Application Number : 1 U54 HG004555-02

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

Quarterly progress report summary (Year 2, Q1: 10/1/08 - 12/31/08)

Manual Annotation

The HAVANA group has completed the de novo annotation on chromosomes 2 which will now be quality checked and then loci tagged for experimental validation. HAVANA have moved onto de novo annotation of chromosomes 3 and 7 and have annotated a total of 947 new loci on these chromosomes. HAVANA have also re-evaluated loci highlighted by the different analysis pipeline (Ensembl cDNAs/Yale pseuodgenes/CONGO) on chromosomes 21 and 22. A total of 192 transcripts were updated as a result with 572 remaining unchanged. Unfortunately recruitment has been a problem within the group resulting in one staff position which has still to be filled.

In addition, the GenTrack tracking system was further extended and now contains data from every GENCODE group. Regular updates of HAVANA data can be run and a number of different problem categories can be flagged automatically. This has improved identification and prioritisation of loci needing reannotation. A stable procedure to produce data freezes has also been developed.

The Ensembl group has continued to collaborate with HAVANA to improve both manual and automated gene annotation. In their collaboration with HAVANA, Ensembl highlighted locations where manual and automated annotation does not agree, these indicate regions that may benefit from revised annotation. Ensembl provided HAVANA with a set of locations where Ensembl and HAVANA's annotation of CCDS genes were out of sync. This includes 34 locations where HAVANA had annotated a CCDS gene that was withdrawn from the CCDS set, and 8 locations where HAVANA had not annotated a gene at a position where Ensembl had predicted a gene with either an HGNC symbol or that is in the CCDS set. Similarly, Ensembl provided HAVANA with a set of locations where Ensembl and HAVANA's annotation of genes tagged with an HGNC symbol were out of sync. Ensembl found 1173 genes where HAVANA had not tagged a gene with an HGNC symbol where Ensembl had, or where the HAVANA gene name was an older synonym of the HGNC symbol currently used by Ensembl. We also found 317 genes where Ensembl had not tagged a gene with an HGNC symbol where HAVANA had. Ensembl also provided HAVANA with a list of 598 Ensembl cDNA alignments that overlap HAVANA-annotated processed and unprocessed pseudogenes. Lastly, Ensembl have provided HAVANA with a list of potential alternate translation start sites (TSS). These are in-frame ATGs within 200 bp upstream of HAVANA's currently annotated coding start site. As part of the CCDS project, Ensembl ensure that their CCDS annotation is up-to-date. This was done for both Ensembl releases this quarter: release 51 and 52.

Ensembl have successfully created a pipeline to merge the HAVANA and Ensembl genes were the annotations agree. This pipeline was run and the results updated for Ensembl releases 51 and 52. Ensembl release 52 held 16937 merged genes, comprising 27441 merged transcripts. Ensembl now also provides a pre-release of the Ensembl gene set to HAVANA. This gene set matches the gene set for the Ensembl release, but is handed to HAVANA approximately one month prior to the public Ensembl release.

Experimental Validation

668 out of the 4042 HAVANA annotated transcripts (1332 loci) on the HSA21 and HSA22 chromosomes belong to the “unknown”, “putative transcripts” and “novel transcripts” categories. 662 of these models (99%) were selected in conjunction with the CRG Barcelona group for experimental verification by RT-PCR in 12 different tissues, verification on gels and direct sequencing. The University of Lausanne were able to design suitable primers for 648 of these models and experimentally verify about 40% (n=163). A subset of the verified transcripts will be selected for further characterization by RACE. To reduce the cost of the verification procedure in the future and increase throughput The University of Lausanne are currently testing the possibility of performing direct sequencing of large pools of RT-PCR reactions with an HTS Illumina platform.

Together with the U12 and selenoprotein pipelines, CRG have been focusing their pipeline on charaterizing GENCODE long non-coding RNAs. Starting with the HAVANA annotation in processed transcripts without any Open Reading Frame, CRG identified more than 3000 ncRNAs located outside of protein coding genes and with evidence for both, alternative splicing events and sequence conservation within the mouse genome. Moreover, the increasing number of RNASeq datasets (transcriptome profiling using deep sequencing technologies) allowed CRG to confirm that most of these ncRNAs are expressed in the cell and that their expression patterns are tissue-specific. The pipeline for experimental validation has been set up and the first set of transcripts to be targeted (n=650) have been tested in Lausanne. According to the results of the experiments, CRG may streamline the conception and design process of their analyses to sustain the production of a reference gene set. In all cases, CRG wish to develop a MySQL database in order to facilitate the storage, query and comparisons of data related to the experimental validation.

