Expansions in putative transcription factor gene families between Heliocidaris *erythrogramma* and Heliocidaris *tuberculata* could help to explain differential developmental patterns

Chuan-Yih Yu, Indrani Sarkar, Nathan Nehrt,
Rahul Gupta, Sashikiran Challa
School of Informatics, Indiana University

September 4, 2009

Abstract

The only two sea urchins from the Heliocidaris genus, H. *erythrogramma* and H. *tuberculata*, have drastically different developmental patterns. H. *tuberculata* goes through a typical pluteus larval stage prior to its adult form while H. *erythrogramma* skips this larval stage, directly developing into the adult form. Many previous studies have attempted to explain these differential development patterns. We utilized a dataset of approximately 17,000 ESTs generated from transcripts expressed at various points in the development of these two species and have found evidence of expansion in several gene families including expansions in two families predicted to be transcription factors. Expansions in these families may help to explain the developmental differences between the species. A secondary goal of our analysis was to identify genes included in the initial assembly of the recently sequenced California purple sea urchin, S. *purpuratus*, genome that were filtered out of the final assembly. Through a comparison of significant BLAST hits of our Heliocidaris proteins to the initial and final purple sea urchin assemblies, we identified 97 genes that should possibly be added back to the final purple sea urchin assembly.

1 Introduction

The two Heliocidaris species H. *erythrogramma* and H. *tuberculata* exhibit very different patterns of development. H. *tuberculata* has an indirect pattern of development shared by the majority of sea urchin species. It goes through a roughly 6 week feeding larval stage before developing into the typical adult sea urchin form [14]. H. *erythrogramma*, however, has a direct pattern of development similar to that of a small number of other sea urchins. It goes through a very short, roughly 4 day non-feeding larval stage before developing the typical adult form [14]. Several theories for the differential development of these two species have been proposed including a theory that certain genes needed for the development of the oral ectoderm tissue necessary for feeding during the larval stage are negatively regulated in H. *erythrogramma* thus requiring a direct pattern of development [14].

We analyzed a dataset of approximately 17,000 ESTs from H. *erythrogramma* and H. *tuberculata* obtained from the Indiana University Center for Genomics and Bioinformatics (CGB) in order to find possibly novel reasons for the developmental differences between the Heliocidaris species. After assembling the ESTs, we performed an initial round of gene finding and annotation (full details of these procedures are provided in the Data and

Methods section below). The gene finding program GLIMMER3 was used to build an initial gene set. We then ran a tblastx search of the assembled EST sequences against the S. *purpuratus* genome (version 2.1) to expand the gene set. A total of 3,306 genes were predicted for H. *erythrogramma* and 2,692 genes were predicted for H. *tuberculata* in the union of the GLIMMER3 and tblastx gene sets. Both sets of predicted genes were then translated to amino acid sequences, and the protein function prediction program PFP was used to annotate the protein sets with predicted GO terms.

We looked primarily for evidence of relative expansions in gene families between H. *erythrogramma* and H. *tuberculata* as possible causes of developmental differences between the two species. To do so, we used the California purple sea urchin, S. *purpuratus*, as a reference species. The purple sea urchin is the nearest relative of the Heliocidaris species with a fully-sequenced genome. The divergence of the Heliocidaris species from the purple sea urchin is estimated to have occurred around 35 to 40 million years ago [10]. The official genome assembly (version 2.1) of the purple sea urchin contains approximately 23,300 genes [12]. Using this total as a very rough estimate for the number of genes in each of the Heliocidaris species, our predicted gene sets contain only around 10 to 15% of the total genes in H. *erythrogramma* and H. *tuberculata*. This low coverage of the genomes makes inferences regarding differences in the two species based on relative gene gain or loss difficult to support. We therefore looked for expansions in gene families relative to the purple sea urchin or for evidence of a large number of copies of proteins (3 or more) in one Heliocidaris species relative to the other to try to overcome the effect differential sampling of the two gene sets. Through gene clustering, we found 7 instances of such expansions in our dataset. Analyses of these clusters are provided in the Results and Discussion sections below.

