Evolutionary analysis of the well characterized *endo16* promoter reveals substantial variation within functional sites

James P. Balhoff and Gregory A. Wray*

Department of Biology, Duke University, Durham, NC 27708

Edited by Roy J. Britten, California Institute of Technology, Corona Del Mar, CA, and approved April 21, 2005 (received for review December 22, 2004)

The evolutionary mechanisms that operate on genetic variation within transcriptional regulatory sequences are not well understood. We present here an evolutionary analysis of an exceptionally well characterized cis-regulatory region, the *endo16* promoter of the purple sea urchin. Segregating variation reveals striking differences in the intensity of negative selection among regulatory modules, reflecting their distinct functional roles. Surprisingly, transcription-factor-binding sites are as polymorphic and as likely to contain fixed differences as flanking nucleotides. Whereas nucleotides in protein-binding sites in the most proximal regulatory module exhibit reduced variation, those in other modules tend to be more polymorphic than putatively nonfunctional nucleotides. Two unrelated large insertions at the same position within the promoter are segregating at low frequencies; one is a strong ectodermal repressor that contains 16 verified transcription-factor-binding sites. These results demonstrate that a simple relationship between conservation and function does not exist within this cis-regulatory region and highlight significant population heterogeneity in the fine structure of a well understood promoter.

cis-regulatory evolution | population genetics | transcription

Detailed information about the nature and dynamics of sequence variation within cis-regulatory regions is necessary both for understanding the genetic basis of evolutionary change in organismal phenotype (1–4) and for developing informatic approaches to identifying transcription-factor-binding sites through interspecies comparisons (5, 6). Genes may acquire new patterns or levels of expression because of cis-regulatory-sequence evolution in several ways. Transcription-factor proteins interact with specific cis-regulatory sequences, which are typically 7–10 nucleotides long. These binding sites are often arranged into modules, clusters that each generate a particular subset of the total expression pattern (7). Individual protein-binding sites may be lost or added through simple nucleotide substitution. In the same way, substitutions within binding sites may modulate the binding intensity of proteins to their target sites. Another possibility is that groups of binding sites, or even entire modules, may be inserted near a gene through recombination or as part of a mobile element, thereby gaining a role in regulation. Each of these possibilities could produce functional variation available for the evolutionary process.

The types of genetic variation segregating within regulatory regions should affect the mechanisms in which genes gain new functions. For instance, if the addition of entire modules were a common component of genetic variation in promoters, it would suggest that genes could "easily" gain new, discrete modes of expression, if favored by natural selection. Likewise, available variation in individual protein-binding sequences might favor more fine-scale adjustments of a gene's expression pattern. To date, however, most of what we know about the population genetics of cis-regulatory sequences is based on fragmentary coverage, focusing on just one or a few transcription-factor-binding sites or a single enhancer directing just one part of a gene's transcription (reviewed in ref. 4). Because many cis-regulatory regions are composed of dozens of protein-binding sites distributed among several modules (7), information on the dynamics of variation within regulatory sequences is best obtained from a gene in which the majority of binding sites have been mapped and empirically validated and in which the mechanisms of cis-regulation are already well understood.

One such gene is *endo16*, which encodes an extracellular matrix protein of the sea urchin *Strongylocentrotus purpuratus*. The cis-regulatory region of this gene has been the subject of a series of detailed studies by Yuh, Davidson, and colleagues (8–11), who have created a fine-scale map of protein–DNA interactions (8, 9) and a detailed model of the regulatory consequences of each interaction (10, 11). *endo16* is expressed in a dynamic pattern during early development, beginning in the vegetal plate endoderm and, later, restricted to the stomach (12). Yuh *et al.* (8) have identified 56 protein-binding sites within a 2.3-kb sequence upstream of the transcriptional start site, which can completely recapitulate the embryonic expression pattern when used in transient expression assays. The *endo16* promoter sequence can be divided into six modules responsible for specific functions, such as early embryonic activation, expression within the archenteron, and repression within ectoderm and skeletogenic cells.

Leveraging the extraordinary level of functional detail available for the *endo16* promoter, we analyzed genetic variation at this locus within *S. purpuratus* and between *S. purpuratus* and its congener *Strongylocentrotus droebachiensis*. Here, we present results that reveal substantial differences in variation within functionally important regions of the promoter and a striking example of a polymorphism encompassing an entire regulatory module.

Materials and Methods

Sample Collection. *S. purpuratus* individuals were provided by C. Biermann (Portland State University, Portland, OR), C. Hollahan (Santa Barbara Marine Biologicals, Santa Barbara, CA), and Marinus (Los Angeles, CA). An individual of *S. droebachiensis* was provided by C. Biermann. Approximately a dozen tube feet were collected from each individual by clipping with scissors. Genomic DNA was extracted from either fresh or EtOH-preserved tube foot tissue by using the DNEasy tissue kit (Qiagen). DNA was stored at −20°C.

