Noncoding regulatory sequences of *Ciona* exhibit strong correspondence between evolutionary constraint and functional importance

David S. Johnson,[1] Brad Davidson,[2] Christopher D. Brown,[1] William C. Smith,[3] and Arend Sidow[1,4]

[1]*Department of Pathology and Department of Genetics, Stanford University Medical Center, Stanford, California 94305-5324, USA;* [2]*Department of Molecular and Cell Biology, Division of Genetics and Development, Center for Integrative Genomics, University of California–Berkeley, California 94720, USA;* [3]*Molecular Cellular and Developmental Biology Department, Neuroscience Research Institute, University of California–Santa Barbara, California 93106, USA*

We show that sequence comparisons at different levels of resolution can efficiently guide functional analyses of regulatory regions in the ascidians *Ciona savignyi* and *Ciona intestinalis.* Sequence alignments of several tissue-specific genes guided discovery of minimal regulatory regions that are active in whole-embryo reporter assays. Using the *Troponin I* (*TnI*) locus as a case study, we show that more refined local sequence analyses can then be used to reveal functional substructure within a regulatory region. A high-resolution saturation mutagenesis in conjunction with comparative sequence analyses defined essential sequence elements within the *TnI* regulatory region. Finally, we found a significant, quantitative relationship between function and sequence divergence of noncoding functional elements. This work demonstrates the power of comparative sequence analysis between the two *Ciona* species for guiding gene regulatory experiments.

[Supplemental material is available online at www.genome.org and http://mendel.stanford.edu/supplementarydata/johnson.]

Ascidians, simple basal chordates popularly known as sea squirts, offer several conventional benefits as experimental model organisms. First, they share with higher vertebrates a conserved program of early embryogenesis that culminates in a larva whose body plan is remarkably similar to that of vertebrates (Satoh 1994). Second, ascidians have small genomes (~180 Mb) with several expressed genes (15,000), similar to that of *Drosophila* (Adams et al. 2000). Third, ascidians have attractive embryological and experimental features that facilitate studies of gene regulation during development, including production of large quantities of embryos, development from egg to larva in 18 h, an invariant cell lineage (Nishida 1987), and the ability of embryos to express genes from reporter constructs transfected in bulk by electroporation (Corbo et al. 1997). Fourth, informational resources that aid computational and experimental analyses were recently generated, the most important of which are the draft genome sequences and assemblies for *Ciona savignyi* (http://www.broad.mit.edu/annotation/ciona) and *Ciona intestinalis* (Dehal et al. 2002). Extensive EST sequence collections and large-scale in situ hybridization data for *C. intestinalis* (Satou et al. 2002) and *Halocynthia roretzi* (Makabe et al. 2001) are also available. These benefits make ascidians exceptionally useful model organism for studies of gene regulation.

We show that the true power of ascidians for gene regulatory studies lies in the relationship between *C. savignyi* and *C. intestinalis*. The distance between the *Ciona* species is so large that unconstrained sequences do not display more similarity than expected by chance, as analysis of 18S rRNA sequences shows that the pairwise divergence of the two *Cionas* is slightly less than that between human and frog and slightly greater than that between human and chick (data not shown). Despite the extensive sequence divergence, the two species are virtually identical in their embryology up to the tadpole stage. In fact, hybrids can be obtained by fertilizing either species' dechorionated eggs with sperm from the other, and are easily reared to the tadpole stage (Byrd and Lambert 2000). This suggests that the essential mechanisms of early development are partially if not fully conserved between the two species. Owing to this relationship, which is currently unique among sequenced metazoans, comparative sequence analyses between the *Ciona* species facilitate discovery of functional sequence elements with remarkable resolution and without knowledge of previously identified motifs.

Here we use comparative sequence analysis between *C. savignyi* and *C. intestinalis* to locate potentially important noncoding sequence elements in several genomic loci. Large-scale global alignments give a bird's-eye view of each locus, and detect longer regions (>100 bp) of high average conservation between *C. savignyi* and *C. intestinalis*. Using these alignments as a guide, we isolated functional promoters for four *C. savignyi* genes, expressed in a variety of embryonic tissues, and show that

[4]**Corresponding author.**
E-mail arend@stanford.edu; fax (650) 725-4905.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2964504. Article published online before print in November 2004.

14:2448–2456 ©2004 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/04; www.genome.org

functional 5′-regions of four previously characterized genes also contain extensive conservation. We then conducted a high-resolution functional analysis of *Troponin I* (*TnI*) (MacLean et al. 1997; Cleto et al. 2003), in which shorter regions of high local similarity in the *TnI* locus guided identification of the minimally sufficient regulatory region (MSRR). The MSRR was then subjected to a saturation mutagenesis. We find that necessary regulatory sequences, including certain predicted myf-binding sites, are highly conserved between *C. savignyi* and *C. intestinalis*, that these sequences are functionally identical between the two species, and that there is a quantitatively robust correlation between sequence conservation and functional importance. We conclude that *Ciona* sequence comparisons can provide remarkable resolution in predicting functional elements.

Results

C. intestinalis and *C. savignyi* alignments reveal conserved elements contained within functional regulatory regions in eight loci

We examined four diverse genes for functional sequence conservation between the two *Ciona* species. We specifically chose genes that are expressed in a variety of tissues and perform a variety of functions in the developing ascidian embryo to demonstrate that our analyses are widely applicable (Fig. 1a–d). Coding exons tend to stand out as peaks of high similarity. Untranslated regions (UTRs), introns, and intergenic regions rarely contain conservation beyond what is expected by chance. Against this background of extremely low noncoding sequence similarity, the intergenic regions of these loci contain short, significantly conserved sequences 5′ to the predicted start of transcription (Fig. 1a–d). Our reporter constructs that contained these sequences faithfully recapitulated endogenous expression:

- *Troponin I* is a highly conserved component of the muscle strand, and transcripts are strongly expressed in larval tail muscle (Cleto et al. 2003). Our TnI construct, which includes noncoding regions of high sequence identity that approach 80%, is expressed strongly in the tail muscle (Fig. 1i). This construct is 1793 bp long, and includes four exons that together account for 12% of the bases.
- *Synaptotagmin* (*Syt*) transcripts are expressed in the neurectodermal lineages of the epidermis, brain, and nerve cord (Cluster 01098; http://ghost.zool.kyoto-u.ac.jp/indexr1.html). Accordingly, the Syt fusion construct is expressed in both epidermal neurons (punctate staining in the tail) (Fig. 1j) and neurons in the spinal cord (Fig. 1k). This construct includes an extensive region of similarity immediately adjacent to the 5′-UTR.