As a secondary goal of our research, we utilized the Heliocidaris datasets to identify potential genes filtered out of the official S. *purpuratus* genome assembly. One of the main reasons the sequencing of the purples sea urchin was undertaken is that the echinoderms are the closest known relatives of the chordates, and thus provide a useful outgroup for analyses (see phylogenetic tree in supplementary material online). The purple sea urchin's genome assembly was released in two sets. The first assembly released in April 2005 was produced via whole genome shotgun sequencing and is referred to as the NCBI version 1.1 build. Various gene prediction tools were used and the results combined to form an estimated set of approximately 28,900 genes. Meanwhile, a second assembly effort using BAC libraries was also underway. This dataset was primarily used to aid in the assembly of repeated sequences and to handle problems arising from a high degree of heterozygosity in the sea urchin genome. In July 2006, a new combined assembly was released as NCBI version 2.1. Genes were filtered from the initial version 1.1 set to remove redundant genes, fragments and pseudogenes . This filtering reduced the gene set to approximately 23,300 genes. Some have suggested that the filtering process was too strict, and that many genes were likely left out of the official gene set. We therefore performed BLAST searches of our Heliocidaris gene sets against the NCBI versions 1.1 and 2.1 gene sets in order to identify potential genes to add back to the official gene set.

2 Results

After clustering the proteins in H. *erythrogramma*, H. *tuberculata*, and S. *purpuratus*, we found 7 gene families showing evidence of expansion between H. *erythrogramma* and H. *tuberculata* as shown in Table 1. We used BLAST and Pfam searches to assign likely functions for the proteins in each of the families. Most interesting were clusters 271 and 375 which have multiple zinc finger domains and appear to be likely transcription factors. Differential gene regulation during the developmental cycle in the two species could be a very likely cause for developmental differences between the species. Another interesting cluster is a cluster which appears to be a cystatin protein family. Cystatin is an immune system receptor protein found very extensively in sea urchins as a whole. It appears there may be a duplication of a particular cystatin protein relative to the purple sea urchin and possibly in H. *erythrogramma* in relation to H. *tuberculata*, though it is

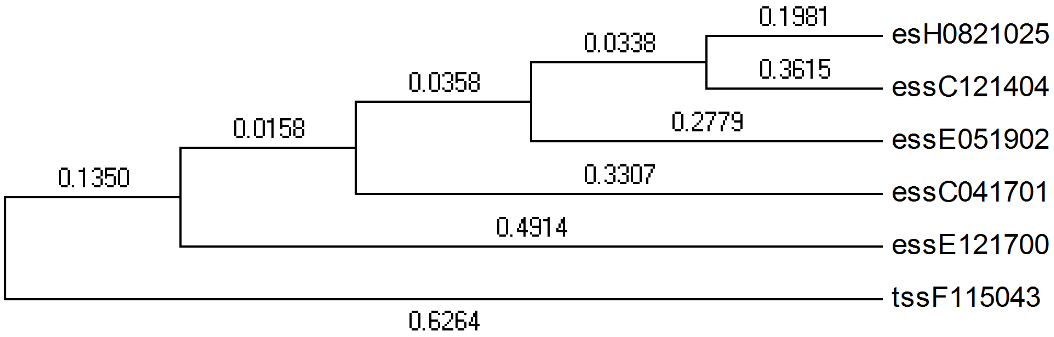

Figure 1: Cluster 271 alignment tree

also possible that our H. *tuberculata* gene set simply does not include an additional copy of the cystatin protein due to sampling differences in the EST sets.