PCR, Cloning, and DNA Sequencing. The following primers were used to amplify five fragments of the *endo16* locus: 5′-CCCTTGTTACGCAGTTTTGTAT-3′/5′-GTTACGGTTTGGTCATTG-3′ (promoter 5′-half), 5′-GGGCACTGCTGGGATGAT-3′/5′-CCAAAACCGCGAACAGCA-3′ (promoter 3′-half), 5′-GGTCGAGGACAGGTCATA-3′/5′-GAGTTAGAGTCATCGTCG-3′ (first exon, intron), 5′-ATCAAG-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: indel, insertion and deletion.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ066774–DQ066866).

*To whom correspondence should be addressed. E-mail: gwray@duke.edu.

GAATGCTGCCTACT-3′/5′-ATTCTTTTCTCCTCGCT-CAT-3′ (fifth intron, 5′-end sixth exon), 5′-CTGTGAGCT-GACTAGCGATT-3′/5′-AGCGCAAATGGCTTTATT-3′ (sixth exon).

PCR was conducted in a 25-μl reaction volume by using either *Pwo* polymerase (Roche) (2.5 μl of 10× PCR buffer, 2.0 μl of $MgSO_4$, 2.5 μl of 2 mM dNTP, 2.5 μl of each 10 μM primer, 0.25 μl of 5 units/μl *Pwo*, 150 ng of DNA template, and 11.75 μl of sterile distilled water) or Phusion polymerase (Finnzymes, Helsinki) (5.0 μl of 5× PCR buffer, including $MgCl_2$, 2.5 μl of 2 mM dNTP, 2.5 μl of each 10 μM primer, 0.5 μl of 2 units/μl Phusion, 150 ng of DNA template, and 11.0 μl of sterile distilled water). *Pwo* reactions were carried out for 1 cycle of 3 min at 94°C followed by 35 cycles of 30 s at 94°C for denaturation, 1 min at 55°C for annealing, and 1 min at 72°C for extension followed by 1 cycle of 10 min at 72°C. Phusion reactions were carried out for 1 cycle of 1 min at 98°C followed by 30 cycles of 10 s at 98°C for denaturation, 20 s at 57°C for annealing, and 30 s at 72°C for extension followed by 1 cycle of 10 min at 72°C.

The PCR product was separated by electrophoresis on a 1.5% agarose gel, and the band was excised and purified by using the QIAquick gel purification kit (Qiagen) and stored at −20°C. Purified PCR products were cloned into the pCR4Blunt-TOPO plasmid by using the Zero Blunt TOPO cloning kit for sequencing (Invitrogen). Cloning followed the manufacturer's protocol, except for the addition of a 30-min incubation of each PCR product with *Pfu* native polymerase (Stratagene) at 72°C just before the cloning reaction (8.25 μl of PCR product, 0.25 μl of Pfu, 1.0 μl of 10× Pfu buffer, and 0.5 μl of 2 mM dNTP).

Plasmids were purified by using the Wizard Miniprep kit (Promega) and sequenced on both strands with both M13 and sequence-specific primers. DNA sequencing was performed with BigDye v3.0 (Applied Biosystems) and run on an Applied Biosystems 3700 automated DNA sequencer. Each clone was sequenced in both directions to ensure accuracy.

Sequence Analysis. PCR products representing the 5′ half of the promoter, the 3′ half of the promoter, the first exon and intron and proximal 55 bp of the second exon, the fifth intron and proximal 165 bp of the sixth exon, and the following 935 bp of the sixth exon were obtained from 20 individuals of *S. purpuratus* (only 12 for intron 5) and 1 individual of *S. droebachiensis*. Comparative analyses of variation reported here were confined to individuals of *S. purpuratus* from Santa Barbara, CA, for which 10 alleles were obtained at all of these locations. Ten additional alleles were sequenced from individuals collected at the other localities. The geographic sampling differed among regions of the locus as follows: the 5′ half of the promoter (5, Long Beach, CA; 5, Friday Harbor, WA), 3′ half of the promoter (2, Long Beach; 8, Friday Harbor), intron 1 and exon 1 (5, Long Beach; 5, Friday Harbor), and intron 5 and exon 6 (10 additional from Santa Barbara). Pooling these additional alleles with those from Santa Barbara did not materially alter the results. Sequence fragments were assembled into contigs by using Sequencher (Gene Codes, Ann Arbor, MI). These sequences have been submitted to GenBank as accession nos. DQ066774–DQ066866. The contigs, along with the published *endo16* cis-regulatory sequence (GenBank accession no. S75835) and sequence 127I21 from the *S. purpuratus* genomic bacterial artificial chromosome (BAC)-sequencing project conducted by Eric Davidson (http://sugp.caltech.edu/), were aligned by using CLUSTALX (13), and the alignments were manually edited with MACCLADE (14) to minimize unnecessary gaps. Sequence 127I21 was used as the site-numbering reference because of the substantial length differences between the published sequence and the most common alleles. The multiple alignments were partitioned into subalignments for each module as well as partitions of only-protein-binding sites and of only-nonfunctional sites. Modules and binding sites were located as described by Yuh and Davidson (10, 11); these partitions, along with the full alignment, were used in the following analyses. The software program DNASP (15) was used to calculate intraspecific variation (π and θ) and interspecific differences and also for computing the significance of McDonald–Kreitman test contingency tables. For the purposes of these analyses, portions of the alignment containing gaps were excluded. Significant difference in variation between sequence partitions was calculated by using a permutation test with 10,000 iterations.

