Genomic and cis-regulatory analyses of *Ciona intestinalis* phase II genes: Glutathione S-transferases

Gerardo M. Nava[1], Joseph Aman[2], Javier H. Ospina[1],
Clare B. Congdon[3], Carolyn Mattingly[2] & H. Rex Gaskins[1]
[1]University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672
[3]Colby College, Waterville, ME 04901

Ciona intestinalis (hereafter *Ciona*) is an invertebrate member of the chordate clade Urochordata (tunicates), which diverged from the last common ancestor of all chordates at least 520 million years ago[14]. Their critical evolutionary position as basal chordates and the simplicity of their embryogenesis have attracted developmental and evolutionary biologists since the turn of the 20th century. Because of this, the genomes of *Ciona* as well as its close relative *Ciona savignyi* were recently sequenced[3]. Adult animals are filter feeders and live attached to submerged substrates in marine waters where they encounter high concentrations of complex polyphenols, halogenated aromatics, methylated sulfides, and some heavy metals. Essentially nothing is known about the mechanisms by which tunicates detoxify or otherwise tolerate these natural toxins in the marine environment. Our recent work is consistent with the hypothesis that *Ciona* relies, to some extent, on bacteria associated with the branchial sac for detoxification of natural marine toxins.[6,7] Accordingly, we further hypothesize that: 1) reliance on bacteria for detoxification has influenced over evolutionary time, the complement of *Ciona* detoxification genes, and 2) given the extent of divergence of *Ciona* from vertebrate chordates, that functional, noncoding elements in sequences of *Ciona* detoxification genes *that are conserved* with higher vertebrates may be particularly useful for the identification of novel regulatory elements that contribute to inter-individual variation in human responsiveness to carcinogens, chemoprotective agents and chemotherapeutic drugs. At least for developmental genes, there is extreme conservation of cis-acting regulatory elements and corresponding trans-acting transcription factors among deeply divergent genomes (e.g., fly and human; ~ 700 Myr[9]). This report describes our initial efforts to determine the extent to which this principle applies to detoxification genes. Because the marine toxins that *Ciona* continuously encounters fall into chemical classes that are well characterized for their induction of phase II genes across the vertebrate chordates, we've focused first on the glutathione transferases (GST).

The *Ciona* genome is comprised of an estimated 155 million base pairs (Mb)[3]. It is approximately one-twentieth the size of the human genome and among the smallest of any experimentally accessible chordate. Approximately, 117 Mb of the *Ciona* genome is composed of nonrepetitive, euchromatic sequence, coding for approximately 16,000 genes, roughly two-thirds of the number present in the human genome. Further, *Ciona* possess approximately 136 genes per MB, while approximately 11 genes are found per Mb of the human genome. The small size of the *Ciona* genome, which provides a distinct advantage for understanding genome organization and gene function, reflects in part the fact that the urochordate lineage diverged before the extensive gene duplications that have occurred in vertebrates[3]. As such, most genes that tend to be members of multigene families in vertebrates often have just a single representative in *Ciona*. This makes *Ciona* attractive for dissecting gene-regulation networks related to cell signaling and development. Again, exploitation of this advantage of the *Ciona* genome has largely been restricted to genes encoding developmental functions.

The cytosolic GSTs have been subdivided into at least 13 distinct evolutionary classes designated Alpha, Mu, Pi, Theta, Sigma, Zeta, Kappa, Omega, Phi, Tau, Delta, Epsilon, and Beta on the basis of their primary structure, immunological properties, and substrate specificities[8]. In addition, each class consists of several isoenzymes with partially overlapping substrate selectivity. These enzymes catalyze the nucleophilic addition of the glutathione (GSH) sulfhydryl group to electrophilic centers of various physiological and xenobiotic substances rendering them more water soluble, and thereby facilitating their excretion. Numerous polymorphic alleles of human GST genes have been identified, many of which appear to segregate with susceptibility to one or more forms of cancer. Much effort is underway to better define molecular details underlying their broad specificity with regard to structurally diverse substrates and the catalytic steps involved in glutathiolation. We anticipate that this effort will be advantaged by comparative analysis of GST genes across evolutionarily divergent chordates.

