Supporting Online Material for

Functional Architecture and Evolution of Transcriptional Elements That Drive Gene Coexpression

Christopher D. Brown, David S. Johnson, Arend Sidow*

*To whom correspondence should be addressed. E-mail: arend@stanford.edu

Published 7 September, *Science* **317**, 1557 (2007)
DOI: 10.1126/science.1145893

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/317/5844/1557/DC1)

Supporting Online Material

1. Muscle Genes Encoded in Multigene Clusters

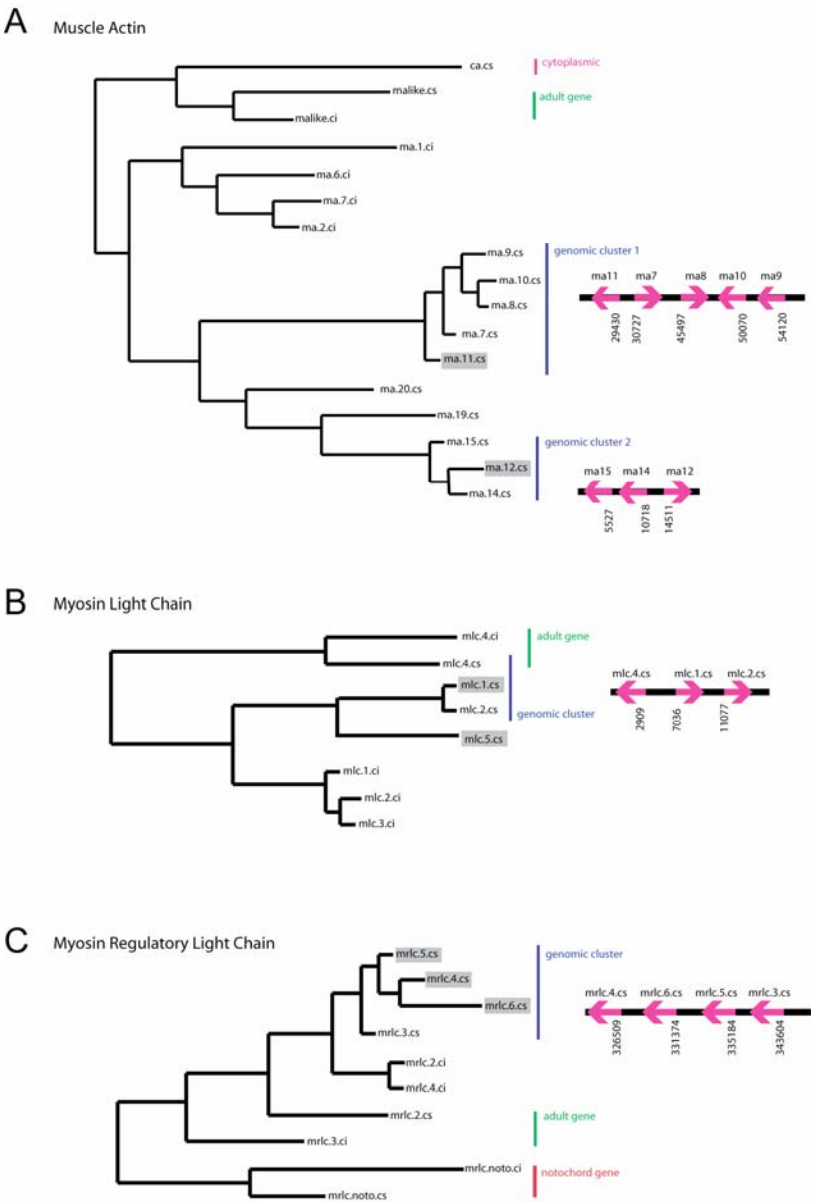


Fig. S1. Evolutionary relationships of **(A)** Muscle Actin, **(B)** Myosin Light Chain, **(C)** Myosin Regulatory Light Chain paralogs. Trees built from third-position coding sequences by maximum likelihood (1). Functional analyses reported in this study were carried out on genes shaded grey. Several 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.

2. Transcription Factor Binding Site Types that Drive *Ciona* Muscle Coregulation.

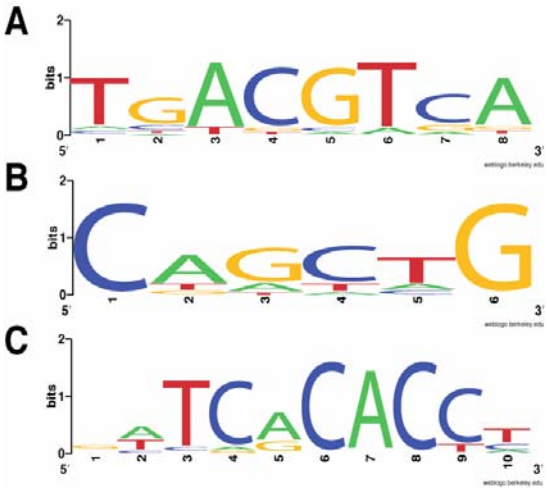


Fig. S2. Logos of PSSMs used in motif finding across all loci. (A) CRE, (B) MyoD, (C) Tbx6.

3. On the use of "Expression Frequency Units" as a Metric for Expression

The activity of reporter constructs electroporated into *Ciona* embryos has traditionally been scored as the percentage of embryos that express the reporter in the cells of interest. We reasoned that scoring transfections based on the percentage of stained cells of interest in each embryo would capture more information. (For methodological details, see supplemental section 9 - Materials and Methods). Similar cell-type specific scoring metrics have been previously employed to score *Ciona* transgene expression (2). We use the term "Expression Frequency Units", or "efu", to describe the metric.

Two lines of evidence suggest that efu is a robust scoring metric. First, an analysis of the distribution of stained cells per embryo demonstrates that efu captures a probabilistic shift in cell-autonomous reporter expression (Fig. S3; see also Methods). Across all transfections, as the percentage of stained muscle cells increases there is a sequential increase in the percentage of embryos that exhibit staining in greater numbers of cells.

Second, the fraction of muscle cells stained (as measured by efu) is directly correlated with the amount of LacZ mRNA in a pool of transfected embryos (as measured by quantitative RT-PCR), with Spearman's $\rho = 0.96$ (Fig. S4).