Calculation of Intraspecific Length Variation. Because standard measures of sequence diversity exclude sites with length polymorphism, we measured intraspecific length variation, "indel π," as the average weighted number of insertion and deletion (indel) differences between allele pairs. A simple count of indel mutations does not take into account the size of each indel; however, length differences can result from both the number of indel mutations and their size (length). In comparing two alleles, it is possible simply to count the number of sites with gaps, but we expect that, as indels within a promoter increase in size, additional length will, at first, cause increased disruption of local protein interactions but have diminishing impact once local interaction is no longer possible. For this reason, we counted each indel as one difference plus a weight corresponding to the natural logarithm of the indel's length. This weighting scheme captures the log–log frequency distribution of indel length characteristic of many organisms, including *S. purpuratus* (16). Because the number of aligned sites depended on the alleles being compared, the number of differences was converted to a per-site value separately for each comparison by using the total alignment length for each comparison as the number of sites. We evaluated indel π for each of our sequence partitions by using a custom PYTHON script, available from the authors.

Results

We sequenced portions of the *endo16* locus, including the entire cis-regulatory region, from 20 individuals of the sea urchin *S. purpuratus* (Fig. 1). The promoter contains >250 SNPs and >40 length polymorphisms in a wide range of sizes within an ≈2.3-kb sequence (Fig. 2). The length variation is composed of isolated indels and four di- and trinucleotide tandem repeats from a few dozen to >300 bases in length. Although the sequence of cis-regulatory regions is expected to be under functional constraint, their organizational structure is distinctly different from that of protein-coding sequences. Because of the interspersion of functional elements among putatively nonfunctional sequences, one would expect promoters to accommodate both SNP and indel variation to a greater degree than that allowed by the genetic code found in protein-coding sequences (4).

Nucleotide Variation Within the *endo16* Locus. We compared sequence variation among different partitions of the *endo16* locus within 10 individuals, all collected near Santa Barbara, CA, by using a standard measure of nucleotide diversity, the average number of differences between allele pairs, π (17), and the heterozygosity measure, θ (18). Although analyses of all 20 haplotypes produced concordant results, we focused on variation within Santa Barbara urchins to exclude the possibility of geographic structure within the data. In general, levels of polymorphism found within *endo16* matched expectations based on functional assignments: highest in introns, intermediate in the cis-regulatory region, and lowest in exon sequences (Table 1). Because π and θ are in broad agreement, we will focus on comparisons of π between functional partitions within the sequence. Overall polymorphism within the *endo16* promoter sequence, with $\pi = 0.040$, is slightly greater than that of the nearby intron 1, $\pi = 0.028$ (Table 2), which has been shown to be without embryonic regulatory function (9). In contrast, nucleotide variation within intron 5, $\pi = 0.060$, is significantly greater than that of the promoter and intron 1 (Table 2). If the level of polymorphism in intron 5 is representative of neutral sequence

