Chromatin Topology and Transcription in Myogenesis

Thesis by
Katherine I Fisher-Aylor

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California
2017
(Defended August 19, 2016)

Katherine Irene Fisher-Aylor
ORCID: 0000-0003-3371-2947

Acknowledgments Page

Abstract

High-throughput sequencing and the resulting development of biochemical “-Seq” experiments such as ChIP-Seq, DNase-Seq, and Methyl-Seq over the past decade has given rise to a wealth of predicted enhancers and other cis-regulatory regions (CRMs). These new assays provide a new opportunity to compare the number, location, and possible nature of CRMs that are predicted by various new biochemical techniques to instances of known CRMs, which until recently have primarily been located—for reasons of technological limitations—at a few tens of highly expressed, mostly developmentally-specific genes and the several kilobases (kb) upstream of their promoters. For example, an early surprise in the first ChIP-Seq experiments was that the number of predicted tissue-specific transcription factor-occupied sites outnumbered the number of tissue-specific genes by at least a factor of 10, and that many of these occupied sites were nowhere near developmentally relevant genes. In this thesis, I use the ChIA-PET technique, which preserves factor-containing physical interactions between loci in the genome that are far from each other (10kb-2Mb), where the factors used in this thesis are RNA Polymerase II (pol2) to capture active genes, and separately the developmental transcription factor Myogenin to additionally capture CRMs not at promoters. Overall, I report that (1) the closer together two occupied regions are, the more likely they are to be connected, and (2) that a gene’s activity level is highly correlated with its likelihood of being physically engaged with a distant occupied locus. These lead to the discoveries that occupied regions tend to engage with the active genes nearest to them regardless of the developmental profile of the genes, that many genes engage with multiple individual loci, and that many occupied regions interact with multiple genes, including genes that are not at all related in terms of their expression patterns. Individual elements that have multiple connections likely represent sequential rather than simultaneous interactions, and developmental genes may require more engaged enhancers than genes that are expressed in all cell types. Most excitingly, it is possible that many genes with unchanging expression patterns, including so-called “housekeeping genes,” use CRMs; very few such genes have ever been assayed with respect to gene regulation, and they are the vast majority of genes in the genome.

Published Content

Ozdemir, Anil*, Katherine I Fisher-Aylor*, et al. (2011). “High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation.” Genome Research **21**(4): 566-577. doi: 10.1101/gr.104018.109

K.I.F-A. prepared the ChIP-Seq data, analyzed the ChIP-Seq and SELEX data, devised a novel methodology for analyzing Illumina/Solexa library content, discovered and quantified biases in sequencing data, participated in the conception of the second set of experiments, and participated in the writing of the manuscript.

Tai, Phillip WL, Katherine I Fisher-Aylor, et al. (2011). “Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer.” Skeletal Muscle **1**:25. doi: 10.1186/2044-5040-1-25

K.I.F-A. conceived of the redesign of the ChIP-Seq fixation -- and was the first person -- to ChIP Mef2 in mouse myocytes; performed, analyzed, and prepared the Mef2 ChIP-Seq data; and drafted portions of the manuscript describing the ChIP-Seq data and other Wold lab data.

Contents

List of Figures

Chapter I: Introduction

I.1: Overview

This project aims to map the physical landscape of DNA interactions that are associated with key regulatory molecules, and to relate the resulting map to changes in gene regulated across muscle differentiation. A specific goal was to use a genomic assay to define the genomic repertoire of distally located candidate cis-acting regulatory elements in myoblasts and myocytes and to learn how they associate with each other physically. To do this, I refined and made much more sensitive and robust an assay that was then in its early stages of development, called ChIA-PET (chromatin interaction analysis by paired-end tag sequencing). This method is designed to detect complexes that join relatively long-distance interactions (10kb-2Mb) and also contain a regulatory protein of interest. Here, the proteins I investigate are RNA polymerase II (pol2), which transcribes protein coding and lnc RNAs; and myogenin, a major tissue specific transcription factor necessary for muscle differentiation.

A longstanding model for transcriptional regulation in large eukaryotic genomes centers on specific physical looping events that are thought to join active transcriptional enhancer elements with their proximal promoters. As summarized below, this textbook model (Fig. I-1) had been built up beginning in the 1980s mainly by dissecting a relatively small set of “model” genes in increasing detail. The focal genes for this work, such as the globins, immunoglobulins, interferon, or in my myogenesis system, muscle creatine kinase (MCK), actin and MyoD, had not been selected randomly from the genome. Rather, they had become central because they were either especially accessible technically or were thought to be exceptionally interesting due to the function of their protein product – most often both.

As I began my thesis work, new methods had just been developed to map the entire genome for biochemical signatures associated with gene regulatory elements in the DNA, and to detect physical contacts between sites on the chromosome. The regulatory elements, called cis-acting regulatory modules (CRM), contain clusters of binding sites for sequence specific DNA binding proteins, and those proteins can in turn engage a variety of cofactors and chromatin modifying enzymes. CRMs are understood to alter transcription at their target promoters through mechanistically diverse actions of their bound protein complexes. As reviewed below, these interactions were thought to be mainly focused on the nearest promoter in DNA space, although it was well appreciated that some long distances in DNA-space could be bridged (1MB for a limb shh enhancer) and there is evidence for a few cases of cross-chromosomal regulation (transvection) which might or might not be a special instances of conventional CRM function. The functional impact of a CRM is defined operationally for each given cell type and state, where its net effect will be to activate (enhance) or repress (silence) productive transcription from a target gene. A third function is insulation. An active insulator can prevent an enhancer positioned on one side from affecting a promoter located on the other side of the insulator.

What I found was exciting because it made me question some established ideas about transcriptional enhancement and also our way of thinking about the distinctions made previously between so-called housekeeping genes and other genes that are aggressively regulated during differentiation.

I.2: The origin of the CRM as a distinct concept

Cis-Regulatory Modules (CRMs) are elements in the DNA which affect transcription of a targeted promoter, but which are not themselves promoters. They are currently divided into three major types: enhancers, insulators, and silencers, and are

made up of DNA binding sites for a variety of sequence-specific DNA binding proteins. These proteins in turn engage a variety of transcriptional co-factors and chromatin modifying enzymes which enable or inhibit transcription. The promoter is the DNA sequence which enables correct transcription of a gene, and the functional impact of a CRM is defined operationally for each given cell type and state (and, perhaps in the future, for each different promoter), where its net effect will be to increase transcription (enhance), decrease transcription (silence), or protect transcription from the effects of enhancers and silencers (insulate).

It is useful to introduce these four entities as distinct concepts because the experimental history distinguishing them underlies much of our current thinking whether we realize it or not: those who worked during this era or have studied the original literature likely realize where the empirical definitions end and the models and their correlates begin. However, in some contemporary papers, it seems that the definitions have drifted in ways that can be confusing or even circular. It is also useful to recognize the molecular basis for our current knowledge before delving into a discussion of how, excitingly, the clear-cut distinctions between the four entities become blurred as we learn more about the complexities of gene regulation. An important example is that it is conceptually difficult at times to draw lines between where a promoter ends and a CRM begins. In the muscle system studied here, some known CRMs have elements of more than one molecular mechanism that operate in different cells or cell states (e.g., Berghella, De Angelis et al. 2008). I find it most exciting that large-scale genome mapping data like mine can be analyzed to build testable models for CRM sub-classes or to, in some cases, refute existing models. It is possible that not all CRMs of a type behave in the same way as each other, and CRMs of one type may have more in common with CRMs of another type, or with promoters, than previously imagined.

I.2.1: Experiments and prior knowledge of the early CRM era

In the early 1980s, at the birth of CRMs as a concept, there were enough biochemical techniques to do excellent single-locus gene studies, but there were no genomics and any sequencing at all was slow and expensive. Some widespread experimental techniques for studying molecular genetics were, for DNA sequencing, gel-style Sanger sequencing (e.g., Schaffner, Kunz et al. 1978); for quantification of transcription at the mRNA level, Rot curves and Northern blot quantification (e.g., Wold, Wigler et al. 1979) and later transfection assays such as CAT; and for gene knock-in and knock-out experiments, transfections of recombinant plasmids into cultured cells (Wigler, Sweet et al. 1979; Wold, Wigler et al. 1979). It was the DNA sequencing and mRNA quantifications that made things most difficult for scientists of the time. This was well before the sequencing of the human, or any, genome, and DNA sequencing was extremely slow. Similarly, the cDNA quantifications necessitated studying genes with a very high level of transcriptional output at the level of mRNA. Taken together, this meant that although many at the time knew this would give them a biased view of regulation across genome, molecular genetics studies had to focus on (1) genes that stood out by classical genetics because their mutation and phenotype allowed their cloning and focal interest; (2) genes with extremely high mRNA levels, making them technically more accessible; (3) genes strongly expressed in organisms or cell types easy to culture; (4) the region of each gene that is the proximal promoter and areas very close to it (usually within 2-5kb), since the search for enhancers typically ended after finding one; and (5) using large cloned regions of DNA, which made it difficult to find small functional elements and to differentiate between the effects of neighboring elements.

Enabling the discovery and working definition of the “enhancer,” the promoter was already a reasonably defined functional DNA element type. In the early 1980s,

promoters had been studied for over a decade in the context of several important genes of simpler organisms. Bacterial promoters such as in the lac operon in E. coli (e.g., Kennell and Riezman 1977), and viral promoters such as lambda phage cI (e.g., Hochschild, Irwin et al. 1983) and T7 phage A3 (e.g., Siebenlist, Simpson et al. 1980) were among the first genes studied. Then genes such as cytochrome C from single-celled eukaryotic yeast (e.g., Faye, Leung et al. 1981) and genes of animal cell viruses such as tk from herpesvirus (e.g., McKnight, Kingsbury et al. 1984) and T-antigen from simian virus 40 (e.g., Gluzman, Sambrook et al. 1980) paved the way for the study of endogenous genes of animals such as H2A in sea urchin (e.g., Grosschedl and Birnstiel 1980a,b) and globin genes in rodents (e.g., Wold, Wigler et al. 1979). The question in these early animal studies was to see how the expression of animal genes in large genomes differed or was similar to smaller genomes like yeast and to constrained genomes in bacteria and viruses. In asking this question, the first enhancer sequences were found.

I.2.2: Promoters

Because CRMs are defined in relation to their own promoter(s), affecting promoters without themselves being promoters, it is necessary first to understand what a promoter is and then to define its specific meaning for the purposes of this work. There are a number of different overlapping and context-specific uses of the term “promoter” when it comes to discussing promoter DNA sequence content: minimal promoter, core promoter, basal promoter, proximal promoter, and more. However, the biochemical definition of a promoter – DNA sequence proximally upstream of a gene that is necessary for the proper mRNA expression of that gene – has remained constant even as we have learned about the diversity of promoter sequences. Another aspect of the

promoter that has remained unchanged throughout history is that promoters are thought to be uni-directional.

An efficient, conserved bacterial promoter had been found (Siebenlist, Simpson et al. 1980), but it was clear that there would be no such consensus sequence for an eukaryotic promoter. The TATA box was discovered (Goldberg, 1979, Stanford PhD thesis) which turned out to be necessary for transcriptional initiation of approximately 30% of genes, including the most developmentally regulated genes and most highly expressed (reviewed in Breathnach and Chambon 1981), while the more numerous TATA-less genes were thought mainly to control low-level housekeeping genes (reviewed by Dynan 1986). In similar genes, it was found that TATA was not sufficient to support transcription on its own (necessary for fixing the proper transcriptional start site and direction, but not sufficient to power transcription), while sequences ~75-150bp upstream were necessary for initiation itself and proper stabilization of the mRNA product (Benoist and Chambon 1981; Dierks, van Ooyen et al. 1981). Various bidirectional (Grosschedl and Birnstiel 1980) sequences such as the CCAAT box and the GC box were found upstream of the TATA box which partially but not completely explained some of these requirements (Grosschedl and Birnstiel 1980a,b; Dierks, van Ooyen et al. 1983).

The lack of a fully functional consensus promoter sequence, or sequences, in large metazoan genomes is due in part to finding many different variations of promoter sequences and corresponding transcriptional machinery that are used to achieve “necessary and sufficient” status for proper transcription initiation. Also, we appreciate now that elongation, termination and control of turnover all affect mRNA output from both native genes and reporter genes. This becomes relevant in future chapters when I test for correlations of candidate CRM and promoter connectivity with measured RNA output.

Things that remain unknown or ambiguous about the full range of promoters in the genome and about how they work mean that in the following chapters I have to select and use the most appropriate definition and then clarify the ramifications of that choice. For the results and conclusions of this thesis, this biochemical definition of the promoter is approximated by the use of the gencode M1 annotated TSSs, which uses protein, mRNA, and ncRNA sequences to make a set of biologically derived, informatically predicted TSSs.

I.2.3: Enhancers

By studying the operation of viral promoters in eukaryotic cells, the first enhancer element was isolated from DNA of simian virus 40 (Banerji, Rusconi et al. 1981; Benoist and Chambon 1981; Gruss, Dhar et al. 1981). It was named as an enhancer and recognized as its own conceptual class of regulatory element (Banerji, Rusconi et al. 1981) based on its action in a gain-of-function plasmid transfection experiment: to increase transcriptional output of a nearby reporter gene. “Enhancer” was then and is still an assay-specific term. Upon studying known enhancers, many properties were found in most or all of them, but it is important to remember the distinction between the definitional description of “enhancer” (an assay-specific biochemical definition) and the characteristics observed to associate with them in subsequent one-off enhancer studies. It is not always clear, when the term “enhancer” is used, which enhancer characteristics are definitional and which are observational. It was found that enhancers bind transcription factors, often more than one type, and that they affect promoters by interacting with them physically. However, the observed characteristics of enhancers do not operate commutatively: not all reproducibly factor-occupied DNA sites and not all sites that interact physically with genes will act as enhancers on their own. This is true both in the modern-day versions of the first enhancer assays, which are gain-of-function

experiments, and in the smaller but growing set of loss-of-function mutational assays. I will therefore refer to factor binding sites and physically associated regions where appropriate as “candidate CRMs” to be agnostic.

The first set of enhancer characteristics were described in 1983: (1) they act in cis- to increase transcription of a promoter, (2) they can be 3’ or 5’ relative to the promoter, (3) their sequences can be flipped; they are non-directional, (4) they are able to act on different promoters (Mercola, Wang et al. 1983). These four descriptive “rules” encompassed the actions of the handful of known enhancers that had been studied in the preceding few years, but did not change the underlying assay-specific definition of “enhancer.” To wit, the first cellular enhancers were then found, proving that there were at least a few enhancers in eukaryotic genomes, and that enhancers were not a virus-specific phenomenon (Mercola, Wang et al. 1983; Neuberger 1983; Weber, de Villiers et al. 1984). Based on these studies, the descriptive “rules” in common between experimentally validated enhancers changed: (1) they act in cis- to increase transcription of a promoter, (2) they can be 3’ or 5’ relative to the promoter, overlapping with promoters (Weber, de Villiers et al. 1984), or in an intron (Gillies, Morrison et al. 1983), (3) non-directional, (4) preferentially stimulate the closest of two promoters, (5) they can be cell type preferential (Gillies, Morrison et al. 1983) and inducible (Serfling, Lubbe et al. 1985; reviewed by Serfling, Jasin et al. 1985). Since only aspects (1) and (3) are definitional—enhancers are non-promoter cis- elements that have a measurably positive effect on promoters in assay—all of the other “rules” were and are observational and are likely to change or to apply to subsets of the overall “enhancer” group as we discover and learn more about thousands of enhancers.

Some of the major questions at the time of enhancer discovery were whether enhancers can or do act in cis- only, or also in trans- (i.e. how important was the DNA

backbone in how enhancers acted?) and also whether enhancers acted by looping or by somehow scanning along the DNA. Evidence for the existence of some scanning-like behavior (then posited to be pol2 crawling along the DNA backbone) included “barrier” experiments that impeded enhancement by positioning a chemical barrier in the DNA between enhancer and promoter (Brent and Ptashne 1984; Plon and Wang 1986) and a persistent bias in enhancers to their action on closer promoters (Muller, Gerster et al. 1988). Scanning via gyrase activity of pol2 had been ruled out (Plon and Wang 1986), but the barrier experiments and the closest-promoter bias were still reasonably strong evidence in favor of some sort of scanning-like behavior, if not by pol2 then by something else. There was also circumstantial evidence that looping was possible in the distances between known enhancers and promoters, though. Looped-out DNA had been visualized near active genes in prokaryotes (Griffith, Hochschild et al. 1986; Kramer, Niemoller et al. 1987) and then in eukaryotes (Theveny, Bailly et al. 1987).

Both of these debates – on the necessity to be linked in cis- on the chromosome, or for polymerase to scan through DNA, separating an enhancer spatially from its target promoter – ended in 1989 when it was shown that enhancers could act in trans- via a protein bridge to the promoter (Muller, Sogo et al. 1989). It was accepted that while some enhancers may possibly act by different methods, the “enhancer effect” did not require cis- (past the fact that its sequence is DNA), only trans-, because enhancers and promoters could loop to interact with each other physically. The looping model of enhancement predominated (Muller and Schaffner 1990). Essentially, enhancers can, at least in some instances, act in cis- or in trans-, in cis- because they are DNA sequences and thus subject to certain constraints of the DNA backbone and location in the nucleus and in trans- because enhancers act on promoters in 3D, the DNA sequences of both

the enhancer and the promoter providing a scaffold for the formation of a protein bridge between them.

Assuming the looping model of enhancement, what variety of cis- elements will be connected to active genes? Will the interactions be simple or complex? Will differentially expressed, highly expressed, or well-studied genes be connected any differently than other genes in the genome?

I.2.4: Transcription Factor Interactions at CRMs and Promoters

Shortly after the discovery of enhancers, it was noticed that a certain motif in the DNA was present in many enhancers as well as in many promoters (McKnight 1984). In studying this motif, it was proposed that some factor might occupy it in order to cause a bridge between the enhancer and promoter. It was then found that a protein, called Sp1, did indeed recognize the identified DNA motif (Gidoni, Dynan et al. 1984). Proteins of this type were called transcription factors, and transcription factors were shortly found in every system studying transcriptional regulation. One interesting seeming paradox that was discovered early (Muller, Gerster et al. 1988) and seldom remarked upon, although it has been demonstrated in a variety of systems, is that general transcription factors can seemingly regulate tissue-specific genes while tissue-specific transcription factors can seemingly regulate constitutively active genes. For example, the muscle-specific factor myoD occupies sites in and near muscle-specific and constitutively active genes alike (Cao, Yao et al. 2010; Kwan G and Wold B, unpublished), and so do more ubiquitous transcription factors such as USF1, Jun/Fos, and E-proteins (Fisher-Aylor K, Marinov G, Kwan G, Desalvo G, Kirilusha A, and Wold B, unpublished; Kirilusha A 2014 thesis). In worm muscle, the myoD homolog HLH1 caused changes in both muscle-specific and constitutively active genes when knocked out (Kuntz, Williams et al. 2012).

Other factors to consider when thinking about chromatin and gene expression

It is not just enhancers that interact with genes – far from it. While enhancers have attracted the most attention, partly because of the relative ease of the most-used assays, two other classes of elements, sometimes co-occurring with enhancer elements in compound CRM sequences, are also important in gene regulation and in the interpretation of data in this thesis.

I.2.5: Insulators

Since the late 1970s, it had been noticed that the chromatin state – active or inactive/silenced – as assayed by nucleosome accessibility and core histone tail acetylation, was closely and perhaps causatively related to gene expression. In studying this, the concept of the insulator was born. The first insulator work was done in *Drosophila*, on scs and scs', two elements marking the end of an active chromatin region around hsp70 (Udvardy, Maine et al. 1985). These elements and a similar element, gypsy, were tested in gain-of-function transfection assay. It was found that these elements had two characteristics in common: they protected elements from their neighboring chromatin environment, and they also protected a promoter from enhancement by an enhancer (Kellum and Schedl 1991).

The ability of certain elements to block enhancement of a promoter had been noticed before: first, in the original enhancer paper by inclusion of a long stretch of un-sequenced DNA (Banerji, Rusconi et al. 1981) and then by two experiments that sought to differentiate between the looping and scanning models of enhancement by adding a biochemical road block in between the enhancer and the promoter (Brent and Ptashne 1984; Courey, Plon et al. 1986). However, this was the first time such an element had been found in an organism itself, with at least one apparent function: serving as a border between chromatin states. The first insulators in vertebrate were also isolated and

described (Chung, Whiteley et al. 1993) from previously described sharp transitions in chromatin state near the chicken β-globin locus (Hebbes, Clayton et al. 1994).

It was found that insulators acted as barriers to polycomb-group repression (Mallin, Myung et al. 1998). They were also found to act as barriers to DNA replication (Wiesendanger, Lucchini et al. 1994), including imprinting-specific replication (Greally, Starr et al. 1998) and that some MARs acted as insulators (Namciu, Blochlinger et al. 1998) and co-localized with the gypsy insulator in the nucleus (Nabirochkin, Ossokina et al. 1998). It did not take much imagination to wonder if this bordering or blocking mechanism acted similarly in all cases, particularly with respect to the blockage of enhancers, and particularly since insulation can be directional (Chung, Bell et al. 1997) and can seemingly prevent enhancers from acting in trans- (Krebs and Dunaway 1998). A major hypothesis of the time was that factors binding to insulators may disrupt the assembly or stabilization of a trans- E:P complex, particularly through the Chip or Ldb1 family of factors (Morcillo, Rosen et al. 1997).

Upon studying insulators, proteins that bound them were found. The first were suppressor of hairy wing (Nabirochkin, Ossokina et al. 1998; Gerasimova, Gdula et al. 1995), bithorax, and trithorax (Gerasimova and Corces 1998) in *Drosophila*. Similar proteins were found in other systems. The role(s) of insulators are therefore similar to other CRMs in the way they bind proteins. The ways that insulators have been demonstrated to affect transcriptional activation are various and conceptually relate transcriptional regulation to chromatin states and chromatin conformation.

I.2.6. Chromatin, LADs/LASs, MARs/SARs

It was later found that SATB1, a homeobox protein which associates with the nuclear matrix, also recognizes a non-specific “secondary” DNA binding motif made up of ~25bp A/T-rich non-palindromic sequence, and that SATB1 is the likely mediator of

DNA-nuclear matrix interactions (Nakagomi, Kohwi et al. 1994; Galande, Purbey et al. 2007). Later molecular methods showed that the nuclear lamina also appears to play a structural role in some chromatin interactions. Genes are often connected to the lamina, although it appears true that in different cell types or organisms, these lamina-associated genes are expressed (Pickersgill, Kalverda et al. 2006), repressed (Guelen, Pagie et al. 2008), or temporally associated with the lamina in a way that establishes lineage commitment programs (Peric-Hupkes, Meuleman et al. 2010). Possibly related to these phenomena is the fact that occasionally gene activity is reported as being related to proximity to telomeres or centromeres in various cells (Brown, Guest et al. 1997; Francastel, Walters et al. 1999; Andrey, Montavon et al. 2013; Robin, Ludlow et al. 2015). In addition to genic interactions that inspired many of these experiments, CTCF was found to be involved in many chromatin interactions (Ling, Li et al. 2006; Splinter, Heath et al. 2006). CTCF is now primarily understood to be an insulator whose pattern of occupancy varies relatively little across cell and tissue types, although it also has additional active and repressive functions at some loci (Kim, Abdullaev et al. 2007), showing that some chromatin interactions other than MARs, namely insulators, might be structural. In fact, it has been shown that many CTCF sites are associated physically with other CTCF sites, suggesting that at least some looping (likely the most invariant loops in the genome) may be mediated by CTCF (Ong and Corces 2014).

I.2.7: Silencers and Compound CRMs

Silencers are cis-regulatory modules that, when brought into proximity of an active promoter, cause decreased expression in a knock-in assay. Silencers bind protein factors called repressors that cause the negative regulation. Some silencers appear to work by binding a factor that disrupts proper binding of an activating transcription factor, whether binding directly to sites that prevent access of necessary

transcription factors (Johnson, Mortazavi et al. 2007) or by physically interacting with scaffold proteins that in turn recruit transcriptional disruptors or otherwise block association of enhancers and promoters (Lupo, Cesaro et al. 2013). In addition to silencers, insulators, promoters, and enhancers, there also exist mixed-category CRMs, sometimes called “compound CRMs,” which have binding sites for a wide variety of protein factors and which may sometimes bind activating factors, acting as enhancers, and sometimes bind repressors, acting as silencers (Davidson 2006).

I.3: Skeletal Myogenesis and the C2C12 model system

Skeletal myogenesis is a highly conserved and ancient developmental process that is characterized in all systems studied by having at least one bHLH Muscle Regulatory Factor (MRF) and usually also involves MADs cofactors such as Myocyte enhancer factors (MEF-2s) and Serum response factors (SRFs). For example, in *Drosophila* and in *C. elegans*, nautilus and bHLH1 are the respective myoD orthologs, and dMef2 is a major MADs family gene in fly.

For over 30 years, mammalian skeletal muscle differentiation has been studied in the mouse using a model cell line called system called C2C12 (Yaffe and Saxel 1977; Blau, Chiu et al. 1983; Blau, Pavlath et al. 1985). That work provides a vast background of molecular biological data to draw on to interpret genome-scale data in this thesis. Most skeletal muscle – the trunk and limb skeletal muscles – all originate from the somites of the developing embryo. The first committed myoblast precursor cells appear in the dermomyotome, and then begin to differentiate and migrate (see (Buckingham and Rigby 2014)). At this time, the somite has flattened and is wrapped around the neural tube in a “C” shape with the notochord (a signaling center) positioned close to its ventral side. The inner portion of the dermomyotome is fated to become the vertebral column, ribs, and tendons of the midsection. The portion of the outer dermomyotome

expresses Pax3, and nearer the center, Pax7 is also detected. Cells in this domain, expressing one or both of these PAX-family factors, comprise a renewing source of myoblasts fated to become skeletal muscles of the trunk. Meanwhile, Pax3-positive cells from the hypaxial (proximal to notochord) dermomyotome migrate to the limb bud, where they later express the three MRF determination factors – MyoD, Myf5, and MRF4 and eventually myogenin – to become skeletal muscles of the limbs.

Much is known about the factors involved in limb skeletal myogenesis. The Pax3 precursor cells begin to migrate, and while doing so, they express c-met and Lbx1, which activates CXCR4 (Buckingham and Rigby 2014). During migration, the cells are prevented from differentiating by factors that repress both the MRF's and their cofactors: Snail represses an important subset of myoD-binding sites, Sim2 represses myoD, Msx1 represses myoD and Myf5, and the Dach factors repress Six and Eya transcription factors. However, upon reaching the limb bud, the migratory limb skeletal muscle precursors express MRF's and begin to differentiate. This is due in part to Shh signaling, which activates Myf5 by way of Gli, but also to Wnt signaling, which activates Pitx2 by way of Tcf4. Pitx2 activates Six and Eya, which in turn activate myoD and Myf5. Meox2 is another transcription factor turned on at this time, and it contributes to activating expression of Myf5 as well.

Once myoD and/or Myf5 are activated, they are able to positively regulate themselves, each other, and some of their precursors such as Pax3 and the Six transcription factors. The same goes for Mrf4 (also known as Myf6 and herculin), the third myogenic determination factor. All three determination factors then switch on myogenin – the only MRF that is solely a differentiation factor – and all four MRFs are then able to activate the myogenic differentiation program. This is actually the conserved part of skeletal myogenesis: myogenesis in the trunk and face use different

cofactors, but once one or all MRF's are turned on, skeletal myogenesis is fated to occur.

The C2C12 muscle cell line is a long-studied (Yaffe and Saxel 1977; Casas-Delucchi, Brero et al. 2011) stable myogenic cell line that is propagated in culture in a committed myoblast-like state. Upon withdrawal from the cell cycle, it undergoes cyto-differentiation to a myocyte-like state and partial fusion into myotube structures (Fig. I-2).

In our C2C12 RNA-Seq data, NRSF, a repressor expressed at 1 copy per cell (Johnson, Mortazavi et al. 2007), has a value of 5FPKM. Using this as an initial practical threshold to define biologically pertinent expression, these cells do not express either Pax3 or Pax7. However, some of the Six and Eya factors are significantly expressed (Six1: 46FPKM blast, 36FPKM cyte; Six4: 5FPKM blast, 4FPKM cyte; Eya3: 5FPKM blast, 7FPKM cyte; Eya4: 18FPKM blast, 16FPKM cyte). Further, Myf5 is expressed at relatively low levels (10FPKM blast, 11FPKM cyte) and Mrf4 is essentially off (0FPKM blast, 3FPKM cyte). By contrast, the system strongly expresses myoD in both myoblasts and myocytes (166FPKM blast, 200FPKM cyte). C2C12s differentiate and ultimately, most will fuse into multi-nucleated "myotubes." In this thesis, the main focus in on gene regulation, so I choose to call all differentiating cells "myocytes," and I base this designation on their gene expression status, whether they are mononucleate or multinucleate. C2C12 myocytes express the well-studied myogenic transcription factors, enzymes, and proteins including myogenin (17 FPKM blast, 1,026 FPKM cyte), myoglobin (2 FPKM blast, 642 FPKM cyte), muscle creatine kinase (1 FPKM blast, 759 FPKM cyte), myosin heavy chain Mybph (1 FPKM blast, 841 FPKM cyte), actin Acta1 (12 FPKM blast, 4,337 FPKM cyte), and myosin light chain Myl4 (3 FPKM blast, 2,787 FPKM cyte). Altogether, C2C12s have 7,325 genes expressed above 5 FPKM in one timepoint or the other, of which 714 genes are up-regulated more than 3x and 363 genes

downregulated more than 3x (Fig. I-3). I use these unambiguous developmentally defined gene sets for clarity of biological analysis when examining the association of connectivity with putatively regulated sets of genes.

I.4: A history of chromatin topology

It is now understood that chromatin structure is a vital component of transcriptional regulation, but at the time enhancers and promoters were being defined, work on chromatin topology and dynamics was conceptually a separate field. Chromatin biology of the time necessarily focused on the major structural changes of chromatin and sought to understand which molecules were responsible. This gene-non-centric approach to understanding the nucleus, however, revealed many events that occurred at the same time as gene expression. Many of these events are now understood to be required for the proper regulation of genes.

From a variety of traditional microscopy and molecular labeling experiments, it has been known since the late 1970s that there are topologically independent loops or small domains of chromatin, perhaps governed by attachment to the nuclear matrix (Berezney and Coffey 1974; Benyajati and Worcel 1976; Cook and Brazell 1978; Lebkowski and Laemmli 1982; Gasser and Laemmli 1986; Gasser and Laemmli 1987). Chromosomes themselves also have relatively distinct domains in the nucleus (Cremer, Cremer et al. 1982), and later methods substantiated this, though noting that there are intermingling interchromosomal interactions between chromosomes.

When chromatin is viewed in bulk, rather than at specific loci, it is apparent that active regions co-localize with each other and repressed heterochromatin also co-localize with each other. This is such a general phenomenon that it is observable at the level of light microcsopy. The interphase nucleus, when stained in certain ways, exhibits a dark and light banding pattern that suggested the nucleus is filled by a series of

alternating discs of more and less dense chromatin. D-bands, which overlap spatially with early DNA replicating Giemsa light bands (Kerem, Goitein et al. 1984), correspond to transcriptionally active (Goldman, Holmquist et al. 1984). At the time, an accepted interpretation of the results was that genes actively co-localized with each other, but another possible interpretation is that highly condensed heterochromatin is segregated. These interpretations are not mutually exclusive, and some current notions, such as the hypothetical “transcription factory” that will be discussed later, question whether co-localization of active genes is an active or a passive process.

More recently, it was shown via DNA-FISH that active and silent genes are in different areas of the nucleus from each other (Kosak, Skok et al. 2002). DamID, a technique that uses an engineered protein by joining DNA adenine methyltransferase (DAM) to the binding portion of a chromatin protein, and which methylates regions of the genome to which the chromatin protein is bound, has also showed that heterochromatin preferentially interacts with hererochromatin (van Steensel and Henikoff 2000); this was later substantiated by 6C (Tiwari, Cope et al. 2008), one of the 3C family of assays – the only assays since FISH that are able to quantify 3D chromatin interactions. One domain of active chromatin corresponding to a nuclear subcompartment was investigated using 3C, and it was found that chromatin within the domain interacted locally with other members of chromatin in the domain, and with no chromatin without (Chubb, Boyle et al. 2002).

I.5: Enhancer and Promoter multiplicity

In studying developmentally regulated genes and in looking for enhancers that recapitulated the native expression patterns, it was soon found that some genes could be influenced by multiple enhancers. Some of this was done in my system, mouse myogenesis. In the case of MYOG, although in some assays it appeared that the

promoter by itself was able to recapitulate proper gene expression (Buonanno, Edmondson et al. 1993), it soon became apparent that there were at least two enhancers required, one for proper expression in the limb and one for proper expression in the somites (Cheng, Wallace et al. 1993). In the case of MyoD, two enhancers are currently known. The DRR enhancer, which is 5 kb away (Tapscott, Lassar et al. 1992), first appeared to drive proper expression of MyoD in adult muscle (Hughes, Taylor et al. 1993), but it eventually became apparent that this enhancer is required for expression in adult muscle but not for developing muscle (Asakura, Lyons et al. 1995; Kablar, Krastel et al. 1997; Chen, Ramachandran et al. 2002). The second enhancer, termed the core enhancer, which is 20 kb away, while insufficient to create proper MyoD expression in adult muscle (Faerman, Goldhamer et al. 1995) is highly active during the course of muscle development and is required for expression of somites, limb buds, and branchial arches (Faerman, Goldhamer et al. 1995; Goldhamer, Brunk et al. 1995; Kablar, Asakura et al. 1998). In a third gene, MCK, which is one of the most highly expressed genes in adult muscle, there are also two known enhancers, one which synergizes with the promoter, although the promoter can function independently at a lower level. However, the small intronic enhancer is necessary for proper expression of the gene in slow-twitch muscle (Tai, Fisher-Aylor et al. 2011).

Studies of other genes in other systems also turned up multiple enhancers for well-studied genes, although sometimes the developmental necessity of these enhancers was not as easily understood. Some people even hypothesized that certain multiple enhancers were redundant (Fiering, Epner et al. 1995; Zakany, Fromental-Ramain et al. 1997; Monroe, Sleckman et al. 1999). More recent assays in the high throughput genomic era identified multiple transcription factor occupied sites, and many labs independently noted and gave names to hypothesized special enhancers. In the

Levine lab, it was noted that “shadow enhancers” occurred near developmentally important enhancers (Zeitlinger, Zinzen et al. 2007; Hong, Hendrix et al. 2008). Meanwhile, the Young and Collins labs took note of regions containing perhaps multiple enhancers or perhaps massive single enhancers binding many different factors, and called them “super enhancers” (Hnisz, Abraham et al. 2013; Whyte, Orlando et al. 2013) or “stretch enhancers” (Collins lab). When multiple putative enhancers occur near a gene, it remains an open question whether or not all of these elements are required throughout the development and lifespan of the animal, whether they may be redundant, or whether they are even enhancers at all.

In addition to studying enhancers in conjunction with single promoters, and as expected from the twin observations that active genes associate with other active genes and many genes co-occur with closely related genes on the genome, it has been observed that multiple promoters can interact with one enhancer. In a type of interaction called “promoter competition,” the handoff of one enhancer from E-globin to B-globin was shown to underlie the developmental switch between these two globins in the developing chicken embryo (Foley and Engel 1992). Likewise, the notion of “enhancer competition,” where one enhancer competes for two different genes, was raised in conjunction with the alternate expression of H19 and IGF2; however, although this enhancer does regulate both of these genes, competition was proven not to be the cause of the handoff, but rather imprinting (Schmidt, Levorse et al. 1999). Other studies of loci containing related genes soon also uncovered single enhancers that interacted with multiple genes, and the majority of these studies hypothesized that the shared enhancers were responsible for coordinately regulating the entire multi-gene locus (IL4 (Loots, Locksley et al. 2000; Mohrs, Blankespoor et al. 2001), Myf5/6 (Carvajal, Cox et al. 2001), DLX (Sumiyama, Irvine et al. 2002), HOX (Spitz, Gonzalez et al. 2003)). One

group, interpreting this conclusion even further, suggested that the co-association of an enhancer with two globin promoters represented an “active chromatin hub” (Tolhuis, Palstra et al. 2002). Still others, looking at the same locus and perhaps drawing on the several anecdotes of single enhancers interacting with multiple promoters, combined all of these observations with the knowledge that certain factors, such as pol2, are localized in specific loci in the nucleus to create the hypothesis of the transcription factory.

In addition to the phenomena of enhancer and promoter competition, there are other factors that could affect how enhancers and promoters join together selectively, given a group of potential matches.

I.6: Promoter and Enhancer agreement

In 1980 an efficient conserved promoter was discovered in bacteria (Siebenlist, Simpson et al. 1980). However, it quickly became apparent that eukaryotic promoters had a much greater range of diversity. The TATA box was the first eukaryotic promoter motif discovered. TATA determines transcription initiation in many non-housekeeping genes (rev. (Breathnach and Chambon 1981) and TATA is conserved from archaea to human (Reeve 2003). However, it is only found in 10 to 15% of mammalian core promoters (Kim, Barrera et al. 2005; Carninci, Sandelin et al. 2006; Cooper, Trinklein et al. 2006). Non-TATA genes have for a while been thought to be constitutively expressed, low output genes with multiple 5’ start sites (Dynan 1986).

In the past three decades, multiple other eukaryotic promoter motifs have been found, such as CCAAT boxes, GC boxes (Grosschedl and Birnstiel 1980a,b; Dierks, van Ooyen et al. 1983; McKnight and Tjian 1986), BRE upstream (Lagrange, Kapanidis et al. 1998), BRE downstream (Deng and Roberts 2005), initiator (Inr) (Smale and Baltimore 1989), motif 10 element (MTE) (Burke and Kadonaga 1997; Kutach and Kadonaga 2000), and the DPE motif (Burke and Kadonaga 1996; Burke and Kadonaga 1997). The

latter three motifs co-occur in promoters with strict spacing requirements (Burke and Kadonaga 1997; Kutach and Kadonaga 2000), while the former motifs co-occur in a functionally different set of promoters. The currently popular view is that TATA and its co-occurring motifs are used to cause focused initiation, which is used in about 35% of vertebrate genes including most of the known differentially regulated genes (rev. (Juven-Gershon and Kadonaga 2010). Also in focused promoters are DPE and its related motifs, Inr and MTE, which co-occur given strict spacing requirements (Kutach and Kadonaga 2000; Lim, Santoso et al. 2004). Certain enhancers preferentially connect to TATA over DPE's promoters and vice versa (Ohtsuki, Levine et al. 1998; Butler and Kadonaga 2001). In fact, this preference is important in how caudal regulates HOX genes (Juven-Gershon, Hsu et al. 2008), and some transcription factors such as NC2 (Willy, Kobayashi et al. 2000) and MOT1 block TATA function while activating DPE and vice versa (Hsu, Juven-Gershon et al. 2008; van Werven, van Bakel et al. 2008). Although these well-characterized instances of promoter and enhancer selectivity occur only in the approximately 30% of genes that are highly expressed or differentially regulated, it stands to reason that similar selectivity may occur in at least a portion of the remaining 70% of genes. Less studied are dispersed promoters, which generally lack both TATA and DPE and their related motifs (Carninci, Sandelin et al. 2006; Sandelin, Carninci et al. 2007), although dispersed initiation is used in the majority of eukaryotic genes (Smale and Kadonaga 2003; Carninci, Sandelin et al. 2006; Juven-Gershon, Hsu et al. 2006; Juven-Gershon, Hsu et al. 2008).

I.7: Enhancer assembly and dynamics

The initial view of enhancer assembly was that the transcription factor would bind to the recognition site and remain there in a static fashion (Becker, Renkawitz et al. 1984). However, transcription factor occupancy, at least in a few specific cases, has

been shown to be dynamic. In a now classic set of experiments done in muscle, it was shown that MyoD requires four binding motifs in order to stably occupy muscle enhancers (4Rcat). Related to this phenomenon is the hit-and-run model of enhancer function, in which the transcription factor does not stably occupy the enhancer but recruits a set of factors that does (Suen, Berrodin et al. 1998). This model of enhancement was demonstrated for the GR transcription factor by photo bleaching GFP-tagged transcription factors and showing that it exchanged rapidly with chromatin regulatory elements (McNally, Müller et al. 2000). Like the 4Rcat experiment, this result suggests that some, perhaps many, initiation complexes are created through an equilibrium reaction of a transcription factor with its DNA motif or motifs rather than being stably occupied.

That is not to say, however, that stable initiation complexes do not exist. Some people have noticed that enhancers can have tightly clustered transcription factor binding sites, whereas other enhancers, termed modular enhancers, have more loosely clustered factor binding sites (Arnosti and Kulkarni 2005). The former case has been hypothesized to result in the formation of a stable enhanceosome structure between the enhancer and promoter (Thanos and Maniatis 1995). The enhanceosome was first characterized by the Maniatis lab using a well-characterized viral inducible enhancer that relies entirely on general transcription factors at the interferon-R gene. This enhanceosome is so stable that it was even able to be partially crystallized so that the structure of three transcription factors bound to half of the enhancer has been completely described (Panne, Maniatis et al. 2007). Perhaps these two different models of initiation complex formation reflect the biological necessity for certain promoters to be regulated slowly and precisely, like interferon-R, while others must assemble quickly and are therefore more structurally loose, like GR and the model of 4Rcat. If many genes

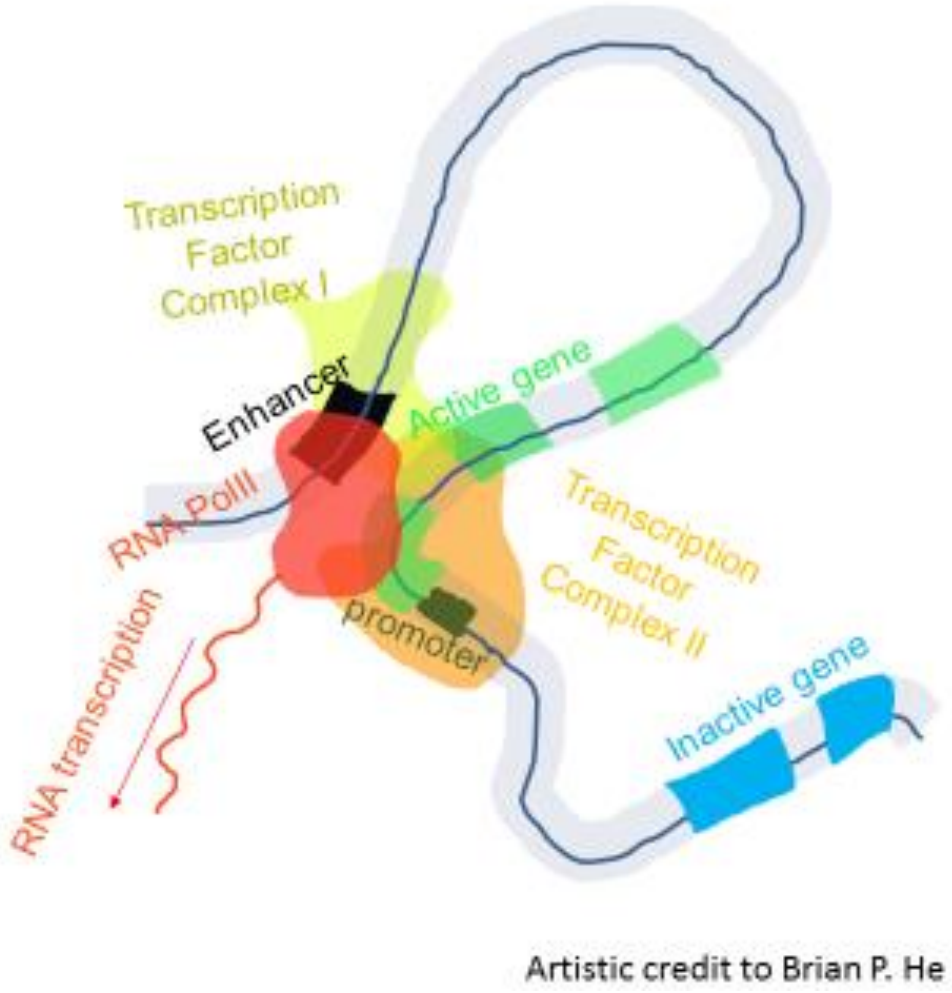
are regulated in the quicker and looser way, it would be reflective in the appearance of multiple different enhancers interacting with one promoter, since some instances of all the dynamic possibilities would be captured.

I.8: Possible sequestration of important transcriptional machinery

The notion of the transcription factory is currently the most popular interpretation of the result from microscopy that certain factors, in this case pol2, occur in a limited number of foci per cell (Osborne, Chakalova et al. 2004). However, other factors also appear to be sequestered in the nucleus. Many transcription factors, such as GR (Htun, Barsony et al. 1996) and Myog (unpublished observations), are also visible in microscopy as bright foci in the nucleus. In addition to these transcription factors, which are found both in enhancers and promoters, it has also been noted that basal transcription factors can be sequestered. During muscle development, for example, the canonical TFIID complex is replaced by a complex made up of TRF3 in place of TBP and TAF3 in place of one of the canonical TAFs (Deato and Tjian 2007; Deato, Marr et al. 2008). These non-canonical basal transcription factors have also been implicated in hematopoiesis (Hart, Raha et al. 2007).

Figures for Chapter I

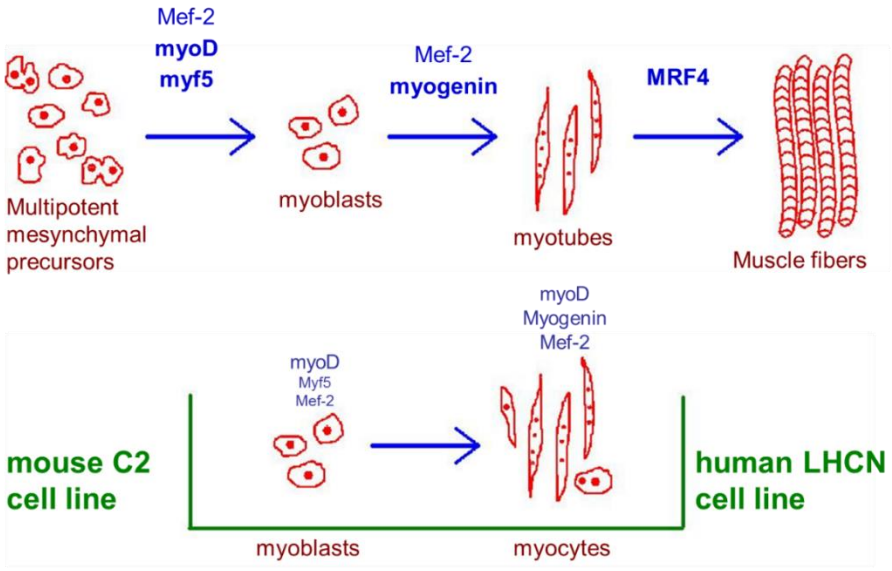
Figure I-1



Artistic credit to Brian P. He

Figure I-1: Schematic depicting the looping model of enhancement. *An enhancer (black square) and the promoter of a gene (green squares) are each bound by sets of transcription factors which interact with each other physically (green and orange blobs). The complex created by this interaction recruits pol2 (red blob), which activates the gene.*

Figure I-2



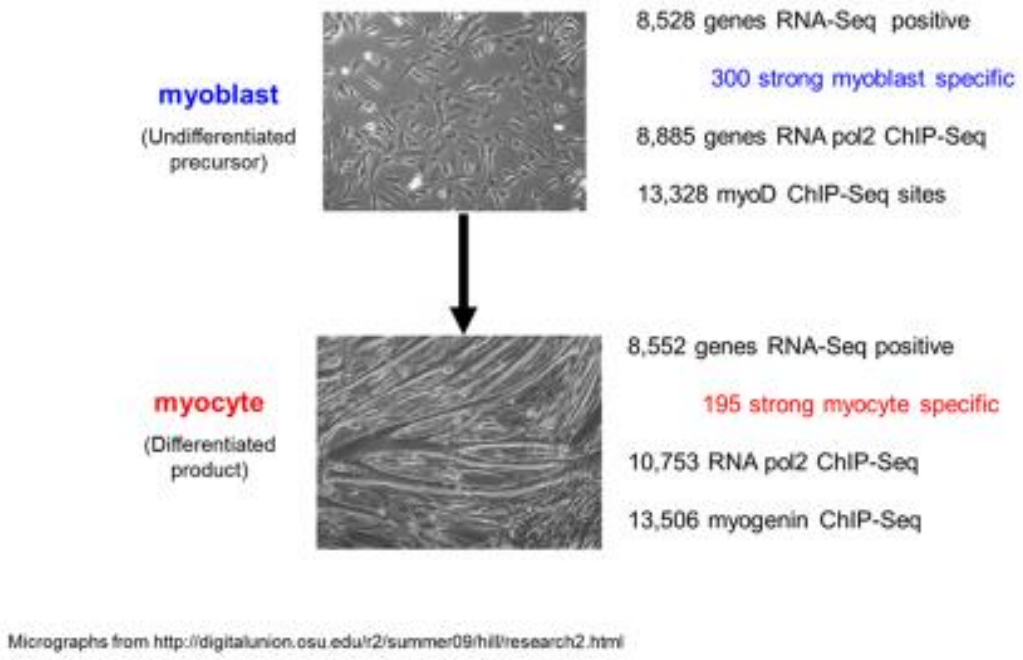


Figure I-2: C2C12 skeletal muscle differentiation. *A. Top: Four major stages of in vivo skeletal muscle development and the transcription factors that are causatively expressed to do this. Bottom: The two stages of C2C12 skeletal muscle development*

and the levels of the same transcription factors. C2C12's approximate the myoblast to myotube phase of terminal skeletal myogenesis. B. The stages of C2C12 skeletal muscle development illustrating different metrics of activity in each stage: RNA expression, pol2 occupancy, and myogenic transcription factor occupancy. Data and numbers courtesy of G Kwan, A Mortazavi, and A Kirilusha.

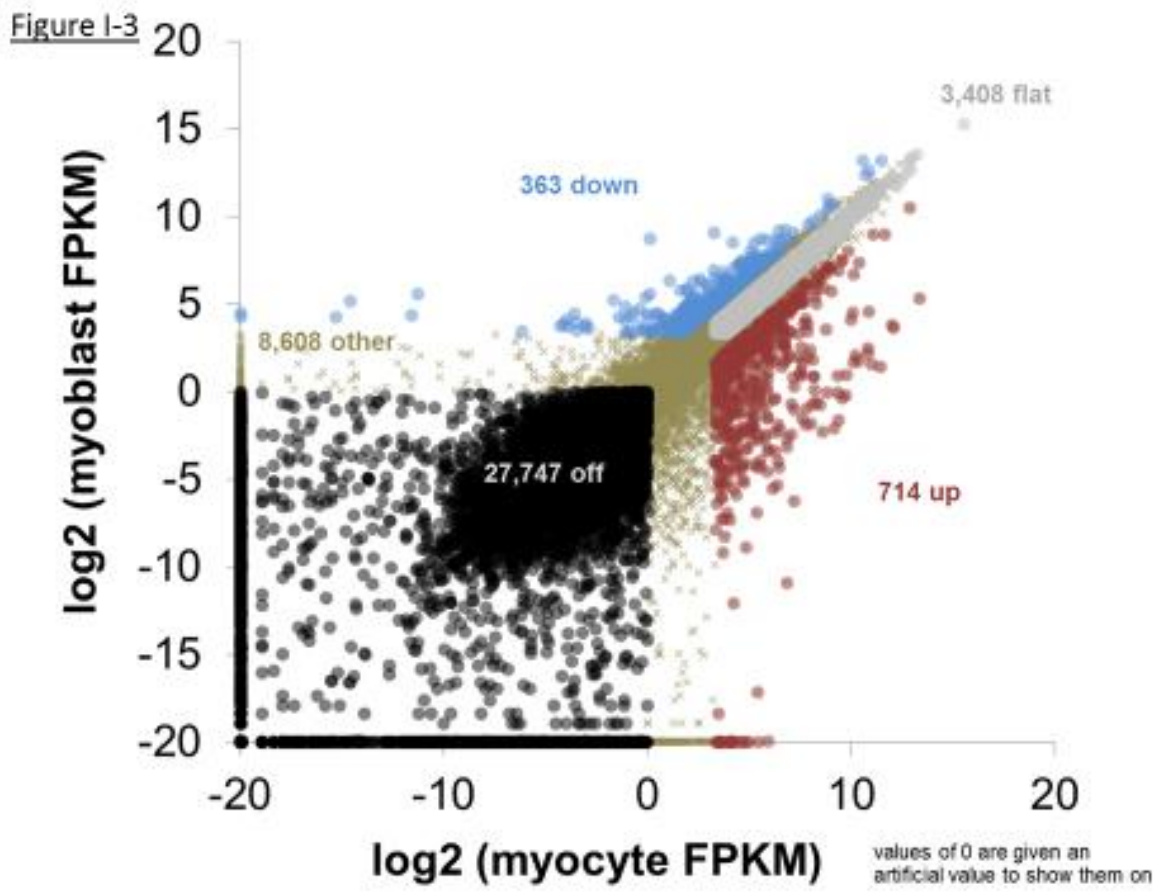


Figure I-3: Expression of all genes in C2C12 myoblasts and myocytes. *With respect to RNA expression analysis presented in later thesis chapters, the values graphed are assigned to gene promoter candidate "Wellington" graph vertices, which are defined in Chapter II. Off genes (black) are never expressed above 1FPKM. The three "special" groups of genes – up (red),*

Sources for Chapter I

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz and D. Duboule (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." Science **340**(6137): 1234167.

Arnosti, D. N. and M. M. Kulkarni (2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" J Cell Biochem **94**(5): 890-898.

Asakura, A., G. E. Lyons and S. J. Tapscott (1995). "The regulation of MyoD gene expression: conserved elements mediate expression in embryonic axial muscle." Dev Biol **171**(2): 386-398.

Banerji, J., S. Rusconi and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Becker, P., R. Renkawitz and G. Schutz (1984). "Tissue-specific DNaseI hypersensitive sites in the 5'-flanking sequences of the tryptophan oxygenase and the tyrosine aminotransferase genes." EMBO J **3**(9): 2015-2020.

Benoist, C. and P. Chambon (1981). "In vivo sequence requirements of the SV40 early promotor region." Nature **290**(5804): 304-310.

Benyajati, C. and A. Worcel (1976). "Isolation, characterization, and structure of the folded interphase genome of Drosophila melanogaster." Cell **9**(3): 393-407.

Berezney, R. and D. S. Coffey (1974). "Identification of a nuclear protein matrix." Biochem Biophys Res Commun **60**(4): 1410-1417.

Berghella, L., L. De Angelis, T. De Buysscher, A. Mortazavi, S. Biressi, S. V. Forcales, D. Sirabella, G. Cossu and B. J. Wold (2008). "A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle." Genes & Development **22**(15): 2125-2138.

Blau, H. M., C. P. Chiu and C. Webster (1983). "Cytoplasmic activation of human nuclear genes in stable heterocaryons." Cell **32**(4): 1171-1180.

Blau, H. M., G. K. Pavlath, E. C. Hardeman, C. P. Chiu, L. Silberstein, S. G. Webster, S. C. Miller and C. Webster (1985). "Plasticity of the differentiated state." Science **230**(4727): 758-766.

Breathnach, R. and P. Chambon (1981). "Organization and expression of eucaryotic split genes coding for proteins." Annu Rev Biochem **50**: 349-383.

Brent, R. and M. Ptashne (1984). "A bacterial repressor protein or a yeast transcriptional terminator can block upstream activation of a yeast gene." Nature **312**(5995): 612-615.

Brown, K. E., S. S. Guest, S. T. Smale, K. Hahm, M. Merkenschlager and A. G. Fisher (1997). "Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin." Cell **91**(6): 845-854.

Buckingham, M. and P. W. Rigby (2014). "Gene regulatory networks and transcriptional mechanisms that control myogenesis." Dev Cell **28**(3): 225-238.

Buonanno, A., D. G. Edmondson and W. P. Hayes (1993). "Upstream sequences of the myogenin gene convey responsiveness to skeletal muscle denervation in transgenic mice." Nucleic Acids Res **21**(24): 5684-5693.

Burke, T. W. and J. T. Kadonaga (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters." Genes Dev **10**(6): 711-724.

Burke, T. W. and J. T. Kadonaga (1997). "The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila." Genes Dev **11**(22): 3020-3031.

Butler, J. E. and J. T. Kadonaga (2001). "Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs." Genes Dev **15**(19): 2515-2519.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Carvajal, J. J., D. Cox, D. Summerbell and P. W. Rigby (2001). "A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and maintenance of gene expression during muscle development." Development **128**(10): 1857-1868.

Casas-Delucchi, C. S., A. Brero, H. P. Rahn, I. Solovei, A. Wutz, T. Cremer, H. Leonhardt and M. C. Cardoso (2011). "Histone acetylation controls the inactive X chromosome replication dynamics." Nat Commun **2**: 222.

Chen, J. C. J., R. Ramachandran and D. J. Goldhamer (2002). "Essential and Redundant Functions of the MyoD Distal Regulatory Region Revealed by Targeted Mutagenesis." Developmental Biology **245**(1): 213-223.

Cheng, T. C., M. C. Wallace, J. P. Merlie and E. N. Olson (1993). "Separable regulatory elements governing myogenin transcription in mouse embryogenesis." Science **261**(5118): 215-218.

Chubb, J. R., S. Boyle, P. Perry and W. A. Bickmore (2002). "Chromatin motion is constrained by association with nuclear compartments in human cells." Current Biology **12**(6): 439-445.

Chung, J. H., A. C. Bell and G. Felsenfeld (1997). "Characterization of the chicken beta-globin insulator." Proc Natl Acad Sci U S A **94**(2): 575-580.

Chung, J. H., M. Whiteley and G. Felsenfeld (1993). "A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila." Cell **74**(3): 505-514.

Cook, P. R. and I. A. Brazell (1978). "Spectrofluorometric measurement of the binding of ethidium to superhelical DNA from cell nuclei." Eur J Biochem **84**(2): 465-477.

Cooper, S. J., N. D. Trinklein, E. D. Anton, L. Nguyen and R. M. Myers (2006). "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome." Genome Res **16**(1): 1-10.

Courey, A. J., S. E. Plon and J. C. Wang (1986). "The use of psoralen-modified DNA to probe the mechanism of enhancer action." Cell **45**(4): 567-574.

Cremer, T., C. Cremer, T. Schneider, H. Baumann, L. Hens and M. Kirsch-Volders (1982). "Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments." Hum Genet **62**(3): 201-209.

Deato, M. D., M. T. Marr, T. Sottero, C. Inouye, P. Hu and R. Tjian (2008). "MyoD targets TAF3/TRF3 to activate myogenin transcription." Mol Cell **32**(1): 96-105.

Deato, M. D. and R. Tjian (2007). "Switching of the core transcription machinery during myogenesis." Genes Dev **21**(17): 2137-2149.

Deng, W. and S. G. Roberts (2005). "A core promoter element downstream of the TATA box that is recognized by TFIIB." Genes Dev **19**(20): 2418-2423.

Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser and C. Weissmann (1983). "Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells." Cell **32**(3): 695-706.

Dierks, P., A. van Ooyen, N. Mantei and C. Weissmann (1981). "DNA sequences preceding the rabbit beta-globin gene are required for formation in mouse L cells of beta-globin RNA with the correct 5' terminus." Proc Natl Acad Sci U S A **78**(3): 1411-1415.

Dynan, W. S. (1986). "Promoters for housekeeping genes." Trends in Genetics **2**: 196-197.

Faerman, A., D. J. Goldhamer, R. Puzis, C. P. Emerson, Jr. and M. Shani (1995). "The distal human myoD enhancer sequences direct unique muscle-specific patterns of lacZ expression during mouse development." Dev Biol **171**(1): 27-38.

Faye, G., D. W. Leung, K. Tatchell, B. D. Hall and M. Smith (1981). "Deletion mapping of sequences essential for in vivo transcription of the iso-1-cytochrome c gene." Proc Natl Acad Sci U S A **78**(4): 2258-2262.

Fiering, S., E. Epner, K. Robinson, Y. Zhuang, A. Telling, M. Hu, D. I. Martin, T. Enver, T. J. Ley and M. Groudine (1995). "Targeted deletion of 5'HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus." Genes Dev **9**(18): 2203-2213.

Foley, K. P. and J. D. Engel (1992). "Individual stage selector element mutations lead to reciprocal changes in beta- vs. epsilon-globin gene transcription: genetic confirmation of promoter competition during globin gene switching." Genes Dev **6**(5): 730-744.

Francastel, C., M. C. Walters, M. Groudine and D. I. Martin (1999). "A functional enhancer suppresses silencing of a transgene and prevents its localization close to centrometric heterochromatin." Cell **99**(3): 259-269.

Galande, S., P. K. Purbey, D. Notani and P. P. Kumar (2007). "The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1." Curr Opin Genet Dev **17**(5): 408-414.

Gasser, S. M. and U. K. Laemmli (1986). "The organisation of chromatin loops: characterization of a scaffold attachment site." EMBO J **5**(3): 511-518.

Gasser, S. M. and U. K. Laemmli (1987). "Improved methods for the isolation of individual and clustered mitotic chromosomes." Exp Cell Res **173**(1): 85-98.

Gerasimova, T. I. and V. G. Corces (1998). "Polycomb and trithorax group proteins mediate the function of a chromatin insulator." Cell **92**(4): 511-521.

Gerasimova, T. I., D. A. Gdula, D. V. Gerasimov, O. Simonova and V. G. Corces (1995). "A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation." Cell **82**(4): 587-597.

Gillies, S. D., S. L. Morrison, V. T. Oi and S. Tonegawa (1983). "A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene." Cell **33**(3): 717-728.

Gluzman, Y., J. F. Sambrook and R. J. Frisque (1980). "Expression of early genes of origin-defective mutants of simian virus 40." Proc Natl Acad Sci U S A **77**(7): 3898-3902.

Goldhamer, D. J., B. P. Brunk, A. Faerman, A. King, M. Shani and C. P. Emerson, Jr. (1995). "Embryonic activation of the myoD gene is regulated by a highly conserved distal control element." Development **121**(3): 637-649.

Goldman, M. A., G. P. Holmquist, M. C. Gray, L. A. Caston and A. Nag (1984). "Replication timing of genes and middle repetitive sequences." Science **224**(4650): 686-692.

Greally, J. M., D. J. Starr, S. Hwang, L. Song, M. Jaarola and S. Zemel (1998). "The mouse H19 locus mediates a transition between imprinted and non-imprinted DNA replication patterns." Hum Mol Genet **7**(1): 91-95.

Griffith, J., A. Hochschild and M. Ptashne (1986). "DNA loops induced by cooperative binding of lambda repressor." Nature **322**(6081): 750-752.

Grosschedl, R. and M. L. Birnstiel (1980). "Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo." Proc Natl Acad Sci U S A **77**(3): 1432-1436.

Grosschedl, R. and M. L. Birnstiel (1980). "Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo." Proc Natl Acad Sci U S A **77**(12): 7102-7106.

Gruss, P., R. Dhar and G. Khoury (1981). "Simian virus 40 tandem repeated sequences as an element of the early promoter." Proc Natl Acad Sci U S A **78**(2): 943-947.

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat and B. van Steensel (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Hart, D. O., T. Raha, N. D. Lawson and M. R. Green (2007). "Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3." Nature **450**(7172): 1082-1085.

Hebbes, T. R., A. L. Clayton, A. W. Thorne and C. Crane-Robinson (1994). "Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain." EMBO J **13**(8): 1823-1830.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke and R. A. Young (2013). "Super-enhancers in the control of cell identity and disease." Cell **155**(4): 934-947.

Hochschild, A., N. Irwin and M. Ptashne (1983). "Repressor structure and the mechanism of positive control." Cell **32**(2): 319-325.

Hong, J. W., D. A. Hendrix and M. S. Levine (2008). "Shadow enhancers as a source of evolutionary novelty." Science **321**(5894): 1314.

Hsu, J. Y., T. Juven-Gershon, M. T. Marr, 2nd, K. J. Wright, R. Tjian and J. T. Kadonaga (2008). "TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription." Genes Dev **22**(17): 2353-2358.

Htun, H., J. Barsony, I. Renyi, D. L. Gould and G. L. Hager (1996). "Visualization of glucocorticoid receptor translocation and intranuclear organization in living cells with a green fluorescent protein chimera." Proc Natl Acad Sci U S A **93**(10): 4845-4850.

Hughes, S. M., J. M. Taylor, S. J. Tapscott, C. M. Gurley, W. J. Carter and C. A. Peterson (1993). "Selective accumulation of MyoD and myogenin mRNAs in fast and slow adult skeletal muscle is controlled by innervation and hormones." Development **118**(4): 1137-1147.

Johnson, D. S., A. Mortazavi, R. M. Myers and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Juven-Gershon, T., J. Y. Hsu and J. T. Kadonaga (2006). "Perspectives on the RNA polymerase II core promoter." Biochem Soc Trans **34**(Pt 6): 1047-1050.

Juven-Gershon, T., J. Y. Hsu and J. T. Kadonaga (2008). "Caudal, a key developmental regulator, is a DPE-specific transcriptional factor." Genes Dev **22**(20): 2823-2830.

Juven-Gershon, T., J. Y. Hsu, J. W. Theisen and J. T. Kadonaga (2008). "The RNA polymerase II core promoter - the gateway to transcription." Curr Opin Cell Biol **20**(3): 253-259.

Juven-Gershon, T. and J. T. Kadonaga (2010). "Regulation of gene expression via the core promoter and the basal transcriptional machinery." Developmental Biology **339**(2): 225-229.

Kablar, B., A. Asakura, K. Krastel, C. Ying, L. L. May, D. J. Goldhamer and M. A. Rudnicki (1998). "MyoD and Myf-5 define the specification of musculature of distinct embryonic origin." Biochem Cell Biol **76**(6): 1079-1091.

Kablar, B., K. Krastel, C. Ying, A. Asakura, S. J. Tapscott and M. A. Rudnicki (1997). "MyoD and Myf-5 differentially regulate the development of limb versus trunk skeletal muscle." Development **124**(23): 4729-4738.

Kellum, R. and P. Schedl (1991). "A position-effect assay for boundaries of higher order chromosomal domains." Cell **64**(5): 941-950.

Kennell, D. and H. Riezman (1977). "Transcription and translation initiation frequencies of the Escherichia coli lac operon." Journal of Molecular Biology **114**(1): 1-21.

Kerem, B. S., R. Goitein, G. Diamond, H. Cedar and M. Marcus (1984). "Mapping of DNAase I sensitive regions on mitotic chromosomes." Cell **38**(2): 493-499.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov and B. Ren (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green and B. Ren (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-880.

Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher and H. Singh (2002). "Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development." Science **296**(5565): 158-162.

Kramer, H., M. Niemoller, M. Amouyal, B. Revet, B. von Wilcken-Bergmann and B. Muller-Hill (1987). "lac repressor forms loops with linear DNA carrying two suitably spaced lac operators." EMBO J **6**(5): 1481-1491.

Krebs, J. E. and M. Dunaway (1998). "The scs and scs' insulator elements impart a cis requirement on enhancer-promoter interactions." Mol Cell **1**(2): 301-308.

Kutach, A. K. and J. T. Kadonaga (2000). "The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters." Mol Cell Biol **20**(13): 4754-4764.

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg and R. H. Ebright (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB." Genes Dev **12**(1): 34-44.

Lebkowski, J. S. and U. K. Laemmli (1982). "Evidence for two levels of DNA folding in histone-depleted HeLa interphase nuclei." J Mol Biol **156**(2): 309-324.

Lim, C. Y., B. Santoso, T. Boulay, E. Dong, U. Ohler and J. T. Kadonaga (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-1617.

Ling, J. Q., T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry and A. R. Hoffman (2006). "CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1." Science **312**(5771): 269-272.

Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin and K. A. Frazer (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-140.

Mallin, D. R., J. S. Myung, J. S. Patton and P. K. Geyer (1998). "Polycomb group repression is blocked by the Drosophila suppressor of Hairy-wing [su(Hw)] insulator." Genetics **148**(1): 331-339.

McKnight, S. and R. Tjian (1986). "Transcriptional selectivity of viral genes in mammalian cells." Cell **46**(6): 795-805.

McKnight, S. L., R. C. Kingsbury, A. Spence and M. Smith (1984). "The distal transcription signals of the herpesvirus tk gene share a common hexanucleotide control sequence." Cell **37**(1): 253-262.

McNally, J. G., W. G. Müller, D. Walker, R. Wolford and G. L. Hager (2000). "The Glucocorticoid Receptor: Rapid Exchange with Regulatory Sites in Living Cells." Science **287**(5456): 1262-1265.

Mercola, M., X. F. Wang, J. Olsen and K. Calame (1983). "Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus." Science **221**(4611): 663-665.

Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, V. Afzal, H. Hadeiba, K. Shinkai, E. M. Rubin and R. M. Locksley (2001). "Deletion of a coordinate regulator of type 2 cytokine expression in mice." Nat Immunol **2**(9): 842-847.

Monroe, R. J., B. P. Sleckman, B. C. Monroe, B. Khor, S. Claypool, R. Ferrini, L. Davidson and F. W. Alt (1999). "Developmental regulation of TCR delta locus accessibility and expression by the TCR delta enhancer." Immunity **10**(5): 503-513.

Morcillo, P., C. Rosen, M. K. Baylies and D. Dorsett (1997). "Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in Drosophila." Genes Dev **11**(20): 2729-2740.

Muller, H. P. and W. Schaffner (1990). "Transcriptional enhancers can act in trans." Trends in Genetics **6**(9): 300-304.

Muller, H. P., J. M. Sogo and W. Schaffner (1989). "An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge." Cell **58**(4): 767-777.

Muller, M. M., T. Gerster and W. Schaffner (1988). "Enhancer sequences and the regulation of gene-transcription." European Journal of Biochemistry **176**(3): 485-495.

Nabirochkin, S., M. Ossokina and T. Heidmann (1998). "A nuclear matrix/scaffold attachment region co-localizes with the gypsy retrotransposon insulator sequence." J Biol Chem **273**(4): 2473-2479.

Nakagomi, K., Y. Kohwi, L. A. Dickinson and T. Kohwi-Shigematsu (1994). "A novel DNA-binding motif in the nuclear matrix attachment DNA-binding protein SATB1." Mol Cell Biol **14**(3): 1852-1860.

Namciu, S. J., K. B. Blochlinger and R. E. Fournier (1998). "Human matrix attachment regions insulate transgene expression from chromosomal position effects in Drosophila melanogaster." Mol Cell Biol **18**(4): 2382-2391.

Neuberger, M. S. (1983). "Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells." EMBO J **2**(8): 1373-1378.

Ohtsuki, S., M. Levine and H. N. Cai (1998). "Different core promoters possess distinct regulatory activities in the Drosophila embryo." Genes Dev **12**(4): 547-556.

Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik and P. Fraser (2004). "Active

genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Panne, D., T. Maniatis and S. C. Harrison (2007). "An atomic model of the interferon-beta enhanceosome." Cell **129**(6): 1111-1123.

Peric-Hupkes, D., W. Meuleman, L. Pagie, S. W. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels and B. van Steensel (2010). "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation." Mol Cell **38**(4): 603-613.

Pickersgill, H., B. Kalverda, E. de Wit, W. Talhout, M. Fornerod and B. van Steensel (2006). "Characterization of the Drosophila melanogaster genome at the nuclear lamina." Nat Genet **38**(9): 1005-1014.

Plon, S. E. and J. C. Wang (1986). "Transcription of the human beta-globin gene is stimulated by an SV40 enhancer to which it is physically linked but topologically uncoupled." Cell **45**(4): 575-580.

Reeve, J. N. (2003). "Archaeal chromatin and transcription." Mol Microbiol **48**(3): 587-598.

Robin, J. D., A. T. Ludlow, K. Batten, M. C. Gaillard, G. Stadler, F. Magdinier, W. E. Wright and J. W. Shay (2015). "SORBS2 transcription is activated by telomere position effect-over long distance upon telomere shortening in muscle cells from patients with facioscapulohumeral dystrophy." Genome Res **25**(12): 1781-1790.

Sandelin, A., P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki and D. A. Hume (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-436.

Schaffner, W., G. Kunz, H. Daetwyler, J. Telford, H. O. Smith and M. L. Birnstiel (1978). "Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing." Cell **14**(3): 655-671.

Schmidt, J. V., J. M. Levorse and S. M. Tilghman (1999). "Enhancer competition between H19 and Igf2 does not mediate their imprinting." Proceedings of the National Academy of Sciences **96**(17): 9733-9738.

Serfling, E., M. Jasin and W. Schaffner (1985). "Enhancers and eukaryotic gene-transcription." Trends in Genetics **1**(8): 224-230.

Serfling, E., A. Lubbe, K. Dorsch-Hasler and W. Schaffner (1985). "Metal-dependent SV40 viruses containing inducible enhancers from the upstream region of metallothionein genes." EMBO J **4**(13B): 3851-3859.

Siebenlist, U., R. B. Simpson and W. Gilbert (1980). "E. coli RNA polymerase interacts homologously with two different promoters." Cell **20**(2): 269-281.

Smale, S. T. and D. Baltimore (1989). "The 'initiator' as a transcription control element." Cell **57**(1): 103-113.

Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annu Rev Biochem **72**: 449-479.

Spitz, F., F. Gonzalez and D. Duboule (2003). "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." Cell **113**(3): 405-417.

Splinter, E., H. Heath, J. Kooren, R. J. Palstra, P. Klous, F. Grosveld, N. Galjart and W. de Laat (2006). "CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus." Genes Dev **20**(17): 2349-2354.

