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, Q1: 10/1/09 - 12/31/09)

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. This is despite the additional work associated with QC and update of existing annotation for which there is no milestone: In the quarter Y3Q1 601 (793 transcripts) loci were revised because of new computational QC analysis, in total 2525 loci were updated in addition to the completely novel loci. This workload is expected to increase as a result of evidence for additional alternative transcript forms from RNAseq experimental data (see 4). Spending is tracking the original budget (entirely salaries).

Milestone 2 (Sheet 1: Experimental Validation) is substantially behind. This is mainly due to production work being halted after an initial validation test on chromosomes 21 and 22 in order to evaluate if transcriptome sequencing using the next generation platforms (RNAseq) could be substituted to allow both cheaper and more comprehensive experimental validation. This is linked to the more general evaluation of automatic annotation using RNAseq currently being carried out (RGASP). 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 it should be noted that the milestone is a percentage, related to a projected total number of pseudogenes (see question below). The decrease of level 1 pseudogenes in Y2Q3-4 was caused by an error in the comparison calculations. 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. Levels 1, 2 and 3 represent annotations with different degrees of verification: Level 3 is evidence based computational annotation (Ensembl automatic pipeline); Level 2 is evidence based manual annotation (Havana); Level 1 is validated Havana. In the original plan, only a subset of level 2 genes, selected by biotype, were to be validated experimentally (Milestone 2) and promoted to level 1 if confirmed. Genes that were not targeted for experimental validation would be promoted to level 1 following QC against multiple computational pipelines if no anomalies were found. However, with the advent of RNAseq, some degree of genome wide experimental QC now seems possible, so it is planned to include this in the criteria for level 1. Criteria will be established partly based on results from the RGASP evaluation of use of RNAseq in annotation. Therefore, up to now, only a small number of genes have been promoted to level 1, based on experimental data.
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; Y3Q1: 6366 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?

The experimental verification process has been significantly delayed. There were initial delays to primer design for RT-PCR. More recently production verification has been suspended as alternative strategies involving next generation sequencing have been tested. 3 way comparisons of (1) RT-PCR + capillary sequencing; (2) RT-PRC + next generation sequencing; (3) whole genome RNAseq transcriptomics are being carried out to assess the different strategies performance for validation. It has been established that RT-PCR + capillary sequencing can be substituted by RT-PRC + next generation sequencing. However currently it appears that directed (pooled RT-PCR) and whole genome RNAseq transcriptomics are complementary. Since (2) is significantly cheaper than (1) it is planned to carry out both (2) and (3) on the range of tissues originally proposed. In order to speed up the optimisation of this pipeline and carry out the above evaluations, we have for several months organised dedicated weekly telephone conferences between Sanger, CRG and Lausanne to monitor and improve progress. This is in addition to the bi-weekly calls of the whole GENCODE project.

A second bottleneck is the use of whole genome RNAseq for gene annotation. It is clear that this data is valuable, but algorithms that use it need calibration to ensure they do not generate large numbers of false positive annotations. To address this we have organised the RGASP competition to evaluate the ability of automatic gene annotation algorithms to use RNAseq data. This evaluation is ongoing with bi-weekly conference calls of the organising committee, a first workshop help in November 2009 and a second planned for late February 2010.

4. How has your pipeline changed since the beginning of the project? Please report for all Quarters since the last quarter we gathered progress reports - Y2Q1.

As already discussed (see above) 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. The tools for

manual annotation are also being continuously improved to allow better QC. One change during this period has been the migration of the entire system including all computational pipelines from the NCBI36 to the GRCh37 assembly. Since ENCODE as a whole still works mainly on NCBI36, output datasets are provided on both assemblies, the NCBI36 version being a projected one.

One change in the overall pipeline has been to the process of generating genome wide data freezes. GENCODE is based around Havana manual annotation, however since this is not yet genome wide, gene annotation from the Ensembl automatic annotation pipeline is merged with it. Prior to Gencode release 3 in late 2009, the merge process was different to the process used by Ensembl itself to incorporate Havana annotation into its geneset (a process that existed prior to the start of the GENCODE project). With Gencode release 3, a single merge process was used resulting in the Ensembl human geneset becoming a release of Gencode. Improvements in the merge process continue, including efforts to improve consistency and completeness between GENCODE, CCDS and Uniprot.

Project-specific questions

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. As mentioned above, in the last 6 months the core annotation systems at Sanger and all external pipelines have migrated from NCBI36 to GRCh37. These include Yale's PseudoPipe and UCSC's RetroFinder for pseudogenes; MIT's CONGO and RFC+CSF pipelines for finding candidate novel exons and assessing the conservation of existing gene annotations; UCSC's ExoniPhy and TransMap pipelines; and WashU's transcriptome alignment pipelines.

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. It is being further optimized for the handling of special cases (e.g. Immunoglobulin genes).

A new computational pipeline from Washu / UCSC is helping to identify errors in the annotation of splice-sites. This was developed following an RNA-seq summit held at UCSC in early 2009 and subsequent effort to evaluate methods and construct pipelines to use RNAseq data to make predictions for RGASP.

MIT have developed a new computational pipeline that can identify regions within human ORFs that appear to have evolutionary constraint on their synonymous sites. The method can locate regions of just nine codons, by analyzing comparative sequence alignments of the genomes of 29 placental mammals. This pipeline has the potential to identify transcripts that encode amino acid sequences in two or more reading frames.

The pseudogene prediction pipelines are being run regularly and have targeted a number of pseudogene families, including ribosomal pseudogenes, glycolytic enzyme pseudogenes, nuclear Receptor pseudogenes, pseudogenes associated with segmental duplication and unitary pseudogenes. A major topic that has involved multiple groups has been the identification of polymorphic pseudogenes, that have both non-functional and functional alleles currently

segregating in the human population. The significance of these cases and how to represent these on the human genome is an ongoing discussion involving 1000 genomes data and the Genome Reference Consortium.

2. What do you believe to be the total number of pseudogenes? (This is the number that figures into the % of pseudogenes annotated on the "Pseudogenes" tab. In future quarters this question will be "Do you still believe *x* to be the total number of pseudogenes?")

The current estimate is 10,000 based on previous analysis (Zheng et al. Genome Res. 2007). Currently this seems a reasonable genome wide estimate, although its possible the final figure for consensus pseudogenes by this criteria will end up slightly higher.