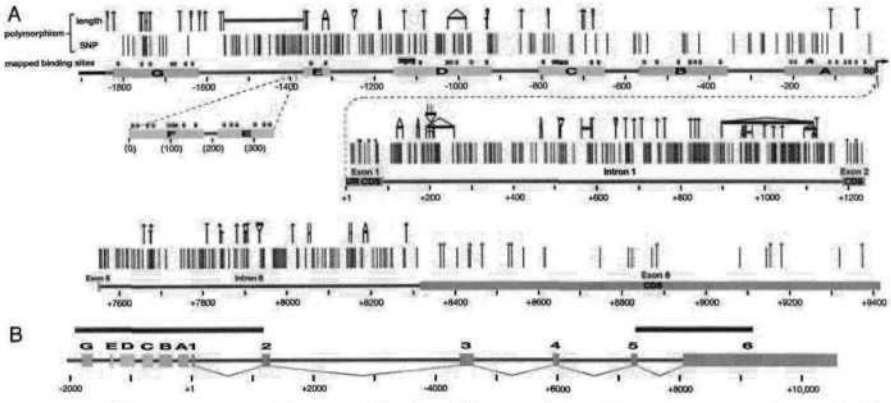

Fig. 1. Structure of the *endo16* promoter and coding sequence. (*A*) Variation within sequenced regions. Light green boxes represent activator modules, light pink boxes represent repressor modules, and light blue boxes represent protein-coding sequence. Hot pink squares represent locations of protein-binding sites. Red bars identify SNPs found within our sample of 20 individuals (12 individuals for intron 5). Within coding sequence, a bar with a ball on the end denotes an SNP that changes an amino acid. Above the SNPs are representations of length variation (blue) and simple tandem repeats (green) among the 20 individuals. Downward-pointing triangles represent insertions relative to a reference bacterial artificial chromosome sequence, whereas upward-pointing triangles indicate deletions. The length of the indel polymorphism is reflected in the width of the triangle. Filled arrowheads reflect a single-base indel. (*B*) Overview of *endo16* locus up to exon 6. Purple bars denote regions used in this study. Intron–exon structure beyond exon 6 has not been determined.

variation, it may suggest that the first intron performs some regulatory function in other parts of the life cycle. The promoter is substantially more variable than the 1,100-bp exon-6 sequence (Table 1). In contrast, the short fragment of exon 2 exhibits much lower levels of conservation.

Because of the exceptional detail in which the structure and function of the *endo16* cis-regulatory region have been characterized, we were able to examine nucleotide polymorphism within the promoter by using a more fine-grained approach. Each of the modules within the *endo16* promoter has been shown to play a distinct role in generating the overall transcriptional profile (8). Additionally, some modules appear more important for function than others. For instance, module A is required for early activation of *endo16* transcription in the endoderm and is also necessary to integrate the effects of the other modules. In contrast, module G provides a simple boosting effect on the general transcriptional output of the promoter. This regional partitioning of function within the promoter may result in distinct evolutionary dynamics among the different modules.

We calculated π for each module separately (Fig. 2*A*) and found distinct differences in the levels of variation. Modules A–C all exhibit $\pi \leq 0.026$, whereas variation more than doubles in the more distal regions of the promoter. The difference in polymorphism within modules B and A compared with G and D is highly significant (Table 2). Modules E and F (Fig. 2, FE region) may not represent functional sequences in our sample because these modules appear to exist as a single insertion polymorphism and are absent from all 20 haplotypes we sequenced (described in more detail below).

Next, we asked whether selective constraint could be detected among polymorphisms in nucleotides belonging to transcription-factor-binding sites, the functional components of cis-regulatory regions. When comparing binding-site nucleotides from all of the modules (excluding modules E and F) with all other module nucleotides (Table 1), there is no significant difference in π (Table 2). Surprisingly, only in module A is a reduced level of polymorphism observed within binding sites as compared with the other nucleotides in the module [Figs. 2 and 3*A*, although the reduction is not significant ($P = 0.121$)]. The other modules, B–D and G, display a significantly elevated level ($P = 0.025$) of nucleotide polymorphism within protein-binding sites in comparison with putatively nonfunctional sites (Table 2). Of the 56 described protein-binding sites within the *endo16* promoter, approximately half (23 sites) bind the DNA-looping factor GCF1. Because the GCF1 site is represented so many times and has a predominantly structural rather than regulatory role (19), one might expect an increased tolerance for variation in these sites. Although GCF1 sites do appear more variable than other protein-binding sites ($\pi = 0.062$ vs. 0.044), this is not a significant difference (Table 2). Even when the more variable GCF1 sites are excluded, the other binding sites are still at least as polymorphic as non-protein-binding sites (not significantly different, Table 2).

***endo16* Divergence Between Species.** When the *endo16* promoter of *S. purpuratus* is compared with the homologous region of *S. droebachiensis*, much the same spatial pattern of variation emerges (Fig. 2*A* and Table 1). Again, the proximal half of the cis-regulatory region appears to be under greater selective constraint than the distal portion. This spatial correspondence in polymorphism and divergence is supported by an examination of π vs. substitutions per site in 17 nonoverlapping 100-bp windows within the promoter, showing a Pearson correlation coefficient of $r = 0.678$ ($P = 0.001$). Comparisons of divergence within binding sites and nonbinding sites again show that only module A has sustained fewer nucleotide substitutions per site within protein-binding sites as compared with nonfunctional sites. In modules B–D and G, a greater number of substitutions is found within protein-binding sites. We performed a modified McDonald–Kreitman test (20) to compare the proportion of variation within binding sites and nonbinding sites within and between species (Table 3). In each module, as well as overall, there is no significant difference in the proportion of variation in the two classes of sites within and between species, consistent with a neutral