Pseudogene Assignment

The Yale PseudoPipe was run again with the most recent data and a new computational consensus set was created. An analysis of the pseudogene overlaps was undertaken to determine the ideal method of combining the different annotation sets. Two papers were sent to press and will be published in early 2009: "Pseudofam: the pseudogene families database" and "Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes." The regular pseudogene conference calls continued with further discussions on the unitary and ribosomal pseudogenes. Progress has been proceeding on the expected pace. No significant bottlenecks have been encountered.

UCSC have taken part in pseudogenes conference calls with the Yale and WTSI (HAVANA) groups. During these calls, UCSC have discussed methods for producing a consensus set of processed pseudogenes for HAVANA manually annotated processed pseudogenes, Yale processed pseudogenes predicted by PseudoPipe and retrogenes predicted by UCSC’s RetroFinder. UCSC have also discussed the annotation of some interesting pseudogene cases as well as the creation of a pseudogene ontology proposed by Yale.

Collaboration continued between HAVANA and Yale’s pseudogene group to analyse 309 unitary pseudogenes automatically predicted which resulted in 87 being manually annotated by HAVANA. Yale has also forwarded HAVANA a small ribosomal pseudogene list to check which lead to a new ribosomal gene being identified.

Gene Prediction

The MIT group is developing computational pipelines that use comparative evidence from whole genome alignments of many mammalian genomes (1) for assessing the quality of existing gene annotations (CSF and RFC evolutionary signatures), and (2) for predicting new exons (CONGO gene finder). In this quarter, MIT extended their pipelines to use new alignments of 29 eutherian mammals, and have transitioned to a finalized alignment dataset for the the 2X mammals project. MIT have also been working closely with the HAVANA annotators to evaluate their initial runs on chr21+22, and using their feedback to finalize their RFC+CSF and CONGO pipelines in order to extend MIT results to the whole genome.

In the previous quarter, MIT encountered a bottleneck of insufficient parallel compute power to train their comparative exon predictor CONGO. MIT therefore obtained an allocation on the NSF TeraGrid supercomputing system to overcome this limitation, and have migrated their software to this system. MIT began making several modifications necessary for the new environment, such as replacing an RPC communication mechanism with MPI, and expect to have it up and running in the next quarter.

Over the course of this quarter, MIT contributed results and data to the group about long noncoding RNA transcripts, an emerging class of likely new regulators that have recently been discovered by several groups and our collaborators at the Broad Institute based on ChIP-Seq chromatin state maps. MIT believe that the annotation of such non-coding transcripts is likely to become an increasingly important part of the GENCODE project, as it becomes clear that the human genome harbours many hundreds or thousands that serve deeply conserved functions in processes such as stem cell differentiation, body plan, development, and innate immunity. To accomplish these goals, and also help with the project's existing goals of protein-coding gene annotation, MIT are working to increasingly integrate chromatin information into their prediction pipelines, as produced by several ENCODE groups.

UCSC's milestones have not been met for the running of the RetroFinder pipelines every six months or the update of ExoniPhy. They are currently working on RetroFinder pipeline to make it easier to run it so that they can accomplish their milestones, but it has been a low priority as the data is not currently required by the rest of the GENCODE group. ExoniPhy annotations have not been updated because this predicts exons based on conservation between human, mouse, rat and dog genomes. Since none of these have been updated recently, there would be no change in the output from this pipeline. Once the new human reference genome assembly is released in 2009, the ExoniPhy pipeline can be run to update the exon predictions on the human genome. Another bottleneck that UCSC have encountered is difficulty in retrieving data from the WTSI DAS server but they are working with people at WTSI to resolve the issue. However, UCSC anticipate that they are able to accomplish their milestones within their budget.

During this quarter University of Washington continued to both apply and evaluate Pairagon, their pair-HMM based aligner, for aligning cDNA sequences to the human genome. A detailed analysis and comparison of aligners revealed that Pairagon is indeed more accurate than all others tested, although the margin is small. This increased accuracy results from a more realistic scoring system that scores indels, mismatches, and introns of all types according to their true frequencies in alignments of cDNA sequences to the human genome. This results in an improved ability to align micro-exons and to determine the boundaries of exons when there are indels or mismatches near the splice sites. Pending the outcome of this study Washington University have been using Pairagon only to align cDNA sequences that "looked bad" when aligned by other programs, since Pairagon requires much more computational power than others. However, their simulations show that it is very difficult to tell the correct from the incorrect alignment, and since Pairagon is right more often, the only solution is to run it on all the cDNA sequences. Washington University plan to attempt that

in the coming quarter, although more computational resources may be required. A publication on the evaluation of Pairagon and other aligners has been prepared and we expect to submit it in the next couple of days. This will be an ENCODE publication.