To quantify RNA expression levels, we followed an RNA extraction protocol based on (3), followed by first-strand cDNA synthesis and quantitative PCR. Briefly, embryos were transfected and allowed to develop as described below. Total RNA was extracted from transfected embryos at 14 hours after fertilization as follows. Approximately 100 embryos were collected in 100 µL of artificial sea water and homogenized on ice after addition of 500 µL embryo lysis buffer (100 mM NaCl, 20 mM Tris, pH 8.0, 10 mM EDTA, 1% SDS, 250 µg/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. RNA was suspended in 50µL 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. 1µg total RNA was used for oligo-dT

primed first-strand cDNA synthesis with SuperScript III reverse transcriptase (Invitrogen). After reverse transcription, reactions were digested with RNAse H, followed by digestion with DpnI, to remove any remaining plasmid DNA. 5% of the resulting cDNA was used for quantitative real-time PCR using the DyNAmo HS SYBR Green qPCR kit (Finnzymes). LacZ transcript levels produced off 6 different constructs was estimated using two lacZ amplicons, each flanking a different DpnI site, and each amplicon quantified in duplicate.

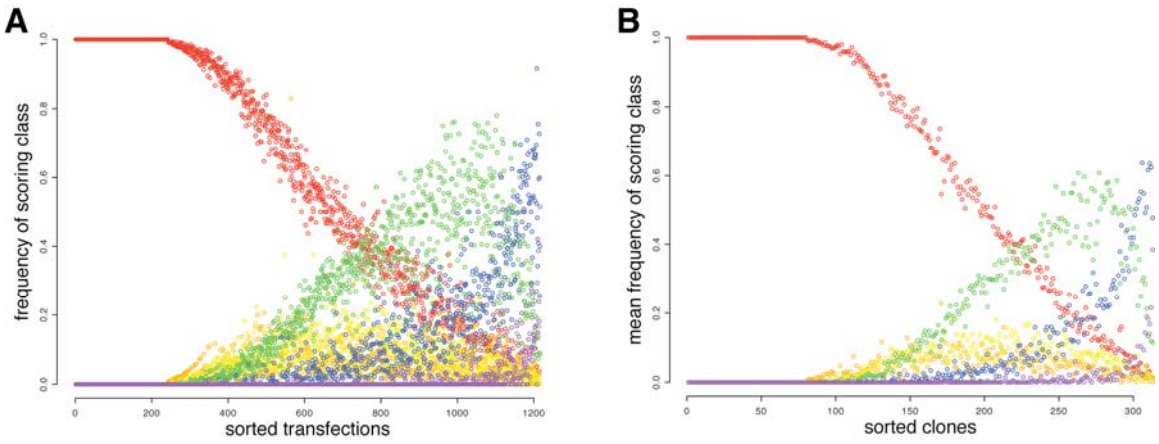


Fig. S3. Distributions of expression frequencies across transfections **(A)** and constructs **(B)**. All transfections (A), or the means of the five replicated transfections of each construct (B), are sorted along the x-axis by the total percentage of stained muscle cells. Each transfection or construct is depicted as a set of six points, one for each scoring class. Scoring classes are the percentage of cells staining in each embryo, specifically: no expression (red), one cell to 20% of cells (orange), 20-40% (yellow), 40-60% (green), 60-80% (indigo), 80-100% (violet). On the left are nonfunctional transfections or constructs, with all transfected embryos in the red class and none in any of the other classes; as constructs get stronger (towards right), the distributions shift until many embryos express the construct in the majority of cells (green, indigo, violet), and few embryos express it in few or no cells (red, orange, yellow).

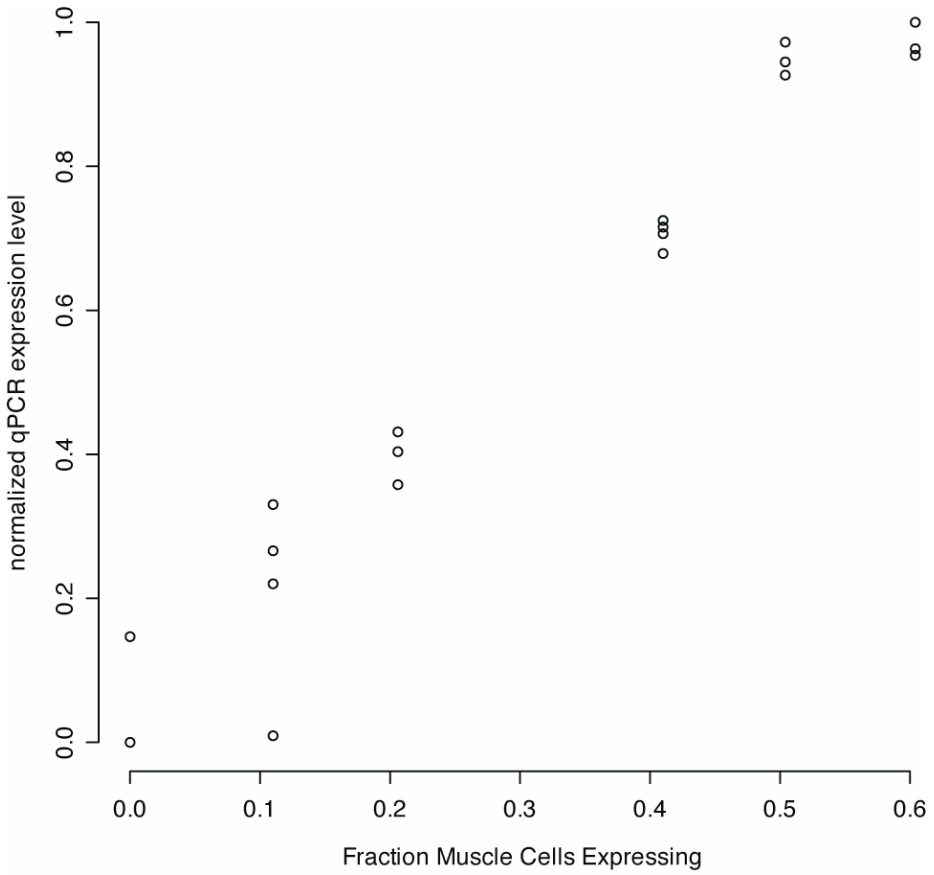


Fig. S4. Fraction of cells expressing a construct is proportional to LacZ RNA expression levels. The LacZ RNA expression levels driven by six constructs of varying activity (in efu, x-axis) were measured by quantitative RT-PCR (y-axis). A strong correlation is evident (Spearman's $\rho = 0.96$).

