Application Number: 1 U54 HG004555-02

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

Deliverables and milestones.

This document was requested following the Notice of Award for Grant Number: 5U54HG004555-02, and provides proposed deliverables and milestones. A list of institute abbreviations used is shown in **Table 1**.

Table 2 lists milestones in terms of annotated genes and fraction of genome annotated in terms of the overall criteria of the project.
(1) indicates gene annotations produced per year by WTSI for chromosomes not previously annotated by them (D.2). Each gene and transcript will be labeled according to the gene categories given in C.1 (Known, Novel_CDS, Novel_transcript, Putative, Pseudogene, TEC) and the transcript categories given in D.2, Figure 7. Genes categorized as putative and novel transcripts will be added to the buffer of targets for experimental validation (2).
(2) indicates the buffer of genes available for experimental validation. It will be initially high since it can be populated from genes already annotated by Havana to the same standards, but outside this project. Genes in this list will be prioritized through ranking from computational pipelines (D4-D7). Computational pipelines will also add some novel genes to this list.
(3) indicates genes verified experimentally per year by Lausanne/CRG (D.3) taken from the experimental validation buffer (2). Raw sequence information generated during validation will be submitted to public databases as well as being fed into the overall evaluation system.
(4) The pseudogene pipeline (D.4) will reconcile the output from the two computational pipelines (D.4a and D.4b) progressively with the set of annotated genes labeled as pseudogenes. The fraction of genome completely evaluated in this way will increase in a different way to (5) since more has already been annotated manually, in most cases it will not depend on the fraction checked by experimental validation and it is anticipated that progressively more work will be needed to reach consensus on difficult cases. The pseudogene pipelines will produce "draft" pseudogenes, which will cover the whole genome. These will be computationally checked and manually annotated to produce finished "quality" pseudogenes. This will be labor intensive. We anticipate that manual annotation and checking will increase the accuracy of the pipelines so that our output of finished pseudogenes will increase more quickly in later years. The cuts in the budget will decrease the frequency with which the pseudogene pipelines can be run and can be compared. It will also decrease the frequency with which it assessed manually. This will decrease the total number of manually annotated pseudogenes and will decrease the quality of the finished "quality" pseudogenes.
(5) While the manual (D.2) and experimental (D.3) pipelines work to progressively generate and evaluate individual annotations, a series of computational pipelines (D4-7) will be run repeatedly over the entire genome. The pipelines will be run at least every 3 months in order to incorporate new external data and to allow assessment of annotation generated by the progressive pipelines. Each pipeline will generate priority lists of existing annotation that needs to be re-checked by curators and new annotation that should be evaluated either manually or experimentally. The fraction of genome annotated indicates the fraction for which computational, manual and (where appropriate) experimental investigation has been carried out and is considered consistent. Every 3 months a complete snapshot of the current consensus annotation in the system will be made and submitted to the DCC. Annotation will be labeled at both gene and transcript level to indicate how it has been classified (as described above in (1)) and the level of completeness of the evaluations. **Tables 3 and 4 list** the Pipeline Deployment and Individual Consortium Member Institution Milestones respectively.

Table 1: Institute Name	Abbreviation
Wellcome Trust Sanger Institute, Hinxton, UK.	WTSI
University of Lausanne, Lausanne, Switzerland.	Lausanne
Centre de Regulacio Genomica, Barcelona, Spain.	CRG
University of California Santa Cruz, Santa Cruz, USA.	UCSC
Washington University, St. Louis, USA.	WashU
Massachusetts Institute of Technology, Boston, USA.	MIT
Yale University, New Haven, USA.	Yale
Spanish National Cancer Research Centre, Madrid, Spain.	CNIO

Table 2: Production milestones (1 – 4)	Year 0	Year 1	Year 2	Year 3	Year 4
(1) New Manual Annotation (D.2)		2250	4500	4500	4500
(2) Buffer of targets for experimental validation	4500	3800	2850	1425	0
(3) Experimental validation (D.3)		0	4000	3000	3000
(4) Cumulative fraction of genome evaluated for "quality" pseudogenes		20%	35%	55%	90%
(5) Cumulative fraction of genome evaluated (D2, 3, 4, 5, 6, 7)		13%	36%	63%	90%

The results of our experimental verification are sequences obtained after RT-PCR verification. The sequences are submitted to GenBank as ESTs, and retrofitted to the GENCODE annotation. RT-PCR results are not considered by themselves sufficient evidence without the corresponding sequence, and no plans have been made to submit them to the DCC. However, this should probably be considered and discussed. Similarly, we could consider submitting the RT-PCR sequences. These were not submitted during the pilot.

