Application Number: 1 U54 HG004555-01

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

Quarterly progress report summary (Q3: 1/4/2008 - 31/6/2008).

Manual Annotation

The Havana group have been focusing on completing the new annotation of chromosome 2 and have annotated 591 new loci so far (including coding genes, pseudogenes and transcripts). The WashU group highlighted 18 new Pairagon cDNA alignments of which 13 were incorporated into our annotation of chromosome 22 and 21. They also reanalysed the data from the Michele Clamp in the recent PNAS paper (PMID 18040051) which highlighted 49 erroneous Havana annotation not supported by conservation, although only 5 annotations changed as a result. Finally they incorporated the recent MGC data from Siepel et al (PMID:17989246) which meant 188 new exons from 562 were incorporated into the annotation. Finally, Havana have also been working on updating CCDS and have resolved 82 new conflict cases in collaboration with RefSeq and UCSC

Experimental Validation

CRG have developed a pipeline to automatically select annotated transcripts for experimental verification, according to the criteria established in our project, identify the most reliable exon pairs within the transcripts, and generate specific primers within the exons to initiate the RT-PCR reactions. They have selected a first set of candidates in chromosomes 21 and 22. CRG has been quite exhaustive in these two chromosomes with the goal of refining the pipeline after the evaluation of the experimental results. A summary of the cases selected appears in Table 1.

Table 1.

	Status to test	with 2 exons sup 50bp		Discard KNOWN loci
Protein coding transcripts	**801**	**627**		**36**
	NOVEL: 277	NOVEL: 234		NOVEL: 11
	PUTATIVE: 510	PUTATIVE: 383		PUTATIVE: 25
	UNKNOWN: 14	UNKNOWN: 10		UNKNOWN: 0
Coding Loci	**371**	**295**		**20**
	KNOWN: 323	KNOWN: 275		KNOWN: 0
	NOVEL: 47	**NOVEL: 20**		**NOVEL: 20**
	UNKNOWN: 1	UNKNOWN: 0		UNKNOWN: 0
	Status to test (PUTATIVE +UNKNOWN)	with 2 exons sup 50bp	Discard Coding loci	Discard KNOWN loci
Processed transcripts	**1680**	**1083**	**375**	**298**
	PUTATIVE: 436	PUTATIVE: 172	PUTATIVE: 76	PUTATIVE: 73
	UNKNOWN: 1244	UNKNOWN: 911	UNKNOWN: 299	UNKNOWN: 225
All Loci	**800 loci**	**492 loci**	**200 loci**	**162 loci**
	KNOWN: 377	KNOWN: 304	KNOWN: 38	KNOWN: 0
Non coding loci	**NOVEL: 196**	**NOVEL: 123**	**NOVEL: 115**	**NOVEL: 115**
	PUTATIVE: 208	**PUTATIVE: 47**	**PUTATIVE:47**	**PUTATIVE:47**
	UNKNOWN: 19	**UNKNOWN: 18**	**UNKNOWN: 0**	**UNKNOWN: 0**

Selected transcripts will be the target of experimental verification by RT-PCR in multiple tissues through designing of specific pairs of primers mapping to two different exons. PCRs will be performed in 384 well plates on a TECAN Freedom EVO 200 robotic platform coupled with a BioRad DNA Engine Tetrad workstation allowing to reach the necessary high-throughput. Reactions will be verified on gels and sequenced. A subset of verified transcripts will be selected for further characterization by RACE.

Lausanne has begun to purchase the necessary RNAs and began to produce the cDNAs required throughout the extend of this project, as well as benchmarked the newly implemented scripts to be used by the robotic workstation for the RT-PCR, gel loading and

RACE procedures. In parallel, it established in collaboration with the CRG Barcelona group the data-tracking system to allow samples to be followed through the multiple procedures required. With the CRG group multiple strategies are being tested to implement an efficient predicted transcripts selection procedure.

Pseudogene Assignment

In addition to the quarterly pipeline run, Yale has been working on developing a consensus set of predicted pseudogenes. They have been discussing strategies for integration with UCSC and Sanger and have examined some preliminary results. Some focus was also spent on examining unitary pseudogenes (pseudogenes whose parent protein is in a separate species), furthering work in the paper, "Analysis of nuclear receptor pseudogenes: how the silent tell their stories." Refinement of PseudoPipe also continued. Yale are modifying the pipeline to accept the latest Sanger annotations through DAS as input instead of using Ensembl releases.