EVOLUTION

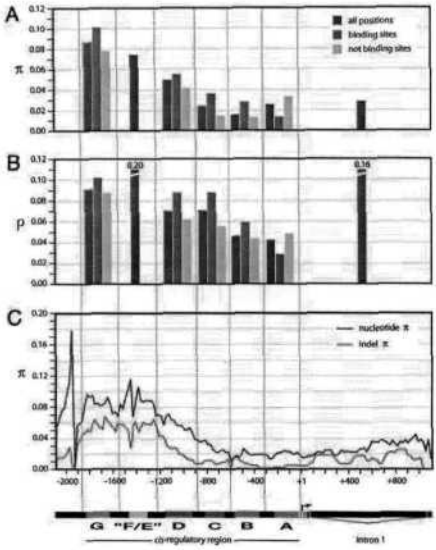

Fig. 2. Heterozygosity and divergence differ among modules, represented in scale for the reference allele (127121) at the bottom of the figure. Within our sample, the promoter ranges from 1,850 to 2,297 bases in length. (*A*) Average pairwise differences (π) within each module. Blue (first bar) represents the entire module, dark green (second bar) includes only nucleotides within binding sites, and light green (third bar) includes all nucleotides excluding binding sites. (*B*) Observed nucleotide substitutions within each module as compared with *S. droebachiensis*. Colors are as described for *A*. (*C*) Sliding 300-base window of indel π (blue) and nucleotide π (red) across the *endo16* promoter, in steps of 30 bases.

model of evolution. Given that we examined only one complete sequence of the promoter from *S. droebachiensis*, it is possible that some of the between-species substitutions could actually be shared polymorphisms. However, inspection of four *S. droebachiensis* partial promoter fragments of 950 bp revealed no such polymorphisms (data not shown).

Length Variation Within *endo16*. Length polymorphisms are rare in coding sequences and are, generally, ignored in studies of molecular population genetics. In cis-regulatory sequences, however, indels and simple sequence repeats are more common, and many are likely to have a functional impact (4). For this reason, we would expect length variation to be affected by local selective constraints in much the same way as the surrounding SNP variation. Natural selection should shape both the size and physical distribution of length polymorphisms with cis-regulatory regions. One might expect fewer indels within modules than between modules because, within modules, transcription factors are more likely to interact in spacing-dependent ways, such as cooperative binding or competitive exclusion. Although there are standard measures used to quantify nucleotide polymorphism in a population, variation in length is more difficult to compare. We devised a simple measure of average pairwise length differences that takes into account both the number of indels between two sequences and the length of those indels (indel π or π_1). We expect the difference in functional impact

Table 1. Nucleotide diversity and fixed differences

Sequence partition	π per site	θ per site	Fixed differences per site*	Length†
Entire promoter	0.040	0.039	0.074	1,729
All modules	0.041	0.040	0.064	980
All intermodules	0.039	0.039	0.090	585
All b.s.	0.049	0.049	0.073	318
All non-b.s.	0.037	0.036	0.060	662
All non-GCF1 b.s.	0.044	0.042	0.072	230
All GCF1 b.s.	0.062	0.068	0.076	88
Module A	0.026	0.023	0.041	184
Module A b.s.	0.014	0.017	0.030	64
Module A non-b.s.	0.033	0.027	0.048	120
Module B	0.016	0.020	0.047	213
Module B b.s.	0.028	0.026	0.060	54
Module B non-b.s.	0.012	0.018	0.043	159
Module C	0.024	0.029	0.070	159
Module C b.s.	0.036	0.037	0.088	66
Module C non-b.s.	0.015	0.023	0.057	93
Module D	0.050	0.053	0.070	227
Module D b.s.	0.064	0.066	0.087	86
Module D non-b.s.	0.041	0.045	0.061	141
FE region	0.075	0.059	0.204	54
Module G	0.086	0.072	0.091	197
Module G b.s.	0.110	0.103	0.102	48
Module G non-b.s.	0.078	0.062	0.088	149
Exon 1	0.009	0.012	0.054	61
Exon 2	0.029	0.026	0.107	55
Exon 6	0.006	0.006	0.045	1,100
Intron 1	0.028	0.031	0.161	966
Intron 5	0.060	0.045	0.137	727

b.s., Binding sites.
**S. purpuratus* vs. *S. droebachiensis*.
†Number of nucleotides excluding indels in population data.

of increasing an indel's length to be more substantial among shorter indels, whereas, beyond a certain length, local interactions will be so disrupted that different lengths have similar impacts. For this reason, each indel is weighted according to its length, but the additional weight diminishes logarithmically as indels get longer.