A secondary finding of our research was the result of a comparison of BLAST searches of our predicted protein sequences from H. *erythrogramma* and H. *tuberculata* against the protein sequences from the NCBI 1.1 and 2.1 versions of the purple sea urchin genomes. We found 97 sequences with significantly similar matches (e^{-5}) in the NCBI 1.1 version of the assembly which did not have matches in the NCBI 2.1 version of the assembly. These sequences were therefore filtered out of the 2.1 version of the assembly. Our findings provide additional evidence that these genes, or a subset of the genes, should possibly be added back to the version 2.1 assembly of the purple sea urchin genome. A further analysis of these genes is provided below in the discussion section.

Table 1: Number of sequences in each species

Cluster No.	H.*ery*	H.*tub*	S.*pur*	Descript
42	2	1	1	Cystatin family (immune system)
130	7	1	2	Ca binding domain
156	2	1	1	LYR Complex 1 (NADH dehydrogenase)
271	5	1	0	Zinc finger (transcription factor)
375	3	0	2	Zinc finger (transcription factor)
516	3	0	25	Reverse transcriptase
589	2	0	1	Ambiguous

3 Data and Methods

EST generation and assembly RNA was collected by the Raff lab at Indiana University in 2006 from each species at selected points during development and was used to by the CGB to generate cDNA libraries. A quality control sample from the sequencing and annotation of the cDNA libraries revealed the genes to be highly unique; only about 40% of the cDNAs yielded significant matches (e^{-4}) to genes in existing databases. An analysis of these similar genes showed the majority to be involved in metabolic functions. A further search of the cDNAs against the S. *purpuratus* genome showed that about 78% of the Heliocidaris genes were shared with S. *purpuratus*. A combined set of 17,113 ESTs (H. *erythrogramma*: 8,599; H. *tuberculata*: 8,514) were then generated and cleaned by the CGB using their EST Analysis Pipeline, ESTPiper [13]. Further details of the production of the cDNA libraries

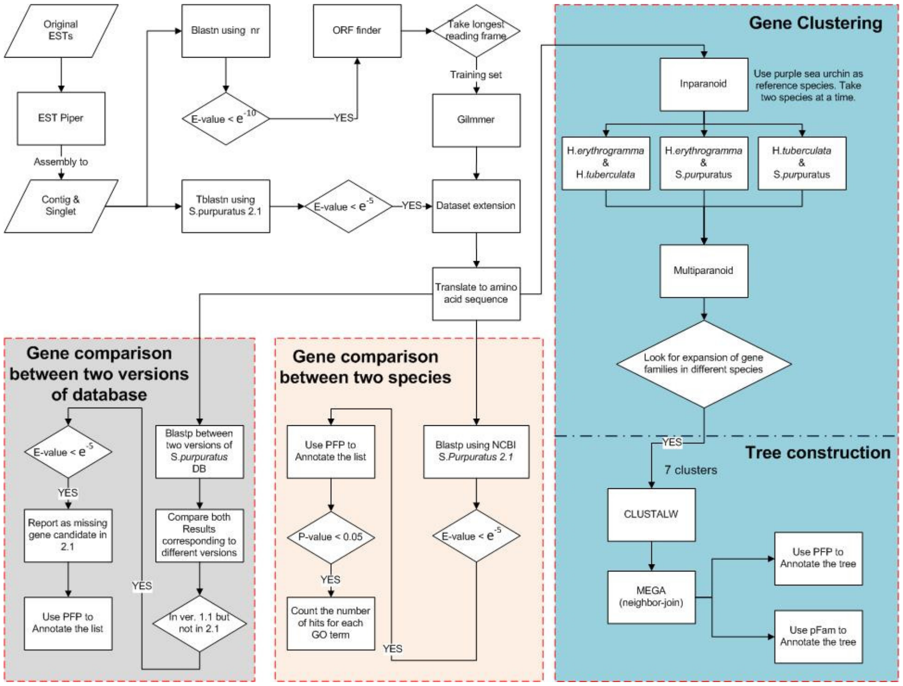
Original ESTs
EST Piper
Assembly to
Contig & Singlet
Blastn using nr
E-value < e^{-10}
YES
ORF finder
Take longest reading frame
Training set
Gilmmer
Tblastn using S.purpuratus 2.1
E-value < e^{-5}
YES
Dataset extension
Translate to amino acid sequence
Gene comparison between two versions of database
E-value < e^{-5}
YES
Report as missing gene candidate in 2.1
Use PFP to Annotate the list
YES
Blastp between two versions of S.*purpuratus* DB
Compare both Results corresponding to different versions
In ver. 1.1 but not in 2.1
Gene comparison between two species
Use PFP to Annotate the list
P-value < 0.05
YES
Count the number of hits for each GO term
YES
Blastp using NCBI S.*Purpuratus 2.1*
E-value < e^{-5}
Gene Clustering
Inparanoid
Use purple sea urchin as reference species. Take two species at a time.
H.*erythrogramma* & H.*tuberculata*
H.*erythrogramma* & S.*purpuratus*
H.*tuberculata* & S.*purpuratus*
Multiparanoid
Look for expansion of gene families in different species
YES
Tree construction
7 clusters
CLUSTALW
MEGA (neighbor-join)
Use PFP to Annotate the tree
Use pFam to Annotate the tree

