Application Number : 1 U54 HG004555-01

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

Quarterly progress report summary (Q4: 7/1/08 – 9/30/08).

Manual Annotation

The HAVANA group has continued the *de novo* manual annotation of chr2 (1176 new loci) and started annotation of chromosomes 3 (26 loci) and 7 (160 loci) this quarter (including pseudogene loci). This has resulted in HAVANA exceeding their milestone of 2250 to a total of 3613 loci this year, although the milestone requirements will escalate in the remaining years of the project, and there will be additional load on manual annotation marrying together contributions from other partners.

HAVANA also started assessing the validity of 514 unitary pseudogene predictions from Yale. 17 of 34 predictions checked were confirmed as unitary pseudogenes (by validating the parent gene in mouse as well as the pseudogene prediction in human) and a further 98 (of 98) predictions of olfactory receptor family pseudogenes were confirmed but excluded from the unitary pseudogene set as orthology could not be determined with confidence.

HAVANA also reviewed RFC (reading frame conservation) and CSF (codon substitution frequency) analysis from MIT of annotated protein coding transcripts. The transcripts with the lowest HC-RFC (RFC in high 5 coverage genomes) score (*ie* most likely to be non-coding) were investigated in detail and in only a single case was the annotation updated to remove the CDS. A selection of higher scoring coding transcripts were also checked and none were identified as requiring changes to CDS. For CCDS contributions see the section below.

The Ensembl group has identified genomic locations that require annotation review by HAVANA for various reasons explained below:

The first set of locations identified were those where HAVANA has annotated more than one protein-coding gene in a region and these protein-coding genes overlap at a coding level. Such cases are currently not allowed in the Ensembl genebuild system and they therefore cause incorrect merging of Ensembl genes during the Ensembl-HAVANA merge process.

The second set of locations contained cases where Ensembl has predicted a protein-coding exon, but HAVANA has not predicted a protein-coding exon at the same position. These locations can identify cases where Ensembl is making use of a new protein sequence that has not yet been annotated by HAVANA. Similarly, introns unique to Ensembl were also identified in an effort to find cases where Ensembl and HAVANA are annotating different splice variants for a coding region. It is hoped that agreement between Ensembl and HAVANA splice variants will increase the number of CCDS candidates.

The final set of locations identified were those cases where Ensembl has aligned a cDNA or EST to the genome and where that alignment produces an exon strucure different to that annotated by HAVANA.

Additionally, Ensembl has an in-house database for storing protein, cDNA and EST sequences that are discarded from the set of sequences used in the genebuild process. This database has been shared with HAVANA in order to maximise on the information generated for future analysis.

Finally, at WTSI the "Gentrack" tracking system has now been successfully designed and implemented for the integration of the data from all groups. The annotators are using it to prioritise annotations to be re-examined and have started to store controlled vocabulary comments were appropriate. WTSI submitted the first combined dataset on time (Oct 1st 08) to the DCC. This included the first set of consensus pseudogenes between Havana, Yale and UCSC on chr21/22 pilot data.

Experimental Validation

To sustain the production of a reference gene set, CRG have streamlined and refined their pipelines for both, the experimental validation of targeted transcripts and the identification of U12 introns. A general workflow (Figure. 1) has been set up based on the HAVANA team annotation for chromosomes 21 and 22. This workflow is also effective for other specific computational pipelines and has already been tested using the CRG U12 introns prediction pipeline. Briefly, the more up to date HAVANA annotation is collected using the DAS server provided by WTSI and all loci exhibiting a « novel » or « putative » status are extracted from this annotation. Then, specific primers are automatically generated (using primer3) within exon pairs of a targeted transcript, and primer sequences are sent to Lausanne to perform RT-PCR experiments. Up until now, approximately 650 primers pairs referring to HAVANA manual annotation have been designed and sent for experimental validation and 31 primers pairs have been generated to test U12 introns predictions (Table 1). Finally, CRG can provide DAS sources for all of the CRG internal pipelines making data available to the rest of the consortium members.

