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, Q3: 04/01/11 - 06/30/11)

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 129% of the original approved milestone. Spending is tracking the original budget (entirely salaries).
Milestone 2 (Sheet 1: Experimental Validation) is still behind, although progress continues to be made.In this quarter a substantial amount of experimental validation data feed into the labelling of genes in the latest GENCODE release and into displays at the DCC (reflected in level 1 labelling of Milestone 4). We unfortunately also found a bug in our reporting which resulted in the 'Completed not submitted' value being 648 higher than was correct between Y3Q3 and Y4Q2. These values have been corrected.
Milestone 3 (Sheet 2: Pseudogene Annotation) has remained at the same percentage (72%).Spending is tracking the original budget.
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 in the last quarter by 5.5% to 9.4%. As noted previously that theoriginal target for loci validated as Level 1 using the experimental pipeline isonly 10%, however since thevalidation rate is currently >50% 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.
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 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 indicated that ENCODE funding will be extended for an additional year. Year 5 funds 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 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 3.9 to 9.4% in the last quarter. Currently we are sequencing 2549 RT-PCR experiments and are designing primersbased on annotations from the new GENCODE 8 release and a set of pseudogenes. In order to keep the experimental verification process on track we continue to have dedicated biweekly conference calls between Sanger, CRG and Lausanne.
We hoped 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). GENCODE 8 has now been released on the www.gencodegenes.orgsite for collaborators to download and will be the default geneset in Ensembl release 63 (June 2011). This geneset contains the new manual annotation for chromosomes 11 and 14 (80%) and increases in the number of lncRNA, including the six different biotypes within this category, as a separate GTF file for collaborators to use and examinee.g. CRG.We have also imported new CAGE clusters from the Riken lab in collaboration with the transcriptomics group as a guide for 5' UTR annotation.We are also starting to look at pseudogenes that have evidence of transcription when comparing with the IlluminaBodyMap RNAseq data and ENA classical "transcriptional" evidence. We have designed primers for a pilot(10) set of expressed pseudogenes to undergo experimental validation in Lausanne.Computational pipelines using RNA-Seqcontinue to be developed by Ensembl and MIT. Both groups have used the IlluminaBodyMap dataset to construct gene models with their specialized methods. MIT have recently identified another small set of high confidence exons missing from the current GENCODE set, based on a combination of RNAseq data, gene model prediction and conservation evaluation (BodyMap2 + Scripture + PhyloCSF Pipeline). These predictions are being assessed by Havana.

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.

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

Batch IV sequencing data was submitted to ArrayExpress, submission ID E-MTAB-684 (ENA Accession: ERP000774) and will be visible from 22nd August. Batch V data was submitted to ArrayExpress with submission ID E-MTAB-737.

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.

Lin et al., PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. doi = 10.1093/bioinformatics/btr209

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

Lin et al. used GENCODE 6 data.