Figure 2: Flow Chart

Table 2: Number of sequences

	H. *erythrogramma*	H. *tuberculata*
No. ESTs	8599	8514
	ESTPiper result	
Total Number	6319	6413
Contig	1247(19.7%)	1261(19.7%)
Singlet	5072(80.3%)	5152(80.3%)
	Predicted Genes	
Glimmer	2290	1776
tblastx	1016	916
	Total	
	3306	2692

and EST generation can be found on the CGB website [3]. We analyzed the length distributions of the initial EST sequences from each species and found the ESTs ranged from about 100 to 800 bases with similar average lengths (H. *erythrogramma*: mean = 624.5, sd = 123.5; H. *tuberculata*: mean = 603.1, sd = 129.1).

We used ESTPiper to assemble the ESTs into a set of non-redundant gene transcripts. We submitted the initial EST sets in fasta format for each species to ESTPiper for de novo assembly. ESTPiper uses the CAP3 sequence assembly program. We did not supply base quality files to aid in the assembly as they were not provided with the EST datasets. ESTPiper returned sets of contigs (consisting of multiple single ESTs) and singlets (single, unassembled ESTs) for each species (H. *erythrogramma*: contigs - 1,247, singlets - 5,072; H. *tuberculata*: contigs - 1,261, singlets - 5,152). After assembly, we analyzed the length distributions of the contigs and found them to be comparable for both species ranging from about 150 bases to 2,050 bases (H. *erythrogramma*: mean = 956.2, sd = 317.8; H. *tuberculata*: mean = 922.9, sd = 296.9). The number of ESTs contained in each contig was similar for each species as well ranging from 2 to 14 ESTs with comparable averages (H. *erythrogramma*: mean = 2.83, sd = 1.37; H. *tuberculata*: mean = 2.67, sd = 1.18). Histograms of these analyses can be found in the supplementary materials online (see the Supplementary Materials section below for the URL).

Gene finding We used GLIMMER3 to search the assembled ESTs that were obtained for H. *erythrogramma* and H. *tuberculata* for protein coding genes. GLIMMER3 builds an Interpolative Markov Model (ICM) [4] on a given training set and predicts genes based on that model. In order to build the training sets for GLIMMER, a BLAST [2] search was run on the EST assembly output files against the NR database. All the contigs and singlets that had a significant BLAST hit (e^{-10}) were taken and were given as input to NCBI ORF finder. NCBI ORF finder is a graphical analysis tool which finds all the open reading frames in a given sequence. It returns the reading frame, start position, end position, and the length of the open reading frames. The ORF with the longest length was chosen for every contig and singlet with a significant BLAST hit. Based on the corresponding start and end positions, sequences were extracted from the EST assembly files. While extracting the sequences, only those sequences with proper start codons ATG, GTG, CTG or TTG were taken. We randomly selected 250 of these extracted sequences for each species to be included in the training sequences for GLIMMER3.

