Published in final edited form as:

Mar Genomics. 2009 June ; 2(2): 85–98. doi:10.1016/j.margen.2009.03.007.

Intragenomic Evolution of a Transcriptional Enhancer in the Genome of *Strongylocentrotus purpuratus*

Takae Kiyama¹, Jiexin Zhang², Shoudan Liang², Shuguang Liang¹, and William H. Klein^{1,3,†}

¹Department of Biochemistry and Molecular Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

²Division of Quantitative Sciences, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

³Training Program in Genes and Development, The University of Texas Graduate School of Biomedical Sciences at Houston, Houston, TX 77030, USA

Abstract

General principles for how genomic regulatory elements evolve to alter patterns of gene expression remain vague. The purpose of this study was to gain insights into the evolution of genomic regulatory elements by investigating the unique features of a transcriptional enhancer that directs *Spec2a* gene expression in *Strongylocentrotus purpuratus*. The *Spec2a* enhancer is embedded in a repetitive sequence family interspersed throughout the genome. We surveyed the genome and identified 274 of these sequences. They displayed a continuum of sequence divergence defining high and low divergence classes. Alignment of 52 most related to the *Spec2a* sequence revealed a complex pattern of rearrangements, insertions and deletions, and base-pair changes. A distance tree for the 52 sequences was constructed and correlated with enhancer activity. Unexpectedly, we found a wide range of activities. Notably, repetitive sequences lacking essential *cis*-elements found in the *Spec2a* enhancer still had strong activity. We identified short, conserved motifs within the repetitive sequences that may represent novel *cis*-regulatory elements. Many repetitive sequences with enhancer activity were found nearby genes, suggesting that they regulate gene expression. The results show that the repetitive sequences are rapidly evolving in the *S. purpuratus* genome and may serve as a renewable pool of transcriptional enhancers.

Keywords

Strongylocentrotus purpuratus genomics; sea urchin transcriptional enhancers; interspersed repetitive sequences; *cis*-regulatory element evolution

1. Introduction

Although it is well established that extragenic genomic sequences containing *cis*-regulatory elements have crucial roles in development and evolution, many questions on how

© 2009 Elsevier B.V. All rights reserved.

[†]Corresponding author. Department of Biochemistry and Molecular Biology, Unit 1000, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX, 77030, USA. Tel.: 713-834-63012; fax: 713-834-6266. whklein@mdanderson.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

regulatory elements diversify and are selected to modulate changing patterns of gene expression over evolutionary time remain unanswered (Davidson and Erwin, 2006; Carroll, 2008). An approach often taken to address these issues is to compare orthologous transcription factors and their associated *cis*-regulatory elements among species where alterations in gene expression patterns are correlated to morphological changes. For example, recent investigations on changes in *cis*-regulatory elements for wing pigmentation pattern involved in male courtship behavior in a *Drosophila* clade (Prud'homme et al., 2006) and gill and skin pigmentation in sticklebacks and humans (Miller et al., 2007) provide strong support for the importance of *cis*-regulatory element variation as a driver of morphological diversity.

In this study, we take a complementary approach in which we investigate the intragenomic changes associated with members of an interspersed repetitive sequence family, termed RSR for repeat-spacer-repeat, in the genome of the sea urchin *Strongylocentrotus purpuratus*. The prototype RSR family member has a transcriptional enhancer within its "S" region that is chiefly responsible for regulating the spatial and temporal expression of the aboral ectoderm-specific *Spec2a* gene during *S. purpuratus* embryo development (Gan et al., 1990a, 1990b; Angerer et al., 2001; Yuh et al., 2001). *Spec2a* is a member of a small gene family that encodes intracellular EF-hand calcium-binding proteins related to calmodulin (Brandhorst and Klein, 1992, 2002). RSR enhancers are found associated with all members of the *Spec* gene family as well as with non-*Spec* genes (Dayal et al., 2004). However, only the *Spec2a* RSR enhancer confers aboral ectoderm-specific expression; other *Spec* genes are also expressed exclusively in the aboral ectoderm but require other regulatory sequences outside of their RSR enhancers to achieve proper spatial expression (Gan et al., 1990a, 1990b; Gan and Klein, 1993; Brandhorst and Klein, 1992, 2002). The *Spec2a* RSR enhancer has been optimized over evolutionary time by the recent acquisition of new *cis*-regulatory elements that function mainly to repress *Spec2a* expression in non-aboral ectoderm cells (Yuh et al., 2001; Dayal et al., 2004). Most notably, a single base-pair change has led to the appearance of a novel *cis*-regulatory element capable of binding to three distinct transcription factors, SpOtx, SpGoosecoid (SpGsc), and SpGataE. SpGataE is a key transcription factor in the endomesoderm gene regulatory network (Davidson et al., 2002), and we have shown that it represses *Spec2a* expression in endoderm cells (Kiyama et al., 2005; Kiyama and Klein, 2007).