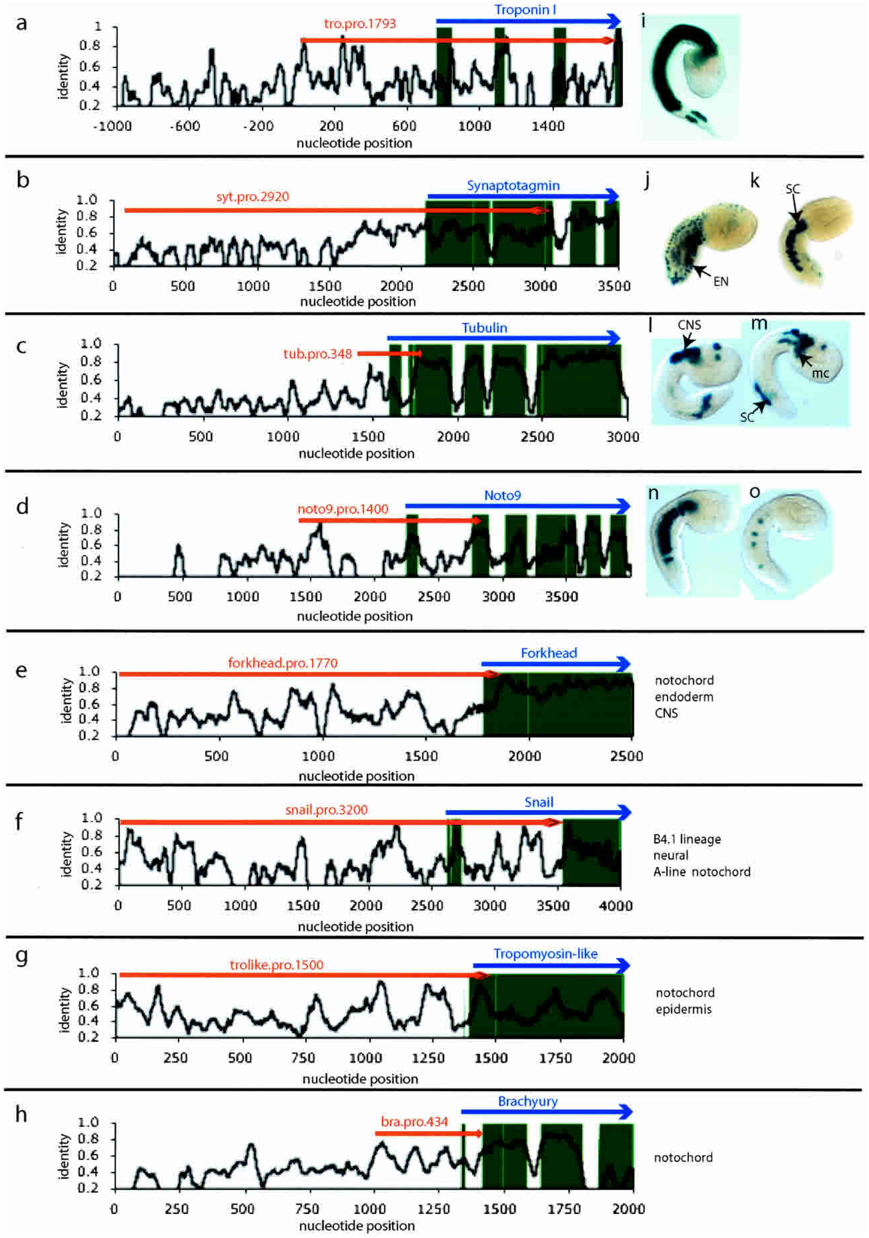


Figure 1. Five prime regions of eight *Ciona* loci. The black plots represent sequence identity between *C. intestinalis* and *C. savignyi* across each region. *C. savignyi* is the reference sequence for *a–d*, and *C. intestinalis* is the reference sequence for *e–h*. Green vertical bars correspond to annotated *C. intestinalis* exons. Blue arrows indicate direction of transcription of each gene. The red arrows indicate the native sequence present in the promoter fusion construct. Each promoter fusion construct contains significant regions of similarity between the two species. Expression patterns for *C. savignyi* constructs (*i–o*). Troponin I is strongly expressed in the tail muscle (*i*), Synaptotagmin is expressed in various neural lineages, including epidermal neurons (EN) (*j*) and spinal cord (SC) (*k*), α-tubulin is expressed in various tissues, including central nervous system (CNS) (*l*) and spinal cord (SC) (*m*), and ectopic expression frequently occurs in the mesenchyme (MC) (*m*), Noto9 is expressed specifically in the notochord (*n*,*o*). All embryos were fixed at the mid–late tailbud stages. Electroporated embryos always exhibit mosaic staining patterns. A variety of typical images are archived at http://mendel.stanford.edu/supplementarydata/johnson.

- α-*Tubulin* transcripts are strongly expressed in the trunk, nerve cord, brain, and epidermis of tadpole ascidians (Cluster 00271; http://ghost.zool.kyoto-u.ac.jp/indexr1.html). Our α-*tubulin* construct recapitulates this expression pattern (Fig. 1l,m). Staining is particularly strong in the hindbrain of the central nervous system (Fig. 1l). Ectopic staining occurs frequently in the mesenchyme (Fig. 1m).
- The *Noto9* gene, originally isolated in a screen for genes regulated by *Ci-Brachyury*, is expressed specifically in the notochord (Takahashi et al. 1999). Our *Noto9* construct (Fig. 1d) includes one major peak of conservation and drives strong expression specifically in both primary and secondary notochord (Fig. 1n,o).

The 5′-regions of the four previously characterized genes also contain regions of strong similarity (Fig. 1e–h). The activities of a variety of deletion constructs of these genes, generated without knowledge of the sequence comparison, are also consistent with a general correspondence between function and conservation (Corbo et al. 1997; Erives et al. 1998; Di Gregorio and Levine 1999; Di Gregorio et al. 2001; data not shown).

We also demonstrated that constructs are reciprocally functional in both species. The Cs-TnI, Cs-Syt, Cs-Tub, and Cs-Noto9 constructs all give conserved expression when transfected into *C. intestinalis* (data not shown). Furthermore, we tested the *Ci-Brachyury* (Corbo et al. 1997) and *Ci-Forkhead* (Di Gregorio et al. 2001) promoters, which give conserved expression when transfected into *C. savignyi*. Thus, the structure and function of all six genes we tested are conserved between the two *Ciona* species.

Troponin I as a case study for the sequence-guided dissection of regulation and expression in *Ciona*

We conducted a more detailed functional and computational analysis for one of these genes, *TnI*, to determine the resolution of sequence comparisons in gene regulatory studies in *Ciona*. *TnI* was chosen because a full-length cDNA and detailed annotation of the genomic locus have been reported previously (Maclean et al. 1997; Chiba et al. 2003; Cleto et al. 2003), and because its strong expression in tadpole muscle permits semiquantitative analysis of variants from wild type.

