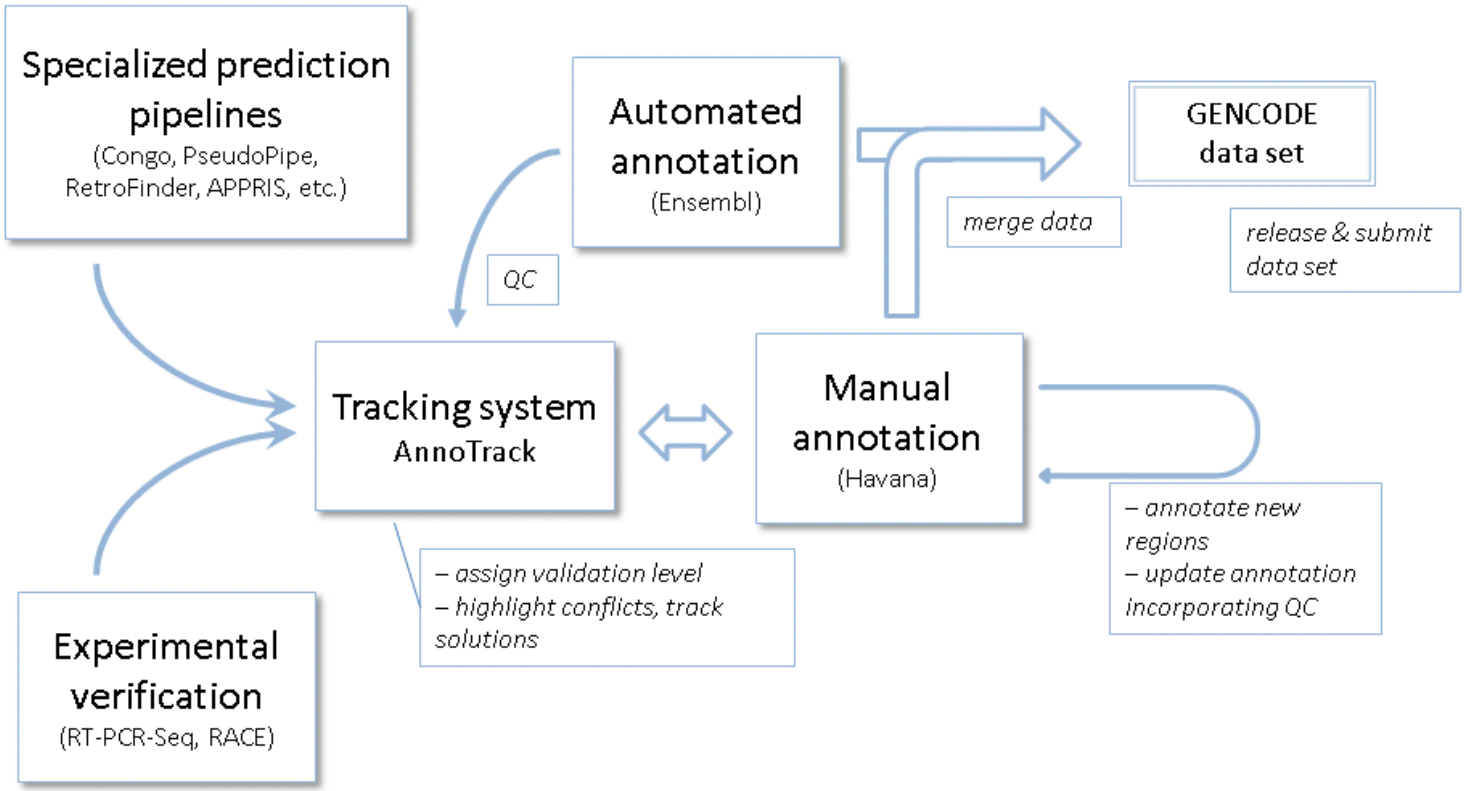
Examining reference gene sets

Jen Harrow
WTSI

What is GENCODE ?

- Human Reference gene set generated for the Encode Project
- Merge of automatic (Ensembl) and manual annotation (Havana) -release 4 May
- Experimental verification by RTPCR and RACE of transcripts(Lausanne)
- Manually annotate all loci biotypes (protein-coding, pseudogenes, “processed-transcripts/non coding RNAs”), small RNAs from Rfam and miRbase mapped to genome by ensembl
- Visualise in UCSC and Ensembl, collaborate with WashU, UCSC, Broad, Yale, CRG

GENCODE pipeline



gencodegenes.org

- Project
- Data
- Participants
- Publications
- RGASP 1/2
- RGASP 3
- Contact us

Statistics about the current Gencode freeze (version 11)

*The statistics derive from the **gtf files**, which include only the main chromosomes of the human reference genome.

Version 11 (October 2011 freeze, GRCh37)

General stats

Total No of Genes	53639	**Total No of Transcripts**	180272
Protein-coding genes	20107	**Protein-coding transcripts**	81040
Long non-coding RNA genes	11600	**- full length protein-coding:**	60661
Small non-coding RNA genes	8801	**- partial length protein-coding:**	20379
Pseudogenes	12761	**Nonsense mediated decay transcripts**	10525
- processed pseudogenes:	9387	**Long non-coding RNA loci transcripts**	18566
- unprocessed pseudogenes:	2446		
- unitary pseudogenes:	156		
- polymorphic pseudogenes:	27		
- pseudogenes:	553		
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments:	370		
- pseudogenes:	192		

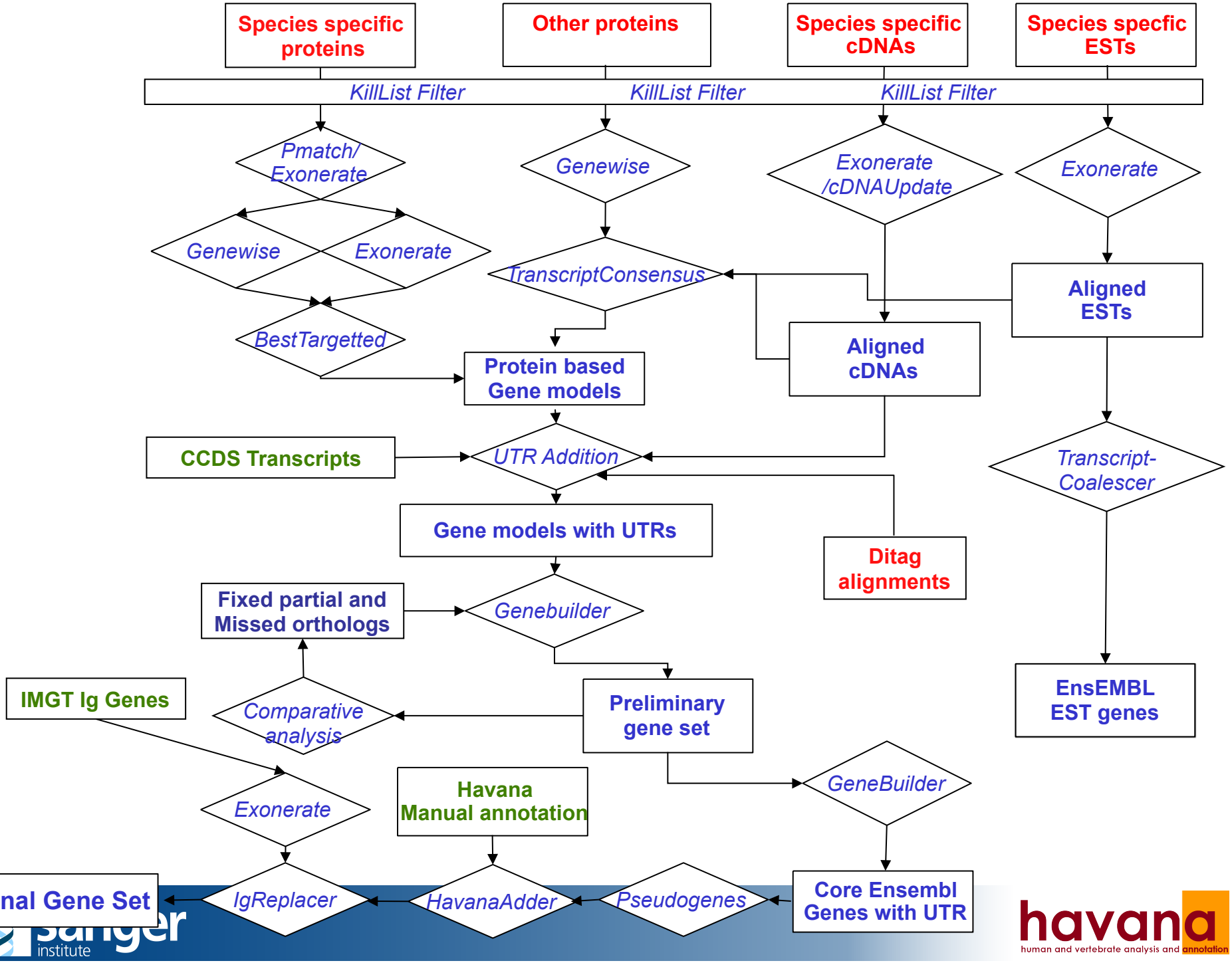
Automatic Annotation vs Manual

Automatic Annotation

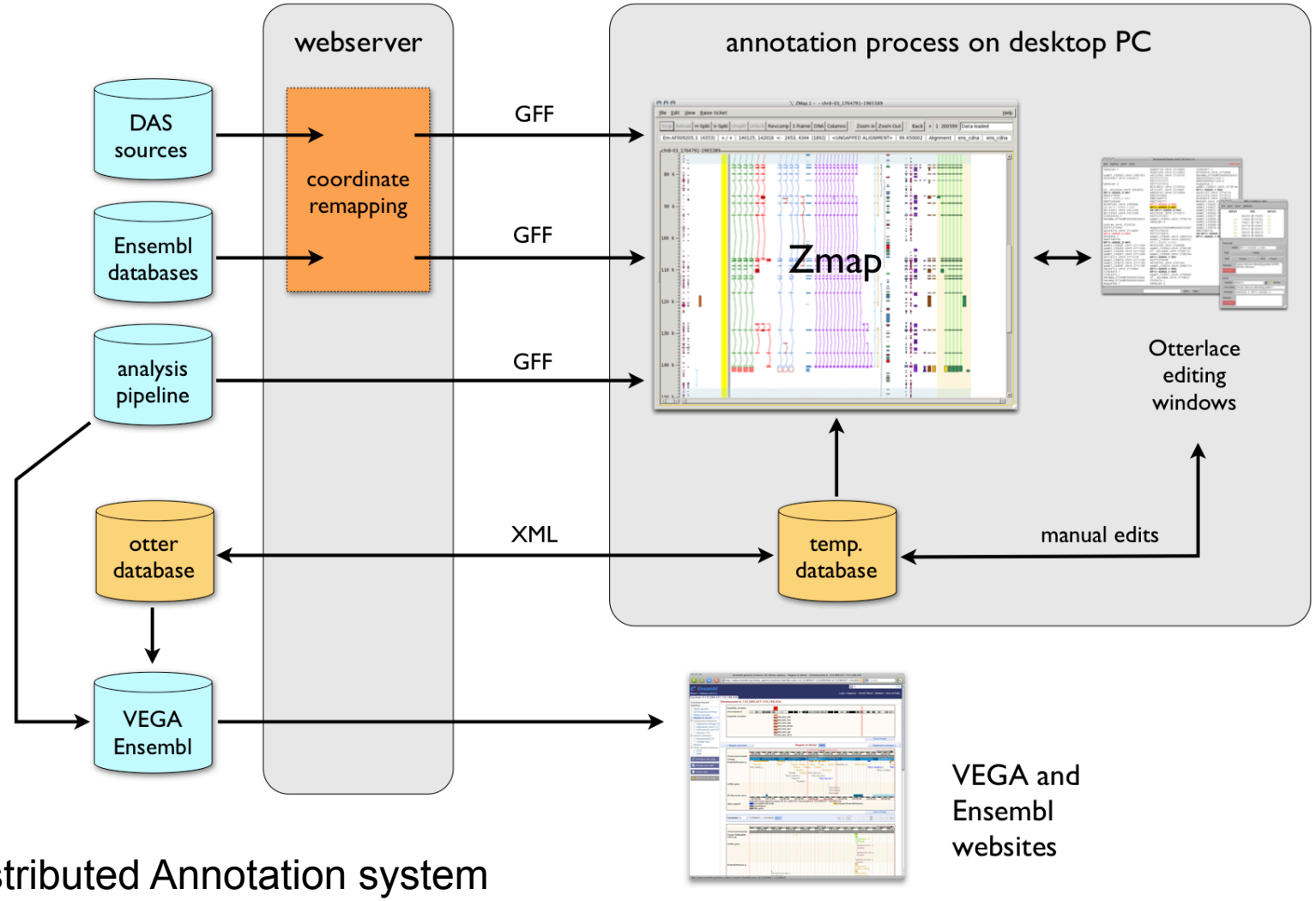
- Quick whole genome analysis ~ weeks
- Consistent annotation
- Use unfinished/illumina sequence/shotgun assembly
- No polyA sites/signals, pseudogene
- Predicts ~80% loci

