FUNCTIONAL ARCHITECTURE AND EVOLUTION OF *CIS*-REGULATORY ELEMENTS THAT DRIVE GENE COEXPRESSION

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF GENETICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFULLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Christopher David Brown

August 2007

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Arend Sidow, Principal Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Richard Myers

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Gregory Barsh

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Dmitri Petrov

Abstract

Transcriptional coexpression of genes that encode interacting products is fundamental to organismal biology, yet the functional architecture and evolution of *cis*-regulatory elements that orchestrate coexpression remain largely unexplored. In this study, I exhaustively mutagenized 19 regulatory elements that drive coexpression of *Ciona* muscle genes, and obtained quantitative estimates of the activity of the 77 transcriptional regulatory motifs that comprise these elements. I find that individual motif activity ranges broadly within and among elements, and among different instantiations of the same motif type. The activity of orthologous motifs is strongly constrained between the two genomically highly divergent *Ciona* species, suggesting that precise control of element function is superbly important for survival of the organism. By contrast, motif arrangement, type, and activity varies greatly among the elements of different coregulated genes. Thus, architectural rules governing regulatory function are flexible, but become highly constrained evolutionarily once they are established in a particular element.

Acknowledgments

I am indebted to the help of many friends and collaborators. My thesis committee, Rick Myers, Greg Barsh, and Dmitri Petrov, provided many helpful insights into this work and provided needed support to push this work to an ambitious conclusion. My advisor, Arend Sidow, provided me an excellent learning environment, with the right balance of support and independence. He has also provided me with an excellent role model for professional and personal life balance. Just as importantly, the members of the Sidow lab over the last six years, including Jon Binkley, David Goode, Matt Hill, Dori Hosobuchi, Kerrin Small, Eric Stone, Anton Valouev, and Mehdi Yahyanejad have been fantastic friends and collaborators. I am particularly indebted to the help Greg Cooper and Cheryl Smith, two coworkers who have influenced all of this work, and Dave Johnson, who started the Stanford *Ciona* project. This manuscript was greatly improved by the criticisms and suggestions of Patrick Collins, Jerel Davis, and Jed Dean. I am also grateful for the excellent technical assistance of Krisztina Perez and Betsy Anton. This work was supported in part by a National Science Foundation Graduate Research Fellowship, the Stanford Genome Training Program (NIH/NHGRI), and grants to Arend from NIH/NHGRI and NIH/NIGMS. Most importantly, I would like the thank Katherine for support and patience, without which none of this would have been possible.

Table of Contents

List of Tables

List of Illustrations

Chapter 1 Introduction

The majority of a complex metazoan genome encodes sequence that is not transcribed into RNA ('non-coding' DNA). A significant fraction of such non-coding DNA is responsible for regulating the spatiotemporal specificity of gene transcription. Such regulation is the first major step in the regulation of the flow of biological information from a genotype, encoded in DNA, a molecule that largely functions as a repository of information, to phenotype, produced by biologically active molecules such as RNA and protein (Jacob and Monod 1961; Britten and Davidson 1969). It is this regulation that controls the development of complex organisms; composed of a multitude of distinct cells types that each contains the same genetic blueprint. Sequence elements that regulate gene expression come in a variety of forms, but here I will be principally concerned with *cis*-regulatory elements that control the complex specificity of gene expression and their associated basal promoters.

The overarching goal of this research is to improve our understanding of the functions that connect genotype to phenotype. More specifically, my interests lie in deciphering the mechanisms by which non-coding DNA produce phenotypes via gene expression and the effect of sequence changes on such phenotypes on macro- and micro-evolutionary timescales. To address these interests, I have used two complementary approaches: a molecular dissection of *cis*-regulatory function and an experimental and data analysis framework that inform the evolutionary process. It is my hope that determining the functional consequences of non-coding sequence changes will address such disparate topics as the genetic basis of butterfly wing

pattern variation or the inherited changes that distinguish humans from our closest primate relatives, in addition to being beneficial for the prediction and amelioration of human disease processes.

Mechanisms of non-coding DNA function

Cis-regulatory elements control the function of DNA by regulating temporal and spatial specificity (Small et al. 1991; Hoch et al. 1992; Corbo et al. 1997), the effects of environmental stimuli on gene expression (Thanos and Maniatis 1992; Garrity et al. 1994), as well as controlling the quantity and specificity of the transcript produced. *Cis*-regulatory elements are typically assembled from compact collections of sequence-specific transcription factor binding sites ('motifs') (Crowley et al. 1997; Berman et al. 2004; Zhou et al. 2004), which represent the fundamental units of *cis*-regulatory function. *Cis*-elements also have an inherent higher-order structure: the presence of particular regulatory motifs is not sufficient for biological activity, but elements require such motifs to be in particular arrangements, with particular spacing, orientation, order, and location. This higher-order arrangement of regulatory motifs is hereby referred to as 'regulatory architecture.' The regulation of gene expression is thought to occur through the binding of transcription factors to regulatory motifs in *cis*-elements that are able to modulate the assembly and activity of the basal transcriptional machinery assembled at the gene promoter. *Cis*-regulatory elements perform a simple, biological computation: based on the protein complement of a particular cell and the specific *cis*-regulatory architecture of an element, they integrate

multiple regulatory inputs to produce specific and distinct gene expression patterns (for an extensive review, please see Davidson 2001).

Experimental dissection of *cis*-regulatory function

The interpretation of the *cis*-regulatory genetic code is hampered by two important characteristics, its degeneracy and apparent redundancy. Degeneracy is evident in the sequence specificity of transcription factor binding sites. While individual transcription factors preferentially bind to particular sequences, this preference is probabilistic. Typically, transcription factors do not simply bind to a single specific sequence but are instead able to bind to groups of similar sequences. I will refer to the probabilistic model of such collections as 'motif types,' whose specificity is often represented in the form of a Position Specific Scoring Metric ('PSSM'). While such sequence differences may result in protein binding of varying affinity and functional consequences, relatively little data exists to quantitatively characterize the effects of binding site degeneracy.

Moreover, the vast majority of sequences within a given genome with significant similarity to a transcription factor binding site are in fact not bound by the protein (I will call such sequences 'false-positive' motif predictions) and thus do not modulate gene expression (Lieb et al. 2001). Therefore, due to the short length (sequence specific transcription factors typically recognize sequences of 4 to 15 bases) and degeneracy of such protein binding sites, the identification of transcription factor binding in large genomes based on primary sequence data alone, or in combination

with low-resolution experimental data, is extremely difficult. The confident annotation of regulatory motifs thus requires one to distinguish functional regulatory motifs from false-positive motif-like sequences with experimental resolution at, or close to, the level of individual transcription factor binding sites.

Secondly, transcription factor binding sites are often difficult to identify because they often appear to act 'redundantly.' The modification or deletion of transcription factor binding sites is often asserted to produce no phenotypic effect in the presence of similar sequences in the local vicinity (Buttgereit 1993; Laney et al. 1996; Belting 1998; Piano et al. 1999; Hersh et al. 2005; Pappu et al. 2005). However, such studies are often hampered by the interpretation of a negative result: the lack of a detectable phenotype might be attributable to insufficient functional sensitivity or simply ignorance as to the nature of the phenotype and/or a proper assay for it. An alternative hypothesis tested in this work is that individual regulatory motifs are responsible for non-redundant quantitative fractions of the total regulatory activity of a *cis*-element (e.g., Galant et al. 2002). Regardless of the specific mechanism producing small or undetectable phenotypic changes, they nonetheless further increase the difficulty of the identification and characterization of *cis*-regulatory elements and their constituent motifs.

However, given proper experimental resolution and sensitivity to quantitative differences in phenotype such redundancy, a specific type of genetic interaction, can be rigorously characterized. Genetic and functional interactions between individual transcription factor binding sites are critical for a comprehensive understanding of *cis*-regulatory element function, as evidenced by the complex interactions (of a form

different from 'redundancy') shown to dominate the molecular function of well-characterized regulatory elements such as the *eve* stripe 2 element (S2E). At S2E, multiple binding sites for overlapping activators and repressors function together to produce a novel expression domain (Stanojevic et al. 1991; Small et al. 1991; Small et al 1992; Arnosti et al. 1996; Ludwig et al. 2000;).

The exquisite experimental resolution characterizing the molecular architecture at S2E has resulted in its treatment as a paradigm for *cis*-regulatory element function. While this paradigm might hold as an excellent example for many other regulatory elements involved in the process of developmental pattern formation, it may not represent a model for other classes of *cis*-elements. For example, *even-skipped* enhancers establish new expression domains by integrating the regulatory inputs of multiple positive and negative factors, some functioning as morphogens, that are distributed across a syncytial embryo. It may be possible, however, that the regulatory architecture of other classes of *cis*-elements differs from this model, for example those elements that regulate genes expressed in differentiated cell types after initial developmental pattern formation has been established. Such genes may be regulated by transcription factors whose expression patterns have already been confined to the same tissue-restricted expression patterns. Given that the regulatory computation necessary at terminal differentiation *cis*-elements appears much simpler, might the elements driving such expression patterns be governed by a different set of architectural rules?

The analyses presented in this study aim to address the molecular degeneracy and redundancy of *cis*-regulation with increased experimental resolution and a novel

analytical framework. This approach allows one to unambiguously distinguish functional regulatory motifs from false-positive motif-like sequences due to motif-level experimental dissection of *cis*-elements and the quantitative assessment of the phenotypic effects of such perturbations. Moreover, the quantitative gene expression measurements produced in this study allow for the rigorous analysis of genetic interactions between regulatory motifs within a *cis*-element.

Lastly, the resolution and quantification produced by this study allow one to analyze a fairly novel quantity: the *cis*-regulatory activity of individual regulatory motifs. Experimentally based measurements of individual motif activity permit the analysis of regulatory architecture at an unprecedented level of resolution. Estimation of this quantity also permits the characterization of the effects of differential motif activity on the selective pressure exerted on the motif. Such analyses are a necessary step towards a more comprehensive understanding of the effects of sequence changes on non-coding DNA.

Coordinate gene regulation

Another interesting mechanistic aspect of *cis*-regulation is the phenomenon of coordinate gene regulation. Quite simply, genes whose products are involved in the same molecular process must be expressed at the same time and place in order for their products to functionally interact. Such coordinately expressed gene sets, often called ‘gene batteries,’ largely determine the function of the cell types they are expressed in; for example, muscle cells are muscle cells in part because the genes of

the multiprotein complex of the muscle strand are expressed together in large quantities in this cell type. The fundamental importance of such coexpressed gene sets is underscored by their conservation across large evolutionary distances (Stuart et al. 2003). Despite its centrality to organismal biology, the mechanistic basis of gene coexpression remains poorly characterized.

Coordinate regulation of genes that must be expressed in the same patterns is generally thought to be achieved by *cis*-regulatory elements that share functional characteristics and therefore respond at the transcriptional level to similar regulatory inputs (Davidson 2001; Yilpel et al. 2001; Berman et al. 2002; Zhou et al. 2004; Segal et al. 2003; Johnson et al. 2005). Whether these shared characteristics are encoded by similar element architectures, what the underlying structural and functional commonalities are, and how the function of coregulatory elements evolves has been largely unexplored.

One approach that has enlightened this question is systems-level analysis of gene expression data and computational analyses adjacent non-coding sequences. Starting from the hypothesis that genes with similar expression patterns are likely to be regulated by similar regulatory inputs, several studies have identified the regulatory factors responsible for the gene expression pattern as well as their constituent regulatory motifs (Pilpel et al. 2001; Segal et al. 2003; Johnson et al. 2005). These studies have suggested that groups of coexpressed genes are typically regulated by small groups of transcription factors. The regulatory motifs identified in such analyses generally correspond to transcription factor binding sites whose instantiations within

individual regulatory elements provide the statistical signal for the identification of the module.

One interpretation of the success of such systems-level studies, which usually do not involve experimental dissection of individual *cis*-elements for functional components, is that a collection of co-expressed *cis*-regulatory elements may adhere to a defined grammar and that there may exist distinct syntactical rules for the responding *cis*-elements to achieve coregulation. Such rules might involve the necessity of a stereotypic collection of transcription factors providing the regulatory input, and commonalities in the *cis*-regulatory architectures of co-expressed genes (as suggested in Senger et al. 2004).

Alternatively, given that selection acts via phenotype (in this case, the precise timing and pattern of gene expression) on the genotype (the primary sequence of the *cis*-element), it is conceivable that coregulation is not achieved with strictly defined rules but with a diversity of mechanisms that all generate the same output. Such flexibility would be possible if diverse *cis*-regulatory architectures could produce the same phenotypic output, if regulatory inputs were interchangeable, and if functional interactions between regulatory motifs were either flexible or relatively unimportant. The resolution of functional studies in multicellular systems has thus far been inadequate to definitively rule out either of these two alternatives, though I note that among the diversity of different coregulation modules there will likely be a variety of behaviors, not all of which will be clearly 'strict' or 'loose'. The quantitative characterization of *cis*-regulatory architecture in this study permits a comparison,

across coregulated *cis*-elements, of regulatory motif content, activity, arrangement, and interactions.

Evolution of *cis*-regulation

A second major area of inquiry into mechanisms of *cis*-regulation concerns the functional evolution of each *cis*-element that interprets regulatory inputs. After their initial establishment, over subsequent evolutionary time, *cis*-elements continue to be subject to the forces of mutation and selection. To date, the vast majority of molecular evolutionary analyses have been conducted on coding DNA. However, it was suggested over twenty years ago that diverging gene expression mechanisms might account for a large proportion of the phenotypic differentiation between species (King and Wilson 1975). This hypothesis has been supported in recent years with a number of lines of evidence.

First, it now appears that in many metazoan genomes, there are a larger number of functional bases in the non-coding portion of the genome than the coding. Based on the work of Siepel and Cooper (Cooper et al. 2005; Siepel et al. 2005) one can estimate the fraction of non-coding sequence that is evolving under selective constraint. This fraction could be an over estimate, due to the presence of other non-coding sequence types that may be evolving under purifying (e.g., unannotated non-coding RNAs) and it may also be and underestimate, if, as has been recently suggested, that pervasive adaptive evolution is responsible for a large fraction of

sequence changes. However neither of these concerns seems likely to significantly bias a comparison to coding DNA.

Second, a handful of experimental studies have demonstrated the phenotypic consequences of non-coding sequence changes. In particular the work of Sean Carrol's lab has characterized several species-specific non-coding sequence changes producing phenotypic effects among the drosophilids (Gompel et al. 2005; Hersh 2005; Jeong 2006; Prud'homme et al. 2006). Similarly, recent work has suggested that non-coding sequence variants are responsible for ecologically important phenotypic differentiation in sticklebacks (Shapiro et al. 2004) as well as the extreme variation in size of dog breeds (Sutter et al. 2007). In addition, several recent studies have identified non-coding sequence variants as causative lesions for human phenotypes and diseases, including lactase persistence (Tishkoff et al. 2007), polydactyly (Lettice et al. 2002), Hirschsprung disease (Emison et al. 2005), and cancer (Rioux et al. 2007).

Thirdly, several recent studies have addressed the most immediate phenotypic effect of cis-regulatory change, gene expression patterns. Across multiple inter-species comparisons, analyses of genome-wide (or nearly so) expression levels have suggested that the expression patterns of many genes have evolved under varying levels of stabilizing or purifying selection (Rifkin et al 2003; Khaitovich et al 2005; Lemos et al. 2005; Gilad et al. 2006), thereby demonstrating the evolutionary relevance of quantitative changes in *cis*-regulatory function.

An understanding of the changes permitted by the evolutionary process and their resulting phenotypic effects will, in combination with high resolution functional data,

enlighten efforts to understand the relationship between non-coding genotype and phenotype, as well as enhance our ability to predict the functional consequences of extant sequence variation (for example, in large scale disease association studies).

Evolutionary constraint in *cis*-regulatory sequences

The phenotypic effects resulting from sequence changes in functional non-coding sequences suggests that, as a class, functional non-coding sequences will evolve at reduced rates due to the effects of purifying selection. In the case of genes that produce the components of a universally important and highly conserved function (such as the muscle strand proteins encoded by the genes of this study), purifying selection dominates over positive selection because a vastly larger number of mutations that affect the system are deleterious than are advantageous. The predominance of purifying selection leads to a reduction in the evolutionary rates of functionally important *cis*-elements, which is referred to as 'evolutionary constraint' (Kimura 1983).

Numerous investigations have leveraged the signal of constraint to identify *cis*-regulatory elements (e.g., Gibbs 2004; Woolfe et al. 2005; Pennacchio et al. 2006), suggesting that constraint is the norm in *cis*-regulatory regions. Similarly, several analyses of functionally verified regulatory elements have shown them to be more constrained than 'background' genomic DNA (Moses et al 2003; Encode Project Consortium 2007). The genomic distribution of conserved noncoding sequences has also hinted at their functional relevance. Such sequences are overrepresented near

genes with complex expression patterns (Nelson et al. 2004; Woolfe et al. 2006), in long introns and intergenic regions (Halligan and Keightly 2006). They are also clustered in primary sequence (Bergman et al. 2002; Webb et al. 2002) and such clustering is conserved (Bergman et al. 2002). Based on inter and intra-specific sequence comparisons, it is also clear that the decreased evolutionary rate in conserved non-coding sequences is due to purifying selection, as opposed to mutation rate heterogeneity (Drake et al. 2006; Casillas et al. 2007).

Sequence turnover in *cis*-regulatory sequences

Sequence constraint does not, however, imply rigid conservation and a number of theoretical (Stone and Wray 2001), comparative (Richards et al. 2005; Margulies et al. 2007), and experimental analyses (Dermitzakis and Clark 2002; Dermitzakis et al. 2003; Moses et al. 2006) have suggested that a significant fraction of putative transcription factor binding sites might evolve quickly under little evolutionary constraint. The theoretical analyses of Stone and Wray suggested that due to the short length and degeneracy of *cis*-regulatory motifs, similar motifs might arise *de novo* via local point mutations within regulatory regions relatively often. Through evolutionary simulations, the authors demonstrated that this process might produce binding site flux, in which the creation of new binding sites allows for the accumulation of sequence substitutions in older regulatory motifs, eventually leading to their functional replacement by such new motifs.

Several comparative sequence analysis based studies have also suggested that *cis*-regulatory DNA is evolving under low levels of selective constraint. Most prominently, the sequencing and analysis of the *Drosophila pseudoobscura* genome demonstrated that, as an annotation class, *cis*-regulatory DNA is only slightly more constrained than random intergenic DNA. Recently, the ENCODE consortium (Margulies et al. 2007; ENCODE Project Consortium 2007) produced estimates of mammalian evolutionary rates and generated experimental data pertaining to *cis*-regulatory function from multiple high-throughput experimental platforms. Analysis of the overlap of these data sets also suggests the existence of a large amount of functional non-coding DNA evolving under only minimal amounts of selective constraint.

Several more directed experimental studies have suggested the existence of large-scale regulatory motif turnover. Two surveys conducted by Dermitzakis and Clark identified potential cases of regulatory motif turnover in Drosophila and mammalian genomes (Dermitzakis and Clark 2002; Dermitzakis et al. 2003). More recently, Alan Moses, Mike Eisen, and colleagues conducted an analysis of *Zeste* binding across the *Drosophila melanogaster* genome, complemented by extensive comparative sequence, evolutionary, and computational *cis*-regulatory analyses, which concluded that 5% of *Zeste* binding sites are turned over within a survey of four Drosphilids (Moses et al. 2006). Perhaps the best-studied example is that of the *Drosophila eve* stripe 2 element, which during the course of evolution has maintained its precise phenotypic output despite significant functional sequence turnover (Ludwig et al. 2000; Ludwig et al. 2005).

Evolution of *cis*-coregulatory mechanisms

No case study has so far experimentally determined the functions of a large number of *cis*-coregulatory motifs that mediate gene expression in the tissues of higher organisms, and analyzed them in light of their evolutionary trajectories. The contrast between studies that suggest the predominance of constraint or motif turnover undoubtedly reflects the vast diversity of possible outcomes of evolution's experiments; on the other hand, the diversity of conclusions underscores that our insights into the architecture and evolution of *cis*-regulatory function are based on either low-resolution data across many loci or on higher resolution experiments on a handful of single elements (such as the *eve* stripe 2 enhancer). To my knowledge, there exist no high-resolution, quantitative studies of *cis*-regulatory function whose conclusions are supported by many loci that have evolved under similar evolutionary constraint.

I reasoned that a high-resolution experimental characterization of *cis*-regulatory architecture, in which the activities of individual regulatory motifs are characterized, might shed light on many of the unanswered aspects of *cis*-regulatory evolution. Such data may inform us about the underlying functional causes of motif conservation or turnover in several ways. First, the experimental resolution provided will allow the unambiguous differentiation of functional and false positive motif sequences, thereby reducing sequences that may bias estimates of sequence evolution. Secondly,

comparisons of motif activity with the evolution of their sequence and function will inform us about the determinants of motif turnover.

Therefore, to address certain fundamental properties of coregulation, namely its regulatory architecture, its evolution of function, and its sequence turnover in *cis*-elements, I embarked on a comprehensive, high-resolution, functional and evolutionary study of 19 genes coregulated by the *Ciona* muscle module (Johnson et al. 2005). The muscle module directs specific expression of these genes in the 36 muscle cells of the developing tadpole larva. 17 of the genes function in the same macromolecular complex, the muscle filament, emphasizing the requirement for tight coregulation of these genes that is also evident from whole mount *in situ* hybridization time courses (Johnson et al. 2005).

Ciona lends itself to quantification of the function of tissue-specific positive regulatory elements, as each transfection with a reporter construct usually results in more than fifty, and often more than a hundred, transgenic animals (for a thorough review of *Ciona* as an experimental model system please see Johnson 2005). This allows for the rapid production of *in vivo* expression measurements of reporter constructs in the proper developmental context. Statistical analyses of control transfections showed that each muscle cell decides autonomously whether to express a reporter gene, which led us to devise a scoring scheme that estimates the fraction of muscle cells expressing the reporter in a field of transfected embryos. In conjunction with stereotyped and reproducible transfection and assay conditions, this provides for a unified system of quantification across all loci and across all constructs that measures the probability of any given cell expressing the transgene. The unit of the

scores I report here is therefore "muscle cell expression probability". The scores of all constructs can be directly compared and provided as input into statistical modeling of function. The combination of quantification of expression probability and multivariate regression modeling allowed the estimation of the functional contribution of individual regulatory motifs. The high resolution of the regulatory constructs facilitated motif-level characterization of the molecular architecture of *cis*-coregulatory function and of the evolutionary dynamics of this system.