Figure 1 : Experimental validation workflow

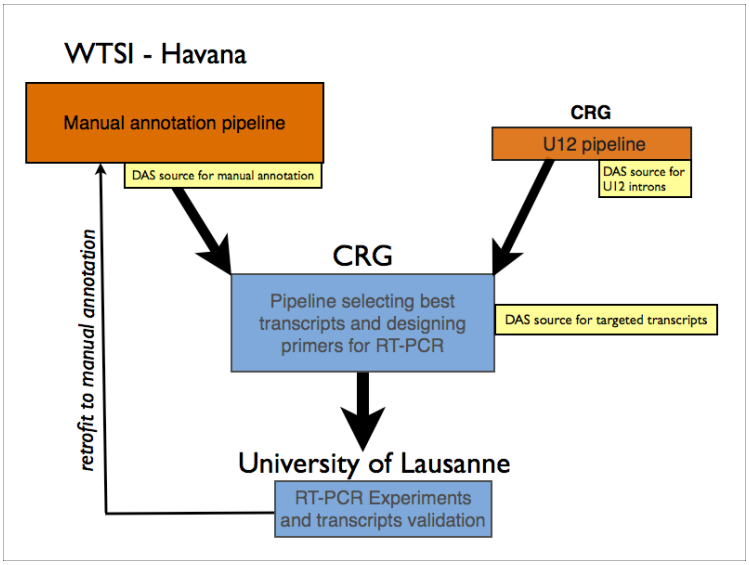

Table 1 : Status of the experimental validation

	HAVANA MANUAL PIPELINE				U12 Pipeline	TOTAL
Havana Biotype Tag	protein_coding	processed_transcripts	processed_pseudogene	Tobe Experimentall Confirmed (TEC)	NA	NA
Number of loci	655	445	228	4	NA	NA
Number of transcripts	2130	1680	228	4	46	4088
Number of loci with Status NOVEL and PUTATIVE	49	395	220	4	NA	NA
Number of primers designed	44	380	220	4	31	679
Number of primers sent to Lausanne	44	380	220	4	0	648

The University of Lausanne have 640 novel transcripts which are currently in the experimental pipeline. Severe delays in primer delivery from the manufacturer have introduced a large shortfall in the production of RT-PCR results which have lead to a knock-on effect on the analysis from this portion of the project. A combination of experimental strategies has been devised in this quarter to fulfil future deadlines and also in the long term to deliver results of higher quality than previously planned. Analysis on the 640 novel transcripts will be assessed by both WTSI, Solexa and 454 sequencing to allow for a systematic approach to decide on the optimal pipeline for the remaining verifications.

Pseudogene Assigment

Through Yale's biweekly pseudogene conference calls with UCSC and WTSI, Yale have continued to examine several problematic pseudogenes, particularly unitary pseudogenes. A consensus list of computationally discovered pseudogenes (combining Yale's PseudoPipe and UCSC's RetroFinder lists) was developed and shared with the manual annotators. Several areas were flagged where the computational methods agreed on a pseudogene, but HAVANA did not annotate. Yale hope to use these flagged areas to improve both the computational methods and the annotator's criteria. A web tool to help track Yale's and other members of the pseudogene sub-group progress with these lists was developed and deployed at http://gencode.gersteinlab.org/consensus/.

As mentioned above UCSC have taken part in pseudogene conference calls with the Yale and WTSI (HAVANA) groups. UCSC set up an area on the ENCODE wiki hosted at UCSC for the posting of pseudogenes conference call minutes and presentations. It can be found in the Gene annotation section. Most recently, UCSC discussed the use of a pseudogene ontology as proposed by the Yale group and the production of a Yale/UCSC consensus processed pseudogenes set. The UCSC retroposed genes (predicted by RetroFinder) track was updated to version 3.0 (see Retroposed Genes 3.0 on the human hg18 assembly at http://hgwdev-gencode.cse.ucsc.edu). Additional features were added to the track to allow various mRNA rendering modes to be turned on; these may be viewed as one zooms in to base level. These modes are: translation of the CDS to genomic codons, translation to mRNA codons, nonsynonymous mRNA codons, all mRNA bases, mRNA bases that are different from the reference genome. Nonsynonymous codons are displayed with the one-letter code for the amino acid encoded by the mRNA codon and the codon is colored yellow if the change results in an amino acid with similar physicochemical properties and red if it is different. The parent gene CDS is projected on to the retroposed gene in the Retroposed Genes 3.0 track display on the UCSC Genome Browser. Using the projected CDS region, 965 retrogenes that have no CDS, could be detected. A subset of these retroposed genes (764) have parent genes that do not have a CDS annotated in their GenBank record. For these retrogenes, we do not know if they contain the parent CDS or whether are composed of only UTR from the parent gene. There are 201 retrogenes that are predicted to be all UTR (the parent has an annotated CDS and the retrogene has no CDS

projected from the parent gene). This set of 965 retrogenes were sent to the HAVANA group at for further evaluation.

A set of 12,079 retrogenes (predicted by RetroFinder), with a score greater than 650, were sent to Yale for comparison to the Yale processed pseudogenes (predicted by PseudoPipe) in order to create a consensus pseudogenes set. Yale have now created the first set of consensus processed pseudogenes based on these annotations.