All milestones were projected to ramp over the first year to a sustained level of output for the remaining 3 years. Most of the ramping in year 1 involved recruitment, training and setup of infrastructure (*i.e.* pipeline setup and integration with DAS servers) in the first 9 months as outlined in Table 2.

Table 3: Pipeline deployment milestones			
Component	0-9 Months	9-15 months	15 Months-4 Years
Manual annotation pipeline (D.2)	Annotator recruitment and Training Setup of otter DAS, Configuration of otterlace client to view DAS sources	Tracking system setup finalizing and data loading from manual annotation into tracking system. Integration of data from external DAS sources into tracking system. Checking for inconsistencies between tracking systems. New annotation and re-evaluations of previous annotation.	New manual annotation to meet milestones in Table 2 (1) Revised annotation to meet milestones in Table 2 (4)
Experimental validation pipeline (D.3)	Recruitment; setup of DAS servers for aligned sequence products. Configuration of RTdb and integration to output status properties via DAS	Setup of pipeline for selection of transcripts to be experimentally verified in collaboration with CRG.Results generated from first batch of experimental validations.	Experimental validation to meet milestones in Table 2 (3) and (4).
Pseudogene assignment pipeline (D.4) De novo gene prediction pipelines (D.5a5, D.5b2, D5c) Evaluation of existing annotation (D5a6, D5b1) CDS analysis pipeline (D.6) Annotation Quality Control and analysis (D.7)	Recruitment; setup of DAS servers for coordinate based annotation with integration of generated consensus status; setup to collect input annotation from external DAS servers	Evaluation of pilot runs on chr21+chr22 by HAVANA curators, and incorporation of their feedback and suggestions to ensure high quality and usefulness of the production pipelines. Updated pipeline runs at least once every 3 months to make use of new data and take into account additional annotation (D.2) to meet milestones in Table 2 (4), see month 15 – 4 years milestones. Comparison of different pseudogene pipeline results, combining and annotating them into a single consensus data source. Flagging of problematic pseudogene cases.	Updated pipeline runs at least once every 3 months to make use of new data and take into account additional annotation (D.2) to meet milestones in Table 2 (4)

Table 4. Milestones for each Consortium Institution	
Consortium Institution	**Project Component Milestones**
WTSI	Havana annotation : Month 6 :recruitment and new annotation of 800 loci month 9: design database structure and set up tracking system load data from manual annotation into tracking system integrate data from external DAS sources into tracking system add and refine data sources to tracking system, identify and resolve conflicts in the data run tracking scripts nightly to check for inconsistency between data sources Month 12: 1450 new annotation re-evaluations 400 add further data sources to tracking system as required Month 18: 2250 new annotation re-evaluations 800 Month 24: 2250 new annotation re-evaluations 1000 add and refine data sources in tracking system, identify and resolve conflicts in the data run tracking scripts nightly to check for inconsistency between data sources Month 30: 2250;re-evaluation of 1000 add data dumping functionality and other required analysis functions to tracking system Month 36: 2250;re-evaluation of 1000 Month 42: 2250;re-eveluation of 1000 Month 48: 2250;re-evaluation of 1000 Ensembl: Run CCDS automatic pipeline every 6 months Run Ensembl-Havana gene merge every 4 months Run cDNA/exon/intron evaluation pipeline to highlight differences between Havana and Ensembl transcripts every 3 months
Lausanne	Month 9: - Setup of pipeline for selection of transcripts to be experimentally verified in collaboration with CRG. Month 12: - Experimental validation of first 1'500 exon-junction of predicted transcript models. Month 15: - Assessment of selection process (and its eventual modification) following first year experimental success rate in collaboration with CRG. Month 18: - Determination of transcript sequence of selected set of experimentally verified first 1'500 exon-junction. Month 24: - Experimental validation first 4'000 exon-junction of predicted transcript models. Month 30: - Determination of transcript sequence of selected set of experimentally verified first 4'000 exon-junction. Month 36: - Experimental validation first 7'000 exon-junction of predicted transcript models. Month 42: - Determination of transcript sequence of selected set of experimentally verified first 7'000 exon-junction. Month 45 - Experimental validation first 10'000 exon-junction of predicted transcript models. Month 48: - Determination of transcript sequence of selected set of experimentally verified first 10'000 exon-junction.
CRG	Month 9: - Setup of pipeline for selection of transcripts to be experimentally verified and primer design in collaboration with Lausanne and University of Washington. Month 12: - Submission to Lausanne for experimental validation of the first 1'500 exon-junction of predicted transcript models. Evaluation of the U12 intron, non-canoncial splicing and selenoprotein annotation of the ongoing gencode annotation. Month 15: - Assessment of selection process (and its eventual modification) following first year experimental success rate in collaboration with Lausanne.