UCSC have added annotation tracks for Havana pseudogenes (May 2008) and Yale pseudogenes (based on Ensembl Build 49) on a virtual host of the UCSC Genome Browser at http://hgwdev-gencode.cse.ucsc.edu. As a first step towards creating a consensus set of processed pseudogenes, an analysis was performed to compare the processed pseudogenes from UCSC (RetroFinder), Havana (manual annotation at WTSI) and Yale (PseudoPipe). The overlapping sets may be viewed on the hgwdev-gencode Genome Browser as well as being available as a DAS server. For this analysis, they have used UCSC processed pseudogenes (retrogenes) with a score of at least 700. Overlapping pseudogenes were clustered; overlap was defined to be at least 50% for pairwise comparisons of pseudogenes. Of 22,616 clusters defining loci, Havana and UCSC pseudogenes shared 872 loci with 1762 overlapping pseudogenes; Havana and Yale shared 119 loci with 241 overlapping pseudogenes; UCSC and Yale shared 4899 loci with 9825 overlapping pseudogenes while all three sets shared 1754 loci with 5279 overlapping transcripts. The majority of loci consist of only those processed pseudogenes from UCSC (39%) or only from Yale processed pseudogenes only (26%). Havana have annotated a fraction of the pseudogene loci so only about 1% of loci contained only Havana pseudogenes.

Gene Prediction

CRG has been streamlining their pipeline to identify U12 introns, and to characterize non-coding RNAs.

WashU have continued to make progress on improving alignment accuracy and communications infrastructure within the GENCODE subgroup. They have retrained Pairagon on a set of alignments at various percents identity as follows. WashU started with high-confidence annotations, created artificial-cDNAs that are 100% identical to the reference genome, mutated the genome using a mutation simulator, and trained alignments to the mutated genome. The result was a new set of parameters that produces more accurate alignments than any other cDNA aligner below 95% identity. The original parameters still perform slightly better at high identity levels, but this new training approach may improve performance at the high end, too. By combining the two parameter sets, they expect to have improved on the previous state-of-the art in cDNA alignment. A publication on this system is in preparation. They are also working with Felix Kokocinski at the Sanger to

identify limitations in the current DAS specification and to specify it more fully. As mentioned in the 'Manual Annotation' section, WashU alignments on chr21 and chr22 have been used to refine the annotation.

UCSC have been working on the DAS server for sharing data with the rest of the Gencode project participants. The DAS server at UCSC now serves up annotations at a speed of about 30 times faster than before the upgrade. The TransMap pipeline was also run to generate cross-species alignments between the human genome and UCSC Genes, RefSeq Genes and GenBank mRNAs and spliced ESTs from other vertebrate species

The MIT group is using evolutionary signatures of protein-coding selection in the many available mammalian genomes to inform the human gene annotation, and in particular to support the efforts of the Havana curators in identifying new genes and exons, and in revising existing gene annotations. This quarter they have applied these pipelines in pilot analyses and produced preliminary datasets for two analyses. Firstly, they performed an initial genome-wide run of our comparative protein-coding exon predictor, CONGO. With very stringent thresholds, this resulted in the recovery of 80% of known exons, and the prediction of ~10,000 new exons, which we believe are of high quality (other available comparative de novo gene prediction sets make 3-4 times as many new predictions at the same recovery rates). Secondly, they made a pilot run of their gene evaluation pipeline on chromosome 21 and 22 gene annotations, using the RFC and CSF evolutionary signatures we have previously published. As expected, our tests strongly confirmed the vast majority of existing annotations, as demonstrating evolutionary signatures of protein-coding conservation. However, a small fraction of annotations (~5%) were indistinguishable from random non-coding regions, suggesting their merit revisiting. These may represent spurious annotations or non-coding RNA genes without a protein product. In both cases, they made their prediction datasets available to the HAVANA curators and the rest of the GENCODE team through DAS sources, to help guide the curation and experimental validation pipelines.

CCDS update

For CCDS, 55 have been reviewed by UCSC during this quarter. UCSC took part in the QA of the latest human CCDS build which was released in May 2008 with the re-annotation of the human genome (NCBI Build 36.3). This added 2,151 new CCDS IDs and 1,249 Genes into the human CCDS set (total: 20,159 CCDS IDs, 17,052 GeneIDs). UCSC have set up automatic updates of their local CCDS database and loading of the data into a UCSC Genome Browser track. This facilitates keeping in synchrony with NCBI's CCDS database. Also, a tool is being developed to aid the CCDS curation process; this extends transcripts using clustered ESTs and searches for upstream translation initiation sites in order to define an ORF. For each ORF, the encoded protein will be analysed for various features. Tepresentatives from UCSC, Havana and Ensembl met with the RefSeq group at NCBI in June 2008 to discuss the CCDS project and its future direction as well as working on a publication for CCDS.