Comparative-grade finished sequences and putative homologs of GSTs were retrieved from whole genomes of chordates that are phylogenetically flanked by *Ciona* and human from the Ensembl project database[1]. Target species included Human (*Homo sapiens*), Dog (*Canis familiaris*), Mouse (*Mus musculus*), Frog (*Xenopus tropicalis*), Zebrafish (*Danio rerio*), *Fugu* (*Takifugu rubripes*), and *Ciona* (*C. intestinalis*). Candidate genes were identified initially by searching reference sequences (human, mouse, and rat nucleotide and protein GSTs retrieved from the NCBI database) against all target chordate species via the BLAST similarity search of Ensembl BLASTView. Sequence alignments with highest score, E-value, percentage of identity or similar predicted protein were chosen. Reference, genomic, predicted transcripts, and predicted amino acid sequences were retrieved from Ensembl project website for additional analysis of putative homology with human, mouse, and rat genes[1]. Additionally, a systematic process was followed to identify and assemble orthologous gene sets. First, standard BLASTX and BLASTP analysis were used to back-search the full retrieved nucleotide sequence and amino acid against GenBank and GenPept, respectively. If the candidate gene sequence is indeed a homolog of the target gene, the same gene identity should be returned. Second, reverse position specific (RPS) BLAST analysis was performed with the amino acid sequences against the NCBI Conserved Domain Search[11] to detect structural and functional domains in protein sequences. Third, Smith-Waterman pairwise alignment[5] of amino acid sequences and Jalview analysis[4] were carried out to calculate percentage of identity, degree of conservation, and consensus between all putative sequences. Finally, probability of random shuffle analysis was performed for each of the putative amino acid sequences[5,12]. Using this approach of curation, which requires similarity with high quality human, rat, and mouse reference sequences, 22 putative GST sequences were identified in the *Ciona* genome, which appear to be members of the Alpha (6 sequences), Mu (6), Theta (4), Omega (4), Kappa (1) and Pi (1) families.

Two approaches were used to identify putative regulatory elements in Ciona GST Mu 1 (GSTM1) as a prototypical GST gene. First, publicly available web tools that predict cis-regulatory regions were used to test the feasibility of finding conserved transcription factor binding sites (TFBS) in phylogenetically divergent chordate genomes. Orthologous gene sets for GSTM1 were first assembled for 7 chordate species according to methodologies described in detail above. TFBSs were then predicted within the 4 kb genomic regions of GSTM1 with P-Match, a bioinformatics tool that combines pattern matching and positional weight matrix analysis with a cut-off score and matrix similarity of 97%. Thirty-eight distinct TFBSs were identified among the 7 GSTM1 sequences with approximately half of these consistently predicted in most of the chordate species (**Table 1**). Surprisingly, human and *Ciona* share 16 TFBSs within the 4 kb upstream region of GSTM1. Also of interest is the fact that none of these particular TFBSs have been investigated for GSTM1 for any species according to published literature. Similar results were also obtained with ConSite software which is based on the integration of binding site prediction generated with high-quality TFBS models and cross-species comparison filtering (phylogenetic footprinting).[13]

As a second approach we used GAMI (see reference 2 and Congdon et al., this issue), a recently-developed a non-alignment-based genetic algorithms (GA) approach that uses a fixed window size to search for putative motifs across a set of non-coding sequences. The prediction tools used in the first approach (Table 1) rely on multispecies genomic sequence alignments to locate putative regulatory motifs in conserved regions. Global sequence alignment is computationally expensive, particularly as the number and length of the sequences increases; furthermore, the approach is problematic because data from species that have diverged over evolutionary time are resistant to global alignment, especially in non-coding sequences. GAMI was used to search for 10mers in 4 kb of both 5′ and 3′ flanking sequences of GSTM1 from the seven chordate species described above. The strongest patterns (putative motifs) found are illustrated in **Figure 1**. The motifs identified by both approaches are in silico predictions and many are likely false positives[10].

Table 1: TFBSs in the upstream region (4 kb) of GSTM1 among various chordates.

	Human	Mouse	Dog	*Xenopus*	Zebrafish	*Fugu*	*Ciona*
AREB6	6	11	6	2	8	17	3
ATF	1	2	1	3	2	3	7
CDP CR3	4	3		7	3		2
CREB	3	3		6	5	6	6
c-Rel	14	25	15	23	17	15	23
E47	2	3	1		1	3	2
Evi-1	8	15	7	4	11	9	21
HFH-1	1	2	3	2	6		2
HLF	1	2	2		2	2	2
HNF-4alpha1	2	1	1				1
Nkx2-5	4	1	2	4	5	2	8
OCT-1	2	2		3	2	1	1
RORalpha1	4	3	2	3	2	4	1
STATx	1	2	2		3	4	2
XFD-2	1	1	1	1	1		3
ZID	1	2	2	1		1	1