We have hypothesized that RSR repetitive sequence family arose through amplification-dispersal-divergence mechanisms at the emergence of the Strongylocentrotidae family (Dayal et al., 2004; Villinski et al., 2005). RSRs are found in the genome of *Strongylocentrotus franciscanus*, which diverged from *S. purpuratus* 13-18 million years ago, but are not found in the genomes of *Helliocidaris erythrogramma* and *Helliocidaris tuberculata*, which diverged from *S. purpuratus*, 30-35 million years ago or *Lytechinus pictus*, which diverged from *S. purpuratus* 35-50 million years ago (Dayal et al., 2004; Villinski et al., 2005). We have further posited that a genomic region containing a transcription enhancer within what is now the S region was amplified along with adjacent sequences to generate the RSR family in an ancestral species that gave rise to the Strongylocentrotidae. An RSR family member may have inserted upstream of an ancestral *Spec* gene to become associated with the expression of this gene family (Dayal et al., 2004). In the case of *Spec2a*, the RSR enhancer acquired aboral ectoderm specificity through the creation of the Otx-Gsc-GataE *cis*-regulatory element. It is possible that other RSR family members near genes other than *Spec* genes also retained their enhancer activities and were optimized under different selective pressures governed by their genomic location. Conversely, some RSRs might not be under selective pressure to retain enhancer activity. The RSR family therefore provides a novel means to investigate the function and evolution of *cis*-regulatory elements within a single genome. A comprehensive analysis of the RSR

repetitive sequence family is now feasible as the *S. purpuratus* genome has been sequenced in its entirety and extensively annotated (Sea Urchin Genome Sequencing Consortium, 2006).

In this report, we surveyed the *S. purpuratus* genome and identified 274 S regions belonging to members of the RSR repetitive sequence family. The S regions display a broad continuum of sequence divergence that roughly separate into high and low divergence classes. Alignments of 52 S regions most related to that of *Spec2a* reveal a complex pattern of rearrangements, insertions and deletions (indels), and base-pair changes. We constructed a distance tree for the 52 S regions and correlated their positions on the tree with their enhancer activity. Unexpectedly, we find a wide range of activities for all branches of the tree. Notably, S regions lacking the essential *cis*-elements found in the *Spec2a* enhancer still have strong activity. We identify short, highly conserved motifs within the S region that may represent novel *cis*-regulatory elements. Many RSRs with enhancer activity are found nearby genes, suggesting that they are involved in regulating gene expression. The results show that the RSR family members are rapidly evolving in the *S. purpuratus* genome and may serve as a renewable pool of transcriptional enhancers that contribute to the regulation of a large and functionally diverse set of genes.

2. Materials and Methods

2.1 Surveying the genome for RSRs and S regions and constructing a divergence plot for S regions

The sequence from the S region of *Spec2a* was searched against the *S. purpuratus* genome (Spur 1.1 assembly released on 04/13/2005; <http://www.hgsc.bcm.tmc.edu/projects/seaurchin/>) using BLASTN. We used the default parameter settings except we did not use any low-complexity filter. Instead of using megablast or discontinuous megablast, we chose BLASTN for its higher sensitivity because of its shorter word length compared to megablast. After obtaining the BLASTN results, we performed the following filtering and grouping: (1) filtered out matches with E-values > 1; (2) separated matches according to whether a match was on the forward or reverse strand of the genome sequence; (3) if small matched blocks were close to each other (< 250bp), and matched to different parts of the S region, we grouped them together to represent a gapped match compared to the S region as a whole. A sequence with a large match to the S region was placed into a matched group by itself; (4) after grouping by these criteria, we removed matched sequences that covered less than 30 bp of the S region. This left 274 matched groups.

We defined the divergence score (DS) for each matched groups as the following: $DS = \text{number of matched base pairs} = \text{total alignment length} - \text{number of mismatches} - \text{number of gaps}$.

The 274 S regions were then graphically sorted from highest to lowest DS with each S region assigned an index number ranging from highest (#1) to lowest (#274) DS (Fig. 1).

2.2 Aligning S regions and constructing an S region distance tree

Alignments and distance trees were constructed only for the top 52 matches according to their DS. In some of the matched groups, two blocks were found to be reversed with respect to the *Spec2a* S region (see Fig. 2B). We artificially switched the blocks in order to construct meaningful alignments and distance trees.

ClustalW software (<http://www.ch.embnet.org/software/ClustalW.html>) was used for multiple alignment using the full sequence option, DNA identity matrix. The gap open penalty was set to 15-gap extension penalty to 6.66, and gap separation penalty to 4.

We used the DNA maximum likelihood program (<http://cmgm.stanford.edu/phyliip/dnaml.html>) to generate a sequence distance tree for the top 52 matches. This program implements the maximum likelihood method for DNA sequences. Parameters used were: transition/transversion ratio, 2.0; time of randomization of input order of sequences, 10.

