Searle *et al. Genome Biology* 2010, **11**(Suppl 1):P36
http://genomebiology.com/2010/11/S1/P36

Genome **Biology**

POSTER PRESENTATION **Open Access**

The GENCODE human gene set

S Searle[1*], A Frankish[1], A Bignell[1], B Aken[1], T Derrien[5], M Diekhans[7], R Harte[7], C Howald, F Kokocinski[1], M Lin[3], M Tress[2], M Van Baren[4], I Barnes[1], T Hunt[1], D Carvalho-Silva[1], C Davidson[1], S Donaldson[1], J Gilbert[1], M Kay[1], D Lloyd[1], J Loveland[1], J Mudge[1], C Snow[1], J Vamathevan[1], L Wilming[1], M Brent[4], M Gerstein[6], R Guigó[5], M Kellis[3], A Reymond[8], A Zadissa[1], A Valencia[2], J Harrow[1], T Hubbard[1]

From Beyond the Genome: The true gene count, human evolution and disease genomics
Boston, MA, USA. 11-13 October 2010

The GENCODE consortium is a sub group of the ENCODE consortium. Its aim is to provide complete annotation of genes in the human genome including protein-coding loci, non-coding loci and pseudogenes, based on experimental evidence. The final aim is for the HAVANA team to manually annotate the complete genome. This is a time-consuming process which will be completed over the course of the ENCODE project. Currently, to provide a set of annotation covering the complete genome, rather than just the regions that have been manually annotated, a merge of manual annotation from HAVANA with automatic annotation from the Ensembl automatically annotated gene set is created. This process also adds unique full-length CDS predictions from the Ensembl protein coding set into manually annotated genes, to provide the most complete up to date annotation of the genome possible. Also included in the set are short and long ncRNA genes predicted by the Ensembl prediction pipelines and a consensus set of pseudogene predictions agreed between Havana, Yale and UCSC. The CCDS set is also fully represented within the GENCODE set. The GENCODE set is the default annotation available in Ensembl and is also available in the UCSC genome browser. All the annotation is tagged as to whether it is produced by manual annotation alone, automatic annotation alone, or by both approaches. We are currently working to provide confidence levels for annotation, based on depth and type of evidence supporting it.

There are several analysis groups in the GENCODE consortium that run pipelines that aid the manual annotators in producing models in unannotated regions and to identify potential missed or incorrect manual annotation, including completely missing loci, missing alternative isoforms, incorrect splice sites and incorrect biotypes. These are fed back to the manual annotators using a tracking system. Some of these pipelines use data from other ENCODE subgroups including RNASeq data, histone modification and CAGE and Ditag data. RNAseq data is an important new source of evidence, but generating complete gene models from it is a difficult problem. As part of GENCODE, a competiton was run to assess the quality of predictions produced by various RNAseq prediction pipelines. To confirm uncertain models, GENCODE also has an experimental validation pipeline using RNA sequencing and RACE.

Author details
[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [2]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [3]MIT Computer Science and AI Laboratory, Broad Institute, Cambridge, MA, USA. [4]Lab. for Comp. Genomics and Dept. of CS, Washington Univ, St. Louis, Missouri, USA. [5]Centre for Genomic Regulation, Barcelona, Catalonia, Spain. [6]Department of Molecular Biophys. and Biochem. Yale University New Haven, CT USA. [7]Center for Biomolecular Science and Engineering, UCSC, CA, USA. [8]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland.

Published: 11 October 2010

doi:10.1186/gb-2010-11-S1-P36
Cite this article as: Searle *et al.*: The GENCODE human gene set. *Genome Biology* 2010 **11**(Suppl 1):P36.

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
Full list of author information is available at the end of the article