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, Q2: 01/01/12 - 03/31/12)

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 110% 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) has now been passed as is ahead of target at 121% although the percentage released by the DCC has not changed in the last quarter. Spending this quarter is over the original budget.
Milestone 3 (Sheet 2: Pseudogene Annotation) continues to increase and is now at 78%. 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 now 100%. The fraction of genes classified as Level 1 is still behind target but has increased to 19.4% 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, now that things have improved in the experimental verification process. This quarter we are over budget for milestone 2 (Sheet 1: Experimental Validation), however overall we are still under our original budget due to earlier problems and delays.

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. We continue to evaluate cufflinks models from the RNAseq data produced by the Gingeras lab using the validation experimental pipeline and have now added these models to our milestone.

Project-specific questions

1. **What is the success rate of your experimental validation experiments and how will that feed into your peak calling?**

Experimental success rate = 79%.

2. **In your Y5Q1 narrative, you report finding lncRNAs and additional Body Map/cufflinks predicted loci from RNA-seq data on chromosomes 17, 18 and 19. Do you expect the gene number to go up given that you only report cufflink gene loci for some of the smaller chromosomes?**

In our Y5Q1 we believe the response may have been misunderstood. There are additional lncRNAs on all chromosomes but we are currently finishing the manual annotation of chromosomes 17, 18 and 19 and thus at present we are only integrating the models on these chromosomes, since each model currently requires manual review.

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

During this quarter GENCODE 11 was released on the www.gencodegenes.org site for collaborators to download and is the default geneset in Ensembl release 66. We have also started to analyse the

454 RACE data from testes and brain, generated from lncRNAs without CAGE data or polyA signals. We have found a few examples of extensions and now have to systematically analyse the data once the alignment parameters have been improved.

4. 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.

5. Please provide a list of accession numbers for any new ENA RACE and RTPCR submissions (please also make sure this data is being submitted to both ArrayExpress and GEO).

Batch VIII sequencing data is currently being submitted to ArrayExpress. Accession numbers for all batches submitted are shown below:

	ENA	ArrayExpress	GEO
Batch 1	ERP000605	E-MTAB-612	GSE30619
Batch 2	ERP000367	E-MTAB-407	GSE25711
Batch 3	ERP000509	E-MTAB-533	GSE30612
Batch 4	ERP000774	E-MTAB-684	GSE34797
Batch 5	ERP000781	E-MTAB-737	GSE34820
Batch 6	ERP000972	E-MTAB-831	GSE34821
Batch 7	ERP001145	E-MTAB-935	GSE37592

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.

Ezkurdia et al., Comparative Proteomics Reveals a Significant Bias Toward Alternative Protein Isoforms with Conserved Structure and Function, doi: 10.1093/molbev/mss100

Frankish et al., The importance of identifying alternative splicing in vertebrate genome annotation, doi: 10.1093/database/bas014

Harte et al., Tracking and coordinating an international curation effort for the CCDS Project, doi: 10.1093/database/bas008

MacArthur et al., A systematic survey of loss-of-function variants in human protein-coding genes, doi: 10.1126/science.1215040

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

Ezkurdia et al., used GENCODE 3C data, Frankish et al., used GENCODE 7 statistics, Harte et al., references GENCODE 11 and relationship with CCDS and MacArthur et al., used GENCODE 8 data.