4. Epistasis Analyses

4.1. Introduction

To arrive at specific estimates of motif activity, which would be needed for all downstream analyses, we searched for a realistic statistical modeling framework. In order to determine what types of statistical approaches would best allow quantitative modeling of regulatory function within the *cis*-elements, we conducted an analysis of genetic interactions among a suitable subset of the fine-scale mutants. Such genetic interactions might result from, for example, cooperative binding of transcription factors to a *cis*-element or the epistatic intersection of two parallel signaling pathways at a *cis*-element. 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 we chose had been successfully used in the quantification of interactions between gene deletions or amino acid substitutions (4), in regulatory network analysis (5), and in theoretical evolutionary and population genetics (6). Quantitative comparisons of the expression frequencies of each member of the set allows determination as to whether the individual mutations genetically interact.

4.2. Two Plausible Models

We examined the distribution of interaction terms under an additive model and a multiplicative model (7). 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. S5).

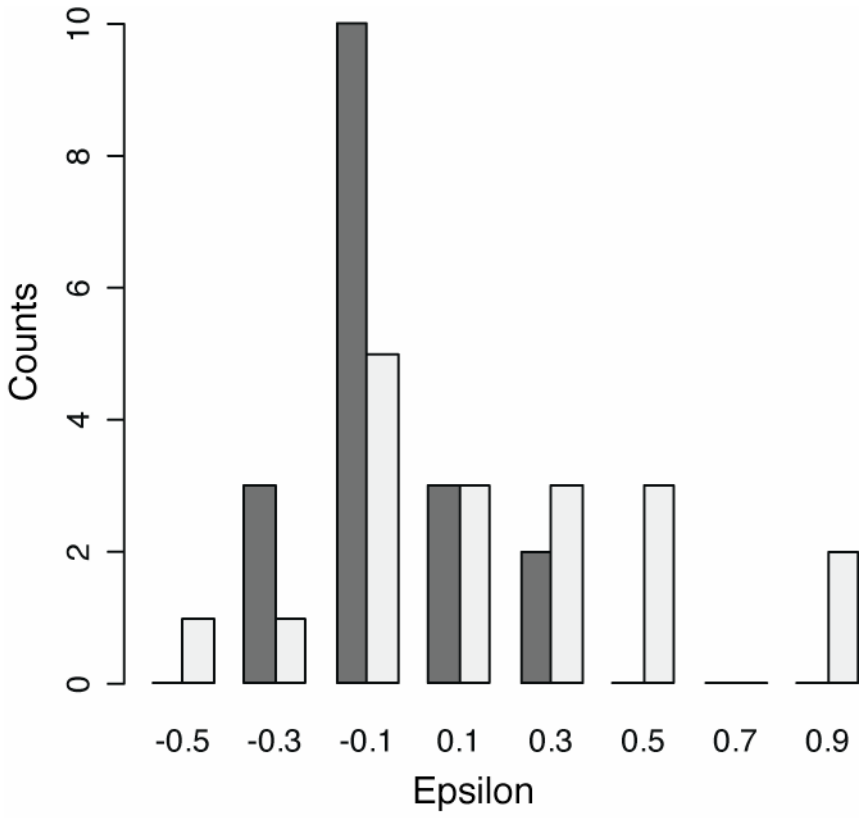


Fig. S5. Genetic interactions between regulatory motifs. Histogram of ε values for each of 18 possible comparisons. Interactions were calculated with additive (white bars) and multiplicative (grey bars) models of epistasis.

4.3. Conclusion

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.

5. Motif Substitution Experiments

We performed substitution experiments at three loci to test, independently from the epistasis analyses, whether genetic interactions among motifs are important. We reasoned that substituting one motif for another if specific interactions are required would not result in rescue. We 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 we selected loci that had the greatest chance of providing evidence for interactions.

Muscle-specific gene expression is indeed rescued when a different type of motif is inserted into a construct that had been rendered nonfunctional by a knockout of the endogenous motif. Every motif substitution we attempted resulted in rescue of muscle-specific expression. At the *C. intestinalis* CK locus, scrambling of the Tbx6 motif at position -268 results in a significant decrease in expression frequency. If the site is instead mutated to either its reverse complement or to a MyoD or CRE motif, muscle-specific expression is partially restored (SFig. 6A). More dramatically, when the Tbx6 motif at -108 in *C. intestinalis* AT2 is exchanged for a MyoD motif, expression is fully rescued (SFig. 6B). Similar results were also obtained with the Tbx6 motif at -89 in *C. savignyi* AT1. A requirement for motif-specific interactions appears therefore unlikely. These results also underscore that the three motif types are at least partially functionally equivalent, and that each motif transmits similar regulatory information to the transcriptional machinery.

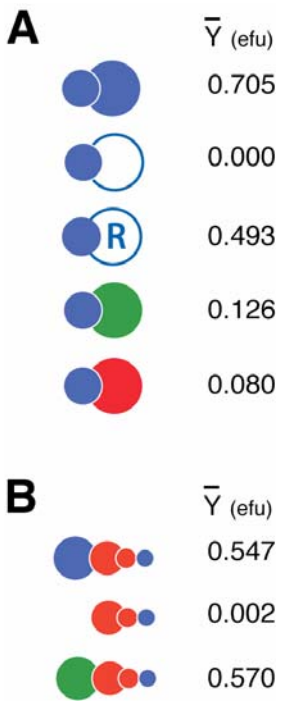


Fig. S6. Motif substitutions at the *cis*-elements of *C. intestinalis* CK (A) and AT2 (B). Color indicates motif type, with area proportional to activity as in Fig. 2B: red, CRE; green, MyoD; blue, Tbx6. Each row is a construct, with the endogenous arrangement at top and mutants below. Open circle is a scrambled sequence, "R" the reverse complement of the Tbx6 site. Mean muscle cell expression frequency is at right.

6. Regression Analyses

6.1. General Approach and Experimental Data

Inherent in our experimental design is the repeated testing of the functionality of individual motifs in multiple independent constructs. Because of this redundancy, the functional contribution of each motif could be estimated more accurately than with single data points. Thus, for each locus, we had between 6 and 30 distinct constructs for which expression frequency was measured, and which had particular combinations of motifs present in wild type form, or either deleted or mutagenized.

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.

All data analyses were conducted using R (8) and custom perl scripts. 14 outlier transfections, as identified by Dixon's test (9), were removed from the total of 1237 quantitatively assayed transfections. Multivariate regression models were constructed for each locus, for each homolog, independently. For simplicity, we 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 over-parameterization. All data analyses presented in the text are therefore based on the 175 constructs used to build the final models.

6.2. Four Types of Regression Models

We explored four different modeling scenarios, whose results are summarized in Table S3 and Fig. S7.

