November 2009 Newsletter

http://Spbase.org

Newsletter. This newsletter coincides with the new build of SpBase. Described below are a series of changes to the both the operation of the databases and the included data. Four of us, Autumn Yuan, Dong He, Manoj Samanta and Andy Cameron attended the recent SUDBXIX meeting in Woods Hole and we have attempted to include suggestions we gathered from the users at that time.

Gene Annotation. A little over 21,000 of the 28,944 GLEAN3 gene models have now been assigned names. The last increment has been inferred from electronic annotation (IEA) by Ung-jin Kim. Using a combination of BLAST alignments to a non-redundant protein database, matches to conserved domain data and comparison to orthologs, a putative gene identity is assigned. Currently, we have annotated more than 7,000 additional gene models by this method. Roughly 40% have been identified to be homologs of known genes, 30% assigned to a "hypothetical" protein group, 20% assigned to anonymous proteins named after their structural motifs, human open reading frame number or other arbitrary cDNA identifiers. Approximately 3% of the newly examined gene models have no convincing match. The recent increment of annotations must still be considered tentative since we have not verified them with expressed sequence evidence.

Along with the manual examination of the remaining GLEAN gene models we have begun to reduce the redundancy in the official gene set. If a coding sequence is perfectly matched between two models or nearly identical including the 3-prime UTR we tag these as duplicates and remove the details from one of the pair leaving a comment to indicate a duplicate. To facilitate this process we have added all of the 3' UTR sequences identified in Manoj Samanta's whole genome tiling array.

Our literature annotation effort continues. Now 4,942 gene models have been linked to corresponding articles in PubMed.

BioMart. In order to support data mining at SpBase, Dong He has implemented BioMart, the GMOD query-oriented data integration system. It supports scalable large scale querying of individual databases as well as query-chaining between them. The Biomart page allows the user to query the main database using any of the various fields in an individual record and produce lists of output with user-defined columns. For example, a search for all bHLH transcription factors can be tailored to return gene ID, gene name and scaffold position. This feature gives the user an easy to use interface to the sequence data without cutting and pasting it from individual graphical pages.

Gene Web Pages. We have spent some time to remove inconsistencies in the various data tables making the gene annotation database more robust. Also, Autumn Yuan has changed some details on the gene search pages. One can search with multiple terms and Boolean relationships now. Popup windows make sequence downloads for individual genes easier. A new UPDATE/REVISE page series has been added as well. After registration a user can update the annotation for an existing model or enter the annotation for a new gene not included in the GLEAN set. We hope that you, the members of the sea urchin research community, will join us in this effort.

Future Plans. We are expecting a new genome assembly and many additional expressed sequences from the various RNA-seq projects already underway. We will focus on perfecting the official gene set and mapping it to the newest assembly as these results become available in the coming months.

You are receiving this newsletter because you have previously been included in a SUDB mailing. Please let us know if you would like to be removed from our list for these very occasional newsletters. Issue #2, 10/28/09

Email to: donghe@caltech.edu