Grant Number: U54 HG004555-04S1
Project Title: Integrated human genome annotation: generation of a reference gene set(GENCODE)
Quarterly progress report - *Narrative Questions* (Year 5, Q1: 10/01/11 - 12/31/11)

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 103% of the revised milestone, with 84% of this figure now released by the DCC. Spending is tracking the original budget (entirely salaries).
Milestone 2 (Sheet 1: Experimental Validation) continues to increase and is now at 78% of the revised milestone. Spending is now tracking the original budget.
Milestone 3 (Sheet 2: Pseudogene Annotation) continues to increase and is now at 76%. Spending is tracking the original budget.
Milestone 4 (Sheet 1: Overall Gene Annotation) continues to increase as the fraction of genes classified as Levels 1 + 2 is 95%. The fraction of genes classified as Level 1 is still behind target but has increased to 12.4% this quarter. 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 revised project objective of a 100%, 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 anticipate reaching our milestone by the end of the funding period.

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 now on target. In order to keep the experimental verification process on track we continue to have dedicated biweekly conference calls between Sanger, CRG and Lausanne. After initial delays we are currently evaluating cufflinks models from the RNAseq data produced by the Gingeras lab using the validation experimental pipeline. As they are a new source of novel genes, once evaluated we propose to add these models to our milestone.

Project-specific questions

1. **Under your "Overall Gene Annotation" section, your data totals to more than 100% for the percentage of genome evaluated for genes. Please explain this discrepancy.**

The percentage in this category is based on a figure of 30000 total genes, excluding pseudogenes and RNA genes. This figure was based on an estimate and after re-evaluating the total number of genes we propose to increase this figure by 2500 to 32500 as a revised estimate. This includes the increased number of lncRNAs we have found as well as additional Illumina Body Map/cufflinks predicted loci from RNAseq data on chromosomes 17, 18 and 19 which we will incorporate in the completion of first pass annotation.

2. What is the status of your computational predictions?

During this quarter GENCODE 10 was released on the www.gencodegenes.org site for collaborators to download and is the default geneset in Ensembl release 65. This geneset represents an incremental improvement on GENCODE 9 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 Illumina Body Map RNAseq data and ENA classical "transcriptional" evidence. Development of computational pipelines using RNAseq by Ensembl to predict novel transcript structures; using comparative genome alignments by MIT to predict unannotated coding regions and of a confidence level pipeline by UCSC and Ensembl are continuing.

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

Currently this still seems a reasonable genome wide estimate of the consensus set; although it is currently unclear how many of these will turn out to be transcribed or translated. It will be interesting to assess this further using RNAseq and proteomics data.

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

Batch VII sequencing data was submitted to ArrayExpress, submission ID E-MTAB-935. After discussion with the DCC we agreed that all our experimental validation data should be submitted directly to GEO, who will liaise with ArrayExpress to get our data incorporated.

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.

Djebali et al., Evidence for Transcript Networks Composed of Chimeric RNA in Human Cells, doi: 10.371/journal.pone.0028213

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

Djebali et al. used GENCODE 3C data