Many of these expectations are borne out by the levels of length variation within the *endo16* promoter. Within the cis-regulatory region, indels range from 1 to 340 bases in length; in contrast, the coding regions sampled have no indel variation (Fig. 1). The levels of π_1 closely follow nucleotide π (π_n), possibly reflecting a higher level of constraint on length variation in modules A and B as compared with the more distal portions of the promoter (Fig. 2*C*). This correlation suggests that indels, despite having a mutational basis independent from nucleotide polymorphisms, respond similarly to localized patterns of selective constraint. Although the intermodule regions did not show increased nucleotide polymorphism relative to the modules, they did display an increased level of interspecific substitutions (Table 1); in the case of indel polymorphism, we see a slightly higher level of π_1 when the intermodule regions are grouped together (0.019) vs. the modules alone (0.015). This finding may suggest that the spacing between binding sites within modules is important. There are also several runs of simple sequence repeats whose length varies among the sampled individuals. Interestingly, all four simple sequence repeats fall within the repressor and booster modules and not within modules A and B. Overall, there seems to be greater constraint on length polymorphisms within the activator modules.

A Polymorphism Encompassing an Entire Module. None of the 20 alleles of the *endo16* cis-regulatory region that we sampled

Table 2. Comparisons of sequence partitions for differences in nucleotide diversity

Compared partitions	π per site	P
Promoter	0.040	<0.0001*
Exon 6	0.006	
Promoter	0.040	0.005*
Intron 1	0.028	
Promoter	0.040	0.001*
Intron 5	0.060	
Intron 1	0.028	<0.0001*
Intron 5	0.060	
All modules	0.041	0.394
Intermodule regions	0.039	
Proximal modules (B, A)	0.021	<0.0001*
Distal modules (G, D)	0.067	
All b.s.	0.049	0.086
All non-b.s.	0.037	
All b.s., excluding module A	0.058	0.025*
Non-b.s., excluding module A	0.038	
GCF1 b.s.	0.062	0.148
Non-GCF1 b.s.	0.044	
Non-GCF1 b.s.	0.044	0.335
All non-b.s.	0.037	

b.s., Binding sites.
*Significant difference.

Table 3. McDonald–Kreitman tests for protein-binding sites and nonbinding sites

Module	Class	Polymorphic sites	Fixed changes	P*
A	Binding	3	1	1.000
	Nonbinding	9	3	
B	Binding	4	2	1.000
	Nonbinding	8	5	
C	Binding	7	4	1.000
	Nonbinding	6	3	
D	Binding	15	2	1.000
	Nonbinding	18	3	
G	Binding	14	1	0.647
	Nonbinding	26	5	
All	Binding	43	10	0.675
	Nonbinding	67	19	

*P value by Fisher's exact test.

contained sequences corresponding to modules E or F as described by Yuh and Davidson (8). We therefore surveyed 60 additional individuals by PCR for length polymorphisms across this region and recovered two alleles whose length suggested the presence of the missing modules (Fig. 3). When sequenced, these two alleles were found to contain unrelated, ≈350- and 250-bp sequences, respectively, within the same GT repeat (because of polymorphisms within this repeat, the exact breakpoints could not be identified but are probably very close). Both sequences

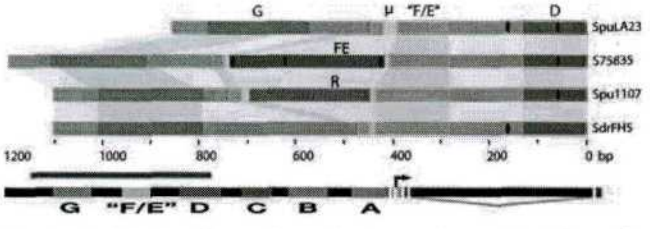

Fig. 3. Modules E and F are the result of a novel insertion, polymorphic within *S. purpuratus*. Four promoter variants are "right-justified" within module D to detail the large length differences among them. Sequences corresponding to modules E and F (blue) are found only in the published sequence (GenBank accession no. S75835), embedded within a dinucleotide repeat (yellow). This insertion is the location of 16 verified binding sites. Most alleles (e.g., *SpuLA23*) possess only the dinucleotide repeat and a small section of module E containing two GCF1 sites, as does the sequence from *S. droebachiensis*. Interestingly, one allele we sequenced (*Spu1107*) contains an unrelated insertion (purple) within the same repeat sequence. Potential cAMP response element-binding protein (CREB)-binding sites are indicated by black ovals (see text).

are clearly insertions (by outgroup comparison with *S. droebachiensis*) and seem to be segregating at low frequencies within the population (<2%). One of these insertions is similar in sequence to the published modules E and F, which function as strong ectodermal repressors (8). Remarkably, the other insertion is a completely unrelated sequence.