We compared in situ hybridizations to the lacZ reporter activity (Fig. 2) to ensure that the reporter construct tn.pro.1793 contains the regulatory elements for expression of *TnI* in the ascidian embryo. *TnI* transcript is first evident in the late gastrula (Fig. 2a) and strengthens in the neurula (Fig. 2b). From the tailbud stage to the fully developed tadpole stage, *TnI* is expressed in the tail muscle (Fig. 2b–e). Both primary and secondary muscle lineages express *TnI* transcripts (Fig. 2c–e). Heart precursor cells, the trunk ventral cells (TVCs), also express *TnI* (Fig. 2d, arrowhead). After settling, staining persists in the juvenile heart and body wall muscle (Fig. 2f–h). In summary, *TnI* transcripts are found in all adult and larval muscle cells and their precursors, starting in early zygotic development. Temporally and spatially, tn.pro.1793 mimics the endogenous expression of *TnI*. It expresses strongly in the tail muscle (Fig. 2i–m) and the TVCs (Fig. 2m), from the early tailbud until the fully developed tadpole. Ectopic mesenchymal expression, which is common after electroporation of *Ciona* expression constructs (Corbo et al 1997; Harafuji et al. 2002), is also seen occasionally (cf. Fig. 2i, arrowhead). Thus, expression of tn.pro.1793 faithfully recapitulates expression of the *TnI* transcript.

Both function and functional noncoding sequences are conserved between the two *Ciona* species

We also designed a construct equivalent to tn.pro.1793 from the orthologous region in the *C. intestinalis* genome. Since the *C. intestinalis* transcript was well characterized (Cleto et al. 2003), we were able to design a shorter fusion construct. As expected, this construct demonstrates the same pattern of staining for lacZ as seen for the Cs-TnI construct (data not shown; cf. Fig. 2i–m). In addition, some embryos electroporated with tn.ci.pro.921 were reared to the juvenile stage to verify that the reporter also faithfully reproduces TnI expression in the heart and body wall muscle (Fig. 2n). Furthermore, there are no detectable differences in expression upon electroporation of either construct into either species.

The conservation of function encoded on both tn.ci.pro.921 and tn.pro.1793 indicates the presence of conserved regulatory sequences necessary and sufficient for expression of TnI. We reasoned that local regions of high sequence conservation between the two species might contain such regulatory sequences. Indeed, from ~400 bp to 800 bp upstream of first known exon of TnI, there are three peaks of conservation (starting at position 0, Fig. 1A) that clearly stand out against the low amount of similarity in the remainder of the noncoding sequences, and appear to be as highly conserved as exons that code for functionally

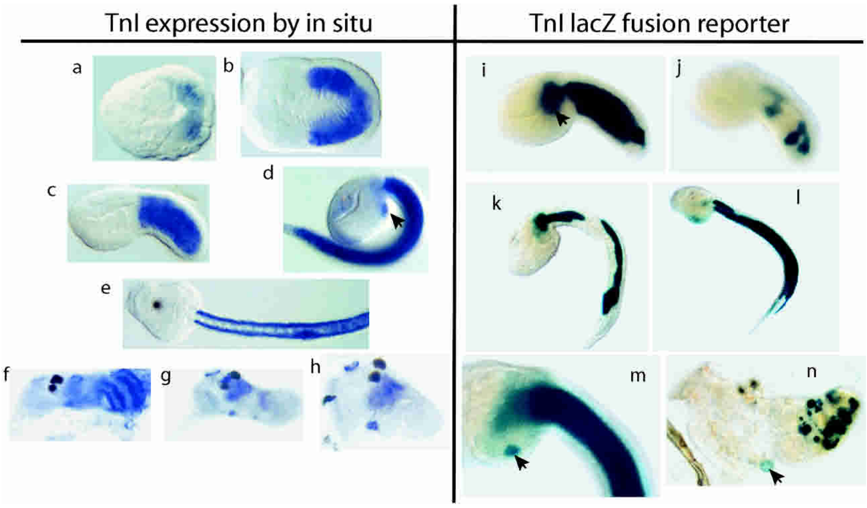


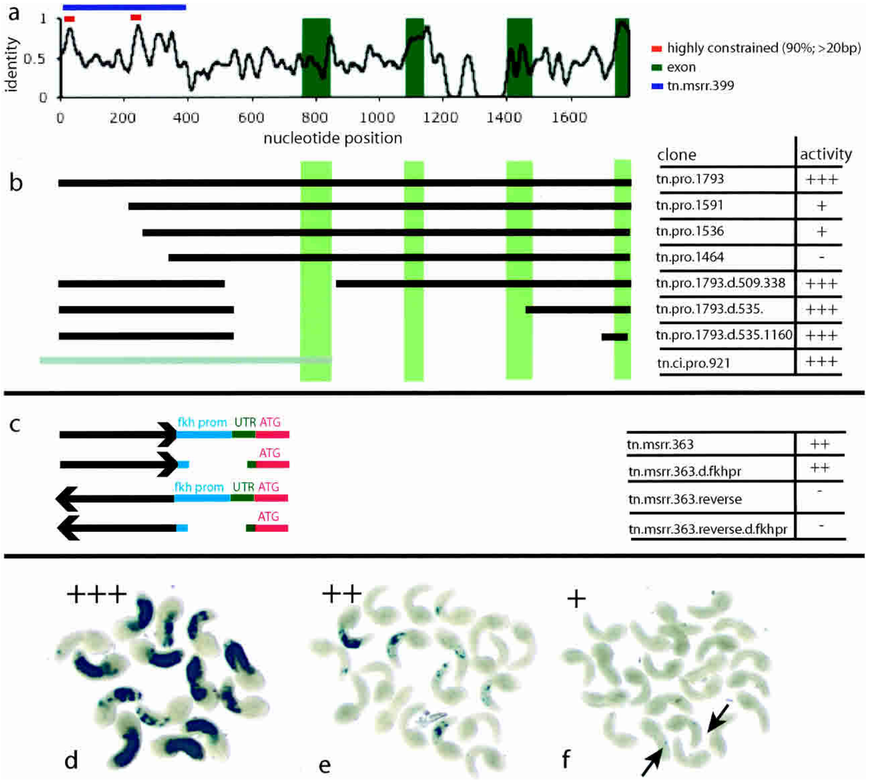
Figure 2. *Ciona* Troponin I gene expression. (*a–h*) Expression of Troponin I transcript as determined by in situ hybridization. Expression is weak as early as gastrulation (*a*), is very strong by initial tailbud (*b*), and continues in the muscle through mid-tailbud (*c*), late tailbud (*d*), tadpole (*e*), and juvenile stages (*f–h*). Note the expression in the trunk ventral cells (TVC; adult heart precursors), during the late tailbud stage in *e*. (*i–n*) Expression of lacZ reporter construct tn.pro.1793 as determined by X-gal staining. Expression is strong in mid tailbud (*i,j*), and continues through late tailbud (*k*), early tadpole (*l*), and juvenile stages (*m*). Ectopic expression in the mesenchyme is common (arrowhead, *i*). Expression continues in the tail muscle from the late tailbud (*k*) to the early tadpole (*l*). Expression also occurs in the TVC (arrowhead, *m*) and juvenile heart (arrowhead, *n*).

important regions of the protein. The most strongly conserved subregions demonstrate >90% nucleotide identity over >20 bp (Fig. 3a). As no other presumably noncoding region in the entire locus appears as conserved as this (cf. Fig. 1a), we next tested whether it was sufficient and necessary for proper expression of TnI.

