IMGC 2014
Bioinformatics Workshop

Laurens Wilming – Mark Thomas – Terry Meehan
lw2@sanger.ac.uk

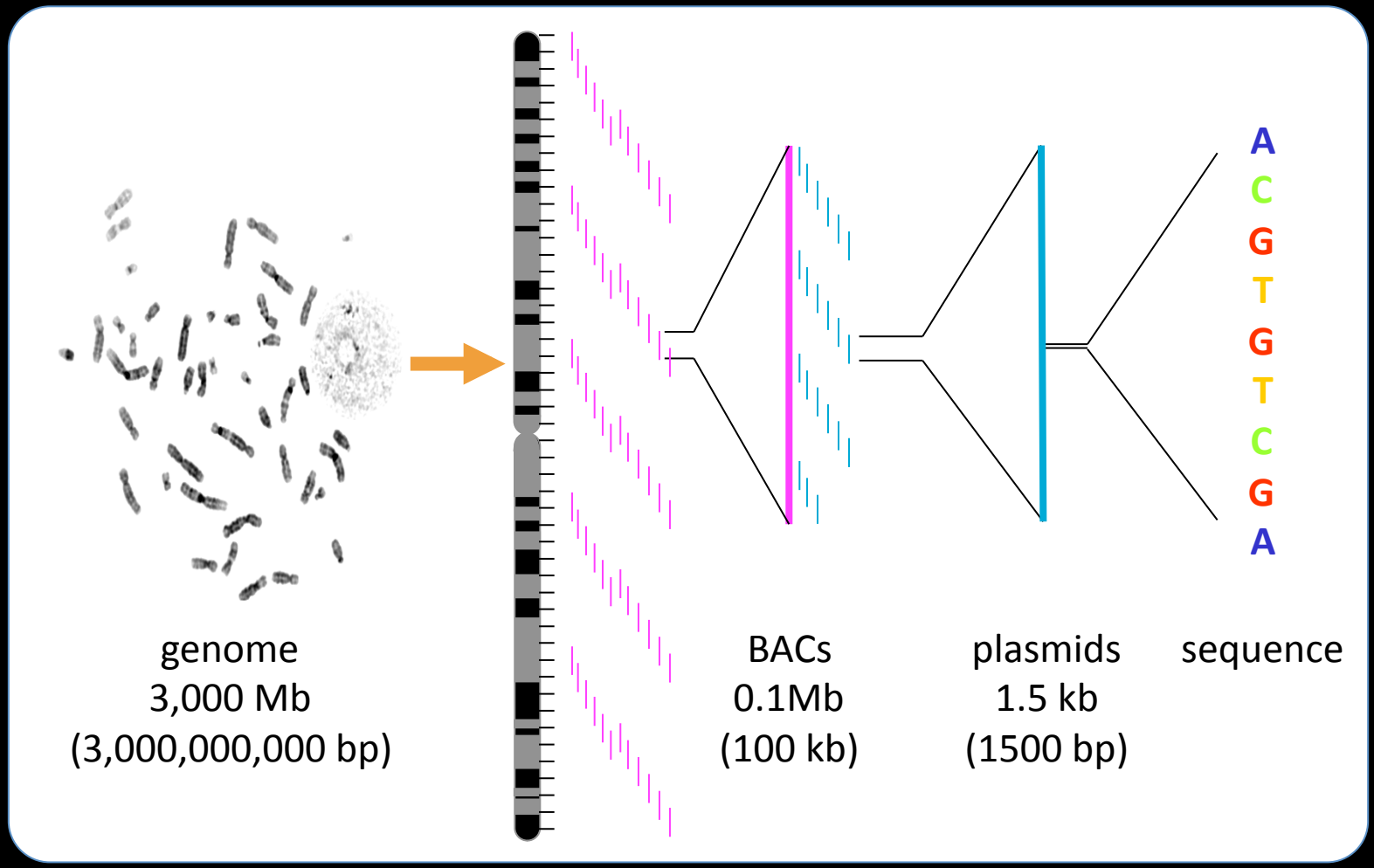
wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

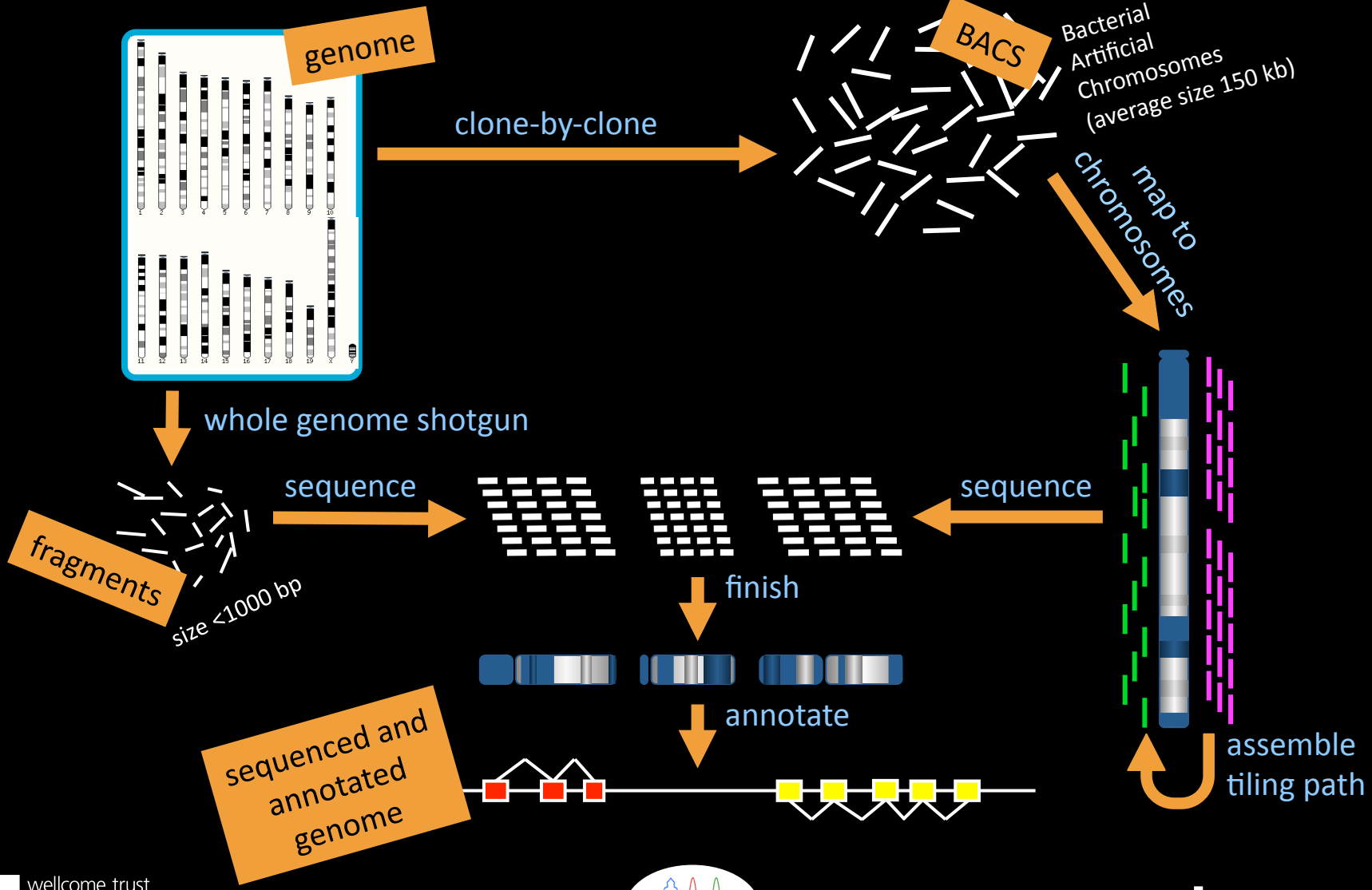
What's Coming Up

- background to origin of genomic sequence
- genes
- annotation
- browsers
- gene sets

Genome Sequencing - HGP



Genome Sequencing

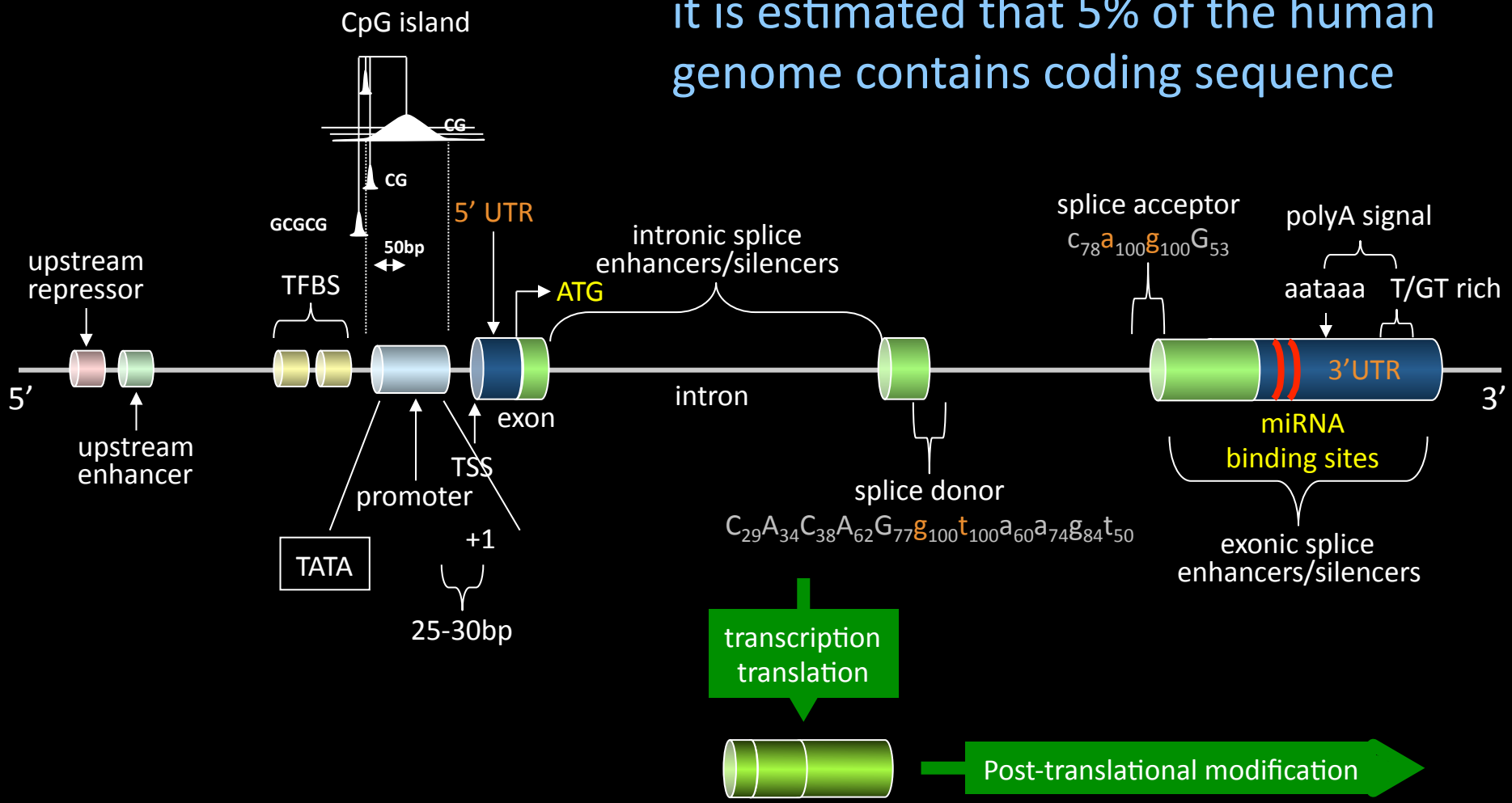


wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

What Is a Gene?

it is estimated that 5% of the human genome contains coding sequence



wellcome trust sanger institute

havana human and vertebrate analysis and annotation

ENCODE

Encyclopedia of DNA Elements

- experimentally and informaticly study all aspects of genes and genomes
  - gene expression
  - gene regulation
  - genome structure
  - etc.