Manual Annotation

- Extremely slow~3 months Chr 6
- Need finished (high quality) seq
- Flexible, can deal with inconsistencies in data
- Most rules have exception
- Consult publications as well as databases

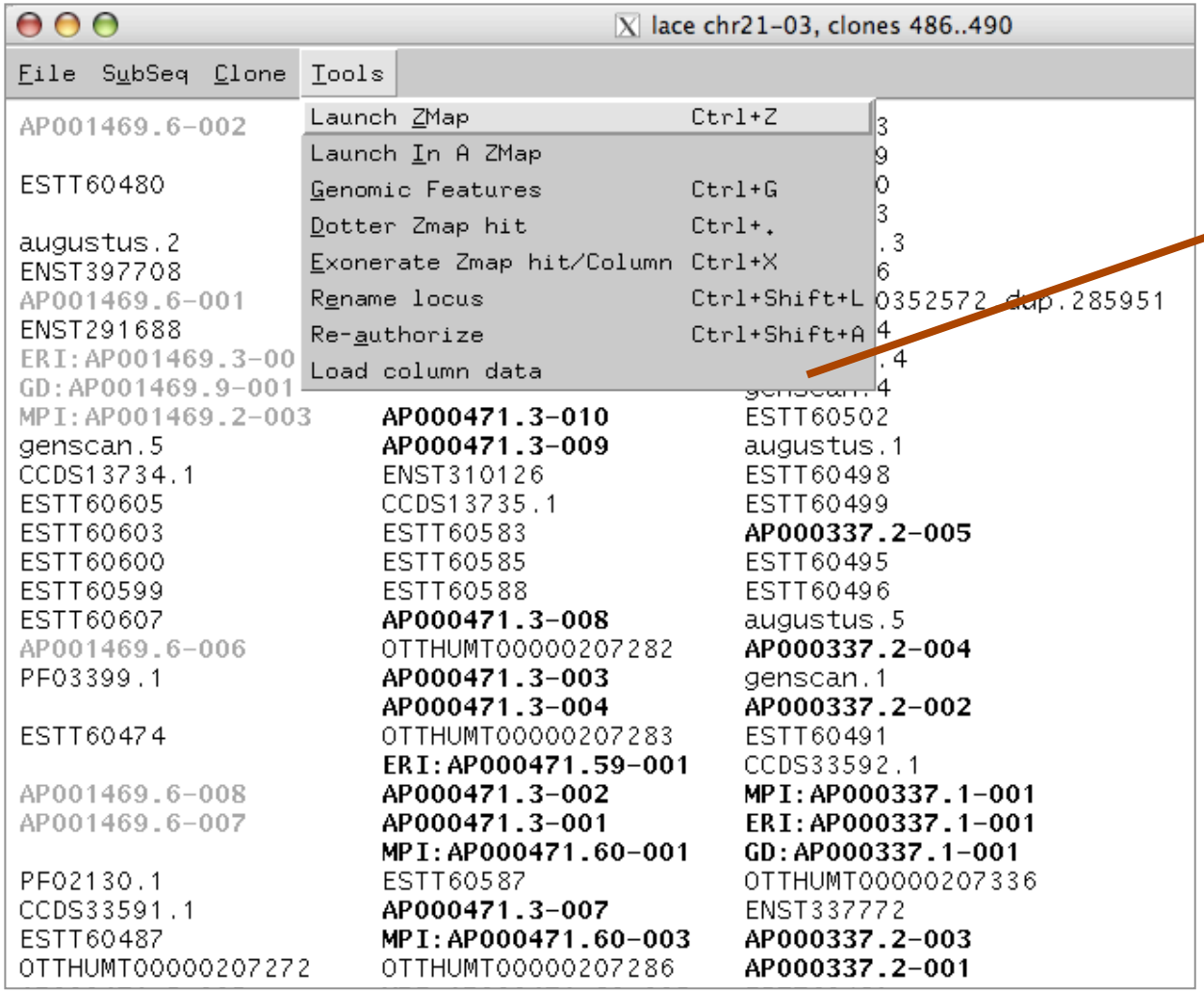

Species specific proteins
Other proteins
Species specific cDNAs
Species specfic ESTs
KillList Filter
KillList Filter
KillList Filter
Pmatch/ Exonerate
Genewise
Exonerate /cDNAUpdate
Exonerate
Genewise
Exonerate
TranscriptConsensus
Aligned ESTs
BestTargetted
Aligned cDNAs
Protein based Gene models
CCDS Transcripts
UTR Addition
Transcript-Coalescer
Gene models with UTRs
Ditag alignments
Fixed partial and Missed orthologs
Genebuilder
IMGT Ig Genes
Comparative analysis
Preliminary gene set
EnsEMBL EST genes
GeneBuilder
Exonerate
Havana Manual annotation
Final Gene Set
IgReplacer
HavanaAdder
Pseudogenes
Core Ensembl Genes with UTR
sanger institute
havana
human and vertebrate analysis and annotation

Otterlace pipeline

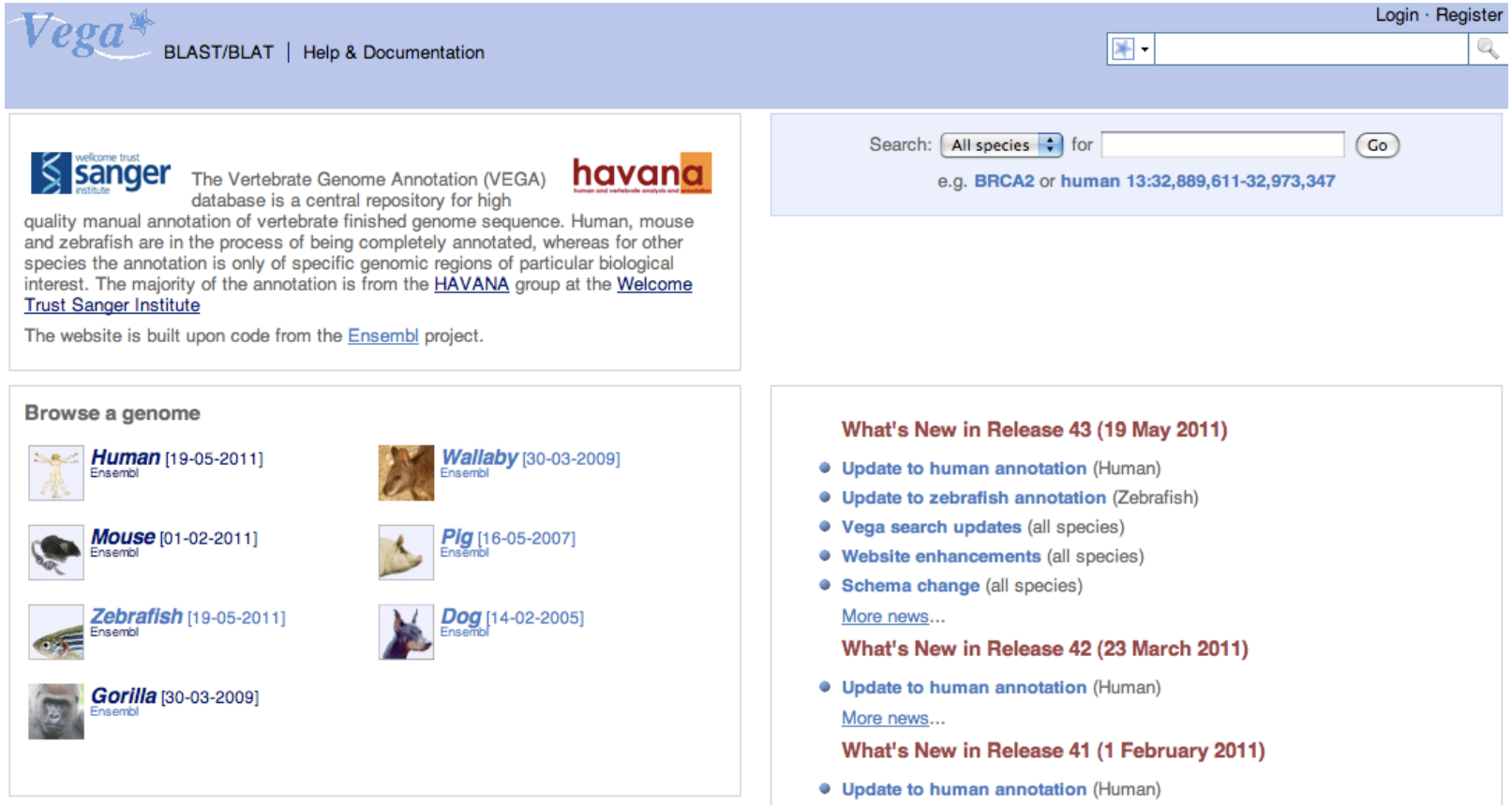


DAS=Distributed Annotation system

DAS source visible in Zmap



Manual annotation browser: Vega



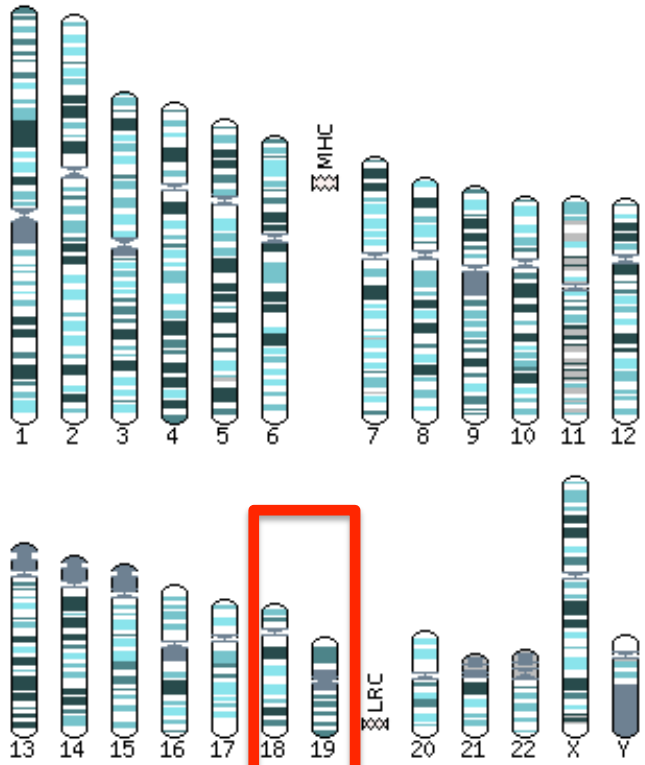
Human annotation update

Statistics

Last update	3 April 2012
Annotated Vega genes:	46,369
Protein coding	19,557
Processed transcripts	12,165
Pseudogenes	13,362
IG & TR Genes	631
Annotated transcripts	166,571
Annotated exons	1,017,199
Total Bases	3,274,812,713
Golden Path Length	3,096,295,592