Our coregulated gene set comprises six single-copy genes from *C. savignyi* and their six orthologs from the sister species, *C. intestinalis*: α-Tropomyosin 1, α-Tropomyosin 2, Myosin Binding Protein, Troponin I, Troponin T, and Creatine Kinase. In addition, seven genes of *C. savignyi* that belong to multicopy gene families were dissected: Muscle Actin (MA), Myosin Light Chain (MLC), and Myosin Regulatory Light Chain (MRLC). MA is encoded by at least twelve presumably isofunctional genes in the *C. savignyi* genome, and I dissected the regulatory regions of two of them. Similarly, MLC is encoded by at least four genes (I dissected two) and MRLC is encoded by at least six genes (I dissected three). 17 of the 19 genes function in the same macromolecular complex, the muscle filament, emphasizing the requirement for tight coregulation of these genes that is also evident from whole mount *in situ* hybridization time courses (Johnson et al. 2005).

Chapter 2 High resolution dissection of *cis*-regulatory function

Identification of *cis*-elements

All constructs utilized in this study are based on initial wild type constructs that contain 2-5kb of upstream sequence from each gene, the endogenous promoter, the start codon, and small amounts of exonic sequence fused in frame to the lacZ reporter gene. In order to make use of a wide dynamic range of expression probability for assaying the function of the mutagenized constructs, I tuned the transfection protocol (see Methods) so that most wild type constructs drove expression in over 30% but less than 80% of muscle cells, as opposed to the 100% that would be the norm for the endogenous locus.

I built hundreds of constructs, assayed in well over 2,000 transfections, to define, for each locus, the *cis*-regulatory elements responsible for the majority of the transcriptional activity. Such *cis*-element identification was achieved with the production, at each locus, of a series of deletion constructs that were transfected in replicate and scored qualitatively. Putative *cis*-elements were identified as sequences that, when deleted, resulted in a significant decrease in the expression probability of the reporter.

Deletion constructs removed sequences from the 'full-strength' constructs, typically from the end distal to the endogenous promoter, although a small number of deletions were created from the promoter proximal end or internally. This distal deletion bias is the result of a technical concern: Because the purpose of this study

was to characterize the nature of tissue-specific positive *cis*-regulatory elements; and due to the relatively unannotated state of the *Ciona* genomes, I was cautious not to delete either unrecognized coding sequences or sequences of the basal promoter. The accidental deletion of coding sequences could produce frame shift mutations resulting in 'false-negative' expression measurements. Similarly, the deletion or mutation of sequences in the basal promoter could decrease or eliminate expression in a non-tissue-specific manner. The compact *cis*-elements identified explained between 55% and 100% of the function in the wild type construct.

Quantitative fine-scale dissection of compact *cis*-elements

The deletion series described above identified compact *cis*-regulatory elements of approximately 100 nucleotides. Even at this level of resolution, however, it remained impossible to distinguish functional regulatory motifs from motif-like sequences with no regulatory activity. As a result, I could not make firm conclusions about the detailed molecular architecture underlying gene coexpression or about the extent of homologous motif turnover. To improve the dataset, I applied two approaches. First, I developed a quantitative assay and scoring system that allowed us to characterize gradations of regulatory activity (as opposed to binary calls of active/non-active) in a robust fashion. Second, I refined the resolution of our experiments by conducting a high-resolution mutagenesis scan, guided in part by predictions of motif sequences.

Dissection of the fine-scale molecular architecture of each *cis*-regulatory element was performed with a combination of small deletions (5-10bp) and site-directed

mutageneses that removed specific motifs or short sequences not matching the motifs. The three motifs utilized here are the Cyclic AMP Response Element (CRE; Kusakabe et al. 2004; Chen et al. 2005; Johnson et al. 2005), the *Ciona* MyoD motif (Blackwell and Weintraub 1990; Johnson et al. 2004; Meedel et al. 2007), and the *Ciona* Tbx6 motif (Yagi et al. 2005). All had been previously shown to be involved in muscle gene expression (see above references), and the present study provides no evidence of any other motifs involved in this process. These short subsequences were mutagenized in isolation or in a large number of combinations to produce 220 constructs that form the basis for all quantitative analyses in this study. Mutagenesis was carried out using two methods: (1) Fine-scale deletions, of approximately 5 to 10 nucleotides, that deleted individual putative regulatory motifs from the distal end of the construct, and (2) Site-directed mutations that scrambled the sequence of a motif, while maintaining local GC content and spacing between adjacent sequences (Fig. 2.1).

Initial analyses of five independent transfections and assays of the same constructs ("biological replicates") showed that the results were remarkably reproducible presumably because thousands of cells are assayed in each transfection, and because I had developed stereotypic transfection conditions. The replicates also resulted in stable estimates of activity for each construct, as revealed by the standard deviation of the fraction of expressing cells for each construct (mean SD = 0.074 efu, median SD = 0.064 efu). I therefore assayed, for nearly every fine-scale construct, at least five biological replicates (mean = 5.04). In the end, the dataset upon which quantitative

analyses were performed consisted of 1237 transfection assays that yielded a total of 85,506 transgenic embryos (Fig. 2.2).

At each locus, the data clearly distinguish functional regulatory sequences from non-functional background sequences. The identified regulatory sequences control ~85% percent of assayed activating function across all loci. The subset of constructs that targeted individual motifs modified the expression frequency by 0% to 75%, leading to two important conclusions. First, many motif-like sequences were deleted or mutated with no significant functional consequences detectable in our assay. Such sequences therefore represent false-positive motif predictions that cannot be distinguished from functional motif predictions on the basis of primary sequence alone. Second, there exists a broad, quantitative gradient of function among regulatory motifs. Therefore, functional motifs, even those matching the same consensus binding site, vary in the amount of regulatory function they contribute to the locus. The exact determinants of individual motif activity remain unknown, but, given that activity is not significantly correlated with strength of match to the motif consensus matrix, local or regional sequence context is certainly important.

Quantitative framework for the estimation of motif function

The density and depth of the dataset presented the opportunity to address several novel questions about *cis*-regulatory architecture. However, a quantitative and biologically meaningful representation of the functional architecture of each *cis*-element required an analysis framework to estimate the activity of each motif. To

develop a framework I needed: (a) a robust metric for the measurement of reporter expression level, (b) an analysis of the basic descriptive statistics of our data set, (c) an analysis of the importance of genetic interactions versus independence among motifs, and (d) a proper mathematical framework to relate explanatory (motifs) and dependent (expression measurements) variables.

Expression frequency scoring

The transfection of most functional reporter constructs produces a collection of embryos that express the reporter at varying levels (see Methods). The sources of this embryo-to-embryo variation in reporter expression are ambiguous, but at least partly due to transgene mosaicism (Zeller 2004; Zeller et al. 2006) and cell autonomous stochasticity of gene expression (Fiering et al. 2000; Raser and O'Shea 2004; Raser and O'Shea 2005). Traditionally, electroporated *Ciona* embryos have been scored by the percentage of embryos that express the reporter in the cells of interest. However, given the embryo-to-embryo variation mentioned above, I reasoned that scoring transfections based on the percentage of stained cells of interest would be a more informative metric (note that such a visual scoring system allows the experimenter to ignore ectopic expression). Similar cell-type specific scoring metrics have been previously employed to score Ciona transgene expression (Bertrand et al. 2003; Oda-Ishii at al. 2005).

Several lines of evidence suggest that this scoring metric is robust. First, a comparison of embryo based and cell cased scores (Fig. 2.3) reveals that the cell based

scoring metric has more resolving power for moderate to strong reporter constructs. This increased resolution appears to result from the early saturation of the embryo based score; all embryos have a stained muscle cell well before all muscle cells stain. While it might be the case that increased high-end resolution comes at the cost of decreased low-end resolution (note Fig. 2.3 A change in slope), greater resolving power for moderate to strong constructs is of more practical interest given structure of this particular data set.

Second, an analysis of the distribution of stained cells per embryo demonstrates that a cell based scoring metric captures a cell autonomous probabilistic shift in the frequency of reporter expression (Fig. 2.3; see also Methods). Across all transfections, as the percentage of stained muscle cells increases (or the percentage of class 0 embryos decreases) there is a sequential increase in the percentage of embryos stained in greater numbers of cells. Again, this suggests that an increase in the percentage of embryos staining is driven by an increased percentage of muscle cells staining, which are effectively randomly distributed across a collection of embryos.

Third, a cell-based scoring metric produces results with decreased variance across replicated transfections of the same construct. This trend is seen in both the raw variance (Fig. 2.4 A) as well as variance as a fraction of the mean expression frequency of a construct (Fig. 2.4 B). This suggests that cell-based scoring metrics are a more accurate summary of the underlying biological reality. Fourth, construct activities measured as the percentage of muscle cells stained are directly proportional to lacZ RNA levels as measured by quantitative RT-PCR (Fig. 2.5).

Descriptive statistics of data set

Due to the stereotyped transfection and scoring techniques developed over the course of this investigation, the electroporation of Ciona embryos produces expression measurements suitable for quantitative downstream analyses. Overall, the mean variance of all ~1200 replicated transfections is 0.011efu. This equates to a mean variance as a fraction of the mean expression frequency of 0.044. The overall distributions of these two statistics can be seen in Fig. 2.4 A, B, and C. 89% of replicated transfections have variances that are less than 10% of the mean expression frequency of the construct (Fig. 2.4).

As seen in Figure 2.2 C, the variance in expression frequency varies as a function of the mean, with decreased variability in the tails of the distribution of means and greater variability seen at moderate expression levels. This distribution suggests that the distribution of stained muscle cells might be best modeled as a draw from a multinomial distribution (Sokal and Rohlf 1995). As predicted for percentage based scoring metrics, some of this variance-mean dependence is removed by subjecting the raw expression measurements to the angular transformation (Fig. 2.4 C and D).

Due to the structure of the data set and the relative paucity of true experimental replicates (~5 per construct) I assessed the normality of the data by constructing a distribution of standard deviates (Fig. 2.6; see also Methods). The untransformed distribution of standard deviates exhibits a significant rightward skew and platykurtic dispersion. Both of these deviations from normality probably result from

measurements at or near the bounds of the frequency distribution used in this study. These deviations are partly reduced by either a square root or angular transformation. The distribution of variance and individual deviates are further discussed below, with regards to model choice.

Analysis of genetic interactions

To arrive at specific estimates of motif activity, which would be needed for all downstream analyses, I searched for a realistic statistical modeling framework. To accurately model the function of *cis*-elements at the level of individual regulatory motifs, I needed to determine whether *cis*-elements are built from motifs that each function largely independently or whether *cis*-element function can only be accurately described using models that account for functional interactions between individual motifs. The presence of motif interactions would necessitate statistical frameworks, such as ANOVA, that accommodate interaction effects between explanatory variables. However, if motifs appear to act independently, simpler approaches, such as linear regression, that require the estimation of fewer parameters, might be appropriate.

In order to determine what types of statistical approaches would best allow quantitative modeling of regulatory function within the *cis*-elements, I conducted an analysis of genetic interactions among a suitable subset of the fine-scale mutants. An assessment of the frequency and magnitude of genetic interactions is necessary to determine if statistical analyses of *cis*-element function must account for inter-motif interaction effects, or if simpler models assuming motif independence are sufficient.

The pertinent subset of data from our experiments were the expression values for 18 sets of constructs, where a set is defined as two constructs that each contain a single motif mutant, one construct that contains the double mutant, and relevant wild type constructs. The approach used had been successfully used in the quantification of interactions between gene deletions or amino acid substitutions (Tong et al. 2004), in regulatory network analysis (Segre et al. 2005), and in theoretical evolutionary and population genetics (Elena et al. 1997). Quantitative comparisons of the expression frequencies of each member of the set allow determination as to whether the individual mutations genetically interact.

I examined the distribution of interaction terms under an additive model and a multiplicative model (Cordell 2002). In the multiplicative model, the relationship between the functional consequences of a double mutant, W_{xy}, and the product of the single mutants, W_xW_y, defines the genetic interaction of the two mutations, denoted as $\varepsilon_m = W_{xy} - W_xW_y$. In the additive model, interactions are defined as $\varepsilon_a = (1 – W_x) + (1 - W_y) - (1 - W_{xy})$. In our study, W is the expression frequency of double or single mutant constructs relative to the expression driven by the wild type construct. Across 18 such comparisons, ε_m varies from –0.39 to +0.26, with 10 comparisons ranging between 0 and -0.1 ($\varepsilon_{mean} = -0.039$, $\varepsilon_{median} = -0.0034$, $\varepsilon_{variance} = 0.023$). Slightly larger interaction effects were observed for ε_a ($\varepsilon_{mean} = 0.20$, $\varepsilon_{median} = 0.15$, $\varepsilon_{variance} = 0.15$) (Fig. 2.7).

Two principal conclusions emerge from this analysis: First, neither 'buffering' nor 'antagonistic' interactions between regulatory motifs is a pervasive functional feature of *Ciona* muscle *cis*-regulatory elements. Second, the constituent motifs of an element

appear to function with a range of interactive effects. Such interactions appear small enough to model *cis*-element function, to a first approximation, with models that assume genetic independence of individual regulatory motifs. Thus, while all *cis*-elements of this study are built from clusters of regulatory motifs, such clustering is apparently not a requirement imposed by genetic interactions between the motifs themselves.

Statistical modeling of regulatory motif function

Based on the results of our interaction studies, I chose models without interaction terms, which have the added benefit of avoiding over parameterization. Inherent in our experimental design is the repeated testing of the functionality of individual motifs in multiple independent constructs. For each locus, I had between 6 and 30 distinct constructs for which expression frequency was measured, and which had particular combinations of 2-6 motifs present in wild type form, or either deleted or mutagenized. Because of this redundancy, the functional contribution of each motif could be estimated more accurately than with single data points. It should be noted that for each parameter added to the models (e.g., interaction terms) this redundancy is decreased and the risk of model over fitting is increased.

For the regression analyses, every tested motif becomes a categorical explanatory variable that contributes some frequency of muscle cell expression, with the wild type motif encoded as presence of the variable, and the mutagenized or deleted motif

encoded as its absence. The regression then provides estimates of each motif's activity by producing the best fit of the data to the model.

I explored four different modeling scenarios, whose results are summarized in Fig 1.8, Fig. 2.9, and Table 1.1.

Additive model:

$$y = a + b_1x_1 + b_2x_2 + ... + b_ix_i \quad (1)$$

Angular transformation, additive model:

$$\arcsin\left(\sqrt{y}\right) = a + b_1x_1 + b_2x_2 + ... + b_ix_i \qquad (2)$$

Multiplicative model:

$$(1 - y) = (1 - a)(1 - b_1x_1)(1 - b_2x_2)...(1 - b_ix_i) \quad (3)$$

which were log transformed and solved as linear models.

Logistic model:

$$y = \frac{e^{a + b_1x_1 + b_2x_2 + ... + b_ix_i}}{1 + e^{a + b_1x_1 + b_2x_2 + ... + b_ix_i}} \qquad (4)$$

Logistic models (eq. 4) were attractive for two principal reasons: proper treatment of bound frequency distributions and use of binomial error functions. As a result, logistic models predict the activity of minimally sufficient clones well (Fig. 2.9).

However, direct estimation of individual motif activity with logistic regression models is difficult to interpret biologically. Multiplicative models (eq. 3) also seemed a reasonable choice given the genetic independence of the data under a multiplicative estimate of epistasis (Fig. 2.7). Such models, after logarithmic transformation, could be solved by simple linear regression. However, multiplicative models consistently explained less expression variation than additive models (Fig. 2.9, Fig. 2.10, and Table 1.1). Models built from angular transformed expression frequencies (eq. 2) were appealing because they removed some of the dependence of expression variance on the mean (Fig. 2.4) and explained slightly more of the experimental variance than non-transformed additive models(Fig. 2.10). In practice, all four model types performed quite well (Fig. 2.9) and I therefore chose to focus on the simplest additive model (eq. 1) due to its methodological transparency and the inherent interpretability of its measurement (muscle cell expression frequency).

Therefore the majority of the data presented in this text (except where specifically noted) are derived from non-transformed additive multivariate linear regression models.

Quantification of individual motif function

The distribution of regulatory function produced by the regression models definitively shows, as qualitatively suggested by the initial large deletion series, that the majority of regulatory function rests in compact elements (min. 25 bp, max. 151

bp, mean 49bp). In fact, across all 19 loci, the mean fraction of function attributed to specifically interrogated motifs is 82%.

The residual function not accounted for by specific regulatory motifs most likely represents additional regulatory motifs that I failed to identify and characterize. Therefore, only 18% of the total *cis*-regulatory function I detected is not attributed to one of the three types of specific muscle regulatory motifs. I have no evidence for any additional regulatory motif types with a quantifiable effect despite extensive experimental testing of these loci, though I cannot formally rule out their existence.

Across 19 *cis*-regulatory elements, I quantified the function of 77 putative regulatory motifs, which effect the probability of muscle cell expression to varying degrees, from –0.14 to 0.45 efu (Fig. 2.8 B). One motif appears to have a repressive function (-0.14 efu), and three motifs have a slightly negative value that is statistically indistinguishable from 0. 39 motifs have significantly non-zero activating function. Importantly, this analysis suggests the existence of numerous false positive motif predictions (i.e., motifs whose activity is indistinguishable from zero) that, by definition, would not be identified based on primary sequence information alone.

Model Accuracy and Robustness

As discussed above, the data from each of the 19 loci was modeled independently under four different modeling scenarios. Comparison of explanatory variable coefficient estimates across models demonstrates surprising robustness to model assumptions. As graphically illustrated in Fig. 2.8A and numerically in Table 1.1, all

pairwise comparisons of model coefficients are significantly correlated (Spearman's rho 0.77-0.97).

Model robustness and overall accuracy is even better demonstrated by comparisons between the observed reporter activity of each construct and the activity predicted by each of the model types. As seen in Fig. 2.9 and Table 1.1, each of the four model types produce predicted construct activities that are highly correlated with the observed data (Spearman's rho 0.88-0.95). While all four models perform quite well, the additive and logistic models produce the best fit to the data, and the multiplicative model consistently performs the worst.

Careful examination of Fig. 2.10A reveals several important aspects of model fit to and deviation from the observed data. First, both additive models, as well as the multiplicative model, often generate predictions of negative expression measurements for constructs of weak observed activity. This results from the differing assumptions of the underlying modeling processes. All of the linear models assume normally distributed, unbounded data and are therefore able to produce such predictions. In contrast, logistic regression models (and our experimental scoring metric) assume a frequency distribution bounded at zero and one. Whether this deviation represents a biologically significant result (e.g., a net repressive effect of some of the constructs with zero measurable activity) or a modeling flaw could be experimentally addressed. Second, several loci are modeled relatively poorly by all model types. Close inspection of the underlying data at these loci suggests that the *cis*-elements might be better modeled by accounting for inter-motif interactions.

The quality of the regression models can be assessed by the fraction of experimental variance they explain (Fig. 2.10 A). As mentioned briefly above, the average coefficient of multiple determination (R^2) across all 19 models is 0.82. These R^2 values also reinforce the trends observed from Fig. 2.9, and demonstrate the differences in data fit across loci.

Examination of each of the models and their underlying constructs allows us to estimate the fraction of regulatory activity that the models attribute to regulatory motifs, as opposed to function attributed to larger stretches of DNA that has not been resolved (Fig. 2.10 B). The combination of constructs, expression measurements, and statistical models typically assigns 70-80% of the total regulatory activity of the locus to individual motifs that have been experimentally resolved. Troponin T exists as a significant outlier in this regard, as less than 30% of its *cis*-regulatory activity has been assigned to specific motifs. In this instance, the data suggest the existence of an additional region of compact regulatory activity, approximately 500 nucleotides upstream of the well-dissected region.

Lastly, I was able to assess the distribution of the residual error from each model (Fig. 2.11). Importantly, model residuals appear to be distributed fairly normally and without a dependence on the observed activity of a clone.

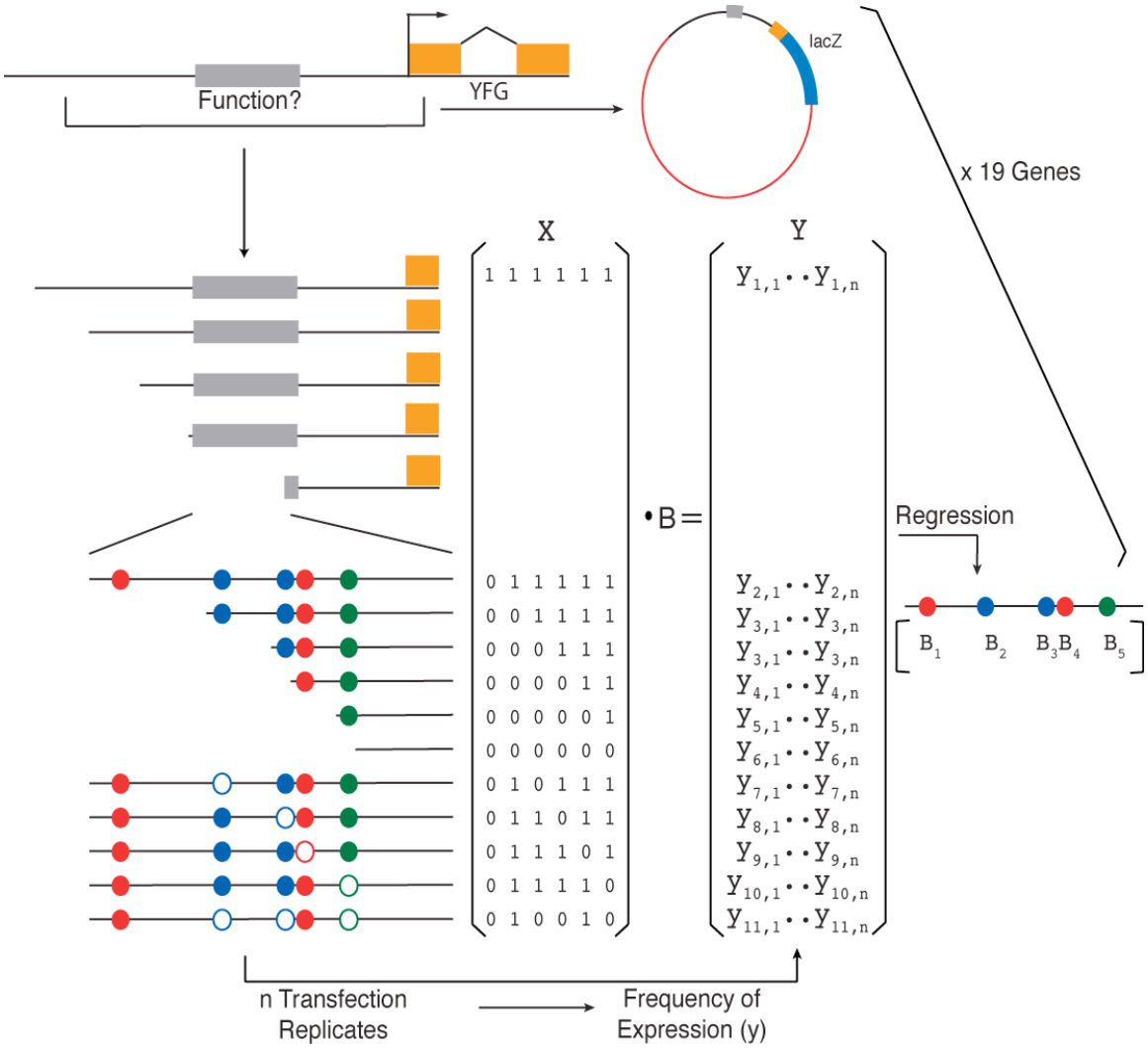

Figure 2.1. Quantitative dissection of *cis*-regulatory architecture.