Identification of the minimally sufficient regulatory region of *TnI*

We conducted a deletion analysis of tn.pro.1793 in order to identify approximate locations of regulatory regions (Fig. 3b). Truncating 202 bp from the 5′-end of tn.pro.1793, resulting in tn. pro.1591, significantly decreases expression, as does a 257-bp truncation (tn.pro.1536). Removal of 329 bp to make tn. pro.1464 completely abolishes expression. This demonstrates that the conserved region of ~400 bp is necessary for expression in larval muscle.

We next determined which portions of tn.pro.1793 are sufficient for expression, using a semiquantitative classification of staining patterns (Fig. 3d–f). Internal deletions that remove any of the first three exons of *TnI* have no effect on expression whatsoever (Fig. 3b), as shown by tn.pro.1793.d.509.338, tn.pro.1793.d.535.861, and tn.pro.1793.d.535.1160. This suggested the presence of a promoter- and tissue-specific regulatory regions on tn.pro.1793.d.535.1160. To assuage the concern that an artifactual promoter resided within the native *TnI* sequence of tn.pro.1793.d.535.1160, we switched to a reporter plasmid (pCES) (Harafuji et al. 2002) that does not require *TnI* exonic sequence to produce functional LacZ protein. An insert of only 363 bp that contains all the highly conserved noncoding sequences present on tn.pro.1793 is sufficient for strong muscle-specific expression in the embryo (tn.MSRR) (Fig. 3c). Further experiments demonstrated that activity of tn.MSRR is not dependent on the Forkhead promoter present on the plasmid (Fig. 3c). Additionally, tn.MSRR subcloned in the reverse orientation, both with and without the Forkhead promoter, fails to show expression. We conclude that the MSRR contains a promoter- as well as tissue-specific regulatory sequences sufficient to convey proper spatiotemporal expression.



clone	activity
tn.pro.1793	+++
tn.pro.1591	+
tn.pro.1536	+
tn.pro.1464	-
tn.pro.1793.d.509.338	+++
tn.pro.1793.d.535.	+++
tn.pro.1793.d.535.1160	+++
tn.ci.pro.921	+++

tn.msrr.363	++
tn.msrr.363.d.fkhpr	++
tn.msrr.363.reverse	-
tn.msrr.363.reverse.d.fkhpr	-

Figure 3. Activity of Troponin I–lacZ promoter constructs bearing various deletions, in context with sequence identity and exon structure of the locus. (*a*) Plot of sequence identity between *C. savignyi* and *C. intestinalis* in the promoter construct pro.1793 (cf. Fig 1a). Green shading represents exons, the blue bar represents the MSRR, and the red bars indicate highly constrained regions (90% id; >20 bp). (*b*) Activity of deletion constructs, with tn.pro.1793; black lines denote sequence present in the constructs. The table lists activities of the constructs as explained below. Truncation of 329 bp from the 5′-end of the tn.pro.1793 construct completely eliminates expression in the tail muscle. Deletion of internal regions, including annotated exon 1, exons 1 and 2, and exons 1–3, has no effect. The gray line (tn.ci.pro.921) denotes the construct derived from *C. intestinalis*. (*c*) Activity of minimally sufficient regulatory region (MSRR) constructs. Fusion of the 5′-most 363 bp of tn.pro.1793 to a heterologous Forkhead basal promoter has strong activity. Deletion of the Forkhead basal promoter has no effect on lacZ expression. The reversed insert has no activity. (*d–f*) Relative activities of constructs. We classify constructs as (*d*) +++, representing 50%–100% animals staining, with many of the animals staining strongly in a majority of the tail muscle cells; (*e*) ++, representing 25%–50% of the animals staining in the tail muscle, and the minority of these staining a majority of the tail muscle cells; and (*f*) +, with <25% of the animals staining, and none of the animals staining the majority of the tail muscle cells; "−" indicates a construct that never showed staining. Constructs that were mostly "−" but stained weakly on a single occasion are denoted +/−.

Functional substructure in the TnI MSRR correlates with evolutionary conservation

Having identified the MSRR, we wanted to find smaller regions within it that are necessary for tissue-specific expression. We first asked whether it contained any conservation that was not found by the global alignment, such as short sequences that had been rearranged since the last common ancestor of *C. intestinalis* and *C. savignyi*, by applying the CHAOS algorithm (Brudno et al. 2003a) to compare the Cs-TnI MSRR against the entire *TnI* locus of *C. intestinalis*. The four top-scoring CHAOS alignments, which we will call CH1-CH4, all fell within the MSRR in conserved order, spacing, and orientation (Fig. 4a).

Using tn.msrrr.399 as the base plasmid for altering the sequence of the MSRR, we tested a variety of truncation and deletion constructs to determine the regions in the MSRR that are necessary for expression in the tail muscle (Fig. 4b). CH1–CH4 were used as a guide for the design of these constructs. The sequences contained in CH1 and CH2 are necessary for activity, as their respective deletion constructs give expression only extremely rarely. In contrast, deletions of either CH3 or CH4 have no effect on activity, but a deletion of both (tn.msrr.301) abolishes function of the MSRR, suggesting that CH3 and CH4 are