Table 4. Milestones for each Consortium Institution	
	Month 18: - Analysis of the transcript sequences obtained by Lausanne from selected set of experimentally verified first 1'500 exon-junction. Month 24: - Submission to Lausanne for experimental validation of the first 4'000 exon-junction of predicted transcript models. Specific analysis of the patterns of gene expression of the verified non-coding RNAs and their comparison with the protein coding RNAs. Month 30: - Analysis of the transcript sequences obtained by Lausanne from the selected set of experimentally verified first 4'000 exon-junction. Month 36: - Submission to Lausanne for experimental validation of the first 7'000 exon-junction of predicted transcript models. Month 42: - Analysis of the transcript sequences obtained by Lausanne from the selected set of experimentally verified first 7'000 exon-junction. Month 45: - Submission to Lausanne for experimental validation of the first 10'000 exon-junction of predicted transcript models. Month 48: - Analysis of the transcript sequences obtained by Lausanne from the selected set of experimentally verified first 10'000 exon-junction. Evaluation of the U12 intron, non-canoncial splicing and selenoprotein annotation of the ongoing gencode annotation.
UCSC	Month 9: Setup system to allow GENCODE pipeline UCSC browser tracks to also be accessible as DAS sources. Month 9-48: • Run the retroFinder pipeline every 6 months to generate processed pseudogene predictions which will be consolidated with the Yale pseudogene set. • Quality controls will be applied to CCDS during build updates, and CCDS will be curated as required when flagged for update and withdrawal. The CCDS set will be further evaluated to identify problem CCDS that require further biological data as evidence so that these can be targeted by the experimental group. Additionally, CCDS will be expanded to cover more loci. • Re-run TransMap pipeline every 6 months for UCSC genes, RefSeqs, mRNAs and ESTs
WashU	Month 18: - Realign all human cDNAs and ESTs that do not concur according to the PASA pipeline with Pariagon Month 36: - List aligned mRNAs and cDNAs which are out of alignment with the annotation Month 18-48:- Work with annotators in selecting pertinent genomic regions
MIT	Month 9: - Develop comparative features for new exon predictions. Scale up probabilistic framework for exon identification. Month 12: - Initial predictions on chromosomes 21 and 22, ENCODE pilot. Initial evaluation of Havana exons on chromosomes 21 and 22, pilot. Month 18:- Integrate low-coverage mammalian genomes for predictions. Integrate additional features for splice sites and single-species. Revised predictions for Chr 21, Chr 22, ENCODE pilot regions. New evaluation for HAVANA exons and CCDS. Month 24: - Incorporation of experimental results for feature selection and method refinement. Second generation prediction pipeline. Revised prediction set for chromosomes and regions selected by GENCODE for validation pipeline. Prediction of di-cistronic genes, stop-codon read-through, programmed frameshifts, and other unusual gene structures. Month 36: - Incorporation of additional mammalian and vertebrate genomes in the prediction pipeline. Continued integration with experimental results and refinement of exon predictions, and evaluation of exons lacking experimental support. Genome-wide predictions of new exons and genome-wide exon evaluations. Month 42: - Final set of revised genome-wide predictions and evaluations. Evaluation and refinement of all genome-wide prediction sets from all groups towards a unified annotation. Month 48: - Final evaluation of genome-wide gene set.
Yale	Month 9: - Setup systems to put results of pipeline runs into various useful formats (e.g. onto the web, into a DAS server, and onto a UCSC track). Month 9-48:

Table 4. Milestones for each Consortium Institution	
	• Run the Yale pseudogene pipeline with updates to improve its accuracy and write efficiency every 6 months. • Ongoing comparisons of the results of pipeline runs with those of UCSC and Sanger. Have conference calls and systematic discussions with Sanger and UCSC on this comparison and use these to divide the results of a pipeline run into sure ("easy") cases and more problematic ones. • Use these problematic cases to develop better approaches for finding them and integrating these approaches into the pipeline. This may include developing custom pipelines for particular types of pseudogenes and specific pseudogene families.
CNIO	Performing a collaborative role within the project consortium, providing annotation from protein analysis pipelines supported externally.

December 2008

Dr Tim Hubbard
(Principle Investigator)
Head of Informatics
Wellcome Trust Sanger Institute

Mr David Davison
(Applicant Organisation Official)
Director of Corporate Services
Wellcome Trust Sanger Institute