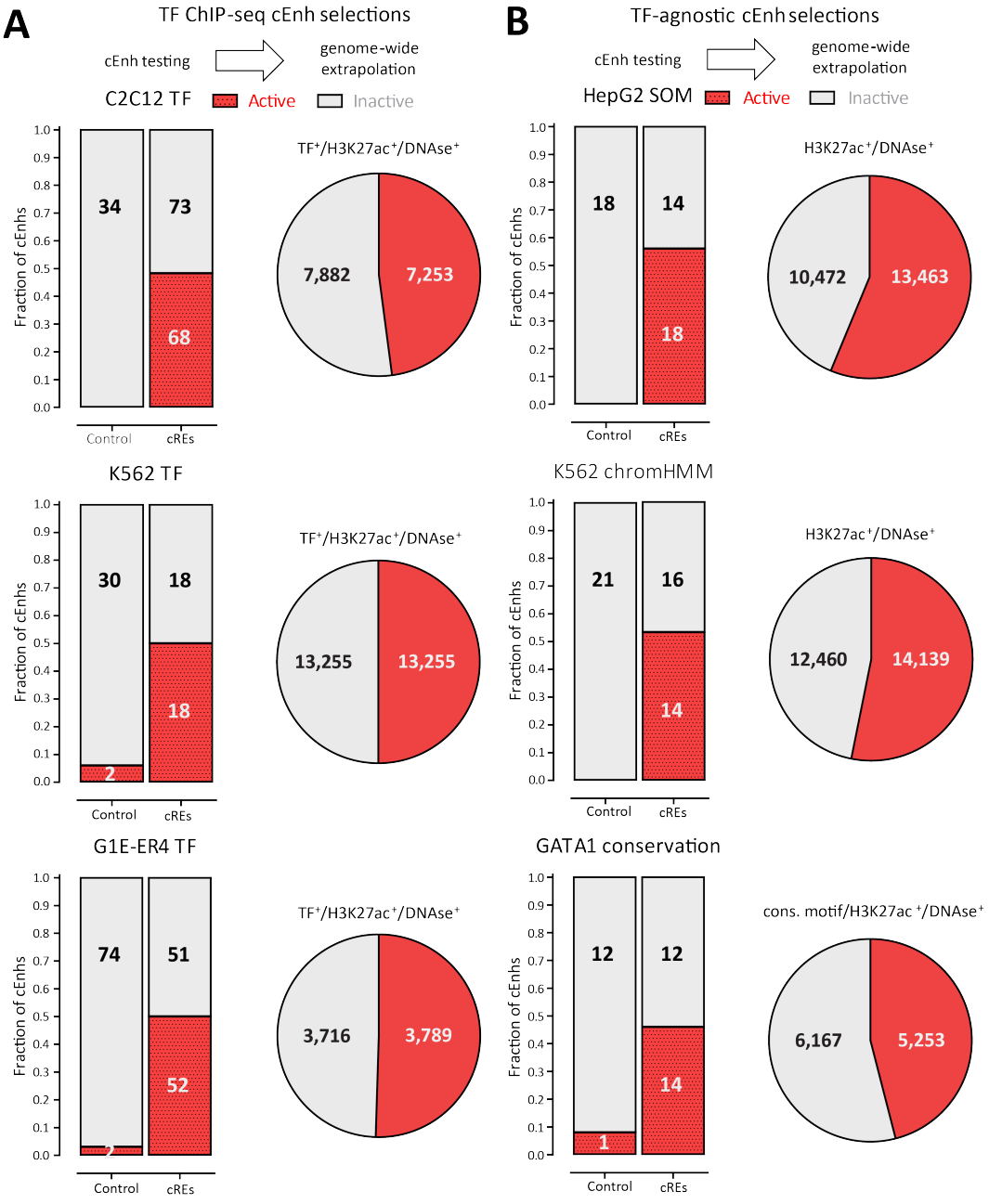


Figure 3.3: (Caption on next page.)

Figure 3.3: Summary of cEnh activity predictions by different selection criteria. (A) TF occupancy-centered selections. Tested eEnhs selected on the basis of TF occupancy in the context of mouse muscle differentiation and erythropoiesis and in human K562 cells were further subselected with the additional requirement of exhibiting DNAse hypersensitivity and the H3K27ac histone mark. The fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of TF^{+}/$DNAse^{+}$/$H3K27ac^{+}$ regions in the genome; (B) TF-occupancy agnostic selections. Tested eEnhs selected using Self-Organizing Maps in HepG2 cells, chromHMM in K562 cells, and evolutionary conservation of GATA1 motifs in G1E cells were further subselected with the additional requirement of exhibiting DNAse hypersensitivity and the H3K27ac histone mark. The fraction of active constructs in negative controls and cEnhs are shown on the left. The expected number of active cEnhs genome-wide is extrapolated on the left based on the number of $DNAse^{+}$/$H3K27ac^{+}$ (for HepG2 SOM and K562 chromHMM selections) $DNAse^{+}$/$H3K27ac^{+}$ regions with a conserved GATA1 motif (for GATA1 conservation selections) in the genome.

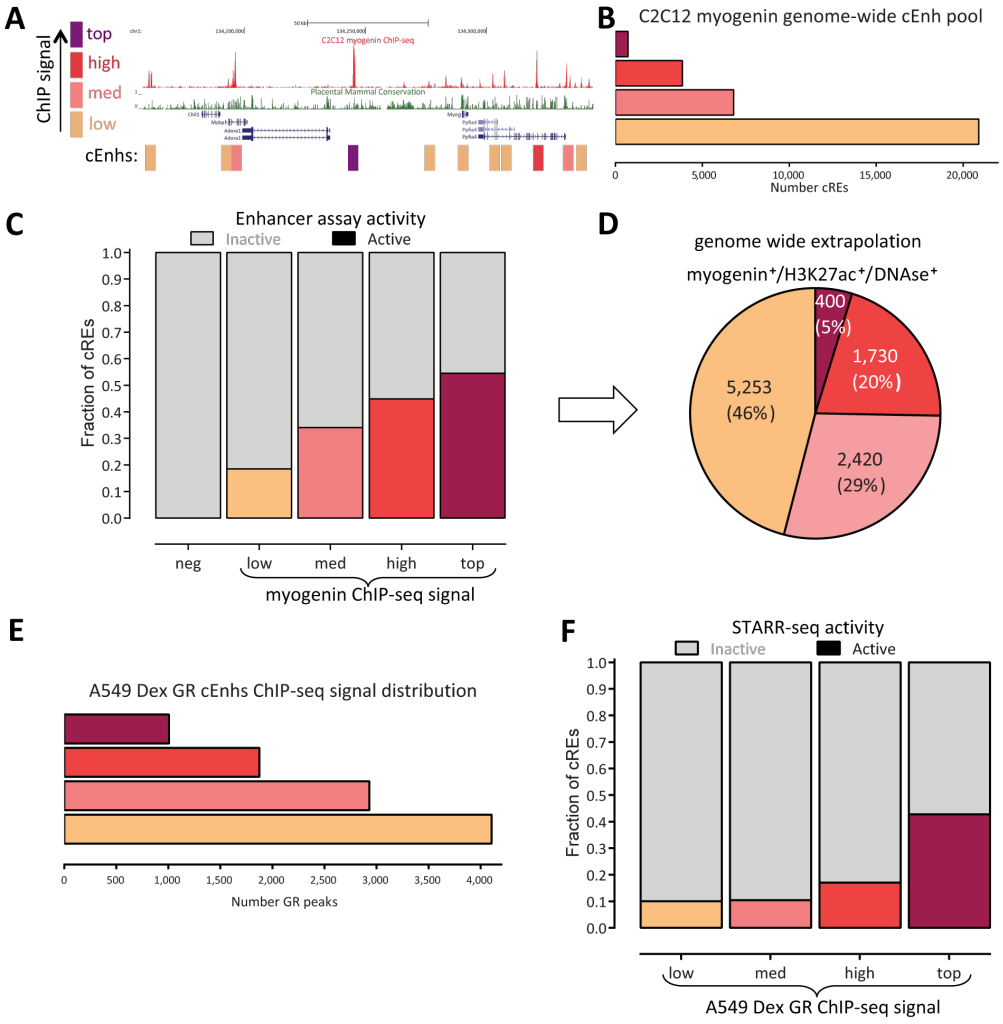


Figure 3.4: (Caption on next page.)

Figure 3.4: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) cEnhs (rectangle boxes) belonging to different signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": RPM $\geq$ 10; "high": RPM$\in$ [5, 10]; "medium": RPM$\in$ [2.5, 5]; "low" RPM$\leq$ 2.5) in the neighborhood of the mouse *Myog* gene; (B) Genome-wide distribution of cEnhs in different signal classes based on ChIP-seq data for myogenin in C2C12 myocytes; (C) Fraction of active enhancers in different cEnh signal classes (based on ChIP-seq data for myogenin in C2C12 myocytes; "top": $n = 66$; "high": $n = 49$; "medium": $n = 45$; "low" $n = 27$) as well as in negative controls (with not myogenin occupancy; $n = 34$). Only cEnhs positive for myogenin, DNAse and H3K27ac were included; (D) Extrapolated number of active enhancers in C2C12 belonging to each signal strength class based on the genome-wide number of myogenin$^+$/DNAse$^+$/H3K27ac$^+$ regions. (E) Genome-wide distribution of cEnhs in different signal classes based on the set of GR ChIP-STARR-seq cEnhs in A549 cells ("top": A549 Dex GR ChIP-seq RPM $\geq$ 10; "high": RPM$\in$ [5, 10]; "medium": RPM$\in$ [2.5, 5]; "low" RPM$\leq$ 2.5). Only GR ChIP-seq regions significantly represented within STARR-seq libraries (i.e. with sufficiently many reads to score as active if they were in fact active) are shown for each signal class. (F) Fraction of cEnhs exhibiting significant activity in the GR ChIP-STARR-Seq assay in stimulated A549 cells for each signal strength class.

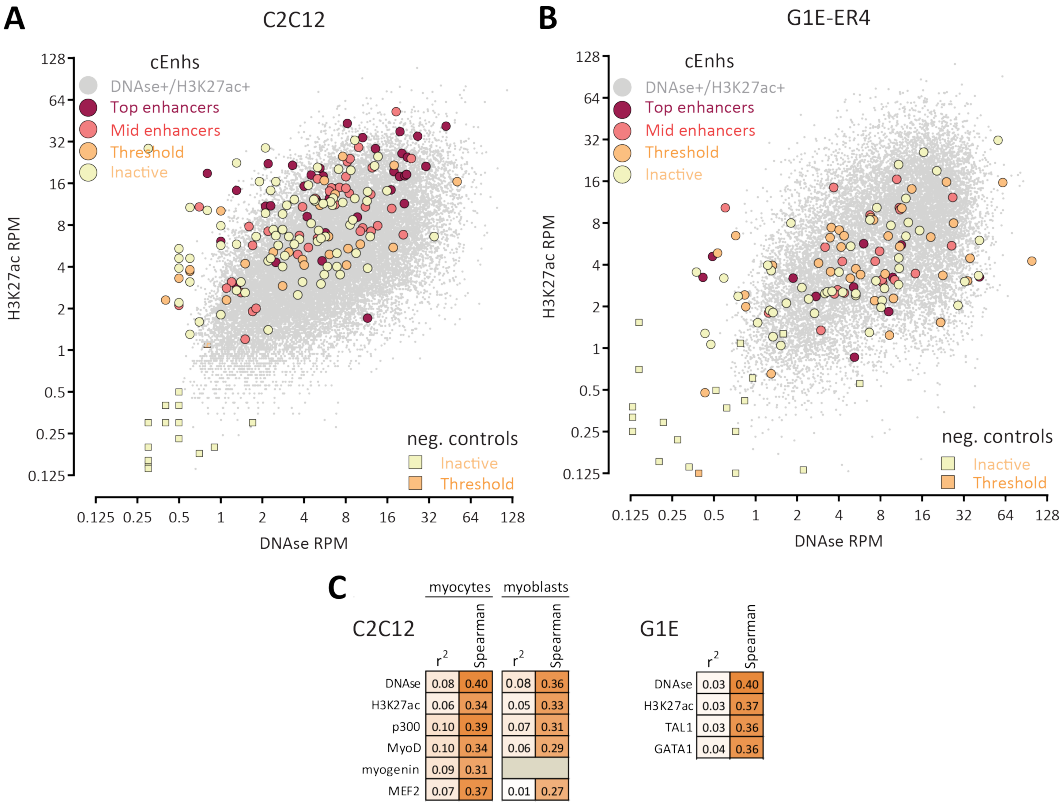


Figure 3.5: Absence of general strong correlation between biochemical signal strength and enhancer activity of cEnhs. (A) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27ac signal distribution in C2C12 myocytes. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEns tested for activity in C2C12 myocytes (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (B) Distribution of tested cEnhs relative to the genome-wide DNAse and H3K27ac signal distribution in G1E-ER4 cells. Shown are DNAse and H3K27ac RPM values for all DNAse$^+$/H3K27ac$^+$ regions as well as for cEnhs tested for activity (outlined circles) and for occupancy negative control (outlined squares), with tested cEnhs separated into four classes based on their measured enhancer activity, from dark red (most active) to yellow (inactive). (C) Correlation between biochemical signals and measured enhancer activity in C2C12 and G1E cells. See also Supplementary Figures 3.4, 3.8, and 3.12 for more details.

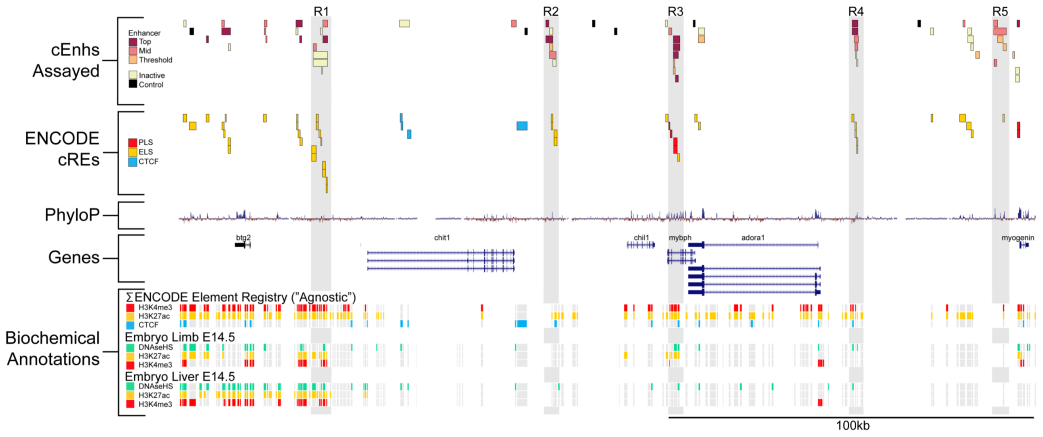


Figure 3.6: Summary of activity for cEnhs (A) in the genomic span between the BTG2 and myogenin loci with the corresponding ENCODE cRE types annotations from the ENCODE encyclopedia (B), PhyloP conservation track, coupled with the underlying biochemical annotations from "agnostic"; Embryo Limb and Liver at E14.5 (C). Candidate Enhancer elements (A) of identical size that appear twice represent both possible orientations relative to the promoter being tested.

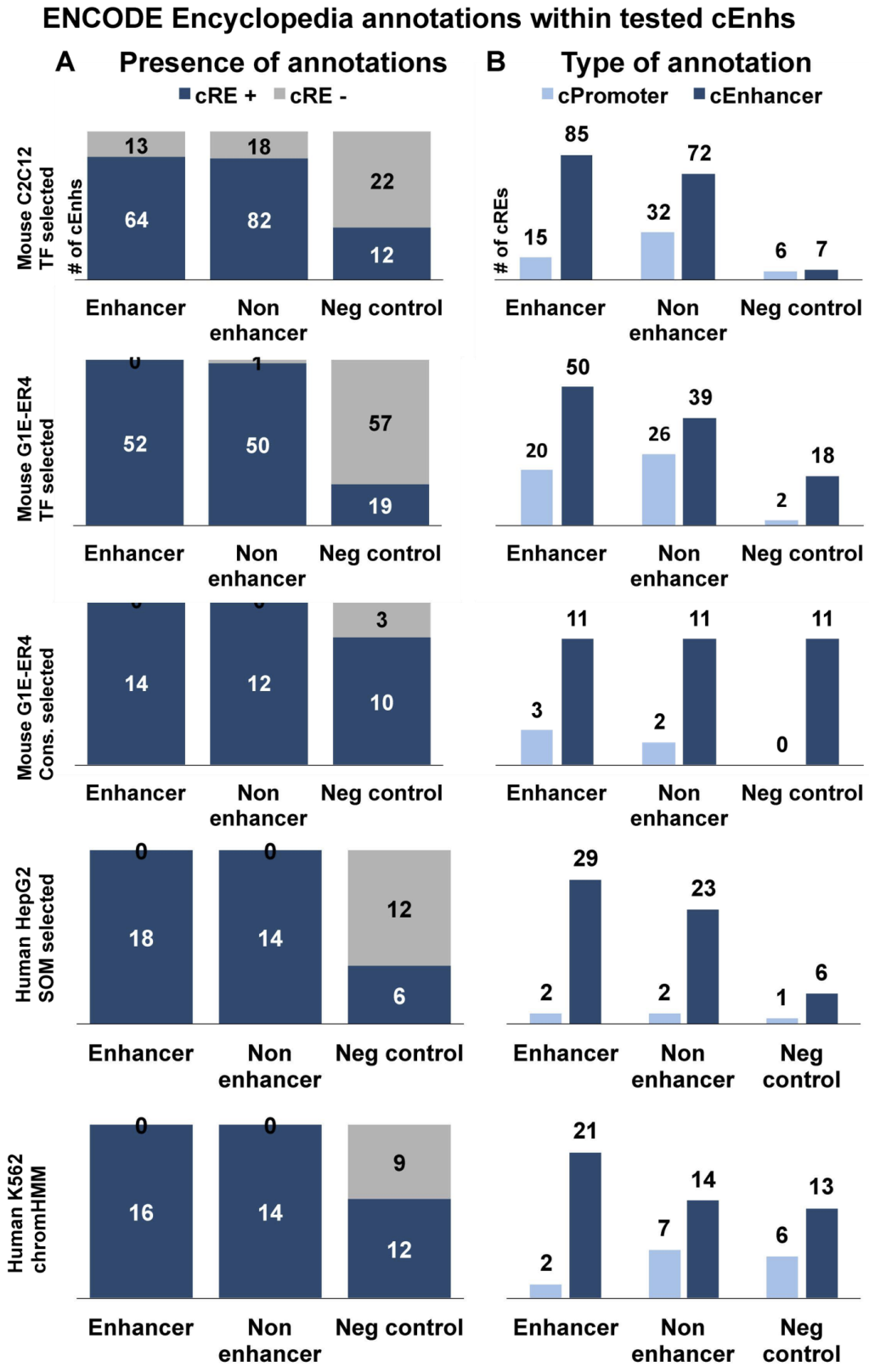
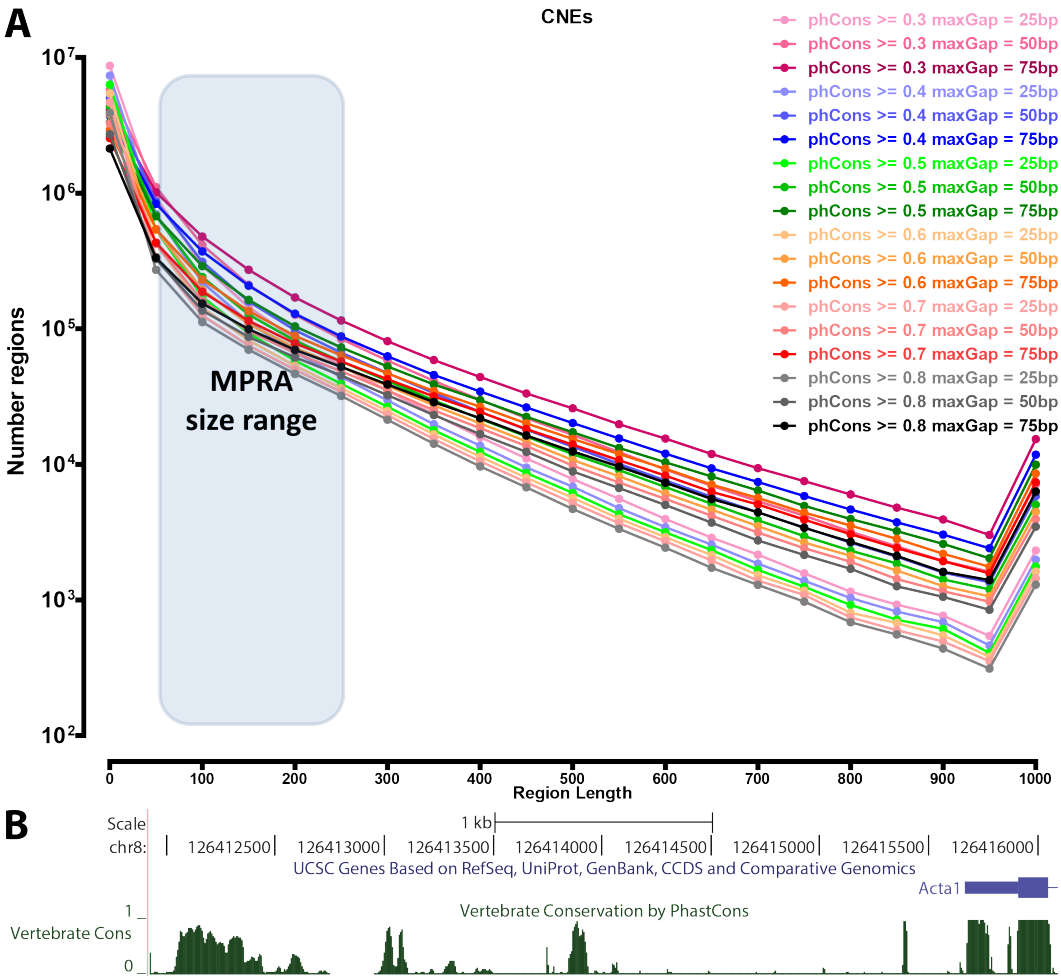


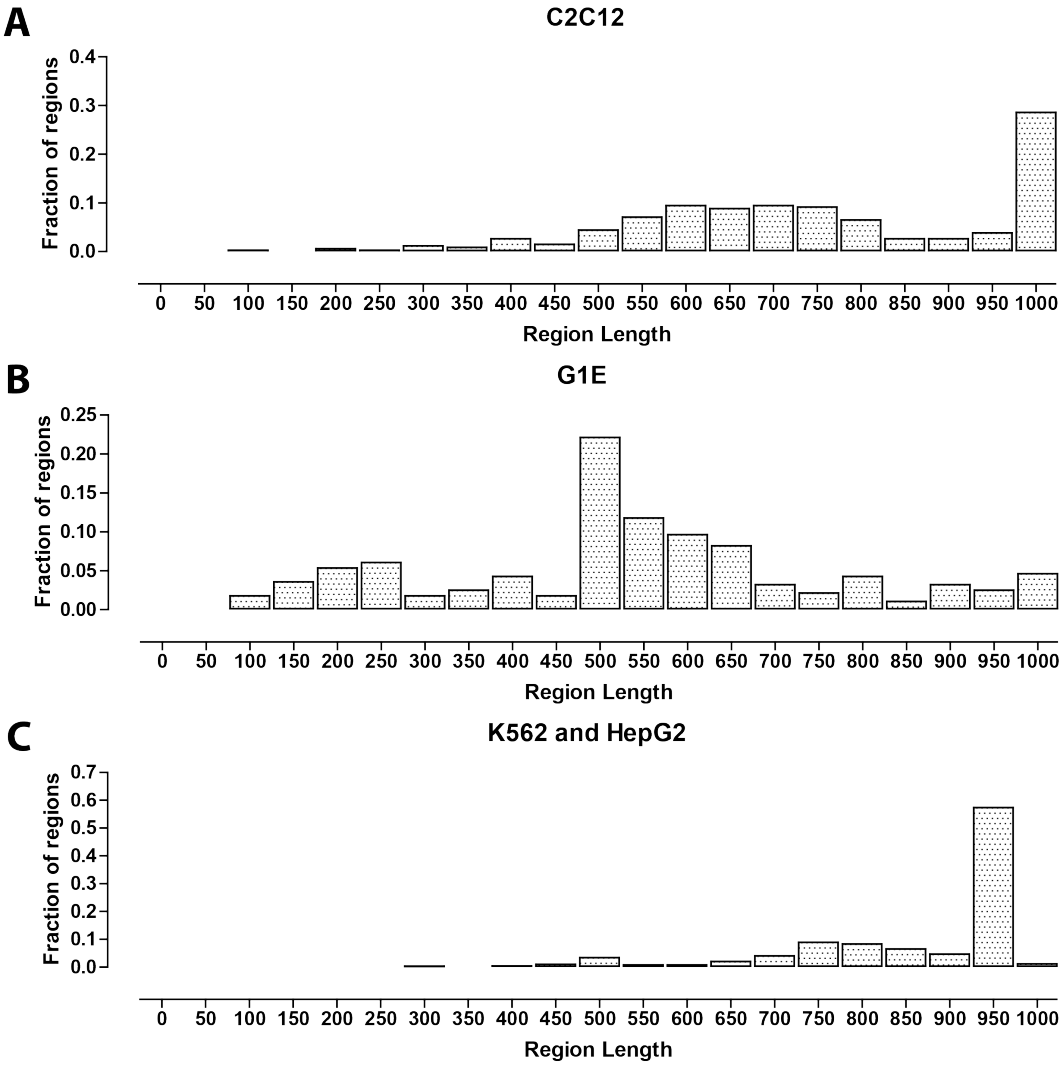
Figure 3.7: (Caption on next page.)

Figure 3.7: Summary of cEnhs from our experiments compared to ENCODE cRE annotations. Elements tested in our selections from both mouse and human genomes were classed by whether they are scored as a biologically active/inactive element and scored for presence (A) and type (B) of ENCODE cREs in each class. Negative control elements from each study were included for comparison.

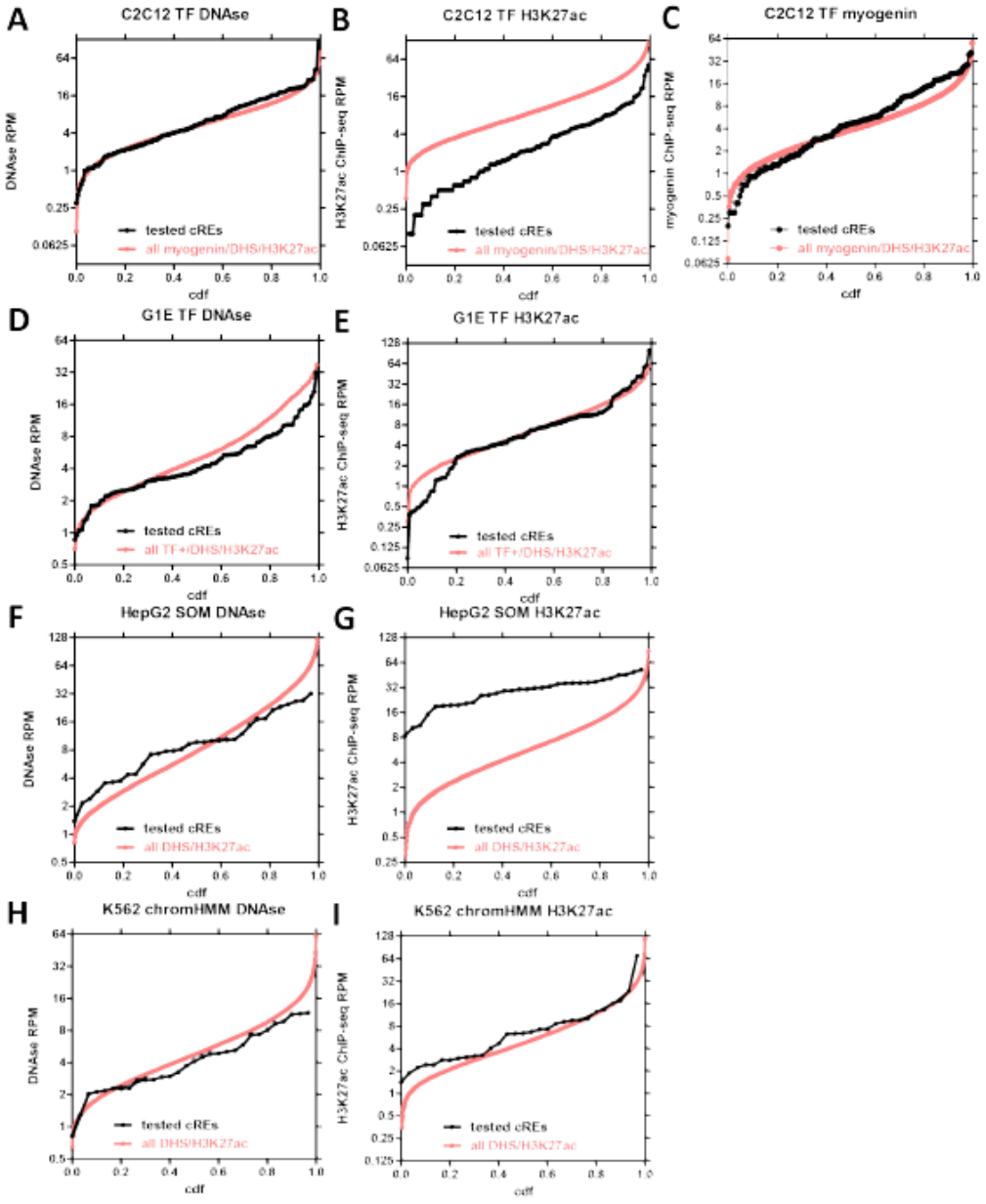
3.13 Supplementary Figures



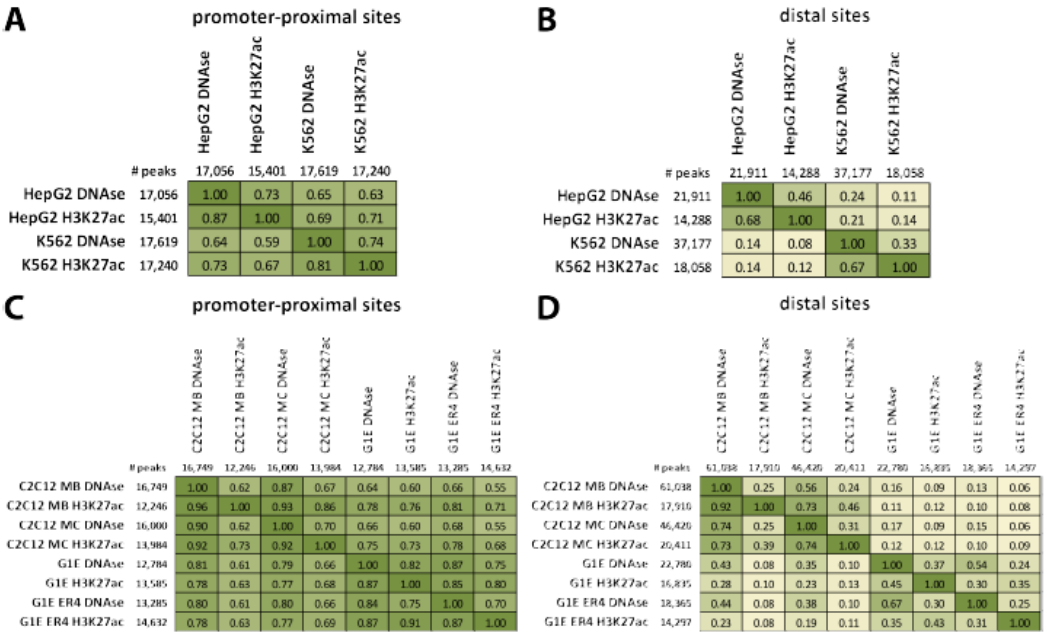
Supplementary Figure 3.1: The length of thousands of conserved noncoding elements in mammalian genomes greatly exceeds the size range of MPRA constructs. (A) The length distribution of conserved noncoding regions in the human genome. The `phastCons100way` conservation track for the `hg20` version of the human genome was downloaded from the UCSC Genome Browser. Blocks of conservation, in which all nucleotides have `phastCons` scores higher than the indicated minimum (`phCons`), were identified, and then merged into larger regions if the length of the gaps between them was smaller than the indicated `maxGap` parameter. The distribution of the lengths of the resulting sets of regions was plotted. This approach captures the properties of enhancer elements observed in the genome, which often consist of multiple blocks of highly conserved sequences separated by gaps of less conserved sequences, resulting in an enhancer element of up to a few hundred base pairs in length or more. (B) Such an example is shown for the *Acta1* gene in mouse.



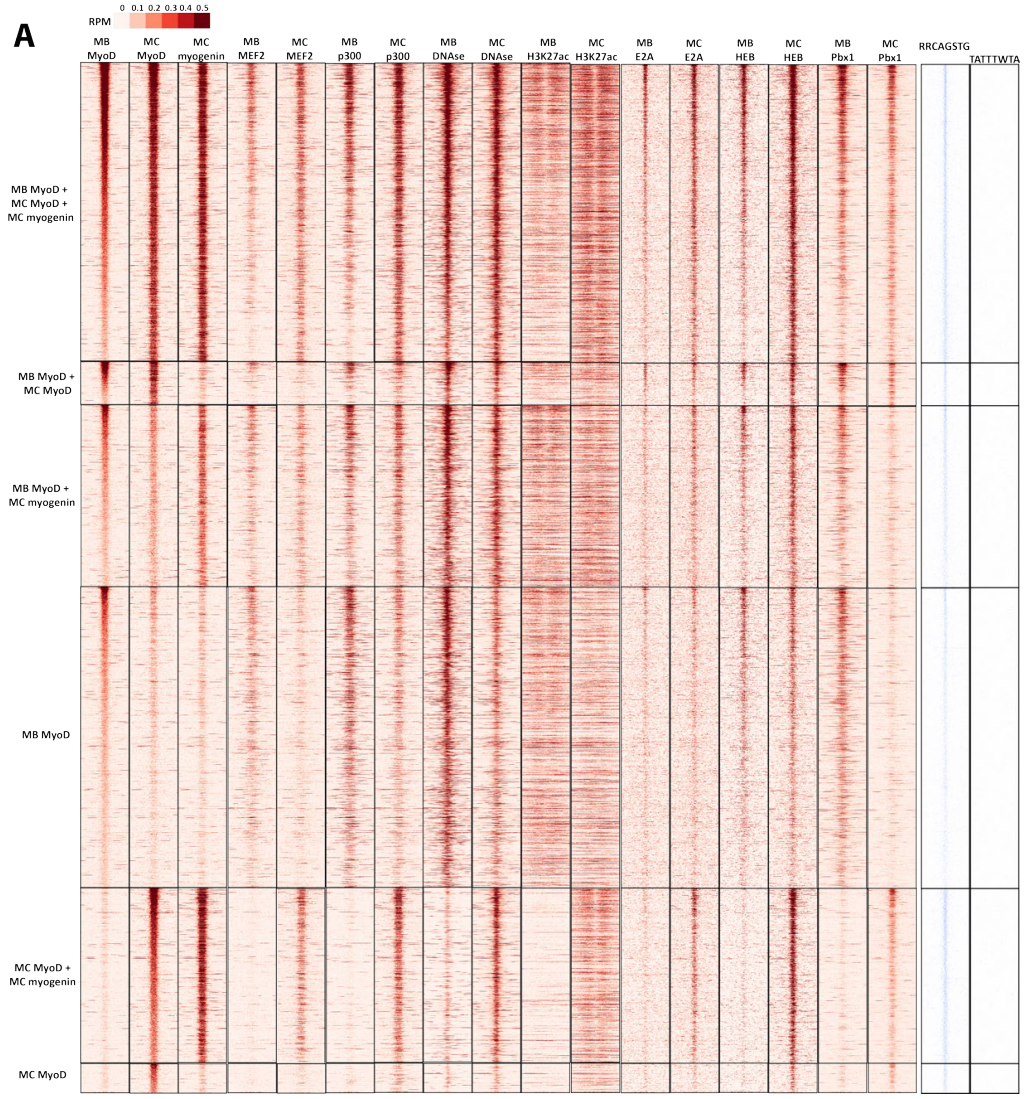
Supplementary Figure 3.2: Length distribution of functional assays constructs used to test cREs in this study. (A) Distribution of functional assay construct lengths tested in this study in C2C12 cells. (B) Distribution of functional assay construct lengths tested in this study in G1E cells. (C) Distribution of functional assay construct lengths tested in this study in K562 and HepG2 cells.

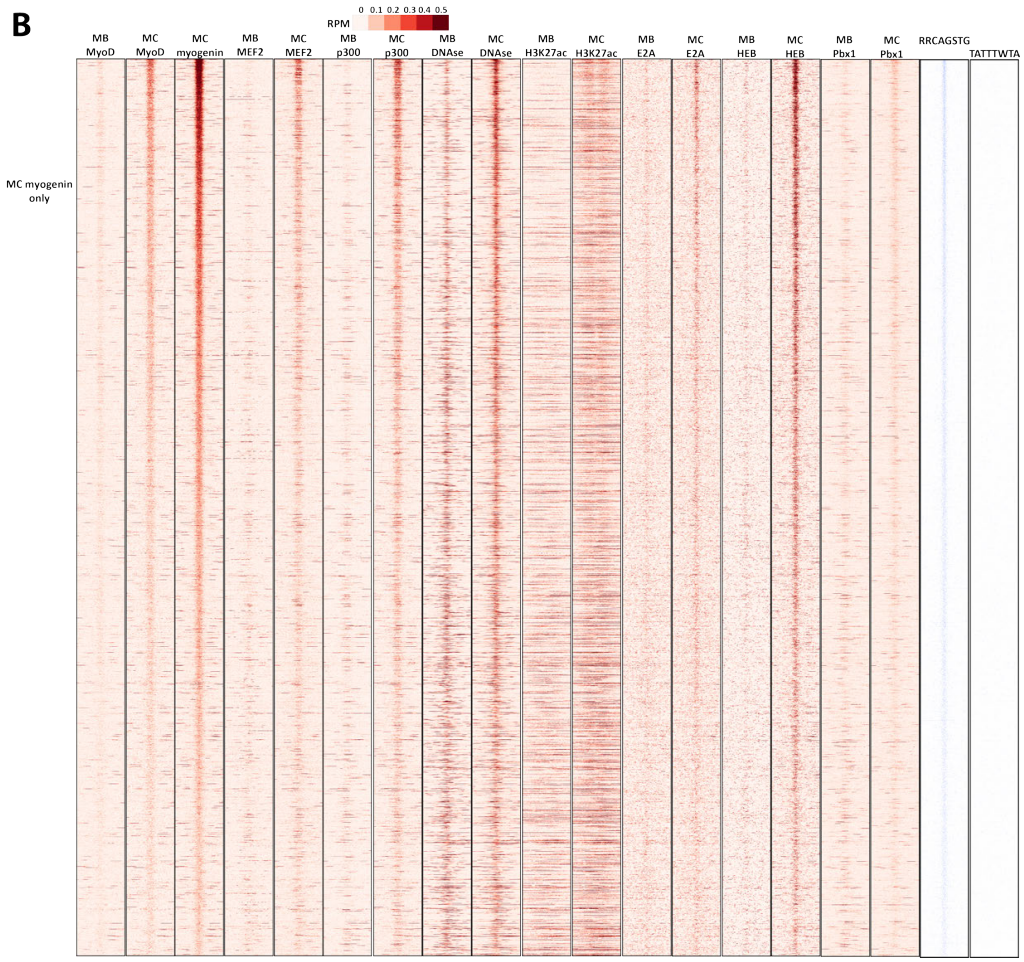


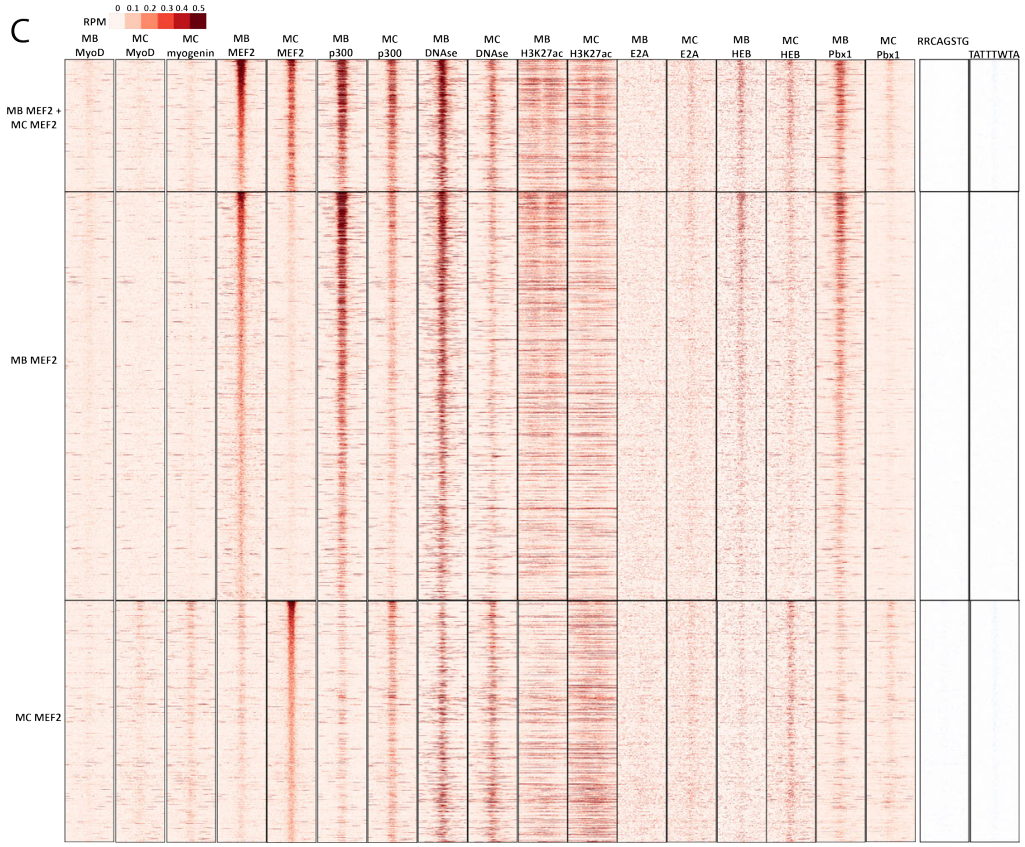
Supplementary Figure 3.3: Distribution of biochemical signal in tested cEnhs and genome-wide. Shown is the distribution of ChIP-seq or DNAse-seq RPM values for the set of cEnhs tested and for the genome-wide set of cEnh with similar biochemical signatures shown in Figure 3.3.