6.2.1. Additive model:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_i x_i \quad (1)$$

6.2.2. Angular transformation, additive model:

$$\arcsin\left(\sqrt{y}\right) = a + b_1 x_1 + b_2 x_2 + ... + b_i x_i \quad (2)$$

6.2.3. Multiplicative model:

$$(1 - y) = (1 - a)(1 - b_1 x_1)(1 - b_2 x_2)...(1 - b_i x_i) \quad (3)$$

which were log transformed and solved as linear models.

6.2.4. Logistic model:

$$y = \frac{e^{a + b_1 x_1 + b_2 x_2 + ... + b_i x_i}}{1 + e^{a + b_1 x_1 + b_2 x_2 + ... + b_i x_i}} \quad (4)$$

6.3. Comparison of Results from the Models

Logistic models (6.2.4.) 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. S7A). However, direct estimation of individual motif activity with logistic regression models is not transparent. Multiplicative models (7.2.3.) also seemed a reasonable choice given the genetic independence of the data under a multiplicative estimate of epistasis and the possibility of cooperative activation of cis-elements by clustered regulatory motifs. Such models, after logarithmic transformation, could be solved by simple linear regression. However, multiplicative models consistently explained less expression variation than additive models (Fig. S7 A, C-D). Models built from angular transformed expression frequencies were appealing because they removed some of the dependence of expression variance on the mean (Fig. S11B) and explained slightly more of the experimental variance than non-transformed additive models(Fig. S7C). In practice, all four model types performed quite well (Fig. S7; Table S3) and we therefore chose to focus on the simplest additive model (6.2.1.) due to its methodological transparency and the inherent interpretability of its measurement (muscle cell expression frequency).

Therefore all data presented in the main text are derived from non-transformed additive multivariate linear regression models. We 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. We considered motif activity to be statistically significant at $p < 0.05$.

Models 6.2.1-6.2.3 were built using the R lm function. Logistic models were built by maximum likelihood using the R glm function with a binomial error distribution.

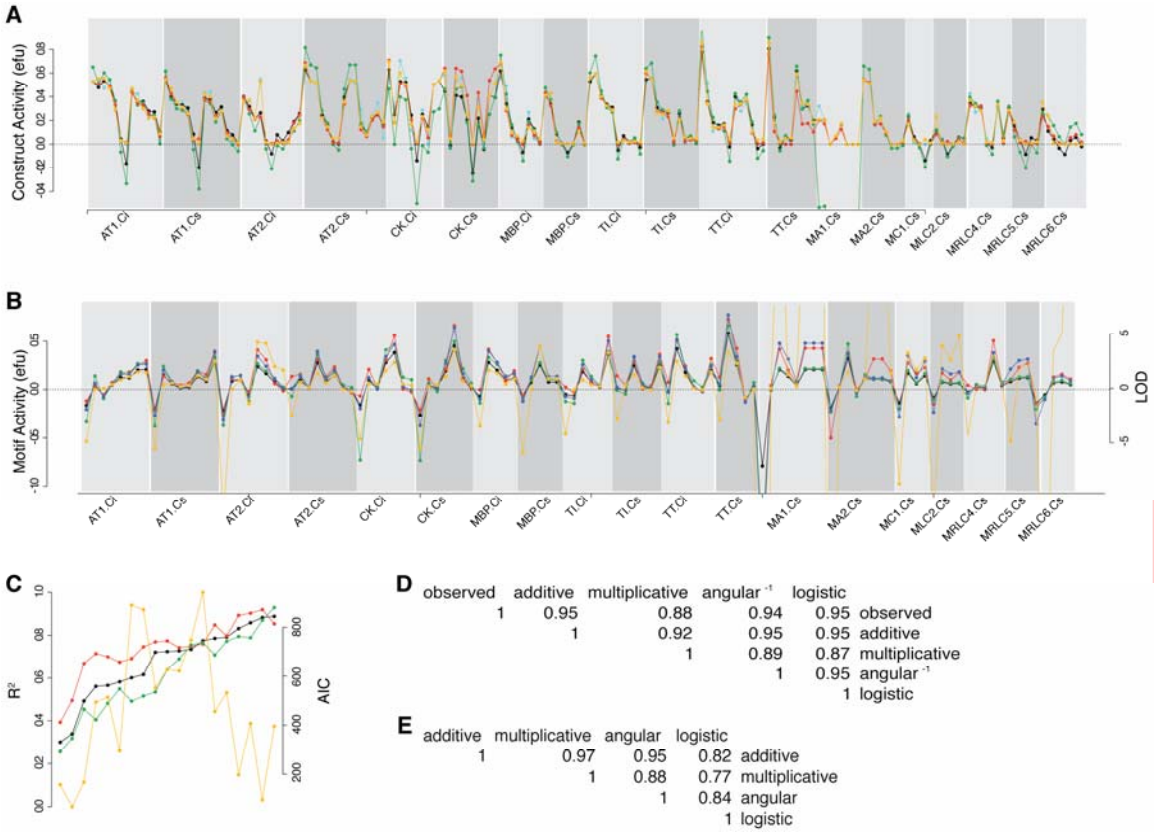


Fig. S7. Comparison of regression models. **(A)** Construct-by-construct comparison, for each gene, of mean observed activity (cyan) or activity predicted by four different regression models: additive (black), angular transformed/additive (red), multiplicative (green), or logistic (orange). Constructs sorted (along x-axis) by gene as in Table S2. Mean expression measurements or estimates in expression frequency units plotted along y-axis. **(B)** Gene-by-gene comparison, for each regression coefficient (including intercepts), as estimated by four different regression models. Black, red, green, and orange as in (A), blue depicting data resulting from an additive model in which data were first normalized by the strongest clone at a particular locus. Coefficients sorted (along x-axis) first by gene and subsequently by the order (5' to 3') of each explanatory variable (intercept estimates plotted first for each gene). **(C)** Model performance for each regression model type. Each point represents the regression estimated coefficient of multiple determination (R^2) or Akaike's Information Criterion (AIC) for a combination of regression model type and gene. R^2s or AIC for each gene plotted along the y-axis, sorted by gene along the x-axis by additive model R^2. Colors as in (A). **(D)** Pairwise correlation (Spearman's ρ), for each combination of observed or predicted functional measurements. **(E)** Pairwise correlation (Spearman's ρ), for each combination of models, of coefficient estimates.

7. Sequence Analyses

7.1. Alignments and Conservation of Homologous Sequences

All local alignments were constructed as reported previously (10). Scaffold-level alignments for each orthologous locus were collected from LBNL (http://pipeline.lbl.gov/data/Cioin2_cioSav2/).

To estimate the amount of identity in motif-like sequences anywhere in the genome, we 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. S8B).

To calculate sequence conservation at motif-adjacent positions we 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. S8B).