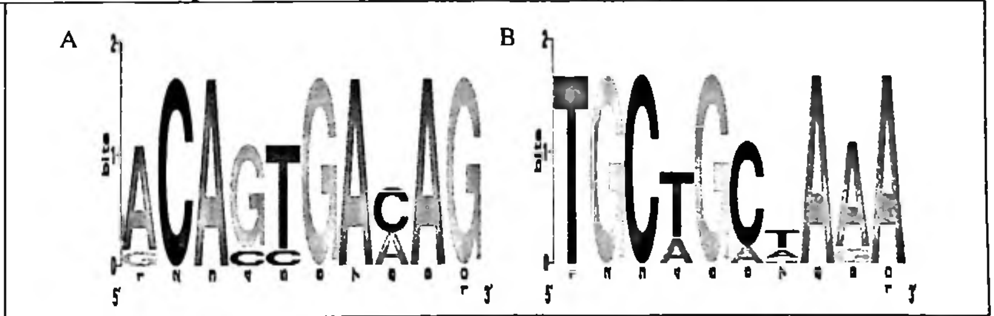

Figure 1: Example motifs found with GAMI using 4 kb up- and downstream GSTM1 sequences from human, mouse, dog, *Xenopus*, zebrafish, *Fugu*, and *Ciona* genomes. (A) 5′ upstream and (B) 3′ downstream motifs.

Two approaches are planned to further identify which (if any) of these motifs are indeed regulatory elements. First, P-Match will be used to determine if any of the GAMI predictions correspond to known TFBSs. Second, a computational approach is being developed to prioritize the corresponding human TFBSs according to the extent to which they overlap with SNPs or SNP blocks in regulatory elements of environmentally responsive genes in the human genome using the extensive data available in the dbSNP and GeneSNPs databases. Together these approaches will prioritize regulatory elements worthy of biochemical studies to examine functionality.

Supported by a New Investigator Award (HRG) with funds from MDIBL's NIEHS Center for Membrane Toxicity Studies (ES03828-19) and NIH Grant Number P20 RR-016463 from the INBRE Program of the National Center for Research Resources.

1. **Congdon C and Septor K** Phylogenetic Trees Using Evolutionary Search: Initial Progress in Extending Gaphyl to Work with Genetic Data. In *Congress on Evolutionary Computation (CEC-2003)*, Canberra, Australia. 2003.
2. **Congdon CB, Fizer CW, Smith NW, Gaskins HR, Aman J, Nava G, Mattingly C.** Preliminary Results for GAMI: A Genetic Algorithms Approach to Motif Inference. *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*: 97-104. 2005.
3. **Dehal P, Satou Y, Campbell RK, Chapman J, et al.,**. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298: 2157-2167, 2002.
4. **Fogel G and Corne D**. *Evolutionary Computation in Bioinformatics*. San Francisco, CA: Morgan Kauffmann Publishers., 2003.
5. **Frietas A**. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Berlin: Springer-Verlag., 2002.
6. **Gaskins HR, McCray NR, Collier CT, King DE, Thurmond JE and Mackie RI.** Molecular ecological and phylogenetic analyses of the intestinal microbiota of the ascidian tunicates *Boltenia echinata, B. ovifera, Halocynthia pyriformis and Ciona intestinalis. Bull. Mt. Desert Isl. Biol. Lab.* 43: 68-71, 2004
7. **Gaskins HR, Thurmond JE, Nava GM, and Mackie RI.** The *Ciona intestinalis* branchial sac and its microbiota. *Bull. Mt. Desert Isl. Biol. Lab.* 44: 80-83, 2005
8. **Hayes JD, Flanagan JU and Jowsey IR**. Glutathione transferases. *Annu Rev Pharmacol Toxicol* 45: 51-88, 2005.
9. **Jarving R, Jarving I, Kurg R, Brash AR and Samel N**. On the evolutionary origin of cyclooxygenase (COX) isozymes: characterization of marine invertebrate COX genes points to independent duplication events in vertebrate and invertebrate lineages. *J Biol Chem* 279: 13624-13633, 2004.
10. **Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N and Wasserman WW**. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2: 13, 2003
11. **Mitchell M**. *An Introduction to Genetic Algorithms*. Cambridge, MA.: MIT Press., 1996.
12. **Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET and Ruan Y**. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2: 105-111. 2005.
13. **Sandelin A, Wasserman WW and Lenhard B**. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res* 32: W249-252, 2004.
14. **Satoh N, Satou Y, Davidson B and Levine M**. *Ciona intestinalis*: an emerging model for whole-genome analyses. *Trends Genet* 19: 376-381, 2003.