Annotation progress

Click on a chromosome to browse:

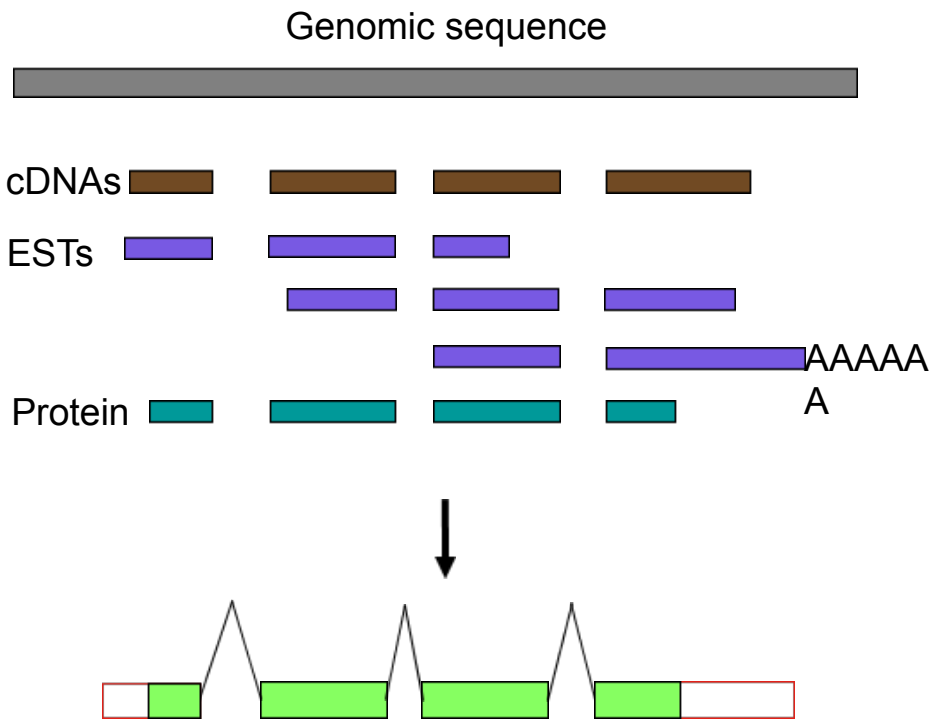


Shading indicates the karyotype bands

To be annotated by Havana

Manual Annotation and Biotypes

Annotation based on transcriptional evidence.



Protein Coding

- Known_CDS
- Novel_CDS
- Putative_CDS
- Nonsense_mediated_decay

Transcript

- Non_coding
- Antisense
- Retained_intron
- Putative
- Artefact

Pseudogene

- Processed_pseudogene
- Unprocessed_pseudogene
- Transcribed_processed
- Transcribed_unprocessed
- Unitary_pseudogene
- Polymorphic_pseudogene

Havana Biotypes now in Ensembl

Show/hide columns Search:

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
SHQ1-001	ENST00000325599	2879	ENSP00000315182	577	Protein coding	CCDS33788
SHQ1-004	ENST00000463369	2065	ENSP00000417452	549	Protein coding	-
SHQ1-005	ENST00000482785	502	ENSP00000418398	111	Protein coding	-
SHQ1-003	ENST00000444040	2844	ENSP00000402447	50	Nonsense mediated decay	-
SHQ1-006	ENST00000471526	553	ENSP00000417739	63	Nonsense mediated decay	-
SHQ1-002	ENST00000468371	4024	No protein product	-	Processed transcript	-
SHQ1-007	ENST00000468347	134	No protein product	-	Processed transcript	-
SHQ1-008	ENST00000475558	666	No protein product	-	Processed transcript	-

NB NMD variants coding unlike RefSeq

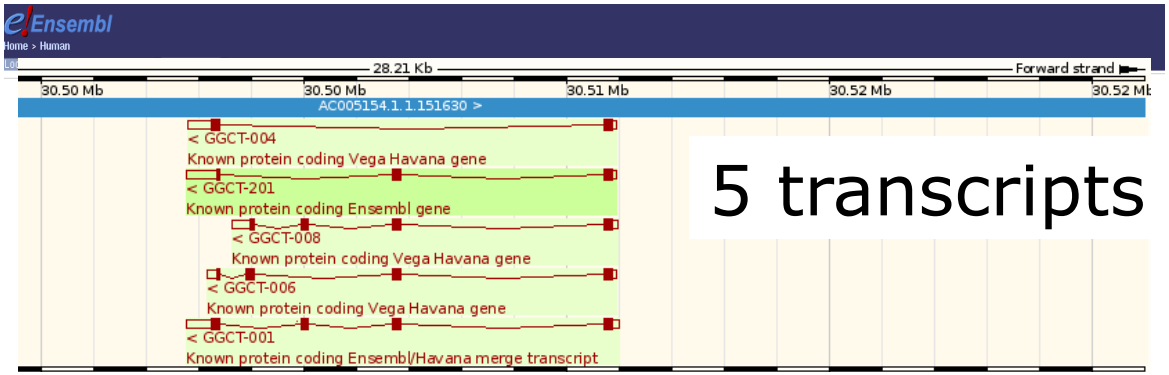
Status (known/novel/putative CDS/transcript) not taken from Havana currently gives some indication of confidence.

New confidence levels to be investigated

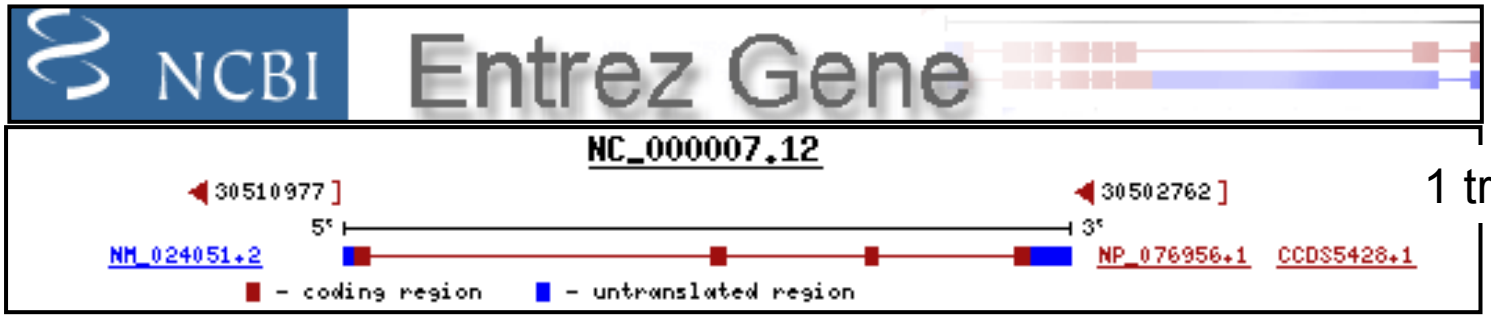
Different gene sets ensembl/ UCSC/NCBI(Refseq)

	human	Mouse
RefSeq	20,669	23,090
Ensembl	21,297	21,111
UCSC	20,930	21,637

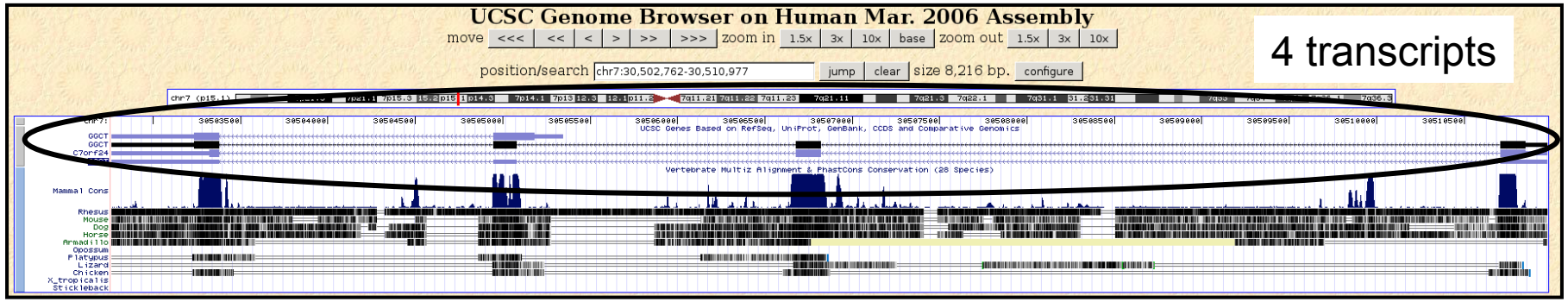
GGCT in Ensembl/Refseq/UCSC



5 transcripts



1 transcript



4 transcripts

Consensus CDS protein set
CCDS Database
EBI • NCBI • UCSC • WTSI

What is CCDS

- Consensus coding sequence project
- UCSC, Ensembl, RefSeq and Havana
- Produce reference CDS set ATG-STOP on human and mouse genome must agree
- 1st rel 2005 13,142 genes 14 795 IDs
- Rel sept 2009 18,177 genes 23 739 IDs
- High quality but few alt splice variants and no UTRs, slow to increase

CCDS website: GATA3 gene ATG->STOP

NCBI — Consensus CDS protein set — CCDS Database — EBI • NCBI • UCSC • WTSI

PubMed | Entrez | Gene | BLAST | OMIM

Search CCDS ID for CCDS15674.1 in All Organisms and Current Builds Go Clear

CCDS
- Home
- FTP
- Process
- Statistics
- AUG-guidelines

Collaborators
- EBI
- NCBI
- UCSC
- WTSI

Contact Us
- GenComp eMail

Genome Displays
- E Ensembl
- U Genome Browser
- N Map Viewer
- V VEGA

Related Resources
- Entrez Gene
- HomoloGene
- RefSeq
- UniGene

Report for CCDS ID CCDS15674.1

CCDS	Status	Species	Chrom.	Gene	NCBI Builds	Links
15674.1	Public	*Mus musculus*	2	Gata3	36.1 - 37.1	H G G

Sequence IDs included in CCDS 15674.1

Original	Current	Source	Nucleotide ID	Protein ID	Status in CCDS	Seq. Status	Links
✔	✔	EBI,WTSI	ENSMUST00000102976	ENSMUSP00000100041	Accepted	alive	N P N P
✔	✔	EBI,WTSI	OTTMUST00000026063	OTTMUSP00000011932	Accepted	alive	N P N P
✔		NCBI	NM_008091.2	NP_032117.1	Updated	not alive	N P N P
	✔	NCBI	NM_008091.3	NP_032117.1	Accepted	alive	N P N P B

Chromosomal Locations for CCDS 15674.1

On '-' strand of Chromosome 2 (NC_000068.6)

Genome Browser links: N U E V

Chromosome	Start	Stop	Links
2	9779997	9780281	N N U E V
2	9784722	9784847	N N U E V
2	9790388	9790533	N N U E V
2	9796016	9796552	N N U E V
2	9798979	9799216	N N U E V

CCDS Sequence Data

Blue highlighting indicates alternate exons.

Red highlighting indicates amino acids encoded across a splice junction.