7.2. Motif Analyses

Initial position specific scoring matrices (PSSMs) (Fig. S2) were generated as follows. MyoD and CRE matrices were built from CisModule predictions (10) that were modified to be symmetrical because of their presumed palindromic nature. The Tbx6b/c matrix was built from *in vitro* binding data (11). All three PSSMs included 1% added pseudocounts. Motif predictions were calculated as LOD scores (12)

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

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 we 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; 13) there is no sequence-specific information beyond the border of the motif (Fig. S8C-E). Comparison of the motif portions of these logos with those of the initial logos reveals, as expected, close similarity of the PSSMs.

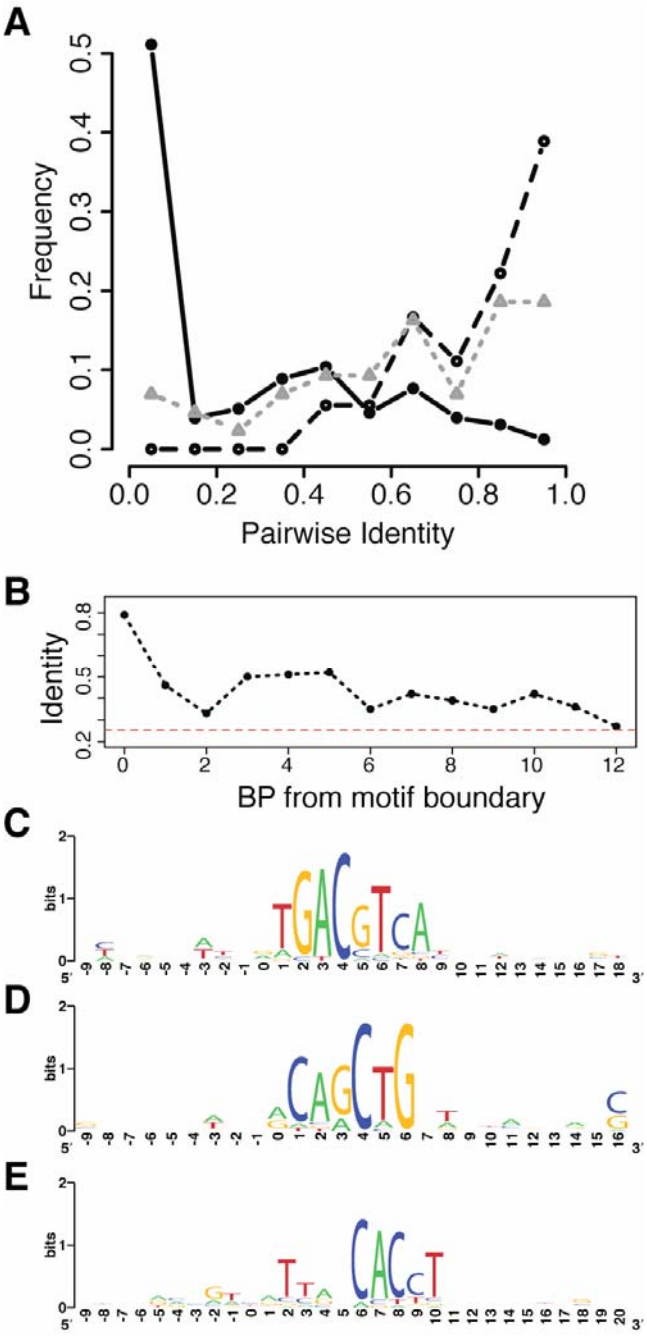
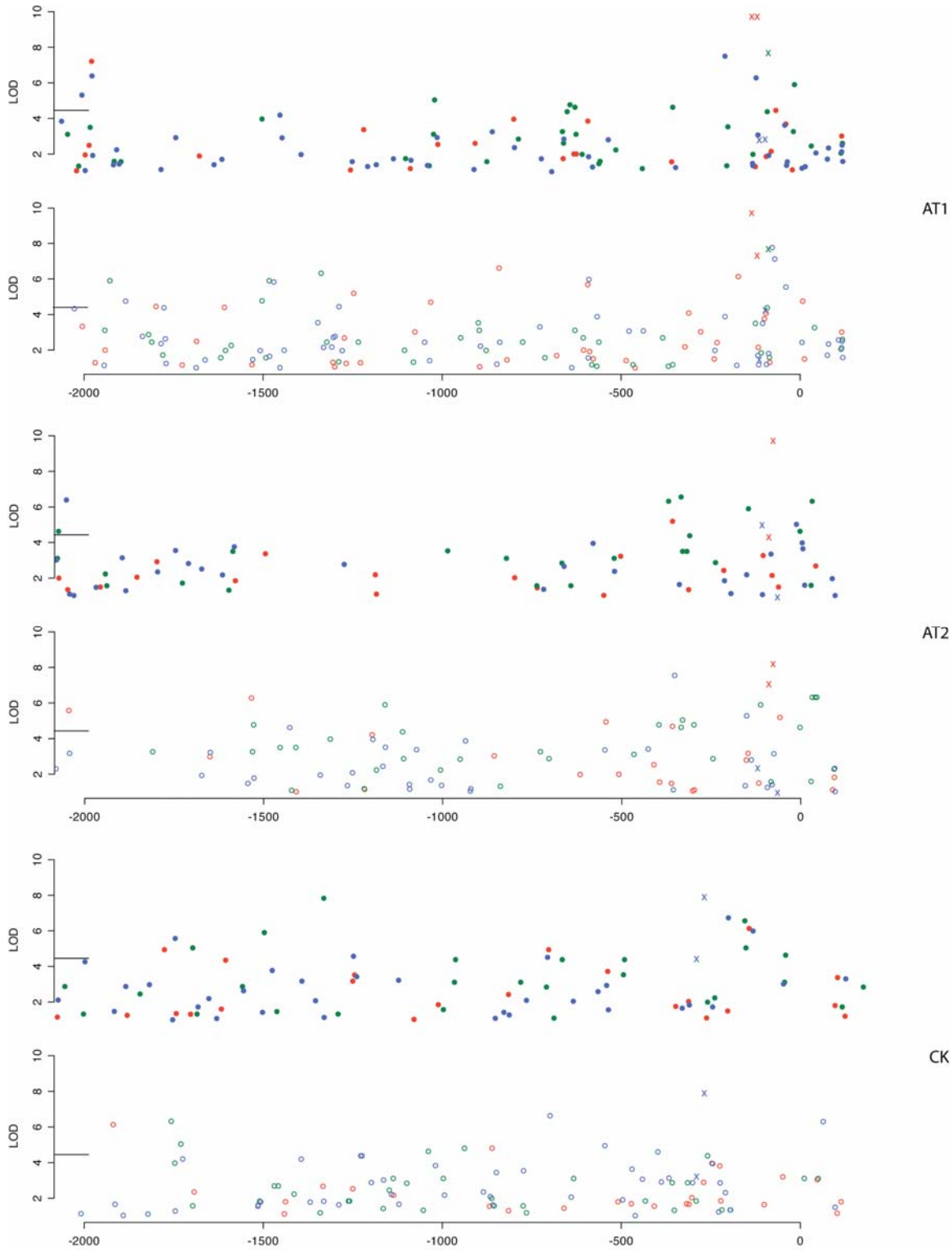