Suen, C. S., T. J. Berrodin, R. Mastroeni, B. J. Cheskis, C. R. Lyttle and D. E. Frail (1998). "A transcriptional coactivator, steroid receptor coactivator-3, selectively augments steroid receptor transcriptional activity." J Biol Chem **273**(42): 27645-27653.

Sumiyama, K., S. Q. Irvine, D. W. Stock, K. M. Weiss, K. Kawasaki, N. Shimizu, C. S. Shashikant, W. Miller and F. H. Ruddle (2002). "Genomic structure and functional control of the Dlx3-7 bigene cluster." Proc Natl Acad Sci U S A **99**(2): 780-785.

Tai, P. W. L., K. I. Fisher-Aylor, C. L. Himeda, C. L. Smith, A. P. MacKenzie, D. L. Helterline, J. C. Angello, R. E. Welikson, B. J. Wold and S. D. Hauschka (2011). "Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer." Skeletal Muscle **1**:25.

Tapscott, S. J., A. B. Lassar and H. Weintraub (1992). "A novel myoblast enhancer element mediates MyoD transcription." Mol Cell Biol **12**(11): 4994-5003.

Thanos, D. and T. Maniatis (1995). "Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome." Cell **83**(7): 1091-1100.

Theveny, B., A. Bailly, C. Rauch, M. Rauch, E. Delain and E. Milgrom (1987). "Association of DNA-bound progesterone receptors." Nature **329**(6134): 79-81.

Tiwari, V. K., L. Cope, K. M. McGarvey, J. E. Ohm and S. B. Baylin (2008). "A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations." Genome Res **18**(7): 1171-1179.

Tolhuis, B., R.-J. Palstra, E. Splinter, F. Grosveld and W. de Laat (2002). "Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus." Molecular Cell **10**(6): 1453-1465.

Udvardy, A., E. Maine and P. Schedl (1985). "The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains." J Mol Biol **185**(2): 341-358.

van Steensel, B. and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase." Nat Biotechnol **18**(4): 424-428.

van Werven, F. J., H. van Bakel, H. A. van Teeffelen, A. F. Altelaar, M. G. Koerkamp, A. J. Heck, F. C. Holstege and H. T. Timmers (2008). "Cooperative action of NC2 and Mot1p to regulate TATA-binding protein function across the genome." Genes Dev **22**(17): 2359-2369.

Weber, F., J. de Villiers and W. Schaffner (1984). "An SV40 'enhancer trap' incorporates exogenous enhancers or generates enhancers from its own sequences." Cell **36**(4): 983-992.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee and R. A. Young (2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes." Cell **153**(2): 307-319.

Wiesendanger, B., R. Lucchini, T. Koller and J. M. Sogo (1994). "Replication fork barriers in the Xenopus rDNA." Nucleic Acids Res **22**(23): 5038-5046.

Wigler, M., R. Sweet, G. K. Sim, B. Wold, A. Pellicer, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Transformation of mammalian cells with genes from procaryotes and eucaryotes." Cell **16**(4): 777-785.

Willy, P. J., R. Kobayashi and J. T. Kadonaga (2000). "A basal transcription factor that activates or represses transcription." Science **290**(5493): 982-985.

Wold, B., M. Wigler, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Introduction and expression of a rabbit beta-globin gene in mouse fibroblasts." Proc Natl Acad Sci U S A **76**(11): 5684-5688.

Yaffe, D. and O. Saxel (1977). "Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle." Nature **270**(5639): 725-727.

Zakany, J., C. Fromental-Ramain, X. Warot and D. Duboule (1997). "Regulation of number and size of digits by posterior Hox genes: a dose-dependent mechanism

with potential evolutionary implications." Proc Natl Acad Sci U S A **94**(25): 13695-13700.

Zeitlinger, J., R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young and M. Levine (2007). "Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo." Genes Dev **21**(4): 385-390.

Chapter II: Chromatin topology

II.1: Introduction

3C is a powerful assay, but it is difficult to perform and it only asks about one-to-one interactions. The mid-2000s saw a development of more and more high-throughput chromatin capture assays alongside the "next-generation" sequencer revolution, sparked by ChIP-Seq (Johnson, Mortazavi et al. 2007) and RNA-Seq (Mortazavi, Williams et al. 2008). 4C in various forms found one-to-many interactions with about a 50kb resolution (Simonis, Klous et al. 2006; Zhao, Tavoosidana et al. 2006; Wurtele and Chartrand 2006; Branco and Pombo 2006). 5C (Dostie, Richmond et al. 2006) and 6C (Tiwari, Cope et al. 2008), two different assays that are able to detect many-to-many interactions, found that many-to-many interactions are possible, instead of single CRM to single CRM.

Two genome-wide many-to-many chromatin capture assays hit the scene in 2009: Hi-C, which detects physical interactions agnostically with a 1Mb detection threshold (Lieberman-Aiden, van Berkum et al. 2009), and ChIA-PET, which improves resolution (5kb) at the expense of agnosticism by detecting only the physical interactions that also contain a ChIP-pable factor (Fullwood, Liu et al. 2009). These two assays do not agree as much as one would suspect. The Liberman-Aiden group found gene-rich euchromatin regions interacting with each other separately from gene-poor heterochromatin domains, and they reported lots of structure was constant between cell lines. The Fullwood group, which analyzed interactions containing an activating transcription factor, reported that interactions depend on factor occupancy and insinuated that interactions are transient with respect to development.

II.1.2: Major questions about general chromatin topology

When I began this project, there were thought to be three canonical types of chromatin loops: enhancer to gene, insulator to insulator, and gene start to genes end (de Wit and de Laat 2012). However, there was also hints of other, structural interactions such as those involving the nuclear matrix (Galande, Purbey et al. 2007) and the nuclear lamina (Peric-Hupkes, Meuleman et al. 2010). The field of chromatin structure has not historically interacted with gene regulation, though many elements of chromatin organization seem to be organized around active and inactive genes. In fact, it is unknown to what elements, overall, active genes connect physically. Are enhancers most common? Insulators or silencers? What about structural elements? Are the connected elements local, or are 1Mb interactions like at Shh common? And how often do multiple genes share an enhancer?

I created a set of pol2 ChIA-PET datasets in two cell states in order to detect physical interactions of distal elements with active genes. In order to determine which detected connections exist independently of a pol2 ChIP-Seq, as well as to better detect classically defined myogenic enhancers, I also created a separate ChIA-PET dataset for the myogenic transcription factor myogenin (Fig. II-1).

II.2: Results

II.2.1: Simplifying ChIA-PET data to elucidate the most reproducible connections

ChIA-PET poses a particular problem above and beyond the ordinary noisiness of genome-wide data in that certain areas of the genome have connectivity patterns that are extremely complex (Fig. II-2).

Other ChIA-PET bodies of work have done little to elucidate what is occurring at these complex loci (Fullwood, Liu et al. 2009; Handoko, Xu et al. 2011; Li, Ruan et al. 2012; Chepelev, Wei et al. 2012). To simplify the ChIA-PET raw data, I took a graph

theoretical approach. First, I specified a set of candidate vertices out of regions of the genome likely to be connected and removed all PETs that do not have both ends in a vertex (Fig. II-3).

I used an independent genome-wide assay, DNase-Seq, as the source of the candidate vertices, along with all annotated TSSs. This narrows the pool of connected regions of the genome to genes and occupied putative CRMs, more easily interpretable by current knowledge in the era of ENCODE. Second, for pol2 ChIA-PET, I reported as edges only the places where there were two individual occurrences of ChIA-PET raw paired tags between candidate vertices. In order to focus on the most reproducible, highest-confidence set of interactions, I performed two separate biological and technical ChIA-PET experiments for each condition and I only reported the edges found in both experiments (Fig. II-1). This purposefully sacrifices weak signal at the threshold of noise for high-confidence, reproducible signal so that I can be certain of the existence of the connections I report. A third dataset, myogenin at the myocyte timepoint, has no replicate. It will be used to determine which aspects of the pol2 ChIA-PETs are factor-dependent. Both the raw and processed data are shown for the CIG containing MyoD, one of the master regulators of myogenesis (Fig. II-4), since MyoD is a locus representative of a medium-sized one gene CIG.

A ChIA-PET edge means that there is evidence of a single physical complex that contains two regions of DNA and the factor for which the ChIP was done. Lack of a ChIA-PET edge suggests that either there is no physical connectivity between the regions, or that connectivity occurs without the presence of the ChIPped factor. A common misconception of ChIA-PET data is that it represents a complete physical connectivity map; it does not (Fig. II-5).

II.2.2: ChIA-PET general characteristics

ChIA-PET connectivity is particularly striking at the myogenic locus containing myogenin (Myog) and myosin binding heavy protein H (Mybph) (Fig. II-6).

At the myoblast timepoint, when both myogenic genes are unexpressed, no connections are recovered. However, they connect to each other as well as many nearby myog+ and myog- DNase-hypersensitive vertices. This locus with around 60 interconnected vertices is in fact spectacularly large compared to most other loci in the genome. Most CIGs are small, though large, multi-genic CIGs like myogenin number in the hundreds (Fig. II-9).

Most CIGs contain at least one gene, but surprisingly, there are CIGs that have no annotated genes. This does not appear to be a characteristic of data stringency (data not shown), so the most likely explanations are that some vertices are unannotated genes (though I used gene models bordering on the extensive), or that pol2 sometimes comes into contact with regions of the genome that don't have genes.

As for the edges themselves, most are local, and strength is inversely correlated with distance (Fig. II-7). However, there are some long edges over 50kb, even a rare few as long as the 1Mb Shh to enhancer interaction. One related property of these local edges is that the ChIA-PET CIGs themselves are relatively localized (Fig II-9). The elements that the edges connect, gene-vertices and distal-vertices, tend to be wider than unconnected candidate vertices, and gene-vertices are also wider than distal-vertices (data not shown). This is due to the merging algorithm in the creation of candidate vertices: some regions of the genome, particularly the bodies of active genes, have multiple DNase regions blanketing a small area. I have standardized edge weights to account for the differing vertex widths (and therefore edge capture likelihood) by normalizing on the basis of the connected vertex widths.

All of these properties are true for the myogenin ChIA-PET as well. However, there is one notable difference between myogenin and pol2 ChIA-PET edge strengths. Pol2 edges are strongest when they involve genes (Fig. II-8, top), and myogenin edges are strongest when they involve non-genic elements (Fig. II-8, bottom). Though there is little relationship between ChIA-PET signal and the antecedent ChIP signal (data not shown), it is likely this means that ChIA-PET signal strength is partially influenced by factor occupancy: pol2 at genes and myog at enhancers.

Since ChIA-PET is an assay done in bulk on a large cell population, there is a major question to ask: when a vertex has connections to multiple other vertices, are the interactions simultaneous or sequential? Is there any evidence for the promoter factory hypothesis (Osborne, Chakalova et al. 2004) and if so, is this the exception or the rule? I chose to use the graph theoretical concept of the clique as a way of determining the likelihood of having simultaneous interactions. A clique is a set of vertices where every vertex is connected to every other vertex. (Fig. II-10A, middle; Fig. II-10B, purple). If there are simultaneous interactions captured by ChIA-PET, they would show up as cliques, though not all cliques need be simultaneous interactions (Fig. II-11A). However, because cliques and non-cliques alike are just as susceptible to the rigorous data treatment, it is not their absolute number but the ratio between their numbers that will tell us which type of interaction is most common. This ratio is 8 to 92% regardless of the data set and data treatment (Fig. II-11B; some analyses not shown). There are indeed cliques in the ChIA-PET data, including a clique of the classic MRF myogenin connected to two other upregulated genes (Mybph and Ppfia4) and a few distal elements (Fig. II-10A, left). However, there are surprisingly few cliques genome-wide, only a few hundred overall (Fig. II-10A, right). In fact, it appears to be a general principle of these data that there is a very narrow range of observed connectivity: most CIGs have about one extra edge per three

vertices above the absolute minimum level of connectivity (Fig. II-10B, red). Taken all together, the most likely explanation for these phenomena are that most multiple interactions in the nucleus are sequential rather than simultaneous, and that instances such as the promoter factory are the exception rather than the rule.

II.3: Discussion/Conclusions

Most connections are local; Shh-length cases (1Mb) are seen but are rare. This is consistent with the notions that chromatin movement is restricted to certain subdomains such as nuclear compartments (Chubb, Boyle et al. 2002; Noordermeer, Branco et al. 2008; Noordermeer, de Wit et al. 2011).

Complexity – in the sense of there being multiple overlapping edges within a genomic region – varies widely among active genes. Most interactions are simple paired edges, but there are nevertheless hundreds of multiply interacting CIGs containing more than one gene, more than one putative CIG, and sometimes tens of each. Though most genes connect to one other element, if there is a detected connection > 10kb at all, there are hundreds of cases of genes interacting with multiple non-genic elements and of single non-genic elements interacting with multiple genes. However, complexity in the sense of the ratio of edges to vertices is surprisingly simple and invariant, suggesting that little simultaneity within these multiply connected regions is possible.

Given the first principle that most edges are local, it seems that gene neighborhood could be very important when predicting – or in enabling – the physical interactions of most active genes.

Figures for Chapter II

Figure II-1

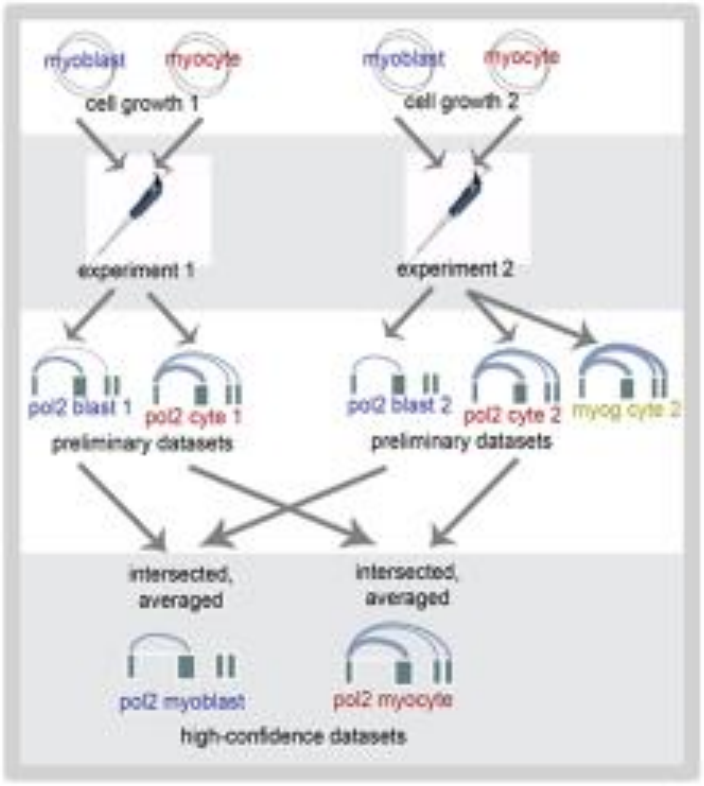


Figure II-1: Experimental design. *Two biological and technical replicates for RNA pol2 were performed for each of the developmental states, myoblast and myocyte. To ensure the edges I analyzed were real, I chose to take the high-confidence step of analyzing only the intersect edges between the two replicates. A third dataset, myogenin in myocytes, was performed in one library and analyzed to determine which data properties are in common between the pol2 and transcription factor ChIA-PET experiments*.

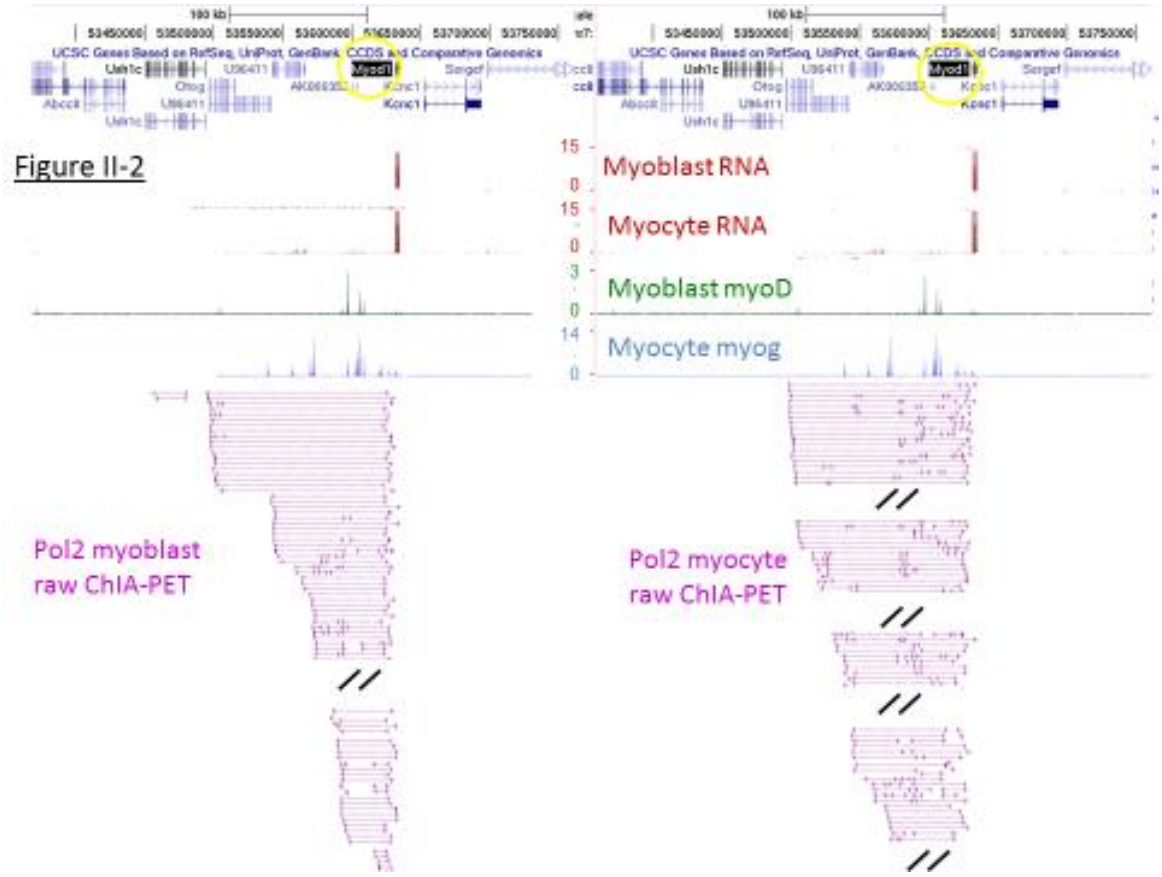


Figure II-2: ChIA-PET raw data. *ChIA-PET individual paired-end tags (light purple) are shown for pol2 myoblast (left) and myocyte (right) at the myoD locus.*

Figure II-3: ChIA-PET data treatment

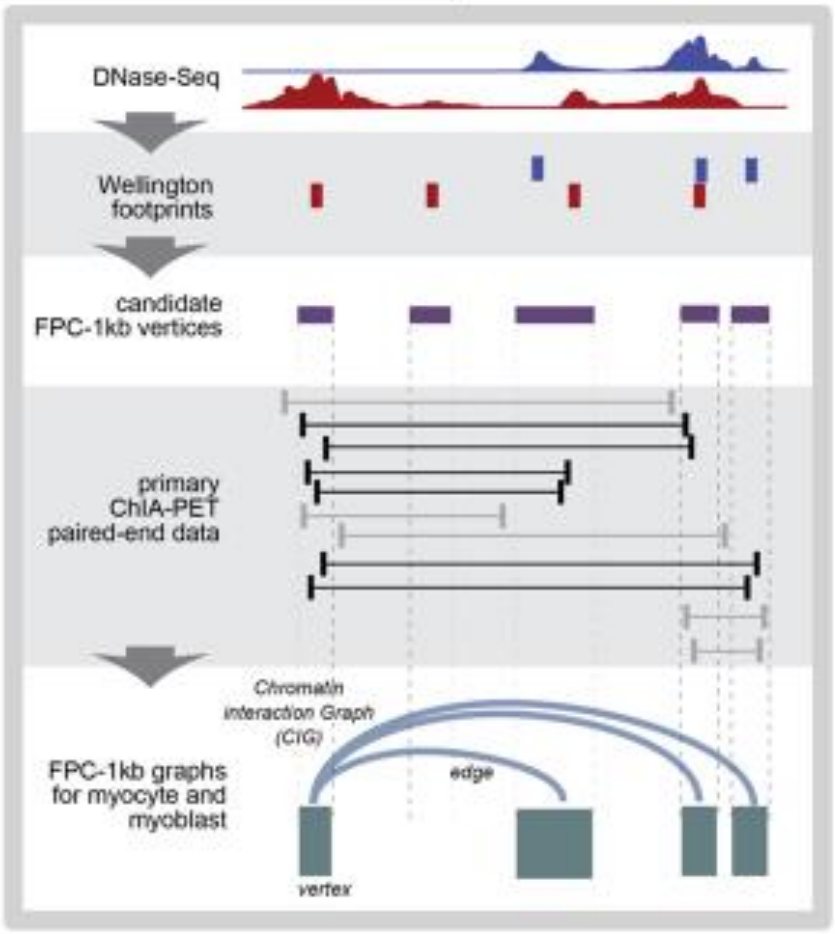


Figure II-3: ChIA-PET data processing. *DNase-Seq data was collected for C2 myoblast and myocyte timepoints (blue and red) and were used to call Wellington digital footprints. These footprints were then expanded to 1kb and combined with annotated TSS's in the genome (see methods) to create a set of candidate vertices (purple). The ChIA-PET raw paired-end tags for each timepoint (black and gray) were then mapped onto the candidate vertices to create interconnected CIGs (gray figure at bottom). Unless otherwise specified, the subgraphs reported are the intersect set of two individual ChIA-PET biological replicates.*

Figure II-4a: simplification of myoblast data at myoD

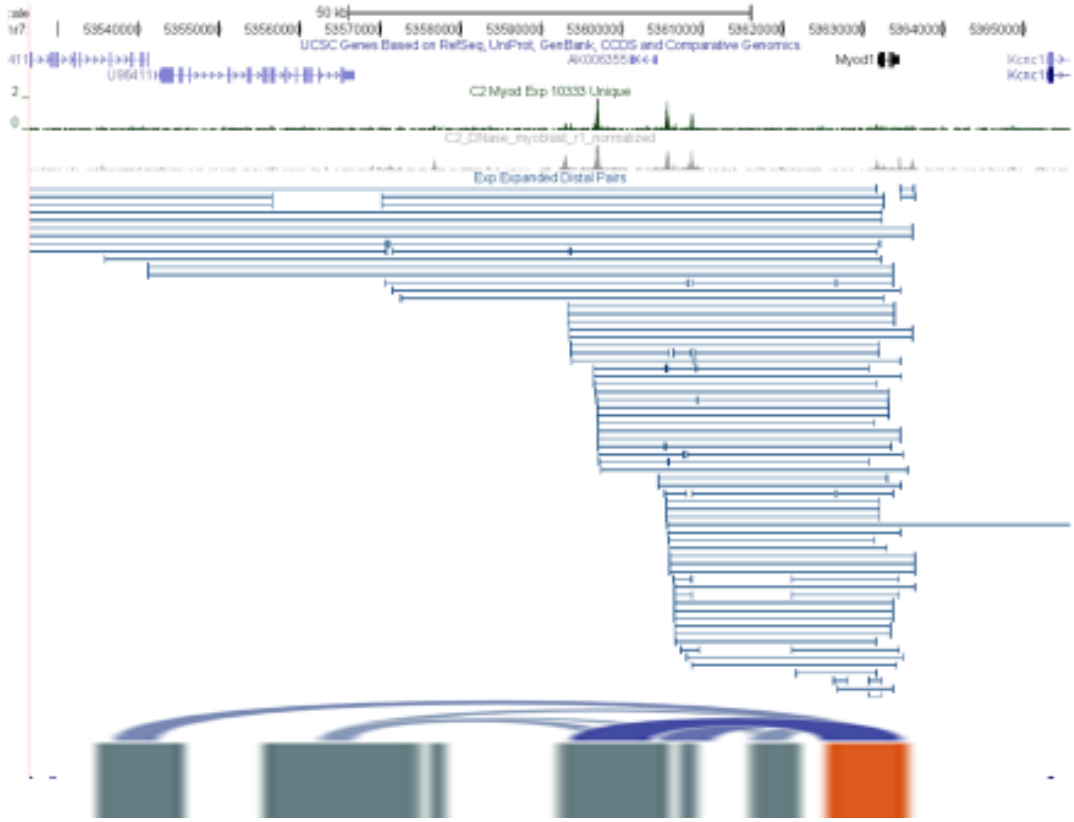


Figure II-4A: ChIA-PET data processing example: myoD*. Myoblast factor occupancy (green: myoD; gray: DNase) is shown over raw ChIA-PET paired-end tags (light blue) at the myoD locus. The numerous ChIA-PET tags are reduced into a 6-vertex CIG showing myoD connecting to 5 nearby occupied regions.*

Figure II4b: simplification of myocyte data at myoD

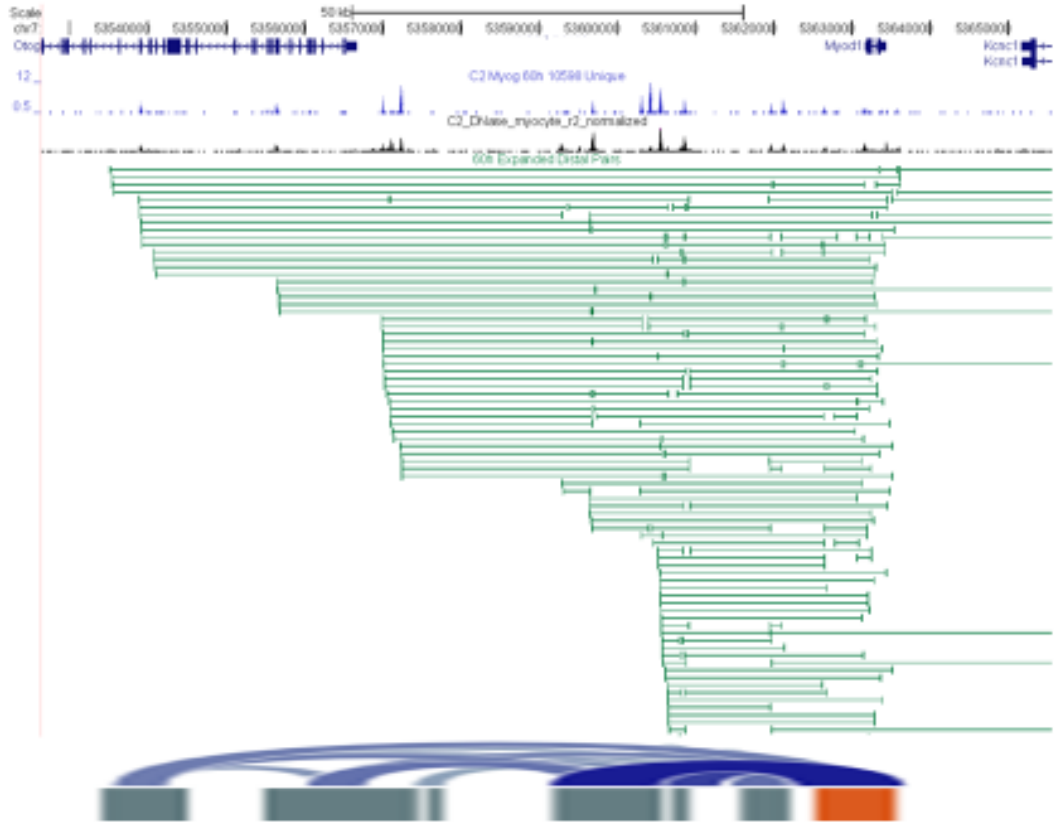


Figure II-4B: ChIA-PET data processing example: myoD. *Myocyte factor occupancy (blue: myogenin; black: DNase) is shown over raw ChIA-PET paired-end tags (green) at the myoD locus. The numerous ChIA-PET tags are reduced into a 9-vertex CIG (the leftmost 7 vertices are shown here) showing myoD connecting to 8 nearby occupied regions.*

Figure II-5

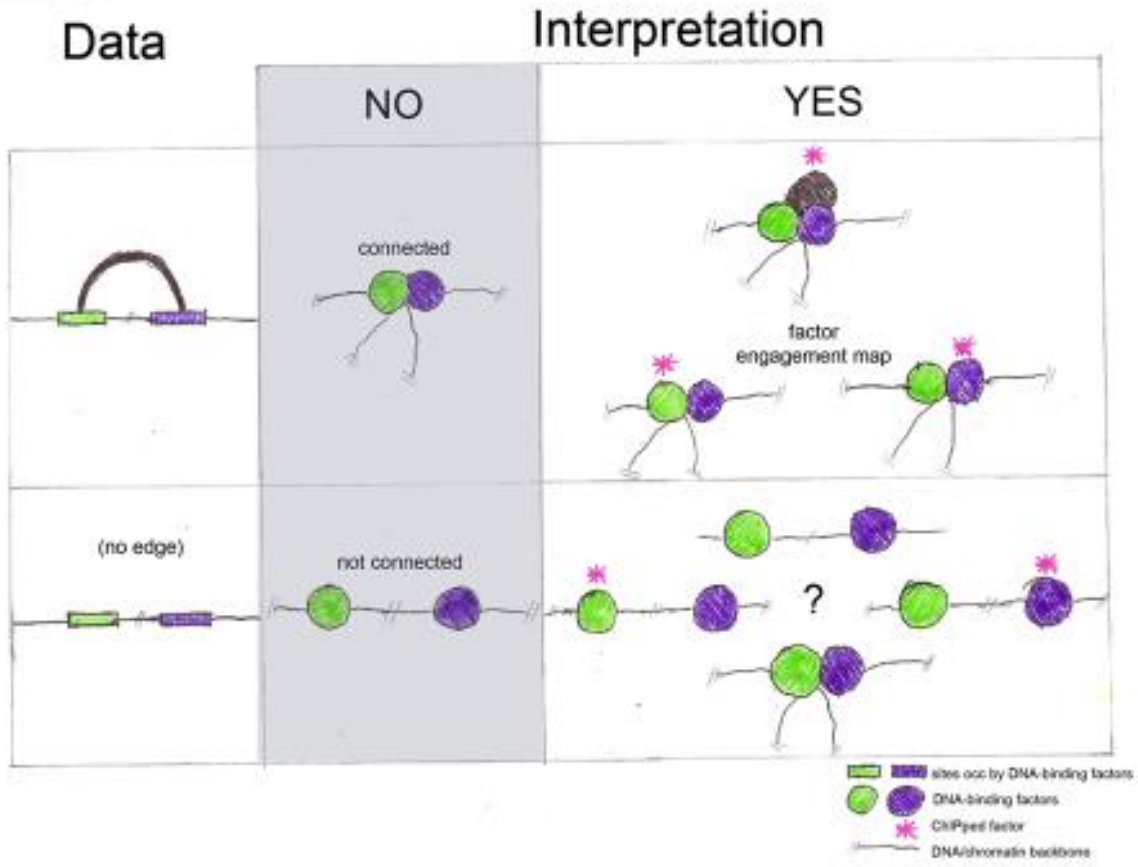


Figure II-5: How to interpret ChIA-PET data*. A ChIA-PET edge means that there is evidence of a single physical complex that contains two regions of DNA and the factor for which the ChIP was done (top right). Lack of a ChIA-PET edge suggests that either there is no physical connectivity between the regions, or that connectivity occurs without the presence of the ChIPped factor (bottom right). A common misconception of ChIA-PET data is that it represents a complete physical connectivity map (middle column); it does not.*

Figure II-6

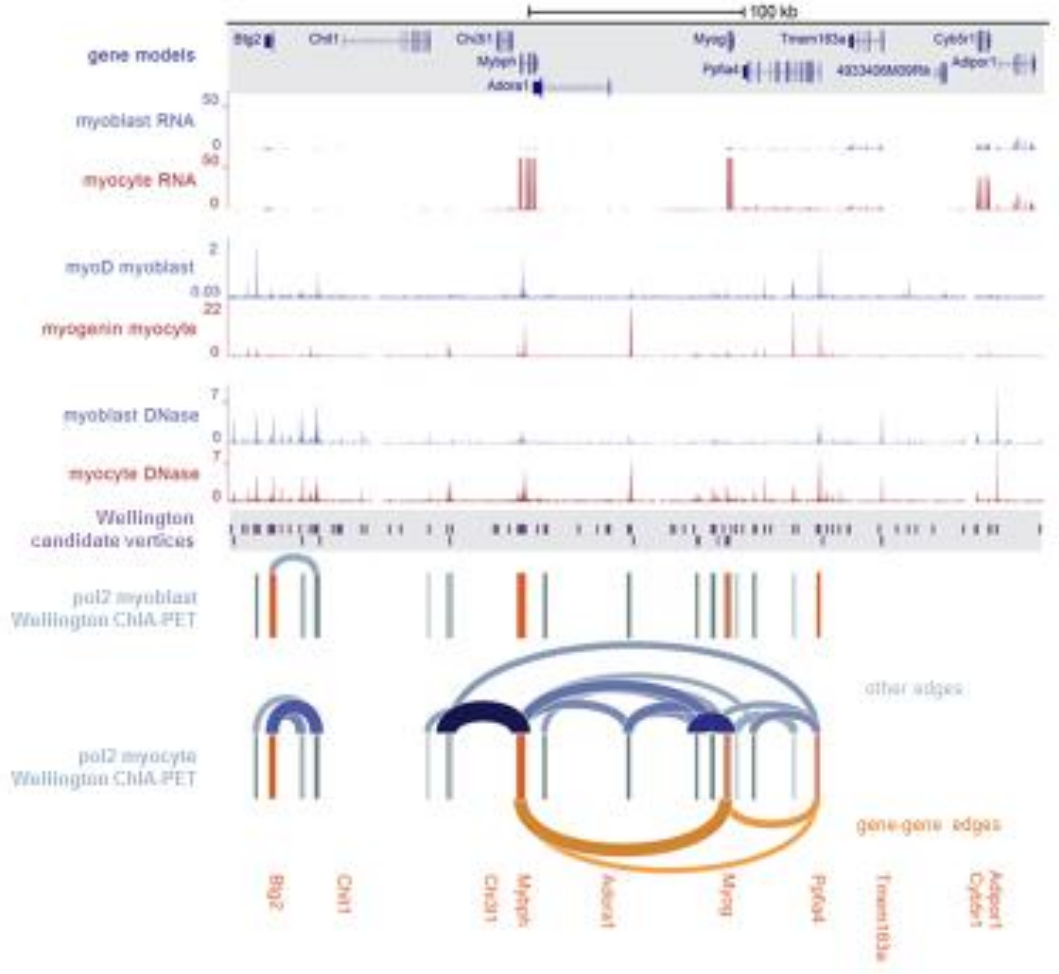


Figure II-6: Simplification of ChIA-PET data at the myogenin locus. *ChIA-PET subgraphs at the myogenin gene locus exemplify the way in which ChIA-PET data relate to other data types. Top: DNase-Seq, ChIP-Seq for muscle regulatory transcription factors (MRF's), and RNA-Seq data for myoblast and myocyte timepoints are shown. Bottom: Two sets of ChIA-PET analyses are shown. For each, candidate vertices are shown above connected myoblast and myocyte pol2 ChIA-PET CIGs. In the CIGs, the orange vertices represent TSS-containing gene vertices and the orange edges gene-to-gene connections. The blue vertices represent distal vertices and the blue edges represent distal-distal and gene-distal connections. The width and darkness of an edge represents the edge strength, which is the number of raw ChIA-PET reads contributing to the edge. There are no reported connections to myogenin in the myoblast footprint resolution dataset. CIG art at the bottom courtesy of Santiago Lombeyda.*

Figure II-7

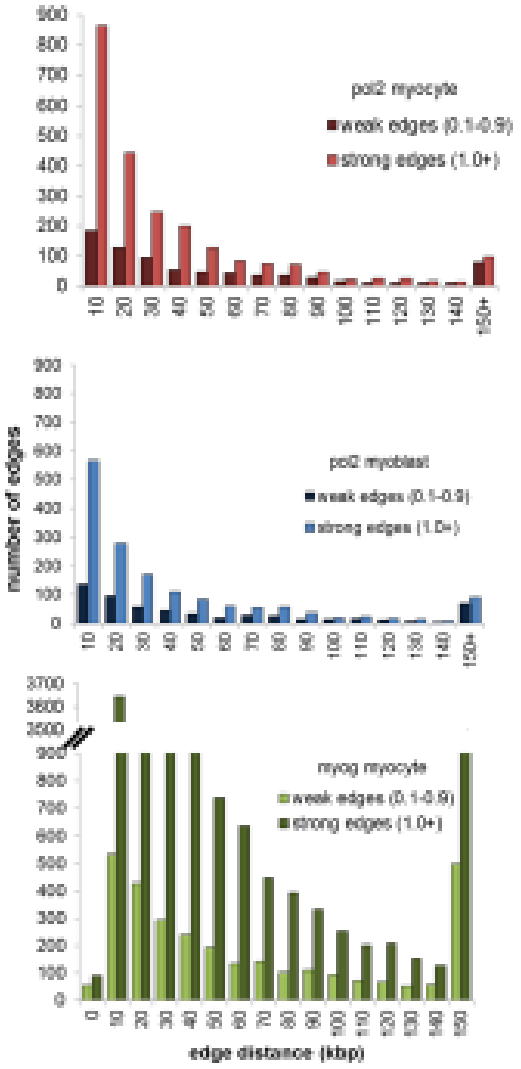


Figure II-7: ChIA-PET edge distance. *Distribution of pol2 (red, blue) and myogenin (green) ChIA-PET edge distances for weak edges <1.0 EPK (light) and strong edges (dark). Over 300 hundred ChIA-PET pol2 connections are > 100kb, and of these about half are in the high edge-weight group. Two thirds of edges are between 10-and 50kb in length, and 10kb is the threshold for inclusion of raw PETs in the analysis. There is only one myogenin dataset, while the intersects of two pol2 datasets each are shown.*

Figure II-8

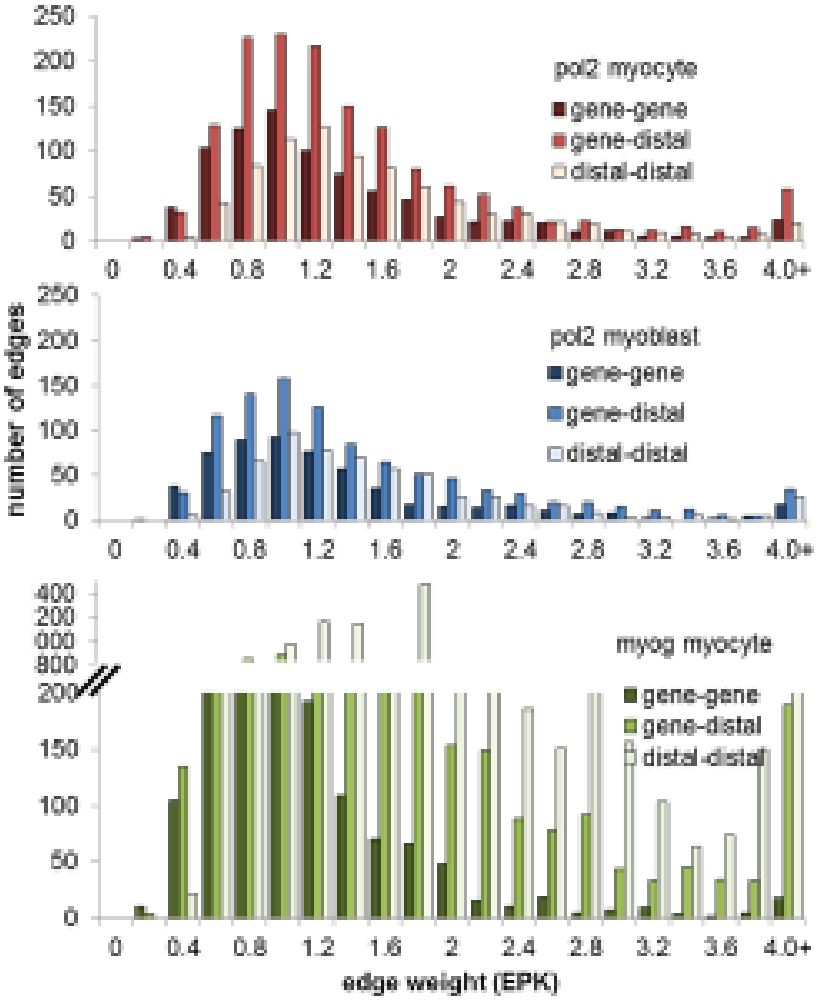


Figure II-8: ChIA-PET edge strengths. *Pol2 edges (red, blue) are strongest and most numerous for gene-vertex-containing edges (darker two colors). Myogenin edges (green) are the opposite: strongest and most numerous for edges that do not contain gene-vertices (light green).*

Figure II-9

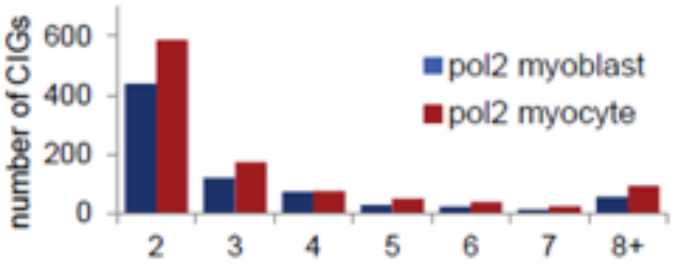
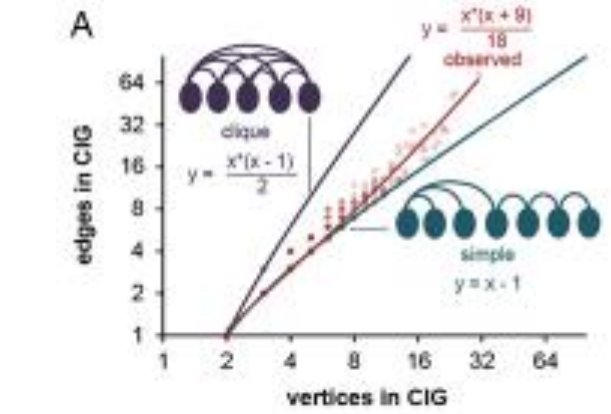


Figure II-9: General CIG characteristics. *The vast majority of CIGs are paired edges, and CIG size decreases monotonically with CIG number. Nevertheless, there are hundreds of complex CIGs with tens of vertices.*

Figure II-10



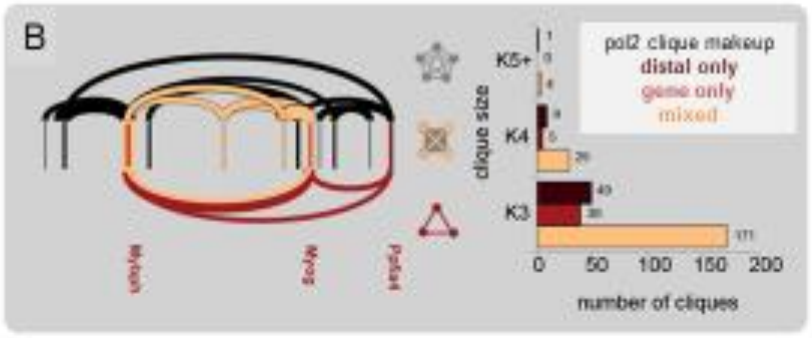


Figure II-10: Cliques and complex interactions in ChIA-PET. *(A) There is a very tight observed relationship between the number of edges and the number of vertices. Although cliques are theoretically possible, they are rarely found and instead, there is roughly one extra edge every three edges over the bare minimum connectivity.*

(B) Cliques, special CIGs in which every vertex is connected to every other vertex (middle column), and exemplified in the myogenin locus (left, red and orange edges) are present in ChIA-PET data but rare. Overall, there are only a few hundred cliques in the genome (right). Wellington pol2 myocyte intersect data are shown.

Figure II-11

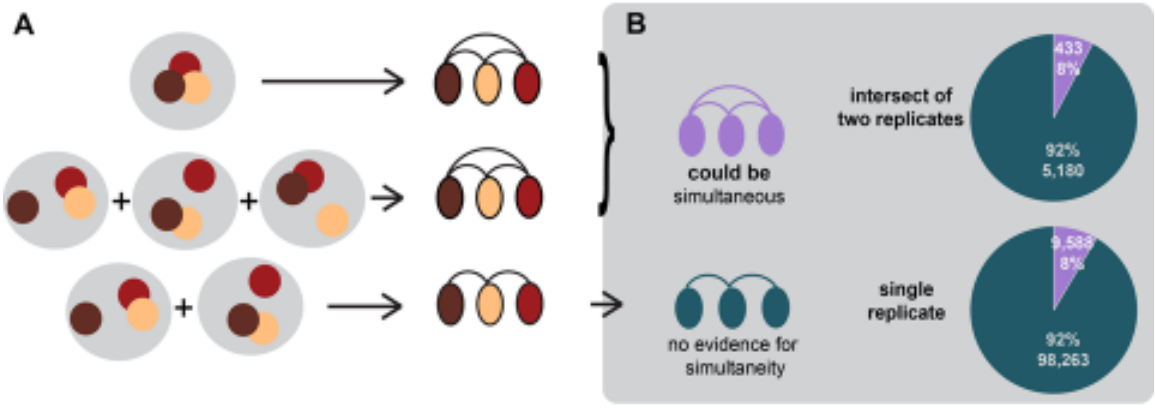


Figure II-11: Complex interactions are rare. *(A) Possible arrangements of chromatin in individual nuclei within a cell sample (left), and what the resulting CIGs would look like (right).*

(B) The number of vertex triplets that are fully connected cliques (purple) is only 8%, and this is true even when looking at less stringent ChIA-PET data (bottom). Wellington pol2 myocyte data are shown.

Sources for Chapter II

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz and D. Duboule (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." Science **340**(6137): 1234167.

Bailey, A. M. and J. W. Posakony (1995). "Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity." Genes Dev **9**(21): 2609-2622.

Bailey, S. D., X. Zhang, K. Desai, M. Aid, O. Corradin, R. Cowper-Sal Lari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains and M. Lupien (2015). "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters." Nat Commun **2**: 6186.

Bailey, T. L., N. Williams, C. Misleh and W. W. Li (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-373.

Banerji, J., S. Rusconi and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Benoist, C. and P. Chambon (1981). "In vivo sequence requirements of the SV40 early promotor region." Nature **290**(5804): 304-310.

Benyajati, C. and A. Worcel (1976). "Isolation, characterization, and structure of the folded interphase genome of Drosophila melanogaster." Cell **9**(3): 393-407.

Berezney, R. and D. S. Coffey (1974). "Identification of a nuclear protein matrix." Biochem Biophys Res Commun **60**(4): 1410-1417.

Berghella, L., L. De Angelis, T. De Buysscher, A. Mortazavi, S. Biressi, S. V. Forcales, D. Sirabella, G. Cossu and B. J. Wold (2008). "A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle." Genes & Development **22**(15): 2125-2138.

Blau, H. M., C. P. Chiu and C. Webster (1983). "Cytoplasmic activation of human nuclear genes in stable heterocaryons." Cell **32**(4): 1171-1180.

Blau, H. M., G. K. Pavlath, E. C. Hardeman, C. P. Chiu, L. Silberstein, S. G. Webster, S. C. Miller and C. Webster (1985). "Plasticity of the differentiated state." Science **230**(4727): 758-766.

Branco, M. R. and A. Pombo (2006). "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations." PLoS Biol **4**(5): e138.

Breathnach, R. and P. Chambon (1981). "Organization and expression of eucaryotic split genes coding for proteins." Annu Rev Biochem **50**: 349-383.

Brent, R. and M. Ptashne (1984). "A bacterial repressor protein or a yeast transcriptional terminator can block upstream activation of a yeast gene." Nature **312**(5995): 612-615.

Brown, K. E., S. S. Guest, S. T. Smale, K. Hahm, M. Merkenschlager and A. G. Fisher (1997). "Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin." Cell **91**(6): 845-854.

Buckingham, M. and P. W. Rigby (2014). "Gene regulatory networks and transcriptional mechanisms that control myogenesis." Dev Cell **28**(3): 225-238.

Casas-Delucchi, C. S., A. Brero, H. P. Rahn, I. Solovei, A. Wutz, T. Cremer, H. Leonhardt and M. C. Cardoso (2011). "Histone acetylation controls the inactive X chromosome replication dynamics." Nat Commun **2**: 222.

Chambeyron, S. and W. A. Bickmore (2004). "Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription." Genes Dev **18**(10): 1119-1130.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang and K. Zhao (2012). "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization." Cell Res **22**(3): 490-503.

Cheutin, T., M. F. O'Donohue, A. Beorchia, C. Klein, H. Kaplan and D. Ploton (2003). "Three-dimensional organization of pKi-67: a comparative fluorescence and electron tomography study using FluoroNanogold." J Histochem Cytochem **51**(11): 1411-1423.

Chubb, J. R., S. Boyle, P. Perry and W. A. Bickmore (2002). "Chromatin motion is constrained by association with nuclear compartments in human cells." Current Biology **12**(6): 439-445.

Chung, J. H., A. C. Bell and G. Felsenfeld (1997). "Characterization of the chicken beta-globin insulator." Proc Natl Acad Sci U S A **94**(2): 575-580.

Chung, J. H., M. Whiteley and G. Felsenfeld (1993). "A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila." Cell **74**(3): 505-514.

Ciejek, E. M., M. J. Tsai and B. W. O'Malley (1983). "Actively transcribed genes are associated with the nuclear matrix." Nature **306**(5943): 607-609.

Cook, P. R. and I. A. Brazell (1978). "Spectrofluorometric measurement of the binding of ethidium to superhelical DNA from cell nuclei." Eur J Biochem **84**(2): 465-477.

Courey, A. J., S. E. Plon and J. C. Wang (1986). "The use of psoralen-modified DNA to probe the mechanism of enhancer action." Cell **45**(4): 567-574.

Cremer, T., C. Cremer, T. Schneider, H. Baumann, L. Hens and M. Kirsch-Volders (1982). "Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments." Hum Genet **62**(3): 201-209.

de Wit, E. and W. de Laat (2012). "A decade of 3C technologies: insights into nuclear organization." Genes Dev **26**(1): 11-24.

Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser and C. Weissmann (1983). "Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells." Cell **32**(3): 695-706.

Dierks, P., A. van Ooyen, N. Mantei and C. Weissmann (1981). "DNA sequences preceding the rabbit beta-globin gene are required for formation in mouse L cells of beta-globin RNA with the correct 5' terminus." Proc Natl Acad Sci U S A **78**(3): 1411-1415.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature **485**(7398): 376-380.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green and J. Dekker (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." Genome Res **16**(10): 1299-1309.

Dynan, W. S. (1986). "Promoters for housekeeping genes." Trends in Genetics **2**: 196-197.

Farnham, P. J. and R. T. Schimke (1985). "Transcriptional regulation of mouse dihydrofolate-reductase in the cell-cycle." Journal of Biological Chemistry **260**(12): 7675-7680.

Faye, G., D. W. Leung, K. Tatchell, B. D. Hall and M. Smith (1981). "Deletion mapping of sequences essential for in vivo transcription of the iso-1-cytochrome c gene." Proc Natl Acad Sci U S A **78**(4): 2258-2262.

Filippova, D., R. Patro, G. Duggal and C. Kingsford (2014). "Identification of alternative topological domains in chromatin." Algorithms Mol Biol **9**: 14.

Fisher-Aylor, K. I. (2011). Long distance looping maps: RNA Pol2 during differentiation. Nuclear Structure and Dynamics. L'Isle sur la Sorgue, France, EMBO.

Francastel, C., M. C. Walters, M. Groudine and D. I. Martin (1999). "A functional enhancer suppresses silencing of a transgene and prevents its localization close to centrometric heterochromatin." Cell **99**(3): 259-269.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung and Y. Ruan (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." Nature **462**(7269): 58-64.

Galande, S., P. K. Purbey, D. Notani and P. P. Kumar (2007). "The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1." Curr Opin Genet Dev **17**(5): 408-414.

Gasser, S. M. and U. K. Laemmli (1986). "The organisation of chromatin loops: characterization of a scaffold attachment site." EMBO J **5**(3): 511-518.

Gasser, S. M. and U. K. Laemmli (1987). "Improved methods for the isolation of individual and clustered mitotic chromosomes." Exp Cell Res **173**(1): 85-98.

Gerasimova, T. I. and V. G. Corces (1998). "Polycomb and trithorax group proteins mediate the function of a chromatin insulator." Cell **92**(4): 511-521.

Gerasimova, T. I., D. A. Gdula, D. V. Gerasimov, O. Simonova and V. G. Corces (1995). "A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation." Cell **82**(4): 587-597.

Gillies, S. D., S. L. Morrison, V. T. Oi and S. Tonegawa (1983). "A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene." Cell **33**(3): 717-728.

Gluzman, Y., J. F. Sambrook and R. J. Frisque (1980). "Expression of early genes of origin-defective mutants of simian virus 40." Proc Natl Acad Sci U S A **77**(7): 3898-3902.

Greally, J. M., D. J. Starr, S. Hwang, L. Song, M. Jaarola and S. Zemel (1998). "The mouse H19 locus mediates a transition between imprinted and non-imprinted DNA replication patterns." Hum Mol Genet **7**(1): 91-95.

Griffith, J., A. Hochschild and M. Ptashne (1986). "DNA loops induced by cooperative binding of lambda repressor." Nature **322**(6081): 750-752.

Grosschedl, R. and M. L. Birnstiel (1980). "Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo." Proc Natl Acad Sci U S A **77**(3): 1432-1436.

Grosschedl, R. and M. L. Birnstiel (1980). "Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo." Proc Natl Acad Sci U S A **77**(12): 7102-7106.

Gruss, P., R. Dhar and G. Khoury (1981). "Simian virus 40 tandem repeated sequences as an element of the early promoter." Proc Natl Acad Sci U S A **78**(2): 943-947.

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat and B. van Steensel (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Hakim, O., M. H. Sung, T. C. Voss, E. Splinter, S. John, P. J. Sabo, R. E. Thurman, J. A. Stamatoyannopoulos, W. de Laat and G. L. Hager (2011). "Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements." Genome Res **21**(5): 697-706.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan and C. L. Wei (2011). "CTCF-mediated functional chromatin interactome in pluripotent cells." Nat Genet **43**(7): 630-638.

Harr, J. C., T. R. Luperchio, X. Wong, E. Cohen, S. J. Wheelan and K. L. Reddy (2015). "Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins." J Cell Biol **208**(1): 33-52.

Hebbes, T. R., A. L. Clayton, A. W. Thorne and C. Crane-Robinson (1994). "Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain." EMBO J **13**(8): 1823-1830.

Hochschild, A., N. Irwin and M. Ptashne (1983). "Repressor structure and the mechanism of positive control." Cell **32**(2): 319-325.

Iborra, F. J., A. Pombo, D. A. Jackson and P. R. Cook (1996). "Active RNA polymerases are localized within discrete transcription "factories' in human nuclei." J Cell Sci **109**(Pt 6): 1427-1436.

Ip, Y. T., R. E. Park, D. Kosman, E. Bier and M. Levine (1992). "The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive …." Genes Dev **6**(9): 1728-1739.

Jackson, D. A., A. B. Hassan, R. J. Errington and P. R. Cook (1993). "Visualization of focal sites of transcription within human nuclei." EMBO J **12**(3): 1059-1065.

Johnson, D. S., A. Mortazavi, R. M. Myers and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Kanji, G. K. (1999). 100 Statistical Tests, SAGE Publications Ltd., London, England.

Kellum, R. and P. Schedl (1991). "A position-effect assay for boundaries of higher order chromosomal domains." Cell **64**(5): 941-950.

Kennell, D. and H. Riezman (1977). "Transcription and translation initiation frequencies of the Escherichia coli lac operon." Journal of Molecular Biology **114**(1): 1-21.

Kieffer-Kwon, K. R., Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grontved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan and R. Casellas (2013). "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation." Cell **155**(7): 1507-1520.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov and B. Ren (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Kosak, S. T. and M. Groudine (2004). "Gene order and dynamic domains." Science **306**(5696): 644-647.

Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher and H. Singh (2002). "Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development." Science **296**(5565): 158-162.

Kramer, H., M. Niemoller, M. Amouyal, B. Revet, B. von Wilcken-Bergmann and B. Muller-Hill (1987). "lac repressor forms loops with linear DNA carrying two suitably spaced lac operators." EMBO J **6**(5): 1481-1491.

Krebs, J. E. and M. Dunaway (1998). "The scs and scs' insulator elements impart a cis requirement on enhancer-promoter interactions." Mol Cell **1**(2): 301-308.

Lebkowski, J. S. and U. K. Laemmli (1982). "Evidence for two levels of DNA folding in histone-depleted HeLa interphase nuclei." J Mol Biol **156**(2): 309-324.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. Ruan (2012). "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." Cell **148**(1-2): 84-98.

Li, G. L., X. A. Ruan, R. K. Auerbach, K. S. Sandhu, M. Z. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Y. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Z. Hong, Z. Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. H. Ge, H. E. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. J. Ruan (2012). "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation." Cell **148**(1-2): 84-98.

Li, Y., W. Huang, L. Niu, D. M. Umbach, S. Covo and L. Li (2013). "Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes." BMC Genomics **14**: 553.

Liberman, L. M. and A. Stathopoulos (2009). "Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence." Dev Biol **327**(2): 578-589.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Ling, J. Q., T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry and A. R. Hoffman (2006). "CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1." Science **312**(5771): 269-272.

MacArthur, S., X. Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. Keranen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin and M. B. Eisen (2009). "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions." Genome Biol **10**(7): R80.

Magistri, M., M. A. Faghihi, G. St Laurent, 3rd and C. Wahlestedt (2012). "Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts." Trends Genet **28**(8): 389-396.

Mallin, D. R., J. S. Myung, J. S. Patton and P. K. Geyer (1998). "Polycomb group repression is blocked by the Drosophila suppressor of Hairy-wing [su(Hw)] insulator." Genetics **148**(1): 331-339.

Markstein, M., P. Markstein, V. Markstein and M. S. Levine (2002). "Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo." Proceedings of the National Academy of Sciences **99**(2): 763.

Markstein, M., R. Zinzen, P. Markstein, K. P. Yee, A. Erives, A. Stathopoulos and M. Levine (2004). "A regulatory code for neurogenic gene expression in the Drosophila embryo." Development **131**(10): 2387-2394.

McKnight, S. L., R. C. Kingsbury, A. Spence and M. Smith (1984). "The distal transcription signals of the herpesvirus tk gene share a common hexanucleotide control sequence." Cell **37**(1): 253-262.

Mercola, M., X. F. Wang, J. Olsen and K. Calame (1983). "Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus." Science **221**(4611): 663-665.

Meshorer, E. and T. Misteli (2006). "Chromatin in pluripotent embryonic stem cells and differentiation." Nat Rev Mol Cell Biol **7**(7): 540-546.

Mirkovitch, J., M. E. Mirault and U. K. Laemmli (1984). "Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold." Cell **39**(1): 223-232.

Mitchell, J. A. and P. Fraser (2008). "Transcription factories are nuclear subcompartments that remain in the absence of transcription." Genes Dev **22**(1): 20-25.

Morcillo, P., C. Rosen, M. K. Baylies and D. Dorsett (1997). "Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in Drosophila." Genes Dev **11**(20): 2729-2740.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Muller, H. P. and W. Schaffner (1990). "Transcriptional enhancers can act in trans." Trends in Genetics **6**(9): 300-304.

Muller, H. P., J. M. Sogo and W. Schaffner (1989). "An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge." Cell **58**(4): 767-777.

Muller, M. M., T. Gerster and W. Schaffner (1988). "Enhancer sequences and the regulation of gene-transcription." European Journal of Biochemistry **176**(3): 485-495.

Nabirochkin, S., M. Ossokina and T. Heidmann (1998). "A nuclear matrix/scaffold attachment region co-localizes with the gypsy retrotransposon insulator sequence." J Biol Chem **273**(4): 2473-2479.

Nakagomi, K., Y. Kohwi, L. A. Dickinson and T. Kohwi-Shigematsu (1994). "A novel DNA-binding motif in the nuclear matrix attachment DNA-binding protein SATB1." Mol Cell Biol **14**(3): 1852-1860.

Namciu, S. J., K. B. Blochlinger and R. E. Fournier (1998). "Human matrix attachment regions insulate transgene expression from chromosomal position effects in Drosophila melanogaster." Mol Cell Biol **18**(4): 2382-2391.

Neuberger, M. S. (1983). "Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells." EMBO J **2**(8): 1373-1378.

Noordermeer, D., M. R. Branco, E. Splinter, P. Klous, W. van Ijcken, S. Swagemakers, M. Koutsourakis, P. van der Spek, A. Pombo and W. de Laat (2008).

"Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region." PLoS Genet **4**(3): e1000016.

Noordermeer, D., E. de Wit, P. Klous, H. van de Werken, M. Simonis, M. Lopez-Jones, B. Eussen, A. de Klein, R. H. Singer and W. de Laat (2011). "Variegated gene expression caused by cell-specific long-range DNA interactions." Nat Cell Biol **13**(8): 944-951.

Ogawa, N., and Biggin, M. D. (2012). "High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro." Methods Mol Biol **786**: 51-63.

Ong, C. T. and V. G. Corces (2014). "CTCF: an architectural protein bridging genome topology and function." Nat Rev Genet **15**(4): 234-246.

Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik and P. Fraser (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Palstra, R. J., M. Simonis, P. Klous, E. Brasset, B. Eijkelkamp and W. de Laat (2008). "Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription." PLoS One **3**(2): e1661.

Pepke, S., B. Wold and A. Mortazavi (2009). "Computation for ChIP-seq and RNA-seq studies." Nat Methods **6**(11 Suppl): S22-32.

Peric-Hupkes, D., W. Meuleman, L. Pagie, S. W. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels and B. van Steensel (2010). "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation." Mol Cell **38**(4): 603-613.

Phillips-Cremins, J. E., M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor and V. G. Corces (2013). "Architectural protein subclasses shape 3D organization of genomes during lineage commitment." Cell **153**(6): 1281-1295.

Pickersgill, H., B. Kalverda, E. de Wit, W. Talhout, M. Fornerod and B. van Steensel (2006). "Characterization of the Drosophila melanogaster genome at the nuclear lamina." Nat Genet **38**(9): 1005-1014.

Plon, S. E. and J. C. Wang (1986). "Transcription of the human beta-globin gene is stimulated by an SV40 enhancer to which it is physically linked but topologically uncoupled." Cell **45**(4): 575-580.

Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. S. Wu, O. Denas, D. L. Vera, Y. L. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren and D. M. Gilbert (2014). "Topologically associating domains are stable units of replication-timing regulation." Nature **515**(7527): 402-+.

Robin, J. D., A. T. Ludlow, K. Batten, M. C. Gaillard, G. Stadler, F. Magdinier, W. E. Wright and J. W. Shay (2015). "SORBS2 transcription is activated by telomere position effect-over long distance upon telomere shortening in muscle cells from patients with facioscapulohumeral dystrophy." Genome Res **25**(12): 1781-1790.

Robinson, S. I., D. Small, R. Idzerda, G. S. McKnight and B. Vogelstein (1983). "The association of transcriptionally active genes with the nuclear matrix of the chicken oviduct." Nucleic Acids Res **11**(15): 5113-5130.

Roulet, E., S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod and P. Bucher (2002). "High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites." Nat Biotechnol **20**(8): 831-835.

Schaffner, W., G. Kunz, H. Daetwyler, J. Telford, H. O. Smith and M. L. Birnstiel (1978). "Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing." Cell **14**(3): 655-671.

Serfling, E., M. Jasin and W. Schaffner (1985). "ENHANCERS AND EUKARYOTIC GENE-TRANSCRIPTION." Trends in Genetics **1**(8): 224-230.

Serfling, E., A. Lubbe, K. Dorsch-Hasler and W. Schaffner (1985). "Metal-dependent SV40 viruses containing inducible enhancers from the upstream region of metallothionein genes." EMBO J **4**(13B): 3851-3859.

Shopland, L. S., C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth, C. Cremer, C. J. Bult and T. P. O'Brien (2006). "Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence." J Cell Biol **174**(1): 27-38.

Siebenlist, U., R. B. Simpson and W. Gilbert (1980). "E. coli RNA polymerase interacts homologously with two different promoters." Cell **20**(2): 269-281.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." Nat Genet **38**(11): 1348-1354.

Small, D., B. Nelkin and B. Vogelstein (1985). "The association of transcribed genes with the nuclear matrix of Drosophila cells during heat shock." Nucleic Acids Res **13**(7): 2413-2431.

Splinter, E., H. Heath, J. Kooren, R. J. Palstra, P. Klous, F. Grosveld, N. Galjart and W. de Laat (2006). "CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus." Genes Dev **20**(17): 2349-2354.

Stathopoulos, A. and M. Levine (2002). "Whole-Genome Expression Profiles Identify Gene Batteries in Drosophila." Developmental Cell **3**(4): 464-465.

Stathopoulos, A. and M. Levine (2005). "Genomic Regulatory Networks and Animal Development." Developmental Cell **9**(4): 449.

Stathopoulos, A., M. Van Drenth, A. Erives, M. Markstein and M. Levine (2002). "Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo." Cell **111**(5): 687-701.

ten Bosch, J. R., J. A. Benavides and T. W. Cline (2006). "The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription." Development **133**(10): 1967.

Theveny, B., A. Bailly, C. Rauch, M. Rauch, E. Delain and E. Milgrom (1987). "Association of DNA-bound progesterone receptors." Nature **329**(6134): 79-81.

Tiwari, V. K., L. Cope, K. M. McGarvey, J. E. Ohm and S. B. Baylin (2008). "A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations." Genome Res **18**(7): 1171-1179.

Udvardy, A., E. Maine and P. Schedl (1985). "The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains." J Mol Biol **185**(2): 341-358.

van Steensel, B. and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase." Nat Biotechnol **18**(4): 424-428.

Verschure, P. J., I. van Der Kraan, E. M. Manders and R. van Driel (1999). "Spatial relationship between transcription sites and chromosome territories." J Cell Biol **147**(1): 13-24.

Weber, F., J. de Villiers and W. Schaffner (1984). "An SV40 'enhancer trap' incorporates exogenous enhancers or generates enhancers from its own sequences." Cell **36**(4): 983-992.

Wiesendanger, B., R. Lucchini, T. Koller and J. M. Sogo (1994). "Replication fork barriers in the Xenopus rDNA." Nucleic Acids Res **22**(23): 5038-5046.

Wigler, M., R. Sweet, G. K. Sim, B. Wold, A. Pellicer, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Transformation of mammalian cells with genes from procaryotes and eucaryotes." Cell **16**(4): 777-785.

Wold, B., M. Wigler, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Introduction and expression of a rabbit beta-globin gene in mouse fibroblasts." Proc Natl Acad Sci U S A **76**(11): 5684-5688.

Wurtele, H. and P. Chartrand (2006). "Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology." Chromosome Res **14**(5): 477-495.

Yaffe, D. and O. Saxel (1977). "Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle." Nature **270**(5639): 725-727.

Yee, S. P. and P. W. Rigby (1993). "The regulation of myogenin gene expression during the embryonic development of the mouse." Genes Dev **7**(7A): 1277-1289.

Zeng, M. e. a. (2015). in review.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

Zhao, Z., G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti and R. Ohlsson (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." Nat Genet **38**(11): 1341-1347.

Zinzen, R., K. Senger, M. Levine and D. Papatsenko (2006). "Computational Models for Neurogenic Gene Expression in the Drosophila Embryo." Current Biology **16**(13): 1358-1365.

Chapter III: Transcriptional topology

III.1: Introduction: What we knew about transcriptional topology at the beginning of this project

Transcriptional topology, the portion of chromatin topology involved in transcriptional regulation, has been conceptually differentiated from chromatin topology since the “looping model” of enhancement came to prominence (Muller, Sogo et al. 1989) but genes, their location, and their activity have been studies with respect to chromatin outside of enhancement alone. MARs (DNA regions associating with the nuclear matrix) marking boundaries of active chromatin domains was a popular field in the 80s (see (Mirkovitch, Mirault et al. 1984)), and it was known that active genes associate with MARs in a variety of organisms (Robinson, Small et al. 1983; Ciejek, Tsai et al. 1983; Small, Nelkin et al. 1985; Gasser and Laemmli 1986), and even that MARs in some case overlapped enhancers. In the 2000s, people began to study chromatin conformation and its effects on genes more closely, noting that some genes are able to “loop out” of place upon activation (Chambeyron and Bickmore 2004), and that gene-poor and gene-rich regions separate (Shopland, Lynch et al. 2006), or that only certain classes of genes do this (Simonis, Klous et al. 2006). Along with this line of thinking came the notion of the “transcription factory,” previously noted through microscopy as rare foci of pol2 (Jackson, Hassan et al. 1993; Iborra, Pombo et al. 1996) actively transcribing genes (Verschure, van Der Kraan et al. 1999), as a method for groups of related genes to be expressed, and perhaps a primary mode of transcription. Transcription factories are according to some definitions architecturally unchanging elements since genes can move to and from transcription factories (Osborne, Chakalova

et al. 2004) and their existence does not depend on transcription itself (Mitchell and Fraser 2008; Palstra, Simonis et al. 2008).

The contemporary genome-wide chromatin literature talks much about a potentially related concept called a “topologically active domain” (TAD), a region of chromatin containing active genes capable of interacting with each other and bounded by CTCF and cohesin (Dixon, Selvaraj et al. 2012; Li, Huang et al. 2013). It is unknown how much the modern TAD has in common with earlier understanding of separate gene-rich and gene-poor areas; whether it explains all or only some. In addition to containing active genes, TADs seem to have a role in the timing of cell replication (Pope, Ryba et al. 2014) and are also associated with Lamin-associated domains (Peric-Hupkes, Meuleman et al. 2010). The TAD is by no means the smallest unit of chromatin within which genes preferentially interact, which may be understandable since the TAD was defined by the 1MB-resolution Hi-C method, while other domains, sometimes confusingly called “smaller TADs,” are found with more highly sensitive measures like 5C and with more computational processing (Phillips-Cremins, Sauria et al. 2013; Filippova, Patro et al. 2014).

All of the above now appears quite relevant to this study since most genes that connect to far-distal elements connect to elements within their neighborhood of about 150kb. A gene’s neighborhood must be important with respect to what most genes connect since all of those connections are within that neighborhood, and the above bulk-scale or microscopy assays showing active elements connecting to other active elements are likely related to a subset of the larger CIGs I report. Others have reported that there are some active-gene-poor and some active-gene-rich areas (“ridges”) and it has even been claimed that certain meta-classes of genes such as transcription factors are more likely to be in gene-poor areas versus in gene-rich areas like lineage-specific

genes are (Lercher, Urrutia et al. 2002). However, a similar paper found that developmentally related genes could actually be in either type of area (Versteeg, van Schaik et al. 2003).

Then there are the cases of known, validated E:P “looping” interactions. Considering this background, pol2 ChIA-PET data are expected to identify physical interactions of several different functional classes. I am most likely seeing many classic E:P “looping” interactions, but I am also surely seeing interactions that are primarily involved in the nuclear architecture. Such interactions can be mediated by known DNA-site specific chromatin factors such as CTCF (Splinter, Heath et al. 2006; Kim, Abdullaev et al. 2007; Handoko, Xu et al. 2011; Ong and Corces 2014), ZNF143 (Bailey, Zhang et al. 2015), or YY1 (Harr, Luperchio et al. 2015; Zeng 2015), as well as less sequence-specific factors (Galande, Purbey et al. 2007) and likely less-known factors and RNA components (Magistri, Faghihi et al. 2012) as well. In chapter III, I focused on the general aspects of topology that are consistent across ChIA-PET experiments. In this chapter, I will focus on the topological interactions that are different between genes of different expression classes.

III.1.1: State-to-state changes

Globally, chromatin interactions are not thought to change much (Simonis, Klous et al. 2006; Hakim, Sung et al. 2011). Nevertheless, some E:P interactions have been shown to be transient and dependent on transcription (Cheutin, O'Donohue et al. 2003; Kosak and Groudine 2004; Meshorer and Misteli 2006). One experimentally-driven hypothesis is that rearrangement of CRMs can only occur within certain native active chromatin domains (Noordermeer, Branco et al. 2008). I will report in this chapter on the changes in detectible interactions between two developmental states, but a caveat is that these interactions may well be invisible to us before genes are active, since the

ChIA-PET experiments I undertook only detect interactions that co-occur with ChIPpable factors associated with transcription. Therefore, ChIA-PET cannot tell us where connectivity changes, only where the active use of connected elements may change from state to state.

III.1.2: Housekeeping genes

The notion of the housekeeping gene has been prevalent since the discovery of genes themselves. Since certain enzymes, structural elements, and other core parts of the universal cellular machinery must be expressed at roughly similar levels in every cell, the reasoning goes, these genes don't need to be regulated. Housekeeping genes are often used as a foil or control for developmental genes, which are regulated and differentially expressed in different cell types.

The promoters of some housekeeping genes were investigated during the course of researching promoter and transcription biochemistry, but because of the technologic limits of the day, highly-expressed genes had to be studied and genes involved in disease and development were investigated first. It was probably the fact that TATA is prevalent in the promoters of developmental genes that led to it being the first promoter motif discovered in mammals (Goldberg 1989), a bias that was noticed by the researchers of the time (Breathnach and Chambon 1981). In contrast, the TATA-less genes were generally regarded as housekeeping genes with low expression levels and multiple 5' ends (Dynan 1986).

