

Module 1

de novo Analysis of Sequence and Manual Genome Annotation

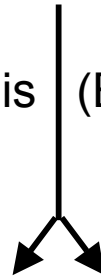
Gavin Laird

Wellcome Trust Sanger Institute

You have obtained a new sequence...

....AATCGTTGCCCAAATTACACGG....

Preliminary analysis (BLAST, etc.)



Great! My sequence has already been analysed and annotated!

But how do I access this data?

manual genome annotation
PART 2

Help! I need do to some analysis by myself.

Which tools do I use and where do I find them?

de novo sequence analysis
PART 1

```

:FTVQRRVEPKVTVYPSKTQPLQHHNLLVCSVSC 531 atggatgtggaggaagacgacttgtgtctcctgacatctctactg
:FTVQRRVQPKVTVYPAKTQPLQHHTLLVCSVNC 576 gaagagaatgaggcagtccttaccttgcagctcagaaaaggataag
:FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC 621 tccttgtctctgggagacggggaccctgatgaatttgatgagctc
:FTVQRRVEPIVTVYPAKTQPLQHHNLLVCSVNC 666 ttgatgctgatggtgatggtgagcttacacagaagaggctggc
:FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC 711 agtggagaagagggcaagactggaaacaggaggaacgtttggcc

```

1. *de novo* Sequence Analysis

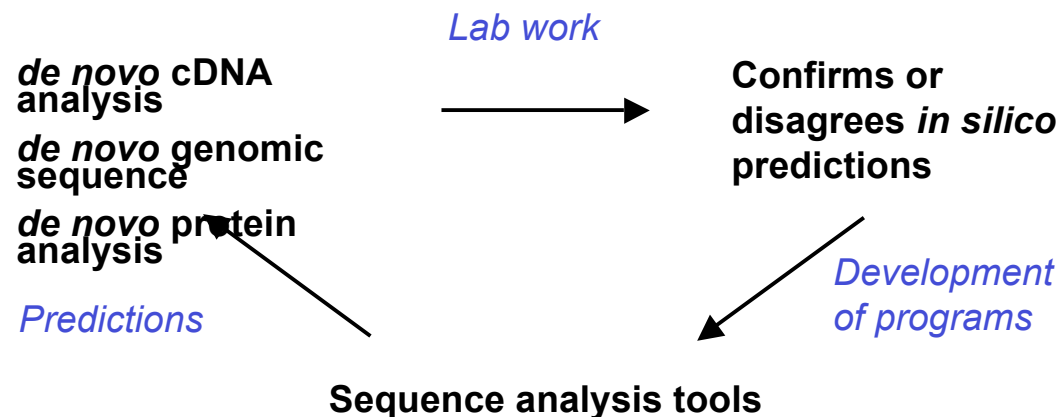
```

531 atggatgtggaggaagacgacttgtgtctcctgacatctctactg  FTVQRRVEPKVTVYPSKTQPLQHHNLLVCSVSC
M D V E E D D L C L L T S L L  FTVQRRVQPKVTVYPAKTQPLQHHTLLVCSVNC
576 gaagagaatgaggcagtccttaccttgcagctcagaaaaggataag  FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC
E E N E A V L P C S S E K D K  FTVQRRVEPIVTVYPAKTQPLQHHNLLVCSVNC
621 tccttgtctctgggagacggggaccctgatgaatttgatgagctc  :FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC
S L S L G D G D P D E F D E L  :FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC
666 ttgatgctgatggtgatggtgagcttacacagaagaggctggc  :FTVQRRVEPIVTVYPAKTQPLQHHNLLVCSVNC
F D A D G D G E S Y T E E A G  :FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC
711 agtggagaagagggcaagactggaaacaggaggaacgtttggcc  :FTVQRRVHPQVTVYPAKTQPLQHHNLLVCSVSC

```

Why do it yourself?