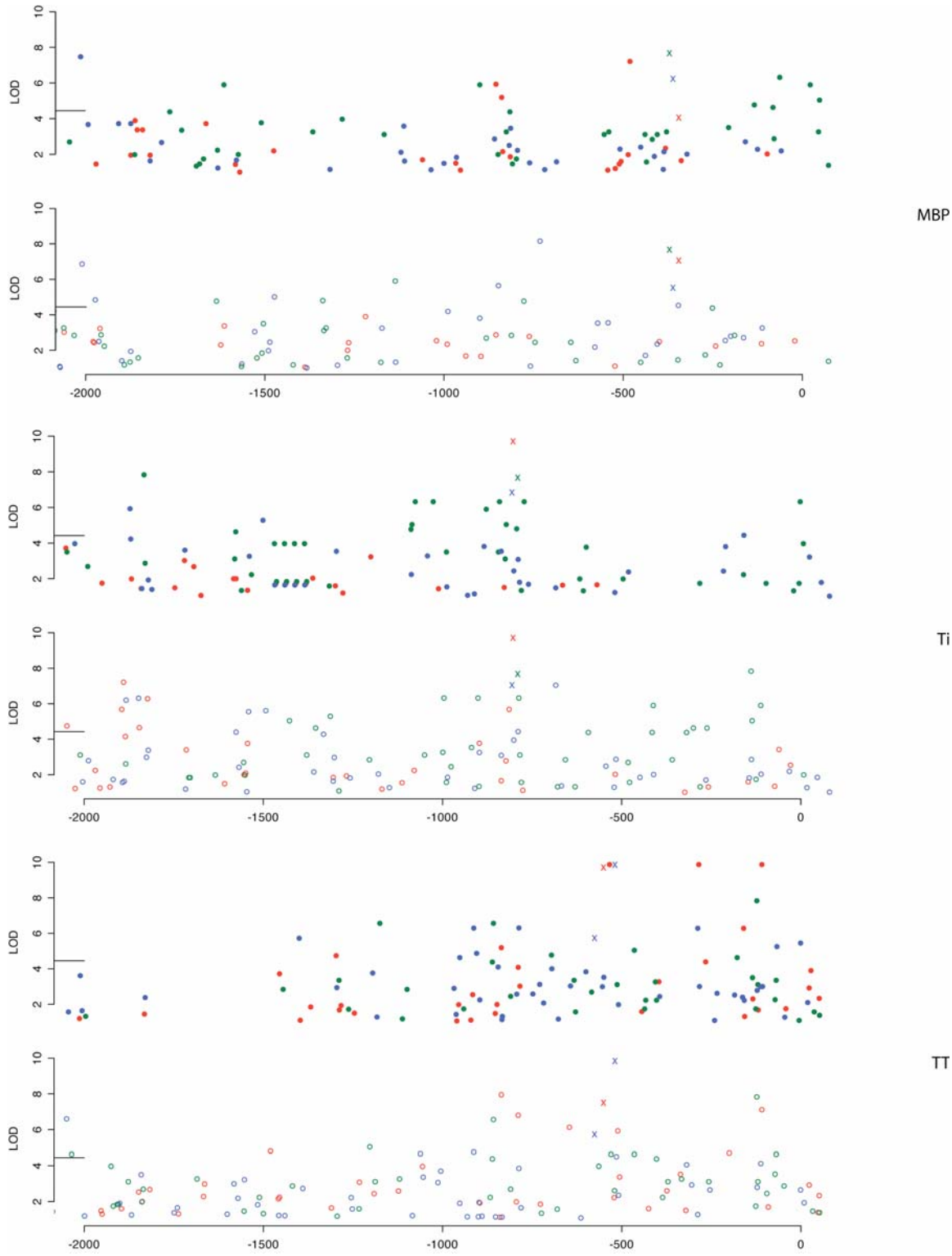
Fig. S8. Sequence conservation at regulatory motifs. **(A)** Histograms of the *C. savignyi*-*C. intestinalis* pairwise percent identity of samples from background genomic DNA (black, filed circles), motif predictions from a 500bp window spanning the functional module (gray, open triangles), and the 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. **(C-E)** 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: **(C)** CRE, **(D)** MyoD, and **(E)** Tbx6. Note the lack of significant sequence specificity outside the originally defined boundaries of each motif.

7.3. Parameters contributing to motif prediction

It should be noted that, due to the short length and degeneracy of the motifs examined in this study, most motif predictions do not contribute regulatory activity (Fig. S9). We have noticed three parameters that help to differentiate functional motifs from their false-positive counterparts. First, functional motifs are significantly more likely to be physically clustered, further solidifying the notion that transcription factor binding site clustering is a hallmark of metazoan *cis*-elements (14-16). Second, functional motifs are preferentially located near the transcription start site. Even within the 2-5kb initial reporter clones built for this study, all functional motifs identified here lie within 1100 bases, and most lie within 400 bases of the transcription start site. Third, the distribution of motif LOD scores (defined above) of functional motifs is significantly higher than a distribution built from false-positive motifs in the region. However, these results should be interpreted with caution, as we also note that within the group of functional motifs, there is no significant relationship between motif activity and : a) LOD score, b) distance from TSS, or c) motif spacing. Moreover, we see no relationship between any of the above mentioned parameters and non-linear deviations from our model-based activity estimates.

Fig. S9 (Spanning following two pages). Motif predictions at each single copy muscle gene. Each panel depicts an individual locus, arranged in orthologous pairs. Motif predictions are represented as closed circles (*C. intestinalis*), open circles (*C. savignyi*), or crosses (functional motifs in either species), colored based on motif type: red (CRE), green (MyoD), or blue (Tbx6). Motif predictions are plotted in two dimensions: aligned bases from the transcription start site are drawn along the x-axis and motif prediction LOD score (calculated as above) along the y-axis. The horizontal line through each plot at LOD=4.45 represents the lower 25^{th} percentile of LOD scores for functional motifs.