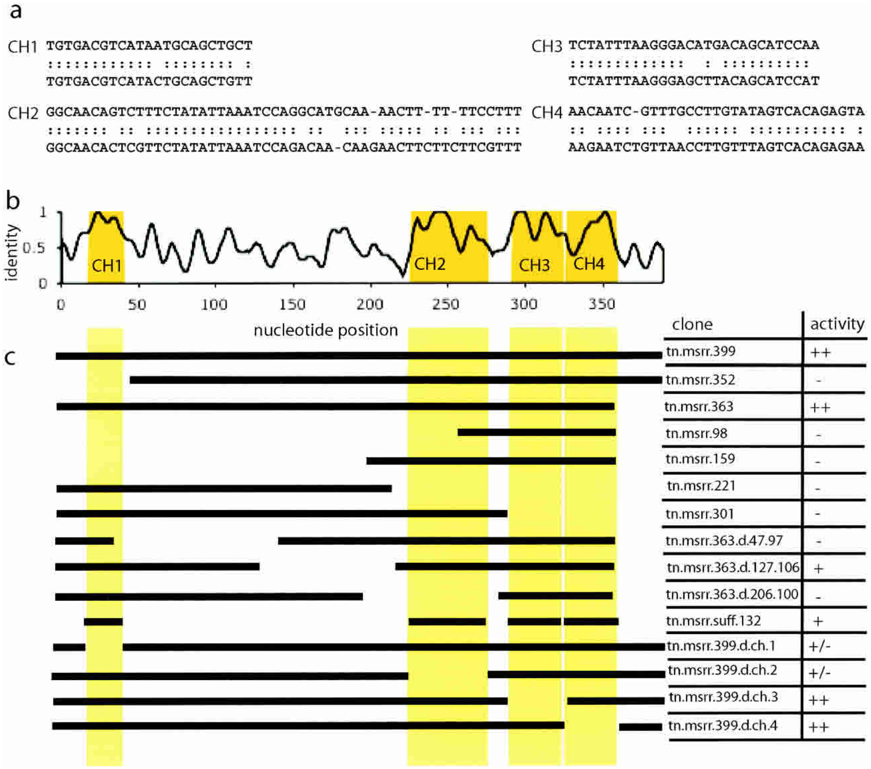


Figure 4. Deletion analysis of the minimally sufficient regulatory region. (*a*) The CHAOS algorithm returns four major regions of local conservation between *C. savignyi* and *C. intestinalis.* Note that this level of conservation stands out over neutrally evolving regions. (*b*) Plot of sequence identity between the *C. savignyi* and *C. intestinalis* TnI enhancer regions. The yellow shaded boxes represent the four CHAOS matches. (*c*) Horizontal lines denote sequence present in the constructs, activity is noted as in Figure 3. CHAOS matches CH1, CH2, and either of CH3/CH4 are necessary for expression of lacZ reporter in larval tail muscle. Concatenation of the four CHAOS matches is sufficient for weak lacZ expression.

redundant. We conclude that the sequences corresponding to the first two CHAOS matches and either CH3 or CH4 are necessary. In addition, the weak activity of construct tn.msrr.363.d.127.106 suggests that there is one small region with little sequence conservation that is necessary for full expression. Concatenation of the four CHAOS matches into a single insert (tn.msrr.suff.132) drove weak expression in the larval tail muscle (Fig. 4c). We conclude that sequence information sufficient to drive expression in the tail muscle is contained within the four short regions of significant similarity.

These results qualitatively demonstrate that identification of conserved regions between Cs-TnI and Ci-TnI by global and local alignment can be useful for guiding functional analyses. To test, at high resolution, for a statistically significant relationship between conservation and function, we conducted a scanning mutagenesis of the MSRR as follows. We selected a tiling path of 20 20-bp windows, each overlapping the next one by 5 bp, across the region. For each window, we built one construct in which the 20 bp of the window was randomized, but the rest of the MSRR was kept intact. This preserved the G/C content within the window but produced a new sequence with minimal sequence identity to the sequence in the wild-type window. We then tested the lacZ activity of each construct. We observed a general trend of windows with high sequence identity between the two species showing lower activity upon mutagenesis (Fig. 5b). To test whether this trend was significant, we conducted a simple statistical analysis. In this analysis, we did not include windows in which the distance between the *C. savignyi* and *C. intestinalis* sequence could not be calculated (because alignment accuracy, and therefore base pair level homology, is questionable in those windows), nor did we include windows that did not significantly reduce function (because our assay is not comprehensive with respect to all functions). Regression on the remaining 12 data points demonstrates a significant correlation between conservation and activity upon mutagenesis ($R^2 = 0.4641$; $P < 0.015$) (Fig. 5c).

Specific sequence features necessary for function of the Troponin I MSRR

Finally, we analyzed the *TnI* locus for the presence of muscle-specific transcription factor binding sites. Because few known binding sites exist for *Ciona*, we used position-specific sequence matrix (PSSM) data derived from known binding sites in the regulatory regions of vertebrate muscle-specific genes (Wasserman and Fickett 1998). These PSSMs were from the Myf family, myocyte-specific enhancer factor 2 (MEF2), serum response factor (SRF), and E-box (Fig. 5a). The Ghost EST database (http://ghost.zool.kyoto-u.ac.jp/indexr1.html) contains significant BLAST matches for vertebrate Myf, MEF2, SRF, and bHLH transcription factors. Of these, the only homolog characterized in *Ciona* is that for Myf, CiMDF (Meedel et al. 2002).

It is apparent from the distribution of predicted sites across the locus that conservation is a better predictor of the location and extent of the TnI MSRR than the predicted binding sites. Most of the predicted binding sites are not conserved. However, three conserved Myf sites are predicted near the 3′-end of CH1 (Fig. 5b), two on the reverse strand and one on the forward strand. Two sites occupy identical positions but on forward and reverse strands. Construct tn.msrr.363.d47.97 (Fig. 4c) eliminates all three Myf sites and is nonfunctional. The three sites were also targeted by two mutagenesis windows (Fig. 5d). Scrambled window 1 (positions 26–45, Fig. 5b) disrupts two of the predicted sites but leaves a third intact. The corresponding construct drives slightly weaker expression than the wild-type sequence. Scrambled window 2 (positions 41–60, Fig. 5b) disrupts all three predicted Myf sites and is almost completely nonfunctional. Restoring a predicted Myf site within this window also restores function (Fig. 5d). We conclude from these results that at least one of these conserved myf sites is necessary for muscle-specific expression of TnI. Therefore, it is likely that CiMDF (Meedel et al. 2002) directly activates transcription of TnI. Alternatively, it is possible that an as-yet-uncharacterized bHLH transcription factor binds this sequence.