Initial reporter constructs (top) were built by fusing 2-5kb of sequence (brackets under locus schematic) immediately 5' of the first identifiable exon (orange boxes) in frame with the lacZ reporter (top right). Constructs with regulatory activity were initially dissected with deletion series (truncated lines) that located regions of concentrated function (grey bar) . Fine-scale deletions and site-directed mutations (open circles) targeted putative motifs (filled circles). Constructs are represented as matrices ('X') of categorical explanatory variables (1s and 0s) whose replicated transfections yield expression frequencies (matrix Y of $y_{n,i}$). Functional contributions of individual motifs (B_1-B_5) are estimated with regression models.

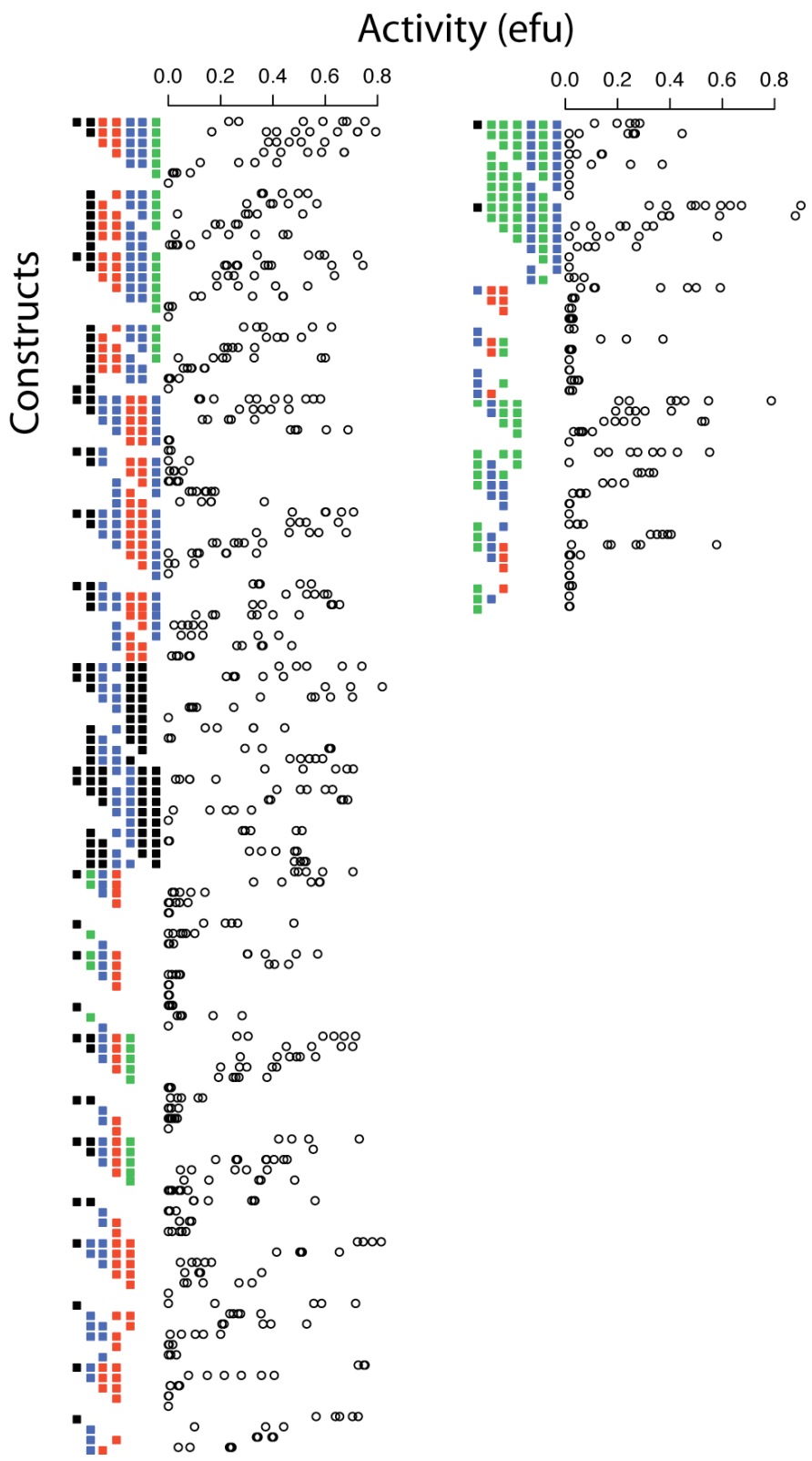
Activity (efu)
0.0
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
Constructs

Figure 2.2. Summary of all quantitative analyses.

The data are presented as two panels, left representing all constructs generated from single-copy orthologous loci, right representing those generated from *Ciona savignyi* specific paralogous loci. Within each panel, two aspects of the data are displayed. At left, each construct is schematized as a series of colored boxes, which represent its particular combination of regulatory motifs (MyoD in green, Tbx6 in blue, and CRE in red)or uncharacterized nucleotide segments (black). Constructs are sorted along the vertical axis by locus, as in Appendix 1. At right, the expression measurement (horizontal axis, in efu) resulting from each transfection of each construct is depicted as an individual circle. Replicate transfections of the same construct are drawn along the same line.

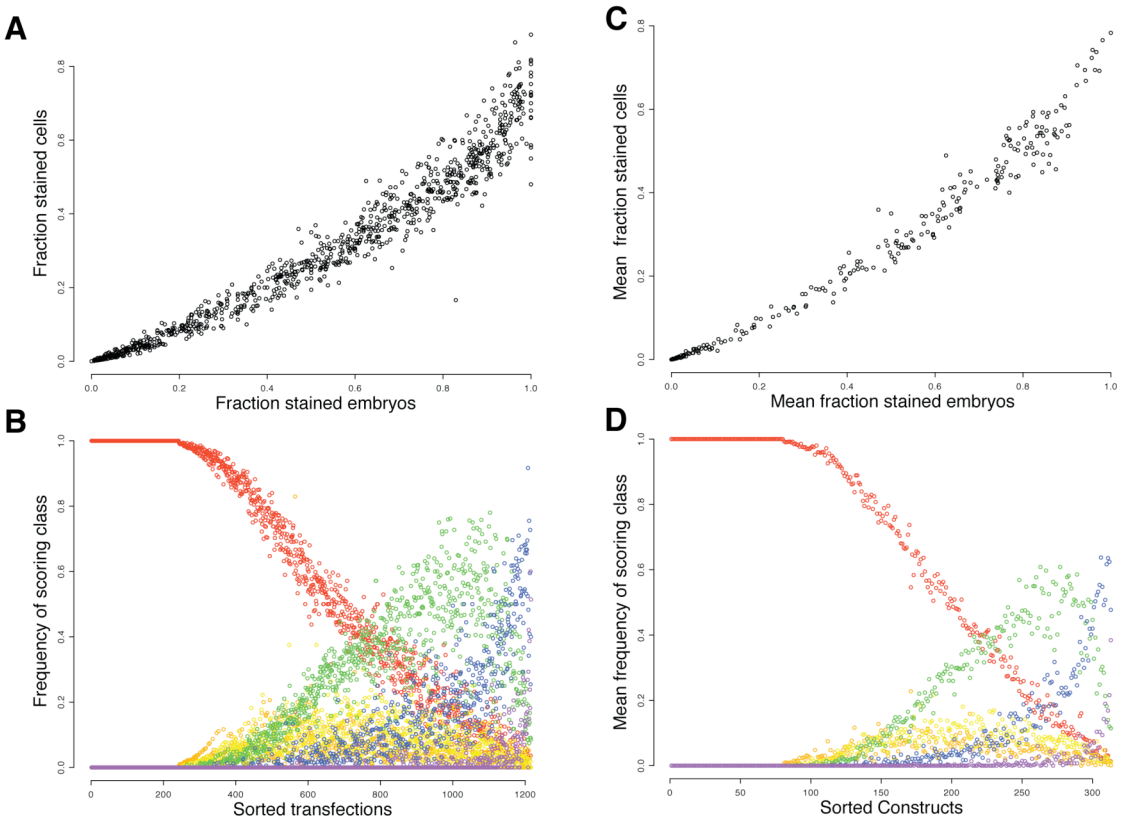

Figure 2.3. Expression Frequency Unit overview.

(A) Comparison of two possible scoring metrics. Each point depicts the expression level of an individual transfection, measured in two scales: percentage of stained embryos (x-axis) and the percentage of stained muscle cells (y-axis). **(B)** Shifting distributions of expression frequencies. Individual transfections (sorted along the x-axis by the mean percentage of stained embryos) are depicted as a set of six points. Each point represents the fraction of embryos from a given transfection in each scoring class: 0 (red), 1 (orange), 2 (yellow), 3 (green), 4 (indigo), 5 (violet). **(C-D)** Depicted as in (A-B), but representing the mean values of replicated transfections for each construct.

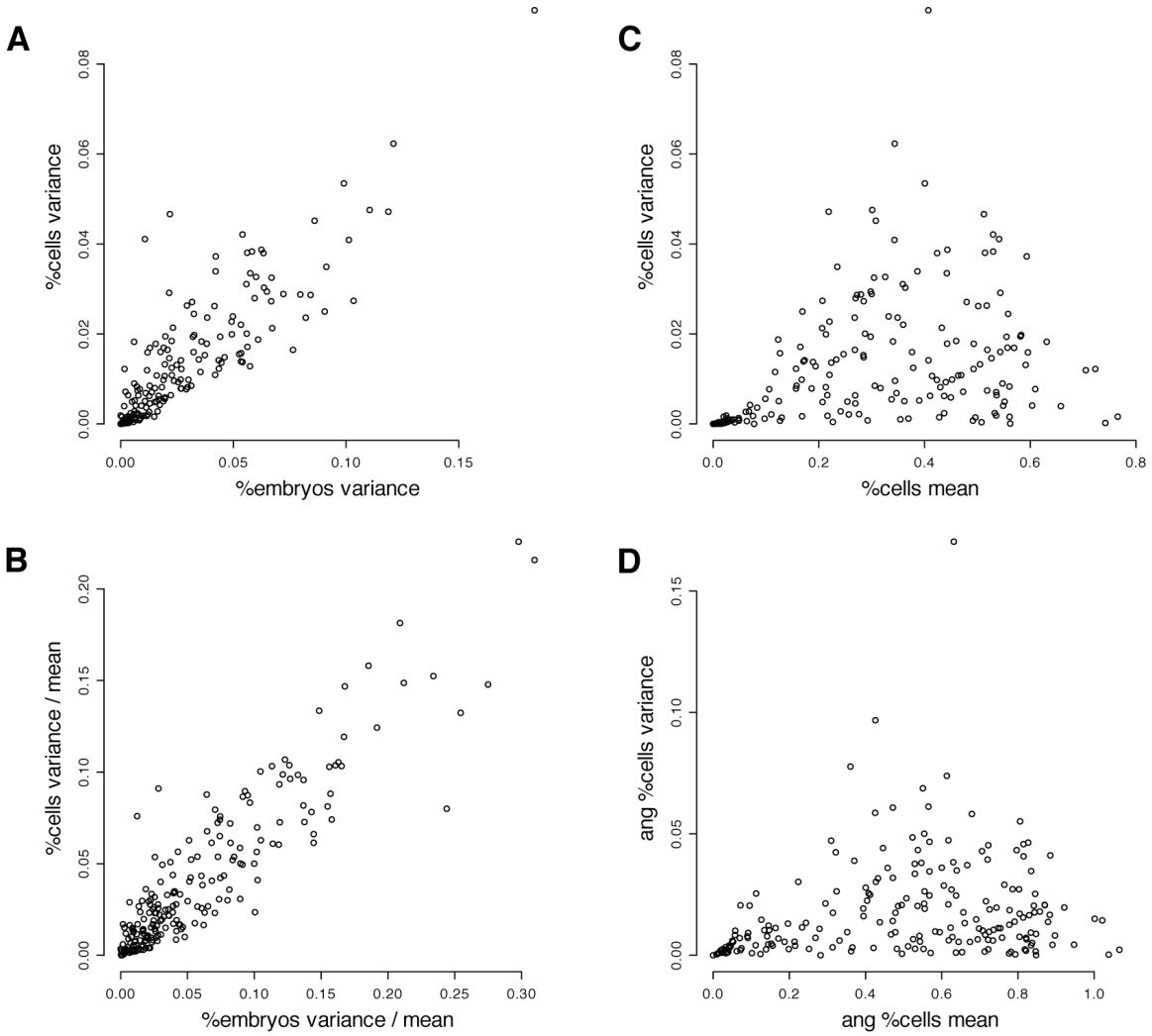

Figure 2.4. Distribution of experimental variance.

(A-B) Decreased variance associated with the efu scoring metric. (A) Each point represents the variance under two scoring metrics for replicate transfections for each construct: percentage of embryos staining (x-axis) and percentage of muscle cells staining (y-axis). (B) Similar to (A), each point depicts the variance as a fraction of the mean for replicate transfections of each construct. **(C-D)** Variance as a function of the mean. (C) Each point depicts two summary statistics for replicate transfections for each construct: mean percentage muscle cells staining (x-axis) and variance of this measurement (y-axis). (D) Layout as in (C), but statistics were calculated after subjecting the raw data to the angular transformation.

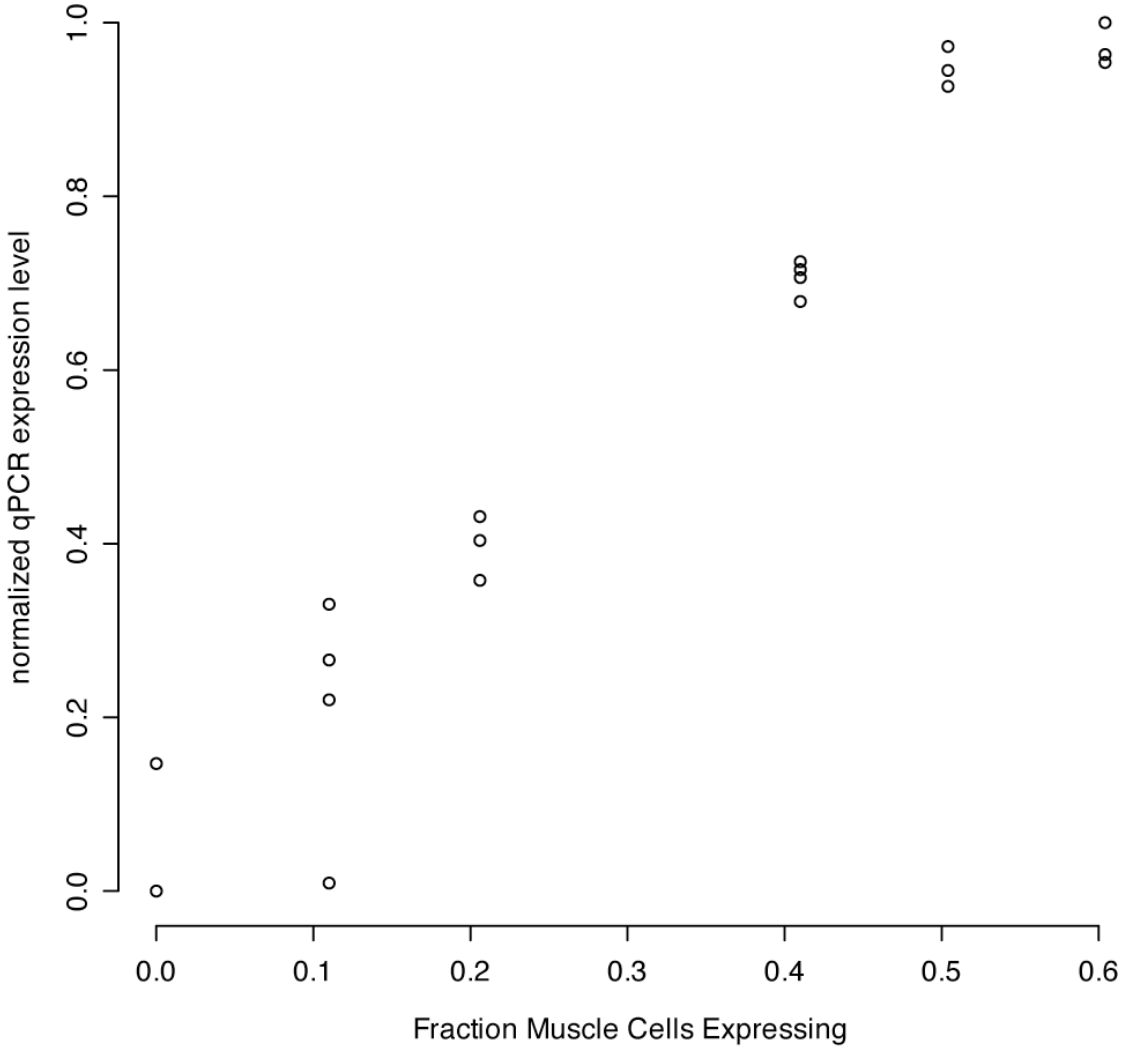

Figure 2.5. Quantitative PCR validation.

The expression level driven by six constructs of varying activity were assayed with two scoring metrics: the fraction of muscle cells stained for the reporter (x-axis) and transcript levels, as determined by quantitative RT-PCR for lacZ RNA (y-axis).

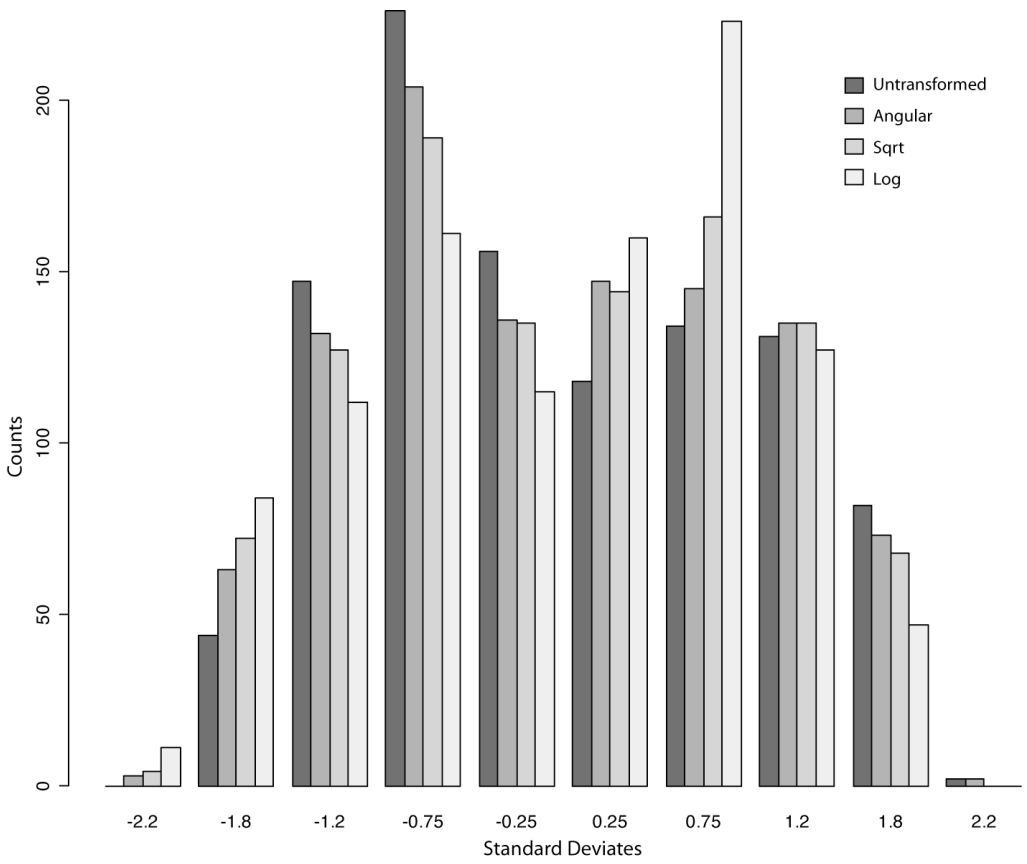

Figure 2.6. Distributions of standard deviates.

Data presented as a histogram, standard deviate bins plotted along the x-axis and counts per bin plotted along the y-axis. Bars of different colors (from dark to light grey) represent untransformed, angular, square root, and log transformed data, respectively.

