Grant Number: U54 HG004555-04
Project Title: Integrated human genome annotation: generation of a reference gene set(GENCODE)
Quarterly progress report - *Narrative Questions* (Year 4, Q1: 10/01/10 - 01/21/11)

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) has been passed and is currently ahead of target at 110% 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 with ~2000 genes entering the experimental pipeline in the last quarter and assignment of confirmation levels of previously tested loci through a re-implemented mapping pipeline. Spending is still lagging the original budget (salaries and experimental reagents).
Milestone 3 (Sheet 2: Pseudogene Annotation) has increased by 3% in the last quarter to 67%. Spending is tracking the original budget (entirely salaries).
Milestone 4 (Sheet 1: Overall Gene Annotation) fraction of genes classified as level 2 increased to 83.3% in the last quarter. The fraction of genes classified as level 1is still behind target, but substantially increased from0.4 to 2.3%. 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. As a first set we have imported the gene model predictions based on the RNASeq data of the HUVEC cell line (normal karyotype) from the Gingeras group into the annotation tool. We have analysed this set for novel regions and are assessing these taking into account the assigned IDR scores. We have also added a CAGE cluster data set (GM12878, Riken) and are working on the best way to visualize this.
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, which was submitted in January. 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 behind target, however more rapid progress is now being made. In the last quarter~2000new genes entered the pipeline and we were able to analysis data generated by the pipeline for genes entered in previous quarters to verify a significant number of gene annotations. This has led to an increase in gene annotations assigned 'Level 1' status in Milestone 4 (Sheet 1: Overall Gene Annotation) from 0.4 to 2.3%. The number of experimentally

tested models is now being increased and should yield higher verification success rates due to better primer design and mapping strategies. While we are now confidence our strategies to use targeted sequencing to verify gene structures, we arealso working on extending the pipeline to use whole transcriptomics RNA-Seq data in an initial verification step to reduce the number of genes that need to be verified using targeted sequencing. Adding this extra pipeline step should allow us to accomplish our milestones quicker. We are testing RNA-Seq data from multiple sources (including the ENCODE transcriptomegroup) for this purpose. 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.

Project-specific questions

1. What is the status of your computational predictions?

The Ensembl-Havana gene model merging pipeline has reached a mature level and anew merge of Havana annotation with a completely newEnsembl automatic genebuildis currently under QC. The genebuild was carried out the GRCh37 assembly incorporating the 2nd set of patches released by the Genome Reference Consortium. It will be released asGENCODE 7 within the next two months.
Computational pipelines using RNA-Seq are being developed by Ensembl and MIT. Both groups have used the IlluminaBodyMap dataset to construct gene models with their specialized methods. These have been imported into the annotation tool and will be visible for future Havana annotation.
The problem leading to a decreased number of pseudogenes from Yale and thereby in the confirmed (level 1) was fixed; the numbers are back on track.
Havana has spent significant time reviewing potential errors in splice-sites identified by the pipeline from UCSC/WashU. So far 3849 cases were rejected but 2707 cases led to an update in the annotation.
The tools for manual annotation(Otterlace) and annotation tracking(AnnoTrack) are being continuously improved to allow better QC.

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

Currently this still 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. Last quarter you told us you are submitting RACE and RR-PCR experiments to the ENA. Please provide a list of accession numbers for these submissions. In addition, how are you coordinating with the DCC to get these data to them?

The batch 1 capillary reads were submitted to EMBL-Bank, submission id Hx2000011242. Batch 2 RNASeq reads were submitted to ArrayExpress, submission idE-MTAB-407 and the ENA, submission id ERP000367 (http://www.ebi.ac.uk/ena/data/view/ERP000367). Batch 3 reads were submitted to ArrayExpress, submission id E-MTAB-533 and ENA, submission id ERP000509 (not visible yet). We have not agreed on the best way of submitting these to the DCC yet.