- Your organism may not be annotated
- You may need to customise analysis to meet your own needs



de novo Analysis Tools

- **ORF finder**
 - identifies open reading frames
- **Spidey**
 - Alignment of cDNAs to genomic sequence
- **ClustalW**
 - Multiple sequence alignments
- **Jalview & GeneDoc**
 - Edit multiple alignments: useful for phylogenetic studies

NCBI ORF finder

NCBI ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION OrFind Clear

or sequence in FASTA format

FROM: TO:

Genetic codes | 1 Standard

Comments and suggestions to: info@ncbi.nlm.nih.gov
Credits to: [Tatiana Tatusov](#) and [Roman Tatusov](#)

Frame	from	to	Length
+2	173..1201	1029	
-2	664.. 888	225	
-1	140.. 343	204	
+3	840.. 974	135	
-2	178.. 312	135	
-1	971..1093	123	
-3	1.. 113	113	

Length: 343 aa

Accept Alternative Initiation Codons

```

172 atggttcaagcaatccgggacaggggtcaaccgtggctcaaaa
   N V Q A S G H R R S T R G S K
218 atggtctctgtccgtgatagcaagatccaggaatactgag
   N Y S W S V I A K I Q E I L Q
263 agpaagatgtggagagttctgtggcagttctcaagacatat
   S K N Y R E F L A I F H S T Y
306 gtcctcagttatctggcctgttccgtggccatctgtttca
   Y S H V F G L G S V A E H V L
353 aataaaaaatatggggtacottggttcaacttgggttttggg
   S K K E G S Y L G V N L G P G
  
```

- Finds all open reading frames, starts and stops in a DNA sequence
- Graphical overview
- Integral BLASTP tool

NCBI Spidey

- Aligns spliced cDNAs, ESTs or mRNAs to genomic sequence
- Uses BLAST algorithm
- Assigns exon-intron boundaries
- Can be used for interspecies alignments

NCBI Spidey

PubMed Entrez BLAST OMIM Taxonomy Structure

Spidey FAQ
Spidey documentation
Spidey executables
Help/Contact

Spidey is an mRNA-to-genomic alignment program. For a complete description of how Spidey works, click [here](#). For an example, click [here](#).

Genomic sequence (FASTA or GI/Accession):

Upload file: Browse...

From: To:

mRNA sequence(s) (One or more FASTA or GI/Accession) [?](#):

Upload file: Browse...

divergent sequences [?](#)
 Use large intron sizes [?](#)

Align
Clear

Minimum mRNA-genomic identity %
Minimum length of mRNA covered %


ClustalW

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW Help
- ClustalW FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours
- Similar Applications
 - Align
 - Kalign
 - MAFFT
 - MUSCLE
 - T-Coffee
- ClustalW Programmatic Access

EBI > Tools > Sequence Analysis > ClustalW

ClustalW

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.
[New users, please read the FAQ.](#)
 >> [Download Software](#)



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text" value="eah@san"/>	<input type="text" value="Etnl"/>	<input type="text" value="interactive"/>	<input type="text" value="full"/>	<input type="text" value="single"/>
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="percent"/>	<input type="text" value="def"/>	<input type="text" value="def"/>
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>
OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
<input type="text" value="aln w/numbers"/>	<input type="text" value="aligned"/>	<input type="text" value="none"/>	<input type="text" value="off"/>	<input type="text" value="off"/>

Enter or Paste a set of Sequences in any supported format:

```
>H2-Eb1
MVMLRPRVPCVAAVILLTLLVLSPPVALVRDSRPFWLEYCKSECHFYNGTQRVLLERYFYN
LEENLRFSDVGEFRAVTELRGPDAAENMNSQPEFLEQKRAEVDVTCRHNYEISDKFLVRR
RVEPTVTVYPTKTQPLEHNNLLVCSVSDFYPGNIEVRFWRNGKEEKTGI VSTGLVRNGDW
TFQTLVHLETVPQSGEYVTCQVEHPSLTDPVTVENKAQSTSAQNKMLSGVGGFVLGLLFL
GAGLFYFRNQKGGQSGLQPTGLLS*

>H2-Eb1-like
MVSLLWLRGLCVAAVILSLMLTPPVILVRDPRPRFLEQLKAECHYFNGKERWWSVTRFI
YNQEEFARFNSDFGKFLAVTELRGPFIVEYLNQKMDLNYRASVDRCRNNDLVDIFMLN
LKAEPKVTWYPAKTQPLEHNNLLVCSVIDFYPGSIEVRFWRNGEKEKTGVVSTGLIQNRD
```