Supplementary Figure 3.4: Differential marking of proximal and distal cREs by DNAse and H3K27ac between different cell types and cell states. (A) Promoter-proximal (within ≤1 kb of an annotated TSS) sites in K562 and HepG2 cells; (A) Distal (≥1 kb from an annotated TSS) sites in K562 and HepG2 cells; (C) Promoter-proximal (within ≤1 kb of an annotated TSS) sites in differentiated and undifferentiated C2C12 and G1E cells; (D) Distal (≥1 kb from an annotated TSS) sites in differentiated and undifferentiated C2C12 and G1E cells. The overlap score (O_{xy}) shown in each cell (x, y) indicates the fraction of peaks in the dataset on the y-axis that are also found in the dataset on the x-axis, i.e. $O_{xy} = |X \cap Y|/|Y|$.

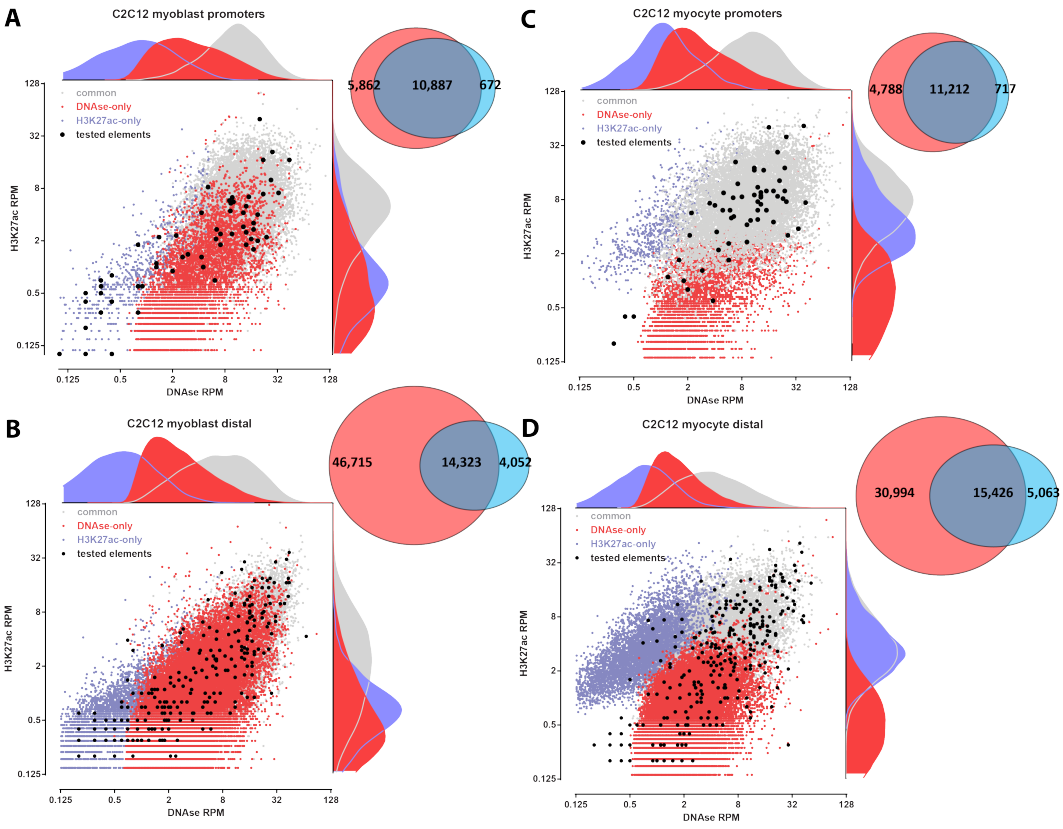

A
RPM
0 0.1 0.2 0.3 0.4 0.5
MB MyoD
MC MyoD
MC myogenin
MB MEF2
MC MEF2
MB p300
MC p300
MB DNAse
MC DNAse
MB H3K27ac
MC H3K27ac
MB E2A
MC E2A
MB HEB
MC HEB
MB Pbx1
MC Pbx1
RRCAGSTG
TATTTWTA
MB MyoD + MC MyoD + MC myogenin
MB MyoD + MC MyoD
MB MyoD + MC myogenin
MB MyoD
MC MyoD + MC myogenin
MC MyoD


B
RPM
0 0.1 0.2 0.3 0.4 0.5
MB MyoD
MC MyoD
MC myogenin
MB MEF2
MC MEF2
MB p300
MC p300
MB DNAse
MC DNAse
MB H3K27ac
MC H3K27ac
MB E2A
MC E2A
MB HEB
MC HEB
MB Pbx1
MC Pbx1
RRCAGSTG
TATTTWTA
MC myogenin only



Supplementary Figure 3.5: (Caption on next page.)

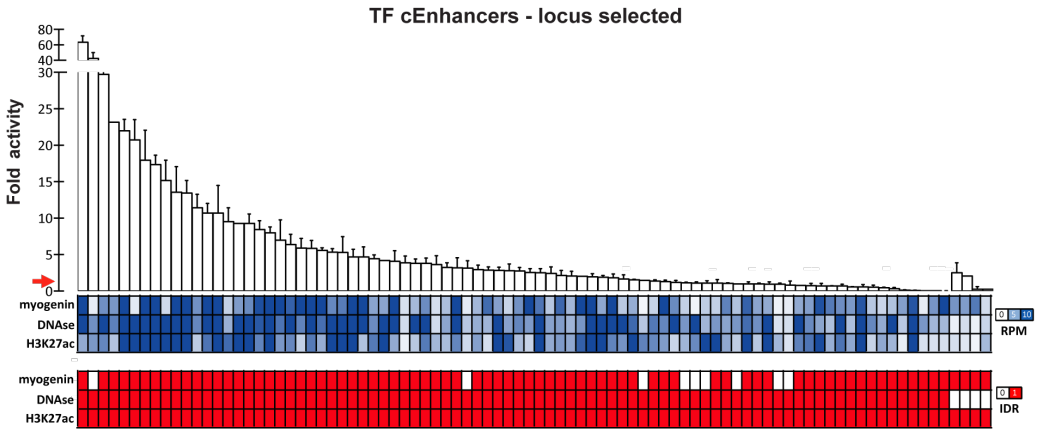
Supplementary Figure 3.5: Regulatory landscape of muscle differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, p300, the MRFs MyoD and myogenin, and cofactors (MEF2, E2A/TCF3, HEB/TCF12, and Pbx1) in undifferentiated (myoblast, or "MB") and differentiated (myocyte, or "MC") C2C12 cells were analyzed. Sites were split into multiple subgroups depending on regulatory factor occupancy (at IDR=0.05) – MyoD-positive (in either condition) sites (A), myogenin-only sites (B), and MEF2-only sites (C) – then sorted by MRF ChIP-seq signal (in the following order of priority: myoblast MyoD, myocyte MyoD, myocyte myogenin, myoblast MEF2, myocyte MEF2); the signal in the 500bp-radius region around the ChIP-seq peak position is shown.
(Full size files available here:
http://woldlab.caltech.edu/~gdesalvo/eLifeFigApp3a-C2C12-heatmaps-MyoD-V2.png
http://woldlab.caltech.edu/~gdesalvo/eLifeFigApp3b-C2C12-heatmaps-myogenin-only-V2.png
http://woldlab.caltech.edu/~gdesalvo/eLifeFigApp3c-C2C12-heatmaps-MEF2-only-V2.png)



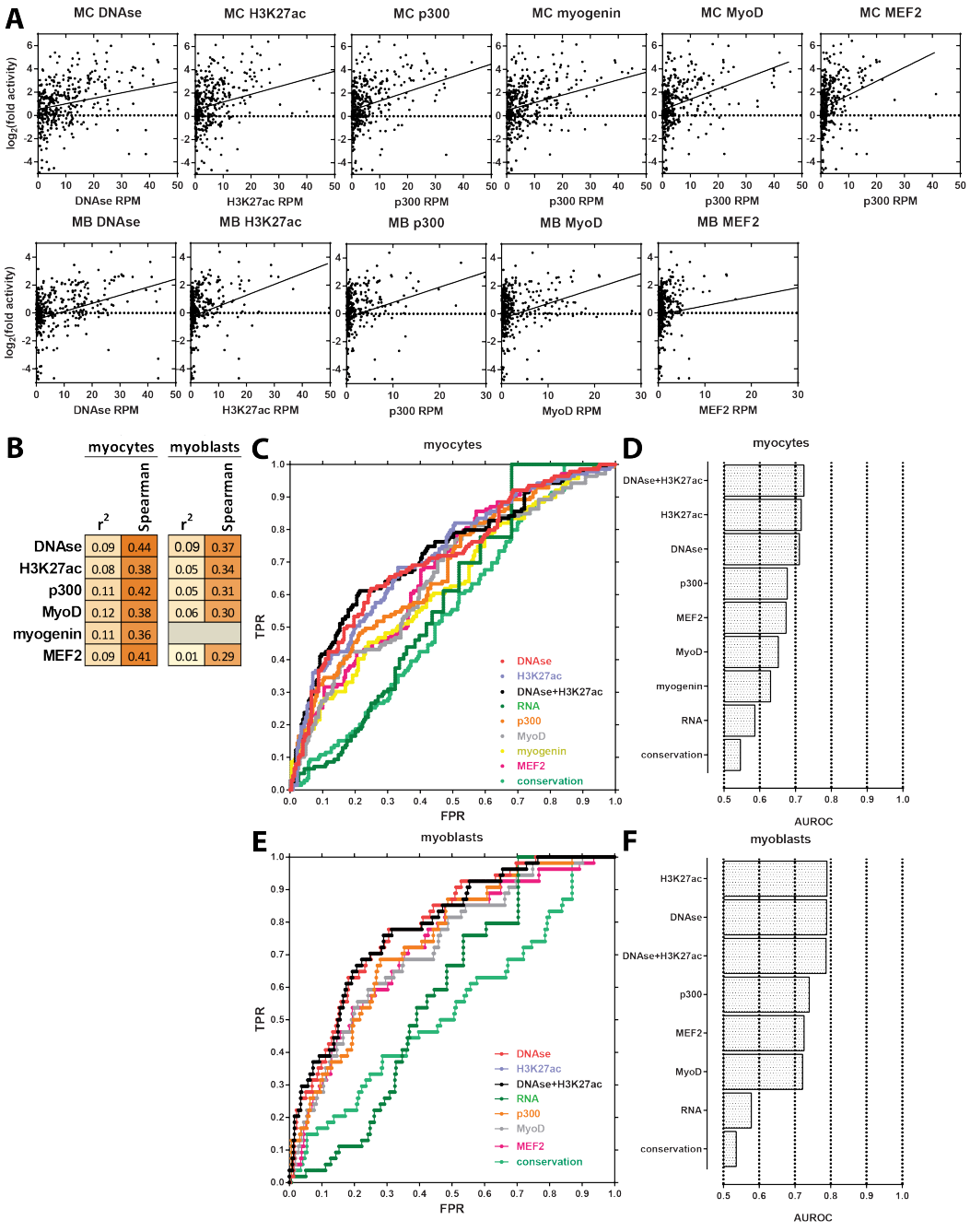
E

	promoter DNAse +			distal DNAse +		
	common	MB only	MC only	common	MB only	MC only
	14,789	1,960	1,697	36,163	24,875	13,791
MB H3K27ac +	66.98%	26.17%	12.91%	33.30%	16.92%	2.46%
MB H3K27ac -	33.02%	73.83%	87.09%	66.70%	83.08%	97.54%
MC H3K27ac +	72.91%	25.51%	35.65%	36.14%	11.25%	21.21%
MC H3K27ac -	27.09%	74.49%	64.35%	63.86%	88.75%	78.79%
	promoter H3K27ac +			distal H3K27ac +		
	common	MB only	MC only	common	MB only	MC only
	10,568	1,678	3,665	8,667	9,243	12,266
MB DNAse +	98.20%	92.25%	81.17%	97.17%	89.69%	58.24%
MB DNAse -	1.80%	7.75%	18.83%	2.83%	10.31%	41.76%
MC DNAse +	97.41%	83.37%	87.15%	90.24%	61.02%	67.46%
MC DNAse -	2.59%	16.63%	12.85%	9.76%	38.98%	32.54%

Supplementary Figure 3.6: Relationship between DNAse hypersensitivity and H3K27 acetylation during muscle differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myoblasts; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in C2C12 myocytes; (C) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in C2C12 myoblasts; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in C2C12 myocytes; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black; (E) Dynamic changes in DNAse hypersensitivity and H3K27 acetylation upon differentiation for promoter-proximal and distal sites.

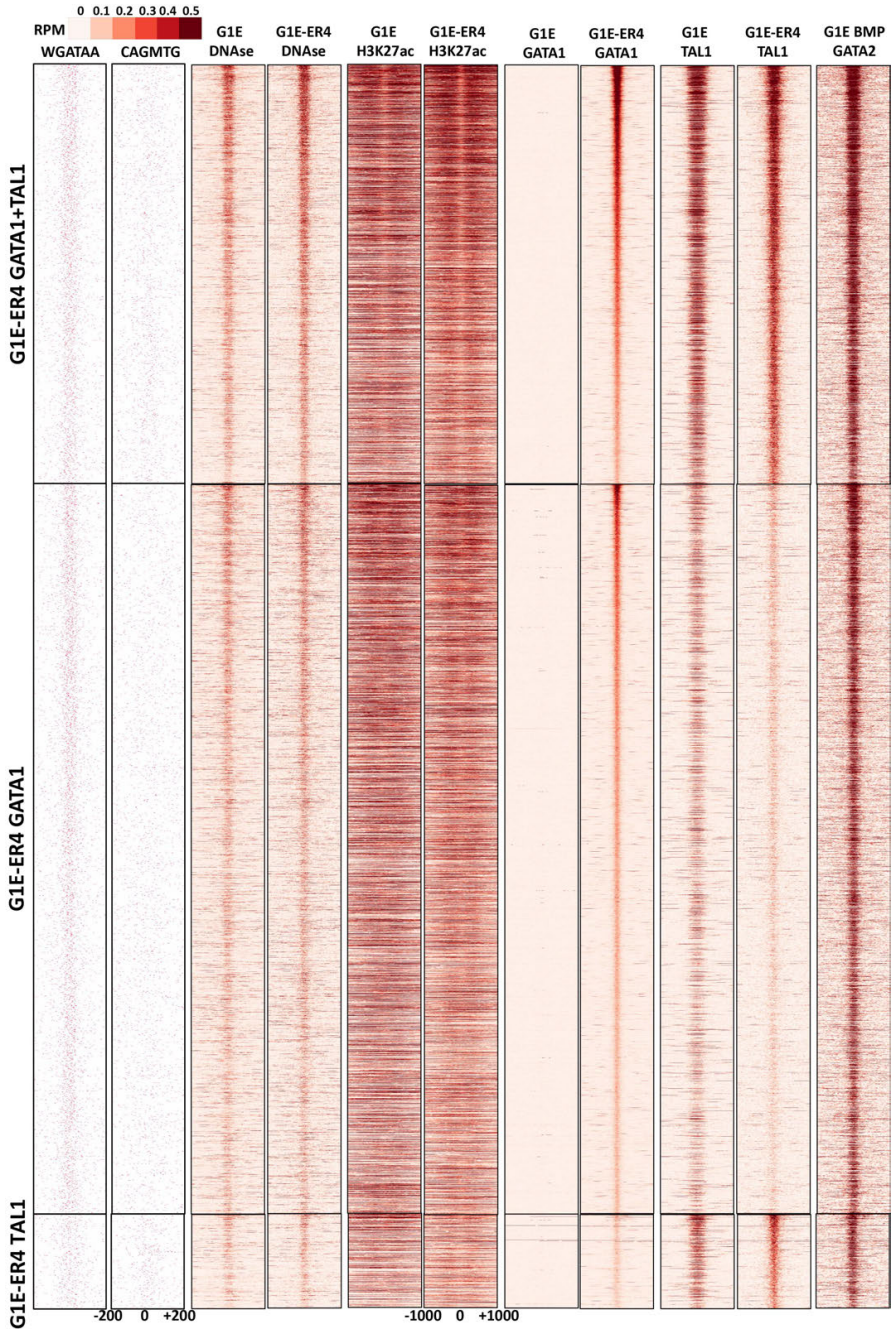


Supplementary Figure 3.7: Functional assay testing of cRE regulatory activity in C2C12 cells. Fold activity in myocytes (top) and myoblasts (bottom) across biological replicates ($n = 4$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity, H3K27ac status, p300, MyoD and myogenin occupancy are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs selected for their physical proximity to loci known for their importance to muscle development ("locus picks"); (B) randomly selected from the genome-wide set of MyoD/myogenin-occupied regions; (C) negative controls.



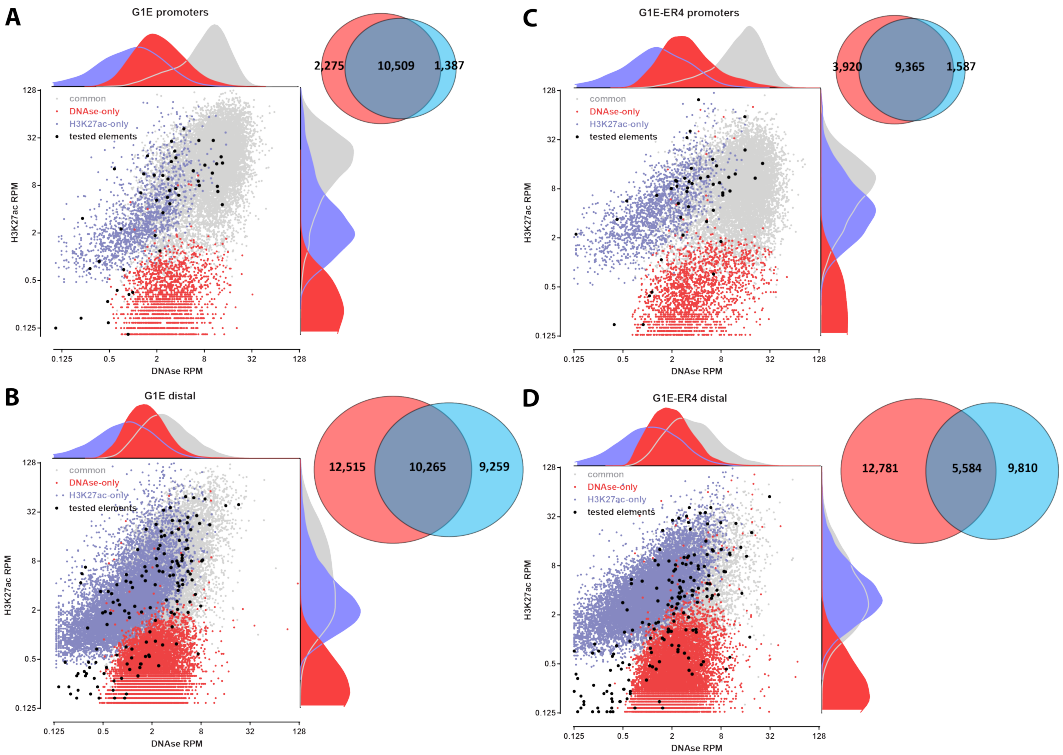
Supplementary Figure 3.8: (Caption on next page.)

Supplementary Figure 3.8: Correlation between regulatory activity and biochemical marks in C2C12 cells. Correlation between regulatory activity and biochemical marks in C2C12 cells. (A and B) Correlation between fold activity and DNAse hypersensitivity, H3K27ac, p300, myogenin, MyoD and MEF2 occupancy in myoblasts and myocytes; (C) ROC curves showing biochemical mark predictivity of cRE fold activity in myocytes; (D) AUROC (area under ROC curve) values for different biochemical marks in myocytes; (E) ROC curves showing biochemical mark predictivity of cRE fold activity in myoblasts; (F) AUROC values for different biochemical marks in myoblasts.



Supplementary Figure 3.9: (Caption on next page.)

Supplementary Figure 3.9: Regulatory landscape of erythroid differentiation. DNAse-seq and ChIP-seq experiments against H3K27ac, GATA1, TAL1 and GATA2 G1E and G1E-ER4 were analyzed. Sites were split into subgroups depending on GATA1 and TAL1 occupancy (IDR=0.05), then sorted by ChIP-seq signal (in the following order of priority: G1E-ER4 GATA1, G1E-ER4 TAL1); the signal in the 500bp-radius region around the ChIP-seq peak position is shown.
(Full size file available here: http://woldlab.caltech.edu/~gdesalvo/eLifeFigApp5-G1E-TAL1-GATA-heatmaps-V2.png)

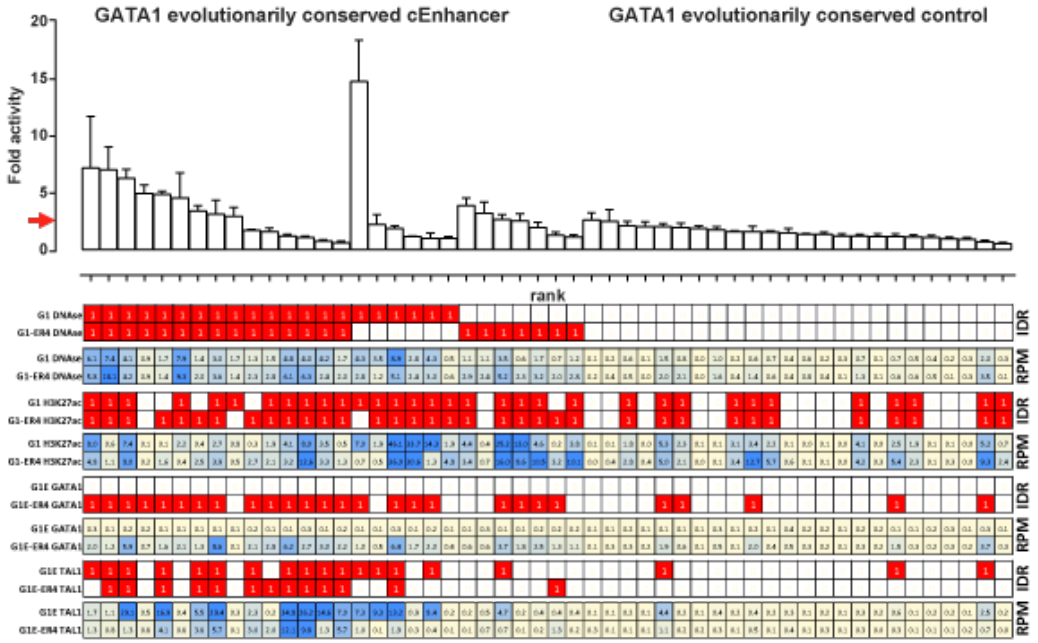


E

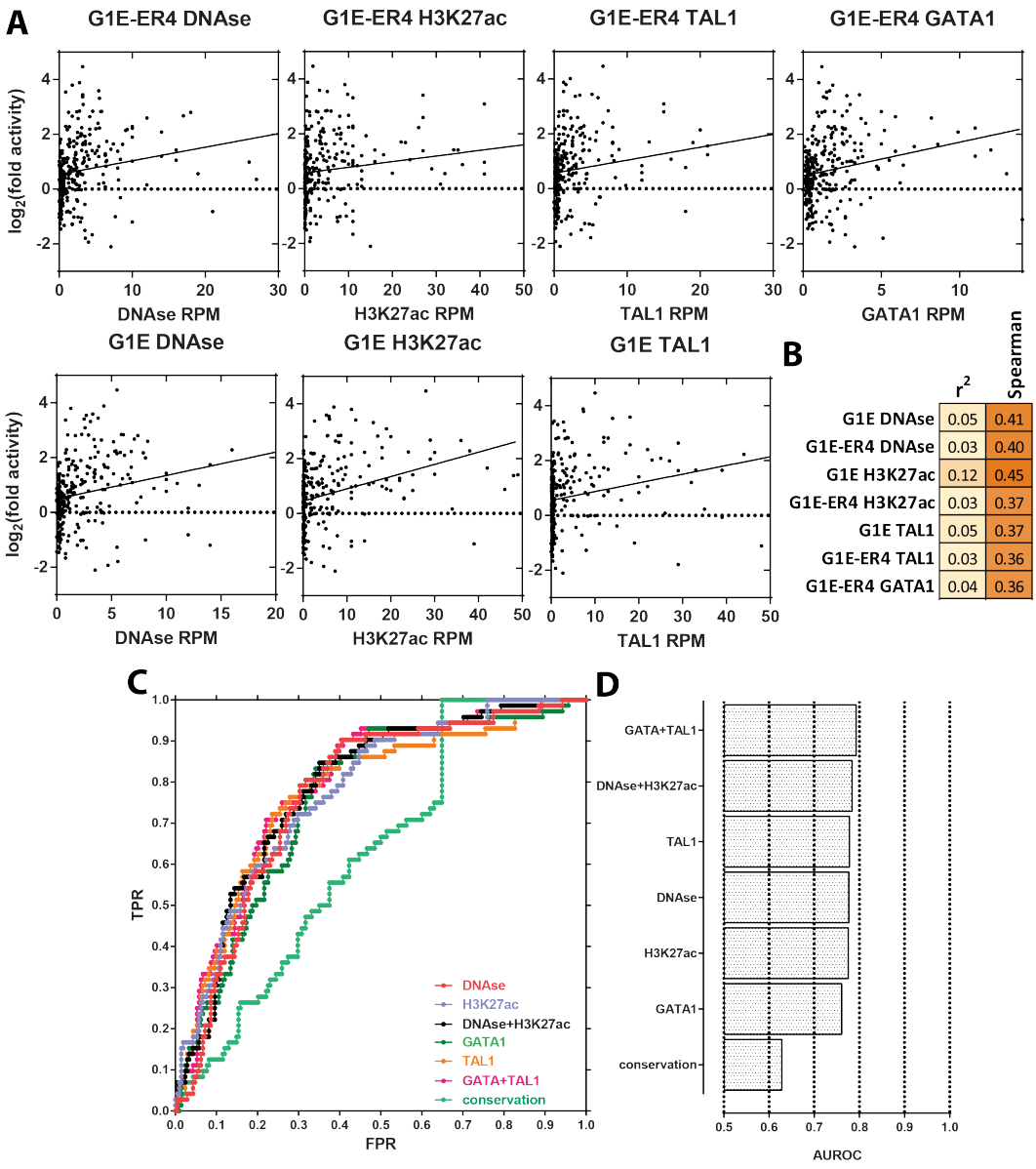
	promoter DNAse +			distal DNAse		
	common	G1E only	G1E-ER4 only	common	G1E only	G1E-ER4 only
	14,789	1,960	1,697	36,163	24,875	13,791
G1E H3K27ac +	86.43%	51.95%	13.33%	44.17%	46.10%	20.02%
G1E H3K27ac -	13.57%	48.05%	86.67%	55.83%	53.90%	79.98%
G1E-ER4 H3K27ac +	80.73%	31.84%	13.63%	33.52%	23.42%	23.65%
G1E-ER4 H3K27ac -	19.27%	68.16%	86.37%	66.48%	76.58%	76.35%

	promoter H3K27ac			distal H3K27ac		
	common	G1E only	G1E-ER4 only	common	G1E only	G1E-ER4 only
	10,568	1,678	3,665	8,667	9,243	12,266
G1E DNAse +	95.78%	63.91%	22.00%	59.51%	35.88%	9.68%
G1E DNAse -	4.22%	36.09%	78.00%	40.49%	64.12%	90.32%
G1E-ER4 DNAse +	94.91%	54.64%	26.98%	48.58%	18.93%	13.64%
G1E-ER4 DNAse -	5.09%	45.36%	73.02%	51.42%	81.07%	86.36%

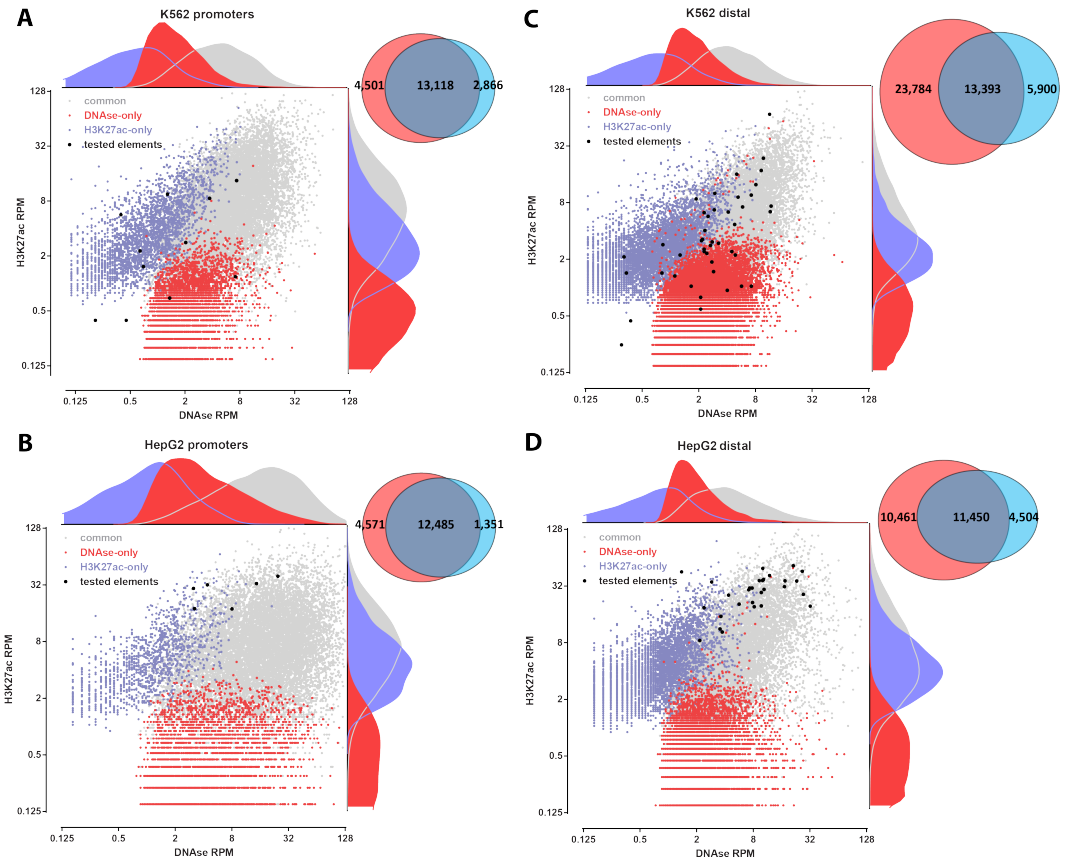
Supplementary Figure 3.10: Relationship between DNAse hypersensitivity and H3K27 acetylation during erythroid differentiation. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in G1E-ER4 cells; (C) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in G1E cells; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in G1E-ER4 cells; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black; (E) Dynamic changes in DNAse hypersensitivity and H3K27 acetylation upon differentiation for promoter-proximal and distal sites.



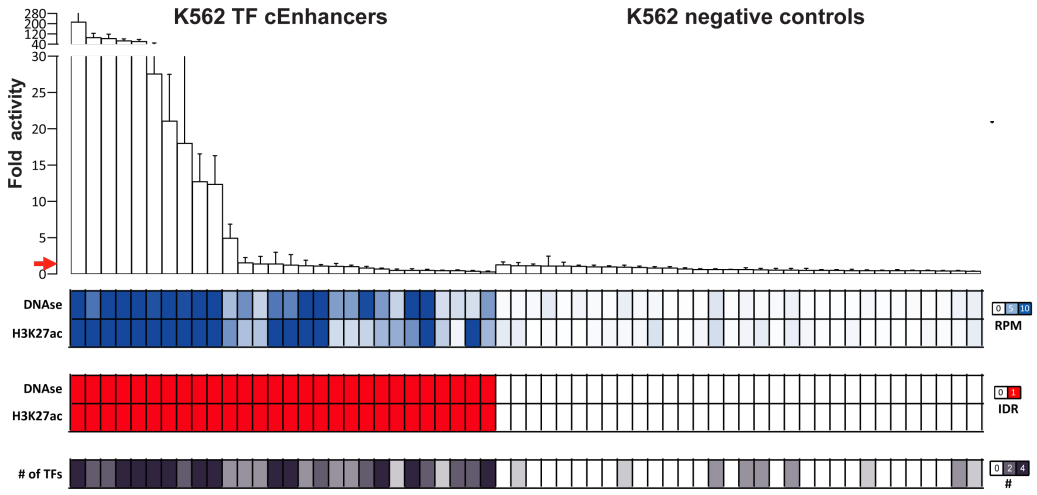
Supplementary Figure 3.11: Functional assay testing of the regulatory activity of erythroid cREs. Fold activity in K562 cells across biological replicates ($n \in [1, 9]$) and technical replicates ($n = 4$ for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity, H3K27ac status, GATA1, and TAL1 occupancy are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs randomly selected from the genome-wide set of GATA1/TAL1-occupied regions; (B) cREs selected among the set of highly evolutionarily constrained non-coding elements that contain a GATA1 motif ("regulatory potential selections").



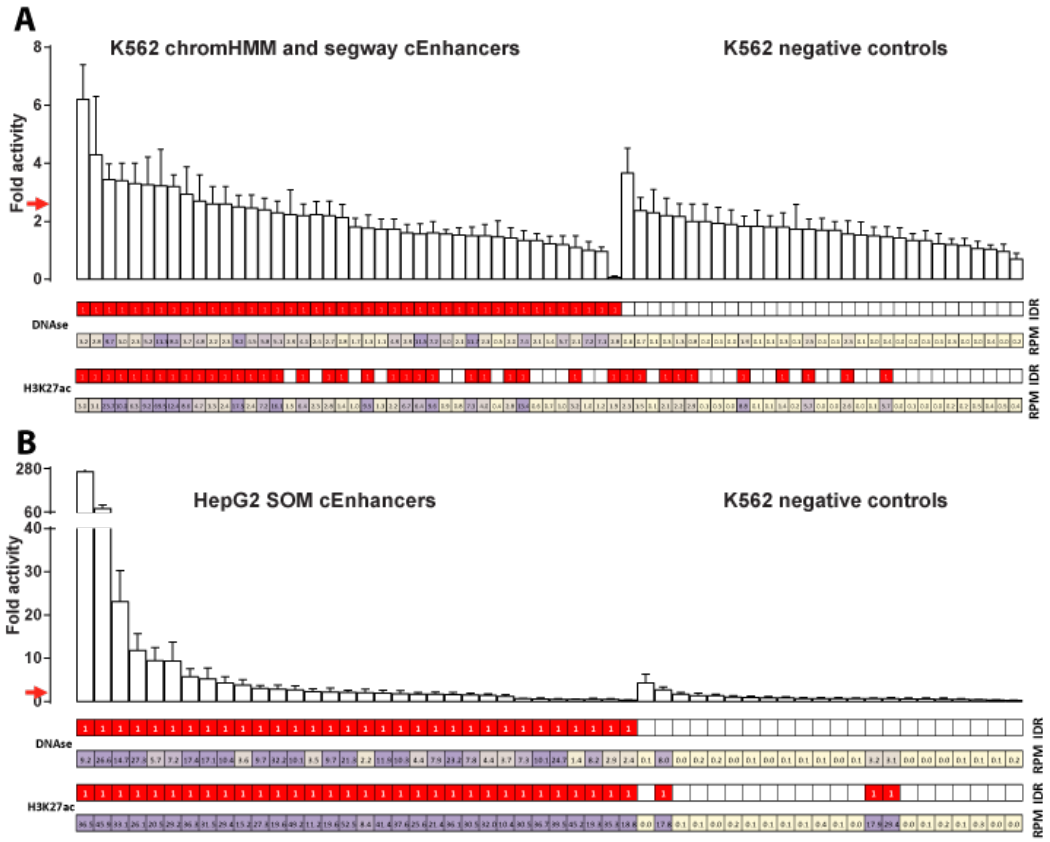
Supplementary Figure 3.12: Correlation between regulatory activity and biochemical marks in erythroid cells. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, H3K27ac, TAL1, and GATA1 occupancy in G1E and G1E-ER4 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity; (D) AUROC (area under ROC curve) values for different biochemical marks.



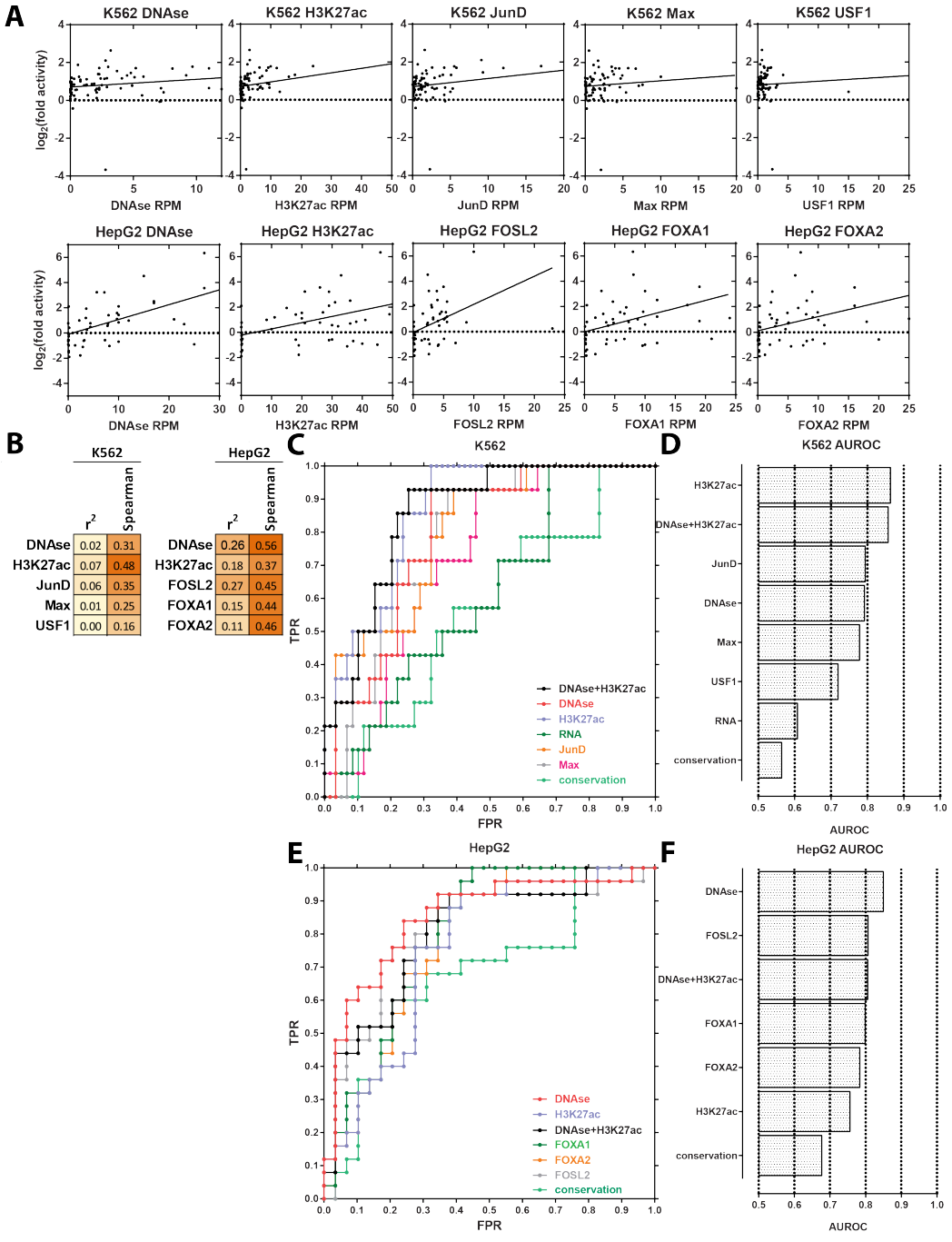
Supplementary Figure 3.13: Relationship between DNAse hypersensitivity and H3K27 acetylation in immortalized human cell lines. (A) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in K562 cells; (B) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in K562 cells; (C) Overlap between DNAse hypersensitive and H3K27ac-positive promoter-proximal regions in HepG2 cells; (D) Overlap between DNAse hypersensitive and H3K27ac-positive distal regions in HepG2 cells; the kernel density of the ChIP-seq/DNAse-seq signal distribution for each class of sites is overlaid over the scatter plots, and the distribution of tested cREs is shown in black.



