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, Q4: 07/01/10 - 09/30/10)

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 99% of the original approved milestone. Spending is tracking the original budget (entirely salaries). The numbers for "overall gene annotation" are showing some fluctuation as we have to use different methods to estimate the numbers where we do not have a data freeze for the given point of time.
Milestone 2 (Sheet 1: Experimental Validation) is still substantially behind, although progress is being made with over 1100 genes tested by experimental methods in the last quarter. Spending is still lagging the original budget (salaries and experimental reagents).
Milestone 3 (Sheet 2: Pseudogene Annotation) is currently ahead of the third year target. 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 and non short RNA genes (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 things continue to improve in the experimental verification process and 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 still being finalised but we hope to have it completed before the end of the year. NHGRI has indicated that ENCODE funding will be extended for an additional year upon demonstration of continued progress and approval of a research plan for this 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. Progress is being made but considerable time has still 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. We are now considering using whole transcriptomics RNA-Seq data (instead of targeted sequencing) to accomplish our milestones and have decided to test this method in our next batch of experiments. However, we are currently waiting on the results of IDR comparison from the transcriptome group before we can continue with this. This is an ongoing process and is helped by continuing dedicated biweekly conference calls between Sanger, CRG and Lausanne. The minutes of these meetings continue to be available on the wiki pages.

In order to improve the biological validation, research is still underway by CNIO and Sanger to investigate translation validation using mass spectrometry.
The experiments for the verifications of selected gene models predicted from RNASeq data by participants of the RGASP was carried out successfully and resulted in high confirmation rates. 107 human and 137 c.elegans junctions were tested.
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. 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.
In collaboration with Barabara Wold's group we are designing a new set of probes for quantification with the Nanostring technology. The targets for this set have been selected from Gencode gene models expressed in all three cell lines GM12878, K562 and HepG2.
We are still investigating 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 known issues to be fixed remaining. The next complete genebuild and merge process has now started and will be released at the end of the year or early next year.
Computational pipelines using RNA-Seq are being developed by Ensembl and MIT. Ensembl has been piloting an ncRNA pipeline. MIT has been exploring the Illumina BodyMap dataset using the Scripture algorithm. As found in the RGASP evaluation, it is difficult to assembly splice isoforms from RNAseq data.
The problem leading to a decreased number of pseudogenes from Yale and thereby in the confirmed (level 1) set was identified and RepeatMasker and the PseudoPipe program were 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.
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.

3. We have one outstanding question from your Y3Q3 progress report: what is the delay with data submission for the experimental validation experiments (RACE and RT-PCR)? There is a large amount of data - 1540 experiments - that have been in the "completed/not submitted" bin of this pipeline since June 2009. Please provide some explanation for this delay in data submission.

So far we felt confident in using the results from the "traditional" Sanger sequencing pipeline for assigning verification status ("level 1") to transcripts only. We have been revising the computational approach to align the sequencing reads from the Illumina pipeline multiple times to rule out errors. The verification rates have been low and we are still investigating this observation.
In parallel we have started submitting the experimental results (sequencing reads) to the ENA and will continue this process.