Upload a file:

- DNA or protein alignments
- Cut and paste sequences or upload files
- Obtain results via email or interactively

ClustalW

CLUSTAL W (1.83) multiple sequence alignment

```

swall|Q8CGP5|H2A1F_MOUSE      SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYSERVGAGAPVY  50
swall|P0C0S9|H2A1_BOVIN      SGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVY  50
swall|P0C169|H2A1C_RAT       SGRGKQGGKARAKAKSRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVY  50
swall|Q96QV6|H2A1A_HUMAN     SGRGKQGGKARAKSKSRSSRAGLQFPVGRVHRLLRKGNYAERIGAGAPVY  50
*****;*;*****;*****;*****;*****

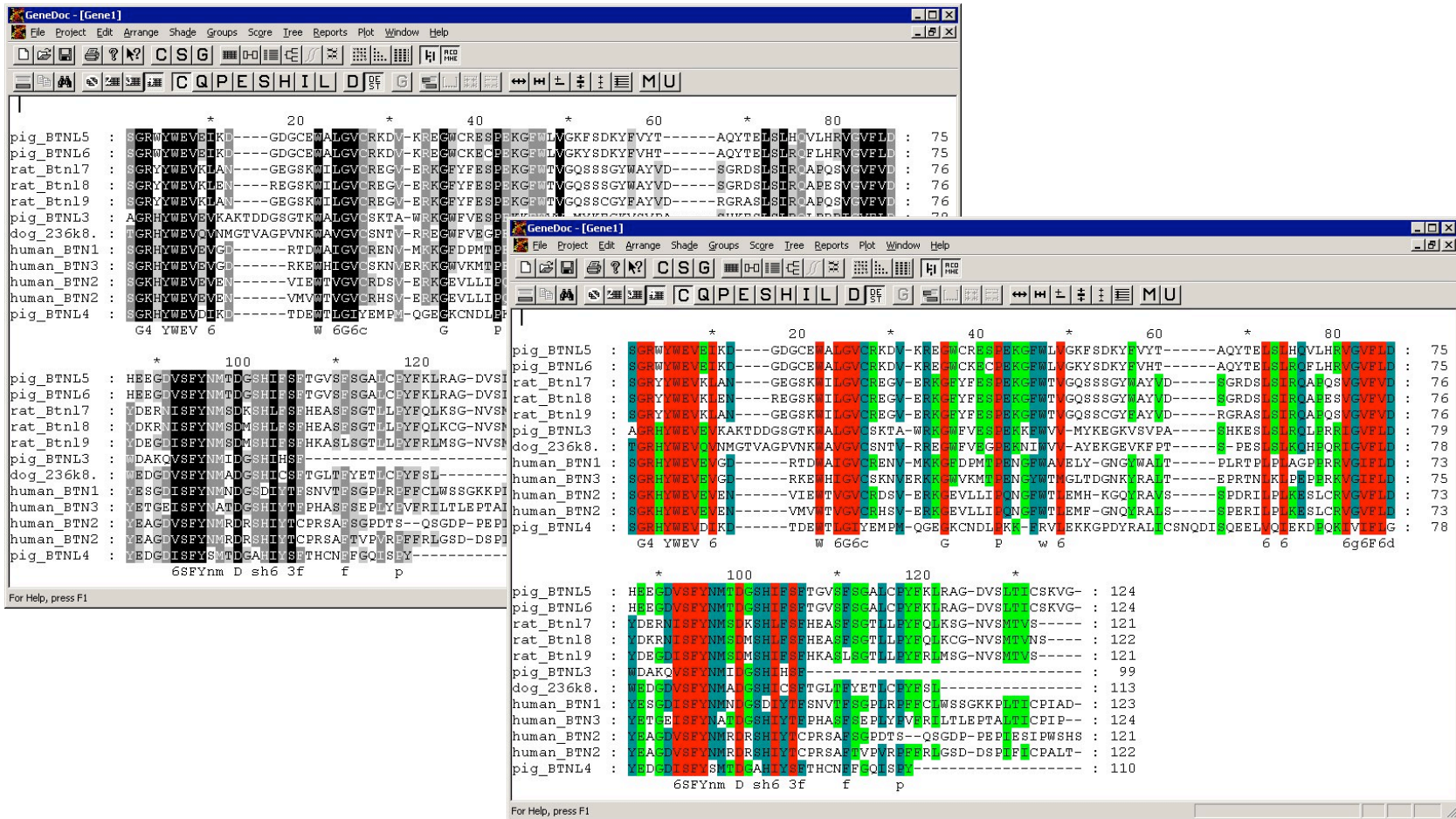
swall|Q8CGP5|H2A1F_MOUSE      LAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLGRV  100
swall|P0C0S9|H2A1_BOVIN      LAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLGKV  100
swall|P0C169|H2A1C_RAT       LAAVLEYLTAEILELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLGRV  100
swall|Q96QV6|H2A1A_HUMAN     LAAVLEYLTAEILELAGNASRDNKKTRIIPRHLQLAIRNDEELNKLLGGV  100
*****;*****