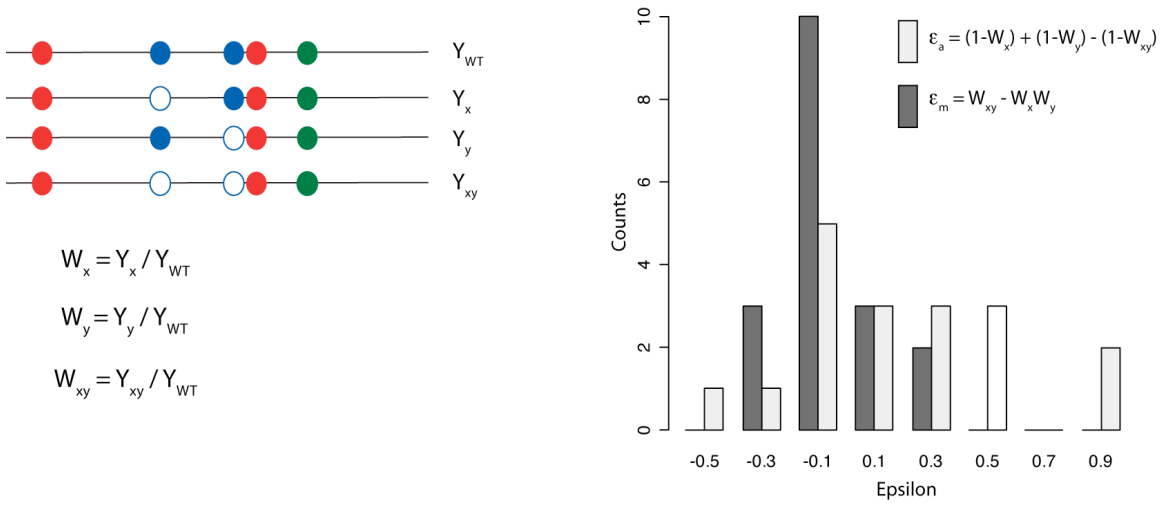

Figure 2.7. Genetic interactions between regulatory motifs.

(A) Sample of constructs required for the assessment of genetic interactions between two regulatory motifs. Each line represents an individual construct. The top line depicts the 'wild type' construct, with 5 regulatory motifs (colored circles). Lines 2 and 3 depict constructs containing site-directed mutations in single regulatory motifs (open circles). Line 4 depicts a double mutant construct. Below the panel of constructs are drawn the equations used for the estimation of the phenotypic effect of single and double mutants. **(B)** Histogram of genetic interactions for each of 18 possible comparisons. Binned ε values plotted along the x-axis and counts per bin along the y-axis. For each set of constructs ε was calculated under a multiplicative (grey bars) and additive (white bars) models of epistasis (relevant equations at top right).

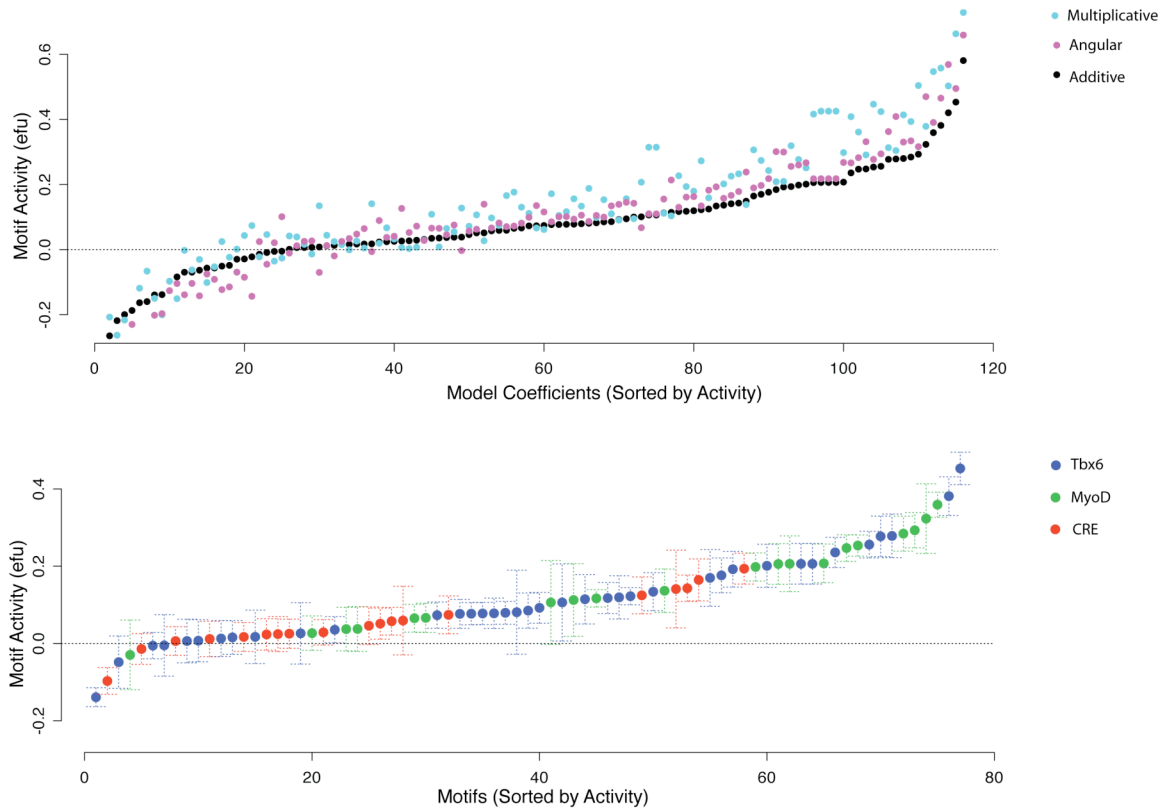

Figure 2.8. Estimation of motif activity.

(A) Robust estimation of model coefficients. Three regression models were built for all 19 genes: additive model (black), angular transformed additive model (magenta), and multiplicative model (cyan). Each explanatory variable is represented as a trio of points, one for the coefficient estimated from each model. Coefficients for each variable were sorted along the x-axis by the activity estimated using the additive model and each coefficient plotted along the y-axis. **(B)** *Cis*-regulatory function of 77 individually resolved motifs. Activity and standard error plotted on y-axis, motifs sorted along x-axis by estimated activity. Color indicates motif type: red, CRE; green, MyoD; blue, Tbx6. Note that panel (A) depicts all model coefficients, while panel (B) only plots variables that represent putative regulatory motifs of 6-10 bases, as opposed to larger stretches of DNA, that are included in panel (A).

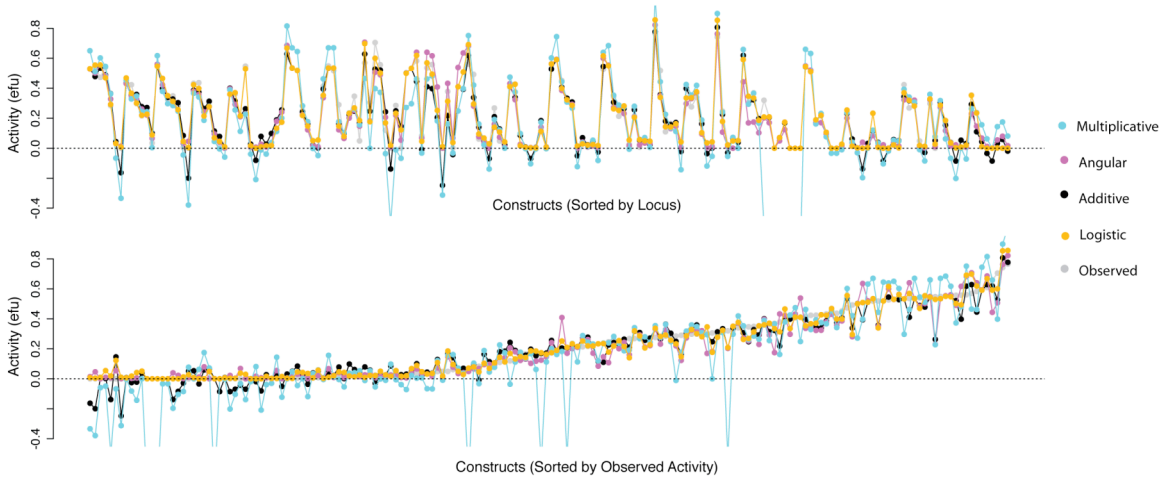

Figure 2.9. Observed and predicted construct activities.

(A) Construct-by-construct comparison, for each gene, of mean observed activity (grey) or activity predicted by four different regression models: additive (black), angular transformed/additive (magenta), multiplicative (cyan), or logistic (orange). Constructs sorted (along x-axis) by gene as in Table S2. Mean expression measurements or estimates in expression frequency units plotted along y-axis. **(B)** Data as in (A), Constructs are sorted along x-axis by measured activity.

Table 1

Non-parametric Spearman's rho correlation coefficients comparing the results of different modeling scenarios on predicted or observed construct activity (left) or explanatory variable coefficients

Constructs					
observed	additive	multiplicative	angular	logistic	
1.00	0.95	0.88	0.94	0.95	observed
	1.00	0.92	0.95	0.95	additive
		1.00	0.89	0.87	multiplicative
			1.00	0.95	angular
				1.00	logistic

Coefficients				
additive	multiplicative	angular	logistic	
1.00	0.97	0.95	0.82	additive
	1.00	0.88	0.77	multiplicative
		1.00	0.84	angular
			1.00	logistic

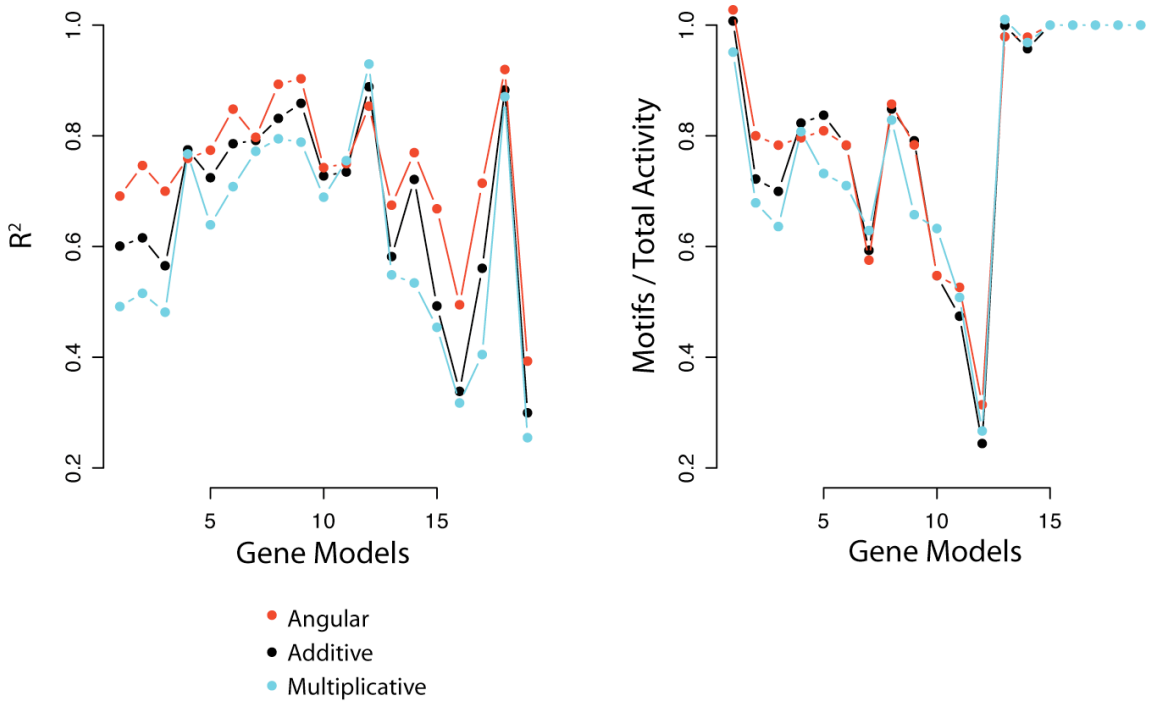

Figure 2.10. Performance of each regression model type.

(A) Variance explained by models. Each point depicts the coefficient of multiple determination (R^2) for a combination of regression model type and gene. R^2s for each gene plotted along the y-axis, sorted by gene along the x-axis. Color depicts model type: additive (black), angular transformed additive (red), multiplicative (cyan). **(B)** Fraction of activity attributed to compact motifs plotted along y-axis, sorted by gene along the x-axis (order as in (A)). Colors as in (A).

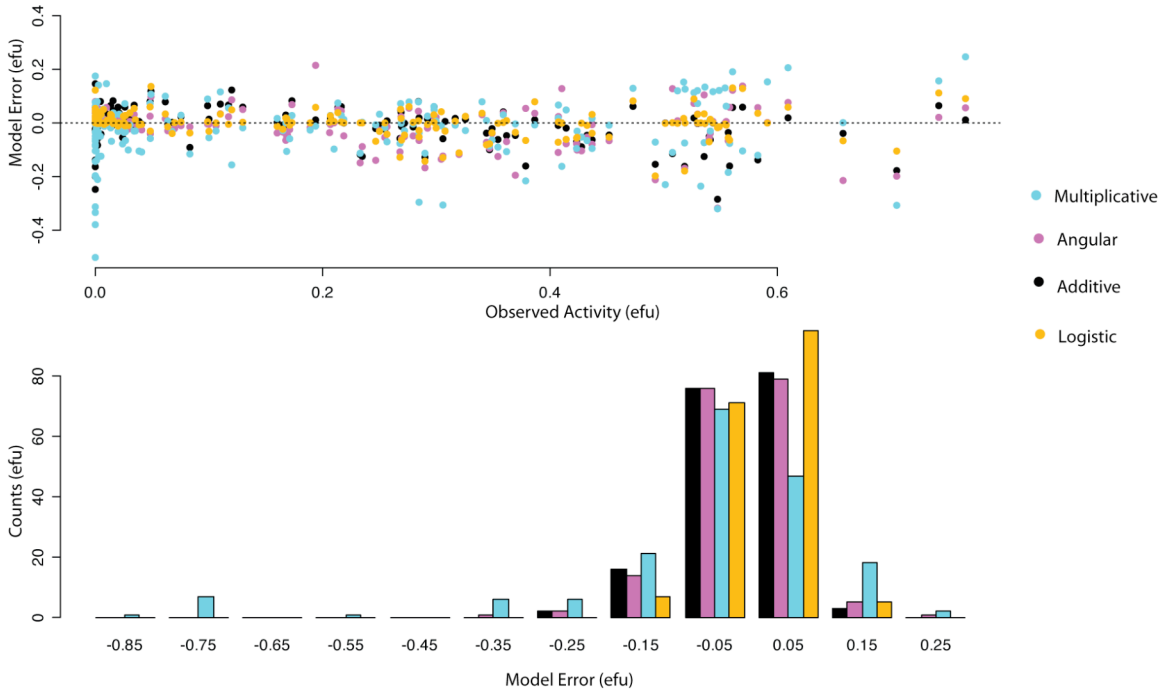

Figure 2.11. Distribution of regression residuals.

(A) Model error as a function of observed construct activity. Each construct is depicted as a set of four points, each point representing the difference between the observed construct activity and the activity predicted by a particular model. Data are plotted by observed construct activity (x-axis) and model error (y-axis). Each model type is symbolized by a different color: additive (black), angular transformed additive (magenta), multiplicative (cyan), and logistic (orange). **(B)** Histogram of model residuals. Model error binned along the x-axis and counts per bin plotted along the y-axis. Colors as in (A).

Chapter 3 Functional architecture and evolution of *cis*-regulatory elements that drive gene coexpression

Diverse regulatory architectures produce similar phenotypic outputs

Having obtained quantitative estimates of motif activity, I examined each *cis*-element's functional architecture. Our data show that expression of each gene is controlled by subsets of three regulatory motif types. These three motif types, Tbx6, MyoD, and CRE, and only these three motif types, are responsible for all the well-resolved *cis*-regulatory activity defined in this study. The existence of a tailbud embryo "muscle module" that confers coregulation, which was initially characterized by the statistical enrichment of motif sequences in muscle-specific promoters (Johnson et al. 2005) is therefore refined and corroborated.

There is striking heterogeneity among the loci: Elements are built from motifs of widely varying activity, from different combinations of motif types, and in diverse arrangements (Fig. 3.12). For example, the *cis*-element at CK spans 31 bp and consists of one intermediate and one strong Tbx6 motif, while the AT1 *cis*-element consists of two weak CRE motifs, followed by two intermediate Tbx6 motifs and a strong MyoD motif, across 35bp. Though motif independence is prevalent, elements do somewhat differ in how much genetic interaction exists. At MBP and AT2, for example, the additive model explains the data very well with high correlations between the predictions and the actual data ($R^2_{Cs\text{-}MBP} = 0.83$, $R^2_{Cs\text{-}AT2} = 0.77$) and with little function unexplained by the model. Function at MA1, by contrast, is not described as well by models without interactions (Fig 1.9 and Fig. 2.10). Thus,

although *Ciona* muscle genes are regulated by a common and restricted set of three transcription factors, their tight coexpression is achieved despite vastly different *cis*-element architectures.

Visual representation strikingly underscores this finding (Fig. 3.12), which presents an interesting contrast: As a co-expressed gene set, *Ciona* muscle genes are coregulated by a common and restricted set of transcription factors, as defined by the co-occurrence of their putative binding sites across genes and by their statistical enrichment in the collection of regulatory elements (Johnson et al. 2005). However, at the resolution of individual *cis*-regulatory elements, the tight coexpression seen for this set of genes is achieved despite vastly different functional motif architectures.

Functional equivalence of motif types

I performed substitution experiments at three loci to test, independently from the epistasis analyses, whether genetic interactions among motifs are important. I reasoned that substituting one motif for another if specific interactions are required would not result in rescue. I had to choose particular constructs from particular loci in which deletion of a single motif could result in complete loss of function, so in effect I selected loci that had the greatest chance of providing evidence for interactions.

Each motif type is capable of conferring spatially and temporally specific function when introduced into the appropriate sequence context. Muscle-specific gene expression is rescued when a different type of motif is inserted into a construct that had been rendered nonfunctional by site-directed mutagenesis or deletion of the

endogenous motif. Notably, every motif substitution I attempted resulted in some degree of rescue. At the *C. intestinalis* CK locus, scrambling of the Tbx6 motif at position -268 results in a significant decrease in expression probability. However, if the site is mutated to either its reverse complement, to a MyoD motif, or to a CRE motif, muscle specific expression is restored (Fig. 3.13 A). Similar results were also obtained with the Tbx6 motif at -108 in the *C. intestinalis* AT2 locus (Fig 2.2 B) and with the Tbx6 motif at -89 in the C. savignyi AT1 locus. This demonstrates that the three motif types are at least partially functionally equivalent and that each motif transmits similar regulatory information to the transcriptional machinery. The interpretation that this reflects a significant degree of functional equivalence, in combination with the lack of prevalent genetic interactions provides a possible mechanistic explanation for the diversity of *cis*-regulatory architectures that control identical gene expression patterns at each locus examined.

Conservation of orthologous motif function

The resolution of our experimental data afforded the opportunity to quantify the functional *cis*-regulatory changes that occurred since the last common ancestor of the two *Ciona* species. In stark contrast to the apparent flexibility of regulatory architecture, there exists little change in motif activity, order, or composition between orthologous elements of *C. intestinalis* and *C. savignyi*. (I note that the *Ciona* species are as genomically divergent as mammals and birds, ruling out the possibility that these sequences have not been afforded enough time to accumulate change.)

At single copy genes, 26 of the 27 motifs with statistically significant activity have a clearly orthologous counterpart. Orthologous motifs drive very similar, in many cases indistinguishable, amounts of activity (Fig. 3.14 A). For example, both MBP orthologs are regulated by a strong MyoD, a weak Tbx6, and a weak CRE motif, with less than 0.039 efu average deviation in individual motif activity. In total, the activity of orthologous regulatory motif pairs is highly correlated between the two species (Spearman's $\rho = 0.61$, $p < 0.005$; Fig. 3.14 B). To our knowledge, this represents the first time the *function* of individual *cis*-regulatory motifs has been quantified and compared with orthologous counterparts. In combination with the low rate of motif turnover, this relationship indicates that the conserved patterns of orthologous gene expression in this study have largely resulted from the maintenance of the ancestral *cis*-regulatory mechanism, as opposed to an evolutionary process that selects for a specific phenotypic output while allowing functional flexibility.

Sequence conservation of functional motif sequences

Strong constraint is evident at the sequence level as well. The high functional resolution of our data afforded the opportunity to quantify the evolutionary processes affecting *cis*-regulatory motifs at base pair level resolution. Functional regulatory motifs exhibit far fewer substitutions than the genome-wide average and therefore, as a group, appear to be evolving under a high level of purifying selection (Fig. 3.15 A; $p < 3.8 \times 10^{-10}$, Wilcoxon Rank Sum Test; see Methods). The pair-wise identity between

C. savignyi / *C. intestinalis* orthologous functional motifs is 79%, whereas the genome-wide background identity is <20% (including indels).

Pair-wise identity between orthologous sequences dramatically drops off outside the boundaries of the functional motifs, decreasing as a function of the distance from motif boundaries (Spearman's $\rho = -0.57$, $p < 0.05$) and nearly reaching genome-wide background levels within 12 bp (Fig. 3.15 B). This demonstrates that the motif sequences are subject to a far greater degree of evolutionary constraint than flanking sequence and that the functional motifs themselves are the units maintained by purifying selection. However, I do note that there is some evolutionary constraint immediately surrounding the motifs, which may be indicative of structure or function that was not assayed in our experiments.

Due to the tight selective footprint and the presence of non-functional motif-like sequences, evolutionary analyses relying solely on motif predictions, as opposed to functionally defined motifs, lead to biases in the direction of overestimating variability. To cement this point, I mimicked lower-resolution data and examined 500 bp regions encompassing our functional motifs. Within these regions I assessed the *C. savignyi*/*C. intestinalis* pair-wise percent identity of high confidence motif predictions in the *C. savignyi* sequence. To generate a conservation distribution of motif predictions, I collected local alignment windows from predicted motif positions within a region of 500 nucleotides encompassing the functional module. To minimize false positive motif predictions, I only assessed predictions with LOD scores > 4.45, representing the 25th percentile of true positive motifs (see Methods).

The distribution of resulting pair-wise identity values is shifted significantly downward relative to the distribution built from functional motifs (Fig. 3.15 A; Wilcoxon Rank Sum Test, $p < 0.05$). Therefore, the motif prediction set contains motifs that are either: (a) False-positives or sequences that, based on primary sequence alone, are indistinguishable from functional motifs, but do not contribute any regulatory activity to the locus, or (b) Extremely weak motifs whose function is below the detectable threshold of the transfection assay.