Supplementary Figure 3.14: Functional assay testing of TF selected cEnhs in human immortalized cell lines. Fold activity across biological replicates (n =?? ???) and technical replicates (n =?? ???) for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity and H3K27ac status are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs tested in K562 cells (B) cREs tested in HepG2 cells.

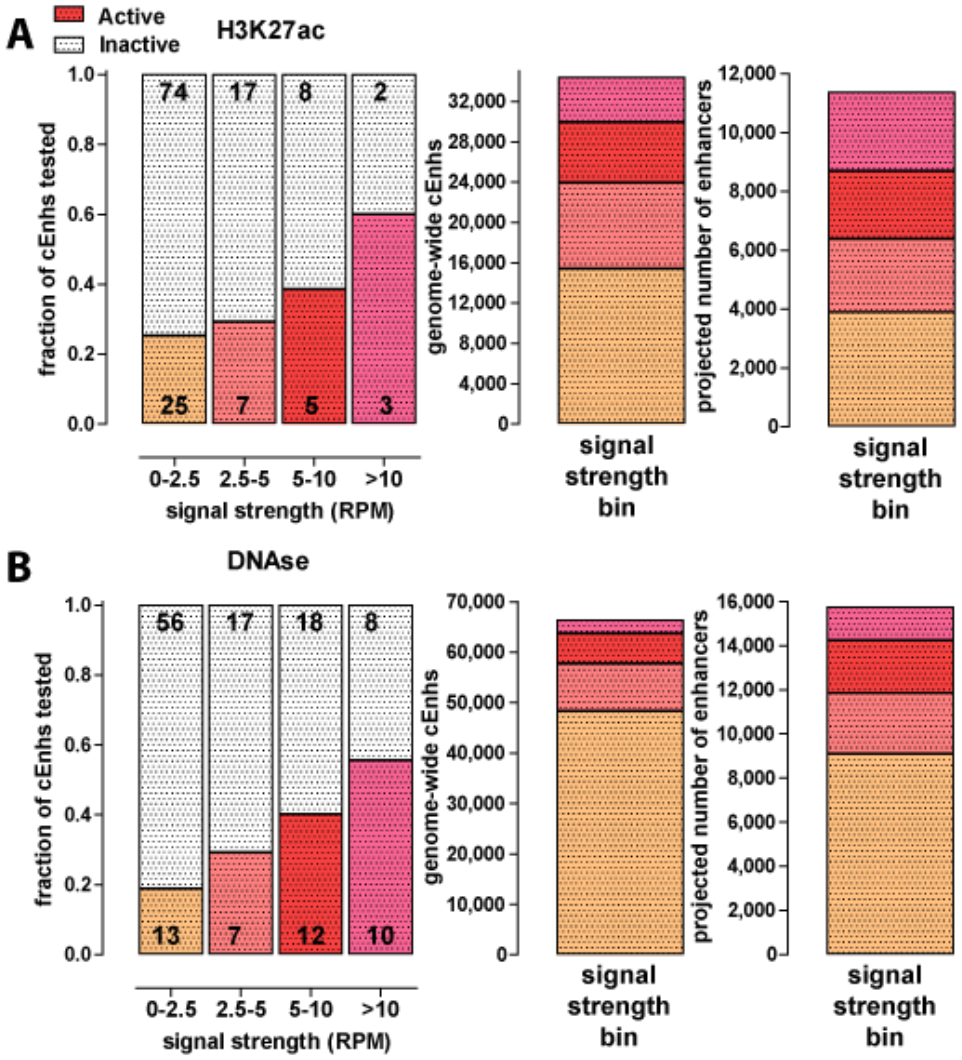


Supplementary Figure 3.15: Functional assay testing of machine learning selected cEnhs in human immortalized cell lines. Fold activity across biological replicates (n =? ???) and technical replicates (n =?? ??? for each biological replicate) is shown. Candidate REs were sorted first by their DNAse status and then by their mean fold activity. The horizontal dotted line corresponds to the mean fold activity threshold above which elements are considered active. In addition, DNAse hypersensitivity and H3K27ac status are shown for each cRE, both as binary (IDR=0.05) calls (red coloring indicates occupancy), and as RPM scores. (A) cREs tested in K562 cells (B) cREs tested in HepG2 cells.

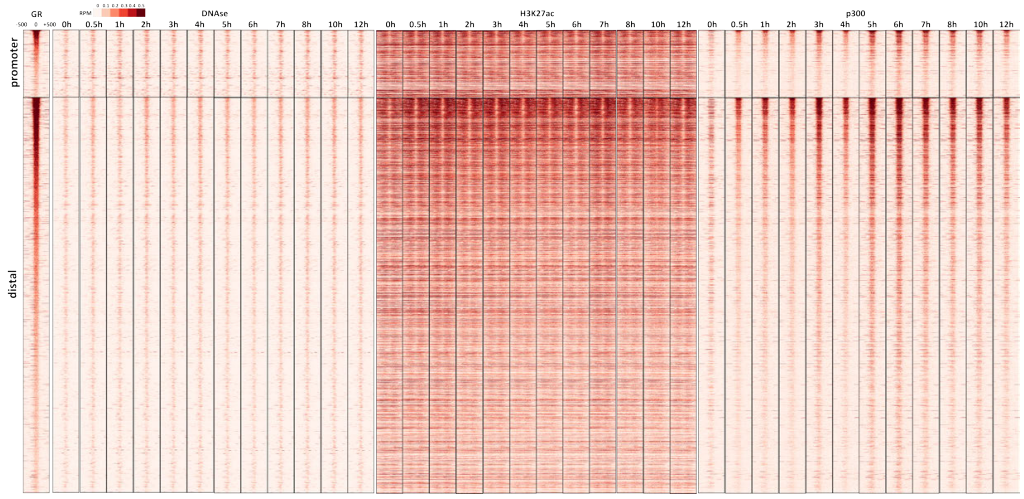


Supplementary Figure 3.16: (Caption on next page.)

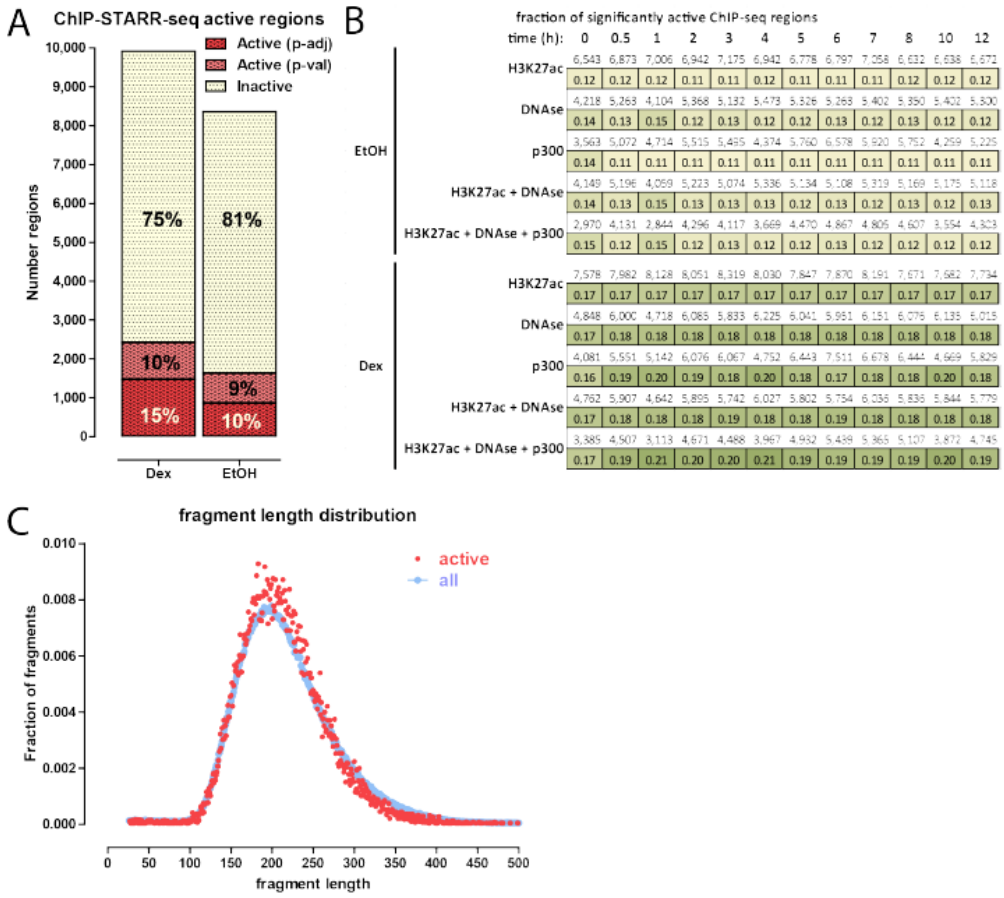
Supplementary Figure 3.16: Correlation between regulatory activity and biochemical marks in human immortalized cell lines. (A and B) Correlation between fold activity in K562 cells and DNAse hypersensitivity, and transcription factor occupancy in K562 and HepG2 cells; (C) ROC curves showing biochemical mark predictivity of cRE fold activity in K562 cells; (D) AUROC (area under ROC curve) values for different biochemical marks in K562 cells; (E) ROC curves showing biochemical mark predictivity of cRE fold activity in K562 cells; (F) AUROC (area under ROC curve) values for different biochemical marks in K562 cells.



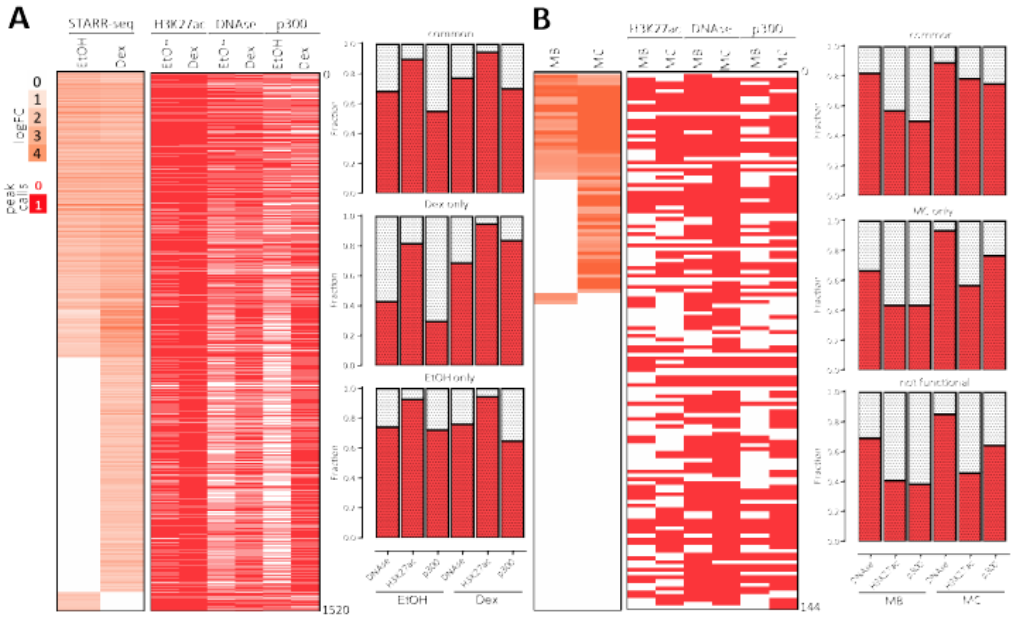
Supplementary Figure 3.17: Enrichment of active cEnhs in different classes of cEnhs defined by the strength of their biochemical signatures. (A) Fraction of active enhancers in different cEnh signal classes based on ChIP-seq data for H3K27ac in C2C12 myocytes. Genome-wide distribution and extrapolated number of active enhancers in C2C12 belonging to each signal strength class. (B) Fraction of active enhancers in different cEnh signal classes based on DNAse hypersensitivity data for H3K27ac in C2C12 myocytes. Genome-wide distribution and extrapolated number of active enhancers in C2C12 belonging to each signal strength class.



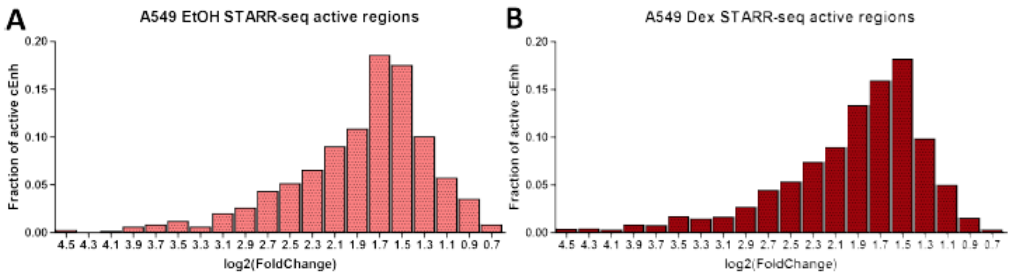
Supplementary Figure 3.18: Regulatory landscape of GR response in A549 cells. DNAse-seq and ChIP-seq experiments against H3K27ac across a 12 hour A549 dex response time course were analyzed within GR occupancy (IDR=0.05). Sites were split using GENCODE annotations into promoter proximal and distal subgroups then sorted by ChIP-seq signal for GR; the signal in the 500bp-radius region around the ChIP-seq peak position is shown. (Full size file available here: http://woldlab.caltech.edu/~gdesalvo/eLifeFigApp6-A549-heatmaps-V2.png)



Supplementary Figure 3.19: Testing of cEnhs for activity using ChIP-STARR-seq for GR in A549 cells with and without Dexamethasone stimulation. (A) Fraction of active cEnhs detected in each condition. Shown is the number of cEnhs that passed the minimum representation threshold (see the Methods section for more details) and were identified as active using DESeq2. (B) Fraction of significantly active (FDR-corrected p-value ≤ 0.05) biochemically marked individually on in combinations by H3K27ac, DNAse, p300. (C) Length distribution of active and inactive STARR-seq fragments as defined by DESeq2.



Supplementary Figure 3.20: Marking of common and cell state-specific active cEnhs by H32K7ac, DNAse and p300. (A) STARR-seq data in A549 cells with and without Dexamethasone treatment (epigenomic datasets from the 3 hour time point were used); (B) Luciferase assay data in differentiated and undifferentiated C2C12 cells.



Supplementary Figure 3.21: Distribution of STARR-seq activity in A549 cells. Shown is the distribution of log_2(FoldChange) values (defined by DESeq2) for STARR-seq experiments in resting EtOH-treated (A) and Dexamethasone-treated (B) A549 cells.

References

Arnold CD Gerlach D, et al. (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." In: *Molecular and Cellular Biology.* 339, pp. 1074–1077.

Cheng Y King DC, et al. (2008). "Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif." In: *Genome Res* 18, pp. 1896–1905.

DellOrso S Wang AH, et al. (2016). "The Histone Variant MacroH2A12 Is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1." In: *Cell Rep.* 14, pp. 1156–1168.

Dogan N Wu W, et al. (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." In: *Epigenetics Chromatin.* 8, p. 16.

ENCODE, Project Consortium (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." In: *PLoS Biol.* 9, e1001046.

– (2012). "An integrated encyclopedia of DNA elements in the human genome." In: *Nature.* 489, pp. 57–74.

Ernst J Kellis M, et al. (2012). "ChromHMM: automating chromatin-state discovery and characterization." In: *Nat Methods.* 9, pp. 215–216.

Ernst J Melnikov A, et al. (2016). "Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions." In: *Genome Res.* 34, pp. 1180–1190.

Feng J Liu T, et al. (2012). "Identifying ChIP-seq enrichment using MACS." In: *Nat Protoc.* 7, pp. 1728–1740.

Garrity PA Wold BJ, et al. (1992). "Effects of different DNA polymerases in ligation-mediated PCR enhanced genomic sequencing and in vivo footprinting." In: *PNAS.* 89, pp. 1021–1025.

Gasper WC Marinov GK, et al. (2014). "Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: identifying ChIP-quality p300 monoclonal antibodies." In: *Sci Rep.* 4, p. 5152.

Gertz J Savic D, et al. (2013). "Distinct properties of cell type-specific and shared transcription factor binding sites." In: *Mol Cell.* 52, pp. 1–2.

Hoffman MM Buske OJ, et al. (2012). "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." In: *Nat Methods.* 9, pp. 473–476.

Kellis M Hardison RC, et al (2014). "Defining functional DNA elements in the human genome." In: *Proc Natl Acad Sci U S A* 111, pp. 6131–6138.

Kharchenko PV Tolstorukov MY, et al. (2008). “Design and analysis of ChIP-seq experiments for DNA-binding proteins.” In: *Nat Biotechnol.* 26, pp. 1351–1359.

Kheradpour P Ernst J, et al. (2013). “Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.” In: *Genome Res.* 23, pp. 800–811.

Kinney JB Murugan A, et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.” In: *Proc Natl Acad Sci U S A.* 107, pp. 9158–9163.

Kwasnieski JC Fiore C, et al. (2014). “High-throughput functional testing of ENCODE segmentation predictions.” In: *Genome Res.* 24, pp. 1595–1602.

Kwasnieski JC Mogno I, et al. (2012). “Complex effects of nucleotide variants in a mammalian cis-regulatory element.” In: *Proc Natl Acad Sci U S A.* 109, pp. 19498–19503.

Landt SG Marinov GK, et al. (2012). “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” In: *Genome Res.* 22, pp. 1813–1831.

Langmead B Trapnell C, et al. (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” In: *Genome Biol.* 10, R25.

Li Q Brown J, et al. (2011). “Measuring reproducibility of high-throughput experiments.” In: *Ann Appl Stat.* 5, pp. 1752–1779.

Love MI Huber W, et al. (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” In: *Genome Biol.* 15, p. 550.

Melnikov A Murugan A, et al. (2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.” In: *Nat Biotechnol.* 30, pp. 271–277.

Moore J Purcaro MJ, et al. (2018). “ENCODE Phase III Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse.” In: *Submitted.* Pp. 1–2.

Mortazavi A Pepke S, et al. (2013). “Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps.” In: *Genome Res.* 23, pp. 2136–2148.

Mueller PR Wold BJ, et al. (1989). “In vivo footprinting of a muscle specific enhancer by ligation mediated PCR.” In: *Science.* 246, pp. 780–786.

Murtha M Tokcaer-Keskin Z, et al. (2014). “FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.” In: *Nat Methods.* 11, pp. 559–565.

Patwardhan RP Hiatt JB, et al. (2012). “Massively parallel functional dissection of mammalian enhancers in vivo.” In: *Nat Biotechnol.* 30, pp. 265–270.

Patwardhan RP Lee C, et al. (2009). “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.” In: *Nature.* 27, pp. 1173–1175.

Pepke S Wold BJ, et al. (2009). “Computation for ChIP-seq and RNA-seq studies.” In: *Nat Methods.* 11, pp. 22–32.

Reddy TE Pauli F, et al. (2009). “Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation.” In: *Genome Res.* 19, pp. 2163–2171.

So AY Chaivorapol C, et al. (2007). “Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid.” In: *PLoS Genet.* 3, pp. 94–95.

Taylor J Tyekucheva S, et al. (2006). “ESPERR learning strong and weak signals in genomic sequence alignments to identify functional elements.” In: *Genome Res* 16, pp. 1596–1604.

Tyner C Barber GP, et al. (2017). “The UCSC Genome Browser database: 2017 update.” In: *Nucleic Acids Res* 45, pp. D626–D634.

Visel A Blow MJ, et al. (2009). “ChIP-seq accurately predicts tissue-specific activity of enhancers.” In: *Nature.* 457, pp. 854–858.

Visel A Minovitsky S, et al. (2007). “VISTA Enhancer Browser-a database of tissue-specific human enhancers.” In: *Nucleic Acids Res* 35, pp. D88–92.

Vockley CM DIppolito AM, et al. (2016). “Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome.” In: *Cell.* 166, pp. 1269–1281.

Wang H Zhang Y, et al. (2006). “Experimental validation of predicted mammalian erythroid.” In: *Genome Res* 16, pp. 1480–1492.

Yue F Cheng Y, et al. (2014). “A comparative encyclopedia of DNA elements in the mouse genome.” In: *Nature.* 515, pp. 355–364.

Zabidi MA Arnold CD, et al. (2015). “Enhancer core promoter specificity separates developmental and housekeeping gene regulation.” In: *Nature.* 518, pp. 556–559.

Chapter 4

DISSECTING THE CONTRIBUTION OF INDIVIDUAL ELEMENTS IN COMPLEX LOCI

4.1 Testing the of entire set of candidate Enhancers affiliated with individual genes

I set out to test the functional contribution of individual candidate regulatory element found within loci of a set of genes that are either important to or have been historically studied in the context of myogenesis. Even though these loci are biochemically complex and score as connected to a large number of candidate Enhancers compared to the run of the mill gene, the mean functional contribution (and overall distribution) of these individually tested elements (Figure 2.10) are not found statistically significantly different at the p=0.05 level (t-test p=0.16; ks test p=0.15) from randomly selected myogenic occupancy in the genome. These landmark loci, dissected by years of studies thus provide a reliable platform on which to extrapolate key mechanistic components on the larger genome.

To this end I tested all of the myogenin occupied and myocyte DNAse hypersensitive regions, using connectivity as a guide from the MYBPH, myogenin, and MyoD1 loci in order to sample all detectable candidate Enhancer elements. I also sampled the myogenin occupancy associated within the connectivity boundaries for ID2, BTG2, TMSB4X, Desmin and the Troponin loci; all genes either important or selected because they closely follow the RNA modulation of these well studied genes across early myogenesis.

4.2 Proportion of enhancer activity predicted by myogenin occupancy is identical irrespective of affiliated gene transcriptional trajectory

Assigning enhancers to a promoter is a difficult task for an algorithm to reliably perform due to the partial redundancy of some signals which are necessary and/or beneficial to a biological system. Although when working with a limited number of genes such as the ones picked above it is possible for a human, based on observation, to assign nearby candidate enhancers with a possible target gene either by the nearest expressed gene (cEnhs within 10kb), or via Chia-pet connectivity (cEnhs >10kb from TSSes) within CTCF boundaries. (Oti M, 2016) Because we work in a differentiation model, which offers a transition between two states, we can sort genes based on their transcriptional trajectory across this stage change. (Figure 1.1) For example, genes that are upregulated more than 3 fold between myoblast and myocytes, genes that "flat" or in between the two other classes, and genes that are downregulated more than 3 fold allowing to contrast the enhancers near genes that have large differences in transcriptional regulation across differentiation of C2C12s. (Kirilusha A, 2014 thesis,Mortazavi A, 2008)

Because genes in the loci we tested tend to fall within stretches of chromosome segmented by repeating CTCF peaks, and within these regions either behave in a manner that is similar following the RNA trajectory or remain silent I was able to assign a gene trajectory to two of our sets of candidate Enhancers. Although the proportion of elements active is significantly different at the p=0.01 level between upregulated and flat genes affiliated enhancers (Figure 4.1- Fisher's Exact test p=0.0034) the statistically insignificant difference at the p=0.05 level in means (t-test p=0.11) can also be attributed both the low sample number and to the large numbers of inactive elements from the MyoD locus (Figure 4.4). In order to obtain a better sampling for this comparison one would want to ideally select a population of candidate Enhancers from a larger number of flat expressed genes. Although

a very small sample (n=11), the activity of enhancers affiliated to downregulated genes observed suggests an important biological role for repression and silencers in order to negatively modulate the transcript level of these genes. (Yokoyama S, 2009)

4.3 Myogenin POLII connected occupied candidate enhancers provide identical proportion and distributions of activity as non myogenic DHS elements similarly selected by POLII connectivity

The myogenin, MyoD and Desmin candidate loci were sampled for myogenin occupancy, and DHS within the boundaries set by the furthest POLII physical connectivity. (Fisher K, 2016 thesis,Li Q, 2011) From these candidates one can compare the functional contribution of myogenin occupied (n=77) vs physically connected non-myogenic (DHS) elements (n=51). (With the caveat that some of these sites might be ChIP-seq non-accessible)

I found that 51% of elements are active in Myogenin based selections contrasted to 47% being active in the DHS candidate Enhancers. There is no statistically significant difference at the p=0.05 level in proportion (Fisher's exact test p=0.72) of active enhancers and a t-test (p=0.13) of means or KS test (p=0.40) of distribution of signal also fail to detect a statistically significant at the p=0.05 level between these two distributions of activity. (Figure 4.2) These regions were tested against an exogenous TK promoter and it is possible that this observation might be different if the candidate Enhancers were paired with a promoter that normally responds (and contains the motifs relevant) in the context of myocytes. (McKnight SL, 1982)

4.4 Individual candidate Enhancer elements and their putative relative locus contributions

The BTG2, MYBPH, Myogenin and TMEM183A loci are located physically proximal to each other in a 300kb span centered near bp 136M of chromosome 1 in the

mouse MM10 assembly. (Figure 4.3) Each sub-locus boundary I assigned is marked by red bars connecting the name box of the locus to the biochemical signal trace of the CTCF ChIP-seq measurement in C2C12 myocytes. The name of the gene represents the trajectory of the RNA expression between myoblasts and myocytes. Genes names are marked as downregulated (blue), flat (purple) and upregulated (red). The top 6 tracks represent specified myoblast (S) while the bottom 7 represent differentiated myocytes (D). The additional track in myocytes is myogenin itself, which is not transcribed in myoblasts, and was therefore not ChIPed in myoblasts.

Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). The measured enhancer activity for each element tested, including dissections of larger elements, are included below. A simplified version of this locus — lacking the C2C12 biochemical measurements — was used to compare our tested elements with the ENCODE encyclopedia cRE annotations from agnostic, E14.5 liver and E14.5 limb. (Figure 3.6)

Although there are other genes in the locus, they behave in a manner that is either consistent, albeit at much reduced level, with the transcriptional behavior of target gene, or remain silent entirely. This implies a strong promoter selectivity to target these enhancers correctly. One notable exception being Chit1, which provides mature macrophages with the ability to hydrolyze chitin in response to fungal diseases. (Kanneganti M, 2012) Expression of this gene is associated with Gaucher's disease and atherosclerosis, and is likely under tight regulation at the promoter. (Grace ME, 2007) I Include descriptions of each sublocus individually for the three most prominent genes in the subsections below.

BTG2 locus

BTG2 is a gene known to function as a co-activator in transcriptional regulation, and is broadly expressed across many tissues. (Passeri D, 2006) It is expressed moderately in both myoblasts and myocytes. The measured enhancer activity (70% active - n=10) within this locus domain includes primarily enhancers that function in both cell states. This locus based scoring includes any partially overlapping elements tested in combinatorics and may appear higher than the non overlapping selections. The exceptions are the two S-MyoD occupied enhancers (Figure 4.3 - BTG Locus cEnhs 3 and 4) which only score as enhancers upon activation of the myogenin expression. Interestingly, this genes has been linked to ERalpha-mediated activation, which makes the Tal1/Gata1 ChIP-sites detected near it difficult to interpret in G1E-ER cells. (Prevot D, 2001)

MYBPH locus

MYBPH, binds to filaments of striated muscle cells, it's function is currently unclear. (Gruen M, 1999) It might helps provide function by locking muscle fibers in place or providing strength, as it is found strongly downregulated in patients with prolapsed pelvic organs. (Hundley AF, 2006) This muscle associated gene is driven by surprisingly strong myocyte specific enhancers (100% active - n=5) within its CTCF domain region. Three of these enhancers are stronger than the ones found between the myogenin locus boundaries, although ChiA-PET data indicates that they are brought into physical contact with the promoter region of myogenin and may function to control both genes. Both of these genes have surprisingly weak promoters (Figure 4.9) and likely rely on the combined effect of these enhancers to achieve their stratospheric transcriptional levels activated during the differentiation between myoblasts and myocytes.

myogenin locus

The myogenin locus includes 6 enhancers that function as individual elements (46% active - n=13), all exclusively in myocytes. Based on their relatively modest enhancer function measured, it is likely that these enhancers in the myogenin locus work best when brought together, or require proximal elements found within their native promoter in order to truly shine. Based on connectivity data, it is likely that myogenin recruits one or more enhancers from the MYBPH locus. The candidate Enhancers affiliated with TMEM183A were only partially sampled but are reported as tested. (20%active - n=5)

4.5 MyoD1 locus

In the MyoD1 locus again promoter contribution is only modest (Figure 4.9) when compared to other promoters tested. All biochemical data for this and all following loci is organized as described in section 4.4 and in the individual figure legends. The activity detected in this locus indicates that a similar mode of enhancer regulation to myogenin that requires combinatorics is likely to be in play (Figure 4.4) especially since the majority of the enhancers tested do not function (16% active - n=19).

4.6 Desmin/Speg locus

The Desmin locus was selected because this gene is similarly transcriptionally regulated to MyoD but not important to the development of muscle. Both the promoter of Desmin, and enhancers 1 and 5 within the Desmin locus behave as one might expect based on the strong transcriptional output of this gene. (29% active - n=7) (Figure 4.5)

The Speg locus enhancers also behave as one might expect based on the weak transcriptional output of this gene including 3 modest enhancers in the region. (75% active - n=4)

4.7 ID2 locus enhancers suggest that combinatorics are key.

Surprisingly several of the moderately biochemically marked elements in this locus, including strong EP300 marked sites, do not function on their own. Similar candidate Enhancers are observed in the myogenin, which when tested combination to their proximal neighbor were able to exert enhancer function. (Figure 3.7)

The findings in this locus show that both biochemically marked regions score as active enhancers when locally combined underlying the necessity of a synergistic model of control for these elements, where multiple factors need to be brought together in order to activate transcription to a high level. (Figure 4.6 - where only the merged enhancers #2 and #5 function on the assay)

These findings, together with the limited observation in Starr-Seq that larger regions tend to score on average as more active, implies that even within our tests we may be too limited to score the context of hundreds of relatively closely spaced regions on a genomic scale. (Vockley CM, 2016) These findings call for testing enhancer pairs; and due to vast differences in relative promoter activity observed, one should also test these enhancers paired with their native promoters. (further discussed in Chapter 5)

4.8 Troponin neighborhood

I designed an experiment in the Troponin loci where I sampled every large enhancer-like signature within approximately a megabase of the locus. (Figure 4.7) I made this collection of high signal myogenin or DHS regions that looked like gold standard enhancers because of their high H3K27ac signal and included a small set of test enhancers that fell within first or second intron enhancers similar to the ones found in CKM (#3,10,12,15,16). Interestingly, 70% (n=17) of these high signal candidate elements worked as expected (Visel A, 2007), except ones found in the first and second introns (#10,12,15,16) or within intronic regions of genes (#3).

An initial key experiment to dissect this finding be to test these cases of first or second intron enhancers that did not function on their own, positioned 3' of a promoter, as the CKM intronic enhancer was found to function irrespective of their orientation. (Tai PWL, 2011)

4.9 TMSB4X locus

The TMSB4X locus was sampled for enhancers at connected regions at the request and to support the work of a fellow graduate student in the lab. (Figure 4.8) Say-Tar Goh aims to show the physical interaction of some of these elements using fluorescent probes for the promoter and distal enhancer (FISH). Although this region spans a gene sparse megabase of the genome, in clear contrast to the gene dense Troponin neighborhood, this gene neighborhood is similarly marked by clear CTCF boundaries as expected and contains similar proportions of active enhancers (67% n=9). The elements found to interact with the promoter by ChIA-pet are for the most part moderately active on their own, in contrast to a gene that is similarly expressed to myogenin. It is possible that these elements will provide much more powerful enhancement if brought together physically, however this gene indicates that these enhancers can act over distances of hundreds of kb, and that our expected locus structure is largely reflected even in this immense locus. (Ghirlando R, 2016; Oti M, 2016)

4.10 Promoter contribution for MRF genes is modest compared to their antagonists (IDs) and similar trajectory genes

After this saturation test of individual candidate enhancers within key loci we tested their target promoter proximal regions in order to compare their contributions. These genes are found to have widely varying degrees of relative promoter contribution. (Figure 4.9) This small set of promoters, while not enough to lay a statistical claim, does provide insights that could be further studied. For example promoters of differ-

entiation specific genes appear to be regulated by the combination of distal elements paired with a relatively weak promoter proximal region. Flat expressed genes have promoters proximal regions that confer comparable transcriptional control in both C2C12s and 10T1/2 fibroblasts. The one striking exception is Desmin which has likely a myogenic responsive element in the promoter region. (Figure 4.9) This proximal enhancer, which likely contributes the up-tick in transcriptional output by the Desmin promoter proximal region detected in the myocytes, shows weak MyoD occupancy in the myoblast, whereas strong occupancy of both MyoD and myogenin in the differentiated myocytes.

4.11 Myogenin promoter function

We tested a dissection of the myogenin promoter, which had been studied in several high profile papers over the last decade. The 1092bp upstream of myogenin promoter in C2C12s roughly matches the function of the elements tested in transgenics showing a specific activation in the myocyte cells. Their GZ133 and GZ188 and GZ1092 constructs match up well in function with our reporter assay (Figure 4.9; MYOG132, MYOG187, MYOG1097) which show modestly increasing amounts of luciferase reporter. (Yee SP, 1993) Removing just a few basepairs (Figure 4.9; MYOG125) of the PBX site occupancy site to 125BP, prevents meis binding and coupling with MyoD to promote myocyte function of the myogenin promoter also behaved as expected killing all activity. (Berkes CA, 2005) Their study also included a second mutant which similarly killed the enhancer function when an Ebox is deleted which makes up the right side of a combined MSY3 motif later identified to act as a competitive repressor for myogenic transcription in late stage differentiation. (Berghella L, 2008) This further supported as the 1092-133 fragment which when tested against an HSP68 promoter loses its muscle specificity instead showing strong ectopic expression. (Yee SP, 1993)

4.12 Effect of a repressive Ebox on the Atoh8 promoter activity

We set out to contrast two different but overlapping Atoh8 promoter proximal regions. The smaller region, which excludes an E-box theorized by a fellow lab member to be a repressive E-box, whereas the second larger construct includes this potentially repressive E-box. (Kirilusha A, 2014 thesis) At least in this particular case it appears that the E-box functions as theorized, resulting in a dampening of the transfection output across all conditions tested. (Figure 4.9)

4.13 Downregulated genes

The downregulated gene affiliated promoters register expression in the myocyte stronger than in the myoblast. This is surprising but may be the result of the cells being transfected and then grown to confluence prior to differentiation stimulus, which possibly gave these constructs more time to accumulate signal. This signal may be great enough in the case of promoters that it is not atoned even by the faster turnover (h-pest) luciferase reporter that we chose for our assay. It is also possible that these constructs rely on distal elements not present on the construct to bring in repressor elements, which would be partly supported by the DNAse patterns remaining unchanged at these promoters of downregulated genes.

4.14 Different genes, different regulatory structures?

These findings support a model where regulation of genes is gated by binding very proximal to the promoter but where distal enhancers, which when acting against the correct promoter can impart a much higher level of expression. (Longabaugh WJ, 2005) This makes physical connectivity maps important in order to be able to in the future test enhancer pairings which are likely to be functional. (Further discussed in chapter 5)

It is also important to note that we selected an heterologous promoter (HSV-TK) which does not contain any myogenic (or known cofactor) occupancy motifs, unlike

several promoters that are myocyte specific. Considering different promoters and pairing sets of enhancers may be important, especially as models are emerging for promoter elements, where a promoter can be a tight compact region, or a more lax promoter where elements are sprawled over a larger region (much like enhancers themselves). (Mikhaylichenko O, 2018) This creates a situation which may be difficult or even impossible to model in bulk, and might be best resolved on a case by case basis as needed for further studies into individual gene functions; although a few potential experiments are proposed in Chapter 5.

4.15 Conclusions

1) Elements affiliated with loci important to myogenesis appear equally predictive as their randomly selected counterparts (Figure 2.10). Similarly, myogenin and DHS connected (no myogenin) elements elements appear to not have statistically significant differences at the p=0.05 level for both the test of proportions (Fisher's exact test p=0.72) and means (t-test p=0.13) of enhancer activity.

2) The proportion of active enhancers is significantly enriched for regions affiliated to upregulated genes (Fisher's Exact test p=0.0034) compared to flat ones, however the means of enhancer activity predicted by myogenin occupancy is statistically insignificant at the p=0.05 level irrespective of affiliated gene transcriptional trajectory. (t-test p=0.11 between upregulated and flat gene affiliated cEnhancers; p=0.36 between upregulated and and downregulated; and p=1 between flat and downregulated gene affiliated). This difference observed is likely due to the large numbers of individually inactive elements found within the MyoD locus, which may not be representative of other flat expressed genes.

3) The function observed when combining our small set of individually inactive regions in these loci suggests (albeit in numbers too low to test statistically) that enhancers likely work in combinations of elements either through short range pairs,

long range pairings or even through recruitment of individually bound (but both necessary) and separate factors such as shown for AP1 and GR upon stimulation by dexamethasone. (Vockley CM, 2016)

4) Although a numerically small set, some of our promoter elements are extremely powerful, whereas the promoters of developmental genes tested in this chapter appear to be relatively weak, possibly relying on the contribution of combinations of distal elements. Further testing could eventually resolve this question, although evidence in drosophila seems to indicate key differences for developmental enhancers which they report as able to act as "weak" promoters. (Mikhaylichenko O, 2018) The observed weak promoter function paired with the significant promoter to promoter connectivity observed in our muscle data raises interesting questions as to the motifs required to support some of these functions, including both stimulation and localization.

5) The overall lack of myoblast exclusive enhancer elements being detected from the occupancy is surprising. (further discussed in Chapter 5) Although exclusively myoblast active elements are not detected, even near downregulated genes, one major pitfall is that our selection did not sample for strong myoblast DHS regions but instead relied on Chia-PET connectivity and myocyte DHS measurements, which are relatively weaker in myoblasts.

4.16 Figures (Chapter 4)

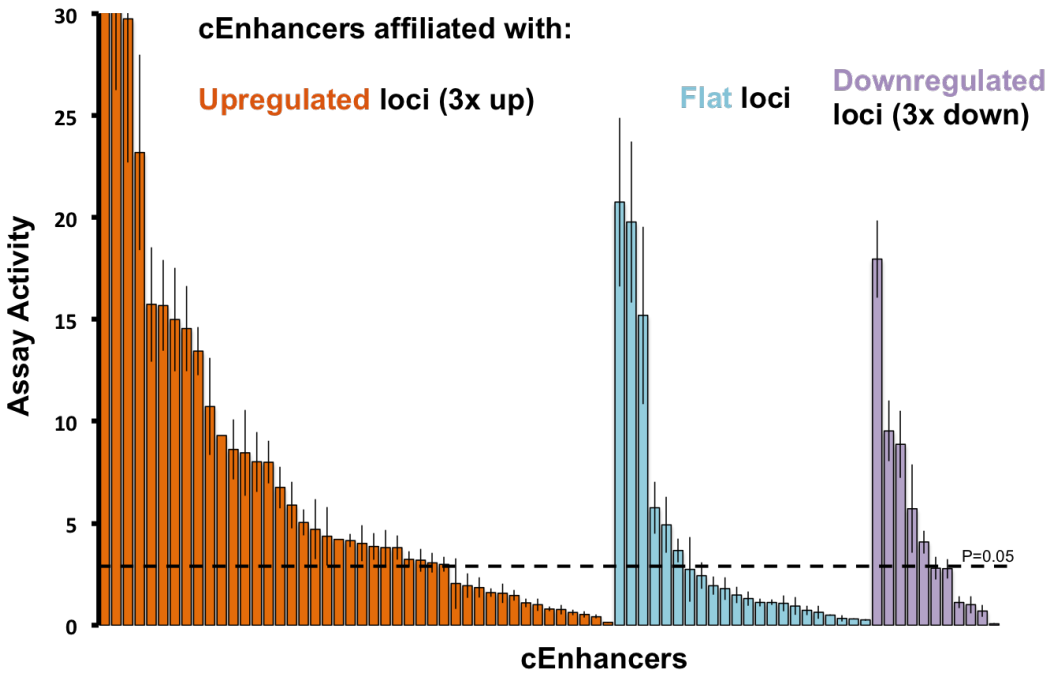


Figure 4.1: Comparing the enhancers of upregulated (n=44), flat (n=22) and downregulated (n=11) genes across differentiation of C2C12 myoblasts to myocytes. Although the proportion of elements active is significantly different between upregulated and flat loci (Fisher's Exact test p=0.0034); the means of signals is not statistically significant at the p=0.05 level (t-test p=0.11 between upregulated and flat gene affiliated cEnhancers; p=0.36 between upregulated and and downregulated; and p=1 between flat and downregulated gene affiliated.)