The "build -icm" program of GLIMMER3 was used to build the Markov model on these training sets. The '-r' option was used to build model in the backwards direction [4]. Finally 'glimmer3' was run on the EST assembly files. The parameters used to run GLIMMER3 were all the default parameters (as given in the GLIMMER manual). The GLIMMER output had the start position, end position, reading frame and per-base raw score [4] for each predicted gene. If a contigs had more than one gene predicted, only that gene with maximum per-base raw score was chosen. Thus, the genes were obtained for both the species. GLIMMER predicted 2,290 genes in H. *erythrogramma* and 1,776 genes in H. *tuberculata*. To ensure that we had all the genes, and to ensure the accuracy of GLIMMER, tblastx was performed on each of the species against the purple sea urchin set (version 2.1). In the case of H. *erythrogramma* 1,016 genes were found to be unique to the tblastx output, and in case of H. *tuberculata* 916 genes were found unique to the tbalstx output. By adding these to the list of genes predicted by GLIMMER, the final sets of predicted genes were obtained. These are sets referred to as the genes union set. See Table 2 for totals.

Gene annotation The union sets of predicted genes were translated into amino acids and the obtained sequences were given as input to PFP (Automatic Protein Function Prediction Server). PFP uses a Function Association Matrix (FAM) to score significantly associating pairs of annotations [6]. PFP annotates all the sequences with GO terms and also the category of the GO term: molecular function (f), biological process (p), or cellular component (c). PFP output also gives a p-value for every term. For every contig, within every GO category, all the GO terms with p-values less than 0.05 were chosen. A total of 175 out of 3,306 H. *erythrogramma* species were not annotated by PFP, and a total of 157 out of 2,692 H. *tuberculata* were not annotated by PFP. The PFP results are provided under the supplementary data on the website.

Gene clustering We used Inparanoid [11] to cluster the predicted genes from H. *erythrogramma* and H. *tuberculata* with those from the S. *purpuratus* assembly (version 2.1). We included the purple sea urchin in the clustering as a reference species to check for expansion in gene families in either H. *erythrogramma* or H. *tuberculata*. Prior to clustering, the genes from the two Heliocidaris species were translated to amino acid sequences to enable alignment to the protein sequences from the purple sea urchin. All three sets of amino acid sequences were provided in fasta format to Inparanoid. Inparanoid performs an all against all BLAST search and identifies bi-directional best BLAST hits as putative orthologs which form the base of the cluster. Additional proteins are added to the cluster if they are more similar to the seed orthologs than to any other proteins in the other proteome. We ran Inparanoid with the BLOSUM80 scoring matrix for the BLAST comparison, bit score cutoff of 50 bits, 0.05% as the confidence cutoff for addition of inparalogs, 0.5 for the sequence overlap cutoff, 0.25 for the segment coverage cutoff, and 0 bits for the grey zone cutoff.

Inparanoid only clusters proteins for two species at a time, so we ran the program three times to build clusters for all combinations of the H. *erythrogramma*, H. *tuberculata*, and S. *purpuratus* protein sets. We then used Multiparanoid [1] to combine the clusters from the three Inparanoid runs. Multiparanoid produced 1086 gene clusters: 501 clusters contained proteins from only H. *erythrogramma* and S. *purpuratus*, 255 clusters contained proteins from only H. *tuberculata* and S. *purpuratus*, 117 clusters with proteins from only H. *erythrogramma* and H. *tuberculata*, and 213 clusters contained proteins from all three species.

