Grant Number : U54 HG004555-03

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

Quarterly progress report - *Narrative Questions* (Year 3, Q3: 04/01/10 - 06/30/10)

General Questions

1. What is your assessment of progress relative to the project's milestones and to the amount of money you have spent?

Milestone 1 (Sheet 1: New Manual Annotation) is still well ahead of target and is currently at 90% of the original approved milestone. Spending is tracking the original budget (entirely salaries).

Milestone 2 (Sheet 1: Experimental Validation) is still substantially behind, although progress is being made. Spending is still lagging the original budget (salaries and experimental reagents). It is anticipated that the milestone can still be achieved once the new strategies are finalised.

Milestone 3 (Sheet 2: Pseudogene Annotation) is back on and ahead of the target. In collaboration between Sanger and Yale the repeat-masking of the genome and the PseudoPipe pipeline were re-run, bringing the numbers of confirmed pseudogenes back up. Spending is tracking the original budget (entirely salaries).

Milestone 4 (Sheet 1: Overall Gene Annotation) is behind in terms of the fraction of genes classified as level 1, however the faction currently classified as level 2 continues to exceed this figure. As for milestone 3, milestone 4 is a percentage, related to the original projected number of non-pseudogenes (30,000). This number was estimated as the total number of protein coding and long non coding genes. Spending is tracking the original budget (entirely salaries).

2. Do you anticipate being able to accomplish your milestones within your budget? If not, what changes are planned?

We anticipate that we will meet the original project objective of a 90%, verified human geneset focusing on protein coding genes presuming that we obtain the 30% of Year 4 funds which is currently being held back. This is being held back pending approval by the ECP of a revised plan for biological validation and incorporation of RNA-Seq data including closer collaboration between the Hubbard and Gingeras groups. This plan is currently been written after discussions at the GENCODE meeting in Hinxton which was held at the end of June. NHGRI has indicated that ENCODE funding will be extended for an additional year. Year 5 funds will allow us to annotate the remaining 10% of the human geneset that was cut from the original proposal (when the GENCODE project started, roughly half of the genome was already partly annotated, so only an additional 40% was targeted to be annotated from scratch over the 4 years).

3. What bottlenecks have you encountered and how are you addressing these? For example, have you made any changes to your production pipeline?

The experimental verification process is still significantly delayed. Considerable time has been spent on checking that the modifications that were made to the pipeline are working. The number of experimentally tested models is being increased now, but by designing the primer sets more stringently, the number of possible targets decreases. This is an ongoing process and is helped by

continuing dedicated weekly conference calls between Sanger, CRG and Lausanne. The minutes of these meetings are now on the wiki pages.

In order to improve the biological validation, research is currently underway by CNIO and Sanger to investigate translation validation using mass spectrometry.

Designing primers for the verification of RGASP predictions also required modifications to the pipeline. The data analysis for the second round of data submission for RGASP is still ongoing but it is anticipated that a third round of RGASP focussing on read mapping will but run at the end of 2010 with a workshop in Barcelona in early 2011.

Currently, the overall structure of our pipeline has not changed. However we are continuing to investigate how to incorporate RNA-Seq data into the pipeline, e.g. to make better use of existing whole-transcriptomic sequencing results for the confirmation of transcript models. This is an ongoing process as it involves a considerable amount of data (different cell lines, compartments, technologies, replicates). Collection of new evidence and the development and refinement of computational methods for the evaluation of the GENCODE annotation by each group continues to be an ongoing feature of the project.

We have started to investigate running RQ-PCR experiments on selected gene models to have independent quantification in parallel to the RNA-Seq and Nanostring results within RGASP.

Project-specific questions

1. What is the status of your computational predictions?

The Ensembl-Havana gene model merging pipeline has reached a mature level with only a few knows issues to be fixed remaining. The frequency of re-running the complete genebuild and merge process is a point of discussion.
The problem leading to decreased number of pseudogenes from Yale and thereby in the confirmed (level 1) set was identified and the PseudoPipe program was re-run.
Potential errors in splice-sites are continuously being reported through a pipeline from UCSC/WashU and integrated into the annotation software Otterlace and the tracking system AnnoTrack as well as output from other computational algorithms as they arise.
CNIO has delivered Mass-Spec data mapped to the GENCODE set which is now used for evaluating gene models. It has also delivered preliminary sets for principal isoform detection.
The tools for manual annotation are being continuously improved to allow better QC.

2. Do you still believe 10,000 to be the total number of pseudogenes?

Currently this seems a reasonable genome wide estimate; although it's possible the final figure for consensus pseudogenes by these criteria will end up slightly higher.