We also attempted to identify the precise location of a promoter within the MSRR, whose presence was first suggested by our first set of constructs (Fig. 3). Using a neural network promoter prediction algorithm (Reese 2001), the strongest predic-

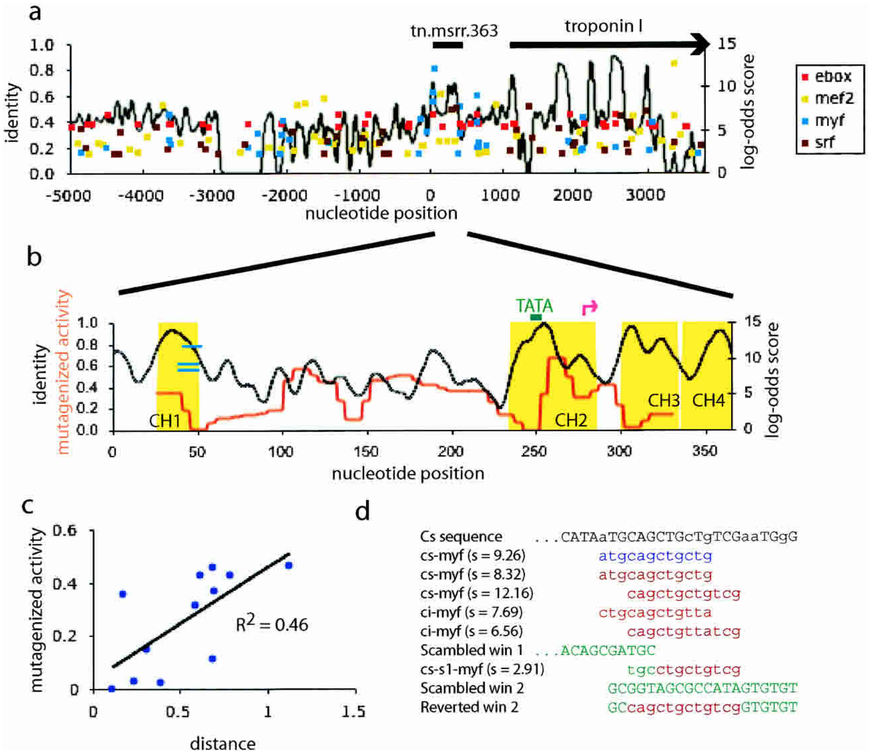


Figure 5. Conservation, predicted binding sites, and activity upon mutagenesis in the TnMSRR. (*a*) Plot of predicted binding sites for four common vertebrate muscle regulators for the *TnI* locus. The right *y*-axis shows the log-odds score of the site. The black line plots sequence identity (*left y*-axis) between *C. savignyi* and *C. intestinalis*. The location of the enhancer (tn.msrr.363) is shown. (*b*) Plot of conservation versus activity upon mutagenesis, with predicted binding sites. The left *y*-axis is sequence identity (black) and activity on mutagenesis (red), and the right *y*-axis is the log-odds score of predicted sites. The black line plots sequence identity between the *C. savignyi* and *C. intestinalis* enhancers. The red line represents the activity (number of embryos staining/total embryos) of the lacZ reporter when a 20-bp window spanning that position is randomized. The yellow shading represents CHAOS hits, the green line is a conserved TATA-box, the pink arrow is the predicted transcription start site, and the blue lines are the three myf predictions. (*c*) Scatterplot of activity upon mutagenesis versus distance. The distance within each 20-mer is calculated by Kimura's two-parameter model. Mutagenized windows with activity >50% are not included, nor are poorly aligned sequences. (*d*) Predicted myf sites in *C. savignyi* and *C. intestinalis*, aligned with the two relevant scrambled windows. A predicted site remains in scrambled window 1, which retains activity, and no sites are predicted in scrambled window 2, which disrupts activity. Reverting one of the predicted sites back to wild type restores function (reverted window 2). Sequence of *C. savignyi* (Cs) is shown on *top*, with capital letters indicating identity with *C. intestinalis*. Blue sites are predicted on the forward strand, red sites on the reverse. All images and sequences are archived for retrieval at http://mendel.stanford.edu/supplementarydata/johnson.

tion (score = 0.92) for a promoter in the locus resides in the most conserved region of CH2. A fully conserved putative TATA-box occurs 32 bp upstream from the predicted transcription start site (base pairs 249–257) (cf. Figs. 3a and 5b). The near-complete lack of activity of the construct bearing scrambled window 15, which disrupts the TATA sequence (Fig. 5b), is consistent with the importance of this region for MSRR function.

Discussion

Our truncations, deletions, and the scanning mutagenesis show that necessary and sufficient elements are usually conserved between the two *Ciona* species. The MSRR, the most conserved large noncoding region of the *TnI* locus, is clearly sufficient for expression; and within the MSRR, there are only a few windows that are at least partially necessary for function but do not appear to be conserved (base pairs 131–150) (cf. Fig. 5b). Conversely, not all conserved elements are necessary for expression in the larval tail muscle. The most obvious example of this is an 11-bp stretch of perfectly conserved sequence in CH1 (TGTGACGTCAT; base pairs 27–38) that appears unnecessary for expression in tail muscle. A preliminary analysis demonstrated that this sequence is necessary for normal expression in the adult body wall, underscoring the possibility that some conserved elements were not detected to be functional because we did not test our constructs for every possible expression domain.

The MSRR also appears to contain a conserved promoter. Intriguingly, its location is 500 bp upstream of the first annotated exon of Cs-TnI (Fig. 3). Why do the most-5′-ends of *Ciona* TnI transcripts, upon which the annotation is based, start there instead of in CH2? TnI is known to be *trans*-leader spliced (Vandenberghe et al. 2001). In *trans*-splicing, the most 5′-part is removed from the nascent transcript, and a leader sequence, transcribed in *trans* at another locus, is ligated to the 5′-end of the truncated transcript (Vandenberghe et al. 2001). This has the effect that the first part of the transcript is not exonic, but equivalent to an intron, and that the first exon does not start at the promoter. The best interpretation of all available data is therefore that the *TnI* promoter is the predicted one in CH2, and that the first exon of *TnI* is, indeed, the annotated one.

It appears that the divergence, developmental biology, and ease of manipulation of *C. savignyi* and *C. intestinalis* are uniquely advantageous for regulatory experimentation guided by comparative sequence analysis. The extent of sequence divergence between the two species is so large that unconstrained sequences do not display more similarity than expected by chance (Fig. 1a–h). Yet, the embryology and biology of the two species are nearly indistinguishable, and, in fact, noncoding sequences that are functional in one species are reciprocally functional in the other species. Here we show that such functional conservation is apparent in the regulatory elements of several genes in a variety of tissues.

Several studies in other systems have leveraged sequence conservation to identify functionally important regions, for example, *C. elegans*/*C. briggsae* (Kirouac and Sternberg 2003), yeast (Kellis et al. 2003), vertebrates (Göttgens et al. 2002), and mouse/human (Nobrega et al. 2003). In *Ciona*, a recent study found conserved GATA transcription factor binding sites in the *Ci-Otx* 5′-region (Bertrand et al. 2003). However, in these studies, the focus has either been on larger regions equivalent in size to the entire TnI MSRR, or on previously characterized short sequence motifs that are likely to bind known transcription factors. Furthermore, both the measure of conservation and the interpretations of the experimental data are usually binary (conserved vs. not conserved; and necessary vs. not necessary, or sufficient vs. not sufficient). In contrast, the strength of the *Ciona* system lies in the ability to efficiently characterize regions of functional significance of any size in a quantitative manner. We show that this can further be combined with other information such as PSSMs for known transcription factors, or promoter predictions, to identify small sequence motifs important for function.