LOD
AT1
AT2
CK
-2000
-1500
-1000
-500
0


LOD
2
4
6
8
10
-2000
-1500
-1000
-500
0
MBP
Ti
TT

7.4. Estimating the Effect of Lower Resolution Functional Data

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, we mimicked lower-resolution data and examined 500 bp regions encompassing our functional motifs. Within these regions we 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, we collected alignment windows from the local alignment at predicted motif positions within a window of 500 bp encompassing the functional module. To minimize false positive motif predictions, we only assessed predictions with LOD scores > 4.45, representing the 25th percentile of true positive motifs. The distribution of resulting values is shifted significantly downward relative to the distribution built from functional motifs (SFig. 8A; 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.

7.5. Purifying Selection in the *C. savignyi* Population on Functional Motifs

To ask whether functional regulatory motifs have been subject to purifying selection in the *C. savignyi* population, we 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 (17). 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% (17). 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. S10). 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.

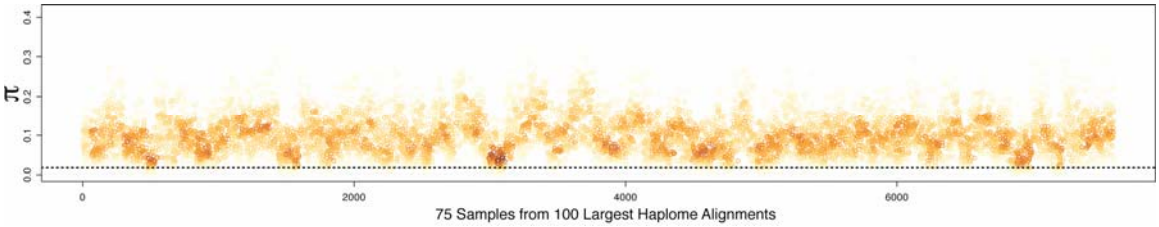


Fig. S10. Reduced polymorphism in functional motifs. Heterozygosity (y-axis) in 7500 samples (75 each from the 100 largest *C. savignyi* haplome alignments; 17), ordered by alignment (x-axis). Each circle represents a single sample of 13 mock motifs. Circles are shaded to highlight overlapping data, from highest (browns) to lowest (yellows) local point density (palette by colorbrewer.org). The heterozygosity within functional motifs is indicated by the dashed line.

8. Materials and Methods

Molecular Biology

Reporter constructs were built using standard PCR cloning techniques (10,18). 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. After the identification of large, functional reporter constructs, we conducted a deletion scan to define the cis-regulatory elements responsible for the majority of the transcriptional activity. Truncation constructs were generated by standard PCR cloning methods, while internal deletions were generated by overlap-extension PCR. We built several hundred such constructs, assayed quantitatively in over 2,000 transfections. Functional *cis*-elements were identified as sequences that, when deleted, resulted in a significant decrease in the expression probability of the reporter. These elements explained between 25% and 100% (mean 83%) of the function in the wild type construct.

We then refined the functional resolution of our analyses by conducting a high-resolution mutagenesis scan, guided in part by predictions of motif sequences. As described in the text, the three motifs utilized here are the Cyclic AMP Response Element (CRE; 10, 12, 19), the *Ciona* MyoD motif (10, 19-21), and the *Ciona* Tbx6 motif (11). All had been previously shown to be involved in muscle gene expression. The majority of the mutageneses carried out were directed at motifs of these 3 types.

In addition, mutageneses were carried out on sequences that did not match MyoD, CRE, or Tbx6. We probed the activity of multiple instantiations of other putative muscle motif types (Motif3 of (10); macho-1 of (22-23)), in addition to numerous sequences without a significant match to a candidate motif, none of which produced significant effects. MyoD, CRE, and Tbx6 are the only motifs for which we have evidence, and other motif types (if indeed present) are unlikely to contribute to function at a similar level of importance (though we cannot formally rule out contributions of other motifs). As noted above, at most loci, a small amount of activity is ascribed to regions outside the small motif clusters we dissected, but these are large and diffuse and we had no experimental power to detect motifs in those.

Site-directed mutagenesis of putative regulatory motifs was carried out 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 (by PCR cloning), and (2) Site-directed mutations that scrambled the sequence of a motif, while maintaining local GC content and spacing between adjacent sequences (by overlap extension PCR) (See Fig.1 for molecular dissection outline and Table S2 for locus-by-locus summary of all relevant constructs). All constructs were verified by 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

C. intestinalis were collected near San Diego, CA, USA by Marine Research and Educational Products. After shipment to Stanford, animals were kept in artificial seawater at 18C under constant light for at least 2 days. Fertilizations and dechorionations were conducted as reported previously (24). Transfections were conducted using a custom built electroporator (10, 25) set at 3000µF and 10ohms. Transfections were carried out in 0.4cm gap cuvettes containing 480µL 0.77M D-Mannitol, 20µL 5µg/µL DNA in TE, and 300µL embryos in artificial seawater. Embryos were reared at 16C for 13.5 to 15.5 hours. After development, batches of 50-200 embryos were transferred to microcentrifuge tubes in a minimal amount (~50-100µL) of seawater. They were then fixed in 300µL 2% paraformaldehyde in 500mM NaCl, 27mM KCl, 2mM EDTA for 30 minutes at room temperature (RT). Embryos were then washed twice in 500µL PBS+1%Triton X-100 for five minutes each at RT. Embryos were then washed once in staining buffer (PBS plus 1mM $MgCl_2$, 3mM $K_3Fe(CN)_6$, 3mM $K_4Fe(CN)_6$, 1% Triton X-100) for five minutes at RT. Embryos were then transferred to staining buffer + 1mM X-Gal and stained for exactly four hours at 37C. Embryos were then washed with PBS and stored at 4C until imaging. Each construct was transfected an average of 5 times, each producing on average ~75 scored embryos.

Embryos were transferred to 12 well tissue culture plates and photographed on a dissecting microscope. Each image was manually scored using the lightweight image scoring utility MARKER (http://mendel.stanford.edu/SidowLab/downloads.html). Each embryo was 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. We then calculated a weighted average, estimating the fraction of muscle cells stained for a given transfection.

To make use of a wide dynamic range of expression frequency for assaying the activity of the mutagenized constructs, we 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. S11).