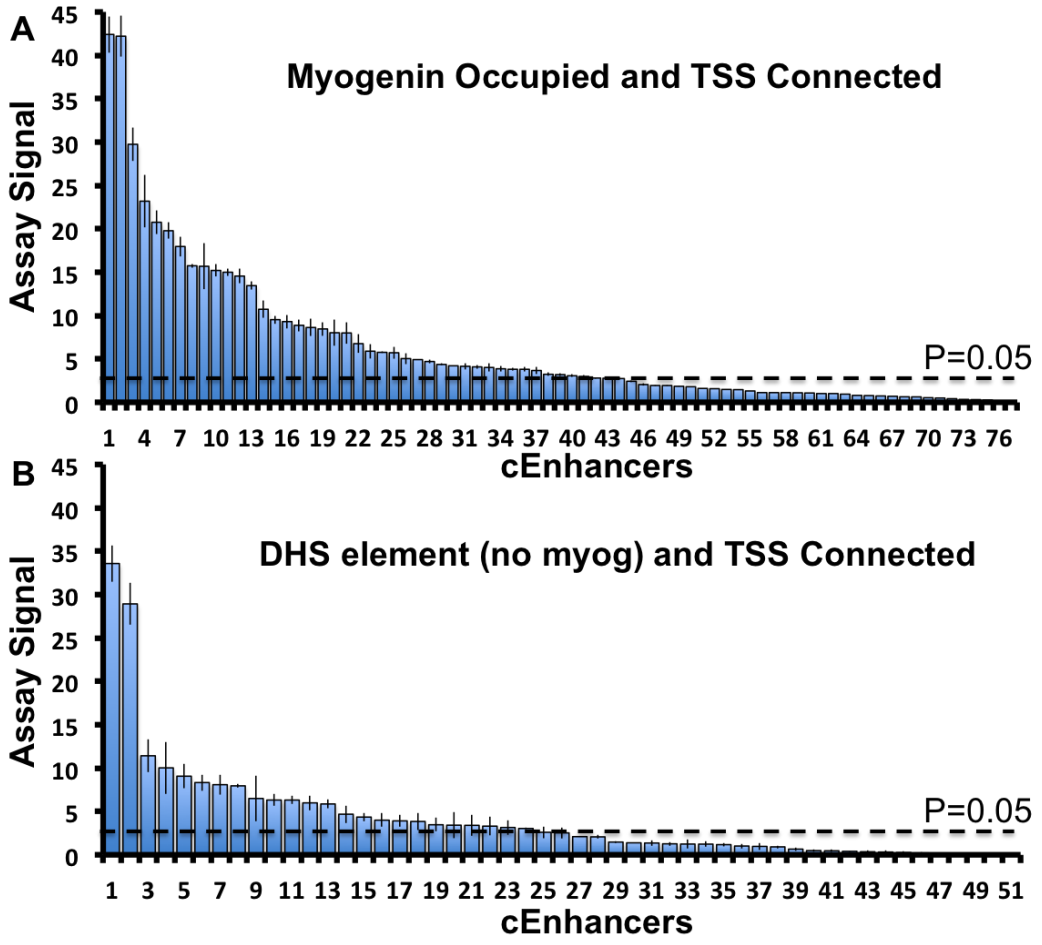


Figure 4.2: Functional comparison of myogenin occupied candidate Enhancers (n=77) and DHS elements (non myogenic candidate Enhancers - n=51) that are connected to a TSS. Fisher's exact test p=0.72 is statistically insignificant at the p=0.05 level for both proportion of active enhancers while the t-test (p=0.13) and KS test (p=0.40) are statistically insignificant at the p=0.05 level for the difference in mean and distribution of signal for both conditions.

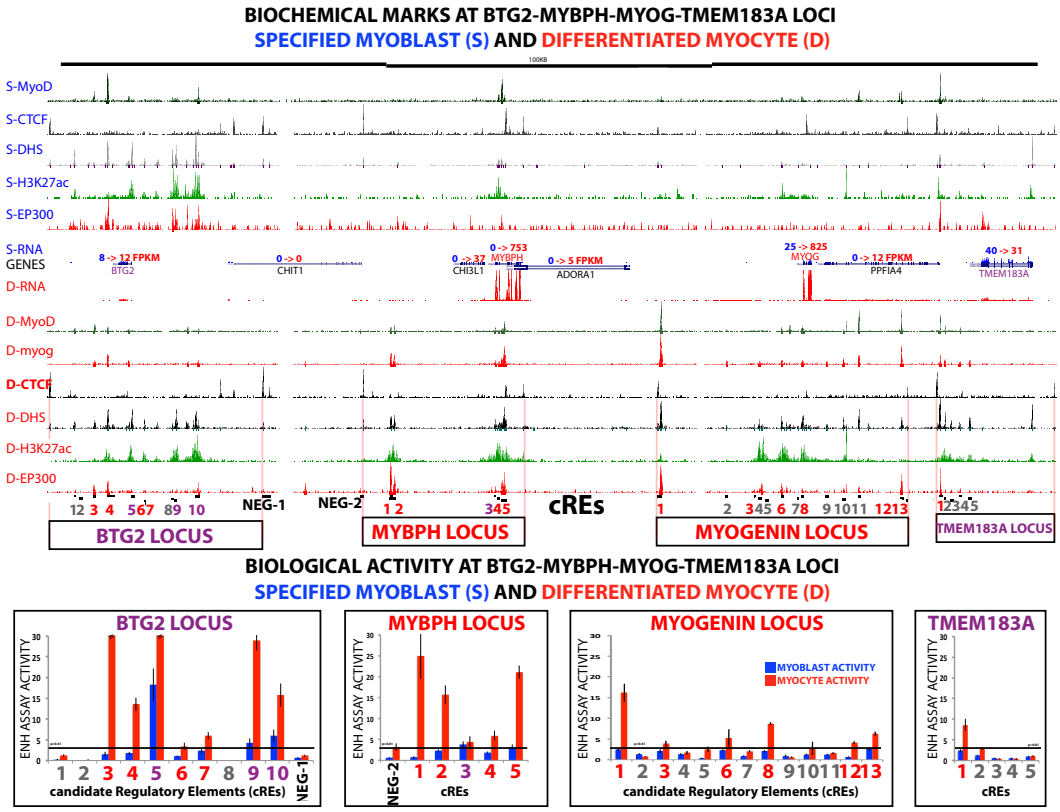


Figure 4.3: BTG2, MYBPH and myogenin locus enhancer activity. Biochemical data and enhancer assay activity provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

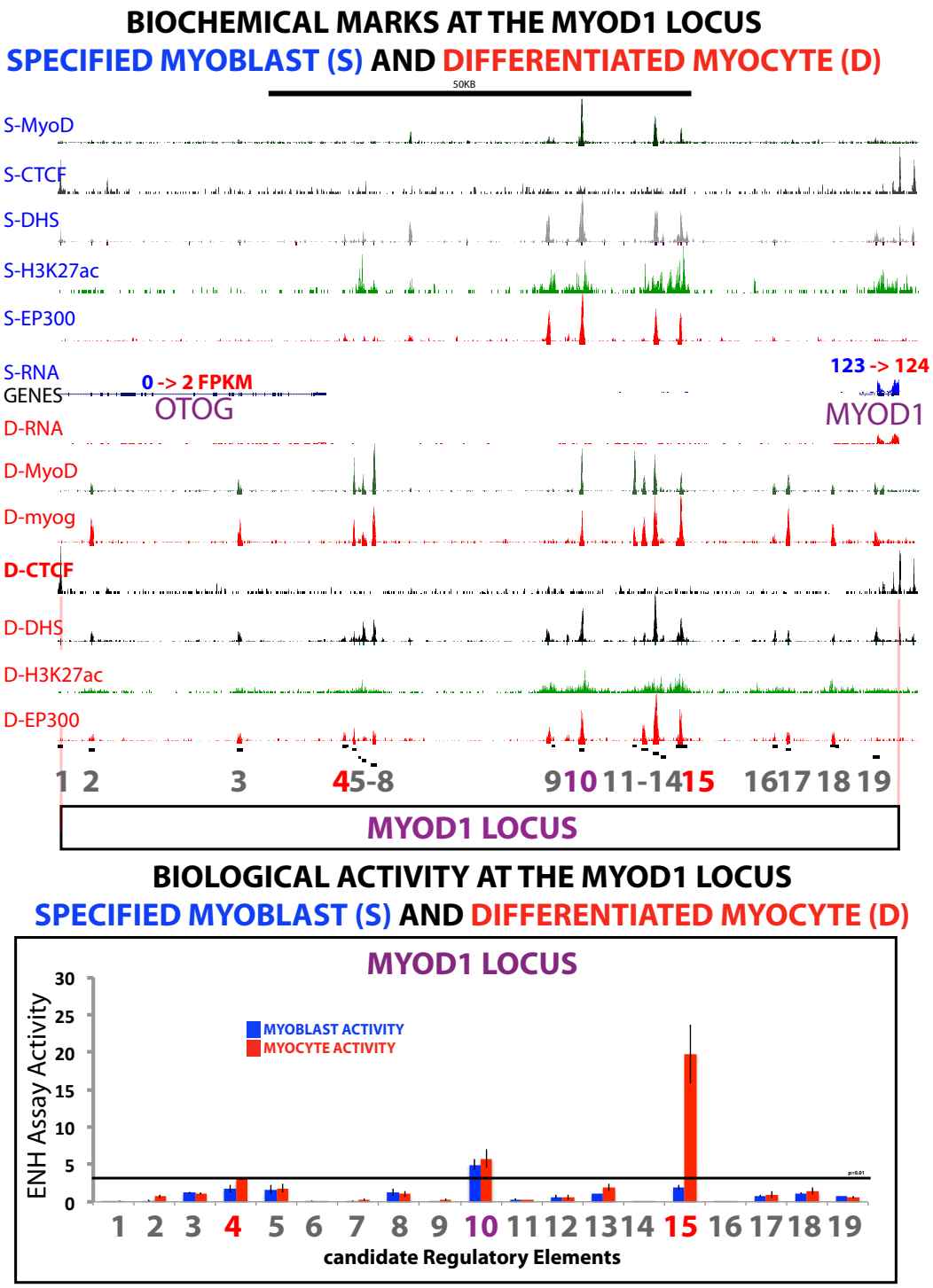


Figure 4.4: (Caption on next page.)

Figure 4.4: MyoD1 Locus enhancer assay activity and Biochemical data provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

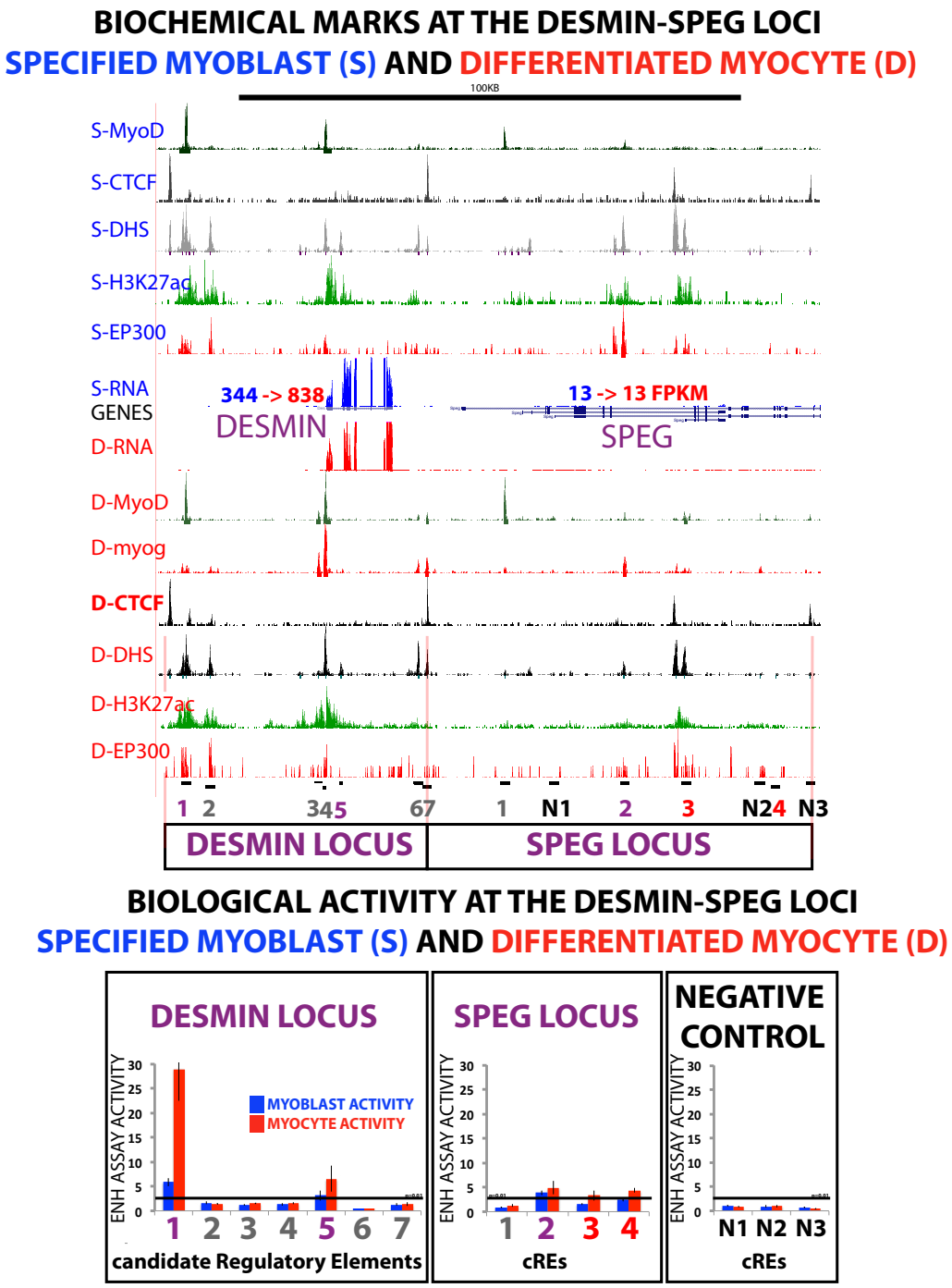


Figure 4.5: (Caption on next page.)

Figure 4.5: Desmin Locus enhancer assay activity and Biochemical data provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

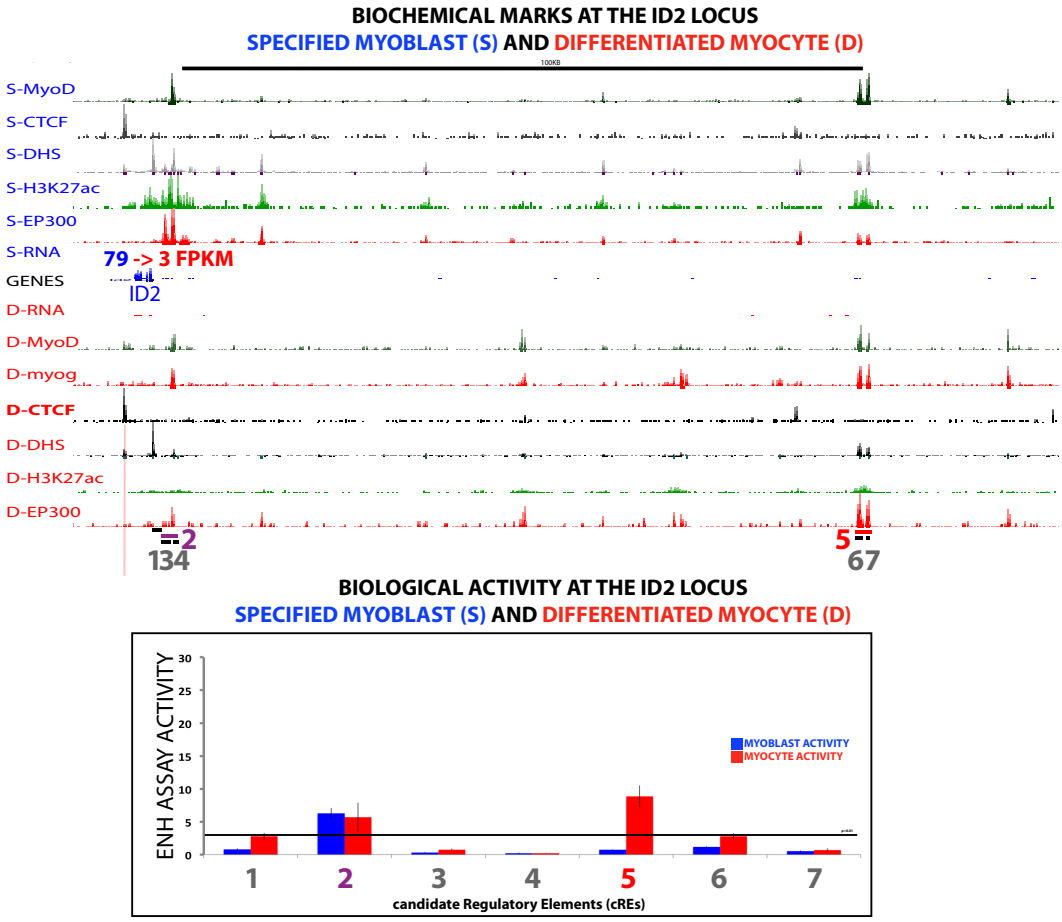


Figure 4.6: LOCUS. Biochemical data and enhancer assay activity provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

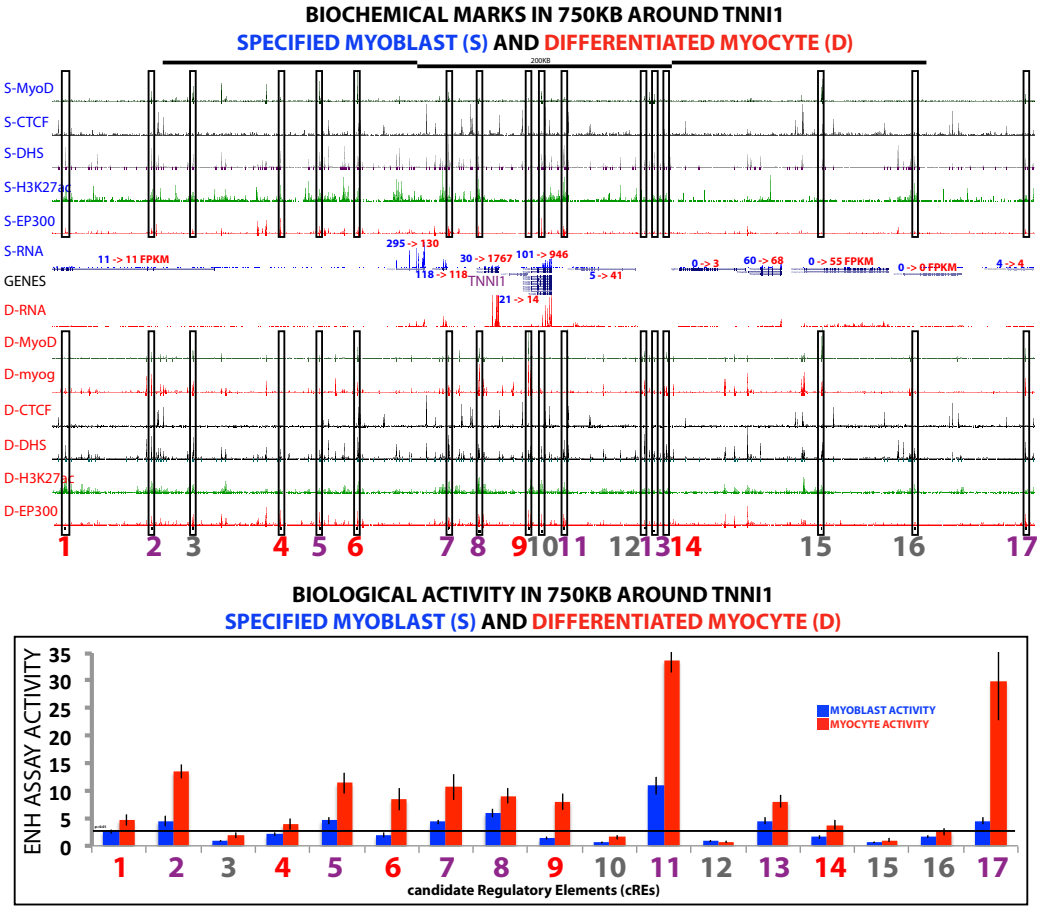


Figure 4.7: 750mb regions around the Tnni1 gene. Biochemical data and enhancer assay activity provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

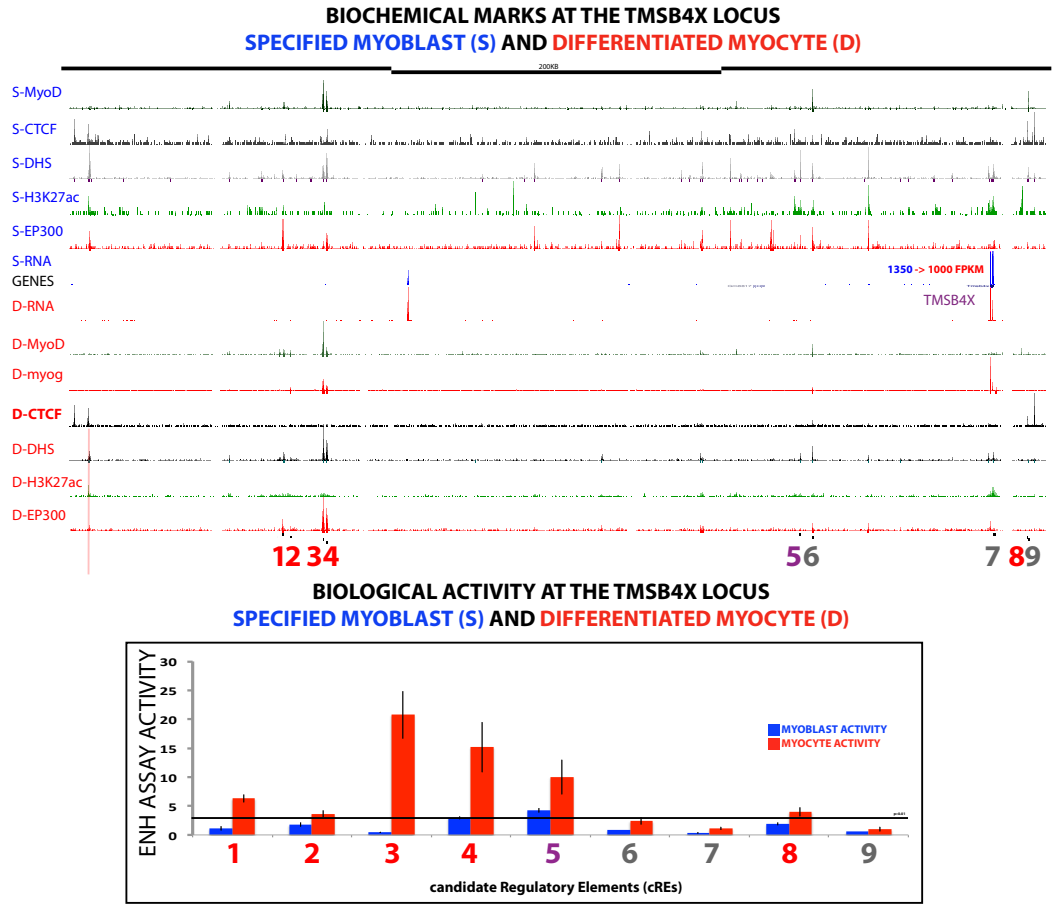


Figure 4.8: TMSB4X Locus. Biochemical data and enhancer assay activity provided for specified myoblasts (blue - S) and differentiated myocytes (red - D). The top 6 biochemical signal tracks represent data including ChIP-Seq and RNA-Seq for specified myoblast (blue - S); while the bottom 7 represent corresponding data for differentiated myocytes (red - D). Tested candidate Enhancer elements are represented by small black boxes below the biochemical tracks, and numbered for each locus with a similar color scheme to represent their enhancer function in myoblasts (blue), myocytes (red) or both (purple). Genes loci are outlined by the black box border, with red lines linking to the observed CTCF boundaries with names encoded to denote the gene as downregulated (blue), flat (purple) and upregulated (red). The measured enhancer activity in myoblast (blue) and in myocyte (red) for each element tested is presented below above its corresponding number. The black bar represents the threshold at which elements are considered an active enhancer.

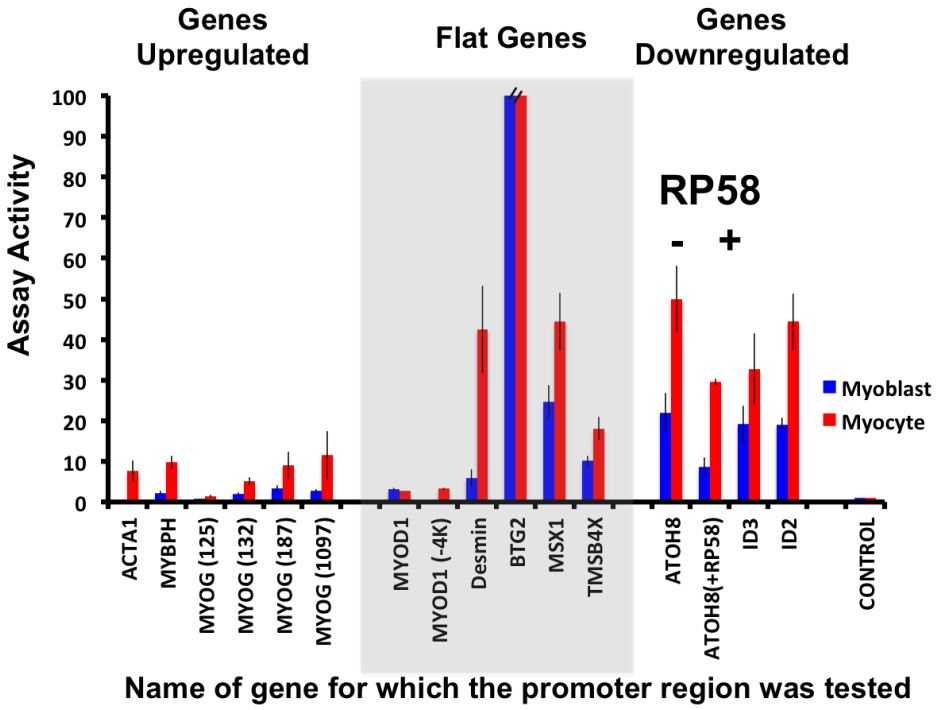


Figure 4.9: Activity of promoter elements for genes upregulated (UP GENES); Flat expressed genes (FLAT GENES) and for genes that are downregulated (DOWN GENES) across the differentiation of C2C12 cells from myoblast to myocyte. No activity threshold is set for these elements as we did not have a distribution of inactive promoters to test against, however the values are normalized to the activity of the TK basal promoter vector.

References

Berghella L De Angelis L, et al. (2008). "A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle." In: *Genes Dev.* 22, pp. 2125–2138.

Berkes CA, et al. (2005). "MyoD and the transcriptional control of myogenesis." In: *Cell and Developmental Biology.* 16, pp. 585–595.

Ghirlando R Felsenfeld G, et al. (2016). "CTCF making the right connections." In: *Genes Dev.* 30, pp. 881–891.

Grace ME Balwani M, et al. (2007). "Type 1 Gaucher disease null and hypomorphic novel chitotriosidase mutations-implications for diagnosis and therapeutic monitoring." In: *Hum Mutat.* 28, pp. 866–873.

Gruen M Gautel M, et al. (1999). "Mutations in beta-myosin S2 that cause familial hypertrophic cardiomyopathy (FHC) abolish the interaction with the regulatory domain of myosin-binding protein-C." In: *J Mol Biol.* 286, pp. 933–949.

Hundley AF Yuan L, et al. (2006). "Skeletal muscle heavy-chain polypeptide 3 and myosin binding protein H in the pubococcygeus muscle in patients with and without pelvic organ prolapse." In: *Am J Obstet Gynecol.* 194, pp. 1404–1410.

Kanneganti M Kamba A, et al. (2012). "Role of chitotriosidase (chitinase 1) under normal and disease conditions". In: *J Epithel Biol Pharmacol.* 5, pp. 1–9.

Li Q Brown J, et al. (2011). "Measuring reproducibility of high-throughput experiments." In: *Ann Appl Stat.* 5, pp. 1752–1779.

Longabaugh WJ Davidson EH, et al. (2005). "Computational representation of developmental genetic regulatory networks." In: *Developmental Biology.* 283, pp. 1–16.

McKnight SL Kingsbury R, et al. (1982). "Transcriptional control signals of a eukaryotic protein-coding gene." In: *Science.* 217, pp. 316–324.

Mikhaylichenko O Bondarenko V, et al. (2018). "The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription." In: *Genes and Development.* 32, pp. 42–57.

Mortazavi A Williams BA, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nat Methods.* 5, pp. 621–628.

Oti M Falck J, et al. (2016). "CTCF-mediated chromatin loops enclose inducible gene regulatory domains." In: *BMC Genomics.* 17, p. 252.

Passeri D Marcucci A, et al. (2006). "Btg2 enhances retinoic acid-induced differentiation by modulating histone H4 methylation and acetylation." In: *Mol Cell Biol.* 26, pp. 5023–5032.

Prevot D Morel AP, et al. (2001). “Relationships of the antiproliferative proteins BTG1 and BTG2 with CAF1, the human homolog of a component of the yeast CCR4 transcriptional complex involvement in estrogen receptor alpha signaling pathway.” In: *J Biol Chem.* 276, pp. 9640–9648.

Tai PWL, et al. (2011). “Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer.” In: *Skeletal Muscle.* 1, pp. 1–25.

Visel A Minovitsky S, et al. (2007). “VISTA Enhancer Browser-a database of tissue-specific human enhancers.” In: *Nucleic Acids Res* 35, pp. D88–92.

Vockley CM DIppolito AM, et al. (2016). “Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome.” In: *Cell.* 166, pp. 1269–1281.

Yee SP Rigby PW, et al. (1993). “The regulation of myogenin gene expression during the embryonic development of the mouse.” In: *Genes Dev.* 7, pp. 1277–1289.

Yokoyama S Ito Y, et al. (2009). “A Systems Approach Reveals that the Myogenesis Genome Network Is Regulated by the Transcriptional Repressor RP58.” In: *Developmental cell.* 17, pp. 836–848.

Chapter 5

DISCUSSION AND POSSIBLE FUTURE DIRECTIONS OF STUDY

5.1 Are biochemical signatures predictive of enhancer activity?

In order to function as an enhancer, a given stretch of DNA must contain the sequence necessary to recruit the TFs that result in biological function (stimulation) plus those that physically target (localization) an element to one or more promoter(s). These localization and stimulation functions need not require separate DNA sequences and might be guided or modified through interactions with non-DNA binding cofactors. (Nakada Y, 2004) It is pertinent in thinking about results and developing future assays that we measured stimulation activity by positioning the test elements adjacent to the reporter gene basal promoter, which renders the measurement indifferent to the localization function, if it exists.

Across all transfection tests in muscle, erythroid and hepatic systems, roughly half of the tested regions functioned significantly while the other half did not. Our muscle and erythroid studies produced hundreds of active elements, allowing a search for underlying distinctions between the groups: for example in sequence motif frequency, biochemical signature, or evolutionary conservation. However, no significant distinction emerged including across biochemical signatures (Figure 3.5); and TF occupied evolutionary conserved regions had similar predictivity rates as randomly selected occupied regions (Figure 3.3, Supplementary Figure 3.11). (Dogan N, 2015) Motif frequency was also not distinguishable between active and inactive elements.

Further, we found that among the significantly active elements, some are quantitatively much stronger than others. Analysis of these top active enhancers in both

muscle and erythroid lineages yielded *de novo* derived motifs but these mapped to both moderately active and inactive regions indicating that they are not a discriminative feature of top enhancers (Appendix B). The enrichment analysis of these expanded motifs relative to their presence in the genome would require isolating a larger set of top enhancers, which are currently numerically limited.

Of course the idea that specific combinations of enhancer elements, including ones that are individually marginal, weak or strong, offers another layer for future testing and integration. Our own initial studies of individual loci, including some results reported in chapter 4, point that this is an important direction to understand the compendium of different functional elements in the genome. Similar conceptual examples have been modeled on the basis of functional data for individual GR and AP1 binding sites coming together - but through looping over greater distances - to form a functional pair by recruiting proximal AP1 sites after GR binding. (Vockley CM, 2016)

In this concluding chapter, I draw on my data and analyses, together with other work in the field, to highlight the questions I currently think are most important together with newly possible approaches to answer them. Overall, my transfection results and comparative analyses across systems raised and consolidated general questions about the relationship of contemporary biochemical and DNA sequence signatures of enhancer capacity and activity state. (Catarino RR, 2018) A second set of questions are more specifically raised in the muscle differentiation system concerning the roles TFs, co-activators, and repressors in establishing and/or maintaining poised and active states.

5.2 Co-factor recruitment and poised elements

EP300, a major co-activating HAT, is required for normal development and has been shown in muscle to recruit the SWI/SNF complex to activate transcription

by acetylating the MyoD TF itself, as well as acetylating histone components of local chromatin conferring modifications associated with opening of chromatin and activity of enhancers. (Albini S, 2010) The presence and magnitude of p300/CBP occupancy, as measured by ChIP, is widely regarded as a superior predictor of an active enhancer in a given cell type or tissue. (Visel A, 2007) As expected from current models, we detect relatively strong EP300 signal at cEnhancers which are only activated after the onset of differentiation, whereas H3k27ac signal is predominant in sites that already measured as (albeit at a lower level) functional in the myoblast. (Appendix Figure B.2 and Appendix B)

The myogenesis field does not yet know if this recruitment is selective for a specific HAT. The lower EP300 signal observed these pan active enhancers (Appendix Figure B.2) is puzzling when both populations are functionally equal in the myocyte and suggests that either some of these elements are only functional in very early steps of differentiation (as the biochemical measurement is temporally decoupled from our enhancer assay by about 36 hours) or that selective HAT recruitment might be an possible mechanism for selecting temporal activity of enhancers.

We also currently do not understand how the enhancer/promoter occupancy, once established, is related to acetylation of its known histone targets and specific TF targets that include MyoD. Some of the experiments presented in the next section aim to dissect some of these functions through mutations. (Albini S, 2010)

Factors that can recognize and bind DNA in previously closed chromatin are operationally known as pioneers. In muscle, PBX1 has been shown to work as a pioneer to help tether MyoD at a specific compound motif in the myogenin promoter just prior to differentiation that requires myogenin expression (Berkes CA, 2004). Supported by my ChIP data, combined with the function assay results, a working model is that PBX1 contributes across the genome to MyoD/E2 mediated pioneering at a subset

of the muscle-class Eboxes (RRCAGGTG). (Nakada Y, 2004; Casey BH, 2018; Fong AP, 2015; Yao Z, 2013)

An attractive idea is that sites in the genome that are pioneered (TF bound but lacking active enhancer-like biochemical signatures) or poised (containing the biochemical signature of an enhancer, but not scoring as functional), are elements that will act elsewhere in development in related cell types/states. In muscle these might include enhancers specific for mature fast-twitch or slow-twitch fiber types, or enhancers specific for different body locations (head versus axial versus limb locations), responsiveness to different signaling pathways, or reserved for adult muscle regeneration. It has also been uncertain if a given bound TF can act simultaneously as both pioneer and activator.

The portion of our functionally tested elements "pioneered" in myoblast (76 of 142 tested - Figure 5.1A, MyoD occupied in myoblast); of these, 80% are "poised" meaning they are either H3K27ac marked or EP300-occupied, scored as either inactive or marginal or the myoblast state. This was a surprise, since the conventional wisdom would be that p300+/H3K27Ac+ is a strong predictor of concurrent enhancer function. These pioneered and (almost entirely) poised elements seen in myoblasts do continue to be marked in myocytes where they can be compared with a second set of elements that are only biochemically occupied by MRFs later, at the onset of muscle differentiation in myocytes (n=66 – Figure 5.1B). Despite their different biochemical occupancy in myoblasts, in the myocyte state the two sets of candidate enhancers were not statistically significant at the p=0.05 for both proportion of active enhancers (Fisher's Exact test p=1) and for the mean and distribution of enhancer activity over each population (ks test p=0.21, t-test p=0.26).

Neither population was strongly active in my transfection assays in the myoblast (though control enhancers were), indicating that while MyoD can pioneer a large

group of sites that will later be active enhancers, MyoD occupancy, even with EP300, is not a powerful activator of transcription until the onset of differentiation. It is worth noting, though some known myoblast enhancers (Figure 4.7 cEnh 11; 4.6 cEnh2, 4.4 cEnh10) and novel (4.3 cEnh 5) that bind MyoD are active in myoblasts and can be used for a first tier stratification analysis against their myocyte exclusive counterparts (Appendix B.2) to inform further experiments.

Among the vast majority that are active only in the myocyte state, the switch could be achieved by a change of heteromeric partner for MyoD; occupancy by a repressor being released (MSY3) or by post-translational activating modification(s), including for example acetylation of the MyoD protein itself or phosphorylations on Mef2. (Albini S, 2010; Molkentin JD, 1996; Black BL, 1998; Berghella L, 2008)

5.3 EP300 interactions with the indirect DNA binding factor MEF2 influence locus activation

A possible model for this EP300 recruitment observed in myocyte can be made from the combination of structural studies and our biochemical observations. Structural studies indicate that EP300 forms up to a potential trimer with the MADS box TFs of the MEF2 group, as each p300 presents 3 pockets for MEF2 binding in its Taz2 domain. (Figure 5.2) (He J, 2011) In differentiated Myocytes MEF2 can occupy DNA directly at canonical MEF2 motifs, but when we examined such class of sites, we found that in myogenesis the vast majority do not recruit EP300 significantly or mediate H3K27ac unless there are also occupying MRF family factors within a given element. (Supplementary Figure 3.5A and C) The occupancy heat maps, paired with analysis of my Mef2 ChIP-seq data (Kirilusha A, manuscript in preparation), show that the majority of Mef2 occupancy in myogenesis is mediated by off-DNA interactions, with directly-bound MRF/E2 at their RRCAGCTG motifs. In striking contrast to DNA-bound MEF2 at its canonical CTAWWWWTAG motif,

these indirect MEF2/MRF co-occupancies avidly recruit EP300 in myocytes, and the regions occupied in this way display the acetylation associated on nearby histones. (Andres V, 1995)

These observations lead me to propose a model in which off-DNA MEF2 is critical for EP300 recruitment and activity at a majority of myocyte enhancer elements. (Figure 5.3). This model covers about 1500 high signal sites, genome-wide, that differ from the numerically minor set of known myocyte enhancers that contain functionally critical MEF2 DNA motifs, such as CKM and myogenin. (Molkentin JD, 1996; Black BL, 1998) These known and well-studied examples are very few in number and happen to be located within a few kilobases of a TSS, qualifying for my definition as "promoter proximal" enhancers. Too few of these MEF2-DNAmotif-requiring elements have been found to clarify if they are a distinct class of element with respect to EP300.

In the myoblast state, the binding profile analysis shows that MEF2 is also recruited to non-MEF2 DNA motif elements, but that this occupancy is independent of active MyoD binding, possibly acting instead through PBX1 occupancy (see MB only section, purple arrows Figure 5.6), where it recruits HATs that acetylate the nearby histones. Note that almost no MEF2 DNA motifs are found in the set of Mef2-only myoblast (MB) genomic sites; whereas the canonical MEF2 sequence motif is present at the center of most myocyte (MC) Mef2A exclusive (no MRF) sites. (Figure 5.6) (Andres V, 1995)

Indeed some of these elements might require promoter proximal regions to facilitate or pioneer in order to allow for occupancy by Mef2 directly on the DNA. It could be informative to translocate some of these promoter proximal elements to a location distal of the TSS and observe their functional output. The opposite, bringing distal elements to a relatively promoter proximal region may prove informative of

cryptic enhancers or repressors if the flanking sequences are made accessible by the translocation. Further, a CRISPR/Cas9 mediated knock-in of a Mef2 DNA motif in a promoter region of a weakly transcribed gene located near function-validated muscle enhancers could be highly informative if controlled by knock-in of a similarly mutated promoter in a known CTCF bounded region generally void of such enhancers (ie globin locus). (Ghirlando R, 2016)

5.4 A mechanistic model for myogenic locus activation

While this explains occupancy, the finding that MEF2 high signal sites are preferentially physically connected in Pol2 Chia-PET data to promoters (Chi Squared p=0.0001; irrespective of their transfection function result or their myogenin occupancy signal), suggests that MRF/MEF2/EP300 interactions may also be explicitly required for bridging distal myogenic elements to other enhancers including promoter proximal ones. (Figure 5.4)

A similar mode of connectivity recruitment to that detected in muscle for regions occupied by Mef2 (Figure 5.3 and 5.4) has been observed in the context of ER where clusters of regions with co-binding by FoxA1 and ER result in a measurable increase of connectivity compared to clusters of regions that only bind ER. Although the combinations of factors in ER detects only 60% of sites as connected, we primarily tested sites that are TSS distal in muscle, also indicated by our 50% baseline of connectivity within tested regions. Because short connections are stripped from such analysis due to technical limits of the connectivity assays the proximal sites may be biased against detecting a connection accounting for the differences noted. (Vockley CM, 2016)

Interestingly, multiple genes in the btg2 - myogenin locus (Figure 4.3) are similarly differentially expressed (albeit at vastly different levels) and in both mouse and human species, even where structural variations have substantially altered the distances

supporting models that the promoter targeting tolerates changes in absolute distance and likely even changes in enhancer order relative to the TSS. (Song G, 2012)

The observation that virtually all myocyte specific gene promoters have at least one MRF proximal occupancy, while 56% of all expressed genes in muscle have a MyoD binding site near the promoter suggests a possible mechanism for promoter targeting through MRF binding. (Figures 5.2 and 5.4) (Cao Y, 2010) A candidate mechanism is that the local chromatin acetylation and basal functional levels require just one Mef2 pocket on EP300 being occupied; as local pioneering opens the chromatin to allow for subsequent factor occupancy (Figure 5.3). Instead, supported by the physical connectivity observed between multiple enhancers, in order to transition to strong enhancement activity in this model, together with long distance physical connectivity with other enhancer and/or promoter elements, multiple pockets of EP300 need to be filled by multiple MEF2 proteins which could each be anchored in a different cis-element. (Figures 5.2 and 5.3) This creates an appealingly flexible model, with the potential to integrate a variable combination of distal and proximal contributions for synergistic activation observed when we combined a few individual regions. It could also explain why biochemical signals from a multi-element complex could be poor predictors of what each individual element can do on its own. (Mikhaylichenko O, 2018; Weintraub H, 1991; Weintraub H, 1990)

A CRISPR-driven point mutation (homozygous) in two of the three EP300 Mef2 pocket(s), assayed with transcriptome and H3K27ac measurements, could test whether chromatin modification and/or transcription can be maintained with just a single pairing possible. Since we have the crystal structure of the interacting domains, we should be able to design disruption of this interaction with relatively minimal effect on overall protein folding. (Figure 5.2)

This proposed model, requiring multiple occupancy of Ep300 pockets, is consistent

with prior observations that 2XMef2 sites can function moderately well to activate the muscle specific promoter of MYH3 in 10T1/2 cells where the p38 MAPK is unlikely to be activated. (Molkentin JD, 1996) In the context of muscle, P38 MAPK and Casein Kinase are involved in the early differentiation of myocytes by targeting Mef2A at its dimerization and DNA binding domains, respectively. ChIP for Mef2, in the presence and absence of P38 MAPK and Casein Kinase inhibitors individually and jointly, could be used to probe the roles of these kinases.