Perhaps it was because the promoters of housekeeping genes seemed more complicated to study than that of the developmental genes, or perhaps because many were expressed at a modest level, or perhaps because developmentally regulated genes have traditionally been more scientifically exciting, but for a type of gene that is often used as a conceptual foil or experimental foil, housekeeping genes have fallen by the

wayside in terms of direct research. One reason that surely has played a part in this lack of research is that, somewhere along the line, “being regulated” became synonymous colloquially with “has enhancer(s).” Housekeeping genes, thought to have steady expression levels, are assumed to maintain these steady expression levels by virtue of a constitutively active promoter. Meanwhile, developmental genes, which have varying levels of expression, need CRMs in order to modulate their levels of transcription over development and in different tissues.

A few people have managed to study the regulation of housekeeping genes, though. One person who studied a vital housekeeping gene, DFHR, was Dr. Peggy Farnham, despite the difficulties of obtaining funding for something assumed not to happen (B. Wold, pers. comm.). Dr. Farnham found that it did, indeed, have enhancers (Farnham and Schimke 1985). However, Dr. Farnham attributed this need for enhancers to the fact that DFHR was known to be differentially regulated in the cell cycle and did not attempt to question whether housekeeping genes broadly had enhancers. Likewise, in the case of *string* in *Drosophila*, in which the gene appeared at the level of tissues to be broadly expressed but was in fact differentially regulated at the level of cells, enhancers were found, but again were written off as a peculiarity of the particular gene studied. Enhancers continued to be studied almost exclusively in the context of developmental genes over the next decade, and the preponderance of developmental gene enhancer literature and absence of housekeeping gene expression literature sometimes seemingly led many to forget the formal possibility that housekeeping genes in general might have enhancers.

Is it really possible for any gene, much less most genes, to truly be unregulated in any way in every cell type? Some cell types like immune cells with their rearranged genomes or neurons and spermatocytes with their uniquely stripped-down metabolic

requirements surely contain a large number of “housekeeping” genes with varied levels of expression relative to the other tissues in the body. Furthermore, even genes that are known to be regulated by enhancers can appear to drive native expression patterns with their proximal promoters alone, as was once the case with myogenin (Yee and Rigby 1993), so the lack of apparent necessity, in assay, for CRMs does not disprove their existence.

III.2: Results

III.2.1: Connectivity and amount of gene expression

Does connectivity, as measured by ChIA-PET, predict level of gene expression? Since the presence of pol2 is required to detect a ChIA-PET connection and is correlated with active genes (Mortazavi, Williams et al. 2008), a positive correlation is expected and is observed. Further, this correlation is quantitative, with the most highly expressed being the most likely to have detectible connections (Fig. III-1). Surprisingly, this is not due to the source ChIP being pol2; it is also true of myogenin ChIA-PET (Fig. III-1, bottom). From a related perspective, larger CIGs are more likely than smaller CIGs to contain highly expressed genes (Fig. III-8).

III.2.2: Gene-distal interactions

Are the majority of gene-to-distal interactions multiple, as in the case of the β-globin LCR, or are genes regulated by many enhancers, either timing-specific or tissue-specific, each? I have shown that the majority of captured ChIA-PET interactions are one-to-one (Fig. II-9), and this likely suggests that most genes, which are modestly expressed, connect to one distal element, but I also discovered large sets of more complex interactions. First, there are hundreds of instances of individual distal elements connecting to multiple genes (Fig. III-3A), which was perhaps less expected than the pattern of multiple distal elements connecting to one gene (Fig. III-3B). Distal elements,

whether they connect with only one gene or multiple genes, all connect to the nearest annotated active gene, even the 25% of elements that connect to multiple genes (Fig. III-2, right). The majority of genes without a ChIA-PET connection are not detectably expressed (Fig. III-1, compare Figures I-3 to III-4). I used available C2C12 active and repressive chromatin mark data, and skipped over genes that lack any of the active or repressive chromatin marks assayed, opposed to connected genes, which have active marks as expected (data not shown). However, due to the relatively little data on repressive marks in our laboratory collection or the literature, I cannot comment on the biochemistry underlying gene-skipping, other than to say that it is consistent with a previous microscope experiment's claim that inactive genes "loop out" of transcription factories (Mitchell and Fraser 2008).

III.2.3: Connectivity and changes in gene expression

I next wanted to know how ChIA-PET connections changed over time. It has been reported that E:P interactions do change with transcription, but also that many are constant across tissue types (Simonis, Klous et al. 2006). In order to determine how change in gene expression relates to ChIA-PET, I chose to create four well-defined trajectories of gene expression – up, flat, down, and off – to analyze with respect to each other, while leaving behind genes that are ambiguously expressed or that have an ambiguous trajectory (Fig. III-4). The change in ChIA-PET connectivity from myoblast to myocyte correlates highly with the change in gene expression (Fig. III-5). This could mean that flat genes are unlikely to change their architecture, while developmentally regulated genes might. Alternatively, heeding the cautions from Fig. II-5, the architecture could remain constant while increased input from CRMs could cause the increase in gene expression. Another, less likely, possibility is that high expression at one promoter

somehow bleeds over into surrounding genomic area, increasing the number of ChIA-PET connections returned.

III.2.4: Distal degree shows a preference for gene type

To further quantify ChIA-PET connectivity within the up, flat, down, and off groups, I measured the degree. The degree of a gene is not as closely correlated with its expression (data not shown) as edge weight is (Fig. III-5). When measuring specifically the distal degree (distal connection number per gene) in linear regression versus gene expression, it became clear that while distal degree is weakly correlated with expression level (data not shown), it is strongly correlated with gene trajectory, specifically, the upregulated genes (Fig. III-7, top). Flat genes have no such correlation (data not shown).

I have shown that the amount of ChIA-PET connectivity is partially related to the quantity of a gene's expression. However, I wanted to determine if the striking results in the upregulated set of genes were due to expression change alone, or if they showed evidence of being connected in a qualitatively different way from other genes. To further explore the distinction between upregulated and flat genes, I asked how for each trajectory group distal connectivity is distributed with respect to RNA amount. For each RNA abundance class, I quantified the global myocyte distal degree. There is a strong distinction ($P<10\text{-}5$) for medium- and high-abundance upregulated genes to have a higher distal degree than flat genes of the same abundance (Fig. III-7, bottom).

III.2.5: Promoter-promoter connections

Almost a third of pol2 ChIA-PET interactions are gene-gene interactions, which in our data define gene vertices centered at transcription start sites (see Materials and Methods). In two contemporary studies, the authors suggested that such ChIA-PET connections reflect, and may even cause, co-regulation (Chepelev, Wei et al. 2012; Li, Ruan et al. 2012; Kieffer-Kwon, Tang et al. 2013). My differentiation system is well-suited

to test this with XX G-G connections overall, YY containing at least one significantly up- or down-regulated gene, and ZZ containing genes of substantial trajectory >50FPKM. I interrogated my data in several different ways to ask if, globally, G-G connectivity predicts co-regulation of the paired genes. I found no statistically significant correlation overall between pairs of genes with respect to their expression levels (data not shown) when I confined the analysis to active genes, since unexpressed genes aren't expected to connect (and therefore will give a false positive significant result when included in the null hypothesis). Similarly, there was no significant correlation between expression trajectory of pairs of connected gene vertices (data not shown). However, the G-G landscape isn't completely random. Flat genes, the most expressed gene type, are connected equally with each other, and with upregulated or downregulated, while the two differential classes are almost never connected (Fig. III-6, left versus center bars).

Although there was no global evidence of co-regulation associated globally with gene-gene pol2 connectivity, inspection of several loci of known biological interest led me to ask whether a more narrowly defined set of development genes, isolating the most extreme expression differences, are overrepresented in certain expression patterns. I compared the activity of genes which directly or indirectly connect to super-differential "seed" genes to all genes within the 2Mb window available to the ChIA-PET connections of the "seed" genes (Fig. III-6, top). I found that members of the group of 252 extremely upregulated muscle genes are more likely to be close to other myogenic genes than other groups of genes, but even taking this into consideration, they were also more likely to connect to other upregulated genes (Fig. III-6, bottom). A similar analysis with downregulated genes just barely failed statistical significance, perhaps because of low n, and there was no such result with flat genes (data not shown). This suggests that only the

small subsets of developmentally regulated genes are candidates for co-regulation, while the majority of G-G interactions reflect co-expression.

For large graph analysis, one interpretation is that the additive effect of enhancers causes high expression. Another interpretation is that expression at some modestly and steadily expressed genes may be a byproduct of being physically connected to an important gene.

III.3: Discussion/Conclusions

First, I reported that the number of multiple gene and distal interactions was unexpectedly high based on the founding ChIA-PET paper (Fullwood, Liu et al. 2009), and then I explored the likewise unexpected number of gene-gene interactions. Separating connected genes according to their behavior, I showed that flat, presumptively "housekeeping" genes connect to all classes of genes with equal frequency. This applies to direct and indirect connections. In contrast, upregulated and downregulated genes never connect directly to each other, and rarely are connected indirectly. Upregulated and downregulated genes share several MRFs (in C2's, myoD; elsewhere, myf5 and mrf4 too), and they show slight genomic separation (Fig. III-6). Perhaps their CRMs must be spatially separated from each other to avoid cross-activation, or maybe the result is an "accident" of evolutionary gene groups being linked on the chromosome.

Next, I explored connectivity as a function of expression amount and as a function of behavior class. The amount of expression appears to predict whether or not a gene will have detectible long-range interactions, and trajectory predicts how many long-range interactions there will be. Developmentally specific genes can be predicted by ChIA-PET edge changes, and this, plus the small number of connected, unexpressed genes, predicts that "poised" pol2-containing structures are minimal and perhaps rare.

A final main conclusion not anticipated by prior ChIA-PET studies is that upregulated myogenic differentiation genes are significantly more likely to be connected to multiple distal elements than are genes in the large “housekeeping” group. This is strongly observed for both highly and moderately active genes, and suggests a subset of myogenic genes are an ancient conserved gene class that is regulated differently than the simple, ubiquitously expressed housekeeping genes or developmental genes that are in gene deserts.

What about transcription factories?

The notion of the transcription factory has been raised in conjunction with ChIA-PET (Li, Ruan et al. 2012). However, the definition of “transcription factory” is not used consistently at all in the literature. The original “transcription factory” was an immobile area of high RNA polymerase II density visible under the light microscope after staining for the protein (Jackson, Hassan et al. 1993; Iborra, Pombo et al. 1996). It was hypothesized that these loci were stationary areas containing transcriptional machinery where multiple genes could physically interact in order to be transcribed (Jackson, Iborra et al. 1998; Cook 1999; Francastel, Walters et al. 1999).

However, others, having studied this phenomenon in conjunction with a few well-known developmental gene loci, particularly the alpha-globin/beta-globin locus, tied a “co-regulation” requirement into the definition, rather than the more agnostic “co-expression” requirement. It was, in fact, a preliminary version of my analysis in Fig. II-6 (Fisher-Aylor 2011) that drove the observations in one of the most recent papers in which I am credited as a minor author (primarily for my contributions to the ChIA-PET protocol of using EGS to stabilize protein-protein interactions, but also for our discovery of the prevalence of gene-to-gene, possibly promoter-to-promoter, interactions, and for this analysis) – an author who did not have any editorial input to the paper to my dismay

– that ChIA-PET interactions represent “co-regulation” (Li, Ruan et al. 2012). That work claimed that an unexpected number of promoters connected to each other and (using “a novel statistical analysis” they coined specifically to support this result in their datasets) that differential genes were likely to be connected to each other, and the senior authors of the paper opted for the connectivity of like genes to be a principal conclusion. My own complete analysis arrived at a much less strong emphasis on the connectivity of genes according to their expression types, though. There are indeed a large number of gene vertices connecting to gene vertices (perhaps promoters connecting to promoters) and enrichment over expectation for like genes to connect to one another, at least in the myocyte state (Fig. III-6); the smaller-n myoblast analysis yielded results that hovered around statistical significance but did not pass my preferred stringent P-value cutoff of 0.005 (data not shown). However, genes of like expression, whether defined by magnitude of or change in RNA output, connecting to one another are not at all what the majority of the ChIA-PET data show. Most genes are expressed similarly in both timepoints (perhaps housekeeping genes); most genes, regardless of type, connect to these flat expressed genes (Fig. III-6; other analyses not shown). This makes sense, given that I have determined how important gene neighborhood is to connectivity, and given what we know overall about gene neighborhoods.

While transcription factories may be a primary mode of transcription for some genes, it has been noted that there are only a limited number of them per cell (Osborne, Chakalova et al. 2004). If all genes use transcription factories, chromatin must rearrange at a higher order of chromatin coiling than current methods can detect since the first of the high-throughput assays suggest that globally, chromatin interactions do not change much (Simonis, Klous et al. 2006; Hakim, Sung et al. 2011).

Figures for Chapter III

Figure III-1

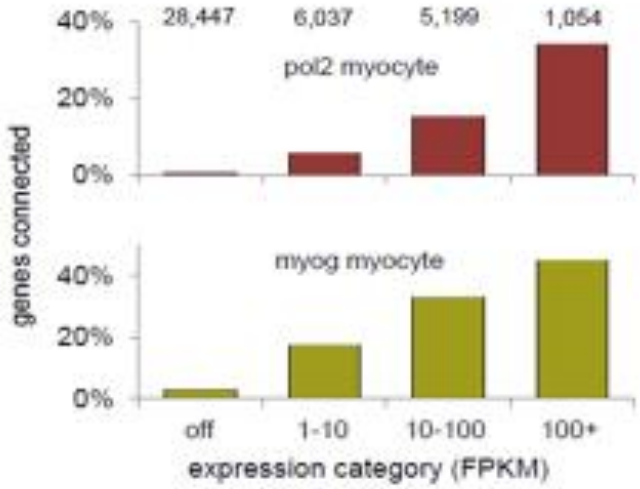


Figure III-1: Most unexpressed genes have no ChIA-PET connectivity, and the more highly expressed a gene is, the more likely it is to be connected in ChIA-PET. *The percent of gene-vertices at each expression level that are connected in pol2 (red) or myogenin (green) ChIA-PET. For a proper comparison to the single myogenin dataset, the experiment-matched single myocyte pol2 replicate dataset was used.*

Figure III-2

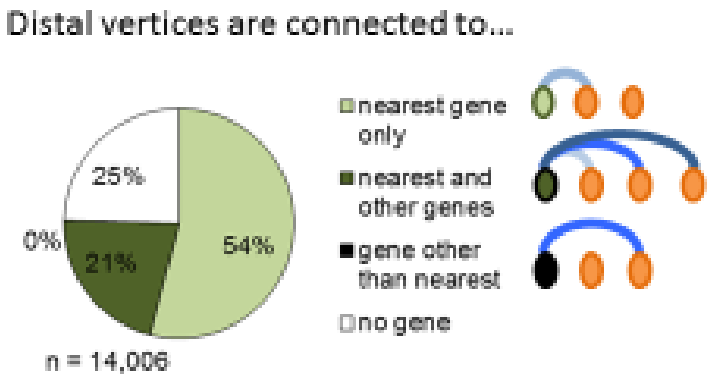


Figure III-2: Distal elements connect to the nearest active genes. *Distal vertices (leftmost, non-orange ovals which correspond to the colors in the pie chart) are shown connecting to genes (orange ovals) in different configurations. All distal vertices that connect to an active gene within the 10kb-2Mb range visible to ChIA-PET are connected to the nearest active gene (light green, dark green). There is no evidence of distal vertices skipping over an active gene without connecting to it (black).*

Figure III-3

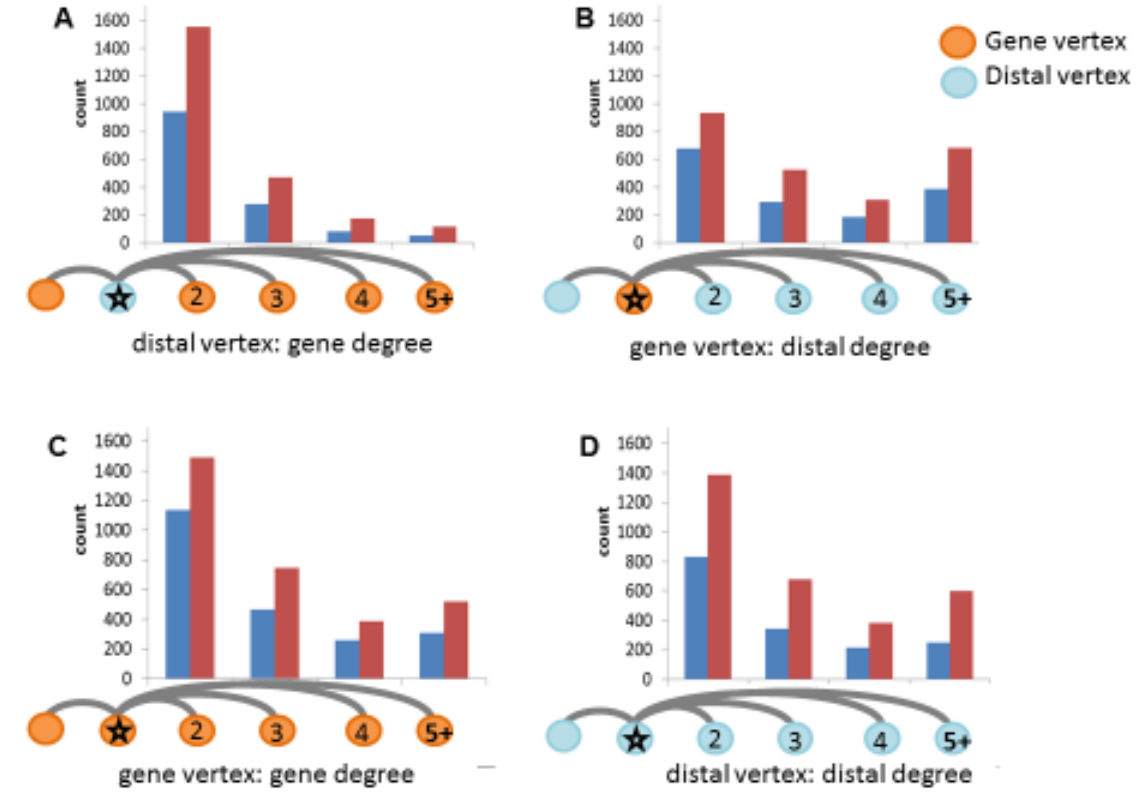


Figure III-3: One-to-many interactions in ChIA-PET data. *This set of figures represents the multiplicity of distal to gene interactions. Cartoon: blue circles represent distal vertices; orange circles represent gene vertices; the single starred vertex in each cartoon represents the "founder" type of vertex for which the connections are being tallied in the accompanying graph. Blue bars: myoblast pol2 edges; red bars: myocyte pol2 edges. (A) The number of distal vertices that connect to multiple gene vertices. (B) The number of gene vertices that connect to multiple distal vertices. (C) The number of gene vertices that connect to multiple other gene vertices. (D) The number of distal vertices that connect to multiple other distal vertices.*

Figure III-4 A

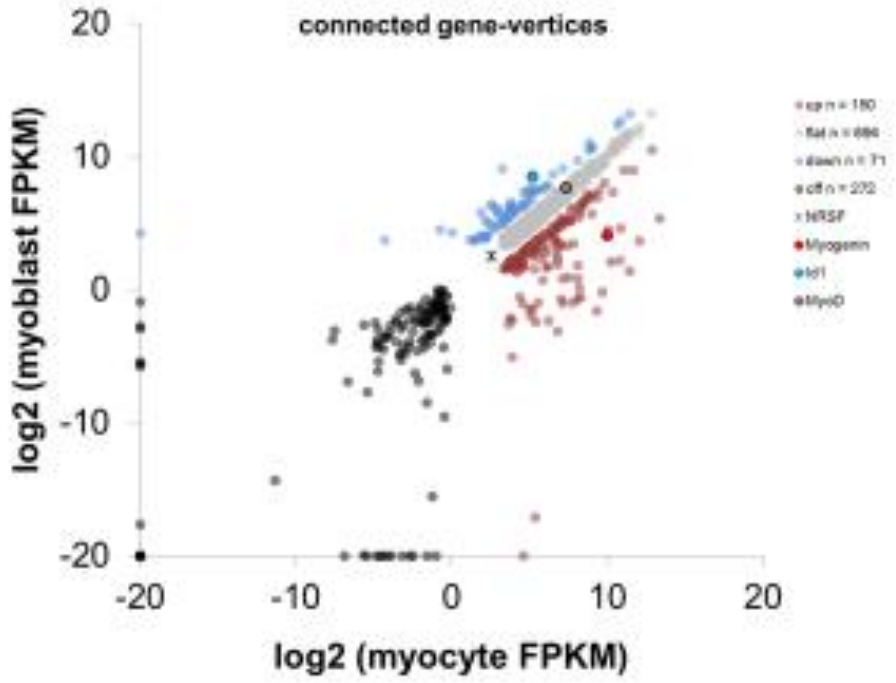


Figure III-4: Gene functional classes as defined in this work. *(A) The myoblast vs. myocyte RNA levels of connected up (red), down (blue), flat (gray), and off (black) gene classes. There are 990 gene-vertices that fall between these stringent categories and that are left out of the expression category analysis. All FPKM values of 0 were replaced with a very small fraction so they could be plotted on the graph's axes.*

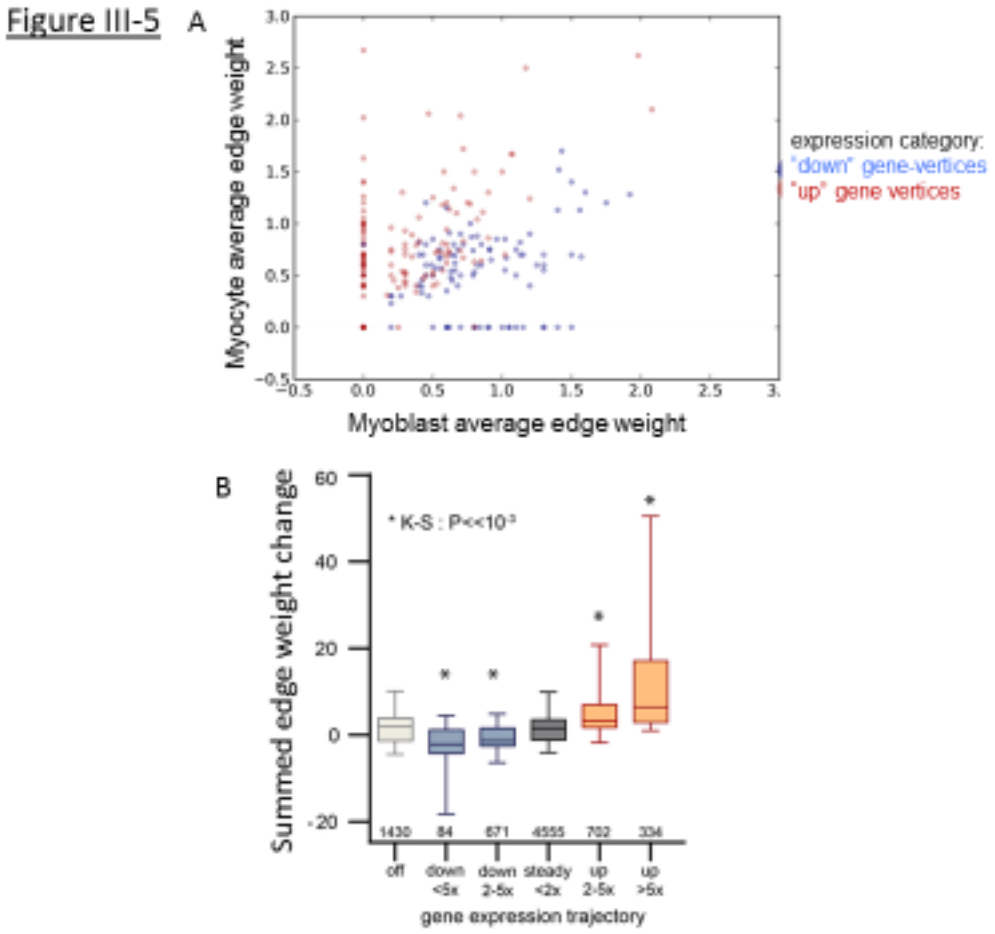


Figure III-5: Change in connectivity predicts developmental classes of genes. *(A) Gene-vertices with myoblast (blue) and myocyte (red) preferential expression have highly variable average edge weight between the two developmental states. (B) Fold change from myoblast to myocyte summed edge weight for two downregulated (blue) and two upregulated (orange) sets of gene-vertices are significantly different from the flat (gray) and off (white) gene-vertices, which do not change much in terms of ChIA-PET connectivity during the developmental transition.*

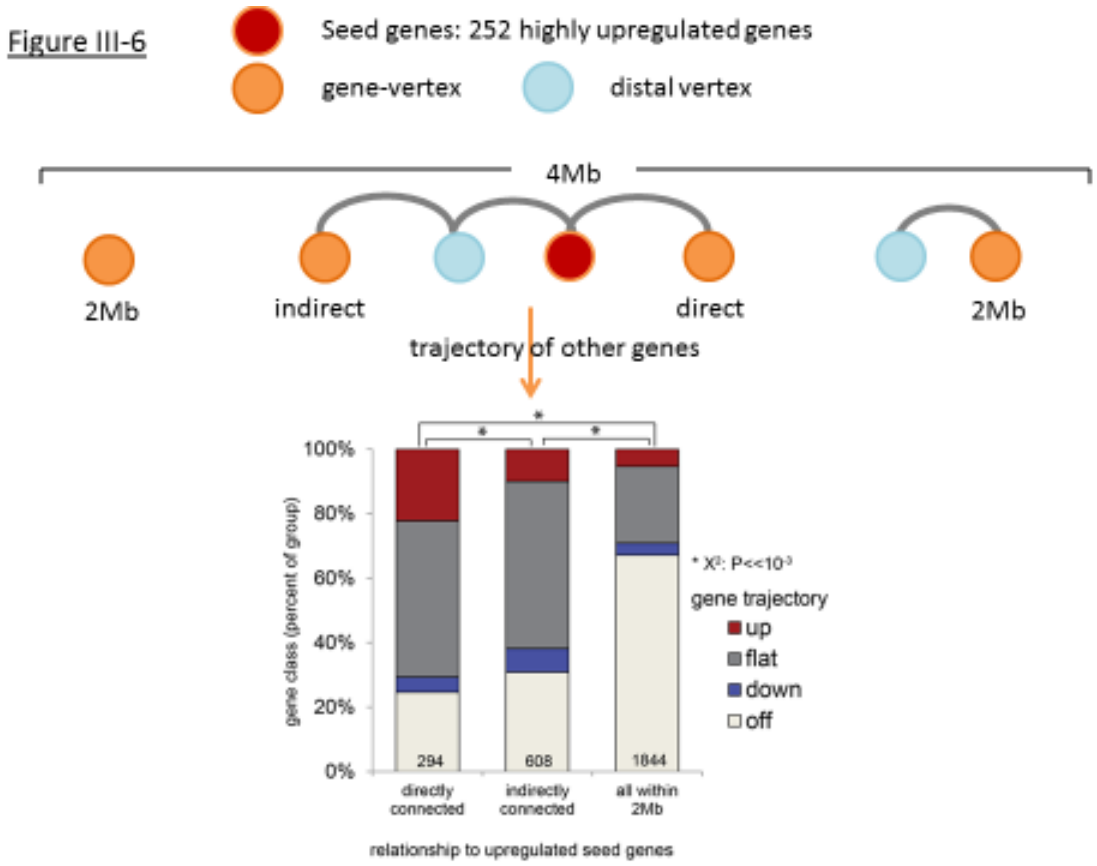


Figure III-6: *The number of unique upregulated (red), flat (gray), downregulated (blue), or off (white) genes that are directly connected to, indirectly connected to, or within 2Mb of 252 highly upregulated seed genes. In the upper cartoon, the red circle is an example upregulated seed gene and the orange circles represent the different reported categories of gene-vertices.*

Figure III-7

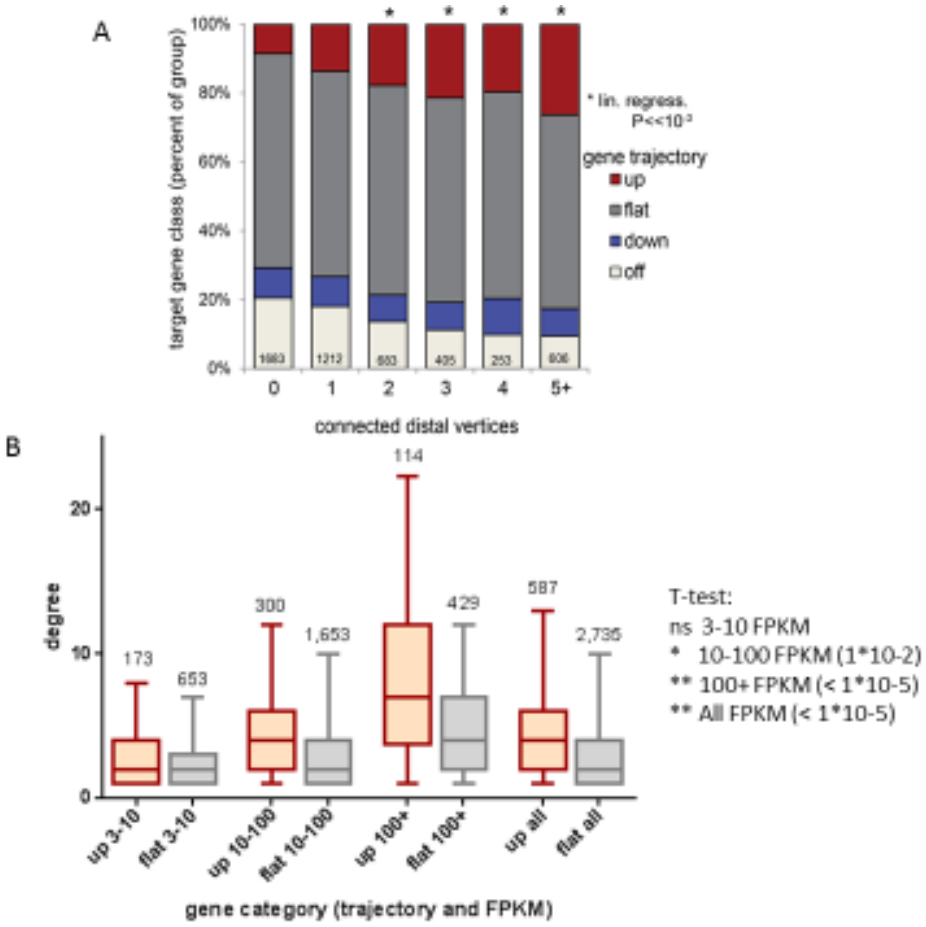


Figure III-7: Upregulated genes have more connected distal vertices than flat expressed genes. *(A) Visualization of a linear regression. Genes are divided according to the number of distal vertices to which they connect (x-axis), and also by their expression pattern: upregulated (red), downregulated (blue), flat (gray), or off (white). (B) Controlling for expression level, upregulated genes (red) have more connected distal elements than flat genes (gray).*

Figure III-8

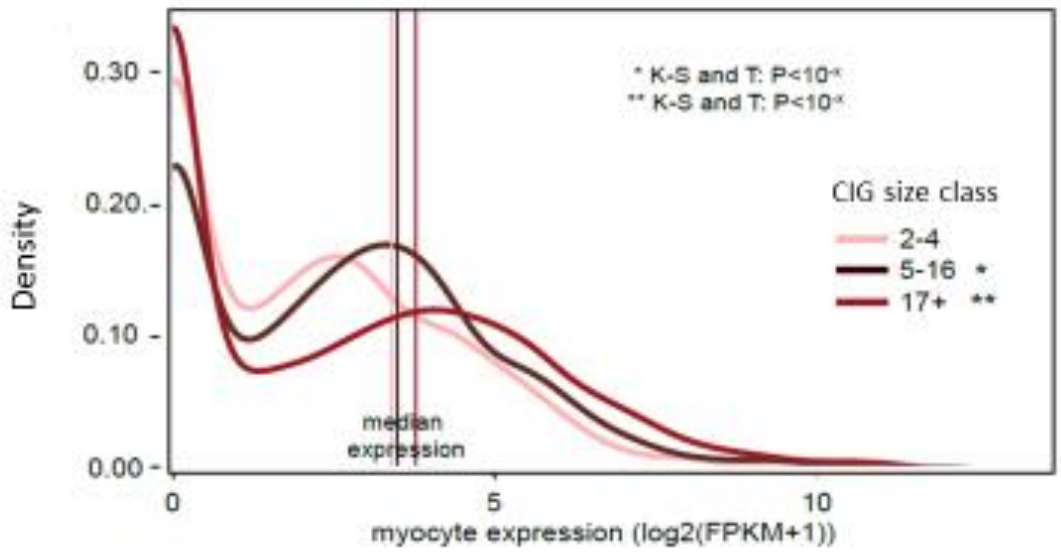


Figure III-8: Highly expressed genes are more associated with large CIGs than with small CIGs. *The distribution of myocyte RNA levels according to the total number of vertices in a gene-vertex's myocyte CIG (CIG size class). Small CIGs (2-4 vertices): pink; medium CIGs (5-16 vertices): dark red; large CIGs (17+ vertices): red. A single asterisk marks a significant difference between CIG classes (P<0.05) in both K-S and T-tests. A double asterisk represents a highly significant difference in both K-S and T-tests (P>1*10^-7). The median RNA levels for gene-vertices in CIGs of each size class (vertical lines) are not significantly different from each other.*

Figure S-1

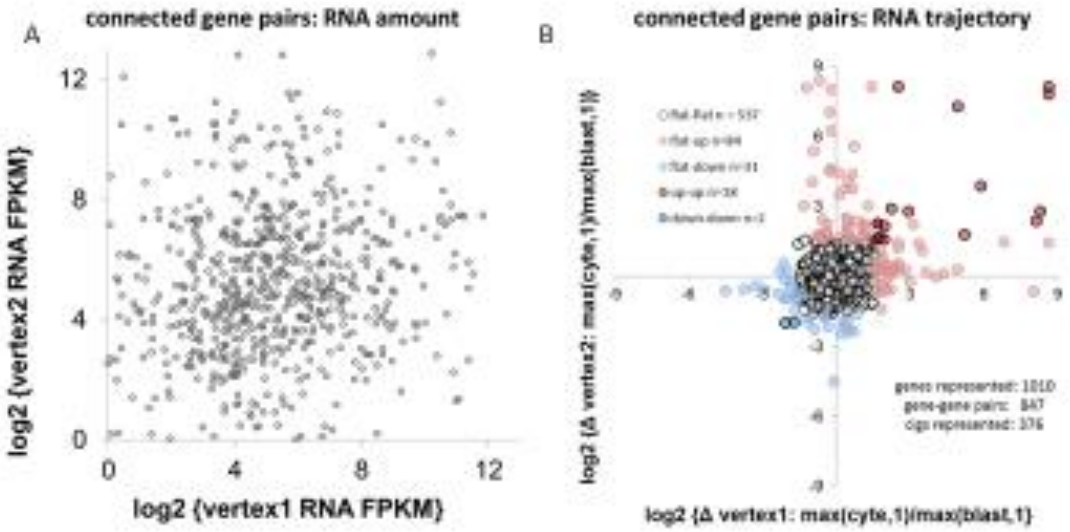


Figure S-1: Lack of correlation between connected promoters. *(A) The amount of myocyte RNA for one gene-vertex vs. the other in every gene-gene pair. (B) The change in RNA for one gene-vertex vs. the other in every gene-gene pair. Data points have been colored according to the paired gene classes: gray: flat-flat; pink: flat-up; light blue: flat-down; red: up-up; dark blue: down-down.*

Sources for Chapter III

Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal Lari, R., . . . Lupien, M. (2015). “ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.” Nat Commun **2:** 6186. doi:10.1038/ncomms7186

Breathnach, R., & Chambon, P. (1981). “Organization and expression of eucaryotic split genes coding for proteins.” Annu Rev Biochem **50**: 349-383. doi:10.1146/annurev.bi.50.070181.002025

Chambeyron, S., & Bickmore, W. A. (2004). “Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription.” Genes Dev **18**(10): 1119-1130. doi:10.1101/gad.292104

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., & Zhao, K. (2012). “Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization.” Cell Res **22**(3): 490-503. doi:10.1038/cr.2012.15

Cheutin, T., O'Donohue, M. F., Beorchia, A., Klein, C., Kaplan, H., & Ploton, D. (2003). “Three-dimensional organization of pKi-67: a comparative fluorescence and electron tomography study using FluoroNanogold.” J Histochem Cytochem **51**(11): 1411-1423.

Ciejek, E. M., Tsai, M. J., & O'Malley, B. W. (1983). “Actively transcribed genes are associated with the nuclear matrix.” Nature **306**(5943): 607-609.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). “Topological domains in mammalian genomes identified by analysis of chromatin interactions.” Nature **485**(7398): 376-380. doi:10.1038/nature11082

Dynan, W. S. (1986). Promoters for housekeeping genes. *Trends in Genetics, 2*, 196-197. doi:10.1016/0168-9525(86)90226-X

Farnham, P. J., & Schimke, R. T. (1985). “Transcriptional regulation of mouse dihydrofolate-reductase in the cell-cycle.” Journal of Biological Chemistry **260**(12): 7675-7680.

Filippova, D., Patro, R., Duggal, G., & Kingsford, C. (2014). “Identification of alternative topological domains in chromatin.” Algorithms Mol Biol **9**: 14. doi:10.1186/1748-7188-9-14

Fisher-Aylor, K. I. (2011). “Long distance looping maps: RNA Pol2 during differentiation.” Nuclear Structure and Dynamics. L'Isle sur la Sorgue, France, EMBO.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., . . . Ruan, Y. (2009). “An oestrogen-receptor-alpha-bound human chromatin interactome.” Nature **462**(7269): 58-64. doi:10.1038/nature08497

Galande, S., Purbey, P. K., Notani, D., & Kumar, P. P. (2007). “The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1.” Curr Opin Genet Dev **17**(5): 408-414. doi:10.1016/j.gde.2007.08.003

Gasser, S. M., & Laemmli, U. K. (1986). “The organisation of chromatin loops: characterization of a scaffold attachment site.” EMBO J **5**(3): 511-518.

Hakim, O., Sung, M. H., Voss, T. C., Splinter, E., John, S., Sabo, P. J., . . . Hager, G. L. (2011). “Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements.” Genome Res **21**(5): 697-706. doi:10.1101/gr.111153.110

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., . . . Wei, C. L. (2011). “CTCF-mediated functional chromatin interactome in pluripotent cells.” Nat Genet **43**(7): 630-638. doi:10.1038/ng.857

Harr, J. C., Luperchio, T. R., Wong, X., Cohen, E., Wheelan, S. J., & Reddy, K. L. (2015). “Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins.” J Cell Biol **208**(1): 33-52. doi:10.1083/jcb.201405110

Iborra, F. J., Pombo, A., Jackson, D. A., & Cook, P. R. (1996). “Active RNA polymerases are localized within discrete transcription "factories' in human nuclei.” J Cell Sci **109**(Pt 6): 1427-1436.

Jackson, D. A., Hassan, A. B., Errington, R. J., & Cook, P. R. (1993). “Visualization of focal sites of transcription within human nuclei.” EMBO J **12**(3): 1059-1065.

Kieffer-Kwon, K. R., Tang, Z., Mathe, E., Qian, J., Sung, M. H., Li, G., . . . Casellas, R. (2013). “Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation.” Cell **155**(7): 1507-1520. doi:10.1016/j.cell.2013.11.039

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., . . . Ren, B. (2007). “Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.” Cell **128**(6): 1231-1245. doi:10.1016/j.cell.2006.12.048

Kosak, S. T., & Groudine, M. (2004). “Gene order and dynamic domains.” Science **306**(5696): 644-647. doi:10.1126/science.1103864

Lercher, M. J., A. O. Urrutia and L. D. Hurst (2002). "Clustering of housekeeping genes provides a unified model of gene order in the human genome." Nat Genet **31**(2): 180-183.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., . . . Ruan, Y. (2012). “Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.” Cell **148**(1-2): 84-98. doi:10.1016/j.cell.2011.12.014

Li, G. L., Ruan, X. A., Auerbach, R. K., Sandhu, K. S., Zheng, M. Z., Wang, P., . . . Ruan, Y. J. (2012). “Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation.” Cell **148**(1-2): 84-98. doi:10.1016/j.cell.2011.12.014

Li, Y., Huang, W., Niu, L., Umbach, D. M., Covo, S., & Li, L. (2013). “Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes.” BMC Genomics **14**: 553. doi:10.1186/1471-2164-14-553

Magistri, M., Faghihi, M. A., St Laurent, G., 3rd, & Wahlestedt, C. (2012). “Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts.” Trends Genet **28**(8): 389-396. doi:10.1016/j.tig.2012.03.013

Meshorer, E., & Misteli, T. (2006). “Chromatin in pluripotent embryonic stem cells and differentiation.” Nat Rev Mol Cell Biol **7**(7): 540-546. doi:10.1038/nrm1938

Mirkovitch, J., Mirault, M. E., & Laemmli, U. K. (1984). “Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold.” Cell **39**(1): 223-232.

Mitchell, J. A., & Fraser, P. (2008). “Transcription factories are nuclear subcompartments that remain in the absence of transcription.” Genes Dev **22**(1): 20-25. doi:10.1101/gad.454008

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” Nat Methods **5**(7): 621-628. doi:10.1038/nmeth.1226

Muller, H. P., Sogo, J. M., & Schaffner, W. (1989). “An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge.“ Cell **58**(4): 767-777. doi:10.1016/0092-8674(89)90110-4

Noordermeer, D., Branco, M. R., Splinter, E., Klous, P., van Ijcken, W., Swagemakers, S., . . . de Laat, W. (2008). “Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region.” PLoS Genet **4**(3): e1000016. doi:10.1371/journal.pgen.1000016

Ong, C. T., & Corces, V. G. (2014). “CTCF: an architectural protein bridging genome topology and function.” Nat Rev Genet **15**(4): 234-246. doi:10.1038/nrg3663

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., . . . Fraser, P. (2004). “Active genes dynamically colocalize to shared sites of ongoing transcription.” Nat Genet **36**(10): 1065-1071. doi:10.1038/ng1423

Palstra, R. J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B., & de Laat, W. (2008). “Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription.” PLoS One **3**(2): e1661. doi:10.1371/journal.pone.0001661

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., . . . van Steensel, B. (2010). “Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.” Mol Cell **38**(4): 603-613. doi:10.1016/j.molcel.2010.03.016

Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., . . . Corces, V. G. (2013). “Architectural protein subclasses shape 3D organization of genomes during lineage commitment.” Cell **153**(6): 1281-1295. doi:10.1016/j.cell.2013.04.053

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W. S., Denas, O., . . . Gilbert, D. M. (2014). “Topologically associating domains are stable units of replication-timing regulation.” Nature **515**(7527): 402-+. doi:10.1038/nature13986

Robinson, S. I., Small, D., Idzerda, R., McKnight, G. S., & Vogelstein, B. (1983). “The association of transcriptionally active genes with the nuclear matrix of the chicken oviduct.” Nucleic Acids Res **11**(15): 5113-5130.

Shopland, L. S., Lynch, C. R., Peterson, K. A., Thornton, K., Kepper, N., Hase, J., . . . O'Brien, T. P. (2006). “Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence.” J Cell Biol **174**(1): 27-38. doi:10.1083/jcb.200603083

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., . . . de Laat, W. (2006). “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).” Nat Genet **38**(11): 1348-1354. doi:10.1038/ng1896

Small, D., Nelkin, B., & Vogelstein, B. (1985). “The association of transcribed genes with the nuclear matrix of Drosophila cells during heat shock.” Nucleic Acids Res **13(**7): 2413-2431.

Splinter, E., Heath, H., Kooren, J., Palstra, R. J., Klous, P., Grosveld, F., . . . de Laat, W. (2006). “CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.” Genes Dev **20**(17): 2349-2354. doi:10.1101/gad.399506

Stathopoulos, A., & Levine, M. (2002). “Whole-Genome Expression Profiles Identify Gene Batteries in Drosophila.” Dev Cell **3**(4): 464-465. doi:http://dx.doi.org/10.1016/S1534-5807(02)00300-3

Verschure, P. J., van Der Kraan, I., Manders, E. M., & van Driel, R. (1999). “Spatial relationship between transcription sites and chromosome territories.” J Cell Biol **147**(1): 13-24.

Versteeg, R., B. D. van Schaik, M. F. van Batenburg, M. Roos, R. Monajemi, H. Caron, H. J. Bussemaker and A. H. van Kampen (2003). "The human transcriptome map

reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes." Genome Res **13**(9): 1998-2004.

Yee, S. P., & Rigby, P. W. (1993). “The regulation of myogenin gene expression during the embryonic development of the mouse.” Genes Dev **7**(7A): 1277-1289.

Zeng, M. e. a. (2015). in review.

Chapter IV: Conclusions

IV.1: Introduction

In this thesis, I have described results that are true in all of the ChIA-PET datasets I have analyzed regardless of the factor ChIPped, the developmental timepoint, or the numerous different analysis approaches that I used. There are two primary principles which in my opinion can explain most of the other ChIA-PET results I described in this thesis. First, physical connectivity between transcription factor- or pol2-occupied elements in the genome is, in terms of the DNA backbone, primarily local; there are fewer than one percent the number of 100kb edges as the number of 10kb edges in every dataset, and fewer than two hundred high-confidence, reproducible edges over 150kb. Second, the amount of gene expression measured by RNA-Seq is highly correlated with the likelihood that a gene will have a ChIA-PET connection. Developmental genes might show slightly different patterns of engagement, and the implications for how genes of average expression profiles might be expressed is that most may differ only in degree (they have fewer enhancers), not in type (only operated by a basal promoter and mostly not having enhancers), though of course much exciting gain-of-function and loss-of-function experimental work is required to substantiate these final two hypotheses.

IV.2: Connectivity of active elements is much more prevalent than expected and most connections are local (<50kb)

When the genome-wide assay ChIP-Seq was invented, one of the first surprises when studying tissue-specific transcription factors was how often they occupied sites in the genome (Cao, Yao et al. 2010; Kwan G, Kirilusha A, Fisher-Aylor K, unpublished). For example, MyoD and Myog occupy more than 14,000 sites in the genome, even though there are only a few hundred muscle-specific genes (also see Fig. I-2). In an

analogous surprise, when I used a genome-wide assay to determine how many active genes and occupied distal elements were physically connected to each other, I found connectivity almost everywhere there were areas of widespread DNase accessibility. This means that the majority of active genes, no matter what type of expression they have, show physical connectivity to active elements that are near to them. Although there very well may be some notable exceptions to this, it appears, based on these data, that what a gene connects to is determined in large part by its genomic neighborhood. It is possible that there are two different approaches to transcription between genes in gene-dense, highly inter-engaged areas and in areas that are gene-poor or otherwise sparse in CRMs.

IV.3: Most ChIA-PET connectivity occurs sequentially rather than simultaneously

One surprise to me in the ChIA-PET data was how little connectivity varied when counting the relationship between vertices and edges across the genome. Although by large the ChIA-PET data consists of simple connections in gene desert areas and extremely large interacting graphs in gene dense areas, the relationship between vertices and edges remains essentially constant. This likely relates to the first principle that shorter edges are more common than longer (Fig. II-7). However, it is likely possible to take this into account in a way that could determine if there are different modes of interactivity in the genome. In order to do this, the existing edge and vertex data could be used to create different models of interaction, such as cooperative binding of multiple elements to a gene, independent binding of multiple elements to a gene, or binding of elements to a gene in a way that is mutually exclusive. Treating these three models as, in effect, null hypotheses for which to create P values on a graph by graph basis would help determine if there are any loci that exhibit classic examples of

cooperative, independent, or mutually exclusive activation of a promoter and if so, whether this co-occurs with different classes of genes.

IV.4: Possible implications for the regulation of developmental and housekeeping genes

Although most detected interactions in the genome are simple, there are nevertheless many thousand instances of one element connecting to multiple other elements. In congruence with the knowledge that many developmental genes used differently across time and tissue type have multiple enhancers, with these enhancers often being identified as active in only one tissue/at only one time, there are hundreds of examples in ChIA-PET of one gene connecting to multiple distal elements. However, there are also hundreds of cases (though fewer cases than the reverse) of one distal element connecting to multiple different genes. There were a very small number of known examples of this phenomenon before the very recent advent of high-throughput connectivity assays, and many known examples involved developmentally related genes. My data show that many enhancers are even shared between genes with different expression patterns; for example, the majority of genes in CIGs containing myogenic genes are expressed at a steady level over development, though I rarely found instances of genes with opposite expression patterns in the same CIGs as each other.

Most expressed genes are expressed at a low and steady level, and a subset of these genes can be thought of as “housekeeping genes”: genes which are on in all cell types (or almost all, if one thinks critically about gene expression in cells such as sperm and eggs) because their products are necessary for a cell’s basic existence. However, many steadily expressed genes have detectible connections to distal elements in both cell states. While it is possible that some of these connections are structural and not

otherwise regulatory, it is also possible that most genes, regardless of what their product is used for and when, might be regulated by enhancers.

In this myogenic system, I did detect two ways in which myogenic genes engage with distal elements in subtly different ways from other genes. First, myogenic genes are slightly more likely than predicted by gene neighborhood to engage with other myogenic genes, although the majority of genes a myogenic gene connects to are steadily expressed. Second, myogenic genes are more likely than flat genes of the same expression levels to engage with multiple distal elements. Together, these observations predict that developmentally active genes may be regulated in a subtly different way from other genes, but in a way that differs in degree rather than type. Perhaps most promoters require input from an enhancer, while developmentally regulated genes require more enhancers in order to change their expression levels across time and tissue.

IV.5: Take-away lessons for other biologists

In the process of completing this larger ChIA-PET project, I have determined some useful relationships and rules for bioinformatics that do not fit within my larger narrative. First, from analyzing tissue-specific transcription factors in two different systems, Twist in very early fly embryos and Myogenin in C2C12 mouse muscle, I found that 400 base pairs is likely the average width of a CRM (S. Pepke, K. Fisher, A. Ozdemir, and A Kirilusha, data not shown). As expected, since these two observations agreed despite how different the systems were, an analysis of genome-wide DNase in this system (C2C12), which looks at many different types of factors simultaneously, also found that 400 base pairs was a good estimate of the average CRM (Ramirez, R unpublished).

In studying transcription factors genome-wide, it is often unclear how to assign a particular instance of occupancy to a gene that it might express. Typically, an occupied site is assigned to its nearest promoter, sometimes including rules such as the promoter being downstream, or the enhancer being within a certain distance of a gene (such as the 2KB window where most enhancers fall). My analysis shows that connectivity can be predicted reasonably well using RNA-seq and a genome-wide footprinting assay such as DNase or ATAC-seq. This is because while most unexpressed genes are not connected, expressed genes connect to many elements that are close to them. It is appropriate in the absence of connectivity data to assign a region occupied by an active transcription factor to the nearest gene that is expressed within about 50 kb (Fig. II-7, Fig. III-2). The only caveat to doing this is that in areas that are gene dense or highly occupied by transcription factors, the site in question may additionally connect to other genes and elements (Fig. III-2).

ChIP-Seq signal size cannot correlate to amount of factor occupancy, as is a common erroneous assumption. In addition to the more likely possibility that sample heterogeneity could also cause differences in signal size from locus to locus, it is a fact that sequence content affects ChIP-Seq results. All samples that are processed on Illumina sequencers, and likely on other types of high throughput sequencers as well, go through a PCR step, called library building, which is necessary in order to obtain enough material to sequence. However, this means that since one library building method is typically used as the standard method, most libraries will exhibit the same bias which is derived from the temperature and time setting of the PCR protocol itself. In the case of Illumina platform sequencing, the PCR protocol appears to have been chosen so that all libraries of 200 base pair fragments return an average sequence content of 65% G/C, which is the overall sequence content of most mammalian genomes, including human.

However, this does a disservice to any 200-bp segments of DNA that are more A/T or G/C rich than the mammalian average. This may explain partially why some factors in our system that occupy A/T rich motifs, such as SRF and MEF2, have been difficult to ChIP. It also explains why some of the existing data sets for factors that bind A/T rich elements, such as anterior-posterior patternin transcription factors in early *Drosophila* development (for review, see MacArthur, Li et al. 2009), contain a modified library-building protocol. Other factors might be difficult to CHIP because the areas they occupy are difficult to shear into 200 bp fragments. This likely occurs for factors that bind in or near repressed regions of the chromosome because all current ChIP-Seq fragmentation methods, including sonication, preferentially cleave active areas of the DNA (Auerbach, Euskirchen et al. 2009). Therefore, if a factor that needs to be ChIPped is thought to be repressive (therefore not easily accessible to any current fragmentation protocol) or binds an extremely A/T or G/C rich motif or area, it will likely be necessary to change the conventional ChIP-Seq protocol.

IV.6: Paths forward

ChIP-Seq showed us that the well-studied instances of the functional MyoD or Myog binding near muscle-specific genes were not incorrect, they just did not appear to be as unique as they previously had been assumed to be. Likewise, the previously characterized functional connections in this system are supported by ChIA-PET, but do not appear special or unique. Highly expressed and differential genes appear to be connected to an exceptionally high degree; however, constitutively expressed and modestly expressed genes also connect to active elements nearby, including elements occupied by muscle-specific transcription factors. Based on these observations, it is quite possible that enhancers are more prevalent than previously thought. However, it would be jumping to conclusions to conclude that a ChIA-PET edge represents a

functional interaction (Li, Ruan et al. 2012). The current claims by my contemporaries who are doing ChIA-PET that ChIA-PET-unconnected elements are far apart from each other are not substantiated by any evidence. Despite suggestions otherwise, which are based on a careful selection of loci and an experimental protocol which involved a non-conventional step to bloat the nucleus and presumably tear apart weak connections, the absence of an edge in ChIA-PET tells you nothing scientifically about connectivity (Fig. II-5).

These uncertainties lead to several obvious scientific questions. What does connectivity look like where it is invisible to ChIA-PET? Which connected elements can be considered classic enhancers? Which connections are functional, as opposed to incidental? The first of these questions can be answered by classic 3C and FISH assays, and in my opinion, the most interesting loci to study first would be the MYOG loci which do not show ChIA-PET connectivity in the preceding myoblast state. It is entirely possible that the connections that seemingly appear upon differentiation are already established in the myoblast state but are invisible to ChIA-PET because they lack pol2. This specific question relates to a deeper one: is physical association with a promoter a cause or an effect of transcription factor occupancy, and is the answer the same for every occupied site? A second related question is when, during the course of development, are physical connections established? The current view of the field that makes the most sense to me is that some connections, such as those thought to occur in classic CTCF insulation, are established early in development in order to mark large active areas of chromatin, but that the interactions that occur within these domains are primarily transient enhancer to promoter interactions (Kim, Abdullaev et al. 2007; Chepelev, Wei et al. 2012). Some of these questions could be addressed using 4C or DNA-FISH at a well-studied and massively up-regulated large genomic locus, like the

area around Myog the gene. Preliminary experiments of both kinds have been executed by my colleague Say-Tar Goh and suggest that the connections within this locus already exist in the myoblast state. To take the question further and ask when these connections are formed, the same experiment could be performed in developmentally earlier cells, such as 10T1/2 or multipotent mesenchymal cells.

Another open question is which elements that connect to genes are functionally connected to these genes. This is a complex question that can only be answered using several different types of experiments, both gain-of-function and loss-of-function. An overview of current classic enhancer assays based on high-throughput genomics data, both published and unpublished, suggests that between 50 to 80% of transcription factor occupied elements (predicted by ChIP-Seq) increase the activity of a generic promoter (Ozdemir, Fisher-Aylor et al. 2011; ENCODE consortium unpublished; Desalvo, G unpublished). In this system, preliminary evidence suggests that about 50% of occupied elements are definitely enhancers, and this is independent of ChIA-PET connectivity. However, gain-of-function assays such as this mean little when they are negative. For example, an element might regulate the promoter to which it is connected in the cell but not the promoter used in the enhancer assay. Other elements, such as the intronic MCK enhancer (Tai, Fisher-Aylor et al. 2011), only function in one orientation, although it is not the promoter. Because of this, it is necessary to use loss-of-function assays also in order to determine how transcription initiation truly occurs. With CRISPR this is now economically feasible. Occupied and connected sites should be deleted, first one-by-one and then in combination, in order to determine which loci affect gene expression.

One of the more interesting types of ChIA-PET edges to investigate is the surprisingly prevalent promoter-to-promoter edges ((Li, Ruan et al. 2012), Fig. II-8). One of the reasons I created the Myog ChIA-PET was to determine which characteristics of

pol2 ChIA-PET were specific to pol2 itself. The fact that numerous promoter-to-promoter edges still occur in the Myog ChIA-PET as well as other characteristics that might have been suspected to be pol2 specific, such as the phenomenon of more highly expressed genes being more likely connected, makes it more likely that these edges are characteristic of active genes. However, since Myog is found at promoters as well as enhancers, it would be useful to determine whether these phenomena exist using a factor for ChIA-PET that occupies enhancers but not promoters.

However, it has been found in the past that promoters can connect to other promoters and in some cases can alter one another functionally. One of the first experiments done upon the discovery of enhancers found a promoter in an enhancer trap experiment (Weber, de Villiers et al. 1984) and then demonstrated that this promoter acted as a functional enhancer (Serfling, Lubbe et al. 1985). Possibly related are the phenomena that a few unexpressed genes show physical connections (Fig. III-1 (Osborne, Chakalova et al. 2004; Sanyal, Lajoie et al. 2012; Noordermeer and Duboule 2013). Perhaps some of these instances result from inactive promoters acting as enhancers rather than enhancers interacting with a promoter before it is expressed, as is commonly assumed. These two possibilities are not mutually exclusive. Genes are thought to cycle between off and on (Ross 1994, Wijarde 1995, Milot 1996, Kimura 2002, Levsky 2002, Osborne 2004), so even if promoters were only able to act as enhancers when they themselves are unexpressed, it is possible that this is indeed happening in the system. Another possibly related phenomenon is that of pol2 pausing. Many unexpressed genes have paused pol2 (Zeitlinger, Stark et al. 2007) especially genes that are tissue specific (Hendrix, Hong et al. 2008). These paused promoters can act as insulators (Core and Lis 2009), so perhaps they could act as other types of CRM as well. Nor is it necessarily true that promoters and enhancers are as different as we

thought from the perspective of transcription biochemistry. When it was first noticed that general transcription factors and pol2 were sometimes found at enhancers (Koch, Fenouil et al. 2011), particularly using strong fixatives (Kwan, G and Fisher-Aylor, K unpublished data), it was assumed that they were a result of indirect binding to the enhancer via a connected promoter. However, recent experiments suggest that some, or perhaps even most, enhancers produce non-coding RNA; in other words, enhancers could at an extreme be stray, non-coding promoters. When different genome-wide measurements of RNA were used, it appeared that, at an extreme, transcription occurred almost everywhere in the genome that was active, coding or not (Consortium 2012; Djebali, Davis et al. 2012). The same group that proposed enhancers may be stray promoters found that enhancer RNAs might contribute structurally to the transcription initiation complex (Lis, Core et al. 2015). A way to determine in this system how important promoter-promoter connections are to gene expression would involve knocking out one promoter at a time and determining what, if any, change in expression occurs in the putatively connected promoters as a result (for example, using CRISPR to create a line of C2s missing one promoter, then using RNA-Seq on the myoblasts and myocytes of the new line compared with our existing C2 RNA-Seq to identify possibly affected promoters).

Sources for Chapter IV

Auerbach, R. K., G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefrancois, K. Struhl, M. Gerstein and M. Snyder (2009). "Mapping accessible chromatin regions using Sono-Seq." Proc Natl Acad Sci U S A **106**(35): 14926-14931.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang and K. Zhao (2012). "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization." Cell Res **22**(3): 490-503.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Core, L. J. and J. T. Lis (2009). "Paused Pol II captures enhancer activity and acts as a potent insulator." Genes Dev **23**(14): 1606-1612.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo and T. R. Gingeras (2012). "Landscape of transcription in human cells." Nature **489**(7414): 101-108.

Hendrix, D. A., J. W. Hong, J. Zeitlinger, D. S. Rokhsar and M. S. Levine (2008). "Promoter elements associated with RNA Pol II stalling in the Drosophila embryo." Proc Natl Acad Sci U S A **105**(22): 7762-7767.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov and B. Ren (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Koch, F., R. Fenouil, M. Gut, P. Cauchy, T. K. Albert, J. Zacarias-Cabeza, S. Spicuglia, A. L. de la Chapelle, M. Heidemann, C. Hintermair, D. Eick, I. Gut, P. Ferrier and J. C. Andrau (2011). "Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters." Nat Struct Mol Biol **18**(8): 956-963.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. Ruan (2012). "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." Cell **148**(1-2): 84-98.

Lis, J., L. Core, A. Martins, C. Danko, A. Siepel, G. Booth, F. Duarte and D. B. Mahat (2015). "A Unified Model Describing The Architecture And Creation Of Promoters And Enhancers." The FASEB Journal **29**(1 Supplement).

MacArthur, S., X. Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. Keranen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin and M. B. Eisen (2009). "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions." Genome Biol **10**(7): R80.

Noordermeer, D. and D. Duboule (2013). "Chromatin looping and organization at developmentally regulated gene loci." Wiley Interdiscip Rev Dev Biol **2**(5): 615-630.

Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik and P. Fraser (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." Nature **489**(7414): 109-U127.

Serfling, E., A. Lubbe, K. Dorsch-Hasler and W. Schaffner (1985). "Metal-dependent SV40 viruses containing inducible enhancers from the upstream region of metallothionein genes." EMBO J **4**(13B): 3851-3859.

Tai, P. W. L., K. I. Fisher-Aylor, C. L. Himeda, C. L. Smith, A. P. MacKenzie, D. L. Helterline, J. C. Angello, R. E. Welikson, B. J. Wold and S. D. Hauschka (2011). "Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer." Skeletal Muscle **1**.

Weber, F., J. de Villiers and W. Schaffner (1984). "An SV40 "enhancer trap" incorporates exogenous enhancers or generates enhancers from its own sequences." Cell **36**(4): 983-992.

Zeitlinger, J., A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine and R. A. Young (2007). "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." Nat Genet **39**(12): 1512-1516.

Chapter V: Materials and Methods

V.1: Cell growth

C2C12 cells were grown and differentiated according to the standard protocol (see supplemental materials and methods). The myoblast cells were harvested at less than 40% confluence, and the myocyte cells were harvested 60 hours after the start of differentiation.

V.2: ChIA-PET

C2C12 cells for ChIA-PET were fixed in 1.5mM EGS/1% formaldehyde (see supplemental materials and methods). The DNA was sheared to an approximate length of 375 bp., and sheared chromatin was ChIPped using an RNA polymerase II antibody (4H8; Millipore) according to the standard ChIP-Seq protocol (Johnson et al. 2007; supplemental materials and methods) with the following modifications: the sonicate from 5*10^7 nuclei were used for each ChIP reaction (with 5ug of antibody), and all incubation times were doubled. After ChIPping, the nuclei from seven individual ChIPs were pooled for one ChIA-PET samples, and were shipped on-bead to Singapore for library building. Two individual ChIA-PET biological and technical replicates were sequenced for each sample. Libraries were sequenced on the Illumina platform using custom primers as previously described (Fullwood, Liu et al. 2009).

V.3: DNase-seq

DNase-seq experiments and primary analysis were performed by Ricardo Ramirez in the laboratory of Ali Mortazavi at UC Irvine, with minor changes in experimental protocol in two biological replicates on C2C12 exponentially growing cells and on 60 hr horse serum treated C2C12 cells (Ko et al. 2013). This section was authored in party by Ricardo Ramirez. Approximately a total of ~600M (130-160M reads per replicate) Illumina Hi-seq 2500 single 50bp DNase-seq reads sequenced. DNase-

seq reads were mapped to the mm9 reference genome using Bowtie (Langmead et al. 2009). Hotspot version 4 (Thurman et al. 2012; John, S et al. 2011) was used to determine DHS peaks for each replicate and the intersection of Hotspot calls (FDR < 1%) was performed in the subsequent analysis. High quality DNase-seq data was determined by the efficiency (fraction of mapped reads in Hotspot calls) or SPOT score as calculated by Hotspot, with data ranging between 48-68% for all replicates. These Hotspot calls were used in the "3kb"/low-resolution version of the analysis (which is not the primary version reported in this work, but was one of several used to substantiate and test the conclusions).

DNaseI footprinting was performed using the Wellington method (Piper et al. 2013) for each cell type by combining DHS reads for both replicates. Approximately 300 million DHS reads were used to compute DNaseI wFootprints genome-wide for each cell type respectively. The resulting footprints were used in the "Wellington" analysis, which is the primary analysis reported in this work.

V.4: ChIA-PET raw data processing

After sequencing, the raw data were stripped of chimeric reads and reads without linkers (Fullwood et al. 2009). Non-chimeric paired reads were then mapped to the mm9 genome at 100% match, and pairs of reads closer together than 10kb or farther apart than 2Mb were discarded.

Paired-reads with concordant half-linkers were mapped onto the UCSC mm9 genome using bowtie 0.12.7 and no mismatches (Langmead et al. 2009). Reads were then processed with ERANGE 3.3 (Mortazavi, Williams et al. 2008), and paired-reads that mapped uniquely on the same chromosome more than 4.510 kb and less than 2 Mb apart were discarded. Connections mapping between different chromosomes were not analyzed. Such interchromosomal interactions primarily fell into areas with many

repeats or areas without evidence of chromatin accessibility. Many such interactions may represent intrachromosomal interactions spanning chromosomal rearrangements in this unsequenced cell line, which might also be pseudotetraploid because similar C2C12s (not the same because the Wold C2C12 lineage is different than the one deposited in the cell bank ACTT) are pseudotetraploid (Casas-Delucchi, Brero et al. 2011; supplemental S1). However, when considering only the interchromasomal PETs that fell within CIG vertices (which are so sparse they number in the low hundreds even when accepting interchromosomal edges supported two PETs in only one library), I noticed that the most highly connected regions in the conventional intrachromosomal analyses were the regions with the most likelihood of containing interchromosomal edges as well.

V.5: Construction of ChIA-PET candidate vertices

To create a set of candidate vertices upon which the ChIA-PET data was mapped, Wellington calls from myocyte and myoblast DNase data were expanded to +/-500bp around their peaks, then pooled together with all annotated Gencode M1 (Coffey, Kokocinski et al. 2011; Derrien, Johnson et al. 2012; Harrow, Frankish et al. 2012; Frankish, Uszczynska et al. 2015) PC and LNC TSS's. TSS's were expanded to -600bp and +400bp based on an in-house analysis which found the majority of DHS signal at annotated promoters between -600 and +400 of the TSSs. Any overlapping regions regardless of their source were merged together. The resulting regions are referred to in this work as "candidate vertices" and the analysis based on these vertices as the "Wellington" analysis.

Two different analyses were also performed to ensure whether and which elements of the ChIA-PET analysis were predicated on assumptions we made early in the analysis process. First, to capture more ChIA-PET ends (at the expense of lower

resolution), we created candidate vertices using HotSpot calls for myoblast and myocyte DNase data. Overlapping regions were merged together into one region, and all regions that were narrower than 3kb were expanded to +/-1.5kb around their midpoint. Second, to determine if and which elements connected by ChIA-PET were being left out of the analysis due to the requirement to be near DNase hypersensitive regions, I used ERANGE on the 25bp PET ends (unpaired) – from the second technical replicates of pol2 ChIA-PET for myoblast and myocyte – with no background library to call "pileups" of ChIA-PET reads. These regions became candidate vertices with no expansion and no addition of annotated TSSs in what I termed the "data-driven" analysis. The "data-driven" analysis had the additional benefit of having highly connected vertices that were less than 1kb wide. These two medium-confidence ("3kb") and extremely high-confidence ("data-driven") analyses showed essentially the same answer for all of the analyses that I have included in this thesis, which is the reason I am so confident that the results I report here represent general characteristics of ChIA-PET data.

V.6: Construction of CIGs

The subset of accepted ChIA-PET raw paired-end tags previously described were mapped onto each set of candidate vertices using the following rules. Pairs falling within the same vertex as each other, or where only one mate fell into a candidate vertex, were discarded. An edge is defined as a pair of vertices that is spanned by a pair of partnered raw reads; edges supported by fewer than 2 individual sets of PETs were discarded. In order to focus only on edges that are most likely to be signal instead of noise, one final intersect set of edges was constructed for each of the two timepoints: these sets require that an edge be present in each of the two biological/technical replicate ChIA-PET datasets, and the final edge weight is the average of its weight in the two individual replicates. We normalized the edge weights by the average length of the

end-point vertices to make edges comparable between vertices of different widths. CIGs were defined as fully connected sub-graphs of the parent set of candidate vertices. The final ChIA-PET ("*.matrix") datasets submitted displays all edges, connected vertices, and CIGs on a vertex-by-vertex basis.

Sources for Chapter V

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Y. Ren, W. W. Li and W. S. Noble (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Research **37**: W202-W208.

Casas-Delucchi, C. S., A. Brero, H. P. Rahn, I. Solovei, A. Wutz, T. Cremer, H. Leonhardt and M. C. Cardoso (2011). "Histone acetylation controls the inactive X chromosome replication dynamics." Nat Commun **2**: 222.

Coffey, A. J., F. Kokocinski, M. S. Calafato, C. E. Scott, P. Palta, E. Drury, C. J. Joyce, E. M. Leproust, J. Harrow, S. Hunt, A. E. Lehesjoki, D. J. Turner, T. J. Hubbard and A. Palotie (2011). "The GENCODE exome: sequencing the complete human exome." Eur J Hum Genet **19**(7): 827-831.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow and R. Guigo (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.

Frankish, A., B. Uszczynska, G. R. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo and J. Harrow (2015). "Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction." BMC Genomics **16 Suppl 8**: S2.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung and Y. Ruan (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." Nature **462**(7269): 58-64.

Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo and T. J. Hubbard (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774.

John, S., P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager and J. A. Stamatoyannopoulos (2011). "Chromatin accessibility pre-determines glucocorticoid receptor binding patterns." Nature Genetics **43**(3): 264-U116.

Johnson, D. S., A. Mortazavi, R. M. Myers and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Ko, I. K., B. K. Lee, S. J. Lee, K. E. Andersson, A. Atala and J. J. Yoo (2013). "The effect of in vitro formation of acetylcholine receptor (AChR) clusters in engineered muscle fibers on subsequent innervation of constructs in vivo." Biomaterials **34**(13): 3246-3255.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3): 1-10.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Piper, J., M. C. Elze, P. Cauchy, P. N. Cockerill, C. Bonifer and S. Ott (2013). "Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data." Nucleic Acids Research **41**(21).

Robinson, M. D. and G. K. Smyth (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics **9**(2): 321-332.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Z. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Y. Song, S. Vong, M. Weaver, Y. Q. Yan, Z. C. Zhang, Z. Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. S. Wu, T. Ryba, R. Sandstrom, Z. H. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L. H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. W. Li, M. A. Bender, M. H. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y. C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. L. De Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. R. Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K. H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X. Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, B. Ren and E. C. Mouse (2014). "A comparative encyclopedia of DNA elements in the mouse genome." Nature **515**(7527): 355-+.

Chapter Supplemental I

Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer

Phillip WL Tai, Katherine I Fisher-Aylor, et al. (2011), Skeletal Muscle **1**:25

S1.1 Abstract

Hundreds of genes, including muscle creatine kinase (*MCK*), are differentially expressed in fast- and slow-twitch muscle fibers, but the fiber type-specific regulatory mechanisms are not well understood.

Modulatory region 1 (MR1) is a 1-kb regulatory region within *MCK* intron 1 that is highly active in terminally differentiating skeletal myocytes *in vitro*. A *MCK* small intronic enhancer (*MCK*-SIE) containing a paired E-box/myocyte enhancer factor 2 (MEF2) regulatory motif resides within MR1. The SIE's transcriptional activity equals that of the extensively characterized 206-bp *MCK* 5'-enhancer, but the *MCK*-SIE is flanked by regions that can repress its activity via the individual and combined effects of about 15 different but highly conserved 9- to 24-bp sequences. ChIP and ChIP-Seq analyses indicate that the SIE and the *MCK* 5'-enhancer are occupied by MyoD, myogenin and MEF2. Many other E-boxes located within or immediately adjacent to intron 1 are not occupied by MyoD or myogenin. Transgenic analysis of a 6.5-kb *MCK* genomic fragment containing the 5'-enhancer and proximal promoter plus the 3.2-kb intron 1, with and without MR1, indicates that MR1 is critical for *MCK* expression in slow- and intermediate-twitch muscle fibers (types I and IIa, respectively), but is not required for expression in fast-twitch muscle fibers (types IIb and IId).

In this study, we discovered that MR1 is critical for *MCK* expression in slow- and intermediate-twitch muscle fibers and that MR1's positive transcriptional activity depends

on a paired E-box MEF2 site motif within a SIE. This is the first study to delineate the DNA controls for *MCK* expression in different skeletal muscle fiber types.

S1.2 Background

Muscle creatine kinase (*MCK*) is among the most abundant transcripts in striated muscle [1]. In differentiating muscle cell cultures, the onset of *MCK* expression occurs shortly after proliferating myoblasts exit the cell cycle [2] and begin to express differentiation-specific transcription factors [3]. In mouse embryos, *MCK* expression is initiated after the activation of myogenic transcription factors. *MCK* mRNA is first detectable in embryonic day 13 (E13) cardiac and skeletal muscles, and its expression is maintained throughout adulthood [4]. The expression of *MCK* between different anatomical muscle groups is quite variable; for example, MCK protein as well as its enzymatic product, creatine phosphate, are about two or three times higher in fast-twitch muscles than in slow-twitch muscles [5,6]. Fiber type-specific muscle regulatory factors (MRFs) have been studied in several other skeletal muscle genes, such as in *MLC2v*, *MLC1/3f* and *aldolase* genes [7-10] and even more extensively in *slow* and *fast troponin I* genes [11-16]. These studies have provided important clues that implicate a variety of transcriptional control mechanisms in muscle fiber type-specific gene expression. Aspects of these mechanisms are both similar to and different from those that regulate *MCK* expression in fast- and slow-twitch fiber types.