```

- Alignments may be viewed and edited in Jalview, GeneDoc and other programs
 - Maximise similarities within columns
- Starting point for phylogenetic analyses

GeneDoc

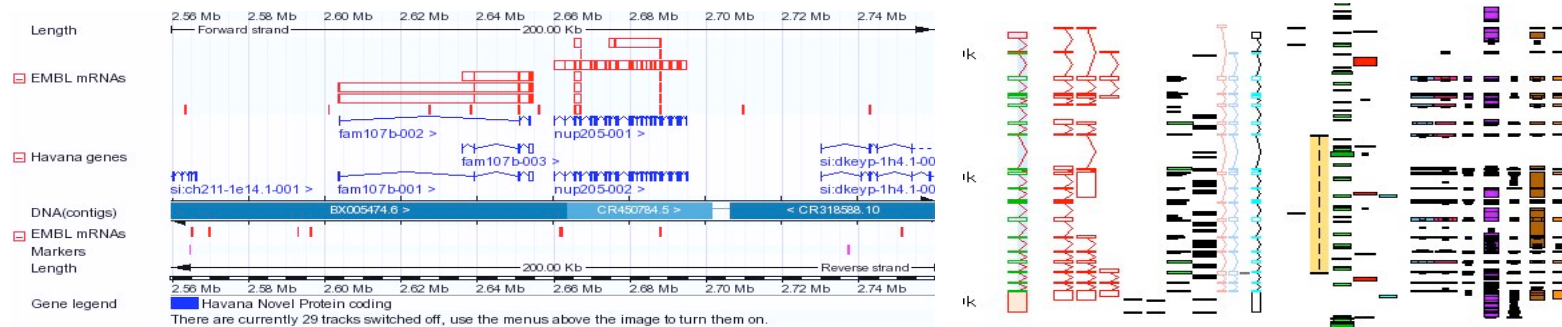
Alignment viewer and editor



Summary

***de novo* analysis of sequence**

- Hunt for open reading frames: [ORF Finder](#)
- Align cDNA to genomic DNA: [Spidey](#)
- Multiple sequence alignments: [ClustalW](#)
- Edit alignments: Jalview, [GeneDoc](#)