2.3 Determining enhancer activities of S regions

S regions from *S. purpuratus* were cloned by PCR using genomic DNA as a template. Luciferase reporter constructs were created by inserting cloned S regions into the *MluI-KpnI* restriction enzyme sites of the *endo16-luciferase* reporter construct described previously (Dayal et al., 2004). Luciferase reporters were introduced by particle gun as described previously (Villinski et al., 2005). Luciferase activities were measured in extracts of blastula stage embryos. Luciferase activities were normalized as described previously. Normalized luciferase activities were standardized to the activity of the luciferase activity from the *endo16-luciferase* reporter construct (Villinski et al., 2005).

2.4 Contrast counting to identify conserved motifs in S regions

For each possible oligomer of length 6-9 bp, we generated all possible motifs, allowing up to 3 IUPAC characters (to count the degenerative motifs). There are approximately 10^9 such degenerate 9-mer motifs with up to 3 IUPAC characters. We counted motifs for both the forward strand and its reverse complement and looked for extreme imbalance in the motif count in the top 53 DS ranked sequences as compared to a random control. The random controls were 53 sequences of the same length as the sample sequences, each cut from a random position in the contig containing the sample sequence. We made certain that the control sequences were not contained within gene-encoding sequences and did not overlap with the sample sequences. After finding the highest ranking motif, which was defined as the one with the best p-value by a chi-square or Fisher exact test, we masked it and its reverse complement so that any sequence overlapping with the motif would not be used in the next counting round. We masked the control by the same procedure.

2.5 Searching for annotated genes near S regions

In order to obtain gene annotations for genomic regions containing S regions, we first identified the contig/scaffold and location for each S region obtained from the *S. purpuratus* genome sequence annotation database (<http://www.hgsc.bcm.tmc.edu/projects/seaurchin/>). We then searched 100 kb upstream and 100 kb downstream for all annotated genes in that 200 kb of genomic sequence.

3. Results

3.1 Surveying the genome for RSRs and S regions and constructing a divergence plot for S regions

Initially, we performed BLAST searches using the entire *Spec2a* RSR as bait against the *S. purpuratus* genome sequence database to identify other RSR sequences. However, we found that RSRs exhibited a great deal of internal sequence rearrangements and large variations in indels, making it difficult to identify beginning and end points for the R regions. In fact, some R regions were found without an adjacent "...SR" or "RS..." portion. An additional complication was that the R regions are divergent direct repeats making interpretations of sequence alignments problematic. However, our knowledge of crucial features of *Spec2a*'s

enhancer allowed us to partially alleviate these problems. We had previously demonstrated that the core enhancer activity associated with *Spec2a* aboral ectoderm-specific expression was a 78-bp sequence extending from base-pair 44 to base-pair 122 in the 251-bp-long S region of the *Spec2a* RSR (Mao et al., 1994; Yuh et al., 2001). The core enhancer contains four *cis*-regulatory elements sufficient for aboral ectoderm expression, namely, proximal Otx-Gsc, OER, CBF, and Otx-Gsc-GataE (Mao et al., 1994, 1996; Yuh et al., 2001; Kiyama et al., 2005; Kiyama and Klein, 2007). Importantly, the *Spec2a* S region contains the Otx-Gsc-GataE *cis*-regulatory element, TAATCT, while most other RSR S regions differ from this motif by a single base pair having a consensus sequence TGATCT (Dayal et al., 2004; Villinski et al., 2005). In contrast to TAATCT, TGATCT cannot efficiently bind to SpOtx, SpGsc, or SpGataE (Dayal et al., 2004; Kiyama et al., 2005; Kiyama and Klein, 2007). Moreover, we found that S regions are far more conserved than complete RSR sequences. We therefore refocused our BLAST searches using the *Spec2a* S region as bait. We used six different E-value cutoff points to capture S region sequences; $E = 10^0$ captured 274 sequences, $E = 10^{-1}$ captured 228 sequences, $E = 10^{-2}$ captured 212 sequences, $E = 10^{-3}$ captured 198 sequences, $E = 10^{-4}$ captured 163 sequences, and $E = 10^{-5}$ captured 132 sequences. Divergence plots generated for all E values had similar features, displaying three distinct slopes. Figure 1 shows the divergence plot for $E = 10^0$ with its set of 274 S regions. The steepest slope corresponded to seven S regions most closely related to the S region of *Spec2a*. These S regions may therefore share close evolutionary ancestry with the *Spec2a* S region, although unexpectedly, none represented S regions associated with other *Spec* genes. The divergence plot revealed that the S regions fell into two general classes; those with high divergence scores (HDS), representing 52 S regions (including the most highly conserved S regions), and those with low divergence scores (LDS) representing the remaining 222 S regions (Fig. 1). The LDS sequences ranged from ones that easily aligned with HDS sequences to very divergent sequences with limited similarity to the *Spec2a* S region and other HDS S regions. The results demonstrated that even by limiting the search to the S sequences, we could not identify a distinct set of S regions. Rather, a broad continuum of sequences was found, suggesting that many S regions were under very mild, if any, evolutionary constraints.