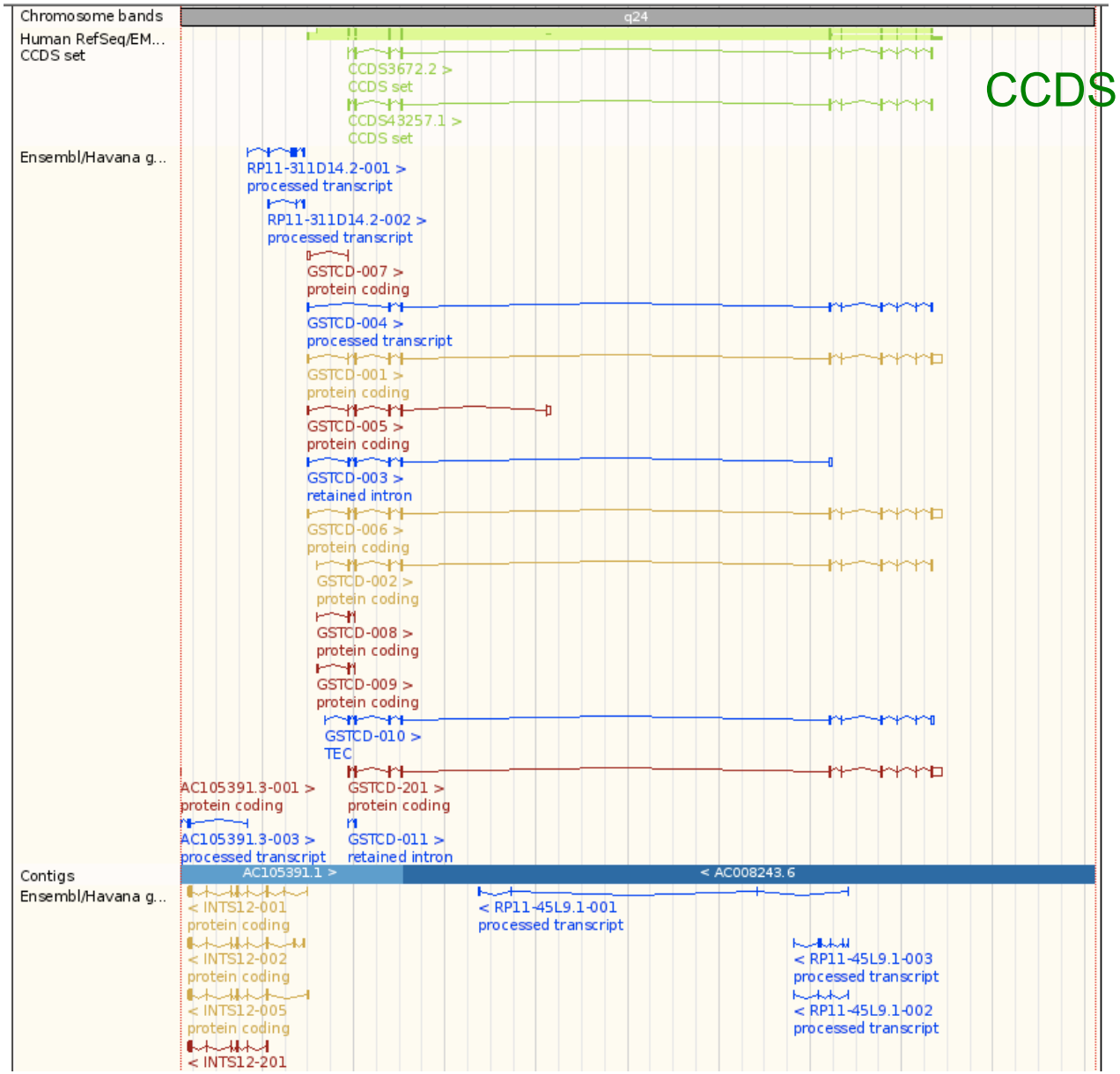
Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (1332 nt):

```
ATGGAGGTGACTGCGGACCAGCCGCGCTGGGTGAGCCACCATCACCCCGCGGTCCTCAACGGTCAGCACC
CAGACACGCACCACCCGGGCCTCGGCCATTCGTACATGGAAGCTCAGTATCCGCTGACGGAAGAGGTGGA
CGTACTTTTTAACATCGATGGTCAAGGCAACCACGTCCCGTCCTACTACGGAAACTCCGTCAGGGCTACG
GTGCAGAGGTATCCTCCGACCCACCACGGGAGCCAGGTATGCCGCCCGCCTCTGCTGCACGGATCTCTGC
CCTGGCTGGATGGCGGCAAAGCCCTGAGCAGCCACCACACCGCCTCGCCCTGGAACCTCAGCCCCTTCTC
CAAGACGTCCATCCACCACGGCTCTCCGGGGCCTCTGTCCGTTTACCCTCCGGCTTCATCCTCTTCTCTG
GCGGCCGGCCACTCCAGTCCTCATCTCTTCACCTTCCCGCCCACCCCGCCGAAAGACGTCTCCCCAGACC
CGTCGCTGTCCACCCCGGGATCCGCCGGGTCGGCCAGGCAAGATGAGAAAGAGTGCCTCAAGTATCAGGT
GCAGCTGCCAGATAGCATGAAGCTGGAGACGTCTCACTCTCGAGGCAGCATGACCACCCTGGGTGGGGCC
TCATCCTCAGCCCACCACCCCATTACCACCTATCCGCCCTATGTGCCCGAGTACAGCTCTGGACTCTTCC
CACCCAGCAGCCTGCTGGGAGGATCCCCTACCGGGTTCGGATGTAAGTCGAGGCCCAAGGCACGATCCAG
CACAGAAGGCAGGGAGTGTGTGAACTGCGGGGCAACCTCTACCCCACTGTGGCGGCGAGATGGTACCGGG
CACTACCTTTGCAATGCCTGCGGACTCTACCATAAAATGAATGGGCAGAACCGGCCCCTTATCAAGCCCA
AGCGAAGGCTGTCGGCAGCAAGGAGAGCAGGGACATCCTGCGCGAACTGTCAGACCACCACCACCACCCT
CTGGAGGAGGAACGCTAATGGGGACCCGGTCTGCAATGCCTGTGGGCTGTACTACAAGCTTCATAATATT
AACAGACCCCTGACTATGAAGAAAGAAGGCATCCAGACCCGAAACCGGAAGATGTCTAGCAAATCGAAAA
AGTGCAAAAAGGTGCATGACGCGCTGGAGGACTTCCCCAAGAGCAGCTCCTTCAACCCGGCCGCTCTCTC
CAGACACATGTCATCCCTGAGCCACATCTCTCCCTTCAGCCACTCCAGCCACATGCTGACCACACCGACG
CCCATGCATCCGCCCTCCGGCCTCTCCTTCGGACCTCACCACCCTTCCAGCATGGTCACCGCCATGGGTT
AG
```

Translation (443 aa):

```
MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLGHSYMEAQYPLTEEVDVLFNIDGQGNHVPSYYGNSVRAT
VQRYPPTHHGSQVCRPPLLHGSLPWLDGGKALSSHHTASPWNLSPFSKTSIHHGSPGPLSVYPPASSSSL
AAGHSSPHLFTFPPTPPKDVSPDPSLSTPGSAGSARQDEKECLKYQVQLPDSMKLETSHSRGSMTTLGGA
SSSAHHPITTYPPYVPEYSSGLFPPSSLLGGSPTGFGCKSRPKARSSTEGRECVNCGATSTPLWRRDGTG
HYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTSCANCQTTTTTLWRRNANGDPVCNACGLYYKLHNI
NRPLTMKKEGIQTRNRKMSSKSKKCKKVHDALEDFPKSSSFNPAALSRHMSSLSHISPFSHSSHMLTTPT
PMHPPSGLSFGPHHPSSMVTAMG
```