The high degree of resolution possible in *Ciona* facilitated the demonstration that there is a quantitative relationship between degree of conservation and functional importance (Fig. 5c). This correlation extends over the entire MSRR, and is unlikely to be fully explainable by known transcription factor binding sites alone. It may also reflect structural, or additional functional, constraints of unknown nature on regulatory regions. Future genome-wide comparative sequence analysis between these two species, coupled with assays for function, may uncover the mechanisms underlying such constraints.

Methods

Comparative sequence analysis

The orthologous genomic region surrounding the *Troponin I* (*TnI*), *Synaptotagmin* (*Syt*), *α-tubulin* (*Tub*), *Noto9*, *Forkhead* (*Fkh*), *Tropomyosin-like* (Tro-like), *Snail*, and *Brachyury* (*Bra*) genes of *C. savignyi* (http://www.broad.mit.edu/annotation/ciona/) and *C. intestinalis* (http://genome.jgi-psf.org/ciona4/ciona4.home.html) were defined by regions of synteny spanning three genes predicted by the JGI annotation and by tBLASTN (Altschul et al. 1997). Each ascidian's gene is the other's reciprocal best hit by tBLASTN. We used MLAGAN (Brudno et al. 2003b), which uses a global alignment algorithm to fill in the regions between local anchors of high similarity, to perform multiple sequence alignments over large genomic regions. *C. savignyi* is used as the reference sequence for *TnI*, *Syt*, *Tub*, and *Noto9*, and *C. intestinalis* is used as the reference sequence for *Forkhead*, *Tropomyosin-like*, *Snail*, and *Brachyury*. We have archived all alignments for retrieval at http://mendel.stanford.edu/supplementarydata/johnson. CHAOS (Brudno et al. 2003a), which is optimized for locating short conserved sequences and can detect similarity on both strands, was used (rescore cutoff 1700) to find short regions of conservation between sequences. EST data from the Kyoto Ciona Ghost database (http://ghost.zool.kyoto-u.ac.jp/indexr1.html) were used to verify the location of exons in the multiple sequence alignment. Plots of sequence identity as a function of position in the alignment were generated in three steps: (1) eliminating all positions that correspond to gaps in the *C. savignyi* sequence; (2) calculating simple sequence identity in a moving window across the alignment; (3) smoothing the estimates for display by arithmetically averaging neighboring values in another round of moving windows. Window sizes were adjusted according to the desired resolution for display: 41 and 11 for Figures 1, a–h, and 5a; 17 and 17 for Figure 3a; and 11 and 7 for Figures 4a and 5b.

Prediction of transcription factor binding sites

Candidate transcription factor binding sites were identified using PSSM (position-specific sequence matrix) data for vertebrate muscle-regulated genes (Wasserman and Fickett 1998; Mount 2000). Each entry in the PSSM is represented by $f(b, i)$, the frequency for base b at position i. The nucleotide background probability is $p(b)$. For each position i, the log-odds value, $m(b, i)$, was calculated according to the formula $m(b, i) = \log[(f(b, i)/p(b)]$ and the total log-odds score for a sequence of length L was calculated by the sum:

$$S = \sum_{i=1}^{L} m(b, i)$$

A log-odds score cutoff of 2 and a background nucleotide probability of 60% G/C and 40% A/T was used for all analyses. For promoter prediction, we used a neural network algorithm (http://www.fruitfly.org/seq_tools/promoter.html) that had been shown to give a false-positive rate of <1% at the stringency we specified (score = 0.8) (Reese 2001).

Troponin I, synaptotagmin, α-tubulin, and Noto9 regulatory region constructs

Sequences for all constructs are available in the Supplemental data (http://mendel.stanford.edu/supplementarydata/johnson). All constructs were verified by sequencing. The native promoter fusion constructs were derived from the Ci-Bra lacZ fusion reporter construct (Corbo et al. 1997). A *C. savignyi* minimal promoter base plasmid (djmcs.lacZ) was derived from this construct by deleting the *Brachyury* upstream region and the original multiple cloning region, then adding XhoI, SalI, NotI, and KpnI sites. The lacZ coding sequence does not contain a start methionine codon, and the plasmid lacks a basal promoter. All minimally sufficient regulatory region (tn.msrr) plasmids were derived from the Forkhead basal promoter construct (pCES) (Harafuji et al. 2002).

Constructs were generated by standard procedures (Sambrook and Russell 2001) using PCR and subcloning. All truncation and deletion constructs were made by amplification off the original minimal promoter construct (tn.pro.1793) using restriction-site tailed-end primers and subcloning into either djmcs.lacZ or pCES. Perl scripts were used to optimize and choose all primer pairs.

Generation of Troponin I deletion and scrambled window constructs

Deletion and mutagenized constructs were made by overlap extension PCR (Sambrook and Russell 2001). For deletion plasmids, the region to be deleted was used as input for a Perl script, and the output provided sequences for optimized overlaps and pairs of forward and reverse primers closest to a T_m of 56°C. For the saturation mutagenesis experiments, the 363-bp construct was mutagenized by scrambling windows of 20 bp at a time. There were 20 windows, each of which overlapped by 5 bp with their neighboring window. Care was taken to minimize the identity between the wild-type sequence and the mutagenized sequence, resulting in all constructs having no more than two identities to the wild-type sequence. This scrambling method offers the advantage of maintaining the G/C content that exists in the wild-type sequence window while maximizing the dissimilarity to the mutagenized sequence. Distances between the *C. intestinalis* and *C. savignyi* sequence in each 20-mer were calculated using the Kimura 2-Parameter correction formula (Li 1997).

Source of animals and husbandry

C. intestinalis and *C. savignyi* were collected in San Francisco Bay, Santa Barbara Harbor, and San Diego Harbor. Animals were held in 50-gallon aquaria filled with fresh natural seawater equipped with a biofilter, a UV sterilizer, and chilled to 14°C. Gametes were dissected from the animals after they were kept in the aquaria under constant light for at least 48 h (Corbo et al. 1997; Nakatani et al. 1999).

Electroporation

Ascidian eggs were fertilized, dechorionated, and electroporated as previously described (Corbo et al. 1997). We used 100 µg of Quantum Maxi (BioRad) plasmid preps resuspended in water to electroporate fertilized *Ciona* eggs for a time constant of 15–20 msec. Embryos were reared to the desired stage of development, at 14°–16°C, fixed, and stained for lacZ (Corbo et al. 1997). Photography was carried out by clearing embryos in glycerol under a 20× microscope or in PBS with a stereoscope. Note that electroporation results in mosaicism such that many of the embryos stain for lacZ in only a portion of the cells that express the native genes, and embryos display a variety of staining patterns within one round of electroporation. However, we demonstrated that semiquantitative comparisons between different constructs are possible by, in a single dechorionation, conducting two electroporations each with four separate plasmid preparations of tn.pro.1793. The average percentage of embryos staining was 55% with a standard deviation of 18.6%. In light of this level of variability, we classify strength of expression as a range. Very strong (+++) represents a range of 50%–100% embryos staining, with most of the animals staining completely (Fig. 3d); strong (++) represents a range of 25%–50% consistently, with the occasional animal staining completely (Fig. 3e); and weak (+) represents <25% of the animals staining, and none of the animals staining strongly or completely. "Weak" staining animals often only stain a single cell in the tail muscle (Fig. 3f). Constructs that were mostly " − " but stained weakly on a single occasion are denoted +/−. Most data points represent at least two rounds of electroporation with more than two dozen embryos developing normally, and most were tested in both *Ciona* species. All photos are archived at http://mendel.stanford.edu/supplementarydata/johnson.