Havana & Vega

- Havana

Human and vertebrate analysis and annotation

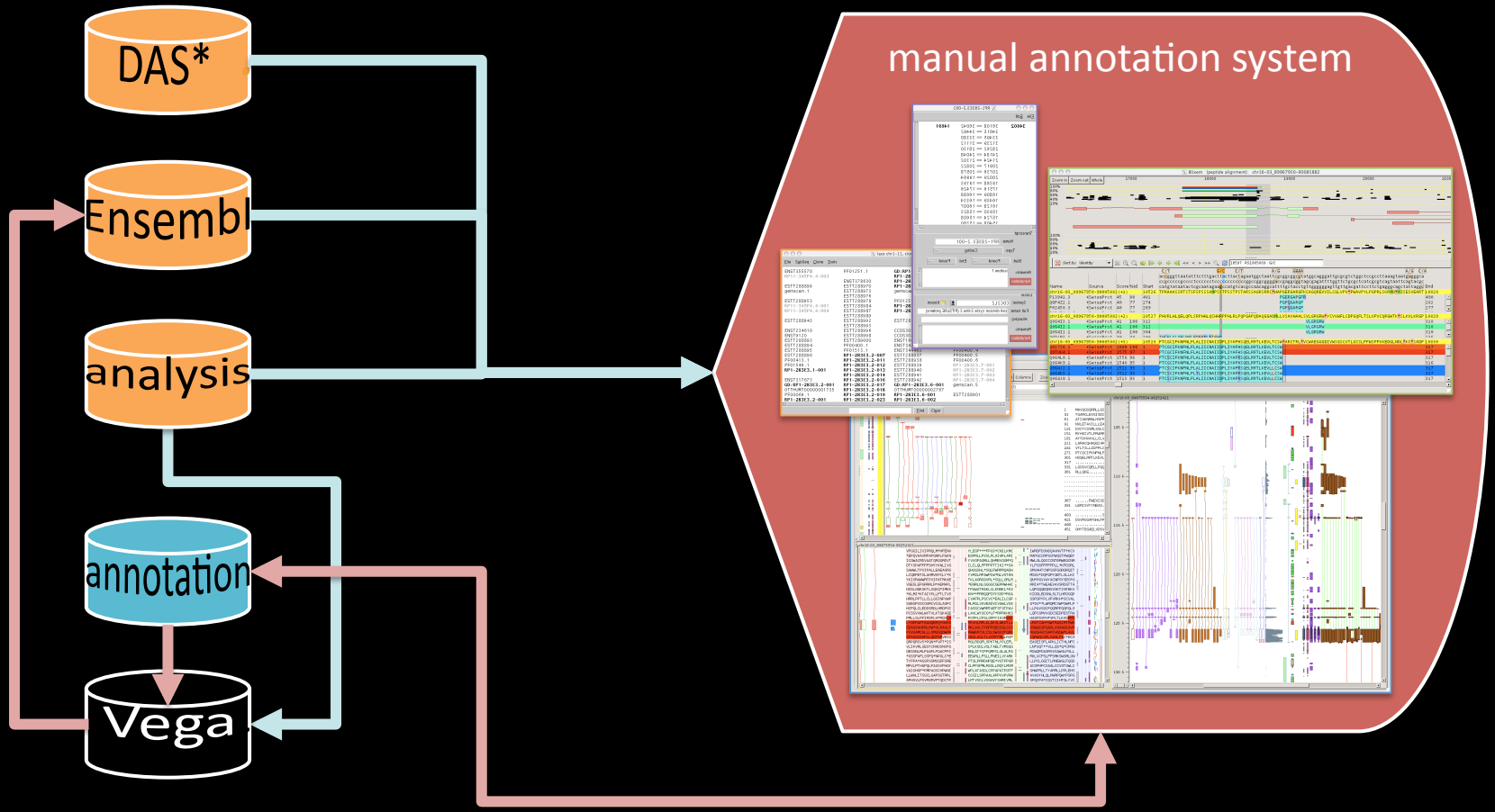
  - manual annotation of human, mouse and zebrafish whole chromosomes or genomes
  - human ENCODE, mouse EUCOMM annotation
  - annotation of specific regions: human MHC and LRC haplotypes, MHC and LRC in multiple species, mouse MUP cluster, ...

- Vega

Vertebrate Genome Annotation

  - Ensembl derived browser focusing on manual annotation

Havana Annotation - Tools

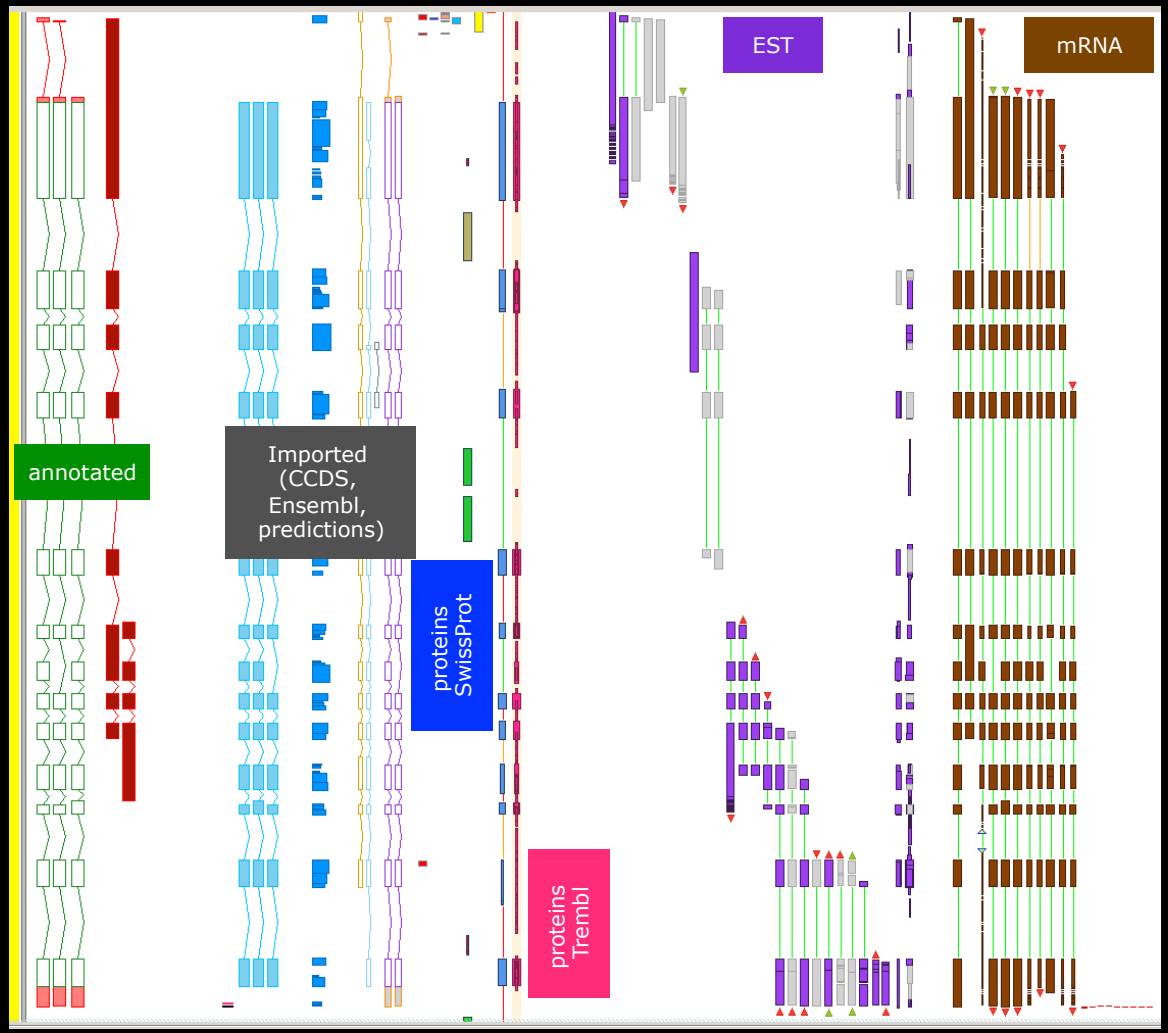


*) DAS = Distributed Annotation System = data hosted remotely by provider.
Shortly to be replaced by Track Hubs.