3.2 Aligning S regions and constructing an S region distance tree

We chose to focus in more detail on the 52 S regions most closely related to the *Spec2a* S region (HDS). Scaffold group ID numbers showing the genomic location for each S region sequence is provided in supplemental Table S1. Alignments of the S region sequences showed that all matched at least part of the *Spec2a* 78-bp core enhancer sequence, although the overlapping portion was not always continuous and displayed frequent occurrences of indels (Fig. 2A). Other portions of the S regions outside of the 78-bp core enhancer also showed indel variations (Fig. 2A). In addition, we identified an unusual rearrangement that further separated the HDS S regions into two groups. Thirty of the S regions aligned sequentially with the *Spec2a* S region, as shown in the top portion of Figure 2B. However, two blocks of sequences within twenty-two of the S regions were reversed with respect to the *Spec2a* S region, as shown in the bottom portion of Figure 2B. This rearrangement needed to be taken into account for meaningful construction of distance trees for the HDS S regions.

Sequence alignment of the 30 sequential S regions, including the *Spec2a* S region, revealed extensive blocks of highly conserved sequences (Fig. 3A). High sequence similarity was found both within and outside of the 78-bp core enhancer of *Spec2a*. The majority of the sequential (forward orientation) S region sequences were continuous with the *Spec2a* S region sequence, although a few contained insertions of varying lengths, as evidenced by common gaps in most of the other sequences, including *Spec2a* (Fig. 3A). Of the four *cis*-

regulatory elements found in the *Spec2a* 78-bp core enhancer, the proximal Otx-Gsc element was conserved in about 50% of the S regions, while the other elements were not well conserved (Fig. 3A). Notably, the Otx-Gsc-GataE element was not conserved. In *Spec2a*, the “TCT” base pairs of the TAATCT element appeared to be generated by an insertion of 7 base pairs not found in most of the other HDS S regions (Fig. 3A). Moreover, as shown previously, most S regions contained a G rather than A in the corresponding second position of the TAATCT element, resulting in low affinity binding of the SpOtx, SpGsc and SpGataE factors (Fig. 3A; Dayal et al., 2004; Kiyama et al., 2005; Kiyama and Klein. 2007).

Sequence alignment of the 22 reverse S regions showed extensive similarity among these sequences with less frequent occurrence of indels than found in the sequential S regions (Fig. 3B). A large indel was found at the beginning of the S region where nine sequences had 55-bp, 57-bp, or 60-bp insertions with respect to the other thirteen (Supplemental Fig. S1). For ease of comparison, Figure 3B shows the alignment of the reverse S regions with this indel omitted. The extensive sequence similarity suggested that the reverse S regions stemmed from a recently rearranged sequential S region sequence that amplified to give rise to a subgroup of closely related S regions. This subgroup also contains a G rather than an A corresponding to the second position of the TAATCT element (Fig. 3B). The proximal Otx-Gsc element (TAATCC) was not found any of the reverse S region sequences further supporting their recent common origin (Fig. 3B, Supplemental Fig. S1).

In order to generate a meaningful sequence distance tree for both the HDS sequential and HDS reverse S regions, we artificially switched the reverse S region sequences to align with the sequential S region sequences (Fig. 3C). This allowed us to evaluate the sequence relationships of all 52 HDS S regions in a single tree. We generated the distance tree by maximum likelihood, which is based on relative sequence distance (Fig. 4). The tree displayed widespread branching with several distinct sequence branches. Unexpectedly, *Spec2a* and *Spec1* S regions mapped to different branches, indicating that a few non-*Spec* S regions were more closely related to the *Spec2a* S region than was the S region of *Spec1* (Fig. 4). The branching pattern indicated that that 6 non-*Spec* gene S regions were more similar in their sequences to that of the *Spec2a* S region than was the *Spec1* S region. The evolutionary significance of this observation is not clear. The tree also shows that the reverse S regions clustered as two closely related branches, supporting the notion that they derived from a very recent rearrangement-amplification event (Fig. 4). At the extreme, we found two sequential S regions, 6a and 7a, that were widely separated from the other S regions, notably from that of the S region of *Spec2a* (Fig. 4).

3.3 Determining enhancer activities of S regions

The distance tree provided a platform on which the relationships between S region sequences and their enhancer activities could be determined. At one extreme, we might expect to find enhancer activities only for S region sequences closely related to that of the *Spec2a* S region. Alternatively, enhancer activities might be associated with all S regions. Furthermore, knowledge of the four *cis*-regulatory elements in the 78-bp core enhancer of *Spec2a* allows us to determine the extent to which enhancer activities from other S regions depend on these elements. We selected ten S regions; 8 represented different branches of the distance tree (including *Spec2a* and *Spec1*) and 2 represented LDS S regions not represented in the tree (Figs. 4, 5). The S regions were cloned into a vector containing the minimal *endo16* promoter and *luciferase* reporter gene. Enhancer activities were determined by transfection of the enhancer constructs into HEK 293 cells.