The discussion above focused on EP300 partly because it is the HAT for which I successfully identified and validated a strong ChIP antibody and generated the occupancy data. In addition to EP300, CBP is a paralogous co-activating HAT expressed in myogenic, erythroid, T-cell and Hepatic lineages, while the myogenic system also uses a third HAT, PCAF, which can acetylate MyoD and TCF3 (Sartorelli V, 1999), though it remains unresolved whether the TF-modifying activity is distinct from EP300's, or is always accomplished in association with it. The presence of the other two related activities, reportedly having many targets that overlap, and some that are thought to be specific or biased, means that the as yet unmapped HATs might complicate genetic experiment and may contribute to discrepancies between EP300 occupancy levels at a given target enhancer, the nearby histone acetylation levels and the corresponding enhancer assay output. Large-scale data projects will likely take on the task of making and testing reagents for PCAF and CBP that could then be applied in myogenesis.

5.5 The BTG-Mybph-myogenin locus as a test case for promoter targeting by enhancers

Biochemical signatures were measured in the context of the entire chromosome, but each element was tested for activity in isolation for all of my transfection experiments. Random interactions of these transfected elements with non-native

elements on the chromosomes cannot however be excluded for elements which can provide localization function. Local combinations of individual elements, including ones that partner inactive elements with stronger elements to test for synergy, offer a second layer of organization to probe. There is already precedent in other systems for synergy among elements (Vockley CM, 2016; Zaret KS, 2016), as well as in some of my results (Chapter 4), and it is attractive to think that a more exhaustive test of compound constructs with combinations of elements will better recapitulate overall locus regulatory activity and correlate better with observed biochemical signatures. In particular, larger protein complexes, including perhaps more off-DNA interactions – may be giving substantial TF signatures to elements lacking individual activity but contributing in the context of larger complexes.

In addition, the differential modulation of individual genes in the BTG2-Mybph-myogenin locus (Figure 4.3) suggests that enhancers I have identified likely interact differently on different target promoters. As discussed, the majority (56%) of myocyte specific promoters contain both an Ebox and MyoD binding within them. A simple model is that paired MRF/MEF2 motif content in a promoter elevates the frequency of interactions with distal enhancer elements that contain MRF-only or other combination in which MEF2 is an off-DNA co-factor or linker protein. The promoter paired motif structure could then account for preferential targeting of distal enhancers in myocytes. One of the questions that remains unanswered, is whether the distal enhancers can provide a similar fold modulation when paired with the nearby muscle (MyoD-occupied M-class Ebox) and non muscle (no Ebox or MyoD occupancy) promoters found in the BTG-Mybph-myogenin locus where CTCF sites in the locus and long distances ranging from up to 100kb have been eliminated by the assay construct. A direct test by cloning a modest set of strong enhancers in front of each promoter type would provide insight into whether the enhancers that we identified do interact differently on each specific promoter type.

These data, together with a modest set of native single cell RNA-seq can be interrogated to tell us if these low, but correlated measurements of secondary genes in the locus are the result of jack-potting where a few connections with the strong muscle enhancers to the wrong genes are giving an overall low, but correlated signal. The indication of a strongly preferential, distance guided system (as ChiaPET connection strength appears to largely be when observed in the local context of CTCF connectivity) indicates that the targeting is a somewhat random process. Indeed modifying CTCF boundaries may only matter in instances when they are necessary for retargeting enhancers to different promoters in the locus and appears to be a common function in multiple systems. (Tang Z, 2015; Oti M, 2016; Vockley CM, 2016) Although CTCF knockout in Zebrafish is lethal, and knockdown of this protein results in significant developmental defects. These findings suggests that it is necessary to maintain order across developmental transitions, but whether this factor is required once the correct connections are already established remains unclear. (CarmonaAldana F, 2018)

5.6 Minimal motif presence and spacing required for biological function

A series of experiments from the early 90s remain relevant. They determined the minimum number of Eboxes required for detectable function in a reporter assay, using multiple copies of the right-side E-box (R-sequence) from the CKM enhancer coupled to a Thymidine Kinase reporter linked with a chloramphenicol acetyltransferase reporter (R-TK-CAT). While a construct containing 2 Ebox motifs (2R) was minimally functional, a dose of 4 was required to recapitulate the native level of activity observed for the CKM TSS proximal enhancer (which natively contains 2 Eboxes (R+L) and a Mef2 motif). (Figure 5.5) (Weintraub H, 1990; Weintraub H, 1991) The proposed model calling for multiple Mef2 proteins being recruited, either directly to the DNA or indirectly through MRFs to interact with EP300 to strongly

activate transcription is consistent with these experiments. My own assay results are consistent with the requirement of multiple occupied E-boxes within a construct being required for function. (Figures 5.3 and 5.2)

The 4R-CAT experiment, when compared to the function observed for 3R, suggestion that a minimum spacing might be preferable for E-box sites to function effectively. This is a suggestion rather than a strong conclusion, due to limitations of the early study: thus the 3R-CAT construct included 4NT spacers in between each RRCAGCTGYY Ebox whereas the 2R and 4R cats included 6NT spacers in between the motifs. (Weintraub H, 1990; Weintraub H, 1991) The spacing bias is consistent with much more contemporary observations that bHLH dimer binding can affect the shape of the DNA flanking the RR nucleotides. (Gordan R, 2013) The impact of DNA shape of the wings might affect the Kd of TFs occupying immediately proximal motifs (in a cofactor- and DNA methylation- dependent manner). (Zuo Z, 2017; Jin J, 2016) Similar spacing is also observed in the context of GR between GRE and JunD motifs and a similar model was proposed by our collaborators where pairs of motifs must either come together locally, or through distal interactions recruited once the primary factor binds. They also demonstrated in a small scale test that the distance between motifs impacts the magnitude of enhancer signal observed on assays similar to my own. (Vockley CM, 2016) However relative orientation (as suggested below), numbers and motif types might all prove to have individual effects on this distance effect that are summed to the enhancer activity observed of each candidate element.

The *in vivo* spacing between Eboxes in MRF-occupied native functional enhancer regions I tested was typically >15 base pairs apart with few exceptions. In cases where Eboxes were more tightly spaced than 15bp, there was always a separate third more distantly spaced Eboxe(s) present. This argues that, at least in muscle, natural enhancers have evolved to use multiple E-boxes with relaxed spacing, although

modern experiments that vary spacing overall large numbers of bar-coded constructs are needed to establish real optima and thresholds. At the observational level, very high quality DHS of chemical footprinting will be useful to confirm which motifs are being accessed most frequently.

In contrast to the general conclusion about MRF-MRF spacing above in the muscle system, motifs spacing in some other systems appear to be tightly spaced possibly by evolutionary constraint, such as the GATA1/TAL1 joint motif in erythropoiesis. (Han GC, 2016) There, two different factor types demonstrate very close spacing between motifs. I was able to derive this paired spaced GATA1/TAL1 motif *de novo* from top erythropoietic enhancers, but it also mapped to several less active, and even inactive elements indicating that it is not deterministic for top enhancer function (Appendix B). This indicates that in this system, like ours, additional epigenetic factors or cryptic repressors are likely at play to control the activity of select elements. Although they are technically challenging, re-ChIP experiments, if successful, could clarify which enhancers have true co-occupancy on the same chromosome at the same time for known factors, versus mutually exclusive patterns that would have quite different mechanistic implications. And stronger evidence for occupancy by unknown or unChIPpable factors (including various repressors) could come from high quality footprinting. These data would also drive mutagenesis studies within elements to better understand the functional capacity of *in vivo* Ebox, MEF and other sites. (Gordan R, 2013)

5.7 The combinatorial control role of paired regions

Physical connectivity data (Pol2 ChIA-PET) shows that some enhancers affiliate with promoters and with other distal elements, although it is not clear that these events are simultaneous in the cell. Such pairing of motifs, whether classical muscle (Ebox+Mef2) or the pairing of 2 Eboxes with Mef2 recruited off DNA, might be

required in order to generate strong biological function and can be used to regulate specific effects. (Liu ML, 1994; Weintraub H, 1990) The simplest model uses the binding of MyoD at the promoter region paired with a singular distal element to provide regulation of which genes are being preferentially activated. More complex modeling can involve multiple distal or potentially multiple promoter regions. My locus-level surveys provide some clues, though in depth dissection will be needed to develop the story. For example, the ID2 locus contains candidate Enhancer 2 (Figure 5.7) which is MyoD-occupied and appears inactive on its own. This element only showed discernible enhancer activity when paired with cRE3, which lacks MyoD binding. A second instance in the ID2 locus is illustrated by cRE4+5, where individually non-functional myogenic cREs become an active enhancer when paired. (Figure 5.7) Combining several relatively closely spaced individual modest enhancers located 3' of Acta1 also results in a very powerful enhancer (not shown).

Although physical connectivity measurements are unable to detect connections between such closely spaced elements, there are many distal candidate enhancers from complex loci that are Pol2 or myogenin ChiA-PET connected between them that are either marginal or sub-threshold on their own. These elements can be tested for enhancer function when brought into a plasmid system together with their connected but distal pairs. Testing combinations of marginal elements, and combinations of moderate enhancers with some of these nonfunctional elements, might also improve our understanding of the combinatoric power of these elements. The relative orientation of individual element pairs is also highly relevant and the possible combinations should be tested.

However, as mentioned above, the vast majority of muscle promoters, and those of a large fraction of active genes include an Ebox motif occupied by MyoD. A paired synergistic activity with MyoD occupied promoters might explain selective strong activation of individual genes within a locus where multiple promoter to promoter

connections are also observed. Several studies have shown that these promoter proximal motifs are necessary to retain the proper developmental patterns of gene expression. Reviewed in (Molkentin JD, 1996) The vast majority of genes also only present either none or a modest 1-2 connected distal regions. Elements of this type that are nonfunctional or marginal on their own, should now be tested against the context of native promoter(s) proximal elements.

5.8 Possible roles for cellular context

The specificity of occupancy patterns could be imparted by the primary pioneering factors, as indicated by the modest 20% overlap between Ascl1/MASH (RRCAGCTG) and myogenin (RRCAGSTG) occupancy, despite a much higher overlap in the motif instances across the genome. (Cao Y, 2010; Casey BH, 2018) Similar numbers were found between B-cell and muscle regions in (Kirilusha A, 2014 thesis). This specificity suggests that some sub-threshold elements might be poised for higher activity later in the differentiation time-course or under alternate metabolic or environmental cues. As noted above, this suggests that, in our system they are actively repressed or are awaiting an additional positive acting factor or activating modification - or both. For these reasons, it is possible that some (maybe a majority of) factor-occupied candidate enhancers that score below threshold in the transfection assay, actually function at higher levels in other muscles and developmental stages (e.g. of the head, face, limb or diaphragm, or during regeneration after injury or in response to exercise). In such cases, the biochemical mark identifies regulatory potential that is not realized in the C2C12 cell line used by us and the broader field to represent “skeletal muscle”. Testing this possibility for putative myogenic elements across the life cycle and anatomy of the mouse, either genome-comprehensively, or even for smaller, yet statistically powered sets of 30-100 elements, is conceptually desirable but not yet technically or economically

feasible using current methods of germ line genome engineering and whole body assays.

In mammalian myogenesis, an initial sense of the answer could come from testing a modest group of connected elements that proved inactive in C2C12 cells, in mouse transgenesis covering most muscle systems up through mid gestation in collaboration with the LBL Mouse resource lab. This small set of threshold elements can be mined for candidates affiliated with genes that are strongly expressed in non-skeletal muscle tissues. (Visel A, 2009) For the broader field, the better answer might be to address this issue in a different system – organism and possibly tissue type – to gain better life-cycle access and transgenesis technology. For example, in Zebrafish, automated high throughput screening platforms have been adapted to track the whole organism in real time. (Walker SL, 2012)

5.9 Does orientation of a DNA element affect its potential for enhancer activity and gene partner selection?

The classical definition of an enhancer element, unlike a promoter, is that it can operate in an orientation independent manner. My enhancer assay, with a few exceptions, only tested TSS distal candidate Enhancers in one orientation. From initial tests done by my collaborator Dr. Christopher Partridge at Hudson Alpha working in the liver cell line system, it appears that the orientation of enhancers relative to the promoter can affect relative activity significantly, but that both orientations typically score as enhancers. I also tested a small set of elements from the myogenin locus in both orientations, with similar results. Candidate elements that failed to function in one orientation, also mainly failed when tested in either orientation. The quantitative differentials for active elements might have improved the correlations of activity level with biochemical marks, but they did not. The results overall did not improve the correlation of measured activity with myogenin occupancy, p300 or

active histone mark levels. Our preliminary results showing significant directional bias do suggest future clarification within constructs and especially at their native locations through CRISPR mediated genome editing. This requires allele-marked reporter genes to quantify impact within the same cell. Because the RR-Ebox is directional, and might be preferentially be occupied on either the Watson or Crick strand, depending on the MRF/E pairing, elements making connections to more than one gene can be tested in a bi-directional promoter-reporter system, along with internal RRCAGSTG orientation flips, to test if a single distal element is potentially able to drive both genes simultaneously, or one preferentially depending on the motif orientations. (Polson A, 2011).

A similar orientation selective mechanism can be gleaned from the bi-directional promoters controlling 10% of all genes in the human genome. These genes share a set of elements that have the potential to act to activate both genes or, depending on context, selectively act on one gene specifically. (Trinklein ND, 2004) Similarly, in ChIA-PET data, some of our tested enhancers are cross-connected or connected to multiple genes over long distances. We noticed that many of our most notable biochemical signals, and a portion of our most active enhancers contain sets of motifs that appear directionally paired in opposing orientations. Although our numbers are too small for any statistical claims, they raise an unresolved question as to whether some of these more complex elements (called by some "super-enhancers") are able to cross-connect multiple genes and/or distal elements to act as a stabilizing bridge. Indeed some of these sites may work to selectively create zones of outstanding activity by creating localized enrichments of TFs. (Hnisz D, 2017)

Similar selectivity of motif orientation was found to be necessary for CTCF mediated to establish looping specificity. (Ghirlando R, 2016; Oti M, 2016) The pair of strong enhancers upstream of myogenin, in the intronic region of MYBPH (Figure 3.6 - R3), and between Adora1 and myogenin (Figure 3.6 - R4), could be inverted by

CRISPR *in vivo* (avoiding the nearby CTCT sites) and may be helpful to further our understanding of the orientation requirement of these key regulatory elements. One could also test the most orientation-sensitive (in transfection) sites identified by our collaborators at Hudson Alpha in their respective native locations. I noticed that these elements were often proximal to mapped CTCF occupancy sites, so one would want to test both an orientation flip of the enhancer alone, and later for both the enhancer and nearest CTCF site together.

5.10 High-throughput methods for enhancer screening

When I began this project I hoped to gain insight into the source(s) of large quantitative differences observed in the enhancer "strength" of different elements. These quantitative differences were reproducible and significant in my survey, spanning nearly two orders of magnitude.

Though the activity distinction is encoded in the transfected DNA sequence, my assays did not include finer scale mapping to locate the source of the difference. New methods now allow the field of cis-regulatory element functional analysis that moves to higher scale numerical tests for both gain-of-function and loss-of-function. While they hold promise for finding more examples, these systems are not a cure-all. New methods that now allow the field of cis-regulatory element functional analysis to move to higher scale numerical tests for both gain-of-function and loss-of-function hold promise, yet these systems are not a cure-all. Scaled-up versions of the assays reported here permit samplings on the order of 1E3-1E6 individual elements. (Arnold CD, 2013; Murtha M, 2014)

While they provide increased diversity of sampling and improved statistical power, their designs retain the effects of transient transfections, including unintended reporter construct issues. (Muerdter F, 2018) They also add some new technical and background challenges. For example, in massively parallel reporter assays, the

size of elements (~80-200bp) is currently smaller than the size range of observed, conserved TF occupied regions of the genome (Supplemental Figure 3.1). These multiplexed assays are also affected by technical issues stemming from competition and or promoter crosstalk from the bacterial origin of replication that is present on the transfection plasmid but this later can likely be resolved in the future by PCR amplification or cleavage of part of the plasmid prior to transfection of DNA. (Muerdter F, 2018) Specifically, Starr-Seq in Drosophila and FIREWACH in human cells both report ~24 and 28% respectively of biochemically marked elements as active enhancers. (Arnold CD, 2013; Murtha M, 2014)

This is in contrast to our assays, across different cell types and gene systems, tested larger regions (~400-1kb+; Supplemental Figure 3.1), and reported that ~50% of elements that have clear biochemical signatures associated with activity showed detectable reporter function. This 50% is similar to the overall enhancer discovery rate in the LANL ENCODE transgenic enhancer assay in which candidate elements were selected for activity primarily from individual tissues, suggesting that this 50% rate is likely accurate at least in the context limited to a single state of development. (Visel A, 2007; Visel A, 2009) However, we also show that the majority of the remaining 50% hover just below the activity threshold of statistical significance. While this creates uncertainty for any one element in the population, the group as a whole is detectably different from a group of known tissue-specific enhancers from the immune and nervous systems.

The power of the higher throughput assays might yet be useful in the context of testing "artificial enhancers" such as the experiment in muscle where a minimum of two RR-Eboxes from the R site of the CKM enhancer were found to be a requirement for basal level biological function. (Figure 5.5) (Weintraub H, 1990) These large numbers of candidate elements permits for tests of relative functional output for all RR-Ebox motif permutations individually (as an internal control) and combination

pairings (with both distance and permutations of motifs) within variable relatively short synthesized oligo sequences (~200bp). (Figure 5.8)

An initial test in the GR system showed that distance between AP1 and GR binding can significantly alter the activity level detected, suggesting an expanded numerical test. This same type of assay can be used to test distance requirements between both pairs of functional Eboxes (for example expanding 2R-cat to the same relative distance between Eboxes as 4R-cat) and distances between classical muscle E+Mef2. (Vockley CM, 2016) Similar tests can be performed in the erythroid system for the more closely paired Tal1/GATA1 motifs. (Han GC, 2016)

Insuring high complexity and proper sequencing of the plasmid pool transfected would also allow for the resolution of preferred enhancer orientation, distance and motif numbers. The results from this experiment could then guide further mutation and orientation flip experiments *in vivo*.

5.11 Searching for active repression

The class of negative regulatory elements termed repressors might provide modulatory input for transcription, although their actual biological contribution remains to be understood. In contrast, "silencers" are able to suppress expression of a gene and have demonstrated biological function. For example, in other systems a silencer has been shown to play a role in lineage selection and determination to CD8+ T cells by suppression of the CD4 gene. (Donda A, 1996; Sawada S, 1994)

Region 1 (R1) figure 3.6 of the BTG2 locus points to repressor content in regulatory genomic regions. The experimental observation was that retaining a sequence located between the two independently active, biochemically marked regions, flat-lined enhancer activity observed for either flanking element. A further example might be the partial repression of activity observed in figure 4.10 for the ATOH8 promoter when a larger region containing a ZNF238/RP-58 susceptible

Ebox (CAGSTGT) is included in the construct. The interpretation of these partial results is however confounded in whole population measurements and the effect in individual cells remains unclear. For example, the transfection results obtained for the ID2 enhancers 2+3 sites appear to be marginally downregulated at 24 hours post differentiation with function in individual subpopulations of cells unknown. (Figure 5.7) In contrast, the function of this site measured over a population is further ablated by RP58 by 60 hours. (Kirilusha A, 2014 thesis, Weintraub H, 1990; So KK, 2017; Han GC, 2016; Blackwell TK, 1990; Yokoyama S, 2009)

Many of these candidate repressors are known to bind DNA at specific sequence motifs. (Mortazavi A, 2006) Interestingly, over the years a variety of primarily activating factors have also been sometimes associated with repressor activity, including MyoD itself. (Jayavelu ND, 2018; Chu C, 1997)

A recent study tested for repressor activity using a STARR-seq library containing genomic regions that contain known repressor as well as putative candidate repressive function for known activators including TCF12 and GATA1-4. (Jayavelu ND, 2018) Even though MyoD was not included, no repressive function was detected for any of the activating TF motifs; although these might not be detectable over a population of cells in an assay primarily designed as a first tier test for enhancer activity. The candidate repressive elements successfully isolated a vast library of validated repressor sites. They also scored for known repressive factor binding and found an identical 50% activity rate to our own. (Jayavelu ND, 2018)

These findings however raise the important and thus far less-studied question of repressor contributions, especially for genes where multiple distal elements are brought into contact. The testing of MRF ChIP-seq sites together with point mutations over the known repressor motifs, and conversely over the Ebox motifs in a high throughput cis-repression element assay would increase our ability to understand

the functional contribution of each of these types of candidate elements individually at first, and later using CRISPR-mediated insertions or deletions, to test the ability of some of these repressors by positioning them proximal to otherwise powerful enhancers.

5.12 The possible role of epigenetic context

Although we have studied and measured the "regulatory potential" of a given DNA segment, whether that regulatory potential is free of epigenetic input remains unclear. Most of the regions that we tested span from 2-8 nucleosomes *in vivo* and could be chromatinized after entering the nucleus. It is possible that some candidate enhancers do not function simply because they lack sufficient histone context without prior sequential pioneering of the given DNA region. (Zaret KS, 2016) Further, in a transfection system DNA methylation might affect the naked DNA entering the cell especially at RP58 susceptible Eboxes which are present in approximately half of the tested elements (both active and inactive). (Pollack Y, 1980)

One unanswered question is whether chromatin, and coherent native histone marks, are present in these transfected regions of DNA. Dr. Nergiz Dogan and Dr. Ross Hardison tested mouse candidate enhancer elements against a human B-globin promoter in K562 cells. Many of the tested candidate enhancers whether functional or not come from *in vivo* highly acetylated locations. This "cross species hybrid" assay presents an opportunity to test the chromatin context of transfected DNA. A ChIP-seq against H3K27ac and H3K27me3 aligned to the mouse genome from a population of transfected cells with a set of strong enhancers can be compared to a second population of cells transfected with strongly acetylated regions that scored inactive on the assay. This should answer, with a relatively modest investment, whether the lack of function of these regions is tied to the native histone context not being reflected selectively. This is especially important because naked DNA is being

introduced, and the histone context may be set simply incorrectly over a secondary site with higher affinity, which might not have been pioneered and accessible *in vivo*. The same strategy applies to accessibility measurements (DHS/ATAC) to confirm that similar sites are being accessed on the plasmid vs *in vivo*. The pairing of these measurements should allow us to assay whether a set of elements are being repressed by binding over unexpected motifs, modified of histone context (if any), DNA methylation, or possible lack of ordered pioneering. (Zaret KS, 2016; Catarino RR, 2018)

DNA methylation might play a large role on modifying the affinity of TFs to seemingly identical DNA sites. (Jin J, 2016; Zuo Z, 2017) This is especially interesting in the context of muscle because the cell line 10T1/2 treated with 5-Azacytidine results in the activation of several mesenchyme derivatives including myocytes. More recently 5-Azacytidine has been shown to provoke the transdifferentiation of cardiac cells to myocytes. In general myoblasts threated with 5-Azacytidine will progress through differentiation more quickly. (Taylor SM, 1979; Kaur K, 2014; Pollack Y, 1980)

I looked through the ENCODE human WGBS data available for human myoblast cells in order to ascertain the methylation status of enhancers which function specifically in the myocyte. Although myoblasts are the precursor state, all of the candidate enhancer regions displayed sparse hypomethylation even when occupied differentially in myocytes only. It is important to note that I could not find instances of differentially methylated regions (DMRs) directly over the Eboxes but the effects of 5-Azacytidine on 10T1/2 cells indicates that hypomethylation might be necessary to achieve differentiation into myocytes. Muscle tissue also has one of the highest levels of non CG methylation (mCH) in adult tissues which likely further affects occupancy patterns. (Schultz MD, 2015)

In the case of the glucocorticoid regulatory element (GRE), methylation on the flanks of the motif results in an improved equilibrium dissociation constant. (Jin J, 2016; Zuo Z, 2017) This might explain why the glucocorticoid receptor (GR) might be able to so effectively bind to "orphaned" H3K27ac sites that are currently excluded from the population in use at the time of dexamethasone induction. (Supplemental Figure 3.20A - Middle histogram; note the low DNAse but relatively high H3K27Ac coverage of these regions) In contrast the estrogen receptor is unaffected by flanking methylation sites but core motif methylation prevents binding. (Jin J, 2016; Zuo Z, 2017) The Ebox of Clock/bmal1 (a bHLH heterodimer) instead is unaffected by central methylation of the core EBOX, but is affected by methylation along the immediate flanking nucleotides of the Ebox. (Jin J, 2016)

This result taken together with the above observations of 5-azacytidine, and the distinct binding profiles of both native and overexpressed TFs such as ASCL2 and MYOD (both of which bind separate populations of an identical RR-CAGCTG Ebox) suggests that the pattern of DNA methylation over these sites is likely constant, and that the pattern of binding is regulated by secondary co-factors developmentally prior to the myoblast state. (Casey BH, 2018) This is likely tuned by co-factors, including splice variants and relative phosphorylation state culminating in the highly selective occupancy observed. However, a deep dissection of the methylation status over the motifs may be necessary to fully understand the control of specific TFs.

5.13 Conclusion

These experiments aim to leverage currently available technology to address questions that remain yet to be answered, including the characteristics of the stimulation functions of enhancers (combinatorics, orientation, cellular context and motif requirements) together with their localization function (promoter targeting preferences) all of which are necessary to maintain; and paired with silencers and

repressors to regulate the correct levels of transcriptional products of genes.

5.14 Figures (Chapter 5)

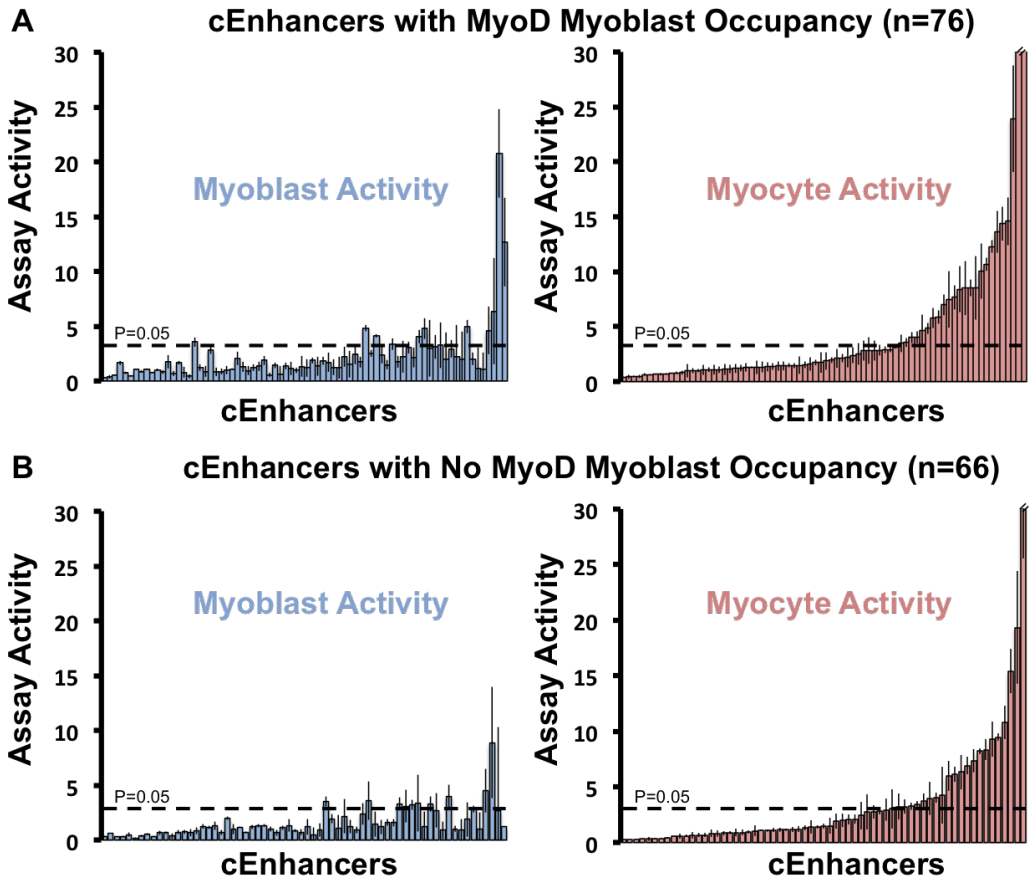


Figure 5.1: Comparison of myoblast and myocyte activity for cEnhancers that are (A) myoblast MyoD occupied (MB+ and MC+) vs (B) cEnhancers only occupied by MyoD upon the onset of differentiation (MB- and MC+).

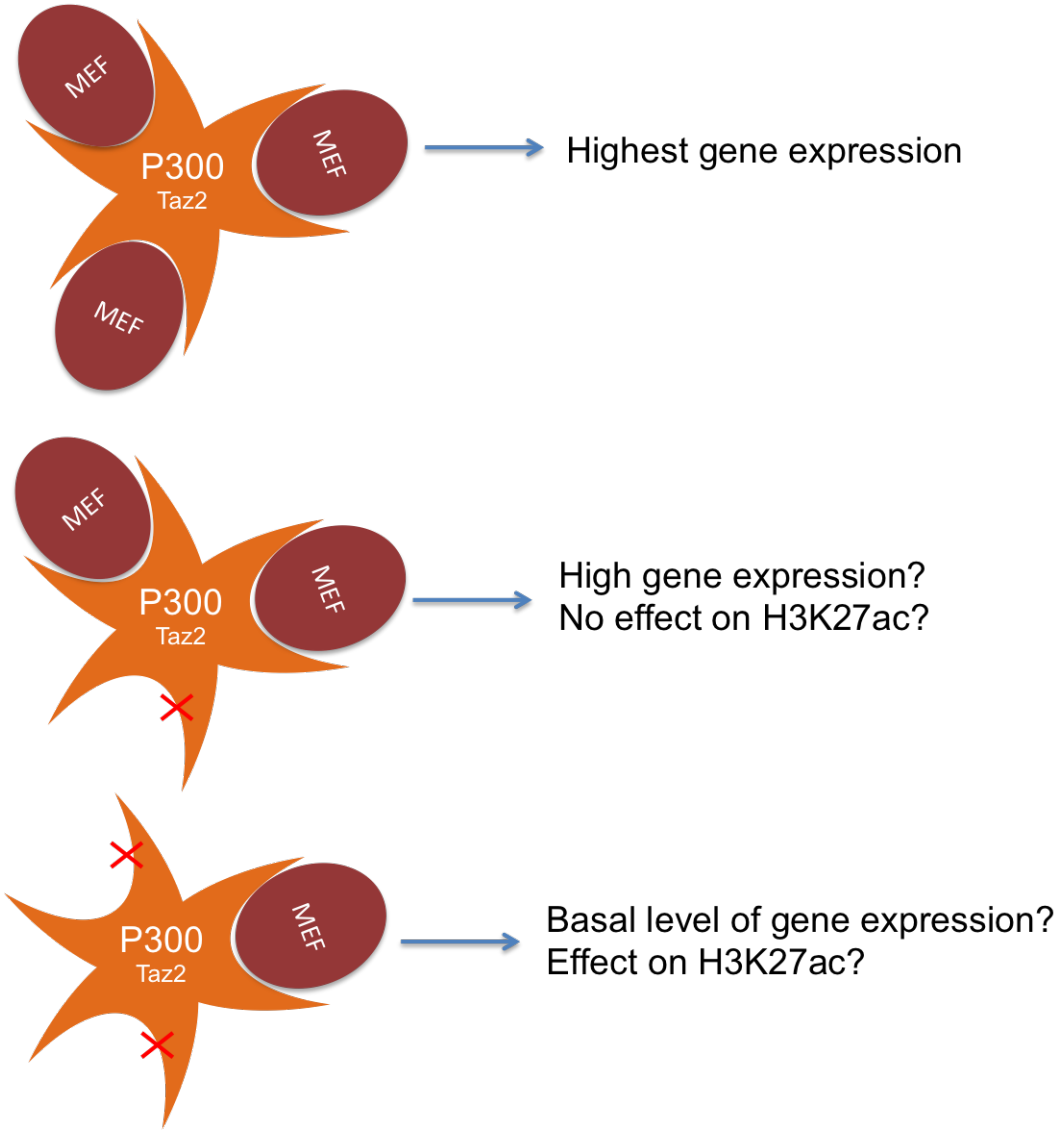


Figure 5.2: Proposed mutation experiment in the Taz2 domain of EP300 affecting the ability to bind multiple Mef2 molecules.

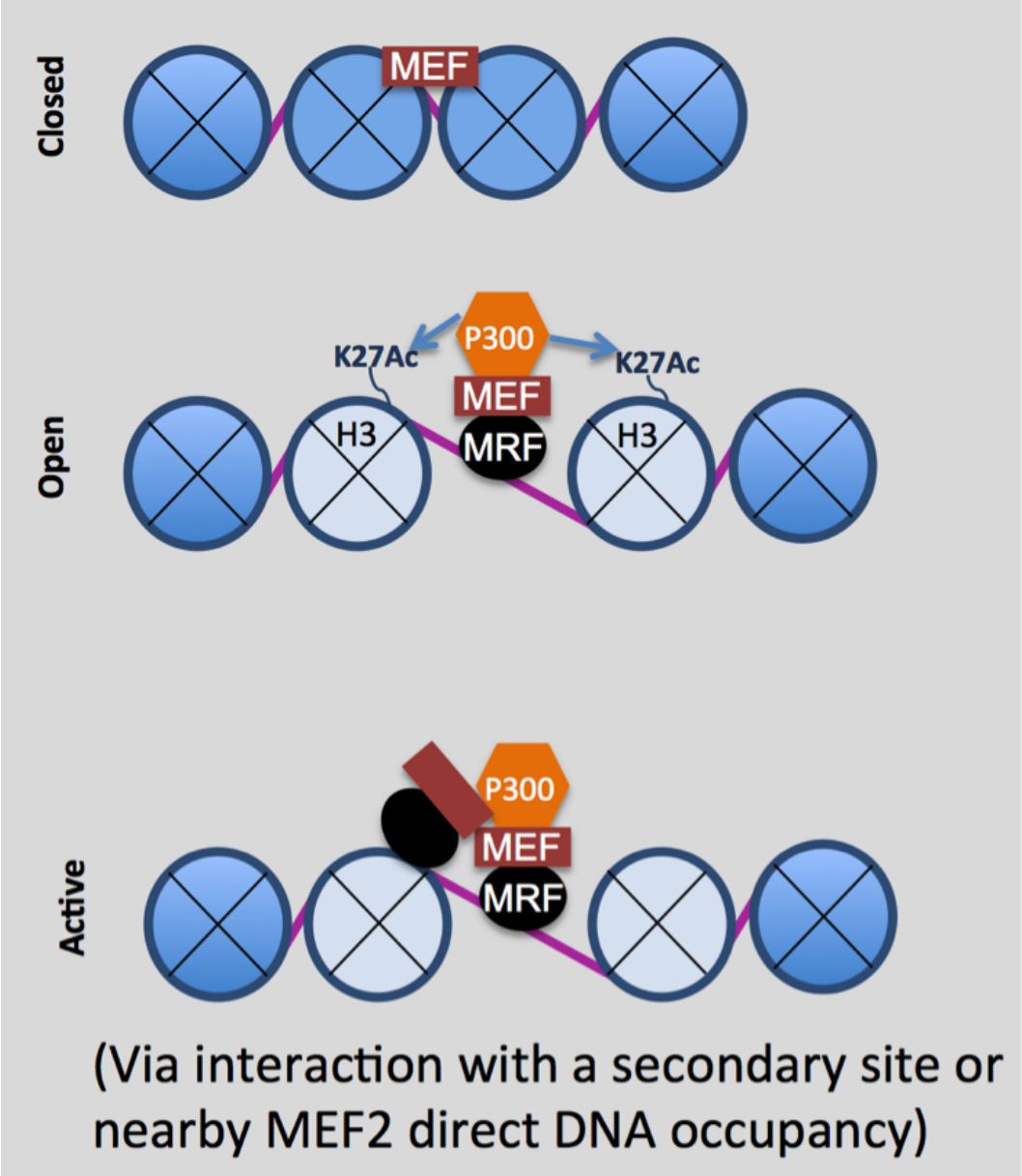


Figure 5.3: Proposed model for minimal requirements of TF occupancy for histone modification and active enhancers in myocytes

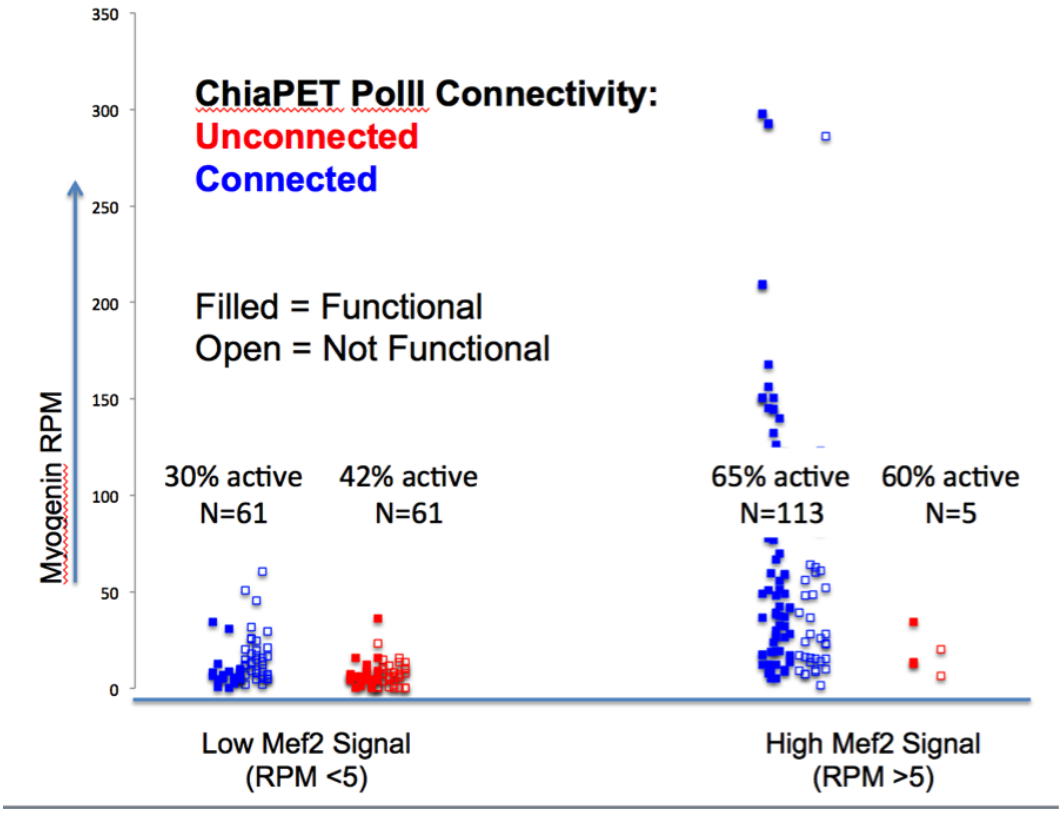


Figure 5.4: POLII Chia-PET connectivity is highly enriched in high vs low Mef2 signal tested cEnhancers

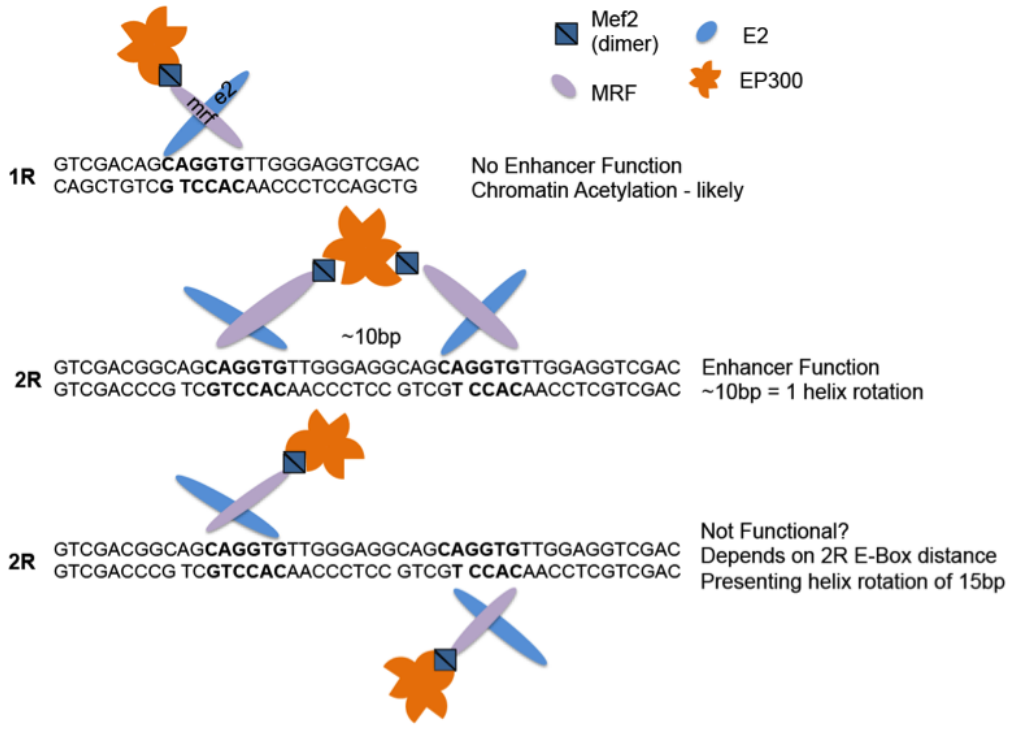


Figure 5.5: Model of 1R vs 2R cat as a demonstration of possible minimal functional requirements of DNA motifs.