In either case, such sequences accumulate substitutions at higher rates than significantly functional motifs; as a result, the inclusion of such sequences will downwardly bias estimates of sequence constraint. This result is of particular interest because *cis*-regulatory data of similar moderate resolution is increasingly common as a result of modern high throughput experimental platforms, such as ChIP-chip or genome-wide DNAseI hypersensitivity data.

Functional regulatory motifs also carry low levels of polymorphism compared to the rest of the genome. Across 13 significantly functional *C. savignyi* motifs at single copy genes, only 2 out of a total of 115 nucleotides were heterozygous in the sequenced *C. savignyi* genome, compared to the genome-wide average neutral heterozygosity of >8% (Small et al. 2007a). This is unlikely to result from stochastic fluctuations in diversity as fewer than 4% of a sample of ~7,500 mock motif sets from across the *C. savignyi* genome display this little polymorphism (Fig. 3.16). Therefore, not only has selection removed *cis*-regulatory motif substitutions over long evolutionary timescales, but it also appears to be acting on extant variation by removing deleterious polymorphism.

Sequence specificity of functional motifs

I also combined all instances of functional motifs of the same type, including their surrounding sequence context, into PSSMs. These functional, *in vivo* sequence specificities were compared to our initial PSSMs, which were built from either statistical over representation in a promoter set or *in vitro* binding data (see Methods). This analysis revealed that the functional *cis*-regulatory motifs identified in this study differ little in nucleotide specificity and lack any significant sequence specificity beyond the original borders of the motifs, as is visually demonstrated by a translation of PSSMs into sequence logos (Fig. 3.17). Thus, neither evolutionary nor informational sequence analysis reveals any sequence-specific information to be contained beyond the edge of the motifs.

Motif activity is correlated with sequence constraint

The sequence changes that have occurred are not distributed evenly among the orthologous functional *cis*-regulatory motifs (Fig. 3.18). Functionally "strong" motifs (those that disproportionately increase the probability of a cell expressing the reporter) have accumulated fewer substitutions than weak motifs. In fact, motif activity is significantly correlated with percent identity in pair-wise *C. savignyi*-*C. intestinalis* alignments(Spearman's $\rho = 0.35$, $p < 0.05$). This is likely due to the fact that strong regulatory motifs are responsible for a larger fraction of the total function of a

regulatory element; substitutions in them will therefore result in greater phenotypic consequences and be subject to stronger levels of purifying selection, decreasing the evolutionary rate of the element. This result demonstrates that: (a) the broad gradient of motif activity estimates I have quantified are biologically meaningful and (b) individual motif activity is a major determinant of motif substitution and turnover.

Regulatory motif turnover

Comparisons of orthologous elements can also identify motifs of significantly differential but complementary function, which represent candidate sites of compensatory evolution. Two such motifs occur at the α-Tropomyosin 2 locus, Tbx6.-120 and Tbx6.-108 (Fig. 3.19 A). At these sites, each species has one functionally strong and one functionally weak motif in a pattern complementary to the other species. These functional differences correlate with substitutions away from or towards the motif consensus. Importantly, the functional differentiation of these *cis*-regulatory motifs is not the result of binding site gain or loss, but rather from fixation of substitutions that have modified their function.

In addition to the analyses of regulatory architecture at the orthologous elements, I also dissected the *cis*-regulatory mechanisms of three groups of isofunctional paralogs in *C. savignyi*. The protein-coding sequences of these genes are evolving under purifying selection (dN/dS << 1) and their expression patterns have remained identical. Based on the topology of third position nucleotide trees built from the coding sequences of these genes, the gene duplication events appear to post-date the

C. savignyi/C. intestinalis speciation event (Fig. 3.20). Alternatively, given that many of the paralogous genes are genomically clustered, gene conversion, or a cycle of repeated duplications and losses may be acting to homogenize paralogous loci.

Regardless of the precise evolutionary history of these loci, the paralogous *cis*-elements present a striking contrast to the highly constrained orthologous *cis*-elements. Their *cis*-regulatory architectures show a high degree of differentiation in the form of module and motif-level sequence turnover as well as functional divergence of well-aligned motifs (Fig. 3.12 and Fig. 3.19 B). For example, the elements regulating Myosin Light Chain 1 and 5 are unalignable and composed of different functional motifs, indicating that they are not homologous DNA. At Muscle Actin 1 and 2, four functional motifs are well conserved while two have accumulated enough substitutions to make assessment of homology difficult. Lastly, the Myosin Regulatory Light Chain (MRLC) 6 regulatory element is composed of functional motifs different from those of either MRLC4 or MRLC5, again suggesting rapid binding site turnover. Given the amount of sequence turnover, it is not surprising that the functions of paralogous regulatory motifs are not significantly correlated.

The combination of functional and sequence analyses provides clear evidence that the evolutionary dynamics of these orthologous and paralogous *cis*-regulatory elements, even for genes operating in the same macromolecular complex, are strikingly different. Thus, while I have shown that purifying selection acting on orthologous *cis*-regulatory motifs is strong enough to maintain conservation of regulatory motif sequence and function over vast evolutionary distances, paralogous *cis*-regulatory motifs exhibit far greater levels of motif turnover. I speculate that this

greater flexibility in elements of clustered multicopy genes is tolerated because changes in the activity of one element have a small effect on the total function of the cluster.

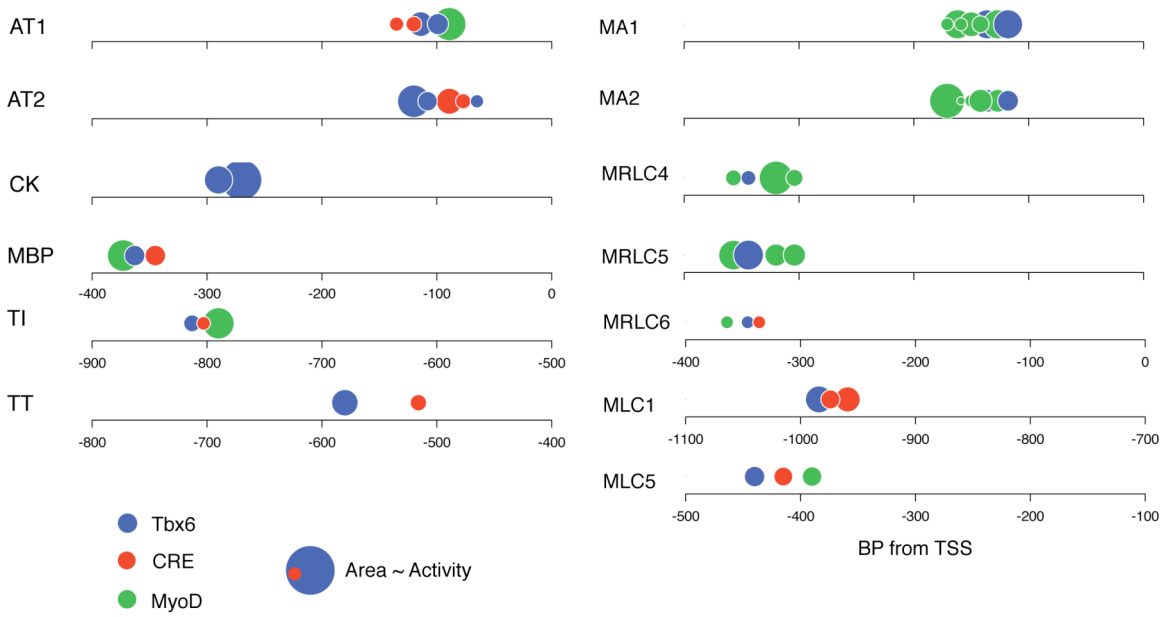

Figure 3.12. Differential *cis*-regulatory architecture.

Each panel depicts the individually resolved *cis*-regulatory motifs at individual *C. savignyi* genes. X-axis represents bases from the predicted transcription start site. Each point represents an individually resolved motif (colors as above), with circle area proportional to motif function. Grey circles represent unresolved *cis*-regulatory function. Also indicated are the expression probabilities of the strongest clone at each locus. Function values and standard error estimates are defined by regression.

Figure 3.13. Functional equivalence of motif types.

Motif substitutions at the *cis*-elements of *C. intestinalis* CK (A) and AT2 (B). Color indicates motif type, with area proportional to activity as in Fig. 2B: red, CRE; green, MyoD; blue, Tbx6. Each row is a construct, with the endogenous arrangement at top and mutants below. Open circle is a scrambled sequence, "R" the reverse complement of the Tbx6 site. Mean muscle cell expression frequency is at right.

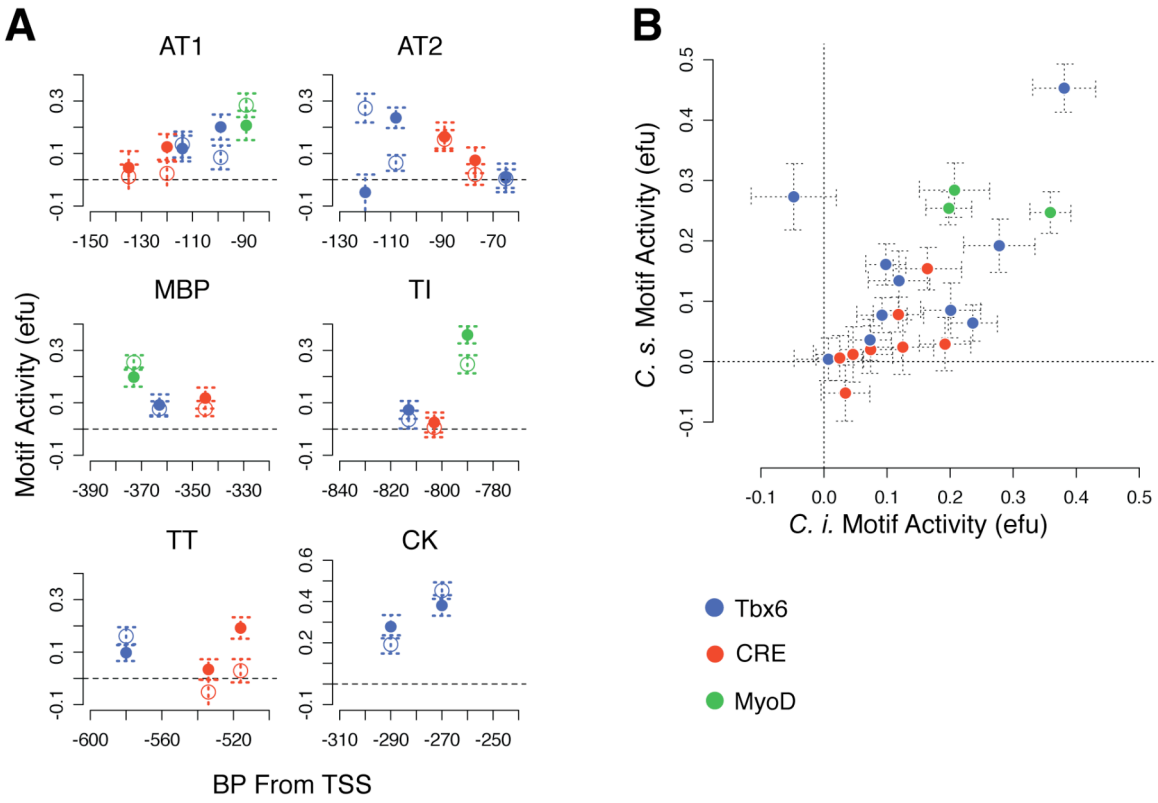

Figure 3.14. Conservation of orthologous motif activity.

(A) Motif-level distribution of regulatory activity at six orthologous gene pairs. Distance from transcription start site and motif activity are plotted along the x- and y-axes, respectively. Open and filled circles represent individually resolved *C. savignyi* and *C. intestinalis* motifs, respectively. Circle color depicts motif type: Tbx6 (blue), CRE (red), and MyoD (green). **(B)** Function of all individually resolved motifs at orthologous loci. X and Y-axes depict estimates of *C. intestinalis* and *C. savignyi* motif function, respectively. Colors as in (A). Motif function and standard errors as estimated from regression.

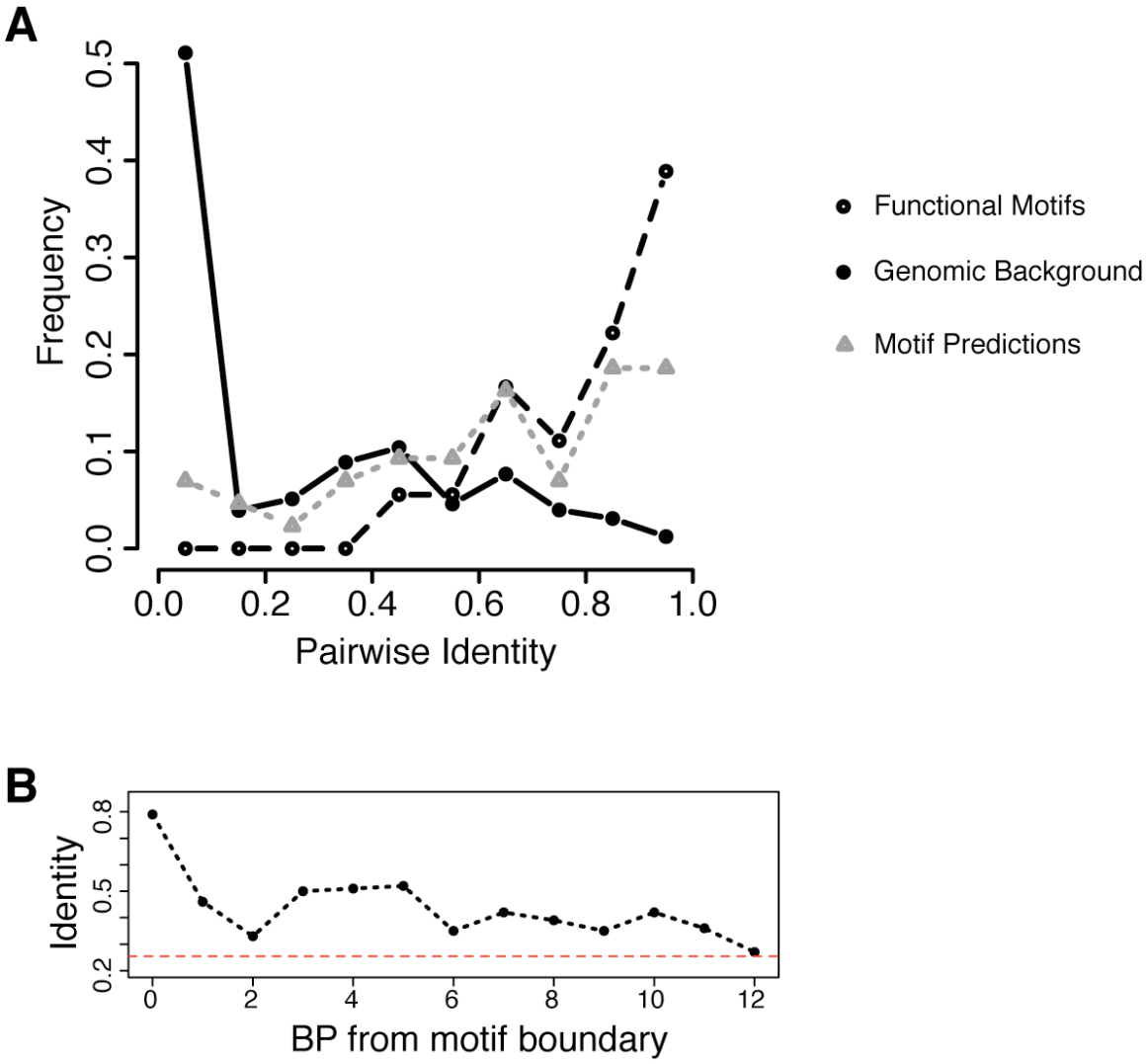

Figure 3.15. Sequence conservation at regulatory motifs.

(A) Histograms of the *C. savignyi*-*C. intestinalis* pairwise percent identity of samples from background genomic DNA (black, filed circles), motif predictions from a 500bp window spanning the *cis*-element (gray, open triangles), and the functional motif set (black, open circles). Note the dilution of conservation signal (high pairwise identity towards the right of the plot) when nonfunctional motifs are included in the analysis. **(B)** Mean pairwise percent identity of orthologous functional motifs at increasing distances from motif boundaries. Position 0 represents the within-motif mean. Red dashed line represents genome-wide mean.

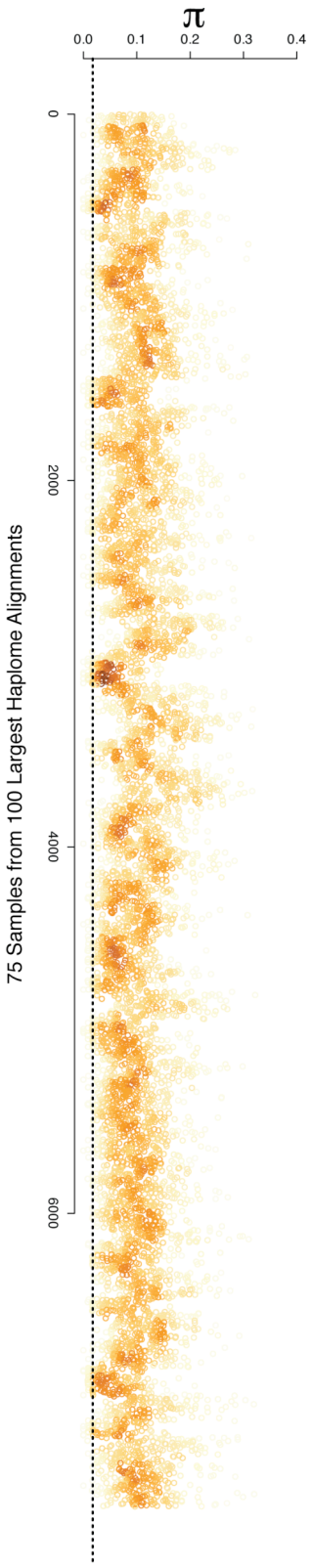
π
0.0
0.1
0.2
0.3
0.4
0
2000
4000
6000
75 Samples from 100 Largest Haplome Alignments

Figure 3.16. Reduced polymorphism in functional motifs.

Heterozygosity (x-axis) in 7500 samples (75 each from the 100 largest *C. savignyi* haplome alignments; 12), ordered by alignment (y-axis). Each circle represents a single sample of 13 mock motifs. Circles are shaded to highlight overlapping data, from highest (browns) to lowest (yellows) local point density (palette by colorbrewer.org). The heterozygosity within functional motifs is indicated by the dashed line.

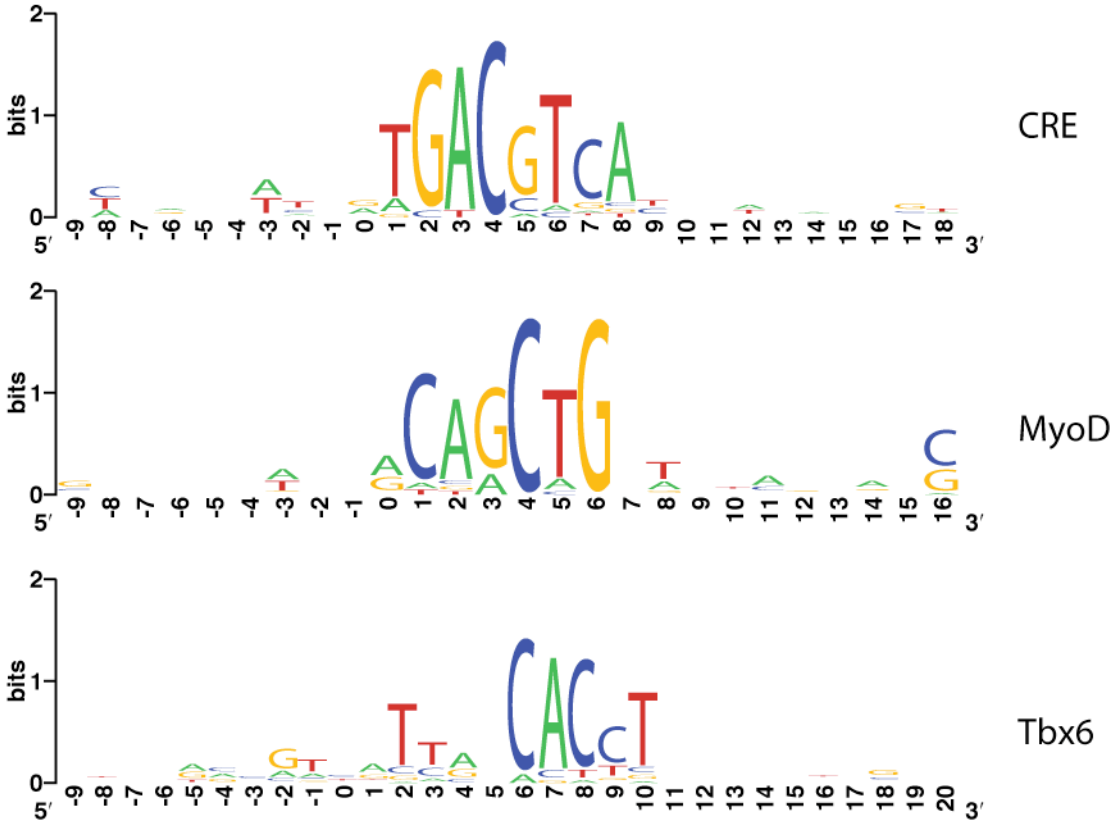

Figure 3.17. Motif specificity.

Sequence specificity of each motif type, represented as sequence logos derived from all functional motifs (plus 10 bases on either side), grouped according to motif type: **(A)** CRE, **(B)** MyoD, and **(C)** Tbx6. Note the lack of significant sequence specificity outside the originally defined boundaries of each motif.