While *MCK* gene expression has been extensively studied [17-22], some of its regulatory regions have yet to be fully characterized. Currently, the 5'-enhancer (-1,256 to -1,050) is the best characterized of the known regions [18,20,23-28]. It has the ability (1) to drive high-level transcription of reporter genes in skeletal and cardiac muscle in both transgenic mice and cell culture and (2) to function with heterologous promoters [29]. Deletion and mutation analyses within this region in cultured skeletal myocytes and

in transgenic mice have defined seven control elements: muscle-specific (CArG) and serum response element promoters, activator protein 2 (AP-2), Six4/5, AT-rich, left and right E-boxes and myocyte enhancer factor 2 (MEF2) [23,24]. The *MCK* proximal promoter (-358 to +1) has also been thoroughly studied. It is active in skeletal and cardiac myocytes in culture and can function independently of the 5'-enhancer. The proximal promoter is also active in transgenic skeletal muscle, and the combination of both the 5'-enhancer and the proximal promoter exhibits significant synergy in both cell culture and transgenic mice. The proximal promoter contains at least four active transcription factor binding sites: p53, E-box, CArG, and MPEX, a recently discovered sequence that recruits both Myc-associated zinc finger protein (MAZ) and Krupple-like factor 3 (KLF3) [30-33]

Studies involving the systemic delivery of expression constructs via adeno-associated vector type 6 vectors and transgenic mice have demonstrated that the *MCK* 5'-enhancer and proximal promoter confer transcriptional activity several orders of magnitude higher in muscles containing primarily fast-twitch fibers, such as the tibialis anterior (TA) and quadriceps, than in muscles containing slow-twitch fibers, such as the diaphragm and soleus [22,34,35]. In contrast, the ratio of endogenous MCK protein levels in fast- to slow-twitch skeletal muscles is only about 2:1 [5,6,36]. The discrepancy between gene construct expression levels and endogenous MCK levels suggests that *MCK* gene transcription in slow-twitch fiber types is partially governed by regulatory elements located elsewhere in the *MCK* locus. This hypothesis is supported by previous transgenic tests of an approximately 6.5-kb mouse *MCK* gene region (-3,349 to +3,236) that was used to express dystrophin in *mdx* mice [37]. While fiber-type expression ratios were not included in these studies, the detection of dystrophin in all fibers implied that one or more subregions within the -3,349 to +3,236 sequence in addition to the 5'-

enhancer and proximal promoter play major roles in *MCK* expression in slow- and intermediate-twitch muscle fibers.

The *MCK* gene locus also contains a less well-characterized 1-kb control region called modulatory region 1 (MR1), which resides within the +740 to +1,721 portion of the gene's first intron. In previous and very preliminary studies, MR1 was shown to promote muscle-specific transcription in skeletal myocyte cultures and in transgenic skeletal muscle [19,22,38]. We began the present study by comparing MR1 sequences among six mammalian species and discovered that MR1 is highly conserved throughout its sequence. Most of the conserved motifs are not sequences known to bind muscle gene transcription factors, but a 95-bp subregion within MR1, the *MCK* small intronic enhancer (*MCK*-SIE), was shown to contain conserved and functional E-box and MEF2 control elements, and chromatin immunoprecipitation (ChIP) assays and ChIP-Seq analyses demonstrate that the *MCK*-SIE's E-box and MEF2 elements interact with MyoD/myogenin and MEF2, respectively. The *MCK*-SIE exhibits much higher transcriptional activity than the entire MR1 in differentiated skeletal muscle cultures, and the SIE's elevated activity is due to removing it from the repressive effects of highly conserved regions flanking the *MCK*-SIE's 5'- and 3'-borders.

Upon discovering the enhancer-like properties of the *MCK*-SIE, and recalling that *MCK* transgenes containing only the 5'-enhancer and proximal promoter regions express relatively poorly in slow- and intermediate-twitch fibers, we hypothesized that expression of *MCK* in these fiber types may require the *MCK*-SIE-containing MR1 region. We therefore generated transgenic mouse lines that carry the 6.5-kb *MCK* regulatory region with or without MR1. Comparison of transgene fiber-type expression patterns between these lines supports our hypothesis. Interestingly, while E-box and MEF2 elements are common to other important regulatory regions in the *MCK*-SIE and the rat slow upstream

regulatory element (SURE) region in *slow troponin I*, the key DNA control elements that ensure slow-twitch muscle fiber expression in the SURE region [11,13,14,39], are not present in the *MCK*-SIE (see Discussion).

S1.3 Results

S1.3.1 Sequence analysis of the intron 1 modulatory region MR1 reveals multiple highly conserved sequence motifs

To begin our characterization of mouse MR1 and its role in *MCK* gene expression, a 1,081-bp region (+740 to +1,721) was aligned to the MR1 regions of five other mammalian species (human, cat, dog, bovine and pig) to reveal the presence of potentially functional control elements (Figure 1 and Additional file 1, Figure S1). This comparison revealed several MR1 subregions containing many highly conserved sequence motifs, which were then compared to a transcription factor binding motif library deposited in the TRANSFAC database [40]. Of particular interest was a 95-bp region (+901 to +995) that was subsequently proven to exhibit the properties of a transcriptional enhancer (Figure 1). The *MCK*-SIE exhibits high sequence conservation and contains four motifs known to control the transcription of many muscle genes: two core E-boxes (CAnnTG) [41,42], a MEF2 site and an overlapping MAF half-site and AP-1 site (Figure 1). Among six mammalian species, 11 to 12 bp of the more 5'-E-boxes conform to the 14-bp MyoD/myogenin consensus binding site: [C/G]N[A/G]$_2$ CA[C/G]$_2$ TG[C/T]$_2$ N[C/G] [17] and 10 to 12 bp of the more 3'-E-boxes conform to the consensus binding sequence. Since the dog and mouse E-box sequences are located further 5' than in the other species (Figure 1), and since the distance between the 5'-E-box and MEF2 site varies from 16 to 40 bp, the precise distances between the four *MCK*-SIE control elements may not be functionally important. The MEF2 motif in all six species conforms fully to the MEF2 consensus sequence ([G/T][C/T]TA[A/T]$_3$ ATA[A/G][A/C/T]) [43]. In

addition, a region located near the 5'-E-box contains partially overlapping sequences that match perfectly with proven MAF and AP-1 binding sites [44]. The clustering of these motifs seems significant, since the combination of a paired E-box and MEF2/AT-rich motif has been observed in many muscle promoters, including the *MCK* 5'-enhancer [45, 46]

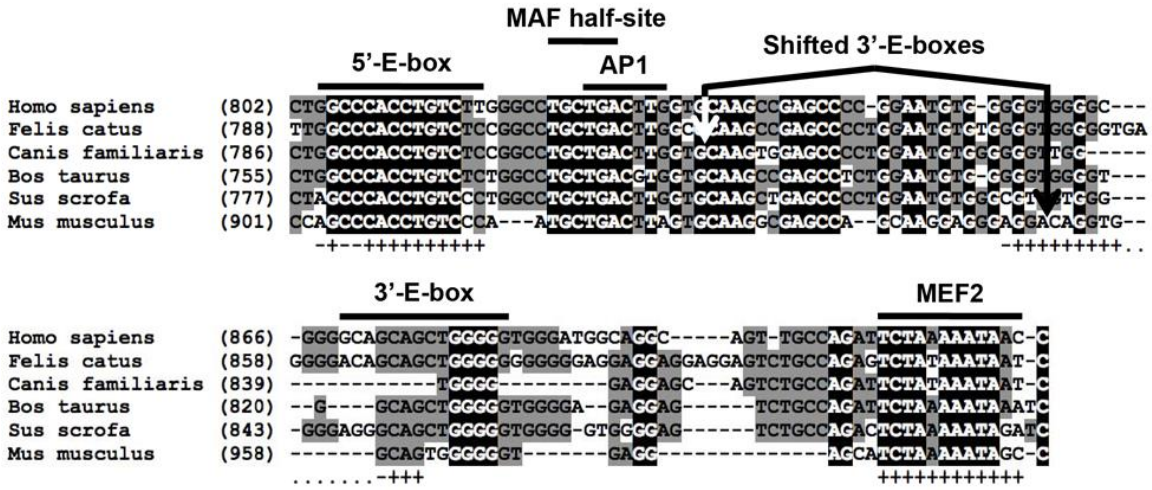


Figure 1. ***Modulatory region 1 (MR1) contains a highly conserved subregion containing known myogenic control element motifs.*** *Sequence alignment of MR1 reveals a highly conserved 95-bp subregion, muscle creatine kinase (MCK) small intronic enhancer (MCK-SIE), that contains five putative control elements: an E-box motif pair, a myocyte enhancer factor 2 (MEF2) consensus motif and partially overlapping sequences that match proven MAF half-site and activator protein 1 (AP-1) sequences (see also Additional file 1 Figure S1). Bases that are identical in all six species (Homo sapiens, Felis catus, Bos taurus, Sus scrota, Canis familiaris and Mus musculus) are shown in black, while bases conserved between at least three species are shown in gray. The 3'-E-box is present in all six species, but is slightly more 5' in the mouse and further 5' in the dog. Conformation of mouse control element sequences to the MyoD/myogenin and MEF2 consensus sequences are indicated below the mouse sequence (+ = conforms, - = differs).*

Figure S1, 1

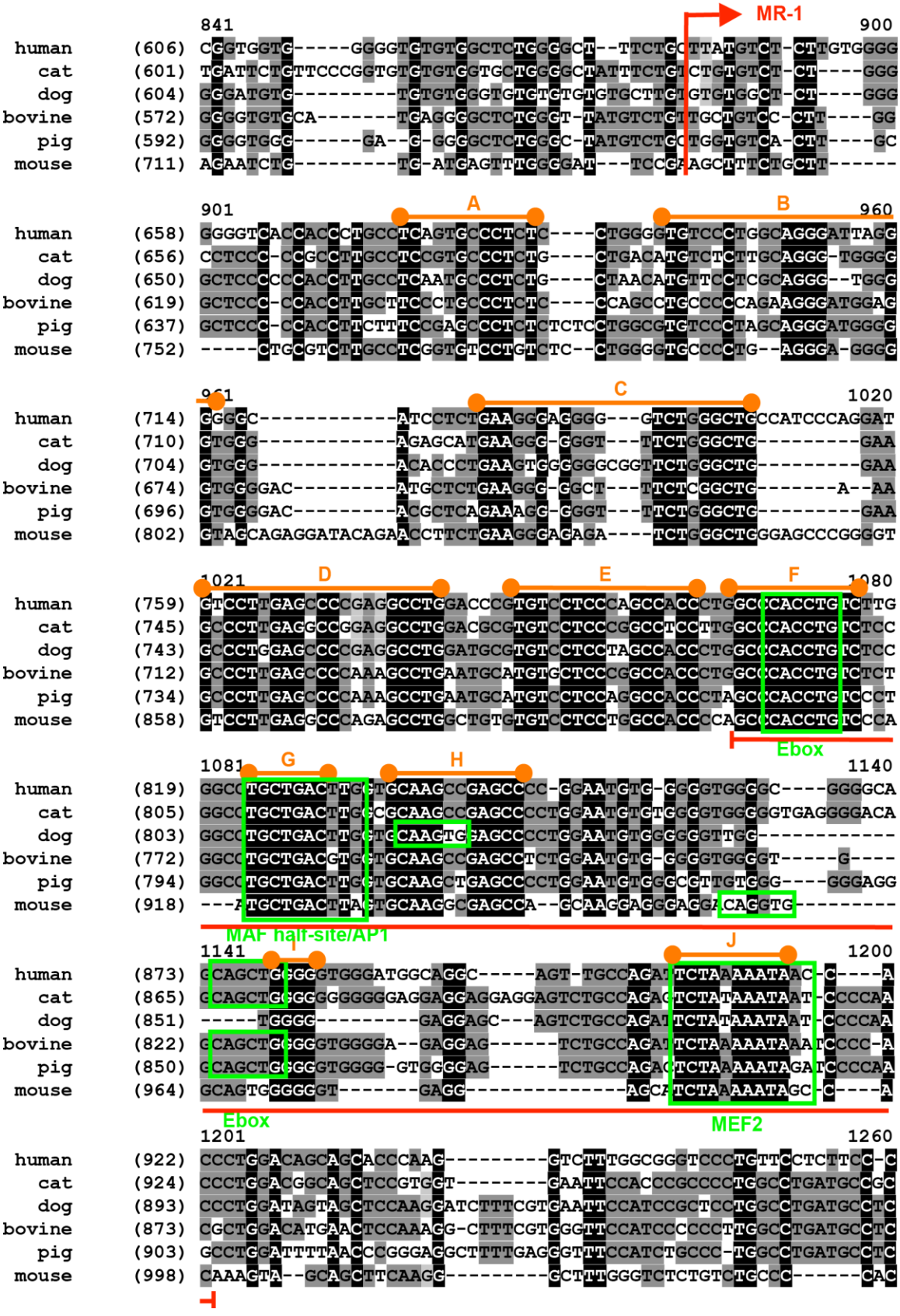

MR-1
A
B
C
D
E
F
G
H
J
Ebox
MAF half-site/AP1
Ebox
MEF2
human
cat
dog
bovine
pig
mouse

Figure S1, 2

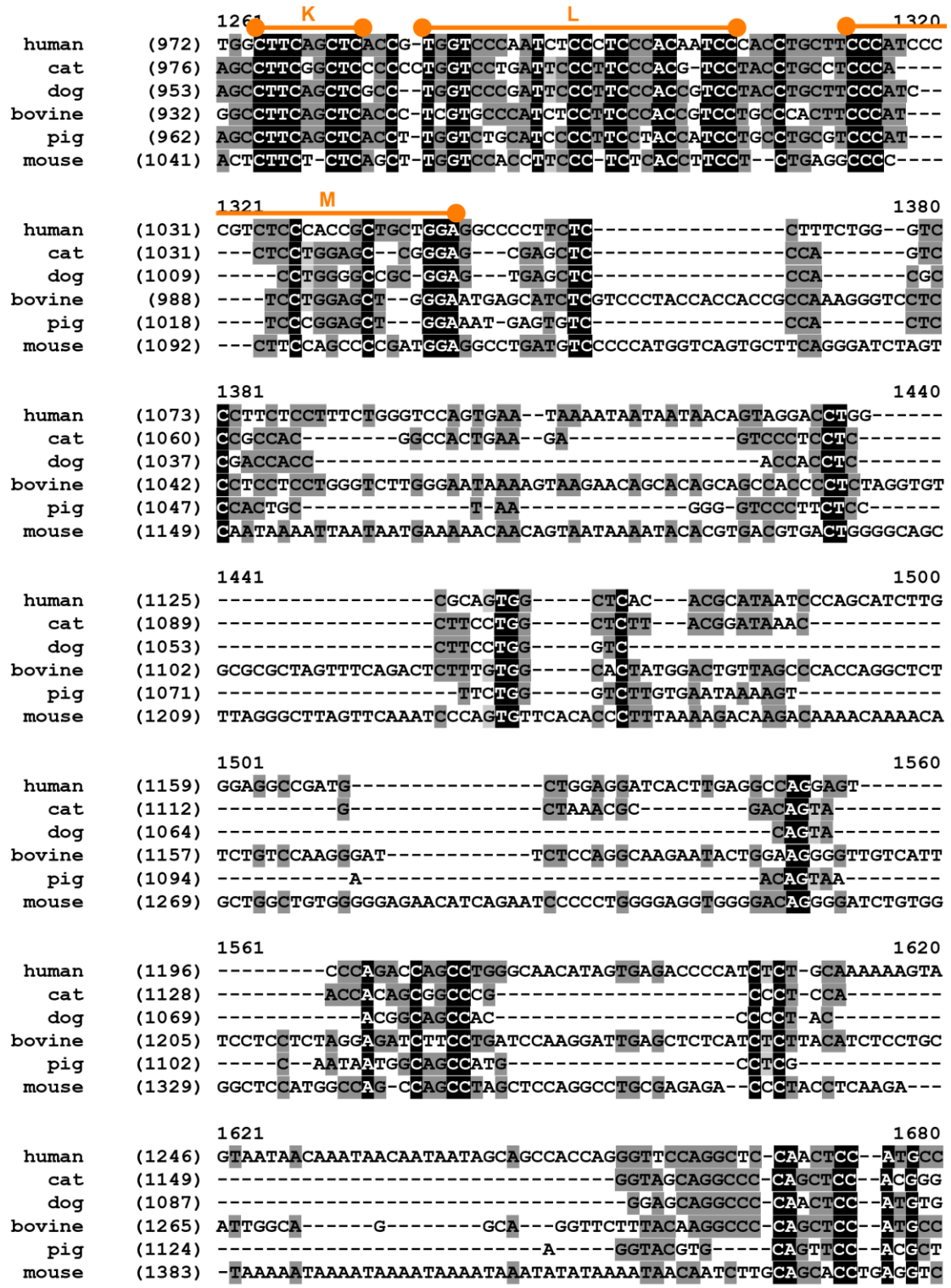

K
L
M
1261
1320
1321
1380
1381
1440
1441
1500
1501
1560
1561
1620
1621
1680
human (972) TGGCTTCAGCTCACCG-TGGTCCCAATCTCCCTCCCACAATCCCACCTGCTTCCCATCCC
cat (976) AGCCTTCGGCTCCCCCCTGGTCCTGATTCCCTTCCCACG-TCCTACCTGCCTCCCA----
dog (953) AGCCTTCAGCTCGCC--TGGTCCCGATTCCCTTCCCACCGTCCTACCTGCTTCCCATC--
bovine (932) GGCCTTCAGCTCACCC-TCGTGCCCATCTCCTTCCCACCGTCCTGCCCACTTCCCAT---
pig (962) AGCCTTCAGCTCACCT-TGGTCTGCATCCCCTTCCTACCATCCTGCCTGCGTCCCAT---
mouse (1041) ACTCTTCT-CTCAGCT-TGGTCCACCTTCCC-TCTCACCTTCCT--CTGAGGCCCC----
human (1031) CGTCTCCCACCGCTGCTGGAGGCCCCTTCTC----------CTTTCTGG--GTC
cat (1031) ---CTCCTGGAGC---CGGGAG---CGAGCTC----------CCA-------GTC
dog (1009) -----CCTGGGGCCGC-GGAG---TGAGCTC----------CCA-------CGC
bovine (988) ----TCCTGGAGCT--GGGAATGAGCATCTCGTCCCTACCACCACCGCCAAAGGGTCCTC
pig (1018) ----TCCCGGAGCT---GGAAAT-GAGTGTC----------CCA-------CTC
mouse (1092) ---CTTCCAGCCCCGATGGAGGCCTGATGTCCCCCATGGTCAGTGCTTCAGGGATCTAGT
human (1073) CCTTCTCCTTTCTGGGTCCAGTGAA--TAAAATAATAATAACAGTAGGACCTGG------
cat (1060) CCGCCAC--------GGCCACTGAA--GA----------GTCCCTCCTC--------
dog (1037) CGACCACC-----------------------------------ACCACCTC--------
bovine (1042) CCTCCTCCTGGGTCTTGGGAATAAAAGTAAGAACAGCACAGCAGCCACCCCTCTAGGTGT
pig (1047) CCACTGC----------T-AA----------GGG-GTCCCTTCTCC------
mouse (1149) CAATAAAATTAATAATGAAAAACAACAGTAATAAAATACACGTGACGTGACTGGGGCAGC
human (1125) ----------CGCAGTGG-----CTCAC---ACGCATAATCCCAGCATCTTG
cat (1089) ----------CTTCCTGG-----CTCTT---ACGGATAAAC----------
dog (1053) ----------CTTCCTGG-----GTC----------
bovine (1102) GCGCGCTAGTTTCAGACTCTTTGTGG-----CACTATGGACTGTTAGCCCACCAGGCTCT
pig (1071) ----------TTCTGG-----GTCTTGTGAATAAAAGT----------
mouse (1209) TTAGGGCTTAGTTCAAATCCCAGTGTTCACACCCTTTAAAAGACAAGACAAAACAAAACA
human (1159) GGAGGCCGATG----------CTGGAGGATCACTTGAGGCCAGGAGT-------
cat (1112) ----------G----------CTAAACGC---------GACAGTA---------
dog (1064) ----------CAGTA---------
bovine (1157) TCTGTCCAAGGGAT----------TCTCCAGGCAAGAATACTGGAAGGGGTTGTCATT
pig (1094) ----------A----------ACAGTAA--------
mouse (1269) GCTGGCTGTGGGGGAGAACATCAGAATCCCCCTGGGGAGGTGGGGACAGGGGATCTGTGG
human (1196) ---------CCCAGACCAGCCTGGGCAACATAGTGAGACCCCATCTCT-GCAAAAAAGTA
cat (1128) ---------ACCACAGCGGCCCG----------CCCT-CCA--------
dog (1069) ----------ACGGCAGCCAC----------CCCCT-AC---------
bovine (1205) TCCTCCTCTAGGAGATCTTCCTGATCCAAGGATTGAGCTCTCATCTCTTACATCTCCTGC
pig (1102) -----C--AATAATGGCAGCCATG----------CCTCG----------
mouse (1329) GGCTCCATGGCCAG-CCAGCCTAGCTCCAGGCCTGCGAGAGA--CCCTACCTCAAGA---
human (1246) GTAATAACAAATAACAATAATAGCAGCCACCAGGGTTCCAGGCTC-CAACTCC--ATGCC
cat (1149) ----------GGTAGCAGGCCC-CAGCTCC--ACGGG
dog (1087) ----------GGAGCAGGCCC-CAACTCC--ATGTG
bovine (1265) ATTGGCA------G--------GCA---GGTTCTTTACAAGGCCC-CAGCTCC--ATGCC
pig (1124) ----------A-----GGTACGTG-----CAGTTCC--ACGCT
mouse (1383) -TAAAAATAAAATAAAATAAAATAAATATATAAAATAACAATCTTGCAGCACCTGAGGTC

Figure S1, 3

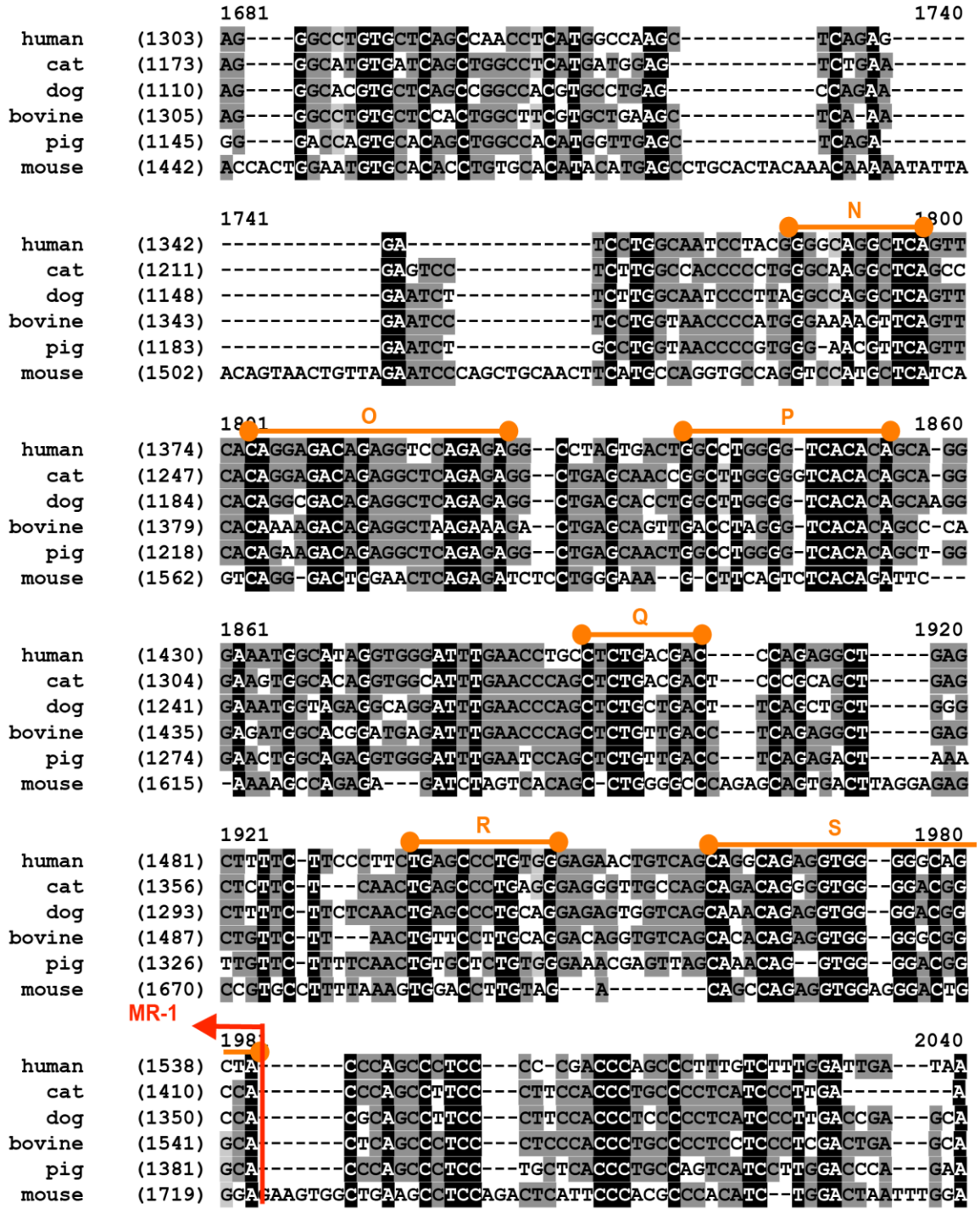


Figure S1. A six-species sequence alignment of modulatory region 1 (MR1), which demonstrates the conserved nineteen subregions throughout the region*. The MR1 sequences of six mammalian species (human, cat, dog, bovine, pig and mouse) were aligned to reveal sequence conservation. Bases that are fully conserved between the six*

species are highlighted in black, while those conserved in three to five species are highlighted in gray. Gaps in the sequence alignment are represented as hyphens. The 5' and 3' flanks of MR1, as defined in this study, are marked with red right-angled arrows. Nineteen conserved subregions (A-S, annotated by orange barbed lines) were tested for transcriptional activity (see Additional file 2, Figure S2). The two E-box elements, the MAF/activator protein 1 (AP-1) site and the myocyte enhancer factor 2 (MEF2) consensus sequence investigated in this study are outlined in green. The 1,081-bp MR1 region (+740 to +1,721) extends slightly more 5' and 3' than the originally described mouse MR1 sequence (+748 to -1,607) [29].

S1.3.2 MR1 is required for high-level MCK gene expression in differentiated skeletal muscle cells, and it contains a highly active SIE

To address the function of MR1 in *MCK* gene expression, the MR1 region was deleted from the entire 6.5-kb *MCK* sequence (Figure 2A, constructs 1 and 2 [6.5*MCK*CAT and 6.5*MCK*ΔMR1-CAT]), and the effect of the deletion was examined in differentiated skeletal myocytes (MM14). To gauge the relative change in transcriptional activity caused by the loss of MR1, we compared 6.5*MCK*ΔMR1-CAT to a construct that contains a deletion of the well-characterized *MCK* 5'-enhancer (Figure 2A, construct 4 [6.5*MCK*ΔEnh-CAT]). Expression from each test plasmid was normalized to the activity of a muscle-specific *MCK* enhancer-driven alkaline phosphatase (AP) reference construct.

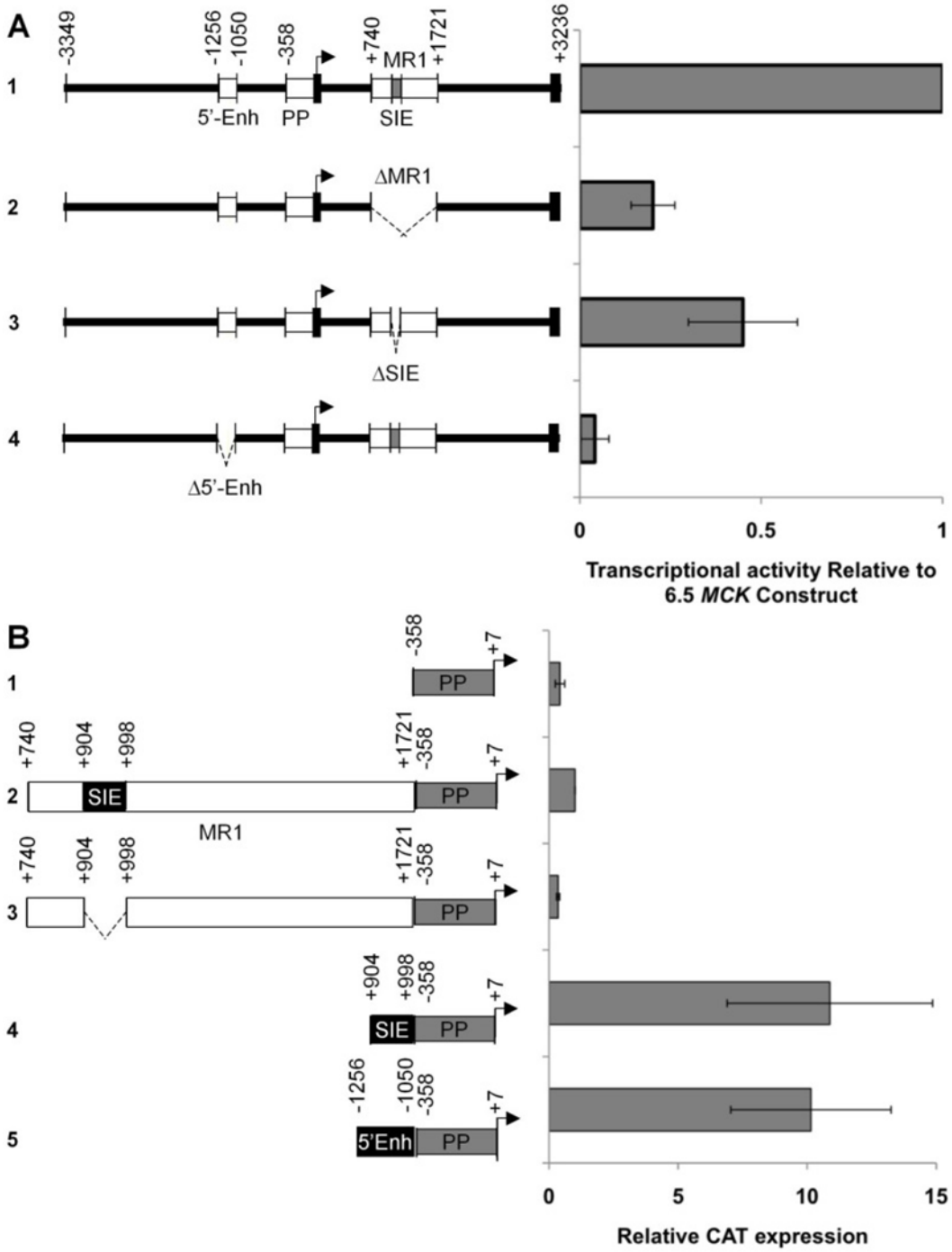


Figure 2. ***MR1 is a positive regulator of*** **MCK** ***transcription****. (A) MM14 skeletal myocytes were cotransfected with an* MCK *enhancer-alkaline phosphatase (AP) reference plasmid and test gene plasmids containing the chloramphenicol acetyl transferase (CAT) reporter gene driven by the full-length 6.5-kb* MCK *construct (6.5*MCK*-CAT, #1), the 6.5-kb construct with MR1 deleted (6.5*MCKΔ*MR1-CAT, #2), the*

6.5-kb construct with the MCK-*SIE deleted (6.5*MCKΔ*SIE-CAT, #3) or, for comparison, the 6.5-kb construct with the 5'-enhancer deleted (6.5*MCKΔ*Enh-CAT, #4). Test construct activities are represented as the average values of relative CAT over AP activity normalized to the activity of 6.5*MCK-*CAT. (B) MR1 is composed of regions that promote transcription in MM14 cultures. Constructs containing the "full-length" MR1 (MR1-PP-CAT, #2), a construct lacking the* MCK-*SIE (MR1*Δ*SIE-PP-CAT, #3) or just the* MCK-*SIE (SIE-PP-CAT, #4) were generated to test the functional activity of the* MCK-*SIE. Activities of these test constructs were normalized to activities of the proximal promoter alone (PP-CAT, #1). The activity of the 5'-enhancer (5'Enh-PP-CAT, #5) is provided for comparison. Each experiment was performed in at least twelve plates in three separate experiments, and activities are averages of those experiments. Error bars represent ±1 standard deviation.*

Deletion of MR1 results in an approximately fivefold lower transcriptional activity in differentiated MM14 cultures than that produced by the entire 6.5-kb *MCK* gene construct ($P < 0.01$) (Figure 2A, constructs 1 and 2), whereas deletion of the *MCK* gene 5'-enhancer results in a greater than 10-fold decrease ($P < 0.01$).

To determine whether the *MCK*-SIE is critical for *MCK* gene transcription, it was deleted from the 6.5*MCK*-CAT construct and the resulting 6.5*MCK*ΔSIE-CAT was tested in differentiated skeletal muscle cultures (Figure 2A, construct 3). The deleted construct exhibited a 60% decrease in transcriptional activity in skeletal myocytes ($P < 0.01$), demonstrating that, in the context of the 6.5-kb *MCK* genomic sequence, the *MCK*-SIE is likely responsible for much of the positive transcriptional activity of MR1.

S1.3.3 MCK-SIE is active in differentiated skeletal muscle cells when placed 5' of the MCK proximal promoter

To facilitate further analysis of MR1 regulatory functions, subsequent studies were carried out in the context of MR1 placed 5' of the highly conserved *MCK* proximal

promoter (Figure 2B (MR1-proximal promoter-chloramphenicol acetyl transferase (MR1-PP-CAT)), construct 2). This test construct frees MR1 from transcriptional effects of the highly active *MCK* 5'-enhancer, which could lead to dampened effects of mutations or deletions within MR1. Importantly, it also avoids potential confounding effects due to cotranscriptional or posttranscriptional events, such as altered splicing efficiency or altered elongation efficiency, which could occur in conjunction with testing MR1 function within its 3' intron 1 location in the native *MCK* gene. In agreement with the decreased activity observed when MR1 is deleted from the 6.5-kb sequence (Figure 2A), MR1-PP-CAT exhibits transcriptional activity in skeletal myocyte cultures that is approximately threefold greater than that of the proximal promoter alone (Figure 2B, compare constructs 1 and 2). MR1's positive activity when moved 5' of the transcription start site also indicates that it has the properties of an enhancer.

Since the *MCK*-SIE had the greatest potential for explaining the positive activity of MR1 (Figure 2A), we tested its capacity to act as an enhancer independent of other MR1 sequences. Deletion of the *MCK*-SIE from MR1 reduces transcriptional activity to a level similar to that of the proximal promoter alone (Figure 2B, construct 3). Conversely, when the *MCK*-SIE was placed directly upstream of the proximal promoter (Figure 2B, M*CK*-SIE-PP-CAT, construct 4), a greater than 10-fold increase in transcription ($P < 0.01$) relative to the MR1-PP-CAT construct was observed. In fact, the *MCK*-SIE synergizes with the proximal promoter, as does the 5'-enhancer (Figure 2B, 5'Enh-PP-CAT, construct 5).

S1.3.4 Two E-box motifs and a MEF2 site are required for full transcriptional activity of the MCK-SIE in skeletal myocytes

To determine the transcriptional activity of the *MCK*-SIE conserved binding site motifs, the 5'- and 3'-E-boxes and MEF2 motifs were subjected to both deletion and

mutation analyses (Figure 3A) in the context of the *MCK*-SIE-PP-CAT construct (Figure 2B, construct 4). In skeletal myocytes, deletion or mutations of the 5'-E-box resulted in approximately 30% reductions in transcriptional activity, whereas deletion or mutations of the 3'-E-box resulted in approximately 65% reductions (Figure 3B), and deletion of both E-boxes caused a nearly 90% decrease in transcriptional activity. Deletion or mutations of the single MEF2 consensus motif also caused an approximately 90% reduction in transcriptional activity (Figure 3B). These data imply that both E-boxes contribute to the *MCK*-SIE's transcriptional activity, but that the 3'-E-box provides most of the activity. Since mutation of the MEF2 site leads to about the same loss in activity as mutation of both E-boxes, and since E-box binding factors are known to synergize with MEF2, it may be that the bulk of the *MCK*-SIE's transcription activity is derived from a single highly active MEF2-MyoD/myogenin complex.

A

	5'-E-box	3'-E-box
Consensus	[C/G]N[A/G] $_2$CA[C/G] $_2$TG[C/T] $_2$N[C/G]	
Mus Wt	ccAGccCACCTG TCCCaa	ggAGGACAGGTG GCAGtg
Del	******------******	******------******
M1	******TG**CA******	******TG**CA******
M2	******GT**AC******	******GT**AC******

	MAF half-site	AP1	MEF2
Consensus	★TGCTGA	★TGACTTA	[G/T][C/T]TA[A/T] $_3$ATA[A/G][A/C/T]
Mus Wt	AATGCTGACTTA GT		CATCTAAAAATAGCCA
Del	**--------*****		**----------****
M1	**CATCAGT*****		****CG****CG****
M2	**ACGACTG*****		****AT****AT****

B

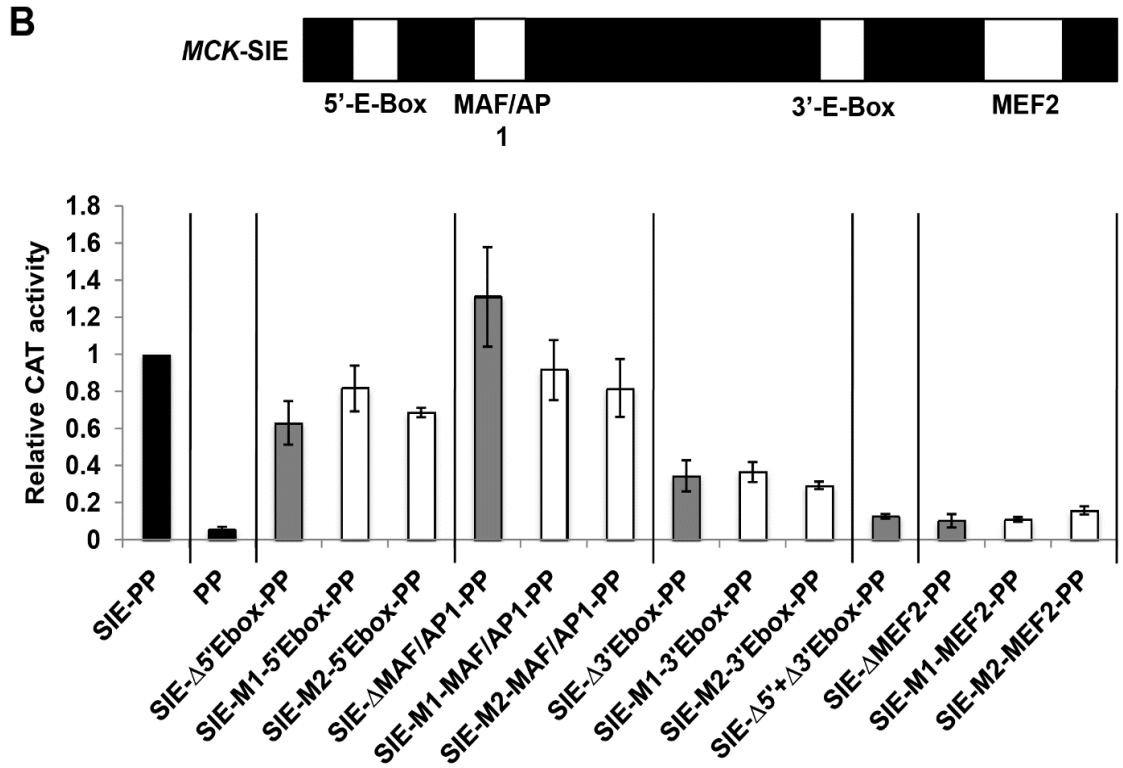


Figure 3. ***Two E-boxes and a MEF2 site are critical for activity of the* MCK-*SIE****.*** *(A) Deletions and mutations tested in* MCK-*SIE. The currently accepted consensus motifs for the E-box and MEF2 motifs are shown. Proven MAF half-site and AP-1 control element sequences are also indicated. Stars indicate sequences that were experimentally proven to recruit the labeled factors and do not represent consensus binding motifs. The wild-type mouse sequences of these elements within the* MCK-*SIE (Wt), the deletion sequences (Del) and two mutation sequences (M1 and M2) used in this study are shown on successive lines. Base pair deletions are indicated as hyphens, point mutations are shown as changed bases and asterisks indicate unchanged bases. (B) Mutational analysis of control elements within the* MCK-*SIE. The E-box, MAF/AP-1 and MEF2 motifs in the* MCK-*proximal promoter-CAT (*MCK-*SIE-PP-CAT) (diagrammed with elements in their relative positions) were deleted (gray bars) or subjected to two mutations (white bars) within core bases (Figure 2A) and were tested for transcriptional activity in differentiated MM14 skeletal myocyte cultures. The relative activities of these constructs were compared to the* MCK-*SIE-PP-CAT construct (scaled to equal 1.0) and PP-CAT alone (black bars). Each construct was tested in twelve plates in three separate experiments, and activities shown are averages of those experiments. Error bars represent ±1 standard deviation.*

The possibility that other control elements may reside in the *MCK*-SIE is raised by the highly conserved TGCTGAC[T/g]T[G/a]G sequence that begins several base pairs 3' of the 5'-E-box (Figure 1). The TGCTGA portion is a perfect match to MAF half-sites [47,48], and the TGACTTA sequence in the mouse *MCK*-SIE is a perfect match to a fully functional noncanonical AP-1 site [49,50]. Deletion and mutations that should have abolished the binding of either MAF or AP-1 (Figure 3A) had little to no effect on transcriptional activity (Figure 3B). This does not negate the possibility that MAF and/or AP-1 interactions within the *MCK*-SIE region play a role in *MCK* gene expression *in vivo*, but such interactions are not important for the *MCK*-SIE's transcriptional activity in differentiating skeletal myocyte cultures.

S1.3.5 Both MyoD and myogenin bind to the MCK-SIE in differentiated skeletal myocytes

On the basis of the rapid onset of *MCK* expression during differentiation, the transcriptional activity of MR1 in myocyte cultures (Figure 2B) and the presence of two active E-box elements within this region (Figure 3B), it seemed likely that MyoD and/or myogenin associate with the *MCK*-SIE. ChIP analysis of differentiating skeletal myocyte cultures was thus employed to determine whether the E-box pair recruits MyoD, myogenin or both MRFs *in vivo*.

One caveat of ChIP data interpretation is that control elements cannot be distinguished with respect to transcription factor binding when they bind the same factors and are close enough that both sites will be present on many of the same randomly sheared chromatin fragments. This would certainly be the case for the *MCK*-SIE E-box pair, where the separation is only 46 bp. Thus, primers that flank the entire *MCK*-SIE were used to detect MyoD- and myogenin-immunoprecipitated chromatin. This issue is also pertinent to ChIP discrimination between occupancy of the *MCK*-SIE E-box pair and other *MCK* E-boxes with proven transcriptional activity. These are centered at -1,175 and +1,152 within the *MCK* 5'-enhancer and at -246 within the proximal promoter [26]. Therefore, in addition to using primers that amplify the *MCK*-SIE, primers for the 5'-enhancer were used as a positive control, since this region is known to contain two functional E-boxes that bind MyoD and myogenin [17,51,52].

Three "negative" control primers were used to rule out the possibility of cross-enrichment from factors binding to non-*MCK*-SIE regions (Figure 4A). The first "negative" control primer set amplifies intron 1 of the MAP/microtubule affinity-regulating kinase 4 (*Mark4*) gene. This sequence is roughly 40-kb 3' of the *MCK*-SIE on mouse chromosome 19 and is within a 1-kb region that entirely lacks the core E-box binding

motif CAnnTG; thus it should serve as a truly negative control for MyoD and myogenin occupancy of the *MCK*-SIE. The second "negative" control primer pair spans the exon 1/intron 1 boundary and amplifies a 217-bp region located 690 bp upstream of the *MCK*-SIE, 242 bp downstream of the active promoter E-box and 1,149 bp downstream of the active *MCK* 5'-enhancer right E-box (Figure 4A). The mouse exon 1/intron 1 boundary region contains two nonconserved E-boxes and also has four nonconserved E-boxes located 52, 67, 97 and 310 bp downstream of its 3'-border. None of these E-boxes have been tested for transcriptional activity, but they are likely to be transcriptionally inactive as they are not conserved in other mammals. Nevertheless, this would not preclude their occupancy by MyoD/myogenin or their function in mouse muscle cells; thus examining this subregion was also of interest in itself. The third "negative" control primer pair spans a 209-bp region starting at exon 2 (Figure 4A). It contains one nonconserved E-box and two other nonconserved E-boxes which are located 36 bp and 638 bp upstream of its 5'-border. MyoD/myogenin binding to any of these exon 2 E-boxes would thus cause an enrichment that would be detected by the exon 2 primer pair. Conversely, if MyoD and/or myogenin occupy the *MCK*-SIE, and if the negative control regions are not occupied, enrichments of the *MCK*-SIE and of the *MCK* 5'-enhancer (positive control) should be significantly greater than those at any of the negative control regions.

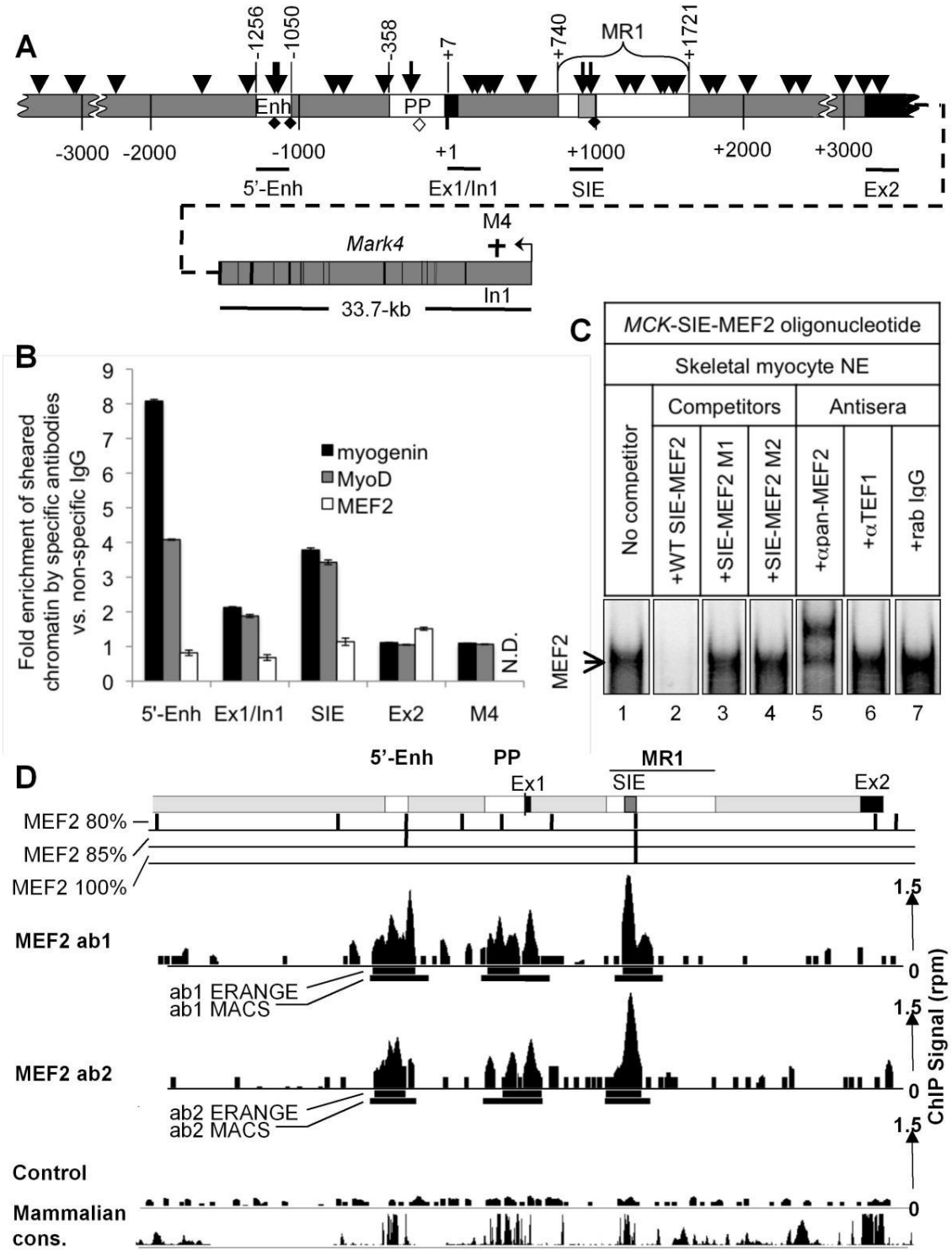


Figure 4. ***MyoD and myogenin are enriched at the*** **MCK-*SIE* in skeletal myocytes**. *(A) Diagram of the 6.5-kb* MCK *regulatory region with the three known active regulatory regions: the 5'-enhancer, PP, MR1 (white boxes), the* MCK-*SIE (light gray box) exons 1*

and 2 (black boxes) and other regions (gray), including the 33.7-kb Mark4 *gene (located approximately 40 kb 3' of the* MCK-*SIE and transcribed in the opposite direction). E-box CAnnTG core motifs (arrowheads) occur throughout the 6.5-kb sequence. Among the thirty-five total E-boxes are two functional E-boxes within the 5'-enhancer, one functional E-box within the proximal promoter and two E-box motifs within the* MCK-*SIE (longer arrows). The less frequent MEF2 motifs (full diamonds) are found only in the 5'-enhancer and MCK-SIE and as a possible nonconsensus MEF2 site (open diamond) in the proximal promoter. The chromatin immunoprecipitation (ChIP) primer pairs (black lines) that span the 5'-enhancer sequence were used as positive controls for MyoD and myogenin binding to functional E-boxes. Negative controls consist of genomic regions containing either no core E-box motifs (region within the Mark4 intron 1 (M4, dagger)) or core E-box motifs with no proven transcriptional function (*MCK *gene exon 1/intron 1 boundary (two E-boxes) and exon 2 (one E-box); see Results, section-5). (B) MyoD and myogenin bind* MCK *gene E-box motifs. ChIP analyses using antibodies for MyoD, myogenin, MEF2 and control immunoglobulin G (IgG) were performed using chromatin from differentiated MM14 cell myocytes. The graph shows data from one of three ChIP experiments that is representative of the enrichment detected at each position by antibodies to myogenin (black bars), MyoD (gray bars) or MEF2 (white bars) over nonspecific rabbit IgG as determined by quantitative polymerase chain reaction (qPCR) assay. Error bars represent ±1 standard deviation of triplicate samples. (C) Electrophoretic mobility shift assay (EMSA) of MEF2 binding to the* MCK-*SIE MEF2 control element. Nuclear extracts from differentiated MM14 cultures were incubated with a* ^{32}*P-labeled probe containing the* MCK-*SIE-MEF2 sequence with no competitor (lane 1), wild-type MEF2 competitor (lane 2), two different mutant MEF2 competitors (lanes 3 and 4), pan-MEF2 antibodies (lane 5), transcriptional enhancer factor 1 (TEF-1)-specific antibodies (lane 6) or nonspecific rabbit IgG (lane 7). Arrows indicate the MEF2-containing complex and free probe. (D) MEF2 ChIP-Seq occupancy at the 6.5-kb* MCK *regulatory region in differentiated* C_2C_{12} *cells shows that MEF2 is present at all three control regions. The 6.5-kb region is shown in schematic at the top (5'-enhancer, proximal promoter and MR1 are shown in white;* MCK-*SIE is shown in gray). Sequences that match the MEF2 canonical motif (CTAWWWWTAG) at the 80%, 85% and 100% thresholds are mapped throughout the 6.5-kb region. The sequenced and mapped ChIP signals (reads per million (rpm)) for the two pan-MEF2 antibodies 1 and 2 and the*

control (input DNA) are indicated as black histograms (scale shown at the right). Two different ChIP-Seq region finders (Model-based Analysis of ChIP-Seq data and Enhanced Read Analysis of Gene Expression) define the sequence range in which MEF2 is predicted to bind (see Materials and methods), and these are shown below each signal track as black bars. Conservation across the regions is shown from the University of California Santa Cruz (UCSC) Genome Browser plot of phastCons scores for the 20 default placental mammals.

Accordingly, ChIP analysis showed that antibodies for both MyoD and myogenin enriched the 5'-enhancer several-fold over nonspecific immunoglobulin G (IgG) (Figure 4B), and both antibodies also enriched the *MCK*-SIE region. In contrast, neither antibody enriched the exon 2 and *Mark4* genomic regions significantly above nonspecific IgG. This demonstrates that MyoD and myogenin bind neither to nonconserved, and presumably nonfunctional, E-box motifs in the regions surrounding the *MCK*-SIE, nor to chromatin regions that lack E-boxes. There is a slight enrichment at the exon 1/intron 1 boundary. However, this could be caused by cross-enrichment due to MyoD and myogenin occupancy of the nearby and functional proximal promoter E-box [26], the 5'-enhancer, the *MCK*-SIE or any combination of these regions. Nevertheless, the enrichment due to MyoD and myogenin occupancy of the *MCK*-SIE region is probably not due to spurious enrichment from amplification of longer sheared chromatin fragments that include the 5'-enhancer or proximal promoter, because the enrichment signal from the exon 1/intron 1 region would then be higher than that of the *MCK*-SIE, and it is not. MyoD and myogenin thus occupy proven functional E-boxes in the 5'-enhancer and the *MCK*-SIE in differentiated skeletal myocytes, and they do not appear to occupy E-boxes in regions flanking the *MCK*-SIE. An additional consistent observation in these studies is that myogenin exhibits an approximately twofold higher occupancy of

the 5'-enhancer than MyoD, whereas both MRFs exhibit equivalent occupancy of the *MCK*-SIE.

S1.3.6 MEF2 interaction with the MCK-SIE in vitro and in vivo

As demonstrated in Figure 3B, the MEF2 site contributes strongly to the transcriptional activity of the *MCK*-SIE region. Since members of the MEF2 superfamily of transcription factors (MEF2A, MEF2B, MEF2C and MEF2D) [53] have previously been shown to play important roles in muscle gene transcription, we asked whether any of the MEF2 family members were associated with the *MCK*-SIE *in vivo*. In initial ChIP analysis, several different MEF2 antibodies unexpectedly failed to enrich the *MCK*-SIE or even the 5'-enhancer (Figure 4B) (see Discussion). Furthermore, antibodies to octamer binding protein 1 (Oct-1) and transcriptional enhancer factor 1 (TEF-1), two factors known to transactivate AT-rich motifs in muscle promoters [54,55] and known to be present in myocyte cultures, also failed to precipitate the *MCK*-SIE when used in ChIP assays (data not shown). This led us to question whether MEF2 in our cell culture model was detectable by immunoassays.

To establish that differentiated MM14 cultures contain MEF2 protein, that MEF2 protein is recognized by the pan-MEF2 antibody used in our ChIP study and that MEF2 can indeed bind to the *MCK*-SIE, we analyzed MEF2 binding by electrophoretic mobility shift assay (EMSA). ^{32}P-labeled *MCK*-SIE-MEF2 sequence probes were generated and incubated with MM14 nuclear extracts. Gel electrophoresis with the *MCK*-SIE-MEF2 probe revealed a single intense band, which implied that either a single or multiple factors of similar size were bound to the *MCK*-SIE-MEF2 probe (Figure 4C). Wild-type competitor oligonucleotides completely abolished this band, whereas two oligonucleotides containing different mutations of the *MCK*-SIE-MEF2 motif had no effect. Furthermore, a partial supershift of the band was caused when the probe was

incubated with nuclear extracts in the presence of a pan-MEF2 antibody, suggesting that the band of interest contains MEF2. The partial shift likely occurred because the entire complex might not be fully and stably accessible by the antibody to produce a consistent supershift. These results argue that MEF2 proteins are present in the nuclei of differentiated MM14 muscle cells, that MEF2 is capable of binding to the *MCK*-SIE probe and that MEF2 antibodies, which did not precipitate *MCK*-SIE-enriched sequences in ChIP analysis, were capable of binding MEF2 oligonucleotide complexes in EMSA studies of similarly differentiated muscle cultures.

Since TEF-1 also binds AT-rich motifs in muscle gene promoters and has been shown to bind the *MCK* 5'-enhancer [55], we asked whether TEF-1 binds to the MEF2 sequence in the *MCK*-SIE. Incubation with TEF-1-specific antisera did not supershift or abolish the "MEF2 complex," whereas it did supershift a TEF-1-specific complex (data not shown). A nonspecific IgG also failed to alter the mobility or intensity of the MEF2-specific band (Figure 4C). The absence of detectable MEF2 binding in our ChIP study (Figure 4) is therefore not likely to be due to competitive *in vivo* occupancy of the MEF2 site by TEF-1.

As MEF2 occupancy of the *MCK* 5'-enhancer has been reported in mouse embryos and in the B22 myogenic cell line following Brahma-related gene 1 and MyoD induction [42], it seemed possible that unknown differences between the myogenic states of the different cell culture models might affect the ability to detect MEF2 occupancy in the *MCK* locus. Fortunately, ChIP-Seq analyses aimed toward identifying genome-wide MEF2 binding events in terminally differentiated muscle cells were being performed in parallel studies by the Wold group (personal communication, B. Wold). We therefore collaborated in analyzing the *MCK* locus. Initial ChIP-Seq experiments in C_2 C_{12} skeletal muscle cells also failed to detect significant MEF2 ChIP signals at the *MCK*

locus or at several other MEF2 target loci, thus suggesting that MEF2 might be inefficiently cross-linked to DNA under standard ChIP conditions. Since members of the MADS family of transcription factors, of which MEF2 is a member, often depend significantly on protein-protein interactions with other DNA-bound factors, and since the MyoD family of factors interact with MEF2 through protein-protein interactions [56], we reasoned that chromatin fixation conditions designed to more efficiently stabilize these interactions might improve ChIP detection (see Materials and methods).

Following the modified fixation procedure, a standard sequencing readout from this material revealed distinct MEF2 signals at the *MCK*-SIE and at the 5'-enhancer (Figure 4D). These signals were very similar in biological replicate chromatin samples that had been immuno-enriched by MEF2 antibodies directed against nonoverlapping epitopes (data not shown). Enrichment over background was more than 10-fold at both sites ($P < 2e-13$ for Model-based Analysis of ChIP-Seq data (MACS) and $P < 8e-7$ for Enhanced Read Analysis of Gene Expression (ERANGE)), and no other site in the *MCK* locus was significantly occupied, except for the dispersed signals observed throughout the *MCK* proximal promoter sequence. Enrichment of MEF2 within the proximal promoter, which contains no sequences that match the canonical motif (although one with 80% similarity is present (Figure 4D)), could be due to any of several possibilities (see Discussion). The observed MEF2 ChIP-Seq peaks overlap regions of high-sequence conservation among placental mammals at the 5'-enhancer, the proximal promoter and the *MCK*-SIE regions as determined by phastCons scores, which predict evolutionarily conserved elements using a 30-species vertebrate sequence alignment and phylogenetic tree information (Figure 4D).

S1.3.7 MR1 contributes to MCK gene expression in slow- and intermediate-twitch fiber types in adult mice

Previous investigations of *MCK* gene regulation in transgenic mice have suggested that the 5'-enhancer and the proximal promoter are highly active in anatomical muscles with predominantly fast-twitch fibers (type IIb and type IId (also called type IIx or type IId/x fibers)) such as the TA muscle. Conversely, the activity levels of the 5'-enhancer and the proximal promoter were at least 10-fold lower in muscles from the same transgenic mice that contained a high proportion of slow-twitch muscle fibers (type I) or intermediate-twitch muscle fibers (type IIa) such as soleus [26,27]. Since the endogenous levels of MCK protein in fast vs. slow muscle fibers differ by only about threefold [5], the previous transgenic studies implied that regulatory regions in addition to the 5'-enhancer and proximal promoter are required for full *MCK* expression in slow-twitch fibers. This led us to hypothesize that MR1 may contribute to *MCK* expression in type I and type IIa fiber types. To test this possibility, we generated transgenic mouse lines containing either the 6.5-kb *MCK* genomic region driving the β-galactosidase (β-gal) reporter gene (6.5*MCK*-β-gal) or the same construct lacking MR1 (6.5*MCK*ΔMR1-β-gal). Adult transgenic mice were killed, and TA and soleus muscles were dissected and cryosectioned. Sections were then X-gal-stained to detect β-gal transgene expression. To identify the specific fiber types expressing β-gal, we adopted a method of visualizing the four distinct fiber types on a single sample section by immunofluorescence tagging of myosin heavy chain (MYHC) isotypes as described by Gregorevic *et al*. [57] (see Discussion for rationale of MYHC vs. histochemical fiber typing). Sister sections were thus immunostained with monoclonal antibodies that recognize the MYHC isoforms found in slow-twitch muscle fibers (type I), intermediate-twitch muscle fibers (type IIa) and fast-twitch muscle fibers (type IIb) (Figures 5A and 5B). Type IId fibers were

identified based on the absence of immunostaining with all of the above-mentioned monoclonal antibodies [58]. It should be noted that the distribution of fiber twitch types assessed by MYHC isotype expression within the anatomical muscles examined among different transgenic lines was qualitatively similar (data not shown). Thus introduction of the transgenes themselves did not alter the distribution of fiber twitch types. Whether expression levels of the wild-type 6.5*MCK*-β-gal and 6.5*MCK*DMR1-β-gal transgenes are differentially affected by the metabolic states within individual muscle fiber types remains to be determined.

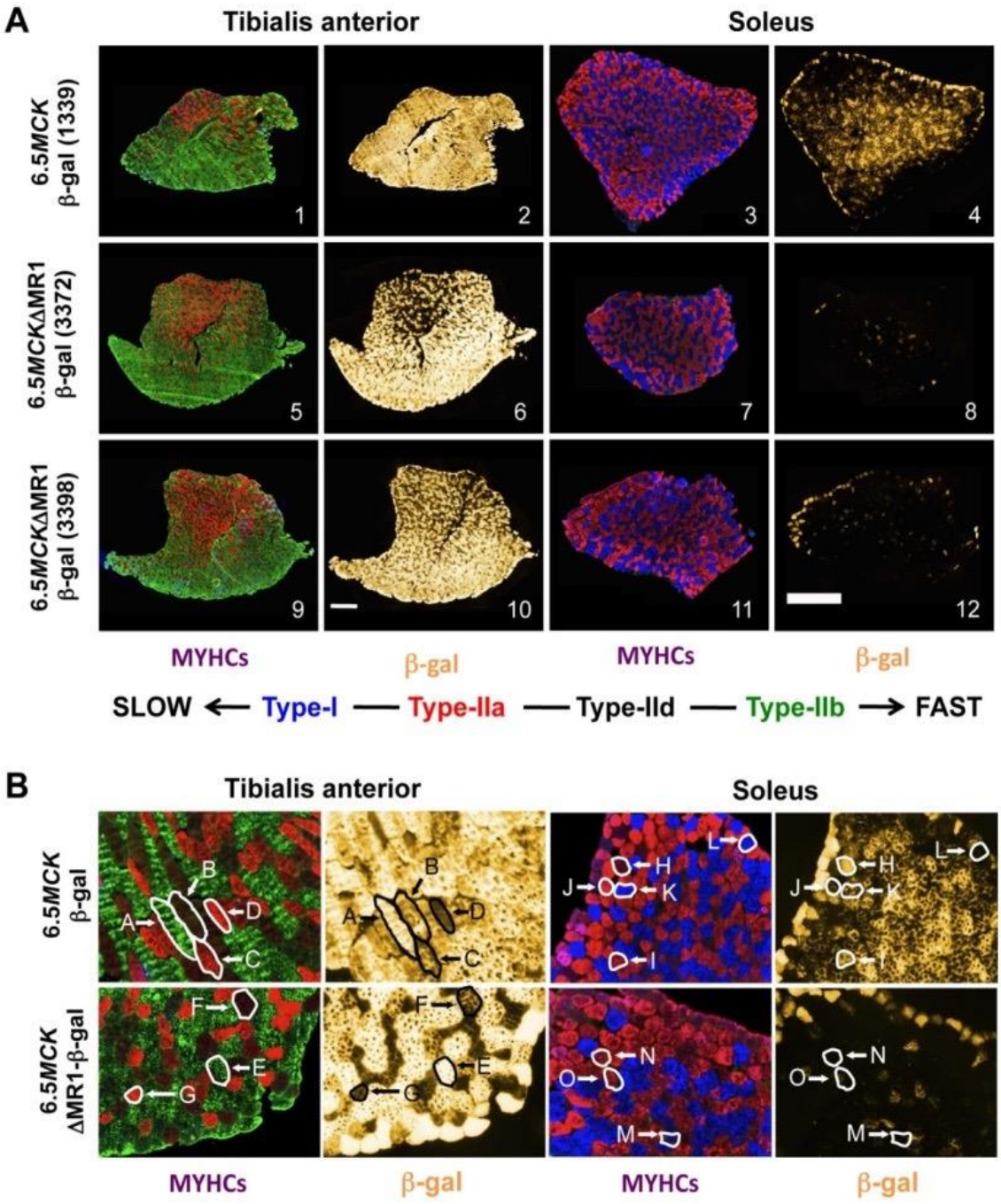


Figure 5. ***MR1 is important for* MCK *expression in slow- and intermediate-twitch skeletal muscle fibers****. (A) Sister sections of tibialis anterior (TA) and soleus muscles from mice carrying the 6.5*MCK-β-*gal or the 6.5*Δ*MR1*-β-*gal transgenes, immunostained with myosin heavy chain (MYHC) fiber type-specific monoclonal antibodies (panels 1, 3, 5, 7, 9 and 11) or activity stained for* β-*galactosidase (*β-*gal) expression (panels 2, 4, 6,*

8, 10 and 12). Antibodies for different isoforms and fluorophore-labeled secondary antibodies mark the fiber types as follows: slow-twitch fibers (type I), blue; intermediate-twitch fibers (type IIa), red; and fast-twitch fibers (types IIb and IId), green and black, respectively (the black appearance of type IId fibers is due to the absence of any type 1, IIa, or IIb antibody binding). Purplish fibers contain both types I and IIa MYHCs (see Figure 5B, soleus), and fibers with weak red or green staining probably contain mixtures of type IId (no color) + type IIa or type IId + type IIb, respectively (see Figure 5B, TA). Sister sections were stained for β*-gal expression (false colored gold). Bars are 0.5 mm. (B) Higher magnification sections indicate differences in* β*-gal expression between fiber types in transgenic lines with and without MR1. Individual fibers, outlined in white or black to show relative differences in X-gal staining between fiber types (type I = K, L and O; type IIa = C, D, G, I and J; type IId = B, F, H and M; and type IIb = A and E), can be cross-referenced to* β*-gal expression in sister sections.*

Comparisons between immunostained and X-gal-stained sister cross-sections of the TA and soleus muscles of mice carrying the 6.5*MCK*-β-gal transgene showed β-gal expression in all fiber types, but there was a clear difference in the distribution of X-gal staining intensities among fiber types in the predominantly fast-twitch TA muscles compared to the predominantly slow- and intermediate-twitch soleus muscles (Figure 5A, panels 2 and 4). As a general rule in TA muscle, type IIb fibers exhibit greater X-gal staining than type IId fibers, and type IIa fibers exhibit the least staining (Figure 5B, TA X-gal panel, fiber staining intensities: A > B > C), whereas in the soleus, type IId and type IIa fibers exhibit the greatest X-gal staining and type I fibers stain the least (Figure 5B, soleus X-gal panel, fiber staining intensities: H > I > K).

Interestingly, fibers that show similar MYHC expression can also vary in X-gal staining intensity (compare TA fibers C with D and soleus fibers I with J and K with L). However, the overall trend found within the same transgenic mouse and even within the same anatomical muscles is that the 6.5*MCK*-β-gal transgene is more active in individual

fast-twitch muscle fibers than in intermediate- and slow-twitch fibers. These β-gal/fiber-type staining patterns were consistent among all mice tested (*n* = 7) in the single 6.5*MCK*-β-gal-transgenic line.

Four transgenic mouse lines that contain the 6.5-kb regulatory region lacking MR1 (6.5*MCK*ΔMR1-β-gal) exhibit a strikingly different β-gal expression profile. In the TA, there is weaker relative X-gal staining in regions of the TA that are dominated by type IIa fibers (Figure 5A; compare panels 5, 6, 9 and 10 with panels 1 and 2). At higher magnification, this difference can be directly correlated with low levels of X-gal staining in type IIa fibers (Figure 5B, TA panels, fiber G and others) and reduced staining in some type IId fibers (Figure 5B, TA panels, fiber F and others). However, in the same TA muscle, type IIb fibers (Figure 5, fiber E and others) stain intensely for β-gal. In the soleus muscle, X-gal staining is relatively weak throughout the section in comparison to similarly treated TA muscle sections (Figure 5A, panels 7, 8, 11 and 12 vs. panels 3 and 4). At higher magnification, both type I and type IIa muscle fibers show very weak X-gal staining (Figure 5B, soleus panels, fibers N, O and others), while the few fibers that express β-gal are type IId fibers (Figure 5B, soleus panels, fiber M and others). These observations were consistent among all mice tested (*n* = 7) from the four independent 6.5*MCK*ΔMR1-β-gal-transgenic lines. This suggests that MR1 contributes strongly to the expression of *MCK* in type I and type IIa fibers, and perhaps weakly in type IId fibers, but that MR1 is not absolutely required for high-level *MCK* expression in type IIb fibers.

Expression levels from the wild-type 6.5*MCK*-β-gal and 6.5*MCK*ΔMR1-β-gal transgenes were also examined in protein extracts from entire anatomical muscles containing different proportions of fast and slow fibers. Extensor digitorum longus (EDL) muscles (primarily fast-twitch fibers) and soleus muscles (primarily slow-twitch and intermediate-twitch fibers) were dissected from four or five mice each from the most

highly active lines carrying each transgene, and β-gal specific activity was determined. In all mice from each transgenic line, EDL extract activities were significantly higher than those from the soleus. However, because absolute expression levels typically differ between individual transgenic mouse lines, owing to variable transgene integration sites and copy numbers [25-27], the β-gal-specific activity levels were then normalized for each line by dividing the EDL levels by the soleus levels. The ratio was three times higher in extracts from the 6.5*MCK*ΔMR1-β-gal-transgenic mice (data not shown). In combination with the much lower transgene expression levels observed within the individual type I and type IIa fibers of 6.5*MCK*ΔMR1-β-gal-transgenic mice (Figure 5), the quantitative data are consistent with the conclusion that the MR1 region plays a relatively more important role in *MCK* gene expression in muscles containing slow and intermediate fiber types than in muscles containing primarily fast fibers.

S1.4 Discussion

In this study, we characterized the *MCK* intronic region MR1 [22] and found that it contains regulatory elements that provide positive transcriptional activity in skeletal muscle cells. Our results argue that MR1 is crucial for the "full" activity of the 6.5-kb *MCK* regulatory region in differentiated skeletal muscle cultures (Figure 2), and they recapitulate those of an earlier study that demonstrated MR1's ability to drive transcriptional activity in a position-independent manner [22]. Additionally, we found that MR1's positive transcriptional activity is conveyed by a highly conserved 95-bp sequence designated the *MCK*-SIE (Figure 1). When separated from its flanking MR1 regions, the *MCK*-SIE synergizes with the proximal promoter to provide transcriptional activity equivalent to that of the highly active *MCK* 5'-enhancer (Figure 2B) [22]. Interestingly, however, the *MCK*-SIE requires the 358-bp *MCK* proximal promoter for its activity,

whereas the 5'-enhancer exhibits high activity with the 80-bp *MCK* basal promoter as well as with the proximal promoter (data not shown).

The *MCK*-SIE's high activity is largely due to the paired E-box and MEF2 motifs, since their mutation or deletion caused a significant decrease in transcription, while mutations affecting the AP-1/MAF half-site motifs did not (Figure 3). Although a TRANSFAC database search of the mouse *MCK* gene's 1-kb MR1 region revealed many possible transcription factor binding motifs, and although many of these overlap with conserved sequences (Additional file 1, Figure S1), deletion of other conserved regions did not disclose a correlation with positive transcriptional activity (Additional file 1, Figure S1, and Additional file 2, Figure S2). While it is also possible that some aspects of MR1-mediated *MCK* expression are regulated by nonconserved control elements, as we have shown is the case for Six4/5 and *MAZ* elements in the 5'-enhancer and proximal promoter [24,32] and as has been shown for other genes [59,60], pursuing this possibility did not seem as immediately fruitful as investigating the SIE's E-box and MEF2 mechanisms. Nevertheless, our studies do not preclude positive transcriptional contributions from other MR1 and SIE sequences.