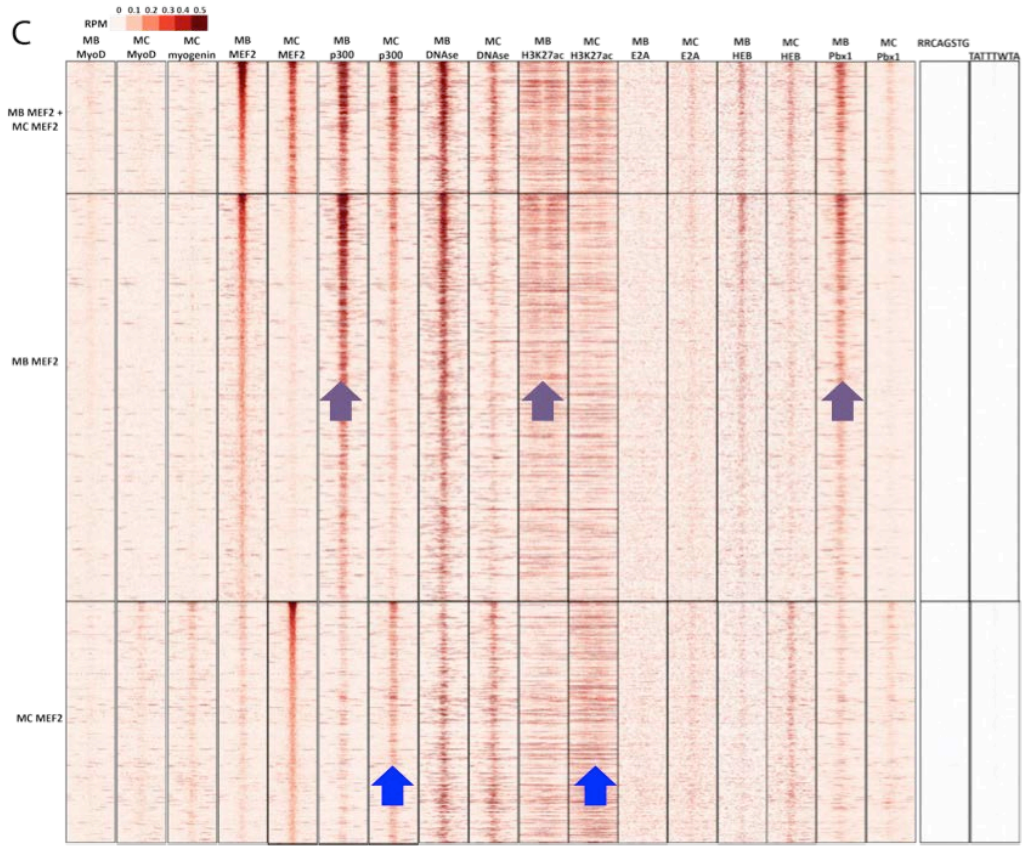


Figure 5.6: Heatmap of ChIP-seq signals for Mef2 exclusively occupied sites of the genome (MRF -)

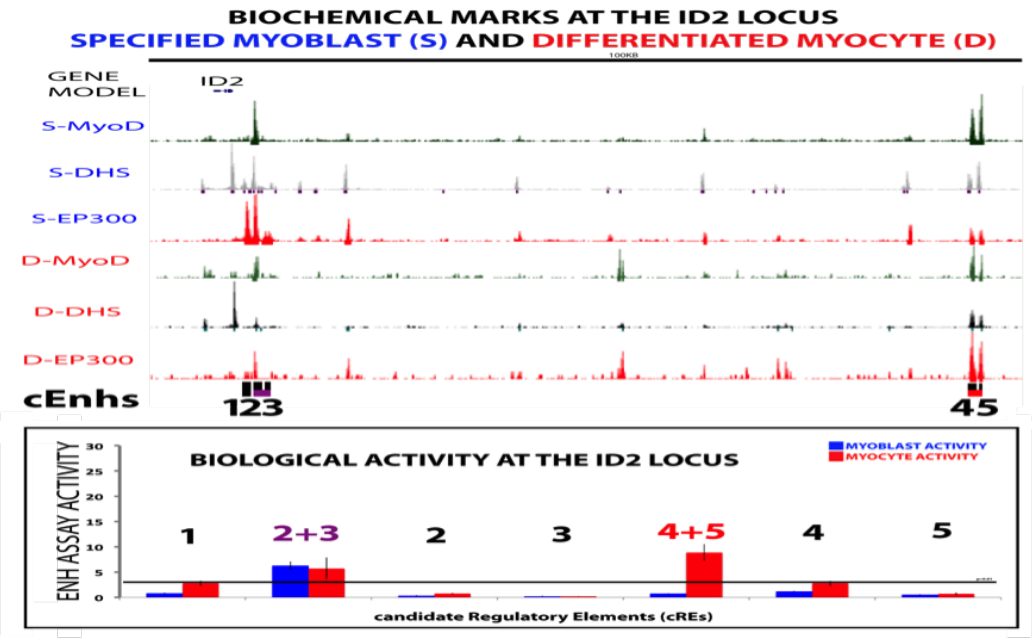


Figure 5.7: Summary of combinatoric Enhancer activity in the ID2 locus.

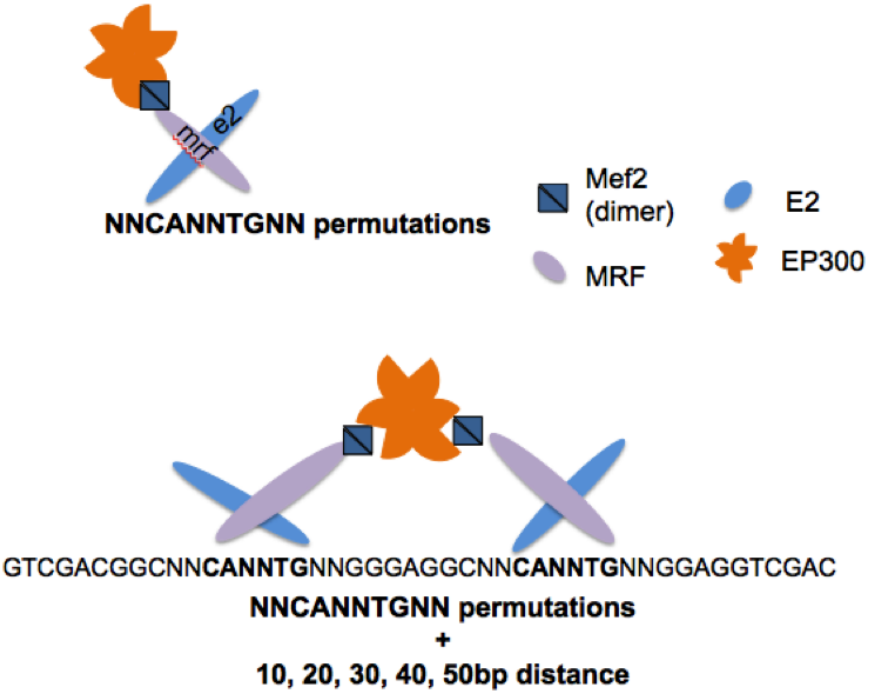


Figure 5.8: High throughput Ebox motif flanks and combinatorics test.

References

Albini S, Puri PL. (2010). "SWI SNF complexes chromatin remodeling and skeletal myogenesis: It s time to exchange!" In: *Exp Cell Res.* 316, pp. 3073–3080.

Andres V Cervera M, et al. (1995). "Determination of the Consensus Binding Site for MEF2 Expressed in Muscle and Brain Reveals Tissue-specific Sequence Constraints." In: *Journal of Biological Chemistry.* 270, pp. 23246–23249.

Arnold CD Gerlach D, et al. (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." In: *Molecular and Cellular Biology.* 339, pp. 1074–1077.

Berghella L De Angelis L, et al. (2008). "A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle." In: *Genes Dev.* 22, pp. 2125–2138.

Berkes CA Bergstrom DA, et al. (2004). "Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential." In: *Mol Cell.* 14, pp. 465–477.

Black BL, Olson EN. (1998). "Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins." In: *Annu Rev Cell Dev Biol.* 14, pp. 167–196.

Blackwell TK, Weintraub H. (1990). "Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection." In: *Science.* 250, pp. 1104–1110.

Cao Y Yao Z, et al. (2010). "Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming." In: *Developmental cell.* 18, pp. 662–674.

CarmonaAldana F Zampedri C, et al. (2018). "CTCF knockout reveals an essential role for this protein during the zebrafish development." In: *Mech Dev.* 2, pp. 0925–4773.

Casey BH Kollipara RK, et al. (2018). "Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors." In: *Genome Res.* 28, pp. 484–496.

Catarino RR, et al. (2018). "Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation." In: *Genes and Dev.* 32, pp. 202–223.

Chu C Cogswell J, et al. (1997). "The Journal of Biological Chemistry." In: *Mol Cell Biol.* 272, pp. 3145–3148.

Dogan N Wu W, et al. (2015). "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility." In: *Epigenetics Chromatin.* 8, p. 16.

Donda A Schulz M, et al. (1996). “Identification and characterization of a human CD4 silencer.” In: *Eur J Immunol.* 26, pp. 493–500.

Fong AP Yao Z, et al. (2015). “Conversion of MyoD to a neurogenic factor binding site specificity determines lineage.” In: *Cell Rep.* 31, pp. 1937–1946.

Ghirlando R Felsenfeld G, et al. (2016). “CTCF making the right connections.” In: *Genes Dev.* 30, pp. 881–891.

Gordan R Shen N, et al. (2013). “Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape.” In: *Cell reports.* 3, pp. 1093–1104.

Han GC Vinayachandran V, et al. (2016). “Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution.” In: *Molecular and Cellular Biology.* 36, pp. 157–172.

He J Ye J, et al. (2011). “Structure of p300 bound to MEF2 on DNA reveals a mechanism of enhanceosome assembly.” In: *Nucleic Acids Res* 39, pp. 4464–4474.

Hnisz D Shrinivas K, et al. (2017). “A Phase Separation Model for Transcriptional Control.” In: *Cell.* 169, pp. 13–23.

Jayavelu ND Jajodia A, et al. (2018). “An atlas of silencer elements for the human and mouse genomes.” In: *bioRxiv.* 2018, pp. 1–42.

Jin J Lian T, et al. (2016). “The effects of cytosine methylation on general transcription factors.” In: *Sci Rep.* 6, pp. 29119–29120.

Kaur K Yang J, et al. (2014). “5-azacytidine promotes the transdifferentiation of cardiac cells to skeletal myocytes.” In: *Cell Reprogram.* 16, pp. 324–330.

Liu ML, et al. (1994). “Myocyte enhancer factor 2.” In: *J. Biol. Chem.* 269, pp. 28514–28521.

Mikhaylichenko O Bondarenko V, et al. (2018). “The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription.” In: *Genes and Development.* 32, pp. 42–57.

Molkentin JD Black BL, et al. (1996). “Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins.” In: *Cell.* 83, pp. 1125–1136.

Mortazavi A Thompson L, et al. (2006). “Comparative genomics modeling of the NRSF/REST repressor network From single conserved sites to genome-wide repertoire.” In: *Genome Res.* 16, pp. 1208–1221.

Muerdter F Boryń M, et al. (2018). “Resolving systematic errors in widely used enhancer activity assays in human cells.” In: *Nat Methods.* 15, pp. 141–149.

Murtha M Tokcaer-Keskin Z, et al. (2014). “FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells.” In: *Nat Methods.* 11, pp. 559–565.

Nakada Y Hunsaker TL, et al. (2004). “Distinct domains within Mash1 and Math1 are required for function in neuronal differentiation versus neuronal cell-type specification.” In: *Development.* 270, pp. 1319–1330.

Oti M Falck J, et al. (2016). “CTCF-mediated chromatin loops enclose inducible gene regulatory domains.” In: *BMC Genomics.* 17, p. 252.

Pollack Y Stein R, et al. (1980). “Methylation of foreign DNA sequences in eukaryotic cells.” In: *PNAS.* 77, pp. 6463–6467.

Polson A Durrett E, et al. (2011). “A bidirectional promoter reporter vector for the analysis of the p53/WDR79 dual regulatory element.” In: *Plasmid.* 66, pp. 169–179.

Sartorelli V Puri PL, et al. (1999). “Acetylation of MyoD directed by PCAF is necessary for the execution of the muscle program.” In: *Mol Cell.* 4, pp. 725–734.

Sawada S Scarborough JD, et al. (1994). “A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development.” In: *Cell.* 77, pp. 917–929.

Schultz MD He Y, et al. (2015). “Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation.” In: *Nature.* 523, pp. 212–216.

So KK Peng XL, et al. (2017). “Whole Genome Chromatin IP-Sequencing (ChIP-Seq) in Skeletal Muscle Cells.” In: *Methods Mol Biol.* 1668, pp. 15–25.

Song G Riemer C, et al. (2012). “Revealing Mammalian Evolutionary Relationships by Comparative Analysis of Gene Clusters.” In: *Genome Biology and Evolution.* 4, pp. 586–601.

Tang Z Luo OJ, et al. (2015). “CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.” In: *Cell.* 163, pp. 1611–1627.

Taylor SM Jones PA, et al. (1979). “Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine.” In: *Cell.* 17, pp. 771–779.

Trinklein ND Aldred SF, et al. (2004). “An Abundance of Bidirectional Promoters in the Human Genome.” In: *Genome Research.* 14, pp. 62–66.

Visel A Blow MJ, et al. (2009). “ChIP-seq accurately predicts tissue-specific activity of enhancers.” In: *Nature.* 457, pp. 854–858.

Visel A Minovitsky S, et al. (2007). “VISTA Enhancer Browser-a database of tissue-specific human enhancers.” In: *Nucleic Acids Res* 35, pp. D88–92.

Vockley CM DIppolito AM, et al. (2016). “Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome.” In: *Cell.* 166, pp. 1269–1281.

Walker SL Ariga J, et al. (2012). “Automated Reporter Quantification In Vivo: High-Throughput Screening Method for Reporter-Based Assays in Zebrafish.” In: *Nucleic Acids Res* 7, e29916.

Weintraub H, et al. (1991). “Muscle-specific transcriptional activation by MyoD.” In: *Genes Dev.* 5, pp. 1377–1386.

Weintraub H Davis R, et al. (1990). “MyoD binds cooperatively to two sites in a target enhancer sequence: occupancy of two sites is required for activation.” In: *PNAS* 87, pp. 5623–5627.

Yao Z Fong AP, et al. (2013). “Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites.” In: *Skeletal Muscle.* 3, pp. 1–8.

Yokoyama S Ito Y, et al. (2009). “A Systems Approach Reveals that the Myogenesis Genome Network Is Regulated by the Transcriptional Repressor RP58.” In: *Developmental cell.* 17, pp. 836–848.

Zaret KS Lerner J, et al. (2016). “Chromatin Scanning by Dynamic Binding of Pioneer Factors.” In: *Molecular Cell.* 62, pp. 665–667.

Zuo Z Roy B, et al. (2017). “Measuring quantitative effects of methylation on transcription factor–DNA binding affinity.” In: *Sci Adv.* 3, pp. 1799–1800.

Appendix A

TRANSFECTION ASSAY DATA FOR ALL 371 REGIONS TESTED IN MUSCLE

The table is available at:

$http://woldlab.caltech.edu/\sim gdesalvo/Wold_transfection_data01152017.xlsx$

The following is a description of the data available in the table: 1. Cell type / Cell line: C2C12 Myoblast C2C12 Myocyte 10T1/2 Fibroblast 10T1/2 Mock Differentiated

2. Transfection protocol: transient transfection Elements were cloned by SwitchGear Genomics 5' of a custom TK promoter driving a PEST containing synthetic luciferase reporter gene. 50ng of fluorometer measured single plasmid was transfected for each technical replicate using the Lipofectamine® LTX with Plus™ reagent.

$https://tools.thermofisher.com/content/sfs/manuals/$

$LipofectamineLTX_PLUS_Reag_protocol.pdf$

3. Assay reagent: Steady-Glo Luciferase Assay System $https://www.promega.com/$

$products/reporter-assays-and-transfection/$

$reporter-assays/steady_glo-luciferase-assay-system/$

4. Element design/prediction A first set of candidate Regulatory Elements were selected to sample occupancy signal in a Myogenin ChIP-Seq experiment.

A second set of candidate Regulatory elements were selected to sample Myogenin occupancy affiliated with key genes that are modulated at the transcriptional level during the differentiation of muscle.

A set of negative control elements were selected from previously characterized T-cell and neuronal enhancer regions that contain a muscle class E-box motif.

5. Replication design: The candidate Regulatory Elements were assayed when possible as biological replicate (transfections performed on different days), with 2-4 underlying technical replicates for each biological replicate.

6. Explanation of numerical assay value, including normalization performed: The units for activity are “fold change relative to activity from parental vector”. Each group of transfections included a transfection with the control parental firefly luciferase plasmid that lacked a test DNA insert.

7. Assignment of Activity bin for each tested DNA segment (also stated in primary data) The rule for declaring a DNA segment an "Enhancer" is that its fold change is at least two standard deviations above the mean of the negative control elements set as a whole. All others are declared "NotEnhancer".

Appendix B

DE NOVO MOTIF ANALYSIS

The vast majority of our candidate Enhancer regions (represented here by a random sample of 77 elements) are only active in the myocyte state. (Figure B.1) Over the entire population of over 300 elements tested a modest set of 24 also function at significant levels in the myoblast state. (Figure B.2 Pan Active Enhancers)

These pan active enhancers are different and perhaps interesting; but end up invariably upregulated in activity upon differentiation raising the possibility that most of the signal is coming form early differentiating cells. These can however be compared to elements which are strictly specific for enhancer function to the myocyte state. (Figure B.2)

The population of pan active enhancers trends towards being more active than the myocyte specific enhancers, but fails to meet statistical significance at the p=0.05 level (t-test p=0.18). This difference in activity observed can be explained because we are assaying at 24 hour post differentiation, and the pan active elements had a head-start in terms of activity in the myoblast which if subtracted would bring the populations in line.

The means of signal of EP300 (Welch t-test p=0.018 - mean of 42.5 (specific) and 22.2 (pan)) and H3K27ac (Welch t-test p=0.013 - mean of 16.5 (specific) and 36.0 (pan)) appear selectively different at the p=0.05 level across these two classes of elements, with EP300 signal being stronger in signal in elements recently accessed, while H3K27ac appears to have on average stronger signal in elements that were already active at the myoblast state. The DHS ratios for myoblast/myocyte across both classes of enhancers are also presented in Figure B.2. Both of these findings

are in accord with current models that EP300 is recruited to sites being activated functions to deposit H3K27ac, which can also lag behind activity. (Zhang JA, 2012; Catarino RR, 2018)

In order to investigate potential motifs selecting for activity in these myocyte specific enhancers, I performed *de novo* motif finding on these two classes of elements. As expected both derived an RRCAGCTG E-box but the elements specific for activity in the myocyte also derived an Ebox-like (but expanded) motif. (Figure B.3)

Observation of ChIP-seq and histone mark data from the myoblast state within the boundaries of the pan active enhancers show that half of these active elements are not likely to be driven by MyoD. (Figure B.4) Interestingly strong H3K27ac is present in some of these low level MyoD/EP300 sites indicating that they are either orphaned H3K27ac locations (ie methylated on the chromosome?) or non MRF enhancers which recruit H3K27ac through another HAT present in our system. If these sites were numerically greater one could score for the presence of other TF motifs (ie AP1) and relative presence or absence of EP300 occupancy compared to the MyoD occupied half. (Figure B.4)

In the regions co-occupied by Tal1 and GATA1 — two key Erythropoietic TFs — a new paired recognition motif has been derived. This motif has a half E-box (CTG) locked in orientation with a GATA motif (AGATAA) about 8 base pairs away. (Han GC et al. 2016) After discovering a novel motif in muscle I used a similar strategy in Erythropoiesis using different classes of active enhancers to derive *de novo* motifs. (figure B.5) I was able derived both a CGT half E-box and a AGATAA motif from average enhancers. In the top enhancers I also derived both a full CATGTG E-box and a GATA motif that is similarly spaced to the paired motif derived from Tal1/Gata1 co-occupied regions.

Unfortunately in both muscle and erythropoiesis these new motifs appear to map

to both active and inactive (pan and specific for muscle) cEnhancers, meaning that they are not deterministic of (strong); or cell type specific enhancer function.

B.1 Figures (Appendix B)

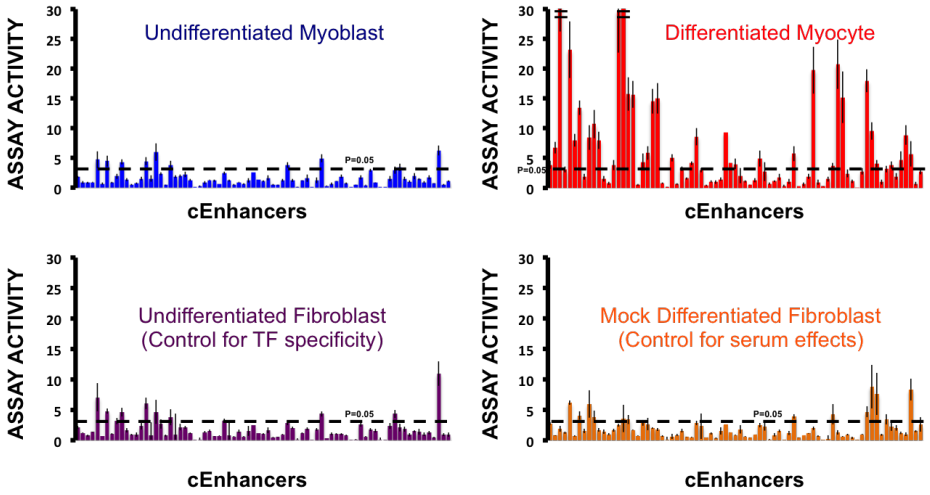


Figure B.1: Summary of cEnhancer activity ranked by affiliated gene RNA transcript ratio across the differentiation of C2C12 cells (High -> Low; myocyte/myoblast ratio). The line on each plot represents the activity threshold at which tested elements are considered enhancers in each cell state. Elements in myocyte are proportionally more active compared to elements in the myoblast (Fisher's exact P=0.01).

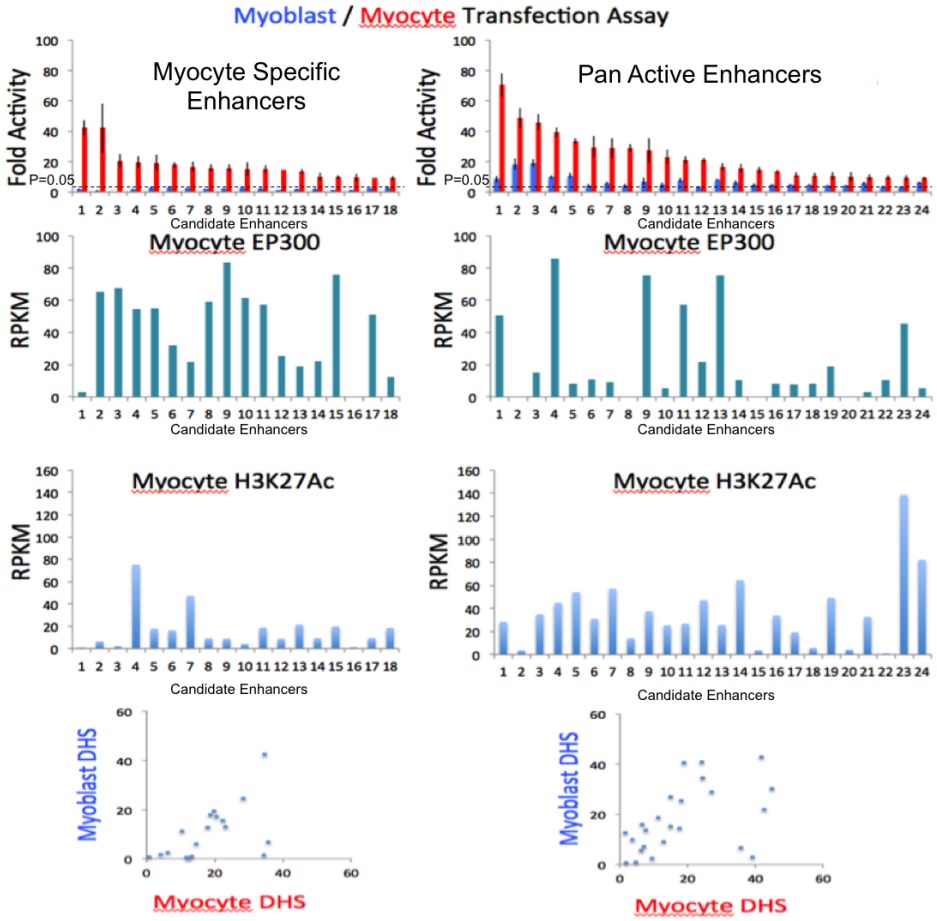


Figure B.2: EP300, H3K27ac and DHS signal in specific vs pan active enhancers across myoblast to myocyte differentiation. The means of signal of EP300 (Welch t-test p=0.018 - mean of 42.5 (specific) and 22.2 (pan)) and H3K27ac (Welch t-test p=0.013 - mean of 16.5 (specific) and 36.0 (pan)) appear selectively different at the p=0.05 level across these two classes of elements, with EP300 signal being stronger in signal in elements recently accessed, while H3K27ac appears to have on average stronger signal in elements that were already active at the myoblast state.

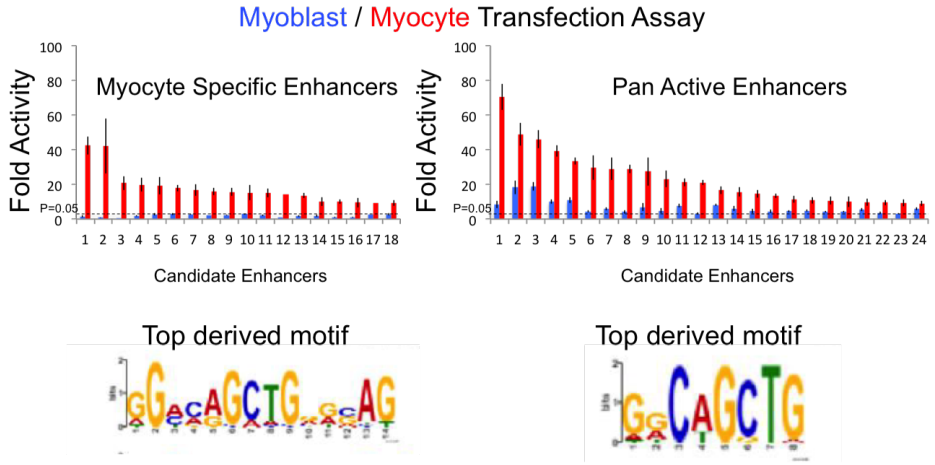


Figure B.3: Motifs derived from specific vs pan active enhancers across myoblast to myocyte differentiation.

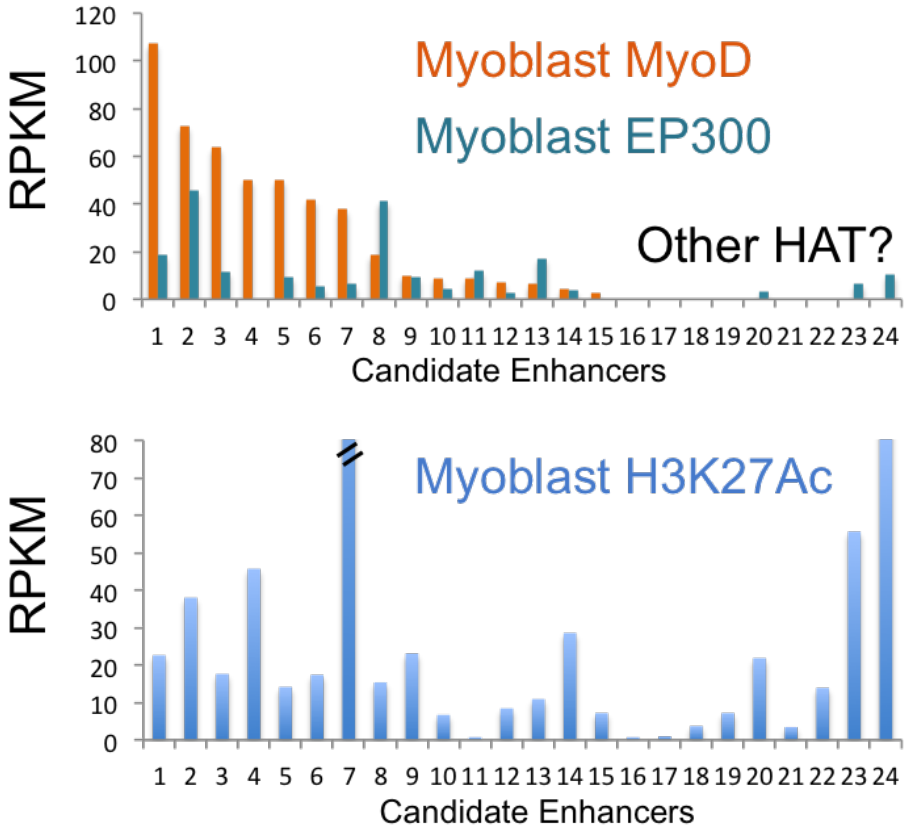


Figure B.4: EP300 and H3K27ac ChIP-seq signal in myoblast for pan active enhancers; ranked by MyoD signal.

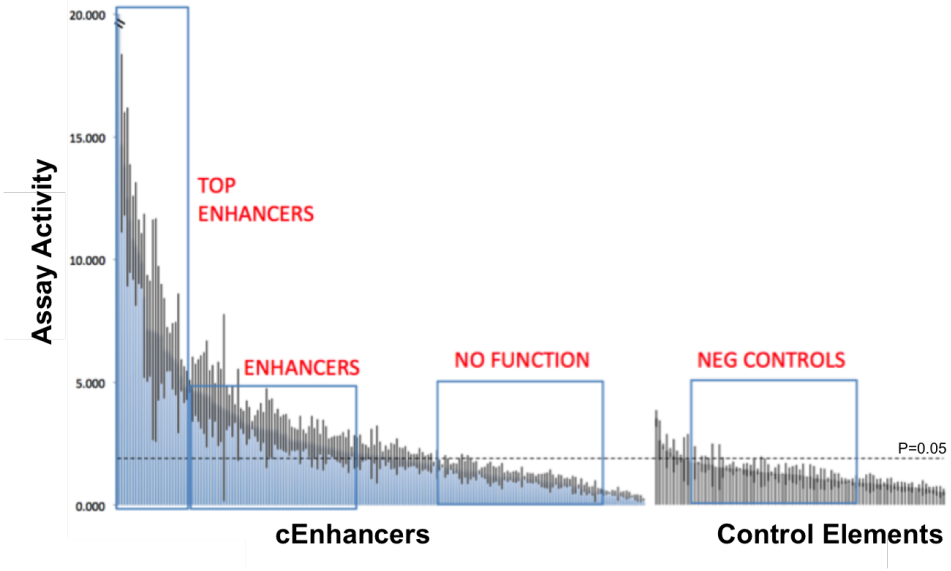


Figure B.5: Erythroid cREs and negative control elements tested in K562 cells, by activity.

Appendix C

ENCODE CHIP-SEQ DATA SUBMISSION

Although these early measurements provided much of the groundwork for selecting candidate elements for my thesis, it was necessary for ChIP-Seq measurements to match the new standards in the field (1PCR) and provide independently grown duplicates for each of these measurements with matched background samples so that the data could be ran through modern IDR based analysis pipelines. With a robotic ChIP-Seq platform developed by Clarke Gasper in the PEC (Gasper WC, 2014), I produced a new set of factor occupancy measurement for several of the key muscle differentiation factors (Mef2, MyoD, Myogenin, E12/HEB) as well as key co-activators such as the HAT EP300. I measured the H3K27ac, H3K4me1 and H3K4me2 histone marks in order to contrast their biochemical occupancy with my enhancer assay results. These ChIP-seq measurements are summarized in the table provided in Appendix C. We have also modern DHS data kindly provided by the Mortazavi Lab for C2C12 cells at both myoblasts and myocytes which was used for much of the analysis presented in Chapter 3. The modern occupancy landscape of these and other factors involved in the differentiation process are illustrated for reference in Supplementary Figure 3.5.

The following tables summarize this ChIP-Seq data:

*http://woldlab.caltech.edu/~gdesalvo/RobotChIPSeq*2015 – 2016.*xlsx*

All of the listed libraries, where possible, were called by Dr. Georgi Marinov using the ENCODE 3 pipeline (*https://www.encodeproject.org/pipelines/*) and are available in on our Lab cluster ChIP repository.

Appendix D

COMBINATORICS (SOM) BASED CANDIDATE ENHANCER ELEMENTS

The following set of Self Organizing Maps (SOMs) were generated using SOMatic (available at *http://crick.bio.uci.edu/SOMatic/*) as part of an effort to analyze my enhancer assay data. Although these maps were not used as part of the manuscript they still represent a resource for future analysis of the C2C12 and multi-tissue comparisons of TF; CTCF, Pol2, DNAse and histone marks. All of these SOMs may be requested.

Custom SOM Viewer

I helped develop a beta version of a custom viewer in collaboration with Santiago Lombeyda at the Caltech CD3 which can be viewed here:

http://woldlab.caltech.edu/~gdesalvo/d3 – viewer – v.0.31/viewer.html

mm9.200.C2Only.HDPP SOM Includes Histone Marks; DNAse; Pol2; CTCF and EP300 in both C2C12 exponential and differentiated states.

