Module 2

The Vega and UCSC Genome Browsers

Using Web Browsers to View Genome Annotation

Jane Loveland PhD
Wellcome Trust Sanger Institute
Hinxton, UK

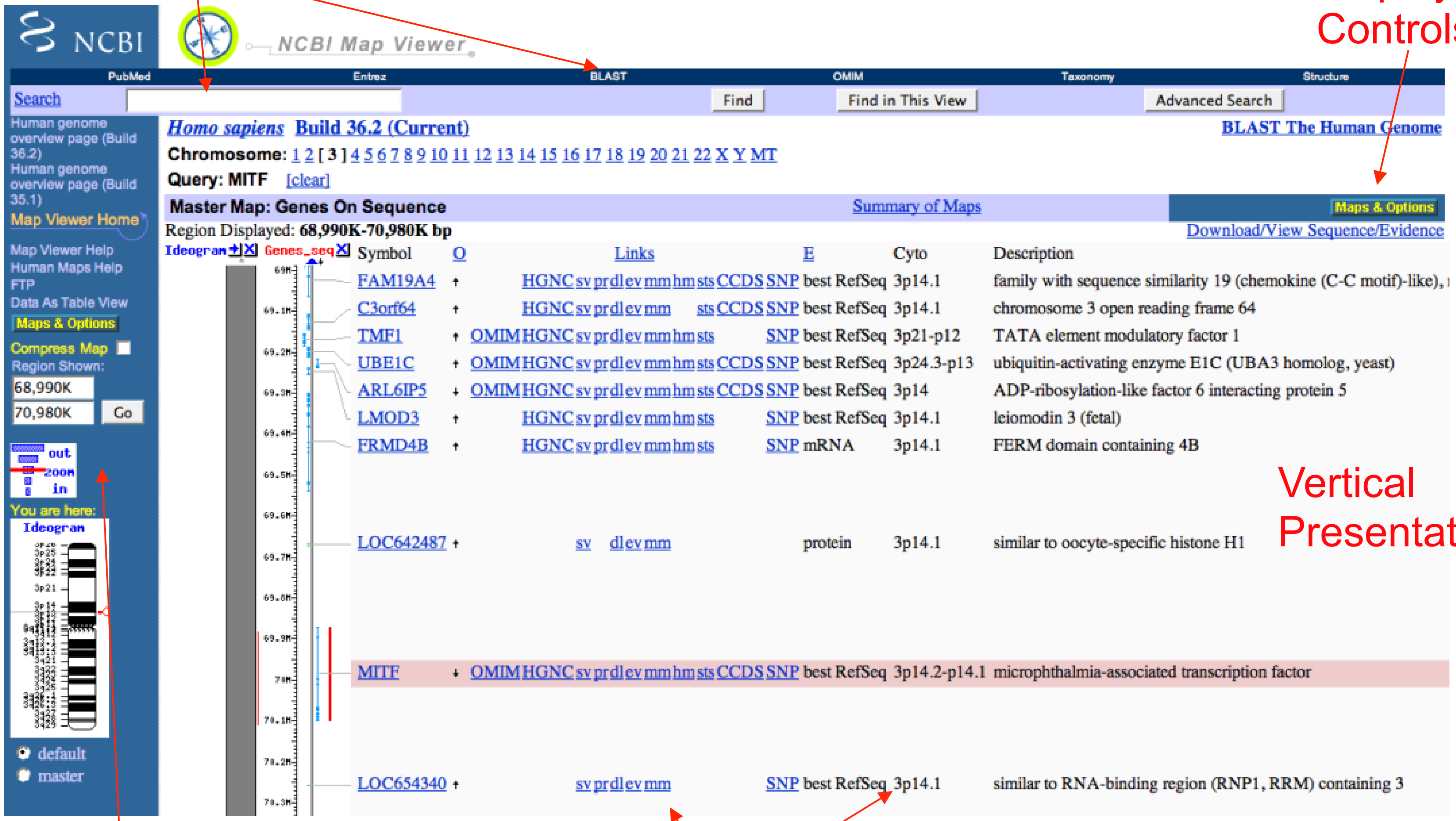
Genome Browsers / Gene Sets

NCBI	http://www.ncbi.nlm.nih.gov/RefSeq/
e!Ensembl	http://www.ensembl.org
Vega	http://vega.sanger.ac.uk
UCSC	http://genome.ucsc.edu/
CCDS Database	http://www.ncbi.nlm.nih.gov/projects/CCDS/ CcdsBrowse.cgi (CDS only)

NCBI Map Viewer

Searching

Display Controls



Homo sapiens Build 36.2 (Current)

BLAST The Human Genome

Chromosome: 1 2 [3] 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Query: MITF [clear]

Master Map: Genes On Sequence

Region Displayed: **68,990K-70,980K bp**

Region Shown: 68,990K 70,980K

Symbol	O	Links	E	Cyto	Description
FAM19A4	↑	HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14.1	family with sequence similarity 19 (chemokine (C-C motif)-like),
C3orf64	↑	HGNC sv pr dl ev mm sts CCDS SNP	best RefSeq	3p14.1	chromosome 3 open reading frame 64
TMF1	↑	OMIM HGNC sv pr dl ev mm hm sts SNP	best RefSeq	3p21-p12	TATA element modulatory factor 1
UBE1C	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p24.3-p13	ubiquitin-activating enzyme E1C (UBA3 homolog, yeast)
ARL6IP5	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14	ADP-ribosylation-like factor 6 interacting protein 5
LMOD3	↑	HGNC sv pr dl ev mm hm sts SNP	best RefSeq	3p14.1	leiomodin 3 (fetal)
FRMD4B	↑	HGNC sv pr dl ev mm hm sts SNP	mRNA	3p14.1	FERM domain containing 4B
LOC642487	↑	sv dl ev mm	protein	3p14.1	similar to oocyte-specific histone H1
MITF	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3p14.2-p14.1	microphthalmia-associated transcription factor
LOC654340	↑	sv pr dl ev mm SNP	best RefSeq	3p14.1	similar to RNA-binding region (RNP1, RRM) containing 3