CCDS

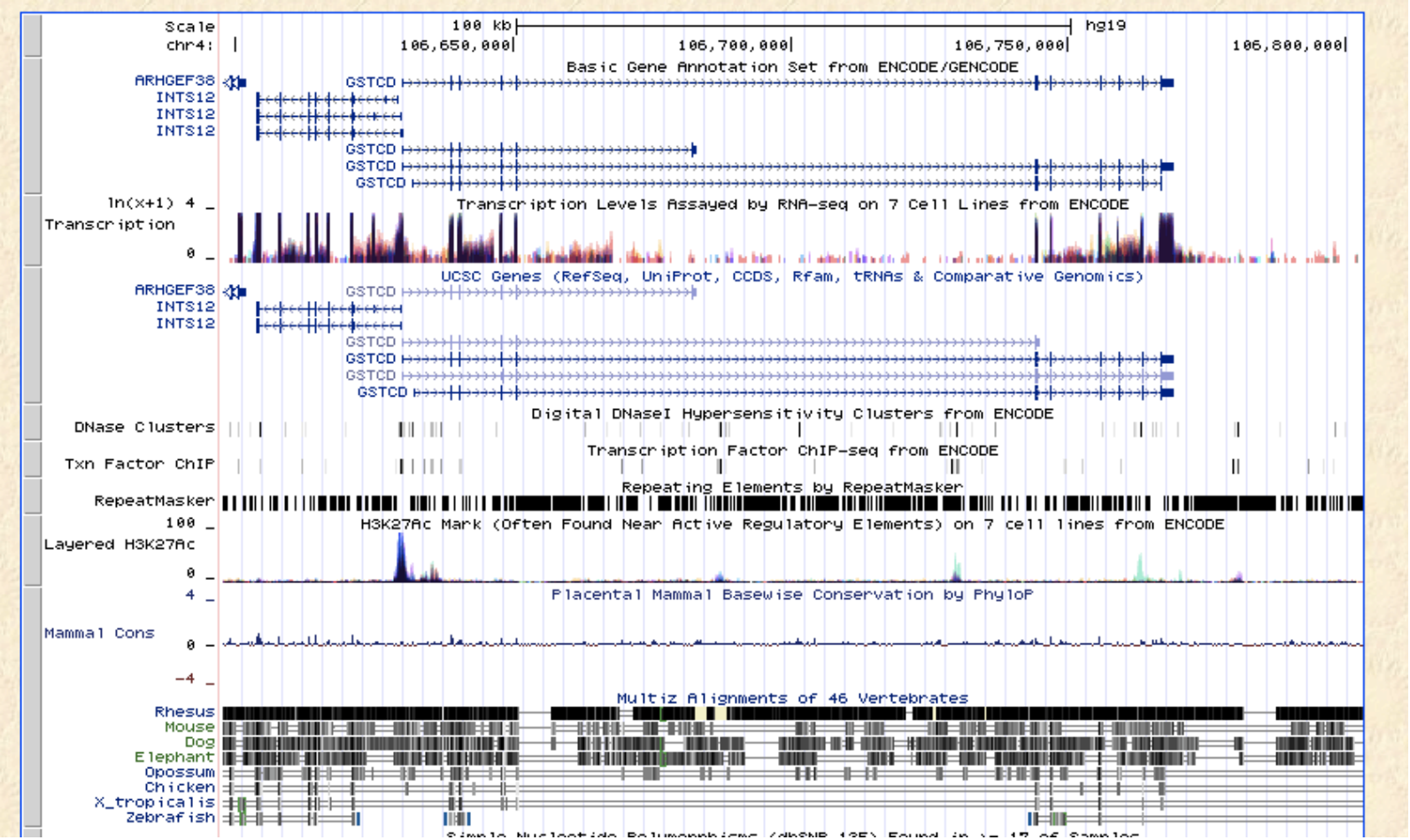
Ensembl view

Gold: agreed ensembl/havana

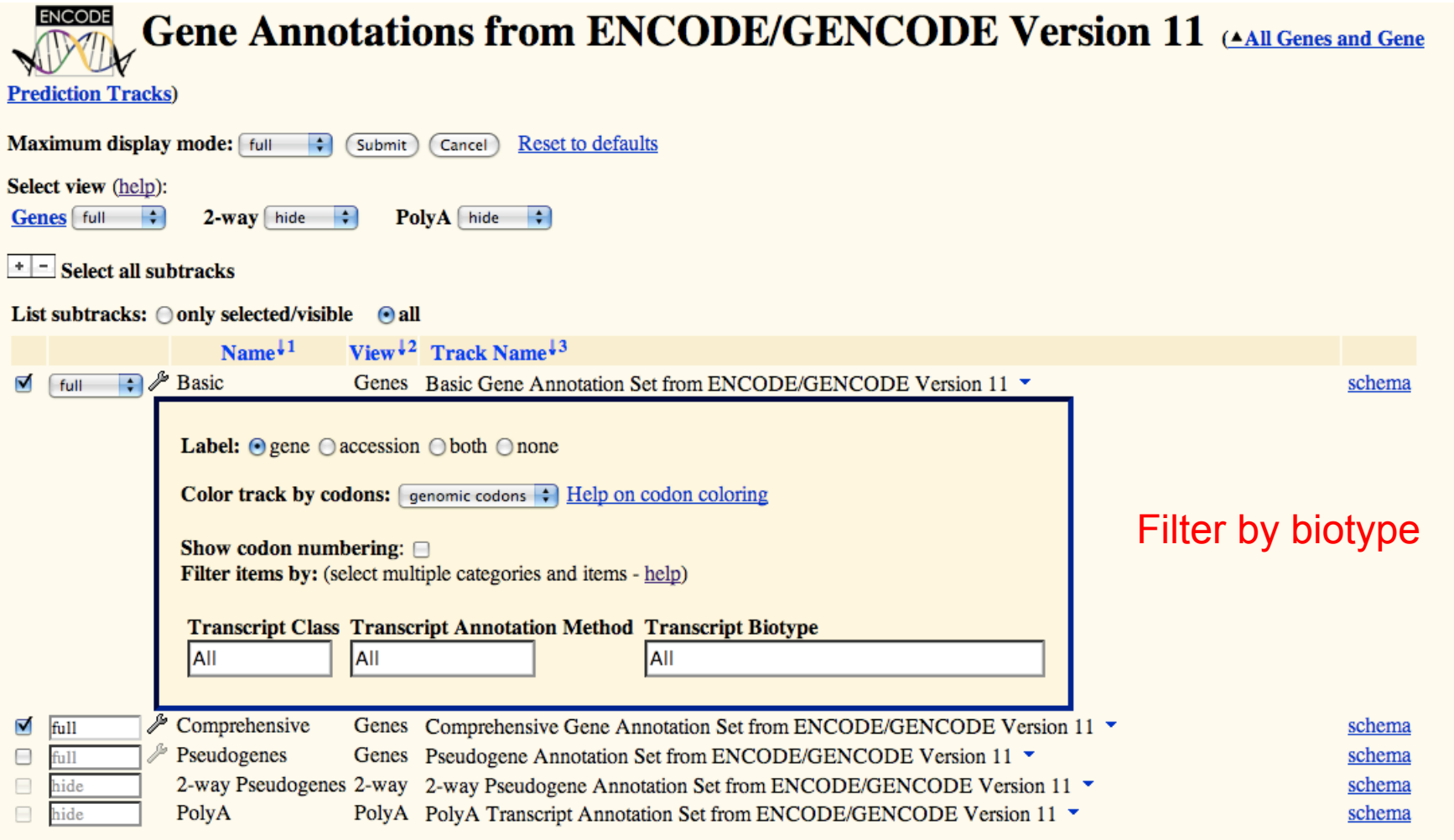
Red : coding (001 havana, 201 ensembl)

Blue :non-coding

UCSC View of GENCODE genes



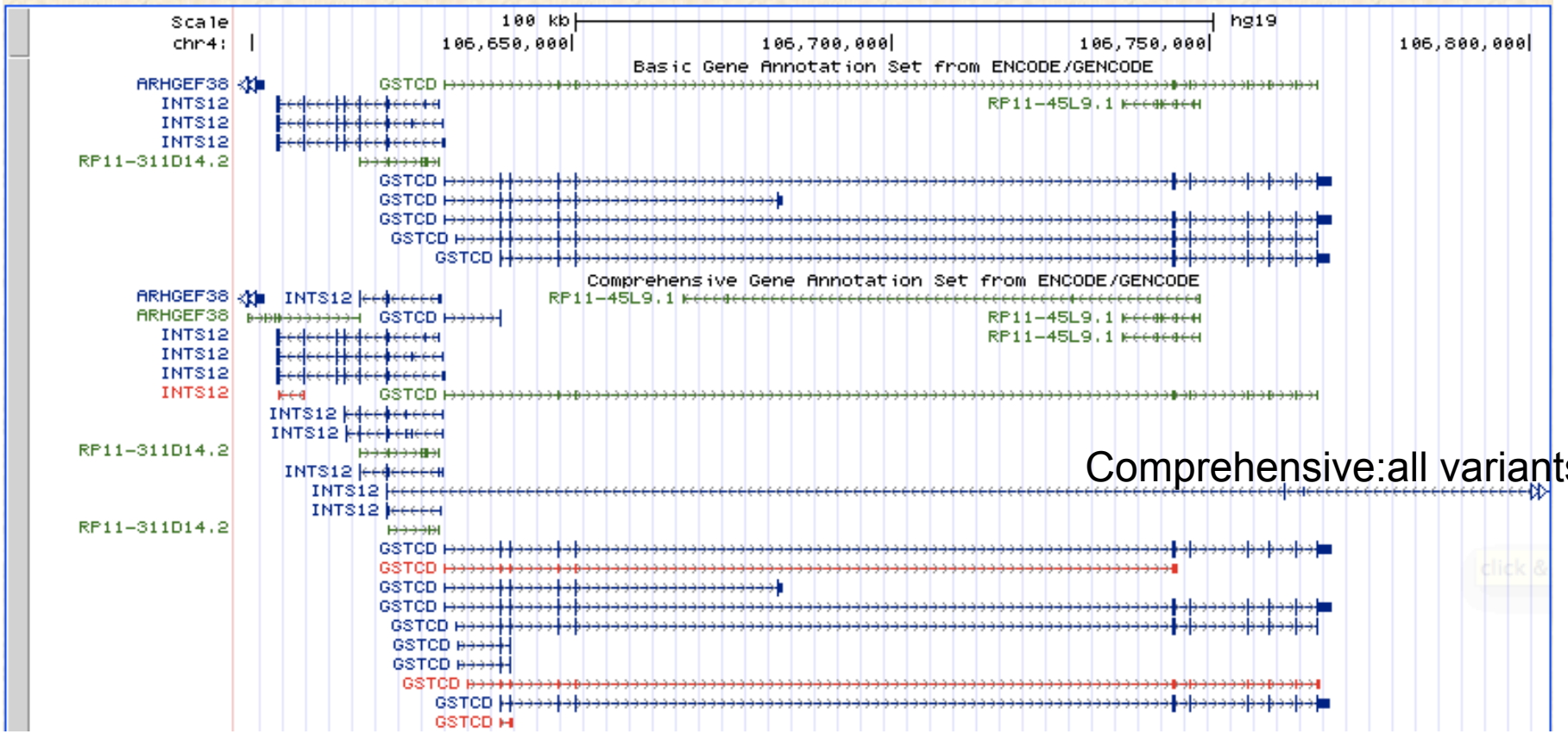
Changing default settings



Filter by biotype

Basic vs comprehensive GENCODE

BASIC:full length



Comprehensive:all variants

Manual and automatic | coding | non-coding | pseudogene | problem

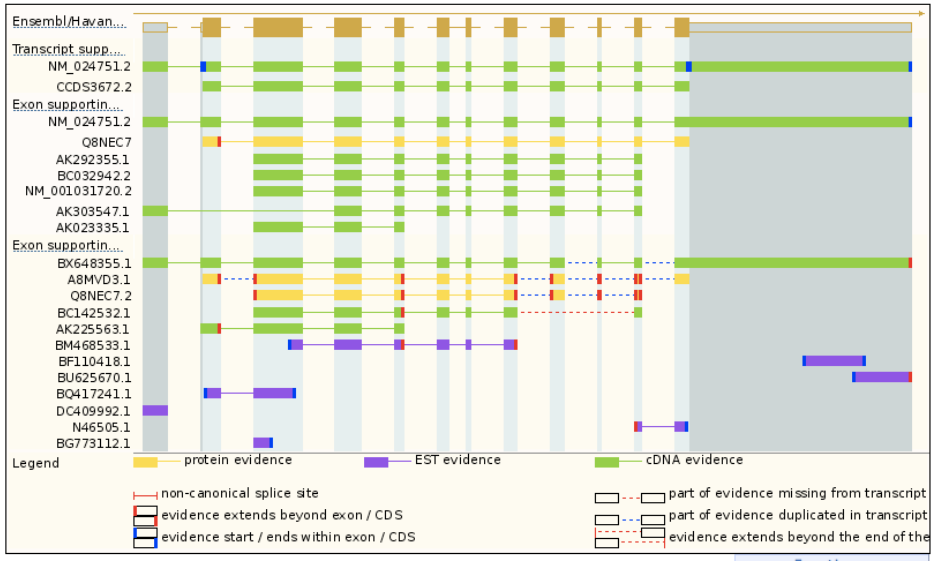
Evidence for annotation

	Transcript	Gene
Gencode id	ENST00000394730.3	ENSG00000138780.9
HAVANA manual id	OTTHUMT00000253947.3	OTTHUMG00000131211.4
Position	chr4:106629941-106768882	chr4:106629941-106768882
Strand	+	
Biotype	protein_coding	protein_coding
Status	KNOWN	KNOWN
Annotation Level	manual (2)	
Annotation Method	manual & automatic	manual & automatic
HUGO gene	GSTCD	
CCDS	CCDS3672.2	

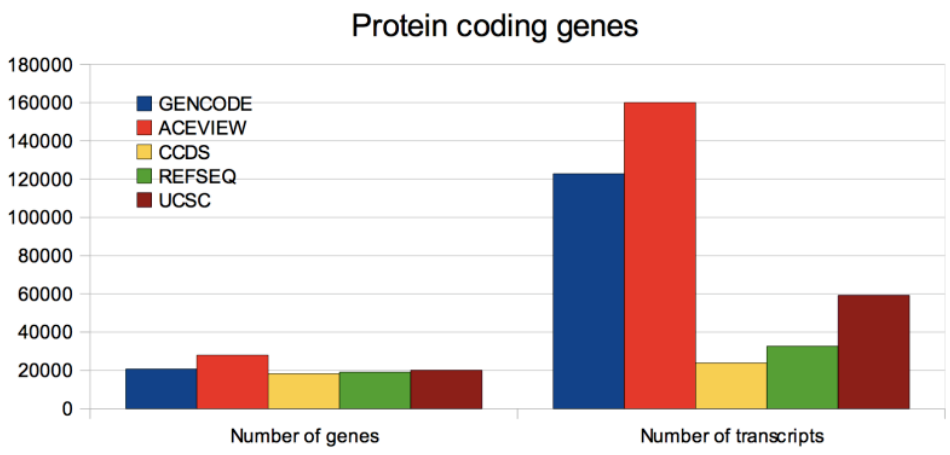
Show All entries Show/hide columns Filter