We looked for evidence of expansion in gene families by identifying those clusters where there were more proteins from either H. *erythrogramma* or H. *tuberculata* in the cluster than there were from the purple sea urchin. We found 5 clusters showing this pattern of expansion. We also looked for clusters where the difference in the number of copies of proteins between H. *erythrogramma* and H. *tuberculata* was greater than or equal to 3. Due to the incomplete coverage of the EST sets for the two Heliocidaris species, it is difficult to tell if small variations in the number of copies between species are due to actual expansions and/or losses in individual species, or are simply due to differential samplings from the genomes of each species. For this reason, we only looked for large differences between copy numbers between the two species. We found 4 clusters showing this pattern of expansion. Two of the initial 9 interesting clusters identified by the two different criteria above turned out to exist in both sets, so this overlap reduced the number of interesting clusters identified to 7.

Multiple sequence alignment and tree building Clustalw[8] was used to produce a multiple sequence alignment for all sequences in each cluster. We generated both nucleotide and amino acid alignments for every cluster. We used the default opening gap and extending gap penalties(15.0 and 6.66 respectively for nucleotide sequences and 10 & 0.2 respectively for protein sequences). We also used the default scoring matrices (Gonnet 250 for protein sequence and DNA identity matrix for DNA sequences). The alignments, scores, and phylogenetic trees for each type of alignment (nucleotide and amino acid) are given in the Supplementary Data section on the website.

The aligned sequences from clustalw were used as the input for MEGA (Molecular Evolutionary Genetic Analyis). MEGA [7] is based on the Neighbor Joining algorithm for tree building. Because the two Heliocidaris species diverged from the purple sea urchin prior to their own divergence, we normally expected to see the purple sea urchin genes as outgroups in most of the trees. In addition, we would also normally expect to see genes duplicated in the purple sea urchin also duplicated in both H. *erythrogramma* and H. *tuberculata*. If not, then we can conclude that either there was an incidence of gene loss or our dataset simply did not contain those genes due to sampling differences in the species. For example, in cluster 130 there might be gene duplication in purple sea urchin (gi|72088920 and gi|11597469). Therefore we should be able to see two copies of these genes in both H. *erythrogramma* and H. *tuberculata*. In this tree we can speculate that one copy of the gene in H. *erythrogramma* has undergone several duplication processes, and H. *tuberculata* has only one copy. We can explain the latter case as the gene loss in H. *tuberculata*. In the case of cluster 156, as purple sea urchin is not the outgroup, we might be able to predict gene loss in both H. *tuberculata* and purple sea urchin.

Identification of genes missing from S. *purpuratus* assembly The second objective of this project was to look for genes present in the NCBI 1.1 version of the S. *purpuratus* assembly, but absent in the NCBI 2.1 version. For achieving this, BLAST was run on the translated predicted gene sets of H. *erythrogramma* and H. *tuberculata* against the NCBI 1.1 and 2.1 versions separately. BLAST results were filtered based on e-values. For the BLAST output against 2.1 version set, all the hits that had an e-value lesser than (e^{-5}), were taken into consideration. For the BLAST output against 1.1 version set, since the size of the set was half (22,000) that of the 2.1 version, all the hits that had an e-value lesser than ($5e^{-6}$) were taken into consideration. The numbers of unique hits for every translated gene (from contig and singlets) were counted. All the contigs or singlets that had significant hits against the 1.1 version set but had zero hits against the 2.1 version set were taken as those that are missing in the 2.1 version set. A total of 39 H. *erythrogramma* protein sequences had at least one hit against the 1.1 version set but had no hits at all against the 2.1 version set. This was confirmed by running BLAST on these 39 sequences a second time against the 2.1 version set. A total of 51 H. *tuberculata* protein sequences had at least one hit against the 1.1 version but had no hits against the 2.1 version set. There were an additional 7 overlapping protein sequences from H. *erythrogramma* and H. *tuberculata* that had hits against the version 1.1 set, but not the 2.1 set. In total, we found 97 protein sequences that had hits against the 1.1 set and not the 2.1 set that could possibly be candidates for addition to the offical 2.1 set.