Washington University also used the transcript track generated by their automatic annotation pipeline to look for unannotated splices in the RNASeq data generated by the Wold lab. Preliminary data shows that there are indeed predicted splices that are only supported by RNASeq data. Since gene predictions are of great use to connect such 'spliced reads', Washington University are currently working with UCSC and the Wold lab to devise a pipeline to incorporate the RNASeq data in the annotation. To achieve this, Washington University will need to identify the best alignment program, filter the data, and identify the most likely gene for every spliced read.

Washington University have been meeting their milestones, so far. Expenditures are on target as predicted. They feel it would be advantage to run Pairagon on all human cDNAs. In order to accomplish this they may need to request funds for additional computing equipment. The requirements of this task are still being assessed. Bottlenecks include communication of genome annotations with some members of the consortium which has required more work than anticipated, due to incompatible methods of representing them. However, this appears to be being resolved. In the first quarter CNIO group have been refining the methods in the protein isoform annotation pipeline. All the methods that make up the pipeline are now able to make predictions. However, the methods still need to be refined in order to improve the reliability of the automatic annotations. CNIO are still on course to be able to add reliable protein structural and functional annotations by early 2009. CNIO have not found any real bottlenecks and will be able to accomplish their milestones within their work budget.

CCDS Update

During the Year 2 Q1 period, 160 CCDS have been evaluated by UCSC. UCSC are on track to being able to curate the proposed CCDS updates and withdrawals and to discuss any open conflict cases before the end of February 2009 in time for the release and annotation of the new human reference genome assembly (NCBI Build 37). The set of guidelines for CCDS translation initiation site annotation established as a result of a discussion between representatives from NCBI, WTSI (Havana group) and UCSC were revised. This has improved UCSC's efficiency and reduced the number of cases that require further discussion. UCSC started to discuss additional guidelines for NMD candidate cases. The publication on the CCDS database and pipeline has been submitted to Genome Research for publication. A pipeline is in production to gather data to facilitate the curation of CCDS. After attending the ENCODE analysis workshop in December 2008, UCSC are using RNA Seq data to confirm non-consensus splice sites that they have identified in the CCDS and GENCODE annotations and are also involved in the task of using RNA Seq data to identify novel splice sites. HAVANA has also continued work on the CCDS collaboration which resulted in an update of 83 genes.

Question 1) What is your assessment of your progress relative to your milestones and to the amount of money you have spent?
Answer: As a whole, spending has been proportional to the financial requirements for the individual institutions within the group.

Manual annotation has progressed well within the anticipated milestones for this quarter in relation to the year in the number of genes annotations produced. Gene and pseudogene prediction is also progressing well within budget. The expenditure for experimental validation has been equivalent to

the results produced but progress has been slower than anticipated, see answer to question 3 below.

Question 2) Do you anticipate being able to accomplish your milestones within your budget? If not, what changes are planned?
Answer: The Gencode group's research activities came within budget for this quarter. No problems can be seen with future funding for the next quarter.

Question 3) What bottlenecks have you encountered and how are you addressing these?
Answer: The University of Lausanne were greatly delayed in their experimental validation of genes due to faults in the primers which were supplied by a commercial manufacturer. These faults were identified and new primers were provided leading to the generation of the first set of results in this quarter, steady progress is now being made.

One member of the HAVANA group left and their position has still yet to be filled. HAVANA are in the process of recruiting for this position.

MIT experienced problems with compute power for their CONGO predictor. They overcame this by obtaining an allocation on the NSF TeraGrid supercomputing system, and are currently customising their predictor to run more optimally within this new system.

UCSC have encountered difficulties in retrieving data from the WTSI DAS server but they are working with people at WTSI to resolve this issue.

Washington Univerisity found a bottleneck in the communication of different types of genome annotations due to incompatible methods of representing them. This appears to be some way to being resolved through increased consultation within the group.

No significant bottlenecks have been found for Pseudogene assignment, collaboration remains strong between Yale, UCSC and HAVANA.