2. Manual Genome Annotation

Transcript Name	ogt1 (ZFIN ID) (to view all Vega genes linked to the name click here)
Transcript information	Exon: 23 Transcript length: 3,859 bp Protein length: 1,132 residues Further Transcript info Epub information
Transcript Class	Known, CDS Definition
InterPro	IPR011717 Tetraosaccharyl TFR_4 View other genes with this domain IPR014445 Tetraosaccharyl TFR_1 View other genes with this domain IPR011105 Tetraosaccharyl TFR_2 View other genes with this domain IPR013028 Tetraosaccharyl region View other genes with this domain
Transcript structure	
Protein features	
Curated Locus	ogt1 (ZFIN ID) (to view all Vega genes linked to the name click here)
Author	This locus was annotated by Havana < vega@sanger.ac.uk >
Locus ID	OTTDARG0000020997
Genomic Location	This gene can be found on Chromosome 14 at location 7,935,304-7,953,328 . The start of this gene is located in Contig BX323628.7.1.206011 .
Gene Type	Known Protein coding Definition
Version & Date	Version 2 Gene last modified on 26/10/2006 (Created on 25/10/2006)
Description	O-linked N-acetylglucosamine (GlcNAc) transferase (UDP-N-acetylglucosamine:polypeptide-N-acetylglucosaminyl transferase) like
Database Matches	This Vega gene corresponds to the following database identifiers: ZFIN: ogt1
Curation Method	Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases as well as a series of ab initio gene predictions (GENSCAN, Fgenes). In addition, comparative analysis using vertebrate datasets is used to aid novel gene discovery. The data gathered in these steps is then used to manually annotate the clone adding gene structures, descriptions and poly-A features. The annotation is based on supporting evidence only.

Automatic annotation

- Fast
 - Unfinished sequence or shotgun sequence
 - Consistent
 - Under/Over-prediction
-




Manual annotation

- Slow
 - Finished sequence
 - Flexible – can deal with inconsistencies
 - Consult publications
-




You may also encounter *ab initio* methods:
genescan, fgenesh

Vertebrate Genome Annotation Database


Home

🔍

[Login / Register](#) | [Docs & FAQs](#)



The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality manual annotation of vertebrate finished genome sequence. Human, mouse and zebrafish are in the process of being completely annotated, whereas for other species the annotation is only of specific genomic regions of particular biological interest. The majority of the annotation is from the [HAVANA](#) group at the [Welcome Trust Sanger Institute](#)


The website is built upon code from the [Ensembl](#) project.


Search Vega


Search: All species for Go


e.g. human gene BRCA2 or mouse X:100000..200000


Browse a genome [\(Log in to reorder this list\)](#)


- 


Human [30-03-2009]
Ensembl
- 

Mouse [30-03-2009]
Ensembl
- 

Zebrafish [30-03-2009]
Ensembl
- 

Gorilla [30-03-2009]
Ensembl
- 

Wallaby [30-03-2009]
- 

Pig [16-05-2007]
Ensembl
- 

Dog [14-02-2005]
Ensembl

What's New in Release 36 (10 July 2009)

- [Schema change](#) (all species)

[More release 36 news...](#)

What's New in Release 35 (30 March 2009)

- [Update to human](#) (Human)
- [Update to mouse including new IDD region](#) (Mouse)
- [Update to zebrafish](#) (Zebrafish)
- [Gorilla annotation](#) (Gorilla)
- [Tamar wallaby annotation](#) (Wallaby)

[More release 35 news...](#)

What's New in Release 34 (18 December 2008)

- [Update to mouse](#) (Mouse)
- [Update to zebrafish](#) (Zebrafish)
- [Addition of new IDD region](#) (Mouse)

