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, Q2: 01/01/11 - 03/31/11)**

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 still ahead of target at 119% of the original approved milestone. Spending is tracking the original budget (entirely salaries).
Milestone 2 (Sheet 1: Experimental Validation) is still behind, although progress is being made.Sequence data for ~2000locithat entered the experimental pipeline in the previous quarteris currently being sequenced. A further ~2500 loci have entered the pipeline in this quarter.
Milestone 3 (Sheet 2: Pseudogene Annotation) has increased by 5% in the last quarter to 72%. Spending is tracking the original budget (entirely salaries).
Milestone 4 (Sheet 1: Overall Gene Annotation): The fraction of genes classified as levels 1+2 is on course to reach 90% by the end of Y4. The fraction of genes classified asLevel 1is currently behind target, but has increased by 1.6% to 3.9%. Note that theoriginal target for loci validated as Level 1 using the experimental pipeline isonly 10%. This is because it was planned to only test ~1/3 of loci and the validation rate was projected to be 30%.In fact the validation rate is currently >50%, so it is anticipated that the final fraction labelled level 1 using this protocol will be at least 15%. 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.
We are currently under our original budget for milestone 2 (Sheet 1: Experimental Validation) due to problems and delays. However, now that we are back on track and have varied the design of the pipeline using cheaper pooled next generation sequencing we anticipate being able to do more experiments to complete the genome and extend the range of transcript types tested.
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 and we are on track to reach our original target.This is reflected in the number of gene annotations assigned 'Level 1' status in Milestone 4 (Sheet 1: Overall Gene Annotation) which increased from 2.3 to 3.9% in the last quarter. Currently we are sequencing 2455 RT-PCR experiments and have primers designed for a further 4794 which have been made from the remainder of the GENCODE 6 release. 2500 of these form the next batch in the experiment pipeline. In order to keep the experimental verification process on track we continue to have dedicated biweekly conference calls between Sanger, CRG and Lausanne.

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 genebuildhas been completed. The merged geneset will be released as GENCODE 7 and as the default geneset in release 62 (April 2011) of the Ensembl browser. It has also been submitted for release through the DCC.
Computational pipelines using RNA-Seq continue to be 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.In addition computational analysis of these models and models from the Transcriptomics group and CRG is being carried out to identify consensus models that are not currently represented in GENCODE as priority candidates for annotation, validation and incorporation. There is also on going work between UCSC and Ensembl to develop a computational pipeline to attach expression level labelling to GENCODE transcripts again using RNAseq data.

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. **How are you coordinating with the DCC to submit your RACE and RT-PCR experiments to them?**

Our understanding, based on the agreed structure of the original quarterly report documents, has been that raw sequence assay data from RT-PCR and RACE experiments only needs to be submitted to public databases (Array Express). Now that we have established a robust criterionfor RT-PCR-seq based validation of GENCODE exon junctions we are discussing with the DCC how best to display this supporting evidence at the DCC.

4. **Please provide a list of accession numbers for any new ENA submissions**

The batch 1 targeted sequencing data was submitted to ArrayExpress, submission ID E-MTAB-612 (visible from 1st May). Submission of batch 4 data to ArrayExpress is in progress.

Additional questions

1. **Have you published any papers on ENCODE data in the past quarter? If so, please list the titles and a doi, if available.**

   - Coffey et al. The GENCODE exome: sequencing the complete human exome. doi = 10.1038/ejhg.2011.28
   - Balasubramanianet al. Gene inactivation and its implications for annotation in the era of personal genomics. doi = 10.1101/gad.1968411

2. **Which ENCODE datasets are published in this paper? Please list DCC submission ID numbers**

Coffey et al. used GENCODE 3c data and Balasubramanian et al. (discussion paper) used GENCODE 6 data.