4 Discussion

The primary goal of our study was to identify possible reasons for the differences in the developmental cycles of the direct-developing species, H. *erythrogramma*, and the indirect-developing species, H. *tuberculata*. As shown in Table 1, we found possible evidence of expansion in 7 gene families by clustering proteins from H. *erythrogramma*, H. *tuberculata*, and S. *purpuratus*. The most interesting of these families by far are clusters 271 and 375 which both contain multiple zinc finger domains and could be possible transcription factors. In both clusters there are more copies in H. *erythrogramma* than in H. *tuberculata* (4 more for cluster 271 and 3 more for cluster 375). It has been shown that the viable H. *erythrogramma* x H. *tuberculata* hybrids regain the oral ectoderm tissue of H. *tuberculata* which is necessary for feeding during the larval stage, implying that the genes responsible for this tissue still exist in H. *erythrogramma*, but are turned off (i.e. negatively regulated) [14]. The additional copies of the potential transcription factors in H. *erythrogramma* could be suppressing the genes responsible for oral ectoderm development, thus necessitating a direct pattern of development.

Also interesting is cluster 42 which appears to be a cystatin family protein. The purple sea urchin genome contains a very large cystatin family compared with many other species [12]. These cystatins are receptor proteins that recognize foreign pathogens or signals from other molecules indicating infection [12]. H. *erythrogramma* appears to have an additional copy of a particular cystatin as compared to purple sea urchin. It may also possibly have an additional copy as compared to H. *tuberculata*, but the additional copy in H. *tuberculata* may also simply be missing from our gene set due to sampling differences between the species. We can only speculate that H. *erythrogramma* might encounter different or increased susceptibility to certain pathogens due to its direct developing pattern that resulted in the selection of individuals with a duplicated copy of the gene.

The remaining clusters are not as interesting. Cluster 130 contains proteins with a calcium binding domain and is likely a calmodulin family. This protein is ubiquitous in all species and is often involved in signaling pathways. It is interesting that we found so many copies in H. *erythrogramma*, and so few copies in purple sea urchin, but are not sure as to why. Cluster 156 is predicted to be a NADH dehydrogenase protein family, a common protein in cellular energy production. Cluster 516 appears to be a family of transposons with a large number of copies in purple sea urchin. Finally, BLAST and Pfam search of the sequences for cluster 589 did not show any distinct protein functions, so it is difficult to tell if it is a family important to the differential development of the Heliocidaris species.

As mentioned earlier, we also found 97 proteins (combined from H. *erythrogramma* and H. *tuberculata*) that had significant matches to proteins in the NCBI 1.1 version of the S. *purpuratus* assembly that did not produce any significant hits against the NCBI 2.1 version protein set. We then attempted to functionally annotate all of these 97 proteins using Pfam [5]. Most of the proteins belonged to one of the following Pfam categories: DNA Replication factor, RNA binding motifs, UDP glucuronosyltransferase, Zinc Fingers, DNA damage regulators, Caspase recruitment domains, Transducins, or Retrovirus-related polyprotein. An analysis of the BLAST results for the nucleotide sequences for these proteins revealed that there were 21 genes that were annotated as 'partial' in the BLAST results. These partial proteins could be pseudogenes. It was mentioned in the purple sea urchin genome assembly paper [9] that some of these pseudogenes were filtered out of the version 2.1 gene set. It could be that these genes that are annotated 'partial' could have been active in the ancestor of the Heliocidaris species and S. *purpuratus*, but evolved into a psuedogene later in the purple sea urchin. Or, they could be transcriptionally active pseudogenes in S. *purpuratus* as well in both H. *erythrogramma* and H. *tuberculata*. In either case it would be worth looking at these possible pseudogenes, to see if they are actually pseudogenes or actually normally functioning genes. If so, they could be added to the NCBI 2.1 version set.