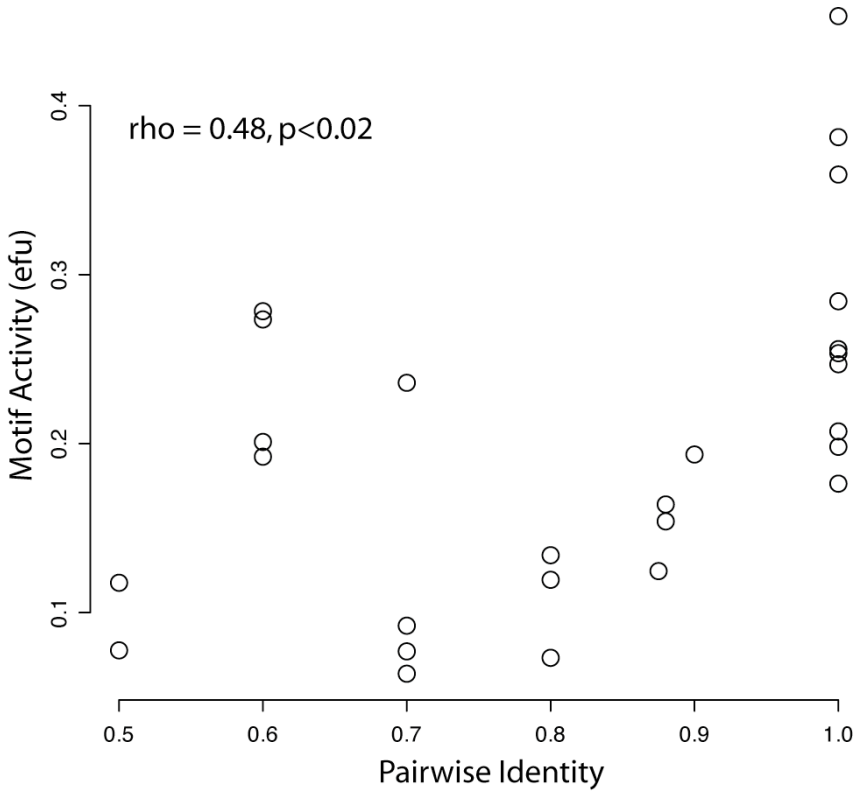

Figure 3.18. Increased sequence constraint at strong regulatory motifs.

All significantly functional regulatory motifs are plotted as single points. X-axis represents the pair-wise identity of the motif in a *C. savignyi – C. intestinalis* alignment. Y-axis represents the activity of the motif, as estimated by regression. As indicated at the top left, the data are significantly correlated by the non-parametric Spearman's rho correlation test.

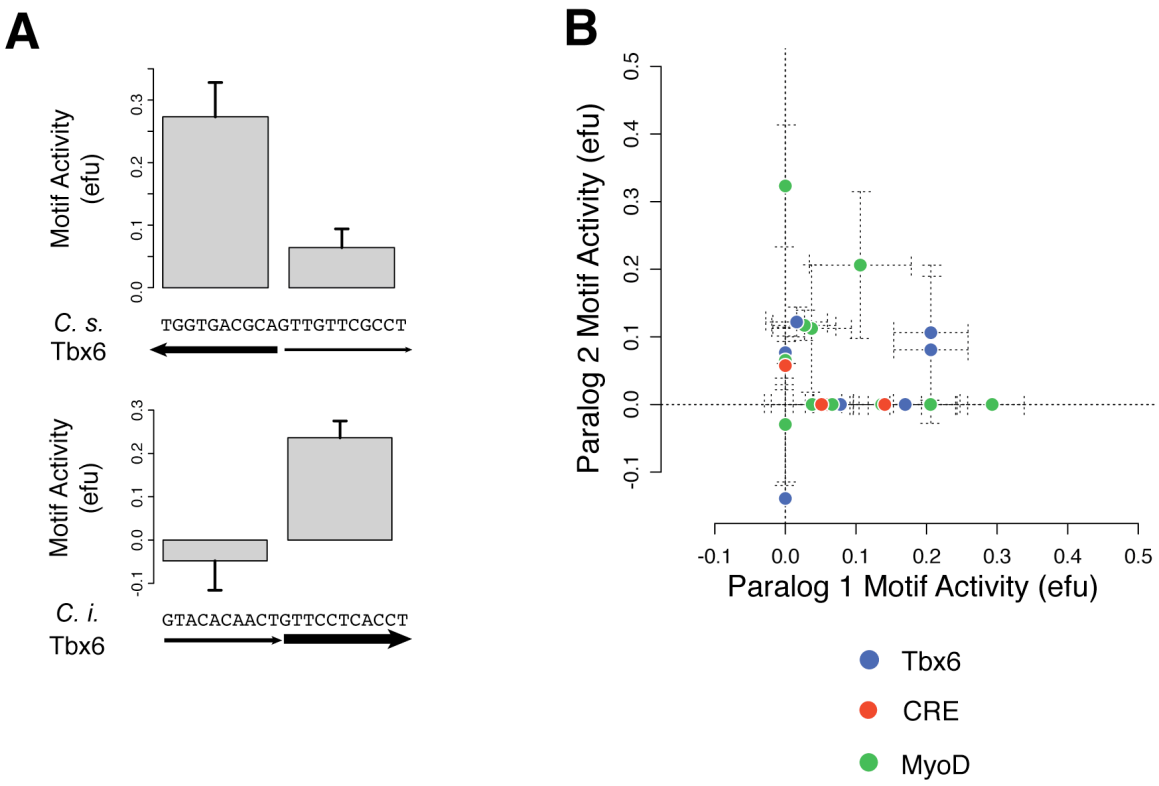

Figure 3.19 Examples of motif turnover.

(A) Compensatory evolution of AT2 regulatory elements. *C. savignyi* at top, *C. intestinalis* below. Arrow direction and thickness represent Tbx6 motif orientation and strength of match to PSSM. Bar plots depict activity of each Tbx6 motif, as estimated from additive regression models. **(B)** Functional turnover in paralogous motifs. Function of all individually resolved motifs at paralogous motifs is depicted on the X and Y-axes. Colors indicate motif type: Tbx6 (blue), CRE (red), and MyoD (green). Motif function and standard errors as estimated from regression.

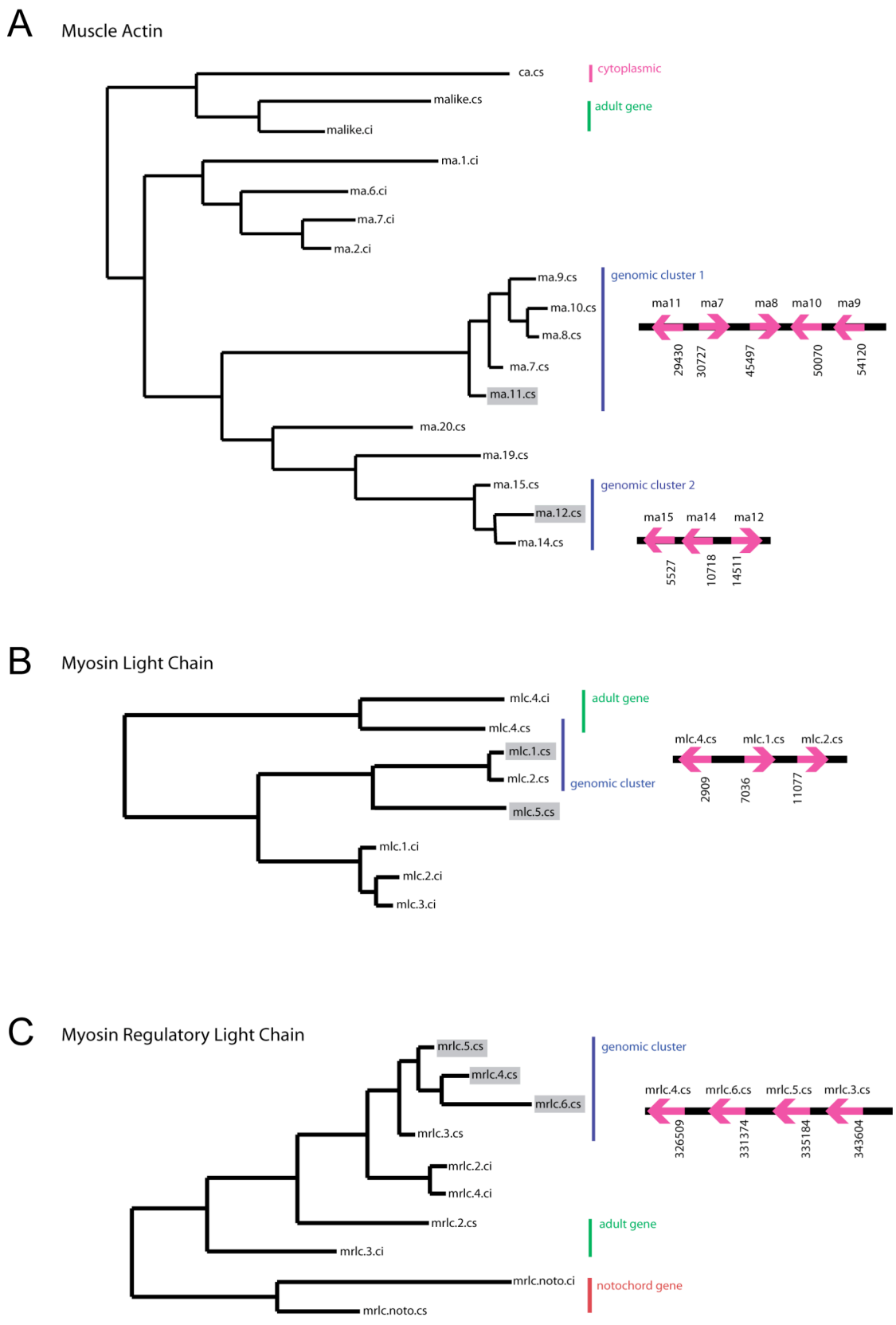
A
Muscle Actin
ca.cs
malike.cs
malike.ci
ma.1.ci
ma.6.ci
ma.7.ci
ma.2.ci
ma.9.cs
ma.10.cs
ma.8.cs
ma.7.cs
ma.11.cs
ma.20.cs
ma.19.cs
ma.15.cs
ma.12.cs
ma.14.cs
cytoplasmic
adult gene
genomic cluster 1
ma11 ma7 ma8 ma10 ma9
29430 30727 45497 50070 54120
genomic cluster 2
ma15 ma14 ma12
5527 10718 14511
B
Myosin Light Chain
mlc.4.ci
mlc.4.cs
mlc.1.cs
mlc.2.cs
mlc.5.cs
mlc.1.ci
mlc.2.ci
mlc.3.ci
adult gene
genomic cluster
mlc.4.cs mlc.1.cs mlc.2.cs
2909 7036 11077
C
Myosin Regulatory Light Chain
mrlc.5.cs
mrlc.4.cs
mrlc.6.cs
mrlc.3.cs
mrlc.2.ci
mrlc.4.ci
mrlc.2.cs
mrlc.3.ci
mrlc.noto.ci
mrlc.noto.cs
genomic cluster
mrlc.4.cs mrlc.6.cs mrlc.5.cs mrlc.3.cs
326509 331374 335184 343604
adult gene
notochord gene

Figure 3.20. Phylogenetic trees of multicopy muscle genes.

Evolutionary relationships of **(A)** Muscle Actin, **(B)** Myosin Light Chain, **(C)** Myosin Regulatory Light Chain paralogs. Trees built from third-position coding sequences by maximum likelihood (1). Functional analyses reported in this study were carried out on genes shaded grey. Several members of each multigene family are present in genomic clusters, schematically indicated on right. Clusters are shown for *C. savignyi* only, as the assembly of these loci in *C. intestinalis* is fragmented. Blue brackets represent genes occurring on the same supercontig. Green brackets represent adult-expressed genes, red brackets indicate notochord genes, and rose represents a cytoplasmic gene. Figure 3.20 is a modification of a figure by Dave Johnson, based on his work and analyses.

Chapter 4 Conclusions

My fine-scale characterization of the molecular architecture of coregulatory elements addresses several questions of basic importance for understanding mechanisms of gene regulation and their evolution: how individual motifs function, how such motifs form *cis*-elements, how *cis*-elements differ across co-regulated genes, and how sequence and function of *cis*-elements evolve at the motif level. Certain findings represent conceptual advances in our understanding of mechanisms and evolution of gene regulation, while others are intriguing examples of biological phenomena that have been anticipated but not rigorously shown before due to a lack of functional resolution. I do note, however, that the insights gained by the comprehensive dissection of one system, in this case the regulatory elements of terminal target genes, will not universally translate to other systems, particular those that involve complex interactions of negative and positive factors that set up embryonic patterning (e.g., Davidson et al. 2002; Stathopoulos and Levine 2005).

I show that *cis*-coregulatory elements that drive muscle-specific gene expression in the developing *Ciona* tail exhibit very diverse architectures. Only two rather general aspects of element architecture are shared among all loci: the elements are compact and the regulatory motifs within them are of one or more of three types (Tbx6, MyoD, or CRE). Beyond these generalities, at least five architectural specifics are heterogeneous among the set, with each element representing a unique instantiation of these characteristics. These are (1) the number of functional motifs; (2) the exact identities of the functional motifs; (3) order and orientation of the functional motifs;

(4) spacing between the functional motifs; and (5) degree of functional contribution of each motif.

In contrast to the heterogeneity in specific regulatory architecture, I find strong conservation of both sequence and function in orthologous regulatory elements, at motif level resolution. To the extent that changes do occur in motif sequences, motif conservation correlates with the strength of motif function. This is a reassuring result, underscoring that a change in a strongly functional motif is likely to have a greater deleterious impact on gene expression (and is therefore more intensely selected against) than a change in a weakly functional motif. A complementary result is that sequence conservation drops off rapidly from the edge of functional motifs, reaching genome-wide background levels within a handful of bases. It is therefore clear that regulatory motif sequences are subject to greater levels of purifying selection than even immediately adjacent nucleotides, underscoring the motif itself as the fundamental unit of regulatory function.

An important practical point emerges from this particular conclusion: Evolutionary analyses are best complemented with high-resolution functional data if conclusions about the evolution of motif function are desired. Due to the tight selective footprint and the presence of non-functional motif predictions, evolutionary analyses relying solely on motif predictions, as opposed to functionally defined motifs, lead to biases in the direction of overestimating variability. To cement this point, I mimicked lower-resolution data and examined 500 bp regions encompassing our functional motifs. Within these regions I assessed the *C. savignyi*/*C. intestinalis* pair-wise percent identity of high confidence motif predictions in the *C. savignyi*

sequence. The distribution of resulting values is shifted significantly downward relative to the distribution built from functional motifs (Wilcoxon Rank Sum Test, $p < 0.05$), illustrating how nonfunctional motifs of a larger region dilute the conservation signal provided by the actually functional motifs.

Notwithstanding the general finding that motif function is highly conserved, our analyses also uncovered an apparent case of compensatory evolution in which sequence substitutions have modified the function of two adjacent regulatory motifs in a complementary fashion, at the α-Tropomyosin 2 locus (Fig. 2.8 A). To our knowledge, this represents the first demonstration in a metazoan of *cis*-regulatory compensatory evolution to be characterized functionally at motif-level resolution.

Precise quantification of motif function, as opposed to the binary encoding of a motif as functional or not, provided novel insights into *cis*-element functional architecture. I have demonstrated that regulatory motifs of an element exhibit a range of functionality, ranging from below the detection limit of our assay to strong function. Importantly, I found little evidence of genetic interactions (redundancy or synergism) among the motifs of an element, and, to a first approximation, the constituent motifs appear to function mostly independently and additively. I note that motif function would have been usually interpreted as 'redundant' if I had used binary encoding (functional versus nonfunctional) instead of a quantitative scale to describe function.

Further confirming additivity and independence of motifs is that each of the three motif types controlling muscle gene coexpression are at least partially functionally equivalent, as shown by the substitution experiments (Fig. 2.2). Motif independence

and functional equivalence suggest that syntactical rules governing the assembly of *cis*-regulatory motifs into elements are quite flexible.

The two most prominent developmental mechanisms that build a multicellular organism are pattern formation and cellular differentiation. Previous studies of regulatory architecture and evolution were conducted in pattern formation systems, either by leveraging sequence comparisons and broad functional genomic data (Yuh et al. 1998; Ludwig et al. 2000; Bertrand et al. 2003; Oda-Ishii et al. 2005) or by studying a single regulatory element in detail (Dermitzakis et al. 2003; Moses et al. 2006; Zeitlinger et al. 2007). By contrast, I here dissect the regulation of coexpression during cellular differentiation, and introduce a quantitative framework for the direct experimental analysis of motif function. Using the *Ciona* muscle system, I demonstrate that coexpression is driven by regulatory motifs of broadly varying activity assembled into a diverse array of *cis*-elements. Despite this flexibility in *cis*-regulatory architecture, motif-level sequence and function are exquisitely maintained in distantly related orthologs. Thus, while a diversity of *cis*-regulatory architectures can generate nearly identical phenotypic outputs, the fitness landscapes separating them appear to be sufficiently rugged to strongly constrain their evolution (Wright 1932).

Our findings have significant implications for our understanding of polymorphisms affecting such coregulatory systems. Polymorphisms in *cis*-elements will range in phenotype, depending on the amount of activity that the affected motif contributes to the function of its element; the most direct evidence for this view is the wide range of effects on *cis*-element function by the individual motif mutants I tested.

Similarly, a polymorphism in a *trans*-acting factor will not affect expression of all targets equally but will instead have a target-specific effect whose magnitude is determined by the architecture of the target's *cis*-element. These conclusions illustrate the challenges that lie ahead for interpretation of genetic variation in gene regulatory systems, particularly in *Ciona*'s most advanced close relatives -- humans.

Chapter 5 Methods

Molecular biology

Reporter constructs were built using standard PCR cloning techniques (Johnson et al. 2005; Sambrook and Russell 2001) using a previously described vector backbone (Johnson et al. 2004; Johnson et al. 2005). At each locus examined, I first built constructs with 2-5kb of endogenous sequence. This sequence was amplified via PCR off genomic DNA: either genomic DNA from the *C. savignyi* individual whose genome was sequenced or off a pool of genomic DNA extracted from 50 unrelated *C. intestinalis* individuals. The use of either DNA of known sequence or pooled DNA was necessary for efficient PCR amplification, due to the high polymorphism rate in *Ciona* (Dehal et al., 2002; Vinson et al. 2005; Small et al. 2007b). All subsequent constructs at each locus were built by modification of these initial reporter constructs, therefore all constructs at each locus are derived from the same haplotype.

Truncation constructs were generated by PCR cloning, internal deletions were generated by overlap extension PCR. Site-directed mutations were also generated by overlap extension PCR so as to scramble the bases of the endogenous sequence while maintaining local sequence length, spacing, and GC content (see Table S2 for locus-by-locus summary of all relevant constructs). All constructs were verified by Sanger sequencing. Reporter constructs were maxiprepped (BioRad Quantum Prep) and concentrations were adjusted to 5µg/ul prior to electroporation. Detailed construct descriptions and primer sequences are available upon request.

Ciona husbandry and transfection

I have expanded the following protocols to provide all details that seemed relevant to me upon completion of my project. This protocol is largely an adaptation of Bob Zeller's methods (Zeller 2004; Zeller et al. 2006; http://www.bio.sdsu.edu/faculty/zeller.html). I have made several significant modifications and provided expanded details. Some of the nitty-gritty details are based on my own anecdotal evidence – adjustments I made because 'they worked.' As with any protocol, some of these details are only important because, if always followed, they result in reproducible data.

Husbandry

Starting during the Winter of '04-'05, all *C. intestinalis* used in my studies were collected near San Diego, CA, USA by Marine Research and Educational Products, run by Steve Le Page. As of May 2007, he can be reached at (510) 782-8936. Information is available on his website: http://www.m-rep.com. When actively using animals, I received weekly shipments of ~40 animals, that were shipped overnight via FedEx, bagged in seawater next to cold packs. Shipments ran ~$125 with shipping. Animals were typically collected over the weekend, allowed to sit in Steve's tanks 1-2 days, and shipped on Monday for a Tuesday delivery.

Upon arrival at Stanford, animals were immediately transferred to our aquarium. Starting August '04, I used a WCA-3 acrylic tray system aquarium for storing *Ciona*, which was purchased for ~$5000 from Sea Life Supply: (831) 394-0848 or http://www.sealifesupply.com/aquaria.htm. The aquarium set up was modified at Stanford by adding a T-valve drain. Prior to the winter of '05-06, this tank was filled with filtered seawater, purchased from PanOcean (I spoke with Will): (510) 782-8936 and http://www.panoceanaquarium.com. They charged $0.75 / gallon, with a minimum >100 gallon purchase (I don't remember the exact cut off). This was a hassle, so after winter '04-'05 I switched to using artificial seawater, which has worked just as well. I've used Instant Ocean brand salt mixture, purchased from Petco. Aquarium salinity was maintained at 30ppt, pH between 8.0 and 8.3, and the temperature was kept at 18C with the aquarium chiller. On occasion, I have used a marine aquarium buffering salt, Kent Marine Superbuffer-dKH, to maintain pH levels. On an approximately monthly basis, I performed ~25% water exchanges. Prior to transfection, animals were kept in the tank under constant light (standard fluorescent bulbs) for >24 hours, usually >48. I typically had gravid animals for 1-2 weeks. For additional thoughts on *Ciona* husbandry, please see Joly et al. 2007.

Transfection

Prior to transfection, have ready:

1. All chemicals, spatulas, dishes, bottles (everything) used for *Ciona* work must be separate from other lab supplies. Soap residue will make your life miserable – do not wash *Ciona* supplies with soap. DO NOT take this lightly. On occasion, everything will stop working – throw out your *Ciona* solutions and start over. Almost always works.
2. Room at 18-20C. I maintained this in R307 with a window AC unit.
3. Incubator at 16C, with room for ~5-60 Petri dishes.
4. Dissecting scope. I've used our Leica MZ95 with a 6x zoom.
5. Custom built electroporator (Johnson 2005; Johnson et al. 2005; Zeller et al. 2006). Set at 3000uF, 10ohms.
6. Mini-centrifuge. For spinning down embryos gently. Fisher cat # 05-090-128.
7. A lot of bench space. Preferably with bench coat, as this protocol is pretty messy.
8. Egg filter device. Cut bottom 1" off a 50mL Falcon tube. Cut a ~2x2" piece of 55 micron filter (Sefar Medifab # 3-100/44). Screw filter onto tube using tube lid. Cut off excess filter.
9. 250mL glass beaker.
10. 4L plastic bucket.
11. Cuvettes. 800µL 4mm gap. I reused these until they were visibly damaged – after use just wash several times with H2O, air dry.
12. Pipettes:
    a. 25mL disposable
    b. 5 ¾" Pasteur, VWR#14673-010

c. 9” Pasteur, VWR#14673-043

d. embryo picking pipette: ~3mm glass tubing. With a flame, draw out one end to a point, break tip to leave a ~1mm opening. Attach ~3mm plastic tubing to other end, fold over and staple to form a small “bulb.”