Transcription of *endo16* in the larval gut has been under stabilizing selection over the past 40 million years despite extensive turnover of cis-regulatory sequences outside of module A (21). The E/F insertion module is compatible with this evolutionarily conserved transcription profile because it represses ectodermal expression; yet, it is missing from the majority of individuals of *S. purpuratus*, suggesting that other cis-acting sequence elements must repress transcription in the ectoderm in most alleles. The trans-acting factor that mediates repression through module F is a cAMP response element-binding protein (CREB)-family protein (8). The functionally characterized CREB-binding site is missing in most individuals because it falls within the insertion. However, two potential CREB-consensus sites are present in and near module D (Fig. 3) and may function in ectodermal repression. The E/F insertion demonstrates that single mutations can transpose functional clusters of transcription-factor-binding sites into a fully operational cis-regulatory system. This length polymorphism illustrates a dramatic way in which transcriptional regulatory sequences of approximately equivalent function might turn over during the course of evolution. Transposons and retroviruses are known, on occasion, to position regulatory sites near genes (22, 23). Sequences similar to the E/F insertion are present in at least five locations throughout the genome (preliminary sea urchin genome assembly, data not shown). However, this insertion does not exhibit any obvious hallmarks of a mobile element, such as direct or inverted repeats. Strikingly, the insertion introduces 16 distinct protein-binding sites into the *endo16* promoter (8), a substantial cluster of novel sites.

Discussion

Evolutionary analyses of promoters can be challenging because of the lack of a genetic code, making the proximate consequences of nucleotide change difficult to predict. More detailed knowledge of the variability of functional sites within promoters will help us develop more accurate models of evolutionary change within cis-regulatory sequences. For this reason, we examined the distribution of natural variation within the *endo16*

EVOLUTION

promoter, perhaps the most thoroughly characterized of any eukaryotic cis-regulatory sequence.

Sequence Variation. Levels of variation within the *endo16* locus were comparable to previous estimates of sequence polymorphism within sea urchins (16, 24). Within the locus, variation within the promoter is, as expected, greater than that in the exons and lower than, or similar to, that in the introns. However, within the promoter, striking patterns of variation emerge when the functional structure of the cis-regulatory sequence is taken into account. Levels of polymorphism among different functional modules within the promoter vary more than 5-fold (Fig. 2), suggesting that different modules experience distinct levels of selective constraint. Although the proximal portion of the promoter is the most conserved, patterns of constraint do not map directly to sites of known functional importance. For instance, promoter modules are no less variable than neighboring stretches of functionless sequence. Furthermore, sequences known to bind transcription-factor proteins exhibit excess variation in comparison with nonfunctional sequences in all modules, with the exception of module A. Divergence between *S. purpuratus* and *S. droebachiensis* is consistent with the patterns of variation seen within *S. purpuratus*. It is clear that only the binding sites in module A exhibit any signature of purifying selection; it is reasonable to suppose that this stems from constraints imposed on module A as an integrator of input from all the other modules (11). Based on sequence data alone, the consequences of the high levels of binding-site sequence polymorphism in the other modules are not clear. It is possible, although unlikely, that the observed changes do not alter the pattern of transcription-factor binding at those sites and, so, are entirely neutral. Other studies of variation in cis-regulatory regions have failed to detect significant patterns of conservation across binding sites (25, 26). Another possibility is that stabilizing selection on transcriptional output allows slightly deleterious mutations to persist, compensated for by adaptive changes elsewhere in the promoter and resulting in continuous binding-site turnover (27, 28). Functional tests, including protein–DNA binding assays and *in vivo* expression assays, provide a means of distinguishing between these possibilities.

Length Variation. The *endo16* promoter harbors substantial amounts of length polymorphism (Fig. 1). Regions of the *endo16* promoter exhibiting the lowest sequence polymorphism also have less length polymorphism (Fig. 2). Although the level of length polymorphism may be influenced by background selection on the surrounding nucleotide sequence (29), the level may also reflect constraints on binding-site spacing within those modules. These constraints include functionality resulting from cooperative binding interactions as well as steric hindrance (30). Expression assays provide a means of exploring this possibility. Besides altering existing regulatory functionality, it is also possible that length variants within promoters could introduce new binding sites. We have documented a striking case in which two of the functionally characterized modules, E and F, are the result of an inserted sequence of more than 300 bases, polymorphic and rare within *S. purpuratus*. This finding raises the question of how *endo16* transcription is repressed in the ectoderm in the majority of segregating haplotypes, which lack modules E and F. Again, experimental analyses provide a way to investigate such questions. Although the insertion of entire functional modules is surprising, it may not be uncommon. Other repetitive sequences within *S. purpuratus*, such as the RSR element controlling transcription of *spec* genes, have been shown to have been recently inserted upstream of genes and then optimized for transcriptional activity (22).