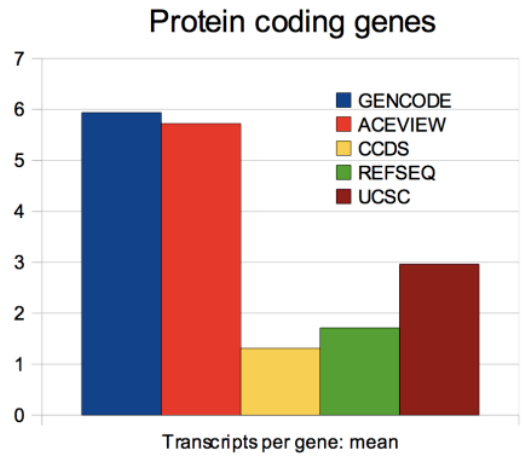
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
GSTCD-001	ENST00000394730	4043	ENSP00000378218	546	Protein coding	CCDS3672
GSTCD-002	ENST00000360505	2154	ENSP00000353695	633	Protein coding	CCDS43257
GSTCD-005	ENST00000507281	2091	ENSP00000422858	340	Protein coding	-
GSTCD-006	ENST00000515279	4273	ENSP00000422354	633	Protein coding	CCDS43257
GSTCD-007	ENST00000512828	572	ENSP00000423639	40	Protein coding	-
GSTCD-008	ENST00000510865	578	ENSP00000423792	166	Protein coding	-
GSTCD-009	ENST00000509336	718	ENSP00000423779	167	Protein coding	-
GSTCD-201	ENST00000394728	4049	ENSP00000378216	633	Protein coding	CCDS43257
GSTCD-010	ENST00000505640	2425	No protein product	-	TEC	-
GSTCD-004	ENST00000515255	1373	No protein product	-	Processed transcript	-
GSTCD-003	ENST00000484843	2228	No protein product	-	Retained intron	-
GSTCD-011	ENST00000503409	653	No protein product	-	Retained intron	-

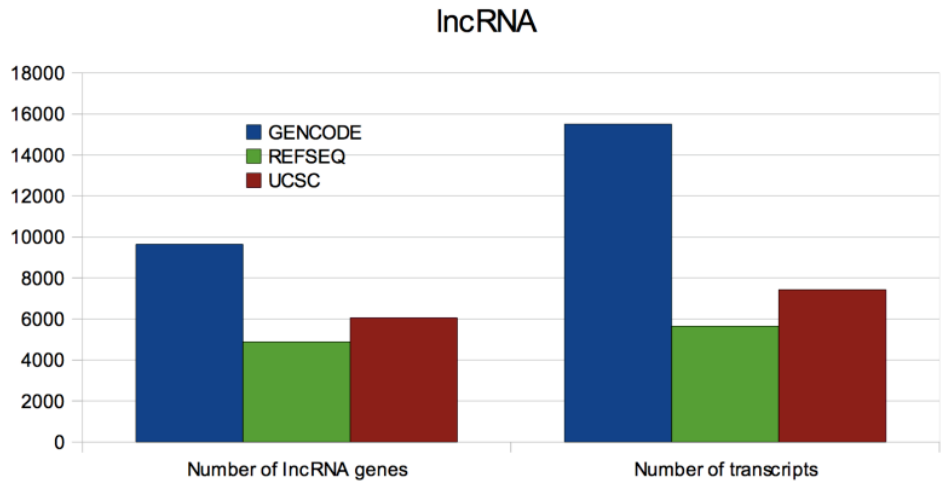
Supporting Evidence			
Source	**Sequence**	**Source**	**Sequence**
CCDS	CCDS3672.2	EMBL	AK023335.1
EMBL	AK225563.1	EMBL	AK292355.1
EMBL	AK303547.1	EMBL	BC032942.2
EMBL	BC142532.1	EMBL	BF110418.1
EMBL	BG773112.1	EMBL	BM468533.1
EMBL	BQ417241.1	EMBL	BU625670.1
EMBL	BX648355.1	EMBL	DC409992.1
EMBL	N46505.1	RefSeq_dna	NM_001031720.2
RefSeq_dna	NM_024751.2	Uniprot/SPTREMBL	A8MVD3.1
Uniprot/SWISSPROT	Q8NEC7	Uniprot/SWISSPROT	Q8NEC7.2

