

## **Introduction:**

An annotator's ultimate goal will be to present the most accurate representation of his/her gene. The official sea urchin gene list is the GLEAN3 list. The coordinates given are based on the assembled Genome Sequence of *Strongylocentrotus purpuratus*, Spur\_v0.5 (April 15, 2005).

In practice, an annotator will either accept the coordinates supplied or modify the coordinates to reflect more accurately their perception of the gene. Totally new genes NOT found on the official list may be added as "novel" genes. **When modifying the official gene or adding a novel gene, submit the DNA and protein sequence.** This annotation is "owned" by the annotator and can be changed only by him/her. Additional information related to gene name, function, time of expression, local of expression and family tree relations should be noted when the information is available.

An annotator **should** indicate which **group coordinator** they are aligned with by selecting his/her name from the pull-down menu opposite "Group Coordinator" on the Gene Information page.

**Caution:** There is a known condition in which GLEAN will predict an exon that spans the "scaffolding" between two contigs (which appears in the scaffold sequence as a long series of N's, which are translated as X's). Sometimes this does reflect biological reality, but sometimes it's an artifact. In order to keep the maximum sequence information, these exons have NOT been disallowed. Please do not be shocked if you encounter such an instance – one can remove the "X's" and "N's" to your taste.

## **Available Tools and Pre-computed information:**

### **A. GENBOREE viewer**

1. The "official sea urchin gene list" is the GLEAN3 track. [All of these gene model coordinates are already loaded into the annotation database for validation by the annotators.]
2. Tracks are available for each of the various gene prediction sets used to create the official list. [FGENESH++ has recently arrived and will be available as well.]
3. Tracks for "novel", non-annotated, genes and modified GLEAN predictions will be added once the database begins to be populated.
4. Tracks containing cDNAs and ESTs. [Splign and Exonerate output tracks].
5. A track with the embryonic expression data derived from an array-based RNA hybridization utilizing 50bp oligonucleotides exists.
  - (a) Move your cursor above the colored box until a hand appears. Left click will pull up a dialog box that will either give you the signal strength, strand, and coordinates or will suggest that you zoom in where the detailed dialog box will be available.
  - (b) There is also an option in the dialog box "Activity Profile" that links to a more detailed view of the data in the region.

6. A track displaying the BAC end sequence placement for the 145kb BAC libraries will be mounted. [Tiling path as well as other available clones will be included.]
7. GENBOREE is linked to the Annotation Database through the GLEAN prediction. [When starting in the database, you will have to login to GENBOREE the first time you activate a link – but NOT later during your session.]
8. DNA sequence for exons, full gene length and genomic regions is available through the GENBOREE viewer: (1) left click on exon of interest ->get genomic DNA-> file options [use “eg” and “?” for definitions of sequence included for each option]. (2) top left button “GetDNA”.
9. URL links to regions of interest may be sent to collaborators: (1) top left button “Full URL” and (2) top right button “email group”.

**B. Blast against the official gene list [GLEAN3] predictions**

1. The Baylor sea urchin BLAST site now allows BlastP and BlastN against the Glean3 predictions:  
<http://www.hgsc.bcm.tmc.edu/blast/blast.cgi?organism=Spurpuratus>
2. Shortly, there should be a link from the Glean3 ID in the Blast result to the corresponding region in Genboree. In addition, the output from a Blast against the S.purpuratus Genome Assembly 2005 07 18 database will have a link to Genboree for the region of the scaffold identified.

**C. Annotation Database pre-computed information**

1. The top ten matches to exons of models derived from proteins of Ciona, mouse and human are available on the database page titled “Glean Prediction Precomputation for GLEAN3\_NNNNN”. The top alignment is available to view. [The goal is to provide additional support for gene predictions.]

Results from comparing each GLEAN3 prediction to PFAM motif features. The results for a particular prediction are listed on the database page titled “Glean Prediction Precomputation for GLEAN3\_NNNNN”. [One can also search for all predictions that contain one or more specified PFAM motifs – start this on the initial database search page.] [The parameters chosen were the default values for a “gathering threshold” analysis.]

## Information to be furnished by annotators:

### A. Account and Log In.

1. The following URL will be the official entrance to the sea urchin annotation database: <http://annotation.hgsc.bcm.tmc.edu/Urchin/> or from ftp main page (click bottom link):  
<http://www.urchingenome.hgsc.bcm.tmc.edu/>
2. Use your registered email address (as in listserv) as username and "analysis" as password to login the annotation submission website. Password is case sensitive.
3. To add an account, send email to [LZHANG5@BCM.TMC.EDU](mailto:LZHANG5@BCM.TMC.EDU) with your preferred email address, first name and last name [this service will be activated after the test period].
4. There is one group account to login Genboree browser: urchin/analysis. It's case sensitive and it is needed only for the first time a user tries to access Genboree within an annotation session.
5. An individual account for Genboree needs registration.
6. The best browser to use Genboree on a MAC is Safari (as tested so far). The Internet Explorer and Firefox are both good choices for PC users.