Gene Prediction

The MIT group has computational pipelines for assessing the quality of existing gene annotations and for predicting new exons, both based on the comparative evidence in whole-genome alignments of many mammalian genomes. In this quarter, MIT worked with the HAVANA curators to assess the results of their pilot runs on chromosomes 21 and 22 that they generated last quarter (and refreshed this quarter). In both cases, this has led to changes in the current HAVANA annotations, flagging of cases for experimental followup, and valuable feedback on how MIT can fine-tune their pipelines and filter the results to maximize their usefulness to the curators. While only a few "major" annotation changes were made so far in this pilot phase (e.g. adding or removing CDS annotations), MIT can expect their results to have a much greater impact genome-wide, since chromosomes 21 and 22 were already very thoroughly annotated.

MIT are on track with their milestones, having achieved their month 12 goals (chromosomes 21 and 22 pilot). One bottleneck that arose recently was a lack of sufficient parallel compute power for their comparative exon predictor, which is based on a probabilistic model with a computationally intensive training procedure. To address this, MIT applied for and received an allocation on the NSF's TeraGrid supercomputing facilities. MIT are currently setting up their software on these systems, and in the next quarter they will run their exon predictor with a more evidence features and more training data, towards their month 18 goals.

At Washington University they have successfully run their annotation pipeline every quarter since spring of 2008. In terms of pipeline runs, Washington University have accomplished their goals. However, they have not achieved as much as they had hoped in terms of submission to the DCC and communication with the WTSI via the Distributed Annotation System (DAS) protocol. The reasons for this include: Washington University did not have a subcontract agreed until Spring of 2008 (they had anticipated starting earlier), a primary programmer for this project required an unanticipated, extended leave this summer, and the amount of programmer effort required exceeded what we had anticipated. The late start and unanticipated leave also led to Washington University being unable to spend a large fraction of the money they were awarded during the first year. Efforts to meet the demands of DCC submission and DAS-based communication are ongoing and Washington University expect that they can be met if they are given permission to pay for additional programmer/analyst time using unexpended funds from the first budget year.

The CNIO has made satisfactory progress towards refining the protein analysis pipeline that will add to the annotation. CNIO are on course to be able to add reliable protein structural and functional annotations by early 2009. CNIO have encountered a series of technical difficulties, but no real bottlenecks. The CNIO is receiving money just for travel, but will be able to accomplish their milestones within their work budget (which is zero).

CCDS update

During the Q4 period, 97 CCDS have been reviewed by UCSC. HAVANA have continued updating CCDS models, have resolved a further 65 conflict cases in collaboration with RefSeq and UCSC. Both groups have been contributing to a publication that is being written to describe the CCDS database and pipeline. A set of guidelines for CCDS translation initiation site annotation were established as a result of a discussion between representatives from NCBI, HAVANA and UCSC. At USCS, a tool (geneStarter) has been under development to aid the CCDS curation process; it extends transcripts using mRNAs (clustered by PASA and provided by the Brent lab at WUSTL) and searches for upstream translation initiation sites in order to define an extended ORF. Predicted proteins from these ORFs were analyzed; HMMER was used to predict PFAM protein domains, SignalPv3.0 was used to predict signal peptides and TMHMMv2.0 was used to predict transmembrane domains. The output of the pipeline is in a format that can be loaded into the UCSC Genome Browser as custom tracks in order to visualize the location of the predicted CDS starts and the predicted domains and signal peptides.

SUMMARY

Question 1) What is your assessment of your progress relative to your milestones and to the amount of money you have spent?
Answer: In terms of manual annotation the HAVANA group has exceeded their year 1 milestones and have achieved this within budget. The milestones set by the experimental group working out of the University of Lausanne have not been reached but they are in the process generating results (see below question 3). In general we are satisfied with progress.

Question 2) Do you anticipate being able to accomplish your milestones within your budget? If not, what changes are planned?
Answer: We are confident that the milestones for the different project components; manual annotation, experimental validation, pseudogene assignment and gene prediction continue be achievable within budget. However, we are investigating whether the use of next generation sequencing technology would improve our experimental validation approach.

Question 3) What bottlenecks have you encountered and how are you addressing these?
Answer: The University of Lausanne switched primer supplier in order to reduce supplies costs. The primer supplier was unable to deliver on time the batch of primers for the first set of experiments leading to a time lag in the production of validation results. Due to the adoption of 386 well plates the throughput will be considerable greater than using the previous 96 well plates, and therefore it is anticipated that the production of experimental results will be back on course during year 2 of this study.

MIT have experienced a lack of sufficient parallel compute power for their comparative exon predictor (see above Gene Prediction section). MIT overcame this by receiving an allocation on the NSF's TeraGrid supercomputing facilities. MIT are currently setting up their software on these systems, and in the next quarter they will run their exon predictor with a more evidence features and more training data, towards their month 18 goals.