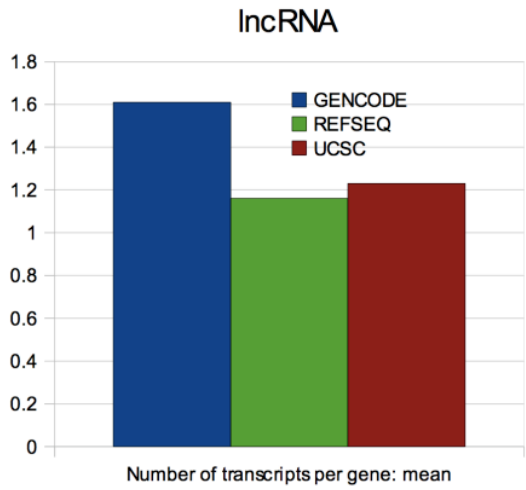


Overview of GENCODE

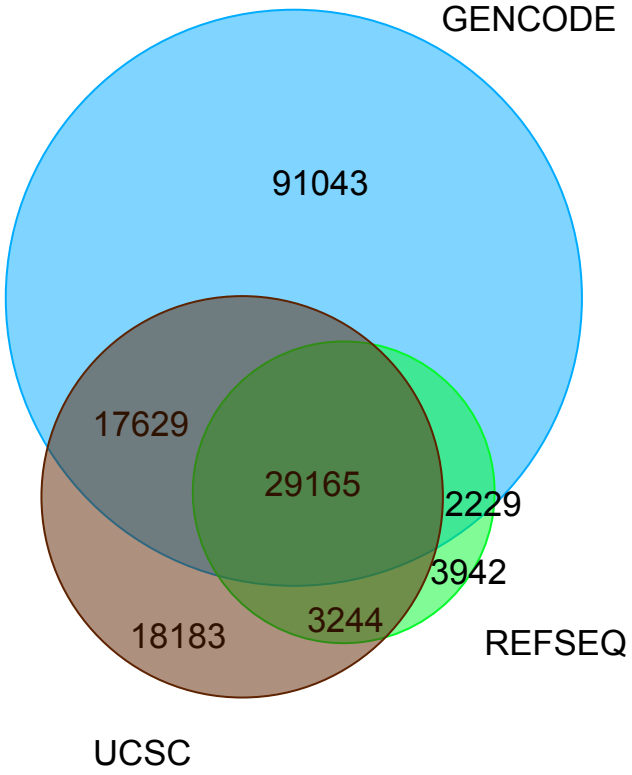




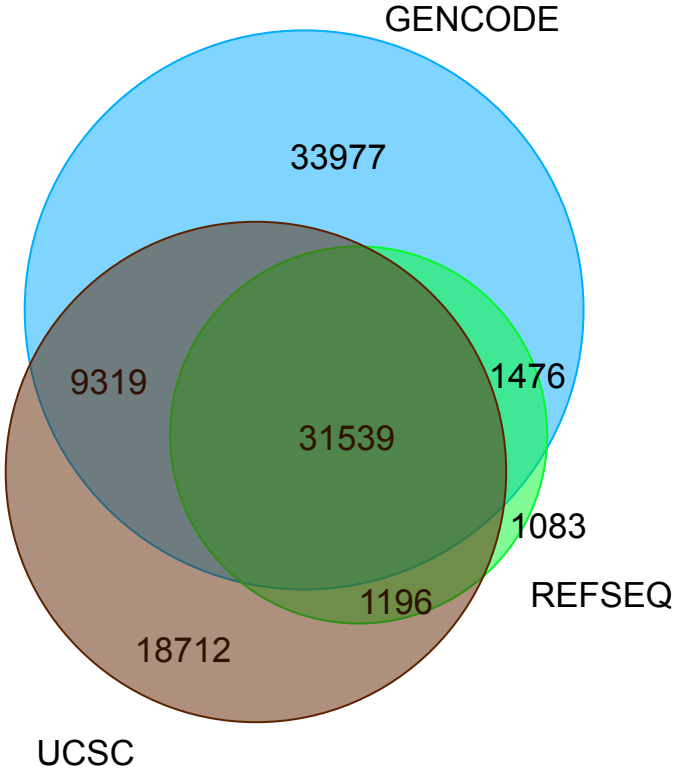




Overlap of exact match transcripts



Transcript

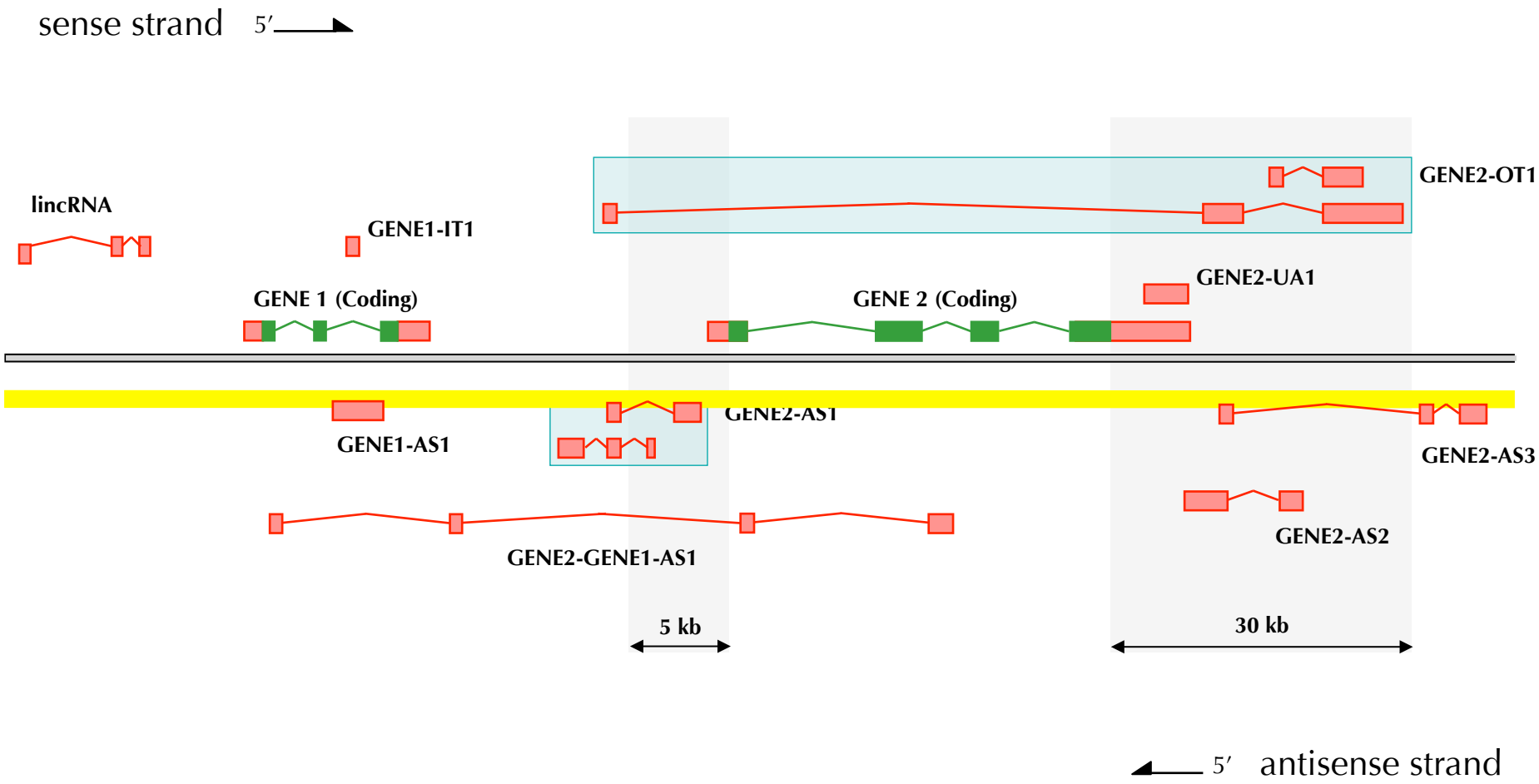


CDS

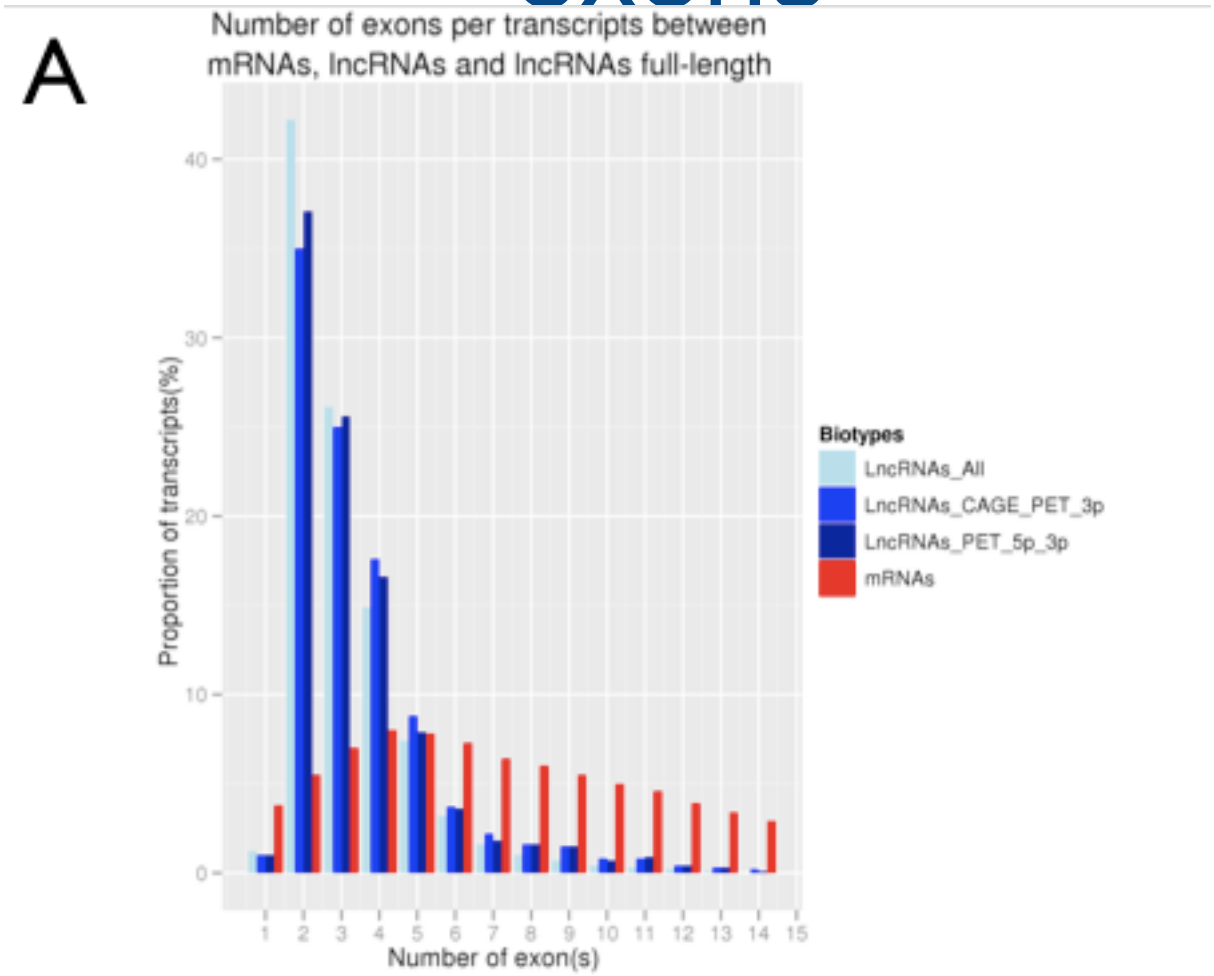
Long noncoding RNA biotypes
(HGNC and John Mattick)

- **lincRNA:**
  - Intergenic >200bp spliced (chromatin signatures observed but not mandatory
- **Antisense:**
  - for transcripts overlapping any part of the genome within 5kb of the start of the CDS and 30 kb of the end of the CDS of a coding locus on the opposite strand.
- **3'_overlapping_ncRNA:**
  - for transcripts where ditag and/or published experimental data strongly supports the existence of short non-coding transcripts transcribed from the 3' UTR.
- **Sense_overlapping:**
  - for transcripts that contains a coding gene in their introns on the same strand.
- **Sense_intronic:**
  - for transcripts that are in introns of coding genes and do not overlap any exons.

Schematic of lncRNA



lncRNA shorter than coding exons



Genome reference consortium (GRC)

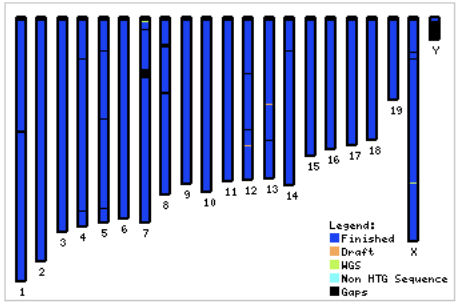
Genome Reference Consortium

GRC Home | Human | Mouse | Help | Report an Issue | Contact Us | Curators Only

Overview | Issues under Review | Assembly Data | Report a problem

Mouse Genome Overview

Information concerning the continuing improvement of the mouse genome.



Mouse Genome Build 37: A graphical representation of the mouse genome in Build 37. The genome is colored with respect to the genomic component used to build the genome assembly at that location. While >95% of the genome is constructed using HTGS phase 3 (finished) sequence, small bits of unfinished and WGS contribute as well.

The most recent assembly for the mouse is Build 37, which was produced by the Mouse Genome Sequencing Consortium (MGSC). This assembly is based on DNA from a single inbred strain (C57BL/6J) and is largely composed of finished clone sequences. There are, however, 105 unspanned gaps and just over 700 spanned gaps remaining in the assembly. Most of the unspanned gaps result from the inclusion of Whole Genome Shotgun (WGS) contigs in the assembly. The inclusion of this sequence was critical in order to preserve genes not represented in the BAC tiling path. Much of the WGS sequence has been placed on the chromosome, but some remains either unplaced or unlocalized .

The GRC is now actively working to close these gaps. Additionally libraries from C57BL/6J have become available and are being screened in an effort to identify clones that span these gaps. Many of the remaining gaps are associated segmental duplication as described in She et al., 2008 . If you know of any remaining issues, please contact us using the links above.

GRC News and Updates

GRCh37 is now available in Map Viewer
Fri, 14 Aug 2009

NCBI has annotated and released the latest version of the public human genome assembly (GRCh37).

GRCh37 now available at Ensembl
Fri, 14 Aug 2009

Ensembl has annotated and released the latest version of the public human genome assembly (GRCh37). This is available as part of Ensembl release 55.

see all

Recently Resolved Mouse Issues

Mouse (MG-40) *Oct 8, 2009*

AC102264.26 is redundant.
Will change the status of AC102264.26 to redundant on sequence curation page.

Mouse (MG-3500) *Oct16, 2009*

Duplicate of MG-3558.

see all

References

Whole Genome Papers

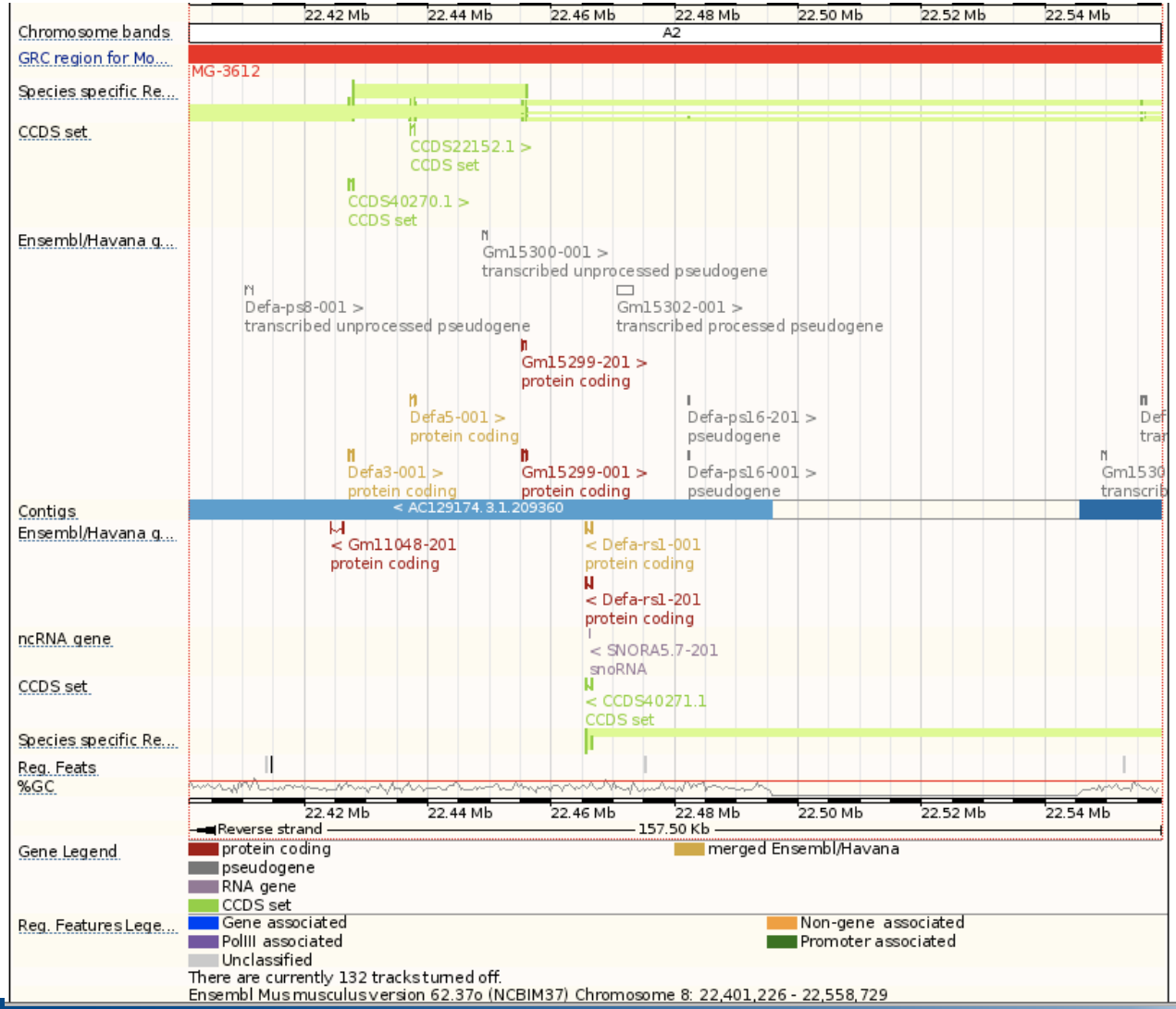
The Mouse Genome WGS Assembly
The Mouse Genome: Clone based assembly

FTP | NHGRI | The Wellcome Trust | HHS | NIH | Accessibility |

Page last updated: Jun 4, 2009

Gaps and assembly issues highlighted which can Affect annotation

GRC problem regions in Ensembl



GRC DAS track
Has link to GRC database

Gap in assembly

Information about assembly errors

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Mouse Overview | Mouse Issues Under Review | Mouse Assembly Data | Report a problem

Issue Report for MG-3612

Category: Gap

Report type: TPF Analysis

Status: Resolved

Description: There is a type 2 reference assembly gap present between components AC129174.3 and AC133094.4

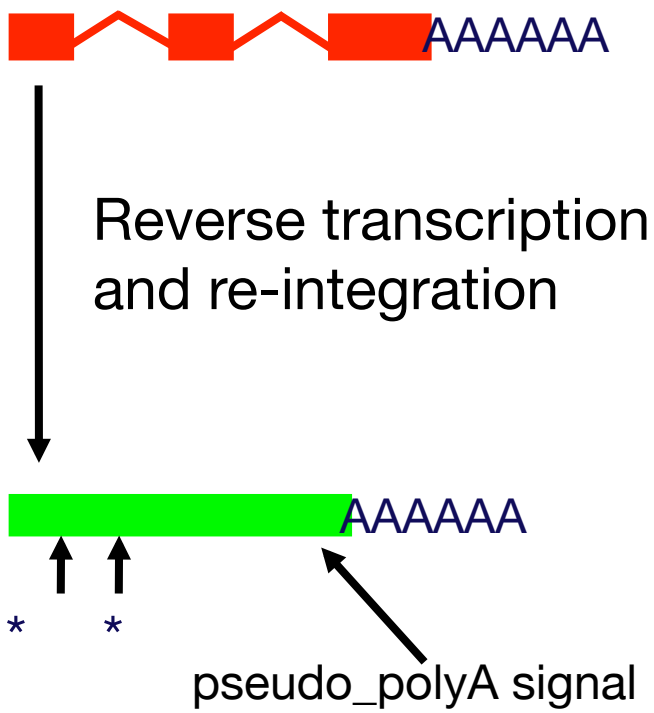
Resolution: RP23-16O24 AC240396.3 and CH36-223O13 AC239604.3 have been selected, sequenced, and submitted, and they close this gap.