### B. Search for Genes and Start Submission Page

1. Both official Glean 3 list and annotated gene list will be searched. The search results from the two lists will be combined to display.
2. Users can select one of the six categories, enter the keyword to search and click "Start Search". For most of the categories, a search will be performed if it returns less than 500 genes.
3. If one needs to use protein sequences or DNA sequence to search for appropriate Glean3 Ids, the Baylor sea urchin Blast site is set up to blast against the Glean3 predictions' DNA or protein sequences:  
<http://www.hgsc.bcm.tmc.edu/blast/blast.cgi?organism=Spurpuratus>
4. Annotator search and Pfam search can use a pop-up select list to make selections. These features are supported by Internet Explorer and Safari.
5. Annotator search can search for one annotator. Pfam search can search multiple selected Pfam domains. When selection is made, the keyword will display in the text field on the search page.
6. The Pfam domain checklist is displayed in Alphabetic order of the Domain Description (not Domain ID). The keyword displayed in search page is the selected Domain ID. Un-checking a box will remove the corresponding Domain ID in the keyword field on search page.
7. Searching on a Group Coordinator will return a list of all genes currently annotated by a member of that group. [Gene ID, common name, synonym, and annotator is reported as well.]
8. **To start the annotation for a gene, first search a particular gene ID to check the existence of this gene ID. The annotator who starts the annotation of a new gene will have the ownership of this gene.**

### **C. Starting to Annotate the Searched Gene from the Search page**

1. If a search is valid and the result is not empty, the search result will display in a table of 7 columns.
2. The first column "Glean Gene" will link to a page to display pre-computed Glean prediction results. (A gene without a valid Glean 3 ID will not have this link.)
3. The second column "Annotated Gene" will link to the gene information page for this annotated gene (for genes with Glean ID or customized ID).
4. The fifth and sixth columns are additional information for annotated genes.
5. The last column indicates the annotation status and ownership for this gene. If the annotation was initiated, the owner should see the "Continue" link. If a gene has not been annotated by any user, the current user will see the "Start" link to begin annotation.

### **D. Annotate an Existing Official Gene:**

1. Follow the link ("Start" or "Continue") on search result page to enter the gene information submission page to start or update the annotation.
2. Ownership will be checked and non-owner users will not have submission buttons on any of the submission forms.
3. The desired gene name (either Glean ID or customized ID) and username will be automatically filled in the gene information submission page and other related forms.
4. Fill in the main gene information page:
  - a. Check the level of detail that you have attained.
  - b. Check any additional evidence that you might have to support the gene model.
  - c. Fill in the empty fields where applicable. [A new feature here allows one to indicate or "flag" some difficult annotation scenarios that can be re-visited.]
  - d. Indicate the Group Coordinator with whom you are collaborating on the gene to be annotated.
5. The gene annotation is created (by clicking "CREATE" button for the first time of submission), the additional forms for gene features, gene expression data, and gene sequences will be available at the top of the page.
6. Gene features, gene expression data, and gene sequences are independent of each other. They all depend on the existence of the general gene information.
7. The gene features of a Glean prediction can be used as a starting point of the annotated gene features by clicking a button "Use Glean Features" on gene feature page.
8. If one accepts the Glean Features, then a new feature table appears that allows one to edit each exon as necessary. Enter changes for an exon or feature, click "Update" and then the "Gene Features" at the top of the page to review your change.
9. If one needs to add exons that fall in sequence gaps edit as follows:

- a. Sequence = Gene ID [Example: GLEAN3\_XXXXX]
  - b. Source = Missing
  - c. Exon Number = [use expected number]
  - d. Coordinate Start = [coordinate in DNA sequence submitted by annotator]
  - e. Coordinate Stop = [coordinate in DNA sequence submitted by annotator]
  - f. Be sure to include DNA sequence file for the modified gene model.**
10. The definition of all columns in gene feature table is borrowed from GFF format definition.
  11. To delete a gene will also delete its additional information including gene features, gene expression data, and gene sequences.
  12. To enter any available information about the developmental time and tissue in which expression occurs click on “Gene Expression” at the top of the main gene information page.
  13. To enter sequence for a gene [CDS, mRNA and peptide] click on “Gene Sequences” at the top of the main gene information page.
  14. To enter information about multiple forms of a gene, i.e., gene duplications, click on “Gene Duplication” at the top of the main gene information page. Add each gene ID in a set of multiple copy genes one at a time to build a complete list. One can also indicate a possible source of the duplication, e.g. haplotype.

#### **E. Add novel genes not in the GLEAN predictions:**

1. If a search finds no “official prediction”, the search result will display an empty table since there are no pre-computed data.
2. To proceed, look for the comment: “No annotation submitted for (gene of interest) Click [here](#) to submit this gene”. Click the “here” to go to the main gene information page for your novel gene.
3. The annotator will have to fill in the scaffold/coordinates on the “Gene Features” page.
4. The annotator should submit the DNA sequence for CDS and mRNA as well as the peptide sequence on the “Gene Sequences” page.
5. Use the comment field at the bottom of the main gene information page to describe how this novel gene was found [i.e., merged Glean3\_XXXXX with GLEAN3\_NNNNN].

#### **Questions asked by annotators:**

1. >when you ask best hit on BLASt do you mean best E value?  
There is a field BEST GENBANK HIT: (Accession Number) . This is to allow someone to trace the gene's connection to published references through the best hit's Accession number.
2. >what does gene model check refer to?  
the GLEAN3 gne prediction

3. >what is a CDS vs. mRNA?

CDS=transcript of coding region without UTR regions

mRNA= full mRNA sequence including UTR regions