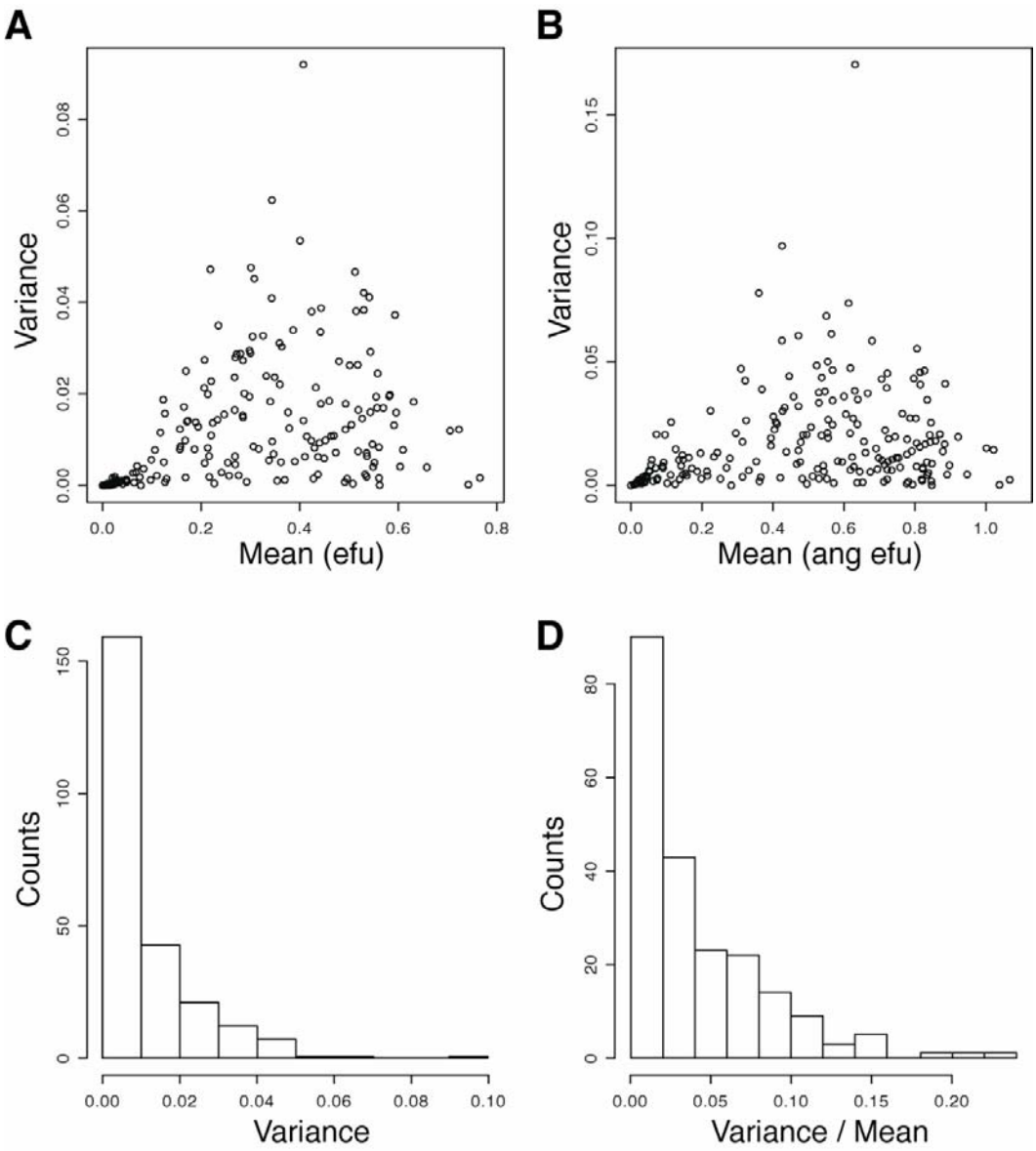


Fig. S11. Transfection data summaries. **(A)** Mean efu and variance of all replicated transfections analyzed in this study. **(B)** Mean efu and variance of angular transformed data. **(C)** Histogram of variance for all replicated transfections. **(D)** Histogram of variance as a function of mean for all replicated transfections.

Supplemental References

1. N. Friedman, M. Ninio, I. Pe'er, T. Pupko, *J. Comput. Biol.* **9**, 331 (2002).
2. Oda-Ishii et al., *Development* **132**, 1663 (2005).
3. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (CSHL Press, Cold Spring Harbor, NY, ed. 2, 1989), pp. 7.16-7.17.
4. H. Y. Tong, *et al.*, *Science* **303**, 808 (2004).
5. D. Segre, A. DeLuna, G. M. Church, R. Kishony, *Nat. Genet.* **37**,77 (2005).
6. S. F. Elena, R. E. Lenski, *Nature* **390**, 195 (1997).
7. H. J. Cordell, *Hum. Mol. Genet*. **11**, 2463.
8. 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.
9. R. R. Sokal, F. J. Rohlf, *Biometry* (W.H. Freeman and Co., New York, NY, 1995).
10. D. S. Johnson, *et al*., *Genome Res*. **15**, 1315 (2005).
11. K. Yagi, N. Takatori, Y. Satou, N. Satoh, *Dev. Biol.* **282**, 535 (2005).
12. D. S. Johnson, B. Davidson, C. D. Brown, W. C. Smith, A. Sidow, *Genome Res*. **14**, 2448 (2004).
13. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res*. **14**, 1188 (2004).
14. E. M. Crowley, K. Roeder, M. Bina, *J. Mol. Biol.* **268**, 8, (1997).
15. W. W. Wasserman, J. W. Fickett, *J. Mol. Biol.* **278**, 167, (1998).
16. B. Berman, et al., *Proc. Natl. Acad. Sci. U.S.A*. **99**, (2002).
17. K. S. Small, M. Brudno, M. M. Hill, A. Sidow, *Proc. Natl. Acad. Sci. U.S.A*. **103**, 5698, (2007).
18. J. Sambrook, D. W. Russell, *Molecular Cloning: A Laboratory Manual* (CSHL Press, Cold Spring Harbor, NY, ed. 3, 2001).
19. T. Kusakabe, R. Yoshida, Y. Ikeda, M. Tsudda, *Dev. Biol.* **276**, 563 (2004).
20. T. K. Blackwell, H. Weintraub, *Science* **250**, 1104 (1990).
21. T. H. Meedel, P. Chang, H. Yasou, *Dev. Biol.* **302**, 333 (2006).
22. H. Nishida, K, Sawada, *Nature* **409**, 724 (2001).
23. K.Yagi, N. Satoh, Y. Satou, *Dev. Biol.* **274**, 478 (2004).
24. R. W. Zeller, *Methods Cell Biol.* **74**, 713 (2004).
25. R. W. Zeller, M. J. Virata, A. C. Cone, *Dev. Dyn.* **235**, 1921 (2006).