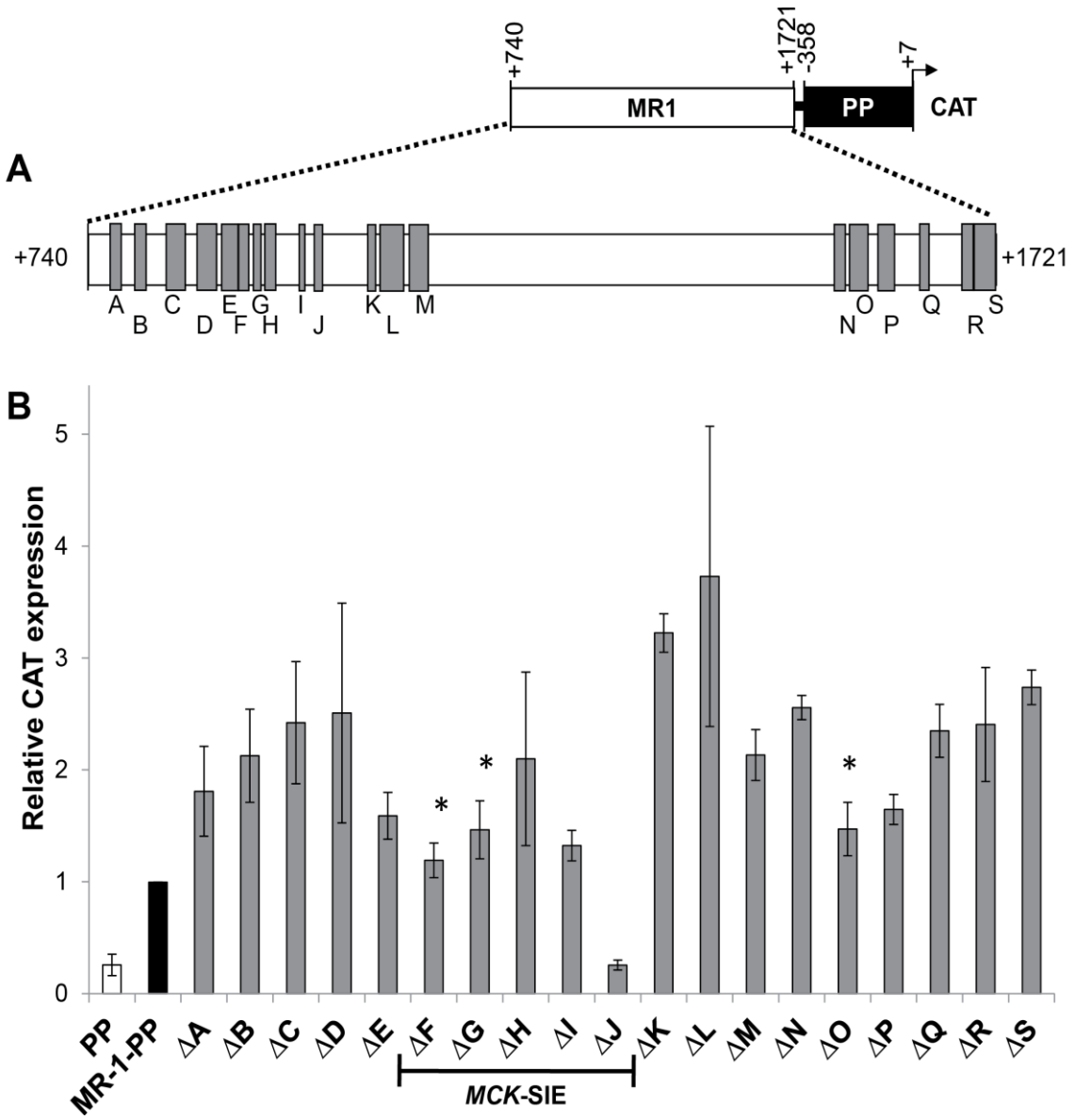


Figure S2. The functional consequence of individual deletions of the conserved 19 subregions throughout MR1*. (A) Conserved regions within MR1 (gray blocks in part A, gray bars in part B) were deleted from MR1-proximal promoter-chloramphenicol acetyl transferase (MR1-PP-CAT) and tested for transcriptional activity in skeletal myocytes (gray bars). (B) MM14 cells were transiently transfected with constructs containing each of the 19 different conserved motif deletions, and cells were harvested as described in the Figure 2 legend. Relative CAT activity was normalized with the* MCK *5'-enhancer alkaline phosphatase (AP) reference plasmid and compared to the intact MR1-PP-CAT (black bar) and to the PP-CAT (white bar). Expression levels of MR1-PP-CAT were*

scaled to equal 1.0. Asterisks indicate constructs that did not result in a statistically significant change in transcriptional activity.

Several ChIP studies have indicated the ability of E-box motifs in skeletal muscle gene promoters to recruit the basic helix-loop-helix factors MyoD and myogenin, and EMSA studies have proven E-box binding by Myf5, MRF4 and E12/47 as well [45]. Analysis of early phases of muscle differentiation also suggests that MyoD may bind muscle gene promoters as a "pioneering" factor [3] that facilitates histone acetylation [45]. As differentiation progresses, MyoD is then replaced by myogenin at the same regulatory regions. This was shown to be the case for the *MCK* 5'-enhancer in E10.5 to E14.5 mouse limb muscles [51]. This transition may be facilitated by decreased levels of Suv39h1, a histone H3 lysine 9-specific methyltransferase that represses myogenin expression via histone and MyoD methylation [61]. However, in our ChIP studies of MM14 muscle cultures harvested four days after the initiation of differentiation, a time at which 90% of the myonuclei are in MYHC-positive cells, both MyoD and myogenin were detected at the 5'-enhancer as well as at the *MCK*-SIE (Figure 4B). These data demonstrate that a rapid and complete MyoD-to-myogenin binding transition is not observed in the cell culture system used in our study. However, it may be informative that we found the ratio of myogenin to MyoD enrichment of the 5'-enhancer to be consistently greater than that of the *MCK*-SIE, where about equal ChIP signals were detected. The biological relevance of this difference in enrichment is not yet understood.

Our *MCK*-SIE ChIP data for differentiating MM14 cultures are generally consistent with ChIP-Seq studies that have probed the entire genomic occupancy of MyoD in differentiated mouse C_2C_{12} myocytes [52] in that both studies detected enriched MyoD occupancy of the *MCK*-SIE, proximal promoter and 5'-enhancer. Our data are also

consistent with a temporal ChIP-Seq data set showing no MyoD or myogenin occupancy of the *MCK*-SIE in replicating C_2C_{12} cells and highly enriched occupancy by both factors in fully differentiated cultures (A. Kirilusha, G. Kwan and B. Wold, personal communication). On the basis of our mutagenesis studies, the *MCK*-SIE 3'-E-box appears to be the more active site, since its deletion caused a greater reduction of transcriptional activity (Figure 3B). This might be attributed to the mouse 3'-E-box's being a closer match (12 of 14 bp) to the overall E-box consensus sequence than the 5'-E-box (11 of 14 bp) (Figure 1C). Alternatively, the closer proximity of the 3'-E-box than the 5'-E-box to the MEF2 site may improve the synergistic interactions between MyoD/myogenin and MEF2 and may lead to greater activity of the 3'-E-box In either case, it is not known whether one or both E-boxes preferentially associate with MyoD or myogenin *in vivo* or whether this might change under different physiological conditions. Ideally, this question could be addressed by ChIP analysis, but because the two E-boxes are only 46 bp apart, their individual occupancies cannot be definitively resolved on the basis of currently available data. Our *MCK* locus-specific MyoD/myogenin ChIP data also concur with the global ChIP-Seq MyoD data [52] with respect to occupied and unoccupied E-boxes in the sense that the strongly preferred sequence for occupied E-boxes in differentiated C_2C_{12} muscle cultures is CAG/cCTG. All of the occupied E-boxes in our study conformed to this sequence, and no unoccupied E-boxes within the *MCK* regions studied had the preferred sequence. It is also worth emphasizing that even though dozens of CAnnTG consensus E-boxes occurred within the 6.5-kb *MCK* genomic region, and while some of these occurred in clusters of two or three E-boxes within a 100-bp region (Figure 4A), neither our study nor the more comprehensive global ChIP-Seq study (personal communication, B. Wold) detected significant MyoD binding at the vast majority of these E-boxes. This indicates that the mere presence of one or more

nearby E-box motifs within transcriptionally active muscle gene promoters does not imply their functionality. Conversely, since our laboratory has proven the function of E-boxes within all three of the *MCK* genomic regions in which ChIP and ChIP-Seq detected significant MyoD binding, the data suggest that the detection of reproducible MyoD ChIP peaks of this type in muscle genes is strongly indicative of transcriptional function of the associated E-boxes. While the ChIP studies implicate MyoD and myogenin as binding to the *MCK*-SIE and 5'-enhancer E-boxes, it is important to point out that cell culture studies are not necessarily indicative of the MRFs that occupy these E-boxes in adult skeletal muscle fibers. In the latter context, it is likely that these E-boxes may be primarily occupied by MRF4, since it appears to be the predominant MRF in adult skeletal muscle [62,63].

The *MCK*-SIE MEF2 site is also critical for transcriptional activity, as removing this sequence is even more deleterious than removing the individual E-boxes (Figure 3B). Consistent with this, we found that MEF2 binds this sequence *in vitro* by EMSA using nuclear extract from MM14 myocytes (Figure 4C). Furthermore, ChIP-Seq studies of differentiated C_2C_{12} muscle cells identified enriched MEF2 occupancy at both the 5'-enhancer and the *MCK*-SIE (Figure 4D), and the fold enrichments at these sites relative to the negative control were more than 10-fold. A diffuse signal over the proximal promoter region was also observed, and this signal may reflect either that binding to a nonconsensus MEF2 site or that MEF2 association with MyoD/myogenin bound to a proximal promoter E-box located at -247 bp provides positive transcriptional activity both *in vitro* and *in vivo* [25,27]. Alternatively, MEF2 enrichment at the proximal promoter may be due to the secondary binding of MEF2 complexes formed at the 5'-enhancer and/or the *MCK*-SIE physically contacting the promoter. Such long-distance interactions of enhancer-affiliated factors with promoter DNA via cross-linking with initiation complex

proteins have been readily detected in standard ChIP reactions during chromatin conformation capture [64].

Overall, we conclude that MEF2 interacts *in vivo* with the *MCK*-SIE complex. The strong dependency of *MCK*-SIE function on the presence of the MEF2 control element (Figure 3B) also supports the hypothesis that MEF2 likely binds directly at this site. The functional synergy of this MEF2 site with E-box control elements bound by MyoD and myogenin is reminiscent of the behavior of an analogous E-box pair and MEF2 site in the *MCK* 5'-enhancer [23] and is consistent with a model of cobinding involving MEF2 and MRFs [46,56,65], although simultaneous occupancy by both factors *in vivo* is inferred and has not been directly measured.

Interestingly, all four isoforms of MEF2 (MEF2A, MEF2B, MEF2C and MEF2D) are present in myocyte cultures [53], but MEF2B is not present in adult mouse muscle [66,67]. The *MCK*-SIE sequence does not predict which, if any, MEF2 isoforms bind preferentially [53], and the antibodies used in our ChIP assays cross-reacted with all MEF2 isoforms. Thus, it is possible that the MEF2 site may be occupied by any of the MEF2 isoforms present in differentiated skeletal muscle cultures. It is also plausible that the MEF2 site can be occupied by other non-MEF2 factors that recognize AT-rich motifs. For example, AT-rich motifs similar to the one found in the *MCK*-SIE are known to bind nuclear factors such as Oct-1, TEF-1 and MHox [24,51,55,68-72], and the *MCK* 5'-enhancer's MEF2 and AT-rich motifs have been shown to recruit MEF2, Oct-1 and TEF-1. In this regard, even though the *MCK* 5'-enhancer and *MCK*-SIE contain similar paired E-box/MEF2 motifs, the *MCK*-SIE fails to bind TEF-1 by EMSA analysis (Figure 4C), whereas the 5'-enhancer MEF2 element binds TEF-1 [55]. Although the functional consequences of this difference are unknown, these data imply that the MEF2 site-

mediated transcriptional activity of the *MCK*-SIE and *MCK* 5'-enhancer may differ in terms of their interactions with non-MEF2 factors.

An intriguing facet of MR1's regulatory function is the discovery that it contains transcriptionally repressive sequences flanking the highly positive *MCK*-SIE. These MR1 regions can repress the *MCK*-SIE's activity via the combined or individual effects of at least 15 highly conserved 9- to 24-bp sequences (Figure 2B and Additional file 1, Figure S1, and Additional file 2, Figure S2). When MR1 constructs containing individual deletions of these motifs were tested in skeletal muscle cultures, most of the deletions resulted in two- to fourfold increases in transcriptional activity (Additional file 2, Figure S2), suggesting that these conserved regions act to repress transcriptional activity. The only deletion that resulted in a significant decrease in activity overlapped the MEF2/AT-rich motif within the *MCK*-SIE region (Additional file 1, Figure S1, and Additional file 2, Figure S2). Interestingly, deletion F, which encompassed the *MCK*-SIE's conserved 5'-E-box, did not cause decreased activity when tested in the context of the entire MR1 region (Additional file 2, Figure S2), but did lead to decreased activity in the context of the isolated *MCK*-SIE (Figure 3B). This may be due to the compensatory functions of other control elements within the entire MR1.

Our studies have also begun to address the *in vivo* function of MR1 in *MCK* gene expression. Comparisons between a transgenic mouse line that contains the 6.5-kb sequence driving β-gal and several lines from which the MR1 region has been deleted revealed differences in transgene expression that indicated a correlation between MR1 function and muscle fiber type. Transgenic lines expressing the 6.5*MCK*ΔMR1-β-gal transgene expressed very low levels of β-gal in slow- and intermediate-twitch fibers (type I and type IIa), while expression levels in fast-twitch fibers (type IIb and type IId) remained high (Figure 5). Although only one wild-type 6.5*MCK*-β-gal-transgenic line was

derived in our own study, an independent transgenic study that employed the same 6.5-kb *MCK* genomic sequence to express the transcriptional enhancer factor domain family member 1 (TEAD1) transcription factor demonstrated high-level transgene expression in the soleus (slow- and intermediate-twitch muscle fibers) as well as in EDL (fast-twitch muscle fibers) [73].

Our transgenic analysis of *MCK* gene regulation has focused on correlations between transgene expression levels and fiber types defined according to their MYHC isotype expression profiles. Since *MCK* functions in an energy transport pathway that is important for optimal contractile function, it might also have been informative to identify fiber types based on metabolic markers such as succinate dehydrogenase and nicotinamide adenine dinucleotide phosphate levels that could be detected via histochemical assays and then to correlate these fiber types with transgene expression levels. This was not done for purely technical reasons, as MYHC immunostaining provided more precise distinctions between fiber types and because the ability to detect four fiber types in a single cryosection facilitated correlations between fiber types and β-gal levels in adjacent sections. Furthermore, since the original investigators of muscle fiber types based on MYHC immunostaining were very careful to ascertain that individual fibers were designated as the same fiber type by both the histochemical and immunostaining protocols [58], it seems likely that our study would have reached similar conclusions regarding the role of MR1 in *MCK* gene expression with either fiber-typing technique.

There is clearly a functional relationship between *Myhc* types and *MCK* gene expression patterns [6,74], but the underlying basis of this regulatory linkage is not known. In this regard, however, the distribution of MYHC isotypes in different anatomical muscle is not altered in *MCK*-deficient mice; rather, the lack of *MCK* appears to be

compensated by an increase in mitochondrial creatine kinase (CK) [75]. Recently, it has also been shown that the expression patterns of myosin isoforms and enzymes involved in muscle fiber energy metabolism can be uncoupled by mutations that affect glycogen storage and sarcoplasmic calcium release mechanisms [76]. These reports suggest that *MCK* transgene expression would not be anticipated to exhibit a strict correlation with muscle fiber types as assessed solely by MYHC fiber typing. This possibility may partially explain why the *MCK*-driven β-gal levels observed in transgenic TA and soleus muscles were not uniform among all fibers of each MYHC-defined type (Figure 5B). These nonuniformities in transgene expression within specific fiber types do not appear to be regulated by MR1, since they are observed in fibers carrying the intact 6.5-kb *MCK* genomic region as well as in those in mice carrying the 6.5*MCK*ΔMR1 transgene. Nevertheless, the MR1 region clearly plays an important role in regulating the steady-state levels of *MCK* gene expression in different anatomical muscles and in different fiber types. In this regard, it has yet to be determined whether MR1 or the *MCK*-SIE alone can drive expression in slow- and intermediate-twitch muscle fibers independently of the 5'-enhancer. It is also not known which physiological signals impinge on the *MCK*-SIE and on the flanking repressive regions within MR1.

Transgenic analysis of fiber type-specific muscle gene expression has also been carried out with the *MLC2v*, *MLC1/3f*, *aldolase A* and *slow troponin I* muscle genes [7-10,14]. Similarly to our studies with *MCK*, E-boxes and MEF2 control elements have been identified within their key regulatory regions. In particular, the *slow troponin I* SURE region contains the critical E-box, MEF2, and a CACC motifs, which in isolation confer pan-muscle expression. Interestingly, the inclusion of a more upstream region within SURE, which contains a *bicoid*-like motif that recruits the general transcription factor 3 (GTF3)/muscle transcription factor II I repeat domain-containing protein 1 (MusTRD1),

restricts activity to slow-twitch muscle [11,14]. A related mechanism may modulate the *in vivo* activity of the *MCK*-SIE, leading to the contribution of MR1 to expression in slow-twitch fibers. However, neither the *bicoid*-like motif (GTTAATCCG) [14] nor the GTF3 consensus DNA binding sequence ($G_{TC}\,G_{A}\,GATTA_{G}\,BG_{A}$) [11] is found in or immediately adjacent to the *MCK*-SIE. In contrast, the fast-twitch activity of the *MCK* 5'-enhancer may be partially due to recruitment of the Six4 transcription factor, since the MEF3 site in the *aldolase A* pM promoter is necessary but not sufficient to drive transcription in some fast-twitch muscle fibers [77].

The contribution of multiple enhancer regions to the expression of striated muscle genes in different fiber types may be a common mechanism. For example, transgenic analysis has demonstrated that the troponin I (fast) enhancer intronic regulatory element (TnIfast IRE), in isolation, results in fast twitch fiber-specific expression in the adult plantaris muscle, where TnIfast IRE elements yield the highest levels of expression in type IIb fibers, intermediate levels in type IId, very low levels in type IIa fibers and no expression in type I fibers [16], while the endogenous *TnIfast* gene is expressed at similar levels in all fast-twitch fiber types [15]. The *MCK* gene MR1 region, although its activity contributes to expression in slow and intermediate fibers, appears analogous to TnIfast IRE in that both regulatory regions provide relatively restricted fiber-type expression patterns and both genes require the contribution of multiple fiber-specific enhancers to achieve pan-skeletal muscle expression. The *MCK* MR1 and 5'-enhancer regulatory regions thus appear to share common mechanisms of transcription with several fast- and slow-twitch muscle genes.

S1.5 Conclusions

This study identifies a regulatory region within the *MCK* gene's intron 1 that plays a major transcriptional role in slow- and intermediate-twitch muscle fibers. This activity

was shown *in vitro* to be dependent on the *MCK*-SIE region, which contains a paired E-box and MEF2 motif. Each motif was shown to be required for full *MCK*-SIE transcriptional activity, and ChIP studies showed that they recruit MyoD, myogenin and MEF2, respectively. It was also shown that the *MCK*-SIE is flanked by repressive regulatory regions containing multiple different negative control elements. The mechanisms and functional purposes of these remain to be determined.

S1.6 Materials and methods

S1.6.1 Sequence analysis

Sequences spanning the TATA box to exon 2 of the *MCK* gene of *Homo sapiens* (human [AC005781.1]), *Felis catus* (cat [GenBank: AC135221.3AC135221.3]), *Canis familiaris* (dog [GenBank: AC137538.2]), *Bos taurus* (bovine [GenBank: AC137535.2]), *Sus scrofa* (pig [GenBank: AC139878.2]) and *Mus musculus* (mouse GenBank: [AC118017.15]) were obtained from compiled genomic sequences in the Entrez Genome Project database and subjected to sequence alignment using ClustalW [78]. The intron 1 sequences of both mouse and human were independently analyzed for putative control element motifs using Match http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi webcite) (Contact B. Wold for specifics: http://woldb@caltech.edu webcite, a matrix search algorithm that scours the TRANSFAC database of transcription factors and their experimentally proven binding sites. Parameters were set to select for vertebrate-only matrices with a 90% core binding similarity to broaden the rate of positive hits.

S1.6.2 Plasmid constructs

A 6.5-kb construct of the mouse *MCK* gene (-3,349 to +3,230) [37] was cloned upstream of the CAT reporter gene 6.5*MCK*-CAT [26]. The 6.5*MCK*ΔMR1-CAT

construct was generated from the 6.5*MCK*-CAT construct by introducing *Cla*I restriction sites 5' and 3' of MR1 (+740 and +1,724) using the QuikChange Site-Directed Mutagenesis Kit (Stratagene, http://www.genomics.agilent.com/ webcite), according to the manufacturer's directions. MR1 was then deleted by digestion of the plasmid with *Cla*I and religation of the remaining vector. The 6.5*MCK*ΔEnh-CAT construct was generated by site-directed mutagenesis to delete the *MCK* 5'-enhancer (-1,256 to -1,040).

The MR1 region was polymerase chain reaction-amplified from the existing 6.5-kb construct with primers containing the restriction sites *Sph*I (5') and *Bst*I (3'). The MR1 amplicon was cloned upstream of the proximal promoter by replacing the 5'-enhancer in the e-358-CAT reporter construct [27] using *Sph*I and *Bst*I. The mouse *MCK* PP region used in these studies extends from -358 to +7. All other deletions and mutations described in this study were generated using the QuikChange Site-Directed Mutagenesis Kit.

S1.6.3 Transient transfections and reporter gene assays

MM14 skeletal myoblasts were cultured as described previously [79]. Collagen-coated 100-mm dishes were inoculated with about 1×10^5 log phase cells/dish and were allowed to proliferate under growth conditions (85% Ham's F10C nutrients + gentamicin, 15% horse serum and 2 ng/mL basic fibroblast growth factor (bFGF) added at approximately 12-hour intervals) for about 24 hours. Myoblasts were cotransfected using a standard calcium phosphate method [23] at about 3×10^5 cells with test constructs driving the expression of the CAT reporter gene and an AP reference plasmid, which contains the 5'-enhancer placed 5' of the basal promoter sequence (-80 to +7). Transfected MM14 cultures were induced to differentiate four hours after beginning the transfection by aspirating the growth medium, rinsing once with saline G, incubating for 2

minutes at room temperature in 15% glycerol 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid-buffered saline, rinsing again with saline G and then adding 10 mL of differentiation medium (98.5% Ham's F10C nutrients + gentamicin, 1.5% horse serum and 1 μM insulin) [79]. Relative enzymatic activities of CAT and AP were determined from extracts as described in previous studies [25]. Since the *MCK* enhancer-AP reference plasmid is expressed only in differentiated muscle cells, it provides a control for plate-to-plate variability in transfection efficiency and extent of muscle differentiation in skeletal myocyte cultures.

S1.6.4 ChIP assays

ChIP assays were performed using a modification of the Fast-ChIP method as described previously [32,80] with the following nuances: 100-mm dishes were plated with about 1×10^5 log phase MM14 cells/dish and grown to near confluence (about 4×10^6 cells/dish), then allowed to differentiate in proliferation medium without additional bFGF for four to six days prior to harvesting. All cultures contained more than 90% terminally differentiated myocytes as assessed by immunostaining a parallel culture with the myosin-specific antibody MF-20. This procedure produced more than 7×10^6 differentiated myonuclei per 100-mm dish. Cells were sonicated with 16 rounds of 15-second pulses with 45 seconds of rest between pulses (four minutes total) on a Model 100 Sonic Dismembrator (Fisher Scientific, http://www.fishersci.com/ webcite) at the highest setting. Antibodies used for immunoprecipitation described in this study were as follows: anti-myogenin (M-225) sc-576 X, anti-MyoD (M-318) sc-760 X, anti-MEF2A (C-21) sc-313 and normal rabbit IgG sc-2027 (Santa Cruz Biotechnology, http://www.scbt.com/ webcite). The primers used in ChIP analyses were *MCK* 5'-enhancer: 183 bp forward: 5'-GCCCATGTAAGGAGGCAAGGCC-3', reverse: 5'-CACCAGGGACAGGGTTATTTTTAGAGC-3', *MCK* exon 1/intron 1 boundary: 217 bp

forward: 5'-GGGTCACCACCACCTCCACAG-3', reverse: 5'-GCCTTGCAAGGAGGGGACACTTG-3', *MCK*-SIE: 168 bp, forward: 5'-CTTGAGGCCCAGAGCCTGGCTG-3', reverse: 5'-GAGACCCAAAGCCCTTGAAGCTGCTAC-3', *MCK* exon 2: 207 bp, forward: 5'-GTCCCAAAGGCCGCCACCATG-3', reverse: 5'-GGGTTGTCCACCCCAGTCTGG-3' *Mark4* gene region: 205 bp, forward: 5'-GGATGCCATGCCTGGTGGCCAT-3', reverse: 5'-GCCATGCAGCTTTCACGCAGAGG-3'.

S1.6.5 EMSA

EMSA was carried out as previously described [32]. Nuclear extracts from differentiated skeletal muscle cultures were prepared as previously described [81] using a cocktail of several protease inhibitors (P8340; Sigma, St. Louis, MO, USA). Total protein in the extracts was quantitated by using the Bradford method [82]. Incubations with antisera or unlabeled oligonucleotide competitors were carried out at room temperature for 20 minutes prior to the addition of probe. The 5' to 3' sequences of the double-stranded probes or competitors with introduced mutations of the sequence underlined are

MEF2 (*MCK*-SIE):

AGGAGCATCTAAAAATAGCCACAAAG

MEF2 (*MCK*-SIE)-M1:

AGGAGCATC

CG

AAAA

CG

GCCACAAAG

MEF2 (*MCK*-SIE)-M2:

AGGAGCATC

AT

AAAA

AT

GCCACAAAG.

Antibodies used for EMSA were anti-MEF2A (pan-MEF2, C-21) (Santa Cruz Biotechnology), anti-TEF-1 (BD Transduction Laboratories, http://www.bdbiosciences.com/home.jsp webcite) and IgG normal rabbit sc-2027 (Santa Cruz Biotechnology, http://www.scbt.com/ webcite).

S1.6.6 ChIP-Seq assays

ChIP assays for MEF2 ChIP-Seq were performed according to the protocol described by Johnson *et al*. [83] with the modifications described in the paragraph below. C_2C_{12} cells were grown at low density on Nunclon 14-cm-diameter plates (Fisher Scientific, http://www.fishersci.com/ webcite) in 20% fetal bovine serum (FBS)/Dulbecco's modified Eagle's medium (DMEM) (#11965; Invitrogen http://www.invitrogen.com/site/us/en/home.html webcite with penicillin and streptomycin and passaged at no more than 50% confluence. Upon reaching confluence, differentiation was induced by switching to 2% horse serum/1 μM insulin/DMEM. After 60 hours of differentiation, the cells were cross-linked with 1% formaldehyde (Avantor Performance Materials, http://www.avantormaterials.com/ webcite) and 0.025% glutaraldehyde (Polysciences, Inc. http://www.polysciences.com/ webcite) for 10 minutes. A total of 2×10^7 cells were fragmented to about 100 to 300 bp using 30-second, 12-W cycles on a Misonix 3000 sonicator

http://www.fishersci.com/ecomm/servlet/fsproductdetail?aid = 2819374&storeId = 10652 webcite for a total sonication time of 15 minutes. The sheared chromatin was immunoprecipitated using 100 μL of sheep anti-mouse IgG M280 beads (Invitrogen) and 5 μg of MEF2 monoclonal antibody (clone B4) from Santa Cruz Biotechnology or 200 μL of sheep anti-rabbit IgG M280 beads and 10 μg of MEF2 polyclonal antibody (clone H300) from Santa Cruz Biotechnology. Illumina libraries for sequencing were made using their ChIP-Seq library kit (Illumina, Inc., http://www.illumina.com/ webcite) as described by the manufacturer, except that a 10-cycle amplification was performed before gel selection according to the method of Johnson *et al*. (library 2) [83]. Library sequencing was performed for 36 cycles on an Illumina Genome Analyzer (Illumina, Inc.), and the resulting reads were mapped to the mouse MM9 genome by using Bowtie software [84]. Mapped reads that permitted up to two mismatches to the reference genome were displayed on the University of California Santa Cruz (UCSC) Genome Browser. ChIP-Seq signals were called using the ChIP-Seq module within the ERANGE version 3.2 software package [85] and were also mapped using the MACS peak caller [86].

S1.6.7 Transgenic mice

The 6.5-kb *MCK* gene sequence and the sequence with MR1 deleted were cloned upstream of the β-gal reporter gene to generate the 6.5*MCK*-β-gal and 6.5*MCK*ΔMR1-β-gal constructs, respectively. DNA for microinjection was prepared by enzymatic digestion to linearize the plasmids and gel-purified by freeze-and-squeeze columns (Bio-Rad Laboratories, http://www.bio-rad.com/ webcite). Transgenic mice were produced using eggs from C57BL/6J × C3H crosses through the University of Washington Transgenic Resource Program. Founders were crossed to C57BL/6J to

generate F1s. Lines of mice analyzed were either F1s or the progeny of F1s (N2 or N3) that were back-crossed with C57B/6J.

S1.6.8 Dissections

Adult mice (1+ months) were killed according to methods approved by the University of Washington Institutional Animal Care and Use Committee. TA and soleus muscles were dissected and mounted in a 2:1 mixture of optimal cutting temperature compound and 10% gum tragacanth in cryomold cassettes. Cassettes were then frozen in liquid nitrogen-cooled isopentane. Tissues contained in blocks were cryosectioned at a thickness of 6 μM at -25°C using a Leica cryostat http://www.leica-microsystems.com/ webcite, mounted onto glass slides at room temperature and then stored at -80°C.

S1.6.9 X-gal staining

Slides were fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) for 15 minutes at 4°C and washed in 100 mM sodium phosphate (pH 7.3), 2 mM $MgCl_2$, 0.01% sodium oxycholate and 0.02% Nonidet P-40 and stained in a standard X-gal reagent solution [87] for 4 hours. After staining, slides were fixed for 15 minutes in 10% formalin and mounted in gelvatol (Sigma-Aldrich, http://www.sigmaaldrich.com webcite). Images were obtained using a Zeiss Axiovert 200 microscope http://www.zeiss.com/micro webcite with a Zeiss AxioCam MRm camera (Zeiss), and acquired using AxioVision software (Zeiss). Images were then uniformly false-colored using Adobe PhotoShop version 7 software (Adobe Systems, http://www.adobe.com/ webcite).

S1.6.10 Immunofluoresence

Monoclonal antibodies specific for myosin isoforms MYHC1, MYHC2A and MYHC2B were produced from cultures of hybridoma lines BA-D5, SC-71 and BF-F3, respectively [58]. These antibodies stain type I, type IIa and type IIb fibers, respectively. Cultures were grown to high density in DMEM High Glucose (HyClone Laboratories, http://www.hyclone.com/ webcite) supplemented with 10% FBS (Gemini Bioproducts, http://www.gembio.com/ webcite)) and penicillin-streptomycin (Sigma). Cultures were then switched to serum-free medium and incubated for two or three days. The medium was collected, centrifuged and filter-sterilized (0.22 μm Stericup; Millipore, http://www.millipore.com/ webcite) and monoclonal antibodies were concentrated by HiTrap column chromatography (GE Healthcare Biosciences, http://www.gelifesciences.com/ webcite). High-protein concentration fractions as determined by the Bradford method [82] were pooled and dialyzed (Slide-A-Lyzer Dialysis Cassettes; Pierce Biotechnology, http://www.piercenet.com/ webcite), and then stored at -20°C. Slides were treated with blocking buffer (1% bovine serum albumin and 0.05% Tween 20 in PBS) and incubated with about 10 μg/mL BA-D5, SC-71 and BF-F3 for 1 hour, washed three times for five minutes in blocking buffer and incubated with goat anti-mouse secondary antibodies IgG2b Alexa Fluor 350, IgG1 Alexa Fluor 594 and IgM Alexa Fluor 488 (Invitrogen) for 30 minutes. Slides were washed as before, rinsed in PBS and mounted in gelvatol. Images were acquired as described above.

S1.7 Abbreviations

AP: alkaline phosphatase; bFGF: basic fibroblast growth factor; β-gal: β-galactosidase; *BRG1*: Brahma-related gene 1; CAT: chloramphenicol acetyl transferase; ChIP: chromatin immunoprecipitation; ERANGE: Enhanced Read Analysis of Gene Expression; KLF3: Krupple-like factor 3; *Mark4*: MAP/microtubule affinity-regulating

kinase 4 gene; MAZ: Myc-associated zinc finger protein; MEF2: myocyte enhancer factor 2; *MCK* and MCK: muscle creatine kinase gene and protein; MCK, *MCK*-SIE: *MCK* small intronic enhancer; MR1: modulatory region 1; MYHC: myosin heavy chain; Oct-1: octamer-binding protein; TA: tibialis anterior muscle.

S1.8 Competing interests

The authors declare that they have no competing interests.

S1.9 Authors' contributions

PWLT carried out the sequence alignments; made the test gene constructs; carried out the transfection assays, the ChIP analyses, the immunohistochemistry and immunofluorescence assays; and drafted parts of the manuscript describing Hauschka Lab data. KIFA conceived of the redesign of the ChIP-Seq fixation, performed and analyzed MEF2 ChIP-Seq and drafted portions of the manuscript describing Wold Lab data. CLH carried out the EMSA study and helped to draft the manuscript. CLS participated in the immunohistochemistry and immunofluorescence assays. APM participated in the transfection analyses. DLH carried out the whole muscle extract transgene expression assays. JCA and REW prepared and labeled the MYHC monoclonal antibodies and participated in the immunohistochemistry assays. BJW conceived of the global ChIP-Seq analysis of multiple myogenic transcription factors, participated in the design and coordination of the MEF2 ChIP-Seq studies and drafted parts of the manuscript describing the Wold Lab data. Together with PWLT, SDH conceived of the overall study, participated in its design and coordination between the two laboratories and played a major role in writing the manuscript. All authors read and approved the final manuscript.

S1.10 Acknowledgements

We thank members of the Hauschka and Jeff Chamberlain laboratories at the University of Washington, Q. Nguyen, D. Helterline, E. Nishiuchi, M. Haraguchi, P. Gregorevic, B. Sharma, A. Zebari and J. Buskin, for their technical assistance and useful comments. We thank members of the Wold Lab at the California Institute of Technology, D. Trout, B. King, H. Amrhein, L. Schaffer and I. Antoschken, for technical assistance, bioinformatics and/or critical discussions. This work was supported by National Institutes of Health (NIH) grant R01-0AR18860 (to SDH), NIH grant 1P01-NS046788 (to SDH), NIH grant 5732-HD07183 Developmental Biology Training Grant (to PWLT) and NIH grant T32-HL007312, Experimental Pathology of Cardiovascular Disease (to CLH). KF is supported by a Graduate Research Fellowship from the National Science Foundation and from the Beckman Foundation at the California Institute of Technology; BJW received funding from NIH grant U54 HG004576.

S1.11 References

1. Welle S, Bhatt K, Thornton CA: **Inventory of high-abundance mRNAs in skeletal muscle of normal men.**

   Genome Res 1999, **9:**506-513. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

2. Chamberlain JS, Jaynes JB, Hauschka SD: **Regulation of creatine kinase induction in differentiating mouse myoblasts.**

   Mol Cell Biol 1985, **5:**484-492. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

3. Tapscott SJ: **The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription.**

   Development 2005, **132:**2685-2695. PubMed Abstract | Publisher Full Text

4. Lyons GE, Muhlebach S, Moser A, Masood R, Paterson BM, Buckingham ME, Perriard JC: **Developmental regulation of creatine kinase gene expression by myogenic factors in embryonic mouse and chick skeletal muscle.**

   Development 1991, **113:**1017-1029. PubMed Abstract | Publisher Full Text

5. Yamashita K, Yoshioka T: **Profiles of creatine kinase isoenzyme compositions in single muscle fibres of different types.**

   J Muscle Res Cell Motil 1991, **12:**37-44. PubMed Abstract | Publisher Full Text

6. Kushmerick MJ, Moerland TS, Wiseman RW: **Mammalian skeletal muscle fibers distinguished by contents of phosphocreatine, ATP, and Pi.**

   Proc Natl Acad Sci USA 1992, **89:**7521-7525. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

7. Wentworth BM, Donoghue M, Engert JC, Berglund EB, Rosenthal N: **Paired MyoD-binding sites regulate myosin light chain gene expression.**

   Proc Natl Acad Sci USA 1991, **88:**1242-1246. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

8. Salminen M, Lopez S, Maire P, Kahn A, Daegelen D: **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.**

   Mol Cell Biol 1996, **16:**76-85. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

9. Lupa-Kimball VA, Esser KA: **Use of DNA injection for identification of slow nerve-dependent regions of the MLC2s gene.**

   Am J Physiol 1998, **274:**C229-235. PubMed Abstract | Publisher Full Text

10. Esser K, Nelson T, Lupa-Kimball V, Blough E: **The CACC box and myocyte enhancer factor-2 sites within the myosin light chain 2 slow promoter cooperate in regulating nerve-specific transcription in skeletal muscle.**

    J Biol Chem 1999, **274:**12095-12102. PubMed Abstract | Publisher Full Text

11. Vullhorst D, Buonanno A: **Multiple GTF2I-like repeats of general transcription factor 3 exhibit DNA binding properties. Evidence for a common origin as a sequence-specific DNA interaction module.**

J Biol Chem 2005, **280:**31722-31731. PubMed Abstract | Publisher Full Text

12. Rana ZA, Gundersen K, Buonanno A: **The ups and downs of gene regulation by electrical activity in skeletal muscles.**

    J Muscle Res Cell Motil 2009, **30:**255-260. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

13. Issa LL, Palmer SJ, Guven KL, Santucci N, Hodgson VR, Popovic K, Joya JE, Hardeman EC: **MusTRD can regulate postnatal fiber-specific expression.**

    Dev Biol 2006, **293:**104-115. PubMed Abstract | Publisher Full Text

14. Calvo S, Vullhorst D, Venepally P, Cheng J, Karavanova I, Buonanno A: **Molecular dissection of DNA sequences and factors involved in slow muscle-specific transcription.**

    Mol Cell Biol 2001, **21:**8490-8503. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

15. Hallauer PL, Bradshaw HL, Hastings KE: **Complex fiber-type-specific expression of fast skeletal muscle troponin I gene constructs in transgenic mice.**

    Development 1993, **119:**691-701. PubMed Abstract | Publisher Full Text

16. Hallauer PL, Hastings KE: **TnIfast IRE enhancer: multistep developmental regulation during skeletal muscle fiber type differentiation.**

    Dev Dyn 2002, **224:**422-431. PubMed Abstract | Publisher Full Text

17. Buskin JN, Hauschka SD: **Identification of a myocyte nuclear factor that binds to the muscle-specific enhancer of the mouse muscle creatine kinase gene.**

    Mol Cell Biol 1989, **9:**2627-2640. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

18. Donoviel DB, Shield MA, Buskin JN, Haugen HS, Clegg CH, Hauschka SD: **Analysis of muscle creatine kinase gene regulatory elements in skeletal and cardiac muscles of transgenic mice.**

    Mol Cell Biol 1996, **16:**1649-1658. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

19. Jaynes JB, Chamberlain JS, Buskin JN, Johnson JE, Hauschka SD: **Transcriptional regulation of the muscle creatine kinase gene and regulated expression in transfected mouse myoblasts.**

Mol Cell Biol 1986, **6:**2855-2864. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

20. Jaynes JB, Johnson JE, Buskin JN, Gartside CL, Hauschka SD: **The muscle creatine kinase gene is regulated by multiple upstream elements, including a muscle-specific enhancer.**

Mol Cell Biol 1988, **8:**62-70. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

21. Johnson JE, Gartside CL, Jaynes JB, Hauschka SD: **Expression of a transfected mouse muscle-creatine kinase gene is induced upon growth factor deprivation of myogenic but not of nonmyogenic cells.**

Dev Biol 1989, **134:**258-262. PubMed Abstract | Publisher Full Text

22. Johnson JE, Wold BJ, Hauschka SD: **Muscle creatine kinase sequence elements regulating skeletal and cardiac muscle expression in transgenic mice.**

Mol Cell Biol 1989, **9:**3393-3399. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

23. Amacher SL, Buskin JN, Hauschka SD: **Multiple regulatory elements contribute differentially to muscle creatine kinase enhancer activity in skeletal and cardiac muscle.**

Mol Cell Biol 1993, **13:**2753-2764. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

24. Himeda CL, Ranish JA, Angello JC, Maire P, Aebersold R, Hauschka SD: **Quantitative proteomic identification of six4 as the trex-binding factor in the muscle creatine kinase enhancer.**

Mol Cell Biol 2004, **24:**2132-2143. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

25. Nguyen QG, Buskin JN, Himeda CL, Fabre-Suver C, Hauschka SD: **Transgenic and tissue culture analyses of the muscle creatine kinase enhancer Trex control element in skeletal and cardiac muscle indicate differences in gene expression between muscle types.**

Transgenic Res 2003, **12:**337-349. PubMed Abstract | Publisher Full Text

26. Nguyen QG, Buskin JN, Himeda CL, Shield MA, Hauschka SD: **Differences in the function of three conserved E-boxes of the muscle creatine kinase gene in cultured myocytes and in transgenic mouse skeletal and cardiac muscle.**

J Biol Chem 2003, **278:**46494-46505. PubMed Abstract | Publisher Full Text

27. Shield MA, Haugen HS, Clegg CH, Hauschka SD: **E-box sites and a proximal regulatory region of the muscle creatine kinase gene differentially regulate expression in diverse skeletal muscles and cardiac muscle of transgenic mice.**

Mol Cell Biol 1996, **16:**5058-5068. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

28. Mueller PR, Wold B: **In vivo footprinting of a muscle specific enhancer by ligation mediated PCR.**

Science 1989, **246:**780-786. PubMed Abstract | Publisher Full Text

29. Jaynes J: *Transcriptional Regulation of the Mouse Muscle Creatine Kinase Gene. Thesis Dissertation.* University of Washington, Department of Biochemistry; 1986.

30. Tamir Y, Bengal E: **p53 protein is activated during muscle differentiation and participates with MyoD in the transcription of muscle creatine kinase gene.**

Oncogene 1998, **17:**347-356. PubMed Abstract | Publisher Full Text

31. Vincent CK, Gualberto A, Patel CV, Walsh K: **Different regulatory sequences control creatine kinase-M gene expression in directly injected skeletal and cardiac muscle.**

Mol Cell Biol 1993, **13:**1264-1272. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

32. Himeda CL, Ranish JA, Hauschka SD: **Quantitative proteomic identification of MAZ as a transcriptional regulator of muscle-specific genes in skeletal and cardiac myocytes.**

Mol Cell Biol 2008, **28:**6521-6535. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

33. Himeda CL, Ranish JA, Pearson RCM, Crossley M, Hauschka SD: **KLF3 regulates muscle-specific gene expression and synergizes with SRF on KLF binding sites.**

Molecular Cellular Biology (In Review) 2010.

34. Dunant P, Larochelle N, Thirion C, Stucka R, Ursu D, Petrof BJ, Wolf E, Lochmuller H: **Expression of dystrophin driven by the 1.35-kb MCK promoter ameliorates muscular dystrophy in fast, but not in slow muscles of transgenic mdx mice.**

Mol Ther 2003, **8:**80-89. PubMed Abstract | Publisher Full Text

35. Salva MZ, Himeda CL, Tai PW, Nishiuchi E, Gregorevic P, Allen JM, Finn EE, Nguyen QG, Blankinship MJ, Meuse L, *et al*.: **Design of tissue-specific regulatory cassettes for high-level rAAV-mediated expression in skeletal and cardiac muscle.**

Mol Ther 2007, **15:**320-329. PubMed Abstract | Publisher Full Text

36. Eppenberger HM, Eppenberger M, Richterich R, Aebi H: **The Ontogeny of Creatine Kinase Isozymes.**

Dev Biol 1964, **10:**1-16. PubMed Abstract | Publisher Full Text

37. Cox GA, Cole NM, Matsumura K, Phelps SF, Hauschka SD, Campbell KP, Faulkner JA, Chamberlain JS: **Overexpression of dystrophin in transgenic mdx mice eliminates dystrophic symptoms without toxicity.**

Nature 1993, **364:**725-729. PubMed Abstract | Publisher Full Text

38. Sternberg EA, Spizz G, Perry WM, Vizard D, Weil T, Olson EN: **Identification of upstream and intragenic regulatory elements that confer cell-type-restricted and differentiation-specific expression on the muscle creatine kinase gene.**

Mol Cell Biol 1988, **8:**2896-2909. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

39. Polly P, Haddadi LM, Issa LL, Subramaniam N, Palmer SJ, Tay ES, Hardeman EC: **hMusTRD1alpha1 represses MEF2 activation of the troponin I slow enhancer.**

J Biol Chem 2003, **278:**36603-36610. PubMed Abstract | Publisher Full Text

40. Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.**

Brief Bioinform 2008, **9:**326-332. PubMed Abstract | Publisher Full Text

41. Wright WE, Binder M, Funk W: **Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site.**

Mol Cell Biol 1991, **11:**4104-4110. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

42. Blackwell TK, Weintraub H: **Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection.**

Science 1990, **250:**1104-1110. PubMed Abstract | Publisher Full Text

43. Fickett JW: **Quantitative discrimination of MEF2 sites.**

Mol Cell Biol 1996, **16:**437-441. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

44. van Dam H, Castellazzi M: **Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis.**

Oncogene 2001, **20:**2453-2464. PubMed Abstract | Publisher Full Text

45. Berkes CA, Tapscott SJ: **MyoD and the transcriptional control of myogenesis.**

Semin Cell Dev Biol 2005, **16:**585-595. PubMed Abstract | Publisher Full Text

46. Molkentin JD, Black BL, Martin JF, Olson EN: **Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins.**

Cell 1995, **83:**1125-1136. PubMed Abstract | Publisher Full Text

47. Igarashi K, Itoh K, Motohashi H, Hayashi N, Matuzaki Y, Nakauchi H, Nishizawa M, Yamamoto M: **Activity and expression of murine small Maf family protein MafK.**

J Biol Chem 1995, **270:**7615-7624. PubMed Abstract | Publisher Full Text

48. Toki T, Itoh J, Kitazawa J, Arai K, Hatakeyama K, Akasaka J, Igarashi K, Nomura N, Yokoyama M, Yamamoto M, Ito E: **Human small Maf proteins form heterodimers with CNC family transcription factors and recognize the NF-E2 motif.**

Oncogene 1997, **14:**1901-1910. PubMed Abstract | Publisher Full Text

49. Miskimins R, Miskimins WK: **A role for an AP-1-like site in the expression of the myelin basic protein gene during differentiation.**

Int J Dev Neurosci 2001, **19:**85-91. PubMed Abstract | Publisher Full Text

50. Zimprich A, Kraus J, Woltje M, Mayer P, Rauch E, Hollt V: **An allelic variation in the human prodynorphin gene promoter alters stimulus-induced expression.**

J Neurochem 2000, **74:**472-477. PubMed Abstract | Publisher Full Text

51. Ohkawa Y, Marfella CG, Imbalzano AN: **Skeletal muscle specification by myogenin and Mef2D via the SWI/SNF ATPase Brg1.**

Embo J 2006, **25:**490-501. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

52. Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, MacQuarrie KL, Davison J, Morgan MT, Ruzzo WL, *et al*.: **Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming.**

Dev Cell 2010, **18:**662-674. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

53. Black BL, Olson EN: **Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins.**

Annu Rev Cell Dev Biol 1998, **14:**167-196. PubMed Abstract | Publisher Full Text

54. Lakich MM, Diagana TT, North DL, Whalen RG: **MEF-2 and Oct-1 bind to two homologous promoter sequence elements and participate in the expression of a skeletal muscle-specific gene.**

J Biol Chem 1998, **273:**15217-15226. PubMed Abstract | Publisher Full Text

55. Karasseva N, Tsika G, Ji J, Zhang A, Mao X, Tsika R: **Transcription enhancer factor 1 binds multiple muscle MEF2 and A/T-rich elements during fast-to-slow skeletal muscle fiber type transitions.**

Mol Cell Biol 2003, **23:**5143-5164. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

56. Black BL, Molkentin JD, Olson EN: **Multiple roles for the MyoD basic region in transmission of transcriptional activation signals and interaction with MEF2.**

Mol Cell Biol 1998, **18:**69-77. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

57. Gregorevic P, Meznarich NA, Blankinship MJ, Crawford RW, Chamberlain JS: **Fluorophore-labeled myosin-specific antibodies simplify muscle-fiber phenotyping.**

Muscle Nerve 2008, **37:**104-106. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

58. Schiaffino S, Gorza L, Sartore S, Saggin L, Ausoni S, Vianello M, Gundersen K, Lomo T: **Three myosin heavy chain isoforms in type 2 skeletal muscle fibres.**

J Muscle Res Cell Motil 1989, **10:**197-205. PubMed Abstract | Publisher Full Text

59. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells.**

Nat Genet **42:**631-634.

60. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, *et al*.: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.**

Science **328:**1036-1040.

61. Mal AK: **Histone methyltransferase Suv39h1 represses MyoD-stimulated myogenic differentiation.**

    Embo J 2006, **25:**3323-3334. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

62. Rhodes SJ, Konieczny SF: **Identification of MRF4: a new member of the muscle regulatory factor gene family.**

    Genes Dev 1989, **3:**2050-2061. PubMed Abstract | Publisher Full Text

63. Hinterberger TJ, Sassoon DA, Rhodes SJ, Konieczny SF: **Expression of the muscle regulatory factor MRF4 during somite and skeletal myofiber development.**

    Dev Biol 1991, **147:**144-156. PubMed Abstract | Publisher Full Text

64. Fullwood MJ, Ruan Y: **ChIP-based methods for the identification of long-range chromatin interactions.**

    J Cell Biochem 2009, **107:**30-39. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

65. Olson EN, Perry M, Schulz RA: **Regulation of muscle differentiation by the MEF2 family of MADS box transcription factors.**

    Dev Biol 1995, **172:**2-14. PubMed Abstract | Publisher Full Text

66. Morisaki T, Sermsuvitayawong K, Byun SH, Matsuda Y, Hidaka K, Morisaki H, Mukai T: **Mouse Mef2b gene: unique member of MEF2 gene family.**

    J Biochem 1997, **122:**939-946. PubMed Abstract | Publisher Full Text

67. Molkentin JD, Firulli AB, Black BL, Martin JF, Hustad CM, Copeland N, Jenkins N, Lyons G, Olson EN: **MEF2B is a potent transactivator expressed in early myogenic lineages.**

    Mol Cell Biol 1996, **16:**3814-3824. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

68. Chang PS, Li L, McAnally J, Olson EN: **Muscle specificity encoded by specific serum response factor-binding sites.**

J Biol Chem 2001, **276:**17206-17212. PubMed Abstract | Publisher Full Text

69. Cserjesi P, Lilly B, Bryson L, Wang Y, Sassoon DA, Olson EN: **MHox: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer.**

Development 1992, **115:**1087-1101. PubMed Abstract | Publisher Full Text

70. Cserjesi P, Lilly B, Hinkley C, Perry M, Olson EN: **Homeodomain protein MHox and MADS protein myocyte enhancer-binding factor-2 converge on a common element in the muscle creatine kinase enhancer.**

J Biol Chem 1994, **269:**16740-16745. PubMed Abstract | Publisher Full Text

71. Gossett LA, Kelvin DJ, Sternberg EA, Olson EN: **A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes.**

Mol Cell Biol 1989, **9:**5022-5033. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

72. Amacher SL: *Myocardiocyte regulatory elements and trans-acting factors of the mouse muscle creatine kinase gene.* University of Washington, Biochemistry; 1993.

73. Tsika RW, Schramm C, Simmer G, Fitzsimons DP, Moss RL, Ji J: **Overexpression of TEAD-1 in transgenic mouse striated muscles produces a slower skeletal muscle contractile phenotype.**

J Biol Chem 2008, **283:**36154-36167. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

74. Watchko JF, Daood MJ, LaBella JJ: **Creatine kinase activity in rat skeletal muscle relates to myosin phenotype during development.**

Pediatr Res 1996, **40:**53-58. PubMed Abstract | Publisher Full Text

75. Ventura-Clapier R, Kuznetsov AV, d'Albis A, van Deursen J, Wieringa B, Veksler VI: **Muscle creatine kinase-deficient mice. I. Alterations in myofibrillar function.**

J Biol Chem 1995, **270:**19914-19920. PubMed Abstract | Publisher Full Text

76. Park SK, Gunawan AM, Scheffler TL, Grant AL, Gerrard DE: **Myosin heavy chain isoform content and energy metabolism can be uncoupled in pig skeletal muscle.**

J Anim Sci 2009, **87:**522-531. PubMed Abstract | Publisher Full Text

77. Spitz F, Salminen M, Demignon J, Kahn A, Daegelen D, Maire P: **A combination of MEF3 and NFI proteins activates transcription in a subset of fast-twitch muscles.**

Mol Cell Biol 1997, **17:**656-666. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

78. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al*.: **Clustal W and Clustal × version 2.0.**

Bioinformatics 2007, **23:**2947-2948. PubMed Abstract | Publisher Full Text

79. Clegg CH, Linkhart TA, Olwin BB, Hauschka SD: **Growth factor control of skeletal muscle differentiation: commitment to terminal differentiation occurs in G1 phase and is repressed by fibroblast growth factor.**

J Cell Biol 1987, **105:**949-956. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

80. Nelson JD, Denisenko O, Sova P, Bomsztyk K: **Fast chromatin immunoprecipitation assay.**

Nucleic Acids Res 2006, **34:**e2. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

81. Dignam JD, Lebovitz RM, Roeder RG: **Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.**

Nucleic Acids Res 1983, **11:**1475-1489. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

82. Bradford MM: **A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.**

Anal Biochem 1976, **72:**248-254. PubMed Abstract | Publisher Full Text

83. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.**

Science 2007, **316:**1497-1502. PubMed Abstract | Publisher Full Text

84. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.**

Nat Biotechnol 2010, **28:**511-515. PubMed Abstract | Publisher Full Text

85. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.**

Nat Methods 2008, **5:**621-628. PubMed Abstract | Publisher Full Text

86. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).**

Genome Biol 2008, **9:**R137. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text

87. Sambrook JFE, Maniatis T: *Molecular Cloning: A Laboratory Manual*. 2nd edition. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1989.

Chapter Supplemental II

High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation

Anil Ozdemir*, Katherine I Fisher-Aylor*, et al. (2011), Genome Research **21**(4): 566-577

SII.1 Abstract

Cis-regulatory modules (CRMs) function by binding sequence specific transcription factors, but the relationship between in vivo physical binding and the regulatory capacity of factor-bound DNA elements remains uncertain. We investigate this relationship for the well-studied Twist factor in *Drosophila melanogaster* embryos by analyzing genome-wide factor occupancy and testing the functional significance of Twist occupied regions and motifs within regions. Twist ChIP-seq data efficiently identified previously studied Twist-dependent CRMs and robustly predicted new CRM activity in transgenesis, with newly identified Twist-occupied regions supporting diverse spatiotemporal patterns (>74% positive, n = 31). Some, but not all, candidate CRMs require Twist for proper expression in the embryo. The Twist motifs most favored in genome ChIP data (in vivo) differed from those most favored by Systematic Evolution of Ligands by EXponential enrichment (SELEX) (in vitro). Furthermore, the majority of ChIP-seq signals could be parsimoniously explained by a CABVTG motif located within 50 bp of the ChIP summit and, of these, CACATG was most prevalent. Mutagenesis experiments demonstrated that different Twist E-box motif types are not fully interchangeable, suggesting that the ChIP-derived consensus (CABVTG) includes sites having distinct regulatory outputs. Further analysis of position, frequency of occurrence, and sequence conservation revealed significant enrichment and conservation of

CABVTG E-box motifs near Twist ChIP-seq signal summits, preferential conservation of ±150 bp surrounding Twist occupied summits, and enrichment of GA- and CA-repeat sequences near Twist occupied summits. Our results show that high resolution in vivo occupancy data can be used to drive efficient discovery and dissection of global and local *cis*-regulatory logic.

SII.2 Background

In animal genomes, *cis*-acting regulatory modules (CRMs) average ~300–500 bp in size and typically contain one or more binding motif instances for several transcription factors (Davidson 2006). DNA binding motif instances can now be readily mapped *in silico* by similarity to a consensus binding motif that has been defined through in vitro methods, or they can be derived from careful functional dissection of a few well-studied CRMs. However, many transcription factors recognize short sequence motifs that occur so frequently in the genome that virtually all gene loci have one or more, raising questions about which of these sites is occupied in the cell and what regulatory impact that occupancy has. We also know that binding motifs in the best-studied CRMs are often clustered (e.g., Ip et al. 1992a; Small et al. 1992; Berman et al. 2002; Markstein et al. 2002), presumably to facilitate coordinated and cooperative interaction among factors and cofactors and to achieve specificity relative to isolated single motif occurrences. However, we do not yet understand the logic by which motif combinations specify the functional output of the vast majority of CRMs in the genome (e.g., Lusk and Eisen 2010), and efficient identification and analysis of many more CRMs are needed to uncover these principles.

Advances in identifying candidate CRMs are coming from whole-genome approaches in which either chromatin immunoprecipitation (ChIP) is employed to find the region of DNA bound by a given transcription factor in vivo (e.g., Zeitlinger et al. 2007;

Zinzen et al. 2009), or high-throughput screening assays are utilized to identify promoter and CRM functions (e.g., Landolin et al. 2010; Nam et al. 2010), although the latter have not yet been widely applied. Global ChIP assays also allow one to define de novo or refine binding motifs used by a factor in vivo and to compare this with in vitro defined motifs. ChIP-seq is a particular form of genome-wide chromatin immunoprecipitation, which can produce high positional resolution of observable DNA binding in vivo (Johnson et al. 2007). In particular, the resolution of ChIP-seq data can be used to infer, within a given binding region, which specific motif occurrence is likely to account for the majority of the observed ChIP signal (Valouev et al. 2008). We refer to the motif instances most likely to drive observed binding as candidate “explanatory” sites, and we explore the value of making explanatory site models for all ChIP signals to guide detailed functional assays.

We apply ChIP-seq and ChIP-chip analyses to Twist, a key transcription factor in the dorsal-ventral (DV) patterning network of the *Drosophila* early embryo. Patterning the DV axis depends partly on Twist, a bHLH transcription factor present at high levels in ventral regions of the embryo (for review, see Chopra and Levine 2009; Reeves and Stathopoulos 2009). Many previous studies have contributed to the current picture of a developmental gene network that describes embryonic DV patterning, in which more than 50 genes and 30 CRMs have been linked (for review, see Stathopoulos and Levine 2005). Previous published ChIP-chip studies conducted using Twist antibodies have demonstrated that its occupancy can be detected in vivo (Sandmann et al. 2007; Zeitlinger et al. 2007). Our goals are to relate the global Twist occupancy pattern to functional CRM activity, as assayed by transgenesis, and to relate the local ChIP-seq profile to specific motif instances and combinations and their contribution to individual CRM activity.

SII.3 Results

SII.3.1 Comparison of ChIP-chip and ChIP-seq in the identification of CRMs

We performed ChIP-chip and ChIP-seq analysis on sheared chromatin isolated from *Drosophila* embryos from 1 to 3 h in age, using an antibody that is specific to Twist protein, and subsequently assessed the overlap between sets of regions identified by each approach (see Supplemental Fig. 1A–C and Methods). For ChIP-chip, we used a script to call peaks based on a minimum signal score, whereas for ChIP-seq, we used the ERANGE software suite to call peaks based on the number, orientation, and ratio of short sequence reads relative to a background control. The results from these methods were compared at several sensitivity thresholds to accommodate different numbers of peaks called by their informatics pipelines (Supplemental Fig. 1D). Given the substantial technical and computational differences between ChIP-chip and ChIP-seq, the fact that the vast majority of ChIP-seq signals overlap with some ChIP-chip regions lends mutual confidence, although a large number of ChIP-chip sites lacked support from ChIP-seq. Inspection of multiple ChIP-seq regions for which Twist activity was previously studied in detail showed that ChIP-seq regions are generally better resolved and provide superior guidance for experimental tests of function that are the central focus of this study (Supplemental Table 1).

SII.3.2 Functional analysis of Twist-occupied regions

We quantified how frequently and strongly ChIP-seq regions function as CRMs at the same time and place in development as the ChIP assays. To first identify a set of known gold-standard Twist CRMs, we applied a conservative standard that allowed only CRMs having prior genetic and molecular evidence. Enhancers (i.e., CRMs supporting gene expression rather than acting as silencers) along the DV axis were categorized as three types: Type I (ventral regions), Type II (ventro-lateral regions), and Type III (dorsal-

lateral and dorsal regions) (Supplemental Table 2B; for review, see Chopra and Levine 2009; Reeves and Stathopoulos 2009). Many enhancers of Types I and II require Twist for expression based on genetic and molecular genetic evidence, but not until recent ChIP-chip analyses was it thought that Twist might function to regulate Type III patterns (Zeitlinger et al. 2007). We observed very strong ChIP signals at *sog* and *brk* Type III CRMs but not at *ind, dpp, zen*, and *tld* (Supplemental Table 2B; Supplemental Fig. 2). When only Type I and II CRMs were considered, 11 of 15 were present in our medium confidence (MC) data set (see Methods). Known CRMs for the four not present (i.e., *Ady43A*, *phm*, *E(spl)*, and *wntD*) had below-threshold or no Twist ChIP-seq signal. The threshold for calling peaks could, of course, be reduced in order to recapture some (e.g., *wntD* and *phm*), but at the expense of increasing the false positive rate. Taken at face value, this gold standard comparison suggests we miss ~25% of true positives at the threshold selected.

Next, we tested 31 new candidate Twist CRMs drawn from the entire ChIP-seq set in a standard reporter gene assay (see Supplemental Table 2A). Of the 31 test regions, 23 (74%) supported expression; 21 supported expression in a classic dorso-ventral pattern or a subregion thereof, and 2 supported distinct patterns (i.e., ubiquitous or purely anterior-posterior) (Supplemental Fig. 3). The 23 new CRMs were distributed throughout the ChIP-seq signal range (Supplemental Fig. 2, “Positive signal”). Peaks near genes *Cyp310a1*, *Traf4, mirror (mirr)*, and *Mef2* were clearly defined by the ChIP-seq data, while the equivalent ChIP-chip data in these regions was much broader and, in some cases, gave multiple peaks, making the location of a candidate CRM ambiguous (see Fig. 1A–D). While Twist ChIP-seq data led to a high recovery rate of CRM detection, surprisingly, only ~25% of the associated genes including *Cyp310a1*, *Asph*, and *emc* (i.e., 3 of 12 assayed) actually required Twist to support expression in embryos.

For instance, *mirr*, *Traf4*, and *Mef2* expression was unaffected in *twist* mutants, even though their Twist-ChIP-seq signals were equally prominent and numerous (data not shown; see Discussion).

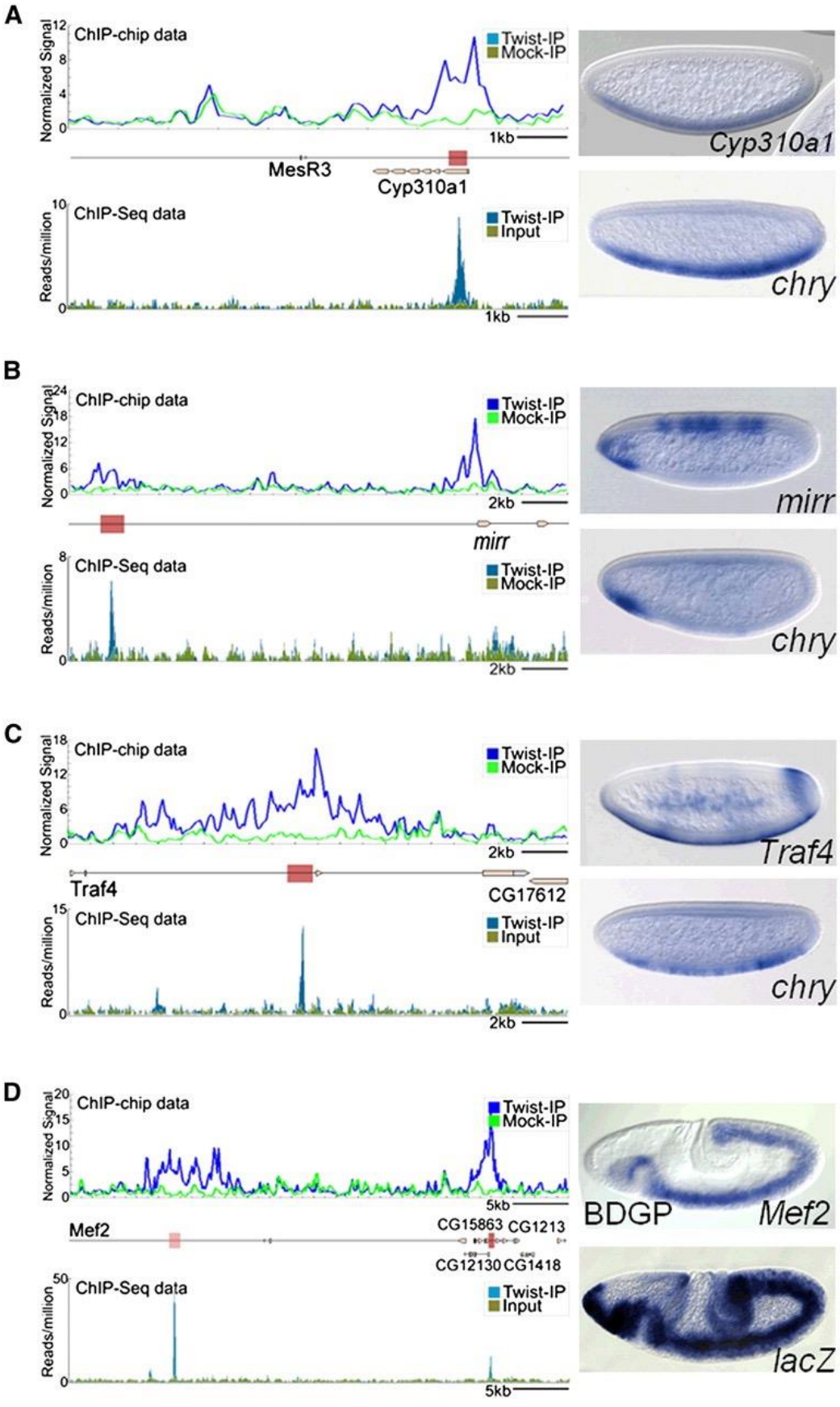

A
ChIP-chip data
Normalized Signal
Twist-IP
Mock-IP
1kb
MesR3
Cyp310a1
ChIP-Seq data
Reads/million
Twist-IP
Input
1kb
Cyp310a1
chry
B
ChIP-chip data
Normalized Signal
Twist-IP
Mock-IP
2kb
mirr
ChIP-Seq data
Reads/million
Twist-IP
Mock-IP
2kb
mirr
chry
C
ChIP-chip data
Normalized Signal
Twist-IP
Mock-IP
2kb
Traf4
CG17612
ChIP-Seq data
Reads/million
Twist-IP
Input
2kb
Traf4
chry
D
ChIP-chip data
Normalized Signal
Twist-IP
Mock-IP
5kb
Mef2
CG15863 CG1213
CG12130 CG1418
ChIP-Seq data
Reads/million
Twist-IP
Input
5kb
BDGP
Mef2
lacZ

Figure 1. *In vivo Twist occupancy supported by Twist ChIP-seq identifies functional CRMs. Representative examples of newly identified enhancers (brown boxes) and those previously identified (pink boxes) are shown for* Cyp310a1 *(*A*),* mirr *(*B*),* Traf4 *(*C*), and* Mef2 *(*D*).* Upper left *panels show ChIP-chip data and* lower left *panels show ChIP-seq data for Twist-IP and control samples. In* upper right *panels, lateral views of whole mount in situ hybridizations of the endogenous genes of stage 5–8 embryos are shown. In* lower right *panels, lateral views of whole mount in situ hybridizations of similar staged embryos containing either* cherry *(for* Traf4*,* mirr*, and* Cyp310a1 *enhancers) or* lacZ *(for* Mef2 *5′ enhancer) reporter constructs.*

SII.3.3 Twist recognition motifs in vivo and in vitro

Twist belongs to a large bHLH family of DNA-binding factors that recognize a core DNA consensus, CANNTG, called an E-box (for review, see Massari and Murre 2000). Prior work using in vitro and in vivo approaches highlighted a subfamily preferred by Twist, led by CATATG (i.e., TA E-box). We asked which, if any, of the 10 possible E-box recognition motifs (counting reverse complements as the same motif) are selectively concentrated within 50 bp of called ChIP-seq signal summits (Fig. 2A). We found that CA and GA core E-boxes were most prominent, while GC, TA, and CG were relatively minor (Fig. 2A, “Twist ChIP-seq”). Compared with regions sampled from ChIP-seq control data or from the entire non-repeat genome, only CA, TA, CG, and GA core E-boxes were statistically enriched in Twist-occupied regions (Fig. 2A, colored slices). When larger radii from the ChIP signal summits were interrogated, the number of E-boxes of all types increased, and the specific enrichment trend was less apparent (i.e., enrichment of CA, TA, CG, and GA core E-boxes). In contrast, when ChIP-chip regions were similarly examined (Supplemental Figs. 5, 6), no specific enrichment of any motif was detected at any radius from the called Twist peaks. Overall, the enrichment and resolution results suggest that the ChIP-seq data could be used to model individual binding domains and causal motif instances in them (see below).

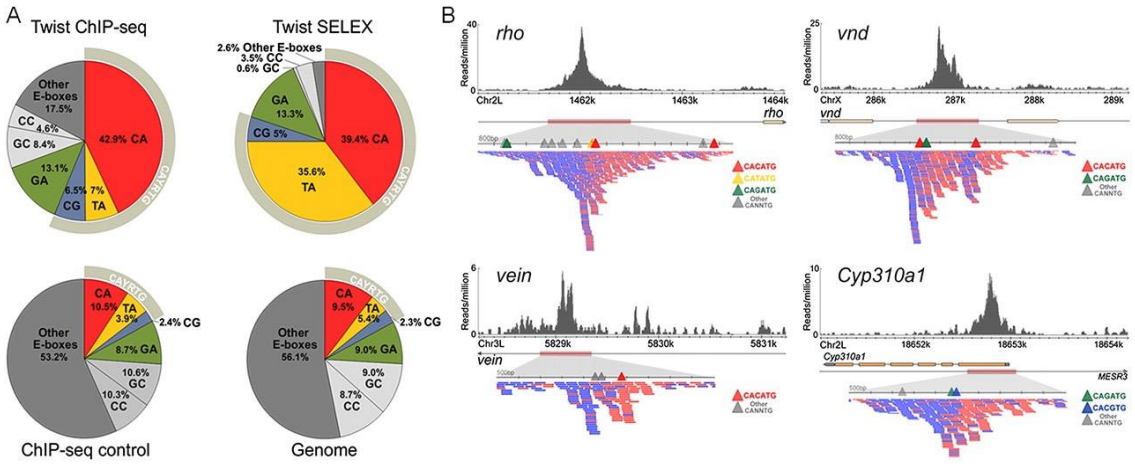


Figure 2. *A comparison of Twist in vivo and in vitro binding preferences. (*A*) The frequency of E-boxes associated with HC twist peaks (±50 bp), SELEX-bound sequences, ChIP-seq enriched control regions (±50 bp of summits), and the non-repeat dm3 genome was calculated. (*B*) Twist ChIP-seq data in the vicinity of CRMs shown to support expression of the genes* rho *(Ip et al. 1992b),* vnd *(Stathopoulos et al. 2002),* vein *(Markstein et al. 2004), and* Cyp310a1 *(this work). The directionality within ChIP-seq sequencing reads points to the position of the "explanatory" site. Blue and red ticks symbolize individual sequencing reads acquired, which match either the Watson or Crick strand.*

Previously published foot-printing data and small-scale SELEX had found that the in vitro Twist protein binding consensus is CAYRTG (i.e., core E-box residues YR = TA, CG, and CA) (Ip et al. 1992b; Zinzen et al. 2006). To test how Twist in vivo binding results relate to in vitro preferences, we determined E-box frequencies in high-throughput Twist SELEX data, and compared them with our ChIP-seq data (see Supplemental Text). For the most part, the same E-boxes were highlighted, except that the TA-core E-box motif, which was the most highly bound by Twist in vitro (35.6% occupancy by SELEX), was less enriched in vivo (7% by ChIP-seq versus 5.3% frequency in the genome). A simple explanation is that there are real differences between the in vivo and in vitro binding conditions that affect Twist motif preference.

Among alternative explanations, one or more species of bHLH heterodimers might be acting in vivo, while only homodimers were assayed in vitro (see Discussion).

SII.3.4 Motif composition of Twist ChIP-seq regions

We examined the positions of all E-box motifs within Twist-ChIP-seq regions (Fig. 2B). The ChIP-seq protocol used here is a standard Illumina platform one that retains information about whether a sequenced fragment end originated from the Watson (red) or Crick (blue) strand (Fig. 2B; Valouev et al. 2008). With appropriate data preprocessing to account for fragment length (for review, see Pepke et al. 2009, see Methods), the summit location within each peak region can be identified computationally. Inspection of known Twist CRMs showed that this agrees well with, on average, 1–2 dominant binding motif instances within ±50 bp (e.g., see Fig. 2B). A subset of previously known Twist-bound regions consists of multiple peaks aggregated together, and these are typically associated with multiple Twist motifs (e.g., see Fig. 2B, *vnd*).