wellcome trust sanger institute

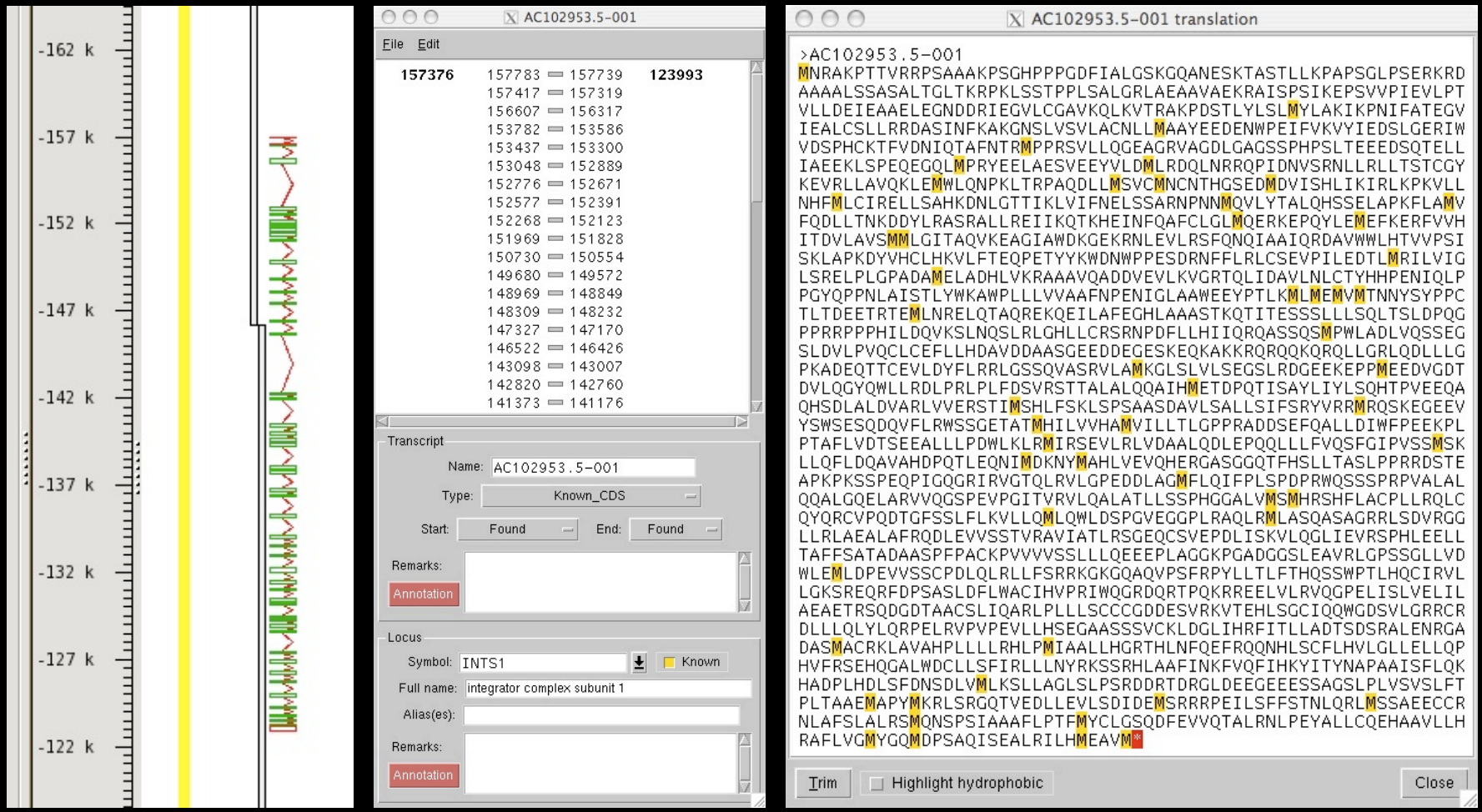
havana
human and vertebrate analysis and annotation

Havana Annotation - Tools



- transcript models based on evidence from matching ESTs, mRNAs, proteins
- taking into account data on TSS, polyA, conservation, repeats, protein domains

Havana Annotation - Tools



editing exon and CDS coordinates, attributes, names, biotype

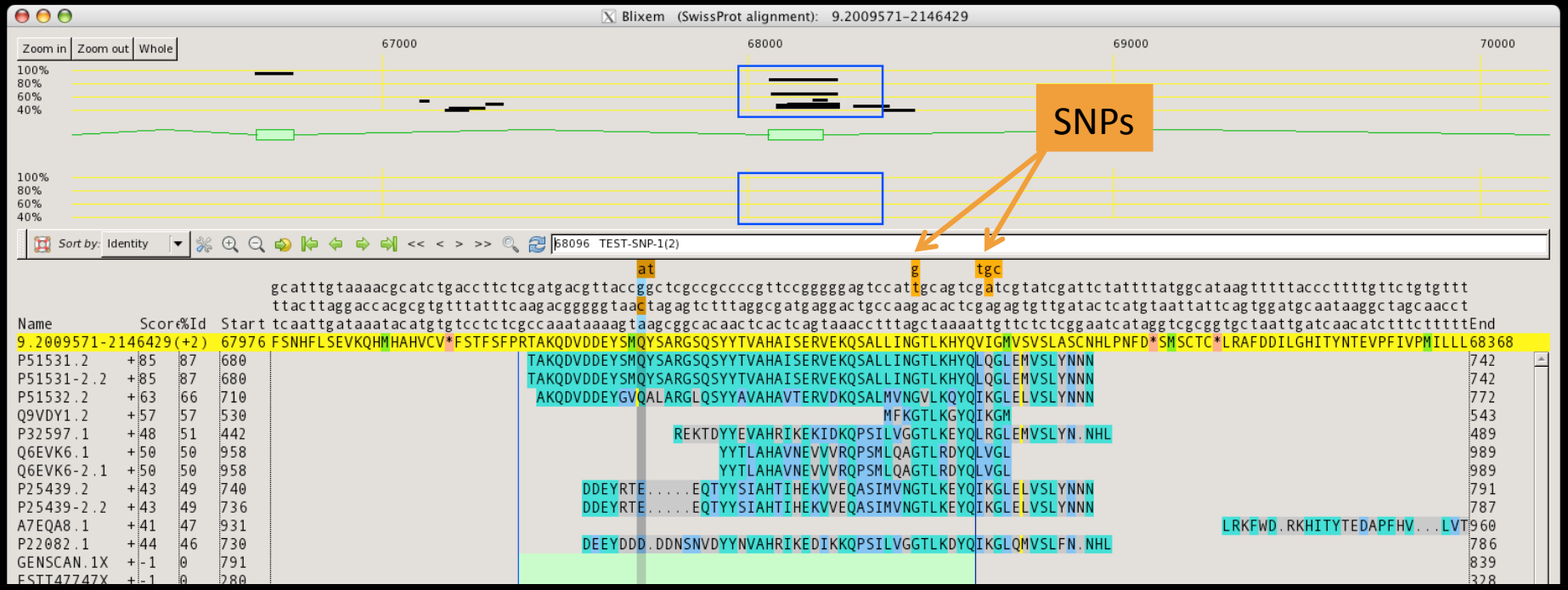
wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Havana Annotation - Tools

looking at alignments in detail
to determine exon boundaries

Havana Annotation - Tools



multiple alignment viewer shows SNPs

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Havana Annotation - Tools

many DAS sources available for display

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Havana Annotation - Tools

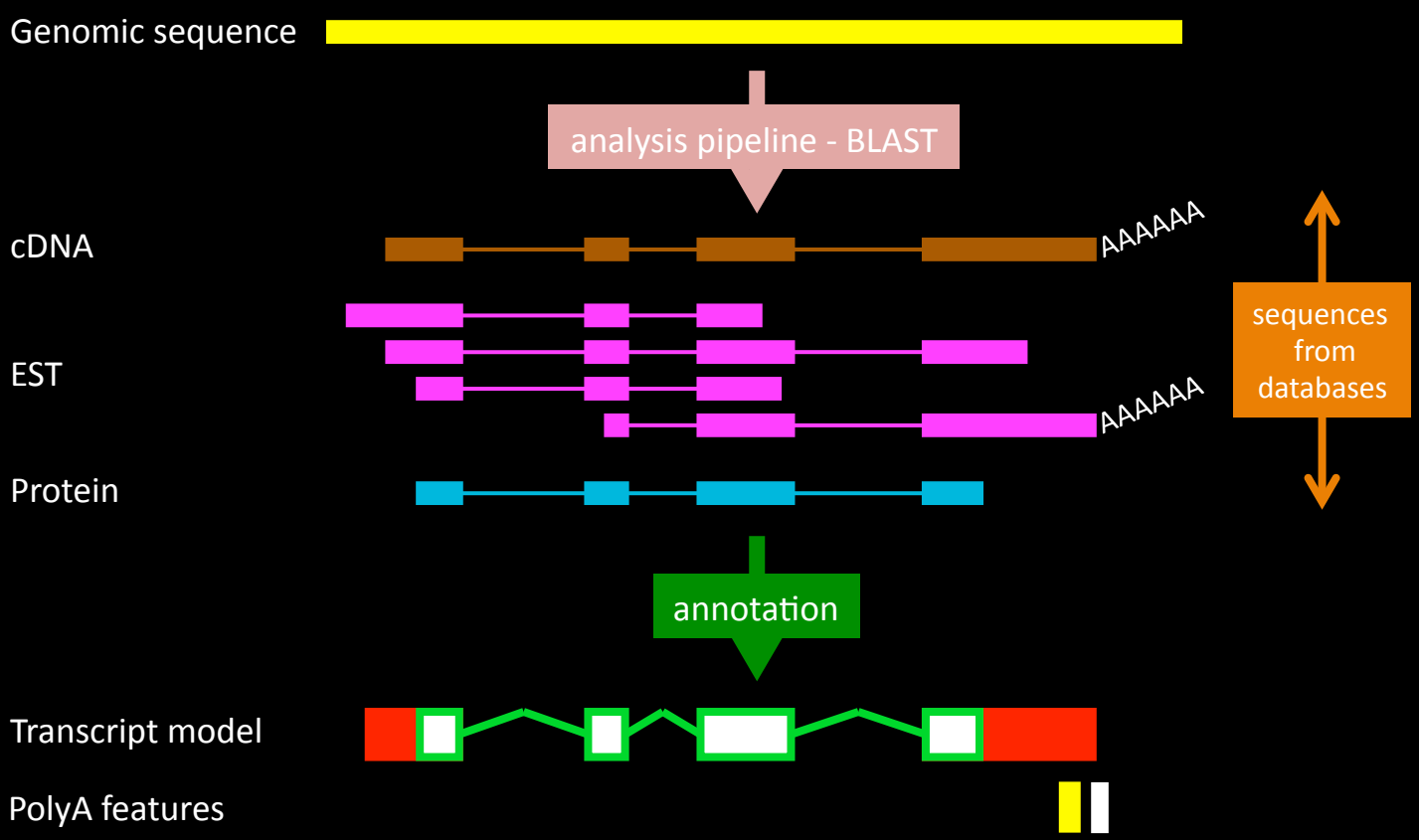
annotating two or more species or strains or haplotypes simultaneously