Figure 5 shows that enhancer activities were found throughout the distance tree and did not correlate with closeness to the *Spec2a*'s S region sequence. Some S region sequences mapping near *Spec2a* did show high activity. In fact, S region 44a had three-fold higher

activity than *Spec2a*'s S region (Fig. 5). However, S region 48a had only modest enhancer activity. S regions 5a and 6a, more distantly related to the *Spec2a* S region, did not contain significant activity above the empty vector control, but S region 45a, also a more distant sequence, had significantly greater activity than *Spec2a*. S regions outside of the tree showed both high (LDS1) and low (LDS2) activity. S region LDS2 is located 1.1-kb upstream of a gene encoding an α -tubulin (Scaffold47534). As we showed in a previous report, the S region from *Spec1* (4a) yielded approximately 50% of the activity observed with *Spec2a* (Fig. 5; Dayal et al., 2004).

The proximal Otx-Gsc and Otx-Gsc-GataE elements both bind to the positive regulator SpOtx, which is the major positive transcription factor responsible for *Spec2a*'s S region enhancer activity (Gan et al., 1990a, 1990b; Gan and Klein, 1993). However, we found no correlation between the number of Otx elements and S region enhancer activity (Fig. 5). For instance, S region 44a has no Otx elements within its sequence and yet showed strong enhancer activity, while S region 5a, has two Otx elements and showed very little activity (Fig. 5).

Despite lack of correlation with enhancer activity, HDS S regions, nonetheless, have strongly conserved Otx-Gsc elements at or very near the position of the corresponding proximal Otx-Gsc element in the S region of *Spec2a* (Fig. 6). Most, though not all, S regions contain this element. The fact that its position varies slightly might be an indication of stabilizing selection in an enhancer element wherein a second site appears letting the first site degenerate (Ludwig *et al.*, 2000). Because the proximal Otx-Gsc element can bind to both a positive regulator (SpOtx) and a negative regulator (SpGsc) (Angerer et al., 2001), the precise function of the conserved proximal Otx-Gsc element in the S regions is difficult to predict. We also observed dual copies of the proximal Otx-Gsc elements in several S region sequences (Fig. 6). Together, the results suggest a highly dynamic pattern of Otx-Gsc element gain and loss in the *S. purpuratus* genome over evolutionary time.

3.4 Contrast counting to identify conserved motifs in S regions

The set of 52 HDS S region sequences provided the opportunity to identify conserved sequence motifs using an unbiased approach. To do this, we developed an algorithm in which we counted the number of times a 6-9 bp motif was found in the HDS S regions compared to random controls. In this case we arbitrarily used the highest DS ranked 53 S regions rather than 52. Control sequences were selected at random from the vicinity of each RSR and outside of any significant open reading frame. We generated all possible sequence motifs for a 6-9 bp oligonucleotide allowing up to three degenerate characters. For example, approximately 10^9 9-mer motifs are possible with three degenerate characters. We counted motifs in forward and reverse orientation and searched for extreme imbalances between the HDS S region set and the random control set. After finding the most frequent motif, we masked the forward and reverse sequence in the HDS S region set and random control set and reiterated the search process. The search was run seven times, at which point motifs with the most significant p values for S regions versus controls were identified, and after which p values became much less significant. Table 1 shows the results from the first six runs. The most frequent motif, AGCAAGA, and its reverse complement, TCTTGCT, was found 100 times in the HDS S regions but only 2 times in the random sequence controls of the same length (Table 1). The next five most frequent motifs also showed extreme deviations from the random controls (Table 1). Although the proximal Otx-Gsc (TAATCC) and Otx-Gsc-GataE (TAATCT) elements together occurred 20 times in the HDS S region set, these motifs did not represent the most frequent using this unbiased approach.

Of the six most frequent motifs, the fourth highest occurring motif (reverse sequence AAAGGY) had similarity to the binding sites of transcription factors, Dof2, Dof3, MNB1A, and PBF (AAAGCY).

3.5 Searching for annotated genes near S regions

As a first step to determine whether any individual S regions play roles as transcriptional enhancers *in vivo*, we searched the *S. purpuratus* genome database to identify annotated genes that had an S region within 20 kb of their translational start sites. As expected, the search yielded the S regions associated with *Spec2a* and *Spec1* (Table 2). Nine other S regions were found upstream of annotated genes at distances ranging from 1.2 to 14 kb (Table 2). In a few cases more than one gene was associated with an individual S region over a 20-kb stretch. In the extreme case, S region 45a was near 12 different genes, although it was closest, 4.8 kb upstream, to *Sp-Slc10a2*, which encodes a putative Ileal sodium/bile acid cotransporter (Table 2). The genes near S regions had no obvious relationships to one another, either structurally or functionally, and it was likely that they represented a random cross section of genes with diverse spatiotemporal expression patterns and functions.