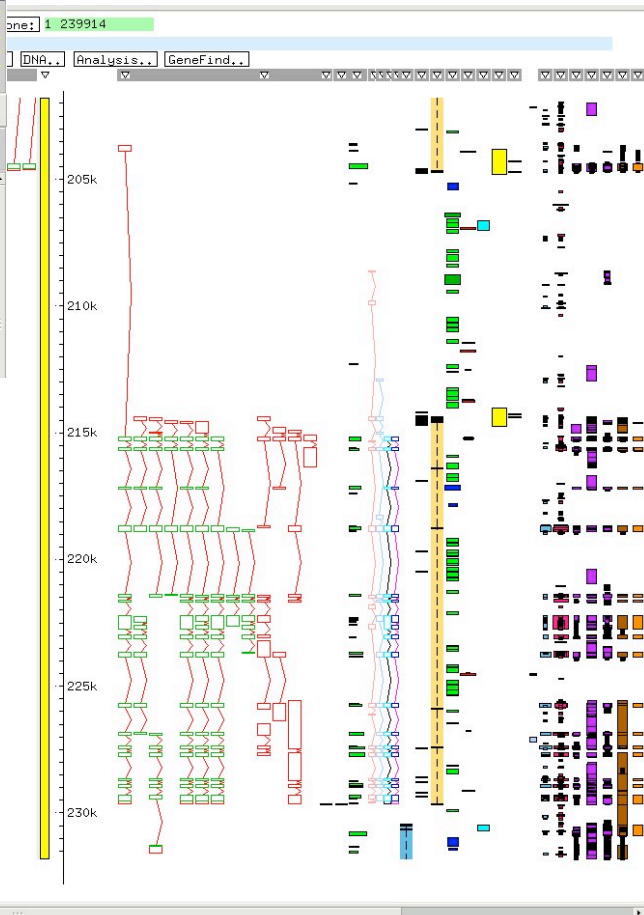
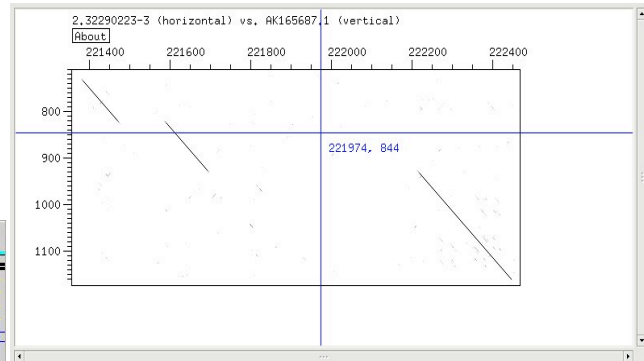
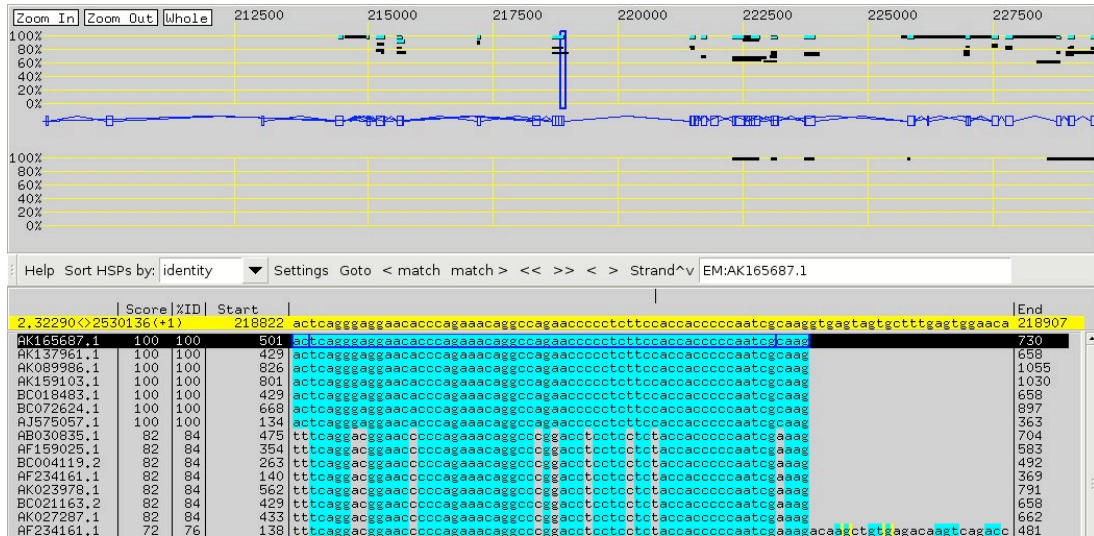
[More release 34 news...](#)