Vertical Presentation

Navigation

“Maps”

NCBI Map Viewer

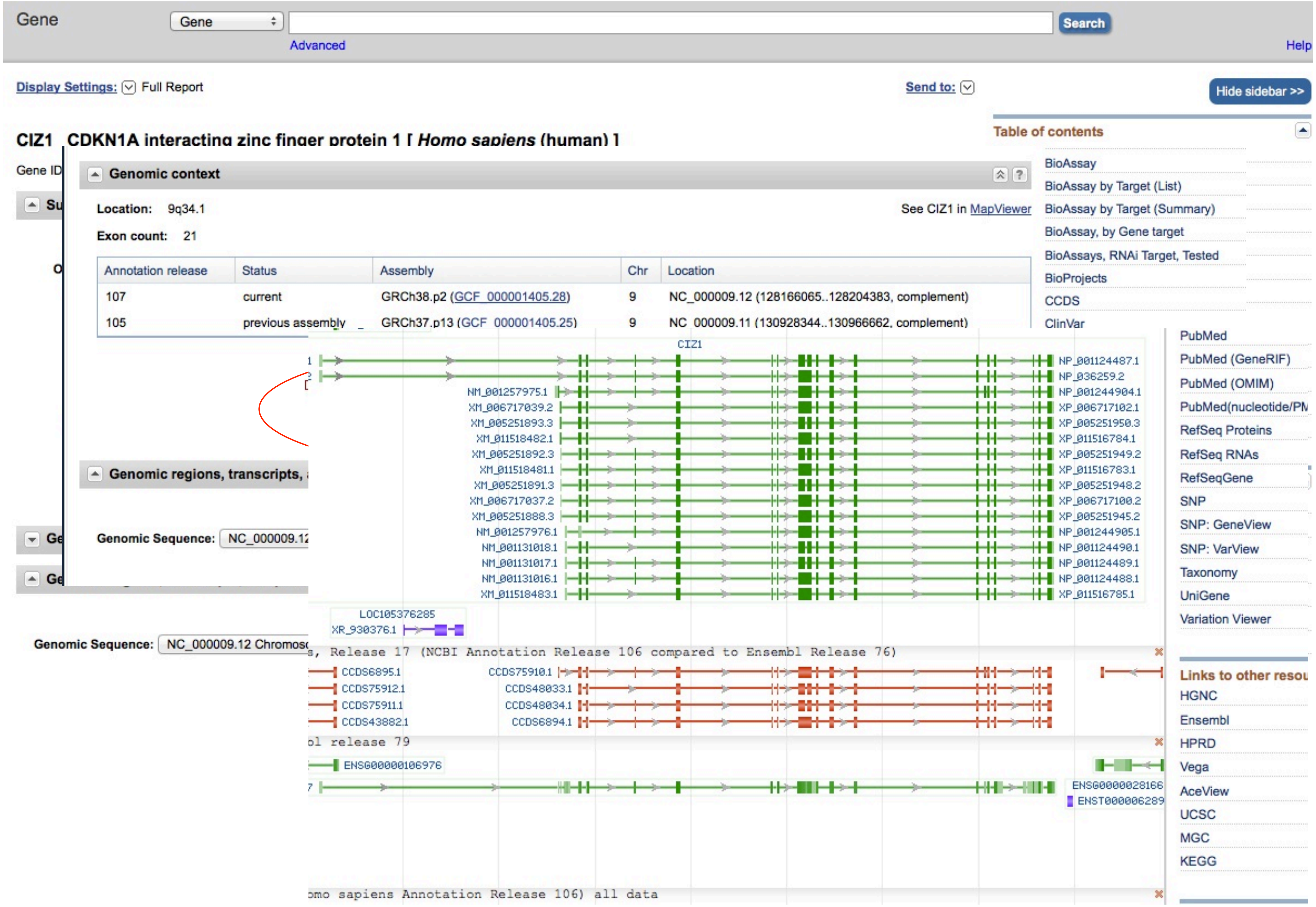
- Excellent integration with other NCBI resources
- Best “map” views of non-sequence maps (i.e. clone maps, genetic maps)
- Includes Celera assembly, alternate haplotypes, assemblies of everything available
- BLAST for sequence searching

NCBI - RefSeq

- Non-redundant gene set
- Accessed via browsers or Entrez Gene
- Accessions for genomic DNA, transcripts and proteins
- Primarily protein-coding
- Semi-curated

	Automated	Curated
Genomic	NC_12345	
mRNA	XM_12345	NM_12345
ncRNA	XR_12345	NR_12345
Protein	XP_12345	NP_12345

http://www.ncbi.nlm.nih.gov/gene



Display Settings: Full Report

Send to:

CIZ1 CDKN1A interacting zinc finger protein 1 [*Homo sapiens* (human)]

Gene ID

Genomic context

Location: 9q34.1

See CIZ1 in MapViewer

Exon count: 21

Annotation release	Status	Assembly	Chr	Location
107	current	GRCh38.p2 (GCF_000001405.28)	9	NC_000009.12 (128166065..128204383, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	9	NC_000009.11 (130928344..130966662, complement)