Enhancer activities for three of the nine non-*Spec* S regions shown in Table 2 were determined in our HEK 293 cell transfection assay (see Fig. 5). These were: 25a, 14 kb from *Sp-122B*, encoding a putative transmembrane protein; 44a, 1.2 kb from *Sp-SRCR-72*, encoding a putative scavenger receptor cysteine-rich domain protein; and 45a, 4.8 kb from *Sp-Slc10a2*, encoding a putative Ileal sodium/bile acid cotransporter. These S regions showed significant enhancer activity over the empty vector control, suggesting that these sequences had the capability of contributing to the transcriptional output of the nearby genes.

4. Discussion

By focusing on the S regions of *S. purpuratus* RSR repetitive sequence family members that are most closely related to the S region of *Spec2a*, we were able to substantially extend our knowledge of RSR structure, enhancer function and genomic organization, as well as provide broader insights into the intragenomic evolution of an unusual short interspersed repetitive sequence element. Unlike mammals where a few short interspersed repeat families tend to dominate the genome, *S. purpuratus* has hundreds of short interspersed repetitive sequence families in its genome (Cameron et al., 2000). We found that S region sequences are highly diverse and appear to be evolving rapidly by rearrangements, indels, and base-pair changes. RSRs are only found in species within the family Strongylocentrotidae and we have shown that in *S. franciscanus*, which diverged from *S. purpuratus* 13-18 million years ago, RSRs have very similar properties (Villinski et al., 2005). It is not known if RSR elements still have mechanisms that permit their excision and reinsertion at different genomic regions or alternatively, they are static in the genomes of extant Strongylocentrotidae species. Other mobile and transposable elements are known to exist in the *S. purpuratus* genome (Britten, 1996). If excision-reinsertion mechanisms still exist for RSR elements, it would suggest that events much like the one we postulated for RSR insertion into an ancestral *Spec* gene at the emergence of the Strongylocentrotidae family are actively occurring within the genomes of extant species.

Many, but not all, of the S regions contain significant enhancer activity. Since it is unlikely that so many RSRs acquired enhancer activity independently, the RSR family is more likely to have amplified from a segment of DNA harboring an enhancer. If true, some of the original *cis*-regulatory elements from a putative ancestral enhancer should have been conserved. One of these is likely to be the proximal Otx-Gsc (TAATCC) element since it is present in the many of the S regions that we examined. Notably, this element is not found in

the HDS reverse S region branch. Although we did not determine the enhancer activity for any reverse S region sequences, some S regions with high enhancer activity do not have Otx-Gsc binding elements, while others with little activity do. SpOtx and SpGsc have antagonistic effects at the TAATCC site (Angerer et al., 2001) making it difficult to identify the precise role an enhancer with a proximal Otx-Gsc element has for any individual S region sequence. In principle, by analyzing S regions with enhancer activity from *S. purpuratus* and other Strongylocentrotidae species, it should be possible to identify core conserved motifs that could represent *cis*-regulatory elements from the ancestral enhancer. Our initial results identify several highly conserved motifs and it would be informative to determine whether these were conserved in the S regions containing of other species.

The *Spec2a* RSR enhancer directs aboral ectoderm-specific expression largely through its Otx-Gsc-GataE element (Dayal et al., 2004; Kiyama et al., 2005; Kiyama and Klein, 2007). S regions from other *Spec* genes do not contain this sequence motif, the difference represented by a single base-pair change (Dayal et al., 2004). These results emphasize the ease in which spatial expression patterns can be altered in major ways by minor changes in *cis*-regulatory elements. Preliminary experiments suggest that non-*Spec* RSR enhancers thus far analyzed are not cell type-specific (Kiyama and Klein, unpublished results). However, this small sampling does not rule out that some RSR enhancers may have acquired cell type-specificity, as did the *Spec2a* enhancer. Stabilizing selection for promoter/enhancers appears to be a widespread phenomenon (Ludwick, et al., 2000). Thus, if an S region with general enhancer activity inserted near a cell type-specific gene, it is possible that minor sequence alterations capable of modifying the repertoire of transcription factors that bind to it might be fixed, resulting in cell type-specificity and thus permitting the original *cis*-regulatory elements to diverge.

RSR elements appear to be widely distributed over the *S. purpuratus* genome. Many containing enhancer activity are near enough to expressed genes to participate in their transcription. The RSR enhancers may be constantly undergoing refinement in their activities based on the selective pressures associated with their location in the genome. Assuming that RSRs had or now have the ability to excise and reinsert into different genomic locations, they may thus serve as a renewable pool of transcriptional enhancers for the *S. purpuratus* and other Strongylocentrotidae genomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Ning Zhang for important technical support and the M. D. Anderson Cancer Center DNA Analysis Facility, supported in part by a National Cancer Institute Center Core Grant (CA016672), for DNA sequencing. We especially acknowledge the valuable advice and guidance provided to us from Yufeng Shen and Erica Sodergren at the Human Genome Sequencing Center, Baylor College of Medicine. This study was supported by a National Institute of Child Health and Human Development Grant (HD022619) to W.H.K. and by the Robert A. Welch Foundation (G-0010).