wellcome trust sanger institute

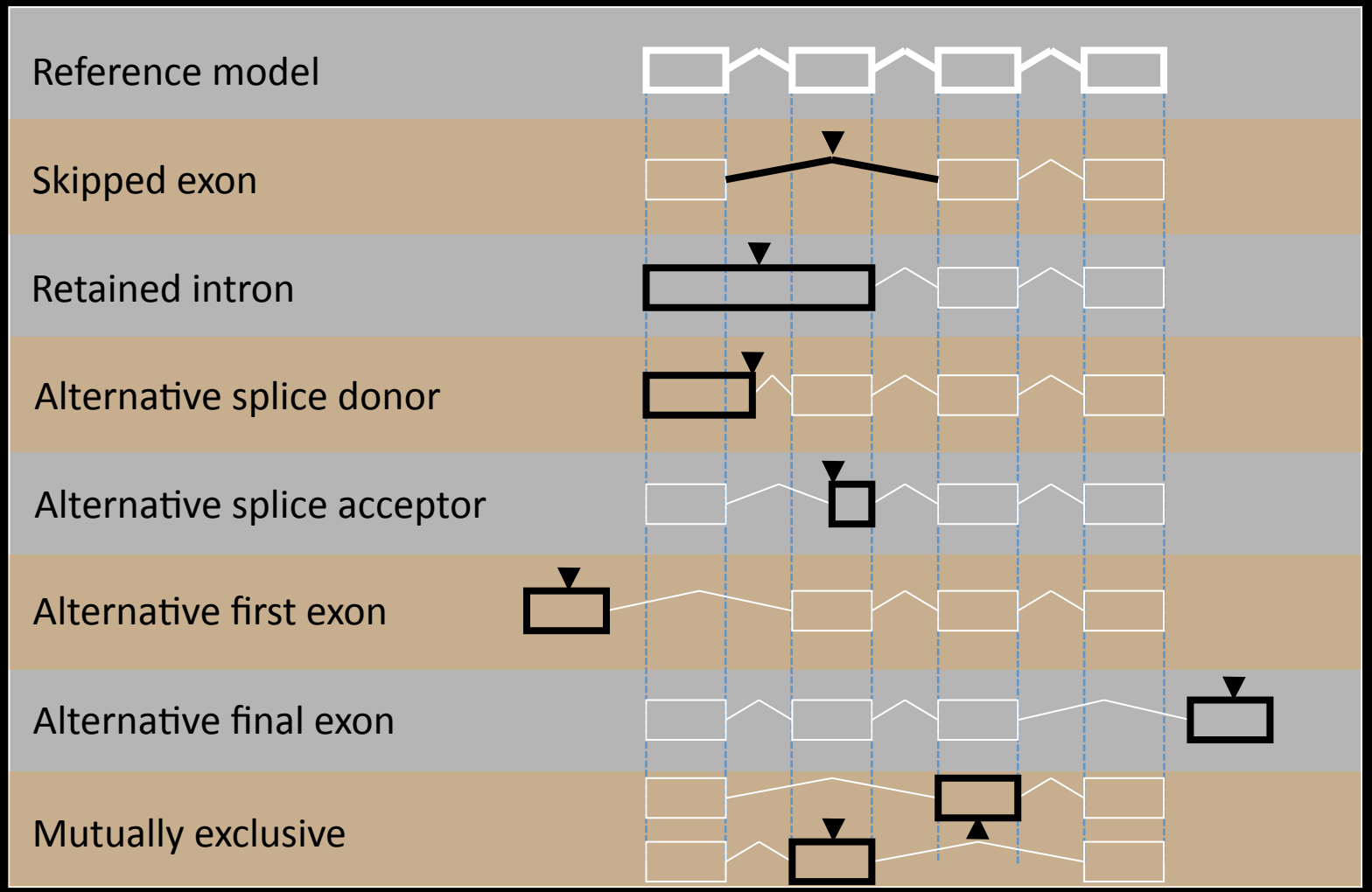
havana
human and vertebrate analysis and annotation

Manual Annotation

Annotation based on transcriptional evidence



Alternative Splicing

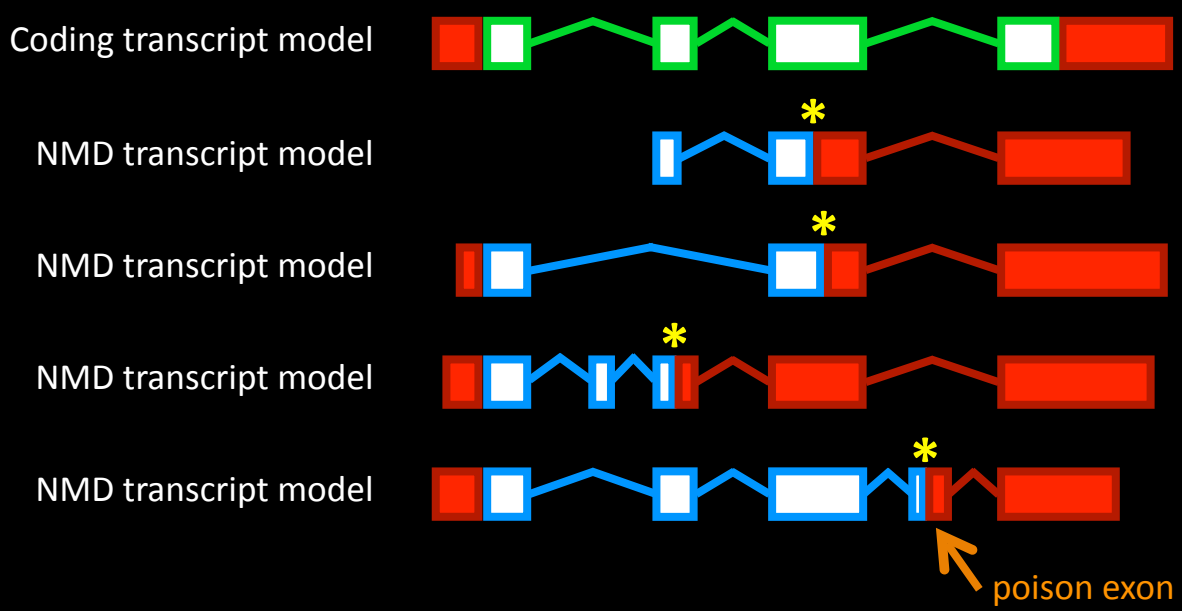


Manual Annotation - Biotypes

Protein Coding	Known CDS
	Novel CDS
	Putative CDS
	Nonsense Mediated Decay
Non-coding	lincRNA
	Antisense
	Sense Intronic
	Sense Overlapping
	3' Overlapping ncRNA
Transcript	Retained Intron
	Putative
Pseudogene	Processed
	Unprocessed
	Transcribed Processed
	Transcribed Unprocessed
	Unitary
	Polymorphic

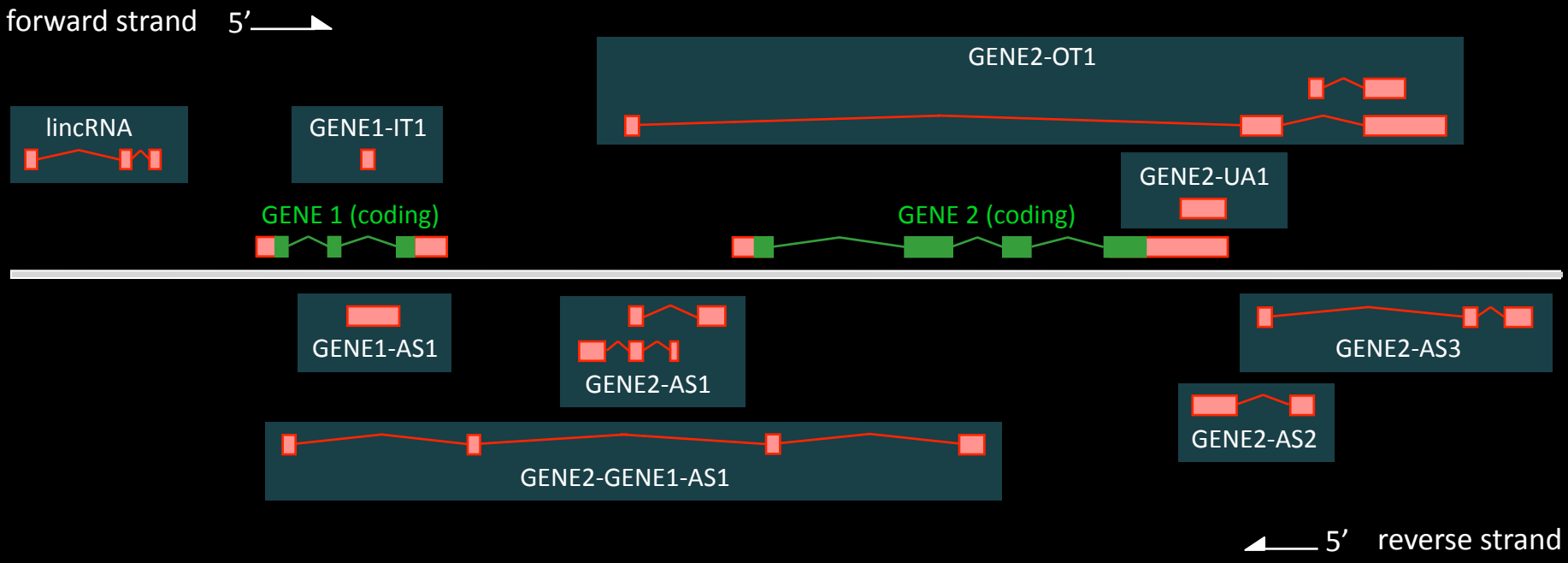
NMD (Nonsense Mediated Decay)

