Application Number: 1 U54 HG004555-01

Project Title: Integrated human genome annotation: generation of a reference gene set (GENCODE)

Quarterly progress report summary (Q2: 1/1/2008 - 31/3/2008).

This report accompanies the Q2 spreadsheet document and the subsections refer to each sheet.
Overall the project is in the infrastructure setup phase, as projected in the milestones plan, although the rate determining production pipeline of manual annotation is running at the levels required to meet 1st year objectives. Informatics infrastructure is being setup to link all the pipelines of the project together. In the mean time pilot datasets are being distributed to allow procedures of different groups to be developed, tested and discussed. Throughout this period there were bi-weekly conference calls of the whole project, interspersed with bi-weekly sub-group conference calls for pseudogene annotation.

Manual Annotation

Manual annotation of chromosomes 2 and 7 has accelerated partly as a result of recruitment and training of Havana annotators (WTSI). During the Q2 period annotation was at a rate sufficient to meet the objectives for the first 12 months of 2250 new loci annotated.
Flat file gff3 dumps of chromosomes 21 and 22, which have been recently manually annotated as part of the extension to the GENCODE pilot project, were initially provided to other groups. By the end of Q2, a DAS source had been setup to provide live access to manual annotation database, with scripts provided to allow other groups to use it to dump gff on demand. A GENCODE project grouping was setup in the DAS registry and DAS sources were registered under its name (WTSI).
No consensus annotations were released to DCC as infrastructure to integrate annotation sources and track annotation consensus status is under construction (WTSI).

Experimental verification

Following recruitment, work started on setup of informatics systems and processes for selection of candidate annotation for experimental verification (CRG). Preparations for experiments have involved purchase of necessary RNA and setup of robotic workstation which will be used for the project (Lausanne).

Pseudogene assignment

Both of the automatic pseudogene prediction pipelines at UCSC and Yale have been run for the first time and DAS servers have been setup. A number of new pseudogene loci have been made by Havana (WTSI) as part of manual annotation. No consensus predictions were made in Q2. Comparisons are ongoing to establish consensus predictions and develop procedures to resolve discrepancies. Yale published an article based on pseudogene analysis related to activities in this project.

Gene prediction

WashU carried out the first N-SCAN_EST production run and are working on setting up DAS sources to make available the results. This will make use of the DAS alignment format. So far Paragon runs have been made only on selected test sets. MIT have carried out the first CONGO exon prediction run and are comparing the results with existing annotation. They will carry out an RFC_CSF run in Q3. DAS servers for each pipeline are being setup. UCSC TransMap and ExonPhy pipelines were not rerun in this quarter, however their existing runs are available via the UCSC DAS server. CRG has carried out their first U12 pipeline run and is similarly in the process of comparison and DAS server setup. Where output from these computational pipelines leads to new consensus annotation this will be recorded. The infrastructure to integrate the output from these pipelines is under construction (WTSI, see ‘Manual Annotation’).

Reporting changes

This project incorporates the activities of the CCDS (Consensus CDS) consortium partners at WTSI and UCSC and so reporting lines have been added to capture activities that revise the CCDS transcript set carried out by annotators at WTSI (Havana) and UCSC. Since part of the process of CCDS updating involves merging the output from Havana annotation and the automatic Ensembl geneset, these merge runs and the comparisons and QC runs carried out by WTSI and UCSC are also reported.