GENCODE: Creating a Validated Manually Annotated Geneset for the Whole Human Genome

A. Bignell[1], A. Frankish[1], B. Aken[1], M. Diekhans[7], F. Kokocinski[1], M. Lin[3], M. Tress[2] , J. Van Baren[4], I. Barnes[1], T. Hunt[1], D. Carvalho-Silva[1], C. Davidson[1], S. Donaldson[1], J. Gilbert[1], E. Hart[1], M. Kay[1], R. Kinsella[1], D. Lloyd[1], J E. Loveland[1], J. Mudge[1], C. Snow[1], J. Vamathevan[1], L. Wilming[1], M. Brent[4], M. Gerstein[6], R. Guigo[5], R. Harte[7], M. Kellis[3], S. Searle[1], J. Harrow[1] and T. Hubbard[1].

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Spanish National Cancer Research Centre, Madrid, Spain. [3]MIT Computer Science and Artificial Intelligence Laboratory, Broad Institute of MIT and Harvard, Cambridge, USA. [4]Laboratory for Computational Genomics and Department of Computer Science, Washington University, St. Louis, USA. [5]Centre for Genomic Regulation, Barcelona, Catalonia, Spain. [6]Department of Molecular Biophysics and Biochemistry, Yale University New Haven, USA. [7]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, USA.

HAVANA

The Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute produces high quality manual annotation of protein-coding, non-coding and pseudogene loci. All HAVANA annotation is supported by transcript (EST, mRNA) and/or protein evidence and provides unparalleled coverage of alternative splicing, untranslated regions, pseudogenes and poly-adenylation features.

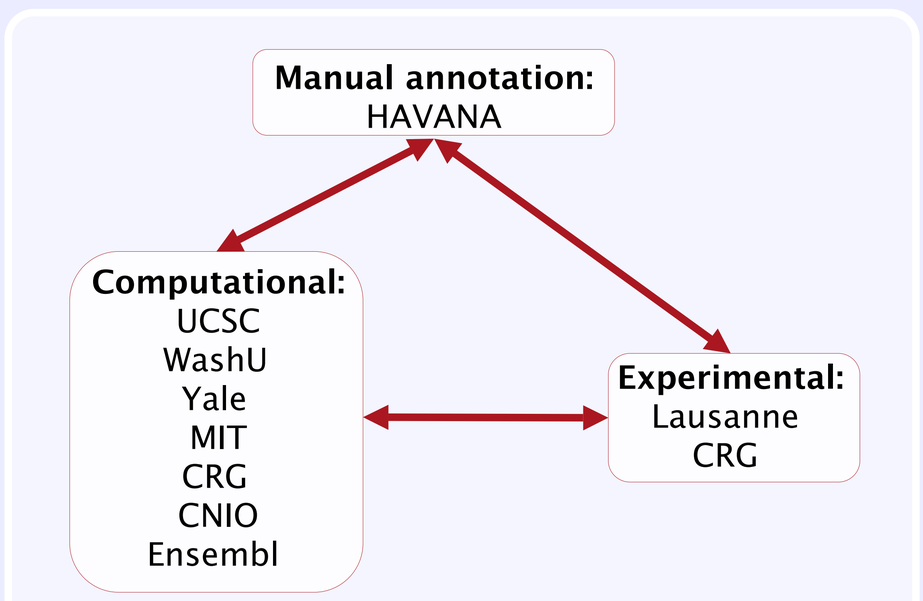
All HAVANA annotation can be viewed on our Vertebrate Genome Annotation browser (VEGA):http://vega.sanger.ac.uk

ENCODE

The aim of the ENCODE (**Enc**yclopedia **o**f **D**NA **E**lements) project is to identify all functional elements in the human genome sequence. During the pilot phase investigating 1% of the human genome HAVANA produced a manually annotated geneset that was validated computationally and experimentally by our collaborators in the GENCODE subgroup.

Following the success of the ENCODE pilot project, GENCODE are reprising their previous role and providing high quality gene annotation for the whole human genome. This geneset will be used in the analyses of all the other members of the ENCODE consortium.

GENCODE

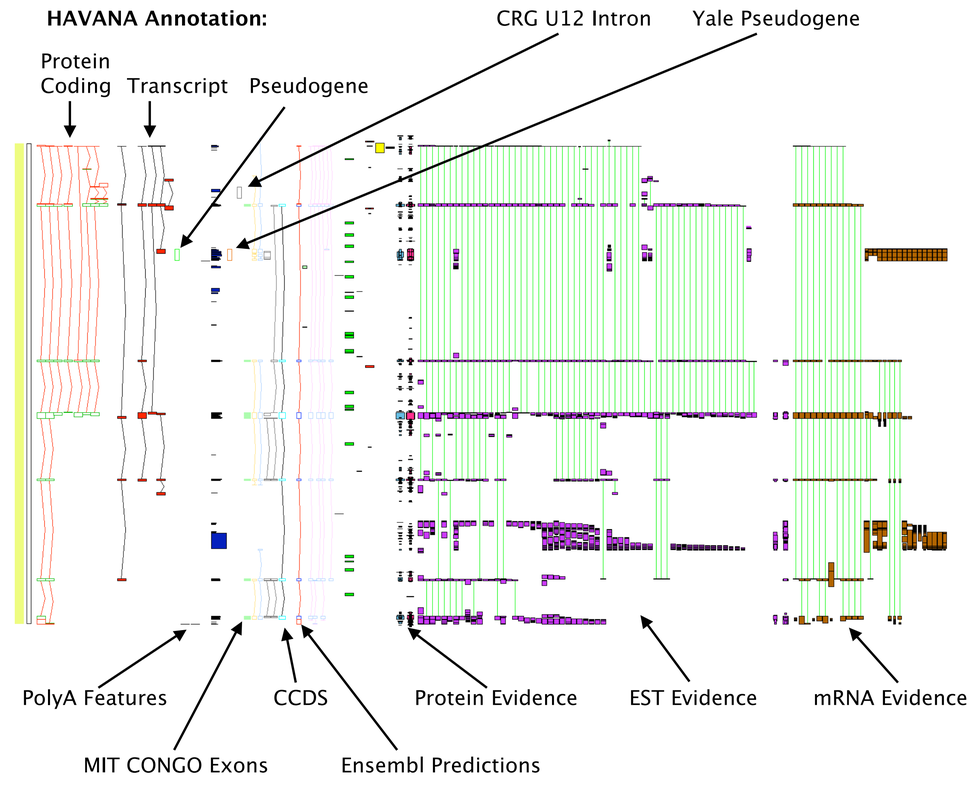


Computational predictions are produced independently of manual annotation and used as both a guide for new annotation and for validation of completed annotation. Potential novel loci and variants are identified by state-of-the-art algorithms for finding exons, splice junctions, transcripts and pseudogenes. The coding potential of all annotated CDSs is assessed by investigating sequence conservation and comparing predicted secondary structures to similar proteins with solved structures.

Although initial experimental validation of transcripts was based on RT-PCR and extension by 5' and 3' RACE, short-read sequences (RNA_Seq) have recently been added to validate annotated splice junctions. RNA_Seq data will also allow the identification of novel transcripts and provide information on tissue specificity of all annotated transcripts. Where novel features are confirmed the annotation is updated.

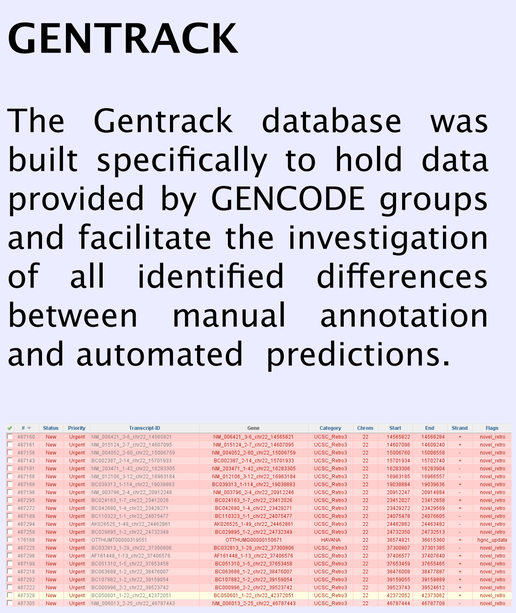
Annotation

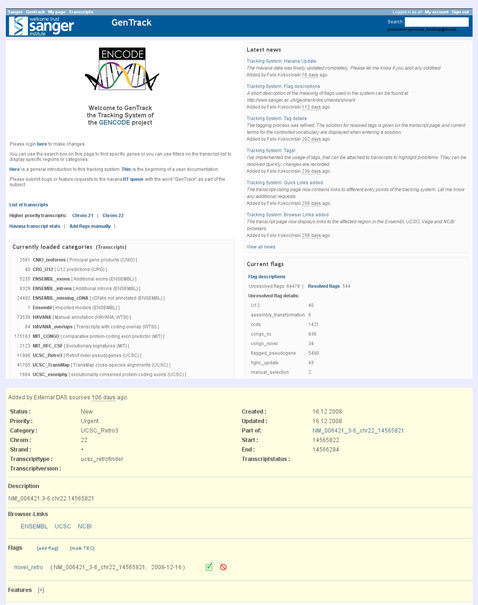
Data from all members of GENCODE are distributed via DAS (Distributed Annotation System) and are now visible in our Zmap annotation interface. For example, the KLHL22 locus shown below contains a U12 intron prediction, CONGO coding exon predictions and an intronic Yale pseudogene prediction.



GENTRACK

The Gentrack database was built specifically to hold data provided by GENCODE groups and facilitate the investigation of all identified differences between manual annotation and automated predictions.





Validation

Computational validation of the manual annotation of chromosomes 21 and 22 demonstrated that while HAVANA annotation is both comprehensive and robust it has been enriched by comparison with good computational predictions.

	Chromosome 21		Chromosome 22	
	Loci	Loci and variants added	Loci	Loci and variants added
Total Genes	582		1,816	
Protein Coding	226	13*	826	6*
Processed Transcripts	223	8*	400	5*
Pseudogenes	129	8	542	18
IG Genes	0		92	