A mechanism to destroy potentially harmful splice variants

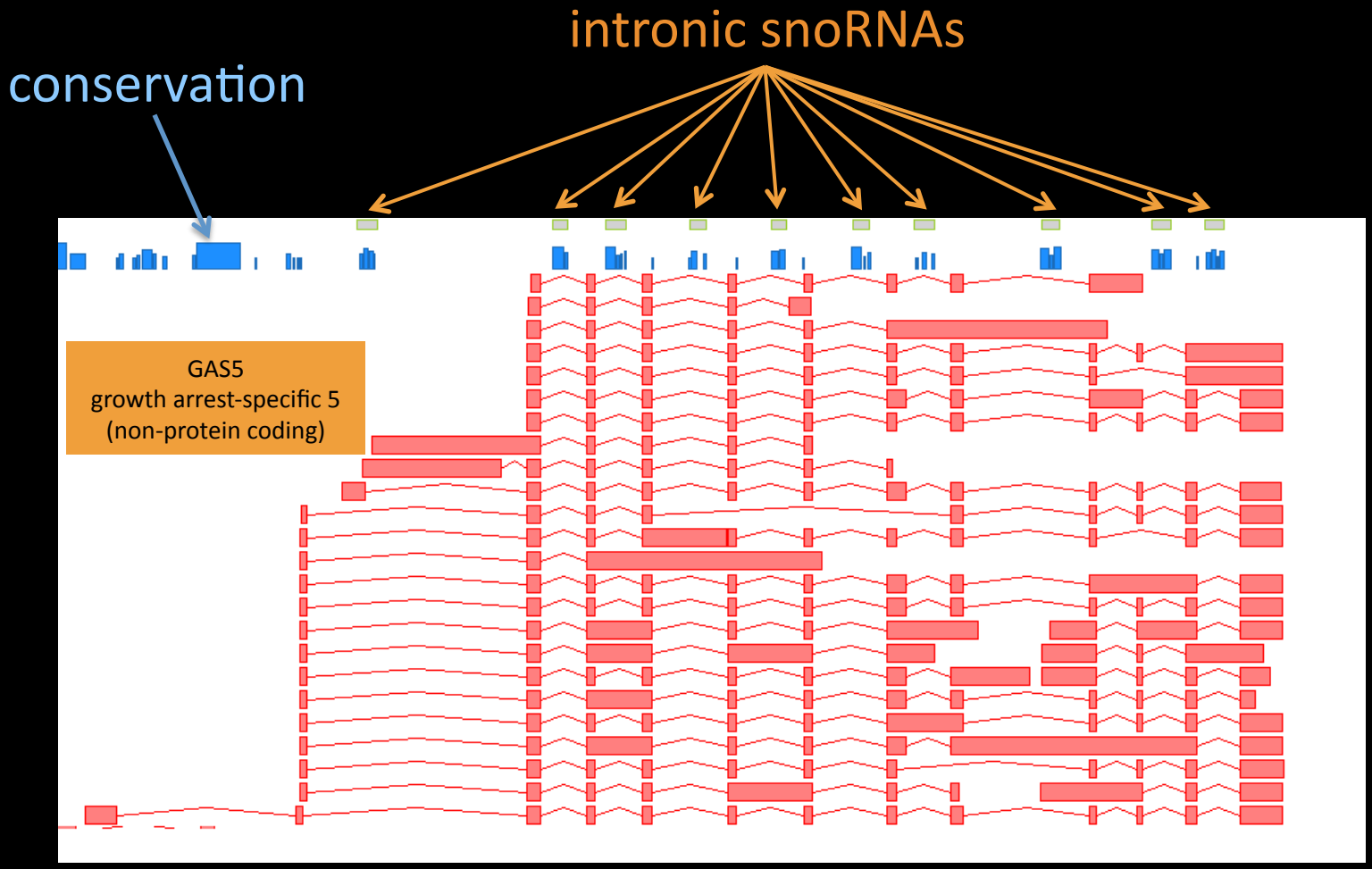


* premature termination codon >50bp upstream of a splice site

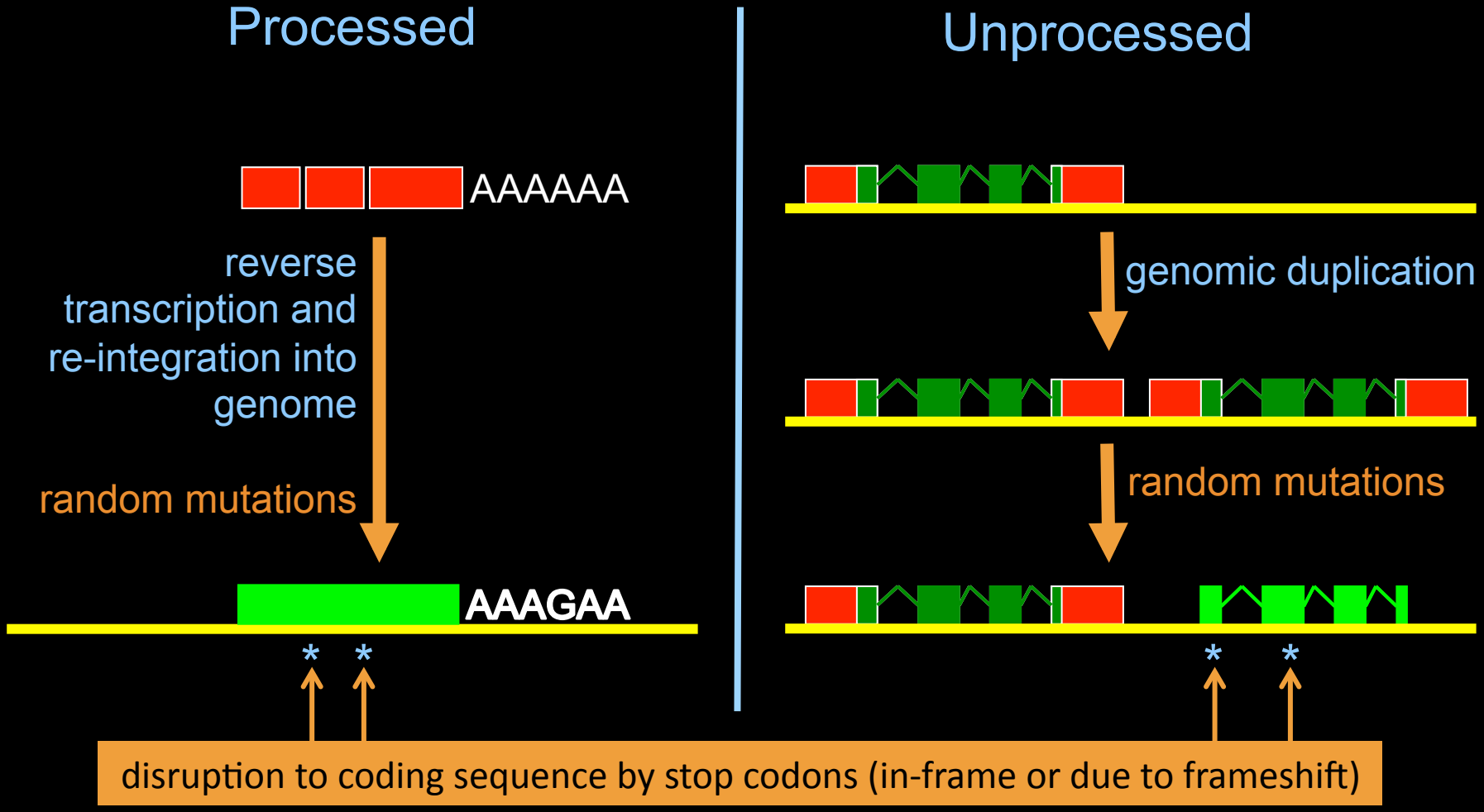
lncRNA



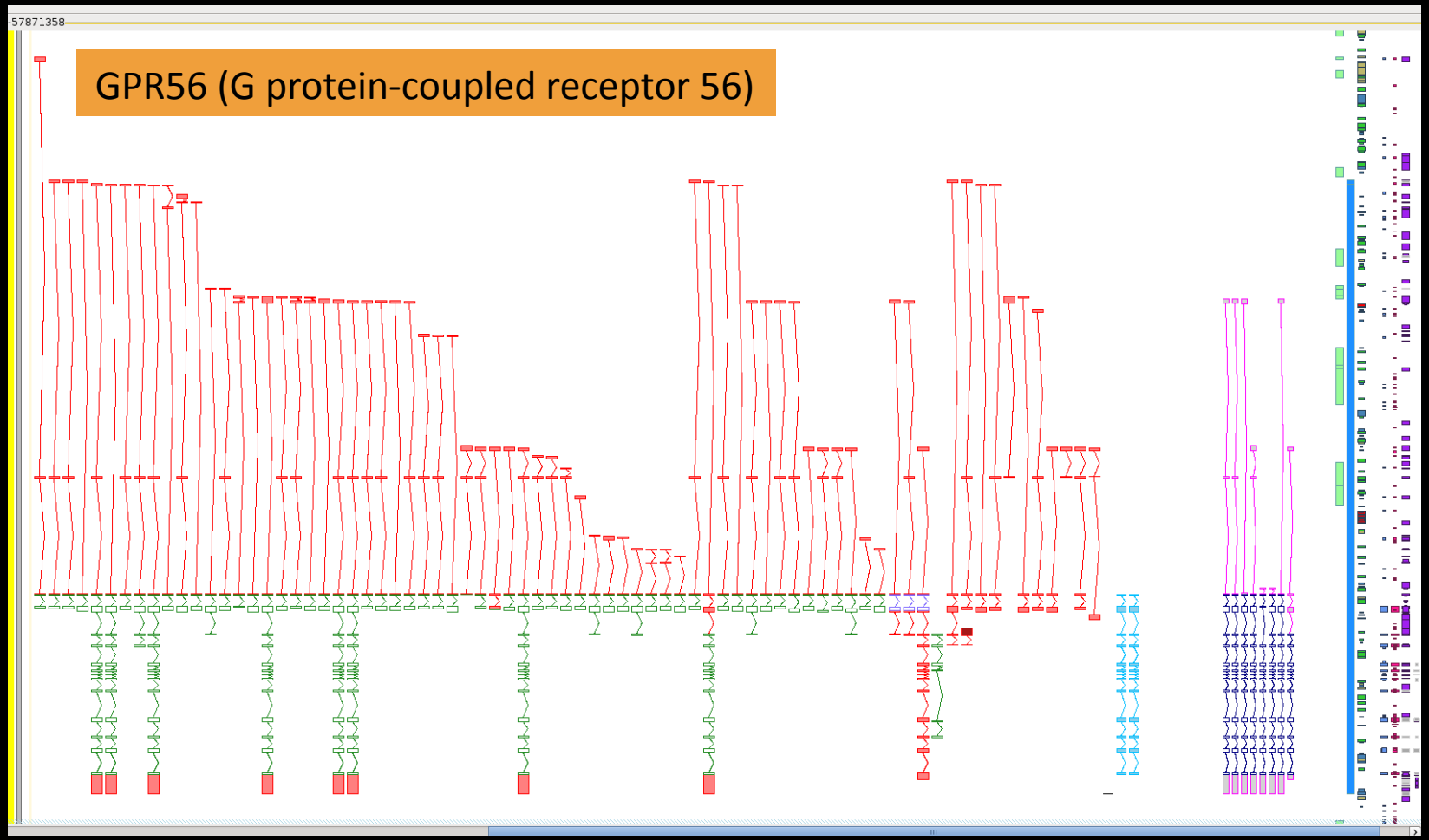
lincRNA - Example



Pseudogenes

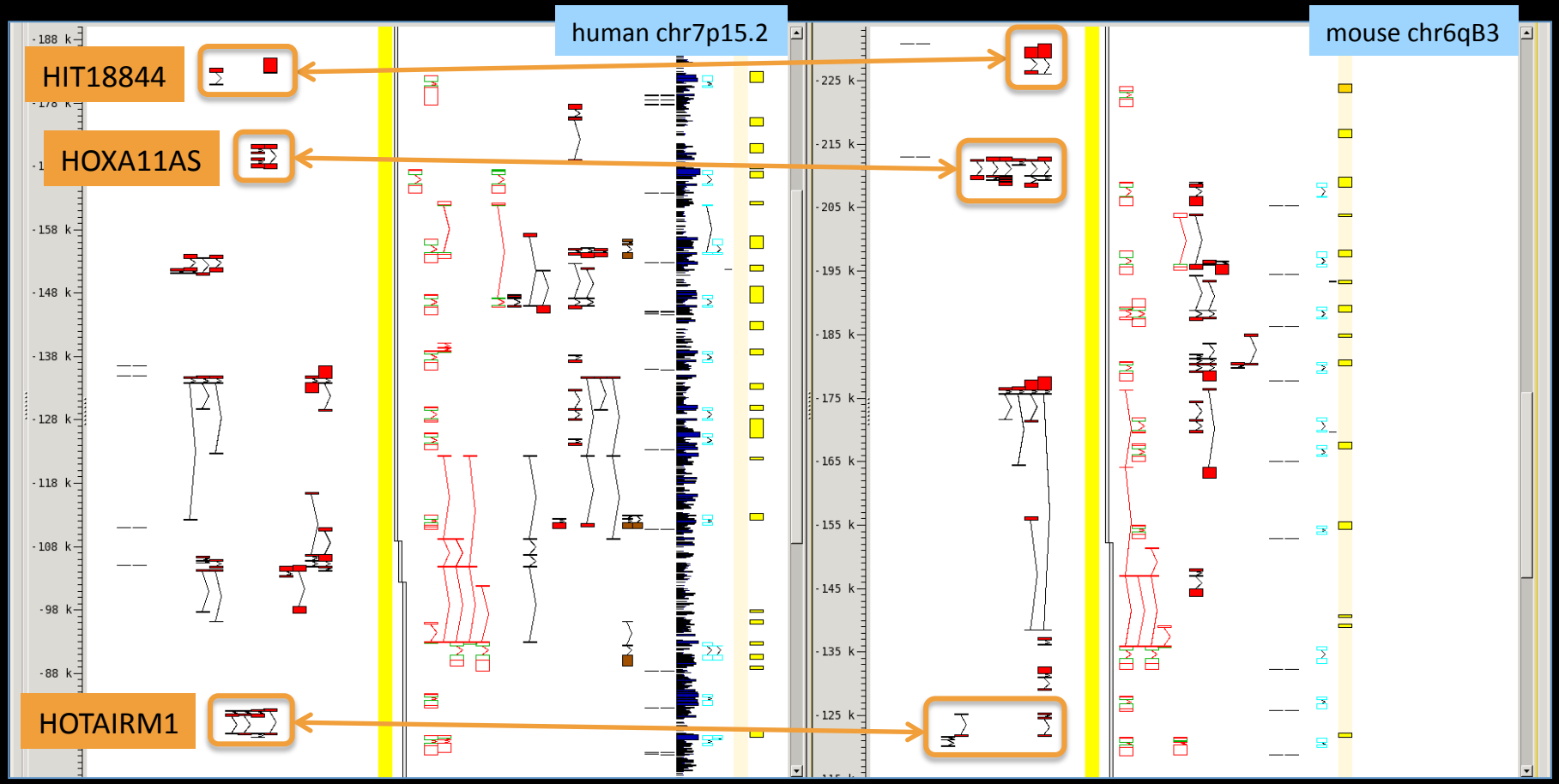


Manual Annotation - Examples



dozens of splice variants – all supported by EST or mRNA homology

Manual Annotation - Examples



long non-coding RNAs are conserved across species and regulate expression of HOX genes

Vega

vega.sanger.ac.uk

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Vega

- displays manual annotation
- displays mouse knock-out transcripts
- has a few non-reference genomic region assemblies that are not in Ensembl/UCSC/ RefSeq
- displays transcript attributes
  - overlapping locus, readthrough transcripts, non-ATG start, etc.

Vega

BLAST/BLAT | Help & Documentation

Login/Register

Vega search...

A repository for high-quality gene models produced by the manual annotation of vertebrate genomes.

wellcome trust sanger institute

havana human and vertebrate analysis and annotation

Search: All species for Go

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

Browse a genome


Human [3-06-2014] [Ensembl]


Mouse [3-06-2014] [Ensembl]


Zebrafish [21-01-2014] [Ensembl]


Pig [23-10-2013] [Ensembl]


Rat [21-08-2013] [Ensembl]

Browse a region


Tasmanian devil [23-10-2013] [Ensembl]


Chimpanzee [12-01-2012] [Ensembl]


Gorilla [30-03-2009] [Ensembl]


Wallaby [30-03-2009] [Ensembl]

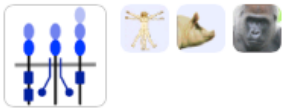

Dog [14-02-2005] [Ensembl]

Major histocompatibility complex (MHC) annotation

Non-reference regions
Human: 6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF
Mouse: NOD/MrkTac, NOD/ShiLtJ
Pig: Large White

Further information on our MHC annotation.

Leucocyte receptor complex (LRC) annotation



Non-reference regions:
Human: COX_1, COX_2, PGF_1, PGF_2, DM1A, DM1B, MC1A, MC1B.

Further information on our LRC annotation.

Our Data

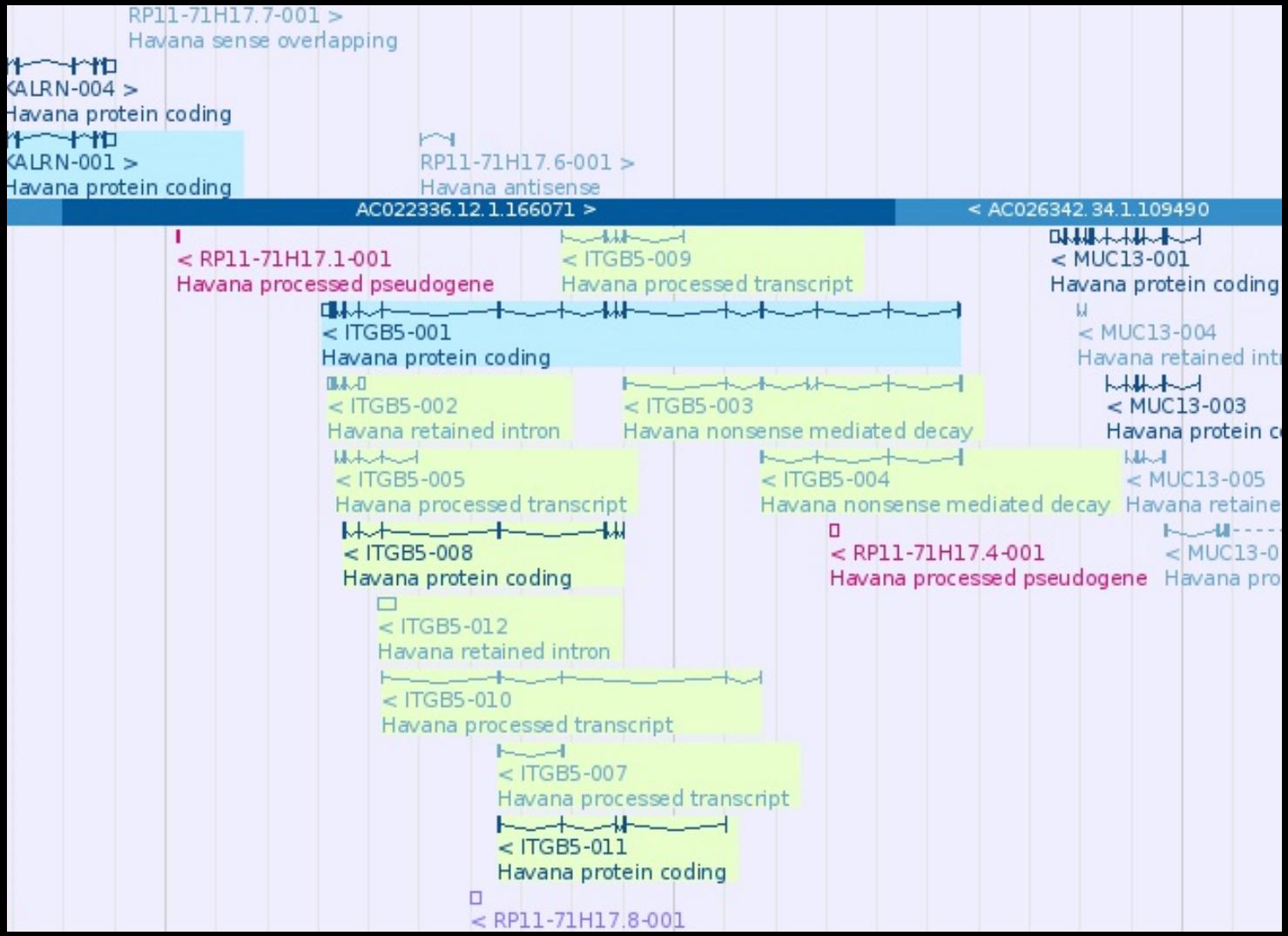
- High-quality manual annotation
- Human annotation incorporated into GENCODE
- Rapid incorporation of new annotation
- Gene sets and regions of particular interest:
  - Genes with mouse knockout and human LOF transcripts
  - MHC and LRC regions
  - *Idd* candidate regions of NOD mice
- Inter- and intra-species comparative genomics
- Cross-referenced to other databases
- Complements Ensembl
- Downloadable datasets

What's New in release 56

- **GRCh38 Human Annotation** (Human)
- **Mouse Annotation updated** (Mouse)

More news...

Vega



wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Vega BLAST/BLAT | Help & Documentation

Mouse (VEGA52) ▼ | Location: 11:70,432,635-70,440,828 | Gene: Arrb2 | Transcript: Arrb2-004

Gene-based displays

- Gene summary
- Splice variants (13)
- Transcript comparison
- Supporting evidence
- Sequence
- External references
- Expression
- Comparative Genomics
  - Genomic alignments (2)
  - Orthologues
  - Alt. alleles (1)
- External data
  - Personal annotation
- Other genome browsers
  - Ensembl

Configure this page

Add your data

Export data

Bookmark this page

Share this page

Gene: Arrb2 OTTMUSG00000006049

Description arrestin, beta 2

Location Chromosome 11: 70,432,635-70,440,828 forward strand.

Transcripts This gene has 13 transcripts (splice variants) Show transcript table

Gene summary

to MGI

to CCDS

Curated Locus Arrb2 (MGI Symbol)

CCDS This gene is a member of the Mouse CCDS set CCDS24946

Gene type Known protein coding [Definition]

Author This gene was annotated by Havana <vega@sanger.ac.uk>

Version & date Version 5, last modified on 10/03/2012 (Created on 01/10/2003)

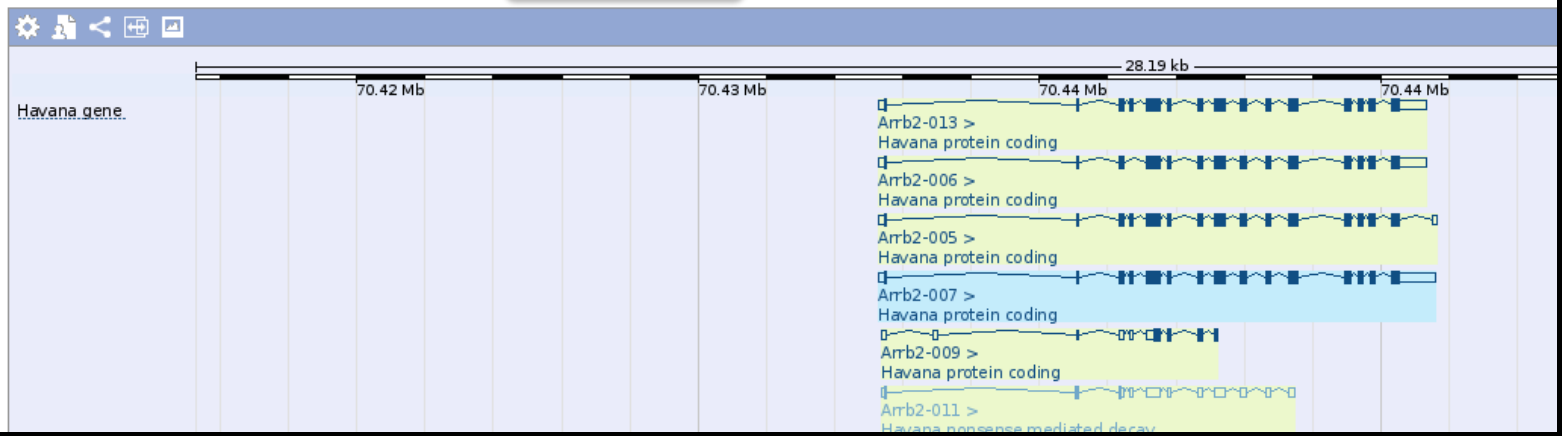
Alternative symbols RP23-42P20.7
bM42P20.7

Other assemblies This gene maps to to 70,432,635-70,440,828 in GRCm38 (Ensembl) coordinates.
Jump to this stable ID in Ensembl

to Ensembl

Curation Method Manual annotation

Alternative genes **Ensembl gene**: ENSMUSG00000060216



wellcome trust sanger institute

havana human and vertebrate analysis and annotation

Vega Evidence Viewer

shows sequences that support the annotation on exon-by-exon basis

Ensembl

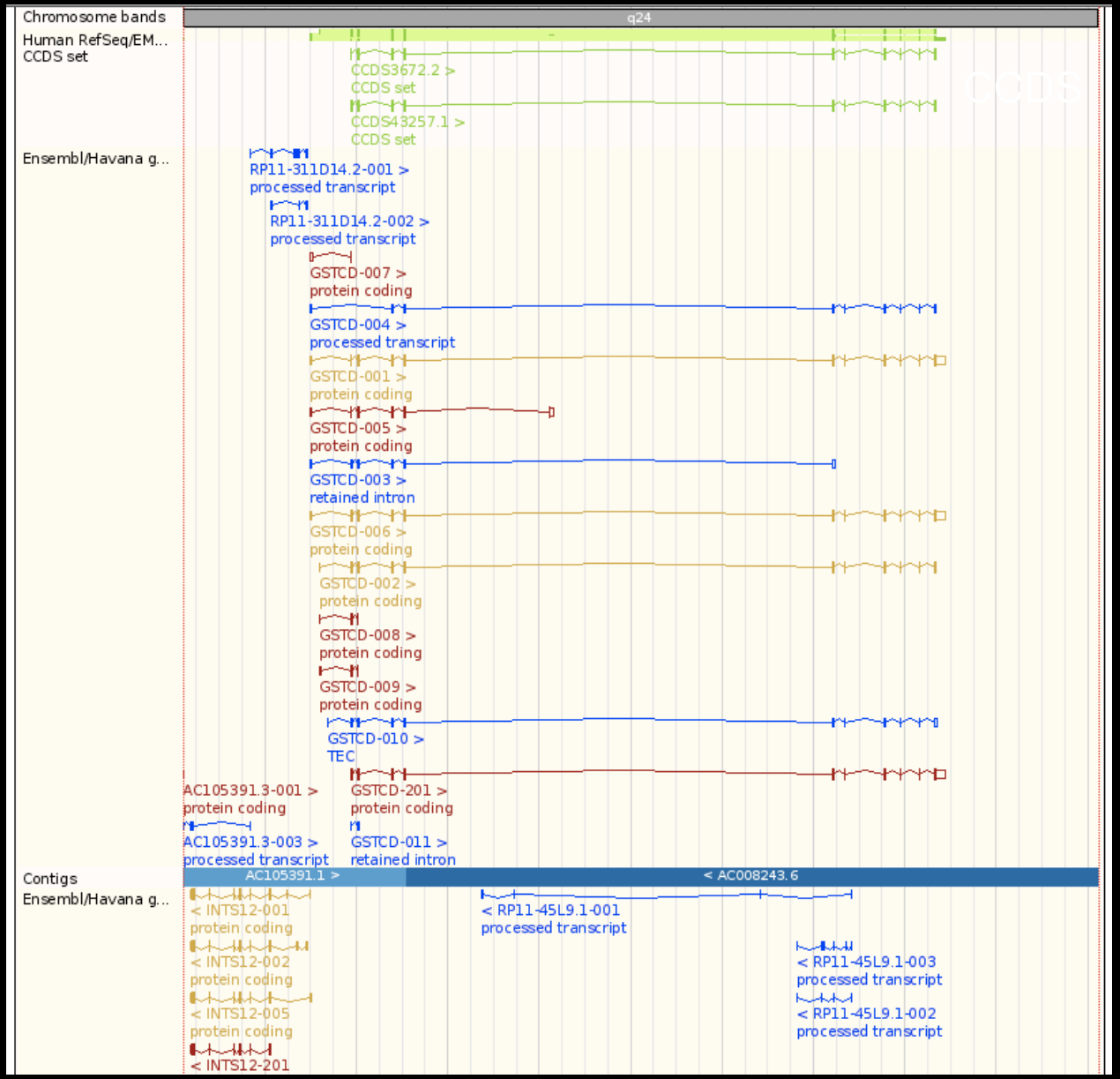
www.ensembl.org

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Ensembl

- automatic annotation
- many species
- imports most Vega annotation
  - but not all features (like attributes)



Ensembl

Gold
merged because identical between Ensembl and Havana

Red
coding
(-001 = Havana only, -201 = Ensembl only)

Blue
non-coding

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

MGI

www.informatics.jax.org

MGI

Mouse Genome Informatics

Community model organism database for the laboratory mouse

Authoritative resource for

- mouse gene, allele and strain nomenclature
- unified mouse genome feature catalogue
- developmental gene expression data
- functional annotation using Gene Ontology
- phenotype annotation using the Mammalian Phenotype Ontology
- mouse models of human disease

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

MGI

- search for genes according to diverse biological attributes
  - genome location
  - expression
  - function
  - phenotype
- find mouse models of human disease
- find mouse and ES cell resources
- compare gene structure and transcript annotation from NCBI, Vega and Ensembl
- link out to information in related resources
  - protein structure
  - gene family/orthology
  - pathways

Gene Sets

Gene Set Sources

www.ncbi.nlm.nih.gov/RefSeq/

genome.ucsc.edu/

www.ensembl.org

vega.sanger.ac.uk

www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi
(*CDSs only*)

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Gene Set Differences

Automatic annotation

- fast
- unfinished sequence or shotgun sequence
- consistent
- under/over-prediction

Manual annotation

- slow
- finished sequence
- flexible – can deal with inconsistencies
- can use publications, personal communications, etc.

NCBI-RefSeq Gene Set

- non-redundant gene set
- accessed via browsers or Entrez Gene
- accessions for genomic DNA, transcripts and proteins
- primarily protein-coding
- semi-curated
- gene centric in stead of genome centric

	automated	curated
genomic	NC_12345	
mRNA	XM_12345	NM_12345
ncRNA	XR_12345	NR_12345
protein	XP_12345	NP_12345

UCSC Gene Set

- non-redundant gene set
- automatic annotation based on BLAT alignments
- transcripts require Genbank accession plus one other supporting feature (e.g. Uniprot)
- includes RefSeq models (require no additional support)
- both protein-coding and non-coding
- data hub for ENCODE data, displays GENCODE gene set (for human)

genome.ucsc.edu

Ensembl Gene Set

- multiple biotypes (Known, Novel, ESTgenes, Pseudogenes)
- automatic annotation based on pair-wise alignment of genome with proteins, mRNAs and ESTs
- transcripts require supporting mRNA and protein evidence
- both protein-coding and non-coding transcripts
- includes merged data from Vega and CCDS and is called the GENCODE gene set

Vega Gene Set

- manually annotated
- based on direct pairwise alignment of genome with mRNA, EST and protein evidence (including cross-species)
- multiple biotypes, reflect confidence levels
- includes additional data sources as DAS tracks (e.g. CTSSs, RNAseq, polyAseq)

Comparing Gene Sets

	Human	Mouse	(June 2013)
RefSeq			
Coding genes	19,176	20,608	
Total transcripts	35,895	27,748	
UCSC			
Coding genes	21,127	21,181	
Total transcripts	80,992	59,121	
Ensembl e72 (Gencode)			
Known coding	20,774	23,139	
Total transcripts	194,846	93,480	
VEGA (unfinished)			
Coding genes	19,711	14,612	
Total transcripts	177,829	73,812	
CCDS			
Coding genes	18,606	19,945 (build 38)	
Total transcripts	27,752	23,027	

wellcome trust sanger institute

havana human and vertebrate analysis and annotation

Comparing Biotypes

- each gene set has defined biotypes
- biotypes can reflect both coding status and confidence levels

RefSeq protein-coding, pseudogene, miscRNA

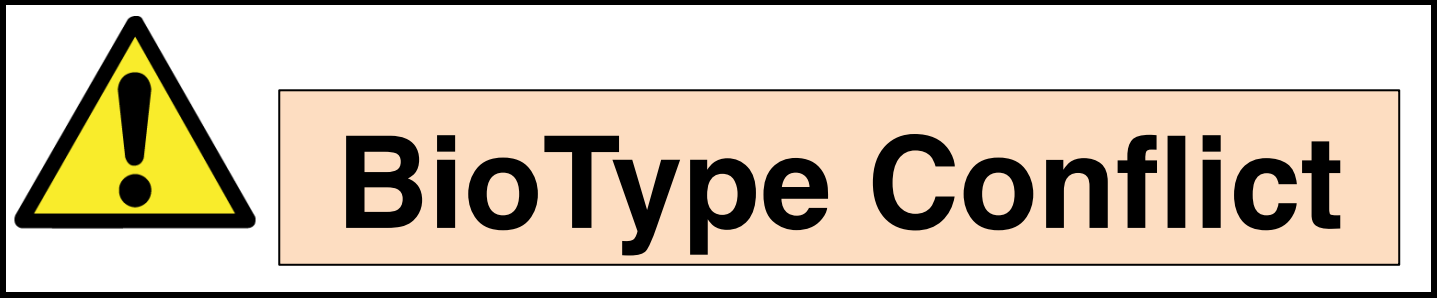
UCSC coding, non-coding, antisense, near-coding

Ensembl protein-coding (*Known/Novel*), processed transcript, pseudogene, polymorphic pseudogene, lincRNA

VEGA protein-coding (*Known/Novel/Putative*), NMD, processed transcript *(non-coding, antisense, lincRNA, retained intron, sense overlapping*), IG Gene, TR Gene, IG pseudogene, processed pseudogene, transcribed processed pseudogene, unprocessed pseudogene, transcribed unprocessed pseudogene, polymorphic pseudogene, unitary pseudogene, transcribed unitary pseudogene, polymorphic pseudogene

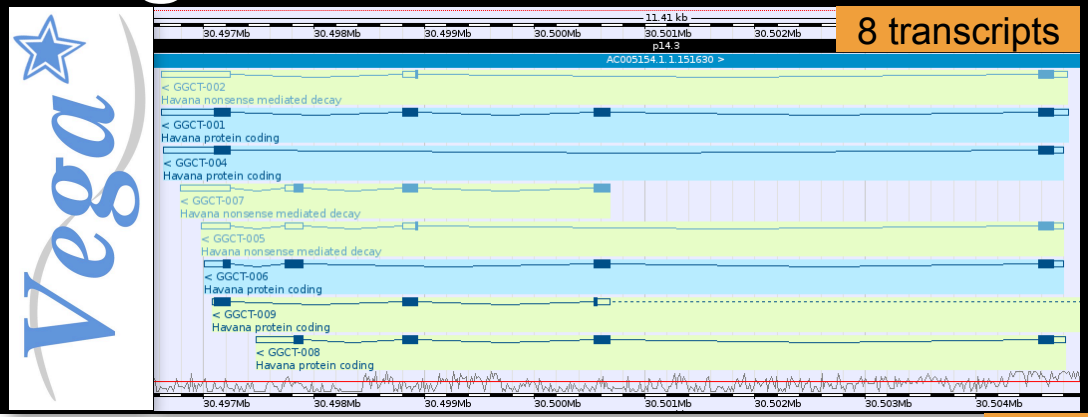
Biotype Conflicts

- there can be biotype conflicts between different gene sets
- the most common biotype conflict occurs with pseudogenes
- biotype conflicts are automatically identified by MGI
- loci with conflicting biotypes are raised internally between MGI, NCBI and Ensembl/Vega
- supporting evidence is reviewed and where possible discrepancies are resolved
- click on the BioType Conflict warning in MGI to see the biotypes by the different groups

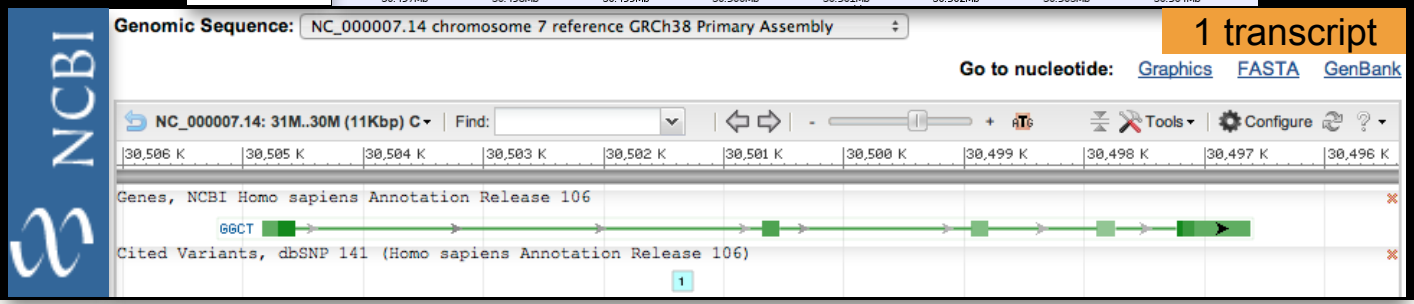


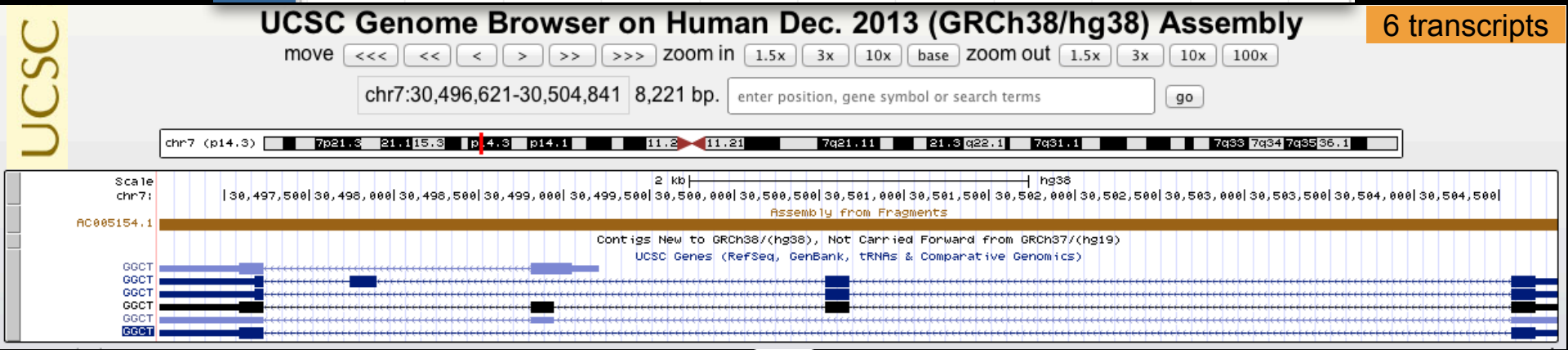
Comparing a Gene Between Browsers

Comparison done using current assembly (GRCh38)



GGCT (gamma-glutamyl cyclotransferase)





wellcome trust sanger institute

havana human and vertebrate analysis and annotation

CCDS

- Consensus CoDing Sequence project
- collaboration between UCSC, Ensembl, RefSeq and Havana groups
- produces reference CDSs: transcripts in CCDS database have a start-stop agreed by all member groups
- high quality
  - different member groups use different annotation methods and standards, so, agreement means high degree of confidence
- few alternative splice variants
- no UTRs
- slow to increase
  - increase is lately more in number of variants rather than loci

Which Gene Set Is Best?

- depends on what you require
- NCBI-RefSeq and MGI provide reference gene sets
- UCSC and Ensembl are larger gene sets, identifying more splice variants and pseudogenes
- VEGA provides the most comprehensive gene set with additional splicing variants, but has lower coverage and is slower to update
  - also, more fine-grained biotypes, plus attributes

Caveat Emptor

- Kim *et al.* (2014) Nature 509: A draft map of the human proteome
  - used RefSeq as gene set
  - extended data figure 4a in paper as example of ‘novel’ downstream ORF in bicistronic CHTF8 gene
  - actually has been in Vega since 2011!

GRC

Genome Reference Consortium

Goal:

- to correct regions in the genome that are currently misrepresented
- to close as many gaps as possible
- to produce alternative assemblies of structurally variant loci where necessary
- scientific community can report loci in need of review
- human, mouse and zebrafish

www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

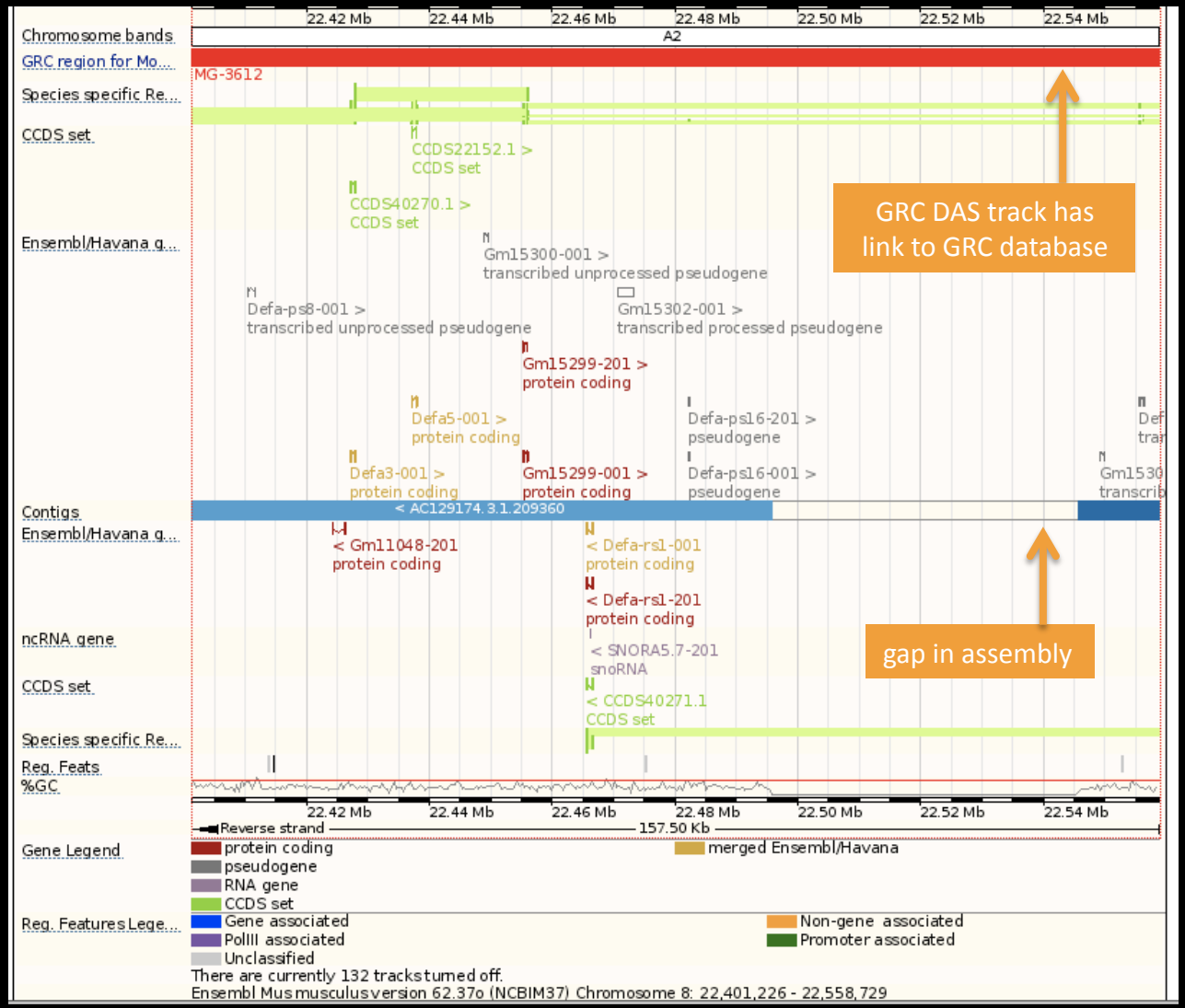
wellcome trust sanger institute

THE Genome CENTER AT WASHINGTON UNIVERSITY

EMBL-EBI

NCBI

GRC in Ensembl



wellcome trust sanger institute

havana human and vertebrate analysis and annotation

LRG

Locus·Reference·Genomic

- maintains representative transcripts of human loci on a stable (old) assembly
- for researchers (mostly from medical fields) that do not want to re-work their data with each new genome assembly

www.lrg-sequence.org

wellcome trust sanger institute

havana
human and vertebrate analysis and annotation

Conclusion

- manual annotation is essential for reference genomes
- automatic gene build essential for rapid updates
- there are four primary gene sets, plus CCDS
- there are benefits and drawbacks to each gene set
- data from all the gene sets is increasingly integrated and can be accessed through most browsers

Acknowledgements

SLIDES

Havana:

Jane Loveland

Charlie Steward

MGI:

Carol Bult

Janan Eppig

ANNOTATION

Havana annotators

RefSeq annotators & gene builders

UCSC annotators & gene builders

Ensembl gene builders

wellcome trust sanger institute

havana human and vertebrate analysis and annotation

MGI Workshops at IMGC

Monday 27 Oct: 14:30 – 16:00 in the Stotesbury Ballroom
Tuesday 28 Oct: 15:00 – 16:30 in the Stotesbury Ballroom
(overlaps with poster sessions)

MGI Workshops at MGI Symposium (at The Jackson Laboratory, following the IMGC)

Thursday 30 Oct: 13:30 – 15:00 in the Staff Conference Room, JAX
Thursday 30 Oct: 15:00 – 16:00 in the Staff Conference Room, JAX

Please sign up at the IMGC registration desk. Attendees are requested to bring laptops.

wellcome trust sanger institute

havana human and vertebrate analysis and annotation