Abbreviations

RSR	repeat-spacer-repeat
indels	insertions and deletions
DS	divergence score

HDS high divergence score
LDS low divergence score

References

- Angerer LM, Oleksyn DW, Levine AM, Li X, Klein WH, Angerer RC. Sea urchin goosecoid function links fate specification along the animal-vegetal and oral-aboral axes. *Development*. 2001; 128:4393–4404. [PubMed: 11714666]
- Brandhorst BP, Klein WH. Molecular patterning along the sea urchin animal-vegetal axis. *Int. Rev. Cytol.* 2002; 213:183–232. [PubMed: 11837893]
- Brandhorst BP, Klein WH. Territorial specification and control of gene expression in the sea urchin embryo. *Semin. Dev. Biol.* 1992; 3:175–186.
- Britten RJ. DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* 1996; 93:9374–9377. [PubMed: 8790336]
- Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, Swartzell S, Wallace JC, Poustka AJ, Livingston BT, Wray GA, Etensohn CA, Lehrach H, Britten RJ, Davidson EH, Hood L. A sea urchin genome project: sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. U. S. A.* 2000; 97:9514–9518. [PubMed: 10920195]
- Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*. 2008; 134:25–36. [PubMed: 18614008]
- Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science*. 2006; 311:796–800. [PubMed: 16469913]
- Dayal S, Kiyama T, Villinski JT, Zhang N, Liang S, Klein WH. Creation of *cis*-regulatory elements during sea urchin evolution by co-option and optimization of a repetitive sequence adjacent to the *spec2a* gene. *Dev. Biol.* 2004; 273:436–453. [PubMed: 15328024]
- Gan L, Klein WH. A positive *cis*-regulatory element with a bicoid target site lies within the sea urchin *Spec2a* enhancer. *Dev. Biol.* 1993; 157:119–132. [PubMed: 8097732]
- Gan L, Wessel GM, Klein WH. Regulatory elements from the related *spec* genes of *Strongylocentrotus purpuratus* yield different spatial patterns with a *lacZ* reporter gene. *Dev. Biol.* 1990a; 142:346–359. [PubMed: 1701736]
- Gan L, Zhang W, Klein WH. Repetitive DNA sequences linked to the sea urchin *spec* genes contain transcriptional enhancer-like elements. *Dev. Biol.* 1990b; 139:186–196. [PubMed: 2328835]
- Kiyama T, Klein WH. SpGataE, a *Strongylocentrotus purpuratus* ortholog of mammalian Gata4/5/6: protein expression, interaction with putative target gene *spec2a*, and identification of friend of Gata factor SpFog1. *Dev. Genes Evol.* 2007; 217:651–663. [PubMed: 17710433]
- Kiyama T, Zhang N, Dayal S, Yun Lee P, Liang S, Villinski JT, Klein WH. *Strongylocentrotus purpuratus* transcription factor GATA-E binds to and represses transcription at an Otx-Goosecoid *cis*-regulatory element within the aboral ectoderm-specific *spec2a* enhancer. *Dev. Biol.* 2005; 280:436–447. [PubMed: 15882584]
- Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*. 2000; 403:564–567. [PubMed: 10676967]
- Mao CA, Gan L, Klein WH. Multiple Otx binding sites required for expression of the *Strongylocentrotus purpuratus Spec2a* gene. *Dev. Biol.* 1994; 165:229–242. [PubMed: 8088441]
- Mao CA, Wikramanayake AH, Gan L, Chuang CK, Summers RG, Klein WH. Altering cell fates in sea urchin embryos by overexpressing SpOtx, an orthodenticle-related protein. *Development*. 1996; 122:1489–1498. [PubMed: 8625836]
- Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM. *cis*-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*. 2007; 131:1041–1043. [PubMed: 18083094]
- Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature*. 2006; 440:1001–1002. [PubMed: 16625184]

- Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science. 2006; 314:941–952. [PubMed: 17095691]
- Villinski JT, Kiyama T, Dayal S, Zhang N, Liang S, Klein WH. Structure, expression, and transcriptional regulation of the *Strongylocentrotus franciscanus spec* gene family encoding intracellular calcium-binding proteins. Dev. Genes Evol. 2005; 215:410–422. [PubMed: 15871032]
- Yuh CH, Li X, Davidson EH, Klein WH. Correct expression of *spec2a* in the sea urchin embryo requires both Otx and other *cis*-regulatory elements. Dev. Biol. 2001; 232:424–438. [PubMed: 11401403]

Figure 1.