Vega Genome Browser release 36 - July 2009
 © 2009 [WTS](#)

<http://vega.sanger.ac.uk>

[Contact Us](#) | [Help](#)

All annotation is supported by a combination of cDNA, EST and/or protein evidence



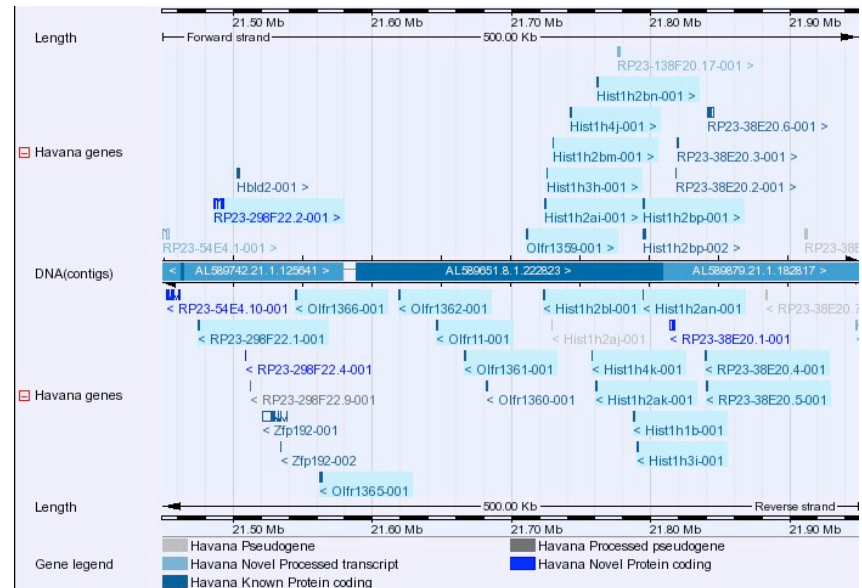
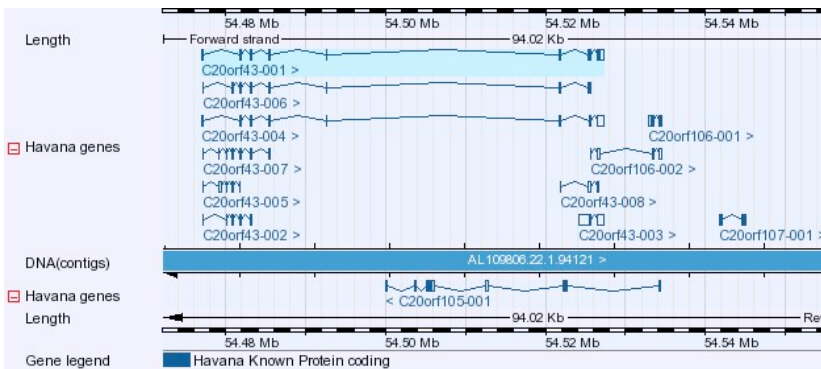
Locus Categories

- Known
- Novel_CDS
- Novel transcript
- Putative
- Pseudogene

Manual annotation is advantageous for:

- Overlapping genes
- Alternative splicing

- Pseudogenes
- Duplications/gene clusters



- Non-coding genes
- Complex loci

- *Anything out of the ordinary*

Summary

Manual genome annotation

- Labour intensive, but the final product is reliable and accurate
- ‘Gold standard’ rating
- VEGA: central repository for manual vertebrate genome annotation that is easy to browse