Assembly Information

MGSCv37 chr8: 22,336,657-22,729,648 (View Region: Ensembl | NCBI | UCSC)

Exercises

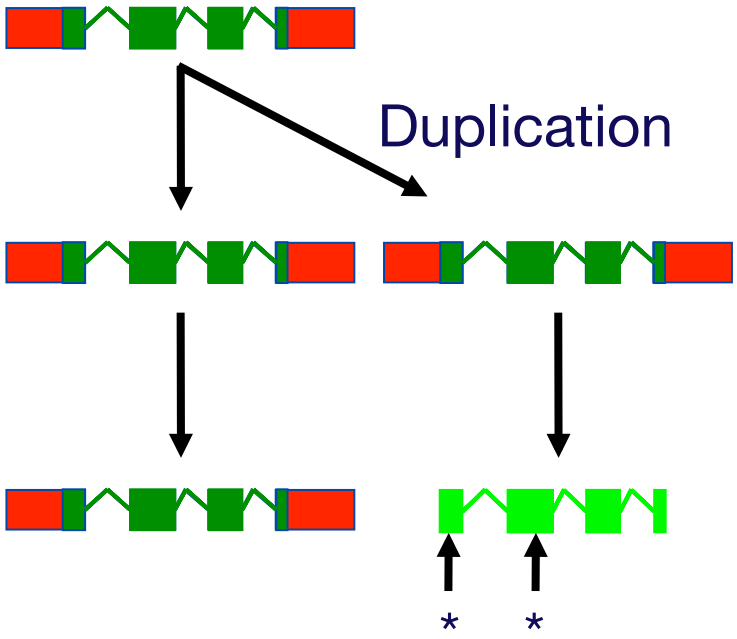
Retrotransposed pseudogene



- mRNA transcript reverse transcribed back into DNA and inserted into chromosomal DNA
- Inserted randomly into genome; introns spliced out; considered "dead on arrival"
- often found by automatic prediction algorithms

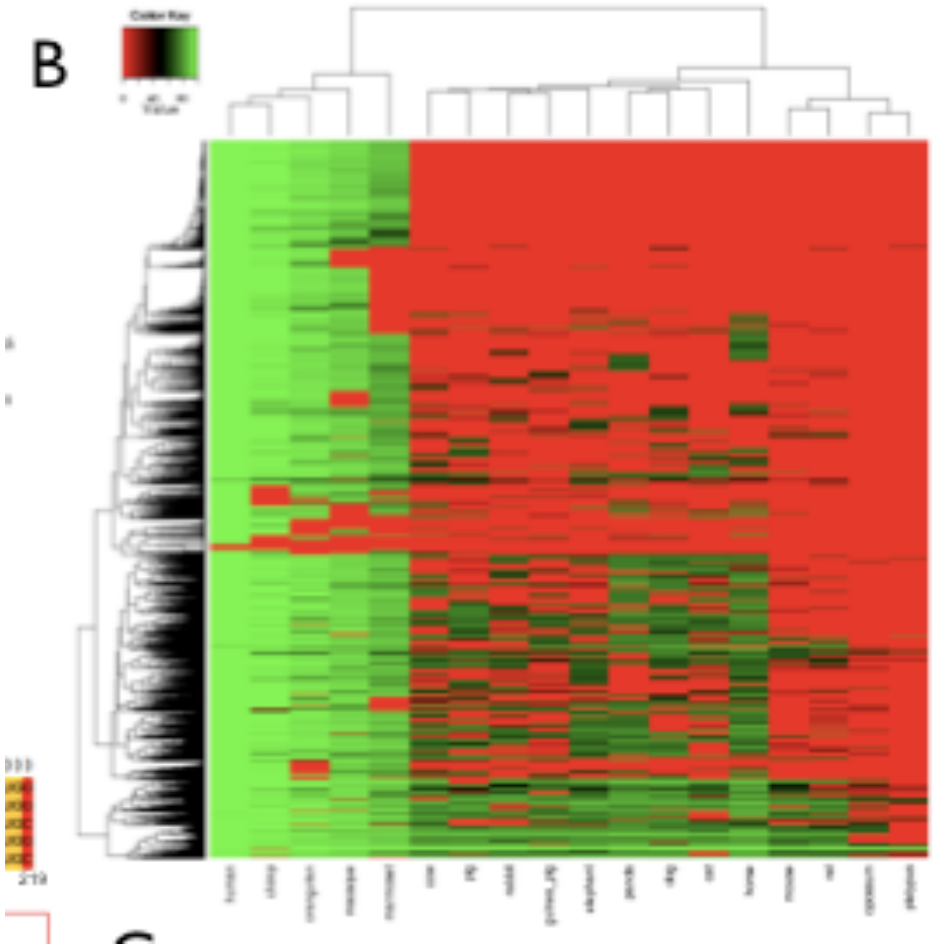
Pseudogene.org

Unprocessed pseudogene



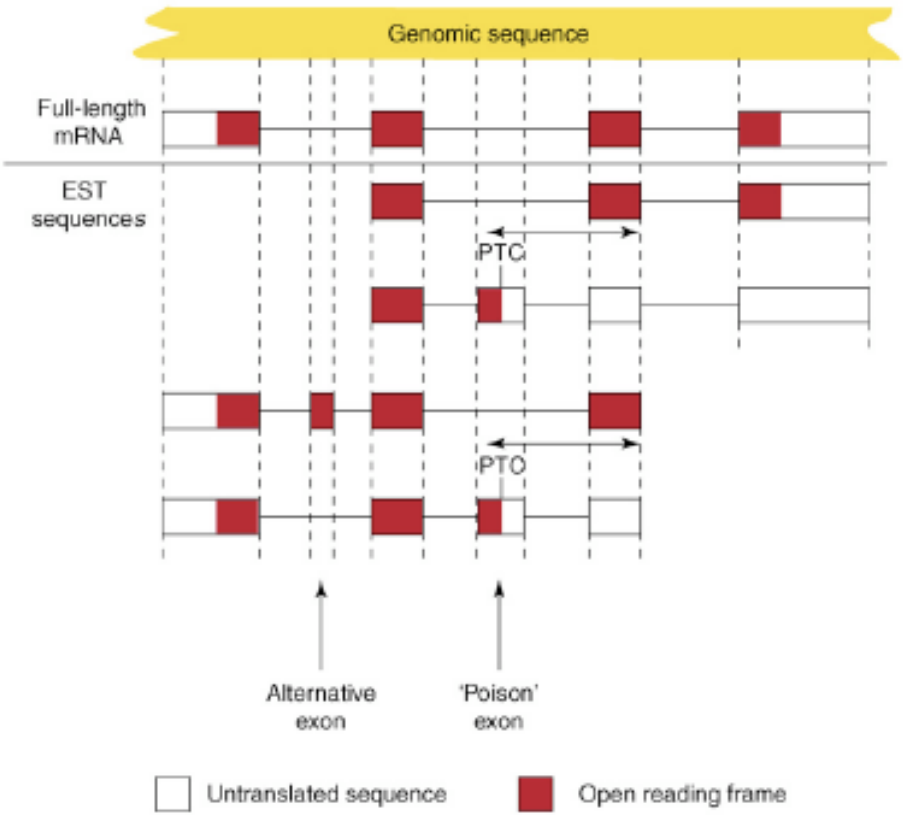
- Gene duplication
- Intact exon-intron structure
- Acquire mutations that result in lose of function
- Identified by lack of CDS

Are lncRNAs conserved?



Guigo's group generated Heat map using blast and exonerate analysis of 9,200 loci from gencode 7 to examine conservation in mammals (platipus/opposum)

Identification of Nonsense mediated decay (NMD) (Stephen Brenner)

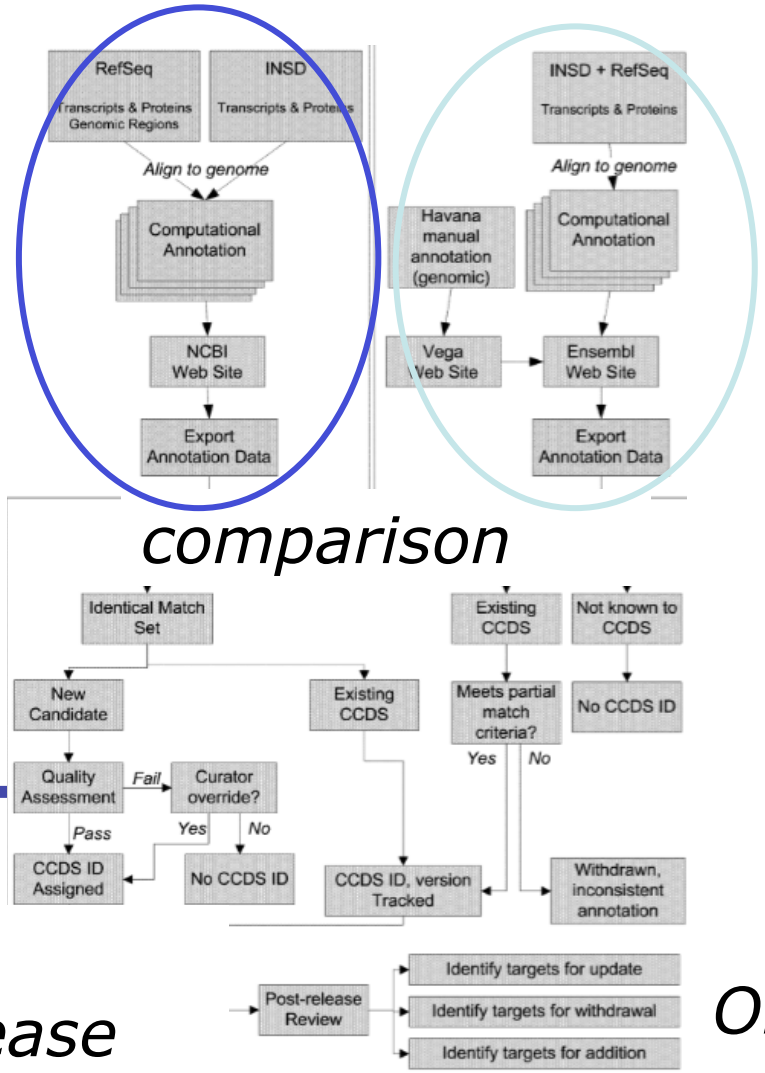


PTC= Premature Termination Codon

TIBs Vol 33:8

CCDS pipeline:producing consensus

NCBI



WTSI

EBI

NMD and ψ are removed

Release

Ongoing revision

Human and Vertebrate analysis and Annotation (HAVANA)group

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation