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

<u>A. Bignell</u>¹, A. Frankish¹, B. Aken¹, M. Diekhans⁷, F. Kokocinski¹, M. Lin³, M. Tress², J. Van Baren⁴, I. Barnes¹, T. Hunt¹, D. Carvalho-Silva¹, C. Davidson¹, S. Donaldson¹, J. Gilbert¹, E. Hart¹, M. Kay¹, R. Kinsella¹, D. Lloyd¹, J E. Loveland¹, J. Mudge¹, C. Snow¹, J. Vamathevan¹, L. Wilming¹, M. Brent⁴, M. Gerstein⁶, R. Guigo⁵, R. Harte⁷, M. Kellis³, S. Searle¹, J. Harrow¹ and T. Hubbard¹.

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







HAVANA

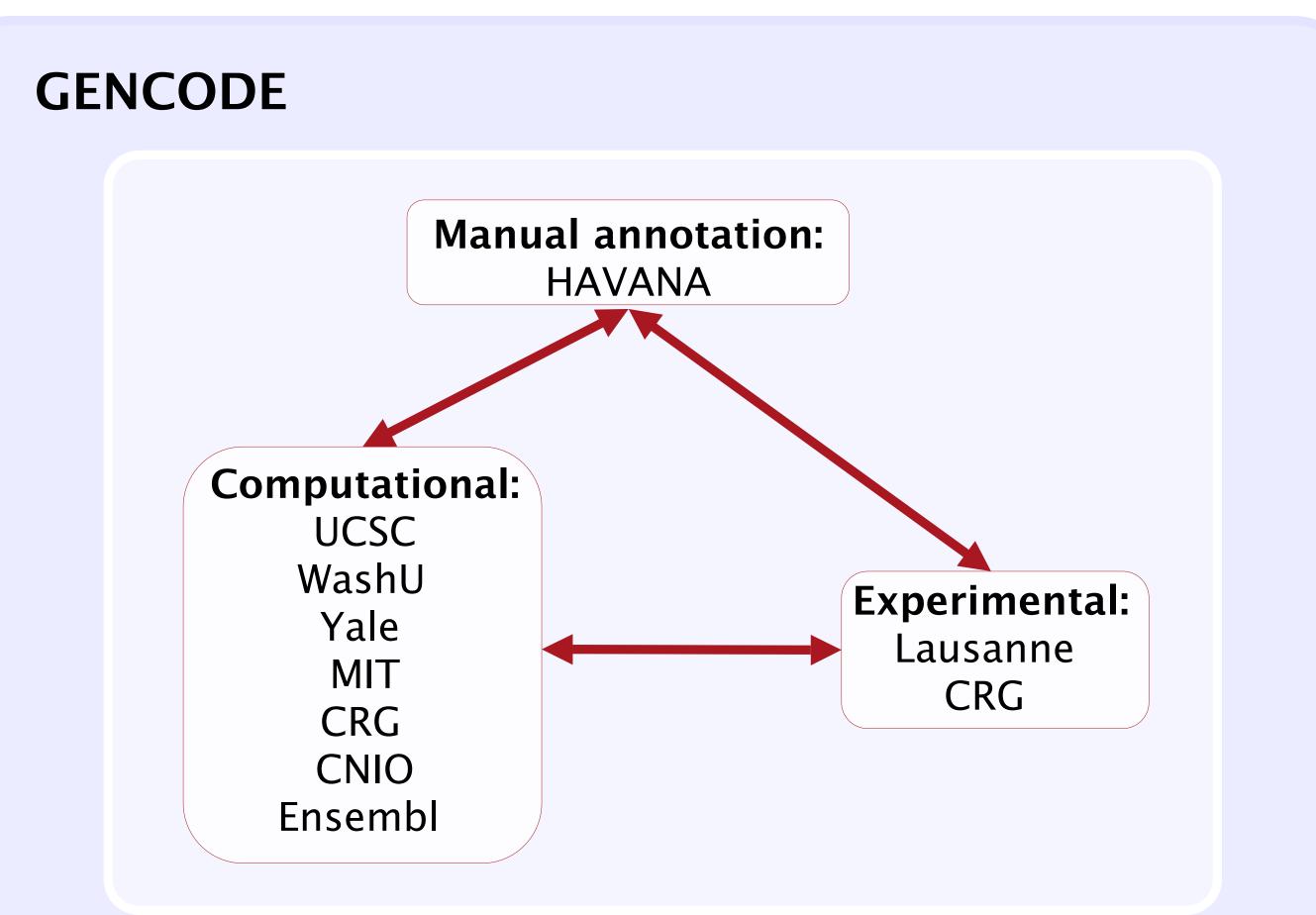
The Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute produces high

ENCODE

The aim of the ENCODE (Encyclopedia of DNA Elements) project is to identify all functional elements in the human genome

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



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.

Annotation

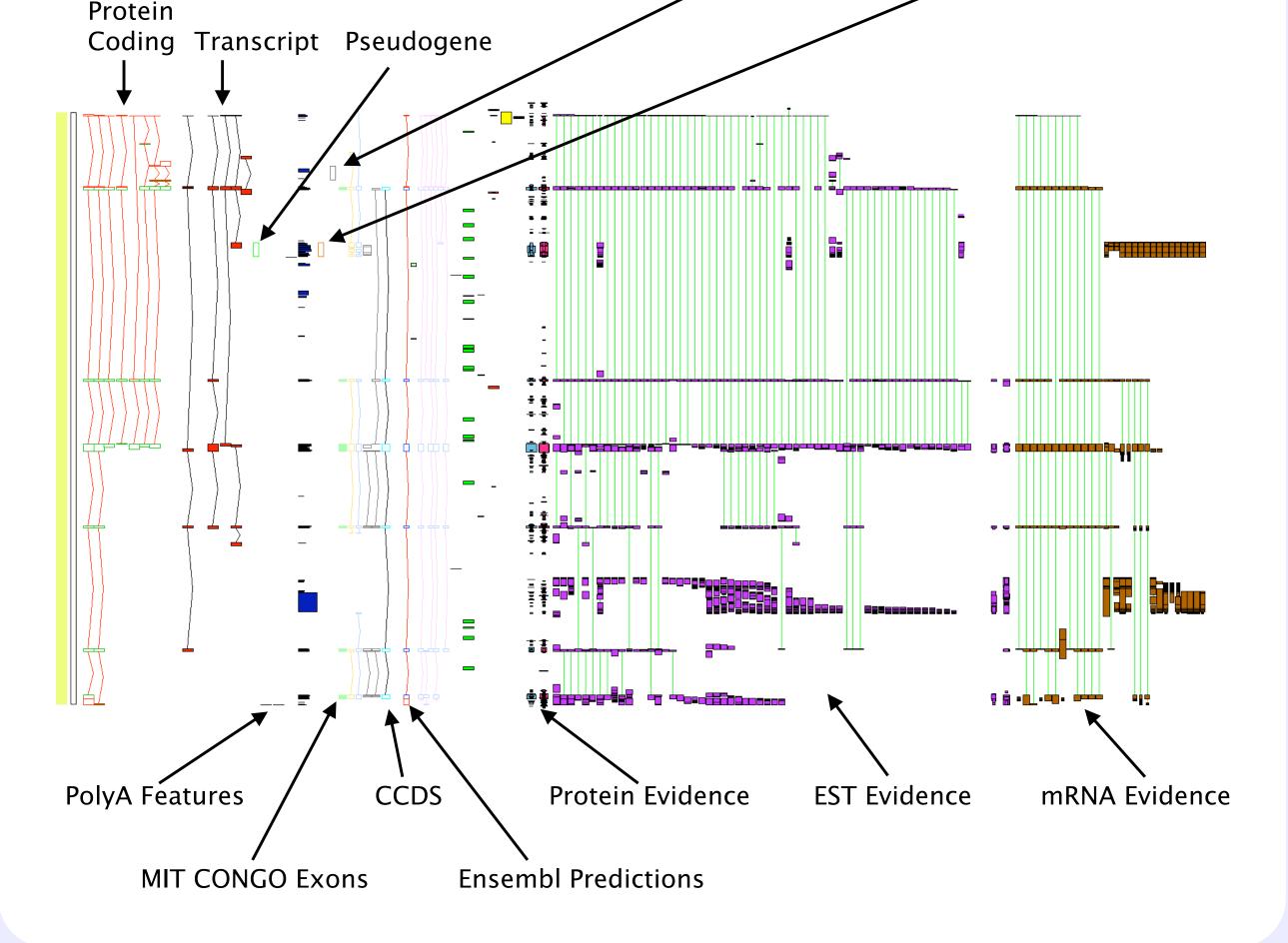
Data from all members of GENCODE are distributed via DAS (Distributed Annotation System) and are now visible in our Zmap annotation interface. For example, the KLHL22 locus shown below contains a U12 intron prediction, CONGO coding exon predictions and an intronic Yale pseudogene prediction.

HAVANA Annotation:

CRG U12 Intron Yale Pseudogene

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.



Validation

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.

~	# -	Status	Priority	Transcript-ID	Gene	Category	Chrom	Start	End	Strand	Flags
	487160	New	Urgent	NM_006421_3-6_chr22_14565821	NM_006421_3-6_chr22_14565821	UCSC_Retro3	22	14565822	14566284	+	novel_retro
Γ	487161	New	Urgent	NM_015124_2-7_chr22_14607095	NM_015124_2-7_chr22_14607095	UCSC_Retro3	22	14607096	14609240	+	novel_retro
Г	487156	New	Urgent	NM_004052_2-60_chr22_15006759	NM_004052_2-60_chr22_15006759	UCSC_Retro3	22	15006760	15008558		novel_retro
Γ	487143	New	Urgent	BC002387_2-14_chr22_15701933	BC002387_2-14_chr22_15701933	UCSC_Retro3	22	15701934	15702740	+	novel_retro
Г	487181	New	Urgent	NM_203471_1-42_chr22_16283305	NM_203471_1-42_chr22_16283305	UCSC_Retro3	22	16283306	16283904	-	novel_retro
Γ	487166	New	Urgent	NM_012106_3-12_chr22_16963184	NM_012106_3-12_chr22_16963184	UCSC_Retro3	22	16963185	16966557		novel_retro
	487168	New	Urgent	BC039313_1-114_chr22_19038883	BC039313_1-114_chr22_19038883	UCSC_Retro3	22	19038884	19039636	+	novel_retro
Γ	487196	New	Urgent	NM_003796_2-4_chr22_20912246	NM_003796_2-4_chr22_20912246	UCSC_Retro3	22	20912247	20914884		novel_retro
	487295	New	Urgent	BC024163_1-7_chr22_23412026	BC024163_1-7_chr22_23412026	UCSC_Retro3	22	23412027	23412658	+	novel_retro
Г	487272	New	Urgent	BC042680_1-4_chr22_23429271	BC042680_1-4_chr22_23429271	UCSC_Retro3	22	23429272	23429569	+	novel_retro
Γ	487189	New	Urgent	BC110323_1-1_chr22_24075477	BC110323_1-1_chr22_24075477	UCSC_Retro3	22	24075478	24076605	-	novel_retro
Г	487294	New	Urgent	AK026525_1-49_chr22_24462861	AK026525_1-49_chr22_24462861	UCSC_Retro3	22	24462862	24463482	-	novel_retro
Г	487259	New	Urgent	BC029895_1-2_chr22_24732349	BC029895_1-2_chr22_24732349	UCSC_Retro3	22	24732350	24732513	-	novel_retro
	176168	New	Urgent	OTTHUMT00000319551	OTTHUMG00000150671	HAVANA	22	36574821	36615360	+	hgnc_update
Γ	487225	New	Urgent	BC032813_1-29_chr22_37300806	BC032813_1-29_chr22_37300806	UCSC_Retro3	22	37300807	37301385		novel_retro
Γ	487296	New	Urgent	AF161448_1-13_chr22_37406576	AF161448_1-13_chr22_37406576	UCSC_Retro3	22	37406577	37407640	+	novel_retro
	487199	New	Urgent	BC051310_1-5_chr22_37653458	BC051310_1-5_chr22_37653458	UCSC_Retro3	22	37653459	37655465	+	novel_retro
Γ	487218	New	Urgent	BC063688_1-2_chr22_38476007	BC063688_1-2_chr22_38476007	UCSC_Retro3	22	38476008	38477887	+	novel_retro
	487202	New	Urgent	BC107882_1-2_chr22_39159054	BC107882_1-2_chr22_39159054	UCSC_Retro3	22	39159055	39159889	+	novel_retro
Γ	487222	New	Urgent	BC000996_2-2_chr22_39523742	BC000996_2-2_chr22_39523742	UCSC_Retro3	22	39523743	39524612	+	novel_retro
Г	487328	New	Urgent	BC050601_1-22_chr22_42372051	BC050601_1-22_chr22_42372051	UCSC_Retro3	22	42372052	42373062	+	novel_retro
Г	487305	New	Urgent	NM_006013_2-25_chr22_46787443	NM_006013_2-25_chr22_46787443	UCSC_Retro3	22	46787444	46787708	-	novel_retro

	f GenTrack		Search: production:gencode_tracking@mcs4a			
	510005	Latest news				
		Tracking System: Havana Update The Havana data was finally updated completely. Please let me know if you spot any oddities! Added by Felix Kokocinski <u>76 days</u> ago				
	Welcome to GenTrack the Tracking System of the GENCODE project	Tracking System: Flag descriptions A short description of the meaning of flags used in the system can be found at http://www.sanger.ac.uk/gentrack/documents/show/4 Added by Felix Kokocinski <u>112 days</u> ago Tracking System: Tag details The tagging process was refined. The solution for resoved tags is given on the transcript page and curre terms for the controlled vocabulary are displayed when entering a solution.				
Please login here to make (happes	Added by Felix Kokocinski	202 days ago			
You can use the search box display specific regions or c	on this page to find specific genes or you can use filters on the transcript list to	Tracking System: Tags! I've implemented the usage of tags, that can be attached to transcripts to highlight problems. They can be resolved quickly, changes are recorded.				
		Added by Felix Kokocinski	239 days ago			
Please submit bugs or featu subject.	ire requests to the havana RT queue with the word "GenTrack" as part of the	Tracking System: Quick Lin The transcript-listing page i any additional requests. Added by Felix Kokocinski 2	now contains links to different entry points of the tracking system. Let me kn			
Higher priority transcripts: Havana transcript stats		Tracking System: Browser Links added The transcript page now displays links to the affected region in the Ensembl, UCSC, Vega and NCBI browsers. Added by Felix Kokocinski 258 days ago				
Currently loaded ca	tegories (Transcripts)	View all news				
2581 CNIO_isoforms	Principal gene products (CNIO)]	Current flags				
40 CRG_U12 [U12]	oredictions (CRG)]	Current flags				
	s [Additional exons (ENSEMBL)]	Flag descriptions				
	IS [Additional introns (ENSEMBL)]	Unresolved flags: 64478	Resolved flags: 544			
	ing_cDNA [cDNAs not annotated (ENSEMBL)]	Unresolved flag details:				
	ted models (ENSEMBL)]	U12	40			
	al annotation (HAVANA, WTSI)]	assembly_transformation	n 6			
	ps [Transcripts with coding overlap (WTSI)]	ccds	1421			
	mparative protein-coding exon predictor (MIT)]	congo_nc	938			
		congo novel	34			
	volutionary signatures (MIT)]	flagged_pseudogene	5490			
	RetroFinder pseudogenes (UCSC)]	hgnc_update	43			
	[TransMap cross-species alignments (UCSC)] [evolutionarily conserved protein-coding exons (UCSC)]	manual_selection	2			
Added by External DAS	sources <u>106 days</u> ago.					
Status :	New	Created :	16.12.2008			
Priority :	Urgent	Updated :	16.12.2008			
Category :	UCSC_Retro3	Part of:	NM_006421_3-6_chr22_14565821			
Chrom :	22	Start :	14565822			
Strand :	+	End :	14566284			
Transcripttype :	ucsc_retrofinder	Transcriptstatus :				
Transcriptversion :						
Description						
NM_006421.3-6.chr22.	14565821					
Browser-Links						
ENSEMBL UCS						
	NUB					

(NM_006421_3-6_chr22_14565821, 2008-12-16)

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	