We mapped and visualized the position of each motif instance relative to its peak summit and calculated the cumulative frequency for each motif type as a function of distance from the peak (Fig. 3). Within the top ranked ~1000 peaks the concentration of CAYRTG motifs was stronger than in lower ranked peaks, with CACATG sites, rather than CACGTG and CATATG, being most prominent near peak summits (Fig. 3B, top). Several criteria, including manual inspection of peaks throughout the ranking and the presence of previously studied Twist-dependent CRMs, led us to define a high confidence (HC) threshold of 513 regions (FDR 1%; see Methods and Supplemental Text); however we also found that binding motif centrality extends to ~1000 sites in the genome, and for most analyses we use this more inclusive set of ~1000 medium confidence (MC) calls (FDR 17%).

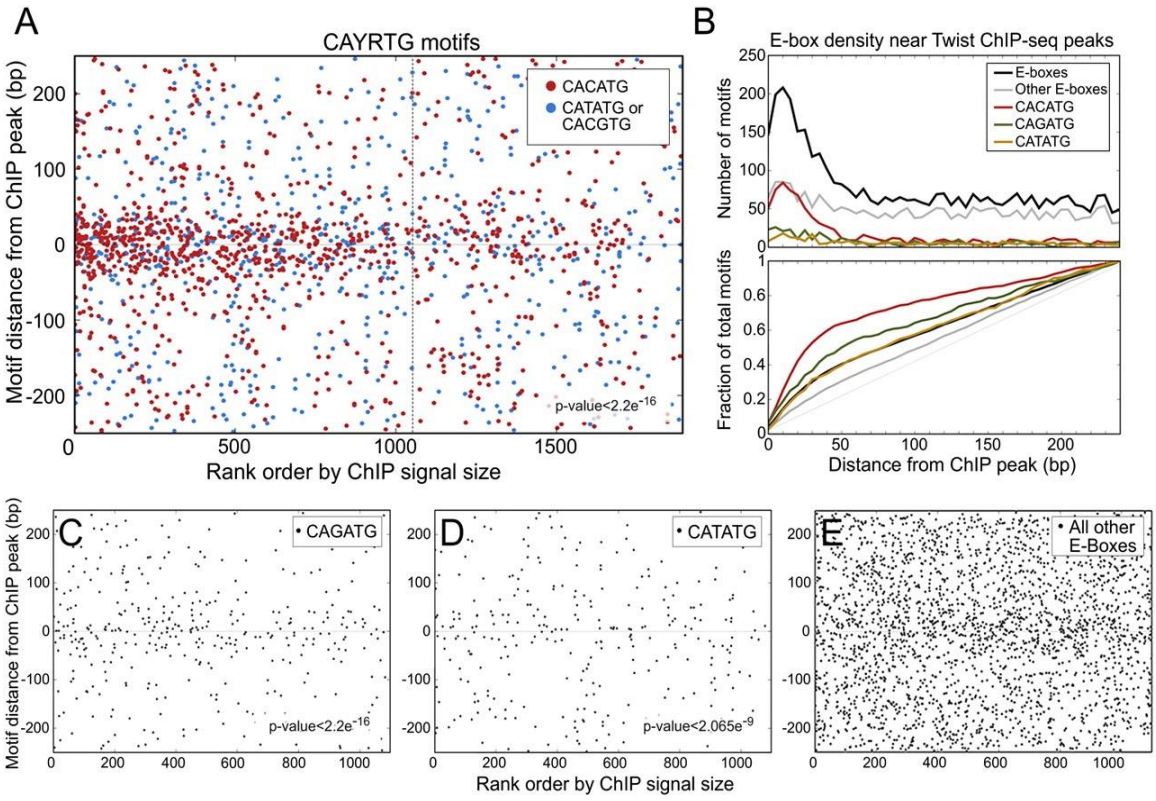


Figure 3. *Motif composition of Twist ChIP-seq regions shows preferential concentration of specific E-boxes near summits. (*A*) Locations of CA*YR*TG = CAC*A*TG CAT*A*TG and CACGTG E-box instances located within ±250 bp of the ChIP-seq peak (ERANGE-shifted called signal summit; see Methods) (*y *axis), plotted as a function of signal intensity rank from highest (1) to lowest (2000) (*x *axis). 1099 MC ChIP-seq data set is shown with a dashed line. CA*C*A*T*G is the most prevalent E-box motif in Twist ChIP regions and it shows the strongest central concentration. (*B*) Direct (*top *panel) and cumulative (*bottom *panel) motif density plots. In the MC data set, 65% of CA*C*A*T*G motifs and 50% of CA*G*A*T*G occur within ±50 bp of Twist peaks. (*C*) CA*G*A*T*G occurs more frequently in Twist ChIP-seq regions and is more centrally localized than (*D*). (*D*) CA*T*A*T*G is the motif most prominent in SELEX data (see text). (*E*) Other E-boxes (defined here as CANNTG motifs where NN is neither CA, GA, nor TA) display a more uniform distribution (*B,E*), though the other CABVTG E-boxes not pictured here (CG, GC, and CC) provide a minor central enrichment (see Supplemental Fig. 8). The number and distribution of explanatory E-boxes changes with ChIP-seq signal strength, suggesting that more E-boxes create a more robust Twist ChIP signal (*A*; Supplemental Fig. 7).*

The accumulation of motif instances as a function of distance from the summit, over the entire set of Twist ChIP-seq regions, was analyzed (Fig. 3B, bottom). Using the K-S test, the *P*-value for CACATG distribution was defined as $<2.2 \times 10^{-16}$ (D = +0.44), meaning that the observed enrichment of CACATG near the peak summit is non-random and highly significant. It suggests that the CA-containing E-box drives in vivo binding at the majority of sites we called. Five other E-boxes also are enriched near summits, though they are less frequent in comparison to CACATG (Fig. 3B, top; Supplemental Fig. 8; Supplemental Table 3). In addition, the highest ranking peaks are associated with 2 or more matches to E-boxes; in particular the CACATG site is prominent (see Supplemental Fig.9).

SII.3.5 CACATG and CATATG motifs are not functionally synonymous

For many ChIP regions, detailed inspection of the primary data displayed in browser format confirms a single explanatory motif (e.g., *vein* CRM, Fig. 2B; Supplemental Fig. 10). However, some CRMs contain two or more closely spaced sites matching the CABVTG consensus, leading us to ask how closely positioned E-boxes interact. The *rho* early embryonic enhancer is such a case, with a highly directional single peak with two E-boxes sites (CATATG, T1, and CACATG, T2) separated by only 5 bp (Fig. 4A). We tested whether a series of enhancer constructs support expression in the lateral domain of the embryo, comparing the wild type CRM with Twist motif mutants.

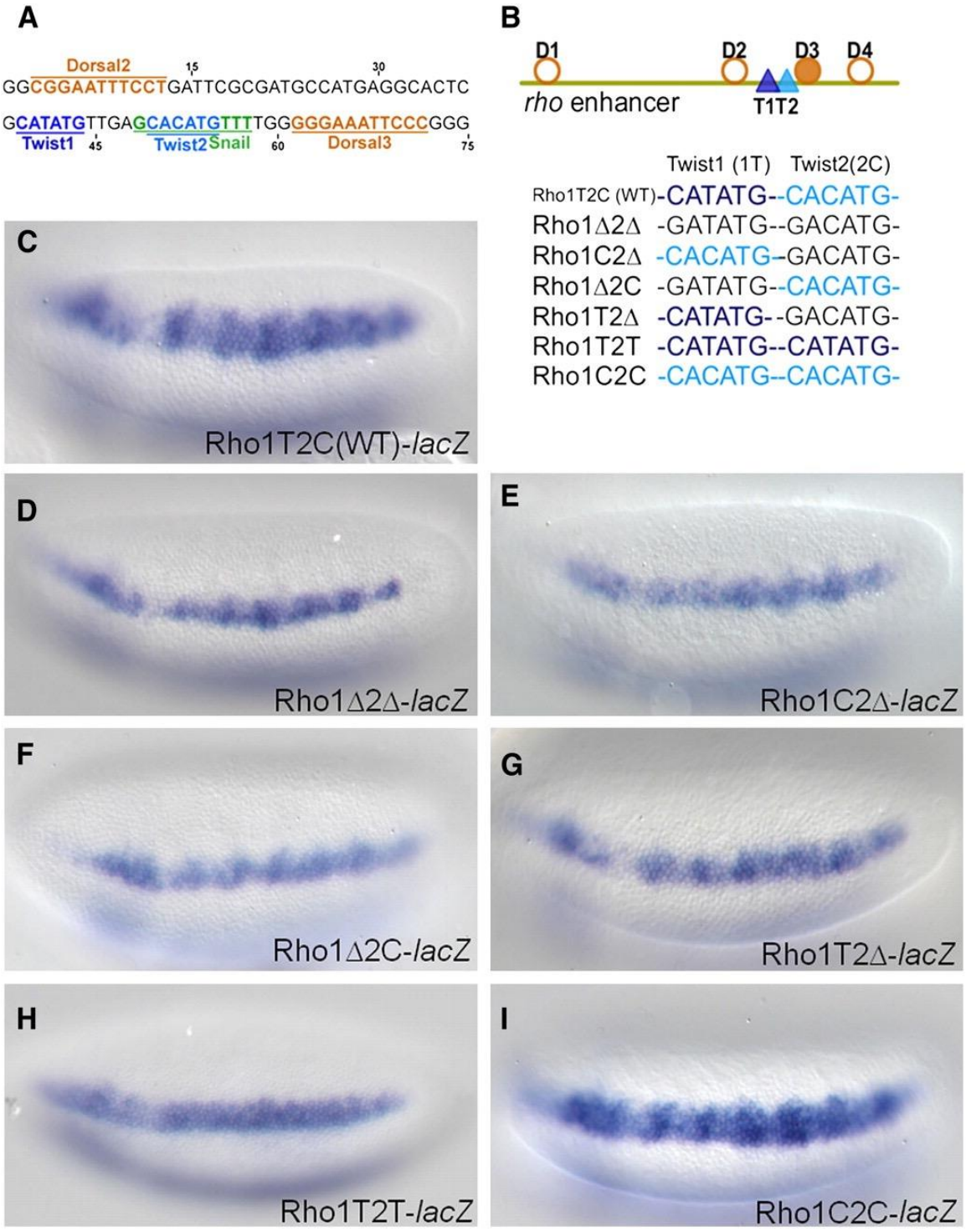


Figure 4. *Mutagenesis of Twist binding sites at the ChIP-seq peak summit of* rho *enhancer. (*A*) The 75 bp sequence from the* rho *minimal enhancer which contains binding sites for Twist as well as for the transcription factors Dorsal and Snail. E-box*

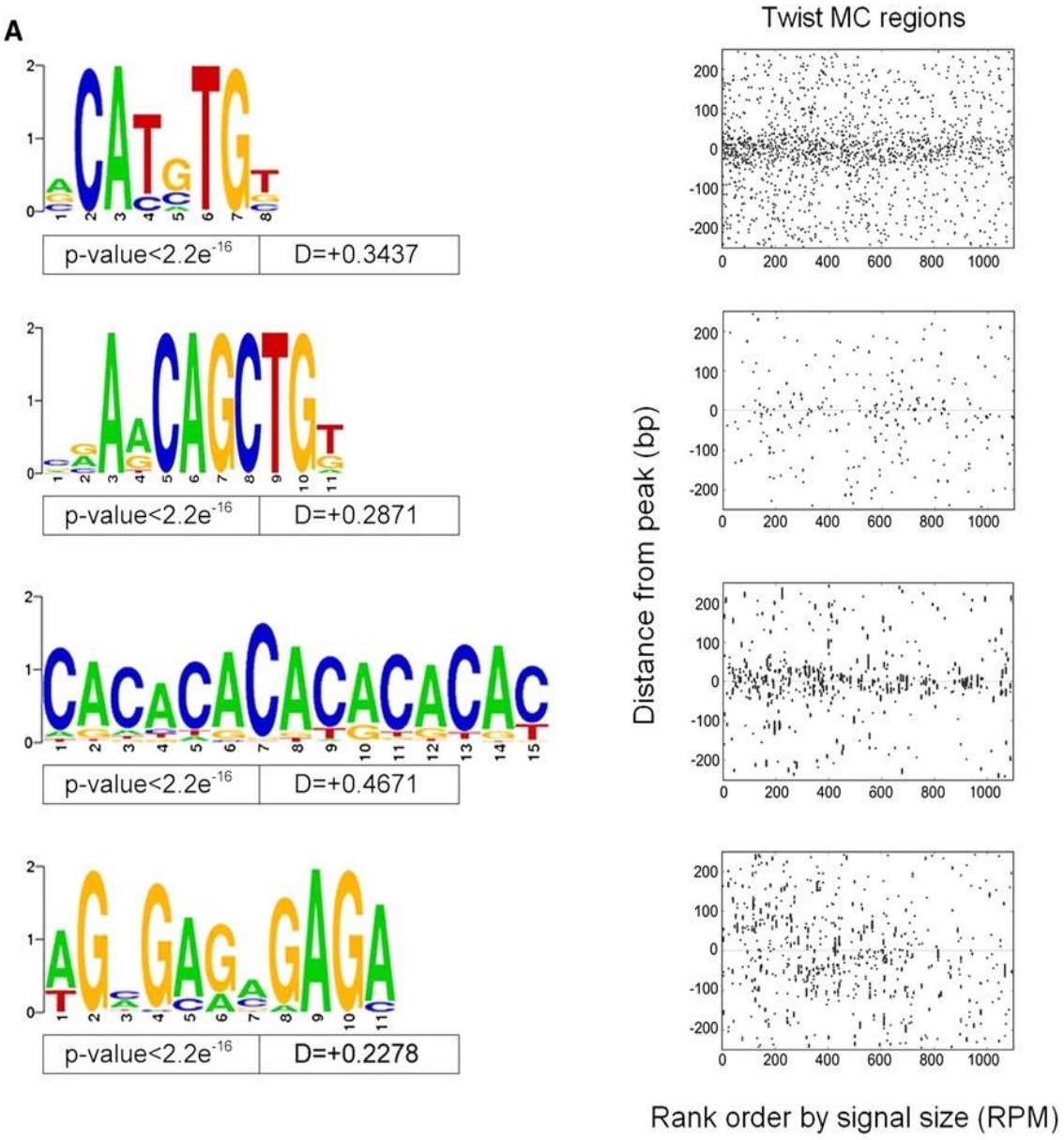
*sequences CA*TA*TG (T1, dark blue) and CA*CA*TG (T2, light blue) are separated by 5 bp, and Dorsal binding sites (orange) are positioned upstream and downstream of Twist sites. A Snail site that overlaps with T2 E-box is shown in green. (*B*) A diagram of the minimal 299 bp* rho *enhancer showing the relative positions of sites for Twist (dark and light blue triangles) and Dorsal (orange circles and filled circles, showing non-canonical and canonical sites, respectively). Lower schematic shows color-coded representations of the WT or mutant Twist binding sites present in various reporter constructs. Single nucleotide mutations were introduced into either T1 or T2 to eliminate binding (black: C*ATATG*>G*ATATG or C*ACATG*>G*ACATG) or to convert one site to the other (light blue: CA*TATG*>CA*CATG or dark blue: CA*CATG*>CA*TATG). (*C*) In situ staining of the wild type construct, minimal* rho *enhancer attached to the* evep.lacZ *reporter. (*D*) The Rho1Δ2Δ double mutant containing point mutations in both of the E-boxes, T1 and T2, supports reporter gene expression that is significantly weakened and more narrow compared to wild type (*C*). (*E–G*) Single mutations support expression that is weaker than wild type (*C*), more similar to the double mutant (*D*). (*H*) When a CATATG E-box is present in both the T1 and T2 positions, this change severely affects the expression domain of the reporter gene, reducing it to levels comparable to those observed in the double mutant Rho1Δ2Δ embryos (*D*). (*I*) When a CACATG E-box is present in both the T1 and T2 positions, the expression supported is comparable to the wild type (*C*).*

Within the *rho* enhancer sequence, we introduced single-nucleotide changes to sites T1 and T2 (CANNTG→GANNTG). These subtle changes abrogated expression, such that instead of supporting expression in a wide domain (~6–8 cells), the mutant enhancer supports expression in a more narrow domain (~3–4 cells) (cf. Fig. 4D,C); this result is comparable to what others have found previously with more severe changes to the T1 and T2 E-box sequence (5 or more changes per site; Ip et al. 1992c). We also found that mutation of either site alone supported reporter gene expression, but neither was as severe as eliminating both (cf. Fig. 4E,F,G and 4C,D). This suggested that Twist binding to both T1 and T2 sites contributes to *rho* expression.

We then asked whether CA and TA E-boxes are interchangeable. When T1 and T2 are both CACATG (i.e., T1 site TA-core was converted into CA-core), reporter expression was comparable to wild type (Fig. 4I). In contrast, replacement of both sites by the CATATG was not sufficient to support expression over the full spatial domain (Fig. 4H); in fact, expression was comparable to the T2 mutant (Fig. 4G). This suggests that the CA E-box can function in both positions, while the TA E-box can function in T1 but not T2.

SII.3.6 Motif discovery in Twist ChIP-seq regions

To uncover possible alternative Twist binding motifs or co-associated motifs for Twist-interacting factors, we used MEME, a motif discovery tool (Bailey et al. 2006), to search for statistically overrepresented motifs in and near Twist-occupied regions. As expected, prominent motifs found by MEME were E-box sequences (Fig. 5A) that overlap with CABVTG defined by our previous analyses (Fig. 3). In addition, MEME output highlighted residues flanking the E-box, such that a leading-A or lagging-T residue is preferred [e.g., CACATG-T (A-CATGTG) or A-CACATG (CATGTG-T)]. In contrast, a lagging A was very rare in Twist regions and in the genome at large (Fig. 5A). Other in vitro and in vivo bHLH binding studies support the idea that flanking bases may influence bHLH DNA binding (Grove et al. 2009; Cao et al. 2010).


A
Twist MC regions
p-value<2.2e-16
D=+0.3437
p-value<2.2e-16
D=+0.2871
p-value<2.2e-16
D=+0.4671
p-value<2.2e-16
D=+0.2278
Distance from peak (bp)
Rank order by signal size (RPM)

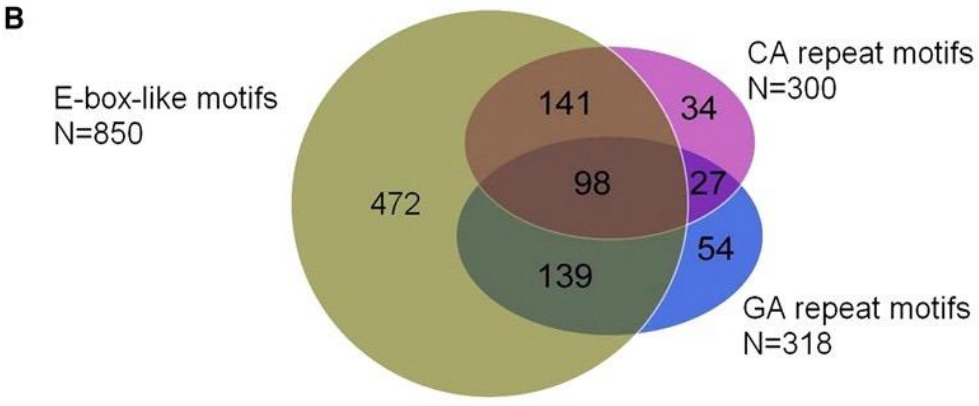

B
E-box-like motifs
N=850
CA repeat motifs
N=300
GA repeat motifs
N=318
141
34
98
27
472
54
139

Figure 5. *Motifs associated with Twist in vivo occupancy identified using MEME. MEME was run on the narrow 50 bp region surrounding each of the 1099 MC ChIP-seq peaks to identify all motifs that are enriched near the point of Twist occupancy. These motifs were mapped back to determine their spatial distribution relative to Twist peaks, and some motifs showing a non-uniform distribution near Twist peaks were selected. (*A*) Variations on CAYRTG and CAGCTG were returned, together specifying CABVTG (*top *two Weblogos). Note that a leading A residue or a lagging T residue is also suggested, which appears preferred by other non-Twist family DNA-binding bHLH factors (K Fisher-Aylor, S Kuntz, and A Kirilusha, unpubl. obs.; Grove et al. 2009). In addition, two simple repetitive sequences (CA and GA) are also significantly enriched at Twist-occupied sites (*bottom *two Weblogos). (*B*) Venn diagram illustrating the relationship between sets of peaks defined as having at least one occurrence of (i) either of the two E-box-like motifs; (ii) the CA-repeat-like sequence; or (iii) the GA-repeat-like sequence.*

Several “simple” repeat sequences were significantly overrepresented in the Twist-occupied regions: the predominant one was a CA-repeat, and a similar GA-repeat sequence was also found (Fig. 5A). Of the 1099 peaks comprising the MC Twist ChIP-seq data set, 850 contain at least one match to either major E-box in the wide area around the peak (±250 bp), and 378 of these (or 44%) also contain at least one CA- or GA-repeat sequence (Fig. 5B). It is possible that the CA- and GA-repeats associated with Twist ChIP-seq peaks play some role in marking or phasing these regions as potentially “open chromatin,” as these same motifs were recently found associated with DNA occupied by Trithorax and Polycomb group/recruitment factors (see Schuettengruber and Cavalli 2009; and Discussion).

Interactions between Twist and other transcription factors might exist, yet not be identified by MEME for various reasons. We therefore tested additional motifs already known to bind transcription factors that pattern the DV axis in the early *Drosophila* embryo. Dorsal is a maternal transcription factor that functions cooperatively with Twist

at some well-studied, closely-spaced sites (e.g., Ip et al. 1992c; Erives and Levine 2004), but the generality of this pattern across other Twist bound regions is not known. We found no significant global correlation between Dorsal motif occurrences and Twist peaks in our data. Among other factors (i.e., Su(H), Zelda, RGGNCAG/unknown, and Snail), only Snail exhibited significant motif co-enrichment in Twist ChIP regions, while Su(H) and RGGNCAG exhibited weak enrichment. The Snail result is neither surprising nor definitive because this factor can bind a sequence similar to that of Twist (Supplemental Fig. 12). Snail is thought to function as a repressor, at least in part, by competitively inhibiting binding of Twist (e.g., Ip et al. 1992b). Perhaps binding of both Twist and Snail to CRMs through the CA-core E-box plays a role that is more widespread than previously appreciated (see Discussion).

Twist-occupied regions were preferentially and significantly concentrated in proximal promoters (Fig. 6A), relative to several control samples, while intronic and intergenic classes were not enriched. Twist regions were slightly, but not significantly, depleted in exons. We tested whether the Twist regions near promoters were, more frequently than any others, lacking an explanatory E-box. This would be expected if many Twist promoter ChIP signals resulted from capture of indirect looping interactions from distant Twist-bound CRMS (e.g., Fullwood and Ruan 2009), rather than from primary motif binding, but it was not observed (Fig. 6B). We also asked if specific E-box motifs are selectively associated with any specific gene region class. Explanatory motifs at promoters showed higher CAGCTG and CACGTG E-box content, relative to intronic and intergenic groups, and a reduction in the dominant CACATG motif (Fig. 6B; Supplemental Fig. 13). These trends were not due to similar changes in the frequencies of GC, CG, or CA dinucleotides in promoters genome-wide (Supplemental Fig. 13). Exons also had distinctive signatures, presumably due to protein coding constraints.

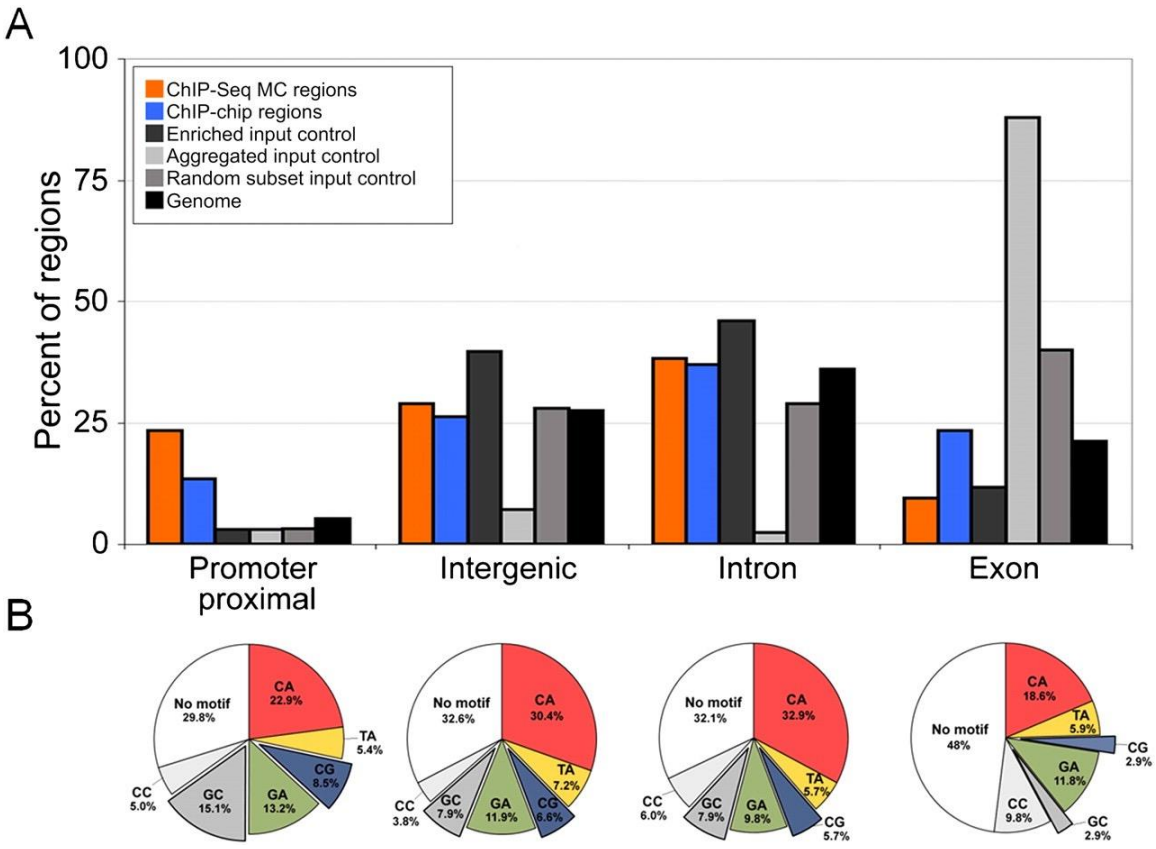


Figure 6. *Enrichment of Twist ChIP-seq summits and explanatory E-box motifs in different genic and intergenic locations.* (A) *Enrichment of Twist ChIP-seq and ChIP-chip summits at particular positions in the genome, relative to a genome random sample and several sequencing negative controls. The genome was segregated into four mutually exclusive categories: promoter proximal (relative to the set of promoters from S. Celniker, including 500 bp upstream), exonic, intronic, and intergenic (see Supplemental Methods). While the majority of Twist regions fall into intergenic and intronic regions, there is a significant overabundance of Twist peaks in promoters relative to the amount of promoters in the genome (24%, or 258 of the ChIP-seq peaks). Intergenic and intronic Twist occurrences are comparable to that expected from a random genomic sample (29%, or 319 intergenic, and 38%, or 420 intronic). The number of summits within exonic regions is relatively disenriched (9%, or 102). In order to assess these numbers compared to expected values, we also compared the same number of Twist ChIP-chip regions (largest by area), the input control DNA regions enriched over Twist, the aggregated input DNA, and a random sampling of sequenced reads mapping uniquely to the genome (see Supplemental Text). We also report the total amount of the genome*

*falling into each of these categories. The aggregated control and, to a lesser degree, the random control reads draw attention to the fact that there are many sequenced reads falling into exons. The enriched control does not show the exon bias perhaps because a directionality requirement was used; there is a mild enrichment of these sequences in the gene flanking category relative to the random genomic sample but a significant depletion in the promoter proximal that is likely due to the fact that Twist peaks are enriched at promoters. (*B*) The frequency of explanatory E-box sequences as a function of position of Twist-bound peaks in the genome (i.e., promoter proximal, intergenic, intronic, and exonic position). The CA, CG, and GA core E-boxes show enrichment in promoter, intergenic, and intronic positions; the GC core E-box is specifically enriched in the promoter proximal position.*

SII.3.7 Evolutionary conservation of ChIP-seq regions and motifs

Preferential sequence conservation is a signature of many biologically-significant regulatory regions and sequence motif instances. On average, our Twist-occupied regions were more conserved over a sequence domain of ~300 bp compared to random genomic background conservation (blue versus red trace, Fig. 7A). In the HC Twist ChIP-seq data set of 513 peaks, conservation was highest over the motif when regions were centered on the explanatory CABVTG instance, and conservation gradually dropped to background levels as a function of distance from the center (green versus blue trace, Fig. 7A). Slight preferential conservation is observed in the background control sequence when they are aligned using the same set of E-boxes (cyan versus red trace, Fig. 7A). This is consistent with E-boxes being targets of a large family of transcription factors that exhibit varying degrees of motif preference. Furthermore, this regional conservation was less prominent in lower ranked peaks, suggesting that the higher ranked peaks are more likely to be functional (see Supplemental Fig. 14).

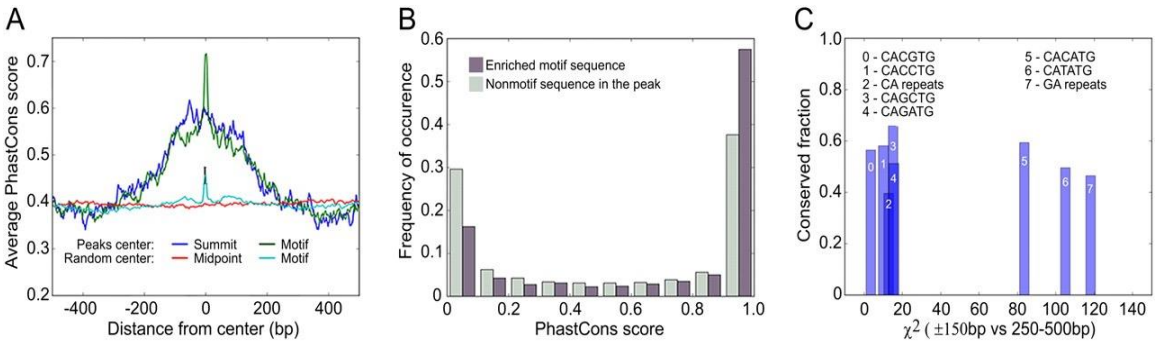


Figure 7. *Conservation analysis of sequences defined by Twist binding. (*A*) Averaged conservation profiles using phastCons scores for ChIP-seq regions and random genome samples. The blue curve shows average conservation in ChIP-seq peak regions is significantly elevated ±150–200 bp from the ChIP-seq signal summit. The green curve shows the same data but with regions recentered over the nearest CABVTG binding motif within 150 bp of the original summit. For the random sample, 500 regions containing one of the motifs were selected with the region start point selected at random for the uncentered distribution. Here "midpoint" refers to the location in the center of the randomly determined region. The error bar shows two standard deviations of 30 trials of 500 samples each. A maximum over the motifs is manifest, though substantially smaller than within the ChIP-seq peak regions. (*B*) Histogram of phastCons scores for bp occurring within the 6 E-box binding motif candidates (gray) compared to that for bp within the ChIP-seq regions, but outside any of the E-box motifs (black). Bp in the motif sites are found to be statistically more conserved than bp outside of motifs (0.005 significance level). (*C*) Fraction of sites in various sequence patterns falling within the top decile of phastCons scores for a 150 bp radius surrounding ChIP-seq summits versus the chi squared statistic for distributions within 150 bp of the summit compared to those of region 250–500 bp from the summit. CACATG, CATATG, and GA repeat sequences exhibit significantly greater conservation in ChIP-seq regions compared to flanking sequence than other motifs (as shown by their clustering at high values of the chi squared statistic), though CATATG and GA repeats do not exhibit high absolute levels of conservation.*

To assess conservation of E-box sites more quantitatively, we compared the distribution of phastCons scores for inferred Twist binding motifs in peak domains (±150 bp from the ChIP-seq summit) to those for other sequences in the same regions (Fig.

7B). E-box motifs were significantly more conserved than the rest of the domain, suggesting that they are more functionally relevant than the average sequence around them. This supports the view that E-boxes in proximity to detected peaks are not only "explanatory" for binding, but that many of these have some function in vivo. The function implied by conservation may or may not occur during the embryonic stage at which we have made our measurements, and it is even possible that some are conserved due to binding by a different bHLH factor during the life of the animal.

We examined the degree of conservation of individual E-boxes of interest relative to one another and to CA and GA repeats that were found to be prevalent in the ChIP-seq signals. We sought to distinguish those with functions associated specifically with the Twist-occupied CRMs by comparison to flanking sequence, by comparing the fraction of conserved (phastCons > 0.9) motif occurrences within ±150 bp of the ChIP-seq summit to those in flanking regions 250–500 bp away from the summit (Fig. 7C); the latter is assumed to be statistically equivalent to genomic background from data in Figure 6A. We find that CATATG, CACATG, and GA repeats stand out in terms of the change in conservation between peak and flanking sequences. In contrast, CAGATG, CACGTG, CACCTG, and CA repeats show minimal change between peak and non-peak sequences.

SII.4 Discussion

This analysis of in vivo Twist occupancy in the developing *Drosophila* embryo provides general and specific insights into relationships of Twist DNA binding motifs and in vivo Twist occupancy with regulatory function. We found that the in vivo consensus binding motif, as derived from Twist ChIP-seq data, is CABVTG (Figs. 2 and 5). Within that subfamily of E-boxes, CACATG is most prevalent within tested CRMs and is occupied preferentially within ChIP-seq defined peaks in general (Supplemental Tables

1 and 2; Fig. 3). Our detailed analysis of the *rho* enhancer showed that within the Twist-subfamily of E-boxes, individual members are not always interchangeable, and this suggests that they can support different functions (Fig. 4). When we searched for other motifs in addition to the E-box sequence that are associated with Twist peaks, we found that two repeat sequences, in particular, are associated with Twist ChIP-seq peaks, CA- and GA- repeat sequences, and that A/T-rich sequences are generally depleted from the region around ChIP signals (Supplemental Fig. 11). E-boxes and the over-represented motifs, in particular CACATG, CATATG, and a GA-repeat, are more conserved within peaks than background, suggesting that they have significant functions, presumably in transcriptional regulation.

We investigated the relationship between Twist occupancy and CRM regulatory activity by conducting functional tests and through analyses of conservation. Because the numbers of Twist-occupied sites we detected (500–1100) is large compared to the number of known Twist-regulated genes, it was not a foregone conclusion that most occupied regions would have any regulatory function. Our observed 74% CRM activity rate (23 positive CRMs of 31 tested) is high, and it argues that ChIP occupancy is efficiently highlighting functional regulatory DNA segments (Supplemental Table 2A); this analysis also captured the majority of gold standard enhancers identified by a number of previous studies (Supplemental Table 2B). Results showing preferential conservation of the Twist-bound cohort provide additional support for the idea that many other candidate regions that we did not test directly for function will also turn out to be CRMs.

A natural question is why the remaining ~25% did not score as active enhancers to support gene expression. Simple biological possibilities are that some Twist occupancy is not associated with any regulatory activity; that the module's regulatory activity is to silence or to insulate, rather than to enhance; that the module is bound but

is not active at this time in development (for review, see Levine and Tjian 2003; Arnosti and Kulkarni 2005; Gurudatta and Corces 2009; Cao et al. 2010). There are precedents for all these possibilities, although not all have been explicitly shown for Twist. Technical explanations are that CRM activity might not have been successfully captured in a segment tested, or that the original ChIP region calls include an unrecognized class of false positives.

Although our ChIP data efficiently identified CRMs, we emphasize that there is a distinction between significant in vivo Twist occupancy, as indicated by the ChIP-seq data, versus significant regulatory dependence on Twist, which appears to be rarer. Lower levels of regulatory dependency are, at present, difficult to measure, and they might be common. At the extreme, Twist-binding at most CRMs could be entirely opportunistic, arising by protein-protein interactions with other already bound factors and cofactors and/or binding to an E-box that has been made accessible by other unrelated factors nearby.

SII.4.2 Incongruity between in vivo and in vitro preferred motifs

Our findings suggest that the TA-core and CA-core E-boxes are similarly preferential for Twist binding in vitro, but in vivo the Twist ChIP-seq explanatory sites are enriched in CA-core E-boxes. If Twist protein sees CA and TA motifs similarly, then the in vivo preference might simply reflect general base composition. When we specifically tested for this, the magnitude of CA enrichment in Twist bound E-boxes was much larger than in the non-coding genome at large (Supplemental Fig. 13). Alternatively, bHLH proteins are known to form heterodimers in addition to homodimers, and an explanation for CA differences is that Twist binding detected in vivo is a combination of homo- and heterodimers (e.g., Murre et al. 1989). The enrichment of CA core E-boxes in vivo could reflect a particular Twist–bHLH heterodimer, since ChIP will, in principle, recover any

Twist-containing complex. In particular, there is some genetic interaction data that suggests that Twist and Daughterless (Da), a bHLH ubiquitously expressed in the embryo, may interact to affect patterning in the early embryo (Jiang et al. 1992; Gonzalez-Crespo and Levine 1993; Stathopoulos and Levine 2002). Other data with forced heterodimers showed that Twist can partner with Da at later stages to influence somatic mesoderm specification (Castanon et al. 2001). When we examined overlap between our Twist ChIP-seq binding events and that of Da ChIP-chip data available (Li et al. 2008), using relaxed criteria for overlap, we found 30% of our high confidence sites have some evidence for Da binding at the same locus. When the explanatory E-box instances for these regions from our data were interrogated, we found no positive correlation with CA core E-boxes and Da, but we did find a positive correlation with GC core E-boxes and possible Twist/Da co-occupancy (data not shown). Since other bHLH factors in the embryo might also partner with Twist, the specific role, if any, of heterodimers in this system will be speculative until the full partnering repertoire for Twist is quantified and characterized. It is also possible that post-translational modifications and local conditions in the nucleus that differ from the in vitro conditions affect DNA binding preferences.

Our mutagenesis experiments with the *rho* CRM further demonstrate that the TA-core and CA-core E-boxes are not equivalent, at least in some instances. What could be different about CA- versus TA-core E-boxes? CACATG and CATATG E-boxes (e.g., T1 and T2; see Fig. 4) were first identified as Twist-binding sites within the *rho* early embryonic enhancer in 1991 by Ip et al. (1992c) using in vitro footprinting. They showed that the CA-core E-box (but not TA-core) can also be bound by the repressor Snail. It is therefore possible that the preference we see for CA core E-boxes near ChIP-seq peaks indicates that Twist/Snail combined sites have been favorably selected, and that this

combination site has a distinct role in regulating the activity of many CRMs in the early embryo. In 2002, the CA-core E-box was also found to be overrepresented in a small group of CRMs that specifically support expression in ventro-lateral domains of the embryo (Stathopoulos et al. 2002), and since then others have studied cooperativity between Twist and Dorsal binding (e.g., Erives and Levine 2004; Zinzen et al. 2006; Crocker et al. 2008). It might follow that the CA-core E-box is generally required to support cooperative interactions with Dorsal or with other collaborating factors, although we did not detect Dorsal motifs in most Twist ChIP-seq defined regions.

We favor the view that in the majority of regions the Twist motif highlighted by ChIP-seq is the one most likely to contribute to regulating gene expression (or other unidentified functions), but we cannot dismiss contributions from other E-box sites present in the region. Our experiments with the *rho* enhancer illustrate this, as both E-boxes CACATG and CATATG, located five nucleotides apart, affect gene expression. Within Twist ChIP-seq peaks, we find that TA core E-boxes are less frequent overall and only weakly enriched under peaks of binding (±250 bp from the peak summit), and as a result they are not often "explanatory" (<±50 bp from the peak summit). Yet these accessory TA core E-boxes may also contribute to regulating gene expression, whether by binding Twist more transiently or by interacting with some other factor. Because the CA core E-box is also bound by Snail, the balance of activation/repression may require that a combination of CA and TA core E-boxes is optimal to support expression. Furthermore, while Twist bound to the explanatory sites may serve a major role in regulating gene expression and these accessory sites may provide less input, even marginal input may be crucial to support gene expression patterns in ways that matter for viability and selection, even though some of these may also be too subtle for our assays to detect.

SII.4.3 Simple sequence motifs and chromatin status

Apart from the CA- and GA-repeat sequences, no motifs other than the E-boxes were found to co-cluster with Twist binding sites in a large fraction of Twist-bound regions even when a wider window around the peaks of detected binding was interrogated. This does not preclude that other factors function in important combinations with Twist, but it suggests that no single transcription factor motif is commonly used in the entire Twist-occupied set. Finding specific combinations will require focus on subsets of regions selected by other criteria, such as expression pattern of nearby genes, performance of CRMs in transgenic assays, or direct binding assays for known or suspected accessory factors.

We do not know the significance of CA- and GA-simple repeat motifs that are enriched in Twist binding regions, but their association in other studies with open chromatin regions is suggestive (Auerbach et al. 2009). We hypothesized that GAGA-binding factor (GAF) which binds to promoters (for review, see Lehmann 2004) might do so here in promoter proximal regions through recognition of the GA-repeats. However, we did not find an enrichment of GA-repeat sequences associated with promoter proximal Twist peaks; the GA-repeats were located in many different positions suggesting a broader role than regulation of promoters, such as making DNA regions accessible.

Depletion of A/T-rich sequences from peaks was striking and it proved to be non-specific, as it is associated with a multitude of ChIP-seq samples. Further analyses showed there is a similar depletion of A/T-rich sequences around ChIP-seq peaks for diverse factors and in multiple genomes, including worm, mouse, and human (Supplemental Fig. 15; K Fisher-Aylor and B Wold, unpubl. obs.). This depletion was also seen when “peaks” of reads were selected from matching control samples of input

chromatin (cross-linked, sheared, and reverse cross-linked). The sonication step associated with ChIP-seq has recently been shown to enrich for promoter regions, DNase I hypersensitive sites, and other "open" chromatin regions (Auerbach et al. 2009), but in that work no specific sequence content biases were reported. The depletion of A/T rich runs might arise from a role these sequences have been suggested to play in nucleosome exclusion and positioning (e.g., Iyer and Struhl 1995; Peckham et al. 2007). Our observations of broad A/T depletion arose from a study of motif representation that happened to be A-rich (Supplemental Fig. 11), and it suggests that careful examination of background input chromatin is needed when evaluating the sequence composition of ChIP regions.

SII.4.4 The conservation profile around explanatory Twist motifs implies CRMs of ~300 bp

The genomes of *Drosophilids* are known to exhibit more conservation, in general, than many other animal species separated by what are thought to be an equivalent length of evolutionary distance. Thus, it has proven difficult to identify putative CRMs based on a simple search for increased local conservation of non-coding DNA sequence among *Drosophilid* genomes. Early comparative studies of enhancer regions in *Drosophila* species suggested that local increases in conservation of non-coding sequence imply regulatory function (Bergman et al. 2002). More recently, it has been suggested that this idea should be narrowed to conservation of specific binding sites only within CRMs or even just conservation of site number without strong primary sequence conservation (Sosinsky et al. 2007; Ho et al. 2009; Liberman and Stathopoulos 2009). Here we provide evidence to support both views: increased general conservation of sequence within putative CRMs relative to genomic background, as well as higher conservation of particular binding sites (Fig. 7). We asked if there is a genome-

wide average conservation signature that would characterize candidate CRMs; ChIP-chip data previously detected a conservation preference but without clarity about the dimensions of regions under selective pressure (MacArthur et al. 2009). Our data suggests that sequences around these motif instances are preferentially conserved compared with genomic background in a window of ~300 bp on average, a size that corresponds well with anecdotal samplings of individual CRMs. We also found evidence that the explanatory sites identified by Twist binding are preferentially conserved compared with their surroundings, arguing for their biological salience.

SII.5 Methods

SII.5.1 Fly stocks and general molecular biology

Drosophila melanogaster fly stocks were reared under standard conditions at 25°C. Transgenic flies were obtained using standard P-element transformation or by site-directed integration. Wild type refers to the background *yw*. P-element transformations were achieved in *yw* flies, while site-directed integration was carried out using *D. mel* stock containing attP insertion at position ZH-86Fb. Enhancer sequences were amplified from genomic DNA (primer sequences are available upon request) and cloned into eve.promoter-LacZ-attB or eve.promoter-cherry-attB vectors (Liberman and Stathopoulos 2009). Anti-sense riboprobes labeled with Digoxigenin-UTP (Roche) were used for in situ hybridization to detect transcripts.

SII.5.2 Chromatin preparation, DNA isolation, amplification, hybridization, and sequencing

Chromatin was prepared as described previously (Sandmann et al. 2006) from 2 g of *yw* embryos of from 1 to 3 h in age. Rat anti-Twist antibody (gift of M. Levine, UC Berkeley) was used for both ChIP-chip and ChIP-seq experiments. For ChIP-chip, the

resulting DNA library was labeled and hybridized to arrays by NimbleGen Systems, Inc.; 10 ng of immunoprecipitated (IP) DNA was amplified using the Whole Genome Amplification kit (Sigma) according to the manufacturer's instructions. The mock ChIP-chip sample used preimmune antibody, rather than anti-Twist. For ChIP-seq, 50 ng of IP material was used to prepare a library (Johnson et al. 2007), and DNA sequencing of samples was performed by the Illumina protocol at Caltech Genome Center. The ChIP-seq input control was processed equivalently to the Twist ChIP-seq sample, except that it was not immunoprecipitated (no antibody or bead processing). Each ChIP-seq library was sequenced to a total of 9 million reads.

SII.5.3 SELEX

SELEX experiments using in vitro binding to a column were carried out as described (Ogawa and Biggin 2011). See the Supplemental Text for more details, including processing of SELEX data.

SII.5.4 Bioinformatics

ChIP-chip and ChIP-seq data processing: Methods used to call ChIP-chip versus ChIP-seq peaks are described in detail within the Supplemental Text. In brief, we used the ERANGE software suite to call peaks based on the number, orientation, and ratio of short sequenced reads relative to a background control. We considered an alternate peak caller (MACS), overlap of ChIP-seq regions with ChIP-chip regions, and the inclusion of known Twist targets to determine the threshold for calling Twist occupied sites (i.e., ChIP-seq signals). We selected a high confidence (HC) set of 513 sites based on high inclusion in ChIP-chip regions (87%), MACS regions (72%), and validated Twist targets (75%). We also selected a medium confidence (MC) set of 1099 regions based on the similarity in motif organization around these peaks (E-box, Fig. 3A).

SII.5.5 ChIP-seq summit refinement

After ChIP-seq enriched regions were identified by the ERANGE program, post-processing was performed to refine the summit location by utilizing directional tag information. For each peak region, plus and minus tags were simultaneously shifted toward the imputed fragment center by a trial amount, ranging from 0 to 100 bp. The shift that maximized area overlap of the plus and minus tag density profiles (i.e., a measure of "directionality") was then implemented prior to calculating the location of the ChIP-seq tag count maximum ("summit").

SII.5.6 Explanatory site interval

The interval for designating "explanatory sites" near ChIP-seq summits was estimated utilizing count statistics for the CACATG motif, due to its being the most prevalent E-box in the set of Twist regions. Specifically, the motif occurrences within increasing radii around peak centers (binned by 5 bp) were compared to the number expected from a Poisson distribution with the mean equal to the genome average density of CACATG motifs. When the probability of the observed number of counts coming from the Poisson model fell below 0.001, the distribution was deemed indistinguishable from random fluctuations, and the boundary of the previous bin was set to be the cutoff for explanatory sites (±50 bp from the summit).

SII.5.7 Conservation analysis

Conservation at each base pair was assessed using phastCons scores (Siepel et al. 2005). Genome-wide scores for the fifteen-way insect alignment including *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi, A.*

gambiae, A. mellifera, and *T. castaneum* were downloaded from the UCSC genome gateway. Statistical analysis of the data is described in the Supplemental Methods.

SII.6 Annotations

Precomputed annotation files for exons and introns were downloaded from the FlyBase website, release 5.27 (Tweedie et al. 2009). Here, exons and introns are mutually exclusive. 5′ UTRs data are from S. Celniker.

SII.5.7 Acknowledgments

We thank the Caltech Jacobs Genome Facility members I. Antoshechkin and L. Schaeffer for library building and DNA sequencing, as well as D. Trout, B. King, and H. Amrhein for primary sequence data processing and visualization. We are grateful to A. Mortazavi and A. Kirilusha (Caltech Biology) for software and discussion of analysis; M. Biggin and S. Celniker (Lawrence Berkeley Lab) for sharing unpublished data; and M. Levine (University of California at Berkeley) for antibodies. K.I.F.-A. was funded by a NSF pre-doctoral fellowship, and S.P. was funded by The Gordon and Betty Moore Foundation. Work at Lawrence Berkeley National Laboratory was conducted under Department of Energy contract DE-AC02-05CH11231. This work was funded by the Functional Genomics Resource Center of the Caltech Beckman Institute, NIH grant R01GM077668 (A.S.), NIH grant U54HG004576 (B.J.W.), and the Bren Chair (B.J.W).

SII.5.8 Footnotes

- [Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE26285, and the sequence data from this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA027330.]

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.104018.109.

SII.5.9 References

1. ↵

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94: 890–898.

Caltech ConnectCrossRefMedline

2. ↵

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci 106: 14926–14931.

Abstract/FREE Full Text

3. ↵

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: Discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–373 (Web Server issue).

Abstract/FREE Full Text

4. ↵

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol 3: RESEARCH0086. doi: 10.1186/gb-2002-3-12-research0086.

Caltech ConnectMedline

5. ↵

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify

cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci 99: 757–762.

Abstract/FREE Full Text

6. ↲

Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, MacQuarrie KL, Davison J, Morgan MT, Ruzzo WL, et al. 2010. Genome-wide MyoD binding in skeletal muscle cells: A potential for broad cellular reprogramming. Dev Cell 18: 662–674.

Caltech ConnectCrossRefMedline

7. ↲

Castanon I, Von Stetina S, Kass J, Baylies MK. 2001. Dimerization partners determine the activity of the Twist bHLH protein during Drosophila mesoderm development. Development 128: 3145–3159.

Abstract/FREE Full Text

8. ↲

Chopra VS, Levine M. 2009. Combinatorial patterning mechanisms in the Drosophila embryo. Brief Funct Genomics Proteomics 8: 243–249.

Abstract/FREE Full Text

9. ↲

Crocker J, Tamori Y, Erives A. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. PLoS Biol 6: e263. doi: 10.1371/journal.pbio.0060263.

Caltech ConnectCrossRefMedline

10. ↲

Davidson EH. 2006. The regulatory genome: Gene regulatory networks in development and evolution. Academic, Burlington, MA.

11. ↲

Erives A, Levine M. 2004. Coordinate enhancers share common organizational features in the Drosophila genome. Proc Natl Acad Sci 101: 3851–3856.

Abstract/FREE Full Text

12. ↲

Fullwood MJ, Ruan Y. 2009. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 107: 30–39.

Caltech ConnectCrossRefMedline

13. ↲

Gonzalez-Crespo S, Levine M. 1993. Interactions between dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in Drosophila. Genes Dev 7: 1703–1713.

Abstract/FREE Full Text

14. ↲

Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell 138: 314–327.

Caltech ConnectCrossRefMedline

15. ↲

Gurudatta BV, Corces VG. 2009. Chromatin insulators: Lessons from the fly. Brief Funct Genomics Proteomics 8: 276–282.

Abstract/FREE Full Text

16. ↲

Ho MC, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C, Bender W, et al. 2009. Functional evolution of cis-regulatory modules at a homeotic gene in Drosophila. PLoS Genet 5: e1000709. doi: 10.1371/journal.pgen.1000709.

Caltech ConnectCrossRefMedline

17. ↲

Ip YT, Levine M, Small SJ. 1992a. The bicoid and dorsal morphogens use a similar strategy to make stripes in the Drosophila embryo. J Cell Sci Suppl 16: 33–38.

Caltech ConnectMedline

18. ↲

Ip YT, Park RE, Kosman D, Bier E, Levine M. 1992b. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. Genes Dev 6: 1728–1739.

Abstract/FREE Full Text

19. ↲

Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M. 1992c. dorsal–twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. Genes Dev 6: 1518–1530.

Abstract/FREE Full Text

20. ↲

Iyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J 14: 2570–2579.

Caltech ConnectMedline

21. ↲

Jiang J, Rushlow CA, Zhou Q, Small S, Levine M. 1992. Individual dorsal morphogen binding sites mediate activation and repression in the Drosophila embryo. EMBO J 11: 3147–3154.

Caltech ConnectMedline

22. ↲

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.

Abstract/FREE Full Text

23. ↲

Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. Genome Res 20: 890–898.

Abstract/FREE Full Text

24. ↲

Lehmann M. 2004. Anything else but GAGA: A nonhistone protein complex reshapes chromatin structure. Trends Genet 20: 15–22.

Caltech ConnectCrossRefMedline

25. ↲

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature 424: 147–151.

Caltech ConnectCrossRefMedline

26. ↲

Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6: e27. doi: 10.1371/journal.pbio.0060027.

Caltech ConnectCrossRefMedline

27. ↲

Liberman LM, Stathopoulos A. 2009. Design flexibility in cis-regulatory control of gene expression: Synthetic and comparative evidence. Dev Biol 327: 578–589.

Caltech ConnectCrossRefMedline

28. ↲

Lusk RW, Eisen MB. 2010. Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. PLoS Genet 6: e1000829. doi: 10.1371/journal.pgen.1000829.

Caltech ConnectCrossRefMedline

29. ↲

MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV, et al. 2009. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol 10: R80. doi: 10.1186/gb-2009-10-7-r80.

Caltech ConnectCrossRefMedline

30. ↲

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci 99: 763–768.

Caltech ConnectMedline

31. ↲

Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A, Levine M. 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. Development 131: 2387–2394.

Abstract/FREE Full Text

32. ↲

Massari ME, Murre C. 2000. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. Mol Cell Biol 20: 429–440.

FREE Full Text

33. ↲

Murre C, McCaw PS, Vaessin H, Caudy M, Jan LY, Jan YN, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, et al. 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell 58: 537–544.

Caltech ConnectCrossRefMedline

34. ↲

Nam J, Dong P, Tarpine R, Istrail S, Davidson EH. 2010. Functional cis-regulatory genomics for systems biology. Proc Natl Acad Sci 107: 3930–3935.

Abstract/FREE Full Text

35. ↲

Ogawa N, Biggin MD. 2011. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. In Methods in molecular biology (ed. Deplanke B), Humana Press, Clifton, New Jersey (in press).

36. ↲

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. Genome Res 17: 1170–1177.

Abstract/FREE Full Text

37. *Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. Nat Methods 6 (11 Suppl): S22–S32.*

Caltech ConnectCrossRefMedline

38. ↲

Reeves GT, Stathopoulos A. 2009. Graded dorsal and differential gene regulation in the Drosophila embryo. Cold Spring Harb Perspect Biol 1: a000836. doi: 10.1101/cshperspect.a000836.

Abstract/FREE Full Text

39. ↲

Sandmann T, Jakobsen JS, Furlong EE. 2006. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. Nat Protoc 1: 2839–2855.

Caltech ConnectCrossRefMedline

40. ↲

Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EE. 2007. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev 21: 436–449.

Abstract/FREE Full Text

41. ↲

Schuettengruber B, Cavalli G. 2009. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development 136: 3531–3542.

Abstract/FREE Full Text

42. ↲

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.

Abstract/FREE Full Text

43. ↲

Small S, Blair A, Levine M. 1992. Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J 11: 4047–4057.

Caltech ConnectMedline

44. ↲

Sosinsky A, Honig B, Mann RS, Califano A. 2007. Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci 104: 6305–6310.

Abstract/FREE Full Text

45. ↲

Stathopoulos A, Levine M. 2002. Linear signaling in the Toll-Dorsal pathway of Drosophila: Activated Pelle kinase specifies all threshold outputs of gene expression while the bHLH protein Twist specifies a subset. Development 129: 3411–3419.

Caltech ConnectMedline

46. ↲

Stathopoulos A, Levine M. 2005. Genomic regulatory networks and animal development. Dev Cell 9: 449–462.

Caltech ConnectCrossRefMedline

47. ↲

Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M. 2002. Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. Cell 111: 687–701.

Caltech ConnectCrossRefMedline

48. ↲

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. 2009. FlyBase: Enhancing Drosophila gene ontology annotations. Nucleic Acids Res 37: D555–D559.

Abstract/FREE Full Text

49. ↲

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829–834.

Caltech ConnectCrossRefMedline

50. ↲

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev 21: 385–390.

Abstract/FREE Full Text

51. ↲

Zinzen R, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the Drosophila embryo. Curr Biol 16: 1358–1365.

Caltech ConnectCrossRefMedline

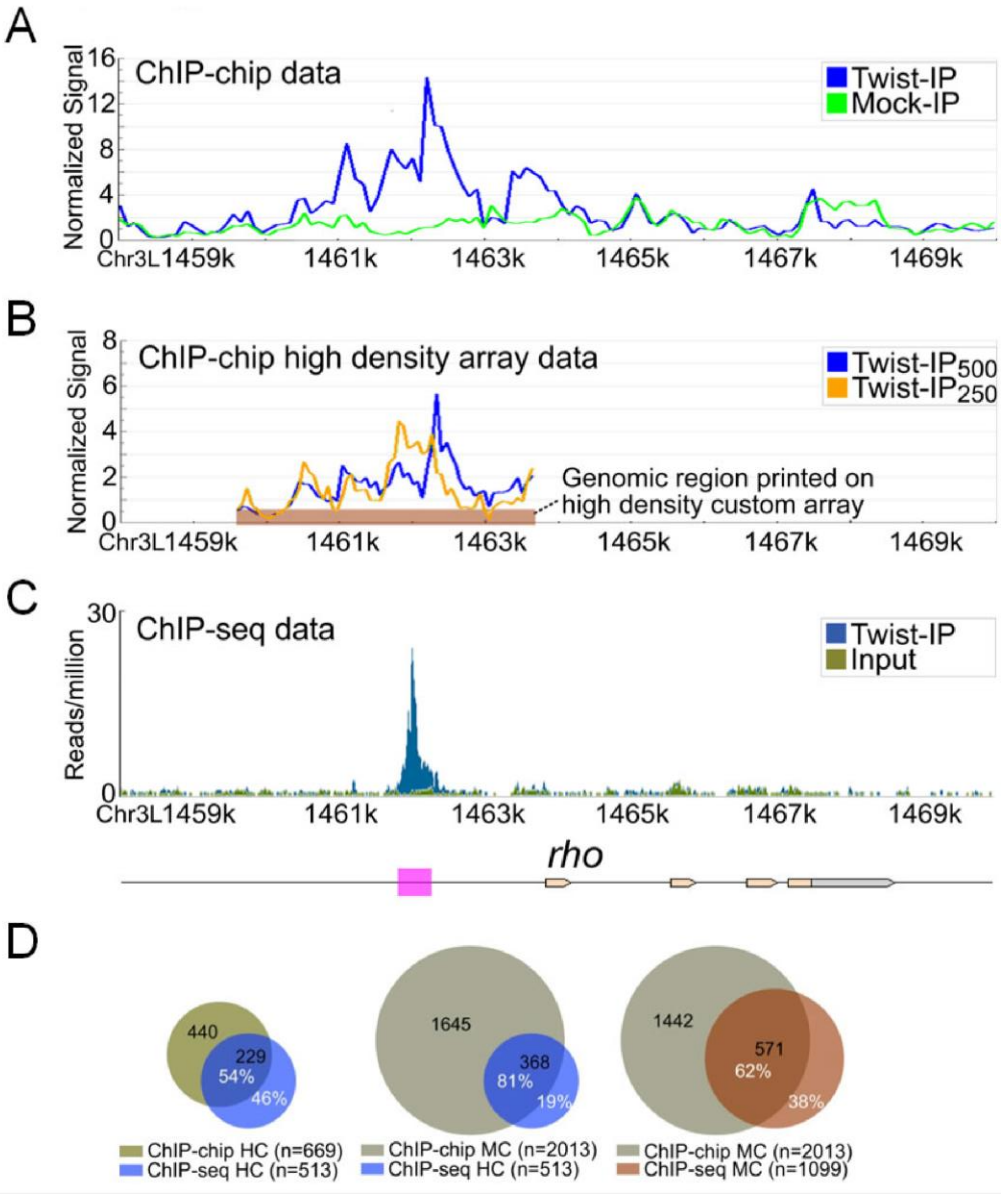
52. ↲

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 462: 65–70.

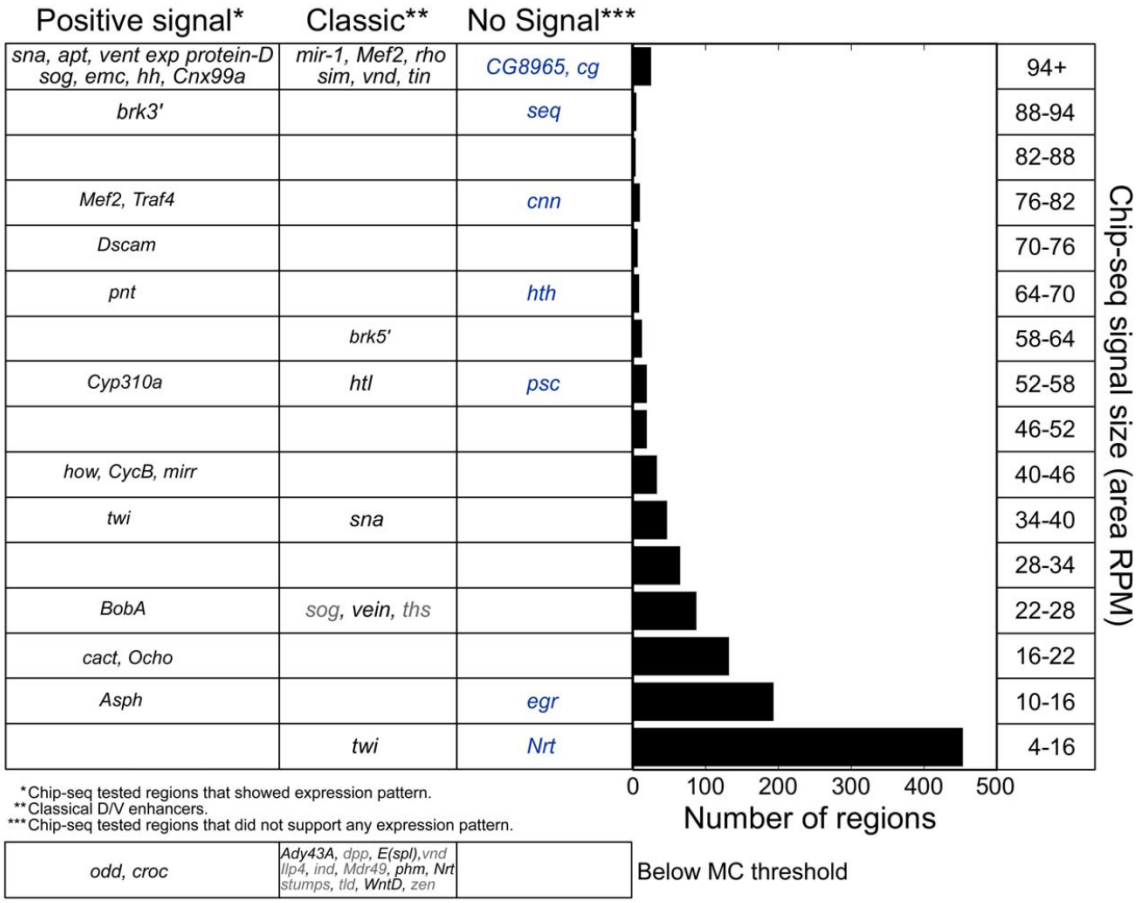
Caltech ConnectCrossRefMedline

SII Supplements for Ozdemir*, Fisher-Aylor*, et al., 2011

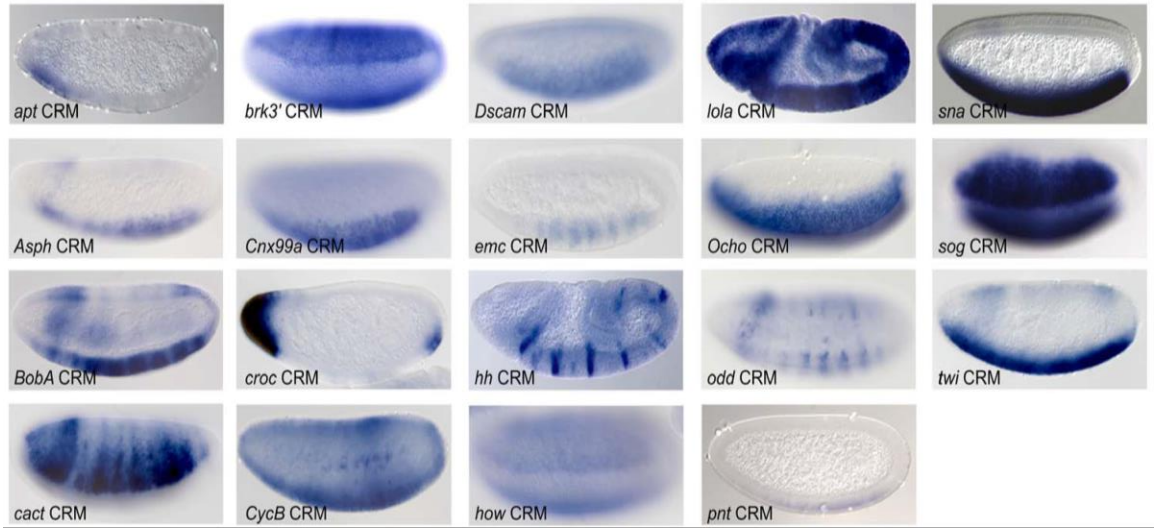
SII.S1 FIGURES



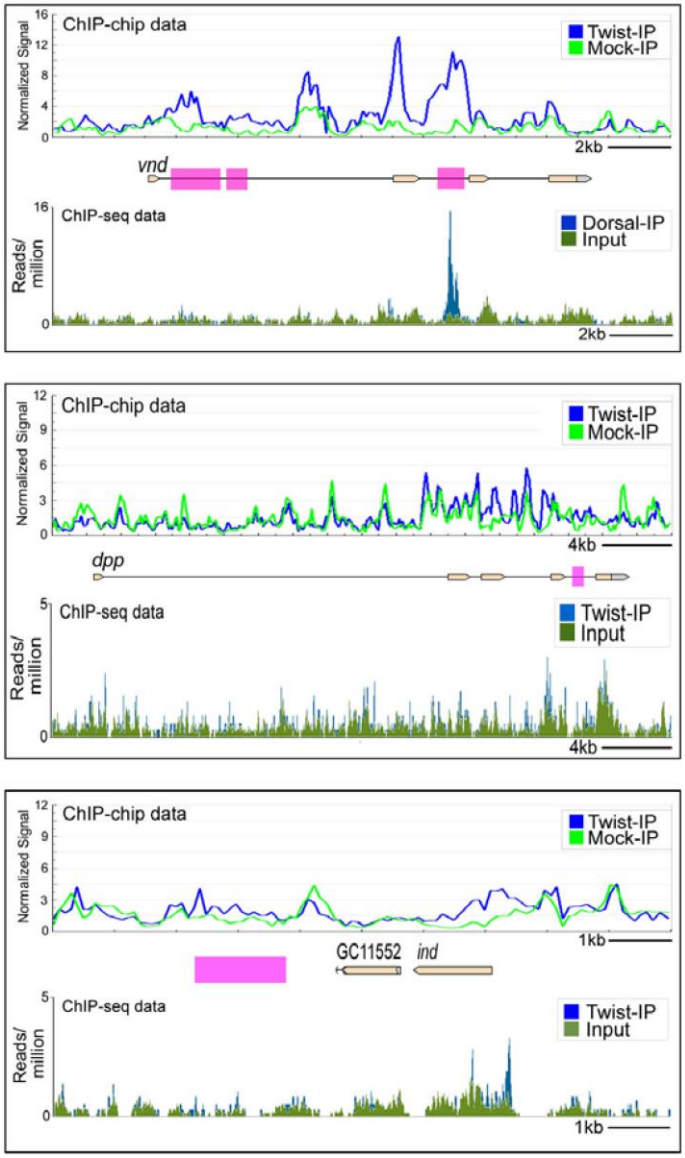
Supplemental Figure 1. In vivo Twist occupancy determined by ChIP-Seq versus ChIP-chip and the isolation of CRMs. (A) *Twist ChIP-chip binding to a standard Nimblegen array at a representative locus, rho, relative to previous characterized early embryonic enhancer (pink box; Ip, Park et al. 1992).* ***(B)*** *Twist ChIP-chip binding to a high-density custom array to same region for same Twist-IP (blue line) as used in (A); differences can be attributed to the assay method and data processing, rather than to the input chromatin lengths or other biological variation. Another independent Twist-IP prepared from smaller chromatin (sheared to ~250bp average) is shown in orange. Brown bar: location of the tiled regions on the custom array.* ***(C)*** *Twist ChIP-Seq-defined occupancy obtained using Twist antibody (blue) compared with sequenced input control DNA (green).* ***(D)*** *Venn diagrams showing the overlap between ChIP-chip and ChIP-Seq datasets of various sizes/FDRs. False Discovery Rate (FDR) of ~1% supported calling 513 high confidence (HC) ChIP-Seq regions and 669 HC ChIP-chip regions. FDR of 17% supported calling 1099 MC ChIP-Seq regions and 2013 MC ChIP-chip regions.*



Supplemental Figure 2: Twist ChIP-Seq signals at known and candidate CRMs from prior studies. *The number of Twist regions is shown ranked by signal size (reads per million in the entire area under the peak). As expected, lower ChIP signal regions are much more numerous than high signal regions. Regulatory regions that have previously been shown to support dorsal-ventral expression in the early embryo correspond to both large and small Twist ChIP-Seq peaks. In addition, regions that have been shown in this study to support expression and regions that failed to do so are distributed over the range of ChIP signal sizes.*

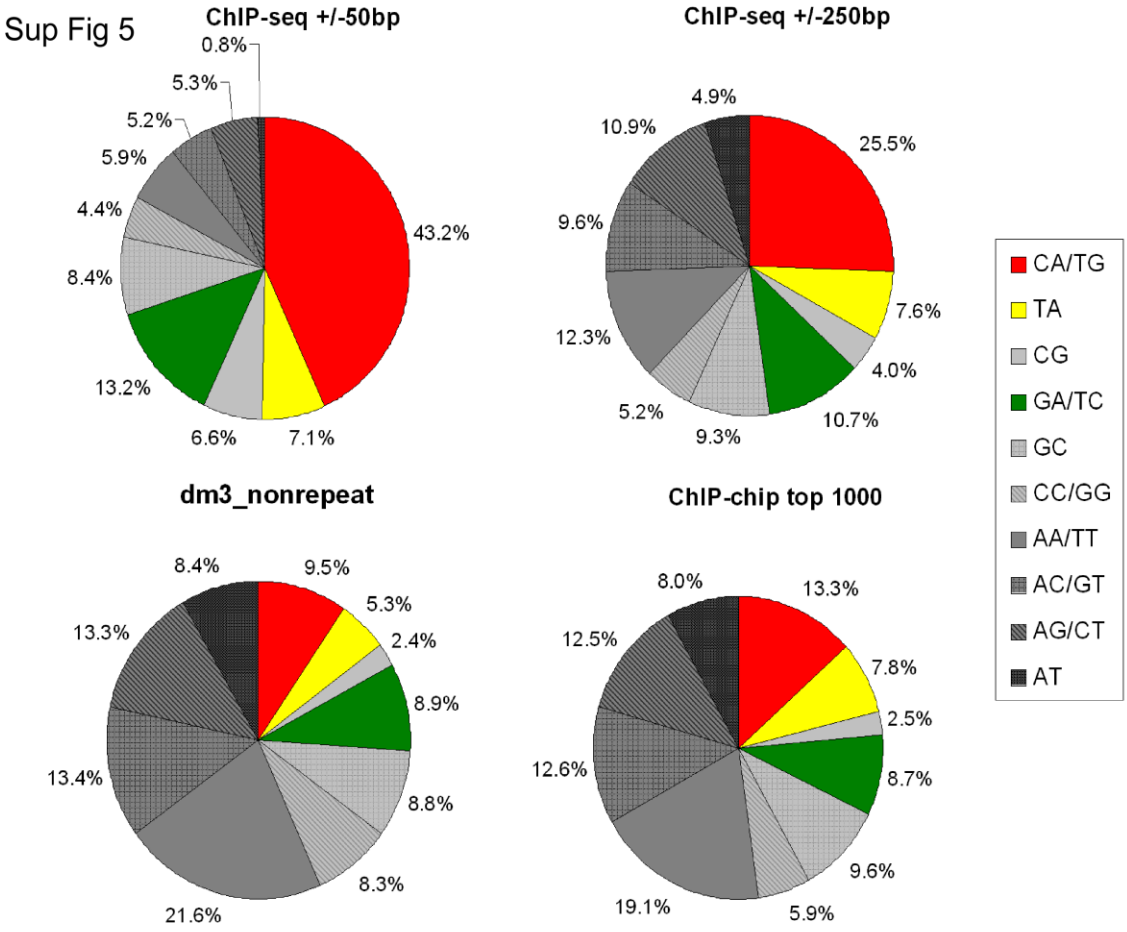


Supplemental Figure 3: Functional analysis of Twist regions by reporter gene assay. *Twist regions were tested for their ability to support gene expression in a standard reporter gene assay using either lacZ or cherry reporter genes. In situ hybridization using riboprobes to lacZ or cherry were used to monitor gene expression supported by these DNA sequences in early embryos. Shown are the 19 of 31 tested regions found to support expression. Closest associated genes are indicated in the bottom corner of each panel; see Table 2a for exact coordinates of the DNA regions tested. Four additional regions found to support expression are shown in Fig. 1, for a total of 23 positives of 31 regions assayed.*



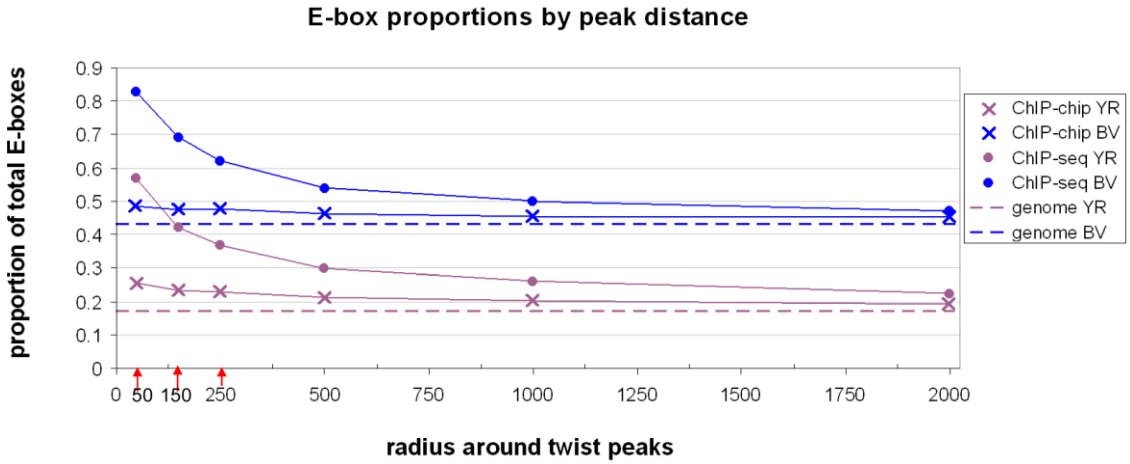
Supplemental Figure 4: Expression activity is not predicted by ChIP-Seq signal size. *ChIP-chip and ChIP-Seq Twist data from this study are shown on the top and bottom of each panel, respectively. Pink boxes mark the locations of previously characterized enhancers. Twist signal is detected at the previously characterized vnd early embryonic enhancer located in the second intron (Stathopoulos, Van Drenth et al. 2002), which is consistent with the early 1-3 hr timepoint assayed in this study. We do not detect significant Twist signal at a second vnd candidate enhancer which was identified more recently by ChIP-chip analyses at a slightly later developmental timepoint (Zeitlinger et al., 2007); perhaps the enhancers in the first intron support later or weaker*

gene expression. In the cases of dpp and ind, the sites shown are candidate enhancers based on motif presence and/or ChIP-chip binding. We did not see significant signals at these sites. dpp and ind are expressed in dorsal and dorsal-lateral regions of the embryo, which are outside the spatial domain of most Twist expression. These therefore fall into the group of previously discussed Twist targets that we call "Type III" (see text).

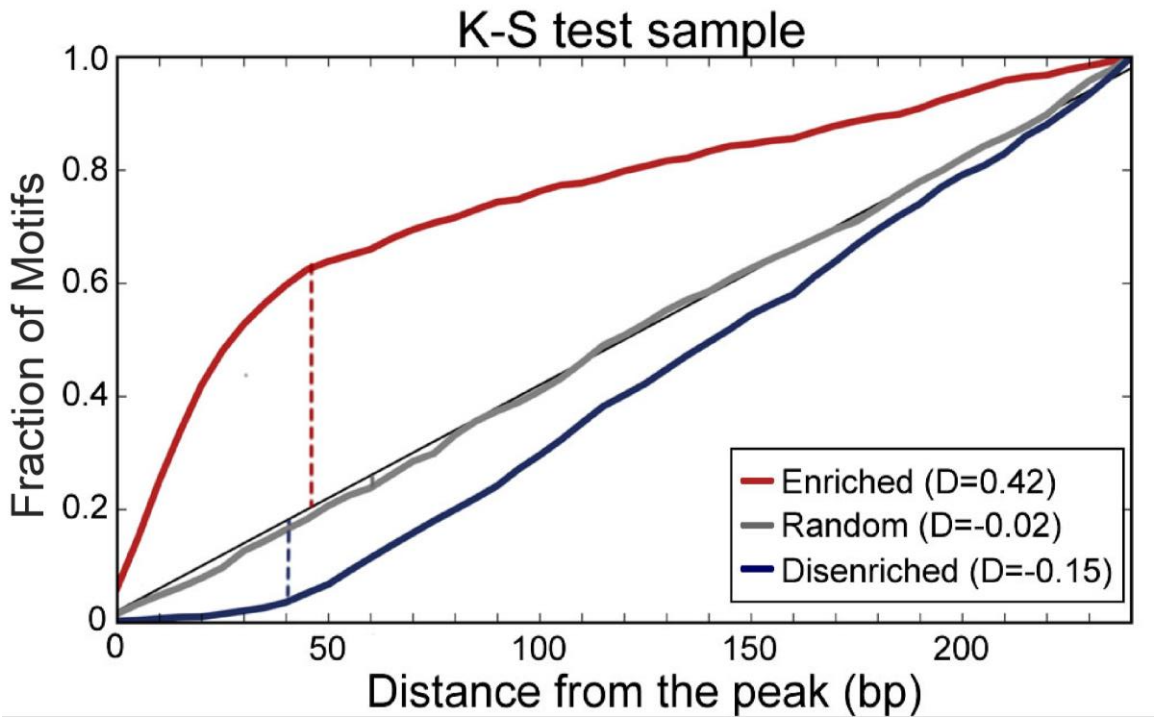


Supplemental Figure 5: Frequency of E-box instances in ChIP-Seq versus ChIP-chip close to the signal summit (±50bp) or at greater distance from it (±250bp). *CANNTG E-boxes were tallied around Twist MC ChIP-Seq peaks, the largest 1,000 MC Twist ChIP-chip peaks, and the non-repeat fly genome. Displayed are the proportions of the different possible interior ten NN base pairs. When the areas very close (±50bp) to Twist ChIP-Seq peaks are compared to the wider ±250bp areas around Twist peaks, CA E-boxes predominate, suggesting that they dominate in supporting ChIP-detectable binding. There is also a distinct lack of AT E-boxes. The proportion of TA E-boxes*

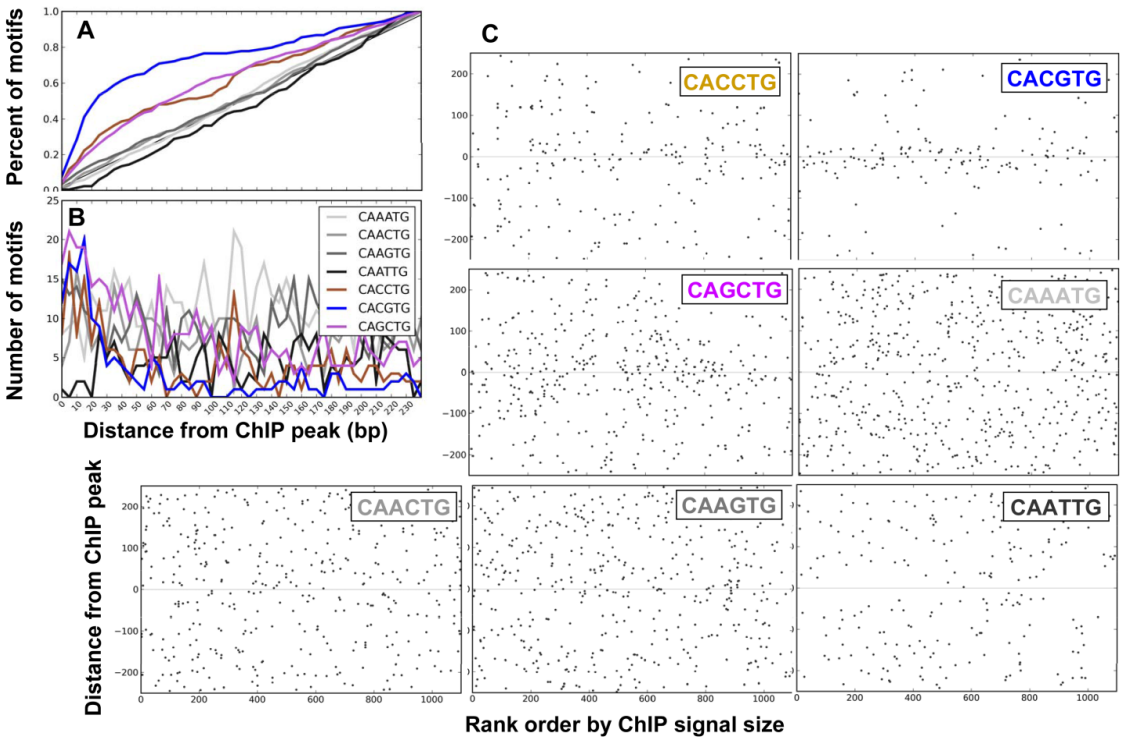
remains relatively steady close to and farther from the peaks. The proportions of E-box cores around ChIP-chip summits are very similar to the genomic background distribution, suggesting that while ChIP-chip tiling arrays find larger domains putatively occupied by Twist, the peak of signal is far less accurate in identifying the explanatory Twist binding sites.



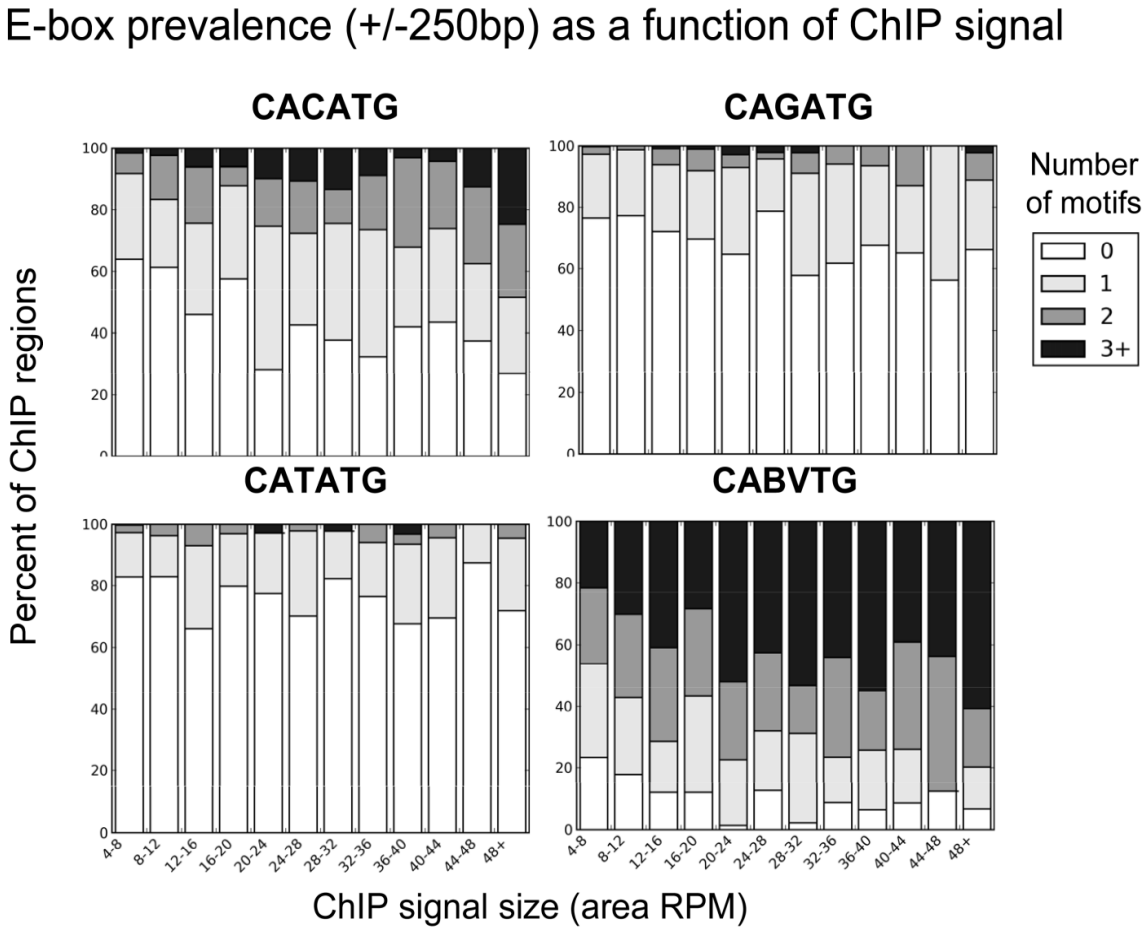
Supplemental Figure 6: Frequency of CAYRTG or CABVTG E-boxes within ChIP-chip or ChIP-Seq data as a function of distance from the summit. *Twist 'explanatory' E-boxes were classed in two ways: the more canonical and stringent CAYRTG-core E-boxes (CA, TA, and CG) as well as the expanded CABVTG core suggested by our data (also including GA, GC, and CC). YR and BV E-boxes as a percent of all 10 possible E-boxes are shown in expanding radii out from the largest 1,000 Twist MC ChIP-chip peaks and the MC Twist ChIP-Seq peaks. They are compared to the distribution in the non-repeat genome. The ChIP-Seq data shows a marked enrichment of both types of explanatory E-boxes within ±50bp of ERANGE peaks (almost 85% of E-boxes are BV and almost 60% are YR) and this drops off exponentially with distance from the peak. The proportion of explanatory E-boxes is slightly greater near ChIP-chip summits as compared to the genomic background distribution.*



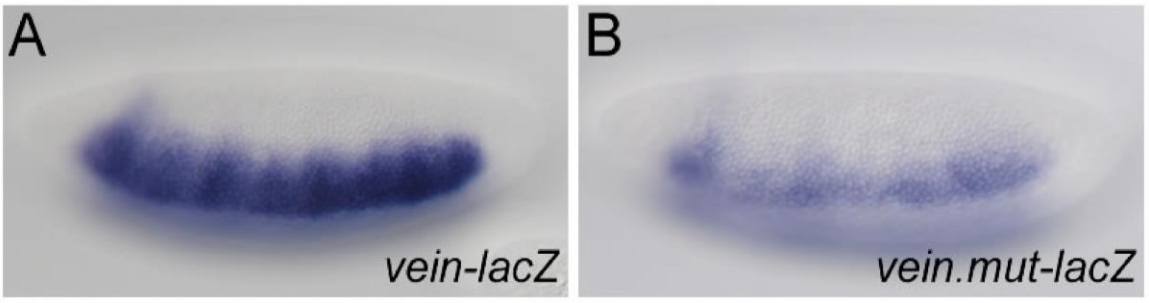
Supplemental Figure 7: Visual example of the K-S test. *The Kolmogorov-Smirnov (K-S) test determines the degree of similarity between two distributions (see Supplemental Methods). In order to determine whether certain motifs were enriched or depleted relative to Twist peaks, their cumulative distributions (red, blue, and grey plots) were compared to the cumulative distribution function of a uniform distribution (black diagonal line). D (dotted vertical line) is the maximum distance between the motif distribution function and the uniform distribution function. While the P-value determines if a distribution is statistically the same as uniform instead of enriched or depleted, the absolute value of D reflects the spatial degree (bp around Twist peaks) of the enrichment or depletion of a motif. A large D absolute value reflects a large degree of enrichment/depletion; enriched motifs have positive D values and depleted motifs have negative D values. P-values reported are in base 10 (i.e. 2.2E-16 means $2.2*10^{-16}$)*



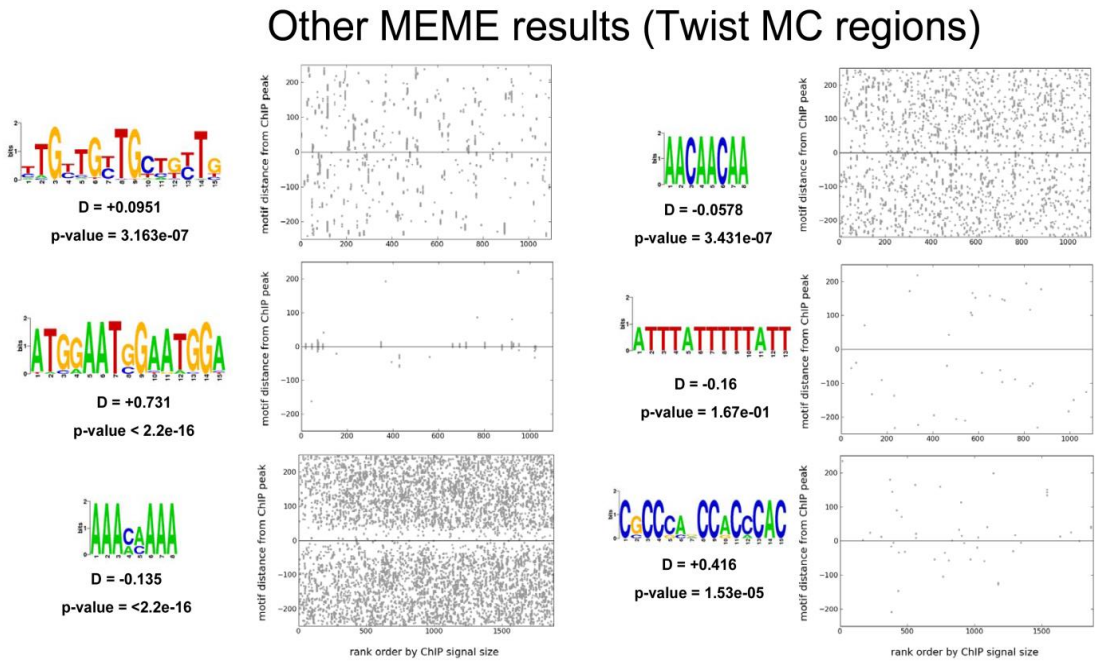
Supplemental Figure 8: Distribution of additional E-boxes within Twist ChIP-Seq data. *The three CABVTG E-boxes not shown in Figure 4: (CACCTG, CACGTG, and CAGCTG) also show some enrichment relative to the peak. Of these, CAGCTG is the most prevalent. CACGTG (the third member of the CAYRTG E-boxes) occurs less frequently but is quite enriched around Twist peaks. The 4 CAANTG E-boxes are not enriched relative to Twist peaks, and in fact, the CAATTG palindrome is weakly depleted. See Supplemental Table 3 for the K-S values.*



Supplemental Figure 9: E-box motif occurrence as a function of Twist ChIP-Seq signal size. *The number of CACATG, CAGATG, CATATG, and CABVTG E-boxes were counted in a ±250bp radius around each Twist peak. MC Twist regions were ranked according to size (area RPM), and the percentage of regions containing 0, 1, 2, or 3 and more motifs is shown for each size category. CACATG motifs occur within about 50% of the whole MC dataset, but the larger peaks are more likely to have multiple occurrences of E-boxes. This trend does not hold true for CATATG and CAGATG, which occur in only about 25% of the peaks, and are most likely to occur singly. Viewed collectively, CABVTG E-boxes are present in the large ±250bp radius around over 90% of Twist peaks and are also more likely to occur multiply near large Twist peaks. This suggests that the largest signal size features are most likely to be driven by multiple binding sites.*

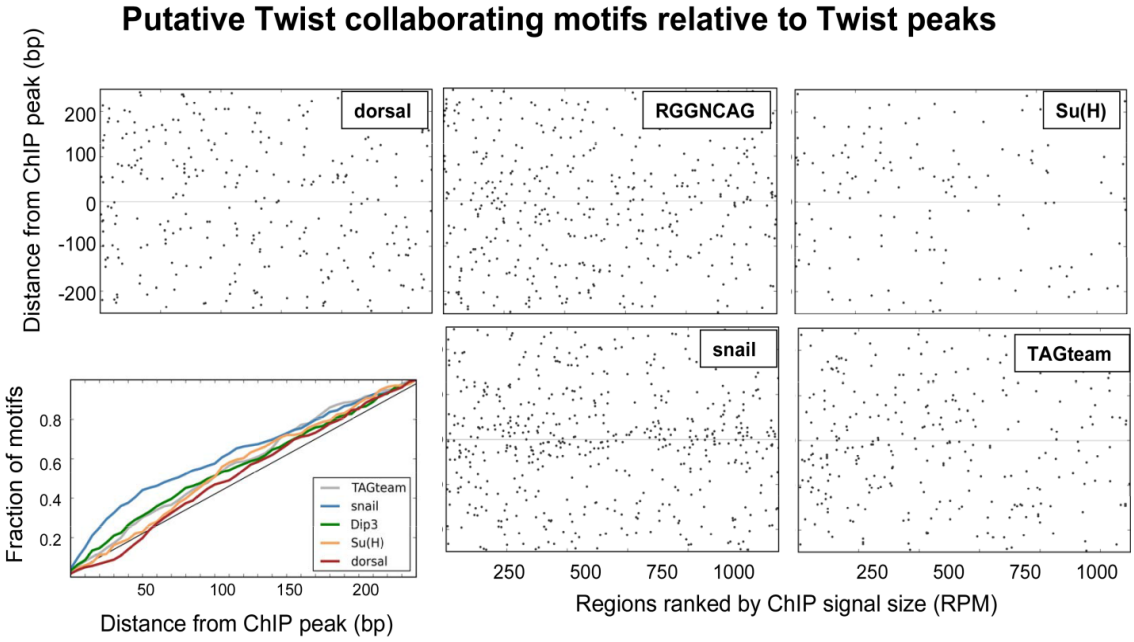


Supplemental Figure 10: vein CRM mutagenesis demonstrates the requirement for the explanatory E-box. *We introduced a single base pair change within potential explanatory sites (CACATG > GACATG) we had defined within the vein CRM (A), characterized previously (Markstein, Zinzen et al. 2004). Mutating the explanatory CA-core E-box in this manner resulted in a dramatic loss of reporter gene expression (B). Reporter gene expression was abrogated such that the expression domain collapsed from 10-12 cells in width to 4-7 cells for the vein CRM; this effect is comparable to the expression of vein genes in twist mutant embryos (data not shown). Previously, the orientation of this same E-box was also shown to be important for vein CRM expression (Zinzen, Senger et al. 2006).*

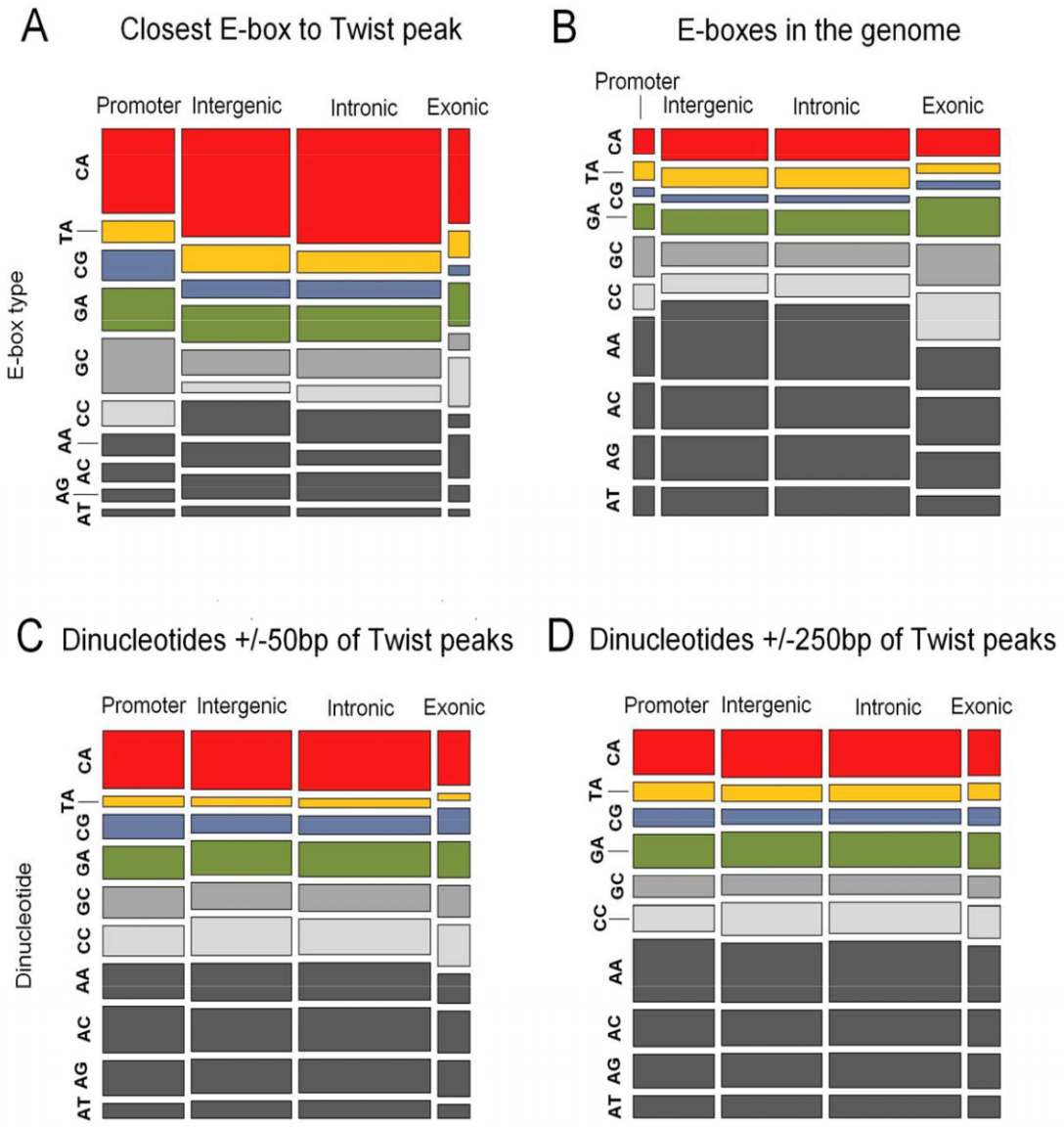


Supplemental Figure 11. MEME outputs. *The other MEME outputs not shown in Figure 6 are displayed here and mapped back onto Twist MC regions at 85% threshold.*

Their K-S values are shown in Supplemental Table 3 where, from top to bottom by column, they are called MEME MC ±50 motifs 3, 4, 6, 7, 9, and 10.

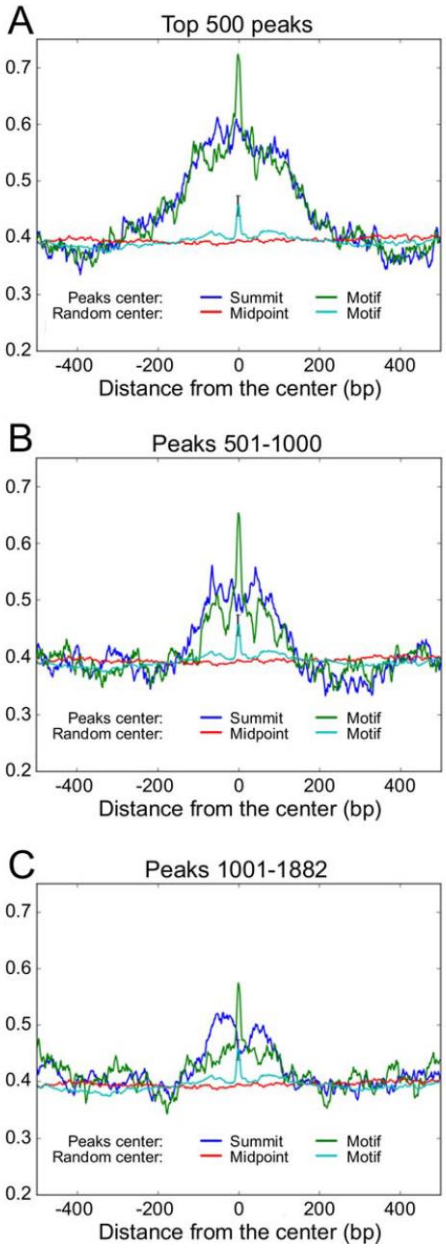


Supplemental Figure 12. Distributions of binding motifs for factors thought to interact with Twist. *The motifs for Dorsal (SELEX – GGG(W3-5)CYV, 100% match) (Markstein, Markstein et al. 2002, Zinzen, Senger et al. 2006, Liberman and Stathopoulos 2009); Zelda (TAGteam – YAGGYAG, 100% match) (ten Bosch, Benavides et al. 2006); Suppressor of Hairless [Su(H) – BRTGRGAAH 90% match] (Bailey and Posakony 1995); RGGNCAG/Unknown (RGGNCAG, 100% match) (Stathopoulos, Van Drenth et al. 2002); and Snail (RCARGWBB, 90% match) (Stathopoulos and Levine 2005) are shown relative to Twist peaks. If these factors interact directly with Twist to support expression through these predicted CRM regions, we would predict enrichment of the binding motifs relative to Twist peaks. The SELEX-derived Dorsal site [GGG(W3-5)CYV (A) as well as other previously described Dorsal sites (data not shown)] and Zelda are not enriched relative to Twist peaks. The Su(H) and RGGNCAG motifs are present and weakly clustered around the Twist peaks (B). Snail exhibits a significantly enriched binding site distribution near Twist summits, yet because the Snail consensus binding sequence overlaps with that of some Twist sites, the interpretation of this result with respect to probable Snail activity is not certain. See Supplemental Table 3 for the K-S values of these motifs.*

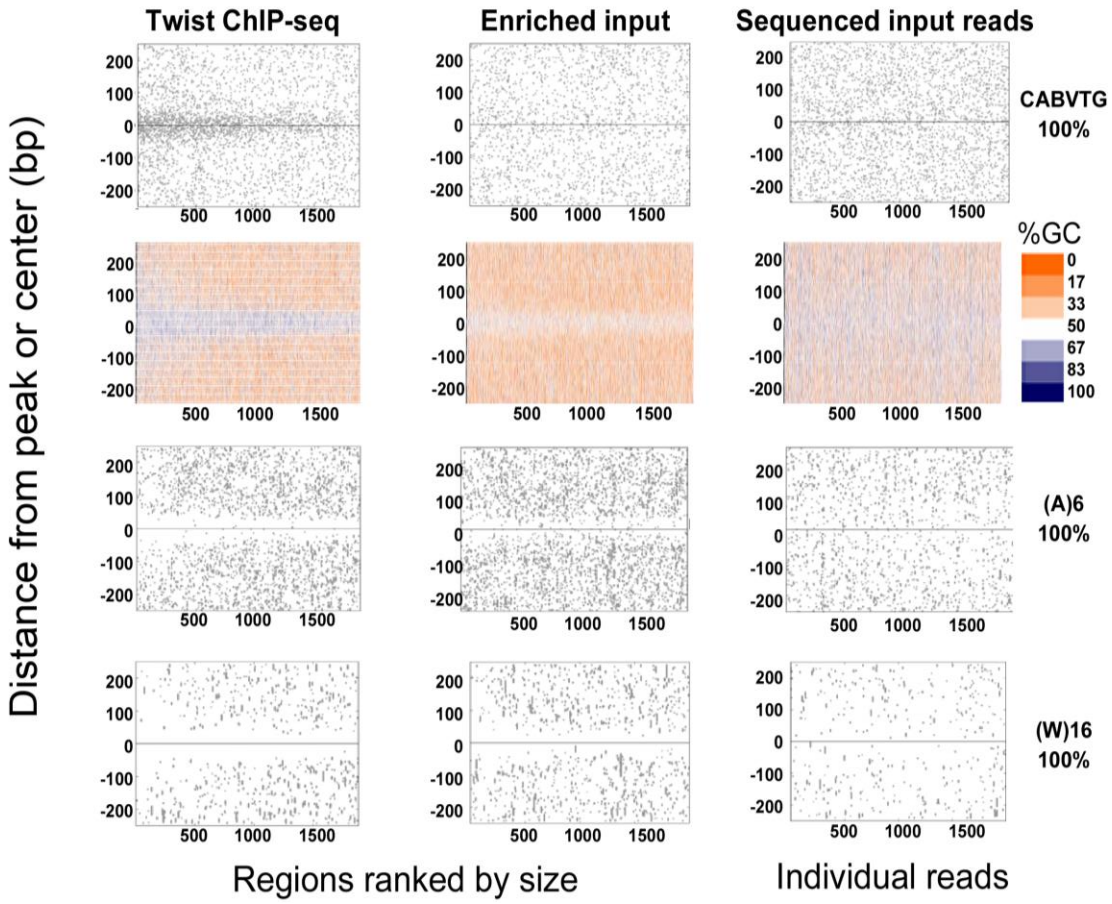


Supplemental Figure 13. E-box and dinucleotide repeat frequencies under Twist ChIP-Seq peaks versus the genome. *(A) Twist MC peaks (i.e. "shifted summits") were classed according to genomic location (as in Fig 6; see Supplementary Methods) and the closest E-boxes within ±50bp of Twist peaks in each category is shown. 23% of promoter proximal, 25% of intergenic, 23% of intronic, and 41% of exonic regions have no E-box within ±50bp. (B) The proportion of E-boxes in all genomic categories is shown. The proportion of CAGCTG E-boxes is greater in promoters than intergenic regions or introns, but it is still not as large as the proportion of CAGCTG E-boxes in Twist regions associated with promoters. In order to determine if the E-box proportions*

under Twist peaks is a direct result of dinucleotide frequencies in different regions of the genome, we analyzed all dinucleotides under the narrow ±50bp around Twist peaks (C) and the larger ±250bp radius (D). There is very little change in the frequency of dinucleotides under Twist peaks falling into different areas of the genome, suggesting that the proportional E-box difference between categories is not due to overall dinucleotide representation. There are slightly fewer A/T-rich dinucleotides very close to Twist peaks, which is consistent with an overall depletion of A/T-rich sequences near peaks (Suppl. Figure 15).



Supplemental Figure 14. Conservation local to summits throughout peak rankings. *The average PhastCon score is shown at every base pair around Twist-occupied sites ("peaks") and compared to average conservation distribution of 30 samples of 500 regions from the non-repeat dm3 genome ("random"). The "summit centered" plots are drawn relative to the shifted ERANGE peaks (Twist) and the "midpoint centered" plots are drawn relative to the centers of the randomly selected genomic background regions. The "motif centered" Twist plot was re-centered on the nearest CABVTG E-box (Twist explanatory motif) within ±150bp of the ERANGE summits, and regions with no such motif were left out. For the "motif centered" random plot, random regions were pre-screened to contain one of the CABVTG motifs. Relative to the genomic background, the entire area around Twist occupied sites is highly conserved in the HC sample (A). This occurs not just in the summits, but out to the broader area ±150bp. This conservation is even more increased when centering on the nearest CABVTG E-box, although the motif-centered random plot shows that CABVTG E-boxes in the Drosophila genome are preferentially conserved relative to the genomic background. The conservation of the 500 peaks added by dropping to the MC threshold is smaller overall (B), and the conservation of the additional 1,000 peaks from the LC threshold is even smaller (C). This may suggest that smaller peaks are less likely to be conserved or it may be a result of having more false positive peaks as the threshold is lowered.*



Supplemental Figure 15. Distribution of motifs within the sequenced input DNA (i.e. sonicated chromatin). *Twist ChIP-Seq regions are significantly depleted in highly A/T-rich sequences. This depletion is not specific to the ChIP because it is also observed for the input control chromatin library. Twist MC ChIP-Seq peaks are shown next to input control data of an equivalent number of regions (1099). See Supplemental methods for the origin of the different control samples shown. "Enriched input" contains regions selected as most significant from the input control over Twist. "Sequenced input reads" reads were randomly selected from all uniquely mapping reads in the input control. For Twist and enriched input, mapping is relative to the shifted summits. For the sequenced input reads, mapping is relative to the center of each 25bp read. Three motifs, the Twist explanatory E-box (CABVTG), AAAAAA [(A)6], and a string of any 16 A's or T's [(W)16] are shown for each dataset and compared to the overall G/C content (averaged in 20bp windows).*

SII.S2 METHODS

ChIP-chip experimental design and processing. Arrays from standard catalog of Roche Nimblegen were used for this experiment covering the entire *Drosophila melanogaster* genome. The set of three arrays (385,000 probes/array) contain 50-mer probes spaced by 48 nucleotides on the genome. Each array was hybridized with two samples - genomic control DNA labeled with Cy3 and experimental sample labeled with Cy5. Two samples were hybridized to the arrays: Twist and mock sample as control (i.e. pre-immune). Each measurement was performed using a single biological replicate. The hybridizations were performed at a Nimblegen facility, and both the raw data and Cy5/Cy3 ratios for each array (Cy5=635 nm, Cy3=532 nm) were made available to us for analysis.

Design of custom array for ChIP-chip experiment. A custom array (Nimblegen 4-plex technology, 72,000 probes,) was designed to confirm the above results and also probe the neighborhoods of high-confidence transcription factors in more detail. Two sets of probes were included in the array: (i) Probes were tiled (60 mer probes, 5 nucleotide spacing) within 6KB upstream and 1KB downstream of ATG sites of 288 high-confidence transcription factors in *Drosophila melanogaster.* The list of transcription factors is available on request; (ii) Probes were also tiled (60 mer probes, 5nucleotide spacing) within 1KB upstream and downstream of 1,600 peaks detected in the earlier ChIP-chip experiments. In total, the array contained 71,000 60-mer probes from the *D. melanogaster* genome and 1000 random sequences as control.

ChIP-chip bioinformatics. The data from all arrays were normalized using quantile normalization procedure. After normalization, ratios of Cy5/Cy3 were taken for each sample for further analysis. The original array design was based on V4 release of the *Drosophila* genome. Therefore, normalized data were mapped on to V5 genome assembly (dm3, April 2006) examined visually for validation.

ChIP-chip peak finding was conducted as previously described (MacArthur, Li et al. 2009). First, quantile normalized data for each probe was replaced by the mean

signal of all probes within +/-350 nucleotides from it. This smoothing step was performed in the logarithmic scale. All probes with normalized smoothed signal above 90th percentile in the array (normalized signal=2, high signal probes) were considered for further analysis. Multiple high signal neighboring probes (maximum gap 200 nucleotides) were combined into summits with height equal to the highest smoothed intensity within the region.

ChIP-Seq bioinformatics. Sequenced reads were trimmed to the first 25 base pairs and mapped onto the dm3 (April 2006, BGDP release 5) *Drosophila melanogaster* genome using bowtie 9.1 (Langmead et al., 2009). No more than two mismatches were allowed. Low-copy multireads (defined as reads mapping in 2 to 10 places) were allowed. Chromosomes U and the Het chromosomes were not used in the downstream analyses.

The ERANGE 3.1 software package was used to identify regions enriched in ChIP-Seq defined Twist occupancy. ERANGE finds areas in the genome that are densely occupied by reads and then identifies those that exceed signals in the background sample (sonicated input DNA) (Pepke, Wold et al. 2009). Regions that do not display proper left/right read directionality are discarded (see also main text). A custom code was used to computationally call a ChIP-Seq signal maximum location (the "shifted summit"), which introduced a shift in the position attributed as the "peak" based on the degree of read directionality. For simplicity, the shifted summit is reported as one nucleotide.

In order to get a broad view of what to expect based on the ChIP-Seq experimental assay as well as the bioinformatics assay, several different types of controls were used. For the genomic background, the dm3 genome was used minus UCSC simple and tandem repeats and minus the Chromosomes U and the Het chromosomes. In order to assay reads that could be sequenced, reads that mapped uniquely to the genome were selected at random ("sequenced control reads"). In order

to determine which places in the genome were sequenced well (“aggregated control”), ERANGE was run on the sonicated input DNA library requiring only two reads per region (no directionality requirement was used and no enrichment relative to another library was required). In order to determine which places in the genome displayed proper read directionality and were overrepresented in the sequenced input control library relative to twist (“enriched control”), ERANGE was run on the input DNA library vs. twist, requiring at least a 1% enrichment per region in the input DNA and a minimum of two reads per region. The directionality filter was used as for Twist regions and the peaks were subsequently shifted using the same algorithm as for the Twist peaks.

A second independent ChIP-Seq algorithm and software package, MACS 1.3.5 (Zhang, Liu et al. 2008), was also used on the same Twist and input control datasets, and we report both sets of "peak calls" (Supplemental Table 4). The effective genome size used was 1.69e8, tag size 25, band width 300, model fold 7, and P-value cutoff 1e-5. There were no major discrepancies between motif occurrences relative to ERANGE and MACS calls nor to the respective MEME outputs (data not shown).

Selection of confidence thresholds. None of the distributions of ChIP signals, under any algorithm, displayed a crisp natural discontinuity that would clearly define "occupied" versus "unoccupied" states. ERANGE was first run on ChIP-Seq data with a stringent gradient of parameters, and the different region sets were evaluated for sensitivity and specificity by their inclusion of (1) validated, functional Twist binding regions; (2) their overlap with an independent region calling algorithm, MACS and (3) the likelihood that the low-confidence end of the region sets were ‘real’ as judged by inspection of the read distribution in ChIP and background data. As a result, we set the ERANGE high-confidence (HC) signal and enrichment thresholds at 14 RPM minimum (reads in the region per million in the dataset), 1 RPM minimum peak height, and 3-fold

enrichment over the control sample), resulting in 513 regions (false discovery rate (FDR) <1%, where the ERANGE FDR reflects the relative number of peaks called when using the same parameters to call the control library over the twist library). Medium confidence (1099 peaks) and lower-confidence (2000 peaks) were called with the same enrichment ratio and minimum peak height but instead using region RPM thresholds of 4 (FDR 17%) and 2 (FDR 83%), respectively. The MC threshold was selected because of the similarity of motif distributions around peaks compared to the HC regions (Fig. 4A), and the LC threshold was selected primarily to demonstrate what happens when selecting a very low informatics threshold (shown in Figure 3A and Supplemental Figure 15).

For comparison sake, HC and MC sets of ChIP-chip regions were defined using equivalent FDR measures as found for ChIP-Seq. To this end, boundaries of ChIP-chip regions were defined using a threshold of 3.8 to identify 669 ChIP-chip regions (HC set; FDR<1%) and a threshold of 6 to identify 2013 ChIP-chip regions (MC set; FDR 17%). We report the MC region boundaries as well as the size and location of the "summit" of each region, defined as the midpoint of the highest part of each region (Supplemental Table 5).

As expected, the weaker ChIP-Seq signals are most numerous in their respective distributions (Sup. Fig. 2), which means that the computational threshold selected for inclusion has a large impact on subsequent VENN comparisons of Twist set membership. ChIP-chip processed data typically identified physically broader regions on the chromosome, partly because array processing algorithms require multiple positive tiles to make a signal call. Furthermore, the array data appear to compress the ChIP signal range compared with ChIP-Seq, bringing the strongest signal closer to the weakest one in the distribution and this, along with other technical differences, may account for the decrease in overlap observed when the HC ChIP-Seq set is compared with MC versus HC ChIP-chip sets (81% versus 54%).

Acquisition of SELEX data and processing. SELEX was performed according to a previously published method (Roulet, Busso et al. 2002) and a standard SAGE protocol

(http://www.sagenet.org/protocol/index.htm) with some exceptions, as follows (for further details see Ogawa 2011). 72 bp DNA oligoes were synthesized with three different end pairs each containing a restriction enzyme site (*Bam*HI, *Bgl*II, or *Hin*dIII) and 20 bp priming sequences for PCR amplification:

Random72: GGATTTGCTGGTGCAGTACAGT-GGATCC-$(N)_{16}$-GGATCC-TTAGGAGCTTGAAATCGAGCAG

Random72R: TCCATCGCTTCTGTATGACGCA-AGATCT-$(N)_{16}$-AGATCT-GTCCTAACCGACTCCGTTGATT

Random72HR: TCCATCGCTTCTGTATGACGCA-AAGCTT-$(N)_{16}$-AAGCTT-GTCCTAACCGACTCCGTTGATT

His-tagged Twist protein was bound to TALON Metal Affinity Resins (Clontech). For the first round of SELEX, 10 ng of the random 72 bp ds DNA oligonucleotides was incubated with the protein bound resin. The input DNA for subsequent rounds of SELEX was derived by PCR amplifying $1/10^{th}$ of the DNA eluted from the previous round.

For all rounds, SELEX-bound DNA was amplified by PCR according to SAGE protocol and then digested with the appropriate restriction enzyme to isolate the 22 bp fragment which includes the Twist-binding sequence. Approximately 1 µg of the 22 bp DNA fragments were ligated to make concatamers in a 10 µl volume at 16˚C overnight. The concatemer DNA was treated with T4 DNA polymerase (NEB) and DNA polymerase I Klenow fragment (NEB) with dNTP mixture at room temperature for 30 min. After heat-inactivation at 65˚C for 5 min, the DNA was separated by 2% agarose (Invitrogen, UltraPure agarose) gel electrophoresis. DNA of 300 to 1000 bp was isolated from the

gel and purified by using QIAquick Gel purification kit (Qiagen). The resulting concatemer DNA was ligated with *Sma*I-digested pUC19 plasmid, and subsequently the ligation mixture was used to transform DH10B E. coli (Invitrogen ElectroMAX cells). Plasmid DNAs from more than 96 clones were sequenced to obtain sequences of over 1,000 individual DNAs. The data presented are 17bp reads, on average (Supplemental Table 6).

Two SELEX experiments were performed to analyze the binding preference for Twist. Each involved 5 rounds of amplifications for a total of 10 total datasets. For experiment one, rounds 4 and 5 were sequenced; for experiment two, rounds 2,3, and 4 were sequenced. The data for these 5 rounds were pooled, and the number of E-boxes in the entire dataset was counted (Figure 2). MEME was run on the SELEX sequences, and in addition to the CATATG/CACATG E-box, an –AYRTG sequence (suggesting a partial E-box) was also returned (data not shown). E-boxes are present in approximately 50% of the SELEX sequences and of the remaining 50%, the majority contain a partial (5-mer) E-box. This may be due to the enzyme cut sites and sequencing or possibly to Twist binding a partial E-box. We see no such representation of the partial E-boxes at ChIP-Seq *in vitro* peaks.

MEME analysis. MEME was run on the MC Twist ChIP-Seq ERANGE regions ±50bp from the peaks (i.e. “shifted summits”) in order to capture the pieces of DNA that show the highest enrichment of explanatory E-boxes (Fig. 2, Fig. 3, Sup. Fig. 8). MEME 3.0.8 was used, using the ”zoops” model, 6 bp minimum, and 15 bp maximum motif widths. MEME finds sequences that are similar to each other but statistically unlikely to be found in the local background of the sample (Bailey, Williams et al. 2006). The MEME results were mapped at 85% match to the output PSFM’s onto the parent set of

Twist regions or the control datasets (Fig. 5, Sup. Fig. 11).

Motif mapping. Scatter plots were made in order to visualize the distribution of motifs relative to Twist peaks (i.e. “shifted summits”). Motifs were mapped on to the genome, and each dot on a scatter plot reflects the distance between the center of the motif and its respective Twist peak. Negative values are to the left of the peaks in the reference genome, and positive numbers are to the right.

Density plots (i.e. Fig. 3B, top panel) were made by taking the absolute distance of each motif from its peak and then summing for the entire dataset the number of motifs in 5bp windows outward from the peaks. Cumulative density plots (i.e. Fig. 3B, bottom panel, Sup. Fig. 7) are another way of reporting the data in the density plots, where the cumulative fraction of the motifs represented in each 5bp window in (from 0 total motifs found at the peak to 100% of the motifs encountered at the maximum 250bp distance from the peak).

A Kolmogorov-Smirnov (K-S) statistical test was performed to determine whether motifs were enriched, depleted or uniformly distributed relative to the set of Twist peaks. This method tests the null hypothesis that a distribution of motif distances relative to Twist peaks is distributed uniformly. Distributions of these distances for motifs that are unrelated to binding are expected to be statistically similar to the uniform distribution; those that are related to binding are expected to be different from uniform. The statistic for testing these hypotheses is the maximum distance between the empirical cumulative distribution function of the distances between motifs and peaks and the cumulative distribution function of a uniform. This distance is known as the "D" value (D values and both types of distributions are illustrated in Supplemental Figure 7). Thus we can obtain P-values for the probability of the null hypothesis and reject the null hypothesis when the

P-value is too small. All regions were made equal length (±250 bp around each peak) for these tests. A small P-value (threshold $1*10^{-3}$) means that a motif distribution is not significantly different from uniform and is instead enriched or depleted relative to Twist peaks.

To relate the K-S test results to a more familiar statistic, we also performed a Student's T-test. The T-test is used here to test whether the mean of the observed motif distance from the peak is equal to the mean of the assumed uniform distribution on the standardized regions. Since we standardized the maximum distance from the peak to 250bp, the mean is 125bp, and so the T statistic reports whether the mean of each motif is different from 125bp. Note that it is possible to have a distribution quite different in shape from the uniform distribution and still have the same mean. The K-S test would determine that the two are significantly different while the T-test would not. In this sense, the K-S test is more powerful than the T-test. In any case, the statistical conclusions from the T-test and the K-S test agree for our observed distributions (see the P-values for both tests in Supplementary Table 3). P-values reported are in base 10 (i.e. 2.2E-16 means $2.2*10^{-16}$)

Genome location analysis. The gene models we used were primarily based on published FlyBase introns and exons but were additionally informed by a set of promoters active in the embryo (generously provided by S.Celniker). We used these data to class the genome into four mutually exclusive categories. "Promoter proximal" refers to any summit that occurs within a Celniker promoter or 500 bp upstream. "Exonic" refers to any FlyBase exon excluding any regions that fall into the promoter proximal category. "Intronic" regions are any regions within the gene body (from FlyBase TSS or Celniker promoter, whichever is upstream, to the last exon) that are not

in the exonic or promoter proximal categories. Intergenic regions are outside of gene bodies and had repeats (from UCSC tandem repeats and repeat masker) removed.

In order to accurately represent the nature of the ChIP-Seq input control data, we used it in three different ways. “Random sequenced input reads” is a set of reads from the input control that map uniquely to the genome. It represents the areas of the non-repeat genome which are able to be sonicated and sequenced. “Aggregated input control” regions were created by allowing ERANGE to run on the input control without a directionality filter or an enrichment requirement. These regions represent places in the genome that have an aggregation of input reads but no other requirements that the reads behave similarly to ChIP-Seq peaks. The “enriched input control” contains regions where the input control library is enriched over Twist and also displays the same left/right read directionality required for Twist (see also main text) .

The number of ChIP-chip and control regions in each dataset was picked to be the same number as MC Twist regions. We chose the largest ChIP-chip and aggregated control regions (by area), the enriched control regions that were most highly enriched over Twist, and a random sample of sequenced control reads. In order to assign regions to each genomic category, we used the shifted summits of Twist ChIP-Seq and enriched control regions, the highest point of the aggregated control regions, the ChIP-chip mock summit (midpoint of the highest part of each regions), and the midpoint of each randomly selected sequenced control read.

Motif conservation analysis. PhastCons scores were obtained (as described in the text) for all base pairs for motif occurrences within +/- 150bp of ChIP-Seq summits and also for those greater than 150 bp but less than 250 bp away from the summits. Number of ChIP-Seq region occurrences for each were CACATG: 396, CACCTG: 74,

CACGTG: 63, CAGATG: 173, CAGCTG: 139, CATATG: 105, CA-repeats (3 or more dyads): 610, and GA-repeats (3 or more dyads): 255. A chi squared statistic corresponding to a one-tailed test for a difference between the two distributions was calculated according to the procedure given in Kanji (Kanji 1999 p.83). The two sample sets were first joined and the median for the combined set calculated. The number of PhastCons scores of the background set that were to the left of the combined set median was calculated and designated nl1; the number to the right of the combined median is designated nr1. The two analogous quantities for the ChIP-Seq region motif set were designated nl2 and nr2 with N = nl1+nr1+nl2+nr2. Then the chi squared statistic is calculated as:

$$N*(| nl1*nr2 – nl2*nr1| - N/2)^2 / ((nl1+nl2)*(nl1+nr1)*(nl2+nr2)*(nr1+nr2))$$

The x-axis in Fig. 7C represents this test statistic for each motif. Because PhastCons scores are the posterior probability of a given bp to belong to a conserved class of bases, we interpret bp with PhastCons scores > 0.9 as almost certainly conserved. The fraction of bp in ChIP-Seq motifs having PhastCons score > 0.9 is represented as the height of the bars.

Sources for SII Supplements

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz and D. Duboule (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." Science **340**(6137): 1234167.

Arnosti, D. N. and M. M. Kulkarni (2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" J Cell Biochem **94**(5): 890-898.

Asakura, A., G. E. Lyons and S. J. Tapscott (1995). "The regulation of MyoD gene expression: conserved elements mediate expression in embryonic axial muscle." Dev Biol **171**(2): 386-398.

Auerbach, R. K., G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefrancois, K. Struhl, M. Gerstein and M. Snyder (2009). "Mapping accessible chromatin regions using Sono-Seq." Proc Natl Acad Sci U S A **106**(35): 14926-14931.

Bailey, A. M. and J. W. Posakony (1995). "Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity." Genes Dev **9**(21): 2609-2622.

Bailey, S. D., X. Zhang, K. Desai, M. Aid, O. Corradin, R. Cowper-Sal Lari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains and M. Lupien (2015). "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters." Nat Commun **2**: 6186.

Bailey, T. L., N. Williams, C. Misleh and W. W. Li (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-373.

Banerji, J., S. Rusconi and W. Schaffner (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Becker, P., R. Renkawitz and G. Schutz (1984). "Tissue-specific DNaseI hypersensitive sites in the 5'-flanking sequences of the tryptophan oxygenase and the tyrosine aminotransferase genes." EMBO J **3**(9): 2015-2020.

Benoist, C. and P. Chambon (1981). "In vivo sequence requirements of the SV40 early promotor region." Nature **290**(5804): 304-310.

Benyajati, C. and A. Worcel (1976). "Isolation, characterization, and structure of the folded interphase genome of Drosophila melanogaster." Cell **9**(3): 393-407.

Berezney, R. and D. S. Coffey (1974). "Identification of a nuclear protein matrix." Biochem Biophys Res Commun **60**(4): 1410-1417.

Berghella, L., L. De Angelis, T. De Buysscher, A. Mortazavi, S. Biressi, S. V. Forcales, D. Sirabella, G. Cossu and B. J. Wold (2008). "A highly conserved molecular switch binds MSY-3 to regulate myogenin repression in postnatal muscle." Genes & Development **22**(15): 2125-2138.

Blau, H. M., C. P. Chiu and C. Webster (1983). "Cytoplasmic activation of human nuclear genes in stable heterocaryons." Cell **32**(4): 1171-1180.

Blau, H. M., G. K. Pavlath, E. C. Hardeman, C. P. Chiu, L. Silberstein, S. G. Webster, S. C. Miller and C. Webster (1985). "Plasticity of the differentiated state." Science **230**(4727): 758-766.

Branco, M. R. and A. Pombo (2006). "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations." PLoS Biol **4**(5): e138.

Breathnach, R. and P. Chambon (1981). "Organization and expression of eucaryotic split genes coding for proteins." Annu Rev Biochem **50**: 349-383.

Brent, R. and M. Ptashne (1984). "A bacterial repressor protein or a yeast transcriptional terminator can block upstream activation of a yeast gene." Nature **312**(5995): 612-615.

Brown, K. E., S. S. Guest, S. T. Smale, K. Hahm, M. Merkenschlager and A. G. Fisher (1997). "Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin." Cell **91**(6): 845-854.

Buckingham, M. and P. W. Rigby (2014). "Gene regulatory networks and transcriptional mechanisms that control myogenesis." Dev Cell **28**(3): 225-238.

Buonanno, A., D. G. Edmondson and W. P. Hayes (1993). "Upstream sequences of the myogenin gene convey responsiveness to skeletal muscle denervation in transgenic mice." Nucleic Acids Res **21**(24): 5684-5693.

Burke, T. W. and J. T. Kadonaga (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters." Genes Dev **10**(6): 711-724.

Burke, T. W. and J. T. Kadonaga (1997). "The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila." Genes Dev **11**(22): 3020-3031.

Butler, J. E. and J. T. Kadonaga (2001). "Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs." Genes Dev **15**(19): 2515-2519.

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

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Carvajal, J. J., D. Cox, D. Summerbell and P. W. Rigby (2001). "A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and maintenance of gene expression during muscle development." Development **128**(10): 1857-1868.

Casas-Delucchi, C. S., A. Brero, H. P. Rahn, I. Solovei, A. Wutz, T. Cremer, H. Leonhardt and M. C. Cardoso (2011). "Histone acetylation controls the inactive X chromosome replication dynamics." Nat Commun **2**: 222.

Chambeyron, S. and W. A. Bickmore (2004). "Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription." Genes Dev **18**(10): 1119-1130.

Chen, J. C. J., R. Ramachandran and D. J. Goldhamer (2002). "Essential and Redundant Functions of the MyoD Distal Regulatory Region Revealed by Targeted Mutagenesis." Developmental Biology **245**(1): 213-223.

Cheng, T. C., M. C. Wallace, J. P. Merlie and E. N. Olson (1993). "Separable regulatory elements governing myogenin transcription in mouse embryogenesis." Science **261**(5118): 215-218.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang and K. Zhao (2012). "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of

interacting genes and modes of higher order chromatin organization." Cell Res **22**(3): 490-503.
Cheutin, T., M. F. O'Donohue, A. Beorchia, C. Klein, H. Kaplan and D. Ploton (2003). "Three-dimensional organization of pKi-67: a comparative fluorescence and electron tomography study using FluoroNanogold." J Histochem Cytochem **51**(11): 1411-1423.
Chubb, J. R., S. Boyle, P. Perry and W. A. Bickmore (2002). "Chromatin motion is constrained by association with nuclear compartments in human cells." Current Biology **12**(6): 439-445.
Chung, J. H., A. C. Bell and G. Felsenfeld (1997). "Characterization of the chicken beta-globin insulator." Proc Natl Acad Sci U S A **94**(2): 575-580.
Chung, J. H., M. Whiteley and G. Felsenfeld (1993). "A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila." Cell **74**(3): 505-514.
Ciejek, E. M., M. J. Tsai and B. W. O'Malley (1983). "Actively transcribed genes are associated with the nuclear matrix." Nature **306**(5943): 607-609.
Cook, P. R. (1999). "The organization of replication and transcription." Science **284**(5421): 1790-1795.
Cook, P. R. and I. A. Brazell (1978). "Spectrofluorometric measurement of the binding of ethidium to superhelical DNA from cell nuclei." Eur J Biochem **84**(2): 465-477.
Cooper, S. J., N. D. Trinklein, E. D. Anton, L. Nguyen and R. M. Myers (2006). "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome." Genome Res **16**(1): 1-10.
Core, L. J. and J. T. Lis (2009). "Paused Pol II captures enhancer activity and acts as a potent insulator." Genes Dev **23**(14): 1606-1612.
Courey, A. J., S. E. Plon and J. C. Wang (1986). "The use of psoralen-modified DNA to probe the mechanism of enhancer action." Cell **45**(4): 567-574.
Cremer, T., C. Cremer, T. Schneider, H. Baumann, L. Hens and M. Kirsch-Volders (1982). "Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments." Hum Genet **62**(3): 201-209.
Davidson, E. H. (2006). CHAPTER 2 - cis-Regulatory Modules, and the Structure/Function Basis of Regulatory Logic. The Regulatory Genome. Burlington, Academic Press**:** 31-86.
de Wit, E. and W. de Laat (2012). "A decade of 3C technologies: insights into nuclear organization." Genes Dev **26**(1): 11-24.
Deato, M. D., M. T. Marr, T. Sottero, C. Inouye, P. Hu and R. Tjian (2008). "MyoD targets TAF3/TRF3 to activate myogenin transcription." Mol Cell **32**(1): 96-105.
Deato, M. D. and R. Tjian (2007). "Switching of the core transcription machinery during myogenesis." Genes Dev **21**(17): 2137-2149.
Deng, W. and S. G. Roberts (2005). "A core promoter element downstream of the TATA box that is recognized by TFIIB." Genes Dev **19**(20): 2418-2423.
Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser and C. Weissmann (1983). "Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells." Cell **32**(3): 695-706.
Dierks, P., A. van Ooyen, N. Mantei and C. Weissmann (1981). "DNA sequences preceding the rabbit beta-globin gene are required for formation in mouse L cells of beta-globin RNA with the correct 5' terminus." Proc Natl Acad Sci U S A **78**(3): 1411-1415.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature **485**(7398): 376-380.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green and J. Dekker (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." Genome Res **16**(10): 1299-1309.

Dynan, W. S. (1986). "Promoters for housekeeping genes." Trends in Genetics **2**: 196-197.

Faerman, A., D. J. Goldhamer, R. Puzis, C. P. Emerson, Jr. and M. Shani (1995). "The distal human myoD enhancer sequences direct unique muscle-specific patterns of lacZ expression during mouse development." Dev Biol **171**(1): 27-38.

Farnham, P. J. and R. T. Schimke (1985). "Transcriptional regulation of mouse dihydrofolate-reductase in the cell-cycle." Journal of Biological Chemistry **260**(12): 7675-7680.

Faye, G., D. W. Leung, K. Tatchell, B. D. Hall and M. Smith (1981). "Deletion mapping of sequences essential for in vivo transcription of the iso-1-cytochrome c gene." Proc Natl Acad Sci U S A **78**(4): 2258-2262.

Fiering, S., E. Epner, K. Robinson, Y. Zhuang, A. Telling, M. Hu, D. I. Martin, T. Enver, T. J. Ley and M. Groudine (1995). "Targeted deletion of 5'HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus." Genes Dev **9**(18): 2203-2213.

Filippova, D., R. Patro, G. Duggal and C. Kingsford (2014). "Identification of alternative topological domains in chromatin." Algorithms Mol Biol **9**: 14.

Fisher-Aylor, K. I. (2011). "Long distance looping maps: RNA Pol2 during differentiation." Nuclear Structure and Dynamics. L'Isle sur la Sorgue, France, EMBO.

Foley, K. P. and J. D. Engel (1992). "Individual stage selector element mutations lead to reciprocal changes in beta- vs. epsilon-globin gene transcription: genetic confirmation of promoter competition during globin gene switching." Genes Dev **6**(5): 730-744.

Francastel, C., M. C. Walters, M. Groudine and D. I. Martin (1999). "A functional enhancer suppresses silencing of a transgene and prevents its localization close to centrometric heterochromatin." Cell **99**(3): 259-269.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung and Y. Ruan (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." Nature **462**(7269): 58-64.

Galande, S., P. K. Purbey, D. Notani and P. P. Kumar (2007). "The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1." Curr Opin Genet Dev **17**(5): 408-414.

Gasser, S. M. and U. K. Laemmli (1986). "The organisation of chromatin loops: characterization of a scaffold attachment site." EMBO J **5**(3): 511-518.

Gasser, S. M. and U. K. Laemmli (1987). "Improved methods for the isolation of individual and clustered mitotic chromosomes." Exp Cell Res **173**(1): 85-98.

Gerasimova, T. I. and V. G. Corces (1998). "Polycomb and trithorax group proteins mediate the function of a chromatin insulator." Cell **92**(4): 511-521.

Gerasimova, T. I., D. A. Gdula, D. V. Gerasimov, O. Simonova and V. G. Corces (1995). "A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation." Cell **82**(4): 587-597.

Gidoni, D., W. S. Dynan and R. Tjian (1984). "Multiple specific contacts between a mammalian transcription factor and its cognate promoters." Nature **312**(5993): 409-413.

Gillies, S. D., S. L. Morrison, V. T. Oi and S. Tonegawa (1983). "A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene." Cell **33**(3): 717-728.

Gluzman, Y., J. F. Sambrook and R. J. Frisque (1980). "Expression of early genes of origin-defective mutants of simian virus 40." Proc Natl Acad Sci U S A **77**(7): 3898-3902.

Goldhamer, D. J., B. P. Brunk, A. Faerman, A. King, M. Shani and C. P. Emerson, Jr. (1995). "Embryonic activation of the myoD gene is regulated by a highly conserved distal control element." Development **121**(3): 637-649.

Goldman, M. A., G. P. Holmquist, M. C. Gray, L. A. Caston and A. Nag (1984). "Replication timing of genes and middle repetitive sequences." Science **224**(4650): 686-692.

Greally, J. M., D. J. Starr, S. Hwang, L. Song, M. Jaarola and S. Zemel (1998). "The mouse H19 locus mediates a transition between imprinted and non-imprinted DNA replication patterns." Hum Mol Genet **7**(1): 91-95.

Griffith, J., A. Hochschild and M. Ptashne (1986). "DNA loops induced by cooperative binding of lambda repressor." Nature **322**(6081): 750-752.

Grosschedl, R. and M. L. Birnstiel (1980). "Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo." Proc Natl Acad Sci U S A **77**(3): 1432-1436.

Grosschedl, R. and M. L. Birnstiel (1980). "Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo." Proc Natl Acad Sci U S A **77**(12): 7102-7106.

Gruss, P., R. Dhar and G. Khoury (1981). "Simian virus 40 tandem repeated sequences as an element of the early promoter." Proc Natl Acad Sci U S A **78**(2): 943-947.

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat and B. van Steensel (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Hakim, O., M. H. Sung, T. C. Voss, E. Splinter, S. John, P. J. Sabo, R. E. Thurman, J. A. Stamatoyannopoulos, W. de Laat and G. L. Hager (2011). "Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements." Genome Res **21**(5): 697-706.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan and C. L. Wei (2011). "CTCF-mediated functional chromatin interactome in pluripotent cells." Nat Genet **43**(7): 630-638.

Harr, J. C., T. R. Luperchio, X. Wong, E. Cohen, S. J. Wheelan and K. L. Reddy (2015). "Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins." J Cell Biol **208**(1): 33-52.

Hart, D. O., T. Raha, N. D. Lawson and M. R. Green (2007). "Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3." Nature **450**(7172): 1082-1085.

Hebbes, T. R., A. L. Clayton, A. W. Thorne and C. Crane-Robinson (1994). "Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain." EMBO J **13**(8): 1823-1830.

Hendrix, D. A., J. W. Hong, J. Zeitlinger, D. S. Rokhsar and M. S. Levine (2008). "Promoter elements associated with RNA Pol II stalling in the Drosophila embryo." Proc Natl Acad Sci U S A **105**(22): 7762-7767.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke and R. A. Young (2013). "Super-enhancers in the control of cell identity and disease." Cell **155**(4): 934-947.

Hochschild, A., N. Irwin and M. Ptashne (1983). "Repressor structure and the mechanism of positive control." Cell **32**(2): 319-325.

Hong, J. W., D. A. Hendrix and M. S. Levine (2008). "Shadow enhancers as a source of evolutionary novelty." Science **321**(5894): 1314.

Hsu, J. Y., T. Juven-Gershon, M. T. Marr, 2nd, K. J. Wright, R. Tjian and J. T. Kadonaga (2008). "TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription." Genes Dev **22**(17): 2353-2358.

Htun, H., J. Barsony, I. Renyi, D. L. Gould and G. L. Hager (1996). "Visualization of glucocorticoid receptor translocation and intranuclear organization in living cells with a green fluorescent protein chimera." Proc Natl Acad Sci U S A **93**(10): 4845-4850.

Hughes, S. M., J. M. Taylor, S. J. Tapscott, C. M. Gurley, W. J. Carter and C. A. Peterson (1993). "Selective accumulation of MyoD and myogenin mRNAs in fast and slow adult skeletal muscle is controlled by innervation and hormones." Development **118**(4): 1137-1147.

Iborra, F. J., A. Pombo, D. A. Jackson and P. R. Cook (1996). "Active RNA polymerases are localized within discrete transcription "factories' in human nuclei." J Cell Sci **109 (Pt 6)**: 1427-1436.

Ip, Y. T., R. E. Park, D. Kosman, E. Bier and M. Levine (1992). "The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive …." Genes Dev **6**: 1728-1739.

Jackson, D. A., A. B. Hassan, R. J. Errington and P. R. Cook (1993). "Visualization of focal sites of transcription within human nuclei." EMBO J **12**(3): 1059-1065.

Jackson, D. A., F. J. Iborra, E. M. Manders and P. R. Cook (1998). "Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei." Mol Biol Cell **9**(6): 1523-1536.

Johnson, D. S., A. Mortazavi, R. M. Myers and B. Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.

Juven-Gershon, T., J. Y. Hsu and J. T. Kadonaga (2006). "Perspectives on the RNA polymerase II core promoter." Biochem Soc Trans **34**(Pt 6): 1047-1050.

Juven-Gershon, T., J. Y. Hsu and J. T. Kadonaga (2008). "Caudal, a key developmental regulator, is a DPE-specific transcriptional factor." Genes Dev **22**(20): 2823-2830.

Juven-Gershon, T., J. Y. Hsu, J. W. Theisen and J. T. Kadonaga (2008). "The RNA polymerase II core promoter - the gateway to transcription." Curr Opin Cell Biol **20**(3): 253-259.

Juven-Gershon, T. and J. T. Kadonaga (2010). "Regulation of gene expression via the core promoter and the basal transcriptional machinery." Developmental Biology **339**(2): 225-229.

Kablar, B., A. Asakura, K. Krastel, C. Ying, L. L. May, D. J. Goldhamer and M. A. Rudnicki (1998). "MyoD and Myf-5 define the specification of musculature of distinct embryonic origin." Biochem Cell Biol **76**(6): 1079-1091.

Kablar, B., K. Krastel, C. Ying, A. Asakura, S. J. Tapscott and M. A. Rudnicki (1997). "MyoD and Myf-5 differentially regulate the development of limb versus trunk skeletal muscle." Development **124**(23): 4729-4738.

Kanji, G. K. (1999). 100 Statistical Tests, SAGE Publications Ltd., London, England.

Kellum, R. and P. Schedl (1991). "A position-effect assay for boundaries of higher order chromosomal domains." Cell **64**(5): 941-950.

Kennell, D. and H. Riezman (1977). "Transcription and translation initiation frequencies of the Escherichia coli lac operon." Journal of Molecular Biology **114**(1): 1-21.

Kerem, B. S., R. Goitein, G. Diamond, H. Cedar and M. Marcus (1984). "Mapping of DNAase I sensitive regions on mitotic chromosomes." Cell **38**(2): 493-499.

Kieffer-Kwon, K. R., Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grontved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan and R. Casellas (2013). "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation." Cell **155**(7): 1507-1520.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov and B. Ren (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green and B. Ren (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-880.

Koch, F., R. Fenouil, M. Gut, P. Cauchy, T. K. Albert, J. Zacarias-Cabeza, S. Spicuglia, A. L. de la Chapelle, M. Heidemann, C. Hintermair, D. Eick, I. Gut, P. Ferrier and J. C. Andrau (2011). "Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters." Nat Struct Mol Biol **18**(8): 956-963.

Kosak, S. T. and M. Groudine (2004). "Gene order and dynamic domains." Science **306**(5696): 644-647.

Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher and H. Singh (2002). "Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development." Science **296**(5565): 158-162.

Kramer, H., M. Niemoller, M. Amouyal, B. Revet, B. von Wilcken-Bergmann and B. Muller-Hill (1987). "lac repressor forms loops with linear DNA carrying two suitably spaced lac operators." EMBO J **6**(5): 1481-1491.

Krebs, J. E. and M. Dunaway (1998). "The scs and scs' insulator elements impart a cis requirement on enhancer-promoter interactions." Mol Cell **1**(2): 301-308.

Kuntz, S. G., B. A. Williams, P. W. Sternberg and B. J. Wold (2012). "Transcription factor redundancy and tissue-specific regulation: evidence from functional and physical network connectivity." Genome Res **22**(10): 1907-1919.

Kutach, A. K. and J. T. Kadonaga (2000). "The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters." Mol Cell Biol **20**(13): 4754-4764.

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg and R. H. Ebright (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB." Genes Dev **12**(1): 34-44.

Lebkowski, J. S. and U. K. Laemmli (1982). "Evidence for two levels of DNA folding in histone-depleted HeLa interphase nuclei." J Mol Biol **156**(2): 309-324.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. Ruan (2012). "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." Cell **148**(1-2): 84-98.

Li, G. L., X. A. Ruan, R. K. Auerbach, K. S. Sandhu, M. Z. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Y. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Z. Hong, Z. Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. H. Ge, H. E. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder and Y. J. Ruan (2012). "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation." Cell **148**(1-2): 84-98.

Li, Y., W. Huang, L. Niu, D. M. Umbach, S. Covo and L. Li (2013). "Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes." BMC Genomics **14**: 553.

Liberman, L. M. and A. Stathopoulos (2009). "Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence." Dev Biol **327**(2): 578-589.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Lim, C. Y., B. Santoso, T. Boulay, E. Dong, U. Ohler and J. T. Kadonaga (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-1617.

Ling, J. Q., T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry and A. R. Hoffman (2006). "CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1." Science **312**(5771): 269-272.

Lis, J., L. Core, A. Martins, C. Danko, A. Siepel, G. Booth, F. Duarte and D. B. Mahat (2015). "A Unified Model Describing The Architecture And Creation Of Promoters And Enhancers." The FASEB Journal **29**(1 Supplement).

Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin and K. A. Frazer (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-140.

Lupo, A., E. Cesaro, G. Montano, D. Zurlo, P. Izzo and P. Costanzo (2013). "KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions." Current Genomics **14**(4): 268-278.