Finally, we attempted a comparison of protein function annotations between H. *erythrogramma* and H. *tuberculata* by first finding the intersection and differences of the GI numbers of the significant hits (e^{-5}) to proteins in the purple sea urchin (version 2.1). We then identified the 10 most frequently matched GI numbers for each set (GIs of hits occurring in both H. *erythrogramma* and H. *tuberculata*, GIs of hits only occurring in only H. *erythrogramma*, and GIs of hits occurring only in H. *tuberculata*). We then used PFP to generate the GO terms for the matching proteins in S. *purpuratus* for all 3 of these sets. The PFP annotations identified the molecular function GO term "RNA-directed DNA polymerase activity" as the most likely for every one of the top 10 most frequently matched GI numbers in purple sea urchin. Thus, we identified a large number of probable reverse transcriptases in our datasets.

5 Supplementary Materials

Supplementary analyses and data can be found online at:
http://mendel.informatics.indiana.edu/~chuyu/Project/
Links to various resources can be found on the right-hand side of the page:
Report 1 - EST assembly and analysis
Report 2 - Gene finding
Report 3 - Gene annotation
Supplementary Data - Initial EST sets for H. *erythrogramma* and H. *tuberculata*, ESTPiper results, GLIMMER results, ORF Finder results, union set of GLIMMER and tblastx predictions (nucleotide and amino acid), BLAST output of H. *erythrogramma* and H. *tuberculata* against NR database, and against purple Sea Urchin NCBI (version 2.1), gene finding and annotation output files, PFP output, KAAS output, Interesting Clusters information (including alignment at protein and nucleotide levels, and phylogenetic trees)

References

[1] Tamas I. Liu G. Sonnhammer E.L.L. Alexeyenko, A. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22:e9–e15, 2006.

[2] Gish W. Miller W Myers E.W Lipman D.J. Altschul, S.F. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[3] et al CGB. Heliocidaris cdna libraries, August 2006.

[4] Arthur .L. Delcher. *GLIMMER Release Notes*. University of Maryland Center for Bioinformatics and Computational Biology, version 3.02 edition.

[5] Tate. J. Mistry J.Coggill P.C. Sammut J.S. Hotz H.R. Ceric G. Forslund K. Eddy E.R. Sonnhammer E.L. Bateman A. Finn, R.D. The pfam protein families database. *Nucleic Acids Research Database Issue*, 36:281–288, 2008.

[6] T. Hawkins and D. Kihara. Pfp: Automatic annotation of protein functions by relative go association in multiple functional contexts. page 117. 13th Annual International Conference on Intelligent Systems for Molecular Biology, 2005.

[7] Dudley J. Nei M. Tamura K Kumar, S. Mega: A biologist-centric software for evolutionary analysis of dna and protein sequences. *Briefings in Bioinformatics*, 9:299–306, 2008.

[8] Blackshields G. Brown N.P. Chenna R. McGettigan P.A. McWilliam H. Valentin F. Wallace I.M. Wilm A. Lopez R. Thompson J.D. Gibson T.J. Higgins D.G Larkin, M. A. Clustal w and clustal x version 2.0. *Bioinformatics*, 23:2947–2948, 2007.

[9] Berney. K. Cameron. R. A. Materna, S. C. The s.pupuratus genome: A comparitive perspective. *Developmental Biology*, 300:485–495, 2006.

[10] P.Z. Myers. Evolution of direct development in echinoderms, June 2007.

[11] C. E. V. Sonhammer E. L.L. Remm. M., Storm. Automatic clustering of orthologs and in-paralogs from pairwise species comparision. *Journal of Molecular Biology*, 314:1041–1052, 2001.

[12] et al Sea Urchin Genome Sequencing Consortium. The genome of the strongylocentrus purpuratus. *Science*, 314:941, 2006.

[13] Choi J. Hemmerich C. Sarangi A. Colbourne J.K. Dong Q.; Tang, Z. Estpiper - a web-based analysis pipeline for expressed sequence tags. *BMC Genomics*, 2009.

[14] M. E.; Raff R. A.; Wilson, K. A.; Andrews. Dissociation of expression patterns of homeodomain transcription factors in the evolution o developmental mode in the sea urchins heliocidaris tuberculata and h.erythrogramma. *Evolution and Development*, 7:5:401–415, 2005.