Genomic regions, transcripts,

Genomic Sequence: NC_000009.12

Genomic Sequence: NC_000009.12 Chromos

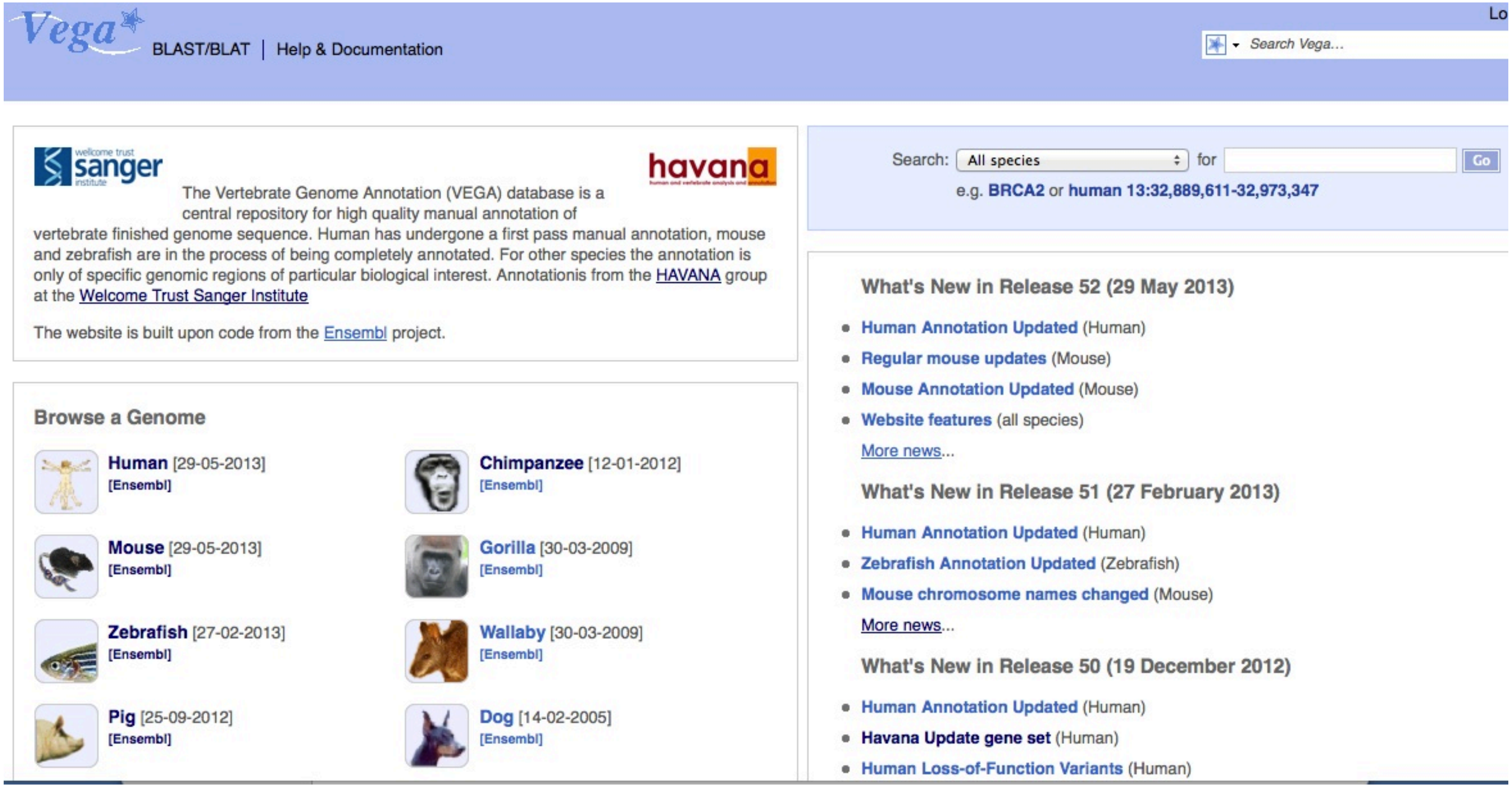
Table of contents

BioAssay
BioAssay by Target (List)
BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Tested
BioProjects
CCDS
ClinVar
PubMed
PubMed (GeneRIF)
PubMed (OMIM)
PubMed(nucleotide/PM
RefSeq Proteins
RefSeq RNAs
RefSeqGene
SNP
SNP: GeneView
SNP: VarView
Taxonomy
UniGene
Variation Viewer

Links to other resou

HGNC
Ensembl
HPRD
Vega
AceView
UCSC
MGC
KEGG

Vertebrate Genome Annotation Database



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

The website is built upon code from the Ensembl project.

Search: All species for Go

e.g. BRCA2 or human 13:32,889,611-32,973,347

Browse a Genome

Human [29-05-2013] [Ensembl]	Chimpanzee [12-01-2012] [Ensembl]
Mouse [29-05-2013] [Ensembl]	Gorilla [30-03-2009] [Ensembl]
Zebrafish [27-02-2013] [Ensembl]	Wallaby [30-03-2009] [Ensembl]
Pig [25-09-2012] [Ensembl]	Dog [14-02-2005] [Ensembl]

What's New in Release 52 (29 May 2013)

- Human Annotation Updated (Human)
- Regular mouse updates (Mouse)
- Mouse Annotation Updated (Mouse)
- Website features (all species)

More news...

What's New in Release 51 (27 February 2013)

- Human Annotation Updated (Human)
- Zebrafish Annotation Updated (Zebrafish)
- Mouse chromosome names changed (Mouse)

More news...

What's New in Release 50 (19 December 2012)

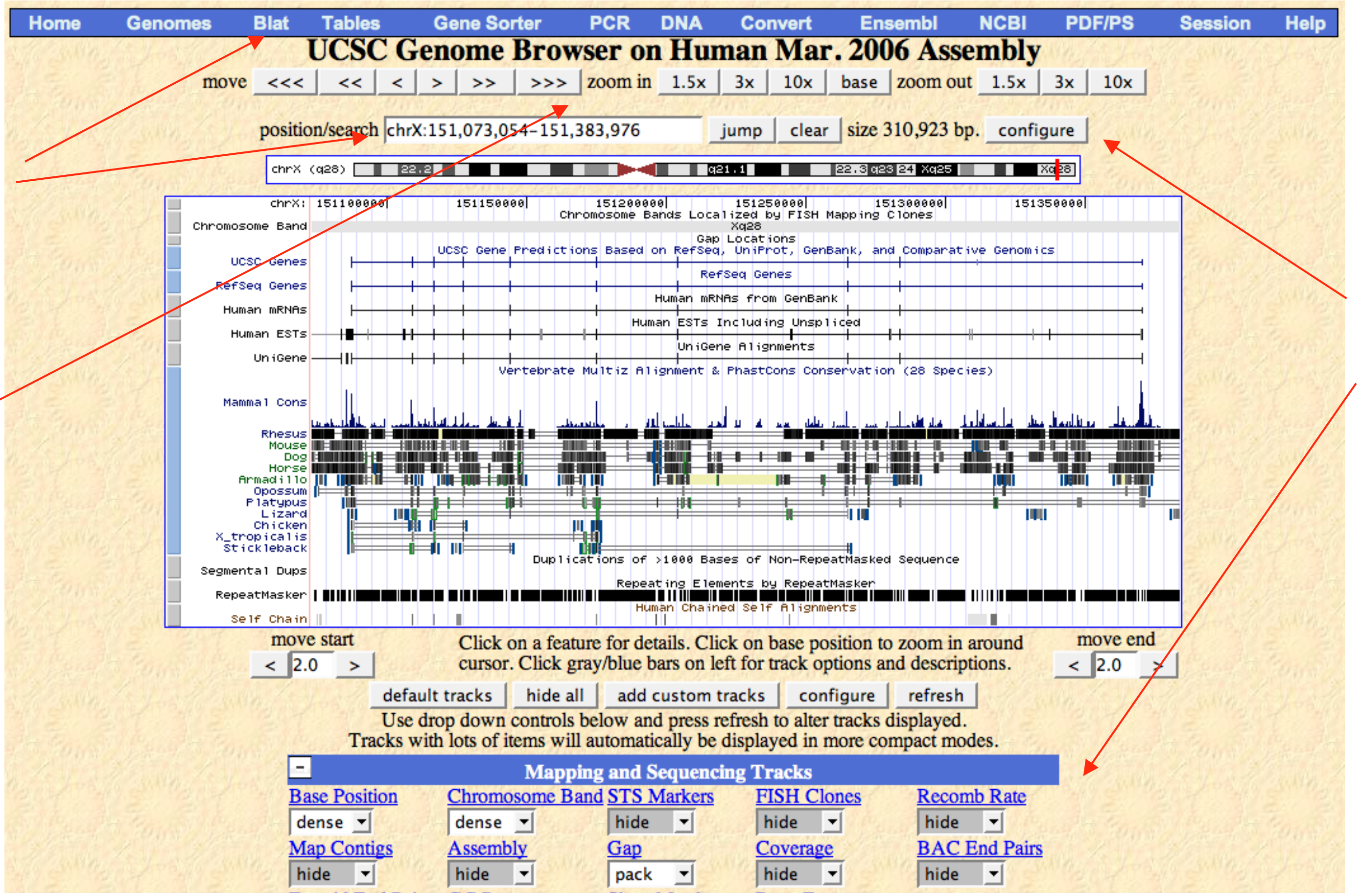
- Human Annotation Updated (Human)
- Havana Update gene set (Human)
- Human Loss-of-Function Variants (Human)

http://vega.sanger.ac.uk

VEGA gene set

- Manually annotated using Otterlace/Zmap annotation software
- Based on direct pairwise alignment of mRNA, EST and protein evidence (including cross-species)
- Multiple biotypes, reflect confidence levels
- Includes additional data sources as DAS tracks (eg. CAGE tags, RNAseq)

UCSC Genome Browser



UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chrX:151,073,054-151,383,976 size 310,923 bp.

Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position	Chromosome Band	STS Markers	FISH Clones	Recomb Rate
dense	dense	hide	hide	hide
Map Contigs	Assembly	Gap	Coverage	BAC End Pairs
hide	hide	pack	hide	hide

Searching

Navigation

Display Controls

Annotations called "tracks"

UCSC Genome Browser

- Straightforward display, easy navigation
- Third-party annotations
- Evolutionary conservation
- “Wiggle” tracks for continuous data
- Fast sequence searching with BLAT
- View your own data
- ENCODE annotations

UCSC gene set

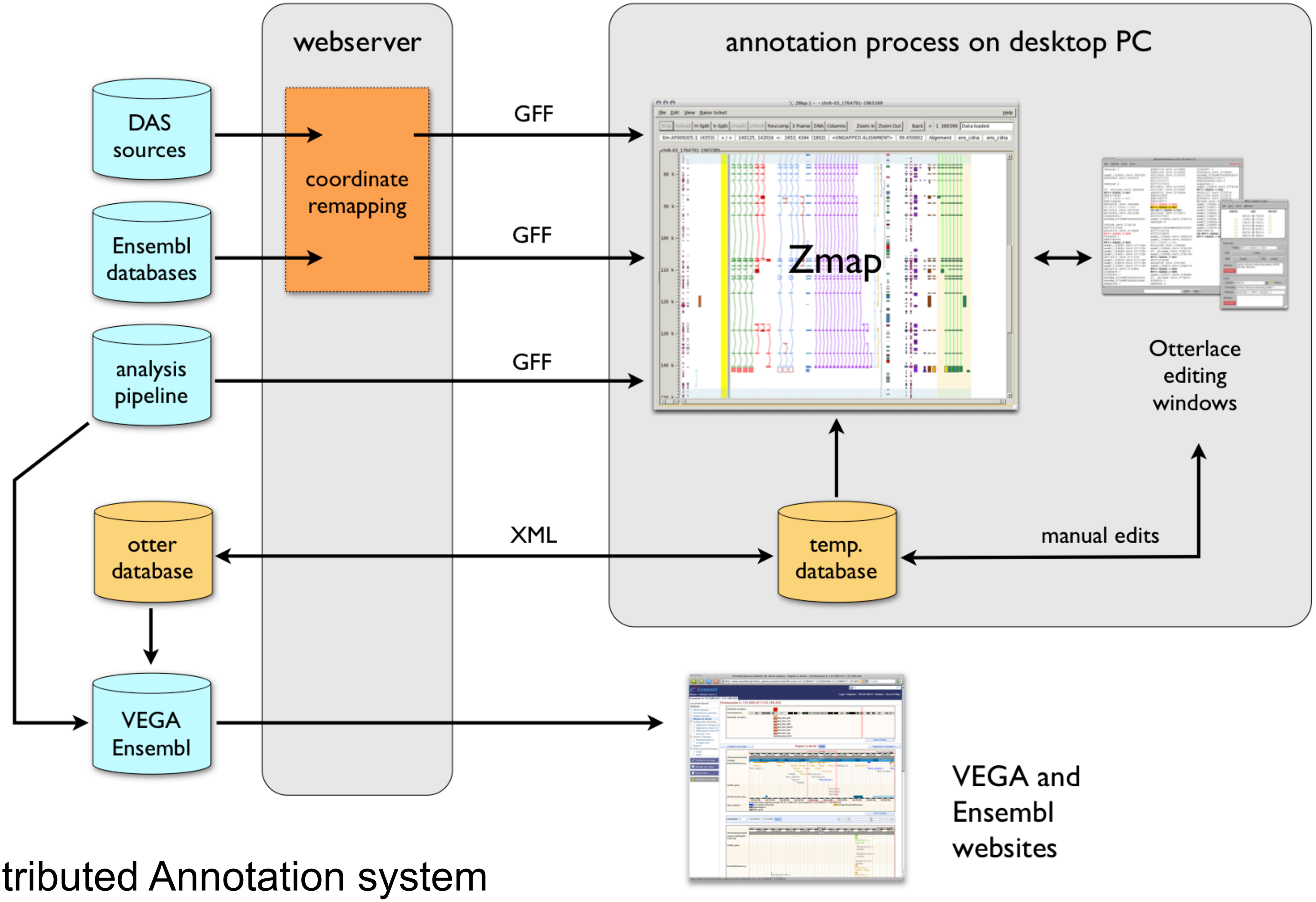
- Non-redundant gene set
- Automatic annotation based on BLAT alignments
- Transcripts require Genbank accession plus one other supporting feature (eg. Uniprot)
- Includes RefSeq models (require no additional support)
- Both protein-coding and non-coding
- Data hub for ENCODE data, displays GENCODE geneset

http://genome.ucsc.edu/

Common Functionality

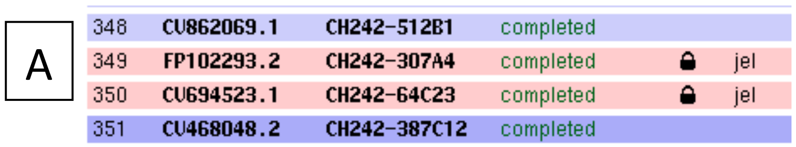
- Navigational tools
  - Searching for markers by name or sequence (BLAST, BLAT)
  - Zooming in and out
  - Choose annotations to display
- Download of annotations
  - Whole genome
  - Specific regions
- Links to other resources
- View own data in browser

Manual Analysis and Annotation pipeline: Otterlace

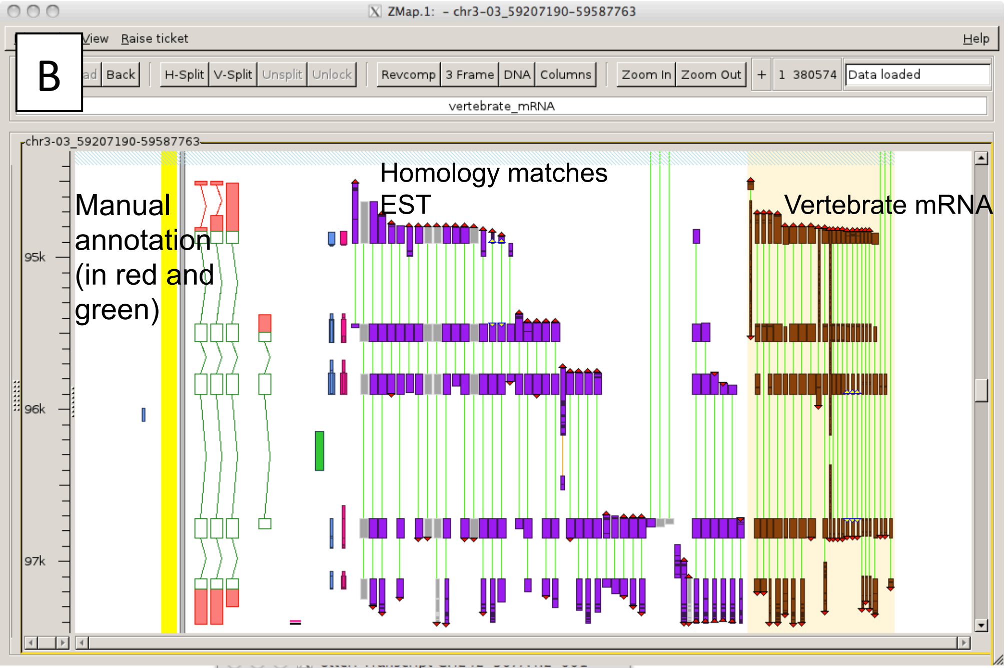


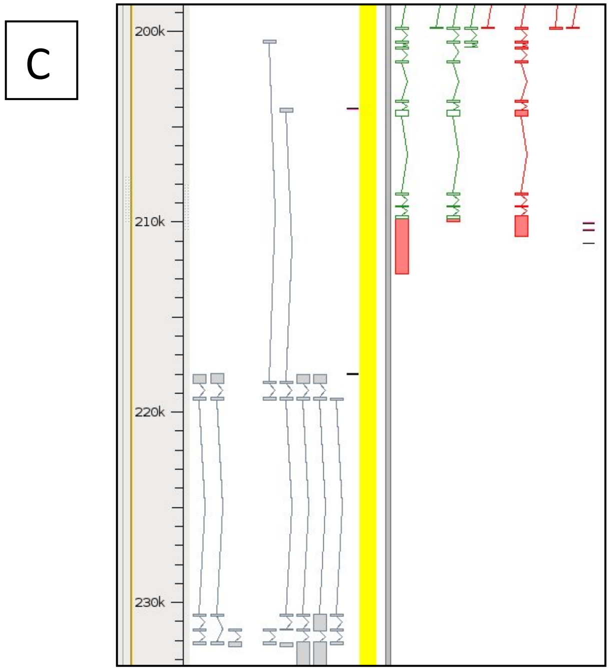
DAS=Distributed Annotation system

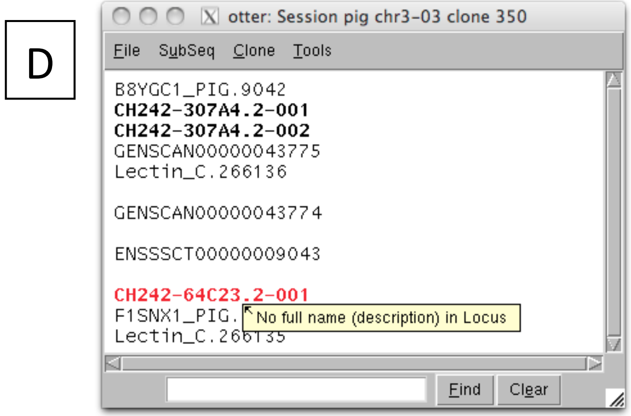
All annotation is supported by a combination of cDNA, EST and/or protein evidence



348	CU862069.1	CH242-512B1	completed		
349	FP102293.2	CH242-307A4	completed	🔒	jel
350	CU694523.1	CH242-64C23	completed	🔒	jel
351	CU468048.2	CH242-387C12	completed		







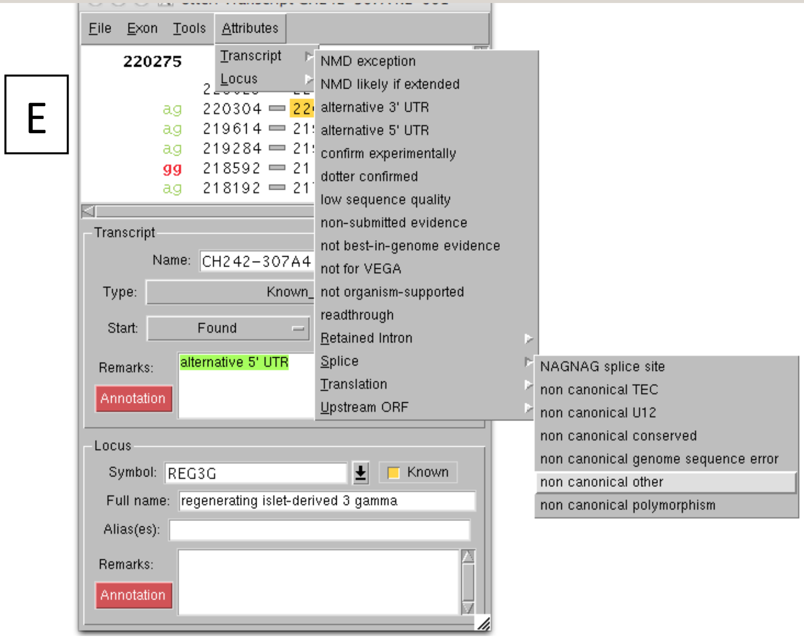
```
B8YGC1_PIG.9042
CH242-307A4.2-001
CH242-307A4.2-002
GENSCAN00000043775
Lectin_C.266136

GENSCAN00000043774

ENSSSCT00000009043

CH242-64C23.2-001
F1SNX1_PIG.
Lectin_C.266135
```

No full name (description) in Locus



220275

ag	220304	22
ag	219614	21
ag	219284	21
gg	218592	21
ag	218192	21

Transcript

Name: CH242-307A4

Type: Known

Start: Found

Remarks: alternative 5' UTR

Locus

Symbol: REG3G ☐ Known

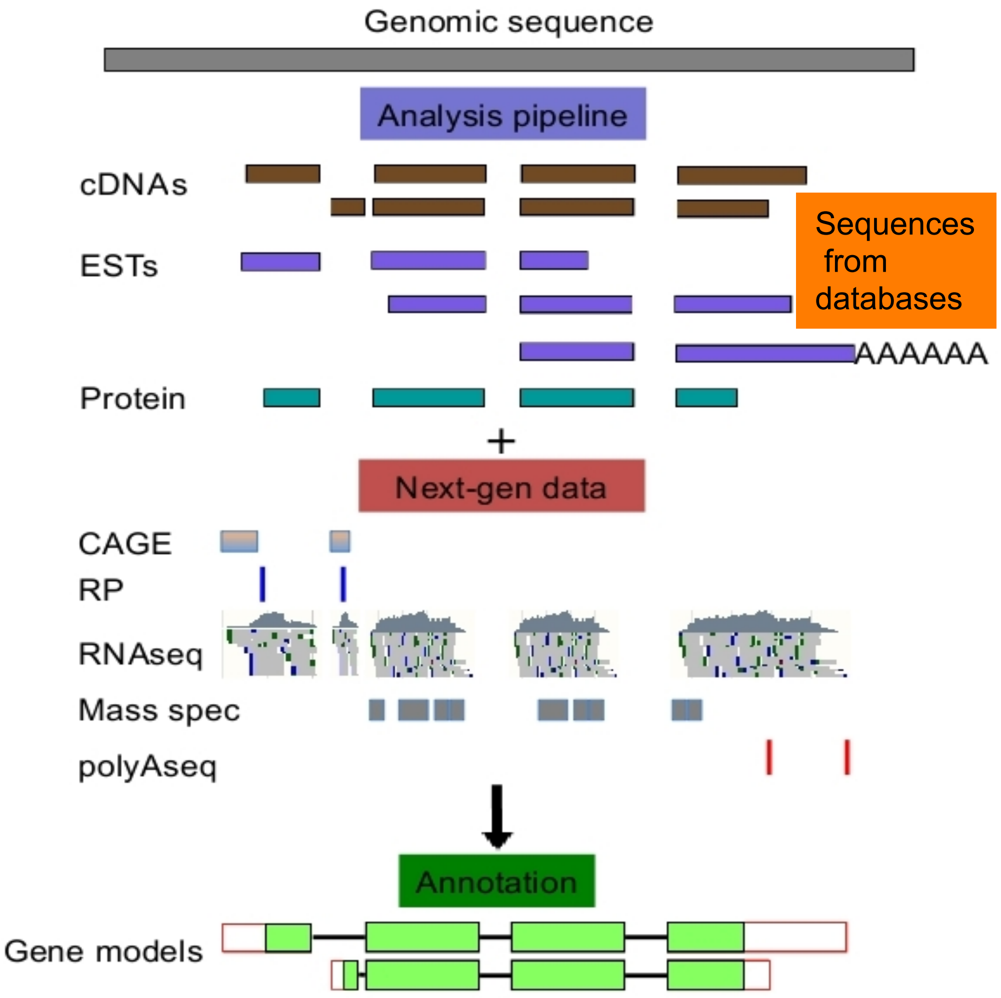
Full name: regenerating islet-derived 3 gamma

Alias(es):

Remarks:

Manual Annotation and Biotypes:

Annotation:
based on transcriptional evidence



Biotypes

Protein Coding
Known_CDS
Novel_CDS
Putative_CDS
Nonsense_mediated_decay

Transcript retained intron
putative

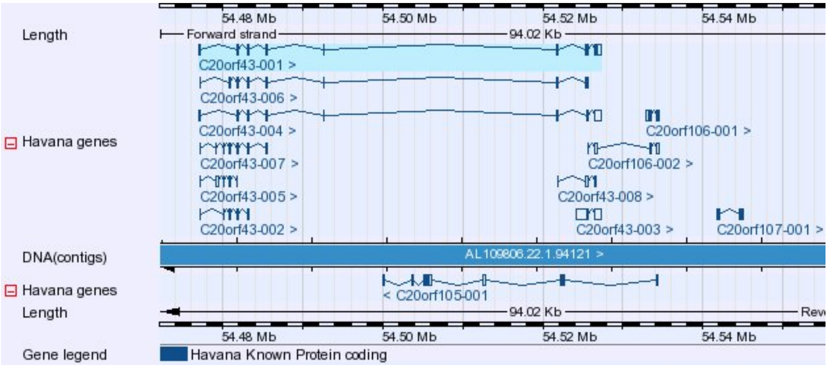
Non-coding lincRNA
Antisense
Sense_intronic
Sense_overlapping
3’_overlapping_ncRNA

Pseudogene
Processed
Unprocessed
Transcribed
Translated
Unitary
Polymorphic

Immunoglobulin
IG_pseudogene
IG_Gene
TR_Gene

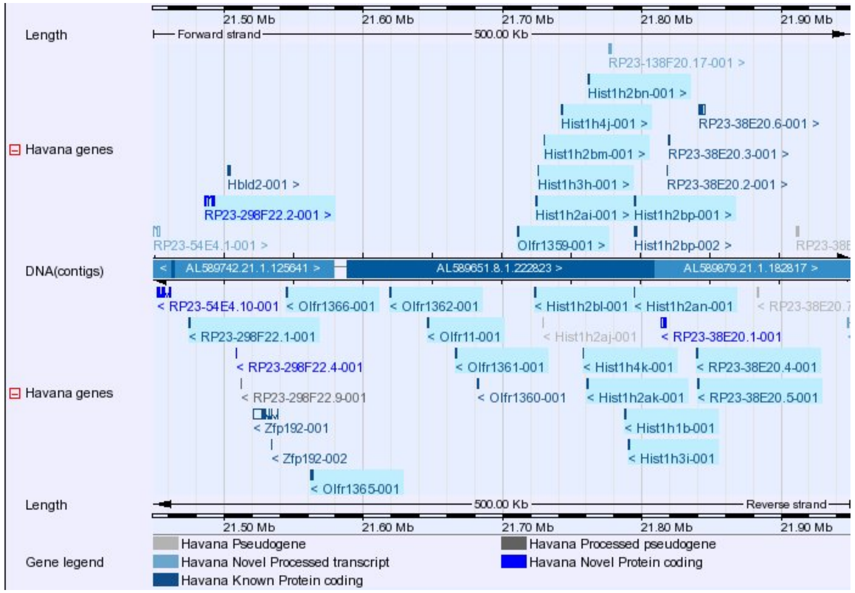
Manual annotation is advantageous for:

- Overlapping genes
- Alternative splicing



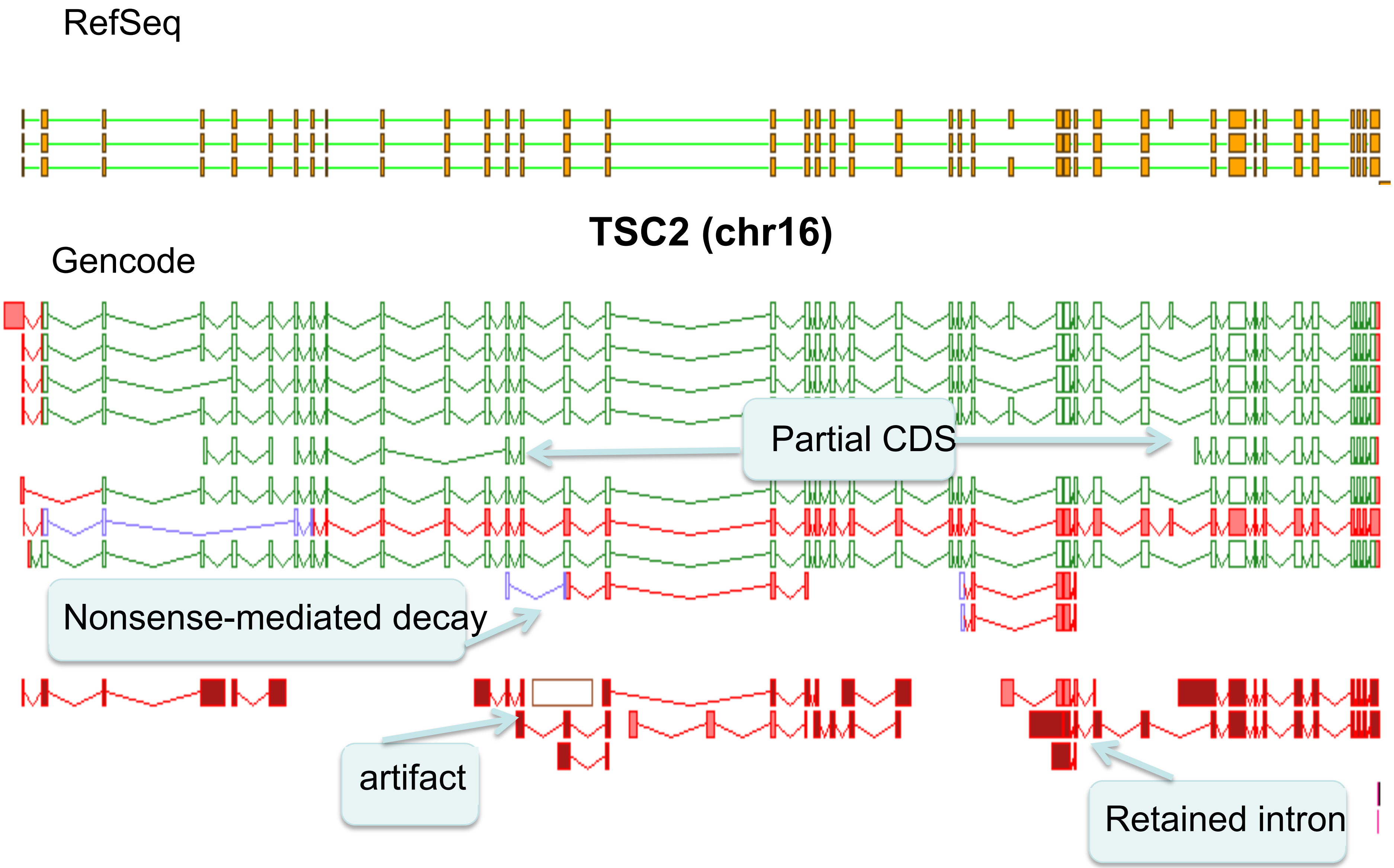
- Non-coding genes
- Complex loci e.g *GNAS*
- RNA seq data

- Pseudogenes
- Duplications/gene clusters



- *Anything out of the ordinary*

Classifying Functional Transcripts within a protein coding gene:



RefSeq

TSC2 (chr16)

Gencode

Partial CDS

Nonsense-mediated decay

artifact

Retained intron

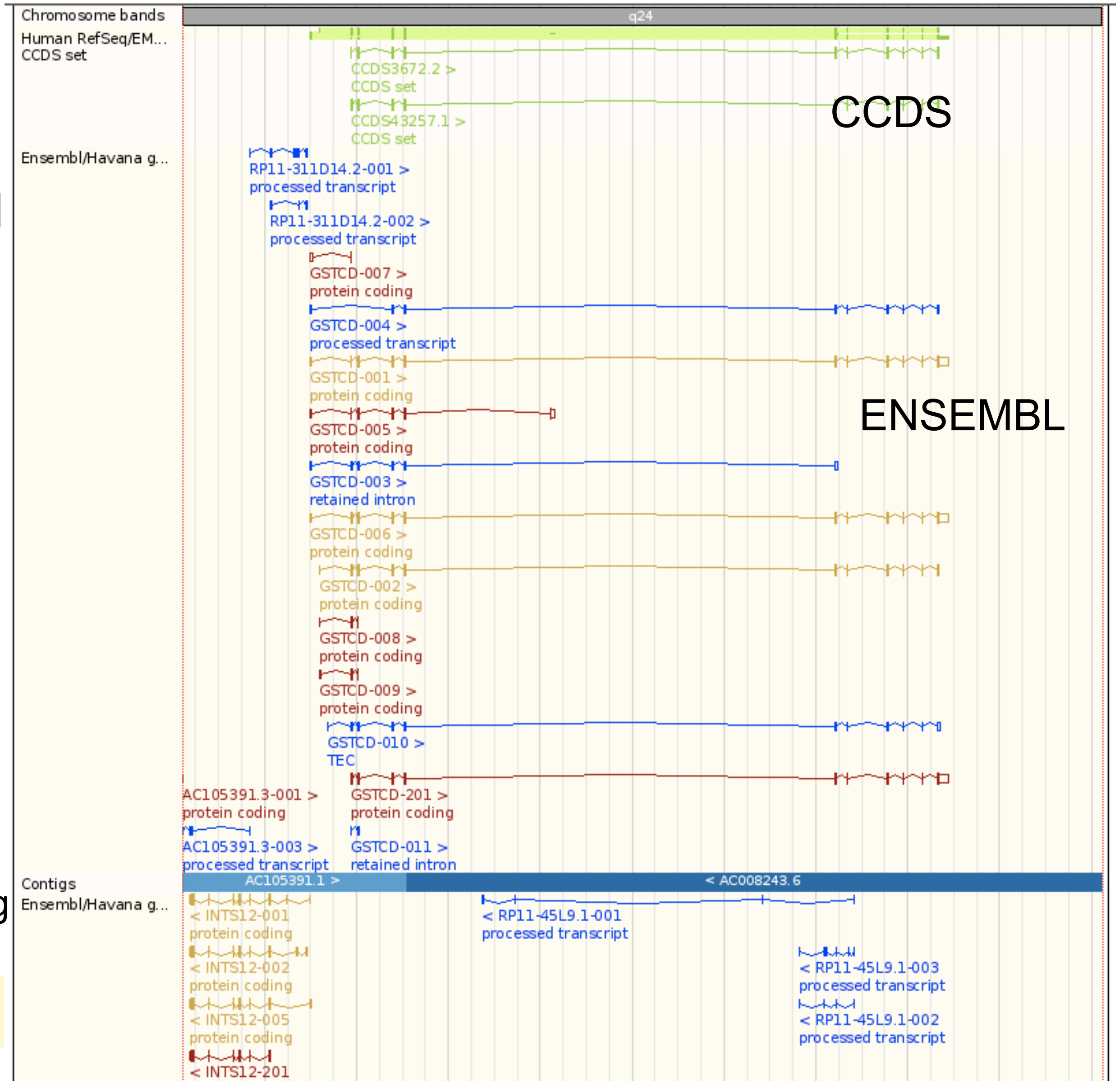
Ensembl view: Gencode geneset

Gold (merged): agreed ensembl/havana

Red: coding (001 Havana, 201 Ensembl)

Blue: non-coding

http://www.gencodegenes.org



Worked Example