Conclusions. The relationship between function and variation within cis-regulatory sequences is complex. Although functional sequences are more conserved within one module of the *endo16* promoter, variation within the rest of the promoter is at least as great as that in putatively functionless sequences. This result indicates that sequence conservation does not always provide a reliable guide to the discovery of cis-regulatory elements. As with coding sequences, understanding the evolutionary mechanisms that determine levels of variation within promoter sequences remains a significant challenge.

We thank Cathy Yuh, Eric Davidson, Christiane Biermann, Jesse Kuhn, William Nielsen, Margaret Pizer, Matt Rockman, Laura Romano, Ann Rouse, Ivy Chen, Brian Hanson, Kitty Oberg, Anjali Patel, and Emily Stammel for contributing resources and other assistance to this project. Julia Bowsher, Dave Des Marais, Matthew Hahn, and one anonymous reviewer provided helpful comments. This work was supported by National Science Foundation Doctoral Dissertation Improvement Grant DEB-02-06660 and a National Defense Science and Engineering Graduate Fellowship (to J.P.B.) and by National Aeronautics and Space Administration Grant NAG-2-1583 (to G.A.W.).

1. Carroll, S. B. (2000) *Cell* **101,** 577–580.
2. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. (2004) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell Scientific, Malden, MA).
3. Davidson, E. H. (2001) *Genomic Regulatory Systems: Development and Evolution* (Academic, San Diego).
4. Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. & Romano, L. A. (2003) *Mol. Biol. Evol.* **20,** 1377–1419.
5. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) *Science* **301,** 71–76.
6. Stormo, G. D. (2000) *Bioinformatics* **16,** 16–23.
7. Arnone, M. I. & Davidson, E. H. (1997) *Development (Cambridge, U.K.)* **124,** 1851–1864.
8. Yuh, C.-H. & Davidson, E. H. (1996) *Development (Cambridge, U.K.)* **122,** 1069–1082.
9. Yuh, C.-H., Ransick, A., Martinez, P., Britten, R. J. & Davidson, E. H. (1994) *Mech. Dev.* **47,** 165–186.
10. Yuh, C.-H., Bolouri, H. & Davidson, E. H. (2001) *Development (Cambridge, U.K.)* **128,** 617–629.
11. Yuh, C.-H., Bolouri, H. & Davidson, E. H. (1998) *Science* **279,** 1896–1902.
12. Ransick, A., Ernst, S., Britten, R. J. & Davidson, E. H. (1993) *Mech. Dev.* **42,** 117–124.
13. Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F. & Higgins, D. (1997) *Nucleic Acids Res.* **25,** 4876–4882.
14. Maddison, D. & Maddison, W. (2000) MACCLADE (Sinauer Associates, Sunderland, MA), Version 4.0.
15. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15,** 174–175.
16. Britten, R. J., Rowen, L., Williams, J. & Cameron, R. A. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4661–4665.
17. Tajima, F. (1983) *Genetics* **105,** 437–460.
18. Watterson, G. A. (1975) *Theor. Popul. Biol.* **7,** 256–276.
19. Zeller, R. W., Griffith, J. D., Moore, J. G., Kirchhamer, C. V., Britten, R. J. & Davidson, E. H. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 2989–2993.
20. McDonald, J. H. & Kreitman, M. (1991) *Nature* **351,** 652–654.
21. Romano, L. A. & Wray, G. A. (2003) *Development (Cambridge, U.K.)* **130,** 4187–4199.
22. Dayal, S., Kiyama, T., Villinski, J. T., Zhang, N., Liang, S. & Klein, W. H. (2004) *Dev. Biol.* **273,** 436–453.
23. Kidwell, M. G. & Lisch, D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7704–7711.
24. Britten, R. J., Cetta, A. & Davidson, E. H. (1978) *Cell* **15,** 1175–1186.
25. Macdonald, S. J. & Long, A. D. (2005) *Mol. Biol. Evol.* **22,** 607–619.
26. Phinchongsakuldit, J., MacArthur, S. & Brookfield, J. F. (2004) *Mol. Biol. Evol.* **21,** 348–363.
27. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. (2000) *Nature* **403,** 564–567.
28. Ludwig, M. Z., Patel, N. H. & Kreitman, M. (1998) *Development (Cambridge, U.K.)* **125,** 949–958.
29. Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993) *Genetics* **134,** 1289–1303.
30. Latchman, D. S. (1998) *Eukaryotic Transcription Factors* (Academic, San Diego).