Divergence plot of S regions identified in the *S. purpuratus* genome. The plot represents 274 S regions with $Ev = 1$. Divergence score is defined as: $DS = \text{number of matched base pairs} = \text{total alignment length} - \text{number of mismatches} - \text{number of gaps}$. Index refers to the individual S region group ID. The Group IDs for the 52 HDS S regions are listed in Supplemental Table S1.

B

Figure 2.

Schematic representation of sequence alignments for the 52 HDS S regions. (A) Sequence alignments. The *Spec2a* S region is shown twice as *Spec2a* and 1a. The red line within the *Spec2a* sequence indicates the 78-bp core enhancer. The purple-shaded area shows the comparable sequence alignments of *Spec2a* with the corresponding regions for the other HDS S regions. Gaps in sequence relative to *Spec2a* are shown as broken lines. (B) Schematic comparison of forward and reverse orientation for HDS S regions. The diagram depicts the orientation of two sequence blocks for 30 HDS S regions (red and blue) that are reversed in 22 HDS S regions.

Figure 3. Nucleotide sequence alignments of the 30 forward (A), 22 reverse (B) and combined (C) HDS S regions. *Spec2a* S region is numbered 1a (sixth row). Individual nucleotides are color coded for ease of visualization. Dashed lines represent gaps. The overall consensus sequence is shown as a histogram and the consensus nucleotides are shown beneath the histogram. The *cis*-regulatory elements found in the 78-bp core enhancer of *Spec2a* are shown as colored bars underneath the consensus sequence. Red, proximal Otx-Gsc and Otx-Gsc-GataE elements; blue, OER element; green, CBF element. Reverse S region alignments in (B) are shown without the first large indel. Reverse S region alignments containing this indel are provided in Supplemental Figure S1. Sequences in (C) were aligned by artificially reversing the reverse orientation sequence blocks in the 22 HDS S regions.

Figure 4. Distance tree for the 52 HDS S regions. The tree was constructed using the maximum likelihood algorithm. Red arrows point to the sequences used for enhancer activity assays.

Figure 5.

Transcriptional enhancer activity of S regions. Luciferase activities were obtained from blastula stage embryo extracts resulting from introducing the *endo16-luciferase* reporter gene constructs into unfertilized eggs. Activities from firefly luciferase are normalized to a co-introduced Renilla luciferase and standardized with the activity from the basal *endo16* promoter set to 1 (basal promoter). LDS1 and LDS2 refer to two low divergence score S regions (group IDs scaffold59358 and scaffold45734, respectively). Numbers on the bottom refer to the number of Otx-Gsc elements found within an individual S region sequence.

Figure 6. Occurrence and relative positions of Otx-Gsc elements within S regions. The nucleotide position with respect to the *Spec2a* (1a) S region is on the Y-axis.

Table 1

High frequency motifs present in the HDS S regions.

Rank	Forward motif	Reverse motif	S region frequency	Adjacent S region frequency
1	AGCAAGA	TCTTGCT	100	2
2	TTCABTGAA	TTCAVTGAA	80	0
3	GTDGTT	AACHAC	139	10
4	RCCTTT	AAAGGY	100	6
5	AAGCTT	AAGCTT	76	4
6	AGCAGA	TCTGCT	49	0

Forward and reverse motifs refer to the forward 5' to 3' sequence and the reverse complementary 5' to 3' sequence. In the analysis we arbitrarily used the top 53 DS S regions rather than 52. Sequencekey: B = not A (i.e., C, G, or T); V = not T; D = not C; H = not G; R = A or G; Y = T or C.

Table 2
Identification of annotated genes neighboring S regions in the *S. purpuratus* genome

S region	Number of neighboring genes	Distance from translational start site (kb)	Closest gene	Closest gene description	Genomic location
1a	1	0.2	<i>Spec2a</i>	EF hand protein	Upstream
4a	4	0.5	<i>Spec1</i>	EF hand protein	Upstream
14b	1	2.6	<i>SpPor</i>	NADPH cytochrome P450 reductase	Upstream
16a	3	9.6	<i>SpOpsin2</i>	Rhodopsin family member	Downstream
25a	1	14.0	<i>Sp122B</i>	Transmembrane protein 16	Sixth intron
29a	1	3.3	<i>SpFoxQ2</i>	Fox head domain protein	Downstream
38b	1	11.0	<i>SpLph</i>	Lactase phlorizin hydrolase	Upstream
40a	2	3.3	<i>SpAgpat3</i>	1 - acetyl-sn-glycerol-3-phosphate acyltransferase gamma	Second intron
41a	4	4.8	<i>SpRdh8</i>	Retinol dehydrogenase 8	Downstream
44a	1	1.2	<i>SpSRCR-72</i>	Scavenger receptor cysteine-rich domain	Upstream
45a	12	4.8	<i>SpSlc10a2</i>	Ileal sodium/bile acid cotransporter	Upstream