13. 2 forceps
14. 1 mini-scissors
15. Maxi-prepped DNA (BioRad Quantum prep), at 5µg/µL. Be careful with concentration/resuspension, this solution is viscous and pippetting is typically inaccurate. Always check the final solution prior to transfection and adjust the volume added to transfection mix if needed (see below).
16. Solutions (store at room temp unless noted):
    a. Protease type XIV solution. Aliquots are 2.5mg protease in 100µL TE pH 7.5. One aliquot per dechorionation. Store at -80C. (Sigma catalog #P5147).
    b. Sodium thioglycolate. 100mg aliquots in 15mL conical tubes. One aliquot per dechorionation. Store at -80C. (Sigma catalog #T0632).
    c. 0.77M D-Mannitol. Sterilized. Aliquot into 50mL conical tubes – enough for ~100 transfections. (Fisher M120-500).
    d. 20x PBS.
    e. 1M $MgCl_2$
    f. 100mM $K_3Fe(CN)_6$. Store at 4C.
    g. 100mM $K_4Fe(CN)_6$. Store at 4C.
    h. Triton X-100

i. 4% XGal in Dimethylformamide. Store at 4C in ~5mL aliquots, wrapped in foil to avoid light. Fisher cat # BP1615-1.

j. PBSTr: 1x PBS with 1% Triton X-100.

k. Paraformaldehyde, 16% solution, EM grade. Store at 4C. Electron Microscopy Sciences, cat # 15710.

l. Embryo fixation solution (-paraformaldehyde). 500mM NaCl, 27mM KCl, 2mM EDTA.

Just prior to transfection prepare:

1. 2L submicron filtered artificial sea water (SMFASW), from the same tank as your animals (This is important – subtle changes in water chemistry seem to disrupt development). I've used Nalgene bottle top filters, from Fisher #291-4545.
2. 1L SMFASW, as above, plus 20µg/mL Kanamyacin and 0.1mM EDTA (SMFASW+). I always kept 1000X stocks in the 16C incubator in 50mL conical tubes.
3. Plates:
    a. Per dechorionation:
        i. 1x 35mm dish (Falcon 35-3001), for dechorionation.
        ii. 5x 60mm dishes (Falcon 35-1007)
            1. 1x for fertilization control
            2. 4x for wash plates

iii. 1x 100mm dishes (Fisher 08-757-12), for dechorionation control.

b. Per transfection:

iv. 1x 100mm dish (Fisher 08-757-12)

c. To prep plates:

v. prepare molten agarose: 1g agarose / 100mL artificial sea water.

vi. Allow to cool to ~55C

a. Pour agarose onto first plate. Transfer agarose to subsequent plates bucket-brigade style until you run out. Repeat. You only need a really thin coat of agarose on each plate.

b. Allow agarose to set ~15 minutes.

c. Add 30mL SMFASW+ to each 100mm plate, 15mL to each 60mm plate. Make sure 35mm plates don't dry out – I keep them upside down in the 16C incubator.

4. Prepare cuvettes:

a. Add 480µL 0.77M D-Mannitol

b. Add 20µL 5µg/µL plasmid DNA in TE pH 8.0, mix well. Check concentration of final solution, adjust volume as needed to achieve 100µg. See notes on constructs.

5. To each aliquot of Sodium Thioglycolate, add:

a. 10mL SMFASW. Vortex to suspend.

b. 25µL 10N NaOH. Invert to mix. Check pH – should be 11 – extremely important.

6. Allow protease aliquots to thaw on ice.
7. Set up 'fertilization plate' – 1x per dechorionation - 1x 100mm plate (no agarose needed), add 30mL SMFASW.
8. Turn on electroporator. Check settings – 3000uF, 10ohms.

Dissection/Fertilization/Dechorionation:

1. Collect animals from tank. I typically use 2-4 gravid animals per dechorionation (depending on the # of transfections I'm trying to squeeze in). I use a mini-cooler (~1L size). I add some water from the tank to the cooler and put the animals in that.
2. In a 100mm Petri dish, remove the tunic from each animal. I use two forceps and a mini-scissors for this. Squeeze the animal away from the base of the tunic, cut ½" off the bottom of the tunic. Make a cut in the tunic from the hole in the base up along one side. Using the forceps, slide the tunic off. Trash the tunic.
3. While animal is in Petri dish, carefully cut the oviduct and vas deferens of each animal (avoid cutting the intestine – seems to poison the reactions). Some sperm and eggs will spill into the dish. Use the forceps to squeeze out the remaining sperm and eggs. Using a 5 ¾" Pasteur pipette, transfer sperm and eggs to fertilization plate.
4. Repeat for each animal, gently swirl fertilization plate.
5. After addition of sperm/eggs from 2nd animal, start timer.
6. Let sit for 1.5 minutes.

7. Place embryo filter column into 250mL beaker, pour fertilization mix into filter column. Water/sperm/etc. will move through the filter, eggs will be held on top. Sometimes the filter will clog – if so, gently tap the filter column on the bottom of the beaker until flow resumes.
8. Discard flow trough to 4L bucket.
9. Wash eggs with 30-40mL SMFASW
10. Repeat 8-9 2x
11. When ~2mL remain on column after 3rd wash:
    a. Transfer ~20 eggs to 60mm fertilization control dish. Lid dish.
    b. Pour enough Na-thioglycolate solution to cover the coated 35mm Petri dish (~1mL).
12. Dump eggs onto 35mm Petri dish by inverting column. Wash remaining eggs onto plate by squirting SMFASW from underside of filter (using fresh pipette).
13. Allow eggs to settle on dish (30 seconds)
14. Remove as much liquid as possible from dish w/o disturbing many eggs. Discard.
15. Wash in ~2mL Na-thioglycolate soln.
16. Repeat 13-15 2x.
17. Remove liquid again.
18. At this point, you should have ~3.5mL Na-thioglyoclate soln left. Add protease aliquot to Na-thioglycolate soln.
19. Wash in Na-Thioglycolate-protease soln. Timer should read 6 minutes.

20. Move dish to dissecting scope – watch through scope to check dechorionation. Watch carefully until you get a feel for the process. First, the outer follicle cells will fall off (should happen at about 9 minutes). Second, the inner chorion cells will fall off. When this happens, the embryos will change color, from dull orange/brown to pink. This step happens slowly at first, but as 12 minutes approaches, will rapidly move to completion.
21. Allow the eggs to settle, gently shake dish, repeat ~3x. I've found that shaking the dish ←→ works better than swirling.
22. When timer reads 9 minutes, gently swirl eggs with pipette, allow to settle, repeat.
23. When timer reads 12 minutes 50-95% of the eggs should have lost their chorions.
24. Allow eggs to settle, swirl dish to collect eggs in center.
25. Using 9" Pasteur pipette, gently transfer eggs to 1st 60mm wash plate. Transfer should be accomplished in 1 or 2 aliquots. Do not allow air bubbles into pipette – it will kill your eggs and you should just start over. Keep pipette as vertical as possible, which seems to keep eggs from sticking to the glass.
26. Allow eggs to settle, swirl to collect at center of plate.
27. Transfer eggs to next wash plate with fresh 9" pipette.
28. Repeat 26-27 2x.
29. Allow eggs to settle, swirl to collect.

Electroporation:

1. Using 1mL adjustable pipette-man with standard plastic tips, collect 300µL eggs/water from last wash dish. Keep tip vertical.
2. Transfer eggs to cuvette, mix twice, gently. Volume should be 800µL. Don't create/add any air bubbles to the mix. If so, remove with pipette.
3. Place cuvette in shock pod.
4. Flip charge switch on.
5. Wait ~3 seconds. I typically take this time to switch pipette tips.
6. Flip charge switch off.
7. Press discharge button.
8. Remove cuvette. (successful electroporation results in the creation of small bubbles in the transfection mix).
9. Pour contents of cuvette into 100mm transfection plate.
10. Repeat 1-9 for each transfection.
11. Add lids to plates, allow post-electroporated embryo plates to sit out for 5 minutes.
12. Carefully transfer plates to 16C incubator.
13. Notes:
    a. Stop when timer reaches 22 minutes. Electroporations after this time point will have reduced efficiency, reduced embryo survival, and delayed embryo growth.
    b. From 4 adults, I have no problem getting 15 transfections – the limiting factor is time.
    c. Always include a positive control plasmid to transfect.

d. Pay attention to order of transfections – note in the morning if any of the plates are 'off'.

Embryo collection/fixation/staining

1. Embryos are fixed at 13.5-15.5 hours post fertilization. Embryo collection typically takes 3-5 minutes per plate, depending on survival rates and practice. Think through when you need to start collecting embryos.
2. Using the embryo picking pipette, transfer 100-500 embryos from each plate to a 1.5 mL eppie tube.
3. Keep eppies with embryos at 16C until fixation.
4. Spin down embryos in mini-centrifuge (<700g) for ~30 seconds.
5. Remove as much water as possible with 5 ¾" Pasteur pipette (down to ~100µL).
6. Add 264µL embryo fixation buffer + 36µL 16% paraformaldehyde to each tube.
7. Invert twice gently.
8. Let tubes sit at room temp x 30 minutes.
    a. While waiting, make embryo staining buffer: PBS plus 1mM $MgCl_2$, 3mM $K_3Fe(CN)_6$, 3mM $K_4Fe(CN)_6$, 1% Triton X-100.
9. Spin down embryos (as above).
10. Remove as much liquid as possible.
    a. Fixation soln must be treated as hazardous waste.
11. Add 500µL PBSTr.

12. Let tubes sit at room temp x 5 minutes.
13. Spin down embryos.
14. Remove liquid.
15. Repeat 11-14.
16. Add 500µL embryo staining buffer.
17. Repeat 11-14.
    a. Staining buffer must be treated as hazardous waste.
18. Add 1.5mL staining buffer + 1mM X-Gal.
19. Mix by gentle inversion.
20. Incubate tubes for 4 hours at 37C.
21. Spin down embryos.
22. Remove as much liquid as possible.
23. Add 1.5mL PBS.
24. Store embryos in eppies at 4C until ready to image.

Imaging/Scoring:

1. Transfer embryos to 12 well non-coated tissue culture plates (Greiner bio-one # 665-180).
2. Take a picture of each well.
    a. I used the Tan/Baker lab dissecting scope. The scope itself was almost the same as ours, but the digital camera was connected to a computer via firewire, so I could make adjustments to focus/lighting/zoom on the fly.

3. Score each image using MARKER
    a. http://mendel.stanford.edu/SidowLab/downloads/Marker/Marker.zip
    b. Each embryo is scored on a 0-5 scale, representing 0%, 1-20%, 21-40%, 41-60%, 61-80%, and 81-100%, respectively, of muscle cells expressing the transgene. Calculate a weighted average, estimating the fraction of muscle cells stained for a given transfection.

The scalability of this protocol:

I routinely use this protocol to do 60 transfections per day, each generating ~75 stained transgenic embryos. This is accomplished in two sets of two batches of dechorionations/transfections. Each batch of transfections takes approximately one hour and I wait about 2 hours in between batches. This delay is necessary to provide enough time the next day for embryo picking, fixing, and staining. I do this 2x2 in order to keep the staging between experiments as tight as possible (with only a pair of fertilizations, the two batches will only be off by 30 minutes).

To make use of a wide dynamic range of expression frequency for assaying the activity of the mutagenized constructs, I tuned the transfection protocol so that most wild type constructs drove expression in over 30% but less than 80% of muscle cells (Table S1, column 3) , as opposed to the 100% that would be the norm for the endogenous locus. Initial analyses of five independent transfections and assays of the same constructs ("biological replicates") showed that the results were remarkably reproducible presumably because of the thousands of cells assayed in each

transfection, and because of stereotypic transfection conditions. The replicates resulted in stable estimates of activity for each construct, as revealed by the standard deviation of the fraction of expressing cells for each construct (mean SD = 0.074 efu, median SD = 0.064 efu; Fig. S8).

Quantitative RT-PCR

Embryos were transfected and allowed to develop as described above. Total RNA was extracted from transfected embryos at 14 hours after fertilization as follows. Approximately 100 embryos were collected in approximately 100uL of artificial sea water. Embryos were then homogenized on ice with a glass dounce after addition of 500uL embryo lysis buffer (100mM NaCl, 20mM Tris, pH 8.0, 10mM EDTA, 1% SDS, 250ug/mL proteinase K, in DEPC treated water). Homogenate was incubated at 42C for one hour followed by two extractions of acidic phenol:chloroform and a final chloroform extraction. RNA was precipitated with sodium acetate and ethanol, after which pellets were then washed in 75% ethanol. RNA was suspended in 50uL DEPC treated water and digested with DNAse I at 37C for 30 minutes, followed by extraction with acidic phenol chloroform. RNA was then precipitated overnight in 4M LiCl at 4C. After washing with 75% ethanol, pellets were resuspended in 20uL DEPC treated water. 1ug total RNA was used for oligo-dT primed first-strand cDNA synthesis with SuperScript III reverse transcriptase (Invitrogen). 5% of the resulting cDNA was used for quantitative real-time PCR using the DyNAmo HS SYBR Green qPCR kit (Finnzymes).

Alignments and interspecific sequence analyses

All local alignments were constructed as reported previously (Johnson et al. 2005). Scaffold-level alignments for each orthologous locus were collected from VISTA-LBNL (http://pipeline.lbl.gov/cgi-bin/gateway2?bg=Cioin2&selector=vista).

To estimate the amount of identity in motif-like sequences anywhere in the genome, I generated a background distribution by sampling. From each scaffold-level alignment, a set of sequences with a size distribution determined by the sizes of the functional regulatory motifs was sampled. In total, 21,000 mock motifs were generated, whose average identity (including insertions and deletions) was 21% (solid line in Fig. S6B).

To calculate sequence conservation at motif-adjacent positions I assessed the average identity at varying distances (pooling both 5' and 3' directions) from all orthologous functional motifs. All flanking positions that were themselves within functional motifs were treated as missing data. Identity within motifs was averaged across all motifs, and yielded a single value of 79% (position 0 in Fig. S6B).

Motif analyses

Initial position specific scoring matrices (PSSMs) (Fig. S2) were generated as follows. MyoD and CRE matrices were built from CisModule predictions (Johnson et al. 2005) that were modified to be symmetrical because of their presumed palindromic

nature. The Tbx6b/c matrix was built from *in vitro* binding data (Yagi et al. 2005). All three PSSMs included 1% added pseudocounts. Motif predictions were calculated as LOD scores (Johnson et al. 2004)

$$S = \sum_{i=1}^{L} \log \frac{f(b,i)}{p(b)}$$

where the motif is of length L, the PSSM is $f(b, i))$, with the frequency f of each base b at each position i. Background nucleotide frequencies, $p(b)$, were taken from the *C. savignyi* genome-wide average, which is 63.8% for G or C and 36.2% for A or T.

To investigate whether there is any sequence-specific signal outside motifs I aligned all functional motif sequences of the same type and built PSSMs and included 10 flanking bases on either side of the motif. As is evident from the sequence logos (built with WebLogo; Crooks et al. 2004) there is no sequence-specific information beyond the border of the motif (Fig. S6C-E). Comparison of the motif portions of these logos with those of the initial logos reveals, as expected, close similarity of the PSSMs.

Intraspecific sequence comparisons

To ask whether functional regulatory motifs have been subject to purifying selection in the *C. savignyi* population, I compared levels of polymorphism in functional motifs to the rest of the genome.

Polymorphism levels were calculated as heterozygosity, by comparison of the two haplotypes of the *C. savignyi* genome assembly (Small, K.S., et al. 2007a). The 13

statistically significantly functional motifs at single copy genes were covered by both haplotypes. Only 2 out of a total of 115 bases of these motifs were heterozygous, compared to the genome-wide average neutral heterozygosity of >8% (Small, K.S., et al. 2007a). This is unlikely to result from stochastic fluctuations in diversity as fewer than 4% of a sample of ~7,500 mock motif sets from across the *C. savignyi* genome display this little polymorphism (Fig. S7). Therefore, not only has selection removed *cis*-regulatory motif substitutions over long evolutionary timescales, but it also appears to be acting on extant variation by removing deleterious polymorphism.

Statistical analyses

All data analyses were conducted using R (R Development Core Team 2006) and custom perl scripts. 14 outlier transfections, as identified by Dixon's test (Sokal and Rahlf 1995), were removed from the total of 1237 quantitatively assayed transfections. Multivariate regression models were constructed for each locus, for each homolog, independently. For simplicity, I refer to 'motifs' in outlining the methodology, though some tested sequences were larger regions not bearing motifs. 45 clearly redundant constructs were consolidated to simplify model building and avoid overparameterization. All data analyses presented in the text are therefore based on the 175 constructs used to build the final models. I call the partial regression coefficient of each explanatory variable 'Motif activity.' Motif activity standard errors and tests of significance are derived from the same models. I considered motif activity to be statistically significant at $p < 0.05$. Linear regression models were built

using the R lm function. Logistic models were built by maximum likelihood using the R glm function with a binomial error distribution.

References

Arnosti, D., Barolo, S., Levine, M., Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122: 205-214.

Belting H.G., Shashikant, C., Ruddle, F. (1998). Modification of expression and *cis*-regulation of *Hoxc8* in the evolution of diverged axial morphology. Proc. Natl. Acad. Sci. U.S.A. 95: 2355-2360.

Bergman, C., Pfeiffer, B., Rincon-Limas, D., Hoskins, R., Gnirke, A., Mungall, C., Wang, A. ,Kronmiller, B., Pacleb, J., Park, S., Stapleton, M., et al. (2002). Assessing the impact of comparative sequence data on the functional annotation of the Drosophila genome. Genome Biol. 3:12.

Berman, B., Pfeiffer, B., Laverty, T., Salzberg, S., Rubin, G., Eisen, M., Celniker, S. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol. 5: R61.

Bertrand, V., Hudson, C. Caillol, D., Popovici, C., Lemaire, P. (2003). Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors. Cell. 115:615-627.

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

Britten, R. and Davidson, E. (1969). Gene regulation in higher cells: a theory. Science. 165: 349-357.

Buttgereit, D. (1993). Redundant enhancer elements guide *beta 1 tubulin* gene expression in apodemes during *Drosophila* embryogenesis. J. Cell. Sci. 105: 721-727.

Casillas, S., Barbadilla, A., Bergman, C. (2007). Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol. Biol. Evol. doi:10.1093/molbev/msm150

Chen, A.E., Ginty, D.D., and Fan, C. (2004). Protein kinase A signaling via CREB controls myogenesis induced by Wnt proteins. Nature. 433:317-322.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 7:901-13.

Corbo, J., Levine, M., Zeller, R. (1997). Characterization of a notochord-specific enhancer from the *Brachyury* promoter region of the ascidian, *Ciona intestinalis*. Development. 124: 589-602.

Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. 11:2463-2468.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004). WebLogo: A sequence logo generator. Genome Res. 14: 1188.

Crowley, E., Roeder, K., Bina, M. (1997). A statistical model for locating regulatory regions of genomic DNA. J. Mol. Biol. 168: 8-14.

Davidson, E.K. (2001). Genomic Regulatory Systems (San Diego: Academic Press).

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A genomic regulatory network for development. Science. 295:1669-16678.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science. 298:2157-2167.

Dermitzakis, E.T., Bergman, C.M., Clark, A.G. (2003). Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol. 20:703-14.

Dermitzakis, E.T. and Clark, A.G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol. 19:1114-1121.

Elena, S.F. and Lenski, R.E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. Nature. 390:195-398.

Emison, E.S., McCallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., NISC Comparative Sequencing Program, Cutler, D.J., Green, E.D., Chakravarti, A. (2005). A common, sex-dependent mutation in a putative RET enhancer underlies Hirschsprung disease susceptibility. Nature 434:857-863.

ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447: 799-816.

Fiering, S., Whitelaw, E., Martin, D.I.K. (2000). To be or not to be active: the stochastic nature of enhancer action. BioEssays. 22:381-387.

Galant, R., Walsh, C., Carroll, S. (2002). Hox repression of a target gene: extradenticle-independent, additive action through multiple monomer binding sites. Development. 129: 3115-3126.

Garrity, P., Chen, D., Rothenberg, E., Wold, B. (1994). Interleukin-2 transcription is regulated in vivo at the level of coordinated binding of both constitutive and regulated factors. Mol. Cell. Biol. 14: 2159-2169.

Gibbs, R.A., Weinstock, G.M., Metsker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen D., Worly, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493-521.

Gilad, Y., Oshlack, A., Smyth, G., Speed, T., White, K. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature. 440: 242-245.

Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., Carroll, S.B. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in Drosophila. (2005). Nature. 433:481-7.

Halligan, D. and Keightly, P. (2006). Ubiquitous selective constraints in the

Drosophila genome revealed by a genome-wide interspecies comparison. Genome. Res. 16: 875-884.

Hersh, B., Carroll, S. (2005). Direct regulation of knot gene expression by ultrabithorax and the evolution of cis-regulatory elements in Drosophila. Development. 132: 1567-1577.

Jeong, S., Rokas, A., Carroll, S. (2006). Regulation of body pigmentation by the Abdominal-B Hox protein and its evolutionary gain and loss in Drosophila. Cell. 125: 1387-1399.

Hoch, M., Gerwin, N., Taubert, H., Jackle, H. (1992). Competition for overlapping sites in the regulatory region of the *Drosophila* gene Krüppel. Science. 256: 94-97.

Jacob, F., Monod, F. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3: 318-356.

Johnson, D.S. (2005). "Yureiboya comparative genomics." Ph.D. dissertation, Stanford University, CA.

Johnson, D.S., Davidson, B., Brown, C.D., Smith, W.C., Sidow, A. (2004). Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. Genome Res. 14:2448-2456.

Johnson, D.S., Zhou, Q., Yagi, K., Satoh, N., Wong, W., Sidow, A. (2005). De novo discovery of a tissue-specific gene regulatory module in a chordate. Genome Res. 15:1315-1324.