C2C12-D-DHS-DnaseHS C2C12-D-HM-H3K27ac C2C12-D-HM-H3K27me3 C2C12-D-HM-H3K36me3 C2C12-D-HM-H3K4me1 C2C12-D-HM-H3K4me2 C2C12-D-HM-H3K4me3 C2C12-D-PM-Pol2-4h8 C2C12-D-TF-CTCF C2C12-D-TF-EP300 C2C12-D-ZI-input C2C12-E-DHS-DnaseHS C2C12-E-HM-H3K27ac C2C12-E-HM-H3K27me3 C2C12-E-HM-H3K36me3 C2C12-E-HM-H3K4me1 C2C12-E-HM-H3K4me2 C2C12-E-HM-H3K4me3 C2C12-E-PM-Pol2-4h8 C2C12-E-TF-CTCF C2C12-E-TF-EP300 C2C12-E-ZI-input

mm9.200.C2Only Includes the following datasets:

C2C12-D-DHS-DnaseHS C2C12-D-HM-H3K27ac C2C12-D-HM-H3K27me3 C2C12-D-HM-H3K36me3 C2C12-D-HM-H3K4me1 C2C12-D-HM-H3K4me2 C2C12-D-HM-H3K4me3 C2C12-D-PM-Pol2-4h8 C2C12-D-TF-CTCF C2C12-D-TF-E2A C2C12-D-TF-EP300 C2C12-D-TF-GABP C2C12-D-TF-HEB C2C12-D-TF-Mef2 C2C12-D-TF-MyoD C2C12-D-TF-myogenin C2C12-D-TF-SP1 C2C12-D-TF-SRF C2C12-D-TF-ZFP143 C2C12-D-ZI-input C2C12-E-DHS-DnaseHS C2C12-E-HM-H3K27ac C2C12-E-HM-H3K27me3 C2C12-E-HM-H3K36me3 C2C12-E-HM-H3K4me1 C2C12-E-HM-H3K4me2 C2C12-E-HM-H3K4me3 C2C12-E-PM-Pol2-4h8 C2C12-E-TF-CTCF C2C12-E-TF-E2A C2C12-E-TF-EP300 C2C12-E-TF-FOSL1 C2C12-E-TF-HEB C2C12-E-TF-Mef2 C2C12-E-TF-MyoD C2C12-E-TF-SRF C2C12-E-TF- ZFP143 C2C12-E-ZI-input

mm9.200.G1EOnly.HDPP

G1E-D-DHS-DNAseHS G1E-D-HM-H3K27ac G1E-D-HM-H3K27me3 G1E-D-HM-H3K36me3 G1E-D-HM-H3K4me1 G1E-D-HM-H3K4me3 G1E-D-PM-Pol2-4H8 G1E-D-TF-CTCF G1E-D-ZI-input G1E-E-DHS-DNAseHS G1E-E-HM-H3K27ac G1E-E-HM-H3K27me3 G1E-E-HM-H3K36me3 G1E-E-HM-H3K4me1 G1E-E-HM-H3K4me3 G1E-E-PM-Pol2-4H8 G1E-E-TF-CTCF G1E-E-ZI-input G1S-M-TF-EP300 mm9.200.G1EOnly G1E-D-DHS-DNAseHS G1E-D-HM-H3K27ac G1E-D-HM-H3K27me3 G1E-D-HM-H3K36me3 G1E-D-HM-H3K4me1 G1E-D-HM-H3K4me3 G1E-D-PM-Pol2-4H8 G1E-D-TF-CTCF G1E-D-TF-GATA1 G1E-D-TF-GATA2 G1E-D-TF-Tal1 G1E-D-ZI-input G1E-E-DHS-DNAseHS G1E-E-HM-H3K27ac G1E-E-HM-H3K27me3 G1E-E-HM-H3K36me3 G1E-E-HM-H3K4me1 G1E-E-HM-H3K4me3 G1E-E-PM-Pol2-4H8 G1E-E-TF-CTCF G1E-E-TF-GATA1 G1E-E-TF-GATA2 G1E-E-TF-Tal1 G1E-E-ZI-input G1S-M-TF-EP300

mm9.200.C2EROnly.HDPP

C2C12-D-DHS-DnaseHS C2C12-D-HM-H3K27ac C2C12-D-HM-H3K27me3 C2C12-D-HM-H3K36me3 C2C12-D-HM-H3K4me1 C2C12-D-HM-H3K4me2 C2C12-D-HM-H3K4me3 C2C12-D-PM-Pol2-4h8 C2C12-D-TF-CTCF C2C12-D-TF-EP300 C2C12-D-ZI-input C2C12-E-DHS-DnaseHS C2C12-E-HM-H3K27ac C2C12-E-HM-H3K27me3 C2C12-E-HM-H3K36me3 C2C12-E-HM-H3K4me1 C2C12-E-HM-H3K4me2 C2C12-E-HM-H3K4me3 C2C12-E-PM-Pol2-4h8 C2C12-E-TF-CTCF C2C12-E-TF-EP300 C2C12-E-ZI-input G1E-D-DHS-DNAseHS G1E-D-HM-H3K27ac G1E-D-HM-H3K27me3 G1E-D-HM-H3K36me3 G1E-D-HM-H3K4me1 G1E-D-HM-H3K4me3 G1E-D-PM-Pol2-4H8 G1E-D-TF-CTCF G1E-D-ZI-input G1E-E-DHS-DNAseHS G1E-E-HM-H3K27ac G1E-E-HM-H3K27me3 G1E-E-HM-H3K36me3 G1E-E-HM-H3K4me1 G1E-E-HM-H3K4me3 G1E-E-PM-Pol2-4H8 G1E-E-TF-CTCF G1E-E-ZI-input G1S-M-TF-EP300

mm9.200.C2EROnly

C2C12-D-DHS-DnaseHS C2C12-D-HM-H3K27ac C2C12-D-HM-H3K27me3 C2C12-D-HM-H3K36me3 C2C12-D-HM-H3K4me1 C2C12-D-HM-H3K4me2 C2C12-D-HM-H3K4me3 C2C12-D-PM-Pol2-4h8 C2C12-D-TF-CTCF C2C12-D-TF-E2A C2C12-D-TF-EP300 C2C12-D-TF-GABP C2C12-D-TF-HEB C2C12-D-TF-Mef2 C2C12-D-TF-MyoD C2C12-D-TF-myogenin C2C12-D-TF-SP1 C2C12-D-TF-SRF C2C12-D-TF-ZFP143 C2C12-D-ZI-input C2C12-E-DHS-DnaseHS C2C12-E-HM-H3K27ac C2C12-E-HM-H3K27me3 C2C12-E-HM-H3K36me3 C2C12-E-HM-H3K4me1 C2C12-E-HM-H3K4me2 C2C12-E-HM-H3K4me3 C2C12-E-PM-Pol2-4h8 C2C12-E-TF-CTCF C2C12-E-TF-E2A C2C12-E-TF-EP300 C2C12-E-TF-FOSL1 C2C12-E-TF-HEB C2C12-E-TF-Mef2 C2C12-E-TF-MyoD C2C12-E-TF-SRF C2C12-E-TF-ZFP143 C2C12-E-ZI-input G1E-D-DHS-DNAseHS G1E-D-HM-H3K27ac G1E-D-HM-H3K27me3 G1E-D-HM-H3K36me3 G1E-D-HM-H3K4me1 G1E-D-HM-H3K4me3 G1E-D-

PM-Pol2-4H8 G1E-D-TF-CTCF G1E-D-TF-GATA1 G1E-D-TF-GATA2 G1E-D-TF-Tal1 G1E-D-ZI-input G1E-E-DHS-DNAseHS G1E-E-HM-H3K27ac G1E-E-HM-H3K27me3 G1E-E-HM-H3K36me3 G1E-E-HM-H3K4me1 G1E-E-HM-H3K4me3 G1E-E-PM-Pol2-4H8 G1E-E-TF-CTCF G1E-E-TF-GATA1 G1E-E-TF-GATA2 G1E-E-TF-Tal1 G1E-E-ZI-input G1S-M-TF-EP300

Muscle (+EP300) vs ER (+EP300) vs embryo

C2C12-D-DHS-DnaseHS C2C12-D-HM-H3K27ac C2C12-D-HM-H3K27me3 C2C12-D-HM-H3K36me3 C2C12-D-HM-H3K4me1 C2C12-D-HM-H3K4me2 C2C12-D-HM-H3K4me3 C2C12-D-PM-Pol2-4h8 C2C12-D-TF-CTCF C2C12-D-TF-E2A C2C12-D-TF-EP300 C2C12-D-TF-GABP C2C12-D-TF-HEB C2C12-D-TF-Mef2 C2C12-D-TF-MyoD C2C12-D-TF-SP1 C2C12-D-TF-SRF C2C12-D-TF-ZFP143 C2C12-D-TF-myogenin C2C12-D-ZI-input C2C12-E-DHS-DnaseHS C2C12-E-HM-H3K27ac C2C12-E-HM-H3K27me3 C2C12-E-HM-H3K36me3 C2C12-E-HM-H3K4me1 C2C12-E-HM-H3K4me2 C2C12-E-HM-H3K4me3 C2C12-E-PM-Pol2-4h8 C2C12-E-TF-CTCF C2C12-E-TF-E2A C2C12-E-TF-EP300 C2C12-E-TF-FOSL1 C2C12-E-TF-HEB C2C12-E-TF-Mef2 C2C12-E-TF-MyoD C2C12-E-TF-SRF C2C12-E-TF-ZFP143 C2C12-E-ZI-input G1E-D-DHS-DNAseHS G1E-D-HM-H3K27ac G1E-D-HM-H3K27me3 G1E-D-HM-H3K36me3 G1E-D-HM-H3K4me1 G1E-D-HM-H3K4me3 G1E-D-PM-Pol2-4H8 G1E-D-TF-CTCF G1E-D-TF-GATA1 G1E-D-TF-GATA2 G1E-D-TF-Tal1 G1E-D-ZI-input G1E-E-DHS-DNAseHS G1E-E-HM-H3K27ac G1E-E-HM-H3K27me3 G1E-E-HM-H3K36me3 G1E-E-HM-H3K4me1 G1E-E-HM-H3K4me3 G1E-E-PM-Pol2-4H8 G1E-E-TF-CTCF G1E-E-TF-GATA1 G1E-E-TF-GATA2 G1E-E-TF-Tal1 G1E-E-ZI-input G1S-M-TF-EP300 H3K27ac-EmbryonicFacialProminence-E11.5 H3K27ac-EmbryonicFacialProminence-E14.5 H3K27ac-Forebrain-E11.5 H3K27ac-Forebrain-E14.5 H3K27ac-Forebrain-P0 H3K27ac-Heart-E11.5 H3K27ac-Heart-E14.5 H3K27ac-Heart-P0 H3K27ac-Hindbrain-E11.5 H3K27ac-Hindbrain-E14.5

H3K27ac-Hindbrain-P0 H3K27ac-Intestine-E14.5 H3K27ac-Intestine-P0 H3K27ac-Kidney-E14.5 H3K27ac-Kidney-P0 H3K27ac-Limb-E11.5 H3K27ac-Limb-E14.5 H3K27ac-Liver-E11.5 H3K27ac-Liver-E14.5 H3K27ac-Liver-P0 H3K27ac-Lung-E14.5 H3K27ac-Lung-P0 H3K27ac-Midbrain-E11.5 H3K27ac-Midbrain-E14.5 H3K27ac-Midbrain-P0 H3K27ac-NeuralTube-E11.5 H3K27ac-NeuralTube-E14.5 H3K27ac-Stomach-E14.5 H3K27ac-Stomach-P0 H3K27me3-EmbryonicFacialProminence-E11.5 H3K27me3-EmbryonicFacialProminence-E14.5 H3K27me3-Forebrain-E11.5 H3K27me3-Forebrain-E14.5 H3K27me3-Forebrain-P0 H3K27me3-Heart-E11.5 H3K27me3-Heart-E14.5 H3K27me3-Heart-P0 H3K27me3-Hindbrain-E11.5 H3K27me3-Hindbrain-E14.5 H3K27me3-Hindbrain-P0 H3K27me3-Intestine-E14.5 H3K27me3-Intestine-P0 H3K27me3-Kidney-E14.5 H3K27me3-Kidney-P0 H3K27me3-Limb-E11.5 H3K27me3-Limb-E14.5 H3K27me3-Liver-E11.5 H3K27me3-Liver-E14.5 H3K27me3-Liver-P0 H3K27me3-Lung-E14.5 H3K27me3-Lung-P0 H3K27me3-Midbrain-E11.5 H3K27me3-Midbrain-E14.5 H3K27me3-Midbrain-P0 H3K27me3-NeuralTube-E11.5 H3K27me3-Stomach-E14.5 H3K27me3-Stomach-P0 H3K36me3-EmbryonicFacialProminence-E11.5 H3K36me3-EmbryonicFacialProminence-E14.5 H3K36me3-Forebrain-E14.5 H3K36me3-Forebrain-P0 H3K36me3-Heart-E11.5 H3K36me3-Heart-E14.5 H3K36me3-Heart-P0 H3K36me3-Hindbrain-E11.5 H3K36me3-Hindbrain-E14.5 H3K36me3-Hindbrain-P0 H3K36me3-Intestine-E14.5 H3K36me3-Intestine-P0 H3K36me3-Kidney-E14.5 H3K36me3-Kidney-P0 H3K36me3-Limb-E14.5 H3K36me3-Liver-E11.5 H3K36me3-Liver-E14.5 H3K36me3-Liver-P0 H3K36me3-Lung-E14.5 H3K36me3-Lung-P0 H3K36me3-Midbrain-E14.5 H3K36me3-Midbrain-P0 H3K36me3-NeuralTube-E11.5 H3K36me3-NeuralTube-E14.5 H3K36me3-Stomach-E14.5 H3K36me3-Stomach-P0 H3K4me1-EmbryonicFacialProminence-E11.5 H3K4me1-EmbryonicFacialProminence-E14.5 H3K4me1-Forebrain-E11.5 H3K4me1-Forebrain-E14.5 H3K4me1-Heart-E11.5 H3K4me1-Heart-E14.5 H3K4me1-Heart-P0 H3K4me1-Hindbrain-E11.5 H3K4me1-Hindbrain-E14.5 H3K4me1-Hindbrain-P0 H3K4me1-Intestine-E14.5 H3K4me1-Intestine-P0

H3K4me1-Kidney-E14.5 H3K4me1-Kidney-P0 H3K4me1-Limb-E11.5 H3K4me1-Limb-E14.5 H3K4me1-Liver-E11.5 H3K4me1-Liver-E14.5 H3K4me1-Liver-P0 H3K4me1-Lung-E14.5 H3K4me1-Lung-P0 H3K4me1-Midbrain-E11.5 H3K4me1-Midbrain-E14.5 H3K4me1-NeuralTube-E11.5 H3K4me1-NeuralTube-E14.5 H3K4me1-Stomach-E14.5 H3K4me1-Stomach-P0 H3K4me2-EmbryonicFacialProminence-E11.5 H3K4me2-EmbryonicFacialProminence-E14.5 H3K4me2-Forebrain-E11.5 H3K4me2-Forebrain-E14.5 H3K4me2-Forebrain-P0 H3K4me2-Heart-E11.5 H3K4me2-Heart-E14.5 H3K4me2-Heart-P0 H3K4me2-Hindbrain-E11.5 H3K4me2-Hindbrain-E14.5 H3K4me2-Hindbrain-P0 H3K4me2-Intestine-E14.5 H3K4me2-Intestine-P0 H3K4me2-Kidney-E14.5 H3K4me2-Kidney-P0 H3K4me2-Limb-E11.5 H3K4me2-Limb-E14.5 H3K4me2-Liver-E11.5 H3K4me2-Liver-E14.5 H3K4me2-Liver-P0 H3K4me2-Lung-E14.5 H3K4me2-Lung-P0 H3K4me2-Midbrain-E11.5 H3K4me2-Midbrain-E14.5 H3K4me2-Midbrain-P0 H3K4me2-NeuralTube-E11.5 H3K4me2-NeuralTube-E14.5 H3K4me2-Stomach-E14.5 H3K4me2-Stomach-P0 H3K4me3-EmbryonicFacialProminence-E11.5 H3K4me3-EmbryonicFacialProminence-E14.5 H3K4me3-Forebrain-E11.5 H3K4me3-Forebrain-E14.5 H3K4me3-Forebrain-P0 H3K4me3-Heart-E11.5 H3K4me3-Heart-E14.5 H3K4me3-Heart-P0 H3K4me3-Hindbrain-E11.5 H3K4me3-Hindbrain-E14.5 H3K4me3-Hindbrain-P0 H3K4me3-Intestine-E14.5 H3K4me3-Intestine-P0 H3K4me3-Kidney-E14.5 H3K4me3-Kidney-P0 H3K4me3-Limb-E11.5 H3K4me3-Limb-E14.5 H3K4me3-Liver-E11.5 H3K4me3-Liver-E14.5 H3K4me3-Liver-P0 H3K4me3-Lung-E14.5 H3K4me3-Lung-P0 H3K4me3-Midbrain-E11.5 H3K4me3-Midbrain-E14.5 H3K4me3-Midbrain-P0 H3K4me3-NeuralTube-E11.5 H3K4me3-Stomach-E14.5 H3K4me3-Stomach-P0 H3K9ac-EmbryonicFacialProminence-E11.5 H3K9ac-EmbryonicFacialProminence-E14.5 H3K9ac-Forebrain-E11.5 H3K9ac-Forebrain-E14.5 H3K9ac-Forebrain-P0 H3K9ac-Heart-E11.5 H3K9ac-Heart-E14.5 H3K9ac-Heart-P0 H3K9ac-Hindbrain-E11.5 H3K9ac-Hindbrain-E14.5 H3K9ac-Hindbrain-P0 H3K9ac-Intestine-E14.5 H3K9ac-Intestine-P0 H3K9ac-Kidney-E14.5 H3K9ac-Kidney-P0 H3K9ac-Limb-E11.5 H3K9ac-Limb-

E14.5 H3K9ac-Liver-E11.5 H3K9ac-Liver-E14.5 H3K9ac-Liver-P0 H3K9ac-Lung-E14.5 H3K9ac-Lung-P0 H3K9ac-Midbrain-E11.5 H3K9ac-Midbrain-E14.5 H3K9ac-Midbrain-P0 H3K9ac-NeuralTube-E11.5 H3K9ac-NeuralTube-E14.5 H3K9ac-Stomach-E14.5 H3K9ac-Stomach-P0 H3K9me3-EmbryonicFacialProminence-E11.5 H3K9me3-Forebrain-P0 H3K9me3-Heart-E11.5 H3K9me3-Hindbrain-E11.5 H3K9me3-Limb-E11.5 H3K9me3-Liver-E11.5 H3K9me3-Midbrain-E11.5 H3K9me3-NeuralTube-E11.5 Input-EmbryonicFacialProminence-E11.5 Input-EmbryonicFacialProminence-E14.5 Input-Forebrain-E11.5 Input-Forebrain-E14.5 Input-Forebrain-P0 Input-Heart-E11.5 Input-Heart-E14.5 Input-Heart-P0 Input-Hindbrain-E11.5 Input-Hindbrain-E14.5 Input-Hindbrain-P0 Input-Intestine-E14.5 Input-Intestine-P0 Input-Kidney-E14.5 Input-Kidney-P0 Input-Limb-E11.5 Input-Limb-E14.5 Input-Liver-E11.5 Input-Liver-E14.5 Input-Liver-P0 Input-Lung-E14.5 Input-Lung-P0 Input-Midbrain-E11.5 Input-Midbrain-E14.5 Input-Midbrain-P0 Input-NeuralTube-E11.5 Input-NeuralTube-E14.5 Input-Stomach-E14.5 Input-Stomach-P0

Appendix E

MUSCLE-SPECIFIC ENHANCER DATABASE

These enhancers were drawn from a resource which I recapitulate here. This resource was a muscle specific database providing a catalog of regulatory elements that was up-kept by James W. Fickett is no longer available on the web.

(Adolph EA, 1993; Amacher S, 1993; Andres V, 1995; Argenin S, 1994; Asakura A, 1993; Baldwin TJ, 1989; Baldwin TJ, 1988; BanerjeeBasu S, 1993; Barbieri G, 1990; Barnea E, 1990; BasselDuby R, 1993; BasselDuby R, 1992; Bauvagnet PF, 1987; Berberich C, 1993; Bergsma DJ, 1986; Bessereau JL, 1993; Biben C, 1994; Bishoprie NH, 1992; Black BL, 1995; Blanchetot N, 1986; Boheler KR, 1992; Bonne G, 1993; Boxer LM, 1989; Brennan TJ, 1990; Buchberger A, 1994; Bucher EA, 1988; Buonanno A, 1993; Buskin JN, 1989; Capetanaki Y, 1989b; Capetanaki Y, 1989a; Carnac G, 1993; Carroll SL, 1988; Catala F, 1995; Chakraborty T, 1991; Cheng TS, 1992; Cheng TS, 1993; Chow KL, 1990; Christensen TH, 1993; Chung AB, 1992; Cogan JG, 1995; Collo G, 1993; Cooper TA, 1985; Corin SJ, 1994; Cribbs LL, 1989; Cserjesi P, 1992; Cserjesi P, 1994; Danilition SL, 1991; Daubas P, 1988; Davey HW, 1995; Dechesne CA, 1994; Deschamps J, 1985; Devlin BH, 1989; Dickson G, 1988; Donoghue M, 1988; Donoghue MJ, 1991; Duan C, 1996; Duclert A, 1993; Dunwoodie SL, 1994; Dürr I, 1994; Dutton EK, 1993; Edmondson DG, 1992; Edwards JG, 1992; Edwards JG, 1994; Eftimie R, 1991; Ernst H, 1991; Essig DA, 1991; Ewart GD, 1991; Ewton DZ, 1995; Fabrizi GM, 1992; Farrell FX, 1990; Fisch TM, 1987; Flink IL, 1990; Flink IL, 1992; Foster DN, 1992; French BA, 1991; FujisawaSehara A, 1991; FujisawaSehara A, 1992; Gardner DG, 1989; Gardner DG, 1988; Garzon RJ, 1994; Getman DK, 1995; Giallongo A, 1993; Gilgenkrantz H, 1992; Gilmour BP, 1991; Gilmour BP, 1995;

Glembotski CC, 1993; Goldhamer DJ, 1995; Goldhamer DJ, 1992; Gorski DH, 1993; Gossett LA, 1989; Graber SG, 1986; Grayson J, 1995; Greenberg ME, 1987; Greenberg ME, 1986; Grichnik JM, 1987; Grichnik JM, 1988; Guicherit OM, 1991; Gupta MP, 1994; Gupta MP, 1996; Gustafson TA, 1987; Han VKM, 1996; Hashimoto N, 1995; Hayes TE, 1987; Hidaka RA, 1993; Horlick RA, 1989; Horlick RA, 1990; Houzelstein D, 1992; Huang CF, 1994; Huang WY, 1997; Iannello RC, 1991; Ip HS, 1994; Izumo S, 1986; James PL, 1993; Jaynes JB, 1986; Jaynes JB, 1988; Joh K, 1991; Johnson JE, 1992; Johnson JE, 1989; Kariya K, 1993; Kariya K, 1994; Karns LR, 1995; Kawamoto T, 1988; Keller A, 1995; Kelly R, 1995; Klamut HJ, 1990; Klarsfeld A, 1987; Klip A, 1994; Knotts S, 1994; Kou K, 1995; KovacicMilivojevic G, 1993; Kovacs AM, 1993; Lamandé N, 1995; Lamandé N, 1989; LaPointe MC, 1988; Lassar AB, 1989; LeeT.C. ChowKL, 1991; Lee TC, 1994; Lee Y, 1997; Leibham D, 1994; Lenka N, 1996; LePage DF, 1994; Leung E, 1993; Lev AA, 1987; Lewis AL, 1996; Li H, 1993; Li H, 1994a; Li H, 1994b; Li K, 1990; Li Z, 1991; Li Z, 1993; Li Z, 1989; Lilienbaum A, 1990; Liu ML, 1994; Liu M, 1994; Liu Z, 1991; Lohse P, 1988; Lompre AM, 1984; Long SD, 1996; Lucas M, 1992; Lyons GE, 1990; Mably JD, 1993; MacLellan WR, 1994; Mahdavi V, 1982; Maire P, 1987; Malik S, 1995; Mar JH, 1986; Mar JH, 1988; Martin KA, 1994; Matsumoto T, 1996; McHugh KM, 1988; McNamara CA, 1995; Mendelzon D, 1994; Merlie PJ, 1989; Merlie PJ, 1994; Mesnard L, 1993; Min BH, 1990; Minty A, 1986; Mitsumoto Y, 1994; Miwa T, 1987; Miwa T, 1991; MohunT. G, 1987; Mohun TJ, 1986; Mohun TJ, 1989; Molkentin JD, 1993; Molkentin JD, 1994a; Molkentin JD, 1996; Molkentin JD, 1994b; Morkin E, 1993; Moss JB, 1994; Moss JB, 1996; MouraNeto V, 1996; Murre C, 1989; Muscat GEO, 1987; Mutero A, 1995; Naidu PS, 1995; Nakamura N, 1995; Nemer M, 1986; Noursadeghi M, 1993; Nudel U, 1988; Numberger M, 1991; Ohshima Y, 1989; Ojamaa K, 1996; Olva D, 1995; Pari G, 1991; Pariasamy M, 1989; Parker TG, 1992; Parmacek MS, 1994;

Parmacek MS, 1992; Parsons WJ, 1993; PedrazaAlva G, 1994; Periasamy M, 1985; Perkins EB, 1995; Peshavaria M, 1991; Peterson CA, 1992; PhanDinhTuy F, 1988; Pieper FR, 1988; Pieper FR, 1987; Pieper FR, 1992; Piette J, 1989; Piette nan, 1990; Piette J, 1992; Prody CA, 1991; Prody CA, 1992; Quax W, 1985; Quax W, 1984; Quitschke WW, 1989a; Quitschke WW, 1989b; Rao MV, 1996; Renz M, 1985; Rindt H, 1993; Rindt H, 1995; Rittling SR, 1989; Rosenthal N, 1990; Rosenthal N, 1989; Rotwein P, 1995; Ryan KJ, 1996; Saggin L, 1990; Sakimura K, 1990; Salminen A, 1991; Salminen M, 1996; Salminen M, 1994; Salvetti A, 1993; Santoro IM, 1991; Sartorelli V, 1990; Sartorelli V, 1992; SassoneCorsi P, 1987; Sawtell NM, 1989; Sax CM, 1989; Sax CM, 1988; Sax CM, 1993; Schlerf A, 1988; Schreier T, 1990; Seidel U, 1989; Seidman DE, 1987; Shillace R, 1994; Shimizu N, n.d.[b]; Shimizu M, 1993; Shimizu N, n.d.[a]; Simon AM, 1993; Skerjanc IS, 1994; Smith EO, 1993; Sommers CL, 1994; Song WK, 1992; Sprenkle AB, 1995; Sternberg EA, 1988; Stover DM, 1992; Su CT, 1995; Subramaniam A, 1990; Sunyer T, 1993; Swiderski RE, 1990; Sympson CJ, 1993; Szucsik JC, 1995; Talib S, 1993; Tanaka M, 1985; Tapscott SJ, 1992; Taylor MV, 1991; Taylor JM, 1995; Taylor MV, 1989; Tebbey PS, 1994; Thompson WR, 1991; Trask RV, 1988; Treisman R, 1986; Trouche D, 1993; Trouche D, 1995; Underwood LE, 1987; FA, 1992; FA, 1994; JJM, 1994; Vosberg HP, 1992; Walke W, 1994; Walsh K, 1988; Walsh K, 1989; Wan B, 1995; Wang G, 1994; Wang XM, 1990; Wang Y, 1988; Watanabe T, 1997; Wefald FC, 1990; Weintraub H, 1990; Weintraub H, 1991; Weller PA, 1986; Wentworth BM, 1991; Wu J, 1989; Yano H, 1994; Yano H, 1995; Yao CC, 1996a; Yao CC, 1996b; Yee SP, 1993; Yi TM, 1991; Yutzey KE, 1990; Zambetti GP, 1992; Zhao J, 1994; Zhu L, 1995; Ziober BL, 1996; Ziober BL, 1993)

References

Adolph EA, et al. (1993). "Role of myocyte-specific enhancer-binding factor." In: *J. Biol. Chem.* 25, pp. 5349–52.

Amacher S, et al. (1993). "Multiple regulatory elements contribute differentially to muscle creatine kinase enhancer activity in skeletal and cardiac muscle." In: *Mol. Cell Biol.* 13, pp. 2753–64.

Andres V, et al. (1995). "Regulation of GAX homeobox gene transcription by a combination of positive factors including myocyte-specific enhancer factor 2." In: *Mol. Cell Biol.* 15, pp. 4272–81.

Argenin S, et al. (1994). "Developmental stage-specific regulation of atrial natriuretic factor gene transcription in cardiac cells." In: *Mol. Cell Biol.* 14, pp. 777–90.

Asakura A, et al. (1993). "MyoD and myogenin act on the chicken myosin light-chain 1 gene as distinct transcriptional factors." In: *Mol. Cell Biol.* 13, pp. 7153–62.

Baldwin TJ, et al. (1988). "Isolation and characterization of the mouse acetylcholine receptor delta-subunit gene: Identification of a 148 bp cis-acting region that confers myotube-specific expression." In: *J. Cell Biol* 107, pp. 2271–9.

– (1989). "Muscle-specific gene expression controlled by a regulatory element lacking a MyoD1-binding site." In: *Nature* 341, pp. 716–20.

BanerjeeBasu S, et al. (1993). "Cis-acting sequences of the rat troponin I slow gene confer tissue-and development-specific transcription in cultured muscle cells as well as fiber type specificity in transgenic mice." In: *Mol. Cell Biol.* 13, pp. 7019–28.

Barbieri G, et al. (1990). "Differential expression of muscle-specific enolase in embryonic and fetal myogenic cells during mouse development." In: *Differentiation* 45, pp. 179–84.

Barnea E, et al. (1990). "Specificity of expression of the muscle and brain dystrophin gene promoters in muscle and brain cells." In: *Neuron* 5, pp. 881–8.

BasselDuby R, et al. (1992). "A 40-Kilodalton Protein Binds Specifically to an Upstream Sequence Element Essential for Muscle-Specific Transcription of the Human Myoglobin Promoter." In: *Mol. Cell Biol.* 12, pp. 5024–32.

– (1993). "Sequence Elements Required for Transcriptional Activity of the Human Myoglobin Promoter in Intact Myocardium." In: *Circ. Res.* 73, pp. 360–6.

Bauvagnet PF, et al. (1987). "Multiple positive and negative 5' regulatory elements control the cell-type-specific expression of the embryonic skeletal myosin heavy-chain gene." In: *Mol. Cell Biol.* 7, pp. 4377–89.

Berberich C, et al. (1993). "Two adjacent E-box elements and a M-CAT box are involved in the muscle-specific regulation of the rat acetylcholine receptor beta subunit gene." In: *Eur. J. Biochem.* 216, pp. 395–404.

Bergsma DJ, et al. (1986). "Delimitization and characterization of cis-acting DNA sequences required for the regulated expression and transcriptional control of the chicken skeletal alpha-actin gene." In: *Mol. Cell Biol.* 6, pp. 2462–75.

Bessereau JL, et al. (1993). "Muscle-specific expression of the acetylcholine receptor alpha subunit gene requires both positive and negative interactions between myogenic factors, Sp1 and GBF factors." In: *EMBO J* 12, pp. 443–49.

Biben C, et al. (1994). "Novel muscle-specific enhancer sequences upstream of the cardiac actin gene." In: *Mol. Cell Biol.* 14, pp. 3504–13.

Bishoprie NH, et al. (1992). "Positive regulation of the skeletal a-actin gene by fos and jun in cardiac myocytes." In: *J. Biol. Chem.* 267, pp. 25535–40.

Black BL, et al. (1995). "The mouse MRF4 promoter is trans-activated directly and indirectly by muscle-specific transcription factors." In: *J. Biol. Chem.* 270, pp. 2889–2890.

Blanchetot N, et al. (1986). "The mouse myoglobin gene: Characterization and sequence comparison with other mammalian myoglobin genes." In: *Eur. J. Biochem.* 159, pp. 469–74.

Boheler KR, et al. (1992). "Cardiac expression of a- and b-myosin heavy chains and sarcomeric a-actins are regulated through transcriptional mechanisms." In: *J. Biol. Chem.* 267, pp. 12979–85.

Bonne G, et al. (1993). "Expression of human cytochrome c oxidase subunits during fetal development." In: *Eur. J. Biochem.* 217, pp. 1099–1107.

Boxer LM, et al. (1989). "The sarcomeric actin CArG-binding factor is indistinguishable from the c-fos serum response factor." In: *Mol. Cell Biol.* 9, pp. 515–22.

Brennan TJ, et al. (1990). "Myogenin resides in the nucleus and acquires high affinity for a conserved enhancer element on heterodimerization." In: *Genes Dev* 4, pp. 582–95.

Buchberger A, et al. (1994). "The myogenin gene is activated during myocyte differentiation by pre-existing, not newly synthesized transcription factor MEF2." In: *J. Biol. Chem.* 269, pp. 17289–96.

Bucher EA, et al. (1988). "Expression of the troponin complex genes: transcriptional coactivation during myoblast differentiation and independent control in heart and skeletal muscles." In: *Mol. Cell Biol.* 8, pp. 4134–42.

Buonanno A, et al. (1993). "Upstream sequences of the myogenin gene convey responsiveness to skeletal muscle denervation in transgenic mice." In: *Nucl. Acids Res* 21, pp. 5684–93.

Buskin JN, et al. (1989). "Identification of a myocyte nuclear factor that binds to the muscle-specific enhancer of the mouse muscle creatine kinase gene." In: *Mol. Cell Biol.* 9, pp. 2627–40.

Capetanaki Y, et al. (1989a). “Expression of the chicken vimentin gene in transgenic mice: efficient assembly of the avian protein into the cytoskeleton.” In: *PNAS* 86, pp. 4882–6.

– (1989b). “Overexpression of the vimentin gene in transgenic mice inhibits normal lens cell differentiation.” In: *J. Biol. Chem.* 109, pp. 1653–64.

Carnac G, et al. (1993). “9-cis-retinoic acid regulates the expression of the muscle determination gene Myf5.” In: *Endocrinology* 133, pp. 2171–6.

Carroll SL, et al. (1988). “A 29-nucleotide DNA segment containing an evolutionarily conserved motif is required in cis for cell-type-restricted repression of the chicken alpha-smooth muscle actin gene core promoter.” In: *Mol. Cell Biol.* 8, pp. 241–50.

Catala F, et al. (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.” In: *Mol. Cell Biol.* 15, pp. 4585–96.

Catarino RR, et al. (2018). “Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation.” In: *Genes and Dev.* 32, pp. 202–223.

Chakraborty T, et al. (1991). “Differential trans-activation of a muscle-specific enhancer by myogenic helix-loop-helix proteins is separable from DNA binding.” In: *J. Biol. Chem.* 266, pp. 2878–82.

Cheng TS, et al. (1992). “Mapping of myogenin transcription during embryogenesis using transgenes linked to the myogenin control region.” In: *J. Cell Biol* 119, pp. 1649–56.

– (1993). “Separable regulatory elements governing myogenin transcription in mouse embryogenesis.” In: *Science* 261, pp. 215–18.

Chow KL, et al. (1990). “A combination of closely associated positive and negative cis-acting promoter elements regulates transcription of the skeletal alpha-actin gene.” In: *Mol. Cell Biol.* 10, pp. 528–38.

Christensen TH, et al. (1993). “Regulation of the human cardiac/slow-twitch troponin C gene by multiple, cooperative, cell-type-specific, and MyoD-responsive elements.” In: *Mol. Cell Biol.* 13, pp. 6752–65.

Chung AB, et al. (1992). “Transcriptional control of nuclear genes for the mitochondrial muscle ADP/ATP translocator and the ATP synthase beta subunit.” In: *J. Biol. Chem.* 276, pp. 21154–61.

Cogan JG, et al. (1995). “Plasticity of vascular smooth muscle alpha-actin gene transcription.” In: *J. Biol. Chem.* 270, pp. 11310–21.

Collo G, et al. (1993). “A new isoform of the laminin receptor integrin alpha-7-beta-1 is developmentally regulated in skeletal muscle.” In: *J. Biol. Chem.* 268, pp. 19019–24.

Cooper TA, et al. (1985). “A single cardiac troponin T gene generates embryonic and adult isoforms via developmentally regulated alternate splicing.” In: *J. Biol. Chem.* 260, pp. 11140–11148.

Corin SJ, et al. (1994). “Structure and expression of the human slow twitch skeletal muscle troponin I gene.” In: *J. Biol. Chem.* 269, pp. 10651–9.

Cribbs LL, et al. (1989). “Muscle-specific regulation of a transfected rabbit myosin heavy chain b gene promoter.” In: *J. Biol. Chem.* 264, pp. 10672–8.

Cserjesi P, et al. (1992). “MHox: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creative kinase enhancer.” In: *Development* 115, pp. 1087–1101.

– (1994). “Homeodomain protein MHox and MADS protein myocyte enhancer-binding factor-2 converge on a common element in the muscle creatine kinase enhancer.” In: *J. Biol. Chem.* 269, pp. 16740–45.

Danilition SL, et al. (1991). “Transcription factor binding and spacing constraints in the human beta-actin proximal promoter.” In: *Nucl. Acids Res* 19, pp. 6913–22.

Daubas P, et al. (1988). “Functional activity of the two promoters of the myosin alkali light chain gene in primary muscle cell cultures: comparison with other muscle gene promoters and other culture systems.” In: *Nucl. Acids Res* 16, pp. 1251–71.

Davey HW, et al. (1995). “The nucleotide sequence, structure, and preliminary studies on the transcriptional regulation of the bovine alpha skeletal actin gene.” In: *DNA Cell Biol* 14, pp. 609–18.

Dechesne CA, et al. (1994). “E-Box- and MEF-2-Independent Muscle-Specific Expression, Positive Autoregulation, and Cross-Activation of the Chicken MyoD.” In: *Mol Cell Biol.* 14, pp. 5474–86.

Deschamps J, et al. (1985). “Identification of a transcriptional enhancer element upstream from the proto-oncogene fos.” In: *Science* 230, pp. 1174–7.

Devlin BH, et al. (1989). “Identification of a Muscle-Specific Enhancer Within the 5‘-flanking Region of the Human Myoglobin Gene.” In: *J. Biol. Chem.* 264, pp. 13896–901.

Dickson G, et al. (1988). “Distinct dystrophin mRNA species are expressed in embryonic and adult mouse skeletal muscle.” In: *FEBS* 242, pp. 47–52.

Donoghue M, et al. (1988). “A muscle-specific enhancer is located at the 3’ end of the myosin light-chain 1/3 gene locus.” In: *Genes Dev* 2, pp. 1779–90.

Donoghue MJ, et al. (1991). “Fiber type- and position-dependent expression of a myosin light chain-CAT transgene detected with a novel histochemical stain for CAT.” In: *J. Cell Biol* 115, pp. 423–34.

Duan C, et al. (1996). “Insulin-like Growth Factor-I.” In: *J. of Endocrynology* 271, pp. 4280–8.

Duclert A, et al. (1993). “An 83-nucleotide promoter of the acetylcholine receptor epsilon-subunit gene confers preferential synaptic expression in mouse muscle.” In: *PNAS* 90, pp. 3043–7.

Dunwoodie SL, et al. (1994). “Multiple regions of the human cardiac actin gene are necessary for maturation-based expression in striated muscle.” In: *J. Biol. Chem.* 269, pp. 12212–9.

Dürr I, et al. (1994). “Characterization of the functional role of E-box elements for the transcriptional activity of rat acetylcholine receptor epsilon-subunit and gamma-subunit gene promoters in primary muscle cell cultures.” In: *Eur. J. Biochem.* 224, pp. 353–64.

Dutton EK, et al. (1993). “Electrical activity-dependent regulation of the acetylcholine receptor delta-subunit gene, MyoD, and myogenin in primary myotubes.” In: *PNAS* 90, pp. 2040–4.

Edmondson DG, et al. (1992). “Analysis of the myogenin promoter reveals an indirect pathway for positive autoregulation mediated by the muscle-specific enhancer factor MEF-2.” In: *Eur. J. Biochem.* 12, pp. 3665–77.

Edwards JG, et al. (1992). “A repressor region in the human b-myosin heavy chain gene that has a partial position dependency.” In: *Biochem Phys Res.* 189, pp. 504–11.

– (1994). “Thyroid hormone influences beta myosin heavy chain.” In: *Biochem Biophys Res Commun* 199, pp. 1482–8.

Eftimie R, et al. (1991). “Myogenin and MyoD join a family of skeletal muscle genes regulated by electrical activity.” In: *PNAS* 88, pp. 1349–53.

Ernst H, et al. (1991). “The myosin light chain enhancer and the skeletal actin promoter share a binding site for factors involved in muscle-specific gene expression.” In: *Mol Cell Biol.* 11, pp. 3735–44.

Essig DA, et al. (1991). “Expression of embryonic myosin heavy chain mRNA in stretched adult chicken skeletal muscle.” In: *Am. J. Physiol.* 260, pp. C1325–31.

Ewart GD, et al. (1991). “Switching of bovine cytochrome c oxidase subunit VIa isoforms in skeletal muscle during development.” In: *FEBS* 292, pp. 79–84.

Ewton DZ, et al. (1995). “IGF Binding Proteins-4, -5, and -6 May Play Speciialized Roles During L6 Myoblast Proliferation and Differentiation.” In: *J. Endocrin.* 144, pp. 539–53.

FA, et al. vandeKlundert (1992). “Identification of two silencers flanking an AP-1 enhancer in the vimentin promoter.” In: *Gene* 122, pp. 337–43.

– (1994). “A proximal promoter element in the hamster desmin upstream regulatory region is responsible for activation by myogenic determination factors.” In: *J. Biol. Chem.* 269, pp. 220–5.

Fabrizi GM, et al. (1992). “Differential expression of genes specifying two isoforms of subunit VIa of human cytochrome c oxidase.” In: *Gene* 119, pp. 307–12.

Farrell FX, et al. (1990). “A negative element involved in vimentin gene expression.” In: *Mol. Cell Biol.* 10, pp. 2349–58.

Fisch TM, et al. (1987). “c-fos sequences necessary for basal expression and induction by epidermal growth factor, 12 -O-Tetradecanoyl phorlol-13-acetate, and the calcium ionophore.” In: *Mol. Cell Biol.* 7, pp. 3490–3502.

Flink IL, et al. (1990). “Interaction of thyroid hormone receptors with strong and weakcis-acting elements in the human a-myosin heavy chain gene promoter.” In: *J. Biol. Chem.* 265, pp. 11233–7.

– (1992). “Characterization of a strong positivecis-acting element of the human b-myosin heavy chain gene in fetal rat heart cells.” In: *J. Biol. Chem.* 267, pp. 9917–24.

Foster DN, et al. (1992). “Positive and negative cis-acting regulatory elements mediate expression of the mouse vascular smooth muscle alpha-actin gene.” In: *J. Biol. Chem.* 267, pp. 11995–12003.

French BA, et al. (1991). “Heterodimers of myogenic helix-loop-helix regulatory factors and E12 bind a complex element governing myogenic induction of the avian cardiac alpha-actin promoter.” In: *Mol. Cell Biol.* 11, pp. 2439–50.

FujisawaSehara A, et al. (1991). “Upstream region of the myogenin gene confers transcriptional activation in muscle cell lineages during mouse embryogenesis.” In: *Biochem* 191, pp. 351–6.

– (1992). “Differential trans-activation of muscle-specific regulatory elements including the myosin light chain box by chicken MyoD, myogenin, and MRF4.” In: *J. Biol. Chem.* 267, pp. 10031–8.

Gardner DG, et al. (1988). “Expression of the gene for the atrial natriuretic peptide in cardiac myocytes in vitro.” In: *Drugs and Therapy* 2, pp. 479–86.

– (1989). “Expression of the atrial natriuretic peptide gene in human fetal heart.” In: *J. Clin. Endocrinol.* 69, pp. 729–37.

Garzon RJ, et al. (1994). “Multiple silencer elements are involved in regulating the chicken vimentin gene.” In: *Mol. Cell Biol.* 14, pp. 934–43.

Gasper WC Marinov GK, et al. (2014). “Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: identifying ChIP-quality p300 monoclonal antibodies.” In: *Sci Rep.* 4, p. 5152.

Getman DK, et al. (1995). “Transcription factor repression and activation of the human acetylcholinesterase gene.” In: *J. Biol. Chem.* 270, pp. 23511–9.

Giallongo A, et al. (1993). “Structural features of the human gene for muscle-specific enolase: differential splicing in the 5’-untranslated sequence generates two forms of mRNA.” In: *Eur. J. Biochem.* 214, pp. 367–74.

Gilgenkrantz H, et al. (1992). “Positive and negative regulatory DNA elements including a CCArGG box are involved in the cell type-specific expression of the human muscle dystrophin gene.” In: *J. Biol. Chem.* 267, pp. 10823–30.

Gilmour BP, et al. (1991). “Multiple binding sites for myogenic regulatory factors are required for expression of the acetylcholine receptor gamma-subunit gene.” In: *J. Biol. Chem.* 266, pp. 19871–4.

– (1995). “Electrical activity suppresses nicotinic acetylcholine receptor gamma subunit promoter activity.” In: *Dev. Biol* 168, pp. 416–28.

Glembotski CC, et al. (1993). “Myocardial alpha-thrombin receptor activation induces hypertrophy and increases atrial natriuretic factor gene expression.” In: *J. Biol. Chem.* 268, pp. 20646–52.

Goldhamer DJ, et al. (1992). “Regulatory elements that control the lineage-specific expression of myoD.” In: *Science* 256, pp. 538–41.

– (1995). “Embryonic activation of the myoD gene is regulated by a highly conserved distal control element.” In: *Development* 121, pp. 637–49.

Gorski DH, et al. (1993). “Molecular cloning of a diverged homeobox gene that is rapidly down-regulated during the G0/G1 transition in vascular smooth muscle cells.” In: *Mol. Cell Biol.* 13, pp. 3722–33.

Gossett LA, et al. (1989). “A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes.” In: *Mol. Cell Biol.* 9, pp. 5022–33.

Graber SG, et al. (1986). “Myoglobin Expression in L6 Muscle Cells.” In: *J. Biol. Chem.* 261, pp. 9150–4.

Grayson J, et al. (1995). “Synergistic Interactions Between Heterologous Upstream Activation Elements and Specific TATA Sequences in a Muscle-Specific Promoter.” In: *Mol. Cell Biol.* 15, pp. 1870–8.

Greenberg ME, et al. (1986). “Effect of protein synthesis inhibitors on growth factor activation of c-fos, c-myc, and actin gene transcription.” In: *Mol. Cell Biol.* 6, pp. 1050–7.

