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, Q4: 07/01/11 - 09/30/11)

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 143%of the original approved milestone, with 85% of this figure now released by the DCC. Spending is tracking the original budget (entirely salaries).
Milestone 2 (Sheet 1: Experimental Validation) is still behind, although progress continues to be made with another 2873 genes added to the pipeline this quarter.We unfortunately also found another bug in our reporting which resulted in elevatednumber of genes for batches 5 and 6. These values have now been corrected.
Milestone 3 (Sheet 2: PseudogeneAnnotation) has increased slightly to 74%.Spending is tracking the original budget.
Milestone 4 (Sheet 1: Overall Gene Annotation)has been passed as the fraction of genes classified as Levels 1 + 2 is now over 90%. The fraction of genes classified as Level 1 is still behind target but has increased slightly to 9.7%. This figure is slightly under the original target for loci validated as Level 1. 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.
We are still under our original budget for milestone 2 (Sheet 1: Experimental Validation) due to earlier problems and delays. However, we are now back on track and have varied the design of the pipeline using cheaper pooled next generation sequencing and by using external whole transcriptomics RNAseq data as supporting evidence.As a result we anticipate being able to do more experiments to complete the genome and extend the range of transcript types tested.
NHGRI has extended ENCODE funding for an additional year. This extra year will allow us to annotate the remaining 10% of the human geneset, focusing on protein coding genes.

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.Currently we are sequencing 2873 RT-PCR experimentsbased on GENCODE 8 annotations anda set of approximately 150 pseudogene models.In order to keep the experimental verification process on track we continue to have dedicated biweekly conference calls between Sanger, CRG and Lausanne.
We still hope to evaluate cufflinks models from the RNAseq data produced by the Gingeras labusing the validation experimental pipeline, but this has been delayed due to procedures for filtering themodels being revised following discussion on IDR thresholds and also the regeneration of the models using an updated version of cufflinks.

Project-specific questions

1. **What is the status of your computational predictions?**

The Ensembl-Havana merged pipeline has been dramatically improved since the release of GENCODE 3c, so GENCODE 7 has now been accepted as the reference annotation for the AWG analysis. Some of the improvements in this genesetover GENCODE 3c include more manual annotation (including non-coding RNA which are difficult to predict through the automatic pipeline). During this quarter GENCODE 9was released on the www.gencodegenes.orgsite for collaborators to download and is the default geneset in Ensembl release 64 (September 2011).This genesetrepresents an incremental improvement on GENCODE 8 with additional manual annotation. We are now analysing imported CAGE clusters from the Riken lab in collaboration with the transcriptomics group to improve 5’ UTR annotation.We are continuing to look at pseudogenes that have evidence of transcription when comparing with the IlluminaBodyMap RNAseq data and ENA classical “transcriptional” evidence.We designed primers for a largercandidate set (150) of expressed pseudogenes which are now undergo experimental validation (sequencing) in Lausanne.Development of computational pipelines using RNAseq by Ensembl and MIT and of a confidence level pipeline by UCSC and Ensemblis continuing.

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 is currently unclear how many of these will turn out to be transcribed or translated.It is interesting to assess this further using RNAseq and proteomics data. For the additional year of funding we propose to increase the milestone to 100% to reflect the genome wide coverage by manual annotation now anticipated. However, this new figure will need to be based on a revised estimate of the total number of pseudogenes in the genome.

3. **Please provide a list of accession numbers for any new ENA RACE and RTPCR submissions**

Batch VI sequencing data was submitted to ArrayExpress, submission ID E-MTAB-831.
We are in discussions with our Data Wrangler to get the experimental validation track submitted to the DCC and are currently waiting for a formal proposal from them regarding the submission format.

Publication Information

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

No

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

N/A