Johnson, D.S. (2005). "Yureiboya comparative genomics." Ph.D. dissertation, Stanford University, CA.

Joly et al. Culture of *Ciona intestinalis* in closed systems. Dev. Dyn. 29 Mar. (2007). Doi: 10.1002/dvdy.21124

Khaitovich, P., Hellman, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachman, M., Päabo, S. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science. 309: 1850-1854.

Kimura, M. (1983). The neutral theory of molecular evolution. (Cambridge: Cambridge University Press).

King, M.C. and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science 188:107-16.

Kusakabe, T., Yoshida, R., Ikeda, Y., Tsuda, M. (2004). Computational discovery of DNA motifs associated with cell type-specific gene expression in Ciona. Dev Biol. 276:563-80.

Laney, J., Biggin, M. (1996). Redundant control of Ultrabithorax by zeste involves functional levels of zeste protein binding at the ultrabithorax promoter. Development. 122: 2303-2311.

Lemos, B., Mieklejohn, C., Caceres, M., Hartl, D. (2005). Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution. 59: 126-137.

Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Disruption of a long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. Proc. Natl. Acad. Sci. U.S.A. 99: 7548 - 7553.

Lieb, J., Liu, X., Botstein, D., Brown P. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet. 28: 327-334.

Ludwig, M.Z., Bergman, C., Patel, N.H., Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 403:564-567.

Ludwig, M.Z., Palsson, A., Alekseeva, E., Bergman, C.M., Nathan, J., Kreitman, M. (2005). Functional evolution of a *cis*-regulatory module. PLoS Biol. 3:e93.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Deway, C.N., Siepel, A., Birney, E, Keefe, D., Scwartz, A., Hou, M., et al. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 17: 760-764.

Meedel, T.H., Chang, P., Yasou, H. (2006). Muscle development in Ciona intestinalis requires the b-HLH myogenic regulatory factor gene Ci-MRF. Dev Biol. 302:333-344.

Moses, A., Chiang, D., Kellis, M., Lander, E., Eisen, M. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol. Biol. 3:19.

Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y., Biggin, M.D., Eisen, M.B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol. 2(10):e130.

Nelson, C., Hersh, B., Carroll, S. (2004). The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 5: R25.

Oda-Ishii, I., Bertrand, V., Matsou, I., Lemaire, P., Saiga, H. (2005). Making very

similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*. Development. 132:1663-16674.

Pappu, K., Ostrin, E., Middlebrooks, B., Sili, B., Chen, R., Atkins, M., Gibbs, R., Mardon, G. (2005). Dual regulation and redundant function of two eye-specific enhancers of the Drosophila retinal determination gene dachshund. Development. 132: 2895-2905.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature. 444:499-502.

Piano, F., Parisi, M., Karess, R., Kambysellis, M., (1999). Evidence for redundancy but not *trans* fractor-*cis* element coevolution in the regulation of drosophila Yp genes. Genetics. 152:605-616.

Pilpel, Y., Sudarsanam, P., Church, G.M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. Nature Genetics. 29:153-159.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.-D., True, J.R., and Carroll, S.B. (2006) Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. Nature, 440:1050-1053.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Raser, J.M. and O'Shea, E. (2004). Control of stochasticity in eukaryotic gene expression. Science. 304:1811-1814.

Raser, J.M. and O'Shea, E. (2005). Noise in gene expression: origins, consequences, and control. Science. 309:2010-2013.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielson, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P. et al. (2005). Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. Genome Res. 15:1-18.

Rifkin, S., Kim, J., White K. (2003). Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. 33: 138-144.

Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverbeg, M.S., Goyette, P., Huett, A., Gree, T., Kuballa, P., Barmada, M.M., Datta, L.W., et al. (2007). A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. Nat. Genet. 39:605-613.

Sambrook and Russell. (2001). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor: Cold Spring Harbor Press).

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics. 34:166-176.

Segre, D., DeLuna, A., Church, G.M., Kishony, R. (2005). Modular Epistasis in yeast metabolism. Nature Genetics. 37:77-83.

Senger, K., Armstrong, G., Rowell, W., Kwan, J., Markstein, Levine, M. (2004). Immunity regulatory DNAs share common organizational features in *Drosophila*. Mol. Cell. 13: 19-32.

Siepel, A., Bejarano, G., Pederson, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al., (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034-1050.

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jonsson, B., Schluter, D., Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 428:717-23.

Small, K.S., Brudno, M., Hill, M.M., Sidow, A., (2007a). Extreme genomic variation in a natural population. Proc. Natl. Acad. Sci. U.S.A. 104:5698-5703.

Small, K.S., Brudno, M., Hill, M.M., Sidow, A., (2007b). A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biology. 13:1297-1305.

Small, S., Kraut, R., Hoey, T., Warrior, R., Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in *Drosophila*. Genes Dev. 5:827-839.

Small, S., Blair, A., Levine, M. (1992). Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. EMBO J. 11:4047-4057.

Sokol, R.R., Rahlf, F.J. (1995). Biometry (New York: W.H. Freeman and Company).

Stanojevic, D., Small, S., Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. Science 254:1385-1387.

Staphopoulos, A. and Levine, M. (2005). Genomic regulatory networks and animal development. Dev. Cell. 9:449-462.

Stone, J.R. and Wray, G.A. (2001). Rapid evolution of *cis*-regulatory sequences via

local point mutations. Mol Biol Evol. 18:1764-1770.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003). A gene co-expression network for global discovery of conserved genetic modules. Science. 302:249-255.

Sutter, N.B., Bustamante, C.D, Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P.G., et al. (2007). A single IGF1 allele is a major determinant of small size in dogs. Science. 316:112-115.

Thanos, D. and Maniatis, T. (1992). The high mobility group protein HMG I(Y) is required for NF-κB-dependent virus induction of the human IFN-β gene. Cell. 71: 777-789

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 39:31-40.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. (2004). Global mapping of the yeast genetic interaction network. Science. 303:808-813.

Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Strange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., et al. (2005).

Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome Res. 15:1127-1135.

Webb, C., Shabalina, S., Ogurtsov, A., Kondrashov, A. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res. 30: 1233-1239.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwan, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3(1):e7.

Wright, S. (1932). On the roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress on Genetics. 355-366.

Yagi, K., Takatori, N., Satou, Y., Satoh, N. (2005). Ci-Tbx6b and Ci-Tbx6c are key mediators of the maternal effect gene Ci-macho1 in muscle cell differentiation in Ciona intestinalis embryos. Dev Biol. 282:535-49.

Yuh, C.H., Bolouri, H., Davidson, E.H., (2001). *Cis*-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development. 128:617-629.

Zeitlinger, B.S., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., 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.

Zeller, R.W., Virata, M.J., Cone, A.C. (2006). Predictable mosaic transgene expression in ascidian embryos produced with a simple electroporation device. Dev Dyn. 235:1921-32.

Zeller, R.W. (2004). Generation and use of transgenic ascidian embryos. Methods Cell Biol. 74:713-30.

Zhou, Q. and Wong, W.H. (2004). CisModule: *De novo* discovery of *cis*-regulatory modules by hierarchical mixture modeling. PROC. NATL. ACAD. SCI. U.S.A. . 101:12114-12119.

Appendix 1 Summary of Regression Models

Gene	Species[1]	Strongest construct (efu)	Function in motifs	Number of explanatory variables (X)[2]	Number of Constructs (Y)[2]	Number of Transfections	Variance explained (R^2)[2]	p-value
AT1	*C. i.*	0.54	101%	7	13	68	0.6	5.50E-10
	C. s.	0.56	72%	7	14	71	0.62	4.90E-11
AT2	*C. i.*	0.55	70%	7	11	62	0.57	5.80E-08
	C. s.	0.61	82%	7	15	82	0.75	2.20E-16
CK	*C. i.*	0.71	87%	7	10	46	0.72	1.50E-09
	C. s.	0.58	96%	7	10	46	0.79	3.30E-14
MBP	*C. i.*	0.56	59%	4	8	39	0.79	1.80E-11
	C. s.	0.43	85%	4	8	40	0.83	1.10E-12
TI	*C. i.*	0.59	79%	5	10	49	0.86	2.20E-16
	C. s.	0.55	55%	5	10	52	0.73	5.80E-12
TT	*C. i.*	0.77	47%	5	12	54	0.73	9.10E-13
	C. s.	0.74	24%	4	9	37	0.89	8.60E-15
MA1	*C. s.*	0.32	100%	7	8	37	0.58	3.00E-04
MA3	*C. s.*	0.54	96%	7	8	40	0.72	2.50E-07
MLC1	*C. s.*	0.25	100%	3	5	19	0.49	1.50E-02
MLC5	*C. s.*	0.23	100%	3	6	21	0.34	6.50E-02
MRLC4	*C. s.*	0.42	100%	4	8	42	0.56	2.80E-06
MRLC5	*C. s.*	0.35	100%	3	6	24	0.88	1.70E-09
MRLC6	*C. s.*	0.23	100%	3	7	32	0.3	1.70E-02
	mean	0.5	0.82	5.21	9.37	45.32	0.67	
	median	0.55	0.87	5	9	42	0.72	
	sum			99	178	861		

1 *C.i., C. intestinalis; C.s., C. savignyi*

2 See Fig. 1 and Supplemental text, section 5.

Appendix 2 Summary of Constructs Used in Quantitative Analyses

Alphatropomyosin1

Aln coords	*C.i.* coords	[B]	[X]												
	Intercept	-0.163													
	4783-6042	0.051	1	0	0	0	0	0	0	0	0	0	0	0	0
	6043-6113	-0.057	1	1	0	0	0	0	0	1	1	1	1	1	1
-135	CRE.6114	0.046	1	1	1	0	0	0	0	0	1	1	1	1	0
-120	CRE.6129	0.125	1	1	1	1	0	0	0	1	0	1	1	1	0
-114	Tbx6.6134	0.119	1	1	1	1	1	0	0	1	1	0	1	1	1
-99	Tbx6.6147	0.201	1	1	1	1	1	0	0	1	1	1	0	1	1
-89	MyoD.6158	0.207	1	1	1	1	1	1	0	1	1	1	1	0	0

Construct#	868	913	1262	954	911	1185	910	1212	1121	1122	1123	1486	1253
[Y]	0.530	0.541	0.473	0.540	0.284	0.039	0.000	0.437	0.423	0.298	0.216	0.272	0.034

Aln coords	*C.s.* coords	[B]	[X]														
	Intercept	-0.200															
	4920-6833	0.149	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	6834-7053	0.059	1	1	0	0	0	0	0	1	1	1	1	1	1	1	0
-135	CRE.7054	0.012	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0
-120	CRE.7065	0.024	1	1	1	1	0	0	0	1	0	1	1	1	0	0	0
-114	Tbx6.7073	0.134	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0
-99	Tbx6.7086	0.085	1	1	1	1	1	0	0	1	1	1	0	1	1	0	0
-89	MyoD.7096	0.284	1	1	1	1	1	1	0	1	1	1	1	0	0	0	Tbx6

construct#	1608	502	1263	951	956	1182	696	1025	1527	1134	1133	1027	1236	1611	1632
[Y]	0.555	0.387	0.326	0.349	0.285	0.005	0.000	0.433	0.437	0.257	0.308	0.100	0.014	0.000	0.011

Alphatropomyosin2

Aln coords	*C.i.* coords	[B]	[X]											
	Intercept	-0.218												
	9787-10935	0.086	1	0	0	0	0	1	0	0	0	0	0	0
-2147	10936-11780	0.100	1	1	0	0	0	1	1	0	0	0	0	0
-120	Tbx6.11776	-0.048	1	1	1	0	0	1	1	0	0	0	0	0
-108	Tbx6.11788	0.236	1	1	1	1	0	0	0	0	1	1	1	MyoD
-89	CRE.11808	0.164	1	1	1	1	1	0	1	1	0	1	1	1
-77	CRE.11820	0.074	1	1	1	1	1	0	1	1	1	0	1	1
-65	Tbx6.11831	0.007	1	1	1	1	1	0	1	1	1	1	0	1

construct#	1575	1280	1708	1489	1211	1594	1452	1504	1517	1438	1406	1695
[Y]	0.359	0.362	0.210	0.547	0.002	0.002	0.020	0.027	0.016	0.130	0.173	0.570

Aln coords	*C.s.* coords	[B]	[X]														
	Intercept	0.004															
	10834-12242	0.099	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	12474-12634	0.015	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0
-120	12665-12674	0.273	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0
-108	Tbx6.12678	0.064	1	1	1	1	0	0	0	0	1	1	0	1	1	1	0
-89	CRE.12697	0.154	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1
-77	CRE.12709	0.020	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1
-65	Tbx6.12720	0.004	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0

construct#	1563	1596	1655	1524	1264	1321	1164	1586	1666	1694	1505	1490	1528	1484	1265
[Y]	0.609	0.536	0.519	0.250	0.160	0.025	0.000	0.414	0.549	0.518	0.290	0.075	0.207	0.347	0.049

Creatine Kinase

Aln coords	*C.i.* coords	[B]	[X]												
	Intercept	-0.160													
	152-1770	0.101	1	1	0	0	0	0	0	0	0	0	0	0	0
-311	1771-1792	0.006	1	1	1	0	0	0	1	1	1	1	1	1	1
-290	1793-1802	0.278	1	1	1	1	0	0	0	1	1	1	1	1	1
-270	1803-1818	0.381	1	0	1	1	1	0	1	0	1	1	RC	MyoD	CRE
-257	1823-1838	0.026	1	1	1	1	1	1	1	1	0	1	1	1	1
-245	1833-1848	-0.005	1	1	1	1	1	1	1	1	1	0	1	1	1

construct#	1576	1584	1214	1656	1305	1432	1275	1276	1278	1279	1404	1437	1405
[Y]	0.569	0.306	0.705	0.557	0.120	0.000	0.285	0.000	0.501	0.533	0.493	0.126	0.080

Aln coords	*C.s.* coords	[B]	[X]									
	Intercept	-0.265										
	360-1683	0.035	1	1	0	0	0	0	0	0	0	0
-311	1684-1727	0.013	1	1	1	0	0	0	1	1	1	1
-290	1706-1721	0.192	1	1	1	1	0	0	0	1	1	1
-268	1722-1737	0.453	1	0	1	1	1	0	1	0	1	1
-257	1738-1749	0.114	1	1	1	1	1	1	1	1	0	1
-245	1750-1764	0.017	1	1	1	1	1	1	1	1	1	0
-230	1765-1780		1	1	1	1	1	1	1	1	1	1

construct#	1577	1585	1658	1659	1347	1348	1660	1661	1662	1663
[Y]	0.583	0.083	0.536	0.558	0.194	0.000	0.379	0.001	0.410	0.508

Myosin Binding Protein

Aln coords	*C.i.* coords	[B]	[X]							
	Intercept	-0.070								
	4076-5167	0.280	1	0	0	0	0	1	0	0
-373	MyoD.5168	0.198	1	1	0	0	0	0	1	0
-363	Tbx6.5179	0.092	1	1	1	0	0	0	0	1
-345	CRE.5196	0.118	1	1	1	1	0	0	0	0

Construct#	1562	881	1177	1128	915	1595	1267	1415
[Y]	0.561	0.493	0.062	0.030	0.001	0.268	0.048	0.005

Aln coords	*C.s.* coords	[B]	[X]							
	Intercept	-0.069								
	12966-14069	0.073	1	0	0	0	0	1	0	0
-373	MyoD.14069	0.254	1	1	0	0	0	0	1	0
-363	Tbx6.14079	0.077	1	1	1	0	0	0	0	1
-348	CRE.14095	0.078	1	1	1	1	0	0	0	0

Construct#	1564	1183	1525	1250	1381	1587	1266	1399
[Y]	0.407	0.428	0.026	0.001	0.001	0.008	0.117	0.000

Troponin I

Aln coords	*C.i.* coords	[B]	[X]									
	Intercept	-0.051										
	8448-9028	-0.063	1	0	0	0	0	0	1	0	0	0
	9029-9687	0.184	1	1	0	0	0	0	1	0	0	0
-813	Tbx6.9688	0.073	1	1	1	0	0	0	0	1	1	0
-803	CRE.9697	0.025	1	1	1	1	0	0	0	0	1	1
-790	MyoD.9710	0.359	1	1	1	1	1	0	0	0	0	0

Construct#	1565	1622	1439	1491	1543	1518	1612	1493	1519	1492
[Y]	0.530	0.591	0.452	0.317	0.269	0.003	0.068	0.010	0.016	0.000

Aln coords	*C.s.* coords	[B]	[X]									
	Intercept	0.018										
	7233-7672	-0.009	1	0	0	0	0	0	1	0	0	0
	7673-8270	0.248	1	1	0	0	0	0	1	0	0	0
-813	Tbx6.8271	0.036	1	1	1	0	0	0	0	1	1	0
-803	CRE.8281	0.006	1	1	1	1	0	0	0	0	1	1
-790	MyoD.8294	0.247	1	1	1	1	1	0	0	0	0	0

Construct#	1561	1621	63	1503	1373	1552	1579	1247	1498	1316
[Y]	0.527	0.554	0.344	0.213	0.280	0.026	0.269	0.007	0.064	0.035

Troponin T

Aln coords	*C.i.* coords	[B]	[X]											
	Intercept	-0.022												
	3797-5450	0.420	1	0	0	0	0	0	1	0	0	0	0	0
-580	Tbx6.5451	0.176	1	1	0	0	0	0	0	1	1	1	0	0
-551	CRE.5481	0.023	1	1	1	0	0	0	0	0	1	1	0	1
-534	CRE.5497	-0.014	1	1	1	1	0	0	0	1	0	1	1	0
-516	CRE.5507	0.194	1	1	1	1	1	0	0	1	1	0	0	0

construct#	1574	1163	1050	929	928	957	1610	1483	1523	1485	1382	1358
[Y]	0.766	0.518	0.110	0.165	0.171	0.000	0.407	0.276	0.340	0.099	0.005	0.014

Aln coords	*C.s.* coords	[B]	[X]								
	Intercept	0.038									
	5270-9859	0.581	1	0	0	0	0	1	0	0	0
-580	Tbx6.9860	0.256	1	1	0	0	0	0	1	1	1
-551	CRE.9878	-0.097	1	1	1	0	0	0	0	0	1
-516	CRE.9910	0.029	1	1	1	1	0	0	0	1	0

construct#	1620	1180	1179	960	952	1623	1500	1499	1501
[Y]	0.742	0.247	0.024	0.000	0.000	0.658	0.304	0.370	0.168

Muscle Actin

Aln coords	MA1/11.coords	[B]	[X]							
	Intercept	-0.792								
-348	851-1018	0.001	1	0	0	0	0	0	0	0
-162	MyoD.1018	0.206	1	1	0	1	1	1	1	1
-150	MyoD.1030	0.136	1	1	1	0	1	1	1	1
-142	MyoD.1038	0.037	1	1	1	1	0	1	1	1
-136	Tbx6.1044	0.206	1	1	1	1	1	0	1	1
-127	MyoD.1053	0.206	1	1	1	1	1	1	0	1
-118	Tbx6.1062	0.206	1	1	1	1	1	1	1	0

Construct#	227	1350	1298	1495	1496	1385	1422	1423
[Y]	0.320	0.206	0.000	0.070	0.169	0.000	0.000	0.000

Aln coords	MA3/12.coords	[B]	[X]							
	Intercept	-0.187								
-239	(943-1011)	0.031	1	0	0	0	0	0	0	0
-171	MyoD.1011	0.323	1	1	0	0	0	0	0	0
-159	MyoD.1023	-0.030	1	1	1	0	0	0	0	0
-142	MyoD.1039	0.112	1	1	1	1	0	0	0	0
-136	Tbx6.1045	0.106	1	1	1	1	1	0	1	1
-127	MyoD.1054	0.106	1	1	1	1	1	1	0	1
-118	Tbx6.1063	0.081	1	1	1	1	1	1	1	0

Construct#	230	1301	1299	1369	1370	231	1516	1494
[Y]	0.543	0.512	0.189	0.219	0.106	0.000	0.000	0.025

Myosin Light Chain

Aln coords	MLC1 coords	[B]	[X]				
	Intercept	-0.138					
-324	Tbx6.664	0.170	1	0	0	0	1
-300	CRE.687	0.059	1	1	0	0	0
-215	CRE.708	0.141	1	1	1	0	0

Construct#	1364	1354	1355	1542	1578
[Y]	0.255	0.016	0.003	0.000	0.004

Aln coords	MLC5 coords	[B]	[X]					
	Intercept	-0.084						
-527	Tbx6.374	0.077	1	0	0	1	1	1
-518	CRE.383	0.058	1	1	0	0	0	1
-505	MyoD.397	0.065	1	1	0	0	1	0

Construct#	1383	1420	1384	1421	1502	1520
[Y]	0.233	0.005	0.000	0.001	0.021	0.004

Myosin Regulatory Light Chain

Aln coords	MRLC4 coords	[B]	[X]							
	Intercept	-0.029								
-357	MyoD.1226	0.027	1	0	0	0	0	1	1	1
-344	Tbx6.1241	0.016	1	1	0	0	0	0	1	1
-320	MyoD.1257	0.293	1	1	1	0	0	1	0	1
-304	MyoD.1271	0.038	1	1	1	1	0	1	1	0

Construct#	1306	1497	1375	1433	1334	1553	1597	1566
[Y]	0.424	0.269	0.300	0.048	0.000	0.298	0.000	0.292

Aln coords	MRLC5 coords	[B]	[X]					
	Intercept	0.082						
-357	MyoD.1151	0.117	1	0	0	0	1	1
-344	Tbx6.1167	0.122	1	1	0	0	0	1
	Tbx6.1263	-0.139	1	1	1	0	1	0

Construct#	1368	1416	1307	650	1402	1488	
[Y]	0.168	0.041	0.001	0.000	0.029	0.354	

Aln coords	MRLC6 coords	[B]	[X]						
	Intercept	-0.057							
-364	MyoD.1289	0.066	1	0	0	0	1	1	1
-346	Tbx6.1296	0.078	1	1	0	0	0	1	0
-336	CRE.1306	0.051	1	1	1	0	1	0	0

Construct#	513	1351	1352	1303	1417	1418	1419
[Y]	0.235	0.010	0.000	0.000	0.003	0.000	0.002