MacArthur, S., X. Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. Keranen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin and M. B. Eisen (2009). "Developmental roles of 21 Drosophila

transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions." Genome Biol **10**(7): R80.
Magistri, M., M. A. Faghihi, G. St Laurent, 3rd and C. Wahlestedt (2012). "Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts." Trends Genet **28**(8): 389-396.
Mallin, D. R., J. S. Myung, J. S. Patton and P. K. Geyer (1998). "Polycomb group repression is blocked by the Drosophila suppressor of Hairy-wing [su(Hw)] insulator." Genetics **148**(1): 331-339.
Markstein, M., P. Markstein, V. Markstein and M. S. Levine (2002). "Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo." Proceedings of the National Academy of Sciences **99**(2): 763.
Markstein, M., R. Zinzen, P. Markstein, K. P. Yee, A. Erives, A. Stathopoulos and M. Levine (2004). "A regulatory code for neurogenic gene expression in the Drosophila embryo." Development **131**(10): 2387-2394.
McKnight, S. and R. Tjian (1986). "Transcriptional selectivity of viral genes in mammalian cells." Cell **46**(6): 795-805.
McKnight, S. L., R. C. Kingsbury, A. Spence and M. Smith (1984). "The distal transcription signals of the herpesvirus tk gene share a common hexanucleotide control sequence." Cell **37**(1): 253-262.
McNally, J. G., W. G. Müller, D. Walker, R. Wolford and G. L. Hager (2000). "The Glucocorticoid Receptor: Rapid Exchange with Regulatory Sites in Living Cells." Science **287**(5456): 1262-1265.
Mercola, M., X. F. Wang, J. Olsen and K. Calame (1983). "Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus." Science **221**(4611): 663-665.
Meshorer, E. and T. Misteli (2006). "Chromatin in pluripotent embryonic stem cells and differentiation." Nat Rev Mol Cell Biol **7**(7): 540-546.
Mirkovitch, J., M. E. Mirault and U. K. Laemmli (1984). "Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold." Cell **39**(1): 223-232.
Mitchell, J. A. and P. Fraser (2008). "Transcription factories are nuclear subcompartments that remain in the absence of transcription." Genes Dev **22**(1): 20-25.
Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, V. Afzal, H. Hadeiba, K. Shinkai, E. M. Rubin and R. M. Locksley (2001). "Deletion of a coordinate regulator of type 2 cytokine expression in mice." Nat Immunol **2**(9): 842-847.
Monroe, R. J., B. P. Sleckman, B. C. Monroe, B. Khor, S. Claypool, R. Ferrini, L. Davidson and F. W. Alt (1999). "Developmental regulation of TCR delta locus accessibility and expression by the TCR delta enhancer." Immunity **10**(5): 503-513.
Morcillo, P., C. Rosen, M. K. Baylies and D. Dorsett (1997). "Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in Drosophila." Genes Dev **11**(20): 2729-2740.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.
Muller, H. P. and W. Schaffner (1990). "Transcriptional enhancers can act in trans." Trends in Genetics **6**(9): 300-304.

Muller, H. P., J. M. Sogo and W. Schaffner (1989). "An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge." Cell **58**(4): 767-777.

Muller, M. M., T. Gerster and W. Schaffner (1988). "Enhancer sequences and the regulation of gene-transcription." European Journal of Biochemistry **176**(3): 485-495.

Nabirochkin, S., M. Ossokina and T. Heidmann (1998). "A nuclear matrix/scaffold attachment region co-localizes with the gypsy retrotransposon insulator sequence." J Biol Chem **273**(4): 2473-2479.

Nakagomi, K., Y. Kohwi, L. A. Dickinson and T. Kohwi-Shigematsu (1994). "A novel DNA-binding motif in the nuclear matrix attachment DNA-binding protein SATB1." Mol Cell Biol **14**(3): 1852-1860.

Namciu, S. J., K. B. Blochlinger and R. E. Fournier (1998). "Human matrix attachment regions insulate transgene expression from chromosomal position effects in Drosophila melanogaster." Mol Cell Biol **18**(4): 2382-2391.

Neuberger, M. S. (1983). "Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells." EMBO J **2**(8): 1373-1378.

Noordermeer, D., M. R. Branco, E. Splinter, P. Klous, W. van Ijcken, S. Swagemakers, M. Koutsourakis, P. van der Spek, A. Pombo and W. de Laat (2008). "Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region." PLoS Genet **4**(3): e1000016.

Noordermeer, D., E. de Wit, P. Klous, H. van de Werken, M. Simonis, M. Lopez-Jones, B. Eussen, A. de Klein, R. H. Singer and W. de Laat (2011). "Variegated gene expression caused by cell-specific long-range DNA interactions." Nat Cell Biol **13**(8): 944-951.

Noordermeer, D. and D. Duboule (2013). "Chromatin looping and organization at developmentally regulated gene loci." Wiley Interdiscip Rev Dev Biol **2**(5): 615-630.

Ogawa, N., and Biggin, M. D. (2011). High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. Methods in Mol. Biol. . B. Deplanke, Humana Press, Clifton, New Jersey**:** in press.

Ohtsuki, S., M. Levine and H. N. Cai (1998). "Different core promoters possess distinct regulatory activities in the Drosophila embryo." Genes Dev **12**(4): 547-556.

Ong, C. T. and V. G. Corces (2014). "CTCF: an architectural protein bridging genome topology and function." Nat Rev Genet **15**(4): 234-246.

Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik and P. Fraser (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Ozdemir, A., K. I. Fisher-Aylor, S. Pepke, M. Samanta, L. Dunipace, K. McCue, L. C. Zeng, N. Ogawa, B. J. Wold and A. Stathopoulos (2011). "High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation." Genome Research **21**(4): 566-577.

Palstra, R. J., M. Simonis, P. Klous, E. Brasset, B. Eijkelkamp and W. de Laat (2008). "Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription." PLoS One **3**(2): e1661.

Panne, D., T. Maniatis and S. C. Harrison (2007). "An atomic model of the interferon-beta enhanceosome." Cell **129**(6): 1111-1123.

Pepke, S., B. Wold and A. Mortazavi (2009). "Computation for ChIP-seq and RNA-seq studies." Nat Methods **6**(11 Suppl): S22-32.

Peric-Hupkes, D., W. Meuleman, L. Pagie, S. W. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels and B. van Steensel (2010). "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation." Mol Cell **38**(4): 603-613.

Phillips-Cremins, J. E., M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor and V. G. Corces (2013). "Architectural protein subclasses shape 3D organization of genomes during lineage commitment." Cell **153**(6): 1281-1295.

Pickersgill, H., B. Kalverda, E. de Wit, W. Talhout, M. Fornerod and B. van Steensel (2006). "Characterization of the Drosophila melanogaster genome at the nuclear lamina." Nat Genet **38**(9): 1005-1014.

Plon, S. E. and J. C. Wang (1986). "Transcription of the human beta-globin gene is stimulated by an SV40 enhancer to which it is physically linked but topologically uncoupled." Cell **45**(4): 575-580.

Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. S. Wu, O. Denas, D. L. Vera, Y. L. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren and D. M. Gilbert (2014). "Topologically associating domains are stable units of replication-timing regulation." Nature **515**(7527): 402-+.

Reeve, J. N. (2003). "Archaeal chromatin and transcription." Mol Microbiol **48**(3): 587-598.

Robin, J. D., A. T. Ludlow, K. Batten, M. C. Gaillard, G. Stadler, F. Magdinier, W. E. Wright and J. W. Shay (2015). "SORBS2 transcription is activated by telomere position effect-over long distance upon telomere shortening in muscle cells from patients with facioscapulohumeral dystrophy." Genome Res **25**(12): 1781-1790.

Robinson, S. I., D. Small, R. Idzerda, G. S. McKnight and B. Vogelstein (1983). "The association of transcriptionally active genes with the nuclear matrix of the chicken oviduct." Nucleic Acids Res **11**(15): 5113-5130.

Roulet, E., S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod and P. Bucher (2002). "High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites." Nat Biotechnol **20**(8): 831-835.

Sandelin, A., P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki and D. A. Hume (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-436.

Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." Nature **489**(7414): 109-U127.

Schaffner, W., G. Kunz, H. Daetwyler, J. Telford, H. O. Smith and M. L. Birnstiel (1978). "Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing." Cell **14**(3): 655-671.

Schmidt, J. V., J. M. Levorse and S. M. Tilghman (1999). "Enhancer competition between H19 and Igf2 does not mediate their imprinting." Proceedings of the National Academy of Sciences **96**(17): 9733-9738.

Serfling, E., M. Jasin and W. Schaffner (1985). "Enhancers and eukaryotic gene-transcription." Trends in Genetics **1**(8): 224-230.

Serfling, E., A. Lubbe, K. Dorsch-Hasler and W. Schaffner (1985). "Metal-dependent SV40 viruses containing inducible enhancers from the upstream region of metallothionein genes." EMBO J **4**(13B): 3851-3859.

Shopland, L. S., C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth, C. Cremer, C. J. Bult and T. P. O'Brien (2006).

"Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence." J Cell Biol **174**(1): 27-38.

Siebenlist, U., R. B. Simpson and W. Gilbert (1980). "E. coli RNA polymerase interacts homologously with two different promoters." Cell **20**(2): 269-281.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." Nat Genet **38**(11): 1348-1354.

Smale, S. T. and D. Baltimore (1989). "The 'initiator' as a transcription control element." Cell **57**(1): 103-113.

Smale, S. T. and J. T. Kadonaga (2003). "The RNA polymerase II core promoter." Annu Rev Biochem **72**: 449-479.

Small, D., B. Nelkin and B. Vogelstein (1985). "The association of transcribed genes with the nuclear matrix of Drosophila cells during heat shock." Nucleic Acids Res **13**(7): 2413-2431.

Spitz, F., F. Gonzalez and D. Duboule (2003). "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." Cell **113**(3): 405-417.

Splinter, E., H. Heath, J. Kooren, R. J. Palstra, P. Klous, F. Grosveld, N. Galjart and W. de Laat (2006). "CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus." Genes Dev **20**(17): 2349-2354.

Stathopoulos, A. and M. Levine (2002). "Whole-Genome Expression Profiles Identify Gene Batteries in Drosophila." Developmental Cell **3**(4): 464-465.

Stathopoulos, A. and M. Levine (2005). "Genomic Regulatory Networks and Animal Development." Developmental Cell **9**(4): 449.

Stathopoulos, A., M. Van Drenth, A. Erives, M. Markstein and M. Levine (2002). "Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo." Cell **111**(5): 687-701.

Suen, C. S., T. J. Berrodin, R. Mastroeni, B. J. Cheskis, C. R. Lyttle and D. E. Frail (1998). "A transcriptional coactivator, steroid receptor coactivator-3, selectively augments steroid receptor transcriptional activity." J Biol Chem **273**(42): 27645-27653.

Sumiyama, K., S. Q. Irvine, D. W. Stock, K. M. Weiss, K. Kawasaki, N. Shimizu, C. S. Shashikant, W. Miller and F. H. Ruddle (2002). "Genomic structure and functional control of the Dlx3-7 bigene cluster." Proc Natl Acad Sci U S A **99**(2): 780-785.

Tai, P. W. L., K. I. Fisher-Aylor, C. L. Himeda, C. L. Smith, A. P. MacKenzie, D. L. Helterline, J. C. Angello, R. E. Welikson, B. J. Wold and S. D. Hauschka (2011). "Differentiation and fiber type-specific activity of a muscle creatine kinase intronic enhancer." Skeletal Muscle **1**:25.

Tapscott, S. J., A. B. Lassar and H. Weintraub (1992). "A novel myoblast enhancer element mediates MyoD transcription." Mol Cell Biol **12**(11): 4994-5003.

ten Bosch, J. R., J. A. Benavides and T. W. Cline (2006). "The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription." Development **133**(10): 1967.

Thanos, D. and T. Maniatis (1995). "Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome." Cell **83**(7): 1091-1100.

Theveny, B., A. Bailly, C. Rauch, M. Rauch, E. Delain and E. Milgrom (1987). "Association of DNA-bound progesterone receptors." Nature **329**(6134): 79-81.

Tiwari, V. K., L. Cope, K. M. McGarvey, J. E. Ohm and S. B. Baylin (2008). "A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations." Genome Res **18**(7): 1171-1179.

Tolhuis, B., R.-J. Palstra, E. Splinter, F. Grosveld and W. de Laat (2002). "Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus." Molecular Cell **10**(6): 1453-1465.

Udvardy, A., E. Maine and P. Schedl (1985). "The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains." J Mol Biol **185**(2): 341-358.

van Steensel, B. and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase." Nat Biotechnol **18**(4): 424-428.

van Werven, F. J., H. van Bakel, H. A. van Teeffelen, A. F. Altelaar, M. G. Koerkamp, A. J. Heck, F. C. Holstege and H. T. Timmers (2008). "Cooperative action of NC2 and Mot1p to regulate TATA-binding protein function across the genome." Genes Dev **22**(17): 2359-2369.

Verschure, P. J., I. van Der Kraan, E. M. Manders and R. van Driel (1999). "Spatial relationship between transcription sites and chromosome territories." J Cell Biol **147**(1): 13-24.

Weber, F., J. de Villiers and W. Schaffner (1984). "An SV40 'enhancer trap' incorporates exogenous enhancers or generates enhancers from its own sequences." Cell **36**(4): 983-992.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee and R. A. Young (2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes." Cell **153**(2): 307-319.

Wiesendanger, B., R. Lucchini, T. Koller and J. M. Sogo (1994). "Replication fork barriers in the Xenopus rDNA." Nucleic Acids Res **22**(23): 5038-5046.

Wigler, M., R. Sweet, G. K. Sim, B. Wold, A. Pellicer, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Transformation of mammalian cells with genes from procaryotes and eucaryotes." Cell **16**(4): 777-785.

Willy, P. J., R. Kobayashi and J. T. Kadonaga (2000). "A basal transcription factor that activates or represses transcription." Science **290**(5493): 982-985.

Wold, B., M. Wigler, E. Lacy, T. Maniatis, S. Silverstein and R. Axel (1979). "Introduction and expression of a rabbit beta-globin gene in mouse fibroblasts." Proc Natl Acad Sci U S A **76**(11): 5684-5688.

Wurtele, H. and P. Chartrand (2006). "Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology." Chromosome Res **14**(5): 477-495.

Yaffe, D. and O. Saxel (1977). "Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle." Nature **270**(5639): 725-727.

Yee, S. P. and P. W. Rigby (1993). "The regulation of myogenin gene expression during the embryonic development of the mouse." Genes Dev **7**(7A): 1277-1289.

Zakany, J., C. Fromental-Ramain, X. Warot and D. Duboule (1997). "Regulation of number and size of digits by posterior Hox genes: a dose-dependent mechanism with potential evolutionary implications." Proc Natl Acad Sci U S A **94**(25): 13695-13700.

Zeitlinger, J., A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine and R. A. Young (2007). "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." Nat Genet **39**(12): 1512-1516.

Zeitlinger, J., R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young and M. Levine (2007). "Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo." Genes Dev **21**(4): 385-390.

Zeng, M. e. a. (2015). in review.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

Zhao, Z., G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti and R. Ohlsson (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." Nat Genet **38**(11): 1341-1347.

Zinzen, R., K. Senger, M. Levine and D. Papatsenko (2006). "Computational Models for Neurogenic Gene Expression in the Drosophila Embryo." Current Biology **16**(13): 1358-1365.

Chapter Supplemental III: Materials and Methods: Experimental Protocols

SIII.1: Cell Growth Protocol and Differentiation treatment for the C2C12 Cell Line

From: Wold mouse ENCODE
Date: May 17, 2011
Prepared by: Katherine Fisher-Aylor and Brian Williams

C2C12 cell culture, differentiation treatment, and cross-linking protocol

The cell line C2C12 is an immortal line of mouse skeletal myoblasts originally derived from satellite cells from the thigh muscle of a two month old female C3H mouse donor 70h after a crush injury (Yaffe and Saxel, 1977; karyotyping available in Casas-Delucchi, 2011). From the C2s the immortal subline C2C12 was selected (Blau et al., 1985). These cells differentiate well into myocytes under appropriate culture conditions given below. The cells are adherent in culture and are grown on Nunc delta surface plastic culture dishes. They grow as undifferentiated myoblasts in growth medium (15-20% fetal bovine serum, with 20% used here). Myogenic differentiation is initiated upon reaching confluence by switching the cells to medium containing 2% horse serum supplemented with insulin. C2C12's are commercially available but because variable handling of this line can select for cells with different kinetics or poor differentiation performance, the Wold lab will provide plugs of these C2C12's upon request.
(See: (1) Yaffe and Saxel, 1977; Nature Vol. 270, 725-727; (2) Casas-Delucchi et al., 2011; Nature Communications Vol. 2, 222. (3) Blau et al., 1985; Science Vol. 230, 758-766).

Cell culture protocol for cycling (exponentially growing) cells

Cells are grown at 37°C in a humidified incubator with 5% CO2.

Myoblast growth medium

	final	stock	example
DMEM			395 mL
FBS (fetal bovine serum)	20%	100%	100 mL
Final			500 mL

Materials

DMEM (high glucose + glutamine, no Sodium Pyruvate)	GIBCO #11965
FBS	HyClone #30071.03

Antibiotics: We use 1X Penicillin/Streptomycin (100X stock = Gibco # 15140). This comes out to final concentrations of 100 units/mL penicillin and 100 ug/mL streptomycin.

Liquid Nitrogen Storage

Freeze cells in growth medium supplemented with 10% (v/v) DMSO in 1 ml aliquots of approximately 0.5-1 x 106 cells. When grown on 15 cm dishes, the cells reach confluence at ~2.6 X 106 cells per dish.

Cell culture and passage

1. Thaw a 1-ml aliquot of cells as quickly as possible in water bath at 37°C. Transfer cells to 24 mL warm media in a 50 mL conical tube. Mix gently. Plate the cells in a 15cm Nunc delta surface plates. Place in incubator. After one day, remove the medium and add fresh media.

2. When cells are 50-60% confluent (meaning that very few of them are physically touching each other), split 1:4 or 1:5 (at most). It is important to not let the cells become fully confluent because they can begin to fuse and partially differentiate upon cell-cell contact. To passage, remove and discard culture medium. Rinse twice with PBS (Calcium and Magnesium free). For a 15 cm dish, add 2.5mL of 0.25% (w/v) trypsin + 0.53 mM EDTA solution (Gibco #25300) prewarmed to 37°C, and observe cells under an inverted microscope until cell aspect changes to round (usually within 60-90 seconds). Aspirate the majority of the trypsin and let stand for an additional 1-2 minutes, then tap the plate to dislodge cells. Add 10mL of myoblast growth medium to the dish, and collect

cells by gently pipetting. (If using 10cm dishes, the volume of trypsin is reduced to 1 mL, and the time is reduced to 1 minute in trypsin). Dilute cells in a larger flask to the appropriate volume using growth media and aliquot to new Nunc dishes. There is no need to feed the cells in between passages. This is a fairly quickly growing cell line (doubling time is approximately 12h); you will need to passage them every 1-2 days.

Differentiation treatment

Differentiate for 24 hours to 7 days by rinsing fully confluent cells once with PBS and adding 25mL of low-serum differentiation medium. Feed with fresh differentiation medium every 24 hours up to the 72h timepoint and after that, every 12 hours (as these cells differentiate, they begin to deplete and acidify the medium more quickly). The timepoints we typically use are 24h, 60h, 5D, and 7D. Feed the cells no closer than 6h before fixation to avoid seeing serum-response effects in the cell prep.

Differentiation medium

	final	stock	example
DMEM			489.5mL
Donor equine serum	2%	100%	10 mL
Insulin (add no more than 24h before use)	1uM	1mM	0.5 mL
Final			500 mL

Materials

DMEM (high glucose + glutamine, no Sodium Pyruvate)	GIBCO #11965
Donor equine serum	HyClone #SH30074.02
insulin	Sigma-Aldrich #I-6634

Insulin: 1,000X stock is 1mg/mL in water with 10-20 µl of acetic acid added to acidify the water so it dissolves (use minimum possible). Filter sterilize with 0.2 um filter. Store at -20°C in small aliquots until use.

Antibiotics: We use 1X Penicillin/Streptomycin (100X stock = Gibco # 15140). This comes out to final concentrations of 100 units/mL penicillin and 100 ug/mL streptomycin.

Cell cross-linking and harvest for ChIP

1. Remove the medium from the culture plates and add a solution of PBS with 1% formaldehyde (Sigma-Aldrich # F87750). Swirl gently, and incubate at room temperature for 10 minutes.

2. Stop the cross-linking reaction by adding glycine to a final concentration of 0.125 M and swirl gently to mix. Use a stock solution of 2.5M glycine dissolved in H20. Incubate for 10 minutes.

3. Remove PBS/FA/glycine from plates and gently wash cells twice with 15 mL room temperature PBS.

4. To detach the cells from the dishes, add dilute trypsin (2mL PBS + 0.4mL of Gibco trypsin+EDTA (Gibco #25300)) for 10 min at 37°C, then quench with 100uL horse serum or FBS. Transfer to ice or 4°C.

5. Add 2 mL of cold PBS and scrape into a 15mL falcon tube; rinse plate once with 5mL of cold PBS and combine.

6. Pellet cells at 360 X g for 5 minutes at 4°C.

7. Aspirate PBS/trypsin solution and resuspend cells in 5 ml cold (4°C) PBS + 1 uM PMSF.

8. Pellet cells at 360 X g for 5 minutes at 4°C.

9. Carefully aspirate PBS and add 6 ml cold (4°C) Farnham lysis buffer (5 mM PIPES pH 8.0 / 85 mM KCl / 0.5% NP-40) + Roche Protease Inhibitor Cocktail Tablet (Complete 11836145001). This step lyses the cell membrane, leaving the nuclear envelope intact.

10. Pellet nuclei at 360 X g for 5 minutes at 4°C.

11. Place the nuclear pellet on ice. Carefully remove supernatant and either proceed to sonication step or snap freeze in liquid nitrogen and store at -80°C or in liquid nitrogen.

RNA yields

A 15 cm dish of undifferentiated cells yields about 20 ugs of total RNA collected with Qiagen RNEasy reagents. A 15 cm dish of differentiated cells yields about 60 ugs of total RNA.

SIII.2: C2C12 fixation protocol
Katherine Fisher-Aylor
last updated 12/2012

RNA extraction (for RNA-Seq)

1. Rinse 2x with PBS
2. Add 2.25mL room temperature mirVana binding + lysis buffer
3. Scrape into a 14mL snap-cap tube
4. Shear 20x through a 21.5g needle using a 3mL syringe.
5. Put in -80C freezer.

ALTERNATE PROTOCOL (better for isolating small RNAs): use MirVana lysis

RIPA lysis (for protein)

1. Rinse 2x with PBS
2. Add 2mL RIPA+PIC
3. Incubate at 4C, tilting occasionally
4. Scrape into 2x Eppendorf tubes
5. Spun at 10Kxg 15 min at 4C
6. Aliquot into 5 Eppendorf tubes and put in -80C freezer

ICC fixation (to determine % myogenin positive nuclei)

1. Add 0.9 mL of 37% formaldehyde (w/MeOH) to 5mL of medium in 6cm plate. Final concentration should be 4% formaldehyde.
2. Put on rotator 20 minutes at room temperature.
3. Rinse 2x with PBS
4. Store wrapped in parafilm in PBS at 4C.

ChIP fixation

1. Rinse with PBS.
2. Add 1% formaldehyde (our stock is Mallinkrodt chemicals 37% FA w/MeOH) diluted in PBS to plates. Final concentration should be 1%. Use enough to cover the surface of the plate (typically 25mL/15cm plate).
3. Incubate on rotator at room temperature for 10-15 minutes.
4. Quench by adding 2mL 2.5M glycine directly to the 25mL fixation solution (7.5g/40mL H2O) and incubate 10 min on rotator at room temp.
5. Rinse 2x with cold PBS

*plates can stay at 4C here for several hours (I've tested up to 12h).

6. Add 2mL PBS+0.4mL trypsin/EDTA (0.05M trypsin) and incubate 10 minutes at 37C (NOT sterile incubator)
7. Add 100uL equine or fetal bovine serum to each dish and tilt to inactivate trypsin.
8. Put plates in refrigerator until scraped; put scraped cells in refrigerator too.

*plates can stay here at 4C for a few hours (I've tested up to 2h)

9. Scrape cells into a 15mL tube.
10. Wash plates with PBS and scrape into the same tube.
11. Spin at 360 rcf, 4C for 5 minutes
12. Remove supernatant and resuspend in 5mL PBS+PMSF (1x)

13. Spin at 360 rcf, 4C for 5 minutes. Remove supernatant.

14. Resuspend in MC lysis buffer (1/2/3mL depending on cell amount)

15. Put at -80C to freeze OR spin at 360 rcf, remove supernatant, and freeze as a pellet. A flash freeze is unnecessary.

Collection of nuclei for genomic DNA assay (to determine the # of nuclei per plate)

1. Rinse 2x in PBS

2. Scrape into an eppendorf tube; store at -20C

SOLUTIONS

RIPA buffer (store at 4C)
1x PBS
1% NP-40
0.5% sodium deoxycholate
0.1% SDS
add Roche complete EDTA-free protease inhibitor cocktail right before using.

MC lysis buffer (store at 4C)
10mM Tris pH 7.5
10mM NaCl
3mM $MgCl_2$
0.5% NP-40
add Roche complete EDTA-free protease inhibitor cocktail right before using.

100x PMSF (store at -20C)
100% EtOH
100mM PMSF

SIII.3: Genomic DNA assay to quantify the number of nuclei per plate of adherent cells
Katherine Fisher-Aylor
last updated 8/1/10

Protocol and reagents are from Epicentre MasterPure Complete DNA and RNA Purification Kit

1. Cells should have been collected by simply scraping them in PBS into an eppendorf. I freeze these samples at -20C.

2. Collect approximately 0.5x10^5 to 2x10^6 nuclei (this is what the volumes in the kit are optimized for). For a 15cm plate of exponential C2's, I use 25% of a plate. For a

15cm plate of packed 60h C2's, I use 1% of a plate. Thaw and resuspend the whole-plate samples, calculate the total volume, and take a percentage accordingly.

3. If the volume of your sub-samples is larger than 30uL, spin them for 5 minutes at 360xg and reduce the volume to 30uL.

4. Add 300uL of T+C lysis buffer to each tube. Then add 2uL of 50ug/uL proteinase K to each tube and vortex briefly to mix.

5. Incubate 30 minutes at 65C, shaking every 10 minutes.

6. Cool tubes to 37.

7. Add 1uL 5ug/uL RNase A and mix. BE CAREFUL WHICH PIPETS/CONTAINERS YOU USE WITH RNASE SINCE OUR LAB WORKS WITH RNA.

8. Incubate 30 minutes at 37C.

9. Put on ice 5 minutes.

10. Add 150uL MPC Protein Precipitation Reagent to each sample and vortex 10 seconds.

11. Centrifuge for 10 minutes at maximum speed, 4C. If the pellet is small or loose, add extra MPC and spin again. Note: the MPC only works when it is kept cold. If the tubes warm up, cool them down and try again.

12. Transfer the supernatant to a new tube and discard pellet.

13. Add 500uL Isopropyl alcohol and mix.

14. Pellet the DNA by spinning 10 minutes at maximum speed, 4C.

15. Remove the supernatant and save the pellet.

16. Rinse the pellet with 75% EtOH by pipetting 200uL into the tube, gently flicking, and removing it. If your pellet dislodges or breaks, spin 1 minute at maximum speed.

17. Remove all residual EtOH. Bench dry 30 minutes or until all liquid is gone.

18. Resuspend the DNA in 35uL TE.

19. Nano-Drop (or even better, Q-bit) the DNA to determine its concentration.

20. Knowing the final concentration, the final volume (35uL), and the % per plate that you started with, you know the amount of DNA per plate. Assuming 6.5pg DNA/nucleus for a mouse cell, you know the approximate number of nuclei per plate. I usually expect about 1x10^6 cells per C2 exponential plate, 5x10^7 cells per C2 24h plate, and between 1 and 3x10^8 cells per C2 60h plate.

Note: this method has its own biases and may even be biased differently for exponential and differentiated cells. For a more certain determination, it is recommended also to count the number of nuclei (with Hoescht or other nuclear stain) under a microscope and extrapolate by area the number of cells per plate. Then average those results with the results of the genomic DNA assay.

SIII.4: ChIP-Seq protocol

Wold lab ChIP protocol
Katherine Fisher-Aylor version
Last updated 2/2013
This is based on G Kwan's ChIP protocol, a derivative of the Johnson et al. 2008 protocol

NOTE 1: "wash" = 5 minutes on magnet followed by incubation on rotator at 4C

NOTE 2: the times listed here are literal i.e. a 5 minutes wash means 5 minutes from the time the solution goes into the tube to the time it is taken out.

<u>Day 1</u>: antibody-bead coupling.

1. Make fresh 5mg/mL BSA (8mL per IP): BSA fraction V in PBS, sterile filtered using a 0.2um syringe filter (cellulose acetate okay). Store at 4C until day 2.

2. Add resuspended (vortexed) magnetic bead slurry to 1.5mL protein low-bind tube. See "technique considerations" for bead amount.

3. Wash 3x for 5-10 minutes in 0.9mL 5mg/mL BSA

4. Resuspend in 1mL BSA and add antibody. See “technique considerations” for antibody amount.

5. Incubate 20-24 hours on rotator at 4C (until Day 2, step 5)[1]

Day 2: sonication and antibody-chromatin binding

** KF note: Sonication is a very important step in the success of a ChIP, but it varies widely from sonicator to sonicator. This and most current protocols call for the majority of the (ChIPpable) DNA to be sheared to an average length of 200bp. Treat this as a “black box” step at your own risk. My best suggestion is to verify your sonication results each time you sonicate by reverse-crosslinking the chromatin and assaying the DNA distribution on a gel. **

1. Resuspend nuclear lysates in 0.5 – 2 mL RIPA+PIC with 5e7 nuclei per tube (more chromatin per volume → more viscosity. You will need to tune this for your sonicator.). Keep the chromatin at 4C for the duration of the sonication.

2. Sonicate:

Misonix 3000 protocol: Sonicate on cold EtOH (-20C). Mix the EtOH with a stir bar during the sonication. Throw out any samples that foam.

a. Unscrew the tip (we use 1/16” tapered microtips) and determine how it looks. A very slight ‘crater’ on the tip is okay. If it has a large hole or multiple holes, it will probably be inefficient. Switch it out for a new one or use extra cycles.

b. Wipe the probe tip with ddH2O, then 75% EtOH

c. Place the tip of the ~4-6cm from the bottom of the tube, near the 300uL mark. Don’t let the probe touch the sides of the tube.[2]

[1] I have let this incubation go up to 48h without having the ChIP fail. However, I do not know if it is better, worse, or the same.

[2] For tubes with 1mL volume, it appears best for the probe tip to be higher. However, I have not done a side-by-side comparison of sonication efficiency vs. probe placement.

d. Sonicate 25 cycles: 30 seconds on with 60 seconds rest in between cycles at setting 3.5. The power output should read 9-12 W. To be sure to avoid foaming, use setting 3.0 (6-9 W) for the first 5 cycles.

Biorupter protocol: Divide each tube of chromatin up into 3-4 TPX (hard plastic) tubes of 200-300 uL. Put them in the machine and fill the extra slots with tubes of water so the efficiency will be the same for all experiments. Make sure the water is exactly at the level marked on the side of the bath. Cycles are 30sec on, 60sec off on 'high' with the chiller running. If the chiller isn't working, add ice every 5-10 cycles and remove the extra water (though we think this massively decreases efficiency and will need ~125 cycles).

3. The volume of sonicate that needs to be added to each tube of ChIP reaction (i.e. to the tubes of rinsed, antibody-coated beads) is 1mL at a concentration of 2.5 * 10^7 nuclei's equivalents of sonicated chromatin. Use RIPA to adjust your concentration and volume accordingly.

4. Centrifuge the sonicates at 14K RPM for 15 minutes at 4C.

5. Meanwhile: wash beads 3x for 5-10 minutes in 5mg/mL BSA. After the last wash, resuspend beads in 100uL BSA.

6. Remove 5%-10% of supernatant for "input DNA" controls for QPCR and ChIPseq. Keep these at the same temperature as the ChIPs for the duration of the protocol and reverse crosslink them at the same time as your ChIP.

7. Add the supernatant from a centrifuged tube of sonicate to the 100uL suspension of beads.

For clarity, one "ChIP reaction" is the sonicate from 2.5 * 10^7 nuclei plus one tube of antibody-coated beads from Day 1.

8. Incubate the ChIP reactions 20-24h on rotator at 4C.

Day 3: reversal of crosslinks

Make sure the water bath is set to 65C and has plenty of water in it.

1. Wash 5x with LiCl wash buffer: 1.2mL volume, 10-15 minute washes on rotator at 4C.

2. Rinse pellet 1x with 1mL TE

3. Resuspend in 200uL IP Elution Buffer at room temperature (this solution precipitates at 4C)

4. Incubate IP's and input DNA at 65C for 1 hour, shaking every 15 minutes to resuspend beads – or put them on the shaking heating block. This dissociates the antibodies from the beads.

5. Spin at 14K RPM for 3 minutes to pellet beads, then remove and save the supernatant. Put the supernatant in DNA low-bind tubes.

6. Add 2ug 50mg/ml proteinase K to the samples.

7. To the input DNA, add the equivalent volume of IP elution buffer to put it in as similar a solution to the ChIPs as possible.

8. Incubate IP's and input DNA at 65C 8-12 hours to reverse formaldehyde cross-links.

Day 4: Cleanup

The columns and reagents used are from the Qiagen Qiaquick PCR cleanup kit unless otherwise noted.

Prep: warm an aliquot of {110 x your sample number}uL EB to 55C.

This temperature is necessary in step 6 to avoid losing small pieces of DNA in the range that ChIP protocols traditionally require.

1. Optional: Add 150 uL of nuclease-free water to the IP's. Extract IP's, depleted DNA, and input DNA with an equal volume of phenol/chloroform/isoamyl alcohol

(25:24:1) by vortexing 20 seconds, then spinning 3 min at 14K RPM. Remove and save the aqueous (top) phase. For lP's, withdraw about 325 uL.s

2. Add 3X the volume of Qiagen buffer PM and mix. To avoid losing long pieces of DNA, bring pH of this solution to 7.

This is necessary because the elution buffer the samples are in is basic, ~pH 10. The kit explains the pH issue and has instructions. Don't add indicator dye to the ChIP samples; rather, use a side sample of input or depleted DNA, or even elution buffer plus RIPA in the same ratio the ChIPs/input DNA are in, to figure out how much acid to use to bring the pH down.

3. Add the sample to a spin column, let sit 2 minutes, then spin 2 min at 14K RPM.. If you have more than a 750 uL volume, add half the sample, spin, dump the liquid, then repeat with the other half of the sample. This binds the DNA to the column.

4. Dump the liquid, then wash the DNA with 750 uL Buffer PE (make sure EtOH has been added to it). Pipette on the buffer, let stand 2 minutes, then spin 2 minutes at 14K RPM.

5. Dump the liquid, then spin 2 minutes at 14K RPM to dry.

6. Pipette 100 uL 55C buffer EB directly onto the column membrane. Let stand 2 minutes, then spin 2 minutes at 14K RPM into DNA low-bind tubes.

7. Optional:[3] re-elute by pipetting the eluate back onto the column membrane and spinning again.

8. Save eluate as your ChIP, input DNA, or depleted DNA. Store at 4C.[4]

[3] Some think this gives a higher yield.

[4] ChIPs should be stable at 4C for a month or so in TE buffer and DNA low-bind tubes. For long-term storage, -80C is best, though you must avoid repeated freeze-thaws.

SOLUTIONS

RIPA buffer (store at 4C)
1x PBS
1% NP-40
0.5% sodium deoxycholate
0.1% SDS -- but better to increase this to 1% SDS for the sonication only
add Roche complete EDTA-free protease inhibitor cocktail right before using.

LiCl IP wash buffer (store at 4C)
100mM Tris
500mM LiCl
1% NP-40
1% sodium deoxycholate

IP elution buffer (store at room temp)
1% SDS
0.1 M NaHCO3

KF note: Technique considerations

1. Do not assume your magnet will instantly clear a sample. The commercial ChIP magnet I use requires up to 5 minutes to fully clear a sample. Test this with your in-house magnets.

2. Use aerosol-barrier tips and sterile solutions.

3. Don't cut the wash times short. In general, if you cannot stick to a listed wash time, longer wash times are better (with the possible exception of the LiCl wash, which I have not varied). Much longer washes don't improve results, but they don't appear to hurt them either.

KF note: Optimization considerations

1. Antibody/bead ratio is a major factor to consider in optimizing ChIP results. In general, 5ug of monoclonal antibodies + 100uL of beads or 10ug of polyclonal antibodies + 200uL of beads works well.

2. Chromatin amount is another major factor to consider. TF's with few binding sites may require different amounts of chromatin as broad-scale chromatin marks.

Ubiquitous TF's and factors such as RNA polymerase II are somewhere in the middle. As a rule of thumb, 2.5*10^7 nuclei per ChIP works well.

3. Sonication: what you are aiming for, especially if you want to sequence your sample on an Illumina sequencer, is having the majority of your sample between 100 and 250bp. Note that the misonix tip looses efficiency after about 10h of sonication, and I imagine other sonicators have similar problems. Again: treat the sonication step as a "black box" at your own risk.

SIII.5: ChIP protocol for successful ChIA-PET experiments

Katherine Fisher-Aylor
Wold lab at Caltech
Written 7/1/2010
Last updated 9/3/2013

ChIP fixation (adherent cells)

WARNING: keep the plates in a fume hood even when you are scraping them. The fumes from the EGS-fixed cells are very dangerous and DO NOT disappear after rinsing plates the way formaldehyde fumes do. I learned this the hard way.

1. Rinse plates 2x with PBS (I don't know if it matters, but the PBS we use has no Ca2+ or Mg2+).

2. Dissolve EGS (Pierce) at 10mM in 50% glacial acetic acid/50% ddH2O. Do this soon before you are ready to use it because EGS hydrolyzes very quickly in solution.[5]

3. Dilute the 10mM EGS to 1.5mM in PBS.

4. Add the 1.5mM EGS to the cells for 30 minutes at room temperature on a rotator.

5. Add formaldehyde[6] to the EGS solution to a final concentration of 1%. Incubate 15 minutes at room temperature on a rotator.

[5] Changing the brand and solvent of the EGS causes the chromatin to appear very different. I am currently investigating different combinations. With this fixative, the chromatin will appear white and fluffy but dense and difficult to resuspend. During sonication, it will leave a very large pellet (I suspect this is cellular debris that has been crosslinked by the EGS) and will be milky white in solution.

6. Quench by adding glycine to a final concentration of 0.2M and incubate 10 min on rotator at room temp. The glycine stock I use is simply dissolved in water at 2.5M and is not buffered.

7. Rinse 2x with cold PBS

8. Add 2mL PBS + 0.4mL 0.05% trypsin/EDTA (Gibco) to each dish and incubate 10 minutes at 37C

9. Add 100uL equine or fetal bovine serum to each dish and tilt to inactivate trypsin.

10. Scrape cells into 15mL tube and put on ice. Keep plates at 4C until scraped.

11. Spin at 360 rcf, 4C for 5 minutes

12. Remove supernatant and resuspend cell pellet in 5mL PBS + 1mM PMSF

13. Spin at 360 rcf, 4C for 5 minutes. Remove supernatant.

14. Resuspend cell pellet in 3mL MC lysis buffer.

15. Spin at 360 rcf, 4C for 5 minutes. Remove supernatant.

16. Repeat steps 15 and 16.

17. Put the nuclear pellet at -80C to freeze.

MC lysis buffer
10mM Tris pH 7.5
10mM NaCl
3mM MgCl2
0.5% NP-40
add Roche complete EDTA-free protease inhibitor cocktail before using.

PMSF: stock solution is 100mM in 100% EtOH. Store at -20C.

ChIP Day 1: antibody-bead coupling

1. Make fresh 5mg/mL BSA (8mL per IP): BSA fraction V in PBS, sterile filtered using a 0.2um syringe filter (cellulose acetate is okay). This can be stored at 4C until day 2.

---

[6] Our stock formaldehyde is 37% containing 10% MeOH from Mallinckrodt Chemicals

2. Add 100uL resuspended (vortexed) magnetic bead slurry (Dynal M280 sheep anti-mouse beads) to 1.5mL protein low-bind Eppendorf tube.

3. Wash 3x for 5-10 minutes in 0.9mL 5mg/mL BSA at 4C on a rotator.

4. Resuspend in 1mL 5mg/mL BSA and add 5ug mouse monoclonal polII CTD4H8 antibody (Millipore).

5. Incubate 48 hours on rotator at 4C (until Day 2, step 5)

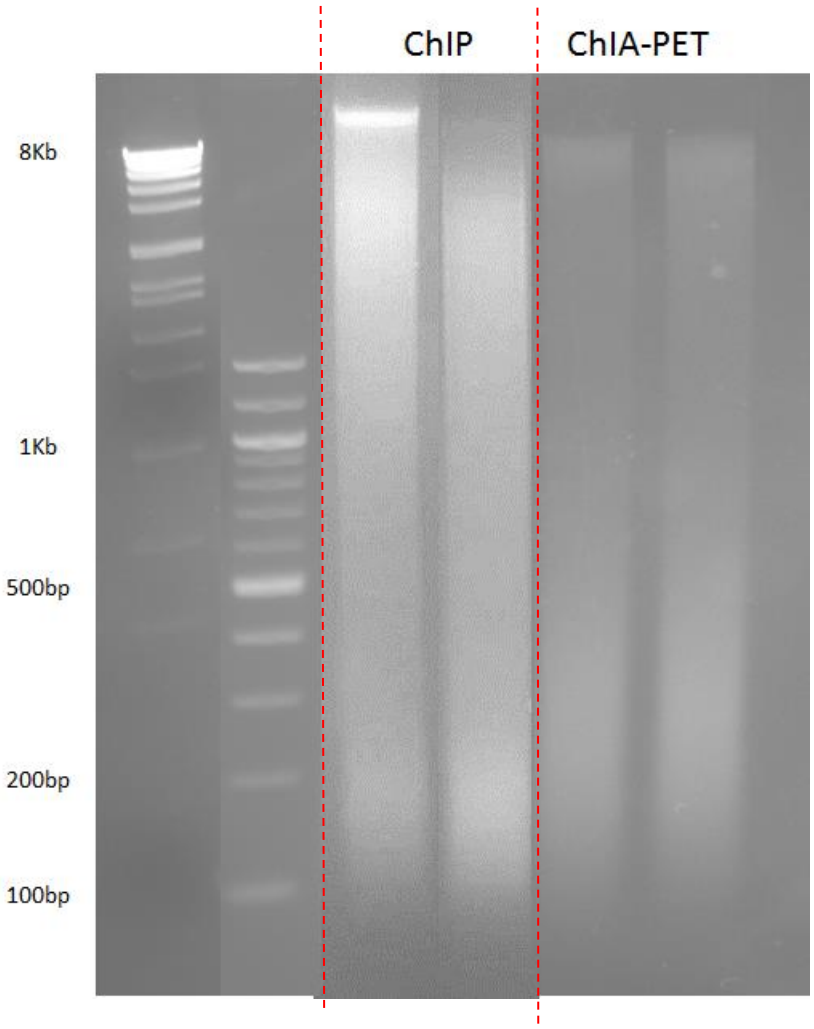
Our standard ChIP incubation time is 20-24h. I do not know if 48h is better; I do it for consistency only because this is how my first successful ChIA-PET was made (it took a long time to sonicate the chromatin properly).

Day 2: sonication and antibody-chromatin binding

1. Resuspend nuclear lysates in 1mL RIPA, with ~1x10^8 nuclei per tube.

2. Sonicate to a length of 300-500bp. I have not yet been able to sonicate EGS-fixed chromatin successfully in the biorupter, so I am using a Misonix 3000 biorupter. I use 25 cycles (30sec on, 60sec) off at 15W**.

** The devil is in the details of this step. In order to get the proper fragment length, I under-sonicate, reverse-crosslink a sub-sample, run it out on a gel, and then repeat until I get the proper fragment distribution. The attached picture is the gel for my first successful ChIA-PET sample compared to one of my successful conventional ChIP samples.



3. Centrifuge at 14K RPM for 15 minutes at 4C.

4. Remove 100uL of supernatant for “input DNA” controls for QPCR and ChIPseq. Keep these at the same temperature as the ChIPs for the duration of the protocol.

5. Meanwhile: wash beads 3x for 5-10 minutes in 5mg/mL BSA. After the last wash, resuspend the beads in 100uL BSA.

6. Add 1 tube of sonicated chromatin each tube of beads (1mL of chromatin originally from 1x10^8 cells per 100uL of beads. *Note, though, that the yield of DNA from sonicated EGS-fixed chromatin appears to be half of that from FA-fixed chromatin*).

7. Incubated 28h on rotator at 4C

Our standard incubation time is 20-24h. This again is for consistency's sake; I do not know if it is a meaningful difference.

<u>Day 3</u>: wash

1. Wash 5x with LiCl wash buffer: 1.2mL volume, 10-15 minute washes on rotator at 4C
2. Rinse pellet 1x with 1mL TE
3. Resuspend in 1mL TE

At this point, I removed 10% of each sample to assay the effectiveness of each tube. I then resuspended each of the 5 best IP's in 200uL TE and combined them into one tube. Each of the samples sent therefore consisted of chromatin from 2.5x10^8 - 5x10^8 myotube cells or 2x10^8 – 4x10^8 myoblast cells

on 500uL beads. These samples were topped off with TE and wrapped up to send to Singapore with 4C cold packs. The libraries were built in Singapore within 45 days and were stored in the dark at 4C until building.

The following is for the ChIA-PET controls

<u>Day 3</u>: reversal of crosslinks

4. Resuspend in 200uL IP Elution Buffer at room temperature (this solution precipitates at 4C)
5. Incubate IP's and input DNA at 65C for 1 hour, shaking every 15 minutes to resuspend beads.
6. Spin at 14K RPM for 3 minutes to pellet beads, then remove and save the supernatant in DNA low-bind Eppendorf tubes.
7. Added 2ug 50mg/ml proteinase K to the IP's and 5ug to the input DNA samples.

8. Incubate IP's, depleted DNA, and input DNA at 65C 8-12 hours to reverse cross-links.

Full disclosure: I suspect the EGS crosslinks are not sufficiently reversed in this method. This does not affect the ChIA-PET since the crosslinks are far from the ligated ends of DNA that are selected, but it does affect making a good EGS-crosslinked ChIP-Seq control. KFA is working on this.

Day 4: Cleanup

Warm the EB at 55C

The columns and reagents used are from the Qiagen Qiaquick PCR cleanup kit

1. Add 3X the volume of Qiagen buffer PM and mix.
2. Add the sample to a spin column, let sit 2 minutes, then spin 2 min at 14K RPM..
3. Remove the liquid, then add 750uL buffer PE, let stand 2 minutes, then spin 2 minutes at 14K RPM.
4. Remove the liquid, then spin 2 minutes at 14K RPM to dry.
5. Pipette 100 uL 55C buffer EB directly onto the column membrane. Let stand 2 minutes, then spin 2 minutes at 14K RPM into DNA low-bind Eppendorf tubes.
6. Save eluate as your ChIP or input DNA. Store at 4C for a few months or -20C long-term.

My successful ChIA-PETs have all come from ChIPs that have at least 0.07ug of DNA per tube, assayed after the above crosslink reversal and purification.

RIPA buffer

1x PBS
1% NP-40
0.5% sodium deoxycholate
0.1% SDS
add Roche complete EDTA-free protease inhibitor cocktail right before using.

LiCl IP wash buffer

100mM Tris
500mM LiCl
1% NP-40
1% sodium deoxycholate

IP elution buffer
1% SDS
0.1 M NaHCO3

SIII.6: Analysis of DNA sonication/fragmentation results

Katherine Fisher-Aylor
Last updated 11/6/13

Summary: purify your DNA and then run 1ug on a 2% agarose gel at low voltage.

1. Reverse-crosslink your sheared chromatin.

    a. To a small sub-sample of chromatin (usually ~5uL of sonicate) add 5uL of 50mg/mL proteinase K plus 1x volume IP elution buffer

    b. Incubate at 65C for 8-12 hours (2 hours is a bit short and will give a lower yield of DNA and a slightly different looking gel).

2. Purify your DNA. Also a few ways to do this.

    a. Column purification. **WARNING**: **some columns impose a size selection. This is highly undesirable for us**, *since we want to know exactly what our fragmented distribution looks like. For the Qiagen columns I list here, I have not noticed any such size selection in the past. But be warned and I'd try one of the below methods from time to time to make sure the columns aren't changing! If you are using a different kit, don't take the manufacturer's word for it – try it yourself using a column vs. one of the non-column methods below. Also note that these columns give you approximately a 25% yield compared to a standard P:C:IAA extraction plus EtOH precipitation.*

        i. The kit is the Qiagen min-elute PCR cleanup kit

        ii. Warm elution buffer to 50C (or you will lose small fragments)

iii. Add 3x volume of buffer PM. Ensure the pH is low enough according to kit directions or you will lose long fragments; remember the IP elution buffer is basic.

iv. Add the above mixture to a column, and let stand 1 minute. Spin 1 minute at maximum speed in a microfuge and discard the eluate. If the starting volume is greater than 750uL, you will need to do this in two batches.

v. To the columns, add 750uL of buffer PE (make sure ethanol has been added). Let stand 1 minute then spin 1 minute at max speed and discard eluate.

vi. Spin again to completely dry the columns. Discard the round-bottom tubes and switch the purple columns to fresh eppendorf tubes.

vii. Pipette 100uL of warmed elution buffer onto the column membranes and let stand 1-2 minutes. Spin 2 minutes at maximum speed into the clean eppendorf tubes. Store at 4C short term, -80C long-term (but avoid repeated freeze-thaws).

b. Ethanol precipitation (Molecular Cloning method). *Note that the phenol extraction also decreases your yield and that phenol contamination can confuse nano-Drop readings and overestimate the amount of DNA you have (its wavelength is 270 compared to DNA at 260, so pay attention to the wavelength curve on the nano-Drop….or use the Qubit).*

i. Add an equal volume of phenol/chloroform/isoamyl alcohol 25:24:1. Vortex to mix, then spin 5 minutes at maximum speed. Keep the upper (aqueous) later by carefully pipetting it into a clean eppendorf tube. *Alternatively, you can extract using phenol/chloroform 1:1 (vortex, spin, remove upper layer) and then chloroform/isoamyl alcohol 24:1 (vortex, spin, remove upper layer). This might decrease phenol contamination.*

ii. Add NaCl to 0.2M. The reason for using this salt as opposed to LiCl or NaOAc is the SDS in the IP elution buffer.

iii. Add 2 volumes cold 100% EtOH and mix.

iv. Put at -80C for 20 minutes

v. Spin 20 minutes at maximum speed, 4C. Keep track of how your tube is oriented in the centrifuge so you know where the pellet will be (in case it is small enough to be invisible).

vi. Pipette off the supernatant and discard. Add 200uL 70% EtOH and mix to rinse the pellet.

vii. Spin 5 minutes at max speed, 4C. Pipette off the supernatant.

viii. Repeat steps vi and vii.

ix. Bench-dry (takes approximately 30 minutes) or dry in the speed-vac.

x. Resuspend in 100uL TE.

c. Singapore/'blue paint' precipitation method

i. Follow step b.i above.

ii. Add 10% volume of 3M NaOAc, pH 5.2, 0.4% volume of 15mg/mL GlycoBlue (Ambion), and 1x volume of Isopropyl alcohol

iii. Follow steps b.iii-b.x.

3. Pour a 2% agarose gel (because we are targeting the 50-500bp length).

a. In a glass Erlenmeyer flask: 50mL 1x TAE + 1g agarose

b. Heat in microwave 1:10, then swirl to dissolve/mix.

c. Once cooled enough to handle, add 4uL 5mg/mL EtBr.

d. Pour into a small casting tray and add the comb with 6 or 10 teeth and let set for 20 minutes.

4. Mix the samples to be run on the gel.

    a. 1ug of DNA

    b. Bring up to 20uL with water

    c. 4uL 6x gel loading buffer – no dye! *The dye will obscure the smooth 'smear' in the gel and your pictures will be misleading. If you are having a hard time loading the gel without dye, add just a tiny amount of buffer+dye to your solutions by dipping the tip of a pipette into it and swirling it in your sample..*

5. Make your ladder DNA. *You must have a ladder that resolves the 100-500bp range (such as the 100bp ladder from New England Biolabs). Ideally, you should also include a larger ladder that will tell you how big your biggest fragments are (I use Roche Marker VII, which goes up to 8.5kb).*

    a. 1ug ladder DNA

    b. bring up to 20uL with water

    c. 4uL 6x gel loading buffer + dye.

    d. It is okay to make this in advance and store it at 4C. However, make sure to dilute the DNA in TE not water!

6. Orient the gel properly in the gel box (DNA runs towards the positive electrode, which is red in US apparati). Cover in 1x TAE containing 8uL EtBr/100mL. *This saves you from having to stain/destain later and having the smaller sizes diffuse or from the EtBr out-migrating the DNA. Don't use TAE that has been used more than twice or that has evaporated because your gel will melt.*

7. Load the 24uL samples onto the gel.

8. Run at a low voltage such as 140 to give the different lengths of DNA time to migrate past each other properly. Run until the lower DNA dye band is ~1.5 inches from the bottom of the gel (approximately 45 minutes at 140).

9. Visualize on a UV light box. For pictures, the exposure time will usually be between 1/4 and 1/8 second.

10. Discard the TAE and gel in an EtBr disposal container.

SOLUTIONS

IP elution buffer
1. 1% SDS
2. 0.1 M NaHCO3

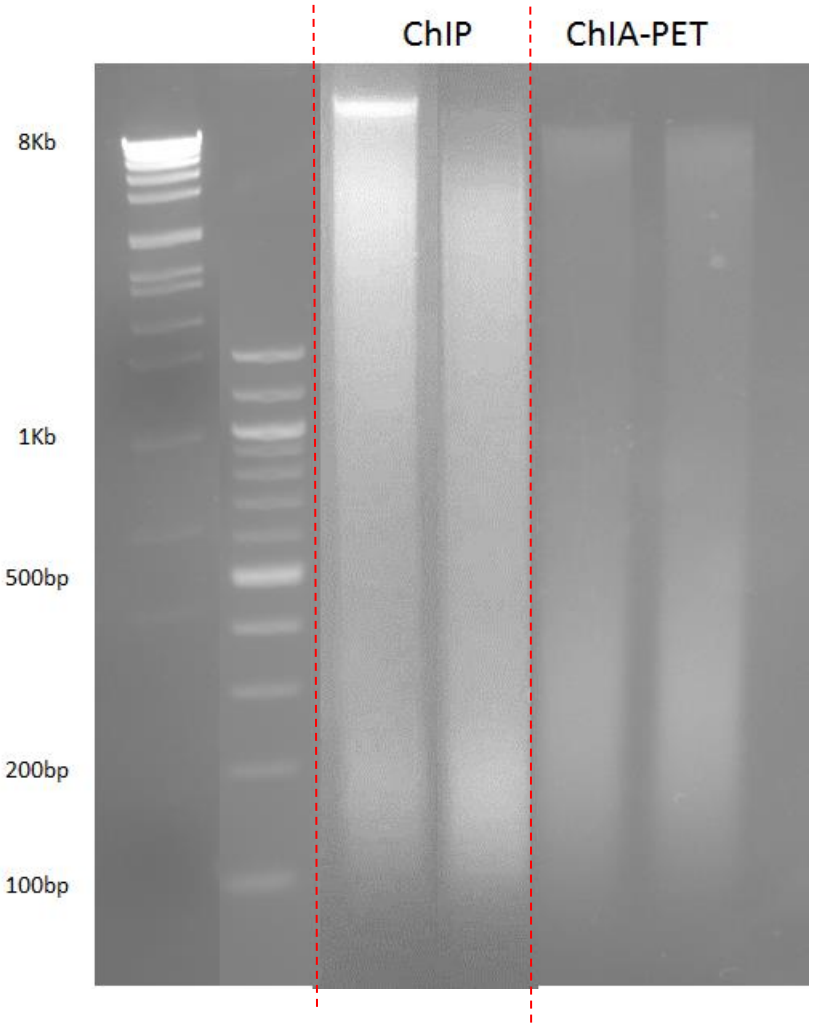
50x TAE
1. 242g Tris base
2. 57.1 mL glacial acetic acid
3. 100mL 0.5 M EDTA (pH 8.0)
4. Bring up to 1L with ddH2O
5. Dilute to 1x with ddH2O when using.

6x gel loading buffer:
1. 15% Ficoll in water. Store at room temperature.

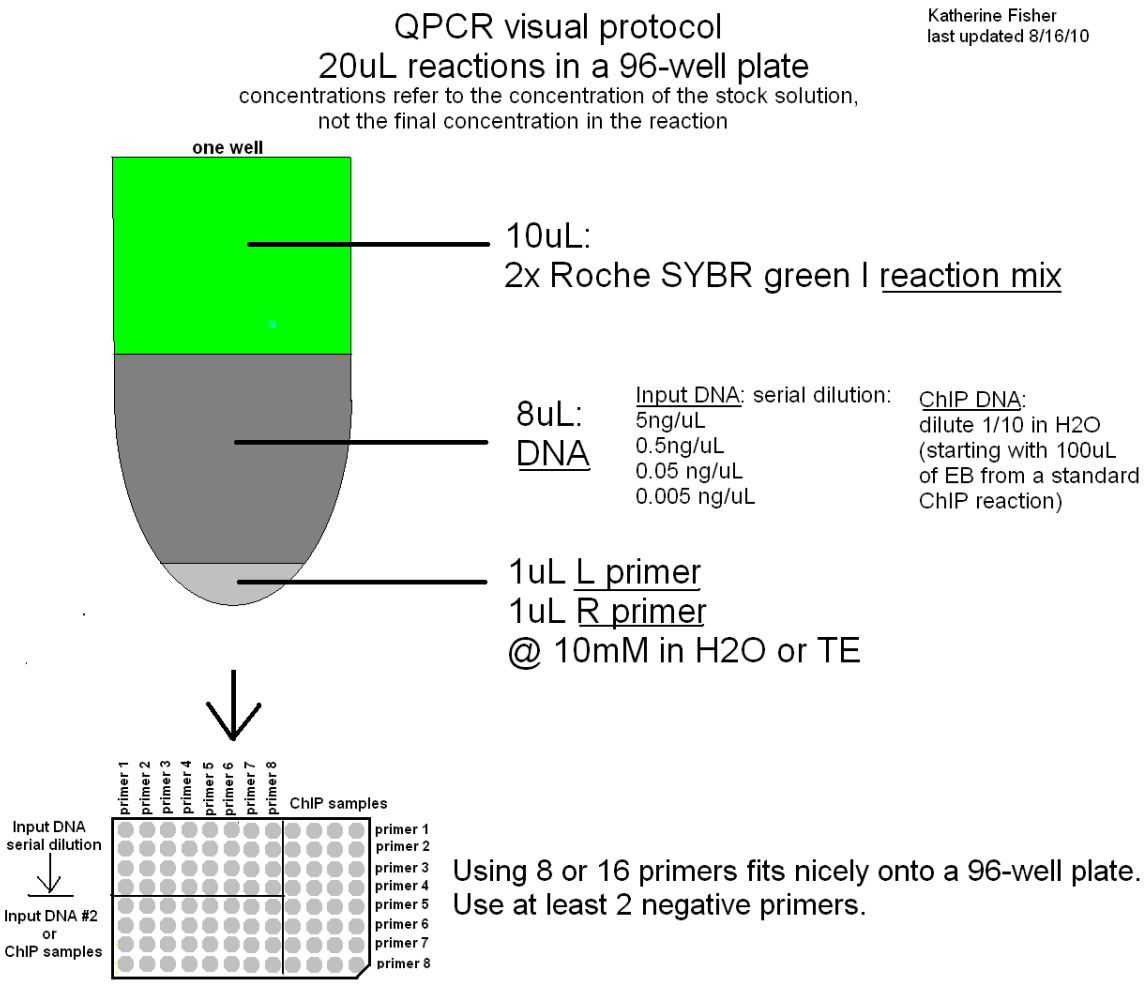
6x gel loading buffer plus dye:
2. 0.25% bromophenol blue (lower ~300bp band. You especially don't want this one obscuring your fragmented samples)
3. 0.025% xylene cyanol FF (upper ~4kb band)

What to expect from a "good" sonication

This varies depending on the application. I've attached a gel of what in my experience makes a good ChIP-Seq library and a good ChIA-PET library, respectively. If you are trying to duplicate someone else's results, ask them for one of their sonication gels (hoping they HAVE one) and don't take their word for what the DNA was sheared to as an "average length". Also please note that the *size distribution matters (another reason to ignore the reported "average length")*: the amount of long chromatin leftover after sonication might be good or bad for different applications. We have some evidence, anecdotal and experimental, that suggests that having long chromatin left behind is GOOD for the activator ChIP-Seq's we've done.


ChIP
ChIA-PET
8Kb
1Kb
500bp
200bp
100bp

SIII.7: QPCR assay



SIII.8: ChIA-PET library building protocol

Protocol from Yijun Ruan @ Genome Institute of Singapore in early 2010

DNA Blunting with T4 DNA Polymerase

1. Resuspend the beads by pipetting up and down. Pellet the beads by centrifugation at 800 rpm, 4 ^{o}C, for 5 min and discard the supernatant.

2. Split the beads into 2 tubes such that the final volume of 100% beads is $\leq$ 150 μl.

   To each tube, add the following:

10×Buffer for T4 DNA Polymerase (Promega)	50 μl

10mM dNTPs	5 μl
T4 DNA Polymerase (Promega, 7.9 u/μl)	6.3 μl
(The concentration of T4 DNA polymerase is 0.1 u/μl)	
H_2O	438.7 μl
	500 μl

3. Resuspend the beads using the above reaction mix. Incubate the beads at 37°C with rotation on the Intelli-Mixer (Skyline, Program F8, 30 rpm) for 15 min.

4. Wash the beads with 1 ml Wash Buffer (10 mM Tris-Cl, pH 7.5; 1 mM EDTA; 500 mM NaCl) 3 times. Mix well by inverting tubes. Pellet the beads by centrifugation at 800 rpm, 4°C, for 5 min after every wash.

Adding hM&M linker (Biotin) to the crosslinked ChIP DNA

5. Prepare the reaction mix as follows: **Mix** water with the linkers well first, then mix in the PEG buffer well before adding ligase.

Biotinylated linkers (200 ng/μl; IDT)	10 μl
5×T4 DNA Ligase Buffer with PEG (Invitrogen)	200 μl
T4 DNA Ligase (30 u/μl, Fermentas)	4 μl (at ≥0.1 u/μl)
H_2O	786 μl
	1,000 μl

6. Resuspend the beads with above reaction mix. Mix well. Incubate at 16°C with rotation on the Intelli-Mixer (Program F8, 30 rpm) for 16 hours.

Remove the excess linkers

7. The beads are then washed 3 times as above to remove the excess linkers.

Add Phosphate group to 5' ends

8. Prepare the reaction mix as follows:

10×T4 DNA Ligase Buffer (NEB)	100 μl
T4 DNA Polynucleotide Kinase (10u/ul, NEB)	20 μl (at ~0.2 u/μl)
H_2O	880 μl
	1,000 μl

9. Resuspend the beads with the above reaction mix. Incubate at 37ºC with rotation on the Intelli-Mixer (Program F8, 30 rpm) for 30 min.

Circularization

10. Add 6 μl of T4 DNA Ligase (30 u/μl, Fermentas) into the above reaction. Mix immediately. Incubate at 22°C with rotation on the Intelli-Mixer (Program F8, 30 rpm) for 24 hours.

Elution and reverse crosslinking

11. Separate the material into 575 μl aliquots. To each aliquot, add 150 μl of ChIP Elution Buffer [final concentration of 1% SDS (Bio-rad), 0.1 M NaHCO3 (Sigma)] and 5 μl of Proteinase K (20 mg/ml, Ambion). Incubate the mixture at 65°C overnight.

DNA purification

12. The beads are pelleted and the supernatant is split into tubes of 500 μl. Do phenol/chloroform (Ambion) extraction with 500 μl phenol/chloroform using Phase-lock gel (Eppendorf) and isopropanol precipitation:

DNA (after phenol extraction)	~500 μl
3M NaOAc, pH5.2 (Ambion)	50 μl
GlycoBlue (Ambion;15 mg/ml)	2 μl
Isopropanol (Sigma)	500 μl

13. Incubate the above at -80ºC for 30 min; pellet DNA by centrifugation at maximum speed, 4ºC for 30 min. Wash the DNA pellet with 750 μl 70% ethanol twice and resuspend DNA in 20 μl of EB buffer (Qiagen).

Nick repair

DNA	20 μl
10×*E. coli* DNA Ligation Buffer (NEB)	5 μl
10 mM dNTPs (Eppendorf)	1 μl
E. coli DNA Ligase (10 u/μl, NEB)	1 μl
E. coli DNA Polymerase I (10 u/μl, NEB)	4 μl
H_2O	19 μl
	50 μl

Incubate at 16 ^{o}C for 16 hours.

14. DNA is adjusted to 200 μl with water and purified with 200 μl of phenol/chloroform using Phase-lock gel. The DNA is then ethanol precipitated as follows:

DNA (after phenol extraction)	~200 μl
3 M NaOAc, pH5.2 (Ambion)	20 μl
Ethanol (Merck)	600 μl

15. Incubate the above at -80 ^{o}C for 30 min; pellet DNA by centrifugation at maximum speed, 4^{o}C for 30 min. Wash the DNA pellet with 750 μl 70% ethanol twice and resuspend DNA in 20 μl of EB buffer.

16. **Mme I digestion to release PETs**

DNA	10 μl
10×NEBuffer 4 (NEB)	4 μl
10×SAM (freshly prepared; NEB)	4 μl
Mme I (2 u/μl, NEB)	1 μl
Unbiotinylated linker (200 ng/μl, to quench the excess MmeI; IDT)	4 μl
H_2O	17 μl
	40 μl

Incubate at 37 ^{o}C for 2 hours.

Prepare the Dynabeads

17. Mix Dynabeads (Invitrogen) well and transfer 50 μl of resuspended Dynabeads to a 1.5 ml tube. Stand for 1 min in the Magnetic Particle Collector (MPC; Dynal/Invitrogen). Remove the supernatant. Wash the beads twice with 100 μl of 2×B&W Buffer (final concentration: 10 mM Tris-HCl pH7.5 (Ambion), 1m M EDTA (Ambion), 2 M NaCl (1st Base)). Each time a wash is done, the following processes should be performed: mix, short spin, stand for 1 min, remove supernatant. When washing, do not let the dynabeads dry out. After removing supernatant from beads, immediately add another batch of supernatant. Do not spin the dynabeads at more than 800 rpm. Resuspend beads in 40 μl of 2×B&W Buffer.

Immobilization of the iPETs

18. Add 40 μl digestion mix (from Step 16) to the resuspended beads and mix well. Incubate at 22°C with rotation on the Intelli-Mixer (Program F8, 30 rpm) for 30 min. With the help of the MPC, wash the beads twice with 100 μl of 1×B&W Buffer (final concentration: 5 mM Tris-HCl pH7.5 (Ambion), 0.5mM EDTA (Ambion), 1M NaCl (1st Base)).

Ligation of modified 454 NN-adapters to the immobilized iPETs

19. Prepare the ligation mix:

Adapter A (200 ng/μl, IDT)	8 μl
Adapter B (200 ng/μl, IDT)	8 μl
10×T4 DNA Ligase Buffer (Fermentas)	5 μl
T4 DNA Ligase (30 u/μl, Fermentas)	1 μl
H_2O	28 μl
	50 μl

20. Resuspend the beads with the above ligation mix. Incubate at 22°C with rotation on the Intelli-Mixer (Program F8, 30rpm) for 16 hours.

Nick translation

21. Wash the beads twice with 100ul of 1×B&W Buffer with the help of the MPC.

22. Prepare the nick translation reaction mix:

10×NEBuffer 2 (NEB)	5 μl
10 mM dNTPs (Eppendorf)	2.5 μl (500 μuM final conc.)
E. coli DNA Polymerase I (10 u/μl, NEB)	4 μl
H_2O	38.5 μl
	50 μl

23. Resuspend the beads with the above reaction mix. Incubate at 22°C with rotation on the Intelli-Mixer (Program F8, 30rpm) for 1h.

Trial/QC PCR

24. Wash the beads twice with 100 μl of 1×B&W Buffer with the help of the MPC. Resuspend the beads in 50 μl of EB buffer. Transfer the mixture to a fresh 1.5 ml tube.

25. For each PCR reaction, in a 0.2 ml PCR tube, add and mix well:

Beads suspension	2 μl
Primer A (100 μM, IDT)	0.25 μl
Primer B-Biotin (100 μM, IDT)	0.25 μl
HotStarTaq Master Mix (Qiagen)	25 μl
H_2O	22 μl
	50 μl

The cycle conditions are:

95 °C, 15 min
94 °C, 30 sec
55 °C, 1 min
72 °C, 1 min
72 °C, 10 min
} 20 cycles

26. Remove the Dynabeads with the help of the MPC. Add 10 μl of 6x loading dye (Fermentas) to the 50 μl reaction and run all in one lane of a 5-well 6% TBE PAGE gel (Invitrogen). Run at 200V for 30 min. Stain with SYBR Green I (Invitrogen) for 15 min to visualize products. View using the Blue-light Darkreader (Clare Chemical).

Large scale PCR and PCR product purification

27. Estimate number of PCR reactions required based on the results from the trial PCR. Scale up accordingly. Pool the PCR products. Remove the Dynabeads with the help of MPC. Purify DNA by isopropanol precipitation with GlycoBlue as described earlier. Resuspend the DNA pellet in 40 μl TE buffer (Qiagen). Add 8 μl of 6x Loading Dye.

28. Run all on two lanes of the 6% TBE PAGE gel (Invitrogen, 5-wells), 200V, for 30 min together with 1μg of 25 bp DNA Ladder (Invitrogen) and 4 μl of Low DNA Mass

Ladder (Invitrogen). Stain with SYBR Green I for 15 min to visualize products. View using the Blue-light Darkreader.

29. The product is expected to be 164-174 bp in length, so a fairly broad smear should be seen. Excise the DNA bands of correct size.

DNA purification using the gel-crush method

30. DNA of interest is excised and collected into 0.6 ml micro tubes that have been pierced at the bottom with a 21G needle (Becton-Dickinson). The pierced tube is placed inside a 1.5 ml screw-cap micro tube, and centrifuged at 13K rpm, 4 ^{o}C for 5 min. The gel slices are thus conveniently shredded and collected in the bottom of each 1.5ml tube.

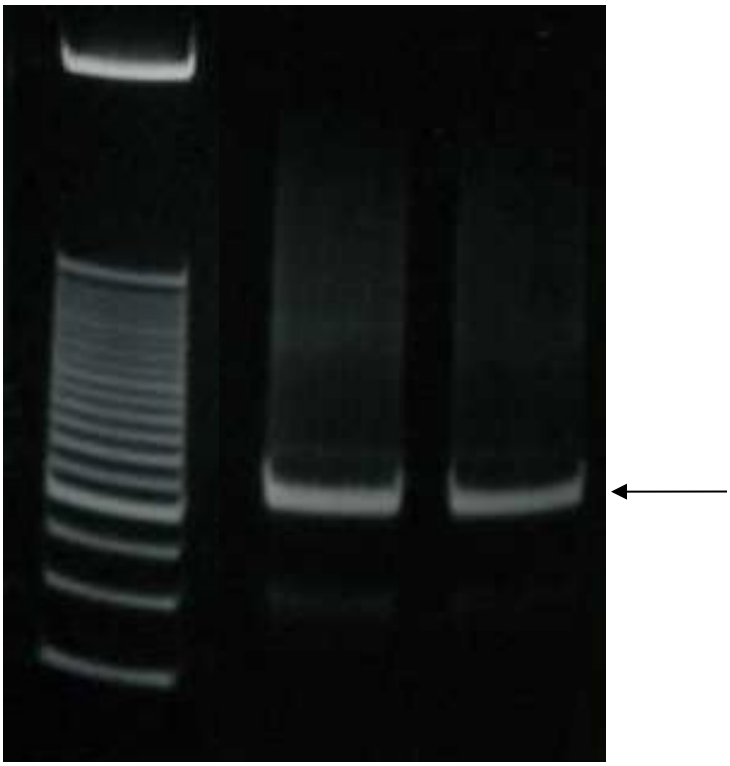
31. Add 200 μl of TE buffer to each 1.5 ml screw-cap micro tube and stir the gel pieces with the pipette tip. Make sure the gel pieces are immersed with the buffer.

32. The 1.5 ml screw-cap micro tubes containing shredded gel are frozen at -80 ^{o}C for 1-2 h, and then transferred directly to 37^{o}C incubation. The shredded gel is thus macerated at 37^{o}C for 16 h.

33. The gel pieces together with the buffer in each 1.5 ml tube are transferred to the filter cup of a SpinX microspin filter unit (Corning) and are centrifuged at 13,000 rpm, 4^{o}C for 10 min. At the same time, rinse the 1.5ml tubes, which have been used to macerate the shredded gel, with 200 μl TE buffer (Qiagen) and collect the liquid by brief spin.

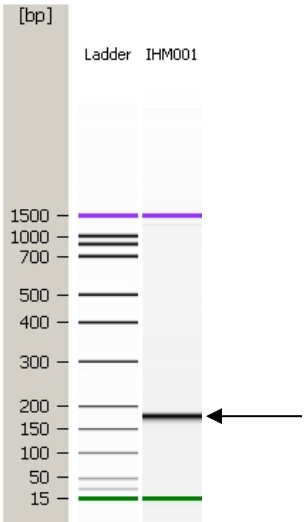
34. When the above centrifugation finishes, add each of 200 μl rinsing buffer to the filter cup of each filter unit. Stir to loose the gel pie with a pipette tip. Centrifuge again at 13,000 rpm, 4^{o}C for 20 min.

35. Pool the filter-through. Perform isopropanol precipitation. Resuspend the DNA in 20 μl of TE buffer.



Gel picture of the MmeI-cut ChIA-PETs.

36. Quality Control: Analyze the DNA using a DNA1000 Labchip (Agilent) using 1 μl of sample to determine the quality and quantity of the recovered DNA.



Agilent Bioanalyzer picture of the MmeI-cut ChIA-PETs.

ChIA-PET DNA sequencing analysis

37. At this point, the purified ChIA-PET templates are ready for multiplex sequencing analysis by GS20 or GSFLX. The sequencing was done following the manufacturer's protocol conditions.