Acknowledgments

We offer many thanks to members of the Smith lab, especially Yuki Nakatani and Di Jiang, for generously making lab space and *Ciona* adults available on numerous occasions, and offering indispensable experimental advice. We also thank Bob Zeller (San Diego State, San Diego, CA) and Dave Keys (Joint Genome Institute, Walnut Creek, CA) for constructs and helpful advice; and Greg Cooper, Kerrin Small, Cheryl Smith, Midori Hosobuchi, and Matt Hill of the Sidow lab for helpful comments and assistance with bioinformatics. We also thank an anonymous reviewer who provided very helpful comments. D.S.J. acknowledges support from the EPA's STAR program and the Stanford Genome Training Program (NIH/NHGRI), A.S. from NIH/NIGMS.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bertrand, V., Hudson, C., Caillol, D., Popovici, C., and 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.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B. 2003a. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4:** 66.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. NISC Comparative Sequencing Program. 2003b. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.

Byrd, J. and Lambert, C.C. 2000. Mechanism of the block to hybridization and selfing between the sympatric ascidians *Ciona intestinalis* and *Ciona savignyi*. *Mol. Reprod. Dev.* **55:** 109–116.

Chiba, S., Awazu, S., Itoh, M., Chin-Bow, S.T., Satoh, N., Satou, Y., and Hastings, K.E. 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. IX. Genes for muscle structural proteins. *Dev. Genes Evol.* **213:** 291–302.

Cleto, C.L., Vandenberghe, A.E., MacLean, D.W., Pannunzio, P., Tortorelli, C., Meedel, T.H., Satou, Y., Satoh, N., and Hastings, K.E. 2003. Ascidian larva reveals ancient origin of vertebrate-skeletal-muscle troponin I characteristics in chordate locomotory muscle. *Mol. Biol. Evol.* **20:** 2113–2122.

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

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.

Di Gregorio, A. and Levine, M. 1999. Regulation of Ci-tropomyosin-like, a *Brachyury* target gene in the ascidian, *Ciona intestinalis*. *Development* **126:** 5599–5609.

Di Gregorio, A., Corbo, J.C., and Levine, M. 2001. The regulation of forkhead/HNF3-B expression in the *Ciona* embryo. *Dev. Biol.* **229:** 31–43.

Erives, A., Corbo, J.C., and Levine, M. 1998. Lineage-specific regulation of the *Ciona* snail gene in the embryonic mesoderm and neurectoderm. *Dev. Biol.* **194:** 213–225.

Göttgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)—Comparative analysis of five vertebrate SCL loci. *Genome Res.* **12:** 749–759.

Harafuji, N., Keys, D.N., and Levine, M. 2002. Genome-wide identification of tissue-specific enhancers in the *Ciona* tadpole. *Proc. Natl. Acad. Sci.* **99:** 6802–6805.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kirouac, M. and Sternberg, P.W. 2003. *cis*-Regulatory control of three cell fate-specific genes in vulval organogenesis of *Caenorhabditis elegans* and *C. briggsae*. *Dev. Biol.* **257:** 85–103.

Li, W.-H. 1997. *Molecular evolution*. Sinauer, Sunderland, MA.

MacLean, D.W., Meedel, T.H., and Hastings, K.E. 1997. Tissue-specific alternative splicing of ascidian troponin I isoforms. Redesign of a protein isoform-generating mechanism during chordate evolution. *J. Biol. Chem.* **272:** 32115–32120.

Makabe, K.W., Kawashima, T., Kawashima, S., Minokawa, T., Adachi, A., Kawamura, H., Ishikawa, H., Yasuda, R., Yamamoto, H., Kondoh, K., et al. 2001. Large-scale cDNA analysis of the maternal genetic information in the egg of *Halocynthia roretzi* for a gene expression catalog of ascidian development. *Development* **128:** 2555–2567.

Meedel, T.H., Lee, J.J., and Whittaker, J.R. 2002. Muscle development and lineage-specific expression of CiMDF, the MyoD-family gene of *Ciona intestinalis*. *Dev. Biol.* **241:** 238–246.

Mount, D. 2000. *Bioinformatics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Nakatani, Y., Moody, R., and Smith, W.C. 1999. Mutations affecting tail and notochord development in the ascidian *Ciona savignyi*. *Development* **126:** 3293–3301.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Nishida, H. 1987. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue-restricted stage. *Dev. Biol.* **121:** 526–541.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Reese, M.G. 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* **26:** 51–56.

Sambrook, J. and Russell, D.W. 2001. *Molecular cloning*, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Satoh, N. 1994. *Developmental biology of ascidians*. Cambridge University Press, New York.

Satou, Y., Yamada, L., Mochizuki, Y., Takatori, N., Kawashima, T., Sasaki, A., Hamaguchi, M., Awazu, S., Yagi, K., Sasakura, Y., et al. 2002. A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis* **33:** 153–154.

Takahashi, H., Hotta, K., Erives, A., Di Gregorio, A., Zeller, R.W., Levine, M., and Satoh, N. 1999. Brachyury downstream notochord differentiation in the ascidian embryo. *Genes & Dev.* **13:** 1519–1523.

Vandenberghe, A.E., Meedel, T.H., and Hastings, K.E. 2001. mRNA 5′-leader *trans*-splicing in the chordates. *Genes & Dev.* **15:** 294–303.

Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.

Web site references

http://genome.jgi-psf.org/ciona4/ciona4.home.html; *C. intestinalis* genome and annotation.

http://ghost.zool.kyoto-u.ac.jp/indexr1.html; *C. intestinalis* in situ expression database.

http://mendel.stanford.edu/supplementarydata/johnson; Source for Supplemental data.

http://www.broad.mit.edu/annotation/ciona/; *C. savignyi* genome sequence.

http://www.fruitfly.org/seq_tools/promoter.html; promoter prediction at Flybase.

http://www.broad.mit.edu/annotation/ciona; *Ciona savignyi* database at the Broad Institute.

Received July 3, 2004; accepted in revised form September 21, 2004.