– (1987). “Mutation of the c-fos gene dyad symmetry element inhibits serum inducibility of transcription in vivo and the nuclear regulatory factor binding in vitro.” In: *Mol. Cell Biol.* 7, pp. 1217–25.

Grichnik JM, et al. (1987). “Tissue restricted and stage specific transcription is maintained within 411 nucleotides flanking the 5’ end of the chicken alpha-skeletal actin gene.” In: *Nucl. Acids Res* 14, pp. 2956–67.

– (1988). “The chicken skeletal alpha-actin gene promoter region exhibits partial dyad symmetry and a capacity to drive bidirectional transcription.” In: *Mol. Cell Biol.* 8, pp. 4587–97.

Guicherit OM, et al. (1991). "Molecular cloning and expression of a mouse muscle cDNA encoding adenylosuccinate synthetase." In: *J. Biol. Chem.* 266, pp. 22582–7.

Gupta MP, et al. (1994). "An E-box/M-CAT hybrid motif and cognate binding protein." In: *J. Biol. Chem.* 269, pp. 29677–29687.

– (1996). "Sympathetic contron of cardia myosin heavy chain gene expression." In: *Mol. Cell Biol.* 157, pp. 117–24.

Gustafson TA, et al. (1987). "Hormonal regulation of myosin heavy chain and a-actin gene expression in cultured fetal rat heart myocytes." In: *J. Biol. Chem.* 262, pp. 13316–22.

Han VKM, et al. (1996). "IGF-Binding Protein mRNAs in the Human Fetus: Tissue and Cellular Distribution of Developmental Expression." In: *Horm. Res.* 45, pp. 160–6.

Hashimoto N, et al. (1995). "Role of tyrosine kinase in the regulation of myogenin expression." In: *Eur J Biochem* 227, pp. 379–87.

Hayes TE, et al. (1987). "Inducible binding of a factor to the c-fos regulatory region." In: *PNAS* 84, pp. 1272–76.

Hidaka RA, et al. (1993). "The MEF-3 motif is required for a MEF-2-mediated skeletal muscle-specific induction of the rat aldolase A gene." In: *Mol. Cell Biol.* 13, pp. 6469–78.

Horlick RA, et al. (1989). "The upstream muscle-specific enhancer of the rat muscle creatine kinase gene is composed of multiple elements." In: *Mol. Cell Biol.* 9, pp. 2396–2413.

– (1990). "Brain and muscle creatine kinase genes contain common AT-rich recognition protein-binding regulatory elements." In: *Mol. Cell Biol.* 10, pp. 4826–36.

Houzelstein D, et al. (1992). "Localization of dystrophin gene transcripts during mouse embryogenesis." In: *J. Biol. Chem.* 119, pp. 811–21.

Huang CF, et al. (1994). "Rapid inhibition of myogenin-driven acetylcholine receptor subunit gene transcription." In: *EMBO J* 13, pp. 634–40.

Huang WY, et al. (1997). "Multiple muscle-specific regulatory elements are associated with a DNase 1 hypersensitive site of the cardiac b-myosin heavy-chain gene." In: *Biochem. J.* 327, pp. 507–12.

Iannello RC, et al. (1991). "Characterization of a promoter element required for transcription in myocardial cells." In: *J. Biol. Chem.* 266, pp. 3309–16.

Ip HS, et al. (1994). "The GATA-4 transcription factor transactivates the cardiac muscle-specific troponin C promoter-enhancer in nonmuscle cells." In: *Mol. Cell Biol.* 14, pp. 7517–26.

Izumo S, et al. (1986). “All members of the MHC multigene family respond to thyroid hormone in a highly tissue-specific manner.” In: *Science* 231, pp. 597–600.

James PL, et al. (1993). “A Highly Conserved Insulin-like Growth Factor-binding Protein.” In: *J. Biol. Chem.* 268, pp. 22305–12.

Jaynes JB, et al. (1986). “Transcriptional regulation of the muscle creatine kinase gene and regulated expression in transfected mouse myoblasts.” In: *Mol. Cell Biol.* 6, pp. 2855–64.

– (1988). “The muscle creatine kinase gene is regulated by multiple upstream elements, including a muscle-specific enhancer.” In: *Mol. Cell Biol.* 8, pp. 62–70.

JJM, et al. vanGroningen (1994). “Rat desmin gene structure and expression.” In: *Biochem Biophys Res.* 1217, pp. 107–9.

Joh K, et al. (1991). “Analysis of upstream regulatory regions required for the activities of two promoters of the rat aldolase A gene.” In: *FEBS* 292, pp. 128–32.

Johnson JE, et al. (1989). “Muscle creatine kinase sequence elements regulating skeletal and cardiac muscle expression in transgenic mice.” In: *Mol. Cell Biol.* 9, pp. 3393–9.

– (1992). “DNA binding and transcriptional regulatory activity of mammalian achaete-scute homologous.” In: *PNAS* 89, pp. 3596–600.

Kariya K, et al. (1993). “Transcriptional enhancer factor-1 in cardiac myocytes interacts with an a1-Adrenergic- and b-protein kinace C-inducible element in the rat b-myosin heavy chain promoter.” In: *J. Biol. Chem.* 268, pp. 26658–62.

– (1994). “An enhancer core element mediates stimulation of the rat b-myosin heavy chain promoter by an a1-adrenergic agonist and activated b-protein kinase C in hypertrophy of cardiac myocytes.” In: *J. Biol. Chem.* 269, pp. 3775–82.

Karns LR, et al. (1995). “M-CAT, CArG, and Sp1 elements are required for alpha1-adrenergic induction of the skeletal alpha-actin promoter during cardiac myocyte hypertrophy.” In: *J. Biol. Chem.* 270, pp. 410–7.

Kawamoto T, et al. (1988). “Identification of the Human beta-Actin Enhancer and Its Binding Factor.” In: *Mol. Cell Biol.* 8, pp. 267–72.

Keller A, et al. (1995). “Differential expression of a- and b-enolase genes during rat heart development and hypertrophy.” In: *Am. J. Physiol.* 269, pp. 1843–51.

Kelly R, et al. (1995). “Myosin light chain 3F regulatory sequences confer region-alized cardiac and skeletal muscle expression in transgenic mice.” In: *J. Biol. Chem.* 129, pp. 383–96.

Klamut HJ, et al. (1990). “Molecular and functional analysis of the muscle-specific promoter region of the Duchenne Muscular Dystrophy gene.” In: *Mol. Cell Biol.* 10, pp. 193–205.

Klarsfeld A, et al. (1987). “A 5’-flanking region of the chicken acetylcholine receptor alpha-subunit gene confers tissue specificity and developmental control of expression in transfected cells.” In: *Mol. Cell Biol.* 7, pp. 951–5.

Klip A, et al. (1994). “Regulation of expression of glucose transporters by glucose: a review of studiesin vivoand in cell cultures.” In: *FASEB J* 8, pp. 43–53.

Knotts S, et al. (1994). “In vivoregulation of the mouse b myosin heavy chain gene.” In: *J. Biol. Chem.* 269, pp. 31275–82.

Kou K, et al. (1995). “Structure and Function of the Mouse Insulin-Like Growth Factor Binding Protein 5 Gene Promoter.” In: *DNA Cell Bio* 14, pp. 241–49.

KovacicMilivojevic G, et al. (1993). “Regulation of the human atrial natriuretic peptide gene in atrial cardiocytes by the transcription factor AP-1.” In: *Am. J. Hypertens.* 6, pp. 258–63.

Kovacs AM, et al. (1993). “Molecular cloning and expression of the chicken smooth muscle g-actin mRNA.” In: *Cell Motil* 24, pp. 67–81.

Lamandé N, et al. (1989). “Murine muscle-specific enolase: cDNA cloning, sequence, and developmental expression.” In: *PNAS* 86, pp. 4445–9.

– (1995). “Transcriptional up-regulation of the mouse gene for the muscle-specific subunit of enolase during terminal differentiation of myogenic cells.” In: *Mol* 41, pp. 306–13.

LaPointe MC, et al. (1988). “Upstream sequences confer atrial-specific expression of the human atrial natriuretic factor gene.” In: *J. Biol. Chem.* 263, pp. 9075–8.

Lassar AB, et al. (1989). “MyoD is a sequence-specific DNA-binding protein requiring a region ofmychomology to bind to the muscle creatine kinase enhancer.” In: *Cell* 58, pp. 823–31.

Lee TC, et al. (1994). “Bifunctional transcriptional properties of YY1 in regulating muscle actin and c-myc gene-expression during myogenesis.” In: *Oncogene* 9, pp. 1047–52.

Lee Y, et al. (1997). “Myocyte-specific enhancer factor 2 and thyroid hormone receptor associate and synergistically activate the a-cardiac myosin heavy-chain gene.” In: *Mol. Cell Biol.* 17, pp. 2745–55.

LeeT.C. ChowKL, et al. (1991). “Activation of skeletal alpha-actin gene transcription: the cooperative formation of serum response factor-binding complexes over positive cis-acting promoter serum response elements displaces a negative-acting nuclear factor enriched in replicating and nonmyogenic cells.” In: *Mol. Cell Biol.* 11, pp. 5090–5100.

Leibham D, et al. (1994). “Binding of TFIID and MEF2 to the TATA element activates transcription on the Xenopus MyoDa promoter.” In: *Mol. Cell Biol.* 14, pp. 686–99.

Lenka N, et al. (1996). “The role of an e box binding basic helix loop helix protein in the cardiac muscle-specific expression of the rat cytochrome oxidase subunit VIII gene.” In: *J. Biol. Chem.* 271, pp. 30281–9.

LePage DF, et al. (1994). “Molecular cloning and localization of the human GAX gene to 7p21.” In: *Genomics* 24, pp. 535–40.

Leung E, et al. (1993). “The mouse b7 integrin gene promoter: transcriptional regulation of the leukocyte integrins LPAM-1 and M290.” In: *Int. Immun.* 5, pp. 551–8.

Lev AA, et al. (1987). “Expression of the Duchenne’s Muscular Dystrophy gene in cultured muscle cells.” In: *J. Biol. Chem.* 262, pp. 15817–20.

Lewis AL, et al. (1996). “Structure and expression of the murine muscle adenylosuccinate synthetase gene.” In: *J. Biol. Chem.* 271, pp. 22647–56.

Li H, et al. (1993). “Regulation of the mouse desmin gene: transactivation by MyoD, myogenin, MRF4 and Myf5.” In: *Nucl* 21, pp. 335–43.

– (1994a). “An E box in the desmin promoter cooperates with the E box and MEF-2 sites of a distal enhancer to direct muscle-specific transcription.” In: *EMBO J* 13, pp. 3580–9.

– (1994b). “Inhibition of desmin expression blocks myoblast fusion and interferes with the myogenic regulators MyoD and myogenin.” In: *J. Cell Biol* 124, pp. 827–41.

Li K, et al. (1990). “OXBOX, a positive transcriptional element of the heart-skeletal muscle ADP/ATP translocator gene.” In: *J. Biol. Chem.* 265, pp. 20585–8.

Li Z, et al. (1989). “Human desmin-coding gene: complete nucleotide sequence, characterization and regulation of expression during myogenesis and development.” In: *Gene* 78, pp. 243–54.

– (1991). “High level desmin expression depends on a muscle-specific enhancerJ.” In: *Biol. Chem.* 266, pp. 6562–70.

– (1993). “Different factors interact with myoblast-specific and myotube-specific enhancer regions of the human desmin gene.” In: *J. Biol. Chem.* 268, pp. 10403–15.

Lilienbaum A, et al. (1990). “Effect of human T-cell leukemia virus type I tax protein on activation of the human vimentin gene.” In: *J. Virol.* 64, pp. 256–63.

Liu M Felsenfeld G, et al. (1994). “Myocyte Enhancer Factor 2 (MEF2) Binding Site Is Essential for C2C12 Myotube-specific Expression of the Rat GLUT4 Muscle-Adipose Facilitative Glucose Transporter Gene.” In: *Journal of Biological Chemistry.* 269, pp. 28514–28521.

Liu ML, et al. (1994). “Myocyte enhancer factor 2.” In: *J. Biol. Chem.* 269, pp. 28514–28521.

Liu Z, et al. (1991). “Importance of the CArG box in regulation of beta-actin-encoding genes.” In: *Gene* 108, pp. 211–7.

Lohse P, et al. (1988). “The down-regulation of the chicken cytoplasmic beta-actin during myogenic differentiation does not require the gene promoter but involves the 3’ end of the gene.” In: *Nucl. Acids Res* 16, pp. 2787–2803.

Lompre AM, et al. (1984). “Expression of the cardiac ventricular a- and b-myosin heavy chain genes is developmentally and hormonally regulated.” In: *J. Biol. Chem.* 259, pp. 6437–46.

Long SD, et al. (1996). “Regulation of GLUT4 gene expression by arachidonic acid.” In: *J. Biol. Chem.* 271, pp. 1138–44.

Lucas M, et al. (1992). “Modulation of embryonic and muscle-specific enolase gene products in the developing mouse hindlimb.” In: *Differentiation* 1, pp. 1–7.

Lyons GE, et al. (1990). “The expression of myosin genes in developing skeletal muscle in the mouse embryo.” In: *J. Cell Biol* 111, pp. 1465–76.

Mably JD, et al. (1993). “Characterization of the CArC motif.” In: *J. Biol. Chem.* 268, pp. 476–82.

MacLellan WR, et al. (1994). “Transforming growth factor-beta response elements of the skeletal alpha-actin gene.” In: *J. Biol. Chem.* 269, pp. 16754–60.

Mahdavi V, et al. (1982). “Molecular characterization of two myosin heavy chain genes expressed in the adult heart.” In: *Nature* 297, pp. 659–64.

Maire P, et al. (1987). “Characterization of three optional promoters in the 5’ region of the human aldolase A gene.” In: *J. Mol. Biol.* 197, pp. 425–38.

Malik S, et al. (1995). “The role of the CANNTG promoter element.” In: *Eur J Biochem* 230, pp. 88–96.

Mar JH, et al. (1986). “A conserved CATTCCT motif is required for skeletal muscle-specific activity of the cardiac troponin T gene promoter.” In: *PNAS* 85, pp. 6404–8.

– (1988). “Analysis of the upstream regions governing expression of the chicken cardiac troponin T gene in embryonic cardiac and skeletal muscle cells.” In: *J. Cell Biol* 107, pp. 573–85.

Martin KA, et al. (1994). “The mouse creatine kinase paired E-box element confers muscle-specific expression to a heterologous promoter.” In: *Gene* 142, pp. 275–8.

Matsumoto T, et al. (1996). “Transcriptional and Post-translational Regulation of Insulin-like Growth Factor-binding Protein-5 in Rat Articular Chondrocytes.” In: *J. Endocrin.* 148, pp. 355–69.

McHugh KM, et al. (1988). “The development expression of the rat alpha-vascular and gamma-enteric smooth muscle isoactins: isolation and characterization of a rat gamma-enteric actin cDNA.” In: *Mol. Cell Biol.* 8, pp. 5224–31.

McNamara CA, et al. (1995). “Nuclear proteins bind a cis-acting element in the smooth muscle alpha-actin promoter.” In: *Am. Physiol. Soc.* 1, pp. 1259–66.

Mendelzon D, et al. (1994). “Phosphorylation of myogenin in chick myotubes: regulation by electrical activity and by protein kinase C.” In: *Biochem* 33, pp. 2568–75.

Merlie PJ, et al. (1989). “Neural regulation of gene expression by an acetylcholine receptor promoter in muscle of transgenic mice.” In: *Neuron* 2, pp. 1295–1300.

– (1994). “Myogenin and acetylcholine receptor alpha gene promoters mediate transcriptional regulation in response to motor innervation.” In: *J. Biol. Chem.* 269, pp. 2461–7.

Mesnard L, et al. (1993). “Molecular cloning and developmental expression of human cardiac troponin T.” In: *FEBS Lett* 328, pp. 139–44.

Min BH, et al. (1990). “The 5’-flanking region of the mouse vascular smooth muscle alpha-actin gene contains evolutionarily conserved sequence motifs within a functional promoter.” In: *J. Biol. Chem.* 265, pp. 16667–75.

Minty A, et al. (1986). “Upstream regions of the human cardiac actin gene that modulate its transcription in muscle cells: presence of an evolutionary conserved repeated motif.” In: *J. Cell Biol* 6, pp. 2125–46.

Mitsumoto Y, et al. (1994). “A long-lasting vitamin C derivative, ascorbic acid 2-phosphate, increases myogenin gene expression and promotes differentiation in L6 muscle cells.” In: *Biochem Biophys Res.* 199, pp. 394–402.

Miwa T, et al. (1987). “Duplicated CArG box domains have positive and mutually dependent regulatory roles in expression of the human alpha-cardiac actin gene.” In: *Mol Cell Biol.* 7, pp. 2803–13.

– (1991). “Structure, chromosome location, and expression of the human smooth muscle.” In: *Mol. Cell Biol.* 11, pp. 3296–306.

Mohun TJ, et al. (1986). “Upstream sequences required for tissue-specific activation of the cardiac actin gene in Xenopus laevis embryos.” In: *EMBO J* 5, pp. 3185–93.

– (1989). “The CArG promoter sequence is necessary for muscle-specific transcription of the cardiac actin gene in Xenopus embryos.” In: *EMBO J* 8, pp. 1153–61.

MohunT. G, et al. (1987). “Xenopus cytoskeletal actin and human c-fos gene promoters share a conserved protein-binding site.” In: *EMBO J* 6, pp. 667–73.

Molkentin JD, et al. (1993). “Myocyte-specific enghancer-binding factor.” In: *J. Biol. Chem.* 268, pp. 19512–20.

– (1994a). “An M-CAT binding factor and an RSRF-related A-rich binding factor positively regulate expression of the a-cardiac myosin heavy-chainin vivo.” In: *Mol Cell Biol.* 14, pp. 5056–65.

Molkentin JD, et al. (1994b). “Transcription factor GATA-4 regulates cardiac muscle-specific expression of the a-myosin heavy-chain gene.” In: *Mol Cell Biol.* 14, pp. 4947–57.

– (1996). “a-myosin heavy chain gene regulation: delineation and characterization of the cardiac muscle-specific enhancer and muscle-specific promoter.” In: *Mol Cell Biol.* 28, pp. 1211–1225.

Morkin E, et al. (1993). “Regulation of myosin heavy chain genes in the heart.” In: *Circulation* 87, pp. 1451–60.

Moss JB, et al. (1994). “The avian cardiac alpha-actin promoter is regulated through a pair of complex elements composed of E boxes and serum response elements that bind both positive- and negative-acting factors.” In: *J. Biol. Chem.* 269, pp. 12731–40.

– (1996). “The myogenic regulatory factor MRF4 represses the cardiac alpha-actin promoter through a negative-acting N-terminal protein domain.” In: *J. Cell Biol* 271, pp. 31688–94.

MouraNeto V, et al. (1996). “A 28-bp negative element with multiple factor-binding activity controls expression of the vimentin-encoding gene.” In: *Gene* 168, pp. 261–8.

Murre C, et al. (1989). “Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence.” In: *Cell* 58, pp. 537–44.

Muscat GEO, et al. (1987). “Multiple 5’-flanking regions of the human alpha-skeletal actin gene synergistically modulate muscle-specific expression.” In: *Mol Cell Biol.* 7, pp. 4089–99.

Mutero A, et al. (1995). “Promoter elements of the mouse acetylcholinesterase gene.” In: *J. Biol. Chem.* 270, pp. 1866–72.

Naidu PS, et al. (1995). “Myogenin and MEF2 Function Synergistically To Activate the MRF4 Promoter during Myogenesis.” In: *Mol Cell Biol.* 15, pp. 2707–18.

Nakamura N, et al. (1995). “Transcriptional regulation of the vimentin-encoding gene in mouse myeloid leukemia M1 cells.” In: *Gene* 166, pp. 281–6.

Nemer M, et al. (1986). “Expression of atrial natriuretic factor gene in heart ventricular tissue.” In: *Peptides* 7, pp. 1147–52.

Noursadeghi M, et al. (1993). “Trans-activation of the murine dystrophin gene in human-mouse hybrid myotubes.” In: *FEBS* 2, pp. 155–9.

Nudel U, et al. (1988). “Expression of the putative Duchenne muscular dystrophy gene in differentiated myogenic cell cultures and in the brain.” In: *Nature* 331, pp. 635–8.

Numberger M, et al. (1991). “Different mechanisms regulate muscle-specific AChR gamma- and epsilon-subunit gene expression.” In: *EMBO J* 10, pp. 2957–64.

Ohshima Y, et al. (1989). “cDNA cloning and nucleotide sequence of rat muscle-specific enolase.” In: *FEBS* 242, pp. 425–30.

Ojamaa K, et al. (1996). “Identification of a contractile-responsive element in the cardiac alpha-myosin heavy chain gene.” In: *J. Biol. Chem.* 270, pp. 31276–81.

Olva D, et al. (1995). “Conserved alternative splicing in the 5’-untranslated region of the muscle-specific enolase gene: primary structure of mRNAs, expression and influence of secondary structure on the translation efficiency.” In: *Eur J Biochem* 232, pp. 141–9.

Pari G, et al. (1991). “Multiple CArG boxes in the human cardiac actin gene promoter required for expression in embryonic cardiac muscle cells developingin vitrofrom embryonal carcinoma cells.” In: *Mol Cell Biol.* 11, pp. 4796–4803.

Pariasamy M, et al. (1989). “Regulation of myosin heavy-chain gne expression during skeletal-muscle hypertrophy.” In: *Biochem Biophys Res.* 257, pp. 691–98.

Parker TG, et al. (1992). “Positive and negative control of the skeletal alpha-actin promoter in cardiac muscle.” In: *J. Biol. Chem.* 267, pp. 3343–50.

Parmacek MS, et al. (1992). “Identification and characterization of a cardiac-specific transcriptional regulatory element in the slow/cardiac troponin C gene.” In: *Mol Cell Biol.* 12, pp. 1967–76.

– (1994). “A novel myogenic regulatory circuit controls slow/cardiac troponin C gene transcription in skeletal muscle.” In: *Mol Cell Biol.* 14, pp. 1870–85.

Parsons WJ, et al. (1993). “Gradients of Transgene Expression Directed by the Human Myoglobin Promoter in the Developing Mouse Heart.” In: *PNAS* 90, pp. 1726–30.

PedrazaAlva G, et al. (1994). “AP-1 binds to a putative cAMP response element of the MyoD1 promoter and negatively modulates MyoD1 expression in dividing myoblasts.” In: *J. Biol. Chem.* 269, pp. 6978–85.

Periasamy M, et al. (1985). “Characterization of cDNA and genomic sequences corresponding to an embryonic myosin heavy chain.” In: *J. Biol. Chem.* 260, pp. 15856–62.

Perkins EB, et al. (1995). “Two homologous enhancer elements in the chicken vimentin gene may bind a nuclear factor in common with a nearby silencer element.” In: *J. Biol. Chem.* 270, pp. 25785–91.

Peshavaria M, et al. (1991). “Molecular structure of the human muscle-specific enolase gene.” In: *Biochem. J.* 275, pp. 427–433.

Peterson CA, et al. (1992). “Beta-enolase is a marker of human myoblast heterogeneity prior to differentiation.” In: *Dev. Biol* 151, pp. 626–9.

PhanDinhTuy F, et al. (1988). “The CCArGG box: A protein-binding site common to transcription-regulatory regions of the cardiac actin, c-fos and interleukin-2 receptor genes.” In: *Eur J Biochem* 173, pp. 507–15.

Pieper FR, et al. (1987). “Upstream regions of the hamster desmin and vimentin genes regulate expression duringin vitromyogenesis.” In: *EMBO J* 6, pp. 3611–8.

– (1988). “5’ Flanking sequence of the hamster desmin gene.” In: *Mol. Biol. Rep.* 13, pp. 59–60.

– (1992). “Regulation of vimentin expression in cultured epithelial cells.” In: *Eur J Biochem* 210, pp. 509–19.

Piette J, et al. (1989). “Interaction of nuclear factors with the upstream region of the alpha-subunit gene of chicken muscle acetylcholine receptor: variations with muscle differentiation and denervation.” In: *EMBO J* 8, pp. 687–94.

– (1992). “Localization of mRNAs coding for CMD1, myogenin and the alpha-subunit of the acetylcholine receptor during skeletal muscle development in the chicken.” In: *Mechanisms of Development* 37, pp. 95–106.

Piette nan, et al. (1990). “Two adjacent MyoD1 binding sites regulate expression of the acetylcholine receptor alpha subunit.” In: *Nature* 345, pp. 353–5.

Prody CA, et al. (1991). “A developmental and tissue-specific enhancer in the mouse skeletal muscle acetylcholine receptor alpha-subunit gene regulated by myogenic factors.” In: *J. Biol. Chem.* 266, pp. 22588–96.

– (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.” In: *Nucl. Acids. Res* 20, pp. 2367–72.

Quax W, et al. (1984). “Intermediate filament cDNAs from BHK-21 cells: Demonstration of distinct genes for desmin and vimentin in all vertebrate classes.” In: *PNAS* 81, pp. 5970–4.

– (1985). “Characterization of the hamster desmin gene: Expression and formation of desmin filaments in nonmuscle cells after gene transfer.” In: *Cell* 43, pp. 327–38.

Quitschke WW, et al. (1989a). “Identification of Two Nuclear Factor-Binding Domains on the Chicken Cardiac Actin Promoter: Implications for Regulation of the Gene.” In: *Mol Cell Biol.* 9, pp. 3218–30.

– (1989b). “The beta-actin promoter: high levels of transcription depend upon a CCAAT binding factor.” In: *J. Biol. Chem.* 264, pp. 9539–46.

Rao MV, et al. (1996). “Distinct regulatory elements control muscle-specific, fiber-type-selective, and axially graded expression of a myosin light-chain gene in transgenic mice.” In: *Mol Cell Biol.* 16, pp. 3903–22.

Renz M, et al. (1985). “Regulation of c-fos transcription in mouse fibroblasts: identification of DNase I-hypersensitive sites and regulatory upstream sequences.” In: *EMBO J* 4, pp. 3711–6.

Rindt H, et al. (1993). “In vivoanalysis of the murine b-myosin heavy chain gene promoter.” In: *J. Biol. Chem.* 268, pp. 5332–8.

Rindt H, et al. (1995). “Segregation of cardiac and skeletal muscle-specific regulatory elements of the b-myosin heavy chain gene.” In: *PNAs* 92, pp. 1540–4.

Rittling SR, et al. (1989). “AP-1/jun binding sites mediate serum inducibility of the human vimentin promoter.” In: *Nucl. Acids. Res* 17, pp. 1619–33.

Rosenthal N, et al. (1989). “Myosin light chain enhancer activates muscle-specific, developmentally regulated gene expression in transgenic mice.” In: *PNAS* 86, pp. 7780–84.

– (1990). “A highly conserved enhancer downstream of hte human MLC1/3 locus is a target for multiple myogenic determination factors.” In: *Nucl. Acids. Res* 21, pp. 6239–46.

Rotwein P, et al. (1995). “Rapid Activation of Insulin-Like GRowth Factor Binding Protein-5 Gene Transcription During Myoblast Differention.” In: *Mol. Endo.* 9, pp. 913–23.

Ryan KJ, et al. (1996). “Muscle-specific splicing enhancers regulate inclusion of the cardiac troponin T alternative exon in embryonic skeletal muscle.” In: *Mol Cell Biol.* 16, pp. 4014–23.

Saggin L, et al. (1990). “Cardiac troponin T in developing, regenerating and denervated rat skeletal muscle.” In: *Development* 110, pp. 547–54.

Sakimura K, et al. (1990). “Structure and expression of rat muscle-specific enolase gene.” In: *FEBS Lett* 277, pp. 78–82.

Salminen A, et al. (1991). “Transcription of the muscle regulatory gene MYF4 is regulated by serum components, peptide growth factors and signaling pathways involving G proteins.” In: *J. Cell Biol* 115, pp. 905–17.

Salminen M, et al. (1994). “Fast-muscle-specific expression of human aldolase A transgenes.” In: *Mol Cell Biol.* 14, pp. 6797–6808.

– (1996). “Fast-muscle-specific DNA-protein interactions occurring in vivo at the human aldolase A M promoter are necessary for correct promoter activity in transgenic mice.” In: *Mol Cell Biol.* 16, pp. 76–85.

Salvetti A, et al. (1993). “Identification of a negative element in the human vimentin promoter: modulation by the human T-cell leukemia virus type I tax protein.” In: *Mol Cell Biol.* 13, pp. 89–97.

Santoro IM, et al. (1991). “Identification of single-stranded-DNA-binding proteins that interact with muscle gene elements.” In: *Mol Cell Biol.* 11, pp. 1944–53.

Sartorelli V, et al. (1990). “Muscle-specific expression of the cardiac alpha-actin gene requires MyoD1, CArG-box binding factor, and Sp1.” In: *Genes Dev* 4, pp. 1811–22.

– (1992). “Myocardial activation of the human cardiac alpha-actin promoter by helix-loop-helix proteins.” In: *PNASU* 89, pp. 4047–51.

SassoneCorsi P, et al. (1987). “Modulation of c-fos gene transcription by negative and positive cellular factors.” In: *Nature* 326, pp. 507–10.

Sawtell NM, et al. (1989). “Cellular distribution of smooth muscle actins during mammalian embryogenesis: expression of the alpha-vascular but not the gamma-enteric isoform in differentiating striated myocytes.” In: *J. Cell Biol* 109, pp. 2929–37.

Sax CM, et al. (1988). “Multiple elements are required for expression of an intermediate filament gene.” In: *Nucl. Acids. Res* 16, pp. 8057–76.

– (1989). “Down-regulation of vimentin gene expression during myogenesis is controlled by a 5’-flanking sequence.” In: *Gene78, 235-42* 78, pp. 235–42.

– (1993). “Functional analysis of chicken vimentin distal promoter regions in cultured lens cells.” In: *Gene* 130, pp. 277–81.

Schlerf A, et al. (1988). “Characterization of two different genes.” In: *EMBO. J.* 7, pp. 2387–91.

Schreier T, et al. (1990). “Cloning, structural analysis, and expression of the human slow twitch skeletal muscle/cardiac troponin C gene.” In: *J. Biol. Chem.* 265, pp. 21247–53.

Seidel U, et al. (1989). “Identification of the functional promoter regions in the human gene encoding the myosin alkali light chains MLC1 and MLC3 of fast skeletal muscle.” In: *J. Biol. Chem.* 264, pp. 16109–17.

Seidman DE, et al. (1987). “Cis-acting sequences that modulate atrial natriuretic factor gene expression.” In: *PNAS* 85, pp. 4104–8.

Shillace R, et al. (1994). “Developmental regulation of tissue-specific isoforms of subunit VIa of beef cytochrome c oxidase.” In: *Biochem Biophys Res.* 1188, pp. 391–7.

Shimizu M, et al. (1993). “Both a ubiquitous factor mTEF-1 and a distinct muscle-specific factor bind to the M-CAT motif of the myosin heavy chian b gene.” In: *Nucl. Acids. Res* 21, pp. 4103–10.

Shimizu N, et al. “Both muscle-specific and ubiquitous nuclear factors are required for muscle-specific expression of the myosin heavy-chain b gene in cultured cells.” In: *Mol Cell Biol.* 12, pp. 619–30.

– “Cis-acting elements responsible for muscle-specific expression of the myosin heavy chain beta gene.” In: *Nucl. Acids. Res* 11, pp. 1793–9.

Simon AM, et al. (1993). “An E box mediates activation and repression of the acetylcholine receptor delta-subunit gene during myogenesis.” In: *Mol Cell Biol.* 13, pp. 5133–40.

Skerjanc IS, et al. (1994). “The E-Box is essential for activity of the cardiac actin promoter in skeletal but not in cardiac muscle.” In: *Dev. Biol* 163, pp. 125–32.

Smith EO, et al. (1993). "Structural organization of the bovine gene for the heart/muscle isoform of cytochrome c oxidase subunit VIa." In: *Biochim* 1174, pp. 63–71.

Sommers CL, et al. (1994). "Regulation of vimentin gene transcription in human breast cancer cell lines." In: *Cell* 5, pp. 839–46.

Song WK, et al. (1992). "H36-alpha7 is a novel integrin alpha chain that is developmentally regulated during skeletal myogenesis." In: *J. Cell Biol* 117, pp. 643–644.

Sprenkle AB, et al. (1995). "Involvement of multiple cis elements in basal- and alpha-adrenergic agonist-inducible atrial natriuretic factor transcription: roles for serum response elements and an SP-1-like element." In: *Circ. Res.* 77, pp. 1060–9.

Sternberg EA, et al. (1988). "Identification of upstream and intragenic regulatory elements that confer cell-type-restricted and differentiation-specific expression on the muscle creatine kinase gene." In: *Mol Cell Biol.* 8, pp. 2896–2909.

Stover DM, et al. (1992). "Identification of a cis-acting DNA antisilencer element which modulates vimentin gene expression." In: *Mol Cell Biol.* 12, pp. 2230–40.

Su CT, et al. (1995). "The depolarization response element in acetylcholine receptor genes is a dual-function E box." In: *FEBS Letters* 366, pp. 131–6.

Subramaniam A, et al. (1990). "Analysis of the upstream regulatory region of a chicken skeletal myosin heavy chain gene." In: *J. Biol. Chem.* 265, pp. 13986–94.

Sunyer T, et al. (1993). "Cell type- and differentiation-dependent expression from the mouse acetylcholine receptor epsilon-subunit promoter." In: *J. Neurosci. Res.* 36, pp. 224–34.

Swiderski RE, et al. (1990). "Precocious appearance of cardiac troponin T pre-mRNAs during early avian embryonic skeletal muscle development in ovo." In: *Dev. Biol* 140, pp. 73–82.

Sympson CJ, et al. (1993). "Cytochalasin D-induced actin gene expression in murine erythroleukemia cells." In: *Exp. Cell Res.* 205, pp. 225–31.

Szucsik JC, et al. (1995). "Cloning and sequence analysis of the mouse smooth muscle g-enteric actin gene." In: *Genomics* 28, pp. 154–62.

Talib S, et al. (1993). "Differential expression of human nicotinic acetylcholine receptor alpha subunit variants in muscle and non-muscle tissues." In: *Nucl. Acids. Res* 21, pp. 233–7.

Tanaka M, et al. (1985). "Switching in levels of teranslatable mRNAs for enolase isozymes during development of chicken skeletal muscle." In: *Biochem Biophys Res.* 133, pp. 868–72.

Tapscott SJ, et al. (1992). "A novel myoblast enhancer element mediates MyoD transcription." In: *Mol Cell Biol.* 12, pp. 4994–5003.

Taylor JM, et al. (1995). “Regulation of the myoblast-specific expression of the human beta-enolase gene.” In: *J. Biol. Chem.* 270, pp. 2535–40.

Taylor MV, et al. (1989). “Muscle-specific (CArG) and serum-responsive (SRE) promoter elements are functionally interchangeable in Xenopus embryos and mouse fibroblasts.” In: *Development* 106, pp. 67–78.

– (1991). “Xenopus embryos contain a somite-specific, MyoD-like protein that binds to a promoter site required for muscle actin expression.” In: *Genes Dev* 5, pp. 1149–60.

Tebbey PS, et al. (1994). “Arachidonic acid down-regulates the insulin-dependent glucose transporter gene.” In: *J. Biol. Chem.* 269, pp. 639–44.

Thompson WR, et al. (1991). “A MyoD1-independent muscle-specific enhancer controls the expression of the b-myosin heavy chain gene in skeletal and cardiac muscle cells.” In: *J. Biol. Chem.* 266, pp. 22687–8.

Trask RV, et al. (1988). “Developmental regulation and tissue-specific expression of the human muscle creatine kinase gene.” In: *J. Biol. Chem.* 263, pp. 17142–9.

Treisman R, et al. (1986). “Identification of a protein-binding site that mediates transcriptional response of the c-fos gene to serum factors.” In: *Cell* 46, pp. 567–74.

Trouche D, et al. (1993). “Repression of c-fos promoter by MyoD on muscle cell differentiation.” In: *Nature* 363, pp. 79–82.

– (1995). “Myogenin binds to and represses c-fos promoter.” In: *FEBS Lett* 361, pp. 140–4.

Underwood LE, et al. (1987). “Pretranslation Regulation of Myoglobin Gene Expression.” In: *Am. J. Physiol.* 252, pp. 450–2.

Vosberg HP, et al. (1992). “The regulation of the human b myosin heavy-chain gene.” In: *Basic. Res. Cardiol.* 1, pp. 161–72.

Walke W, et al. (1994). “Calcium-dependent regulation of rat and chick muscle nicotinic acetylcholine receptor.” In: *J. Biol. Chem.* 269, pp. 19447–56.

Walsh K, et al. (1988). “DNA-binding site for two skeletal actin promoter factors is important for expression in muscle cells.” In: *Mol Cell Biol.* 8, pp. 1800–2.

– (1989). “Cross-binding of factors to functionally different promoter elements in c-fos and skeletal actin genes.” In: *Mol Cell Biol.* 9, pp. 2191–2201.

Wan B, et al. (1995). “Structural characterization and regulatory element analysis of the heart isoform of cytochrome c oxidase VIa.” In: *nan* 270, pp. 26433–40.

Wang G, et al. (1994). “Characterization of cis-regulating elements and trans-activationg factors of the rat cardiac troponin T gene.” In: *J. Biol. Chem.* 269, pp. 30595–603.

Wang XM, et al. (1990). “Expression of the acetylcholine receptor delta-subunit gene in differentiating chick muscle cells is activated by an element that contains two 16 bp copies of a segment of the alpha-subunit enhancer.” In: *EMBO J* 9, pp. 783–90.

Wang Y, et al. (1988). “A cell type-specific enhancer drives expression of the chick muscle acetylcholine receptor alpha-subunit gene.” In: *Neuron* 1, pp. 527–34.

Watanabe T, et al. (1997). “Regulation of Troponin T gene expression in chicken fast skeletal muscle: Involvement of an M-CAT-Like element distinct from the standard M-CAT.” In: *J. Biochem* 121, pp. 212–8.

Wefald FC, et al. (1990). “Functional Heterogeneity of Mammalian TATA-box Sequences Revealed by Interaction with a Cell-Specific Enhancer.” In: *Nature344, 260-2* 87, pp. 5623–7.

Weintraub H, et al. (1990). “MyoD binds cooperatively to two sites in a target enhancer sequence: Occupancy of two sites is required for activation.” In: *PNAS* 87, pp. 5623–7.

– (1991). “The MCK enhancer contains a p53 responsive element.” In: *PNAS* 88, pp. 4570–1.

Weller PA, et al. (1986). “Myoglobin Expression: Early Induction and Subsequent Modulation of Myoglobin and Myoglobin mRNA During Myogenesis.” In: *Mol Cell Biol.* 6, pp. 4539–47.

Wentworth BM, et al. (1991). “Paired MyoD binding sites regulate myosin light chain gene expression.” In: *PNAS* 88, pp. 1242–6.

Wu J, et al. (1989). “Tissue-specific determinants of human atrial natriuretic factor gene expression in cardiac tissue.” In: *J. Biol. Chem.* 264, pp. 6471–9.

Yano H, et al. (1994). “Identification of two distinct promoters in the chicken caldesmon gene.” In: *Biochem Biophys Res.* 201, pp. 618–26.

– (1995). “Transcriptional regulation of the chicken caldesmon gene.” In: *J. Biol. Chem.* 270, pp. 23661–6.

Yao CC, et al. (1996a). “Alpha-7 integrin mediates cell adhesion and migration on specific laminin isoforms.” In: *J. Biol. Chem.* 271, pp. 35598–603.

– (1996b). “Laminins promote the locomotion of skeletal myoblasts via the alpha-7 integrin receptor.” In: *J. Cell Sci.* 109, pp. 3139–50.

Yee SP Rigby PW, et al. (1993). “The regulation of myogenin gene expression during the embryonic development of the mouse.” In: *Genes Dev.* 7, pp. 1277–1289.

Yi TM, et al. (1991). “Rabbit muscle creatine kinase: genomic cloning, sequencing, and analysis of upstream sequences important for expression in myocytes.” In: *Nucl. Acids. Res* 19, pp. 3027–33.

Yutzey KE, et al. (1990). “Differential trans activation associated with the muscle regulatory factors MyoD1, myogenin, and MRF4.” In: *Mol Cell Biol.* 10, pp. 3934–44.

Zambetti GP, et al. (1992). “Wild-type p53 mediates positive regulation of gene expression through a specific DNA sequence element.” In: *Genes Dev* 6, pp. 1143–52.

Zhang JA Mortazavi A, et al. (2012). “Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity.” In: *Cell.* 149, pp. 467–482.

Zhao J, et al. (1994). “Mouse p53 represses the rat brain creatine kinase gene but activates the rat muscle creatine kinase gene.” In: *Mol Cell Biol.* 14, pp. 8483–92.

Zhu L, et al. (1995). “Developmental regulation of troponin I isoform genes in striated muscles of transgenic mice.” In: *Dev. Biol* 169, pp. 487–503.

Ziober BL, et al. (1993). “Alternative extracellular and cytoplasmic domains of the integrin alpha-7 subunit are differentially expressed during development.” In: *J. Biol. Chem.* 268, pp. 26773–83.

– (1996). “Identification and characterization of the cell type-specific and developmentally regulated alpha-7 integrin gene promoter.” In: *J. Biol. Chem.* 271, pp. 22915–22.