Grant Number : U54 HG004555-03

Project Title: Integrated human genome annotation: generation of a reference gene set (GENCODE)

Quarterly progress report - *Narrative Questions* (Year 3, Q2: 01/15/10 - 03/31/10)

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) is progressing well and ahead of target. The additional work associated with QC and update of existing annotation for which there is no milestone resulted in 1097 loci being updated in addition to the completely novel loci. Spending is tracking the original budget (entirely salaries).

Milestone 2 (Sheet 1: Experimental Validation) is still substantially behind, but will be accelerated now. Spending is therefore currently lagging the original budget (salaries and experimental reagents). It is anticipated that the milestone can still be achieved once the new strategy is implemented (see 4).

Milestone 3 (Sheet 2: Pseudogene Annotation) is on target, although the number of 3-way-verified genes in the current freeze data is still held at 32% (since Y2Q3). The cause for this was now identified as a modification in repeat-masking which causes the automatic pipelines to miss many genes. The previous number (51%) was an estimate avoiding this error. Spending is tracking the original budget (entirely salaries).

Milestone 4 (Sheet 1: Overall Gene Annotation) is behind in terms of the fraction of genes classified as level 1, however the faction currently classified as level 2 exceeds this figure. As the statistics for the previous two reports were not based on data freezes, but generated by analysing the live data in the database there are slight fluctuations. This results in the number of level 2 genes decreasing slightly between Y3Q1 and Y3Q2. As for milestone 3, milestone 4 is a percentage, related to the original projected number of non-pseudogenes (30,000). This number was estimated as the total number of protein coding and long non coding genes. The Ensembl pipeline which generates level 3 annotation, identifies a large number of small RNAs (such as microRNAs and matches to Rfam domains) which were not part of this estimate as they were not part of GENCODE milestones (see 2). As such, although GENCODE data releases include them, the total number of loci reported excludes RNA genes that are submitted as level 3 annotation (sheet "Genes", row 82; Y3Q2: 8414 RNA genes). 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 meeting the original project objectives of a complete, verified human geneset focusing on protein coding genes. However during the project the meaning of 'complete human geneset' has been expanded as we anticipate high quality RNAseq experimental data providing evidence for additional alternative transcript forms of existing protein coding genes and increasingly robust evidence for non coding RNA genes (ncRNAs). While we anticipate that our human geneset at

the end of year 4 will incorporate annotation corresponding to a significant amount of this additional evidence, it is unlikely that all will have been incorporated, particularly for short ncRNAs.

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 significantly delayed. Modifications to the experimental procedure had to be made as dimerisation of primer sequences seemed to be influencing the results. Significant time was spent on re-writing the computational primer-selection pipeline and analyzing the sequencing results. This will result in more reliable experimental results as primers are more stringently checked for potential dimerisation and for unique sequences. Mono-exonic transcript can also be handled specifically.

The dedicated weekly telephone conferences between Sanger, CRG and Lausanne to monitor and improve progress of the experimental verification are continuing and have proven valuable.

We also had to spend more time than anticipated on the definition of suitable un-annotated regions for testing RGASP predictions.

The second round of data submission for RGASP was successful and will allow more straightforward comparisons. The data analysis is still ongoing with various group from within and outside of ENCODE focusing on different aspects of the data.

As already discussed previously the experimental verification pipeline has been changed from capillary to next generation sequencing. Evaluation of multiple complementary next generation sequencing is ongoing. Beyond this, the overall structure of the pipeline has not changed. Collection of new evidence and the development and refinement of computational methods for the evaluation of the GENCODE annotation by each group (see below) is an ongoing feature of the project. Updated output from computational algorithms are integrated as they arise in the ANNOTRACK system and the annotation software. The tools for manual annotation are also being continuously improved to allow better QC. Improvements in the merge process continue to improve stability.

<u>Project-specific questions</u>

1. What is the status of your computational predictions?

All computational pipelines continue to be regularly rerun; provided to Sanger via DAS and integrated via the ANNOTRACK system to flag issues with existing annotation and potential missing genes and transcripts.

Some of the improvements to the computational analysis pipelines used in the GENCODE process are as follows:

- The Ensembl automatic gene annotation pipeline is used both for the merge with the manual annotation and as an input to the ANNOTRACK evaluation system, in particular for detecting missing annotation in CCDS genes.
- The new computational pipeline from WashU/UCSC helping to identify errors in the annotation of splice-sites was further improved to also use EST data. This data has proven useful with a current ratio of 48% of flagged cases resulting in an update of the HAVANA annotation.

- The pseudogene prediction pipelines are being run regularly, but currently suffer from modifications to the repeat-masking in the underlying sequence database.

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

Currently this seems a reasonable genome wide estimate, although it's possible the final figure for consensus pseudogenes by this criteria will end up slightly higher.